Machine Learning for Healthcare

6.871, HST.956

Lecture 20: Disease subtyping & progression modeling

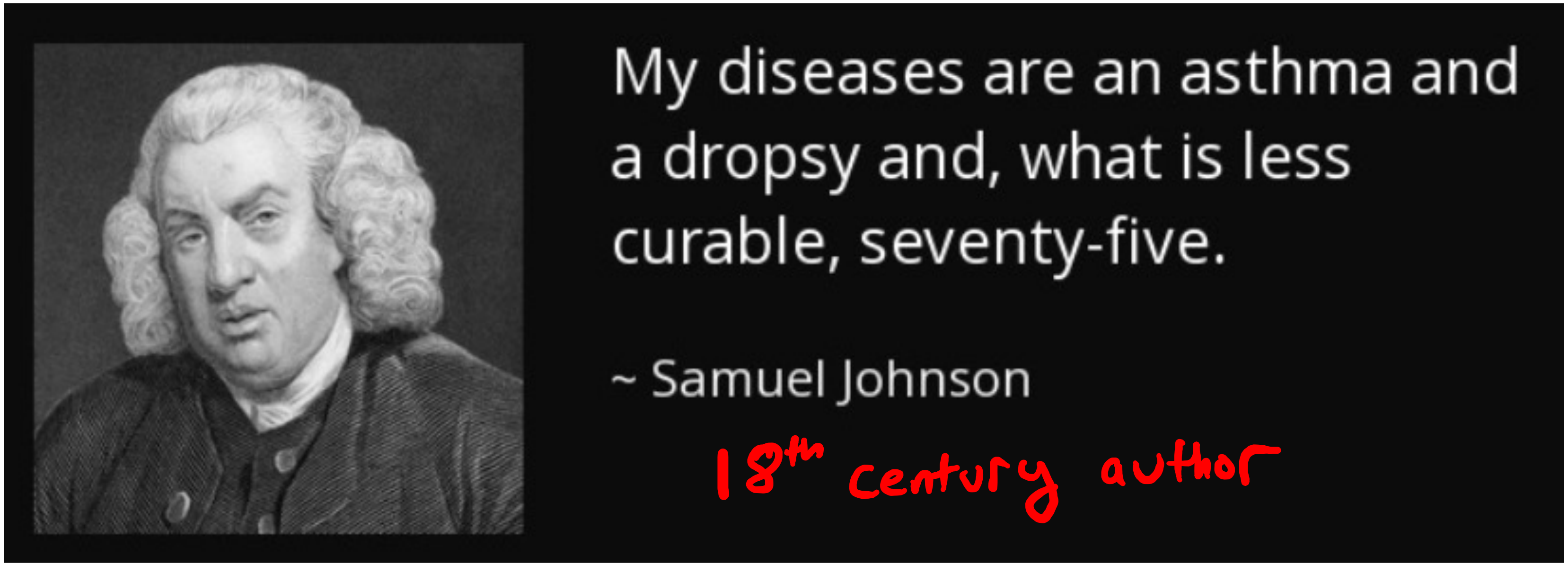
David Sontag

CSAIL

INSTITUTE FOR MEDICAL ENGINEERING & SCIENCE

HEALTH SCIENCES & TECHNOLOGY

How do we define disease & disease subtypes?



My diseases are an asthma and a dropsy and, what is less curable, seventy-five.

~ Samuel Johnson

18th century author

- What is “dropsy”?
  - “water sickness”, “swelling”, “edema”
  - *disease that got Grandma to take to her bed permanently in Victorian dramas*
  - causes: COPD, CHF, CKD, …
  - Last recorded on a death certificate ~1949
- Is “asthma” equally non-specific?

The top ten causes of death recorded in the Leeds General Cemetery burial records (19th c.)

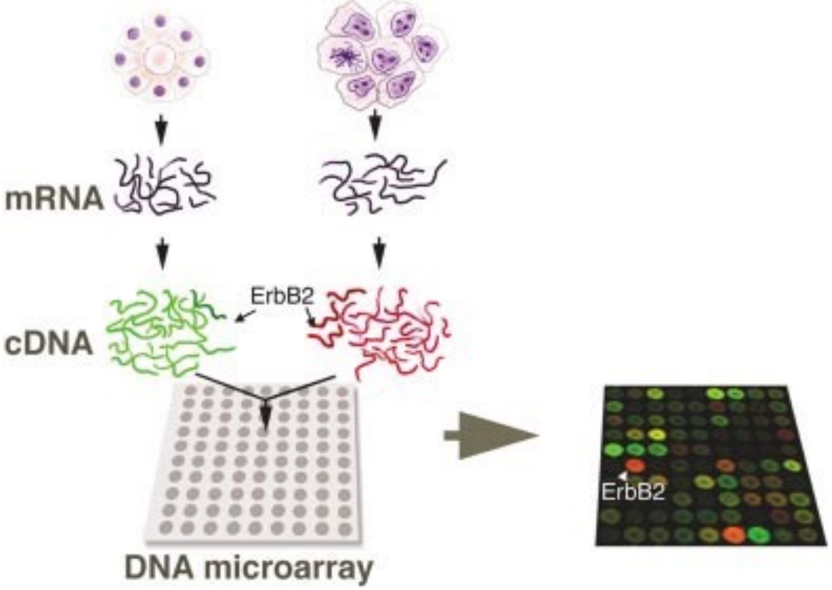
- Unknown
- Stillborn
- Bronchitis
- Consumption
- Convulsions
- Pneumonia
- Inflammation
- Diarrhoea
- Dropsy
- Natural Decay

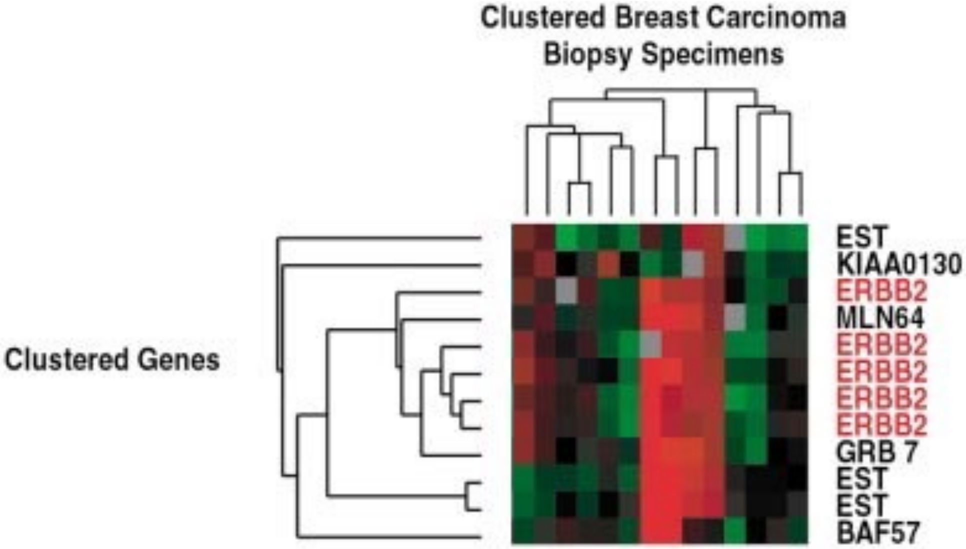
https://livingwithdying.leeds.ac.uk/2017/08/09/top-ten-ways-to-die-in-victorian-britain/

Today's lecture

- Disease subtyping
  - Of breast cancer, using gene expression
  - Of asthma, using clinical data
- Disease progression modeling

Early Efforts to Characterize Disease Subtypes using Gene Expression Microarrays





Schematic representation of a DNA microarray hybridization comparing gene expression of a malignant epithelial cancer with its normal tissue counterpart

These days, we would use RNA-seq

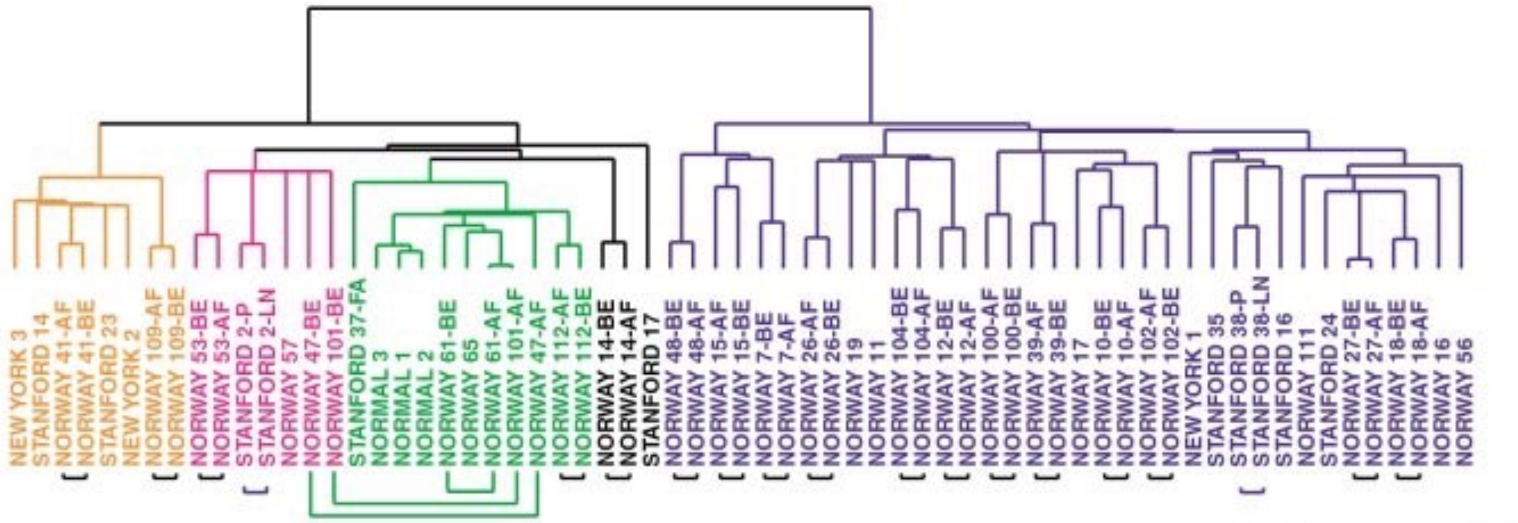
Cluster samples by nearness in gene expression space, genes by expression similarity across samples (bi-clustering)

(This small sample of array data was copied from a much larger data set)

Notice how all five different cDNA clones specific for ERBB2 cluster tightly together

Alizadeh et al., Towards a novel classification of human malignancies based on gene expression patterns, *J Pathol* 2001.

Cluster analysis on 65 breast carcinoma samples



The branching pattern of the dendrogram identifies four groups of breast tumors

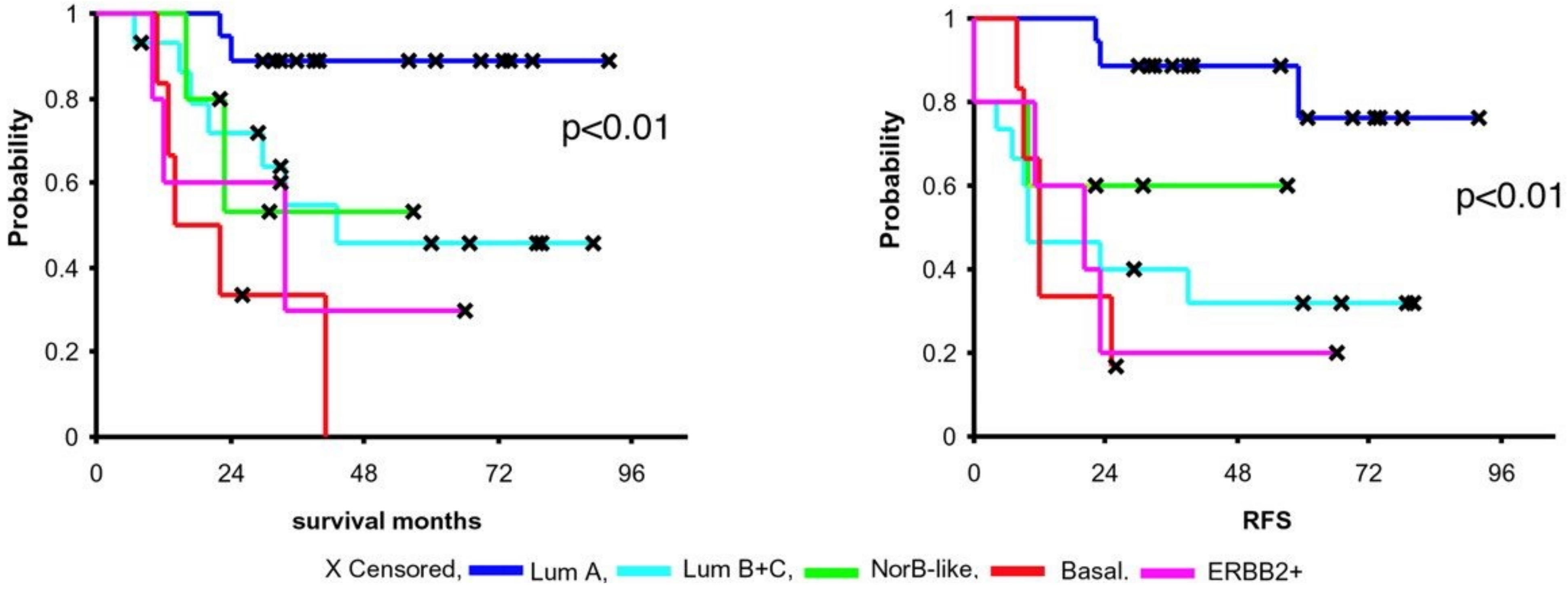
- luminal-epithelial/ER+ } split in two
- ERBB2 and other associated genes
- normal breast
- high-level expression of two clusters of genes that are characteristic of normal breast basal epithelial cells

… found to be statistically significantly associated with differences in overall patient survival and relapse-free survival

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. PNAS, 98(19), 10869–10874. http://doi.org/10.1073/pnas.191367098

Survival of Different Subgroups of Breast Cancer Patients

With a different breast cancer cohort of 49 patients treated uniformly in a prospective study, observe differences in survival across the 5 newly-characterized tumor subtypes:



Overall survival

Relapse-free survival

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. PNAS, 98(19), 10869–10874. http://doi.org/10.1073/pnas.191367098

Today’s lecture

- Disease subtyping
  - Of breast cancer, using gene expression
  - **Of asthma, using clinical data**
- Disease progression modeling

Asthma: the problem

- 5 to 10% of people with severe asthma remain poorly controlled despite maximal inhaled therapy

[Holgate ST, Polosa R. The mechanisms, diagnosis, and management of severe asthma in adults. Lancet. 2006; 368:780–793]

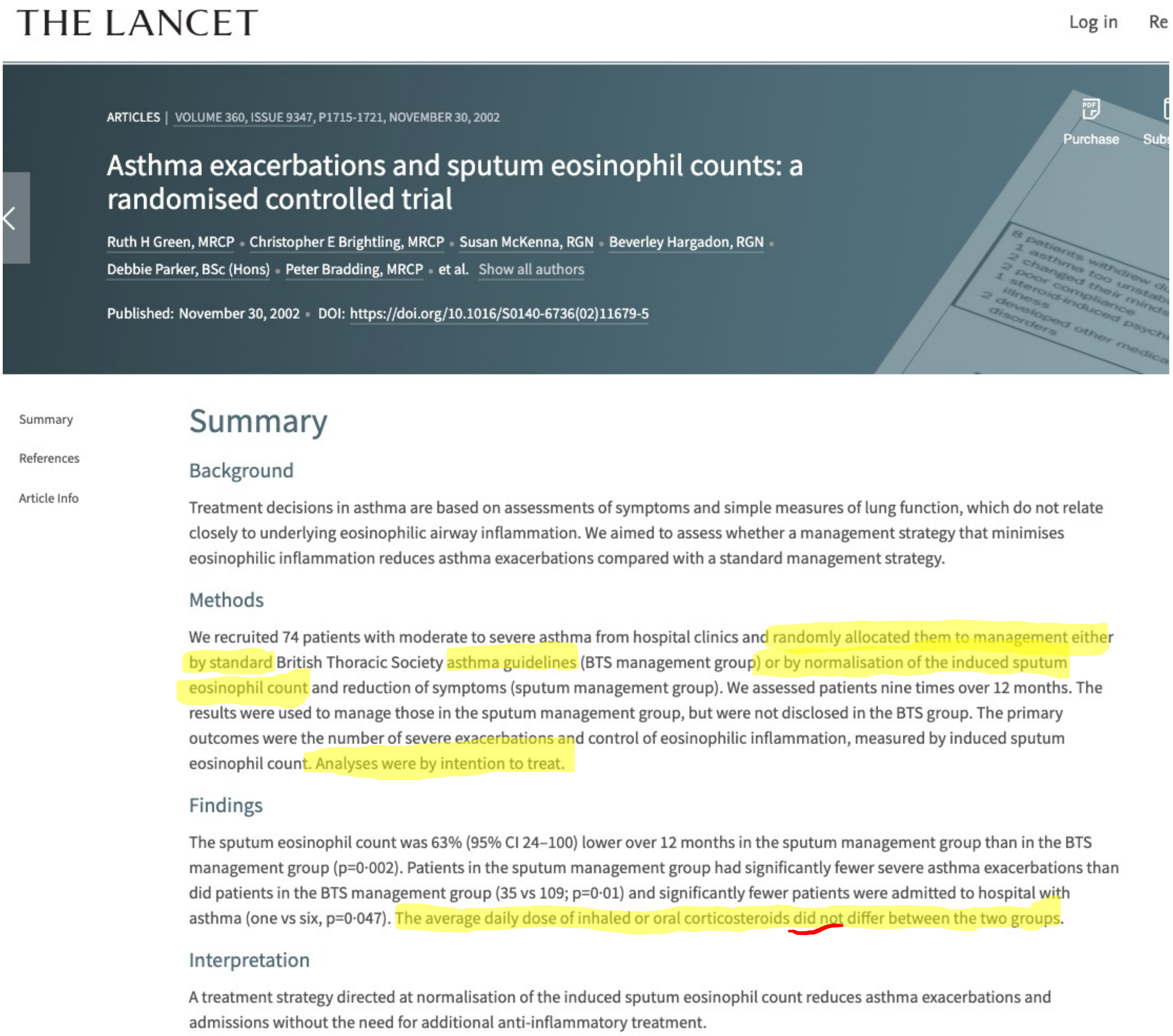
[whatasthmais.com]

Asthma: the question

"It is now recognised that there are distinct asthma phenotypes and that distinct therapeutic approaches may only impinge on some aspects of the disease process within each subgroup"

- What are the processes (genetic or environmental) that underlie different subtypes of asthma?
- Which aspects of airway remodelling are important in disease subtypes?
- What are the best biomarkers of disease progression or treatment response?
- Why are some patients less responsive to conventional therapies than others?

[Adcock et al., "New targets for drug development in asthma". The Lancet, 2008]



ARTICLES | VOLUME 360, ISSUE 9347, P1715-1721, NOVEMBER 30, 2002

Asthma exacerbations and sputum eosinophil counts: a randomised controlled trial

Ruth H Green, MRCP • Christopher E Brightling, MRCP • Susan McKenna, RGN • Beverley Hargadon, RGN • Debbie Parker, BSc (Hons) • Peter Bradding, MRCP • et al. Show all authors

Published: November 30, 2002 • DOI: https://doi.org/10.1016/S0140-6736(02)11679-5

Summary

Background

Treatment decisions in asthma are based on assessments of symptoms and simple measures of lung function, which do not relate closely to underlying eosinophilic airway inflammation. We aimed to assess whether a management strategy that minimises eosinophilic inflammation reduces asthma exacerbations compared with a standard management strategy.

Methods

We recruited 74 patients with moderate to severe asthma from hospital clinics and randomly allocated them to management either by standard British Thoracic Society asthma guidelines (BTS management group) or by normalisation of the induced sputum eosinophil count and reduction of symptoms (sputum management group). We assessed patients nine times over 12 months. The results were used to manage those in the sputum management group, but were not disclosed in the BTS group. The primary outcomes were the number of severe exacerbations and control of eosinophilic inflammation, measured by induced sputum eosinophil count. Analyses were by intention to treat.

Findings

The sputum eosinophil count was 63% (95% CI 24–100) lower over 12 months in the sputum management group than in the BTS management group (p=0·002). Patients in the sputum management group had significantly fewer severe asthma exacerbations than did patients in the BTS management group (35 vs 109; p=0·01) and significantly fewer patients were admitted to hospital with asthma (one vs six, p=0·047). The average daily dose of inhaled or oral corticosteroids did not differ between the two groups.

Interpretation

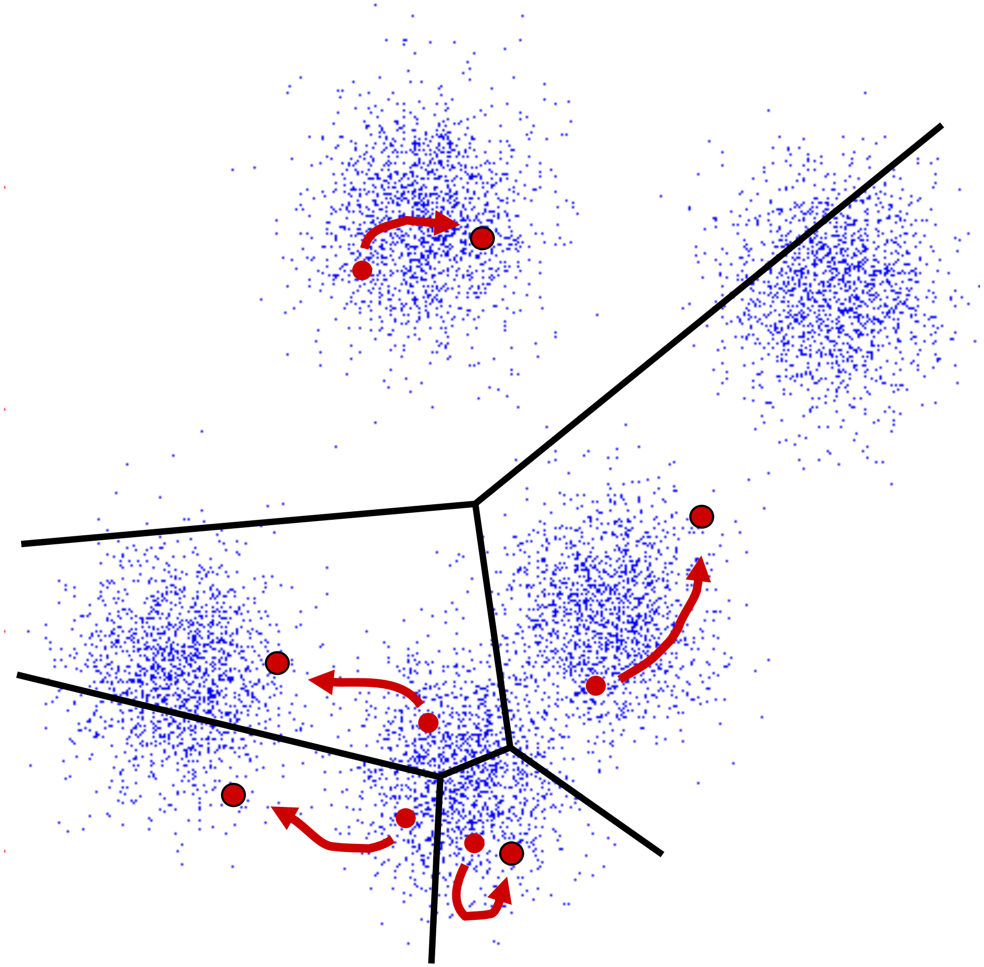
A treatment strategy directed at normalisation of the induced sputum eosinophil count reduces asthma exacerbations and admissions without the need for additional anti-inflammatory treatment.

Might there be heterogeneous treatment effects?

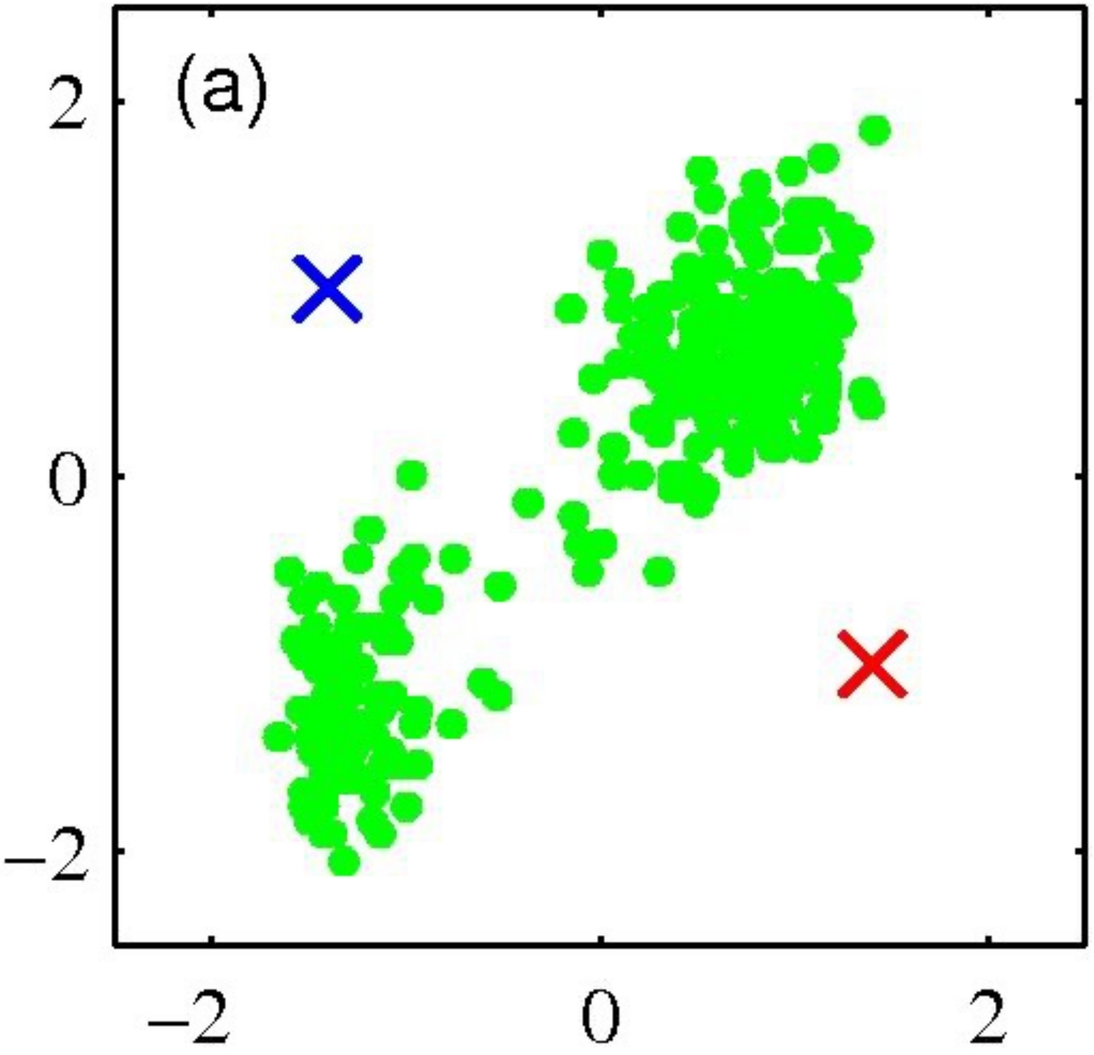
- 74 patients, 2 treatments (A vs B), outcome Y (corticosteroid therapy)
- Using what we learned about causal inference – how can we characterize which patients to use treatment A vs B with?

K-Means

- An iterative clustering algorithm
  - Initialize: Pick *K* random points as cluster centers
  - Alternate:
    1. Assign data points to closest cluster center
    2. Change the cluster center to the average of its assigned points
  - Stop when no points' assignments change



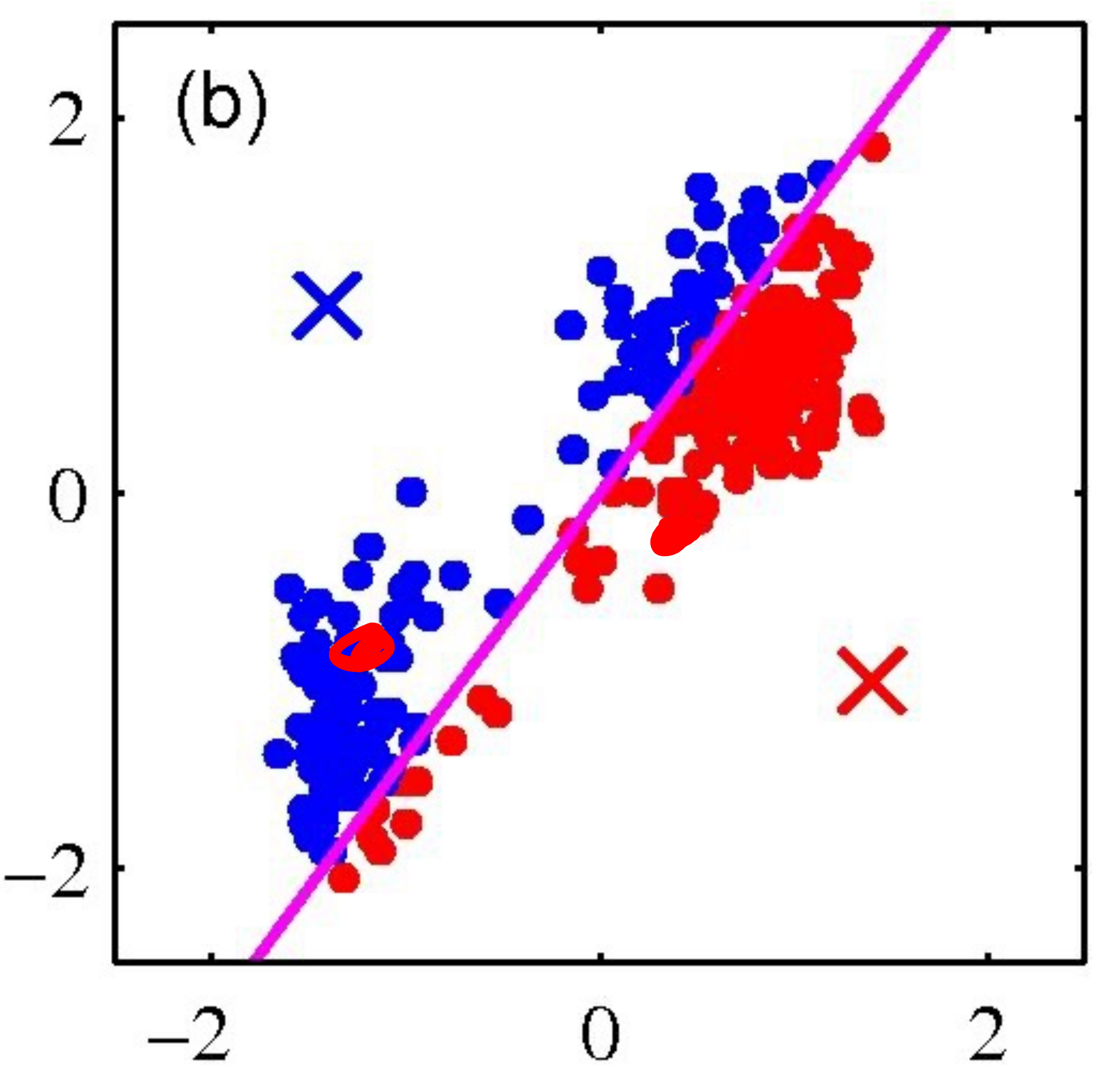
K-means clustering: Example



- Pick *K* random points as cluster centers (means)

Shown here for *K*=2

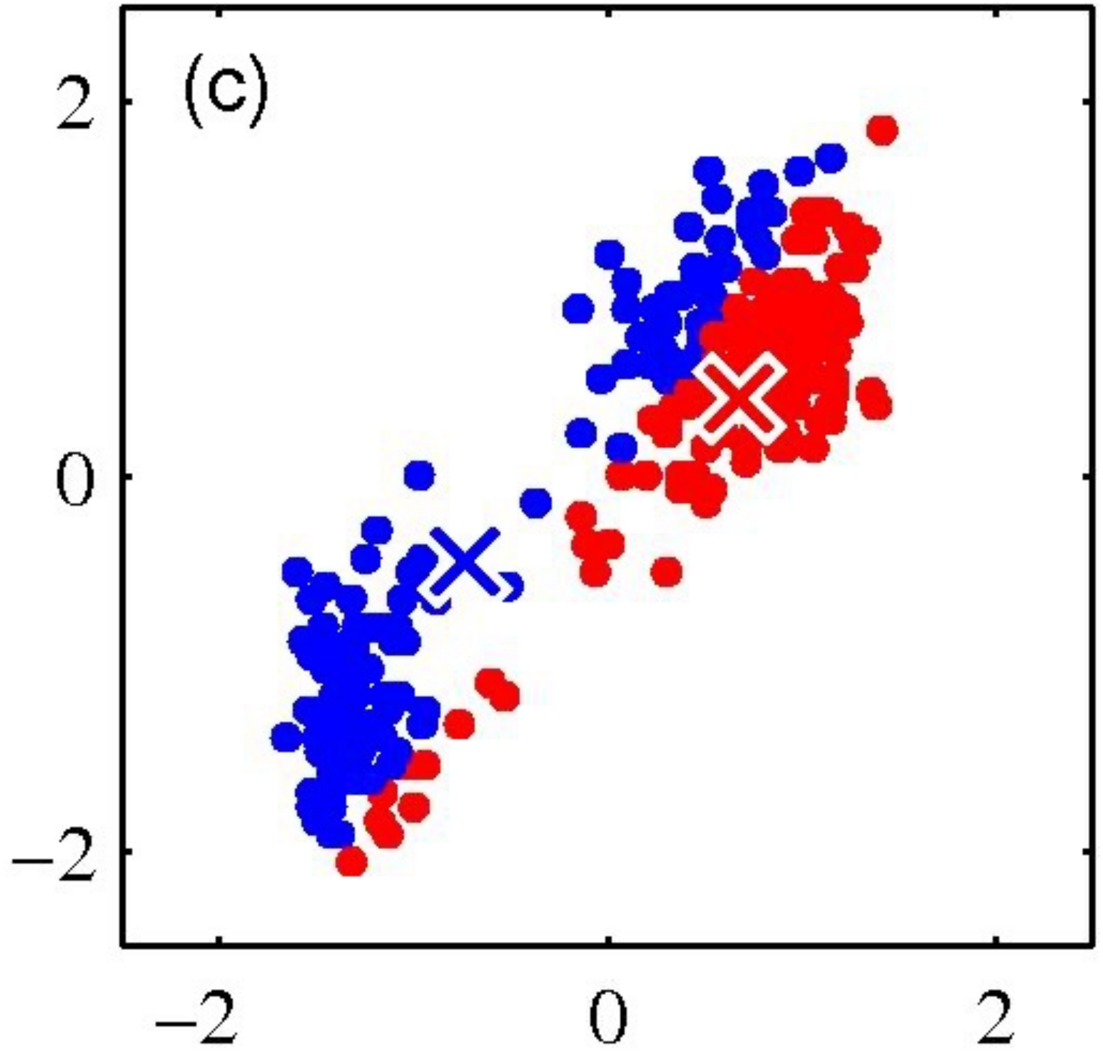
K-means clustering: Example



Iterative Step 1

- Assign data points to closest cluster center

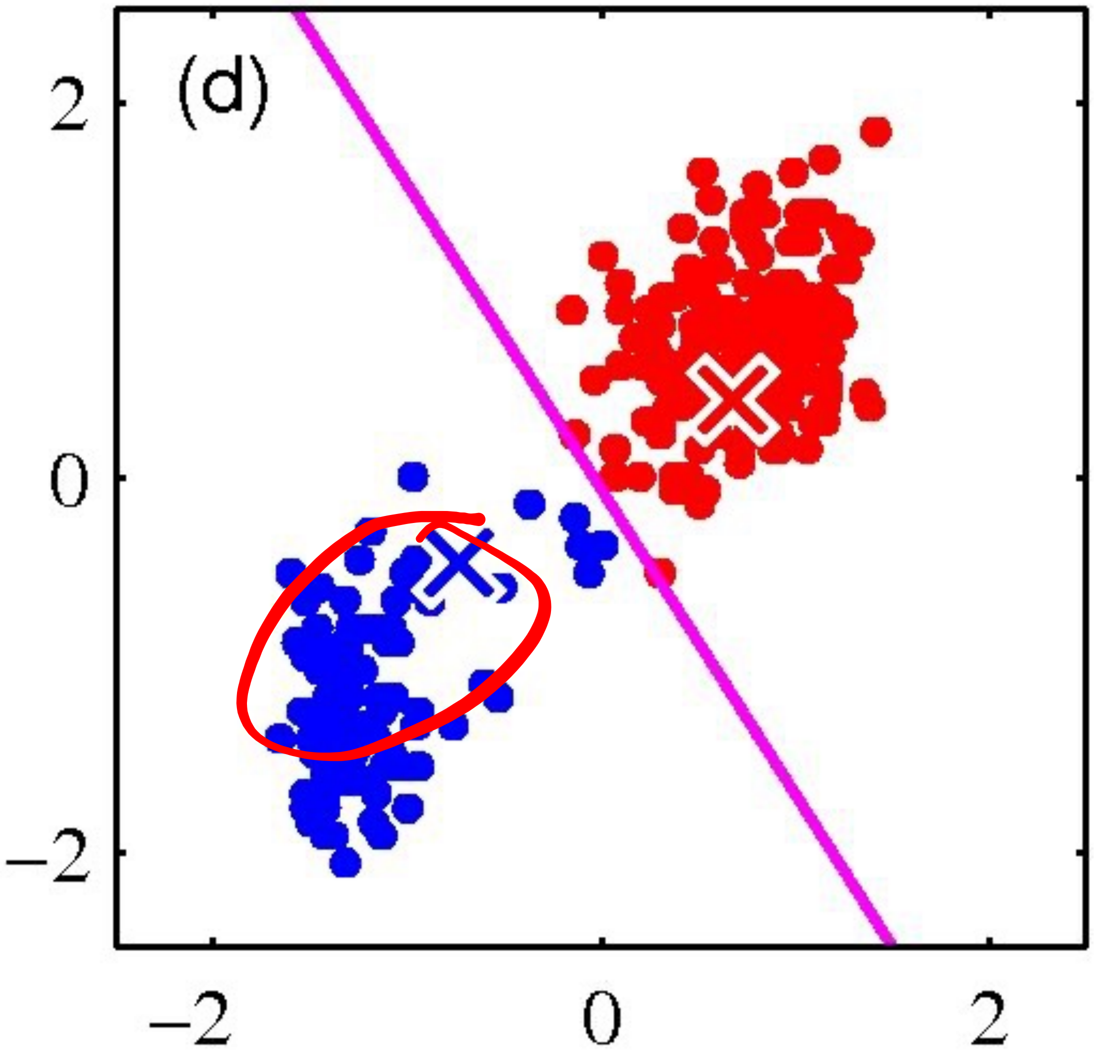
K-means clustering: Example



Iterative Step 2

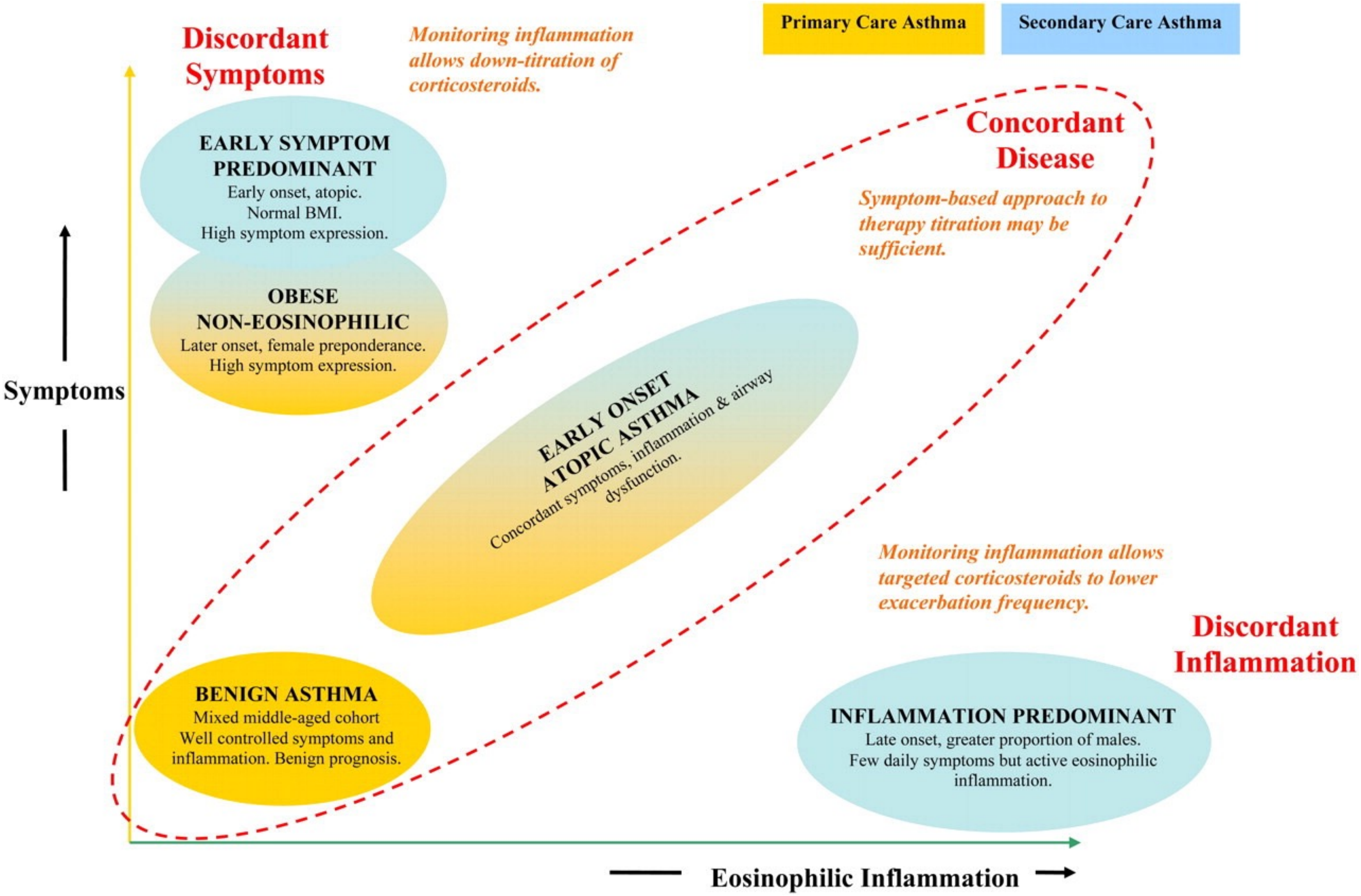
- Change the cluster center to the average of the assigned points

K-means clustering: Example



- Repeat until convergence

Discovering subtypes from data

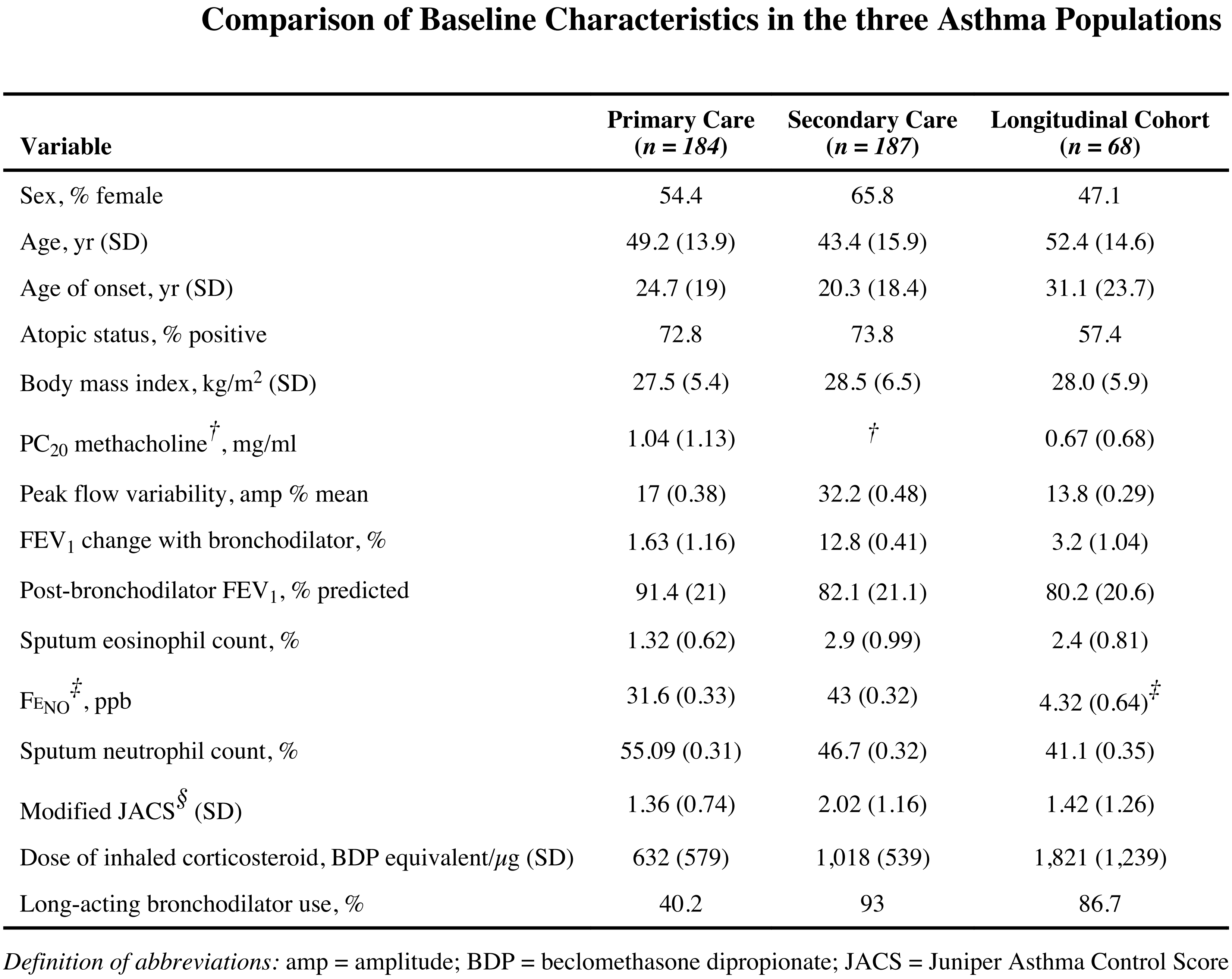

Primary Care Asthma
Secondary Care Asthma
Discordant Symptoms
Monitoring inflammation allows down-titration of corticosteroids.
EARLY SYMPTOM PREDOMINANT
Early onset, atopic.
Normal BMI.
High symptom expression.
OBESE NON-EOSINOPHILIC
Later onset, female preponderance.
High symptom expression.
Symptoms
Concordant Disease
Symptom-based approach to therapy titration may be sufficient.
EARLY ONSET ATOPIC ASTHMA
Concordant symptoms, inflammation & airway dysfunction.
Monitoring inflammation allows targeted corticosteroids to lower exacerbation frequency.
Discordant Inflammation
BENIGN ASTHMA
Mixed middle-aged cohort
Well controlled symptoms and inflammation. Benign prognosis.
INFLAMMATION PREDOMINANT
Late onset, greater proportion of males.
Few daily symptoms but active eosinophilic inflammation.
Eosinophilic Inflammation

[Haldar et al., *Am J Respir Crit Care Med*, 2008]

The data

- All patients had physician diagnosis of asthma and at least one recent prescription for asthma therapy
- All were current nonsmokers
- *Data set #1*: 184 patients recruited from primary-care practices in the UK
- *Data set #2*: 187 patients from refractory asthma clinic in the UK
- *Data set #3*: 68 patients from 12 month clinical study (RCT)
- Features: *z* scores for continuous variables, 0/1 for categorical
  - Some of the continuous variables log-transformed to approximate a normal distribution

[Haldar et al., *Am J Respir Crit Care Med*, 2008]



Comparison of Baseline Characteristics in the three Asthma Populations

Variable	Primary Care (*n = 184*)	Secondary Care (*n = 187*)	Longitudinal Cohort (*n = 68*)
Sex, % female	54.4	65.8	47.1
Age, yr (SD)	49.2 (13.9)	43.4 (15.9)	52.4 (14.6)
Age of onset, yr (SD)	24.7 (19)	20.3 (18.4)	31.1 (23.7)
Atopic status, % positive	72.8	73.8	57.4
Body mass index, kg/m^2 (SD)	27.5 (5.4)	28.5 (6.5)	28.0 (5.9)
PC_{20} methacholine[†], mg/ml	1.04 (1.13)	[†]	0.67 (0.68)
Peak flow variability, amp % mean	17 (0.38)	32.2 (0.48)	13.8 (0.29)
FEV_1 change with bronchodilator, %	1.63 (1.16)	12.8 (0.41)	3.2 (1.04)
Post-bronchodilator FEV_1, % predicted	91.4 (21)	82.1 (21.1)	80.2 (20.6)
Sputum eosinophil count, %	1.32 (0.62)	2.9 (0.99)	2.4 (0.81)
$F_{E_{NO}}$[‡], ppb	31.6 (0.33)	43 (0.32)	4.32 (0.64)[‡]
Sputum neutrophil count, %	55.09 (0.31)	46.7 (0.32)	41.1 (0.35)
Modified JACS[§] (SD)	1.36 (0.74)	2.02 (1.16)	1.42 (1.26)
Dose of inhaled corticosteroid, BDP equivalent/μg (SD)	632 (579)	1,018 (539)	1,821 (1,239)
Long-acting bronchodilator use, %	40.2	93	86.7

Definition of abbreviations: amp = amplitude; BDP = beclomethasone dipropionate; JACS = Juniper Asthma Control Score

[Haldar et al., *Am J Respir Crit Care Med*, 2008]

Clusters in primary care

(found by K-means)

		Cluster 1	Cluster 2	Cluster 3	
Variable	**Primary Care (*n = 184*)**	**Early-Onset Atopic Asthma (*n = 61*)**	**Obese Noneosinophilic (*n* = 27)**	**Benign Asthma (*n* = 96)**	**Significance (*P* Value)[*]**
Sex[†], % female	**54.4**	45.9	81.5	52.1	0.006
Age, yr (SD)	**49.2 (13.9)**	44.5 (14.3)	53.9 (14)	50.8 (13)	0.003
Age of onset[†], yr (SD)	**24.7 (19)**	14.6 (15.4)	35.3 (19.6)	28.2 (18.3)	<0.001
Atopic status[†], % positive	**72.8**	95.1	51.9	64.6	<0.001
Body mass index[†], kg/m^2 (SD)	**27.5 (5.4)**	26.1 (3.8)	36.2 (5.5)	26 (3.6)	<0.001
PC_{20} methacholine[†‡], mg/ml	**1.04 (1.13)**	0.12 (0.86)	1.60 (0.93)	6.39 (0.75)	<0.001
PC_{20} >8 mg/ml, n (%)	**64 (34.7)**	2 (3.3)	6 (22.2)	56 (58.3)	<0.001
Peak flow variability[†‡], amp % mean	**17 (0.38)**	20 (0.47)	21.9 (0.32)	14.8 (0.32)	0.039
FEV_1 change with bronchodilator[‡], %	**1.63 (1.16)**	4.5 (0.91)	1.82 (1.16)	0.83 (1.22)	<0.001
Post-bronchodilator FEV_1, % predicted	**91.4 (21)**	86.9 (20.7)	91.5 (21.4)	94.2 (20.7)	0.107
Sputum eosinophil count[†‡], %	**1.32 (0.62)**	3.75 (0.64)	1.55 (0.51)	0.65 (0.44)	<0.001
FE_{NO}[‡§], ppb	**31.6 (0.33)**	57.5 (0.27)	25.8 (0.29)	22.8 (0.27)	<0.001
Sputum neutrophil count[‡], %	**55.09 (0.31)**	45.87 (0.24)	72.71 (0.13)	57.56 (0.36)	0.038
Modified JACS[†] (SD)	**1.36 (0.74)**	1.54 (0.58)	2.06 (0.73)	1.04 (0.66)	<0.001
Dose of inhaled corticosteroid, BDP equivalent/μg (SD)	**632 (579)**	548 (559)	746 (611)	653 (581)	0.202
Long-acting bronchodilator use, %	**40.2**	34.4	48.2	41.7	0.442
Previous hospital admission or emergency attendance, no. per patient	**0.60 (1.57)**	1.04	0.26	0.20	0.037
Previous outpatient attendance, % attended	**15%**	22%	19%	6%	0.121
Severe asthma exacerbations (requiring oral corticosteroids) in past 12 mo, no. per patient	**1.25 (1.94)**	1.86 (0.32)	1.07 (0.32)	0.39 (0.18)	0.002

Clusters in secondary care

Variable	Secondary Care (*n = 187*)	Cluster 1 Early Onset, Atopic (*n = 74*)	Cluster 2 Obese, Noneosinophilic (*n = 23*)	Cluster 3 Early Symptom Predominant (*n = 22*)	Cluster 4 Inflammation Predominant (*n = 68*)	Significance (*P* Value)*
Sex†, % female	**65.8**			68.2	47.1	<0.001
Age, yr (SD)	**43.4 (15.9)**			35.5 (15.5)	50.6 (15.1)	<0.001
Age of onset†, yr (SD)	**20.3 (18.4)**			12.6 (15)	32.6 (19.1)	<0.001
Atopic status†, % positive	**73.8**			81.8	63.2	0.024
Body mass index†, kg/m^2 (SD)	**28.5 (6.5)**			23.6 (3.1)	27 (3.9)	<0.001
Peak flow variability‡, amp % mean	**32.2 (0.48)**			24.2 (0.65)	27.6 (0.36)	0.002
FEV_1 change with bronchodilator‡, %	**12.8 (0.41)**	24.5 (0.31)	9.3 (0.35)	4.5 (0.33)	9.8 (0.34)	<0.001
Post-bronchodilator FEV_1, % predicted (SD)	**82.1 (21.1)**	79.0 (21.9)	79.0 (18.5)	79.5 (26.1)	87.2 (18.5)	0.093
Sputum eosinophil count†‡, %	**2.9 (0.99)**	4.2 (0.76)	1.3 (1.01)	0.1 (0.9)	8.4 (0.64)	<0.001
FE_{NO}‡§, ppb	**43 (0.32)**	51.2 (0.36)	24.2 (0.27)	22.6 (0.30)	53.1 (0.32)	<0.001
Sputum neutrophil count, %‡	**46.7 (0.32)**	45.4 (0.39)	49.3 (0.22)	51.3 (0.23)	45.9 (0.29)	0.892
Modified JACS† (SD)	**2.02 (1.16)**	2.63 (0.93)	2.37 (1.09)	2.11 (1.11)	1.21 (0.95)	<0.001
Dose of inhaled corticosteroid, BDP equivalent/*μg* (SD)	**1,018 (539)**	1,168 (578)	1,045 (590)	809 (396)	914 (479)	0.008
Long-acting bronchodilator use, %	**93.0**	91.9	95.4	90.9	94.1	0.999

Resembled clusters from primary care – i.e., these are common across spectrum of severity

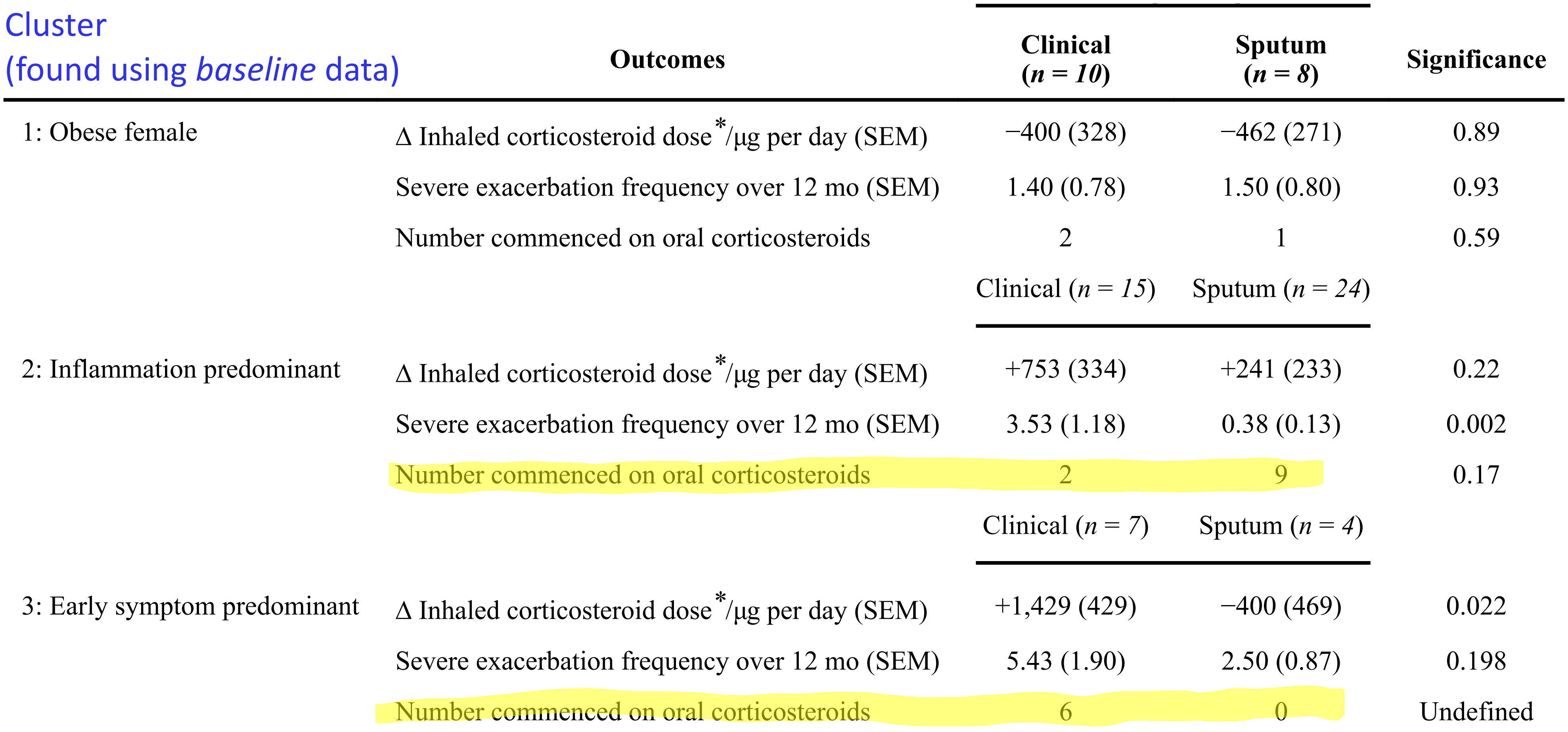
Objective measures of disease severity show more advanced disease

Identifying heterogeneous treatment effects from the RCT

- Now we use the 3rd dataset – 68 patients over 12 months
- Randomized control trial with two arms:
  - Standard clinical care ("clinical")
  - Regular monitoring of airway inflammation using induced sputum, to titrate steroid therapy to maintain normal eosinophil counts ("sputum")
- Original study found no difference in corticosteroid usage
  - But, this could have been explained by heterogeneity in treatment response!

[Haldar et al., *Am J Respir Crit Care Med*, 2008]

Patients in different clusters respond differently to treatment! (analysis using 3rd dataset from 12 month study)

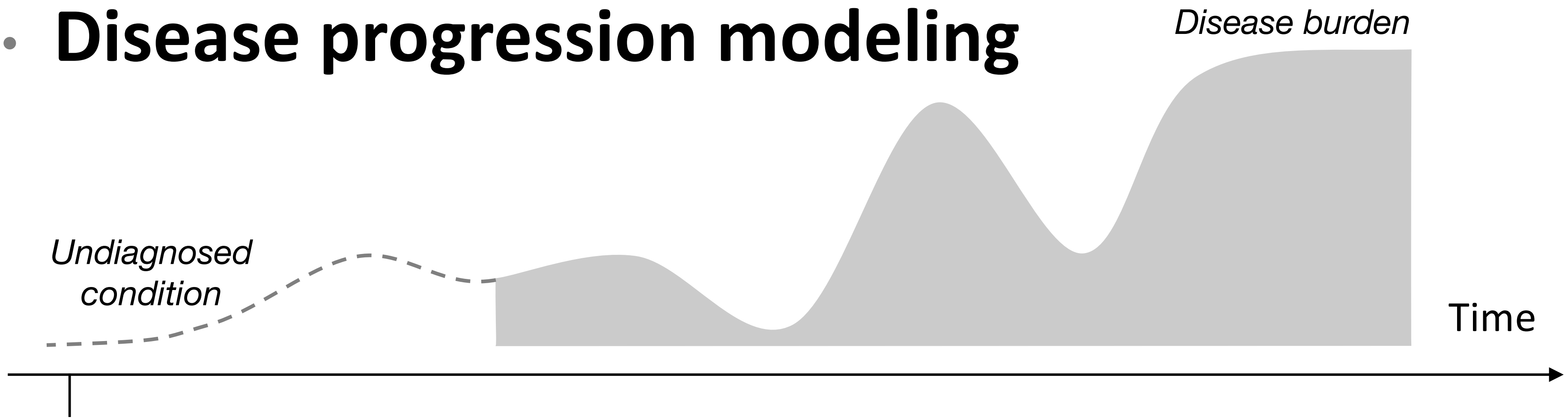


Cluster (found using *baseline* data)	Outcomes	Treatment strategy		Significance
		Clinical **(*n = 10*)**	**Sputum** **(*n = 8*)**	
1: Obese female	Δ Inhaled corticosteroid dose*/μg per day (SEM)	−400 (328)	−462 (271)	0.89
	Severe exacerbation frequency over 12 mo (SEM)	1.40 (0.78)	1.50 (0.80)	0.93
	Number commenced on oral corticosteroids	2	1	0.59
		Clinical (*n = 15*)	Sputum (*n = 24*)	
2: Inflammation predominant	Δ Inhaled corticosteroid dose*/μg per day (SEM)	+753 (334)	+241 (233)	0.22
	Severe exacerbation frequency over 12 mo (SEM)	3.53 (1.18)	0.38 (0.13)	0.002
	Number commenced on oral corticosteroids	2	9	0.17
		Clinical (*n = 7*)	Sputum (*n = 4*)	
3: Early symptom predominant	Δ Inhaled corticosteroid dose*/μg per day (SEM)	+1,429 (429)	−400 (469)	0.022
	Severe exacerbation frequency over 12 mo (SEM)	5.43 (1.90)	2.50 (0.87)	0.198
	Number commenced on oral corticosteroids	6	0	Undefined

[Haldar et al., *Am J Respir Crit Care Med*, 2008]

Today's lecture

- Disease subtyping
  - Of breast cancer, using gene expression
  - Of asthma, using clinical data
- **Disease progression modeling**



Where is a patient in their disease trajectory?
When will the disease progress?
How will treatment affect disease progression?

Goals of disease progression modeling

- **Descriptive:**
  - ***Find markers of disease stage and progression, statistics of what to expect when***
- Predictive:
  - *What will this patient's future trajectory look like?*
  - *How will treatment affect it?*
- Key challenges:
  - Seldom directly observe disease stage, but rather only indirect observations (e.g. symptoms, lab results)
  - Data can be censored – don't observe beginning to end

Example: learning 10-year progression of COPD

- 2-4 years of data for each patient
- High-dimensional, with **lots of missing data**
- No ground truth – not even spirometry

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

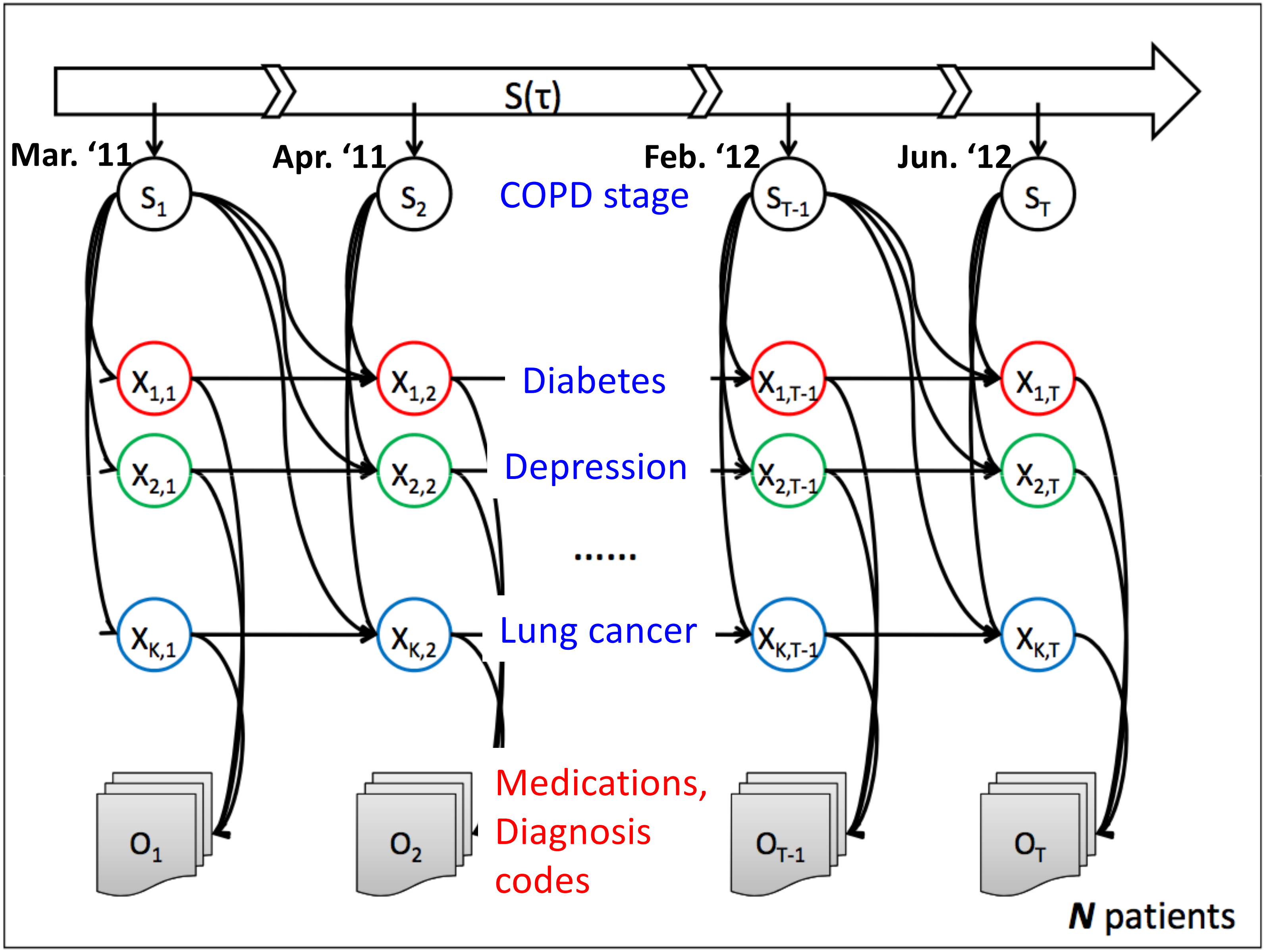
Probabilistic model of disease progression

Markov Jump Process

Progression Stages

Comorbidities, each with its own Markov chain

Observations



COPD stage

Diabetes

Depression

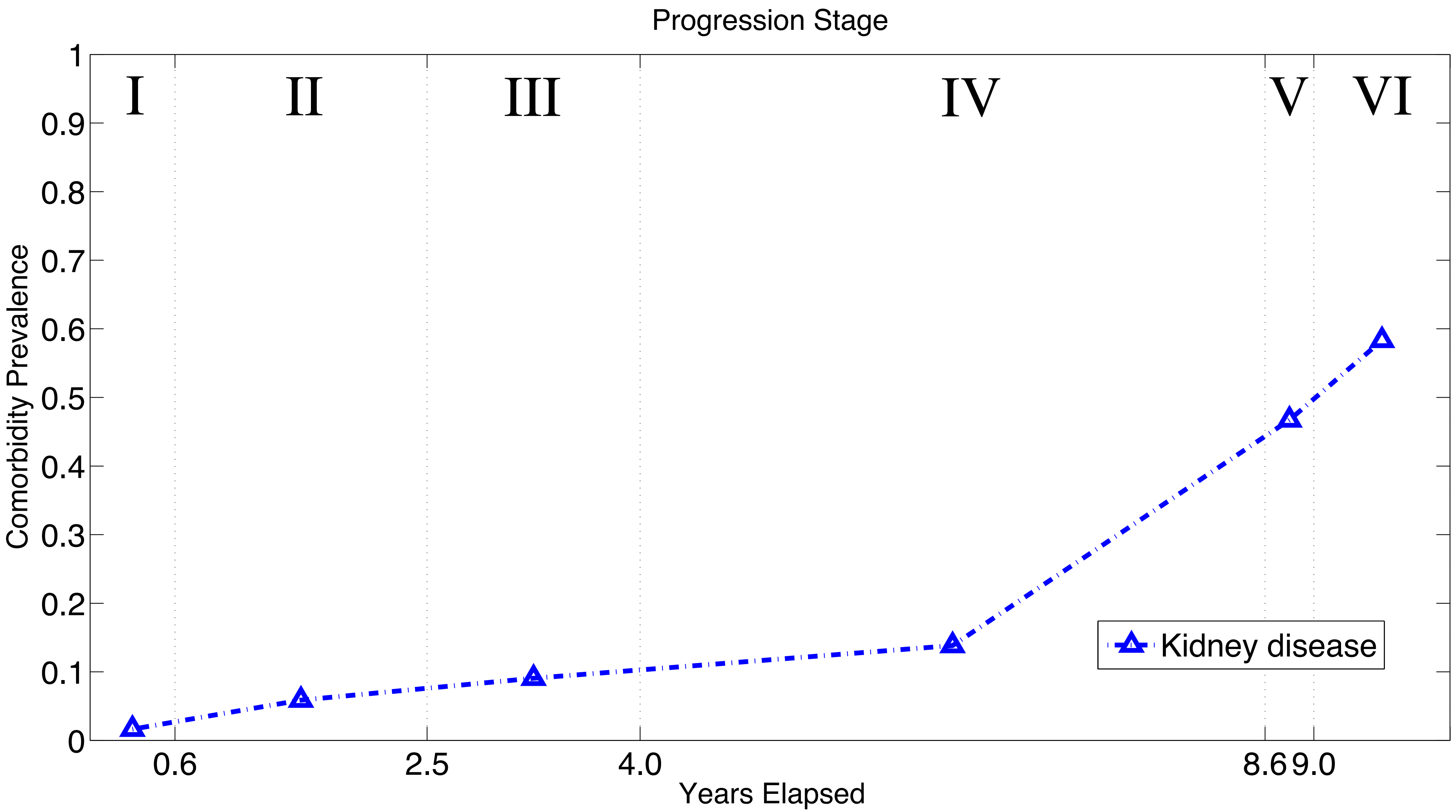
......

Lung cancer

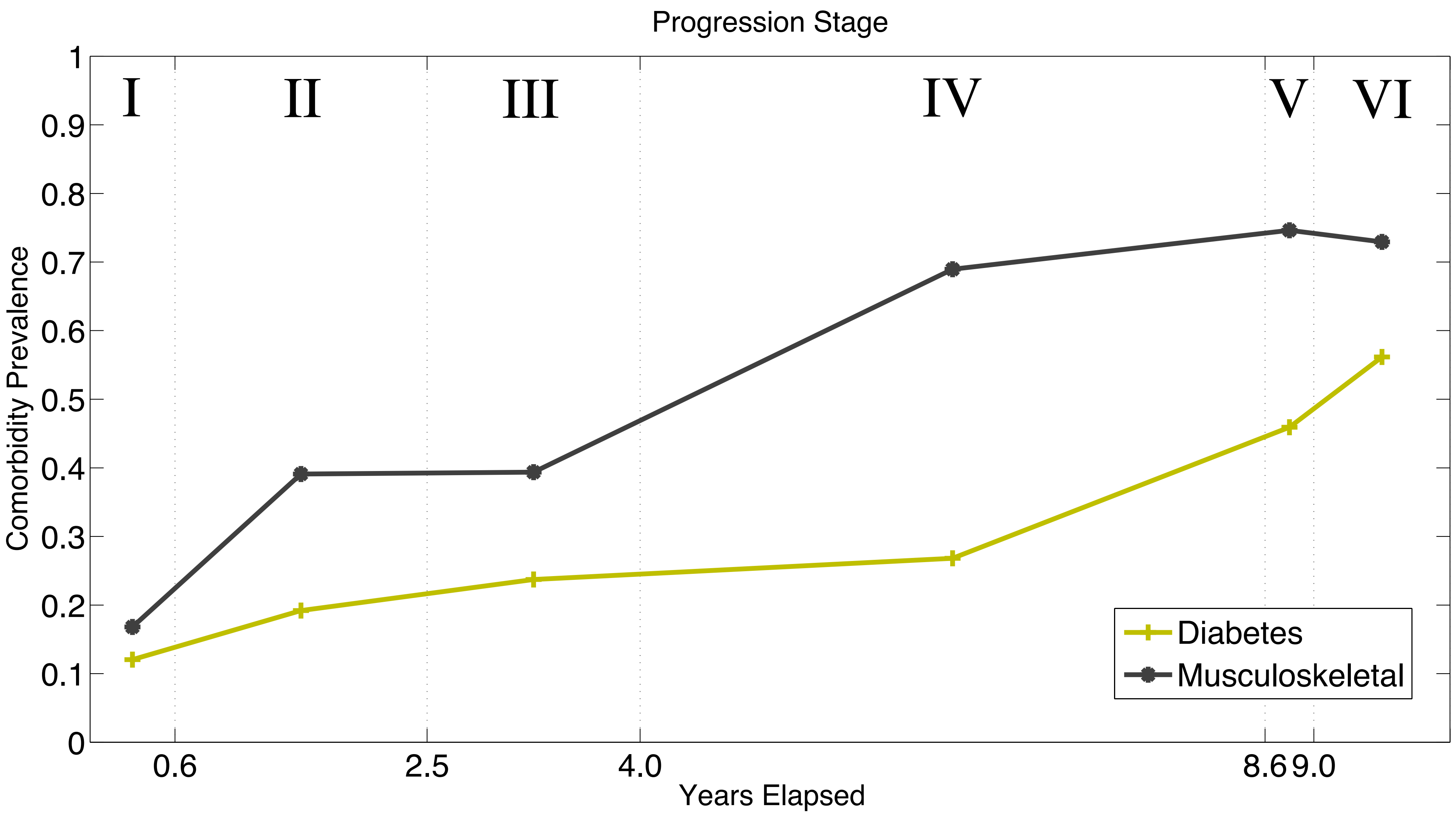
Medications, Diagnosis codes

N **patients**

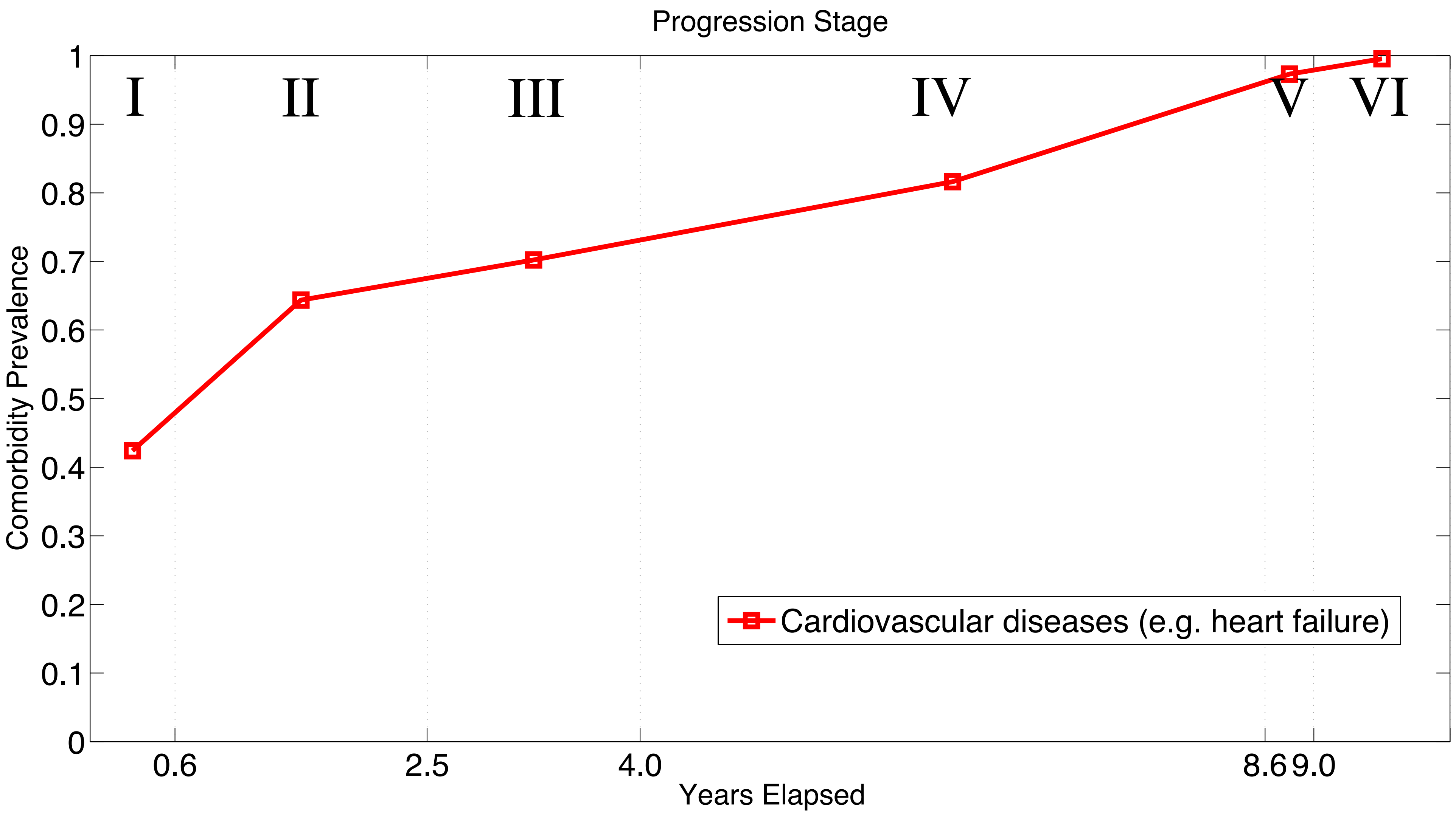
Inferred prevalence of comorbidities across stages (Kidney disease)

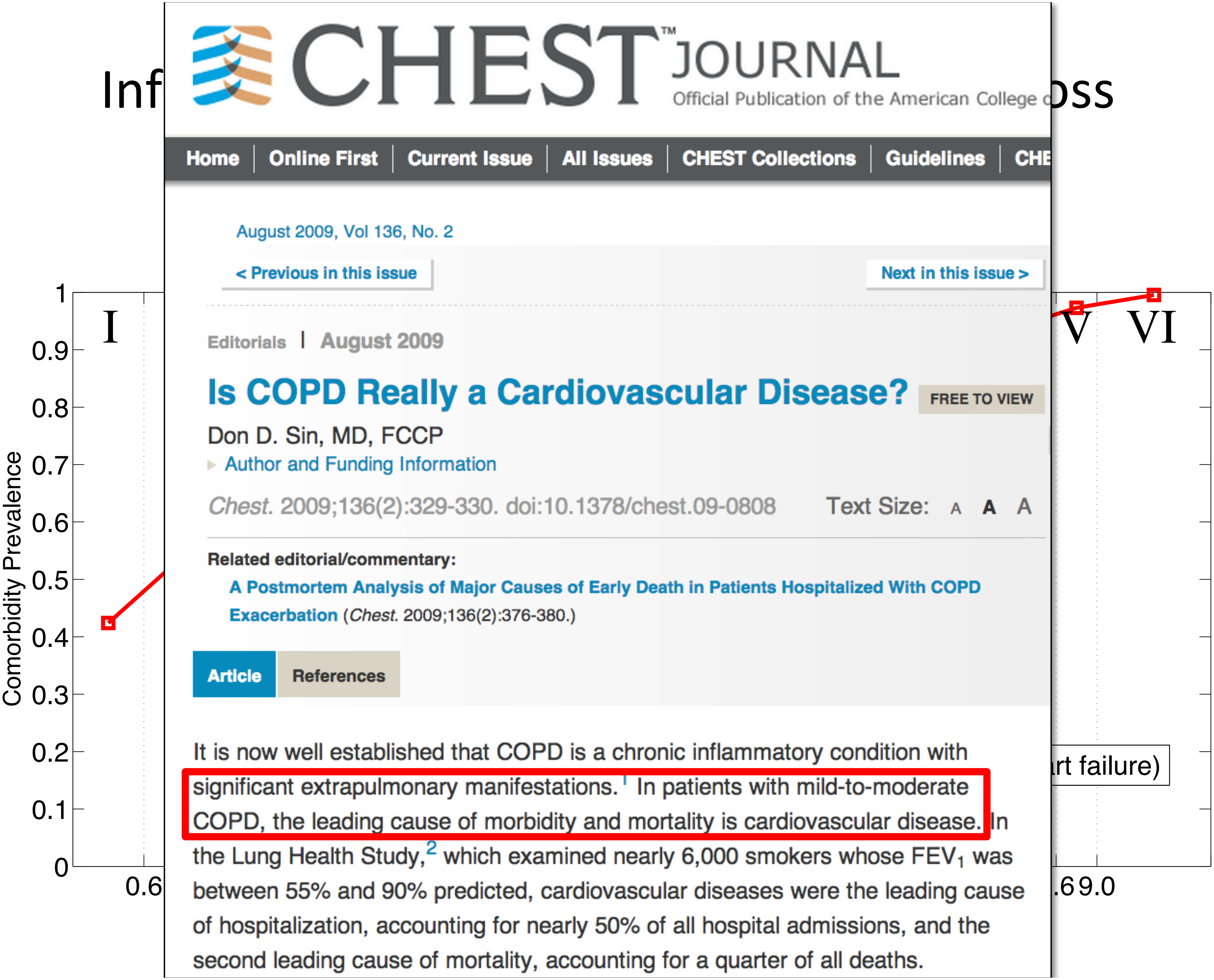


Inferred prevalence of comorbidities across stages (Diabetes & Musculoskeletal disorders)



Inferred prevalence of comorbidities across stages (Cardiovascular disease)





CHEST™ JOURNAL
Official Publication of the American College o

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.)

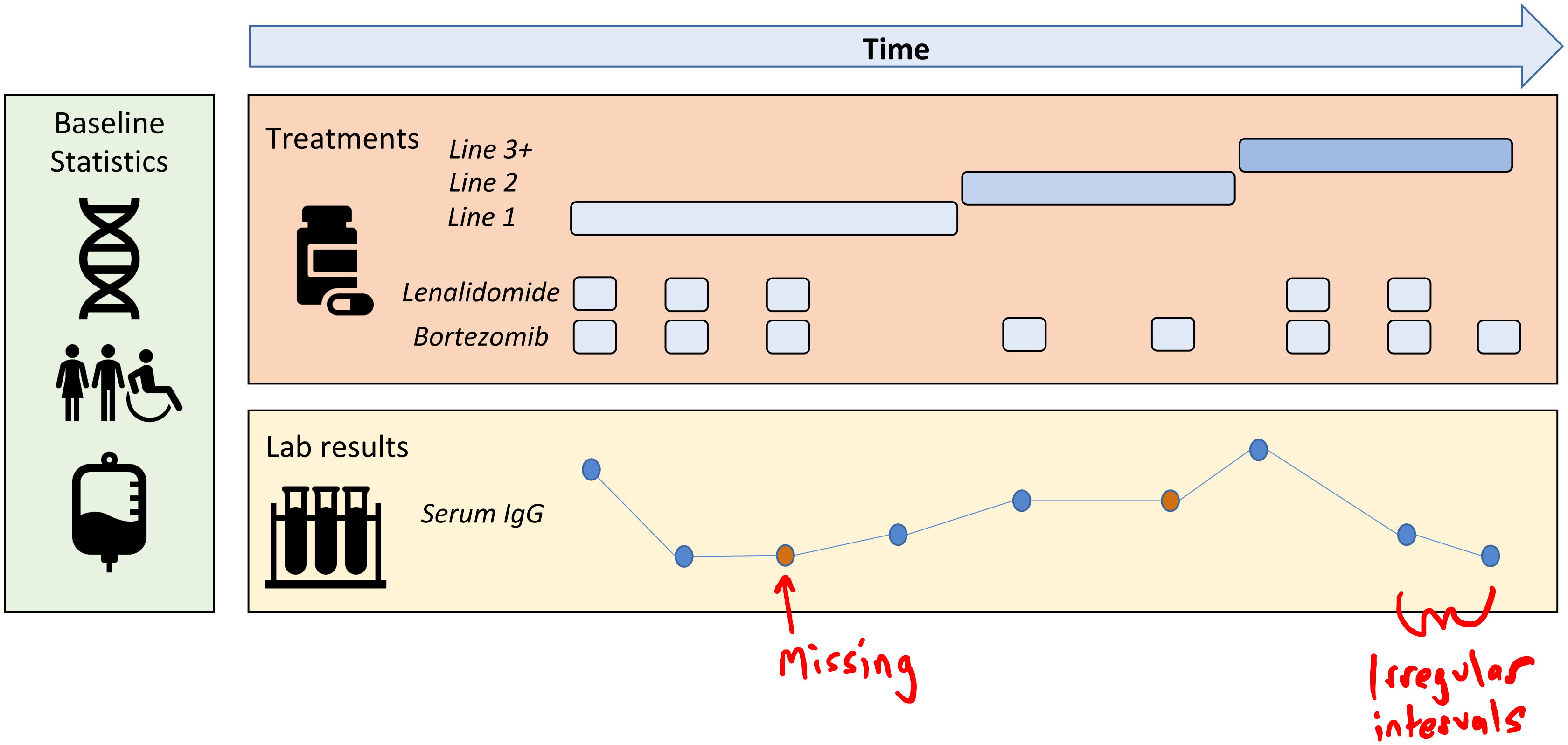
It is now well established that COPD is a chronic inflammatory condition with significant extrapulmonary manifestations.[1] In patients with mild-to-moderate COPD, the leading cause of morbidity and mortality is cardiovascular disease. In the Lung Health Study,[2] which examined nearly 6,000 smokers 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.

Goals of disease progression modeling

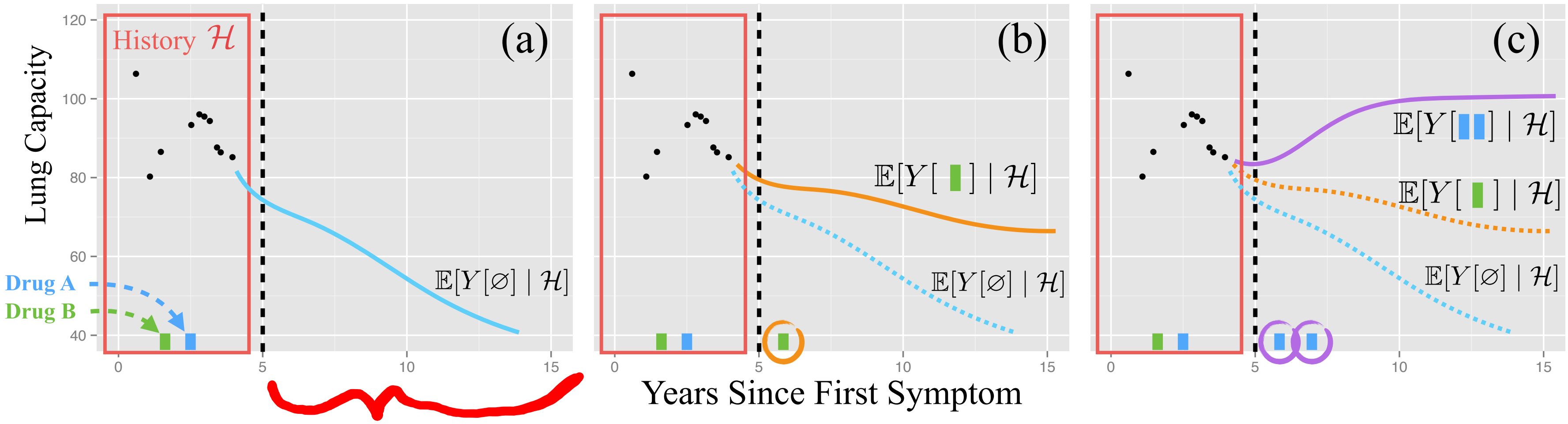
- Descriptive:
  - *Find markers of disease stage and progression, statistics of what to expect when*
- **Predictive:**
  - ***What will this patient's future trajectory look like?***
  - ***How will treatment affect it?***
- Key challenges:
  - Seldom directly observe disease stage, but rather only indirect observations (e.g. symptoms, lab results)
  - Data can be censored – don't observe beginning to end

Challenges for modeling

- Irregular time intervals between observations
- Missing data
- Treatment effects



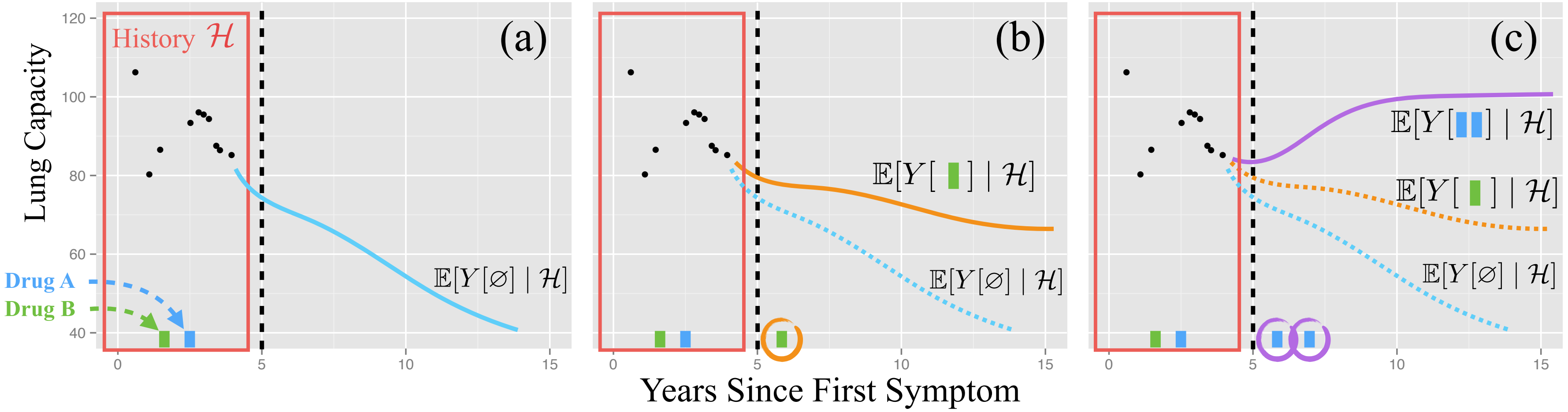
Counterfactual Gaussian Processes



Predict

Schulam & Saria, Reliable Decision Support using Counterfactual Models, NeurIPS 2017

Counterfactual Gaussian Processes



- Causal assumptions:
  - Policy used to choose actions in observational data did not depend on unobserved information that is predictive of future potential outcomes
  - Measurement times independent of measurement values, conditioned on history

Schulam & Saria, Reliable Decision Support using Counterfactual Models, NeurIPS 2017

Counterfactual Gaussian Processes

$$\theta = \Big\{ \underset{i^{th}\ \text{patient}}{h_i} = \Big\{ \big(\overset{\text{Times}}{t_{ij}}, \overset{\text{Values}}{\underset{\in \mathbb{R} \cup \phi}{y_{ij}}}, \overset{\text{Actions}}{\underset{C \cup \phi}{a_{ij}}}\big) \Big\}_{j=1}^{n_i} \Big\}_{i=1}^{m}$$

$m \leftarrow$ # patients

Learning

$$\max \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log P\Big(y_{ij} \,\Big|\, \big\{(t_{ik}, y_{ik}, a_{ik})\big\}_{k<j},\ t_{ij}, a_{ij}\Big)$$

Schulam & Saria, Reliable Decision Support using Counterfactual Models, NeurIPS 2017

Counterfactual Gaussian Processes

$$\theta = \Big\{ \underset{\text{patient}}{h_i} = \Big\{ \big(\overset{\text{Times}}{\underset{\in \mathbb{R} \cup \phi}{t_{ij}}}, \overset{\text{Values}}{y_{ij}}, \overset{\text{Actions}}{\underset{C \cup \phi}{a_{ij}}} \big) \Big\}_{j=1}^{n_i} \Big\}_{i=1}^{m}$$

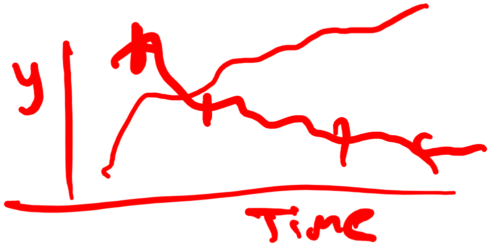
$m \leftarrow$ # patients

Learning

$$\max \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log p\Big(\underline{y_{ij}} \;\Big|\; \{\underline{(t_{ik}, y_{ik}, a_{ik})}\}_{k<j}, \underline{t_{ij}}, \underline{a_{ij}} \Big)$$

How to parametrize? If no actions, use a Gaussian process:

$$\vec{y_i} \mid \vec{t_i} \sim N(\cancel{0}, \Sigma(\vec{t}))$$



Actions using a mean function

For each $c \in C$, $g_c(\delta) \leftarrow \delta$ time since action began

$$\rightarrow \mathbb{E}[y_{ij} \mid \ldots, t_{ij}] = \sum_{k=1}^{j-1} g_{a_{ik}}(t_{ij} - t_{ik}) \qquad g_{\phi k} = 0.$$

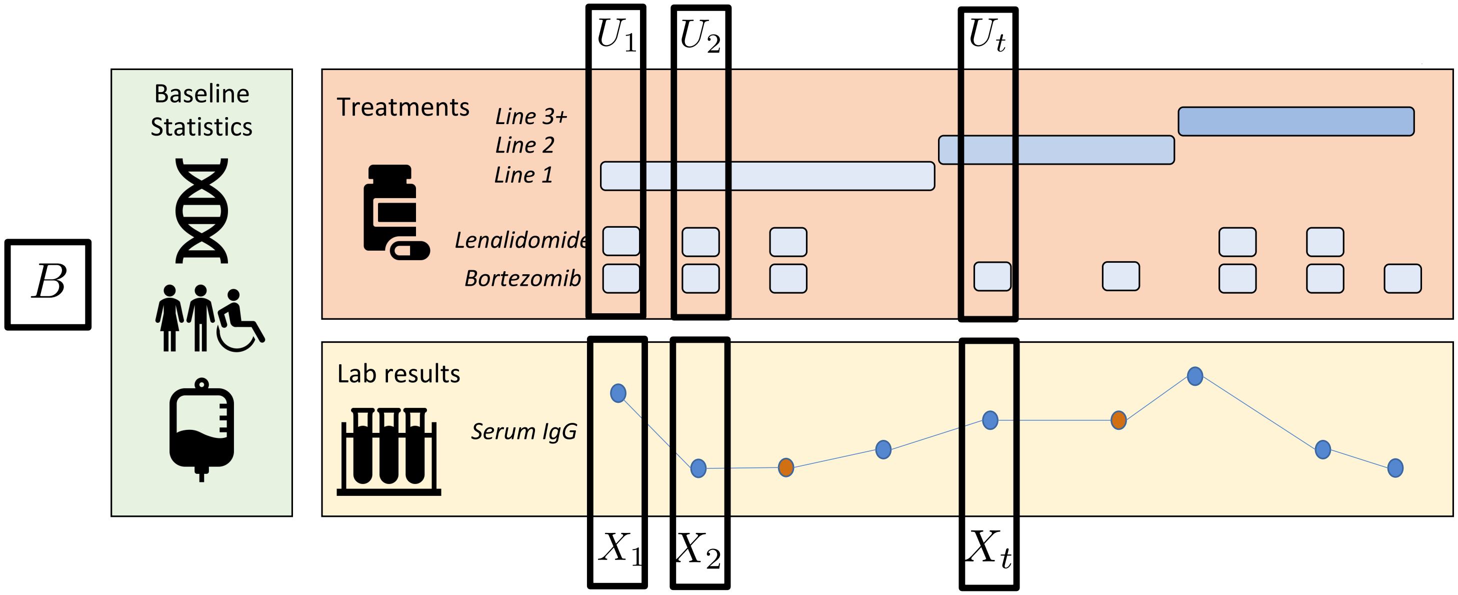
Schulam & Saria, Reliable Decision Support using Counterfactual Models, NeurIPS 2017

Limitations of CGPs

- Models a single biomarker across time
- Limited ability to condition on baseline information
- Treatment response functions are additive

Neural pharmacodynamic state space models

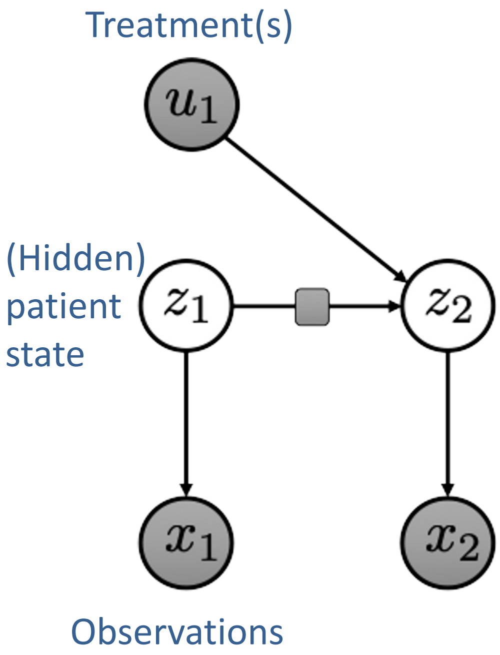
Learn using: $\text{maximize} \sum_{i=1}^{N} \log p(\mathbf{X}^i | \mathbf{U}^i, B^i)$



Hussain, Krishnan, Sontag, Neural Pharmacodynamic State Space Models, ICML 2021

Neural pharmacodynamic state space models

Learn using: $\text{maximize} \sum_{i=1}^{N} \log p(\mathbf{X}^i | \mathbf{U}^i, B^i)$



Treatment(s)

(Hidden) patient state

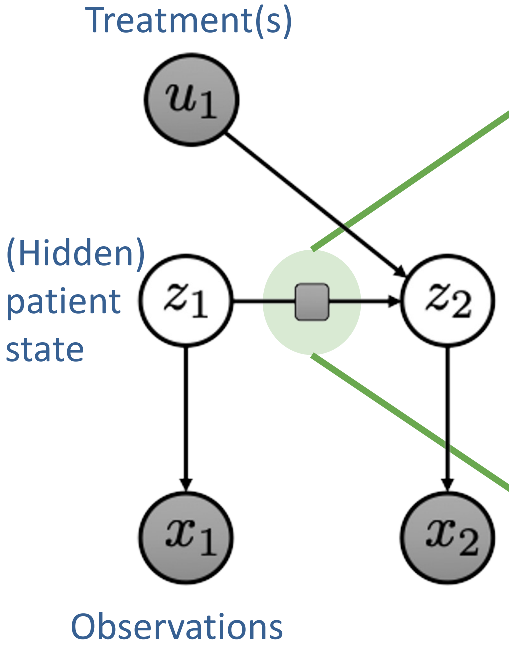
Observations

$$Z_t|\cdot \sim \mathcal{N}(\mu_\theta(Z_{t-1}, U_{t-1}, B), \Sigma_\theta^t(Z_{t-1}, U_{t-1}, B)),$$
$$X_t|\cdot \sim \mathcal{N}(\kappa_\theta(Z_t), \Sigma_\theta^e(Z_t))$$

Krishnan, Shalit & Sontag, Structured inference networks for nonlinear state space models, AAAI 2017

Neural pharmacodynamic state space models

Learn using: $\text{maximize} \sum_{i=1}^{N} \log p(\mathbf{X}^i|\mathbf{U}^i,B^i)$



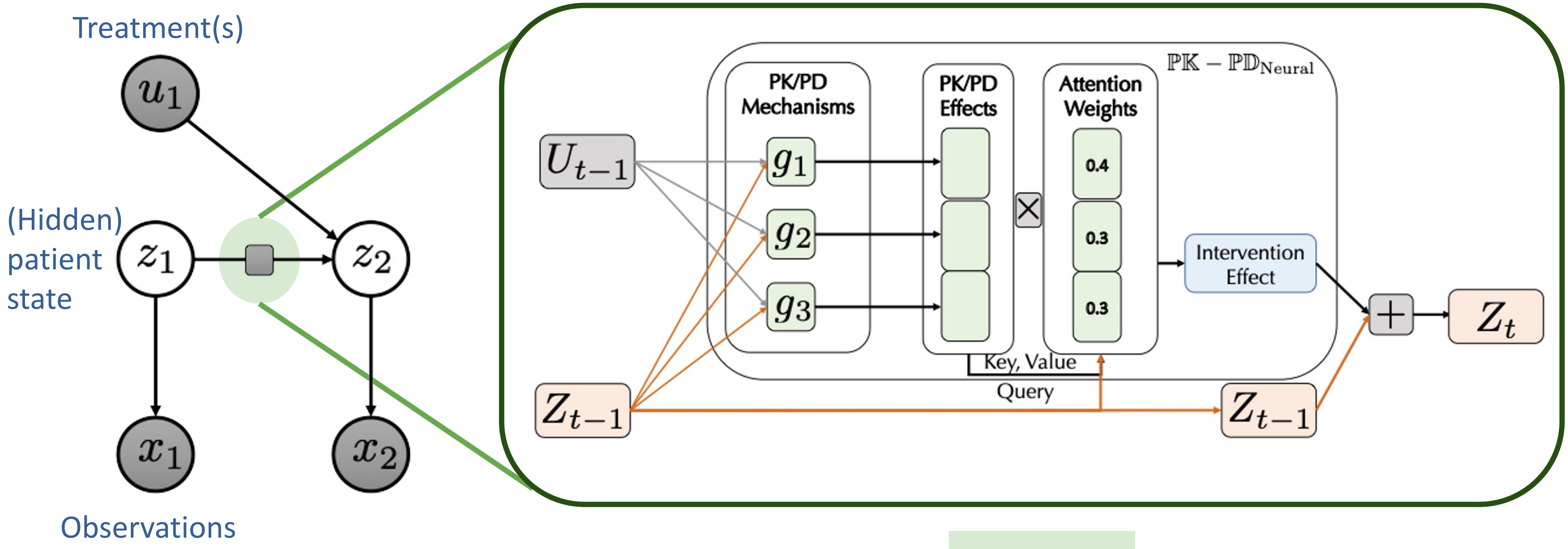
Can we use domain knowledge to parameterize the transition distributions?

- **Lines of therapy**
- **Mechanism of drug-effect**

Krishnan, Shalit & Sontag, Structured inference networks for nonlinear state space models, AAAI 2017
Hussain, Krishnan, Sontag, Neural Pharmacodynamic State Space Models, ICML 2021

Neural pharmacodynamic state space models

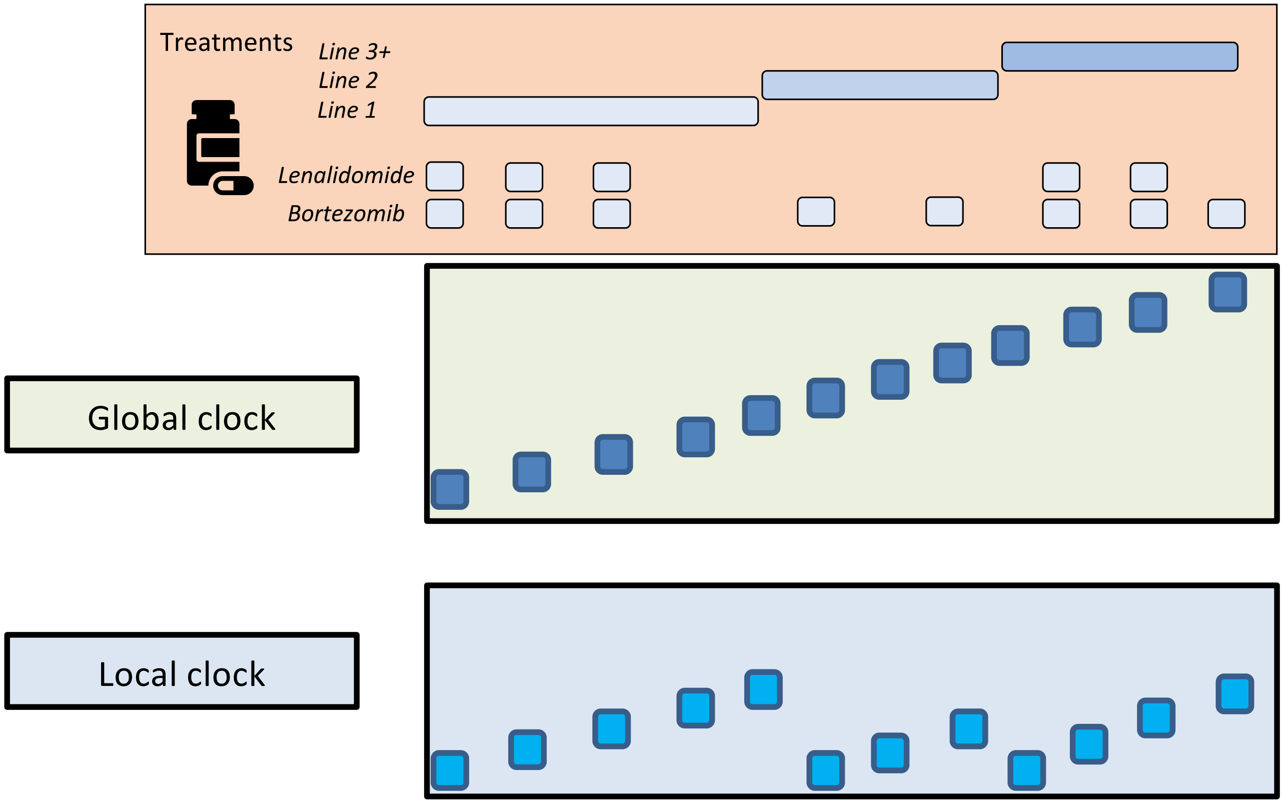
Learn using: $\text{maximize} \sum_{i=1}^{N} \log p(\mathbf{X}^i|\mathbf{U}^i, B^i)$



SSM PK-PD

Krishnan, Shalit & Sontag, Structured inference networks for nonlinear state space models, AAAI 2017
Hussain, Krishnan, Sontag, Neural Pharmacodynamic State Space Models, ICML 2021

From lines of therapy to local and global clocks



Hussain, Krishnan, Sontag, Neural Pharmacodynamic State Space Models, ICML 2021

Neural intervention effect functions

- Modeling baseline conditional variation

$$g_1(Z_{t-1},U_{t-1},B) = Z_{t-1} \cdot \tanh(b_{\text{lin}} + W_{\text{lin}}[U_{t-1},B])$$

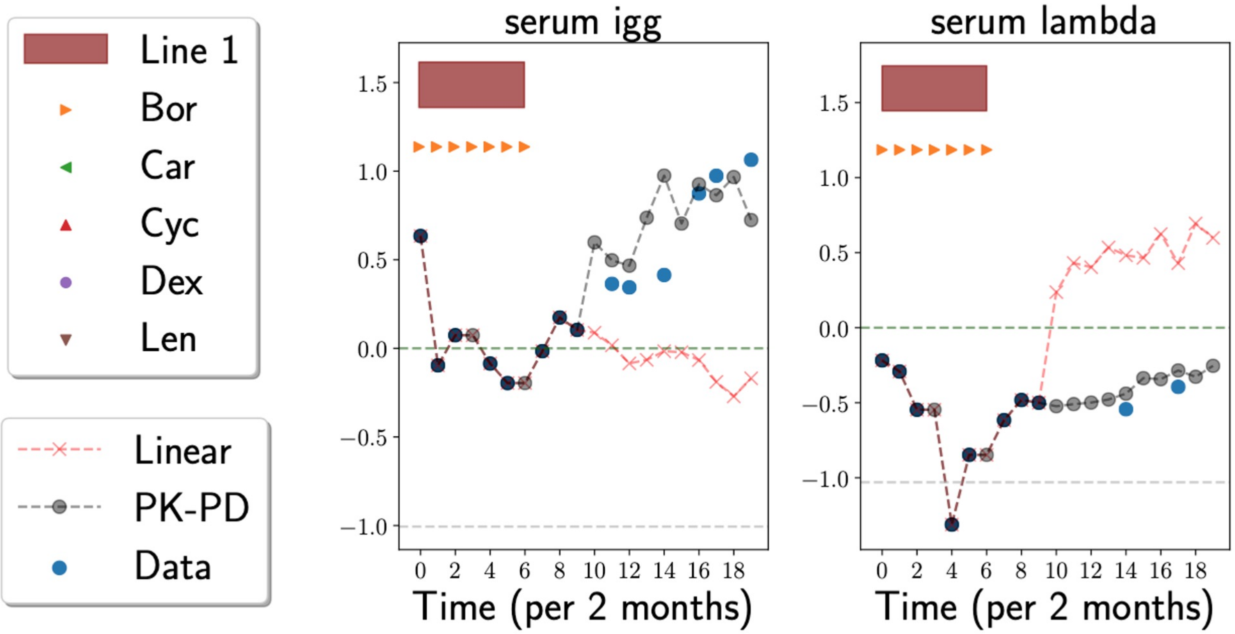
- Modeling slow gradual relapse after treatment
  - Log-cell kill

$$g_2(Z_{t-1},U_{t-1},B) = Z_{t-1} \cdot (1 - \rho \log(Z_{t-1}^2) \\ - \beta \exp(-\delta \cdot \text{lc}_{t-1})),$$

  - Captures rapid variation in representations due to treatment

$$g_3(Z_{t-1},U_{t-1},B) \\ = \begin{cases} b_0 + \alpha_{1,t-1}/[1 + \exp(-\alpha_{2,t-1}(\text{lc}_{t-1} - \frac{\gamma_l}{2}))], \\ \text{if } 0 \leq \text{lc}_{t-1} < \gamma_l \\ b_l + \alpha_{0,t-1}/[1 + \exp(\alpha_{3,t-1}(\text{lc}_{t-1} - \frac{3\gamma_l}{2}))], \\ \text{if } \text{lc}_{t-1} \geq \gamma_l \end{cases}$$

Example of using SSM PK-PD to predict future clinical biomarkers



on real world dataset (multiple myeloma)

(normalized) **healthy maximum** value

(normalized) **healthy minimum** value

forward samples after observing the patient for **15 months**

Hussain, Krishnan, Sontag, Neural Pharmacodynamic State Space Models, ICML 2021

Conclusion

- Many open questions
  - Is it 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

Additional references for disease subtyping

- Cluster Analysis and Clinical Asthma Phenotypes (discussed in class)
  Haldar et al., Am J Respir Crit Care Med. 2008.
  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3992366/pdf/emss-29902.pdf
- Phenomapping for Novel Classification of Heart Failure with Preserved Ejection Fraction
  Shah et al., Circulation 2015
  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4302027/
- Subtyping: What It Is and Its Role in Precision Medicine
  Saria & Goldberg, IEEE Intelligent Systems 2015
  https://www.dropbox.com/s/krofvs7da6u3r4k/Saria_IEEE2015_SubtypingAndPredicionMedicine.pdf
- Comorbidity Clusters in Autism Spectrum Disorders: An Electronic Health Record Time-Series Analysis
  Doshi-Velez, Ge, Kohane. Pediatrics, 2014. https://www.ncbi.nlm.nih.gov/pubmed/24323995
- Learning Probabilistic Phenotypes from Heterogeneous EHR Data
  Pivovarov, et al. Journal of Biomedical Informatics 2015
  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8025140/
- A Bayesian Nonparametric Model for Disease Subtyping: Application to Emphysema Phenotypes
  Ross et al., IEEE Transactions on Medical Imaging, 2017
  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5267575/
- Clustering Interval-Censored Time-Series for Disease Phenotyping. Chen, Krishnan, Sontag. AAAI 2022. https://arxiv.org/pdf/2102.07005.pdf

Additional references for disease progression modeling

- Unsupervised Learning of Disease Progression Models
Wang, Sontag, Wang., KDD 2014
https://people.csail.mit.edu/dsontag/papers/WanSonWan_kdd14.pdf
- Cross-Corpora Unsupervised Learning of Trajectories in Autism Spectrum Disorders
Elibol et al., JMLR 2016
https://www.jmlr.org/papers/volume17/15-431/15-431.pdf
- Modeling Disease Progression via Fused Sparse Group Lasso
Zhou et al., KDD ’12
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4191837/
- Attentive State-Space Modeling of Disease Progression
Alaa & van der Schaar, NeurIPS 2019
https://openreview.net/pdf?id=BkllWHBxUH
- Constructing Disease Network and Temporal Progression Model via Context-Sensitive Hawkes Process
Choi et al., IEEE International Conference on Data Mining, 2015
https://www.cc.gatech.edu/grads/e/echoi48/docs/icdm2015.pdf
- Neural pharmacodynamic state space modeling. Hussain, Krishnan, Sontag. ICML 2021. http://proceedings.mlr.press/v139/hussain21a/hussain21a.pdf