Machine Learning for Healthcare

6.871, HST.956

Lecture 21: Disease progression modeling & subtyping, Part 2

David Sontag

CSAIL

INSTITUTE FOR MEDICAL ENGINEERING & SCIENCE

HEALTH SCIENCES & TECHNOLOGY

Course announcements

- Project touchpoints due Wed 4/29
- Good time to re-engage clinical mentors
  - Schedule meeting with them late this week / early next week
  - E-mail them writeup for touchpoint (CC: TA)
- Class this Thu 4/30 will be student-moderated project discussions

Recap of past two lectures

- How do we define disease?
- Genomics as a driver of major changes in precision medicine
- Clustering with clinical data to discover disease subtypes
- Prediction of disease progression from a single time-point

Outline of today’s lecture

- Deep dive into data commonly used for disease progression modeling
- What can we draw inspiration from, and why they are not good enough
- Probabilistic models of disease progression
- Simultaneous staging & subtyping

Outline of today’s lecture

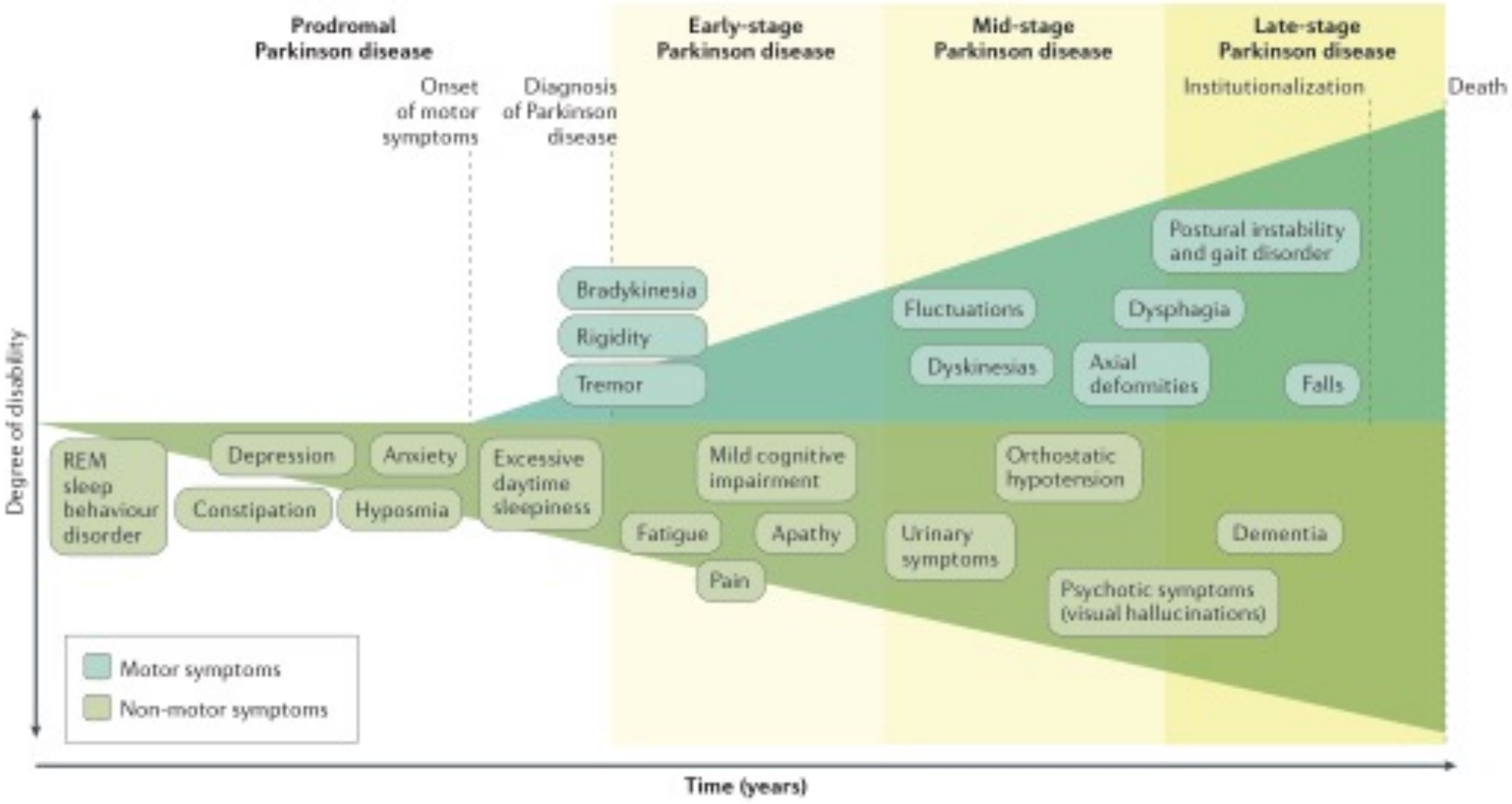
- **Deep dive into data commonly used for disease progression modeling**
- What can we draw inspiration from, and why they are not good enough
- Probabilistic models of disease progression
- Simultaneous staging & subtyping

UK Biobank (from Lecture 19)

- UK Biobank collects data on ~.5M de-identified individuals
  - everyone will have full exome sequencing (50K so far)
  - 100K have worn 24-hour activity monitor for a week, 20K have had repeat measurements
  - on-line questionnaires: diet, cognitive function, work history, digestive health
  - 100K will have imaging: brain, heart, abdomen, bones, carotid artery
  - linking to EHR: death, cancer, hospital episodes, GP, blood biochemistry

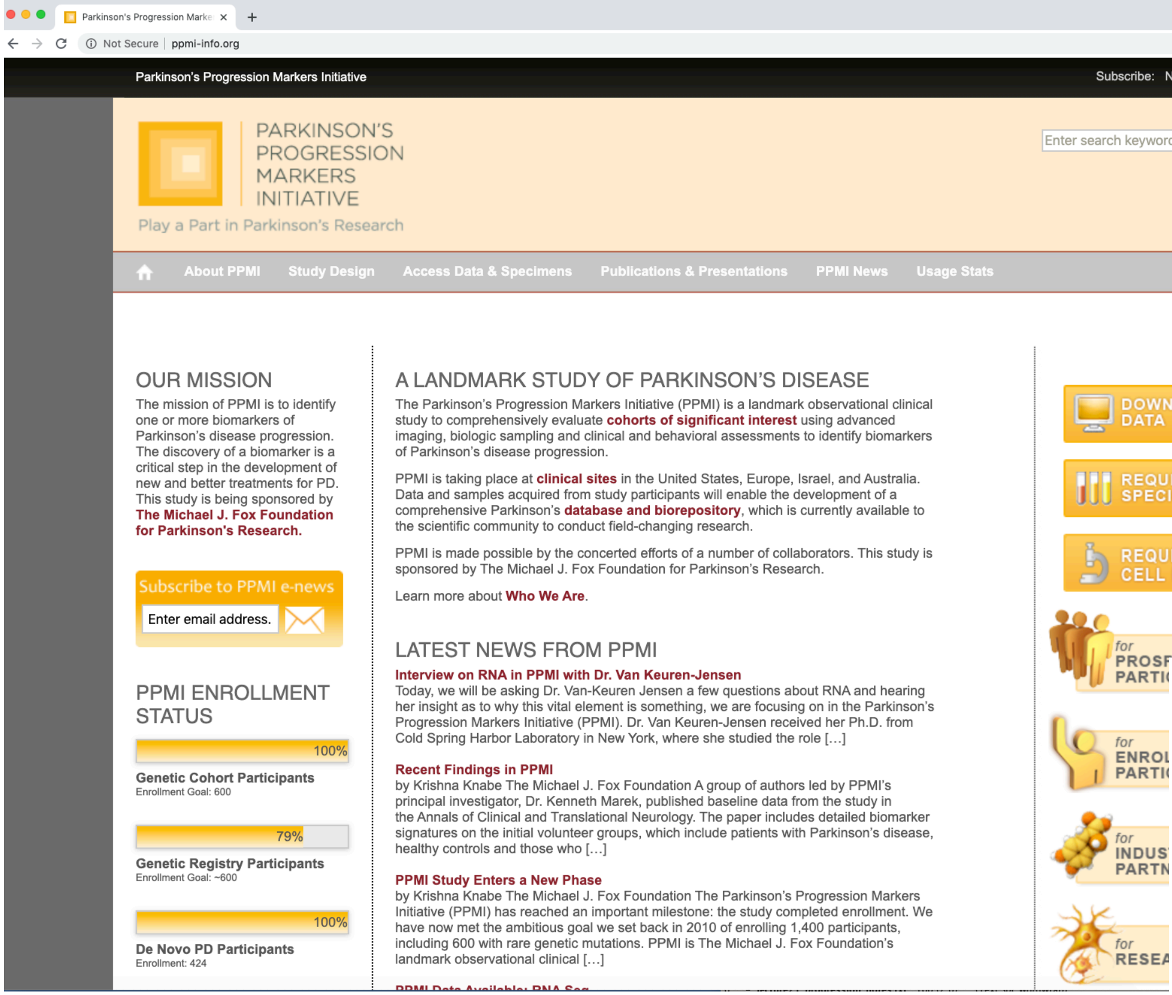
We have similar biobanks in the United States, including Partners Healthcare's biobank (>40K patients), Million Veteran's Program, NIH's All of Us

Parkinson's Progression Marker Initiative (from Lecture 20)


Prodromal Parkinson disease
Early-stage Parkinson disease
Mid-stage Parkinson disease
Late-stage Parkinson disease
Onset of motor symptoms
Diagnosis of Parkinson disease
Institutionalization
Death
Degree of disability
Postural instability and gait disorder
Bradykinesia
Rigidity
Tremor
Fluctuations
Dysphagia
Dyskinesias
Axial deformities
Falls
REM sleep behaviour disorder
Depression
Anxiety
Constipation
Hyposmia
Excessive daytime sleepiness
Mild cognitive impairment
Orthostatic hypotension
Fatigue
Apathy
Urinary symptoms
Dementia
Pain
Psychotic symptoms (visual hallucinations)
Motor symptoms
Non-motor symptoms
Time (years)

[Poewe et al., Parkinson's disease. *Nature Reviews Disease Primers*, 2017]

Parkinson’s Progression Marker Initiative (from Lecture 20)



OUR MISSION

The mission of PPMI is to identify one or more biomarkers of Parkinson’s disease progression. The discovery of a biomarker is a critical step in the development of new and better treatments for PD. This study is being sponsored by **The Michael J. Fox Foundation for Parkinson's Research.**

Subscribe to PPMI e-news

Enter email address.

PPMI ENROLLMENT STATUS

100%

Genetic Cohort Participants
Enrollment Goal: 600

79%

Genetic Registry Participants
Enrollment Goal: ~600

100%

De Novo PD Participants
Enrollment: 424

A LANDMARK STUDY OF PARKINSON’S DISEASE

The Parkinson’s Progression Markers Initiative (PPMI) is a landmark observational clinical study to comprehensively evaluate **cohorts of significant interest** using advanced imaging, biologic sampling and clinical and behavioral assessments to identify biomarkers of Parkinson’s disease progression.

PPMI is taking place at **clinical sites** in the United States, Europe, Israel, and Australia. Data and samples acquired from study participants will enable the development of a comprehensive Parkinson’s **database and biorepository**, which is currently available to the scientific community to conduct field-changing research.

PPMI is made possible by the concerted efforts of a number of collaborators. This study is sponsored by The Michael J. Fox Foundation for Parkinson’s Research.

Learn more about **Who We Are**.

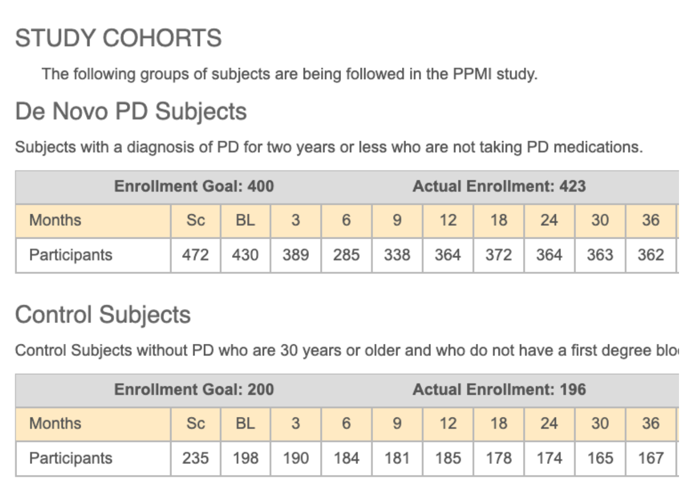
LATEST NEWS FROM PPMI

Interview on RNA in PPMI with Dr. Van Keuren-Jensen
Today, we will be asking Dr. Van-Keuren Jensen a few questions about RNA and hearing her insight as to why this vital element is something, we are focusing on in the Parkinson’s Progression Markers Initiative (PPMI). Dr. Van Keuren-Jensen received her Ph.D. from Cold Spring Harbor Laboratory in New York, where she studied the role […]

Recent Findings in PPMI
by Krishna Knabe The Michael J. Fox Foundation A group of authors led by PPMI's principal investigator, Dr. Kenneth Marek, published baseline data from the study in the Annals of Clinical and Translational Neurology. The paper includes detailed biomarker signatures on the initial volunteer groups, which include patients with Parkinson’s disease, healthy controls and those who […]

PPMI Study Enters a New Phase
by Krishna Knabe The Michael J. Fox Foundation The Parkinson’s Progression Markers Initiative (PPMI) has reached an important milestone: the study completed enrollment. We have now met the ambitious goal we set back in 2010 of enrolling 1,400 participants, including 600 with rare genetic mutations. PPMI is The Michael J. Fox Foundation’s landmark observational clinical […]

PPMI Data Available: RNA Seq



STUDY COHORTS

The following groups of subjects are being followed in the PPMI study.

De Novo PD Subjects

Subjects with a diagnosis of PD for two years or less who are not taking PD medications.

Enrollment Goal: 400						Actual Enrollment: 423					
Months	Sc	BL	3	6	9	12	18	24	30	36
Participants	472	430	389	285	338	364	372	364	363	362

Control Subjects

Control Subjects without PD who are 30 years or older and who do not have a first degree blo

Enrollment Goal: 200						Actual Enrollment: 196					
Months	Sc	BL	3	6	9	12	18	24	30	36
Participants	235	198	190	184	181	185	178	174	165	167

ppmi-info.org

Parkinson's Progression Marker Initiative (from Lecture 20)

Questionnaires

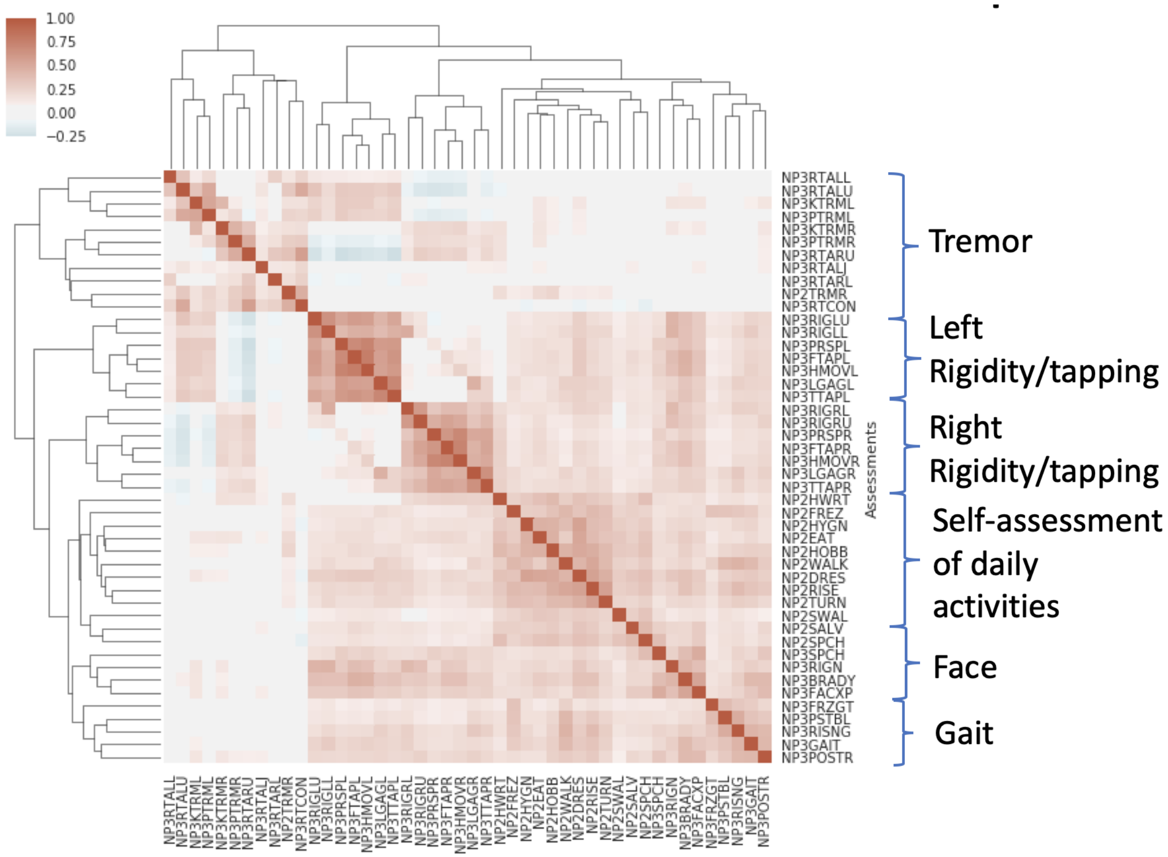
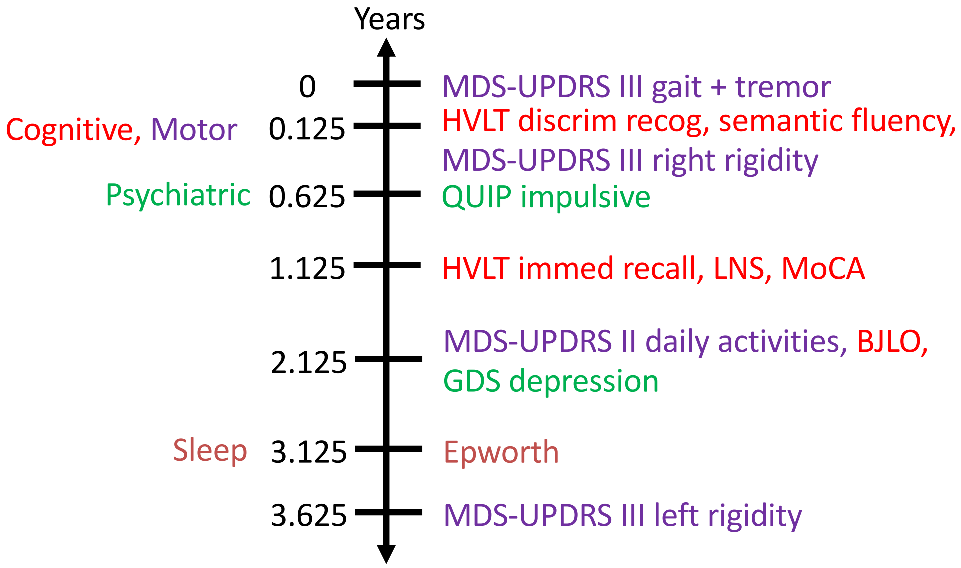


Figure 2-2: **Correlation heatmap of MDS-UPDRS questions** with subtotal annotations on the right.

[Figures from Christina Ji's Master's thesis]

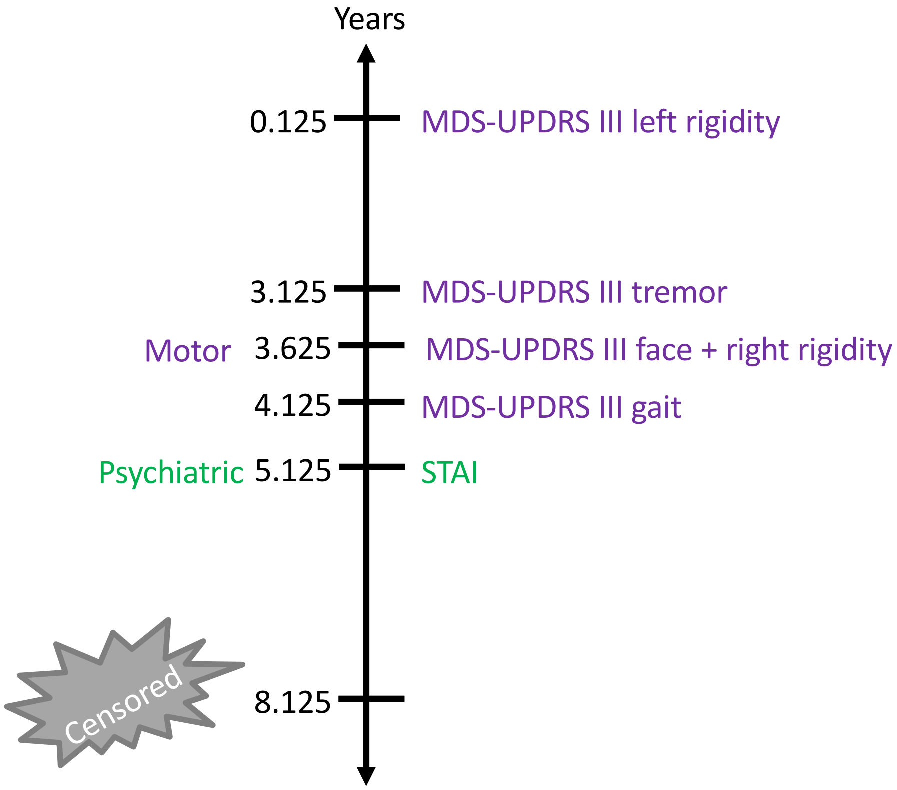
Parkinson’s Progression Marker Initiative (from Lecture 20)

Sample timeline 1



	Years	
	0	MDS-UPDRS III gait + tremor
Cognitive, Motor	0.125	HVLT discrim recog, semantic fluency, MDS-UPDRS III right rigidity
Psychiatric	0.625	QUIP impulsive
	1.125	HVLT immed recall, LNS, MoCA
	2.125	MDS-UPDRS II daily activities, BJLO, GDS depression
Sleep	3.125	Epworth
	3.625	MDS-UPDRS III left rigidity

Sample timeline 2



	Years	
	0.125	MDS-UPDRS III left rigidity
	3.125	MDS-UPDRS III tremor
Motor	3.625	MDS-UPDRS III face + right rigidity
	4.125	MDS-UPDRS III gait
Psychiatric	5.125	STAI
Censored	8.125	

[Figures from Christina Ji’s Master’s thesis]

Parkinson’s Progression Marker Initiative (from Lecture 20)

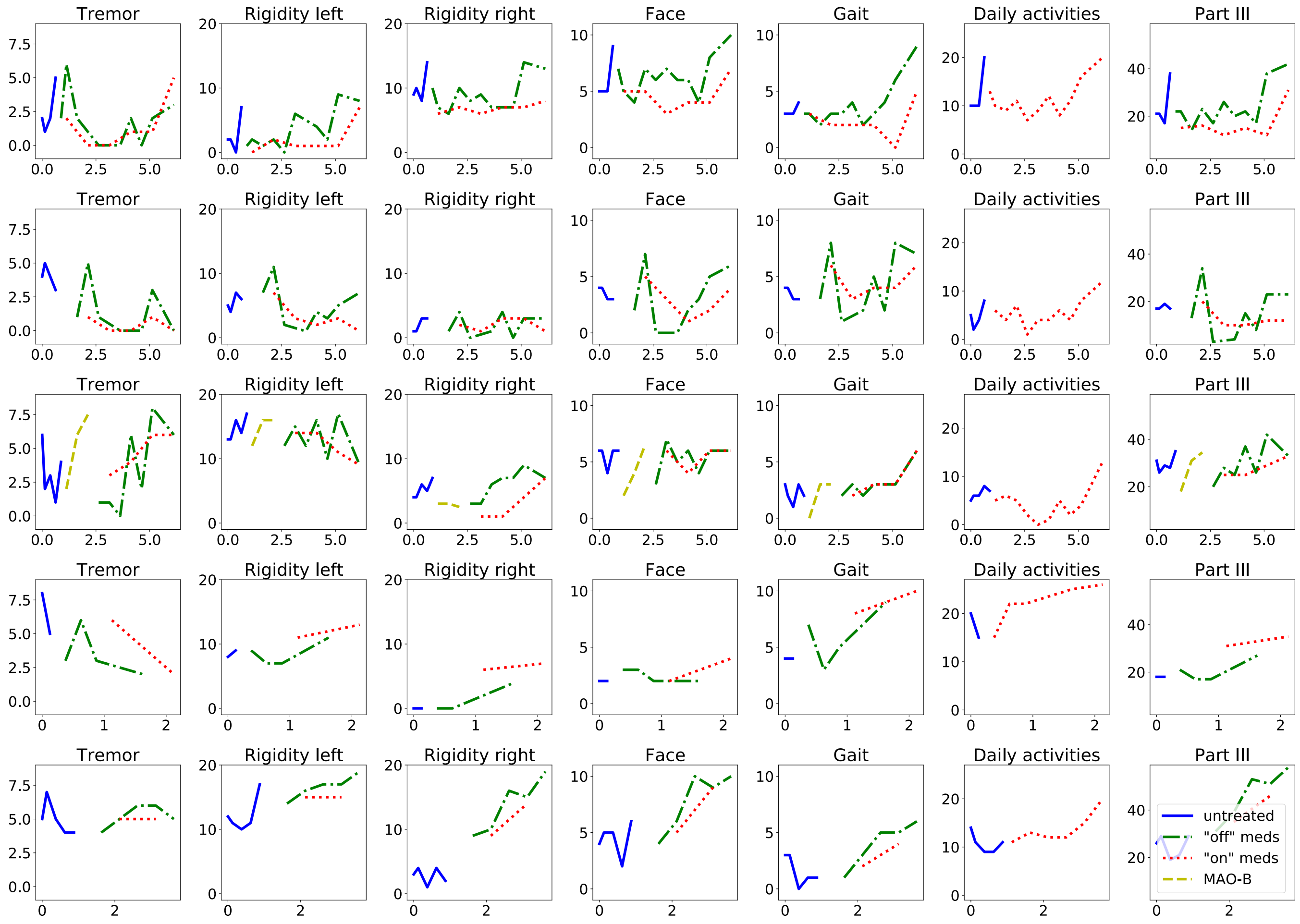
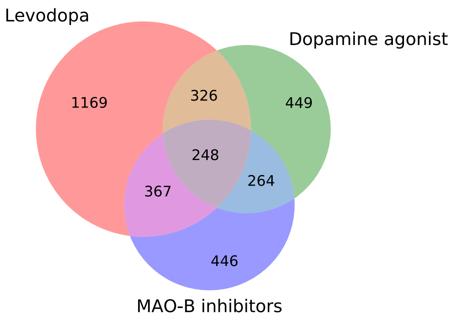


Figure 2-3: **MDS-UPDRS subtotals for 5 PD patients**

Treatment



[Figures from Christina Ji’s Master’s thesis]

Parkinson’s Progression Marker Initiative (from Lecture 20)

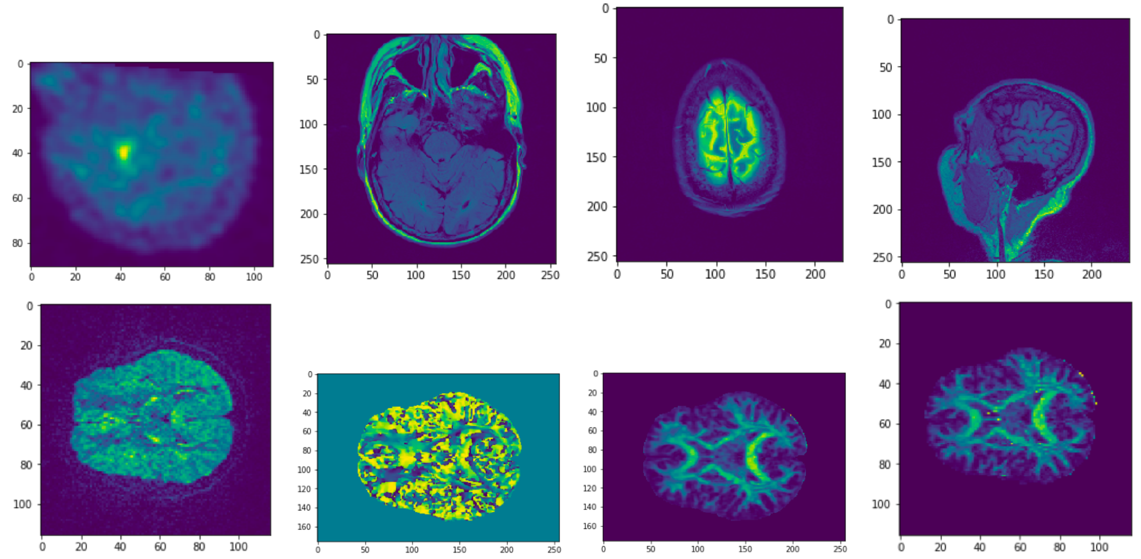


Figure 2-10: Examples of **imaging modalities**. Left to right: Top row: DaTscan, MRI axial fluid-attenuated inversion recovery, MRI axial turbo spin echo, MRI saggital magnetization-prepared rapid gradient echo. Bottom row: DTI 4-d motion trajectory, DTI eigenvectors of MRI, DTI fractional anistrophy of MRI, DTI fractional anistrophy of EPI.

Multi-modal data

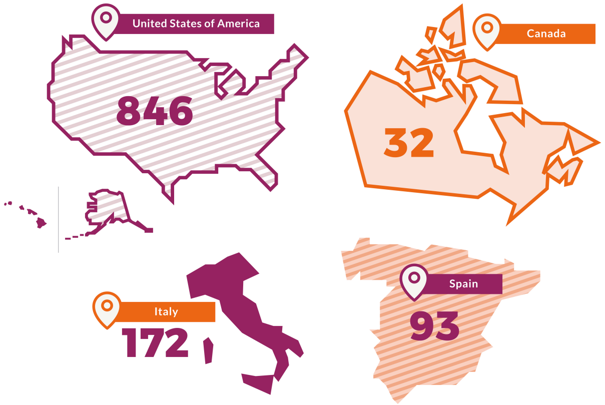
Here, e.g., including imaging

[Figures from Christina Ji’s Master’s thesis]

Multiple myeloma: MMRF CoMMpass

Study population

1150 patients from **90** sites worldwide



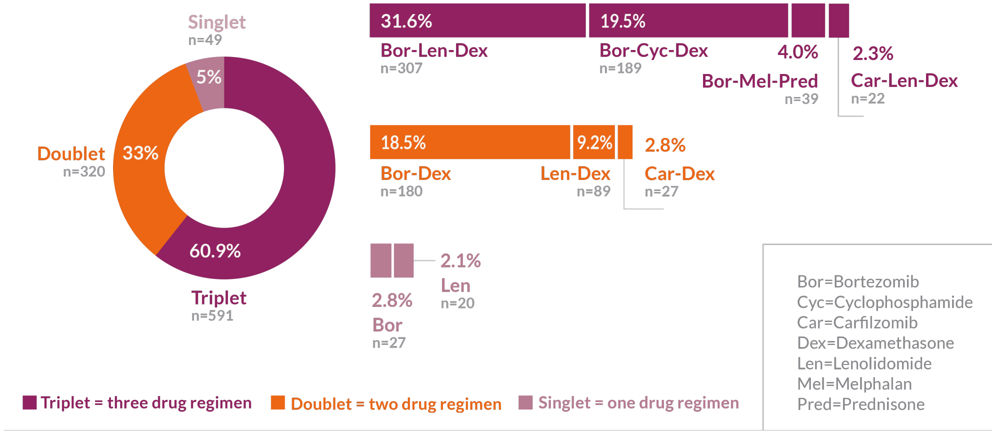
Bone marrow samples were taken

At the start of the study
At response to treatment
At relapse

Each patient was checked on

Every 6 months
For 8 years

First treatment line

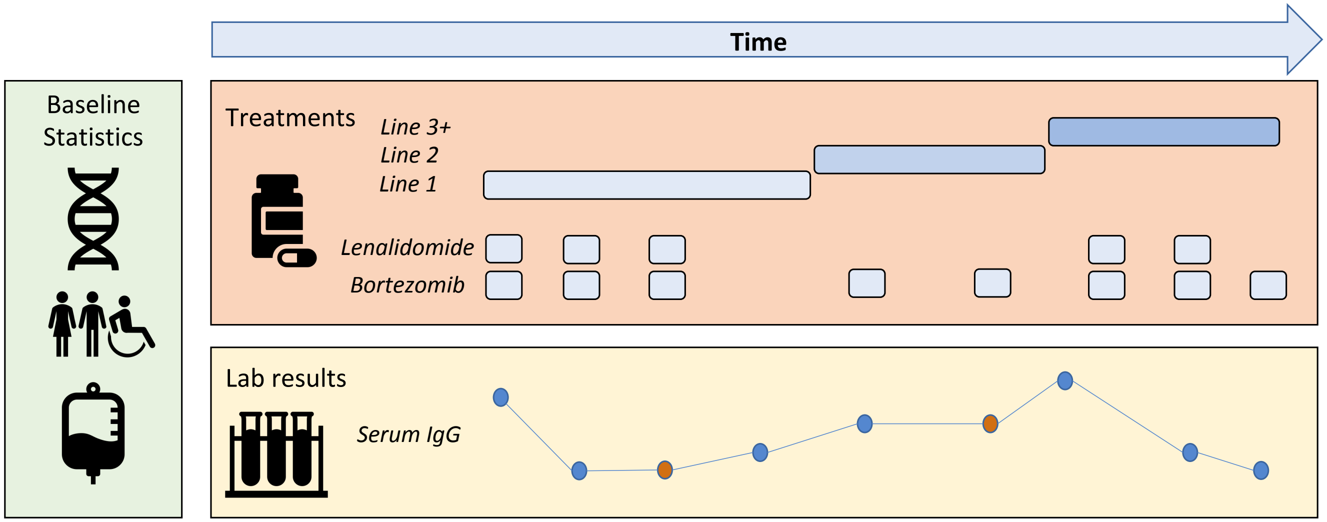


Bor=Bortezomib
Cyc=Cyclophosphamide
Car=Carfilzomib
Dex=Dexamethasone
Len=Lenolidomide
Mel=Melphalan
Pred=Prednisone

https://themmrf.org/we-are-curing-multiple-myeloma/mmrf-commpass-study/

Multiple myeloma: MMRF CoMMpass

Baseline data includes RNA-seq, copy number variations, and gene mutations



At each time step (~3 month intervals), observe blood test results:

- Immunoglobulins and antibodies (IgG, IgA, IgM, kappa chains, light chains)
- M-protein, creatinine, neutrophil count, hemoglobin, platelet count, etc.

Several of these are less frequently measured, so many missing values

[Figures from Rahul Krishnan and Zeshan Hussain]

Summary of challenges

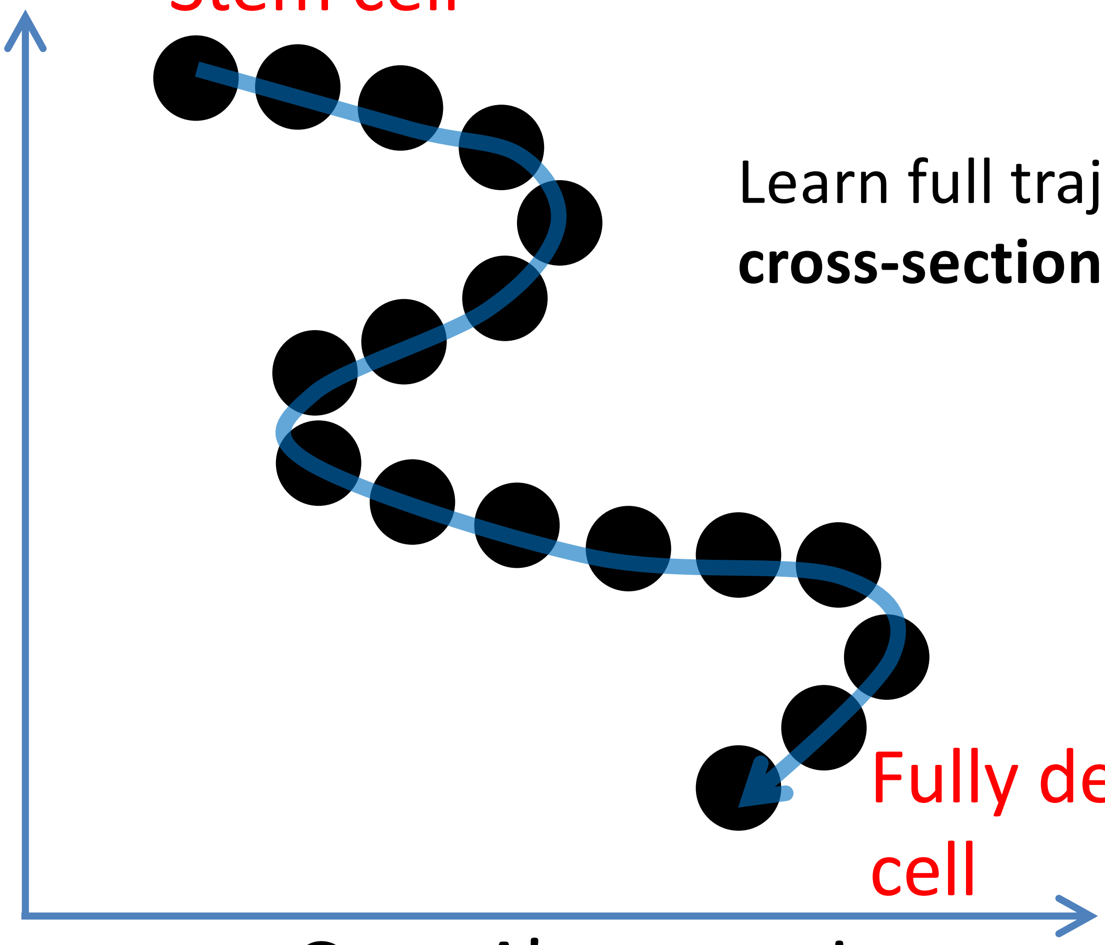
- Censored data – patients come in at various stages of disease progression, and leave studies early
- Irregular time intervals between observations, lots of missing data (potentially biased by healthcare processes)
- Multi-modal data (labs, symptoms, imaging, genomics)
- Limited supervision

Outline of today’s lecture

- Deep dive into data commonly used for disease progression modeling
- **What can we draw inspiration from, and why they are not good enough**
- Probabilistic models of disease progression
- Simultaneous staging & subtyping

Learning “pseudo-time” for single-cell sequencing

Stem cell



Gene B’s expression

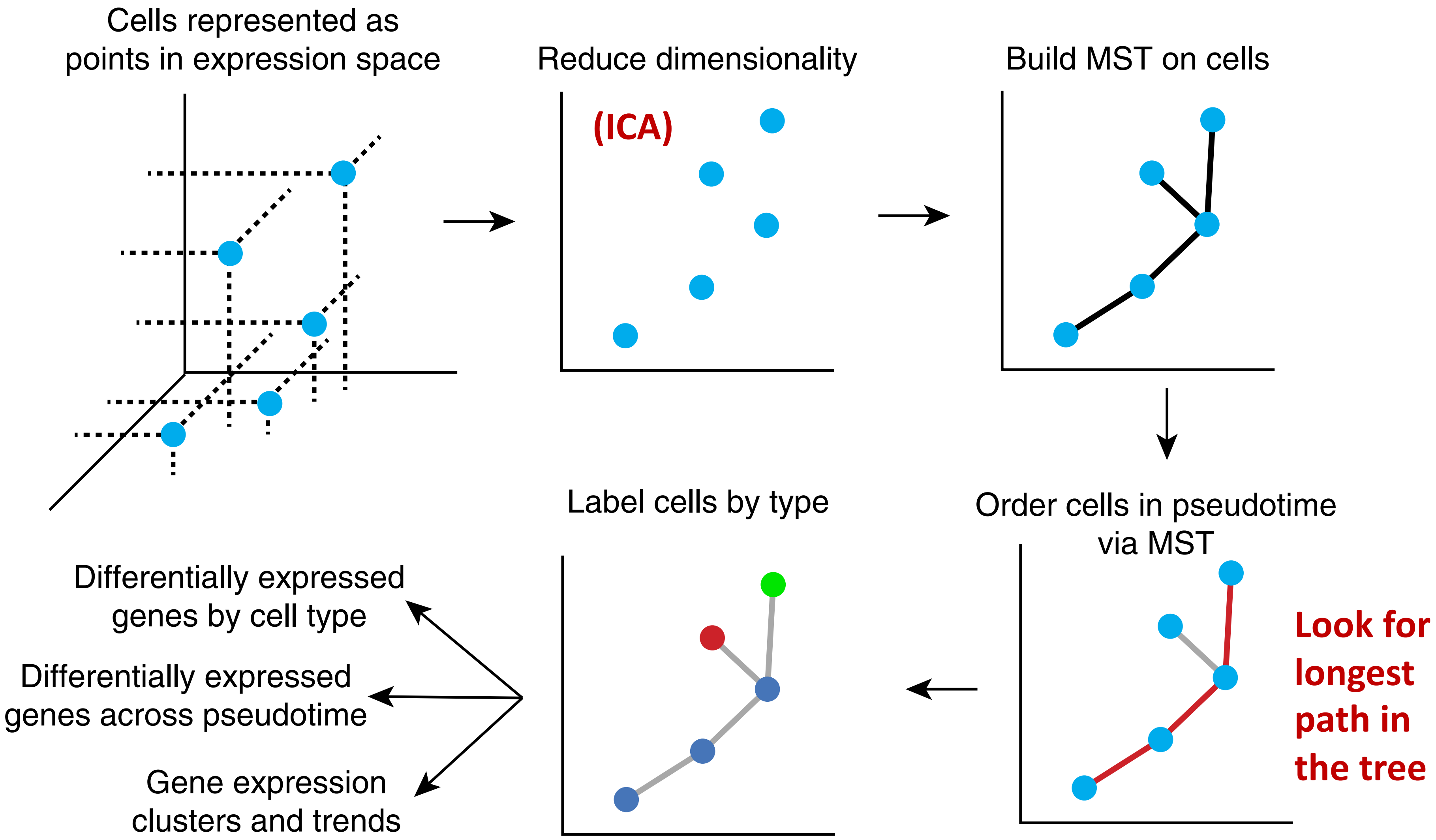
Gene A’s expression

Fully developed B cell

Learn full trajectory from **cross-sectional** data!

[Bendall et al., Cell 2014 (human B cell development)]

Learning “pseudo-time” for single-cell sequencing



Cells represented as points in expression space

Reduce dimensionality

(ICA)

Build MST on cells

Order cells in pseudotime via MST

Look for longest path in the tree

Label cells by type

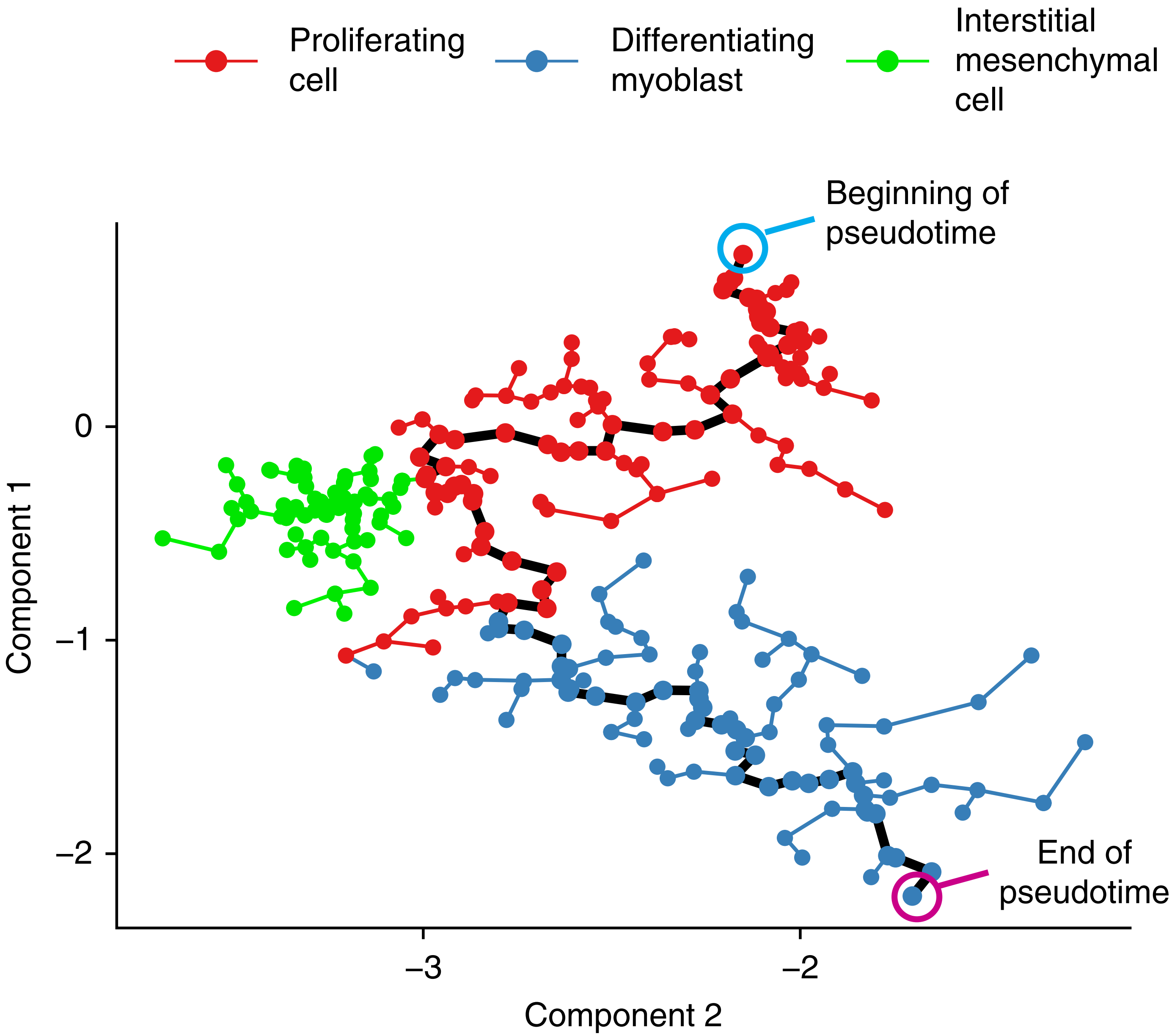
Differentially expressed genes by cell type

Differentially expressed genes across pseudotime

Gene expression clusters and trends

[Magwene et al., *Bioinformatics*, 2003; Trapnell et al., *Nature Biotechnology*, 2014]

MST-based approach (Monocle)



[Trapnell et al., *Nature Biotechnology*, 2014]

RNN language models

- Could use a recurrent neural network as an autoregressive model of the distribution of observations:

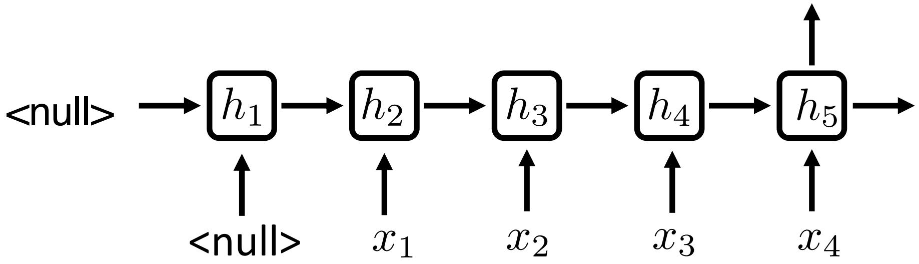
# of time steps

$$\Pr(x_1, x_2, \ldots, x_T) = \Pr(x_1) \prod_{t=2}^{T} \Pr(x_t \mid x_1, \ldots x_{t-1})$$

Labs, symptoms, etc. observed at time 2

- Observations up to time t-1 summarized by RNN's hidden state h$_t$:

$$p_5 = p(X_5 \mid X_1, \ldots, X_4) = p(X_5 \mid h_5)$$



Why these are insufficient for disease progression modeling

- Limitations of (most) pseudo-time methods
  - Good that these handle censored data, but we often have *multiple* observations
  - Needs *lots* of data, but most disease data sets are small (e.g. hundreds of patients)
  - Needs *simple* manifolds embedded in high-dimensions; disease data sets features often low dimensional
- Limitations of (naively) using recurrent neural networks to model the sequence of observations
  - Irregular time intervals between observations*
  - Missing data
  - Must model treatment effects
  - Multi-modal data

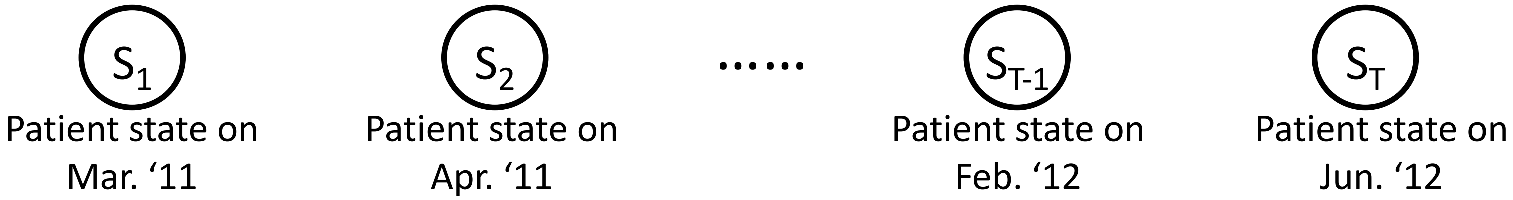
*See Che et al., Recurrent Neural Networks for Multivariate Time Series with Missing Values, Scientific Reports ‘18

Outline of today's lecture

- Deep dive into data commonly used for disease progression modeling
- What can we draw inspiration from, and why they are not good enough
- **Probabilistic models of disease progression**
- Simultaneous staging & subtyping

Key idea: model patient state as a latent variable

- Use a Markov model to describe the joint distribution of patient states over time:



S_1 Patient state on Mar. ‘11

S_2 Patient state on Apr. ‘11

……

$S_{T\text{-}1}$ Patient state on Feb. ‘12

S_T Patient state on Jun. ‘12

- State space of S could be discrete (e.g. take K states) or continuous (e.g. in R^d) – analogous to hidden state of the RNN
- If *regular* time intervals, we model the transition distribution Pr(S_t | $S_{t\text{-}1}$)
- Otherwise, model $P(S_t \mid S_{t-1}, \tau_t - \tau_{t-1} = \Delta)$
- Alternatively, use a Gaussian process or neural ODE to model the joint distribution of *S*[*]

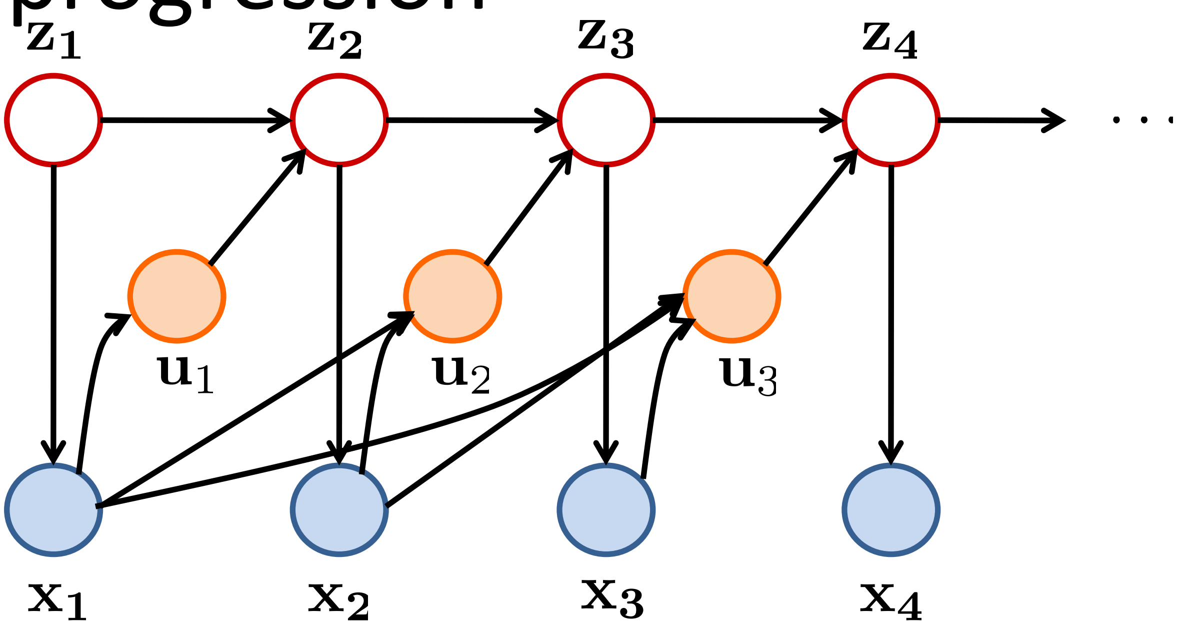
[*] See Schulam & Saria, Reliable Decision Support using Counterfactual Models, NeurIPS 2017 & Chen et al., Neural Ordinary Differential Equations, NeurIPS 2018

Deep Markov models (DMMs) of disease progression

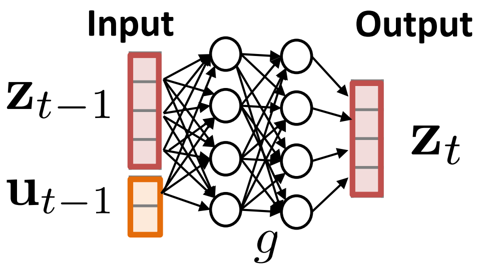
Patient state $\mathbf{z}_t \in \mathbb{R}^{100}$

Actions $\mathbf{u}$
(e.g., medication, surgery)

Observations $\mathbf{x}$
(blood and urine test results, diagnoses, vital signs, …)



- Provides an in-silico model for assessing effect of interventions (actions), by forward sampling in model
- Transition & emission distributions given by deep neural networks:



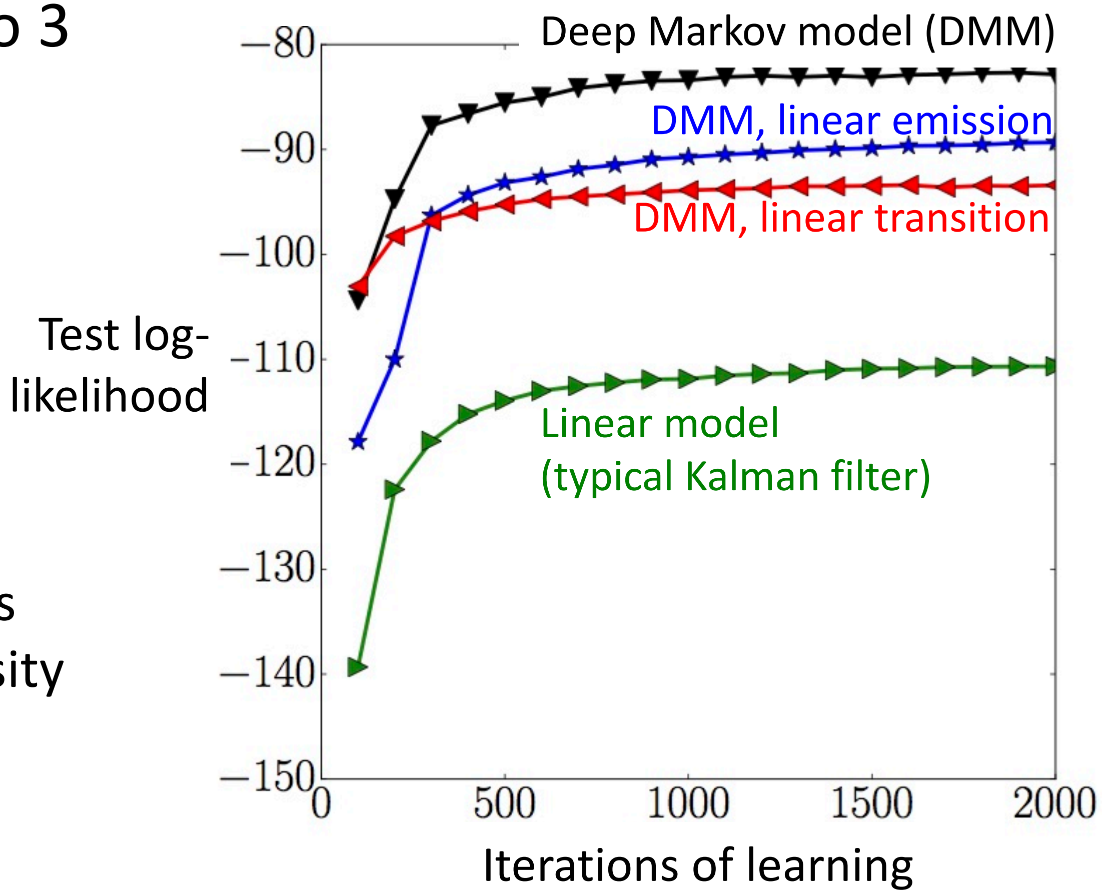
$$\mathbf{z}_t \sim \mathcal{N}(g(\mathbf{z}_{t-1}, \mathbf{u}_{t-1}), s(\mathbf{z}_{t-1}, \mathbf{u}_{t-1}))$$

[Krishnan, Shalit, Sontag, AAAI ‘17]

Progression modeling for diabetes

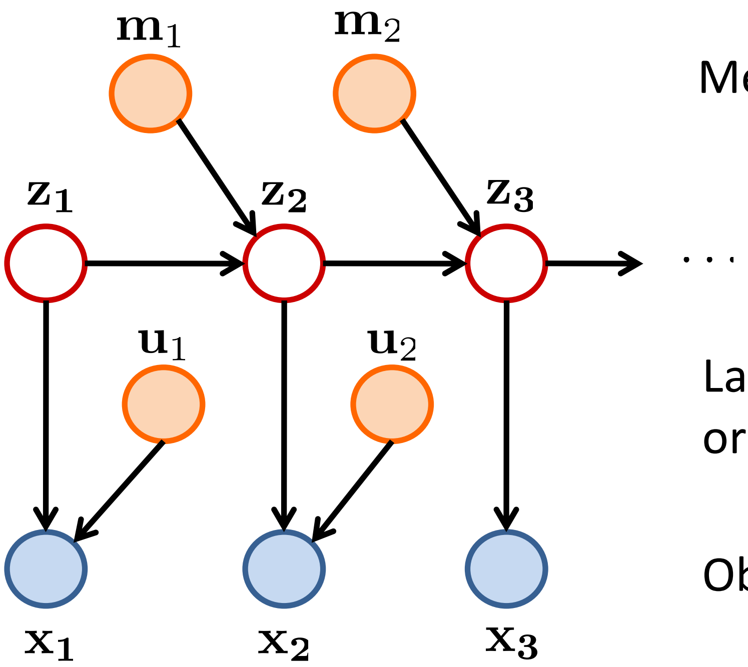
- 8000 diabetic and pre-diabetic patients
- 4 years of data, grouped into 3 month intervals
- **Observations:** 52 binary variables measuring
  - Demographics
  - Laboratory test results (e.g glucose level)
  - Diagnosis codes for conditions such as heart failure and obesity
- 200 latent dimensions for $\mathbf{z}_t$

The non-linearity given by the deep neural networks significantly improves ability to model the data



Learning the effect of diabetic treatments

- Long-term: which diabetes medications work best for whom?
- **Actions:** 9 diabetic drugs including Metformin and Insulin (**m**), lab test orders (**u**)



Medications

Lab test orders

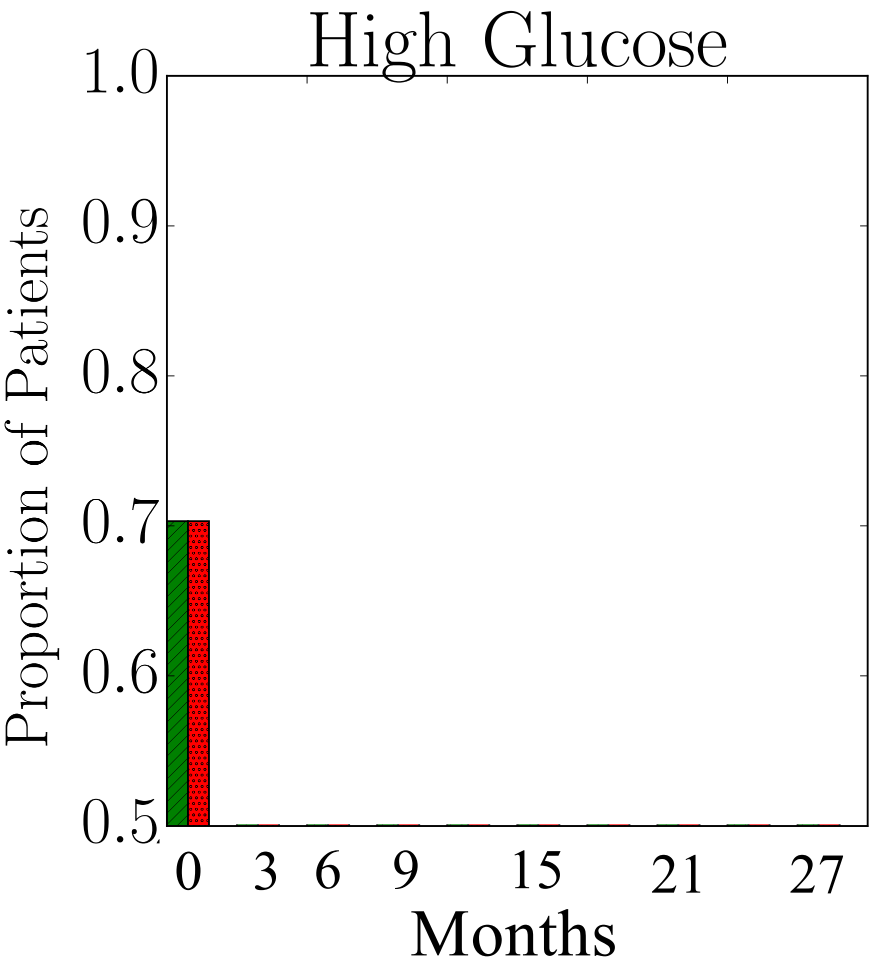
Observations

- *Here we just do a sanity check*

Effect of diabetes treatments on glucose

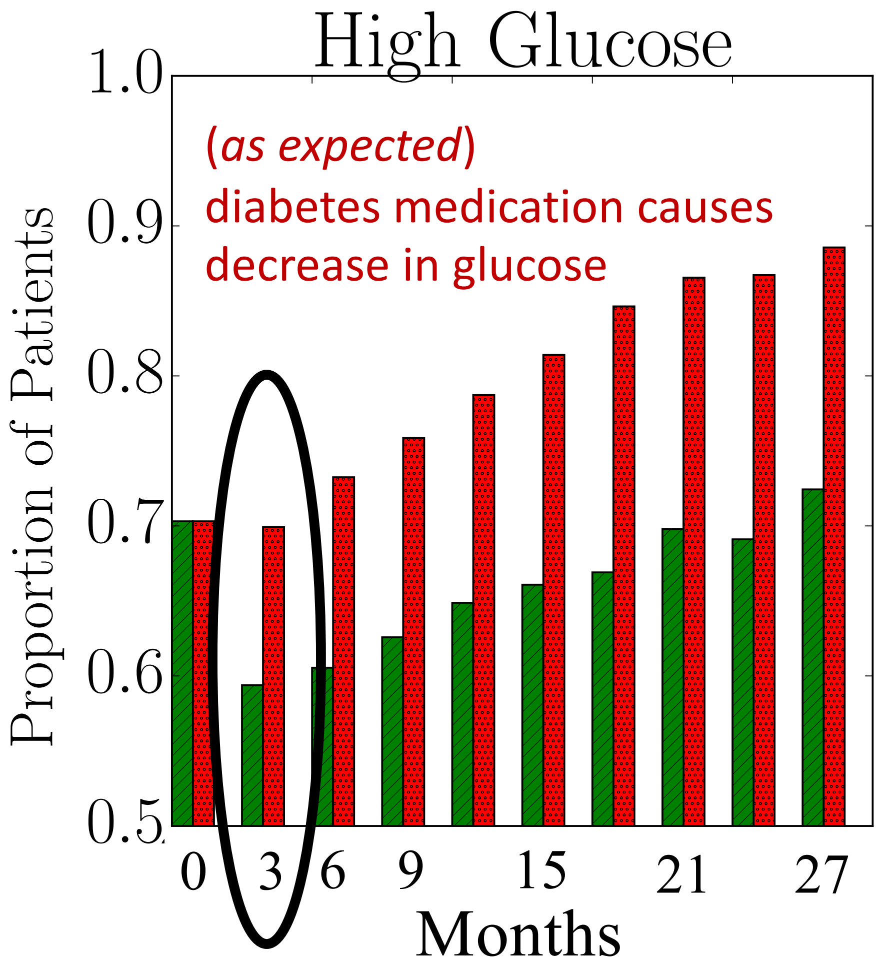
🟩 w/ medication 🟥 w/out medication

1. Align patients by when they were first prescribed Metformin
2. Sample future patient data ***using the medications they truly received***
3. Sample future patient data ***as if they never received medication***



Effect of diabetes treatments on glucose

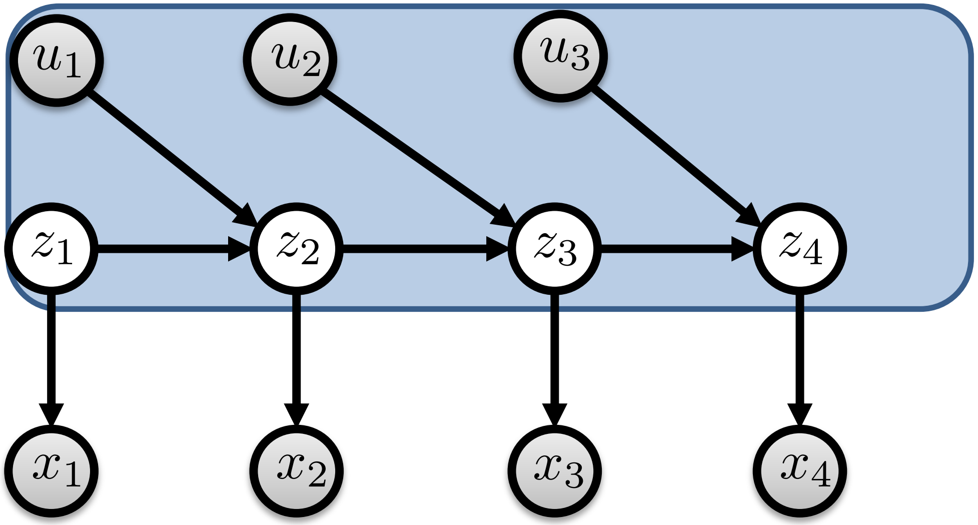
1. Align patients by when they were first prescribed Metformin
2. Sample future patient data ***using the medications they truly received***
3. Sample future patient data ***as if they never received medication***



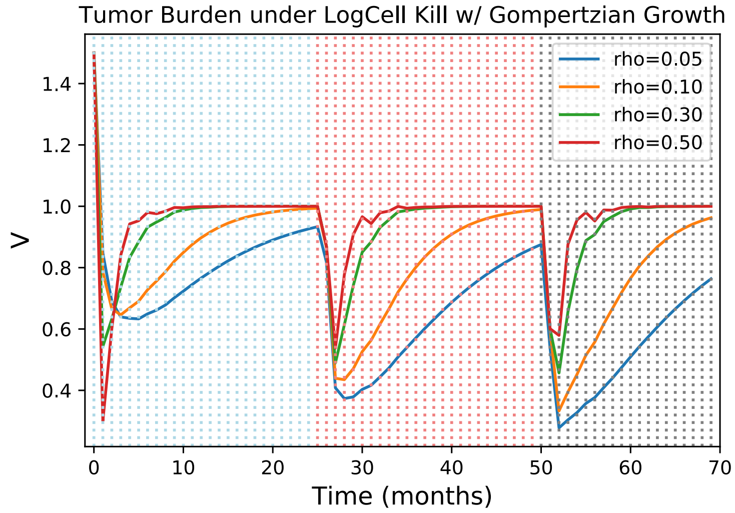
(*as expected*)
diabetes medication causes decrease in glucose

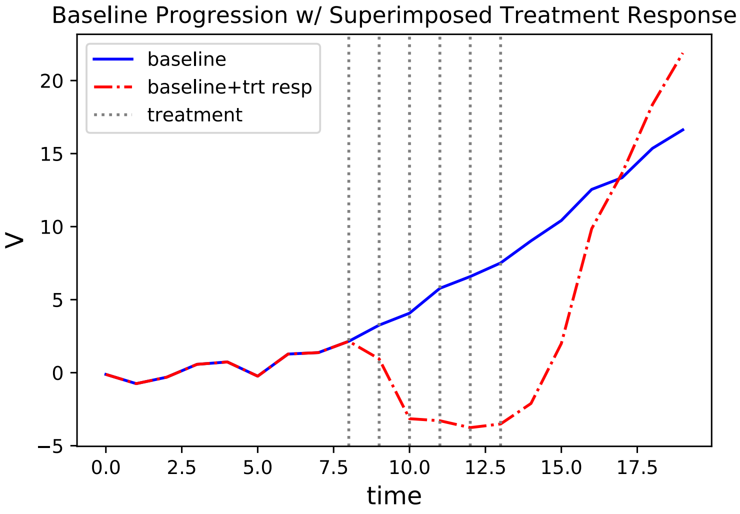
Inductive Biases for Treatment effect

$$p(z_t | z_{t-1}, u_{t-1}; \theta)$$



$$\begin{aligned}
\text{lin}_t &= Z_t \odot \tanh(W_n \cdot [U_t; B] + b_n) \\
\text{logcell}_t &= \text{LC}(Z_t, U_t, t - t_s) \\
\text{te}_t &= E(t - t_s; \alpha_{1t}, \alpha_{2t}, \alpha_{3t}, \gamma_t, b_0, b_l) \\
o_t &= \sigma(\boldsymbol{\delta})_1 \odot \text{lin}_t + \sigma(\boldsymbol{\delta})_2 \odot \text{logcell}_t \\
&\quad + \sigma(\boldsymbol{\delta})_3 \odot \text{te}_t \\
\mu_\theta(Z_t, U_t, B) &= (W_r \cdot Z_t + b_r) + o_t
\end{aligned}$$

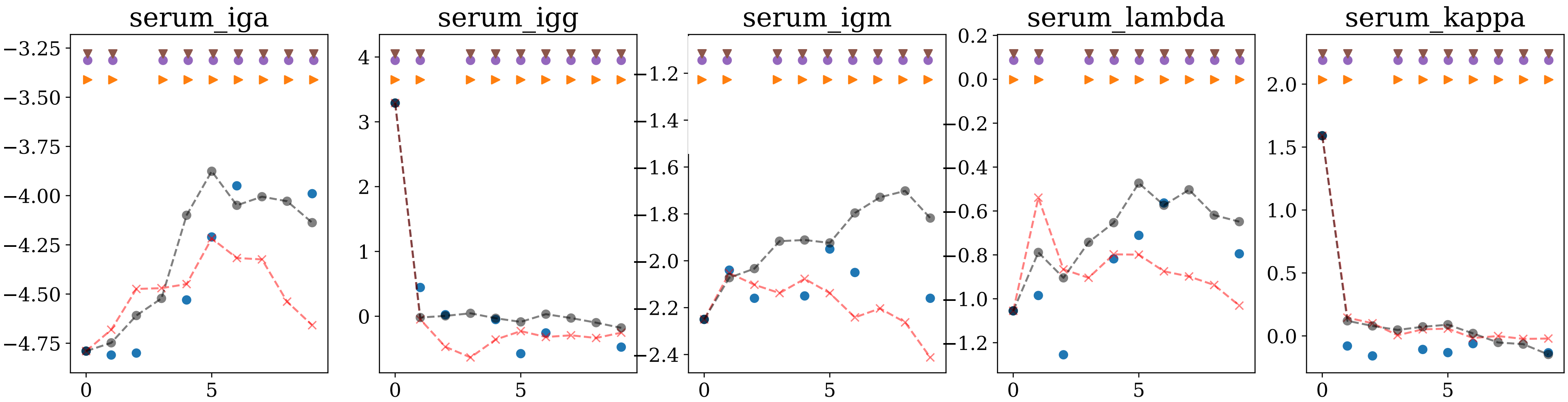




[Recent work by Rahul Krishnan and Zeshan Hussain]

Inductive Biases for Treatment effect

PK/PD DMM better at forecasting patient biomarkers

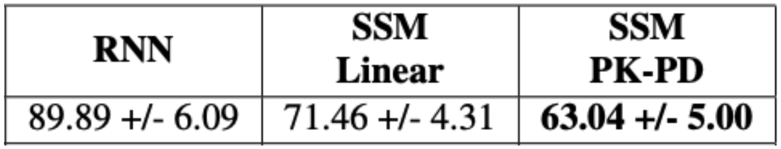


Ground truth

PK/PD DMM

Linear State Space Model

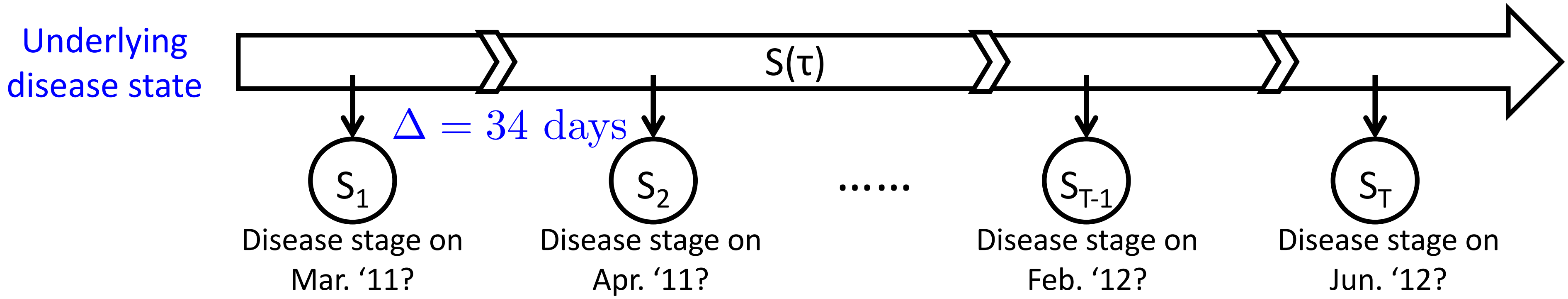
Held-out likelihood:



RNN	SSM Linear	SSM PK-PD
89.89 +/- 6.09	71.46 +/- 4.31	**63.04 +/- 5.00**

[Recent work by Rahul Krishnan and Zeshan Hussain]

Alternative approach: continuous-time Markov model


Underlying disease state
S(τ)
Δ = 34 days
S1
S2
……
ST-1
ST
Disease stage on Mar. '11?
Disease stage on Apr. '11?
Disease stage on Feb. '12?
Disease stage on Jun. '12?

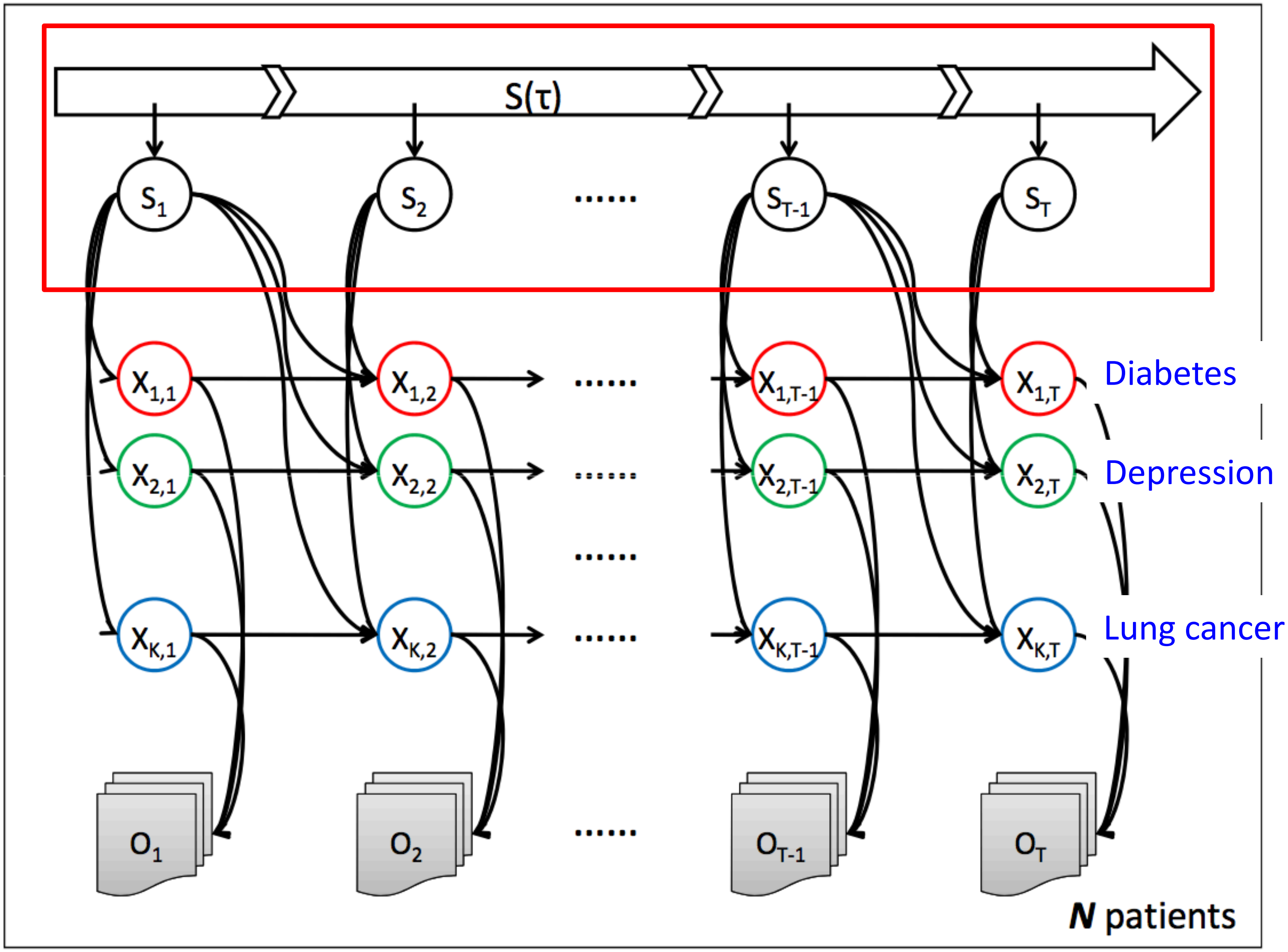
- A continuous-time Markov process with irregular discrete-time observations
- The transition probability is defined by an intensity matrix and the time interval:

$$\begin{aligned} A_{ij}(\Delta) &\triangleq P(S_t = j | S_{t-1} = i, \tau_t - \tau_{t-1} = \Delta; Q) \\ &= \mathrm{expm}(\Delta Q)_{ij}, \end{aligned}$$

Matrix Q: Parameters to learn

[Wang, Sontag, Wang, "Unsupervised learning of Disease Progression Models", KDD 2014]

Generative model for patient data



Markov Jump Process

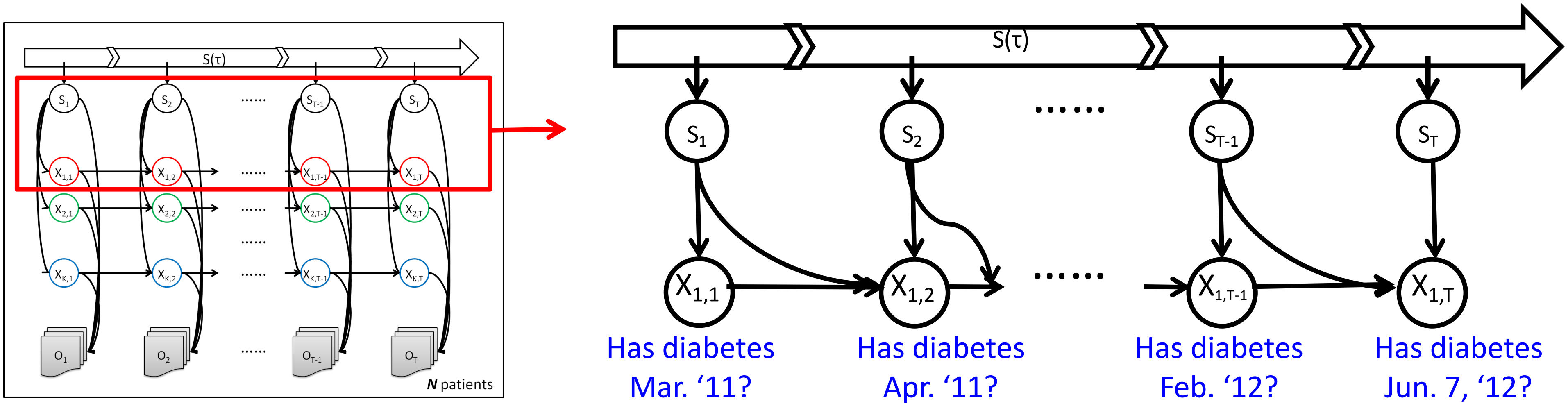
Progression Stages

K phenotypes, each with its own Markov chain

Observations

[Wang, Sontag, Wang, “Unsupervised learning of Disease Progression Models”, KDD 2014]

Model of comorbidities across time



Has diabetes Mar. ‘11?

Has diabetes Apr. ‘11?

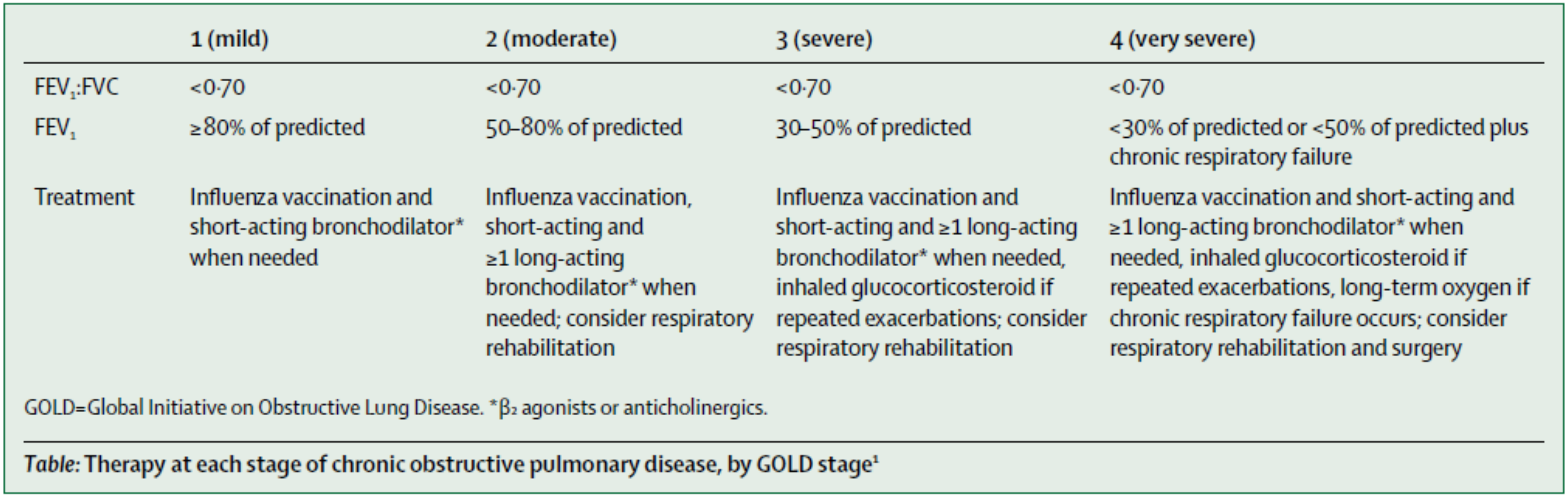
Has diabetes Feb. ‘12?

Has diabetes Jun. 7, ‘12?

- Presence of comorbidities depends on value at previous time step and on disease stage
- Later stages of disease = more likely to develop comorbidities
- Make the assumption that once patient has a comorbidity, likely to always have it

COPD diagnosis & progression

- COPD diagnosis made using a breath test – fraction of air expelled in first second of exhalation < 70%
- Most doctors use GOLD criteria to stage the disease and measure its progression:



	1 (mild)	2 (moderate)	3 (severe)	4 (very severe)
FEV_1:FVC	<0·70	<0·70	<0·70	<0·70
FEV_1	≥80% of predicted	50–80% of predicted	30–50% of predicted	<30% of predicted or <50% of predicted plus chronic respiratory failure
Treatment	Influenza vaccination and short-acting bronchodilator* when needed	Influenza vaccination, short-acting and ≥1 long-acting bronchodilator* when needed; consider respiratory rehabilitation	Influenza vaccination and short-acting and ≥1 long-acting bronchodilator* when needed, inhaled glucocorticosteroid if repeated exacerbations; consider respiratory rehabilitation	Influenza vaccination and short-acting and ≥1 long-acting bronchodilator* when needed, inhaled glucocorticosteroid if repeated exacerbations, long-term oxygen if chronic respiratory failure occurs; consider respiratory rehabilitation and surgery

GOLD=Global Initiative on Obstructive Lung Disease. *β_2 agonists or anticholinergics.

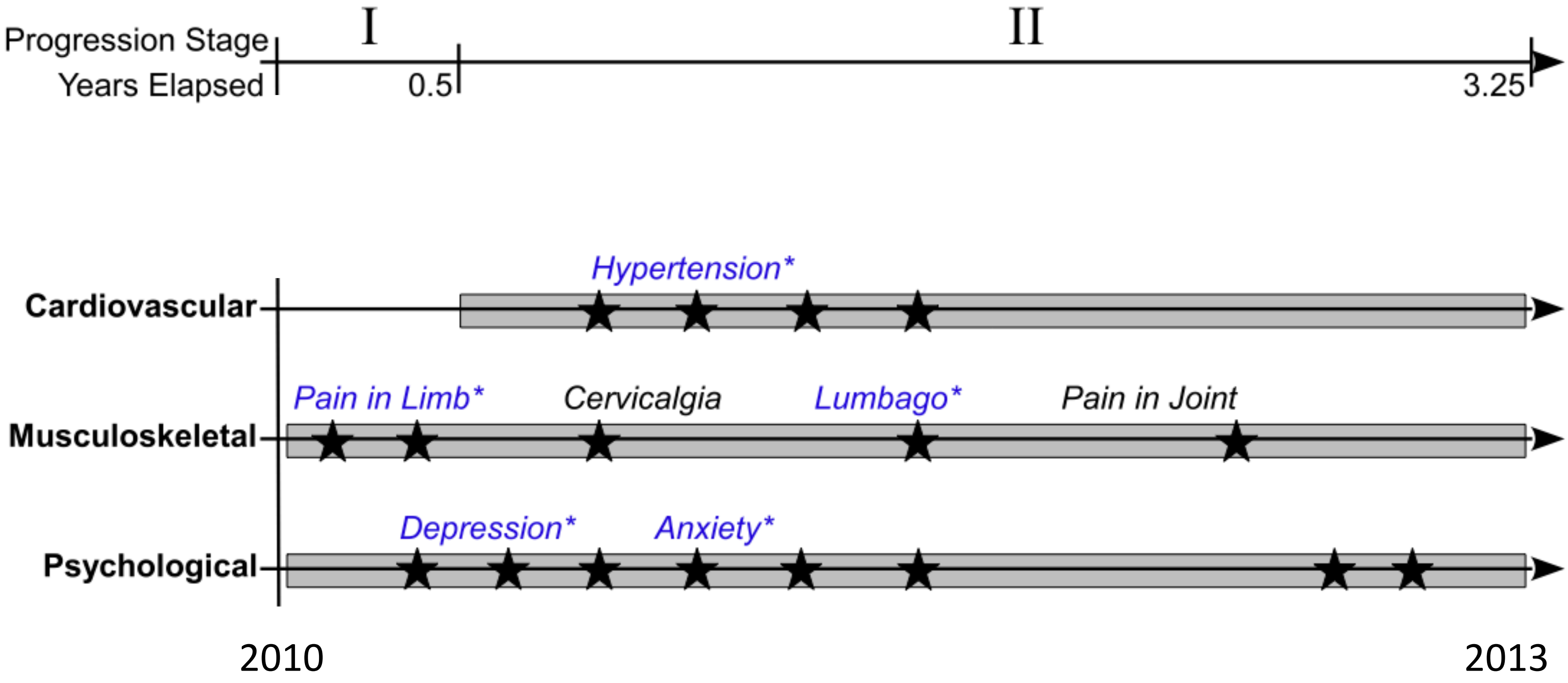
Table: **Therapy at each stage of chronic obstructive pulmonary disease, by GOLD stage**[1]

Chronic obstructive pulmonary disease. The Lancet, Volume 379, Issue 9823, Pages 1341 - 1351, 7 April 2012

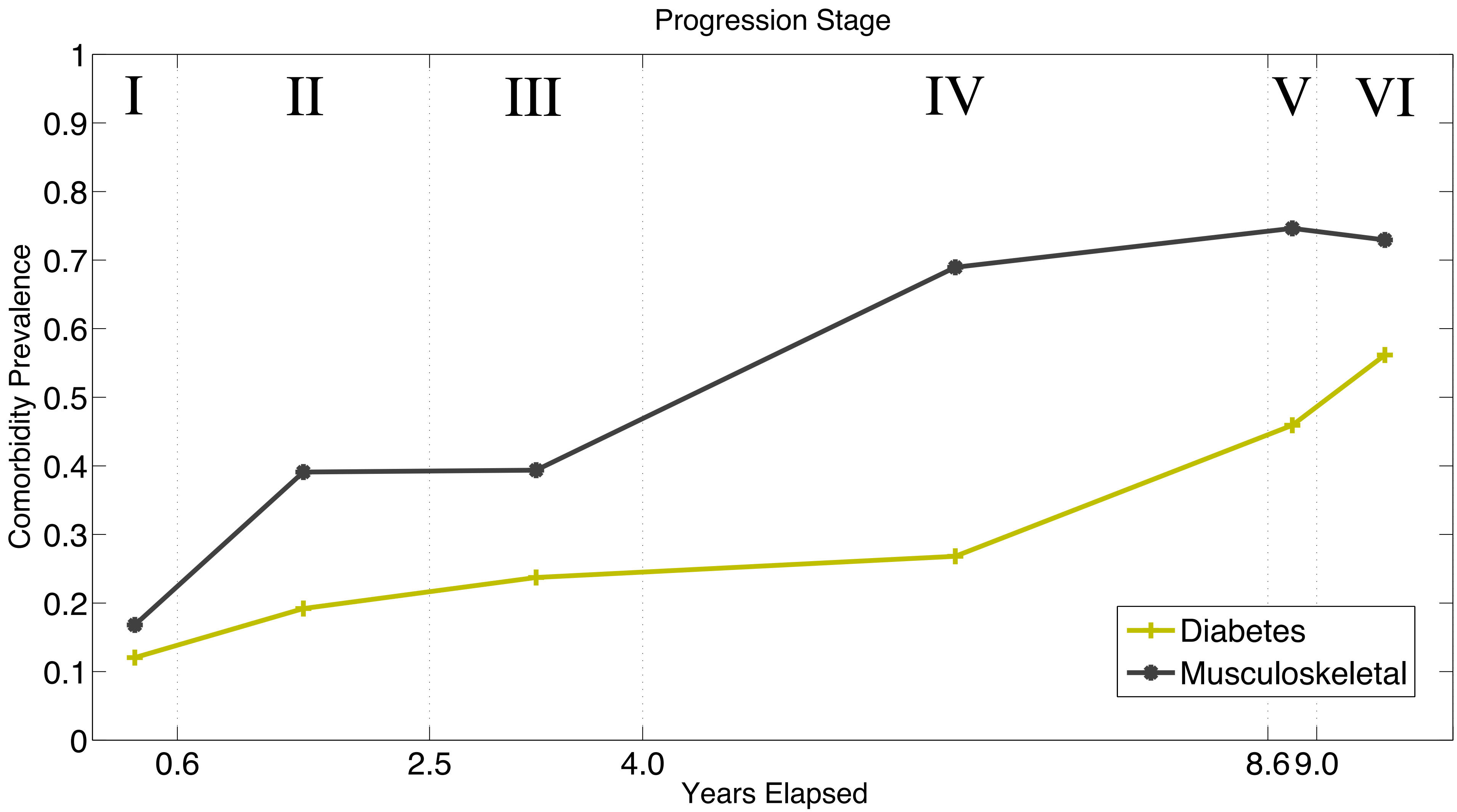
Experimental evaluation

- We create a COPD cohort of 3,705 patients:
  - At least one COPD-related diagnosis code
  - At least one COPD-related drug
- Removed patients with too few records
- Clinical findings derived from 264 diagnosis codes
  - Removed ICD-9 codes that only occurred to a small number of patients
- Combined visits into 3-month time windows
- 34,976 visits, 189,815 positive findings

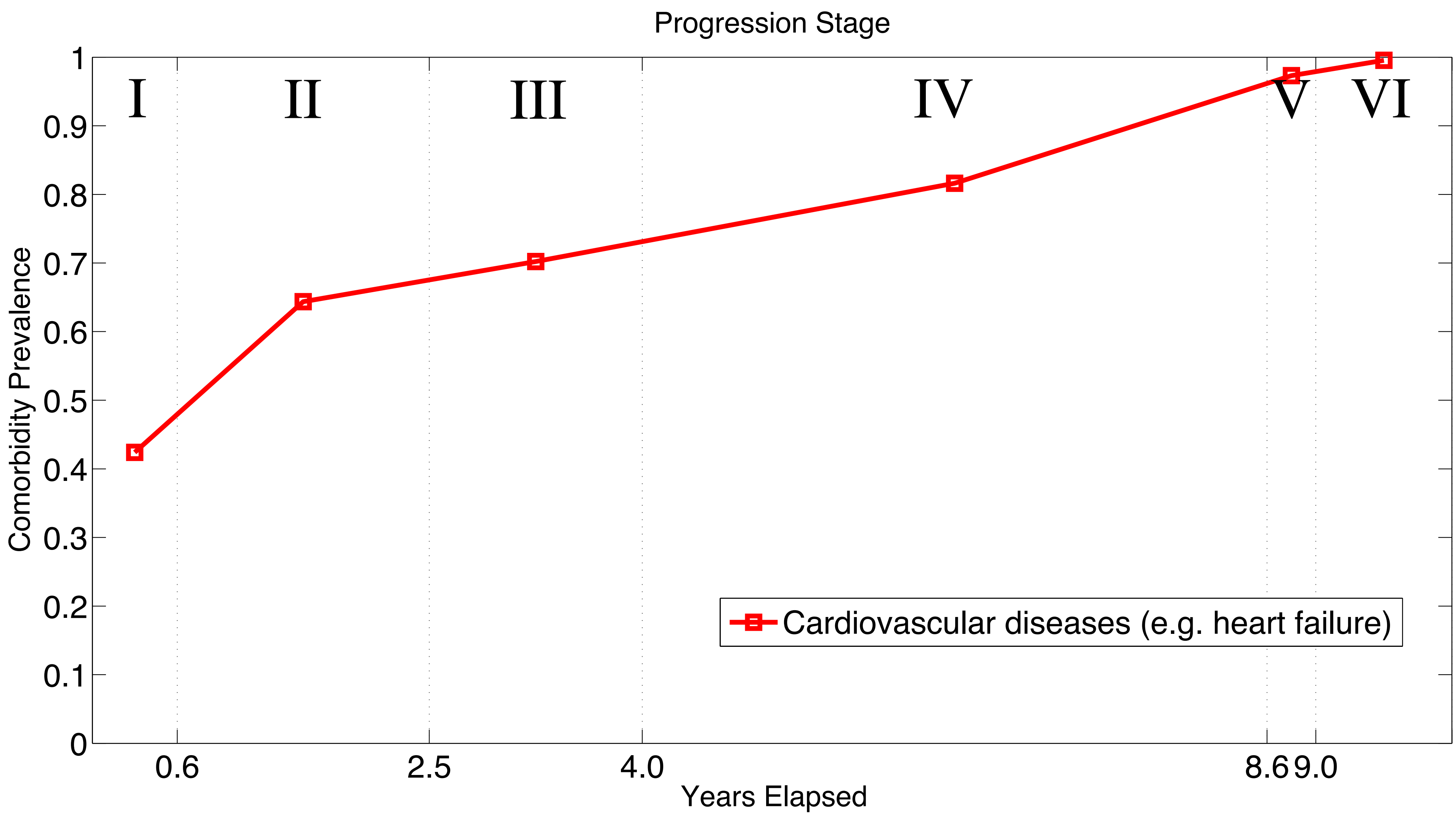
Inferred progression of a single patient

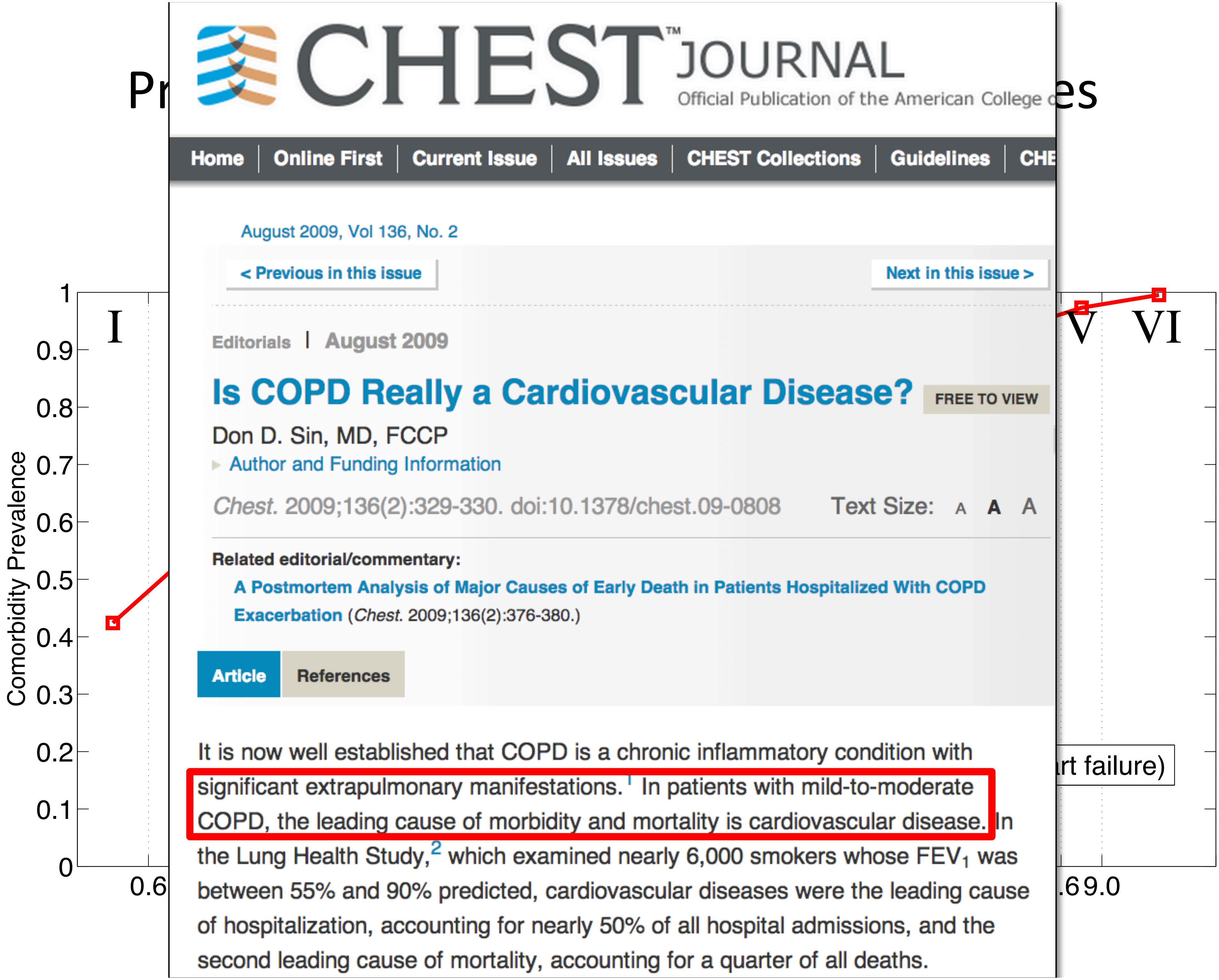


Prevalence of comorbidities across stages (Diabetes & Musculoskeletal disorders)



Prevalence of comorbidities across stages (Cardiovascular disease)





CHEST™ JOURNAL
Official Publication of the American College o

Home | Online First | Current Issue | All Issues | CHEST Collections | Guidelines | CHE

August 2009, Vol 136, No. 2

Editorials | August 2009

Is COPD Really a Cardiovascular Disease?

Don D. Sin, MD, FCCP
▸ Author and Funding Information

Chest. 2009;136(2):329-330. doi:10.1378/chest.09-0808

Related editorial/commentary:

A Postmortem Analysis of Major Causes of Early Death in Patients Hospitalized With COPD Exacerbation (*Chest.* 2009;136(2):376-380.)

It is now well established that COPD is a chronic inflammatory condition with significant extrapulmonary manifestations.[1] In patients with mild-to-moderate COPD, the leading cause of morbidity and mortality is cardiovascular disease. In the Lung Health Study,[2] which examined nearly 6,000 smokers whose FEV$_1$ was between 55% and 90% predicted, cardiovascular diseases were the leading cause of hospitalization, accounting for nearly 50% of all hospital admissions, and the second leading cause of mortality, accounting for a quarter of all deaths.

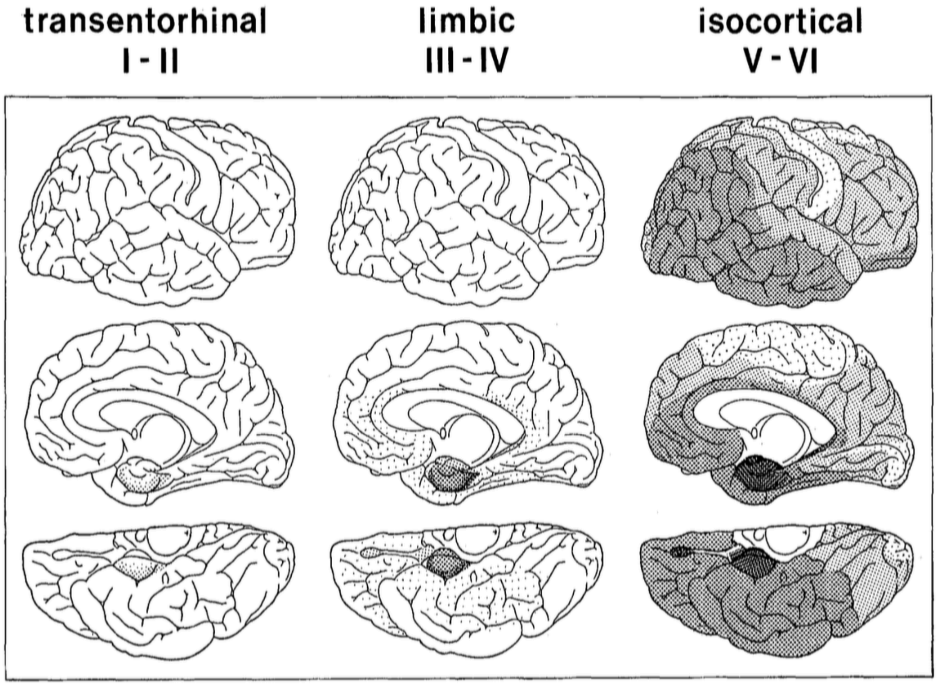
Outline of today’s lecture

- Deep dive into data commonly used for disease progression modeling
- What can we draw inspiration from, and why they are not good enough
- Probabilistic models of disease progression
- **Simultaneous staging & subtyping**

Temporal heterogeneity

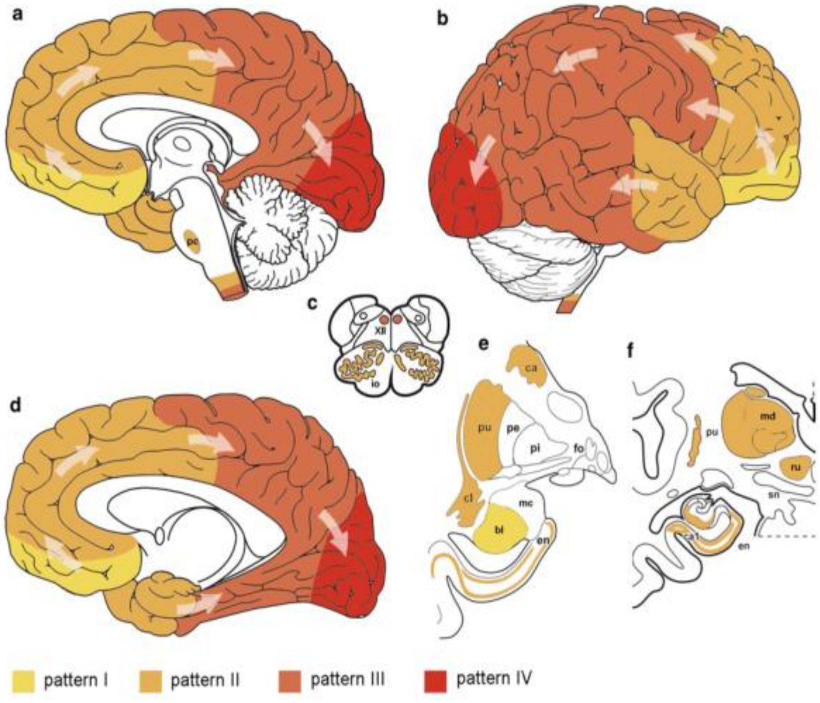
Patients show various disease stages through which patterns of pathology evolve

Alzheimer’s disease



Braak and Braak 1991

Frontotemporal dementia

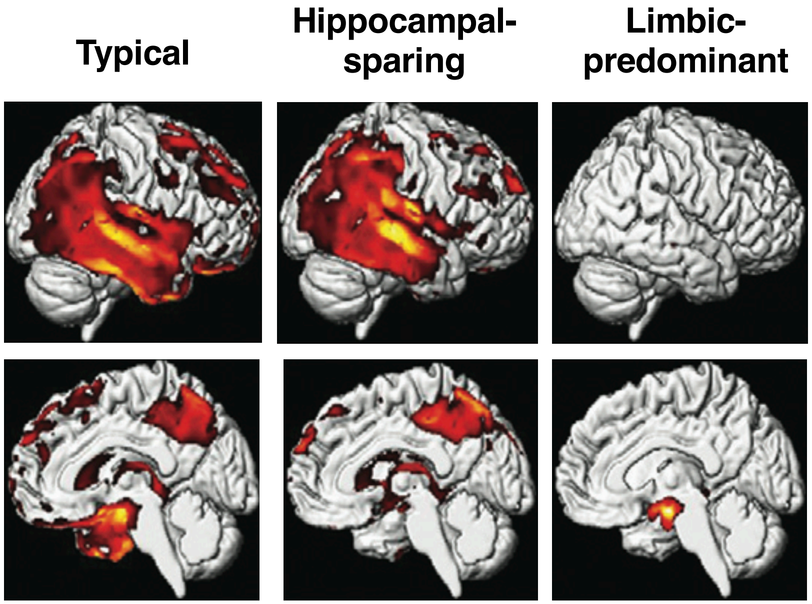


Brettschneider et al. 2014

Phenotypic heterogeneity

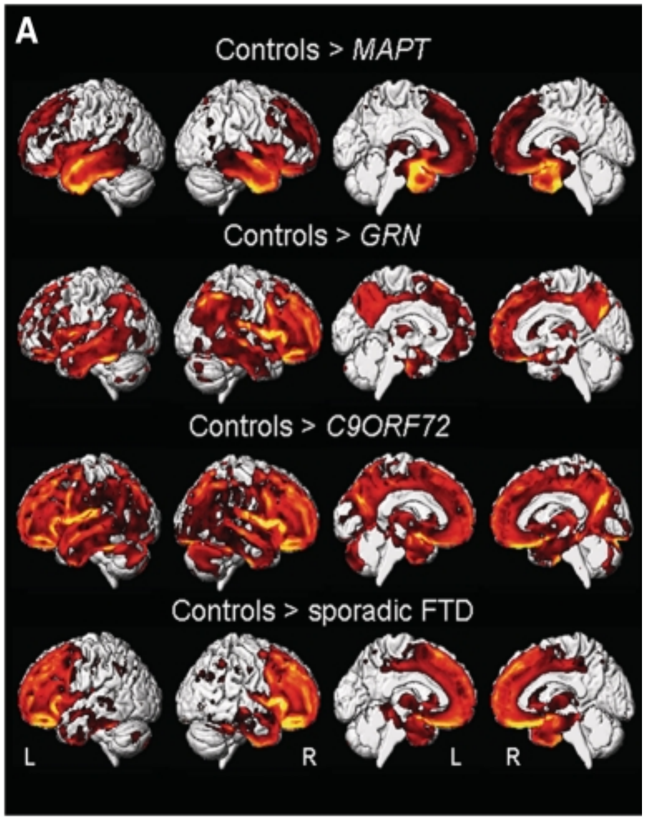
Individuals have different disease subtypes with distinct patterns of pathology

Alzheimer's disease



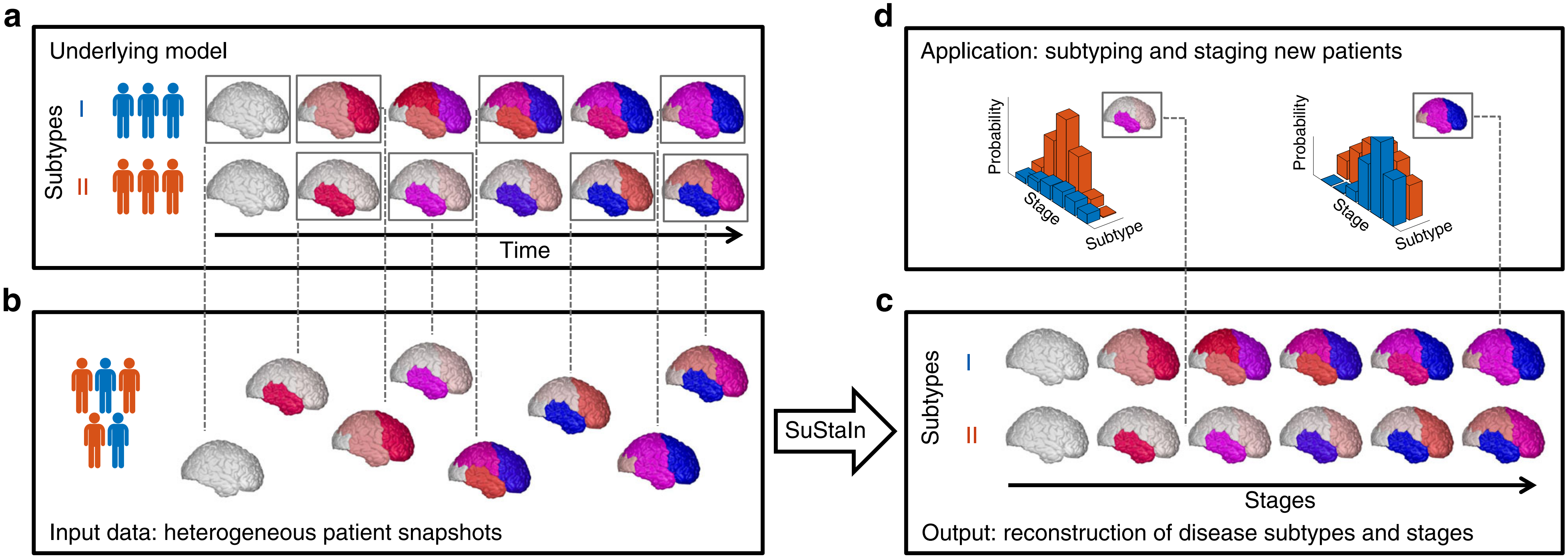
Murray et al. 2011, Whitwell et al. 2012

Frontotemporal dementia



Whitwell et al. 2012

Subtype and Stage Inference (SuStaIn)

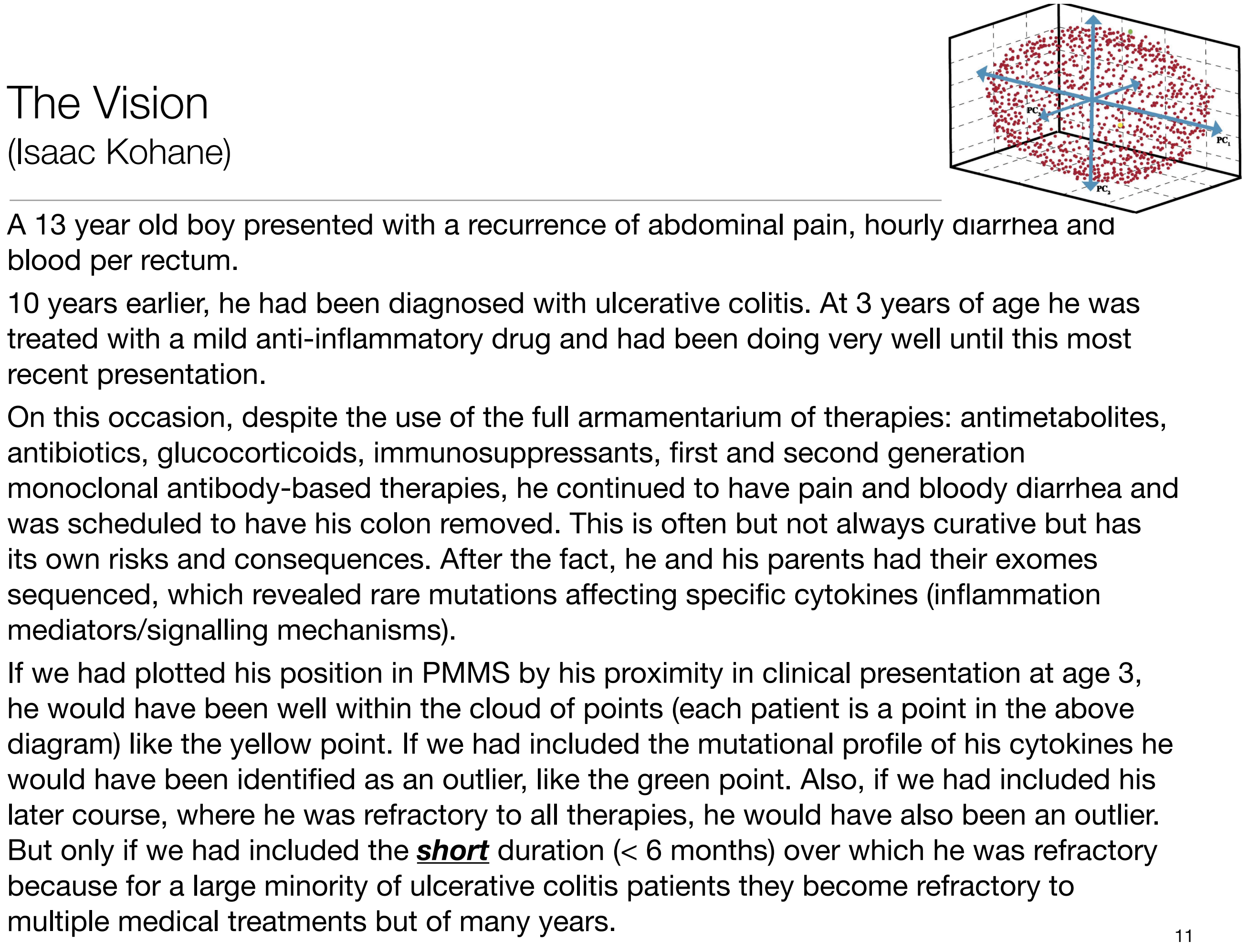

a
Underlying model
Subtypes
I
II
Time
b
Input data: heterogeneous patient snapshots
SuStaIn
d
Application: subtyping and staging new patients
Probability
Stage
Subtype
Probability
Stage
Subtype
c
Subtypes
I
II
Stages
Output: reconstruction of disease subtypes and stages

[Young et al., *Nature Communications* 2018]

Conclusion

- Many open questions
  - What data is sufficient? When is it theoretically possible to disentangle subtype and stage?
  - What are sample efficient learning algorithms, good architectures for multi-modal data, …?
- Next few years, there will be an explosion of patient data from genomics, proteomics, and metabolomics
  - Will help differentiate subtypes where otherwise impossible or very difficult
  - Small sample sizes. Infrequent measurements. Modified by treatment. Confounded by comorbidities. Outcomes must still be derived from clinical data.
  - Incredible opportunity

Returning to “The Vision” from Lecture 19…



The Vision
(Isaac Kohane)

A 13 year old boy presented with a recurrence of abdominal pain, hourly diarrhea and blood per rectum.

10 years earlier, he had been diagnosed with ulcerative colitis. At 3 years of age he was treated with a mild anti-inflammatory drug and had been doing very well until this most recent presentation.

On this occasion, despite the use of the full armamentarium of therapies: antimetabolites, antibiotics, glucocorticoids, immunosuppressants, first and second generation monoclonal antibody-based therapies, he continued to have pain and bloody diarrhea and was scheduled to have his colon removed. This is often but not always curative but has its own risks and consequences. After the fact, he and his parents had their exomes sequenced, which revealed rare mutations affecting specific cytokines (inflammation mediators/signalling mechanisms).

If we had plotted his position in PMMS by his proximity in clinical presentation at age 3, he would have been well within the cloud of points (each patient is a point in the above diagram) like the yellow point. If we had included the mutational profile of his cytokines he would have been identified as an outlier, like the green point. Also, if we had included his later course, where he was refractory to all therapies, he would have also been an outlier. But only if we had included the ***short*** duration (< 6 months) over which he was refractory because for a large minority of ulcerative colitis patients they become refractory to multiple medical treatments but of many years.

How do we achieve this for rare presentations and when we must learn from disparate, sparse, and messy data?