Unsupervised Learning of Disease Progression Models

Xiang Wang (IBM Research)

David Sontag (NYU)

Fei Wang (IBM Research)

The burden of chronic diseases

- Chronic disease is a global burden
  - Hundreds of millions of people
  - Trillions of dollars spent
  - Loss in life expectancy
  - Loss in quality of life

- **Example:** Chronic Obstructive Pulmonary Disease (COPD)
  - Impacts low-income population
  - Key risk factors: smoking and air pollution
  - Causes systemic illness

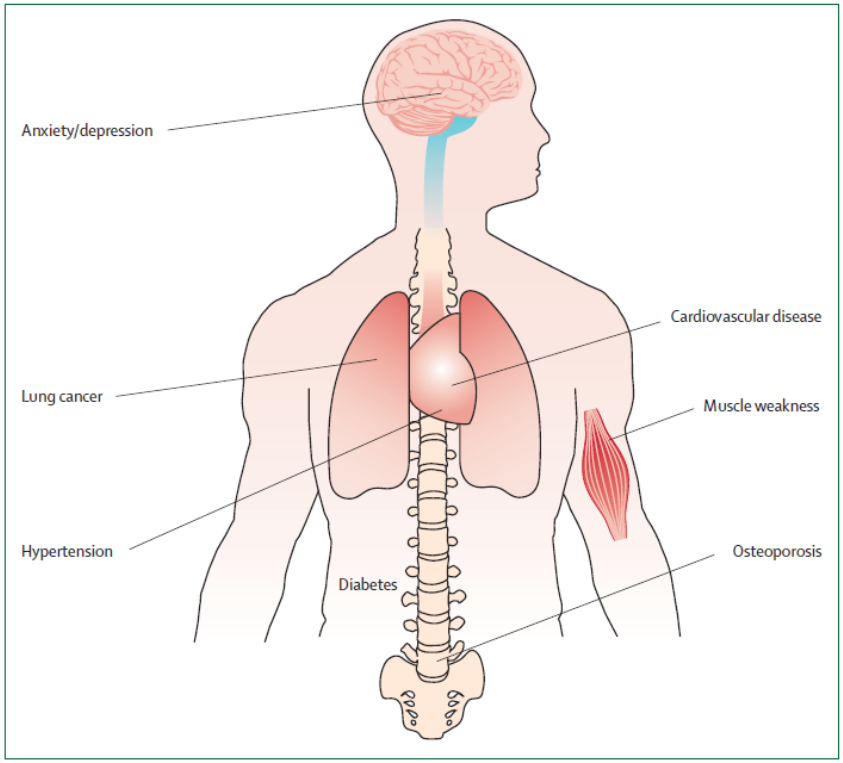


Figure 4: Comorbidities of chronic obstructive pulmonary disease

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

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.

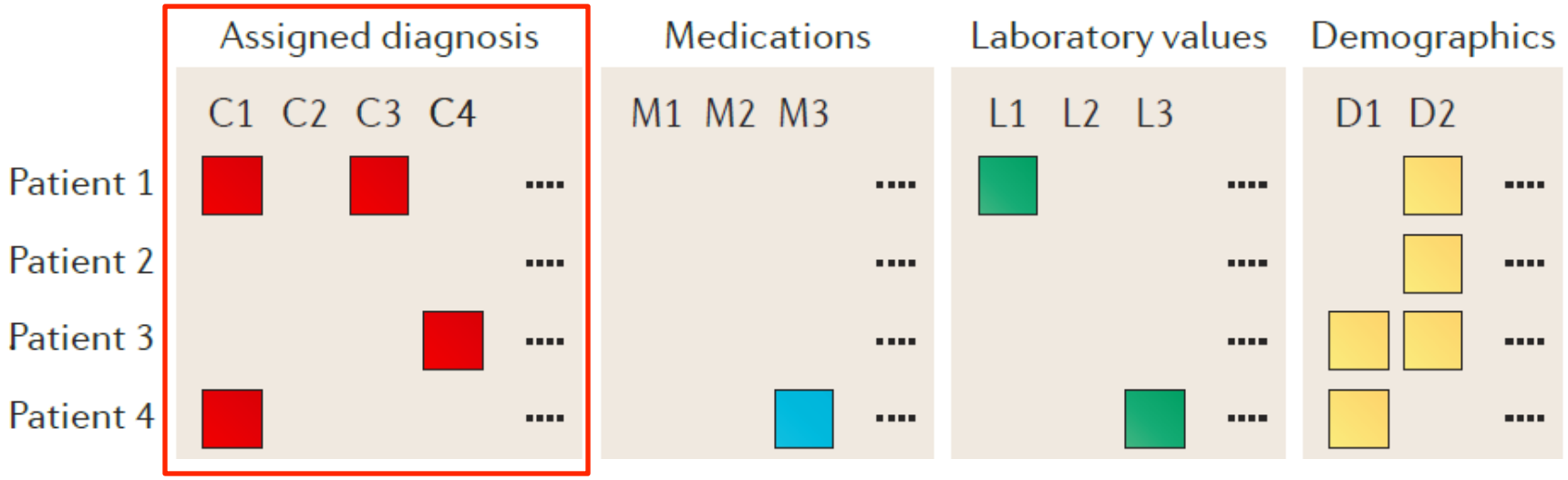
Table: **Therapy at each stage of chronic obstructive pulmonary disease, by GOLD stage[1]**

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

Our contribution

- Algorithm to learn a disease progression model from EHR data
  - No labeled data needed
  - Generative model
- We demonstrate its use in
  - Deriving a meaningful characterization of disease progression and stages
  - Identifying the progression trajectory of individual patients
- More broadly, these models will be used to
  - Provide decision support for early intervention
  - Develop data-driven guidelines for care plan management
  - Align patients across time, by disease stage, to enable comparative effectiveness research (e.g., of medications)

Learn from Electronic Health Records (EHR)



Patient ID	Date	CLINICAL_EVENT	ICD9_LONGNAME
000000	July 1, '11	305.1	Tobacco Use Disorder
000000	July 1, '11	496	Chronic Airway Obstruction, Not Elsewhere Classified
000000	July 1, '11	733	Osteoporosis, Unspecified
000000	July 1, '11	724.2	Lumbago
000000	Aug 15, '11	733	Osteoporosis, Unspecified
000000	Aug 15, '11	733	Osteoporosis, Unspecified
000000	Aug 15, '11	782.3	Edema
000000	Aug 15, '11	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.

Challenges of disease progression modeling

- Multiple covariates
- Progression heterogeneity
  - No natural alignment between records with varied progression rates
- Missing data
  - Doctors only document a subset of what they observe
- Incomplete records
  - Might be only 2-5 years of data available for any one person
- Irregular visits
  - Continuous-time model is needed
- Limited supervision
  - No ground truth regarding the current stage of progression

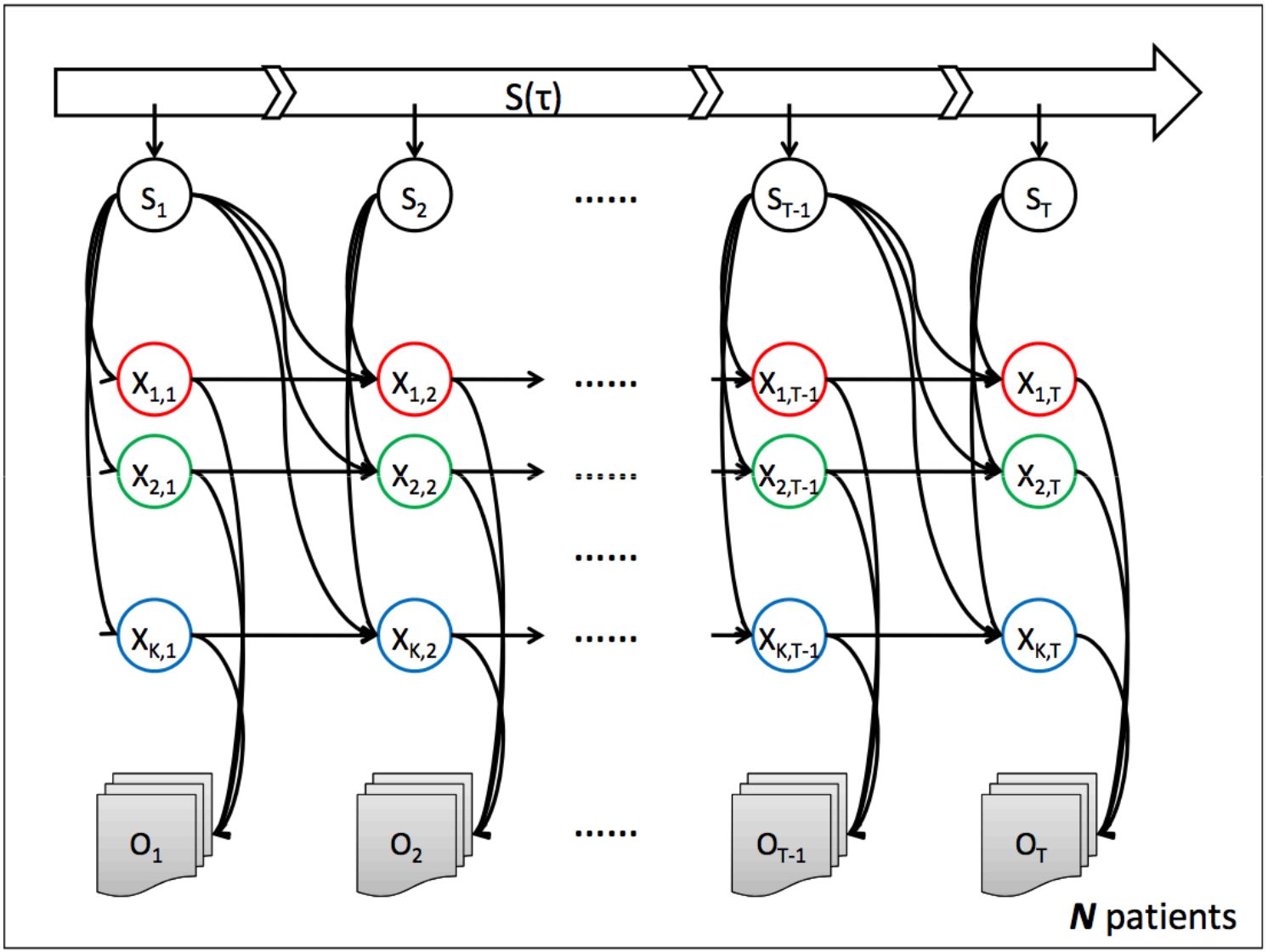
Our overall model

Markov Jump Process

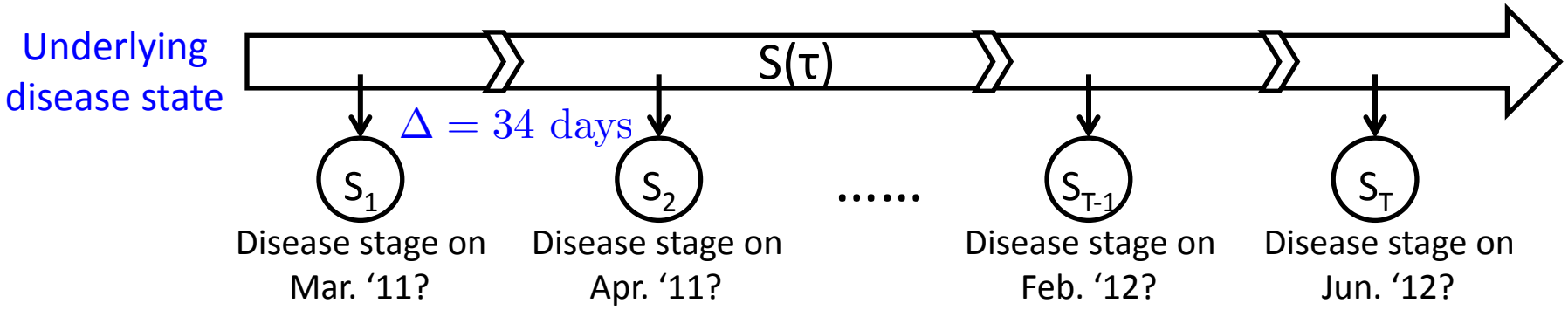
Progression Stages

K Comorbidities, each with its own Markov chain

Observations



Markov Jump Process

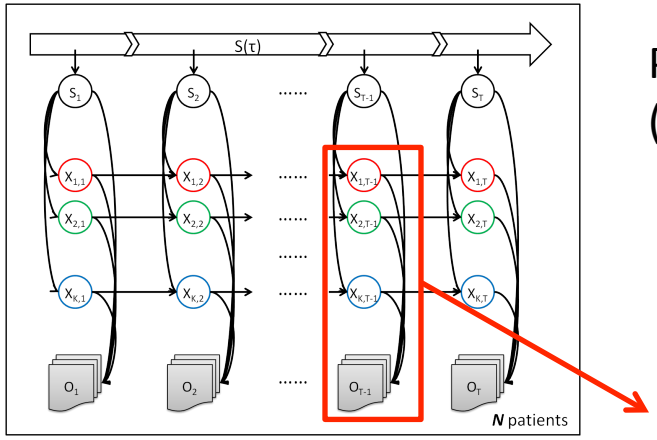


- A continuous-time Markov process with irregular discrete-time observations
- The transition probability is defined by an intensity matrix and the time interval:

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

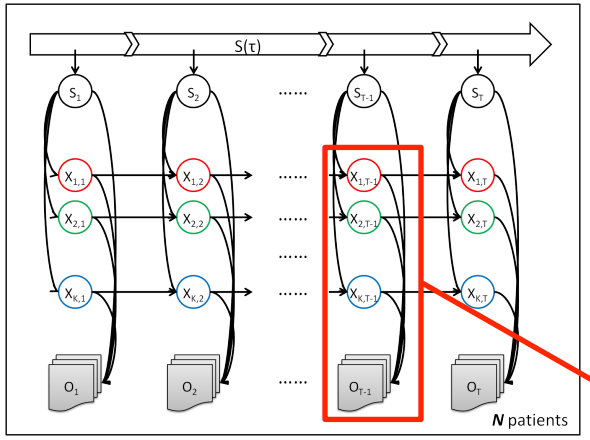
Matrix Q: Parameters to learn

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)

Comorbidities / Phenotypes
(hidden)

"Everything else"
(always on)

Diabetes Depression Lung cancer X_K L

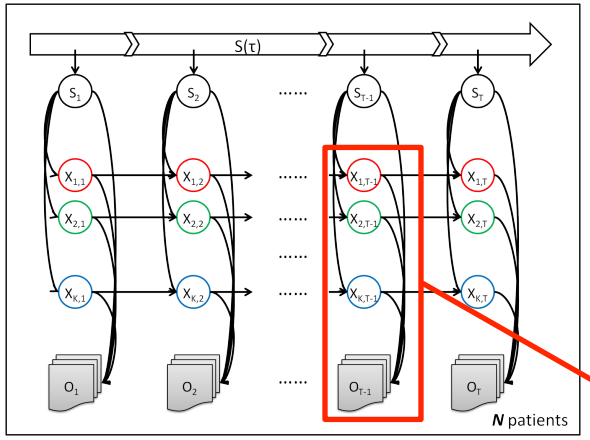
All binary variables

Diagnosis codes, medications, etc.

205.02 296.3 .... Methotrexate O_D

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)

Comorbidities / Phenotypes
(hidden)

"Everything else"
(always on)

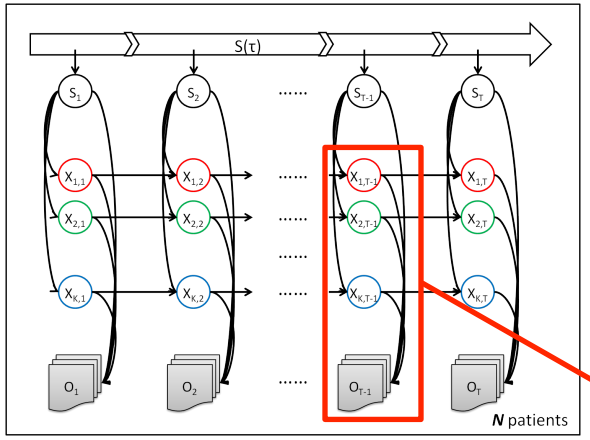
Diabetes Depression Lung cancer X_K L

We learn which edges exist

205.02 296.3 .... Methotrexate O_D

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)

Comorbidities / Phenotypes
(hidden)

"Everything else"
(always on)

Diabetes Depression Lung cancer X_K L

Z_{KD} L_D

We learn which edges exist

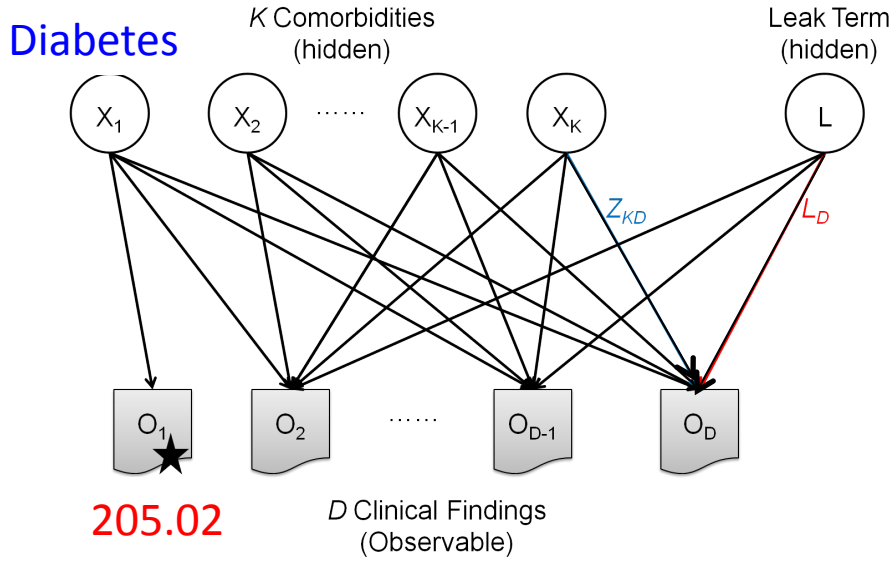
Associated with each edge is a *failure probability*

205.02 296.3 .... Methotrexate O_D

Clinical findings
(observable)

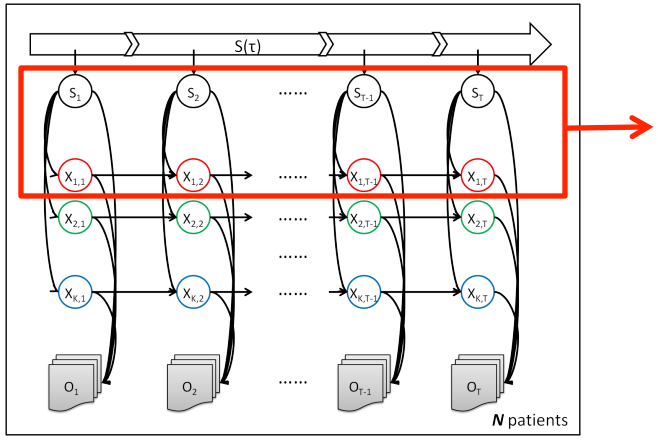
Anchored noisy-OR network

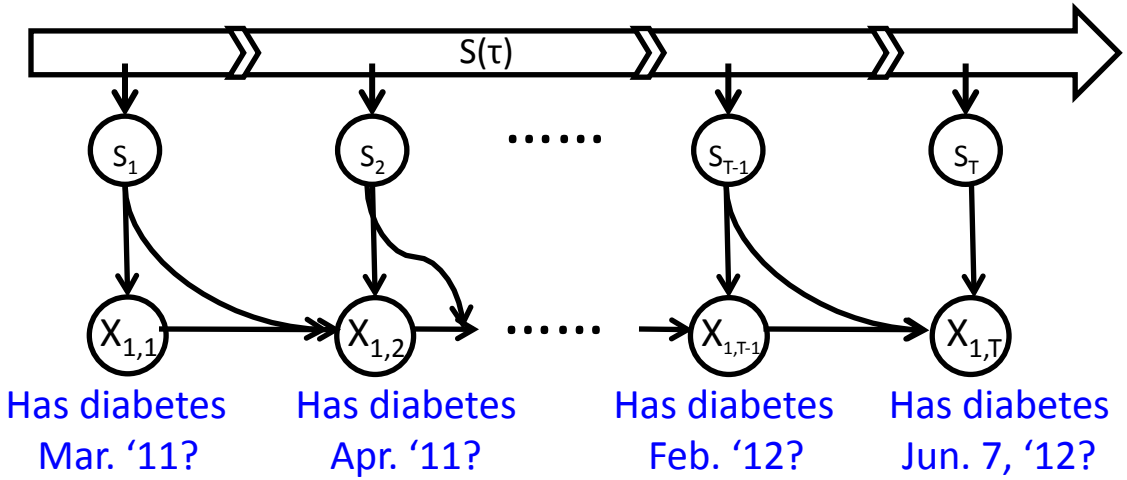
- An *anchor* is a finding that can only be caused by a single comorbidity
- We can specify one or more anchors for each hidden variable
- Use anchors to enable injection of domain expertise



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

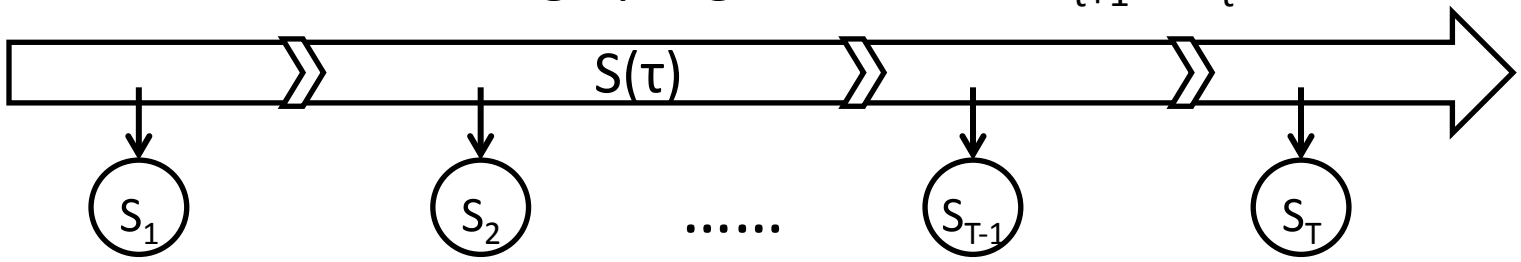
Inference

- 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
    - Block sampling of the Markov chains: decreases mixing time
  - Each update can be performed in time linear in the number of *positive* findings
  - Parallelize over patients and findings: each update takes 3 minutes (using 24 cores)

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.

Customizations for COPD

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



- Enforce monotonicity of likelihood of comorbidities in *initial* time step, i.e. $Pr(X_{j,1} \mid S_1 = 2) \geq Pr(X_{j,1} \mid S_1 = 1)$
  - We solve a tiny convex optimization problem within EM

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

Experimental evaluation

- We create a COPD cohort of 3,705 patients:
  - At least one COPD-related diagnosis code
  - At least one COPD-related drug
- Clinical findings from 264 most common diagnosis codes
- Combined visits into 3-month time windows
- In total: 34,976 visits, 189,815 positive findings

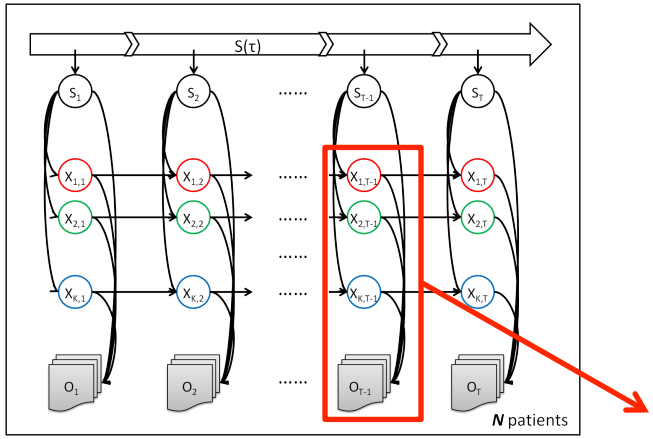
Specifying the latent variables

- 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)

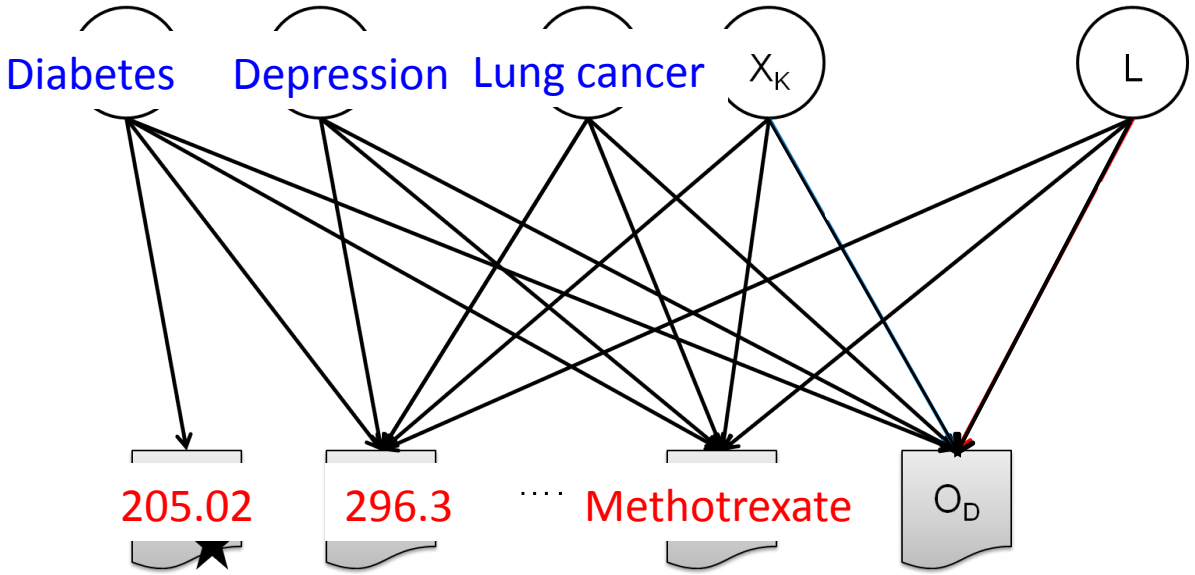
- Can be viewed as a type of weak supervision, using clinical domain knowledge
- Without these, the results are less interpretable

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 Infectio**
*585.4	0.08	Chronic Kidney Diseas
*584.9	0.07	Acute Renal Failure, U
*586	0.07	Renal Failure, Unspeci
782.3	**0.06**	**Edema**
*585.6	0.05	End Stage Renal Disea
593.9	**0.04**	**Unspecified Disorder**
272.4	**0.04**	**Other And Unspecifie**
272.2	**0.03**	**Mixed Hyperlipidemia**

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.

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**

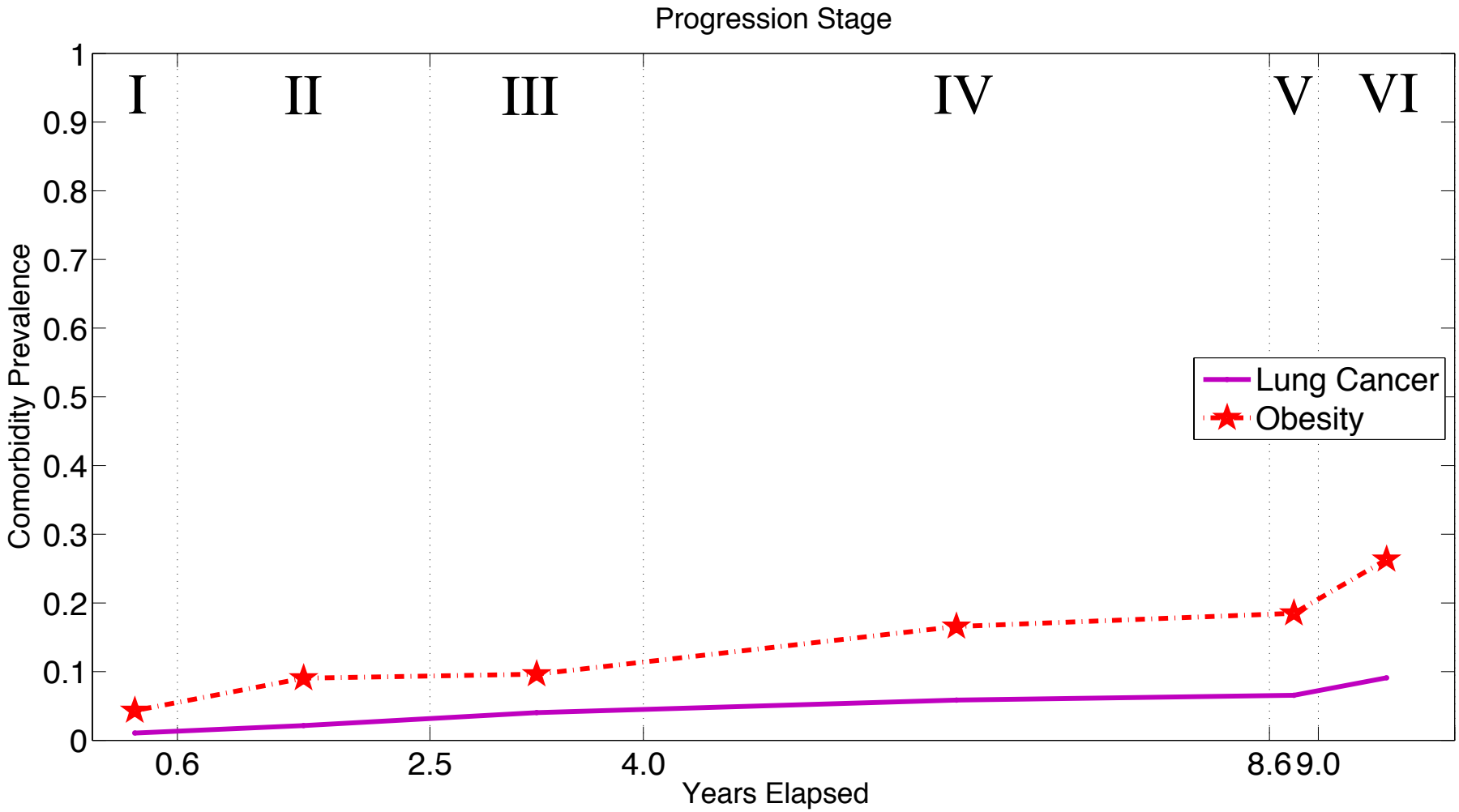
Edges learned for *lung cancer*

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

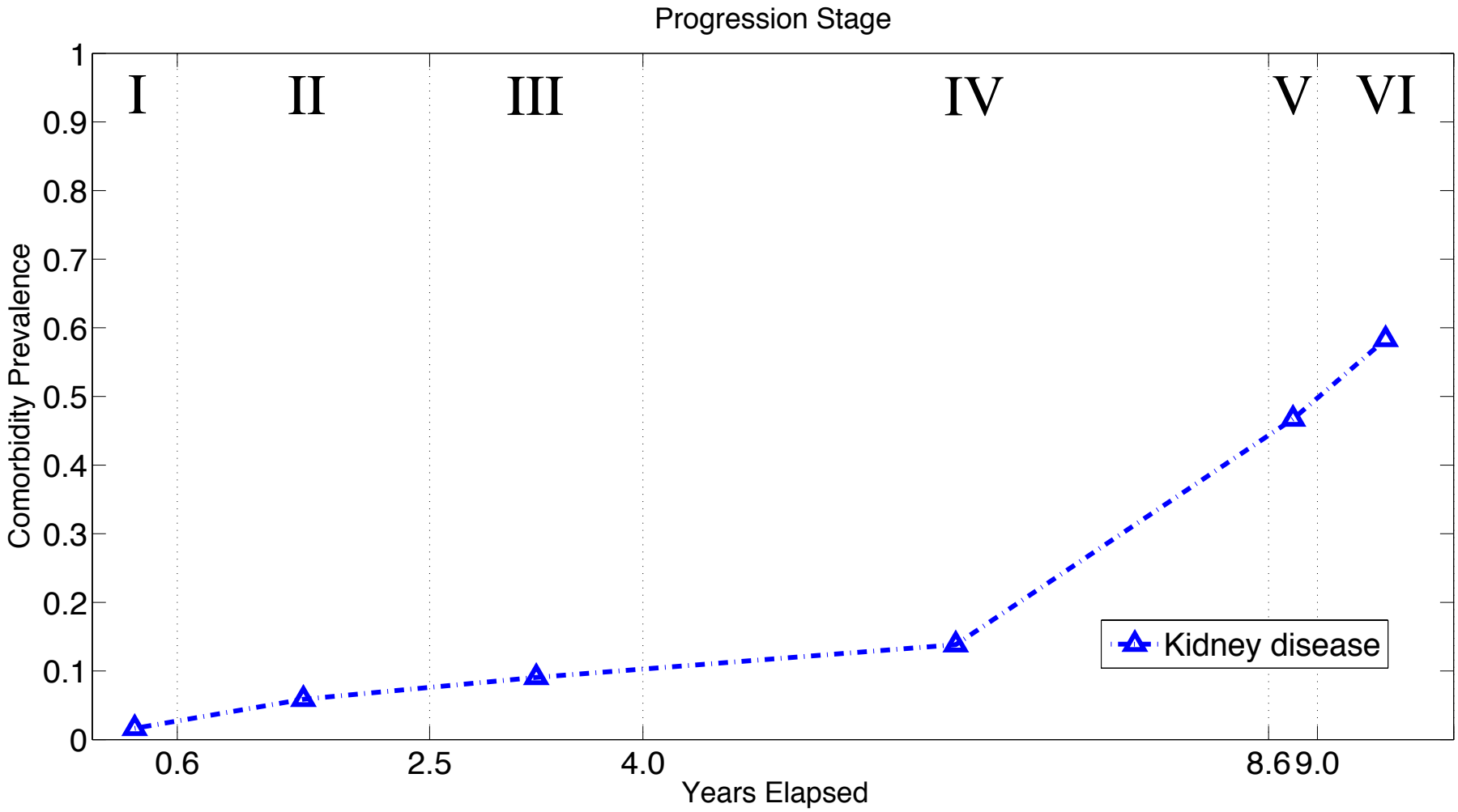
Edges learned for *lung infection*

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

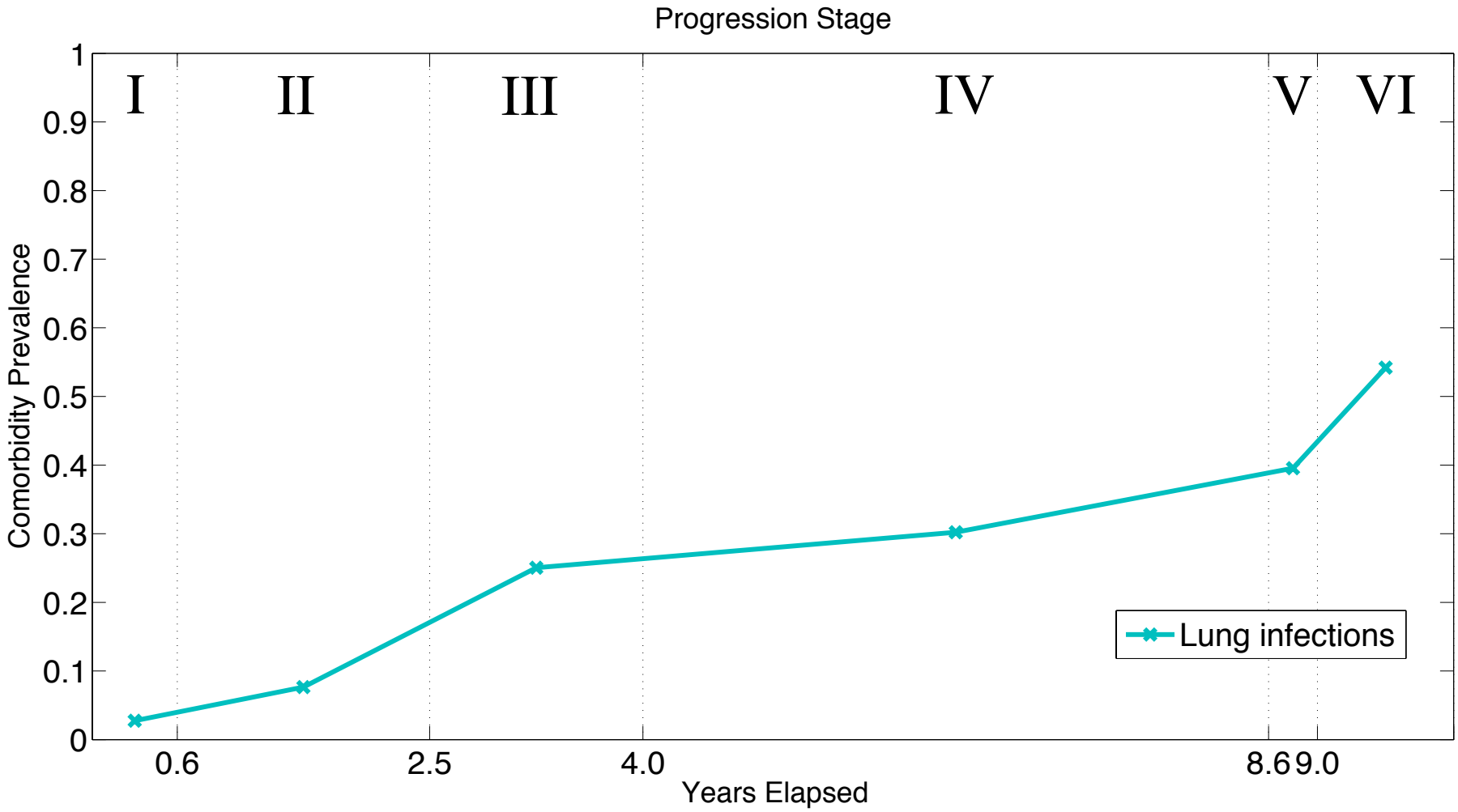
Prevalence of comorbidities across stages (Lung cancer and Obesity)



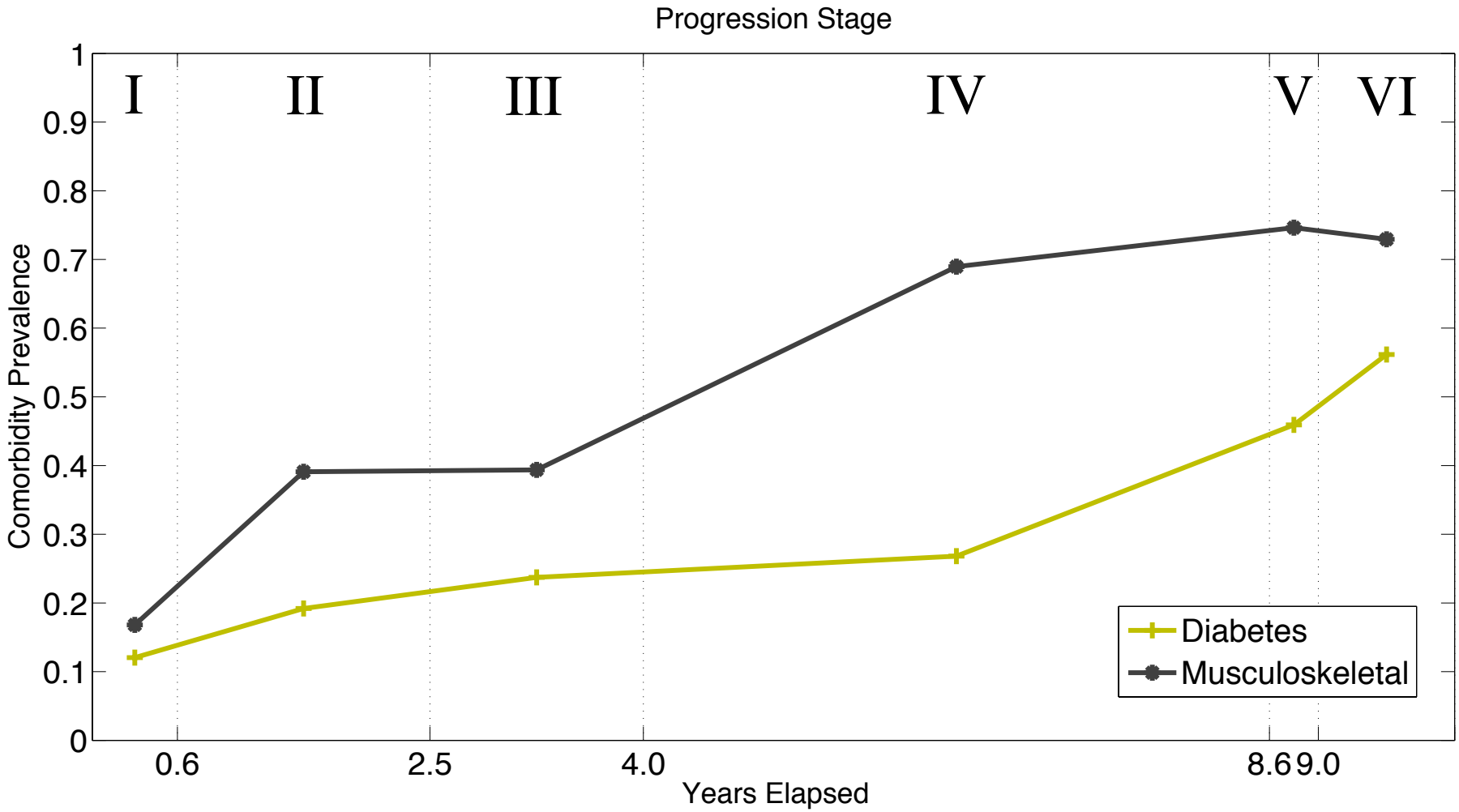
Prevalence of comorbidities across stages (Kidney disease)



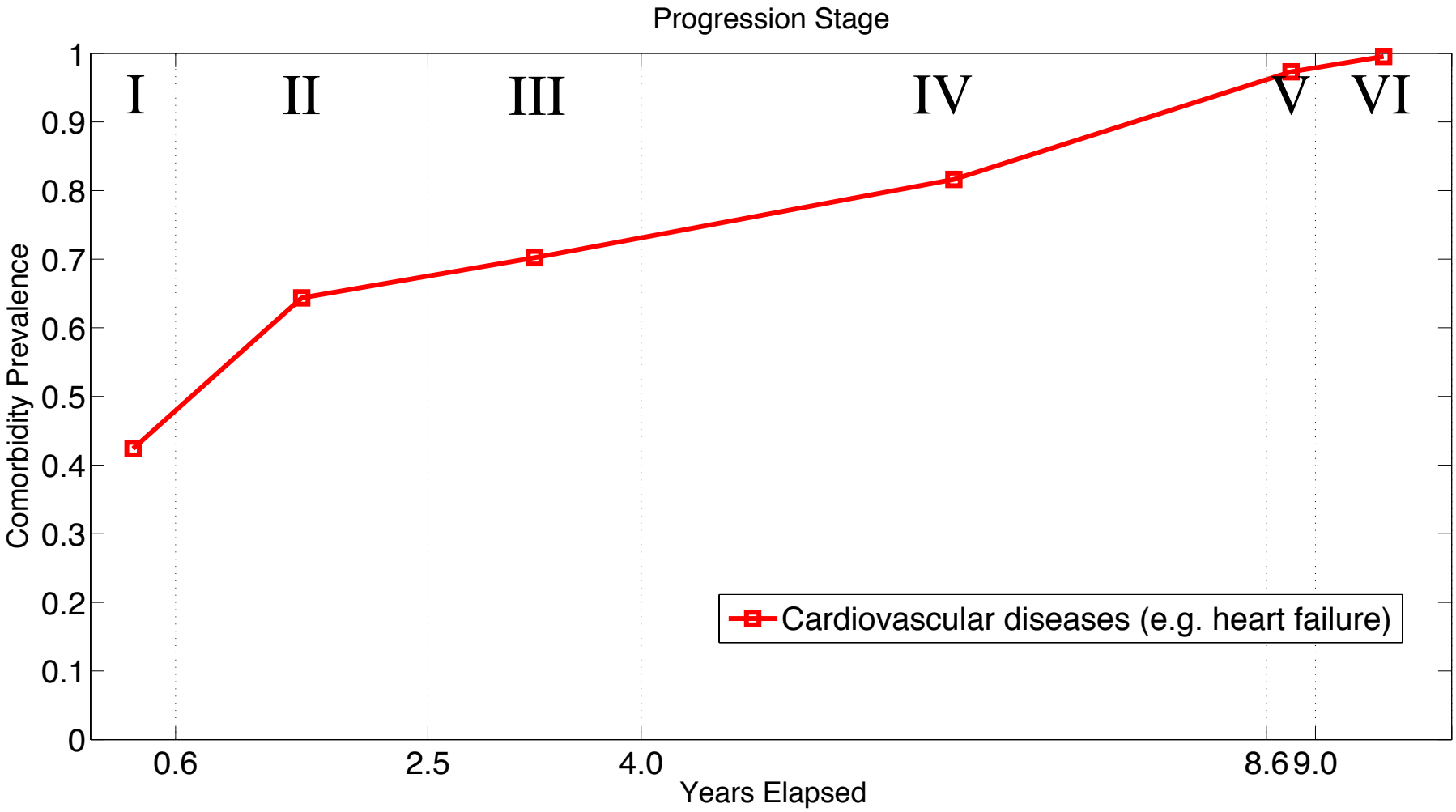
Prevalence of comorbidities across stages (Lung infections)



Prevalence of comorbidities across stages (Diabetes & Musculoskeletal disorders)



Prevalence of comorbidities across stages (Cardiovascular disease)



CHEST™ JOURNAL
Official Publication of the American College o

August 2009, Vol 136, No. 2

< Previous in this issue

Next in this issue >

Editorials | August 2009

Is COPD Really a Cardiovascular Disease?

FREE TO VIEW

Don D. Sin, MD, FCCP

Author and Funding Information

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

Text Size: A A A

Related editorial/commentary:

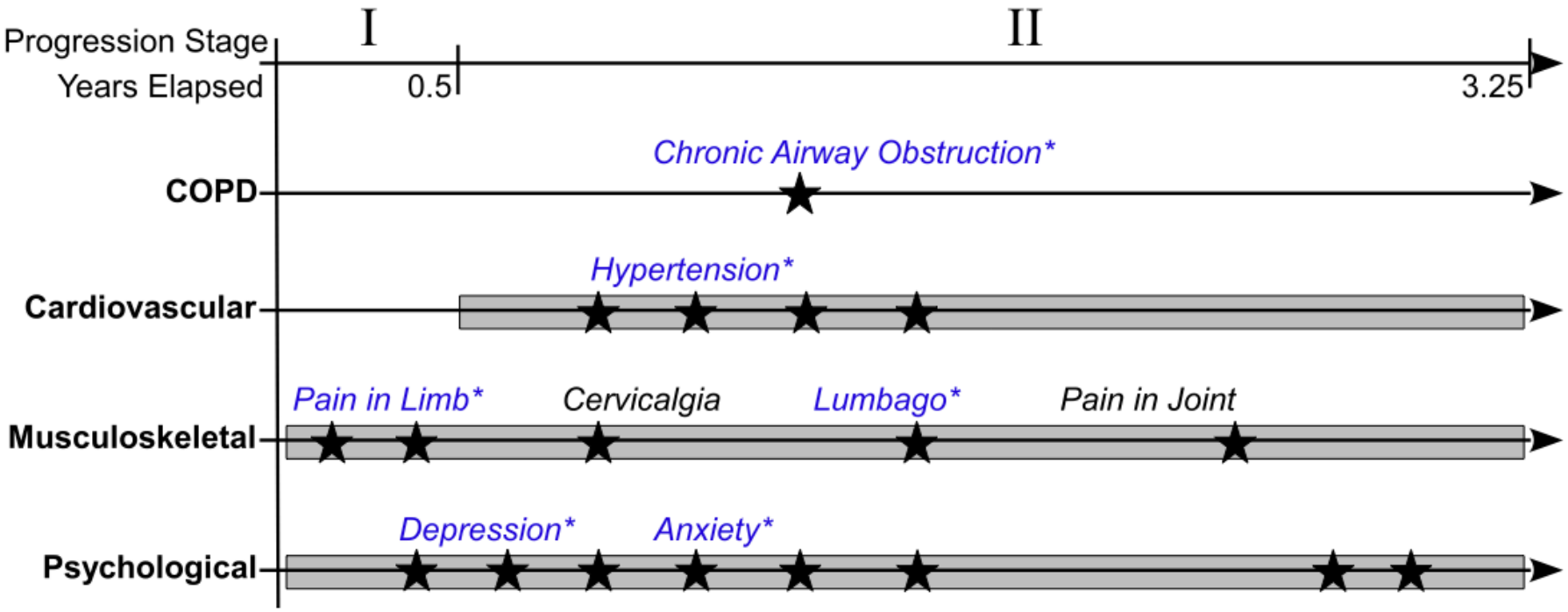
A Postmortem Analysis of Major Causes of Early Death in Patients Hospitalized With COPD Exacerbation (*Chest.* 2009;136(2):376-380.)

Article | References

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.

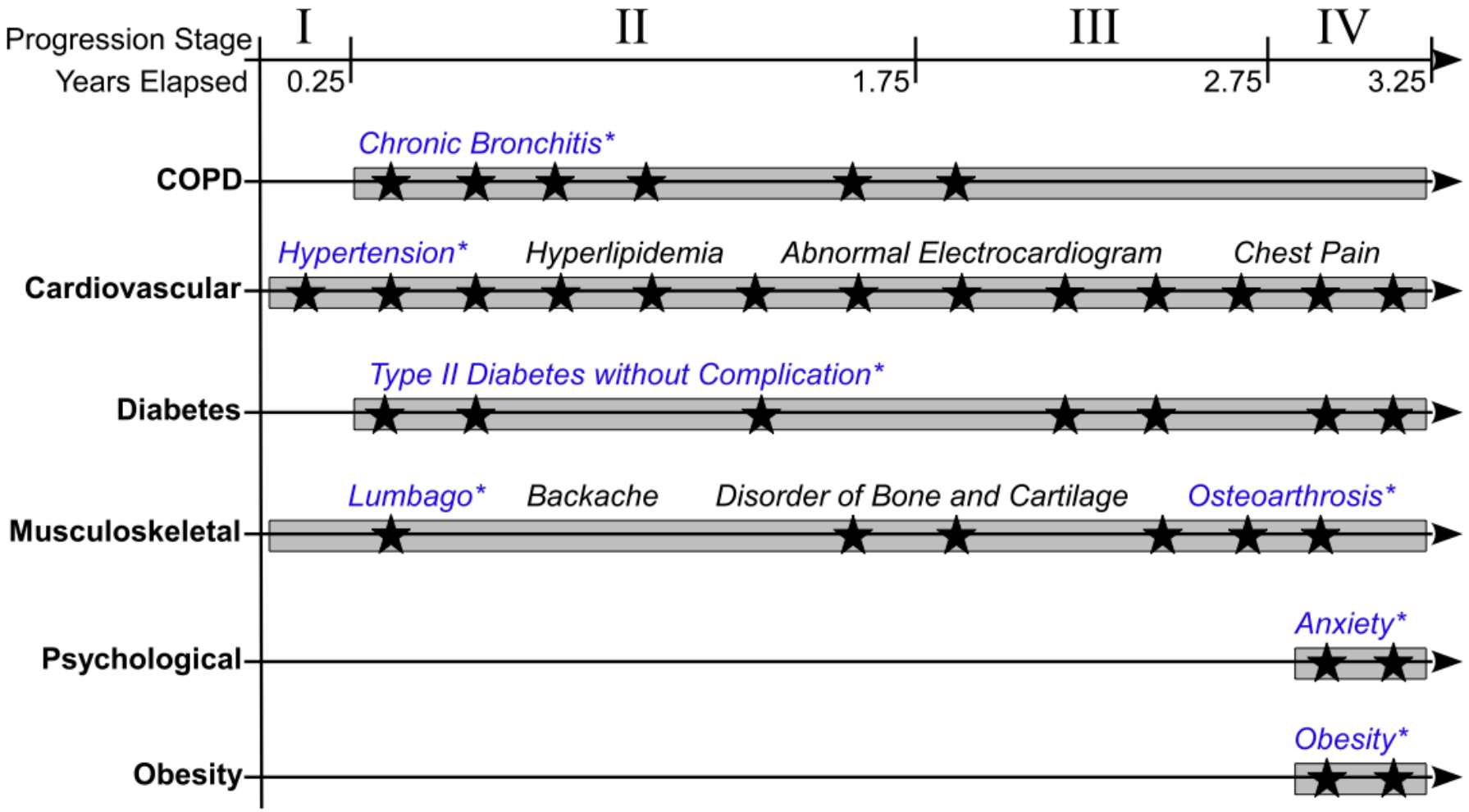
Inference of disease progression

An individual with a stable trajectory



Inference of disease progression

An individual with a progressive trajectory



Conclusion and future work

- We present a continuous-time disease progression model that can learn from censored EHR data with no direct supervision
- We applied it to a real COPD cohort and derived medically meaningful results
- Future work includes:
  - Interaction with physicians for validation and feedback
  - Model multiple trajectories
  - Incorporate medication/treatment data

More info: clinicalml.org