

Machine Learning for Healthcare

HST.956, 6.S897

Lecture 5: Risk stratification (continued)

David Sontag



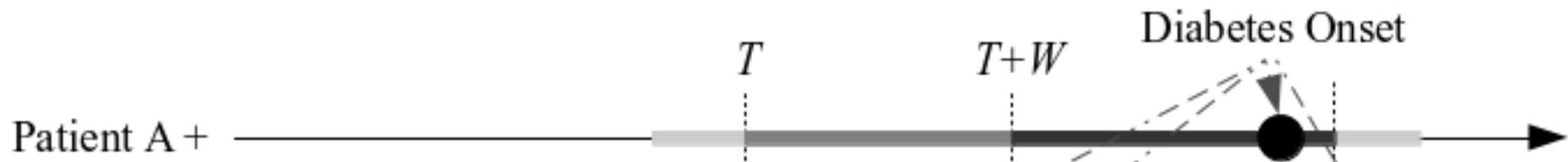
Course announcements

- Recitation Friday at 2pm (4-153) – optional
- PS1 due tonight; PS2 out Tuesday

Outline for today's class

1. Risk stratification (continued)
 - Deriving labels
 - Evaluation
 - Subtleties with ML-based risk stratification
2. Survival modeling

Where do the labels come from?



Typical pipeline:

1. Manually label several patients' data by "chart review"
2. A) Come up with a simple rule to automatically derive label for all patients, **or**
B) Use machine learning to get the labels themselves

Step 1:

Visualization of individual patient data is an important part of chart review

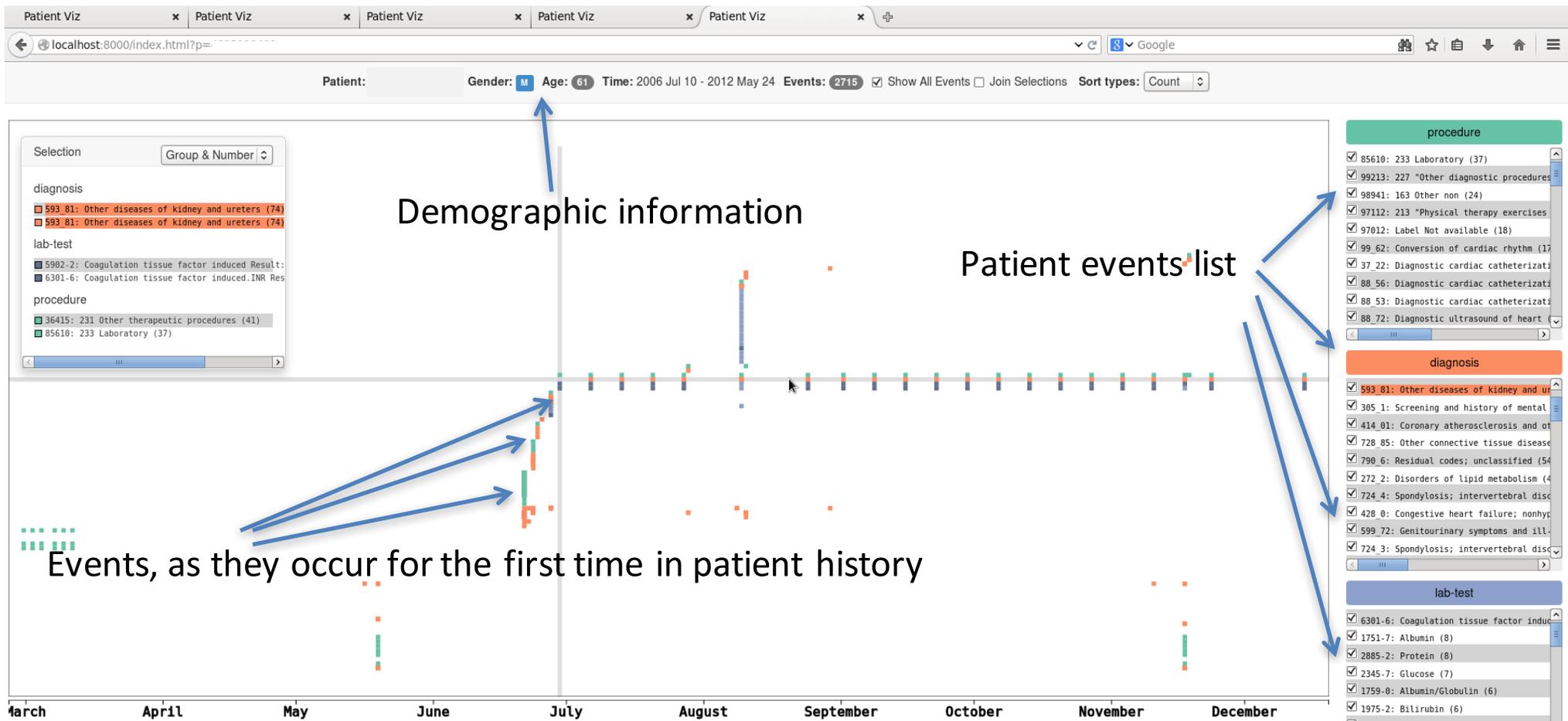
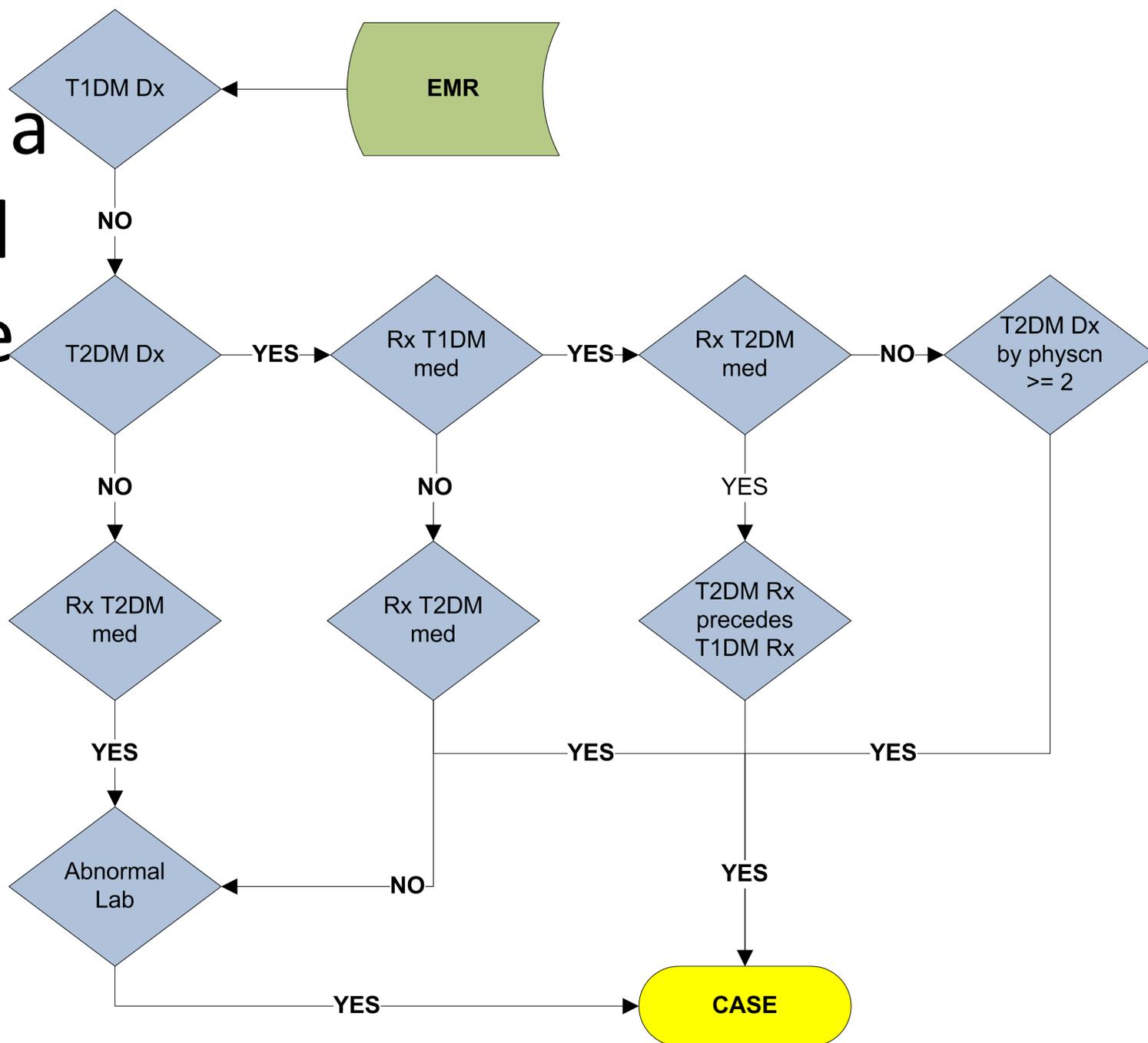


Figure 1: Algorithm for identifying T2DM cases in the EMR.

Step 2: Example of a rule-based phenotype



Step 2: Example of a rule-based phenotype

https://www.phekb.org/phenotypes?field_pgx_type_tid_1=398&field_data_model_value=All

PheKB a knowledgebase for discovering phenotypes from electronic medical records

Login | Request Account

Home | **Phenotypes** | Resources | Contact Us

Public Phenotypes

Public | Collaboration

Public phenotypes are believed to be complete and final by their authors. When you are logged in you can view and edit phenotypes in your groups that are non public and in various stages of development.

[Login To View Private Group Phenotypes](#)

Institution: Type of Phenotype: Owner Phenotyping Groups: View Phenotyping Groups:

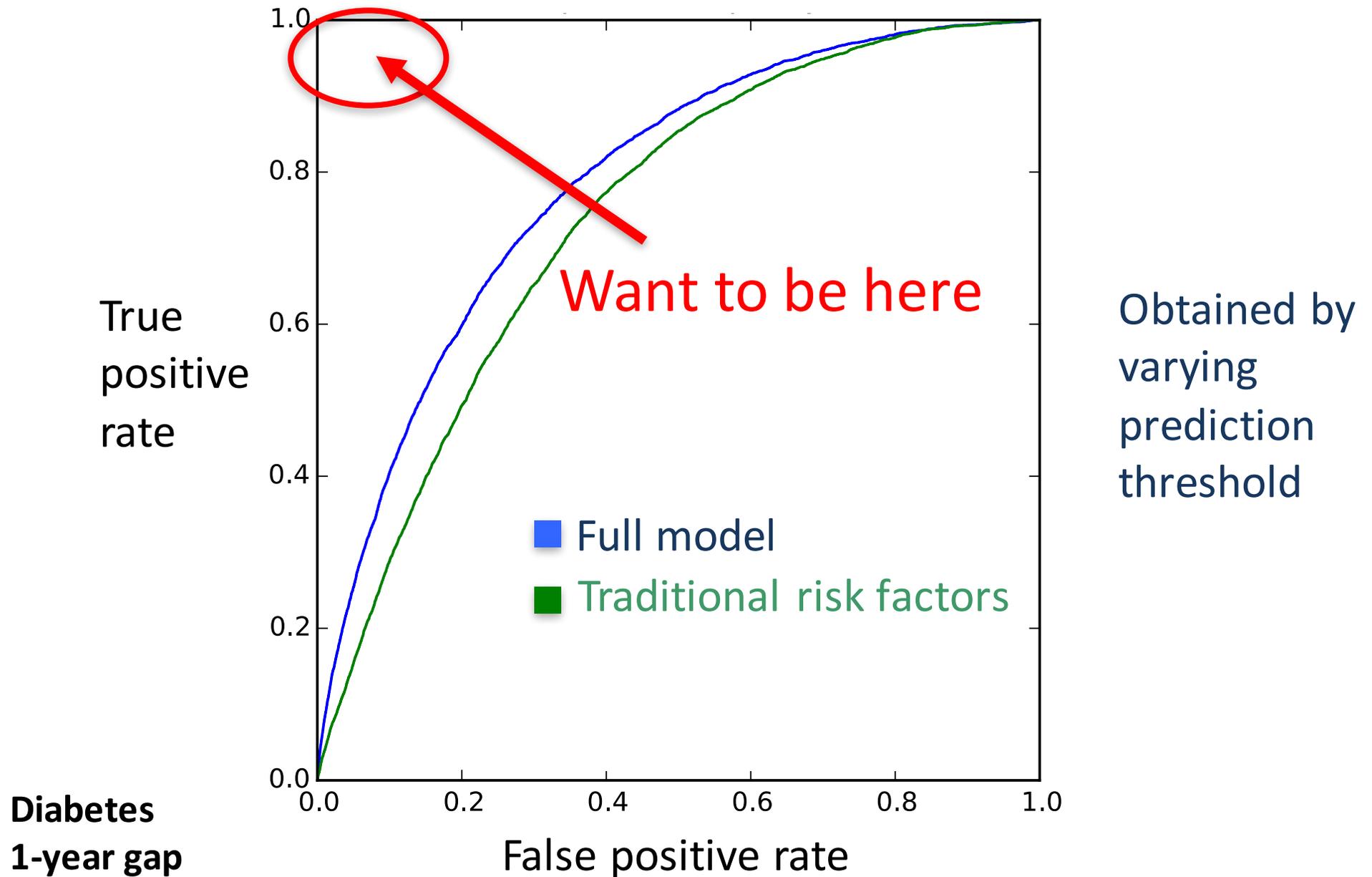
Data Model: [Apply](#)

Title	Institution	Data Modalities and Methods Used	Owner Phenotyping Groups	View Groups	Has new content	Status	Type
Abdominal Aortic Aneurysm (AAA)	Geisinger	CPT Codes, ICD 9 Codes, Vital Signs	eMERGE Geisinger Group	eMERGE Geisinger Group, eMERGE Phenotype WG		Final	Disease or Syndrome
ADHD phenotype algorithm	CHOP	ICD 9 Codes, Medications, Natural Language Processing	eMERGE CHOP Group	eMERGE Phenotype WG		Final	Disease or Syndrome
Appendicitis	Cincinnati Children's Hospital Medical Center	CPT Codes, ICD 9 Codes, Medications, Natural Language Processing	eMERGE CCHMC/BCH Group	eMERGE Phenotype WG		Final	Disease or Syndrome
Atrial Fibrillation - Demonstration Project	Vanderbilt University	CPT Codes, ICD 9 Codes, Natural Language Processing	Vanderbilt - SD/RD Group	Vanderbilt - SD/RD Group		Final	Disease or Syndrome
Autism	Cincinnati Children's Hospital Medical Center	ICD 9 Codes, Medications, Natural Language Processing	eMERGE CCHMC/BCH Group	eMERGE Phenotype WG		Final	Disease or Syndrome
Cataracts	Marshfield Clinic Research Foundation	CPT Codes, ICD 9 Codes, Medications, Natural Language Processing	eMERGE Marshfield Group	eMERGE Phenotype WG		Final	Disease or Syndrome
Crohn's Disease -	Vanderbilt University	ICD 9 Codes, Medications,	Vanderbilt -	Vanderbilt -		Final	Disease

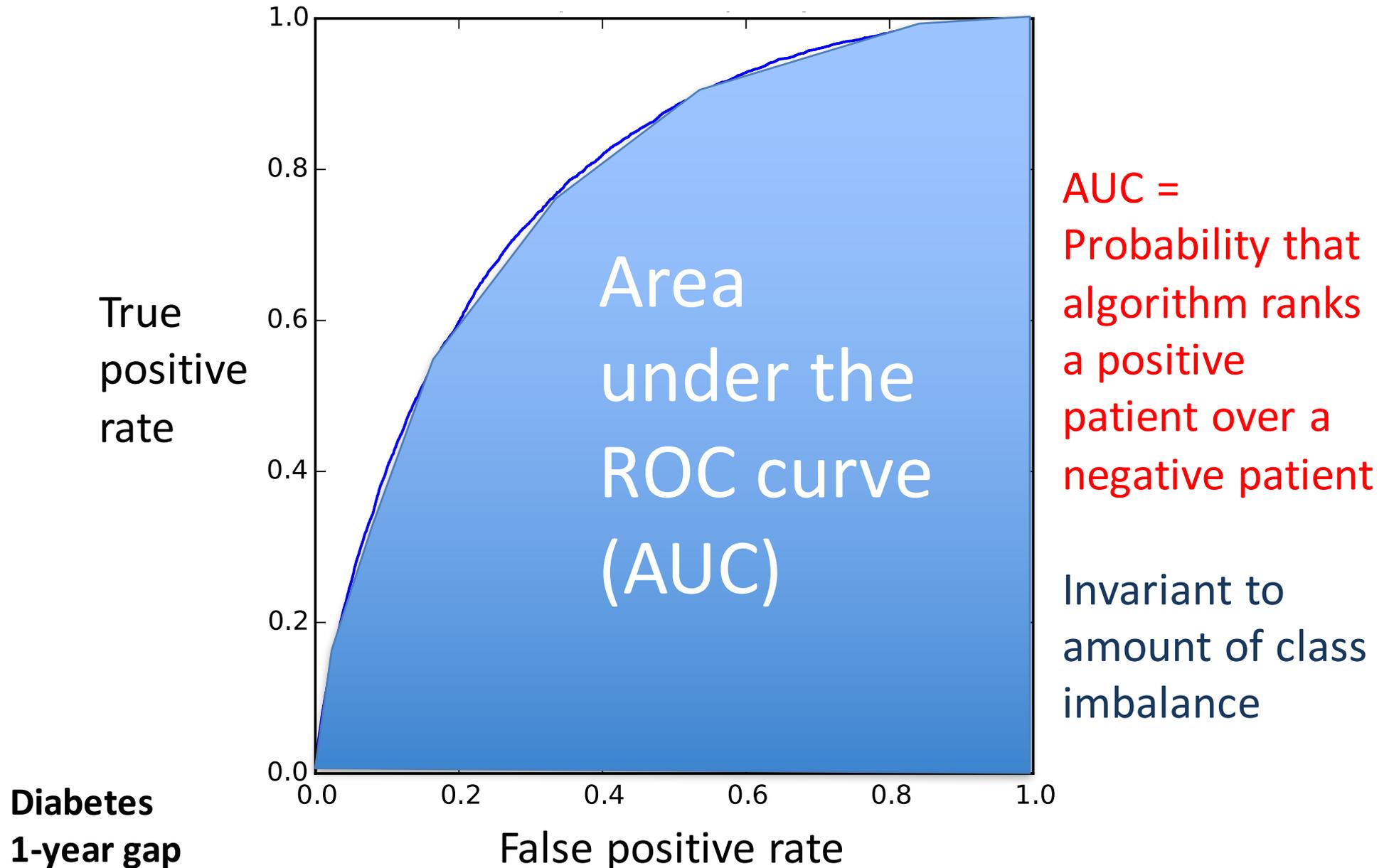
Outline for today's class

1. Risk stratification (continued)
 - Deriving labels
 - **Evaluation**
 - Subtleties with ML-based risk stratification
2. Survival modeling

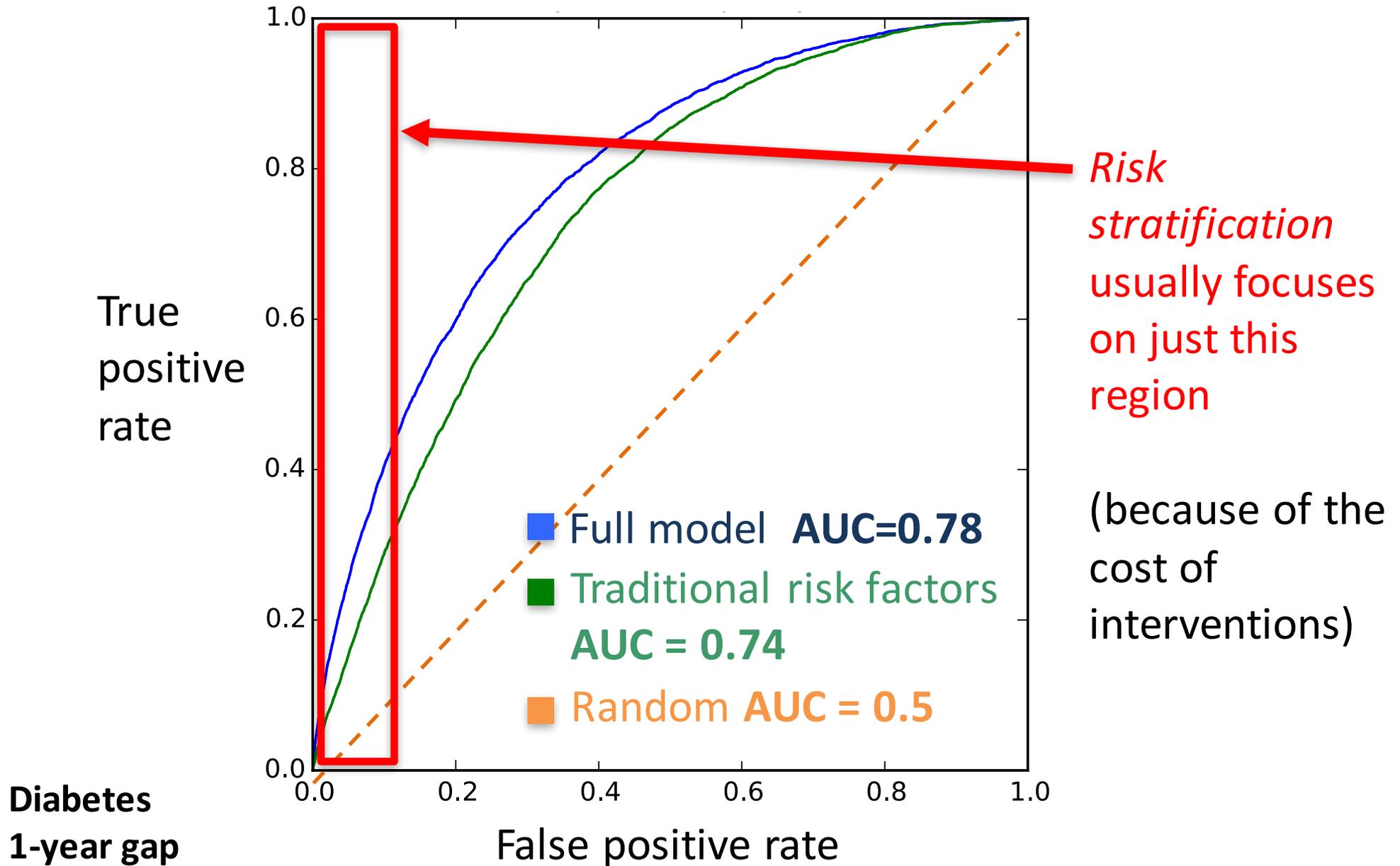
Receiver-operator characteristic curve



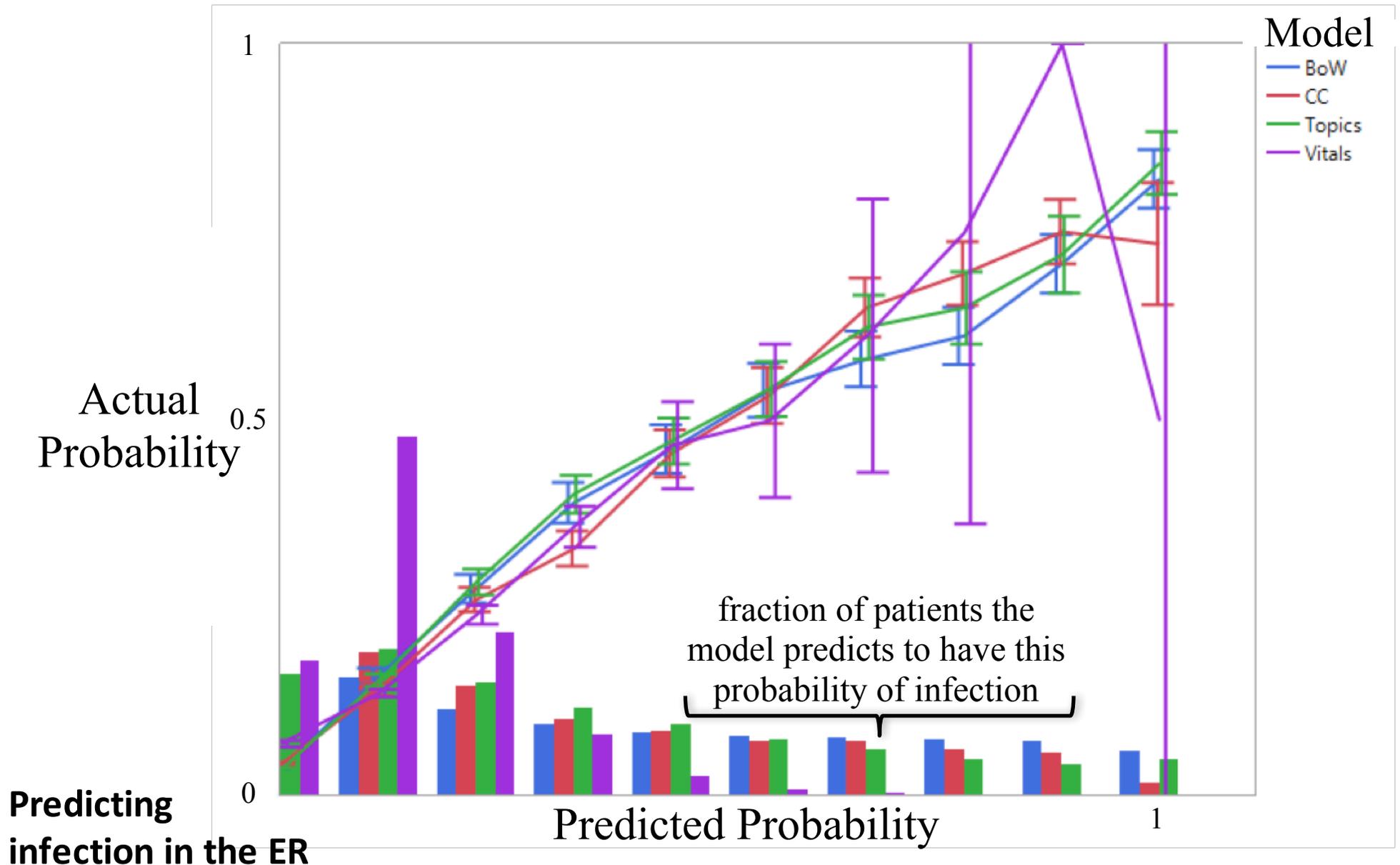
Receiver-operator characteristic curve



Receiver-operator characteristic curve



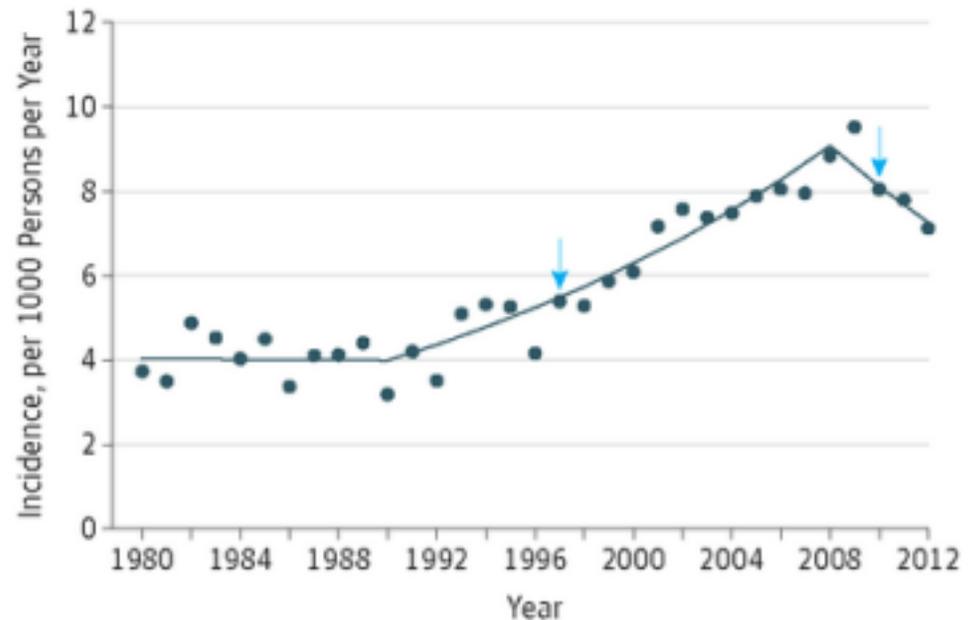
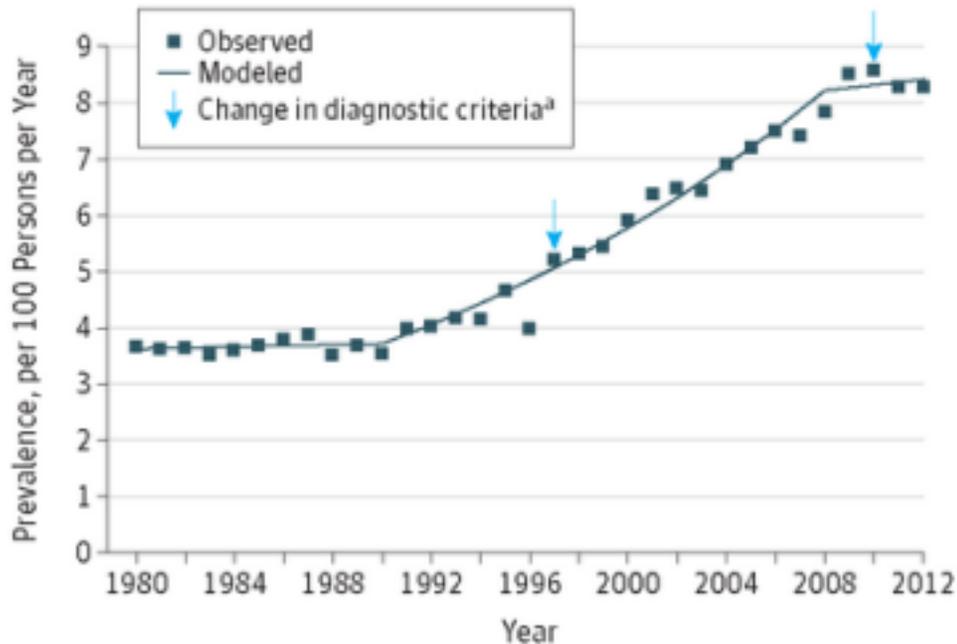
Calibration (*note: different dataset*)



Outline for today's class

1. Risk stratification (continued)
 - Deriving labels
 - Evaluation
 - **Subtleties with ML-based risk stratification**
2. Survival modeling

Non-stationarity: *Diabetes Onset After 2009*

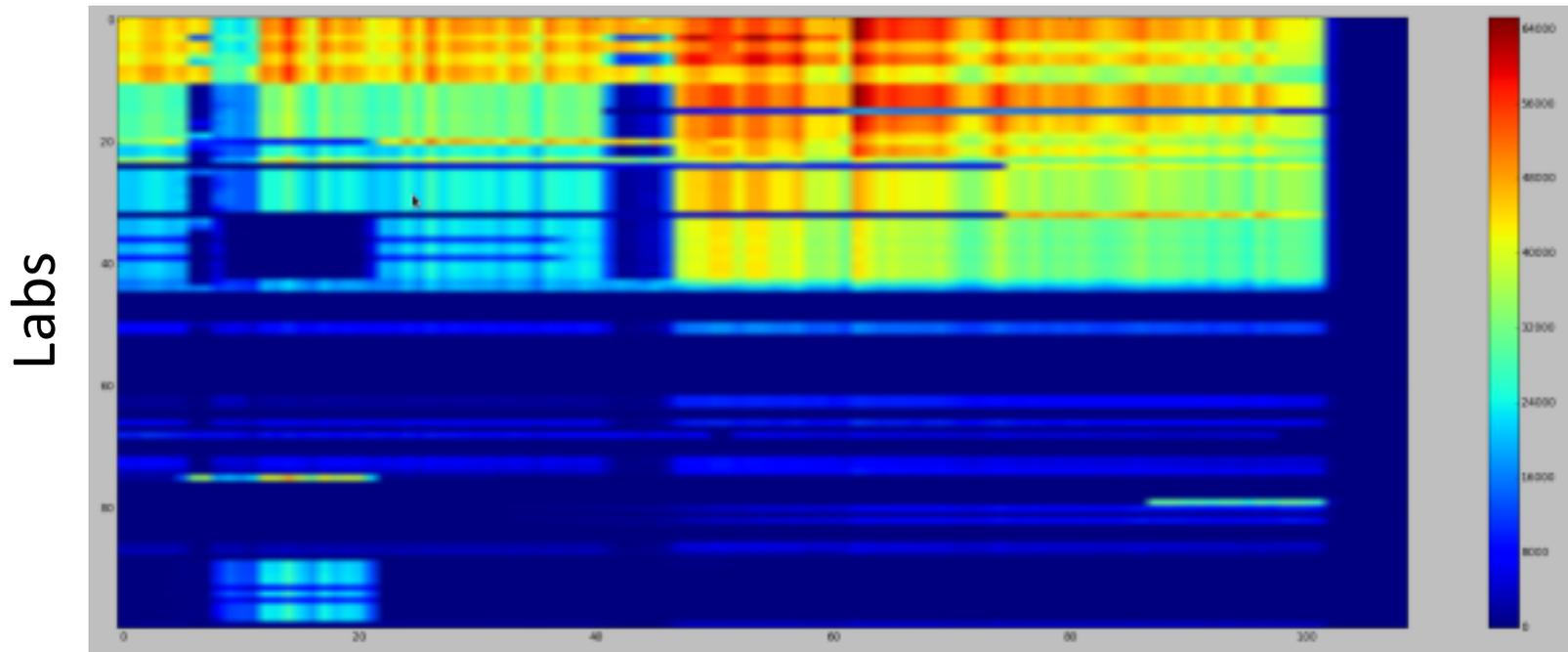


→ Automatically derived labels may change meaning

[Geiss LS, Wang J, Cheng YJ, et al. Prevalence and Incidence Trends for Diagnosed Diabetes Among Adults Aged 20 to 79 Years, United States, 1980-2012. JAMA, 2014.]

Non-stationarity:

Top 100 lab measurements over time

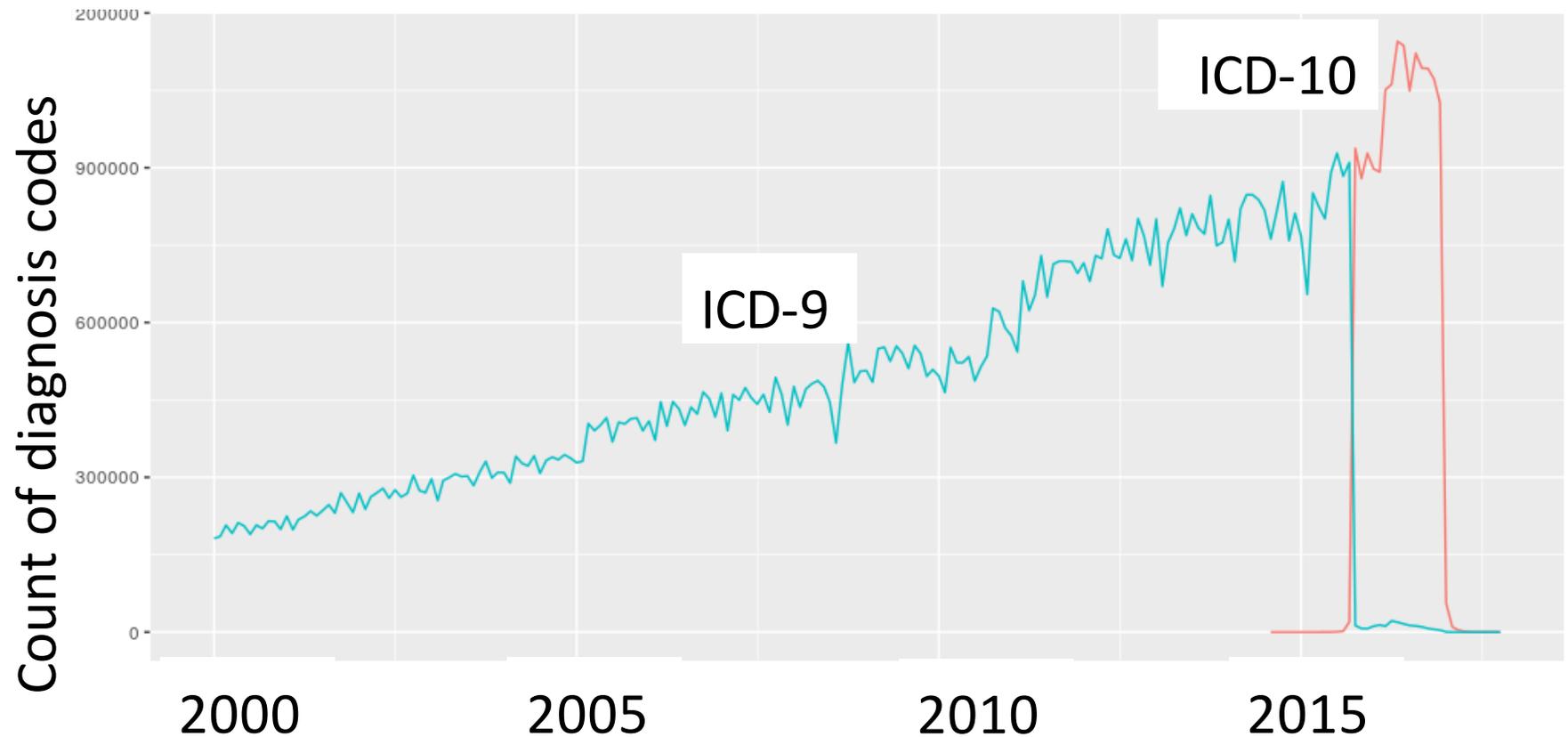


Time (in months, from 1/2005 up to 1/2014)

→ Significance of features may change over time

[Figure credit: Narges Razavian]

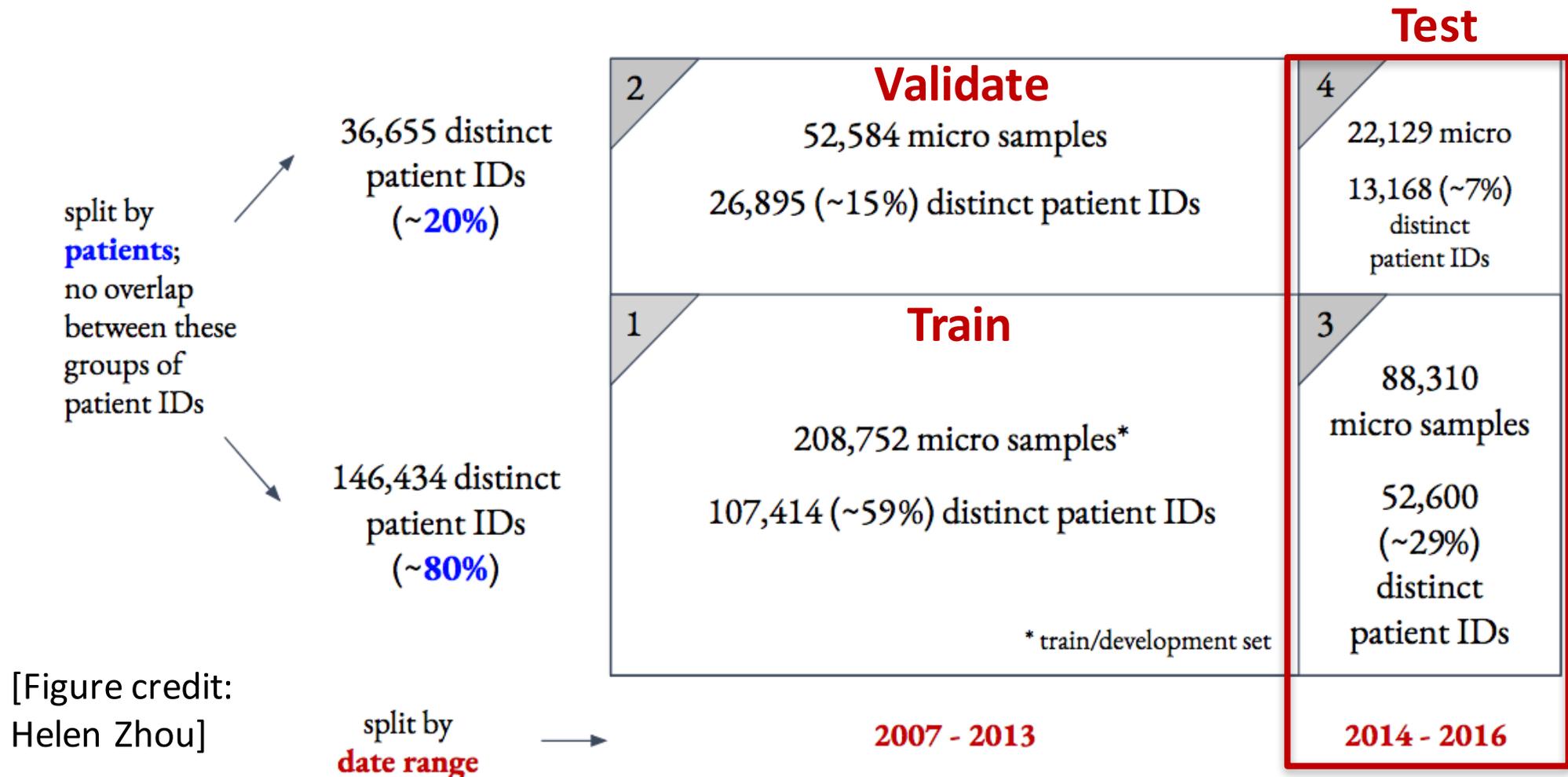
Non-stationarity: *ICD-9 to ICD-10 shift*



→ Significance of features may change over time

Re-thinking evaluation in the face of non-stationarity

- How was our diabetes model evaluation flawed?
- Good practice: use test data from a future year:



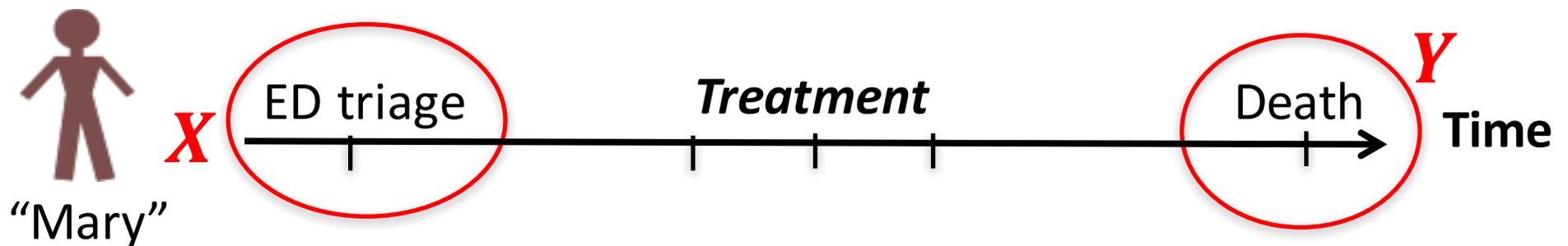
[Figure credit:
Helen Zhou]

Intervention-tainted outcomes

- Example from today's readings:
 - Patients with pneumonia who have a history of asthma have lower risk of dying from pneumonia
 - Thus, we learn: **HasAsthma(x) => LowerRisk(x)**
- **What's wrong with the learned model?**
 - Risk stratification drives **interventions**
 - If low risk, might not admit to ICU. But this was precisely what prevented patients from dying!

Intervention-tainted outcomes

- Formally, this is what's happening:



A long survival time may be because of treatment!

- How do we address this problem?
- First and foremost, must recognize it is happening
 - interpretable models help with this

Intervention-tainted outcomes

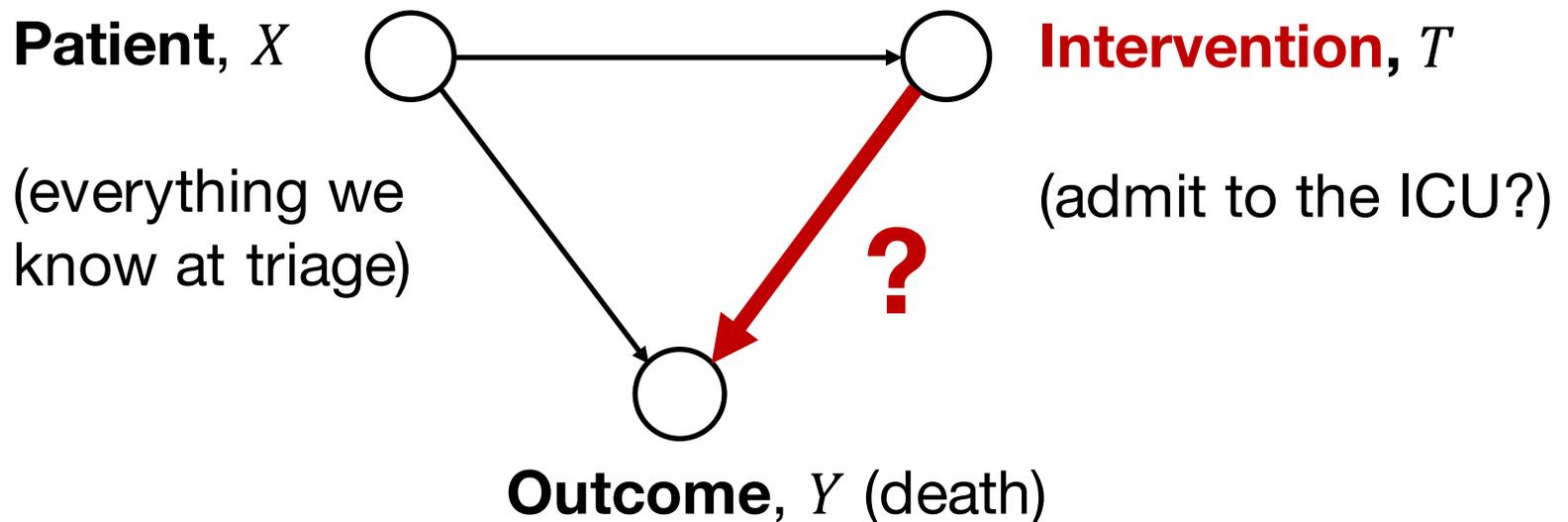
- Hacks:
 1. Modify model, e.g. by removing the **HasAsthma(x) => LowerRisk(x)** rule
I do not expect this to work with high-dimensional data
 2. Re-define outcome by finding a pre-treatment surrogate (e.g., lactate levels)
 3. Consider treated patients as **right-censored** by treatment

Example:

Henry, Hager, Pronovost, Saria. A targeted real-time early warning score (TREWScore) for septic shock. *Science Translation Medicine*, 2015

Intervention-tainted outcomes

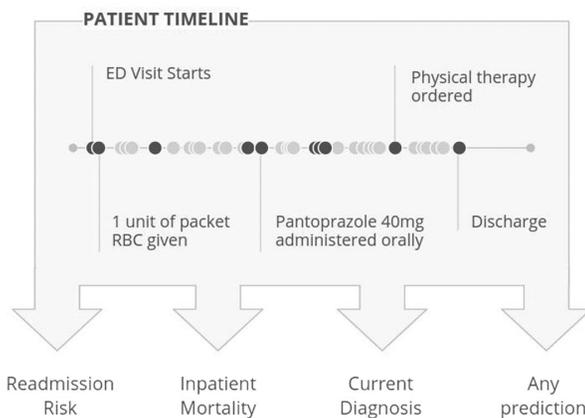
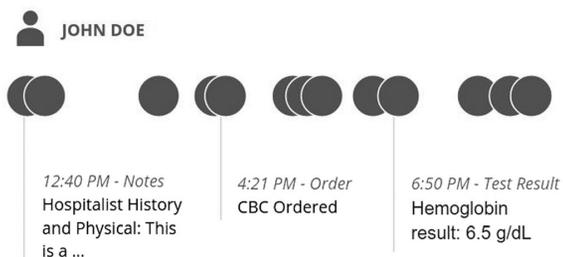
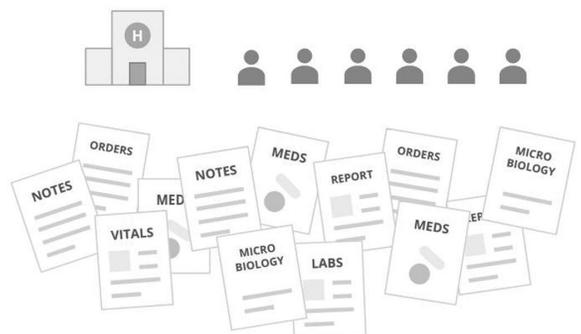
- The rigorous way to address this problem is through the language of **causality**:



Will admission to ICU lower likelihood of death for patient?

- We return to this in Lecture 14

No big wins from deep models on structured data/text



1

Health systems collect and store electronic health records in various formats in databases.

2

All available data for each patient is converted to events recorded in containers based on the Fast Healthcare Interoperability Resource (FHIR) specification.

3

The FHIR resources are placed in temporal order, depicting all events recorded in the EHR (i.e. timeline). The deep learning model uses this full history to make each prediction.

Rajkomar et al., Scalable and accurate deep learning with electronic health records. *Nature Digital Medicine*, 2018

Recurrent neural network & attention-based models trained on 200K hospitalized patients

No big wins from deep models on structured data/text

Supplemental Table 1: Prediction accuracy of each task of deep learning model compared to baselines

	Hospital A	Hospital B
Inpatient Mortality, AUROC¹(95% CI)		
Deep learning 24 hours after admission	0.95 (0.94-0.96)	0.93 (0.92-0.94)
Full feature enhanced baseline at 24 hours after admission	0.93 (0.92-0.95)	0.91 (0.89-0.92)
Full feature simple baseline at 24 hours after admission	0.93 (0.91-0.94)	0.90 (0.88-0.92)
Baseline (aEWS ²) at 24 hours after admission	0.85 (0.81-0.89)	0.86 (0.83-0.88)
30-day Readmission, AUROC (95% CI)		
Deep learning at discharge	0.77 (0.75-0.78)	0.76 (0.75-0.77)
Full feature enhanced baseline at discharge	0.75 (0.73-0.76)	0.75 (0.74-0.76)
Full feature simple baseline at discharge	0.74 (0.73-0.76)	0.73 (0.72-0.74)
Baseline (mHOSPITAL ³) at discharge	0.70 (0.68-0.72)	0.68 (0.67-0.69)
Length of Stay at least 7 days AUROC (95% CI)		
Deep learning 24 hours after admission	0.86 (0.86-0.87)	0.85 (0.85-0.86)
Full feature enhanced baseline at 24 hours after admission	0.85 (0.84-0.85)	0.83 (0.83-0.84)
Full feature simple baseline at 24 hours after admission	0.83 (0.82-0.84)	0.81 (0.80-0.82)
Baseline (mLiu ⁴) at 24 hours after admission	0.76 (0.75-0.77)	0.74 (0.73-0.75)

Comparison to Razavian et al. '15

[Rajkomar et al. '18 **electronic supplementary material**:

https://static-content.springer.com/esm/art%3A10.1038%2Fs41746-018-0029-1/MediaObjects/41746_2018_29_MOESM1_ESM.pdf]

No big wins from deep models on structured data/text

Supplemental Table 1: Prediction accuracy of each task of deep learning model compared to baselines

	Hospital A	Hospital B	
Inpatient Mortality, AUROC¹(95% CI)			
Deep learning 24 hours after admission	0.95 (0.94-0.96)	0.93 (0.92-0.94)	Comparison to Razavian '15
Full feature enhanced baseline at 24 hours after admission	0.93 (0.92-0.95)	0.91 (0.89-0.92)	
Full feature simple baseline at 24 hours after admission	0.88 (0.87-0.89)	0.88 (0.88-0.89)	
Baseline (mLiu ⁴) at 24 hours after admission	0.76 (0.75-0.77)	0.74 (0.73-0.75)	
30-day Mortality, AUROC¹(95% CI)			
Deep learning 24 hours after admission	0.86 (0.86-0.87)	0.85 (0.85-0.86)	
Full feature enhanced baseline at 24 hours after admission	0.85 (0.84-0.85)	0.83 (0.83-0.84)	
Full feature simple baseline at 24 hours after admission	0.83 (0.82-0.84)	0.81 (0.80-0.82)	
Baseline (mLiu ⁴) at 24 hours after admission	0.76 (0.75-0.77)	0.74 (0.73-0.75)	

Keep in mind:

Small wins with deep models may disappear altogether with dataset shift or non-stationarity (Jung & Shah, JBI '15)

[Rajkumar et al. '18 electronic supplementary material:

https://static-content.springer.com/esm/art%3A10.1038%2Fs41746-018-0029-1/MediaObjects/41746_2018_29_MOESM1_ESM.pdf]

No big wins from deep models on structured data/text – why?

- Sequential data in medicine is very different from language modeling
 - Many time scales, significant missing data, and multi-variate observations
 - Likely *do exist* predictive nonlinear interactions, but subtle
 - Not enough data to naively deal with the above two
- Medical community has already come up with some very good features

Outline for today's class

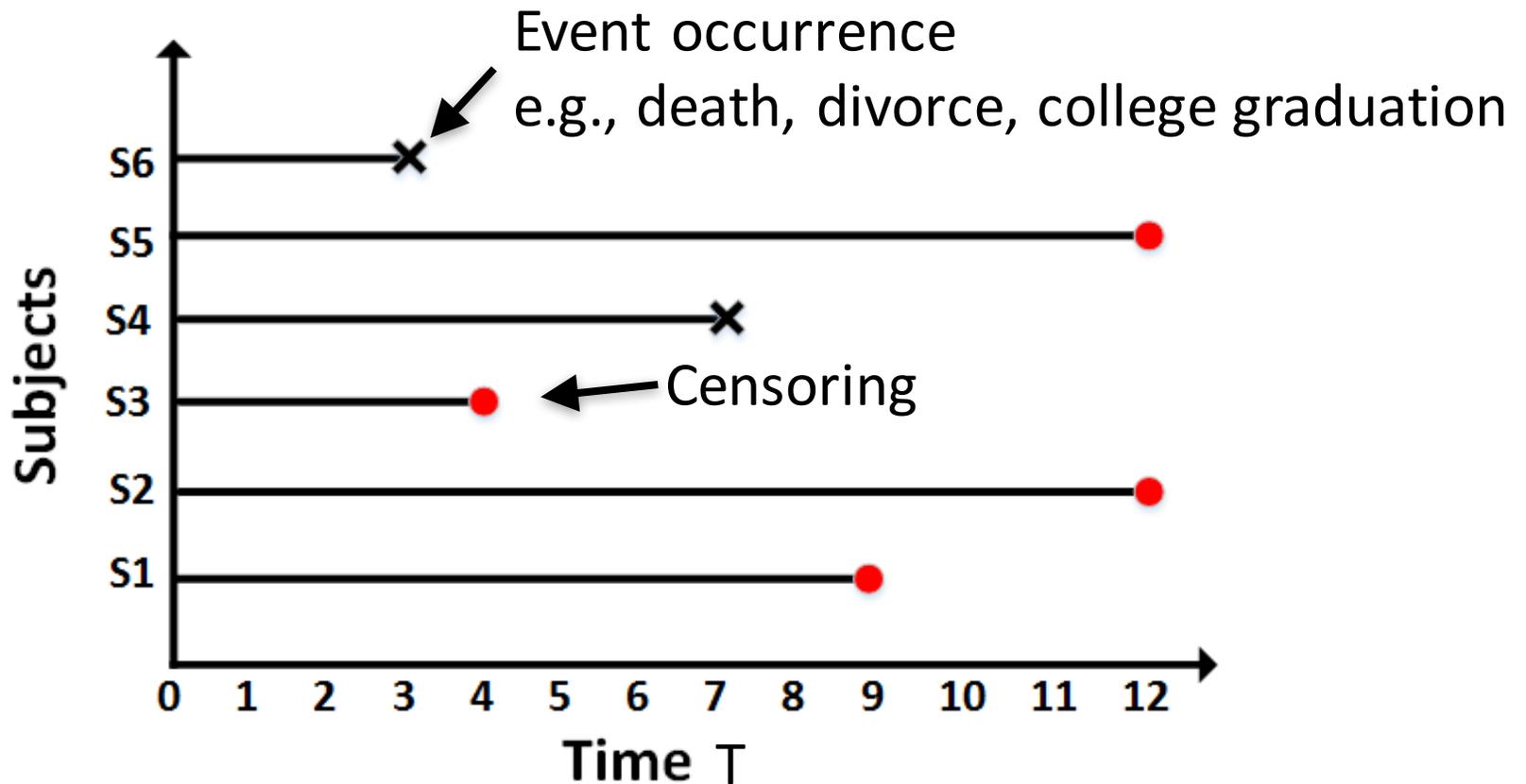
1. Risk stratification (continued)

- Deriving labels
- Evaluation
- Subtleties with ML-based risk stratification

2. Survival modeling

Survival modeling

- We focus on right-censored data:



Survival modeling

- Why not use classification, as before?
 - Less data for training (due to exclusions)
 - Pessimistic estimates due to choice of window
- What about regression, e.g. minimizing mean-squared error?
 - T is non-negative, may want long tails
 - If we just naively removed censored events, we would be introducing bias

Notation and formalization

- Data are (\mathbf{x}, T, b) =(features, time, censoring), where $b=0,1$ denotes whether time is of censoring or event occurrence
- Let $f(t) = P(t)$ be the probability of death at time t
- Survival function: the probability of an individual surviving beyond time t ,

$$S(t) = P(T > t) = \int_t^{\infty} f(x)dx$$

Notation and formalization

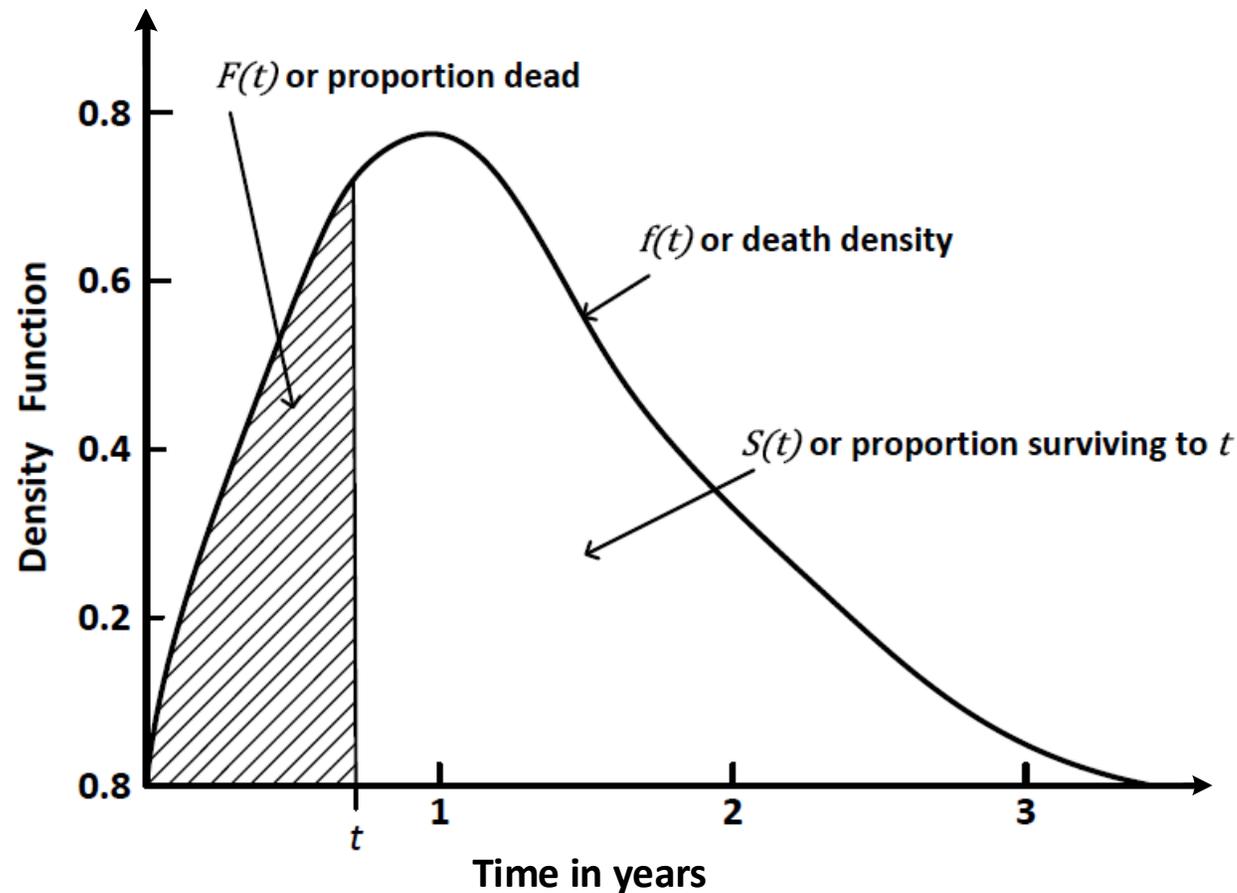
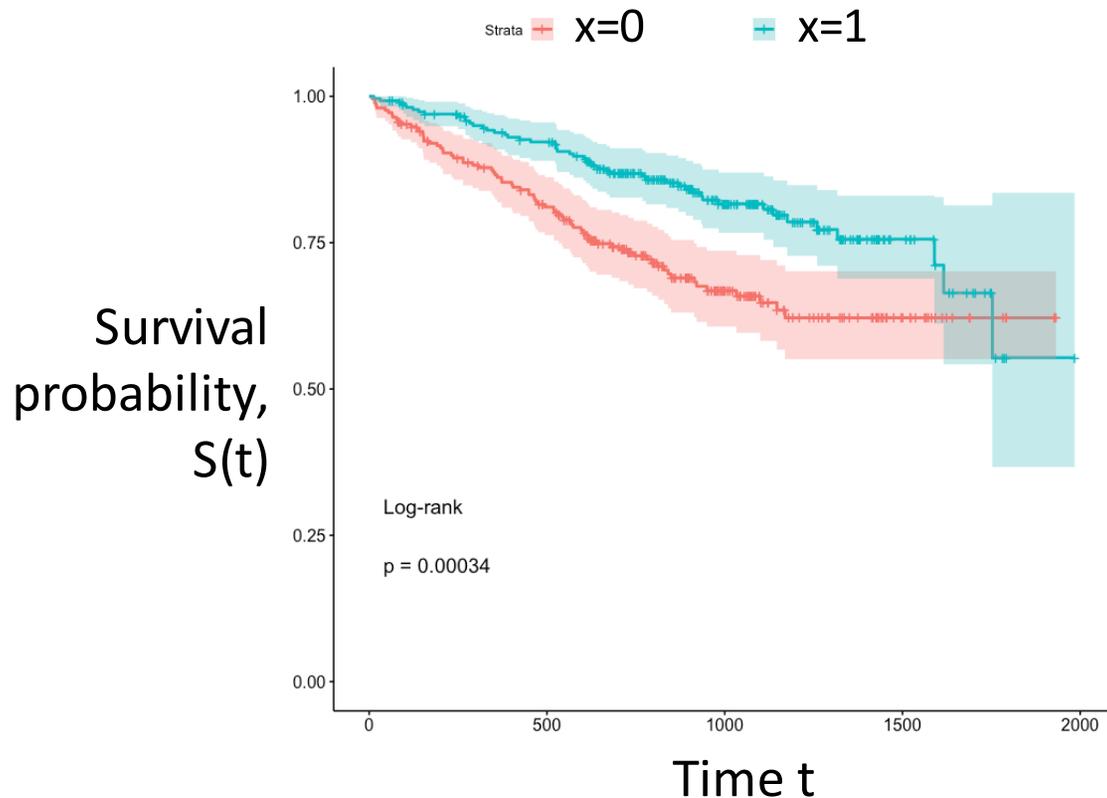


Fig. 2: Relationship among different entities $f(t)$, $F(t)$ and $S(t)$.

[Wang, Li, Reddy. Machine Learning for Survival Analysis: A Survey. 2017]

Kaplan-Meier estimator

- Example of a non-parametric method; good for unconditional density estimation



Observed event times

$$y_{(1)} < y_{(2)} < \dots < y_{(D)}$$

$d_{(k)}$ = # events at this time

$n_{(k)}$ = # of individuals alive and uncensored

$$\widehat{S}_{K-M}(t) = \prod_{k: y_{(k)} \leq t} \left\{ 1 - \frac{d_{(k)}}{n_{(k)}} \right\}$$

Maximum likelihood estimation

- Commonly parametric densities for $f(t)$:

Table 2.1 Useful parametric distributions for survival analysis

Distribution		Survival function $S(t)$	Density function $f(t)$
Exponential ($\lambda > 0$)		$\exp(-\lambda t)$	$\lambda \exp(-\lambda t)$
Weibull ($\lambda, \phi > 0$)		$\exp(-\lambda t^\phi)$	$\lambda \phi t^{\phi-1} \exp(-\lambda t^\phi)$
Log-normal ($\sigma > 0, \mu \in R$)	(parameters can be a function of x)	$1 - \Phi\{(\ln t - \mu)/\sigma\}$	$\varphi\{(\ln t - \mu)/\sigma\}(\sigma t)^{-1}$
Log-logistic ($\lambda > 0, \phi > 0$)		$1/(1 + \lambda t^\phi)$	$(\lambda \phi t^{\phi-1})/(1 + \lambda t^\phi)^2$
Gamma ($\lambda, \phi > 0$)		$1 - I(\lambda t, \phi)$	$\{\lambda^\phi / \Gamma(\phi)\} t^{\phi-1} \exp(-\lambda t)$
Gompertz ($\lambda, \phi > 0$)		$\exp\{\frac{\lambda}{\phi}(1 - e^{\phi t})\}$	$\lambda e^{\phi t} \exp\{\frac{\lambda}{\phi}(1 - e^{\phi t})\}$

Maximum likelihood estimation

- Two kinds of observations: censored and uncensored

Uncensored likelihood

$$p_{\theta}(T = t | \mathbf{x}) = f(t)$$

Censored likelihood

$$p_{\theta}^{\text{censored}}(t | \mathbf{x}) = p_{\theta}(T > t | \mathbf{x}) = S(t)$$

- Putting the two together, we get:

$$\sum_{i=1}^n b_i \log p_{\theta}^{\text{censored}}(t | \mathbf{x}) + (1 - b_i) \log p_{\theta}(t | \mathbf{x})$$

Optimize via gradient or stochastic gradient ascent!

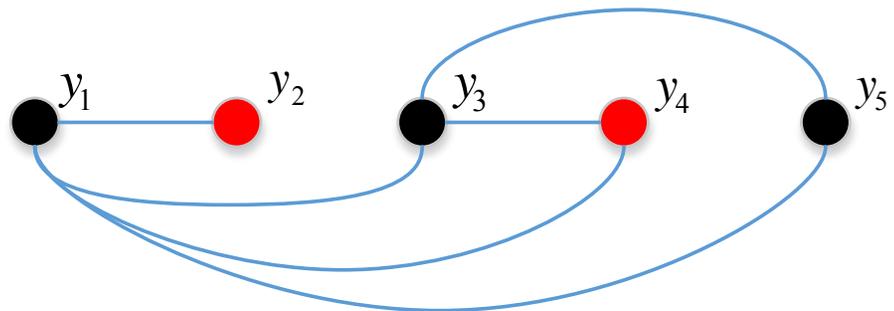
Evaluation for survival modeling

- Concordance-index (also called C-statistic): look at model's ability to predict *relative* survival times:

$$\hat{c} = \frac{1}{num} \sum_{i:b_i=0} \sum_{j:y_i < y_j} I[S(\hat{y}_j|X_j) > S(\hat{y}_i|X_i)]$$

- Illustration – blue lines denote pairwise comparisons:

Black = uncensored
Red = censored



- Equivalent to AUC for binary variables and no censoring

Final thoughts on survival modeling

- Could also evaluate:
 - Mean-squared error for uncensored individuals
 - Held-out (censored) likelihood
 - Derive binary classifier from learned model and check calibration
- Partial likelihood estimators (e.g. for cox-proportional hazards models) can be much more data efficient