MACHINE LEARNING FOR HEALTHCARE

6.S897, HST.S53

Lecture 12: Disease progression modeling

Prof. David Sontag

MIT EECS, CSAIL, IMES

Massachusetts Institute of Technology

Outline of today’s class

1. Multi-task learning of (measurable) disease progression
   - **Application to Alzheimer’s disease (Zhou et al., KDD ‘12)**

2. Discovering fine-grained disease states using hidden Markov models
   - **Application to Alzheimer’s disease (Sukkar et al., IEEE EMBS ’12)**

3. Unsupervised learning of (grounded, multi-dimensional) disease progression models
   - **Application to chronic obstructive pulmonary disease (Wang et al., KDD ‘14)**

Chronic diseases

- A **chronic disease** is a human health condition that persists or otherwise is long-lasting in its effects
- E.g., lasting for more than 3 months
- Common chronic diseases include:
  - Arthritis
  - Asthma
  - Cancer
  - Heart failure
  - Diabetes
  - Hepatitis C
  - HIV/AIDS

[Slide credit: Farzad Kamalzadeh]

Epidemiology

- Chronic diseases constitute a major cause of mortality
  - WHO: 38 million deaths a year to non-communicable diseases
  - United States: 25% of adults have at least two chronic conditions
  - 1 in 2 Americans (133 million) have at least one chronic medical condition
  - 61% of deaths among people older than 65 in the population

- Diabetes
  - 7th leading cause of death in the US
  - Leading cause of complications such as kidney failure, non-traumatic lower limb amputations, blindness
  - Major cause of heart disease

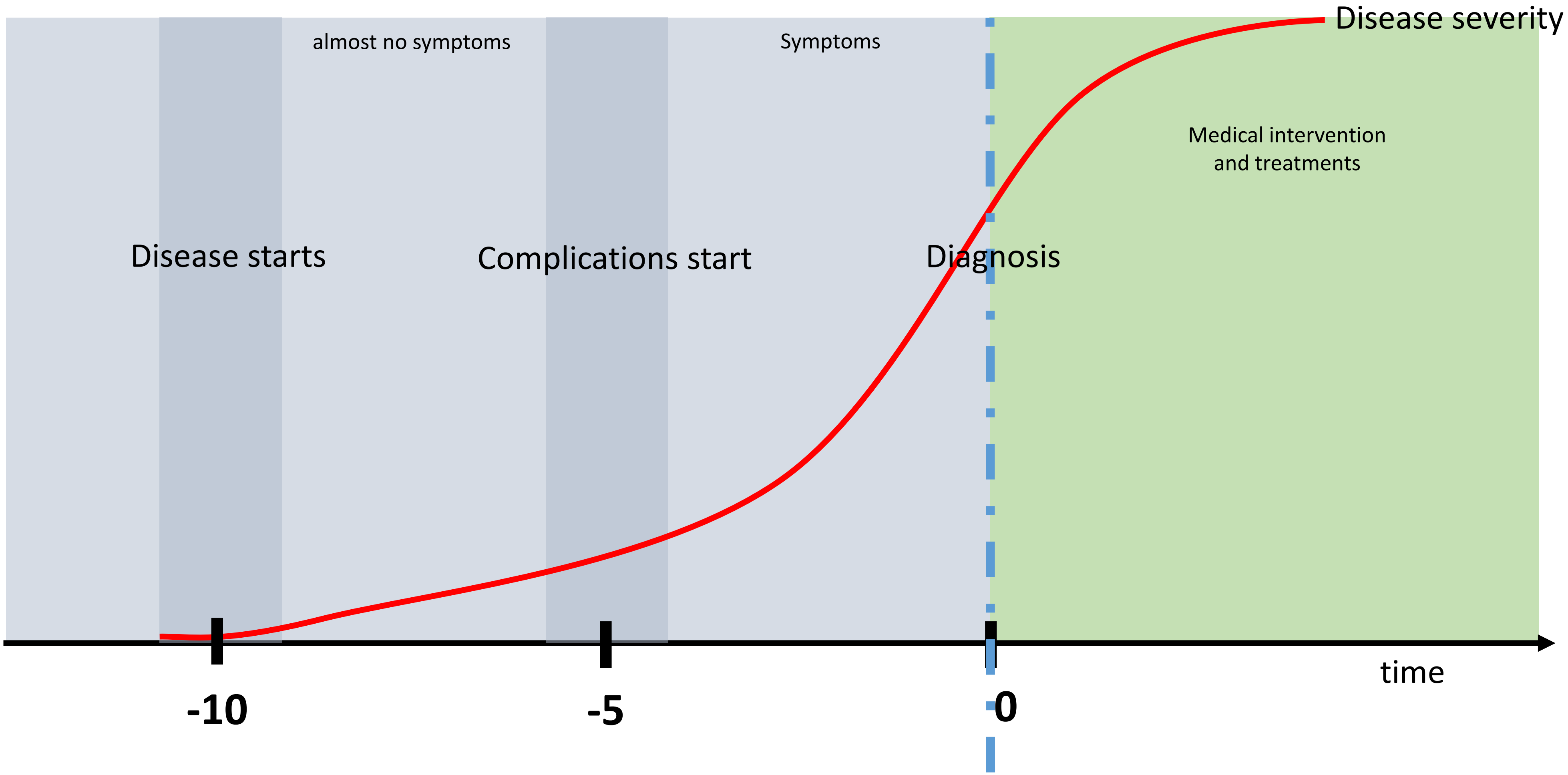
[Slide credit: Farzad Kamalzadeh]

Economic impact

- Chronic diseases constitute a major section of medical care spending (direct costs):
  - **75%** of the $2 trillion spent annually in US medical care
  - Diabetes: $1 in $3 Medicare expenditure
- (indirect costs)
  - Limitations in daily activities
  - Loss in productivity
  - Loss in days of work
- Diabetes: $322 billion per year

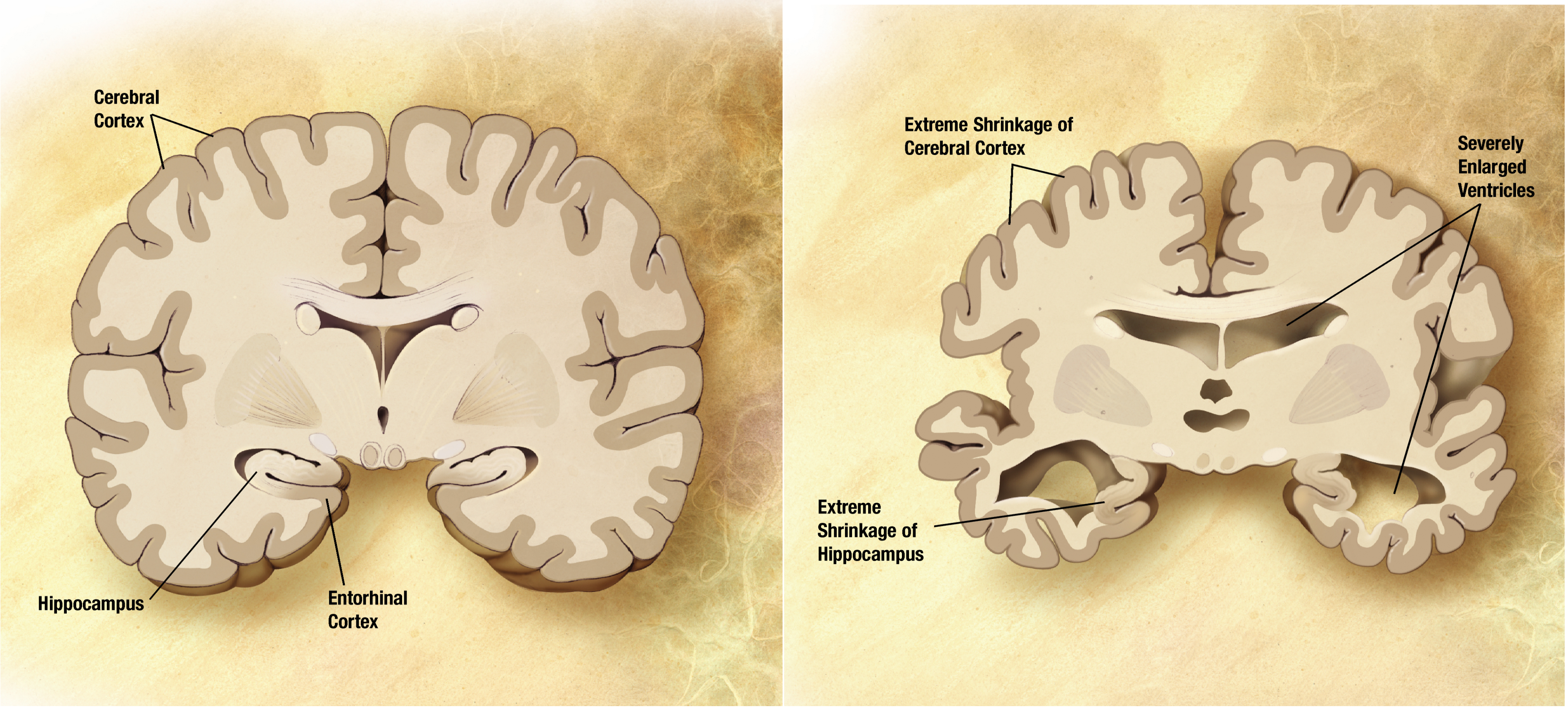
[Slide credit: Farzad Kamalzadeh]

Nature of chronic diseases



[Image credit: Farzad Kamalzadeh]

Predicting disease progression in Alzheimer's disease


Cerebral Cortex
Hippocampus
Entorhinal Cortex
Extreme Shrinkage of Cerebral Cortex
Severely Enlarged Ventricles
Extreme Shrinkage of Hippocampus

[Image credit: Wikipedia; "Alzheimer's Disease Education and Referral Center, a service of the National Institute on Aging."]

Predicting disease progression in Alzheimer's disease

- Goal: Predict disease status in *6 months*, *12 months*, *24 months*, *36 months*…
- Rather than learn several independent models, view as *multi-task* learning:
  - Select a common set of biomarkers for al time points
  - Also allow for specific set of biomarkers at different time points
  - Incorporate temporal smoothness in models

[Zhou et al., KDD '12]

Predicting disease progression in Alzheimer’s disease

- Number of patients X months after baseline (Alzheimer’s Disease Neuroimaging Initiative):

M06	M12	M24	M36	M48
648	642	569	389	87

M06 = 6 months after baseline

[Zhou et al., KDD ’12]

Convex fused sparse group lasso

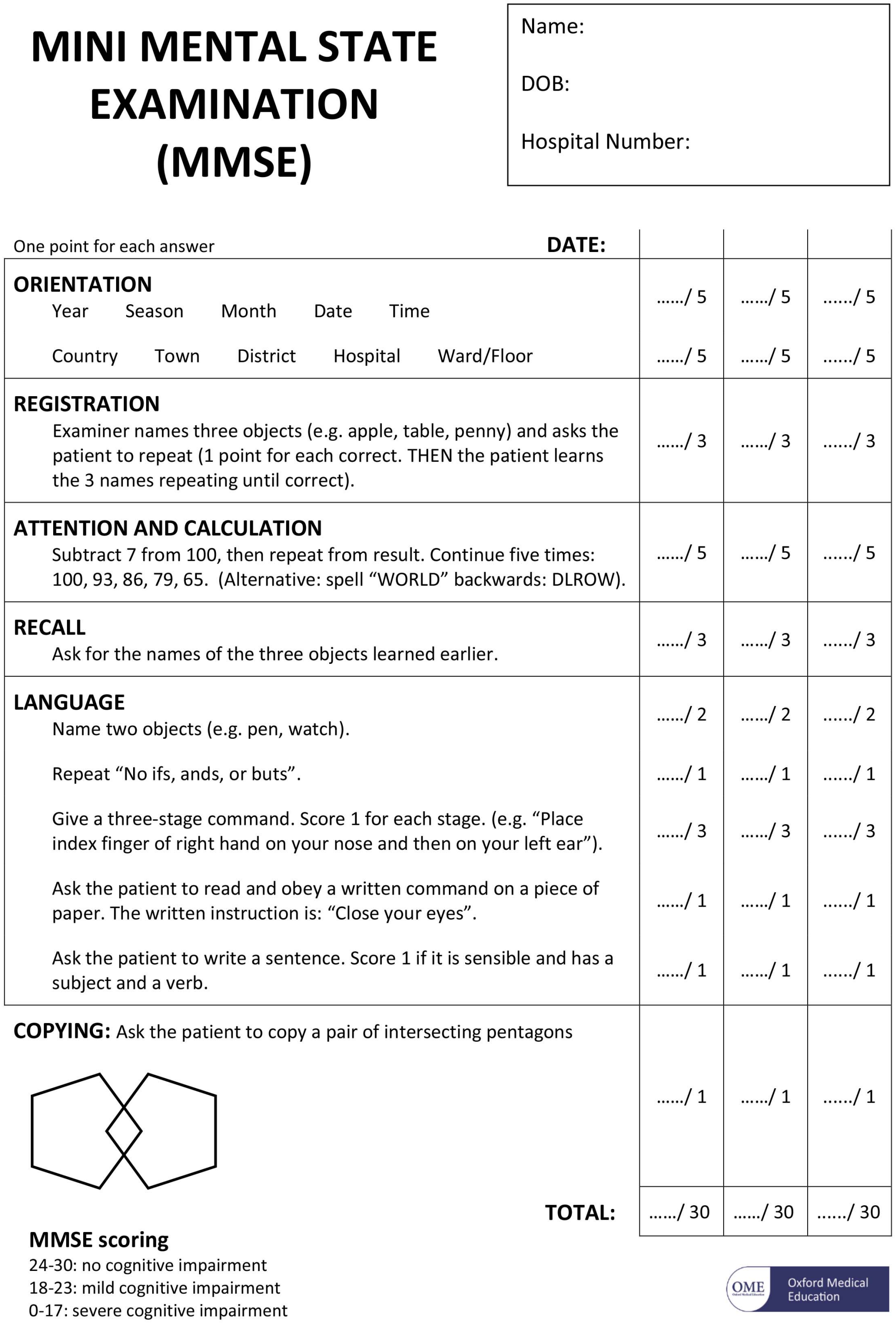
- Simultaneously learn all 5 models by solving the following convex optimization problem:

$$\min_{W} L(W) + \lambda_1 \|W\|_1 + \lambda_2 \left\|RW^T\right\|_1 + \lambda_3 \|W\|_{2,1}$$

- Squared loss: $L(W) = \|S \odot (XW - Y)\|_F^2$
(*S* accounts for labels that might be missing in a subset of the tasks)
- Group Lasso penalty $\|W\|_{2,1}$ given by $\sum_{i=1}^{d} \sqrt{\sum_{j=1}^{t} W_{ij}^2}$
- R = $$\begin{array}{cc} & T \\ T-1 & \boxed{\begin{matrix} 1 & -1 & & \\ & 1 & -1 & \\ & & 1 & -1 \end{matrix}} \end{array}$$

[Zhou et al., KDD ’12]

Outcome (label) derived from clinical score:



MINI MENTAL STATE EXAMINATION (MMSE)

Name:

DOB:

Hospital Number:

One point for each answer **DATE:**			
ORIENTATION Year Season Month Date Time	....../ 5	....../ 5	....../ 5
Country Town District Hospital Ward/Floor	....../ 5	....../ 5	....../ 5
REGISTRATION Examiner names three objects (e.g. apple, table, penny) and asks the patient to repeat (1 point for each correct. THEN the patient learns the 3 names repeating until correct).	....../ 3	....../ 3	....../ 3
ATTENTION AND CALCULATION Subtract 7 from 100, then repeat from result. Continue five times: 100, 93, 86, 79, 65. (Alternative: spell “WORLD” backwards: DLROW).	....../ 5	....../ 5	....../ 5
RECALL Ask for the names of the three objects learned earlier.	....../ 3	....../ 3	....../ 3
LANGUAGE Name two objects (e.g. pen, watch).	....../ 2	....../ 2	....../ 2
Repeat “No ifs, ands, or buts”.	....../ 1	....../ 1	....../ 1
Give a three-stage command. Score 1 for each stage. (e.g. “Place index finger of right hand on your nose and then on your left ear”).	....../ 3	....../ 3	....../ 3
Ask the patient to read and obey a written command on a piece of paper. The written instruction is: “Close your eyes”.	....../ 1	....../ 1	....../ 1
Ask the patient to write a sentence. Score 1 if it is sensible and has a subject and a verb.	....../ 1	....../ 1	....../ 1
COPYING: Ask the patient to copy a pair of intersecting pentagons	....../ 1	....../ 1	....../ 1
TOTAL:	....../ 30	....../ 30	....../ 30

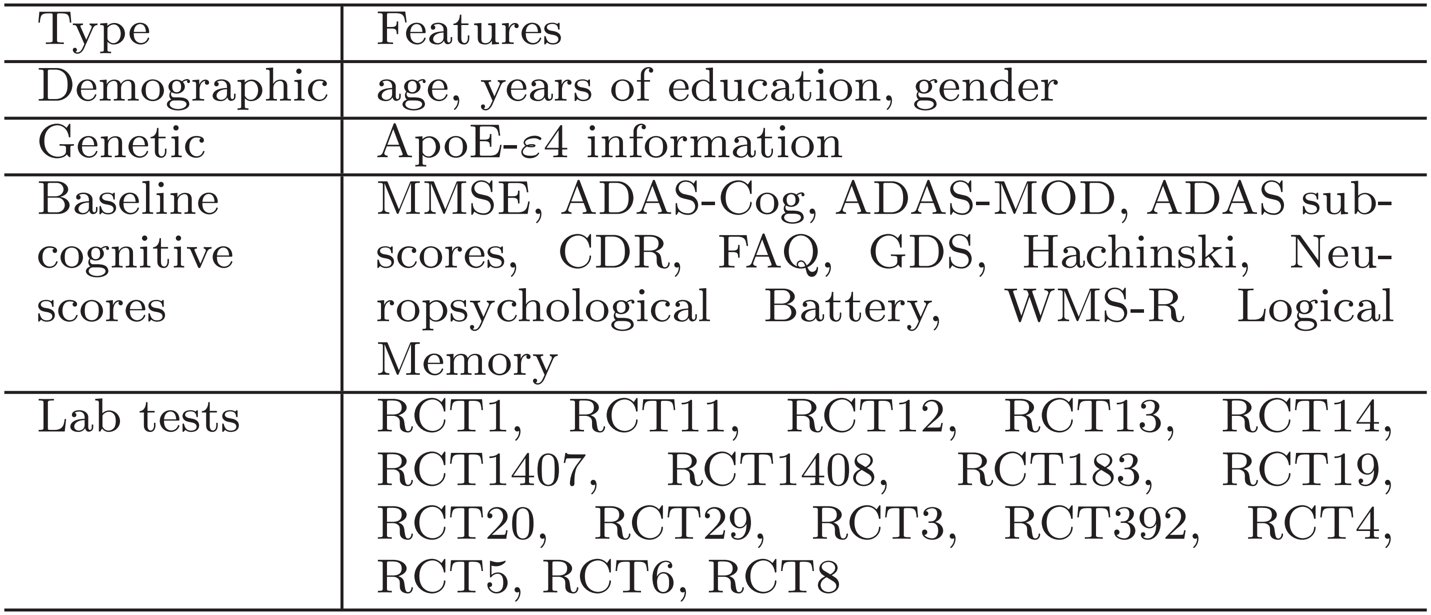
MMSE scoring

24-30: no cognitive impairment
18-23: mild cognitive impairment
0-17: severe cognitive impairment

Oxford Medical Education

Predicting disease progression in Alzheimer's disease

- Features considered:



Type	Features
Demographic	age, years of education, gender
Genetic	ApoE-ε4 information
Baseline cognitive scores	MMSE, ADAS-Cog, ADAS-MOD, ADAS sub-scores, CDR, FAQ, GDS, Hachinski, Neuropsychological Battery, WMS-R Logical Memory
Lab tests	RCT1, RCT11, RCT12, RCT13, RCT14, RCT1407, RCT1408, RCT183, RCT19, RCT20, RCT29, RCT3, RCT392, RCT4, RCT5, RCT6, RCT8

- 306 in total

[Zhou et al., KDD '12]

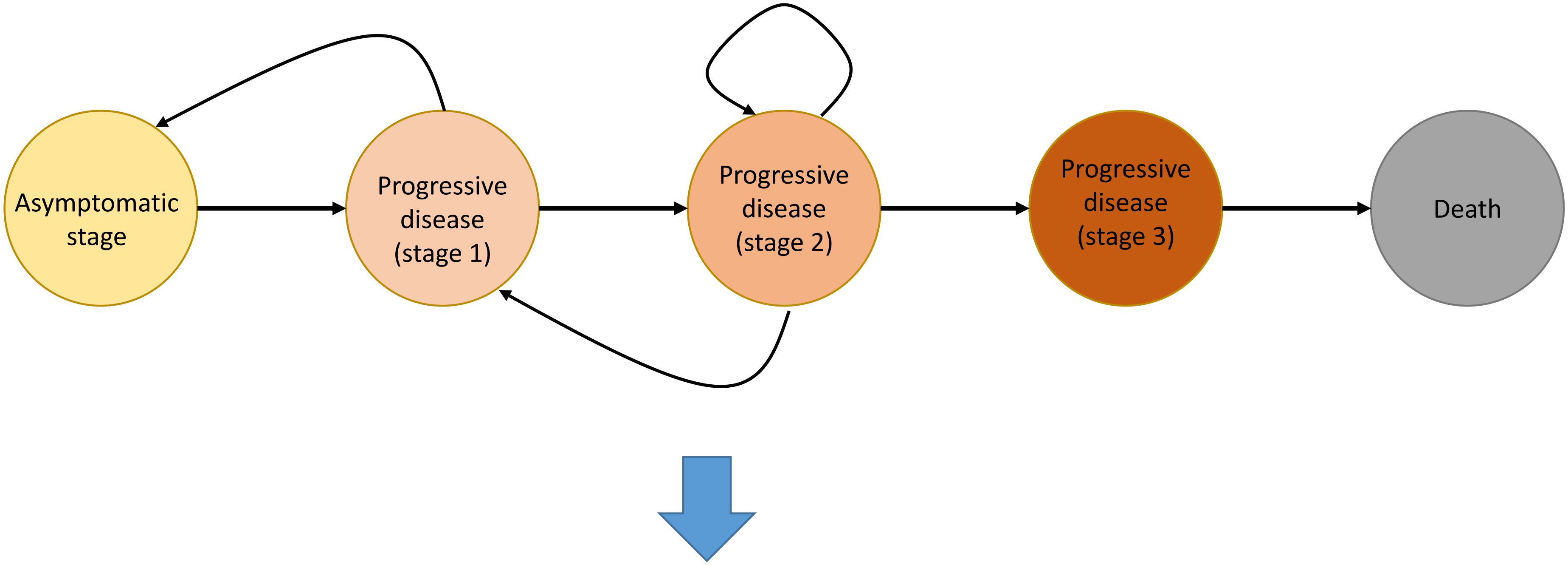
Outline of today’s class

1. Multi-task learning of (measurable) disease progression
   - **Application to Alzheimer’s disease (Zhou et al., KDD ‘12)**

2. Discovering fine-grained disease states using hidden Markov models
   - **Application to Alzheimer’s disease (Sukkar et al., IEEE EMBS ’12)**

3. Unsupervised learning of (grounded, multi-dimensional) disease progression models
   - **Application to chronic obstructive pulmonary disease (Wang et al., KDD ‘14)**

Disease progression

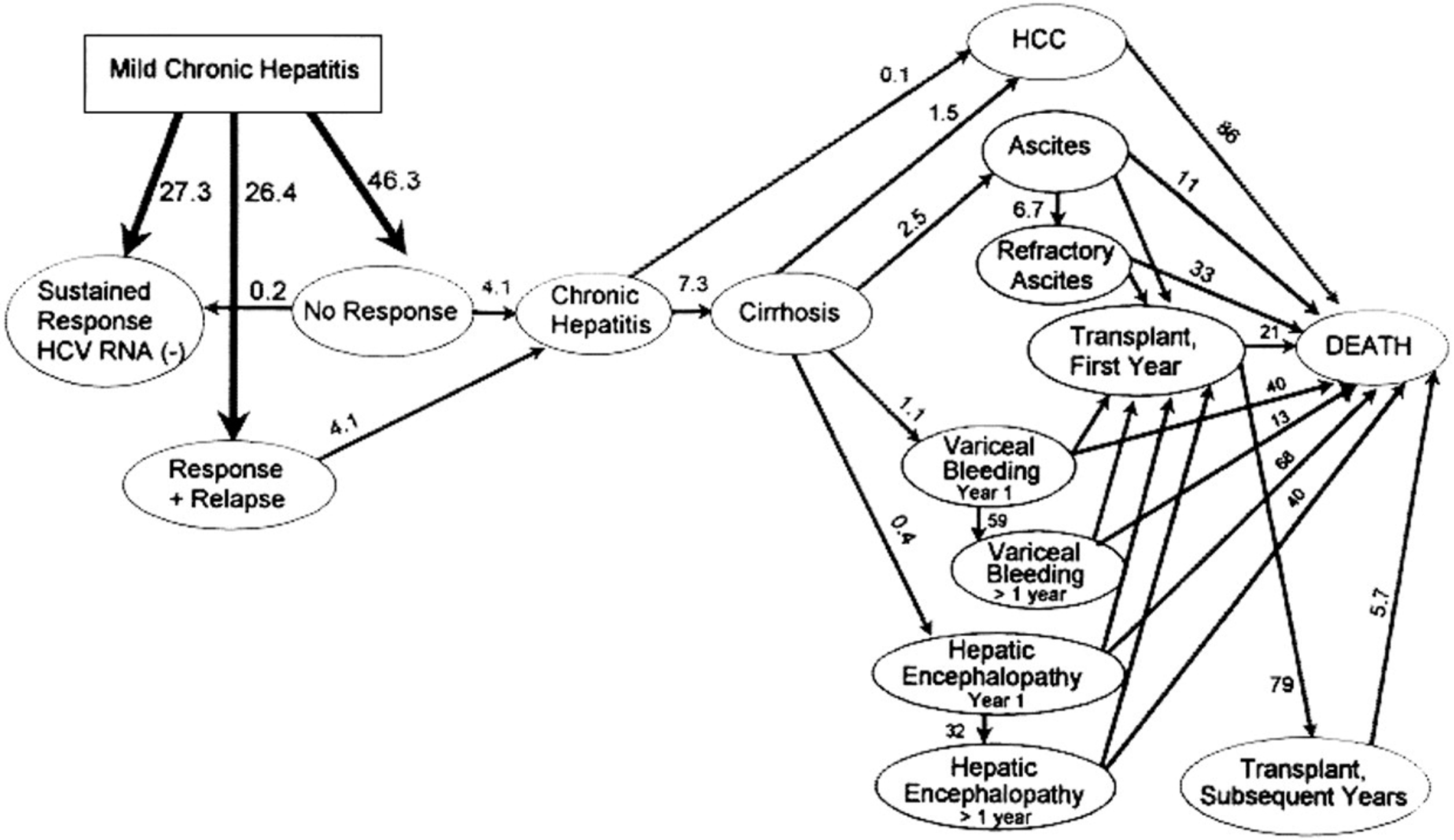

Asymptomatic stage
Progressive disease (stage 1)
Progressive disease (stage 2)
Progressive disease (stage 3)
Death

Markov Chain

[Image credit: Farzad Kamalzadeh]

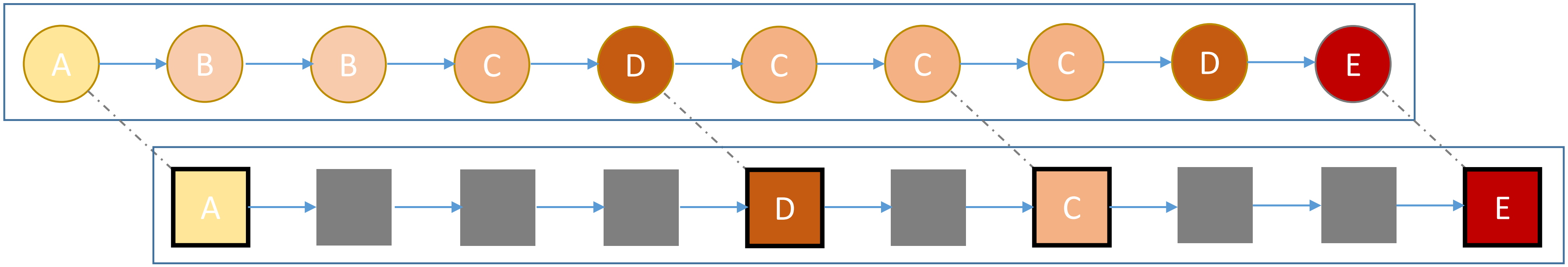
Markov models for disease progression

HCC = hepatocellular carcinoma



[Bennet et al, Estimates of the Cost-Effectiveness of a Single Course of Interferon-α2b in Patients with Histologically Mild Chronic Hepatitis C, *Annals of Internal Medicine*, 1997]

Estimating Markov models when there is missing data: *use Baum–Welch or EM*



	t=1	t=2	t=3	t=4	t=5	t=6	t=7	t=8	t=9	t=10
Patient 1	A				D		C			E
Patient 2	C		D	D		E				
Patient 3	A					B	A		A	
Patient 4		B		C						C
Patient 5	C		C					E		
Patient 6		D					D			E
Patient 7			B			C			B	

[Image credit: Farzad Kamalzadeh]

What if staging system is unknown, or incomplete?

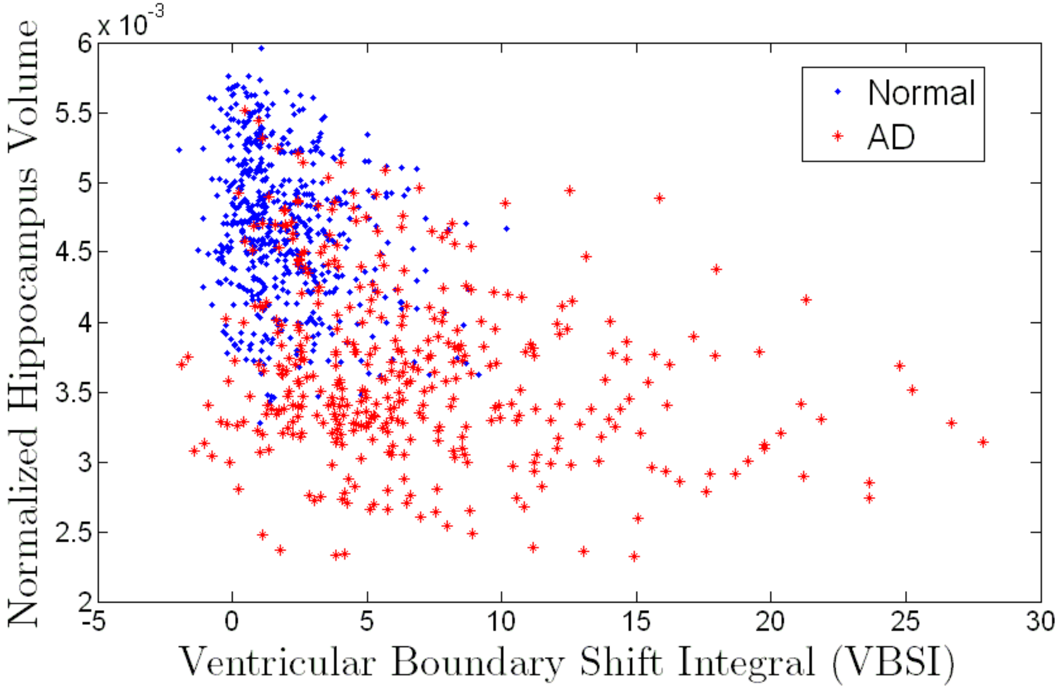
- 3 currently defined clinical stages of Alzheimer’s disease:
  - Normal
  - MCI (Mild Cognitive Impairment)
  - AD (Alzheimer’s disease)
- But, are there really just 3 stages?
- **Goal:** using clinical data, learn a *new* 6 stage system
- How does this relate to disease subtyping as discussed last week?

[Sukkar et al., IEEE EMBS ‘12]

Alzheimer’s disease neuroimaging dataset

- Alzheimer’s disease neuroimaging dataset:
  - 819 subjects
  - 229 “Normal” at beginning, 398 “MCI”, and 192 “AD”
  - Followed for up to 36 months with visits every 6 months

Brain ventricular and hippocampus volumes, as measured by MRI, correlated with AD diagnosis:



[Sukkar et al., IEEE EMBS ‘12]

HMM feature vector

- **We observe four features at each time point:**
  - Ventricular boundary shift integral (VBSI),
  - Hippocampus volume normalized by the skull volume,
  - Change in VBSI between two sucessive visits
  - Change in normalized hippocampus volume between two successive visits

- (A modern version of this study would use a deep generative model directly on the images)

[Sukkar et al., IEEE EMBS ‘12]

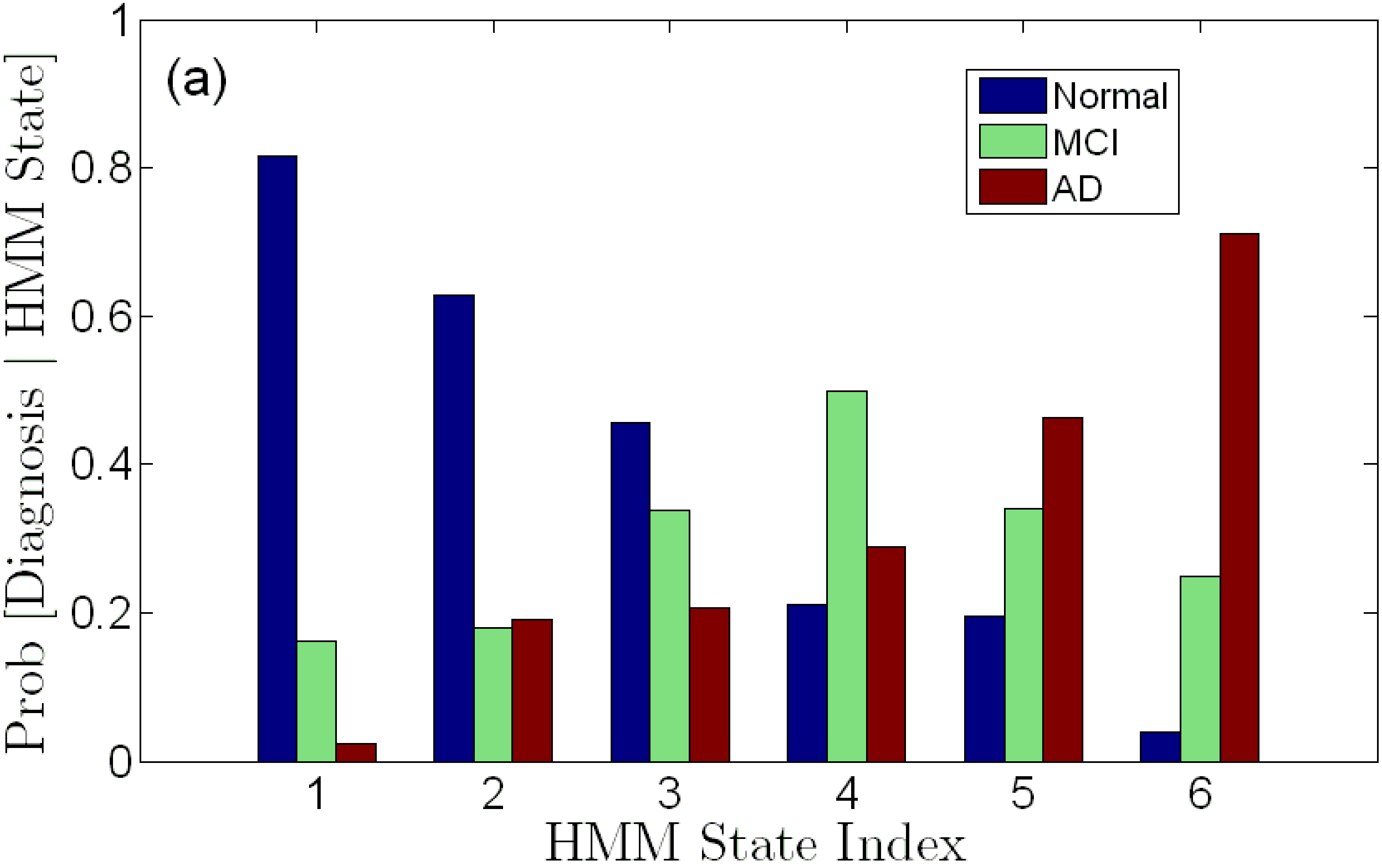
Results

- Each subject *regardless of clinical diagnosis at any of his/her visits* allowed to enter HMM at any state, end at any state
- HMM restricted to only allow transitions between neighboring states, e.g. 1<->2, 2<->3, …

[Sukkar et al., IEEE EMBS ‘12]

Results

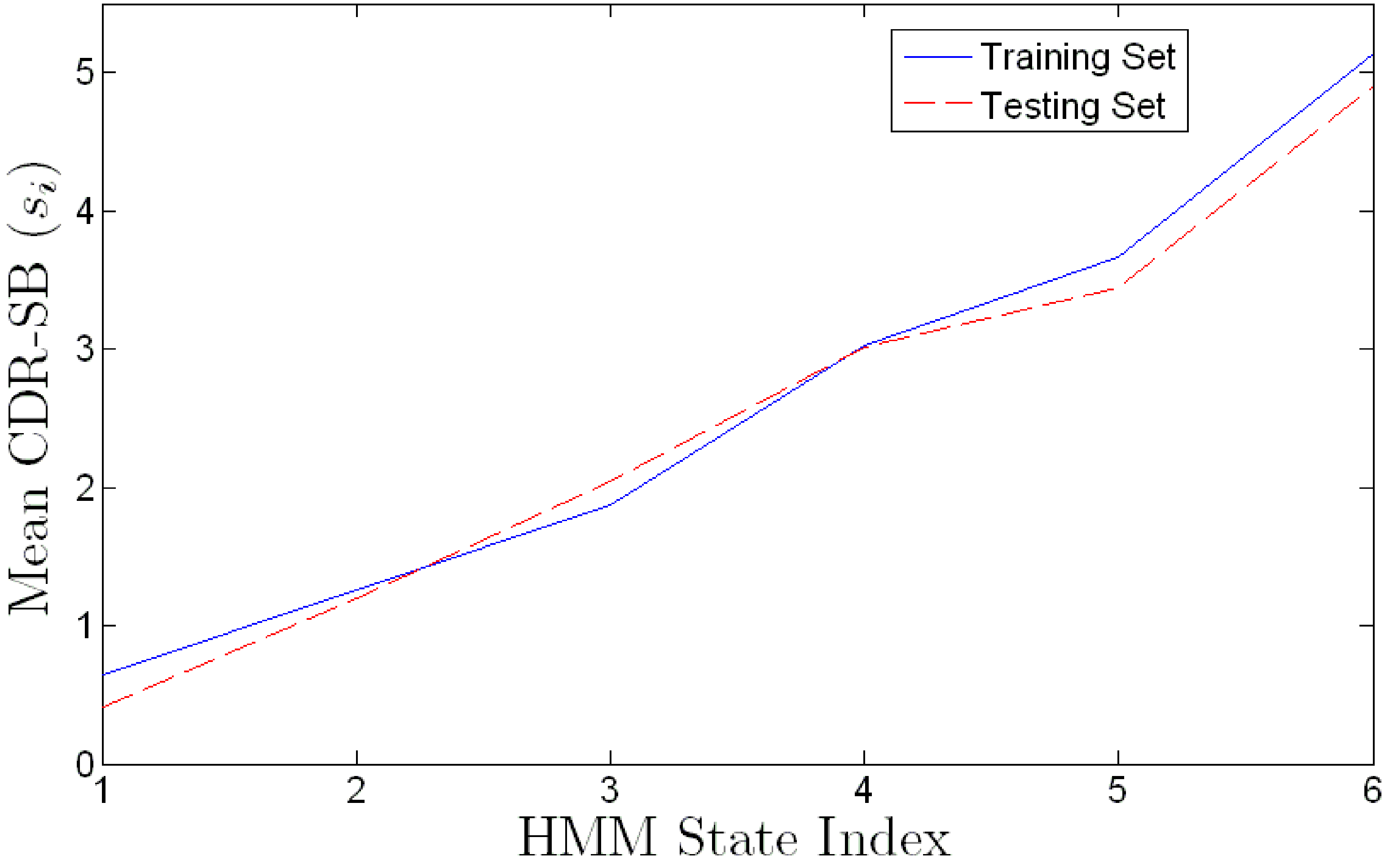
Based on MAP inference on held-out data:



[Sukkar et al., IEEE EMBS ‘12]

Results

Average Clinical Dementia Rating Scale Sum of Boxes (CDR-SB)



[Sukkar et al., IEEE EMBS ‘12]

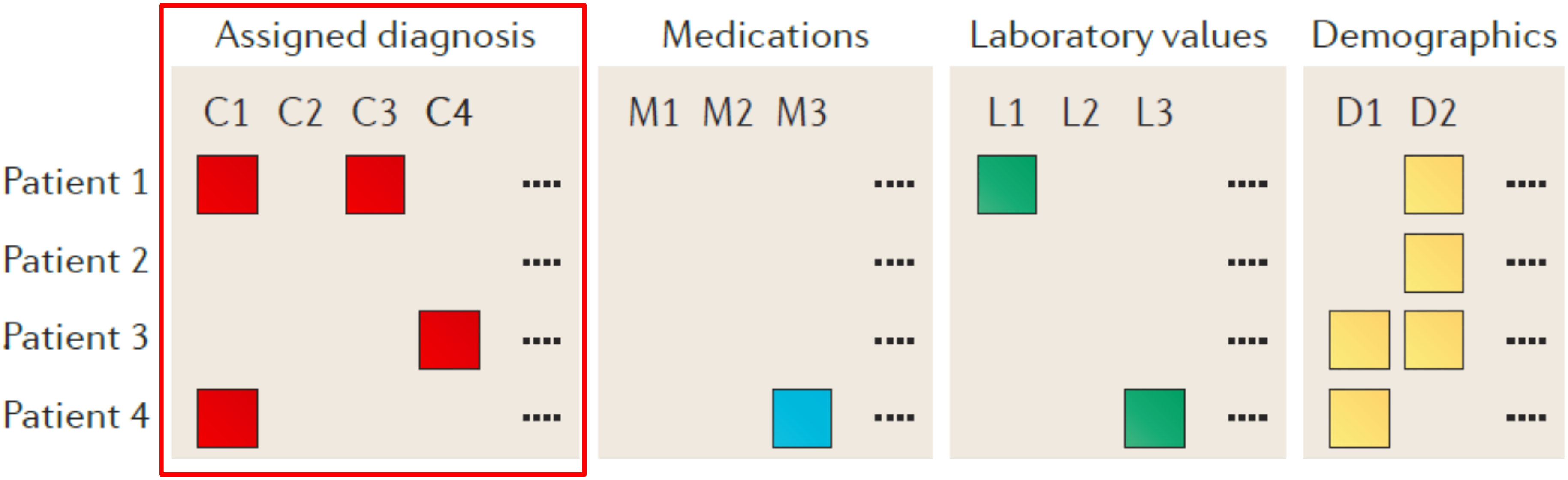
Outline of today’s class

1. Multi-task learning of (measurable) disease progression
   - **Application to Alzheimer’s disease (Zhou et al., KDD ‘12)**

2. Discovering fine-grained disease states using hidden Markov models
   - **Application to Alzheimer’s disease (Sukkar et al., IEEE EMBS ’12)**

3. Unsupervised learning of (grounded, multi-dimensional) disease progression models
   - **Application to chronic obstructive pulmonary disease (Wang et al., KDD ‘14)**

Goal: Learn from Electronic Health Records (EHR)

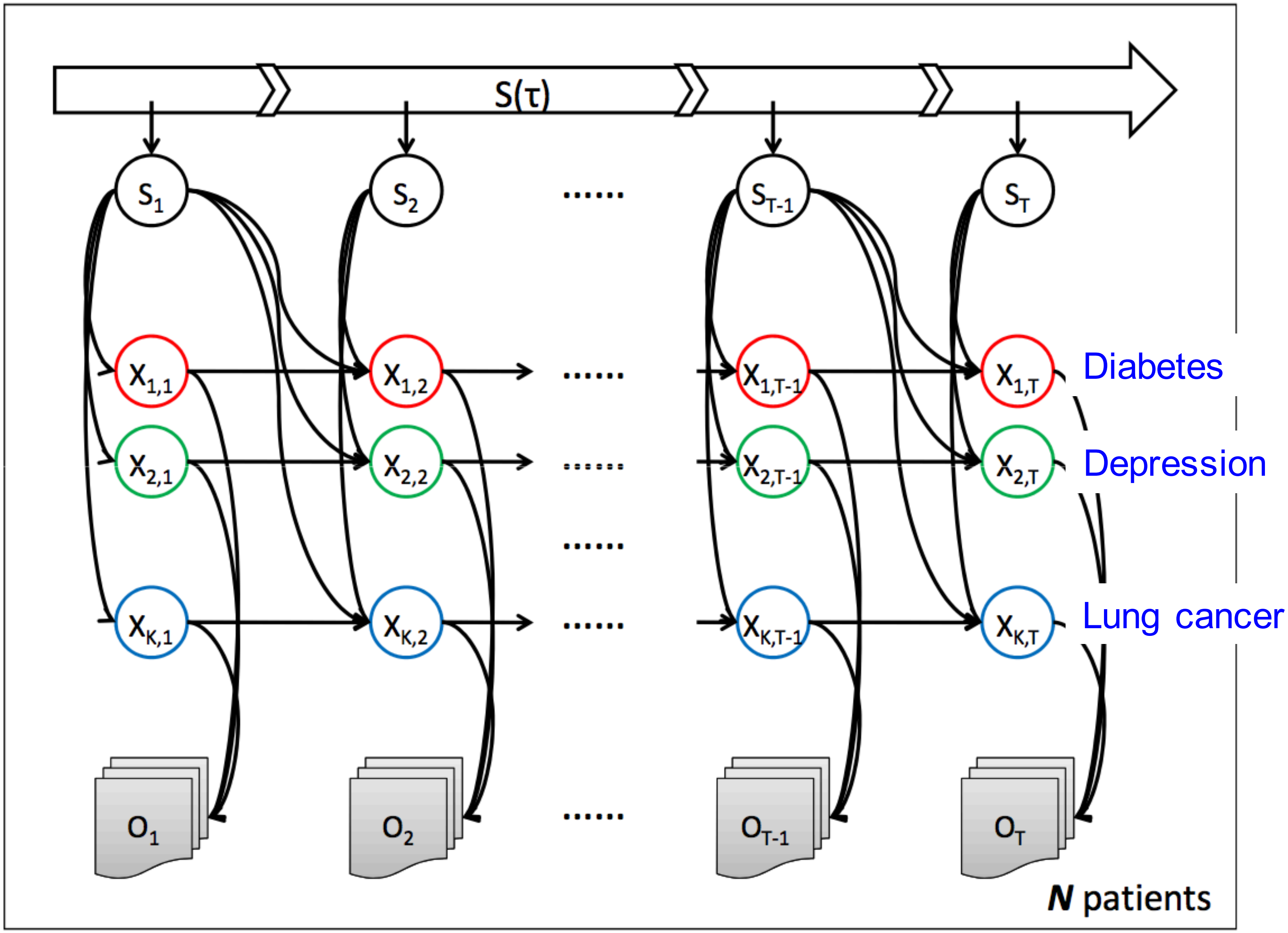


	Assigned diagnosis					Medications				Laboratory values				Demographics		
	C1	C2	C3	C4		M1	M2	M3		L1	L2	L3		D1	D2	
Patient 1	🟥		🟥		....				....	🟩			....		🟨	....
Patient 2					....				....				....		🟨	....
Patient 3				🟥	....				....				....	🟨	🟨	....
Patient 4	🟥				....			🟦	....			🟩	....	🟨		....

PID	DAY_ID	CLINICAL_EVENT	ICD9_LONGNAME
000000	74053	305.1	Tobacco Use Disorder
000000	74053	496	Chronic Airway Obstruction, Not Elsewhere Classified
000000	74053	733	Osteoporosis, Unspecified
000000	74053	724.2	Lumbago
000000	74091	733	Osteoporosis, Unspecified
000000	74148	733	Osteoporosis, Unspecified
000000	74148	782.3	Edema
000000	74148	780.79	Other Malaise And Fatigue

Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012 May 2;13(6):395-405.

The big picture: generative model for patient data



Markov Jump Process

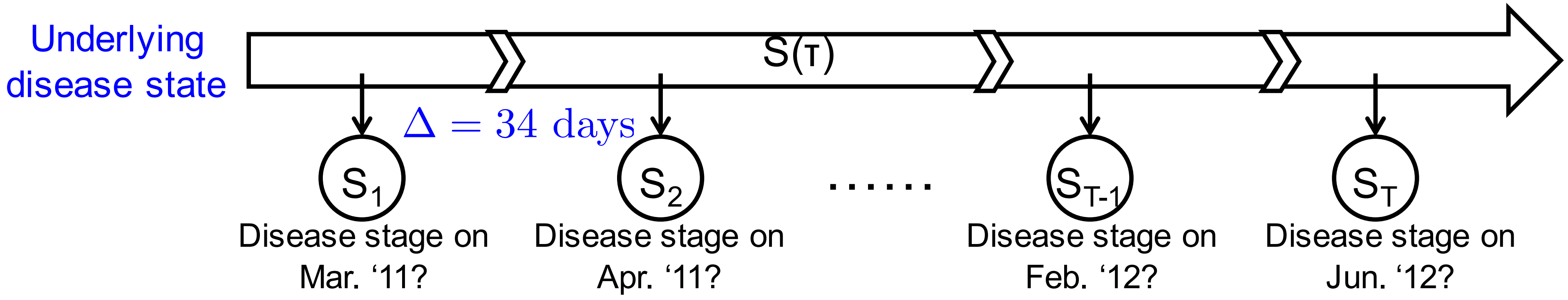
Progression Stages

K phenotypes, each with its own Markov chain

Observations

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

Model for patient’s disease progression across time



Underlying disease state

S(τ)

$\Delta = 34 \text{ days}$

S_1 Disease stage on Mar. ‘11?

S_2 Disease stage on Apr. ‘11?

……

$S_{T\text{-}1}$ Disease stage on Feb. ‘12?

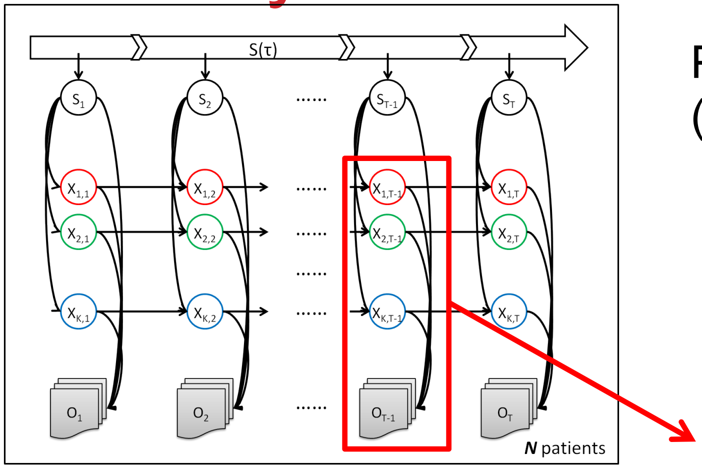
S_T Disease stage on Jun. ‘12?

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

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

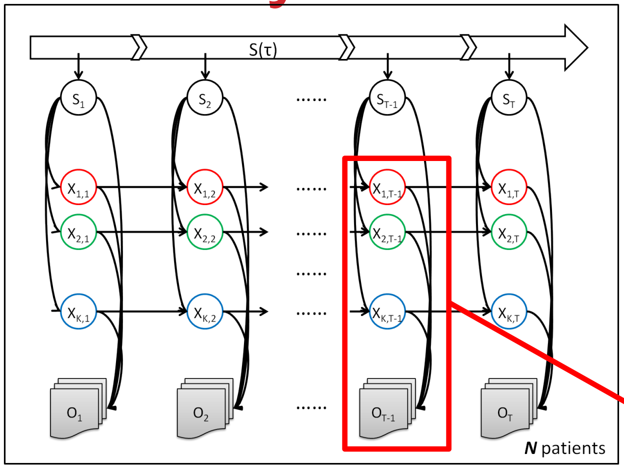
Matrix Q: Parameters to learn

Model for data at single point in time: Noisy-OR network



Previously used for medical diagnosis, e.g. QMR-DT (Shwe et al. ’91)

Model for data at single point in time: Noisy-OR network

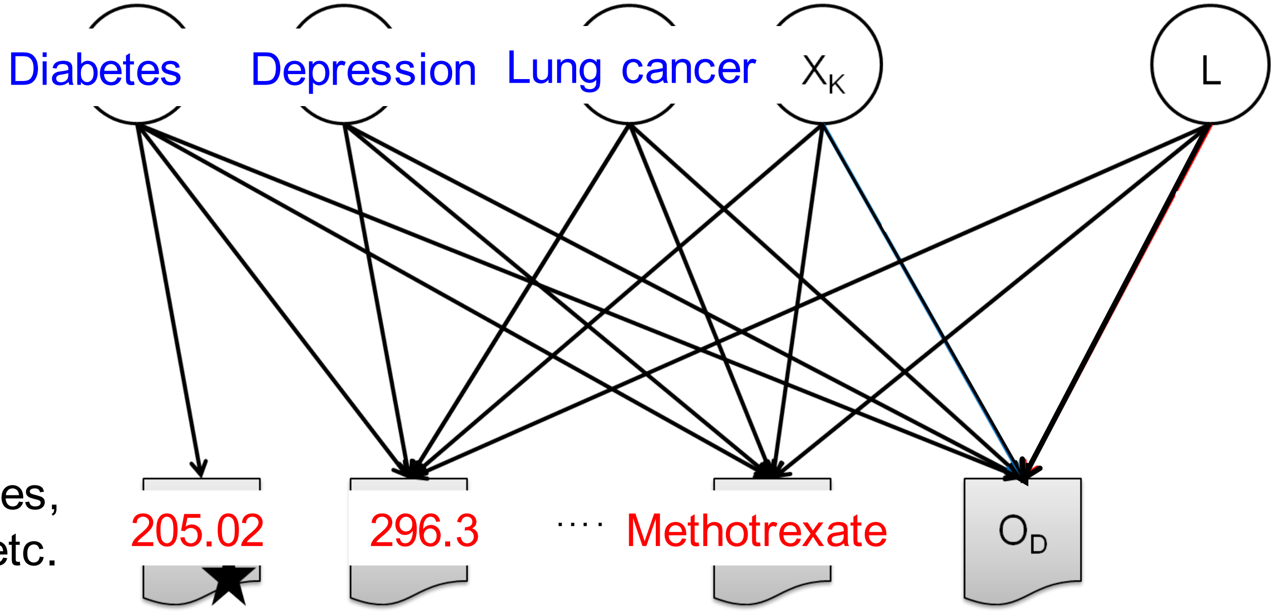


Previously used for medical diagnosis, e.g. QMR-DT (Shwe et al. ’91)

All binary variables

Comorbidities / Phenotypes
(hidden)

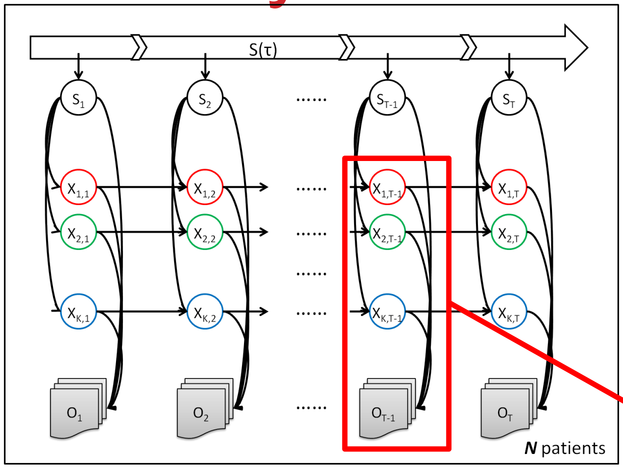
“Everything else”
(always on)



Diagnosis codes, medications, etc.

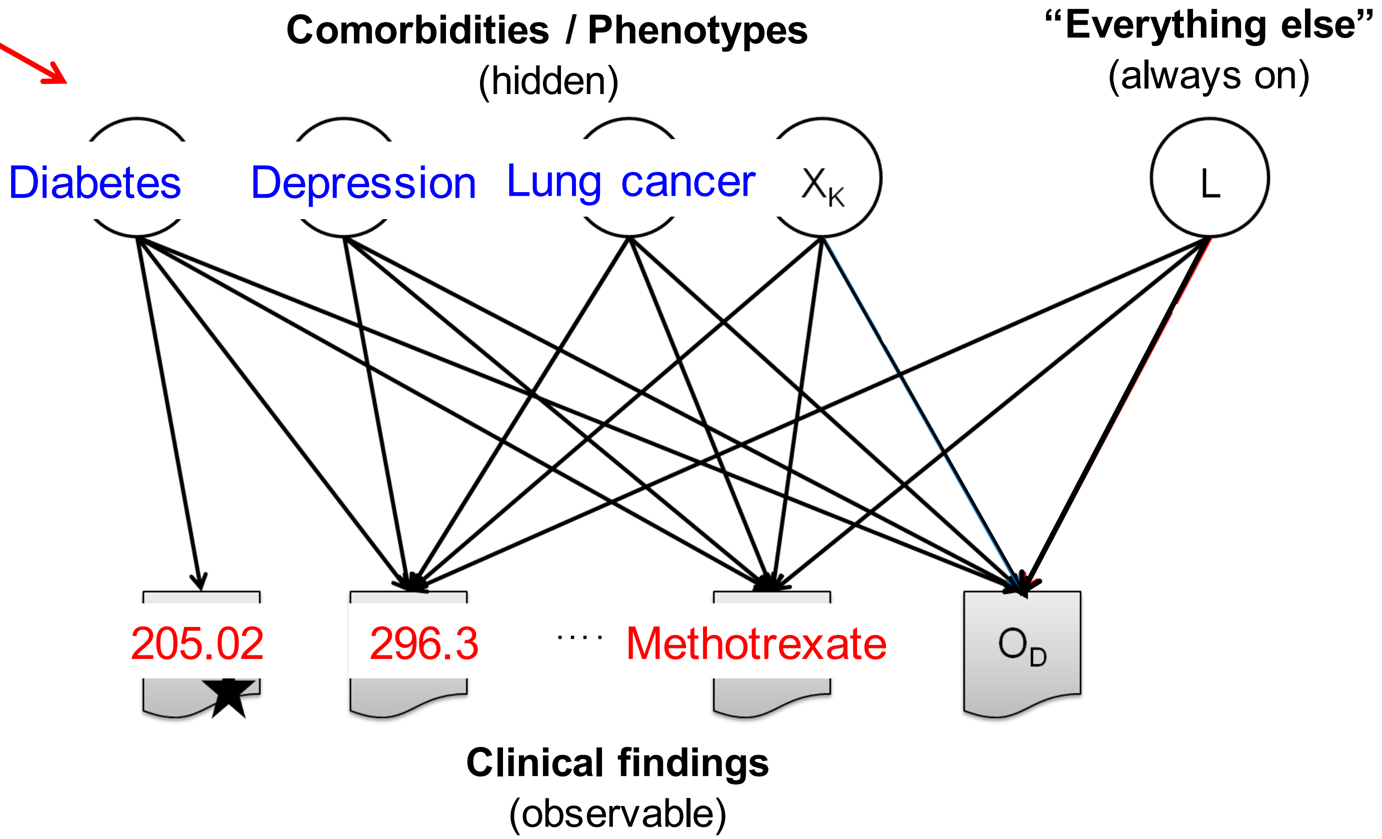
Clinical findings
(observable)

Model for data at single point in time: Noisy-OR network

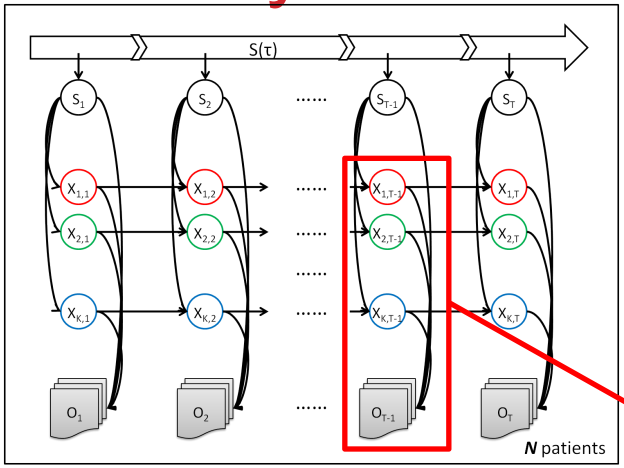


Previously used for medical diagnosis, e.g. QMR-DT (Shwe et al. '91)

We also learn which edges exist



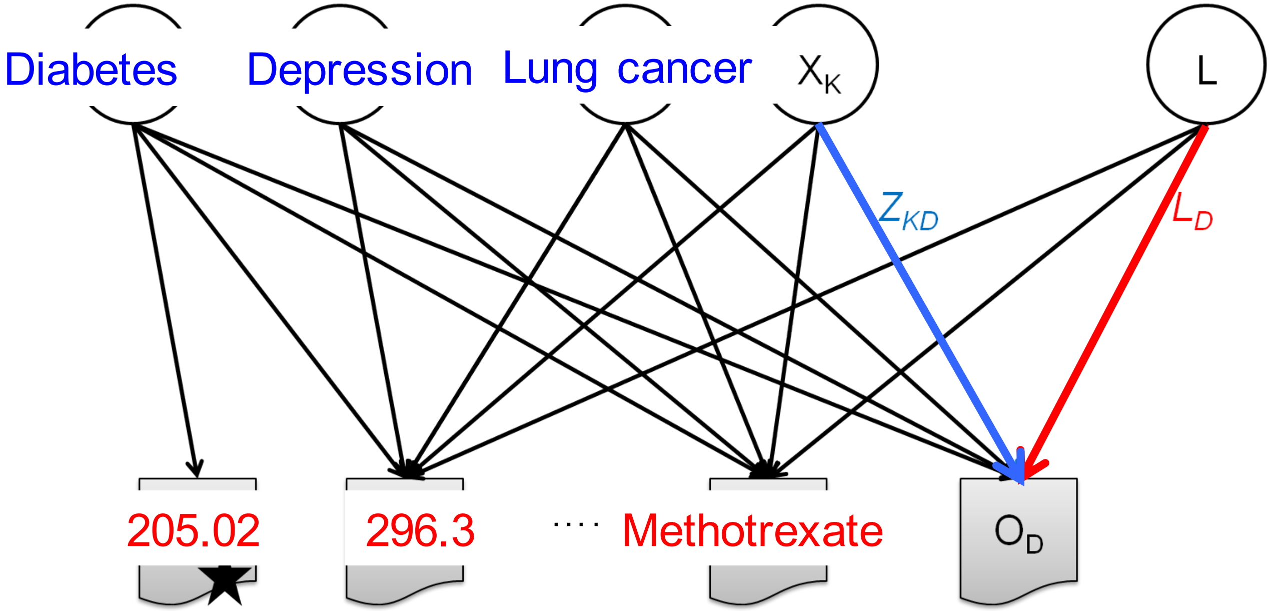
Model for data at single point in time: Noisy-OR network



Previously used for medical diagnosis, e.g. QMR-DT (Shwe et al. '91)

We also learn which edges exist

Associated with each edge is a *failure probability*

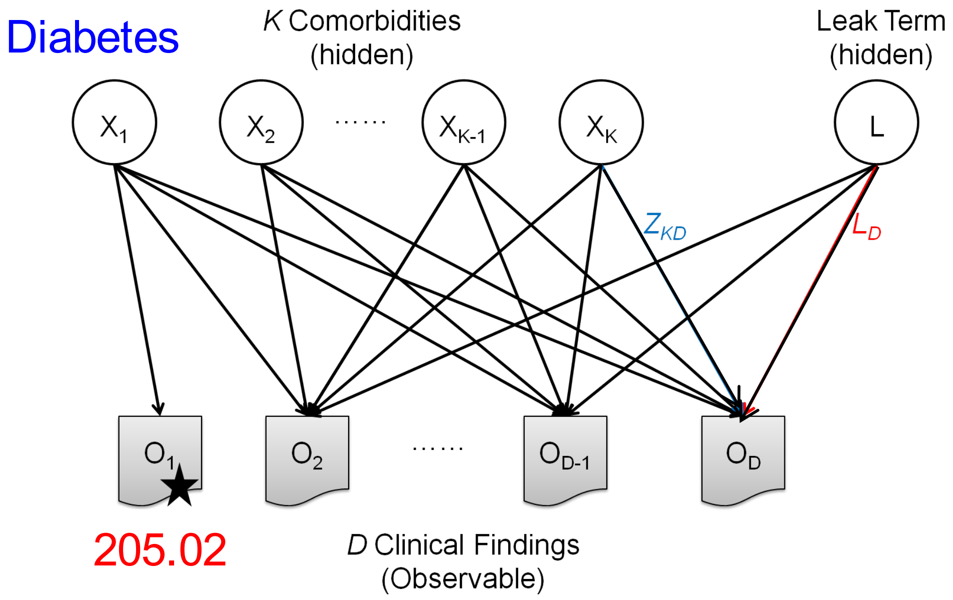

Comorbidities / Phenotypes
(hidden)

"Everything else"
(always on)

Clinical findings
(observable)

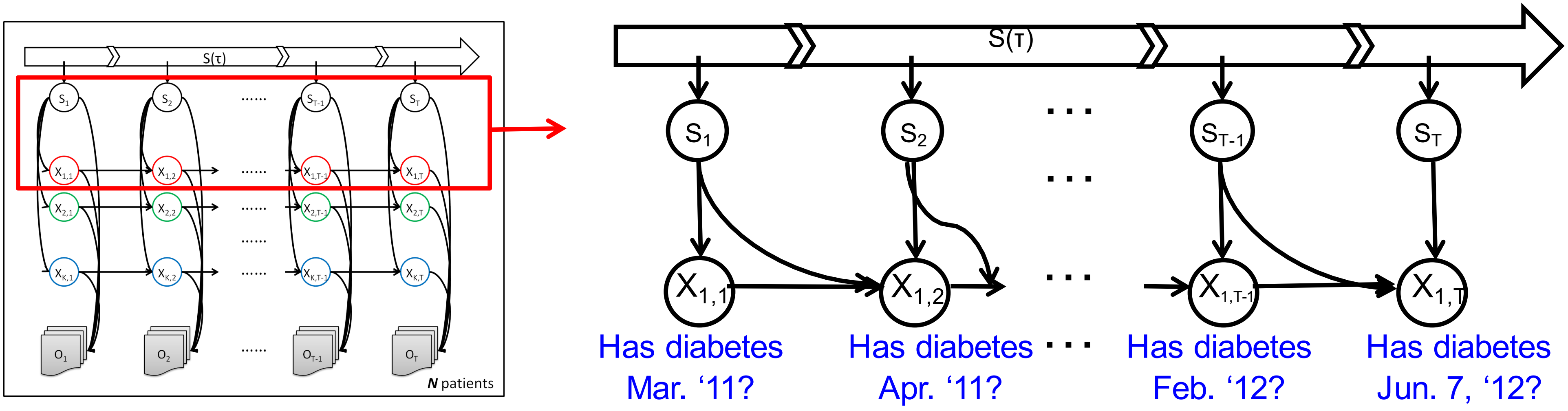
Anchored noisy-OR network

- An *anchor* is a finding that can only be caused by a single comorbidity
- We can specify one or more anchors for each hidden variable
- Use anchors findings to enable injection of domain expertise



Y. Halpern, YD Choi, S. Horng, D. Sontag. Using Anchors to Estimate Clinical State without Labeled Data. To appear in the American Medical Informatics Association (AMIA) Annual Symposium, Nov. 2014

Model of comorbidities across time



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

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

Inference

- Outer loop
  - EM
  - Algorithm to estimate the Markov Jump Process is borrowed form recent literature in physics
- Inner loop
  - Gibbs sampler used for approximate inference
  - We perform block sampling of the Markov chains, improving the mixing time of the Gibbs sampler

P. Metzner, I. Horenko, and C. Schutte. Generator estimation of markov jump processes based on incomplete observations nonequidistant in time. Physical Review E, 76(6):066702, 2007.

Implementation and optimization

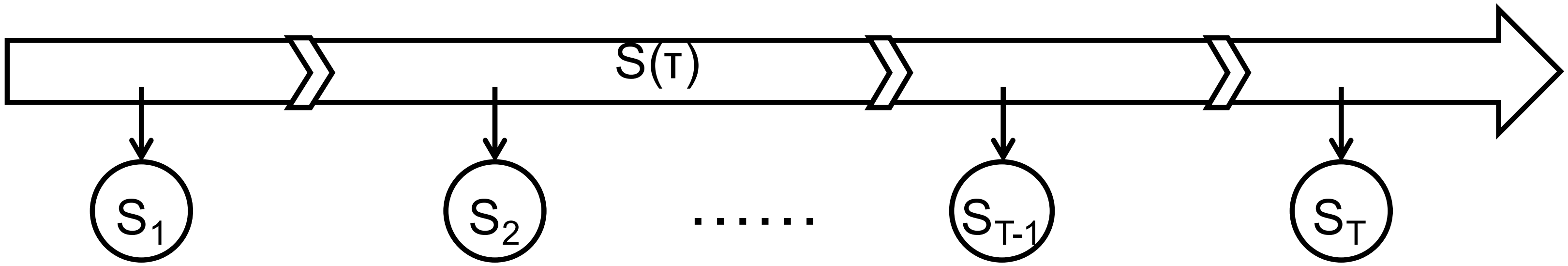
- Implemented in Python
  - Initially, each Gibbs sampling update took hours

- Parallelization
  - Parallelize over patients and findings
  - Almost linear speedup
- Computational tricks
  - Each Gibbs update can be performed in time linear in the number of *positive* findings
  - Caching
  - Pre-compute sufficient statistics

- After these, each update takes < 3 minutes (using 24 cores)

Customizations for COPD

- Enforce monotonic stage progression, i.e. $S_{t+1} \geq S_t$:



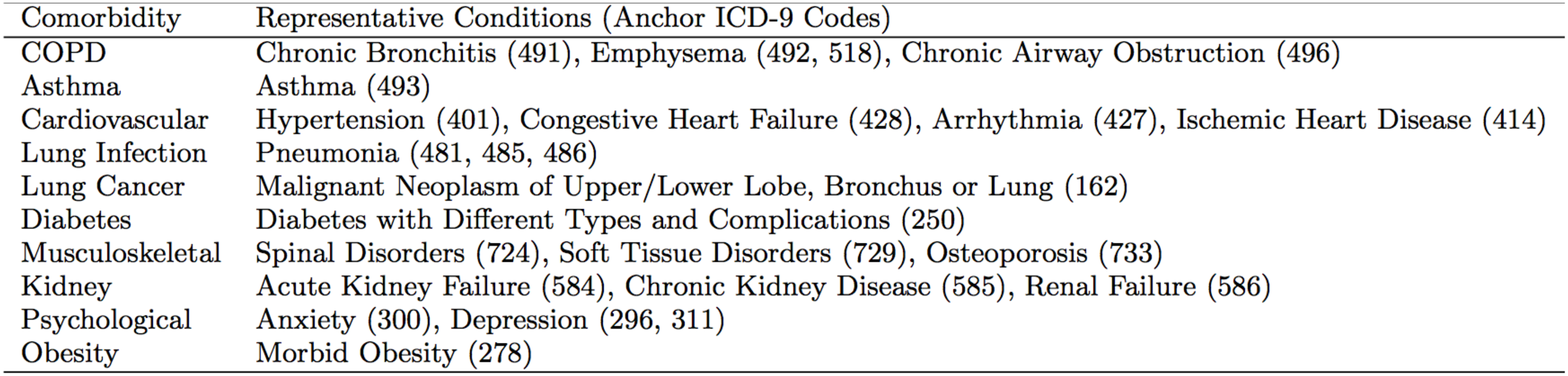
- Enforce monotonicity in distributions of comorbidities in first time step, e.g. $\Pr(X_{j,1} \mid S_1 = 2) \geq \Pr(X_{j,1} \mid S_1 = 1)$
  - To do this, we solve a tiny convex optimization problem within EM

- Enforce that transitions in X can only happen at the same time as transitions in S

- Edge weights given a Beta(0.1, 1) prior to encourage sparsity

Specifying the latent variables

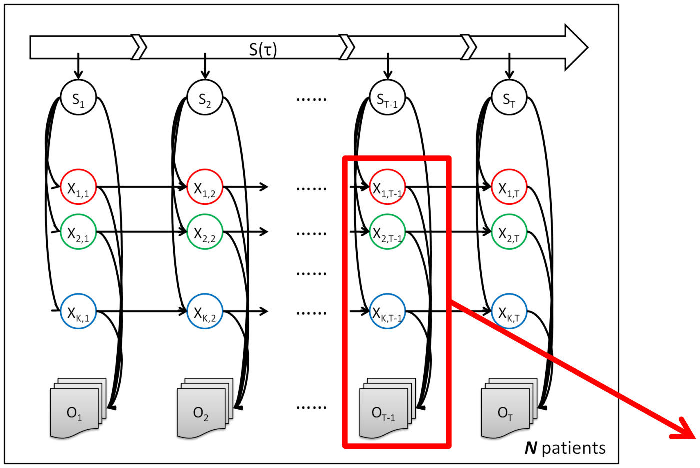
- We provide anchors for each of the comorbidities that we want to model:



Comorbidity	Representative Conditions (Anchor ICD-9 Codes)
COPD	Chronic Bronchitis (491), Emphysema (492, 518), Chronic Airway Obstruction (496)
Asthma	Asthma (493)
Cardiovascular	Hypertension (401), Congestive Heart Failure (428), Arrhythmia (427), Ischemic Heart Disease (414)
Lung Infection	Pneumonia (481, 485, 486)
Lung Cancer	Malignant Neoplasm of Upper/Lower Lobe, Bronchus or Lung (162)
Diabetes	Diabetes with Different Types and Complications (250)
Musculoskeletal	Spinal Disorders (724), Soft Tissue Disorders (729), Osteoporosis (733)
Kidney	Acute Kidney Failure (584), Chronic Kidney Disease (585), Renal Failure (586)
Psychological	Anxiety (300), Depression (296, 311)
Obesity	Morbid Obesity (278)

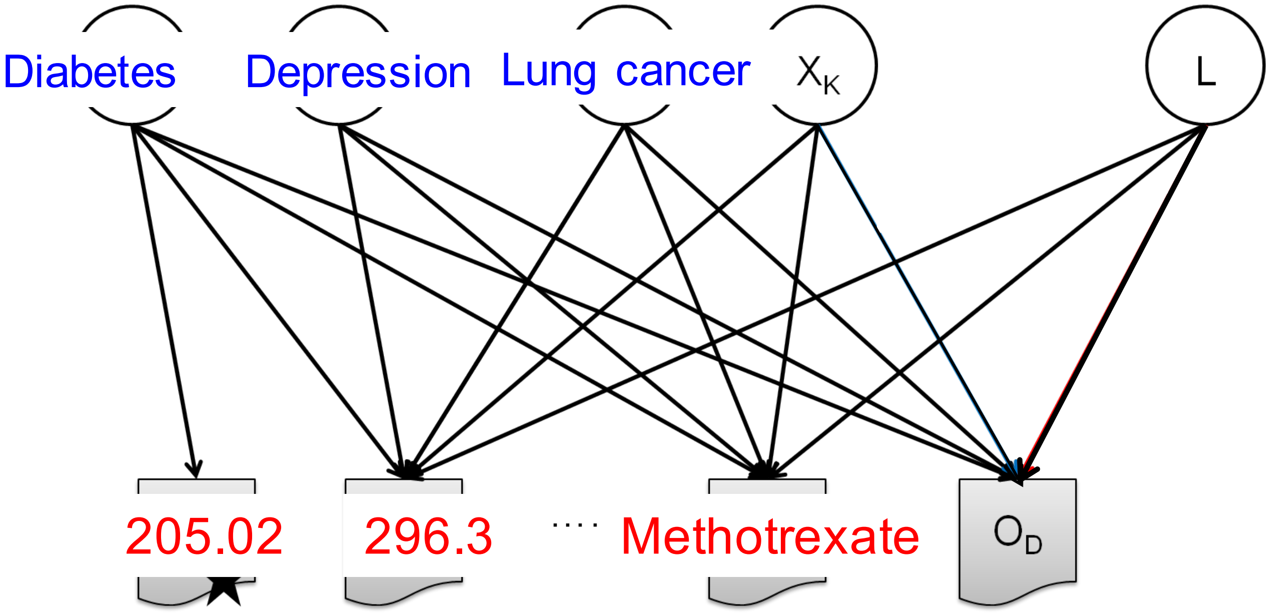
- Can be viewed as a type of weak supervision, using clinical domain knowledge
- Without these, the results are less interpretable

Which edges are learned?



Comorbidities / Phenotypes
(hidden)

"Everything else"
(always on)



Clinical findings
(observable)

Edges learned for *kidney disease*

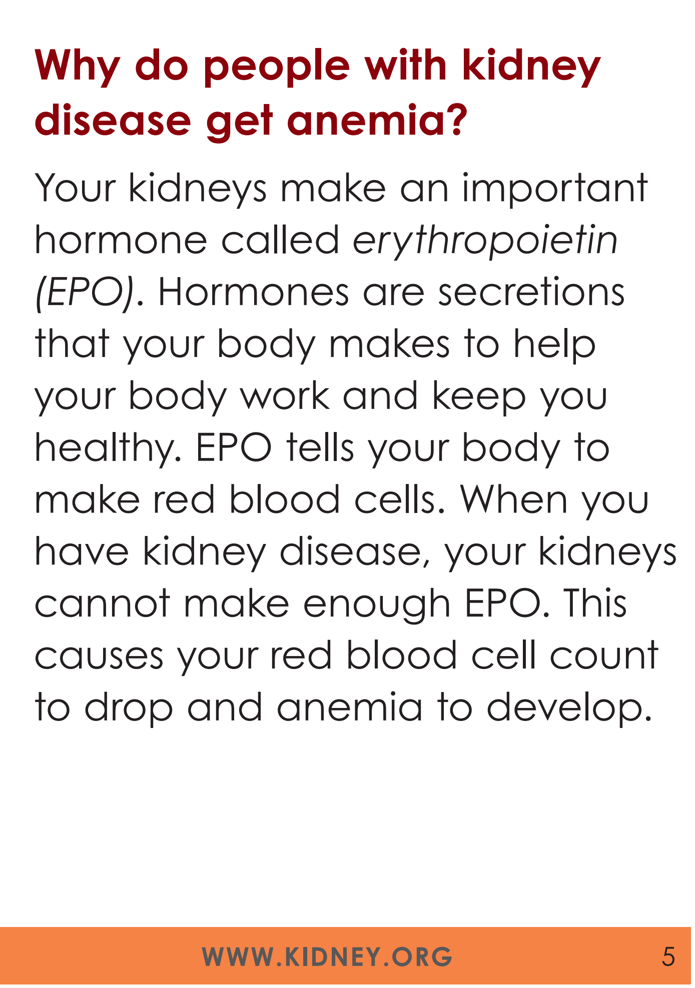
Diagnosis code	Weight	
*585.3	0.20	Chronic Kidney Disease, Stage Iii (Moderate)
285.9	0.15	Anemia, Unspecified
*585.9	0.10	Chronic Kidney Disease, Unspecified
599.0	0.08	Urinary Tract Infection, Site Not Specified
*585.4	0.08	Chronic Kidney Disease, Stage Iv (Severe)
*584.9	0.07	Acute Renal Failure, Unspecified
*586	0.07	Renal Failure, Unspecified
782.3	0.06	Edema
*585.6	0.05	End Stage Renal Disease
593.9	0.04	Unspecified Disorder Of Kidney And Ureter
272.4	0.04	Other And Unspecified Hyperlipidemia
272.2	0.03	Mixed Hyperlipidemia

Edges learned for *kidney disease*

Diagnosis code	Weight	
***585.3**	**0.20**	**Chronic Kidney Disease, Stage Iii (Moderate)**
285.9	0.15	Anemia, Unspecified
***585.9**	**0.10**	**Chronic Kidney Disease, Unspecified**
599.0	0.08	Urinary Tract Infection, Site Not Specified
***585.4**	**0.08**	**Chronic Kidney Disease, Stage Iv (Severe)**
***584.9**	**0.07**	**Acute Renal Failure, Unspecified**
***586**	**0.07**	**Renal Failure, Unspecified**
782.3	0.06	Edema
***585.6**	**0.05**	**End Stage Renal Disease**
593.9	0.04	Unspecified Disorder Of Kidney And Ureter
272.4	0.04	Other And Unspecified Hyperlipidemia
272.2	0.03	Mixed Hyperlipidemia

Edges learned for *kidney disease*

Diagnosis code	Weight	
*585.3	0.20	Chronic Kidney Disease, Stage Iii (Moderate)
285.9	**0.15**	**Anemia, Unspecified**
*585.9	0.10	Chronic Kidney Diseas
599.0	**0.08**	**Urinary Tract Infectic**
*585.4	0.08	Chronic Kidney Diseas
*584.9	0.07	Acute Renal Failure, U
*586	0.07	Renal Failure, Unspec
782.3	**0.06**	**Edema**
*585.6	0.05	End Stage Renal Dise
593.9	**0.04**	**Unspecified Disorde**
272.4	**0.04**	**Other And Unspecifi**
272.2	**0.03**	**Mixed Hyperlipidemi**



Why do people with kidney disease get anemia?

Your kidneys make an important hormone called *erythropoietin (EPO)*. Hormones are secretions that your body makes to help your body work and keep you healthy. EPO tells your body to make red blood cells. When you have kidney disease, your kidneys cannot make enough EPO. This causes your red blood cell count to drop and anemia to develop.

WWW.KIDNEY.ORG 5

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

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**

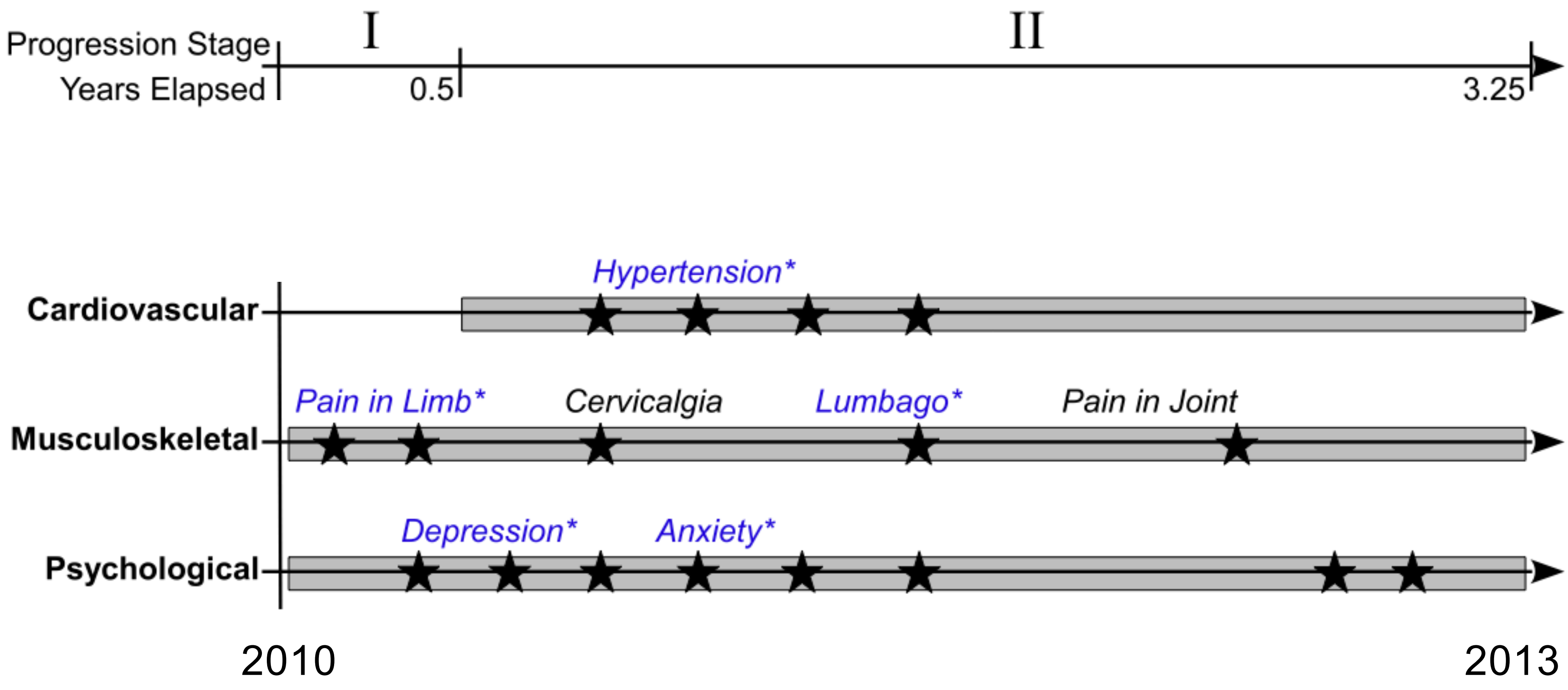
Edges learned for *lung cancer*

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

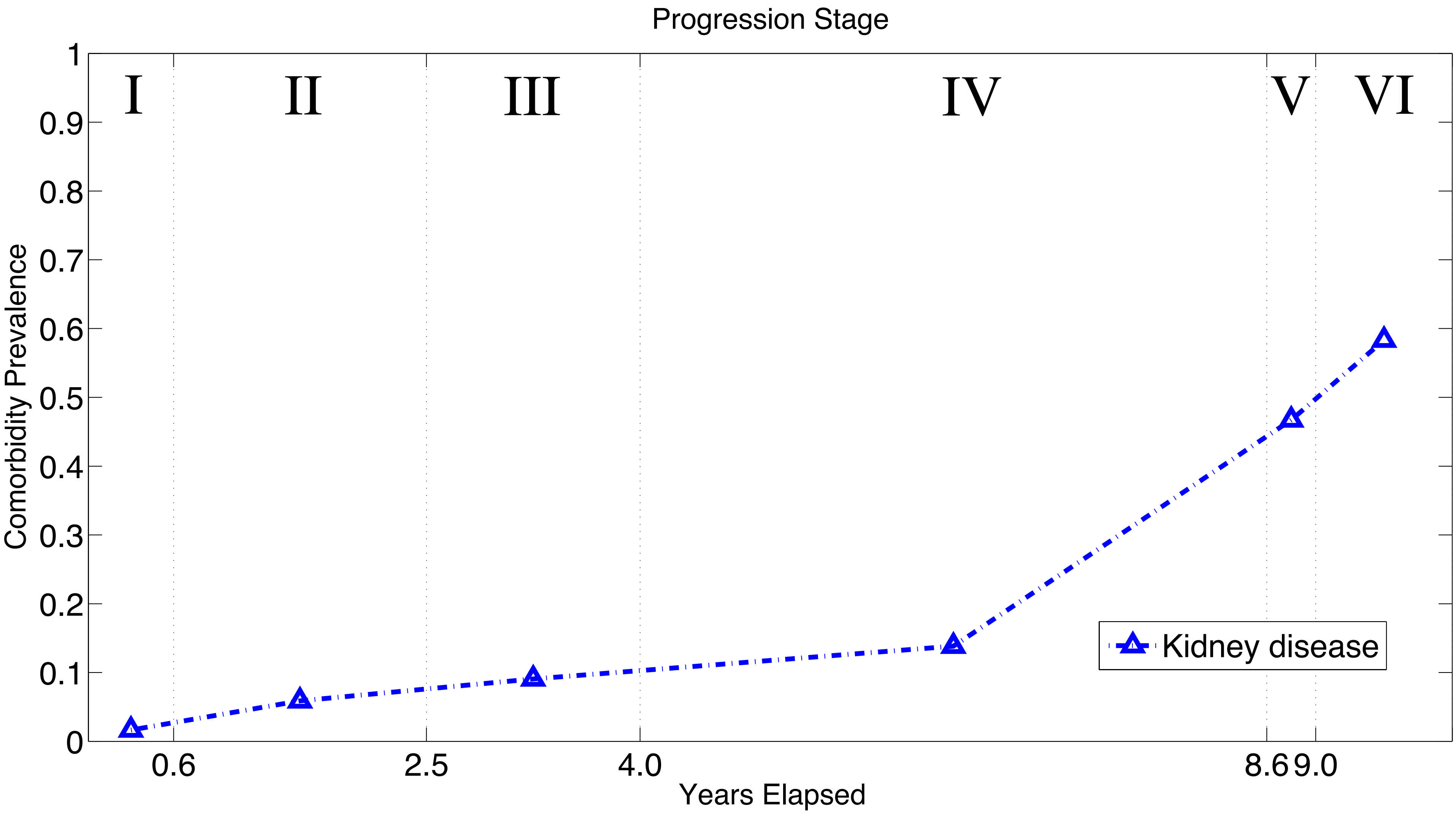
Edges learned for *lung infection*

Diagnosis code	Weight	
*486	**0.30**	**Pneumonia, Organism Unspecified**
786.05	0.10	Shortness Of Breath
786.09	0.10	Other Respiratory Abnormalities
786.2	0.10	Cough
793.1	0.06	Abnormal Findings On Radiological Exam Of Lung
285.9	0.05	Anemia, Unspecified
518.89	0.05	Other Diseases Of Lung, Not Elsewhere Classified
466.0	0.05	Acute Bronchitis
799.02	0.05	Hypoxemia
599.0	0.04	Urinary Tract Infection, Site Not Specified
V58.61	0.04	Long-Term (Current) Use Of Anticoagulants
786.50	0.04	Chest Pain, Unspecified

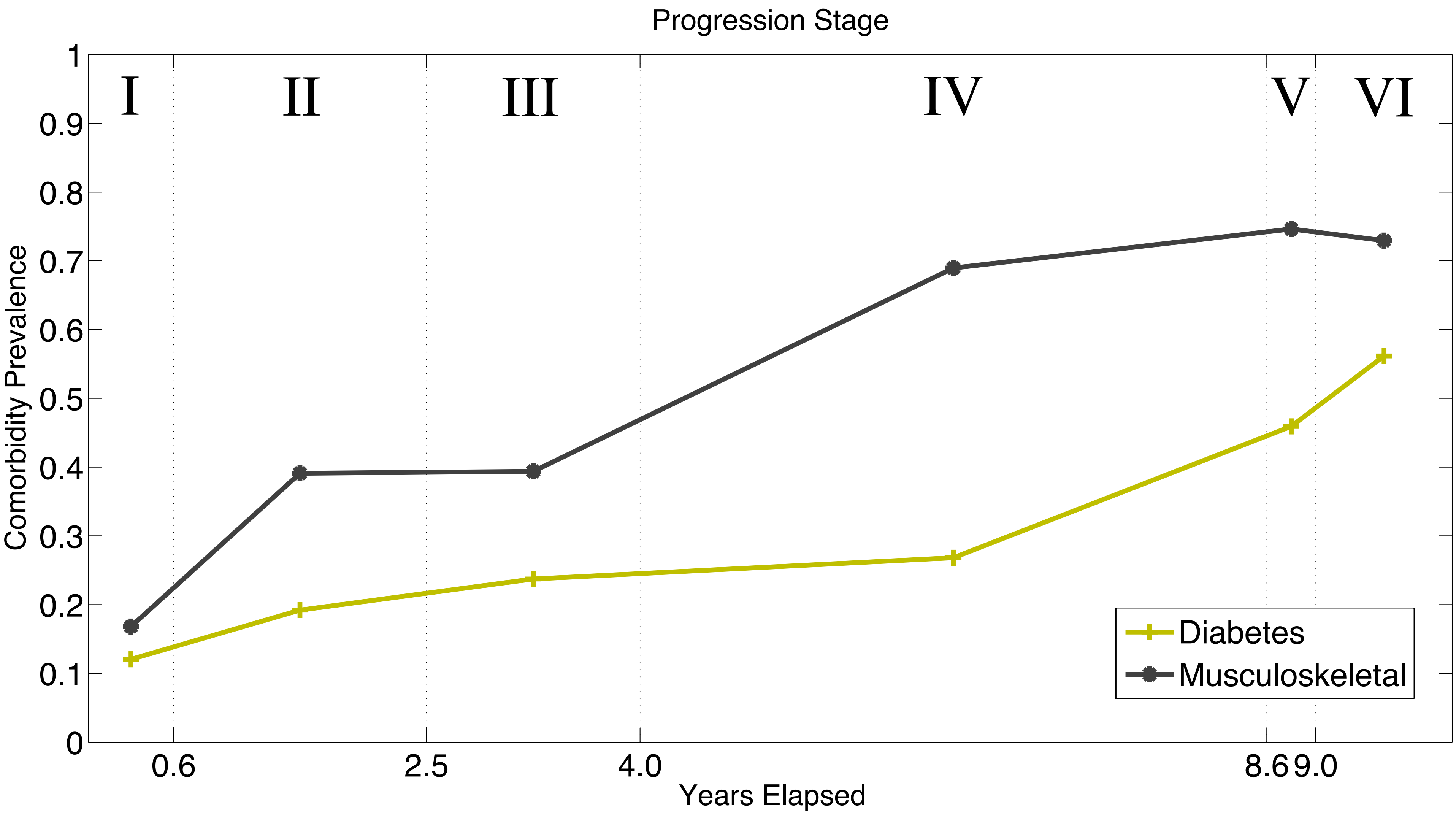
Progression of a single patient



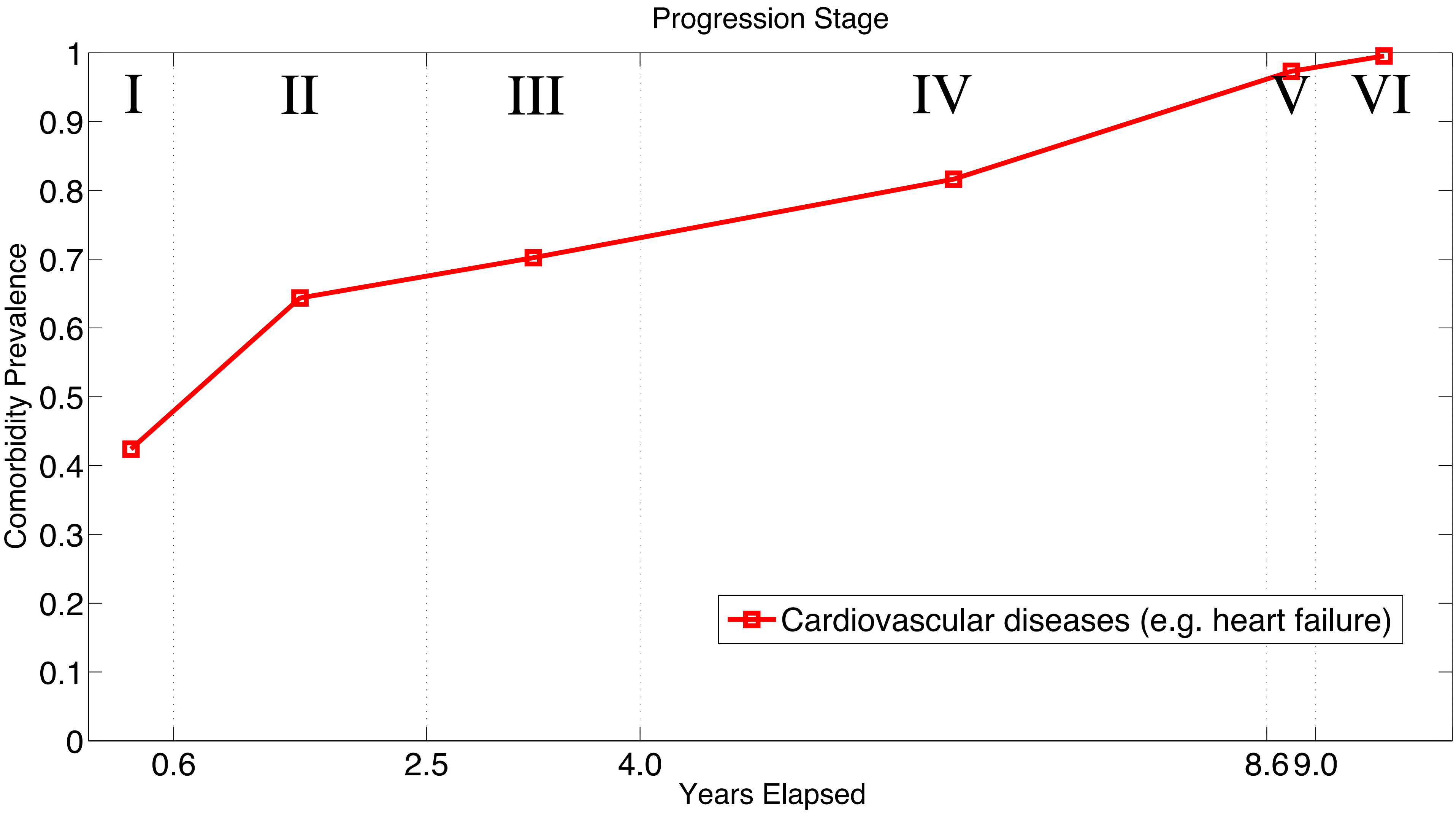
Prevalence of comorbidities across stages (Kidney disease)



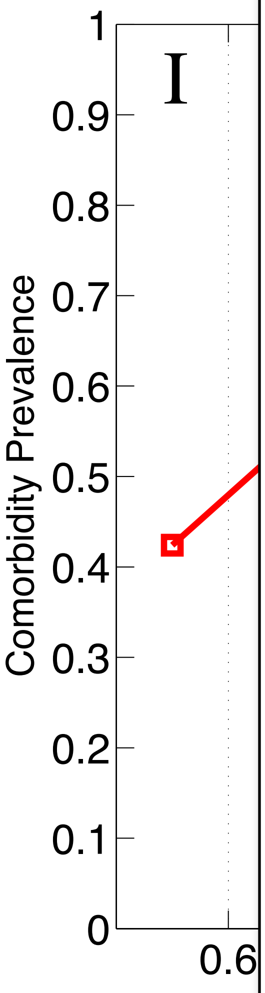
Prevalence of comorbidities across stages (Diabetes & Musculoskeletal disorders)

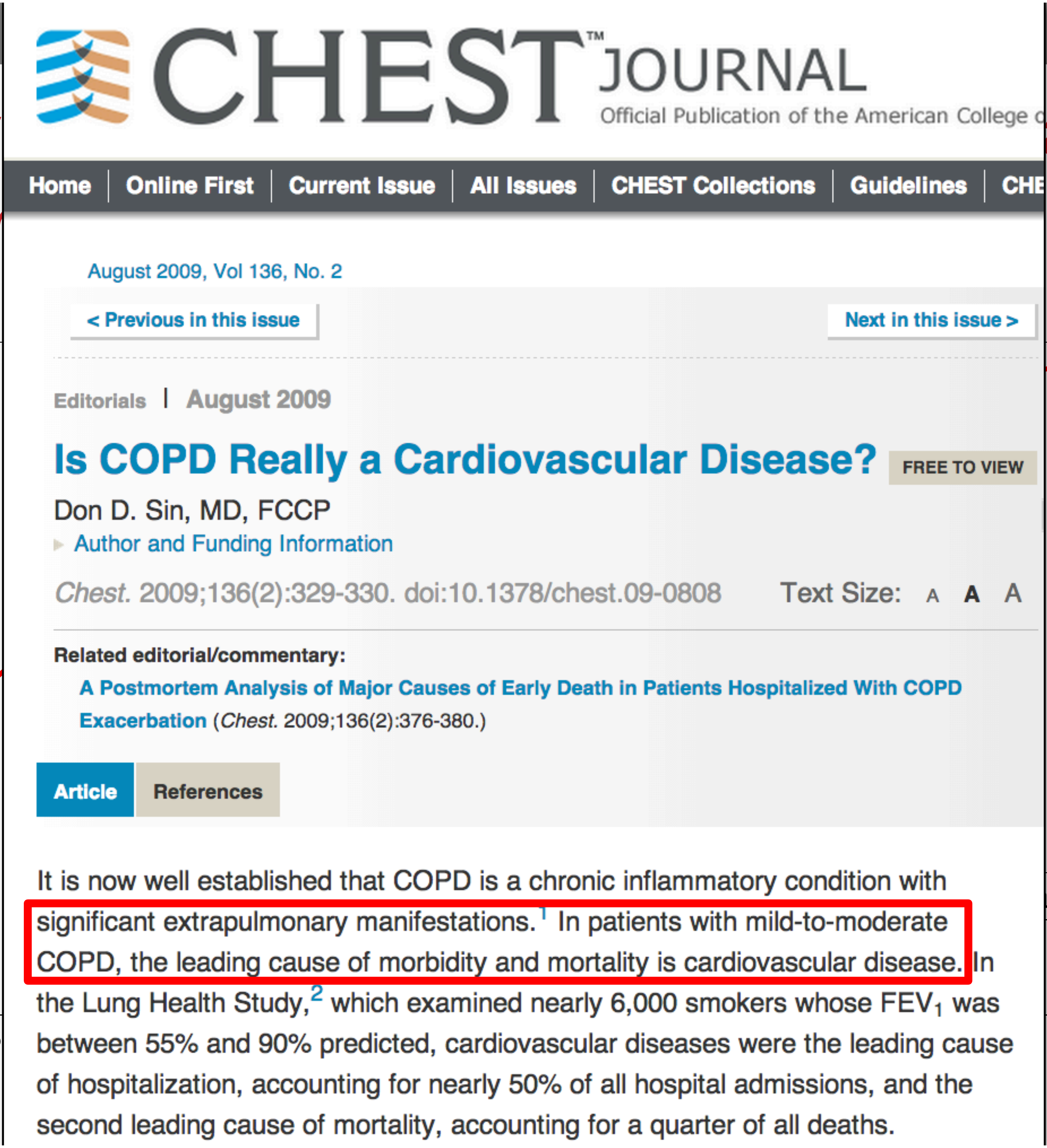


Prevalence of comorbidities across stages (Cardiovascular disease)



Prev ages
(Ca





August 2009, Vol 136, No. 2

Editorials | August 2009

Is COPD Really a Cardiovascular Disease?

Don D. Sin, MD, FCCP

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.