6.S897/HST.956 Machine Learning for Healthcare

Lecture 19: Disease Progression Modeling and Subtyping Part 2

Instructors: David Sontag, Peter Szolovits

1 Recap of goal of disease progression modeling

1.1 Predictive

- What will this patient's future trajectory look like

1.2 Desriptive

- Find markers of disease stage and progression, statistics of what to expect when
- Discover new disease subtypes
- What does it mean for a disease to progress over time

1.3 Key Challenges

- seldom directly observe disease stage, but rather only indirect observations (symptoms)
- Data is censored doesn't observe beginning to end and (ex. patients may be coming in at varying stages of disease)

2 Lecture Overview

- Staging from cross-sectional data
  - Wang, Sontag, Wang KDD 2014
  - Pseudo-time methods from computational biology
- simultaneous staging and subtyping
  - Young et al., Nature Communications 2018

3 Stage vs. Subtype

Staging is sorting patients into early-late disease or severity, i.e. discover the trajectory. The main challenge is that training data is typically cross sectional which means only one single time point (i.e. censored to be a short window) is avaliable. For instance a single vector of lab results on patient, but dont have historical trajectory of the disease. The subtyping algorithms from Thursdays lecture (e.g. k-means clustering) on a single time point individual data would give you clusters that correspond to stages of the disease but would completely ignore subtypes. Naive clustering cant differentiate between stage and subtype (all patients assumed to be aligned at baseline).

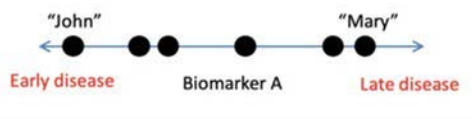

Figure 1: representation of 1-D data

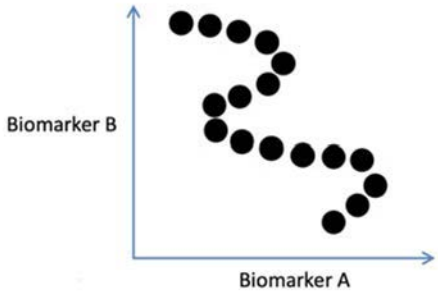

Figure 2: Insight #1

4 Data points in different dimensions

4.1 1-D

In 1-D. might assume that low values correspond to an early disease stage (or vice-versa). Another potential representation of the data in figure 1 is that the normal representation is in the middle and the abnormal is to the right and left.

We can figure out the direction of the biomarker value corresponds with disease progression by correlating with data on how much time the patient has to live. We can also look at the average age of individuals on each side of spectrum and assume the side with a larger average age is later progression.

4.2 Higher Dimensional

Visually this could be a two biomarker scatter plot.

- Insight #1: with enough data it may be possible to recognize a structure to the data. [BDA+14].
- Insight #2: sequential observations from same patient can also help

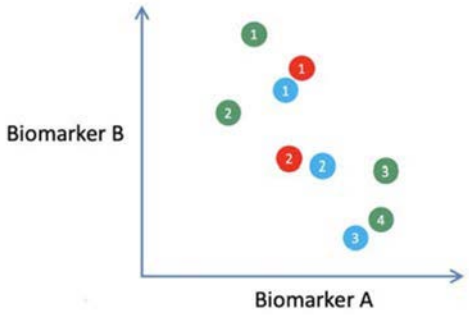

Figure 3: Insight #2

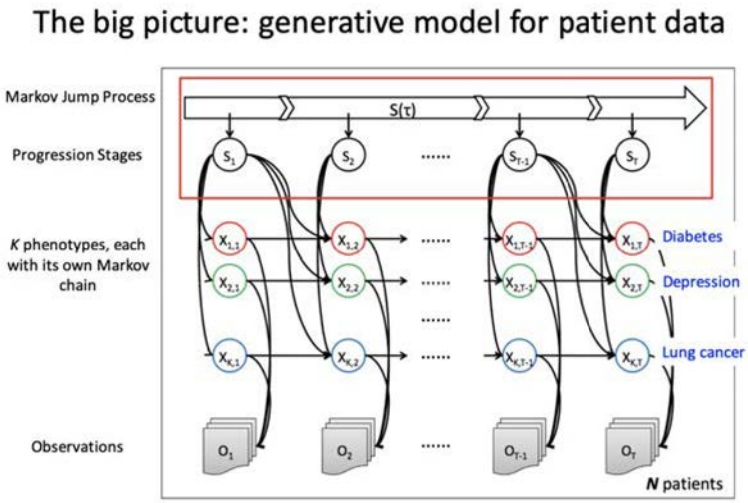

Figure 4: Diagram of the generative model for patient data

Using higher dimensional data can also help to discover disease subtypes

5 COPD Diagnosis & Progression

Chronic Obstructive Pulmonary Disease, COPD, is a condition of the lungs typically caused by air pollution or smoking.

5.1 Staging

COPD also has a good staging method which is to use a spirometry device to measure lung function of the individual through inhale/exhale. The duration of someones exhale is the important information, thus diagnosis are made using a breath test. If a fraction of air expelled in first second of exhalation $< \%70$, then the patient is diagnosed with COPD. Most doctors use GOLD criteria to stage the disease and measure its progression (1-mild to 4-severe). COPD has a reasonably well understood and good staging mechanism which can help to test the algorithms to see if they line up with what is predicted or with the same conclusions other studies have concluded.

5.2 Big picture: generative model for patient data, markov model

This model assumes not a lot of longitudinal data for a patient

Image of model from Wang, Sontag, Wang Unsupervised learning of Disease Progression Models

5.2.1 Markov Jump Process

The Markov Jump Process is a model for patients disease progression across time. It is a continuous-time Markov process with irregular discrete-time observations. One random variable for each point in time with observation of a patient's data. In figure 4,S denotes discrete disease stages in this model (1-early and 10-later stage). The model gives us the probability distribution of transitioning in between sequential stages (S1,S2). The transition probability is defined by an intensity matrix and the time interval: $A_{ij}(\Delta) \triangleq P(S_t = j | S_{t-1} = i, \tau_t - \tau_{t-1} = \Delta; Q)$ or $A_{ij}(\Delta) = expm(\Delta Q)_{ij}$

5.2.2 Noisy-OR network

A Noisy-OR network is used as a model for data at a single point in time.

It is important to understand what makes each disease stage unique, but there is a lot of noise and bias that goes into assigning the disease codes to patients. To fix this problem suppose there is a generative distribution, the diagnosis code for the disease are likely to be observed with edge weight probability given the patient has the disease. This captures the noise rate of assigning codes. Since there are weighted edges the algorithm also learns the transition probabilities to understand which edges exist.

In order to better understand the hidden variables, X, they will be grounded using anchors. An anchor is a finding that can only be caused by a single comorbidity (see Lecture 8). [HCHS14] To use anchors to ground the hidden variables, anchors will be provided for each of the combrodities. Without anchors the results are less interpretable.

5.2.3 Model of comorbidities across time

The presence of comorbidities depends on the value at previous time steps and on the stage of the disease. Later stages of the disease are more likely to develop comorbidities. It is important to make the assumption that once a patient has a comorbidity it is likely to always have it.

5.3 Experiment Evaluation

To evaluate the success of the experiment, the authors created a COPD cohort of 3,705 patients. Within this cohort, each patient was required to have at least one COPD-related diagnosis code and at least one COPD-related drug. In addition, patients with two few records were removed entirely.

The clinical findings in the study were derived from 264 diagnosis codes. ICD-9 codes that only occurred in a small number of patients were removed.

All of the visits recorded were combined into 3-month time windows. This data filtering process resulted in 34,976 total visits and 189,815 positive findings.

5.4 Inference

The authors performed marginal inference over both latent variables and model parameters using an algorithm with nested loops.

In the outer loop of the inference algorithm, the authors used Expectation-Maximization to find a local optimum of the likelihood. The authors then used a Markov Jump Process inspired by recent breakthroughs in physics research [MHS07].

In the inner loop of the inference algorithm, the authors used a Gibbs sampler and performed block sampling of the Markov chains to improve the mixing time of the Gibbs sampler.

Professor Sontag noted that if he were to redo the experiment, he would use variational inference with a recognition network.

5.5 Customizations for COPD

The authors made a number of customizations to the algorithm. First, the authors enforced monotonic stage progression such that later stages of the disease were always subsequent to earlier stages of the disease, i.e. $S_{t+1} >= S_t$.

The authors also enforced monotonicity in the distribution of comorbidities in the first time step, e.g. $Pr(X_{j,1}|S_1 = 2) >= Pr(X_{j,1}|S_1 = 1$. To accomplish this, they solved a tiny convex optimization problem within Expected-Maximization.

In addition, the authors enforced that transitions in X can only happen at the same time as transitions in S.

Finally, the edge weights were given a prior $Beta(0.1, 1)$ to encourage sparsity.

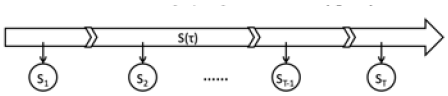

Figure 5: Here is a diagram of the enforced stage progression.

Edges learned for *lung cancer*

Diagnosis code	Weight	
*162.9	0.60	Malignant Neoplasm Of Bronchus And Lung
518.89	0.15	Other Diseases Of Lung, Not Elsewhere Classified
*162.8	0.15	Malignant Neoplasm Of Other Parts Of Lung
*162.3	0.15	Malignant Neoplasm Of Upper Lobe, Lung
786.6	0.15	Swelling, Mass, Or Lump In Chest
793.1	0.10	Abnormal Findings On Radiological Exam Of Lung
786.09	0.07	Other Respiratory Abnormalities
*162.5	0.06	Malignant Neoplasm Of Lower Lobe, Lung
*162.2	0.04	Malignant Neoplasm Of Main Bronchus
702.0	0.03	Actinic Keratosis
511.9	0.03	Unspecified Pleural Effusion
*162.4	0.03	Malignant Neoplasm Of Middle Lobe, Lung

Figure 6: Here are the anchor edge weights learned for lung cancer.

5.6 Results and Analysis

5.6.1 Edges Learned for Different Comorbidities

After running the model, edge weights were learned for different comorbodity variables. There were some edge weights learned for established anchors (highlighted in red) and some weights learned that were automatically associated by the unsupervised learning algorithm (highlighted in blue). Notably all of the edge weights were far less than 1.0, illustrating the significant amount of noise in diagnosis code assignment.

Professor Sontag highlighted several examples of comorbidity edge weights, including kidney disease, lung cancer, and lung infection.

5.6.2 Progression of a Single Patient

From the results of the model the authors could also make predictions about the disease progression of an individual patient. The model could infer the trajectory and time of onset for different predicted comorbidities relative to the predicted onset of different disease stages.

Edges learned for *lung cancer*

Diagnosis code	Weight	
*162.9	0.60	Malignant Neoplasm Of Bronchus And Lung
518.89	0.15	Other Diseases Of Lung, Not Elsewhere Classified
*162.8	0.15	Malignant Neoplasm Of Other Parts Of Lung
*162.3	0.15	Malignant Neoplasm Of Upper Lobe, Lung
786.6	0.15	Swelling, Mass, Or Lump In Chest
793.1	0.10	Abnormal Findings On Radiological Exam Of Lung
786.09	0.07	Other Respiratory Abnormalities
*162.5	0.06	Malignant Neoplasm Of Lower Lobe, Lung
*162.2	0.04	Malignant Neoplasm Of Main Bronchus
702.0	0.03	Actinic Keratosis
511.9	0.03	Unspecified Pleural Effusion
*162.4	0.03	Malignant Neoplasm Of Middle Lobe, Lung

Figure 7: Here are the automatically associated edge weights learned for lung cancer.

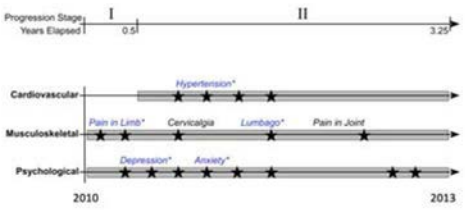

Figure 8: Here is an example projected trajectory of an individual patient. Gray bars indicate predicted comorbidity onset; stars indicate the occurrence of associated ICD-9 codes.

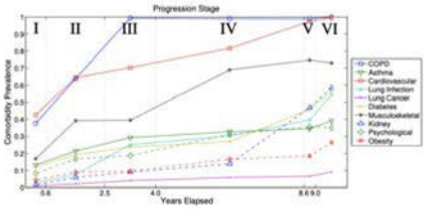

Figure 9: Here are the characterized trajectories of comorbidity progressions relative to state transitions.

5.6.3 Prevalence of Comorbidities Across Stages

The authors could also characterize the overall progression trajectories of the disease and comorbidities. To accomplish this, the authors first generated 10,000 virtual records and averaged the holding time for each state and the prevalence of different comorbidities.

Of particular note was how the comorbidity of cardiovascular disease aligned with COPD progression. The authors confirmed this characterization created by the model by checking against medical knowledge of the two diseases [Sin09].

6 Pseudo-time Methods from Computational Biology

Professor Sontag then discussed how methods from computational biology to characterize progression of single cells over time could be applied to disease progression modeling as well. Due to growing popularity of single-cell sequencing experiments in the biology community in the last five years, many more pseudotime estimation techniques have been developed.

Single-cell sequencing experiments aim to understand on a cell by cell basis what genes are expressed in each cell. The process involves isolating individual cells from tissue, extracting RNA, barcoding each RNA from each individual cell, and then deconvolving to see what the original RNA expression was.

This approach can be applied to disease progression modeling by thinking of each cell as an individual patient and the expression of different genes as the expression of different symptoms. Professor Sontag cited an example Computational Biology paper that compared different single-cell trajectory inference methods as a framework that could be extended from different trajectories to different disease subtypes [SCTS19].

One example of a trajectory that can be applied to a disease progression paradigm is a Bifurcation, e.g. if a set of patients receive a certain treatment and another set don't their outcomes could diverge.

6.1 MST-based Approach (Monocle)

One example of a model to track single-cell trajectories is a Minimum Spanning Tree approach called Monocle [TCG+14]. Monocle represents each cell as a point in the expression space. It reduces the dimensionality

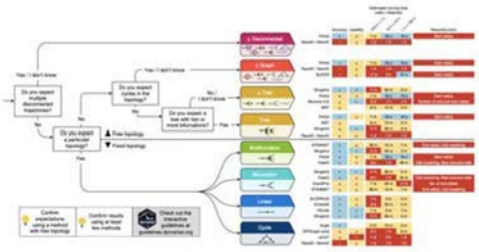

Figure 10: Here are possible trajectory inference methods.

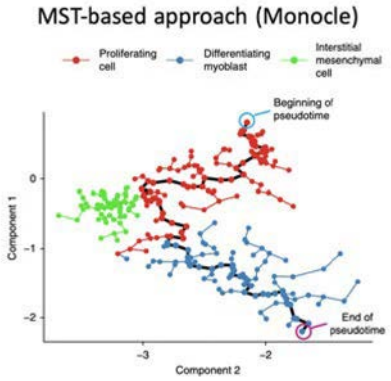

Figure 11: Here is a sample MST produced by Monocle.

of the space and then creates a Minimum Spanning Tree on the cells. The cells can then be ordered in pseudotime via the tree; the researchers find the longest path in the tree as in the Traveling Salesman Problem. One side of the path corresponds to early disease stage and the other to late disease stage. The cells can be labeled by type and colored by what stage of differentiation they are in.

6.2 Statistical Model

One other example of a pseudo-time method is a statistical model for probabilistic pseudotime [CY16]. The researchers defined a Gaussian process μ for a collection of timepoints. They then defined a covariance function k on these timepoints. Mapping this to a disease progression frame, we can assume observations we have are drawn from a Gaussian distribution. We can sample a timepoint from a latent distribution and on a sample μ curve, we can evaluate the expected value for a patient at the sample timepoint.

References

[BDA+14] Sean C Bendall, Kara L Davis, El-ad David Amir, Michelle D Tadmor, Erin F Simonds, Tiffany J Chen, Daniel K Shenfeld, Garry P Nolan, and Dana Peer. Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. *Cell*, 157(3):714–725, 2014.

[BSOM+14] David W Bates, Suchi Saria, Lucila Ohno-Machado, Anand Shah, and Gabriel Escobar. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, 33(7):1123–1131, 2014.

[CY16] Kieran R Campbell and Christopher Yau. Order under uncertainty: robust differential expression analysis using probabilistic models for pseudotime inference. *PLoS computational biology*, 12(11):e1005212, 2016.

[HCHS14] Yoni Halpern, Youngduck Choi, Steven Horng, and David Sontag. Using anchors to estimate clinical state without labeled data. In *AMIA Annual Symposium Proceedings*, volume 2014, page 606. American Medical Informatics Association, 2014.

[MHS07] Philipp Metzner, Illia Horenko, and Christof Schütte. Generator estimation of markov jump processes based on incomplete observations nonequidistant in time. *Physical Review E*, 76(6):066702, 2007.

[SCTS19] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature biotechnology*, page 1, 2019.

[Sin09] Don D Sin. Is copd really a cardiovascular disease? *Chest*, 136(2):329–330, 2009.

[TCG+14] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381, 2014.

[WSW14] Xiang Wang, David Sontag, and Fei Wang. Unsupervised learning of disease progression models. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 85–94. ACM, 2014.

6.S897 / HST.956 Machine Learning for Healthcare

Spring 2019