Name:	Peter Schulam
Affiliation:	Johns Hopkins University
Email:	pschulam@cs.jhu.edu
Presentation type:	Poster

Integrative Analysis using Coupled Latent Variable Models for Individualizing Prognoses

Peter Schulam, MS[1], Colin Ligon, MD MHS[2], Fredrick Wigley, MD[2],
Robert Wise, MD[2], Laura Hummers, MD ScM[2,] Suchi Saria, PhD[1,3]
[1]Johns Hopkins University, Baltimore, MD; [2]Johns Hopkins University School of Medicine, Baltimore, MD;
[3]Bloomberg School of Public Health, Baltimore, MD

Abstract

Personalized medicine hinges on the ability to make accurate, individualized prognoses using the entirety of a patient's clinical history. Common regression techniques used to aid in making diagnoses do not align with this goal for several reasons. First, they typically make predictions of a single future outcome, but a picture of an individual's full future disease course would be more informative in guiding treatment decisions. Second, the predictions they make cannot be updated dynamically using a growing clinical history, which is critical for refining prognoses as more observations are made. Finally, they often use heuristic summaries or simple features to capture information from the clinical history, which may miss subtle cues in past disease expression. We describe an algorithm for making dynamically-updated, personalized prognoses of an individual's full future disease course. We predict a probability distribution over the trajectory of a quantitative clinical marker measuring disease activity (e.g. the percent of predicted forced vital capacity) using an integrative analysis of both baseline information (e.g. demographic characteristics) and the full histories of longitudinally recorded information (e.g. clinical observations and laboratory test results tracked over follow-up). We use our approach to predict lung disease trajectories in scleroderma, a complex autoimmune disease.

Introduction

In personalized (or precision) medicine, the goal is to tailor treatment to a given patient using a personalized prognosis based on all information in the individual's clinical history. In this work, we formalize the problem of personalized prognosis as that of estimating an individualized *function of time* that models the full trajectory of a *target clinical marker* (see Figure 1 for an illustrative example). Clinical markers are quantitative test results used to monitor disease activity and progression in a specific organ system. For example, in scleroderma, a complex autoimmune disease, the percent of predicted forced vital capacity (PFVC) is a clinical marker measuring lung damage severity, and the total modified Rodnan skin score (TSS) is a marker measuring skin disease activity.

In practice, cross-sectional regression models are commonly used to predict clinical marker values at fixed future time points using collections of baseline characteristics and summary statistics of an individual's clinical history from a fixed window (e.g. estimate of marker slope in the past year). These techniques do not align with the dynamic nature of personalized medicine, where new observations are frequently added to a growing clinical history and must be integrated into an updated prognosis. Moreover, predicting a single future outcome may not be sufficiently informative to guide treatment decisions. We describe a model for personalized prognosis that (1) predicts the full future trajectory of a clinical marker, (2) provides uncertainty estimates around the predicted trajectory, and (3) uses a growing clinical history to update prognoses. We demonstrate our approach by using it to predict the course of interstitial lung disease in patients with scleroderma, a complex autoimmune disease.

Methods

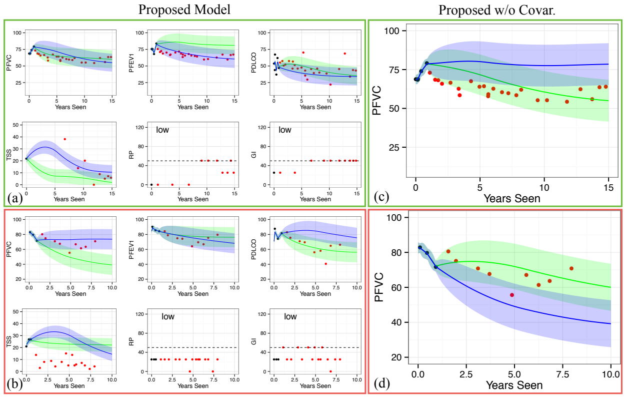
Our approach dynamically updates personalized prognoses using an *integrative analysis* of both baseline characteristics of an individual (e.g. gender) and the time-evolving histories of both the target marker and other *auxiliary clinical markers* tracking related organ systems. To extract information from clinical marker histories, we propose a probabilistic model of clinical marker trajectories that uses both observed and unobserved factors to explain heterogeneous patterns of activity. The form of the probabilistic model and its latent variables are motivated by the idea of *subtypes* (see e.g. Saria and Goldenberg[1]), which have become increasingly important in understanding complex, chronic diseases, and by the idea of *nuisance variability* (see e.g. Lötvall and others[2]), which is individual-specific variation caused by factors orthogonal to disease subtype. Details about the model and its motivation can be found in Schulam and Saria.[3] Our algorithm dynamically estimates unobserved factors using Bayesian inference as more clinical markers are recorded, which act as natural summaries of the clinical history. These inferences are combined using a conditional random field, which is trained to maximize the predictive probability of all future clinical markers (i.e. the future disease activity trajectory). The full integrative technique is described in Schulam and Saria.[4]

To demonstrate our approach, we build a tool to predict lung disease trajectories for individuals with scleroderma. Clinicians use percent of predicted forced vital capacity (PFVC) to track lung disease severity, which is expected to drop as the disease progresses. To train and validate our model, we use data from the Johns Hopkins Scleroderma Center patient registry; one of the largest collections of clinical scleroderma data in the world. We extract the PFVC trajectories of 772 individuals along with five auxiliary markers: % predicted forced expiratory volume in one second (PFEV1), % diffusing capacity (PDLCO), total modified Rodnan skin score (TSS), and two Likert-valued clinical severity scores (one reflecting disease activity affecting the vasculature and one for the gastrointestinal tract). We evaluate our approach by (1) predicting PFVC trajectories and (2) by detecting individuals who will drop by more than 20 PFVC using a score derived from each model's predicted drop in lung capacity. As baselines, we compare our model against a static B-spline regression model, a B-spline regression with an individual-specific Gaussian process allowing for dynamic individualization, a version of our model including subtypes but without any individualization (PwoI), and a version of our model without baseline covariates or auxiliary markers (PwoC).

Results

Figure 1 provides qualitative evidence of the improved performance gained by using an integrative approach. Observed measurements are in black and those to be predicted are in red. The blue trajectory shows the most likely predicted trajectory and the green shows the second most likely. In Figure 1a, we see a 55-year-old white woman who presents with mildly impaired lung function (approximately 65 PFVC), but seems to recover over the course of the first year. The model without time covariates (Figure 1c) predicts that this recovery will stabilize and hold. The model that accounts for the histories of other auxiliary markers, however, correctly predicts that this woman will decline. This is likely due to the early PFEV1 trajectory, which initially dips instead of recovering. In Figure 1b, we see a 75-year-old white woman who initially declines, but later stabilizes. The model that considers the history of PFVC alone over-reacts to this initial decline, whereas the model that includes information from the auxiliary markers correctly predicts that the woman will stabilize after this initial decline.

Figure 2 displays ROCs for the declining population detection task with associated AUCs listed in Table 1. We see that the the proposed model's predictions are more discriminative than the B-spline+GP baseline and the model that depends on the target marker history alone (PwoC).



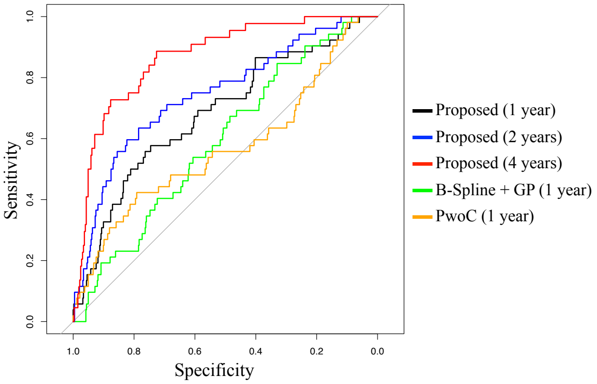


Figure 1. Predictions made using integrative analysis (a, b) and using the target marker history only (c, d). **Figure 2.** ROCs comparing at-risk individual detection across proposed model and baselines.

Table 1. AUCs corresponding to ROCs in Figure 2.

Model / Years of Data	1	2	4
B-Spline+GP	0.59	0.63	0.74
PwoC	0.57	0.71	0.84
Proposed	0.68	0.75	0.87

References

1. Saria S, Goldenberg A. Subtyping: What it is and its role in precision medicine. IEEE Intelligent Systems. 2015 Jul;30(4):70-5.
2. Lötvall J, Akdis CA, Bacharier LB, Bjermer L, Casale TB, Custovic A, Lemanske RF, Wardlaw AJ, Wenzel SE, Greenberger PA. Asthma endotypes: a new approach to classification of disease entities within the asthma syndrome. Journal of Allergy and Clinical Immunology. 2011 Feb 28;127(2):355-60.
3. Schulam P, Saria S. A Framework for Individualizing Predictions of Disease Trajectories by Exploiting Multi-Resolution Structure. In Advances in Neural Information Processing Systems 2015 (pp. 748-756).
4. Schulam P, Saria S. Integrative Analysis using Coupled Latent Variable Models for Individualizing Prognoses. Journal of Machine Learning Research. 2016 (forthcoming).