A Probabilistic Graphical Model for Individualizing Prognosis in Chronic, Complex Diseases

Peter Schulam, MS,[1] Suchi Saria, PhD[1,2]

[1]Dept. of Computer Science, Johns Hopkins University, Baltimore, MD

[2]Dept. of Health Policy and Management, Johns Hopkins University, Baltimore, MD

Abstract

Making accurate prognoses in chronic, complex diseases is challenging due to the wide variation in expression across individuals. In many such diseases, the notion of subtypes—subpopulations that share similar symptoms and patterns of progression—have been proposed. We develop a probabilistic model that exploits the concept of subtypes to individualize prognoses of disease trajectories. These subtypes are learned automatically from data. On a new individual, our model incorporates static and time-varying markers to dynamically update predictions of subtype membership and provide individualized predictions of disease trajectory. We use our model to tackle the problem of predicting lung function trajectories in scleroderma, an autoimmune disease, and demonstrate improved predictive performance over existing approaches.

Introduction

In complex, chronic diseases such as autism, lupus, and Parkinson's disease, the way the disease manifests may vary greatly across individuals. For example, in scleroderma, the disease motivating this work, individuals may be affected across six organ systems—the lungs, heart, skin, gastrointestinal tract, kidneys, and vasculature—to varying extents [1]. For any single organ system, some individuals may show rapid decline throughout the course of their disease, while others may show early decline but stabilize later on. In such diseases, accurate tools for prognosis are especially important. For example, a therapy with harmful side effects may be warranted for an individual who is likely to decline rapidly. Alternatively, if an individual's decline is projected to be gradual, then a more mild drug combined with frequent follow-up may be adequate. In addition, clinical trials in such diseases can be improved by using prognoses as a way of recruiting individuals who are likely to have complications targeted by the therapy under investigation. One strategy for prognosis has been to use understanding of disease mechanism to identify biomarkers that are associated with specific progression patterns. However, such biomarkers may not always be available or may be insufficient to precisely predict course.

In this paper, we focus on the task of predicting disease activity for individual organs in complex diseases. Disease activity is typically measured using clinical or laboratory markers that quantify the health of the organ. For example, in scleroderma, an individual's lung disease is monitored using the percent of predicted forced vital capacity (PFVC), which is a continuous-valued measure of lung volume. Henceforth, we refer to these quantities as *markers* and refer to the hypothetical continuous function representing the value of the marker over time as the *disease trajectory*. Our task is therefore to predict the form of this function in order to forecast an individual's future disease activity.

The main challenge of making prognoses in complex diseases is due to the heterogeneity of disease expression across individuals. To cope with this heterogeneity, clinicians and medical scientists often describe the population in terms of subtypes—groups of individuals that manifest the disease similarly. Subtypes have been explored in a number of complex diseases including Parkinson's disease, heart disease, and autism [2–4]. By stratifying the heterogeneous population into more homogeneous subpopulations, subtype-specific prognostic models can be built to better predict progression. In practice, however, these subtypes are not observed and must be learned from data. Moreover, there may be other sources of variability across individuals within the same subtype. For example, among individuals with a subtype predicting active lung decline, a smoker may have consistently lower lung function than a non-smoker. Similarly, infections may temporarily depress an individual's lung function without affecting overall disease trajectory. Identifying regularities in these individual-specific variations can be used to further *individualize* predictions. For example, after observing that a smoker is always 10 PFVC lower than the subtype trajectory, we can adjust future subtype-specific predictions for the individual by shifting them down. A shift in the opposite direction could occur, for example, if an individual is especially athletic. Finally, modeling sources of short-term variability can allow explaining away that may improve subtype inference. For example, if the individual is experiencing transient decline due to infection, an approach that models this type of variability will consider two possible future progressions—one

in which they continue to rapidly decline, and another in which they are stable but experiencing a transient dip. Thus, predictions of the future trajectory will consider both alternatives instead of incorrectly converging to the more extreme progression pattern of rapid decline. We refer to these types of refinements as *individual-specific adjustments*.

Related work. In scleroderma, there is an active body of work related to predicting outcomes. Many researchers have investigated predictive relationships between measurements collected at baseline (e.g. demographics, clinical observations, and laboratory test results) and a single outcome or event such as death due to interstitial lung disease or time until onset of pulmonary hypertension (e.g. [5]). Others have extended beyond prediction of a single outcome or event, and have studied the association between baseline measurements and rates of decline of markers such as PFVC [6]. There have also been recent attempts to understand how marker measurements observed at follow-up visits may change the predictive relationship between baseline measurements and marker values at future follow-up visits [7]. None, however, incorporate the notion of unobserved subtypes or predict full trajectories. In addition, there are no existing models that dynamically update predictions over time.

More broadly, there is a large body of work on building predictive models for time series data. One natural approach is to construct a model that directly predicts the future value of the time series conditioned on the information in some finite window into the past. For example, order-p autoregressive (AR-p) processes model the distribution of the measurement at the next time step using the previous p measurements. AR-p models can be used to predict full trajectories by recursively applying the model over time [8]. A related, but more flexible approach is to construct a latent Markov process over some state space and model the observed measurements as a function of the current state. Examples of this type of model include hidden Markov models (HMM) and linear dynamical systems (LDS) [9, 10]. These approaches primarily capture local temporal structure and are therefore not as appropriate for modeling chronic disease trajectories where we believe that there is important long-term structure in the trajectory.

Another major thread of work in predicting time series data directly models the global shape of the time series. There are many ways that one can do this, but one popular way is to use Gaussian processes, which place flexible nonparametric prior distributions over functions [11]. By conditioning on previously observed values of the time series, the Gaussian process posterior places high probability on functions that are consistent with the history. Values at future time points of functions with high posterior probability can then be used as predictions (see e.g. [12]). When heterogeneity is expected in the population, mixtures of such models can be used (e.g. [13]). We build on these ideas.

Finally, there has been recent work on tailoring clinical predictions by learning a model that is specific to the individual test case at hand—similar to the ideas used in training a mixture of experts (see e.g. [14]). For example, lazy Bayesian rule learners [15] have been applied to predicting outcomes of community acquired pneumonia [16]. In this work, the authors search over subsets of the training data defined by a conjunctive clause that is consistent with the test case to learn shallow decision trees. Such models are expected to work well in training samples where there is heterogeneity with respect to the relationship between observed covariates and outcome. Similar methods have been applied to learning dynamic Bayesian networks [17]. Although these contributions address the personalization issue, they assume that heterogeneity among individuals can be captured well using observed covariates alone. In diseases such as scleroderma, the sources of variability are often poorly understood and are therefore not fully explained by observed covariates. Thus, we develop an alternate approach to personalization.

Contributions. We describe a probabilistic graphical model that leverages the idea of subtypes and individual-specific adjustment to subtype-specific predictions in order to improve the accuracy of disease trajectory predictions. Furthermore, we present an algorithm that learns subtypes and the associated average disease trajectories automatically from data and discuss the inference procedure within our graphical model that is used to dynamically update an individual's predictions as more marker values are observed over time. Finally, we show how we develop a prognostic tool for Scleroderma. We use data from the Johns Hopkins Scleroderma Center, one of the largest national repositories. By modeling the different types of variability across individuals, we show that our approach produces more accurate predictions of future trajectory compared to models that do not incorporate these types of variability.

Methods: Individualized Trajectory Prediction

Our model makes predictions about future values of clinical or laboratory markers using information available at baseline and any previously observed values of the marker. These pieces of observed information are used to make inferences regarding unobserved random variables that indicate to which subtype an individual belongs and that specify the parameters of an individual-specific adjustment to the subtype-specific model.

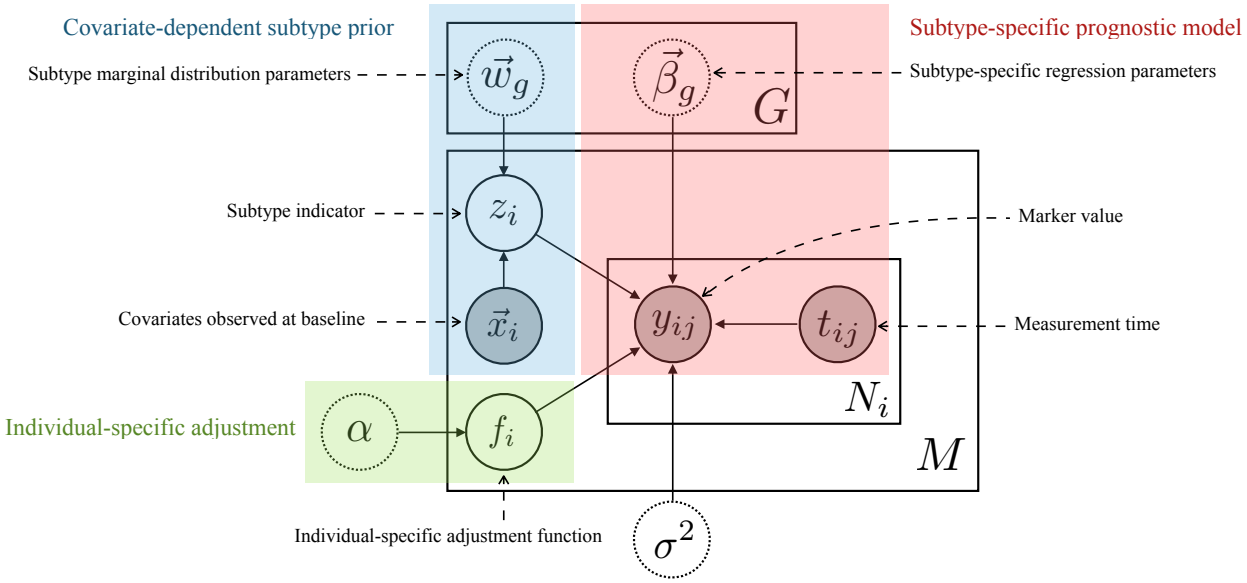

Figure 1: Graphical formulation of the statistical dependencies in the prognostic model. Components of the model are color-coded. Model parameters are enclosed in dashed circles, and random variables in solid circles. Observed random variables are shaded.

We perform these inferences within the framework of probabilistic graphical models. In order to leverage this framework, we must define a joint probability distribution over the marker values and the unobserved random variables. Figure 1 displays the statistical dependencies between all variables in the model, where M denotes the number of individuals being modeled and N_i is the number of markers observed for individual i. We first describe the probability model for a single individual. For individual i, we use $\vec{x}_i \in \mathbb{R}^q$ to denote a vector of baseline covariates, $\vec{y}_i \in \mathbb{R}^{N_i}$ to denote an ordered sequence of observed marker values, and $\vec{t}_i \in \mathbb{R}^{N_i}$ to denote the corresponding measurement times.

Key to our formulation is the idea of an individual's subtype, which we model by introducing an unobserved discrete random variable $z_i \in \{1, \ldots, G\}$, where G denotes the number of subtypes. The subtype variable encodes which of a finite number of discrete disease mechanisms is driving an individual's progression. By explicitly modeling an individual's subtype, we can rely on prognostic models specific to that mechanism to make more tailored predictions. However, because such mechanisms are often poorly understood and difficult to differentiate between in complex diseases, we cannot rely solely on observable characteristics of the individuals to group them into subtypes. We therefore treat the subtype random variable as latent, or unobserved, in the probabilistic model.

In some cases, however, observable characteristics may include some information regarding an individual's mechanism. We therefore assume that the marginal distribution of the subtype z_i depends on the baseline covariates $\vec{x}_i$. In scleroderma, for example, individuals who test positive for Scl-70 antibodies are more likely to develop interstitial lung disease. In our model, an individual who is Scl-70 positive may therefore have a prior distribution over subtypes that places more probability on those that exhibit severe decline in lung function. To formally capture this intuition, we model z_i as a multinomial random variable with distribution

$$z_i \mid \vec{x}_i \sim \text{Mult}\left(\pi_{1:G}\left(\vec{x}_i\right)\right). \tag{1}$$

The covariate-dependent probabilities $\pi_{1:G}\left(\vec{x}_i\right)$ are parameterized using a multinomial logistic regression. For each subtype g the probability $\pi_g\left(\vec{x}_i\right)$ is

$$\pi_g\left(\vec{x}_i\right) = \frac{1}{Z\left(\vec{x}_i\right)} e^{\vec{w}_g^\top \vec{x}_i} \quad \text{, where} \quad Z\left(\vec{x}_i\right) = \sum_{g'=1}^{G} e^{\vec{w}_{g'}^\top \vec{x}_i}. \tag{2}$$

We denote the full set of weights using $\vec{w}_{1:G}$, where the weights of the first or last class are typically constrained to be $\vec{0}$ to ensure model identifiability.

We associate each subtype with its own prognostic model. Instead of estimating a single model of progression for the entire heterogeneous population, the organization of individuals into groups with similar mechanism creates more homogeneous subpopulations. Within each subpopulation, a predictive model is more likely to be able to capture important characteristics of the typical disease progression. For example, in scleroderma there are two classically recognized patterns of skin disease. In the limited cutaneous disease group, we expect individuals to have a mild, constant degree of severity over time. In the diffuse cutaneous group, however, individuals are expected to experience a rapid increase in severity within the first few years, which is then followed by a gradual recovery. Our model discovers distinct patterns of progression such as these automatically, and learns a unique prognostic model for each that characterizes the severity over time.

We represent each subtype's prognostic model non-parametrically using B-splines. We assume that the number and location of the spline knots and the degree of the polynomial pieces are chosen prior to learning. These hyperparameters determine a feature matrix $\Phi(\vec{t}\,)$ which contains in each column one of the P B-spline basis functions evaluated at times $\vec{t}$, written as: $\Phi(\vec{t}\,) = \left[\Phi_1(\vec{t}\,), \ldots, \Phi_P(\vec{t}\,)\right]$.

Subtype-specific prognostic models are then parameterized by a vector of coefficients $\vec{\beta}_g \in \mathbb{R}^P$ for each subtype $g \in \{1, \ldots, G\}$. Given the vector of coefficients for subtype g, we can compute the prognostic model's marker value predictions at a collection of times $\vec{t}$ by taking the dot product of each row of the feature matrix with the coefficients, which we write as: $\hat{\vec{y}}_s = \Phi\left(\vec{t}\,\right)\vec{\beta}_g$.

Despite the benefits realized by using the notion of a subtype as a way to capture the variation in disease mechanism across a heterogeneous population, there are simplifications of the disease process that must be made so that the the subtype-specific models will generalize to unseen individuals. For example, we do not expect that every individual with the same subtype g will have exactly the same marker value after one year of follow-up. In reality, measurements are noisy and individual differences can cause deviations within subtypes despite shared mechanism. To some extent, these concerns may be addressed using noise random variables that are independent of one another and of all other variables in the model. In many cases, however, there is structure in these deviations that can be exploited to further improve prognostic accuracy. For example, an individual's marker values may be consistently 10 units higher than what is predicted by the subtype-specific model. This can occur, for example, if an individual is athletic and has stronger lungs than the average individual within that subtype. Other types of structure may be more transient in nature. An infection, for example, may temporarily inhibit lung function causing a quick dip in the observed forced vital capacity over the course of a year. If this behavior differs from what might be estimated using the last few years of observations, we may expect that the dip will correct itself in the next few months.

To account for and exploit the possible structure in these deviations, we introduce an additional unobserved random variable f_i, which is a function that specifies an additive offset from the subtype-specific model's prediction at a given time. We refer to this random function as the *individual-specific adjustment*. Conditioned on an individual's subtype and adjustment, we can express the predicted marker values at a vector of times $\vec{t}$ as: $\hat{\vec{y}}_{s+a} = \Phi\left(\vec{t}\,\right)\vec{\beta}_g + f\left(\vec{t}\,\right)$.

We model the adjustment f_i nonparametrically using a Gaussian process prior [11] with zero-valued mean function and covariance function $K_\alpha(\cdot, \cdot)$, where α denotes the set of hyperparameters used to specify the covariance function kernel. The prior over f_i is marginally independent of all other variables in the model as we assume that patient-specific adjustments cannot be predicted using baseline information. We use the covariance kernel to capture two types of individual-specific adjustments. The first are *long-term* adjustments, which are individual-specific constant offsets from the subtype. We model this using a constant covariance function, which is equivalent to including individual-specific intercepts. The second are *short-term* or *transient* adjustments, which we model using the Ornstein-Uhlenbeck kernel [11]. The Ornstein-Uhlenbeck process is mean-reverting, and so captures the intuition that transient deviations should return to the long-term mean. We combine these additively to form the following kernel with hyperparameters $\alpha = \{\nu_b, a^2, \ell\}$:

$$K_\alpha(t_1, t_2) = \nu_b + a^2 \exp\left\{\frac{|t_1 - t_2|}{\ell}\right\}. \tag{3}$$

In the expression above, the hyperparameter ν_b controls the amount by which an individual's trajectory can vertically shift from the subtype's, a controls the amplitude of observed transient deviations, and ℓ controls the duration over which transient deviations occur.

Finally, we model an individual's marker values $\vec{y}_i$ as random variables that depend on both the subtype membership z_i

and the individual-specific adjustments f_i. We assume that they are conditionally independent and normally distributed with variance σ^2:

$$\vec{y}_i \mid z_i, f_i, \vec{x}_i \sim \mathcal{N}\left(\Phi\left(\vec{t}_i\right)\vec{\beta}_{z_i} + f_i\left(\vec{t}_i\right), \sigma^2 \boldsymbol{I}\right). \tag{4}$$

Let $\Theta = \{\vec{w}_{1:G}, \vec{\beta}_{1:G}, \alpha, \sigma^2\}$ denote the model parameters introduced above. Taken all together, we obtain the following joint probability over both observed and unobserved random variables for a collection of M individuals:

$$P\left(\vec{y}_{1:M}, z_{1:M}, f_{1:M} \mid \vec{x}_{1:M}, \Theta\right) = \prod_{i=1}^{M} P\left(z_i \mid \vec{x}_i, \vec{w}_{1:G}\right) P\left(f_i \mid \alpha\right) P\left(\vec{y}_i \mid z_i, f_i, \vec{x}_i, \vec{\beta}_{1:G}, \sigma^2\right). \tag{5}$$

We can simplify this expression further by taking advantage of the fact that the conditional probability $P\left(\vec{y}_i, f_i \mid z_i, \Theta\right)$ can be written as the prior and likelihood of a Gaussian process regression model

$$P\left(\vec{y}_i, f_i \mid z_i, \Theta^{(t)}\right) = P\left(f_i \mid \alpha\right) \prod_{j=1}^{N_i} P\left(y_{ij} \mid \Phi(t_{ij})\vec{\beta}_{z_i}, f_i(t_{ij}), \sigma^2\right) \tag{6}$$

$$= \mathrm{GP}\left(f_i \mid 0, K_\alpha(\cdot,\cdot)\right) \prod_{j=1}^{N_i} \mathcal{N}\left(y_{ij} \mid \Phi(t_{ij})\vec{\beta}_{z_i} + f_i(t_{ij}), \sigma^2\right). \tag{7}$$

Standard results allow us to easily integrate out f_i to obtain the conditional density of $\vec{y}_i$ given z_i. The integration admits a closed form expression

$$P\left(\vec{y}_i \mid z_i, \Theta\right) = \mathcal{N}\left(\vec{y}_i \mid \Phi\left(\vec{t}_i\right)\vec{\beta}_{z_i}, K_\alpha\left(\vec{t}_i, \vec{t}_i\right) + \sigma^2 \boldsymbol{I}\right), \tag{8}$$

which is an N_i-dimensional multivariate normal with mean values specified by the subtype-specific model and the covariance matrix specified by the sum of the kernel of the individual-specific adjustment and an independent noise term. Furthermore, after integrating out f_i, the observed-data likelihood for each individual reduces to a mixture of multivariate normals differing only with respect to the mean (notice that the covariance of the multivariate normal above does not depend on z_i). The likelihood for M individuals is therefore:

$$P\left(\vec{y}_{1:M}, z_{1:M}, f_{1:M} \mid \vec{x}_{1:M}, \Theta\right) = \prod_{i=1}^{M} P\left(z_i \mid \vec{x}_i, \vec{w}_{1:G}\right) P\left(\vec{y}_i \mid z_i, \vec{x}_i, \vec{\beta}_{1:G}, K_\alpha(\cdot,\cdot), \sigma^2\right). \tag{9}$$

Going forward, we will refer to this joint probability as the *complete-data* likelihood, and the expression obtained by marginalizing over $z_{1:M}$ as the *observed-data* likelihood.

It is important to point out that we do not explicitly model treatment in the formulation presented. In scleroderma, no drugs have been shown to have course-altering effects, i.e. change an individual's long-term progression pattern. There are therapies that temporarily relive symptoms. For example, steroids are often given to relieve lung inflammation. Since the timing of administration of these drugs are not always reliably recorded, we treat the effects of these drugs as a latent source of transient variability that is handled by the individual-specific adjustments made by our model. In diseases for which course altering drugs do exist, the effects must be carefully considered when both learning the model and using it to make predictions, but we do not address this in the present work.

Learning. To learn the parameters of the model $\Theta = \{\vec{w}_{1:G}, \vec{\beta}_{1:G}, \alpha, \sigma^2\}$, we treat learning for the kernel parameters α and noise variance σ^2 separately from the remaining parameters $\{\vec{w}_{1:G}, \vec{\beta}_{1:G}\}$. We use clinical priors to restrict the optimization of α and σ^2 to a finite set of parameter combinations. For example, transient variations in PFVC are typically expected to revert to the mean long-term trajectory in 2-3 years. In Figure 2 (top row), the marker trajectory for that individual shows several instances of such variations. Similar reasoning was used to determine the range for the other parameters. Note that when such priors are unavailable, cross-validation is an alternative technique that is frequently used to set hyperparameters.

This amounts to performing a grid search. Within each iteration of the grid search, we learn the remaining parameters $\{\vec{w}_{1:G}, \vec{\beta}_{1:G}\}$ by optimizing the observed-data likelihood. The observed-data likelihood includes a summation over

$z_{1:M}$ that makes direct optimization difficult. Instead, we use the expectation maximization (EM) algorithm, a commonly employed learning technique for models such as ours which contain latent (hidden) variables. For the interested reader, we derive the EM update equations for our model in the appendix.

Inference. There are two scenarios for inference: at baseline before having seen any marker values (i.e. the only information available is through baseline covariates), and at some point during follow-up after having observed marker values at previous follow-up visits. In both scenarios, we compute a distribution over the marker value at a future time t_i^*. The difference between the two scenarios is whether previous observations $\vec{y}_i$ observed at times $\vec{t}_i$ are included in the conditional statement in the predictive distribution.

At baseline, previous marker values are not included and the predictive distribution can be written and expanded using the sum rule of probability

$$P(y_i^* \mid t_i^*, \vec{x}_i, \Theta) = \sum_{z_i=1}^{G} P(z_i \mid \vec{x}_i, \Theta) \int_{f_i} P(y_i^* \mid z_i, f_i, \vec{x}_i, \Theta) P(f_i \mid \Theta)\, d\, f_i \tag{10}$$

$$= \sum_{z_i=1}^{G} \pi_{z_i}(\vec{x}_i)\, \mathcal{N}\left(y_i^* \mid \Phi(t_i^*)\, \vec{\beta}_{z_i}, K_\alpha(t_i^*, t_i^*) + \sigma^2\right). \tag{11}$$

Intuitively, this expression computes the probability of a given marker value by taking a weighted vote from the subtype-specific models. The weights are determined by the baseline covariates, and additional variance is added to each multivariate normal that accounts for uncertainty in the individual-specific adjustments.

Computing the posterior predictive after having observed one or more marker values $\vec{y}_i$ at follow-up times $\vec{t}_i$ is done similarly, but distributions over the latent variables are updated in light of the available information collected over the course of treatment. By including the marker values and observation times in the conditioning statement and again using the sum rule of probability, we have

$$P(y_i^* \mid t_i^*, \vec{y}_i, \vec{t}_i, \vec{x}_i, \Theta) = \sum_{z_i=1}^{G} P(z_i \mid \vec{y}_i, \vec{t}_i, \vec{x}_i, \Theta) \int_{f_i} P(y_i^* \mid z_i, f_i, \vec{x}_i, \Theta) P(f_i \mid z_i, \vec{y}_i, \vec{t}_i \Theta)\, d\, f_i \tag{12}$$

$$= \sum_{z_i=1}^{G} \pi_{z_i}^*(\vec{x}_i)\, \mathcal{N}\left(y_i^* \mid \vec{\mu}_{z_i}^*, \Sigma_{z_i}^*\right). \tag{13}$$

Again, we see that the probability of a future marker value is made using a weighted vote from subtype-specific models. There are, however, two critical differences. First, the prior over subtype membership has been replaced by the posterior over subtype membership given the observed marker values (denoted using $\pi_{z_i}^*(\vec{x}_i)$). Note that this is identical to the distribution used in the expectation step during learning (see Equation 16). The weights, therefore, are a function of how consistent the observed data is with what is predicted by the subtype-specific regression models.

Second, the distribution over f_i depends on both the subtype z_i and the previous marker values. This can be seen by noting that z_i and f_i are marginally independent but are common dependencies of $\vec{y}_i$, and are therefore rendered dependent upon observing the common effect $\vec{y}_i$. Intuitively, this captures the notion that individual-specific adjustments should explain whatever is not explained by the subtype-specific model. From a mechanical standpoint, the posterior over f_i can be carried out using Gaussian process inference. The likely individual-specific adjustments conditioned on a subtype are those that both match the characteristics encoded by the kernel $K_\alpha(\cdot, \cdot)$ and explain the residual variation of the marker values remaining after subtracting the subtype predictions. The posterior over f_i is another Gaussian process and so can be integrated out to obtain the subtype-conditional posterior predictive over y^*. The subtype-conditional posterior predictive is a multivariate normal with mean and covariance

$$\vec{\mu}_{z_i}^* = \Phi(t^*)\vec{\beta}_{z_i} + K_\alpha(t^*, \vec{t}_i) K_\alpha(\vec{t}_i, \vec{t}_i)^{-1}(\vec{y}_i - \Phi(\vec{t}_i)\vec{\beta}_{z_i}) \tag{14}$$

$$\Sigma_{z_i}^* = K_\alpha(t^*, t^*) - K_\alpha(t^*, \vec{t}_i) K_\alpha(\vec{t}_i, \vec{t}_i)^{-1} K(\vec{t}_i, t^*). \tag{15}$$

Experimental Methods

Data. We use data from the patient registry at the Scleroderma Clinic at the Johns Hopkins Hospital. In these experiments, we predict future values of the percent of predicted forced vital capacity (PFVC), which is a measure of lung

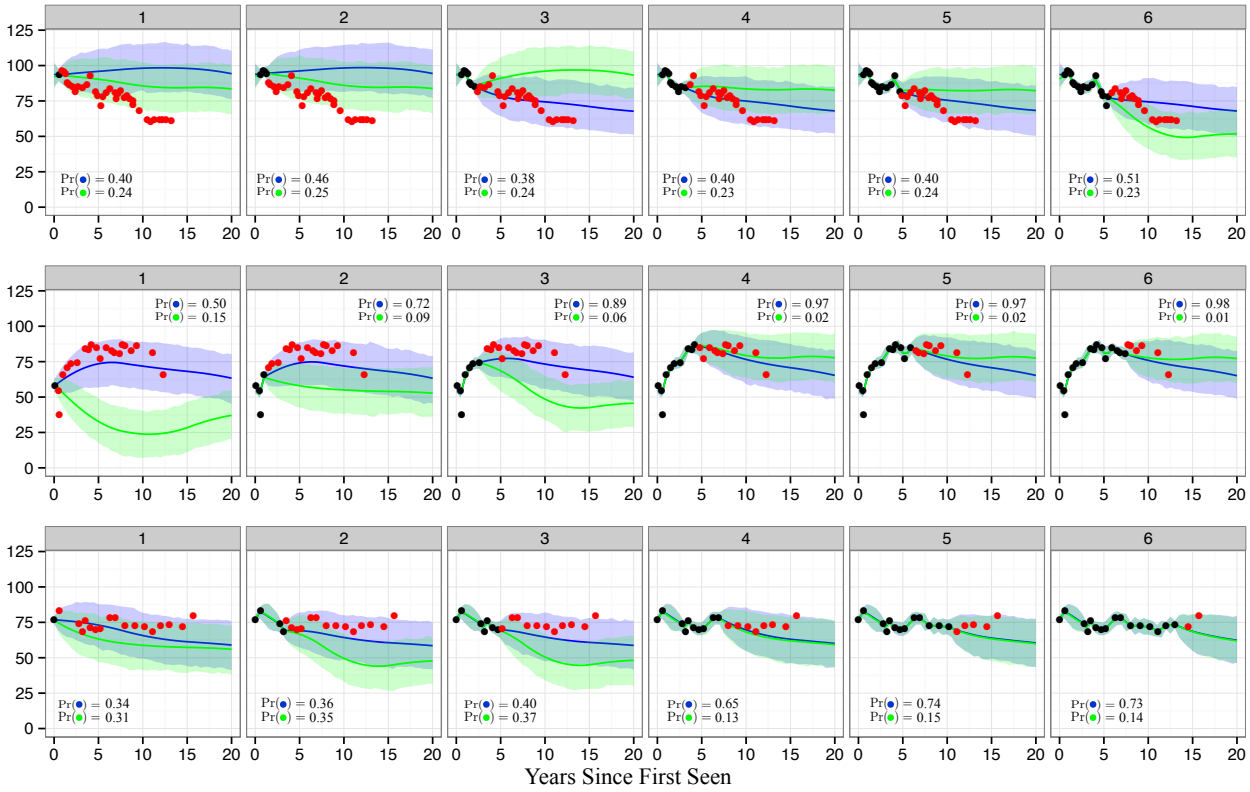

Figure 2: Illustration of PFVC trajectory predictions using our model for three individuals conditioned on different amounts of data. Observed markers are shown in black. Future marker values are shown in red. Blue trajectories are most likely, green are second most likely. The probability of each trajectory is reported in the legend of each cell.

function used to monitor the progression of lung scarring due to scleroderma. Accurately forecasting PFVC trajectories is critical in the treatment of scleroderma lung disease. Although there are no treatments that have been proven to alter course, there are a number of drugs that clinicians believe can taper the inflammatory process that leads to scarring. These drugs have varying levels of toxicity, where the drug believed to be most effective, Cyclophosphamide, has the most harmful side effects. An accurate prediction of an individual's progression can therefore help caregivers to decide which treatment is most appropriate.

To select individuals from the patient registry for this study, we used the following criteria. First we included individuals on whom data early in their course of disease progression were available. Thus, we only include individuals who were seen within two years of their first scleroderma-related symptom. Second, we exclude all individuals with fewer than two PFVC measurements after first being seen by the clinic. Third, antibody measurements were made on only a subset of the individuals. These measurements are assumed to be missing completely at random. As a simplification, we exclude individuals on whom we do not have this data. Finally, we exclude individuals who received a lung transplant. Prediction for these individuals is no longer clinically useful because their trajectory is driven by the transplant rather than the subtype.

The criteria above yield a dataset containing 672 unique individuals and a total of $4,992$ PFVC measurements. The minimum number of follow-up visits is 2 and the maximum is 63. Among all individuals in our dataset, 25% have 3 or fewer PFVC measurements, half have fewer than 6 measurements, and 75% have fewer than 10.

Baseline models. We compare our model against three natural baselines. The first is a B-spline regression that models dependence between observed covariates and expected PFVC trajectory. The effects of covariates are included by first creating second-order conjunctive features among a collection of binary baseline covariates (e.g. one feature may indicate that the individual is female and not Scl-70 antibody positive). We then parameterize the regression by including interactions between the B-spline bases and the second-order conjunctive features. An individual's predicted

Predictions using 1 year of data				
Model	$t \in (1, 2]$	$t \in (2, 4]$	$t \in (4, 8]$	$t \in (8, \infty]$
B-spline Regression with Baseline Covariates	13.55	13.86	14.10	13.23
B-spline Mixture Model (9 subtypes)	7.32	8.79	11.95	14.12
B-spline with Individual Adjustment	6.68	9.16	12.13	13.30
Proposed (9 subtypes)	**6.35**	***8.20**	***10.33**	***12.30**
Predictions using 2 years of data				
B-spline Regression with Baseline Covariates		13.86	14.10	13.23
B-spline Mixture Model (9 subtypes)		7.55	10.03	13.53
B-spline with Individual Adjustment		7.42	10.98	12.70
Proposed (9 subytpes)		***6.91**	***9.03**	***10.63**
Predictions using 4 years of data				
B-spline Regression with Baseline Covariates			14.10	13.23
B-spline Mixture Model (9 Subtypes)			8.52	10.60
B-spline with Individual Adjustment			8.17	11.46
Proposed (9 subtypes)			***6.56**	***9.13**

Table 1: Mean absolute errors of PFVC predictions for the three baselines and the proposed model in all three contexts (after 1, 2, and 4 years of data). Bold numbers indicate best performance across models with * marking the improvement as statistically significant.

trajectory is therefore parameterized by an additive combination of the B-spline coefficients associated with each of the non-zero conjunctive features.

The second is nearly identical to the proposed model, but does not include the individual-specific adjustments. We use a mixture of B-spline regression models. To model the association between baseline covariates and each mixture component, we use a multinomial logistic regression to model the marginal probability of component membership for a given individual. Similarly, the third baseline is a restricted form of the proposed model. We use a single B-spline regression model (i.e. there are no subtypes) that includes individual-specific adjustments f_i.

For the B-spline regression, B-spline mixture, and proposed model we use the same set of baseline covariates: gender, African American race, presence of ACA antibodies, and presence of Scl-70 antibodies. These covariates have been shown to be related to certain lung-disease profiles in the scleroderma literature[1] and are therefore potentially important covariates for prediction. For the B-spline mixture model and proposed model we use nine components/subtypes, which was chosen based on prior work.[18] Finally, for the covariance kernel and random noise in the B-spline with individual adjustments and the proposed model, we chose $\{\nu_b = 16.0,\ a^2 = 64.0,\ \ell = 4.0,$ and $\sigma^2 = 1.0\}$ using domain knowledge.

Metrics. Prediction accuracy in all experiments is measured using the absolute value of the difference between the true and predicted PFVC (absolute error). We compare models by computing mean absolute error of predictions made within certain intervals of time since baseline. Specifically, we group predictions made in the first year of follow-up ($t \in (0, 1]$), in the second year of follow-up ($t \in (1, 2]$), in the third and fourth year of follow-up ($t \in (2, 4]$), fourth to eighth year of follow-up ($t \in (4, 8]$), and beyond the eigth year of follow-up ($t \in (8, \inf]$). These intervals roughly correspond to the quintiles of observation times in our data. All predictions are made using 10-fold cross validation at the level of individuals. That is, the predictions made for a given individual are made using a model trained with her data held out. Summary statistics are then computed across all folds. We make predictions in three contexts: after observing one year of marker values, after observing two years, and after observing four years.

Results

We begin by discussing qualitative results. In Figure 2 we present an illustration of predictions being dynamically updated for three patients (one per row). Each column in each row shows trajectory predictions updated after observing additional data. Each cell in the figure shows two predicted trajectories: blue indicates the most likely trajectory and green indicates the second most likely. Studying the way in which the model reasons about an individual's future course over time yields interesting insights about the strengths of our model. We discuss each of the three individuals in detail.

We see that the first individual (row 1) is initially predicted using less than two years of data to follow a stable or mildly declining trajectory (columns 1 and 2). This is likely because many individuals who begin with PFVC around 100% maintain healthy lungs over the course of the disease. Interestingly, however, after a few more observations collected over 6 months (column 3) the individual begins to show evidence of decline and the model adjusts its predictions accordingly. Based on this prognosis, a clinician may choose to trigger treatment earlier than they might have otherwise.

The second individual (row 2) actually recovers lung capacity over time rather than losing it; a progression that our clinical collaborators found to be especially surprising (further discussion of this trajectory can be found in previous work [18]). It is interesting, however, that the model is fairly confident that the individual will recover after observing only a single PFVC marker. It is especially surprising because the individual starts out with a very low baseline PFVC (one that may alarm clinicians). If, however, the model's predictions were considered when deciding how to proceed with management, a clinician may choose to be less reactive than she would have otherwise using the initial PFVC value alone.

Finally, for the third individual (row 3) we see a rather mild progression. The individual presents with an initial PFVC of 75, which is not alarming but may cause clinicians to monitor the individual carefully. After observing three years of data (column 2), we see that the individual begins to show evidence of decline. A prediction based solely on observed data in column 2 may predict further decline, and indeed our model considers this progression. Interestingly, however, we see that it places nearly equal probability on a more mild progression. Looking ahead, we see that the individual does indeed stabilize, which suggests that our model may be helpful in avoiding false alarms that might be caused by interpreting a rough estimate of rate of decline.

We conclude our experimental results with a quantitative evaluation. We report mean absolute errors for the three baselines and the proposed model in Table 1. The sub-tables show the predictions made using one year of data, two years of data, and four years of data from top to bottom respectively. Within each column of each sub-table we show the mean absolute error achieved by each of the models within the corresponding time interval. For all time intervals and all amounts of observed data, we see that our proposed model outperforms the baselines. In all but one (predictions between years one and two given one year of data), these improvements are statistically significant as computed using a one-sided two-sample t-test. These results suggest that it is indeed important to model both subtypes and individual-specific adjustments as the proposed model outperforms both baselines that include only one of these mechanisms. Furthermore, these positive results reinforce the qualitative evidence discussed above suggesting that our model makes strong predictions of future trajectory.

Conclusion

In chronic, complex diseases such as scleroderma, clinical trials often fail due to heterogeneity across individuals; accurate prognostic models can help to improve trial efficiency by recruiting individuals who are likely driven by the mechanism targeted by the drug under investigation. Moreover, such models can provide valuable guidance in planning therapy.

In this work, we have described a probabilistic model for making individualized predictions about the trajectory of organ function in chronic, complex diseases such as scleroderma. In particular, our solution builds upon the idea of subtypes, which are used to stratify a heterogeneous population into more homogeneous groups. Two challenges with using subtypes is that their characteristics are often unknown and that they do not always account for all observed variation in an individual's progression. We have addressed both issues. The first we address by using electronic health data to *discover* subtypes automatically. The second we address by including an *individual-specific adjustment* term in the probabilistic model that is dynamically learned as more marker values from a given individual are observed. Finally, we presented qualitative results that suggest our model may have important implications for the way individuals are managed in the treatment of scleroderma. In addition, we provide quantitative results demonstrating that our model makes more accurate predictions than three strong baselines.

Our probabilistic approach to modeling heterogeneity allows for ways to readily incorporate other types of variations across individuals. For example, though we we did not explicitly model the different types of variability that may occur due to treatments, these can be incorporated as new forms of individual-specific adjustments.

References

[1] John Varga, Christopher P Denton, and Fredrick M Wigley. *Scleroderma: From Pathogenesis to Comprehensive Management*. Springer Science & Business Media, 2012.

[2] SJG Lewis, T Foltynie, AD Blackwell, TW Robbins, AM Owen, and RA Barker. Heterogeneity of parkinsons disease in the early clinical stages using a data driven approach. *Journal of Neurology, Neurosurgery & Psychiatry*, 76(3):343–348, 2005.

[3] Gilles W De Keulenaer and Dirk L Brutsaert. The heart failure spectrum: Time for a phenotype-oriented approach. *Circulation*, 119(24):3044–3046, 2009.

[4] Matthew W State and Nenad Šestan. The emerging biology of autism spectrum disorders. *Science*, 337(6100):1301, 2012.

[5] Svetlana I Nihtyanova, Benjamin E Schreiber, Voon H Ong, Daniel Rosenberg, Pia Moinzadeh, J Gerrard Coghlan, Athol U Wells, and Christopher P Denton. Prediction of pulmonary complications and long-term survival in systemic sclerosis. *Arthritis & rheumatology*, 66(6):1625–1635, 2014.

[6] Dinesh Khanna, Chi-Hong Tseng, Niloofar Farmani, Virginia Steen, Daniel E Furst, Philip J Clements, Michael D Roth, Jonathan Goldin, Robert Elashoff, James R Seibold, et al. Clinical course of lung physiology in patients with scleroderma and interstitial lung disease: analysis of the scleroderma lung study placebo group. *Arthritis & Rheumatism*, 63(10):3078–3085, 2011.

[7] Shervin Assassi, Roozbeh Sharif, Robert Lasky, Terry McNearney, Rosa Estrada-Y-Martin, Hilda Draeger, Deepthi Nair, Marvin Fritzler, John Reveille, Frank Arnett, et al. Predictors of interstitial lung disease in early systemic sclerosis: A prospective longitudinal study of the genisos cohort. *Arthritis Research and Therapy*, 12(5):R166, 2010.

[8] Souhaib Ben Taieb and Rob Hyndman. Boosting multi-step autoregressive forecasts. In *Proceedings of The 31st International Conference on Machine Learning*, pages 109–117, 2014.

[9] Md Rafiul Hassan and Baikunth Nath. Stock market forecasting using hidden markov model: a new approach. In *Intelligent Systems Design and Applications, 2005. ISDA'05. Proceedings. 5th International Conference on*, pages 192–196. IEEE, 2005.

[10] John A Quinn, Christopher KI Williams, and Neil McIntosh. Factorial switching linear dynamical systems applied to physiological condition monitoring. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(9):1537–1551, 2009.

[11] Carl E. Rasmussen and Christopher K. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[12] Nicolas Chapados and Yoshua Bengio. Augmented functional time series representation and forecasting with gaussian processes. In *Advances in Neural Information Processing Systems*, 2007.

[13] Jürgen Wiest, M Hoffken, Ulrich Kresel, and Klaus Dietmayer. Probabilistic trajectory prediction with gaussian mixture models. In *Intelligent Vehicles Symposium (IV), 2012 IEEE*, pages 141–146, 2012.

[14] Lei Xu, Michael I Jordan, and Geoffrey E Hinton. An alternative model for mixtures of experts. *Advances in neural information processing systems*, pages 633–640, 1995.

[15] Zijian Zheng and Geoffrey I Webb. Lazy learning of bayesian rules. *Machine Learning*, 41(1):53–84, 2000.

[16] Shyam Visweswaran and Gregory F Cooper. Patient-specific models for predicting the outcomes of patients with community acquired pneumonia. In *AMIA Annual Symposium Proceedings*, volume 2005, page 759. American Medical Informatics Association, 2005.

[17] Emily Watt, James W Sayre, and Alex AT Bui. Applying an instance-specific model to longitudinal clinical data for prediction. In *Healthcare Informatics, Imaging and Systems Biology (HISB), 2011 First IEEE International Conference on*, pages 81–88. IEEE, 2011.

[18] Peter Schulam, Fredrick Wigley, and Suchi Saria. Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery. In *Proceedings of the Twinty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Appendix: EM Update Equations

In the expectation step, we compute the posterior over the latent variables $z_{1:M}$ using the current parameter estimates at iteration t denoted $\Theta^{(t)}$. Above, we have shown that the likelihood for each individual is a mixture of multivariate normals. Computing the posterior therefore simply involves computing the joint probability of the marker values and each subtype assignment and normalizing:

$$P\left(z_i \mid \vec{y}_i, \vec{x}_i, \vec{t}_i, \Theta^{(t)}\right) \propto \pi\left(\vec{x}_i\right) \mathcal{N}\left(\vec{y}_i \mid \phi\left(\vec{t}_i\right) \vec{\beta}_{z_i}, K_\alpha\left(\vec{t}_i, \vec{t}_i\right) + \sigma^2 \boldsymbol{I}\right). \tag{16}$$

In the maximization step, we use the posterior over z_i to maximize the expected value of the complete-data log-likelihood. Recall that the expectation is taken with respect to the posterior over z_i. First, the complete-data log likelihood of the data after marginalizing f_i is

$$\sum_{i=1}^{M} \log P\left(z_i \mid \vec{w}_{1:G}\right) + \log P\left(\vec{y}_i \mid z_i, \vec{\beta}_{1:G}, K(\cdot,\cdot), \sigma^2\right). \tag{17}$$

To maximize this objective with respect to $\vec{w}_{1:G}$, we can focus on the first term in the sum above. By writing the density out fully using the multinomial logistic regression formulation above we have

$$\sum_{i=1}^{M} \vec{w}_{z_i}^\top \vec{x}_i - \log\left(\sum_{g'=1}^{G} e^{\vec{w}_{g'}^\top \vec{x}_i}\right) = \sum_{i=1}^{M}\sum_{g=1}^{G} \mathbb{I}\left(z_i = g\right)\left[\vec{w}_{z_i}^\top \vec{x}_i - \log\left(\sum_{g'=1}^{G} e^{\vec{w}_{g'}^\top \vec{x}_i}\right)\right]. \tag{18}$$

Taking the expectation of this expression with respect to the posterior over z_i amounts to replacing the indicator function $\mathbb{I}\left(z_i = g\right)$ with the posterior probability $P\left(z_i = g \mid \vec{y}_i, \vec{x}_i, \vec{t}_i, \Theta^{(t)}\right)$. We can maximize this expression with respect to the multinomial logistic regression parameters $\vec{w}_{1:G}$ using gradient-based methods.

Maximizing the second term in the likelihood can be done by solving a weighted least squares problem. To see this, we first write out the second term in Equation 17 using the log of the multivariate normal density. Let $K'_\alpha\left(\vec{t}_i, \vec{t}_i\right)$ be the full covariance term in the multivariate normal above: $K_\alpha\left(\vec{t}_i, \vec{t}_i\right) + \sigma^2 \boldsymbol{I}$. We can then write the second term out as

$$\sum_{i=1}^{M} -\frac{N_i}{2}\log 2\pi - \frac{1}{2}\log\left|K'_\alpha\left(\vec{t}_i, \vec{t}_i\right)\right| - \frac{1}{2}\left(\vec{y}_i - \Phi\left(\vec{t}_i\right)\vec{\beta}_{z_i}\right)^\top K'_\alpha\left(\vec{t}_i, \vec{t}_i\right)^{-1}\left(\vec{y}_i - \Phi\left(\vec{t}_i\right)\vec{\beta}_{z_i}\right). \tag{19}$$

Maximizing this expression with respective $\vec{\beta}_g$ for each $g \in \{1, \ldots, G\}$ is equivalent to minimizing the negative value of the quadratic in the third term in the sum above. Let $W_i = K'_\alpha(\vec{t}_i, \vec{t}_i)^{-1}$, then we can write the negative of the quadratic term as a weighted least squares objective with weight matrix W_i. The sufficient statistics required to solve for $\vec{\beta}_{z_i}$ are therefore $\eta_{z_i}^{(i1)} = \Phi^\top\left(\vec{t}_i\right) W_i\, \Phi\left(\vec{t}_i\right)$ and $\eta_{z_i}^{(i2)} = \Phi^\top\left(\vec{t}_i\right) W_i\, \vec{y}_i$. To maximize $\vec{\beta}_{1:G}$ we leverage two facts from statistics. First, the sufficient statistics required to compute the maximum likelihood estimate from M independent weighted linear regressions is simply the sum of the individual sufficient statistics. Second, when maximizing an expected complete-data log-likelihood, we can replace the sufficient statistics with expected sufficient statistics. In this case, we multiply $\eta_{z_i}^{(i1)}$ and $\eta_{z_i}^{(i2)}$ by the posterior probability over z_i. We can therefore compute the maximum for $\vec{\beta}_g$ using

$$\left(\sum_{i=1}^{M} P(z_i \mid \vec{y}_i, \vec{x}_i, \vec{t}_i, \Theta^{(t)})\Phi^\top\left(\vec{t}_i\right) W_i\, \Phi\left(\vec{t}_i\right)\right)^{-1}\left(\sum_{i=1}^{M} P(z_i \mid \vec{y}_i, \vec{x}_i, \vec{t}_i, \Theta^{(t)})\Phi^\top\left(\vec{t}_i\right) W_i\, \vec{y}_i\right). \tag{20}$$