Machine Learning for Healthcare

HST.956, 6.S897

Lecture 5: Risk stratification (continued)

David Sontag

CSAIL

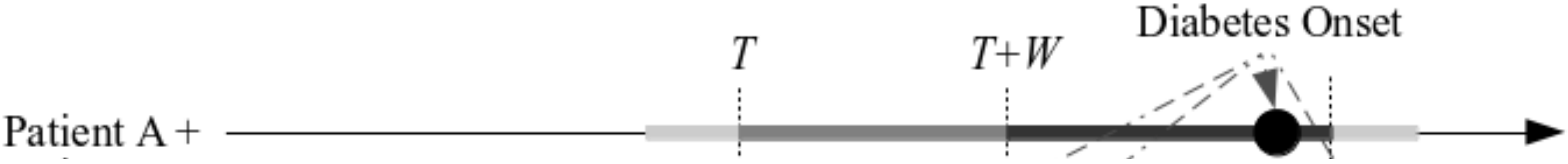
INSTITUTE FOR MEDICAL ENGINEERING & SCIENCE

HEALTH SCIENCES & TECHNOLOGY

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?

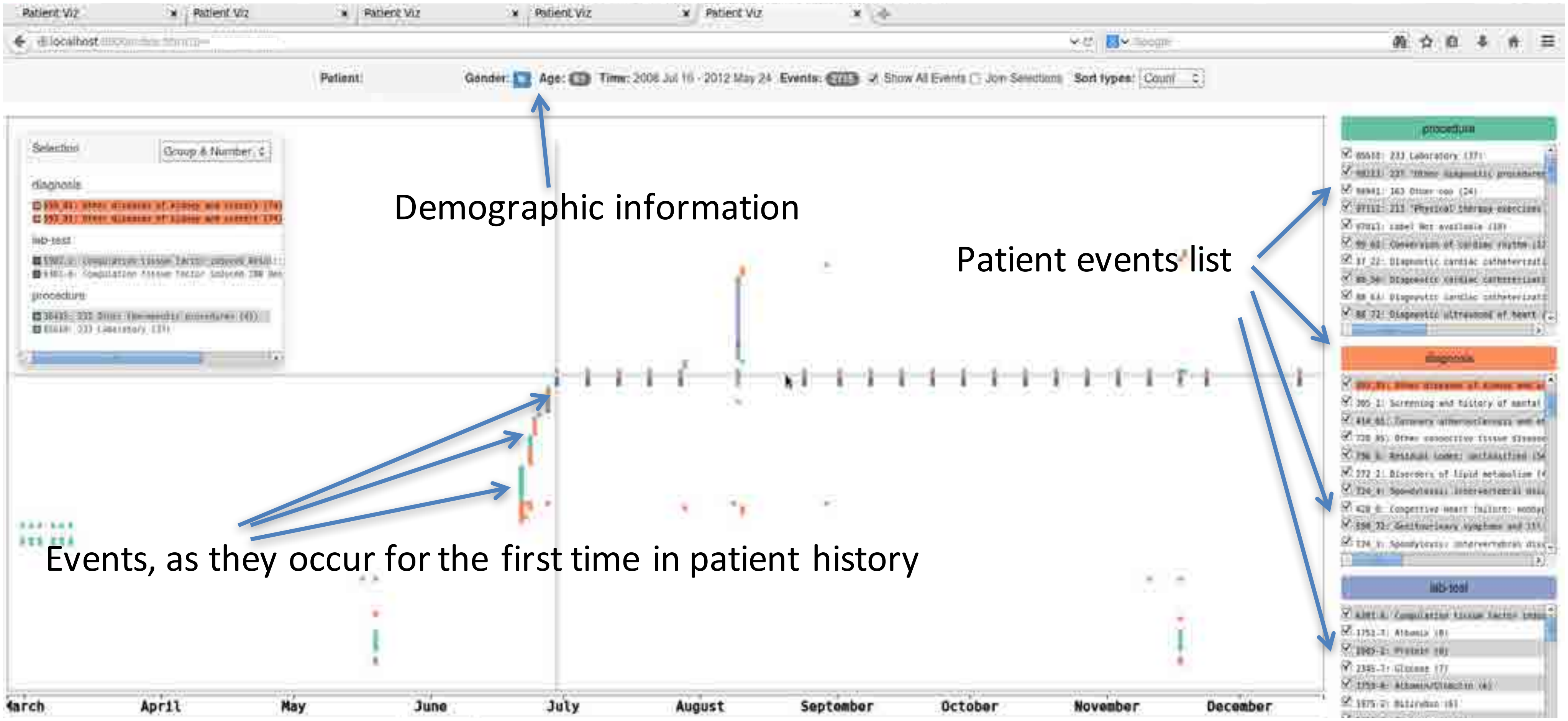


Typical pipeline:

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



- ☑ 85610: 233 Laboratory (37)
- ☑ [illegible]: 227 Other diagnostic procedures
- ☑ [illegible]: 163 Other ooo (24)
- ☑ 97112: 213 Physical therapy exercises
- ☑ [illegible]: Label Not available (18)
- ☑ 99_61: Conversion of cardiac rhythm (1[illegible]
- ☑ 37_22: Diagnostic cardiac catheterizati[illegible]
- ☑ 88_56: Diagnostic cardiac catheterizati[illegible]
- ☑ 88_63: Diagnostic cardiac catheterizati[illegible]
- ☑ 88_72: Diagnostic ultrasound of heart (

- ☑ [illegible]: Other diseases of kidney and ur[illegible]
- ☑ V15_1: Screening and history of mental
- ☑ 414_01: Coronary atherosclerosis and ot[illegible]
- ☑ 710_85: Other connective tissue disease
- ☑ 796_1: Residual codes; unclassified (34
- ☑ 272_2: Disorders of lipid metabolism (4
- ☑ 724_4: Spondylosis; intervertebral disc
- ☑ 428_0: Congestive heart failure; nonhyp[illegible]
- ☑ 598_72: Genitourinary symptoms and ill-[illegible]
- ☑ 724_5: Spondylosis; intervertebral disc

- ☑ 6301-6: Coagulation tissue factor induc[illegible]
- ☑ 1751-7: Albumin (8)
- ☑ 2885-2: Protein (8)
- ☑ 2345-7: Glucose (7)
- ☑ 1759-0: Albumin/Globulin (6)
- ☑ 1975-2: Bilirubin (6)

Demographic information

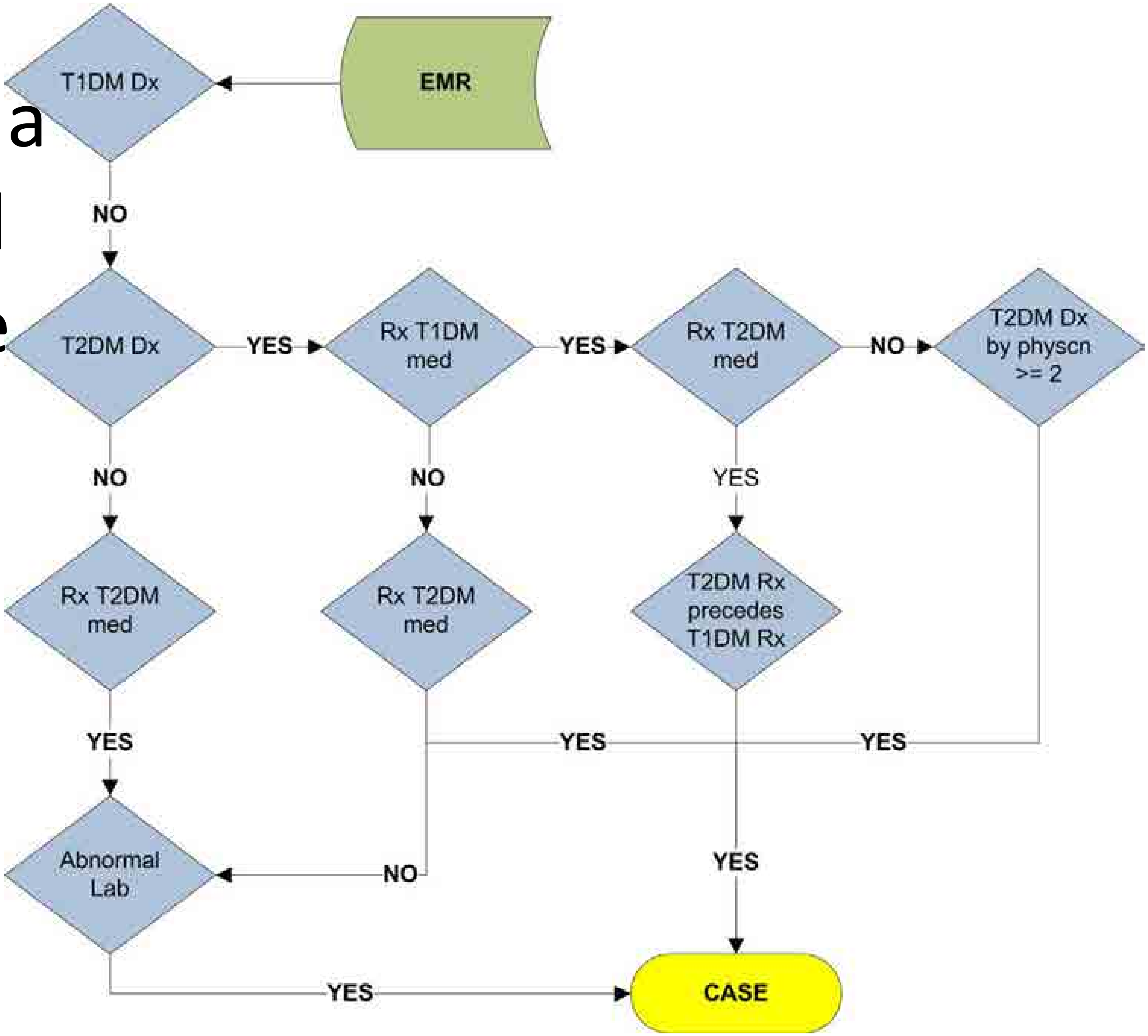
Patient events list

Events, as they occur for the first time in patient history

https://github.com/nyuvis/patient-viz

Step 2: Example of a rule-based phenotype

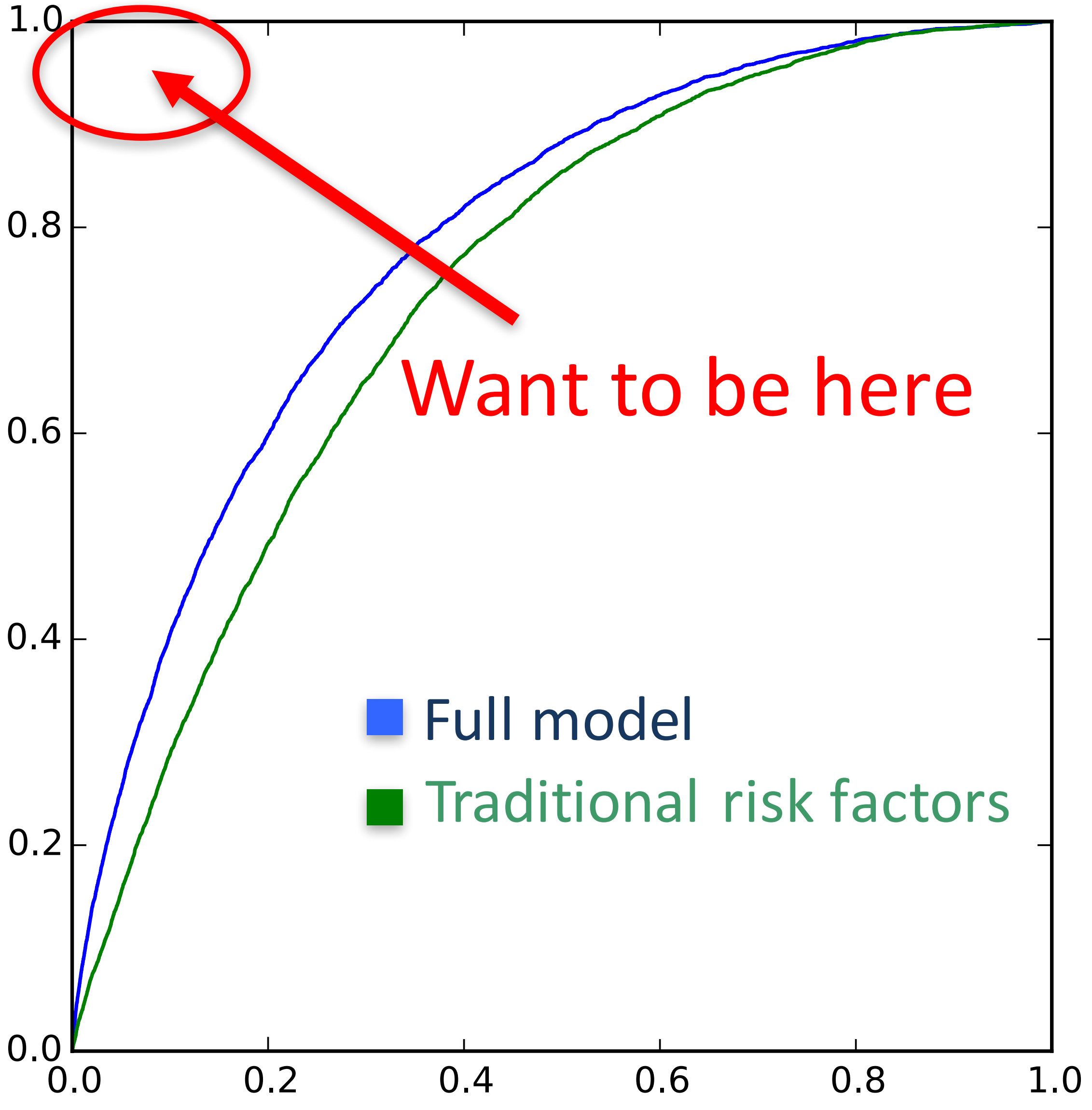
Figure 1: Algorithm for identifying T2DM cases in the EMR.



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

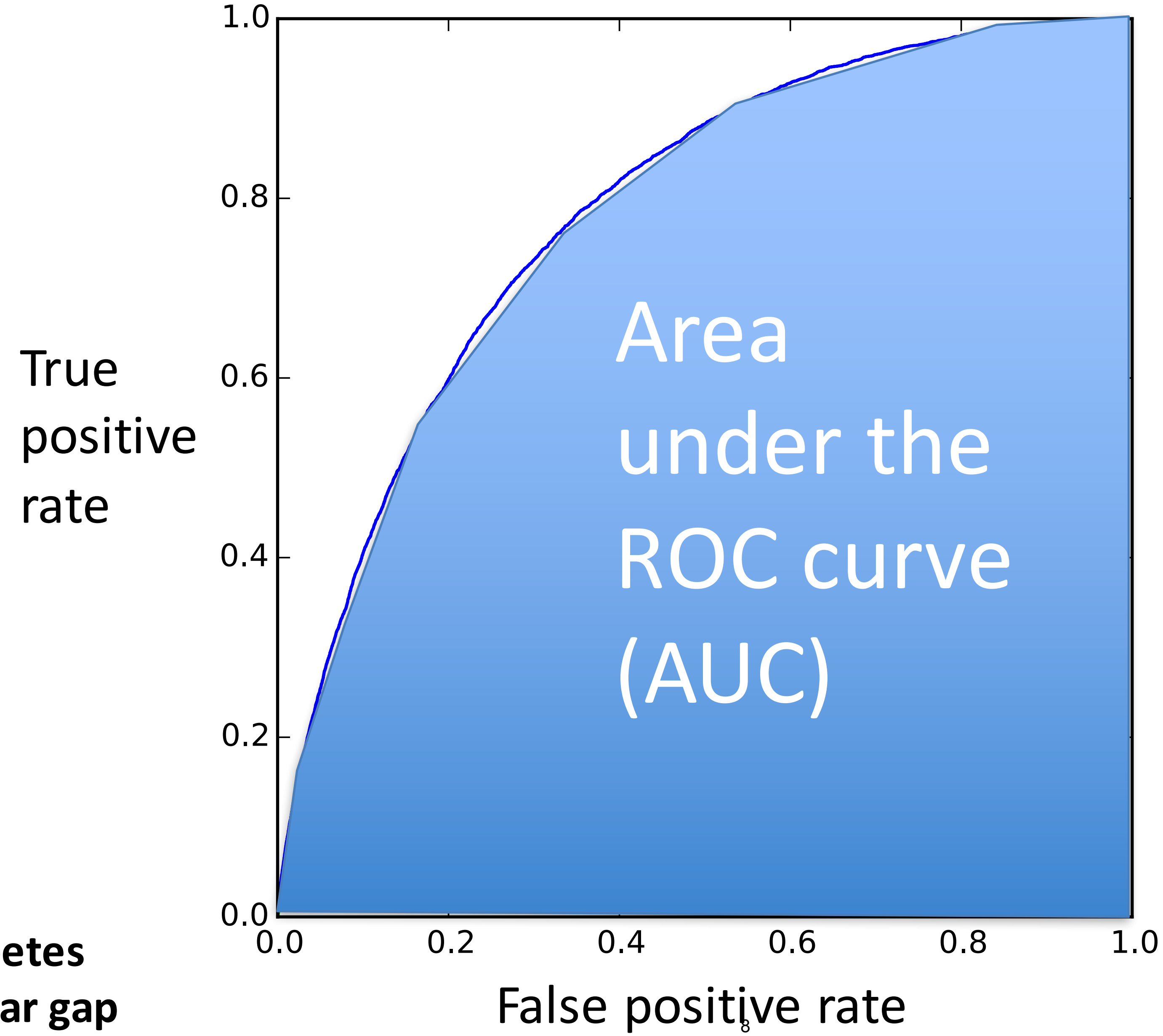


Want to be here

Obtained by varying prediction threshold

Diabetes 1-year gap

Receiver-operator characteristic curve

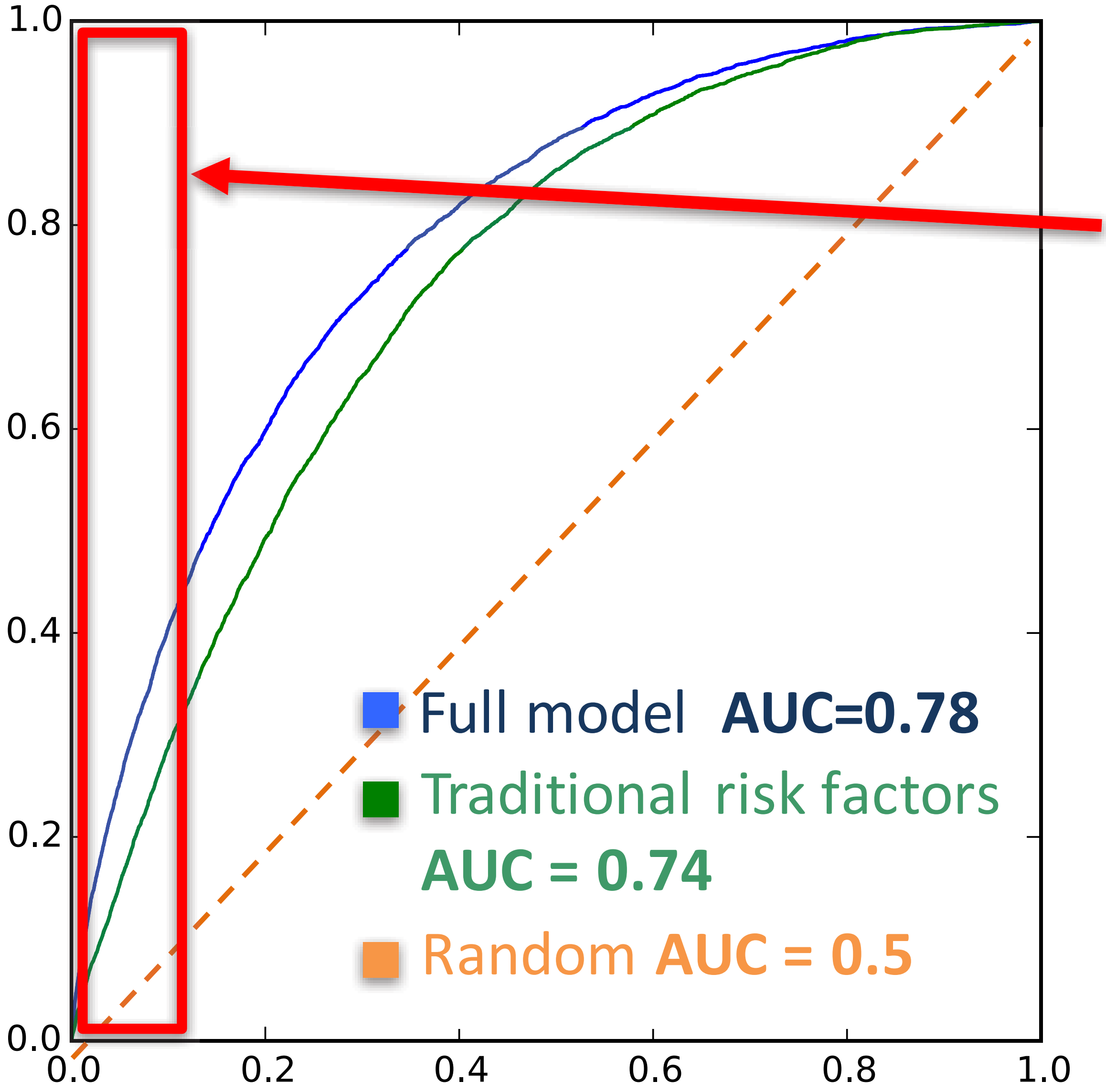


Diabetes 1-year gap

AUC = Probability that algorithm ranks a positive patient over a negative patient

Invariant to amount of class imbalance

Receiver-operator characteristic curve



True positive rate

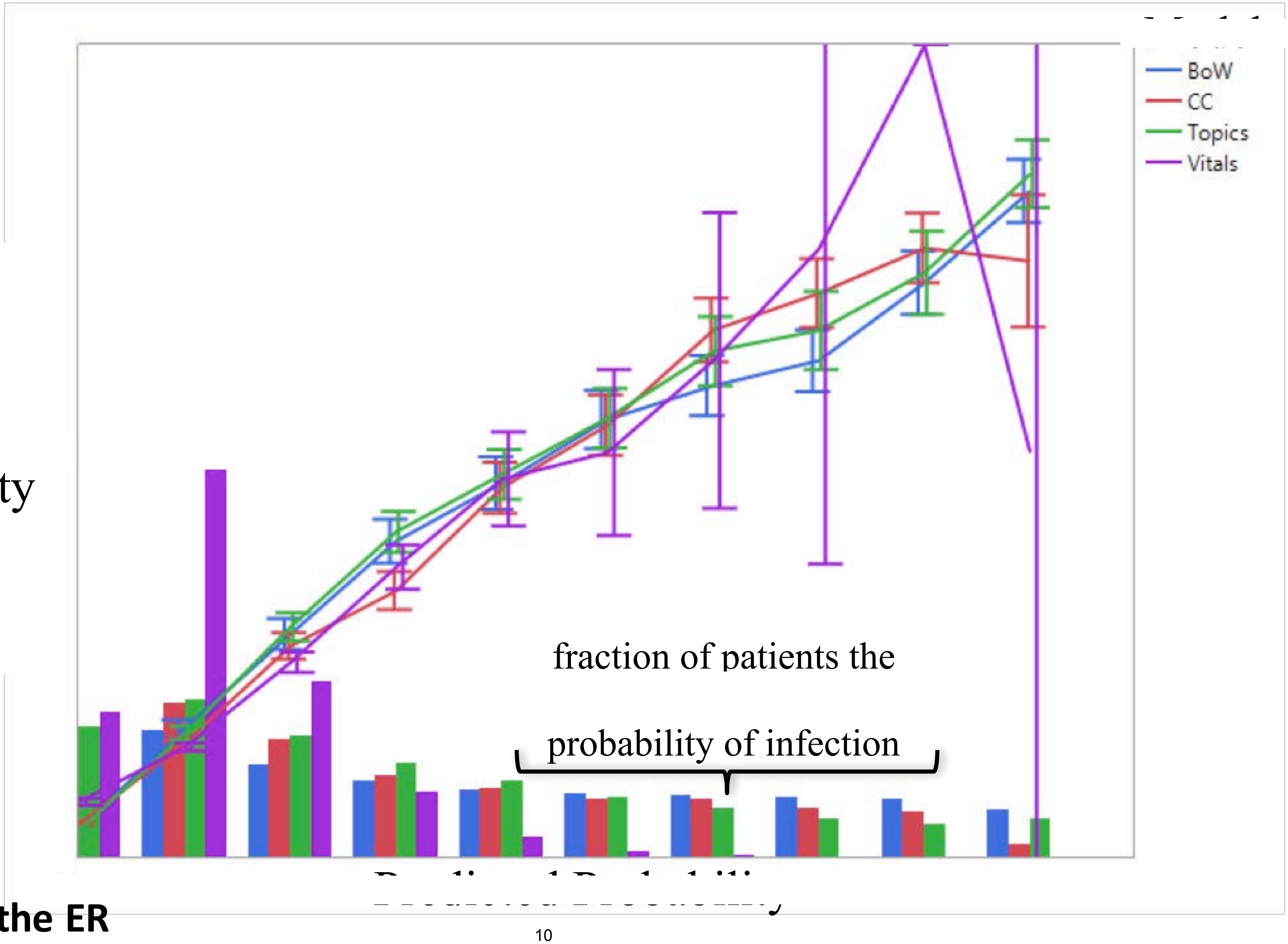
False positive rate

Risk stratification usually focuses on just this region

(because of the cost of interventions)

Diabetes 1-year gap

Calibration (*note: different dataset*)



Predicting infection in the ER

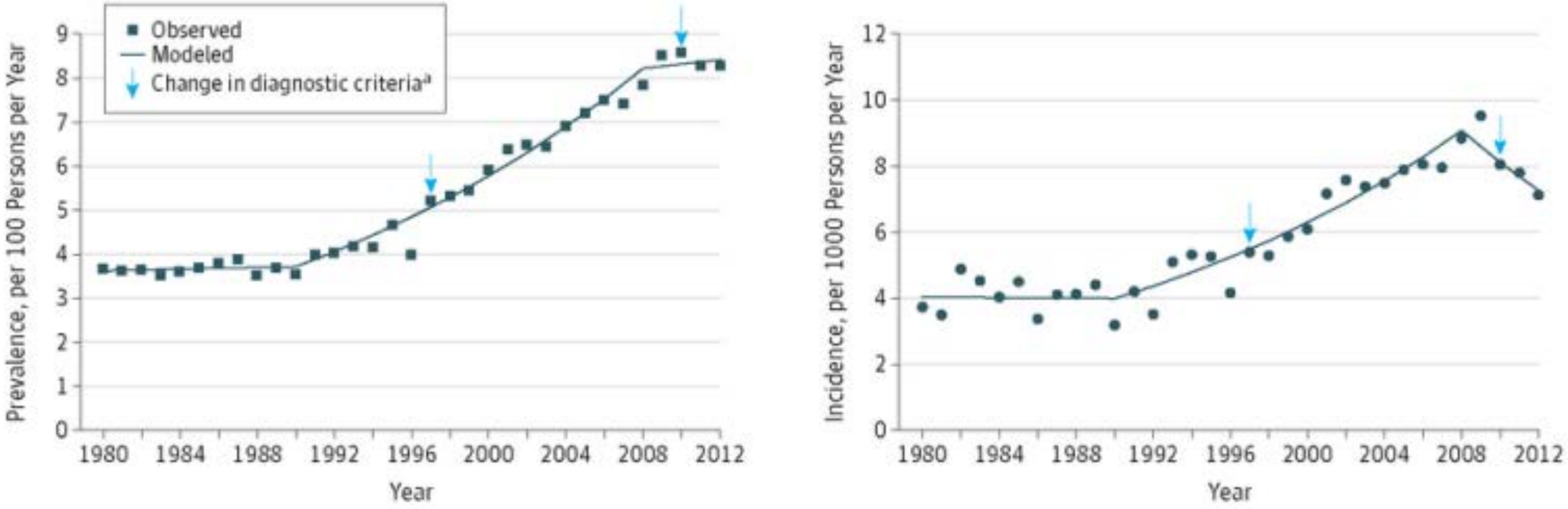
Courtesy of Horng et al. Used under CC BY.

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



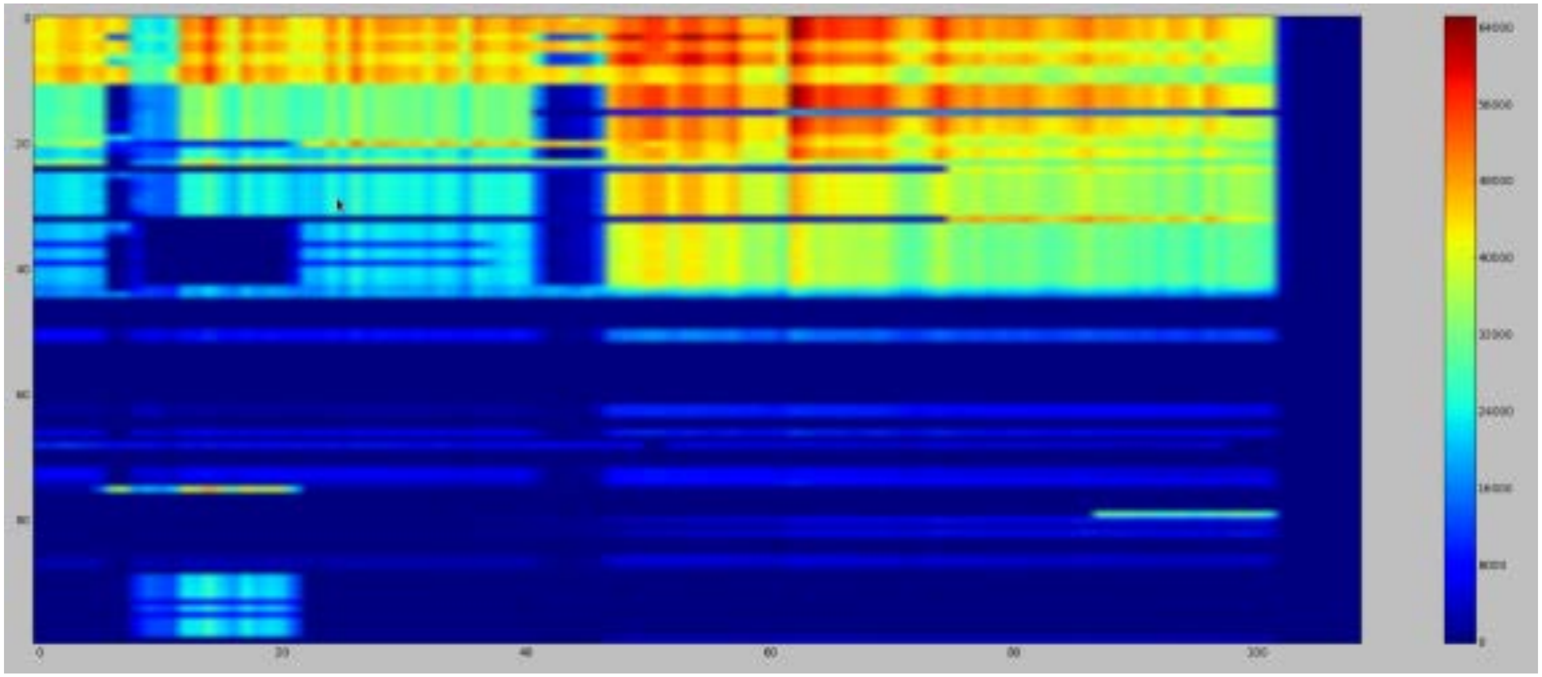
→ Automatically derived labels may change meaning

[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

Labs



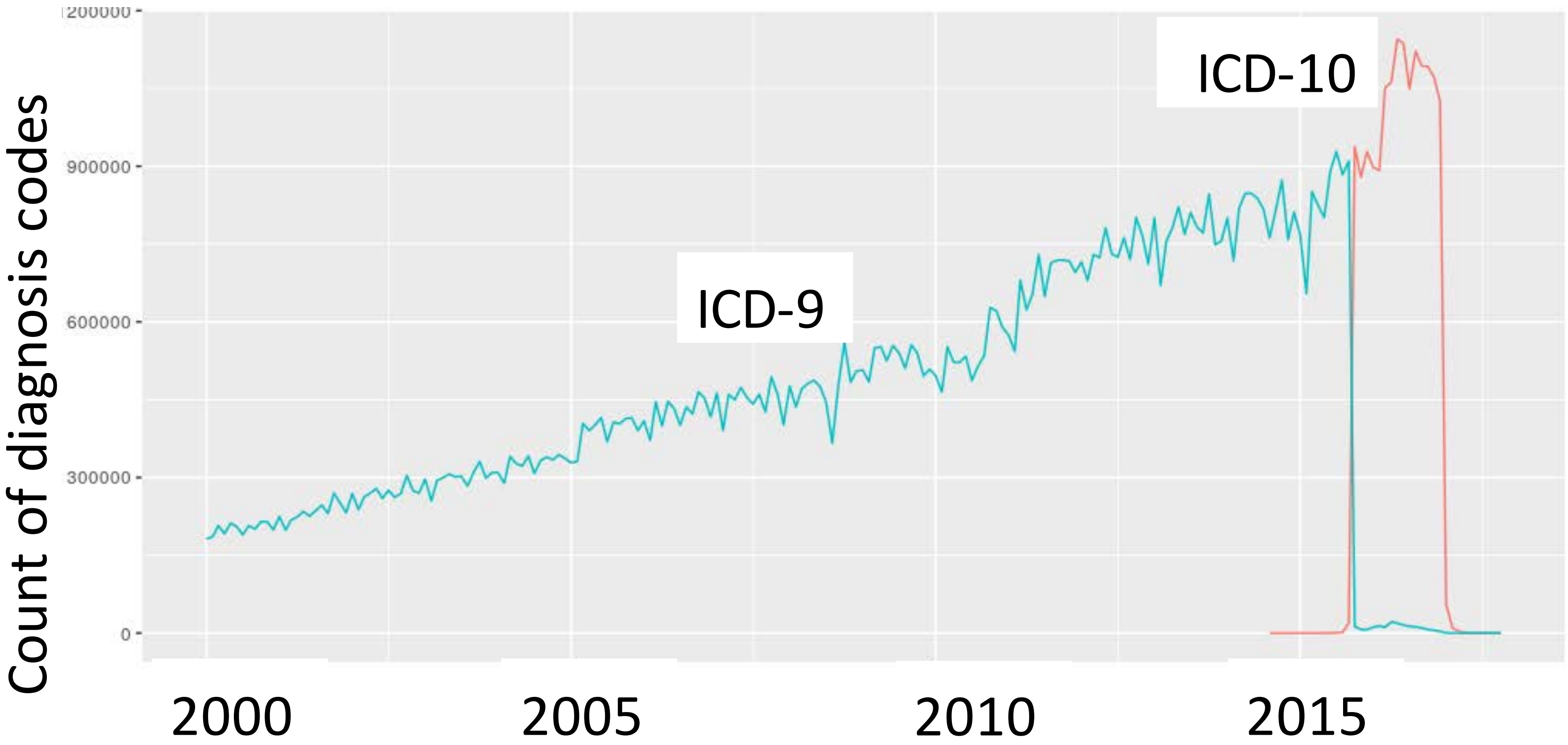
Time (in months, from 1/2005 up to 1/2014)

→ Significance of features may change over time

[Figure credit: Narges Razavian]

Non-stationarity:

ICD-9 to ICD-10 shift

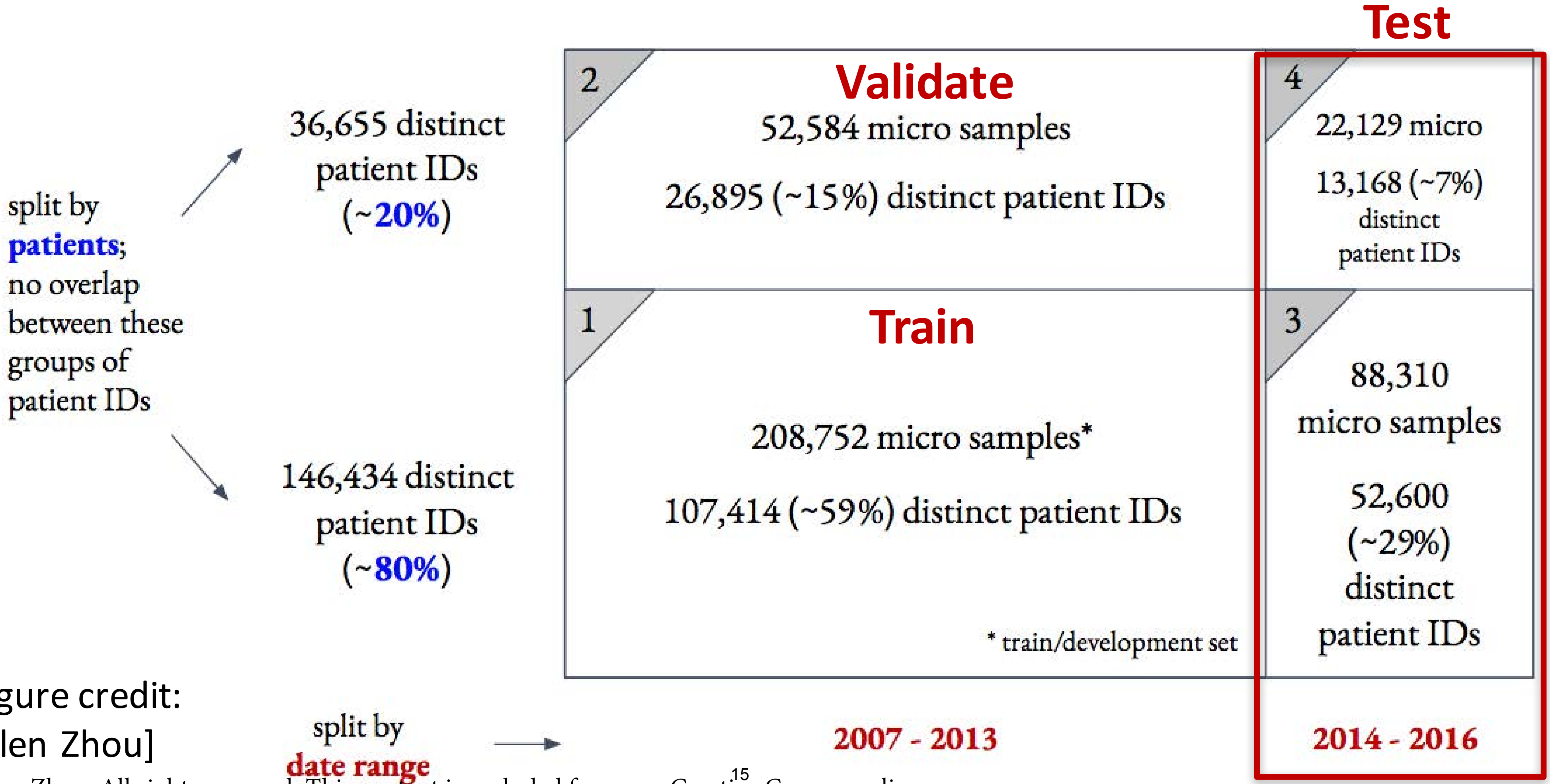


→ Significance of features may change over time

[Figure credit: Mike Oberst]

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:



Test

split by **patients**; no overlap between these groups of patient IDs

36,655 distinct patient IDs (~**20%**)

146,434 distinct patient IDs (~**80%**)

2 **Validate**
52,584 micro samples
26,895 (~15%) distinct patient IDs

4
22,129 micro
13,168 (~7%) distinct patient IDs

1 **Train**
208,752 micro samples*
107,414 (~59%) distinct patient IDs
* train/development set

3
88,310 micro samples
52,600 (~29%) distinct patient IDs

split by **date range**

2007 - 2013

2014 - 2016

[Figure credit: Helen Zhou]

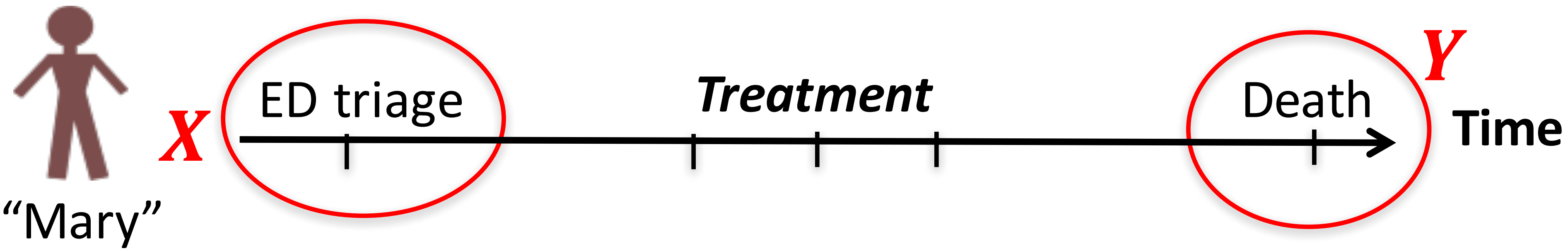
Intervention-tainted outcomes

- 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!

[Caruana et al., Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission. KDD 2015.]

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

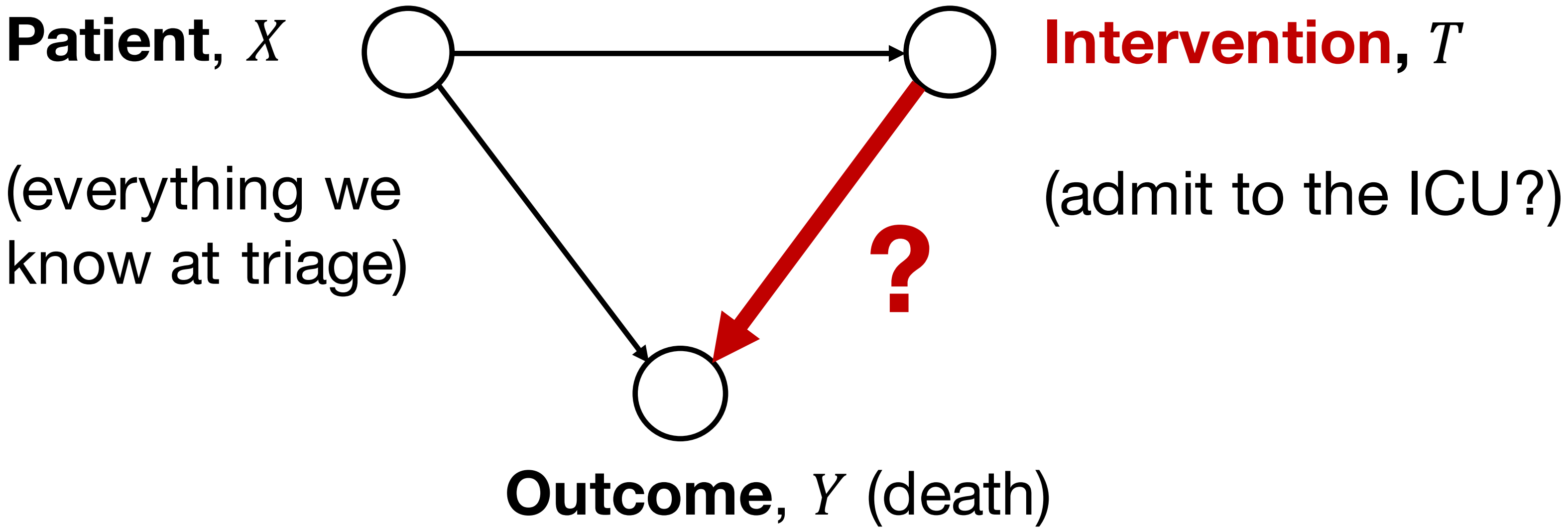
Intervention-tainted outcomes

- 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:**



Patient, X

(everything we know at triage)

Intervention, T

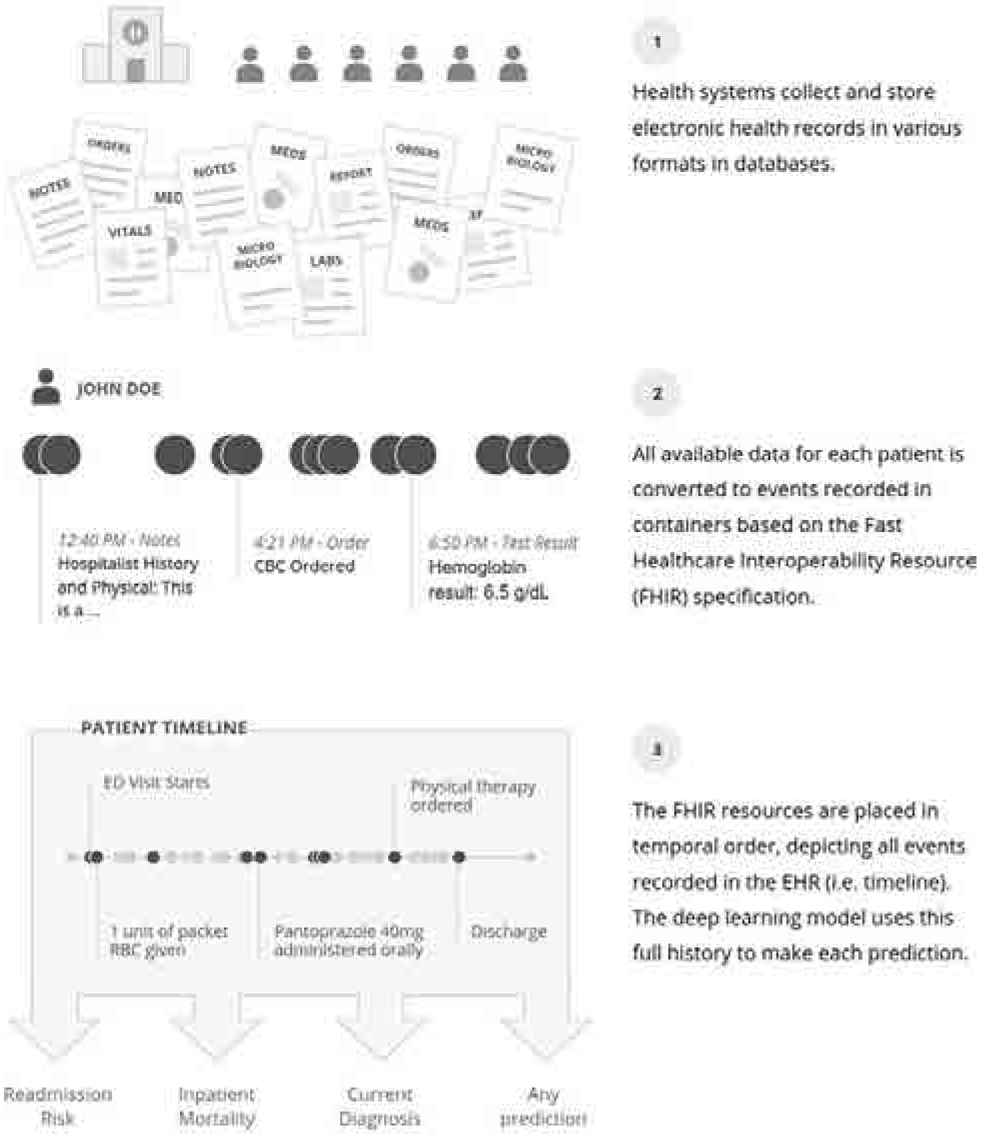
(admit to the ICU?)

Outcome, Y (death)

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

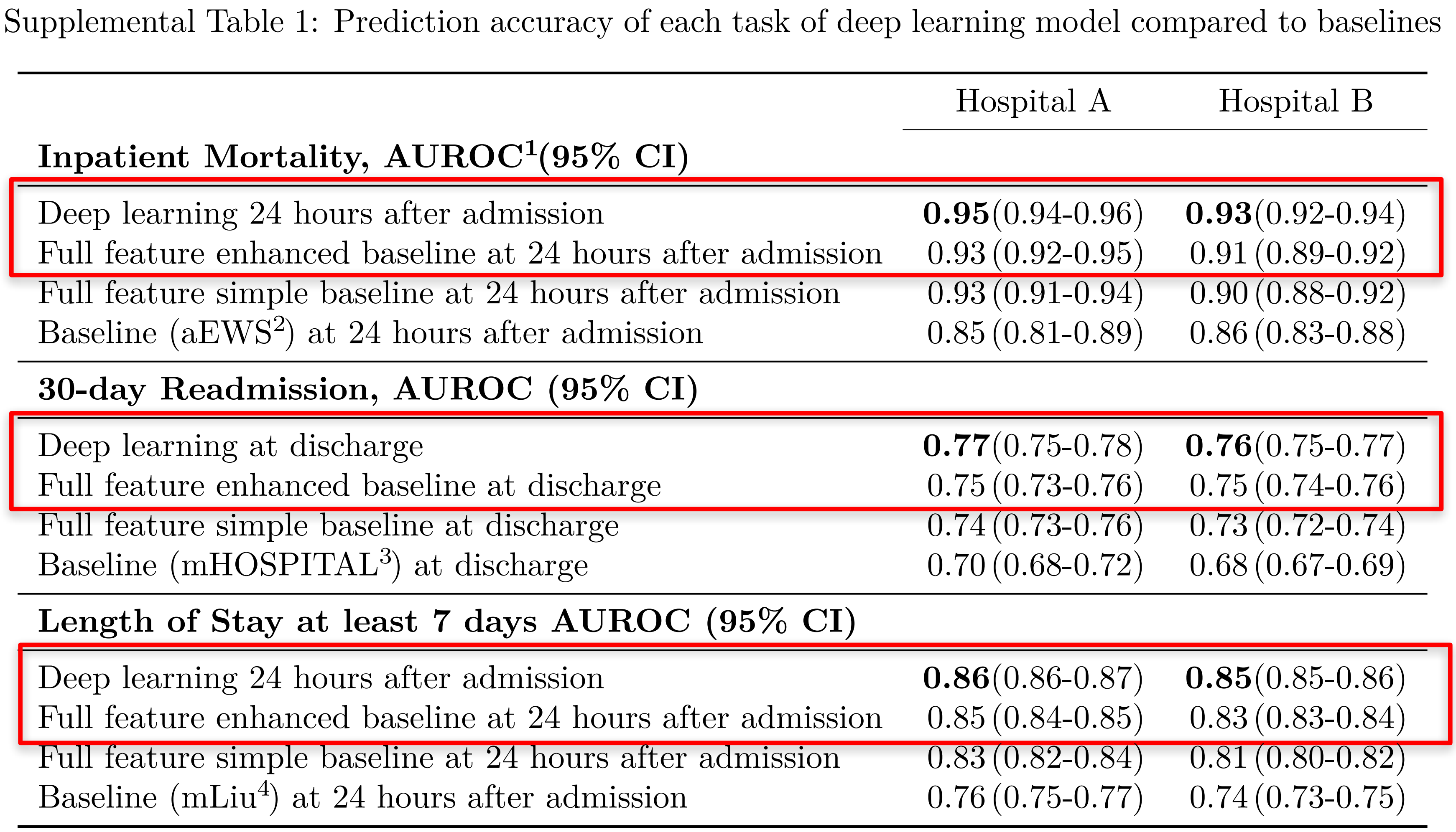
The FHIR resources are placed in temporal order, depicting all events recorded in the EHR (i.e. timeline). The deep learning model uses this full history to make each prediction.

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

Courtesy of Rajkomar et al. Used under CC BY.

No big wins from deep models on structured data/text



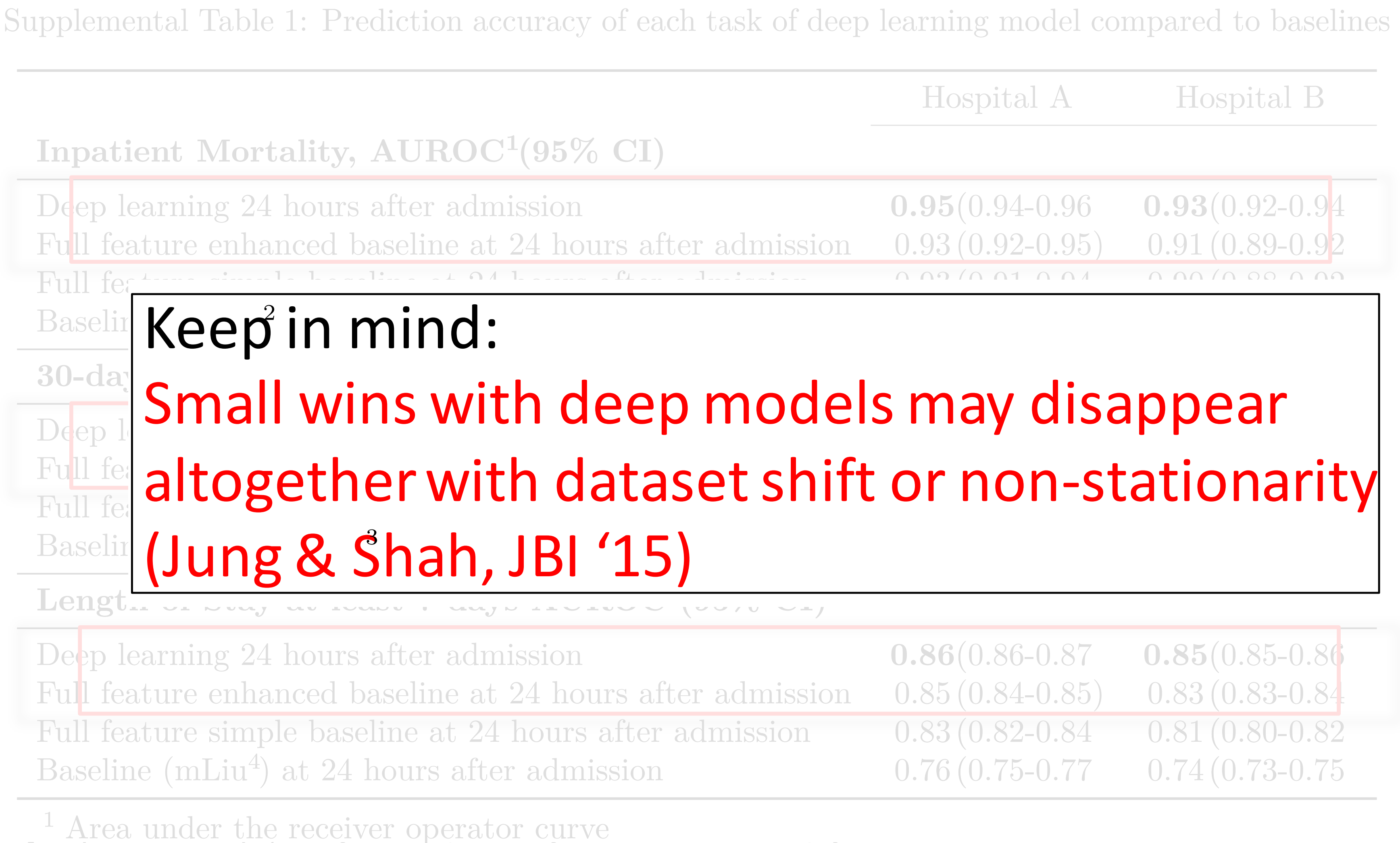
Supplemental Table 1: Prediction accuracy of each task of deep learning model compared to baselines

	Hospital A	Hospital B
Inpatient Mortality, AUROC[1](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

Courtesy of Rajkomar et al. Used under CC BY.

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, AUROC[1](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 fea	0.93 (0.91-0.94	0.90 (0.88-0.92
Baselir		
30-day		
Deep l		
Full fe		
Full fe		
Baselir		
Lengt		
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

[1] Area under the receiver operator curve

Comparison to Razavian ‘15

Keep in mind:

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

Courtesy of Rajkomar et al. Used under CC BY.

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

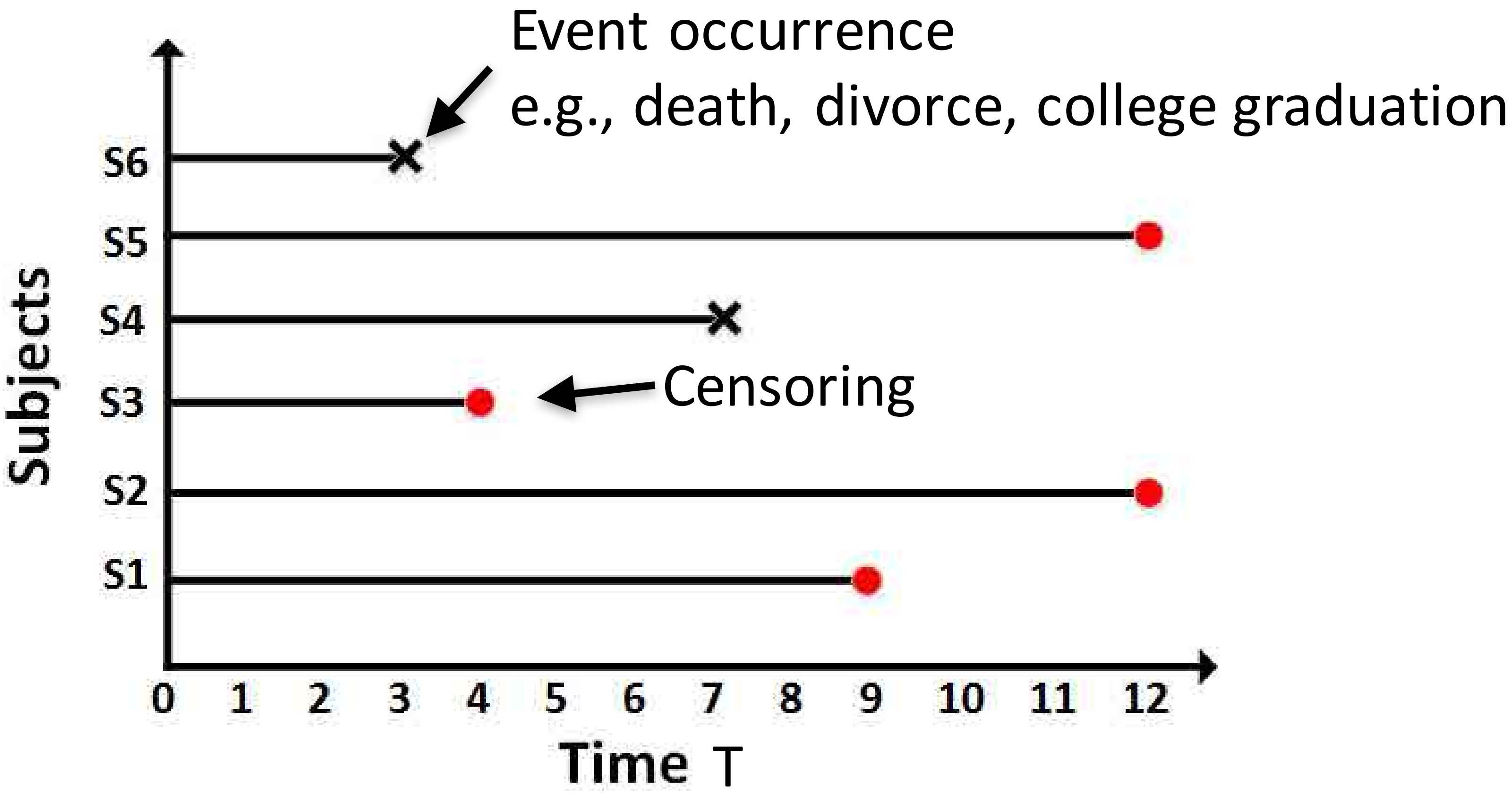
- Sequential data in medicine is very different from language modeling
  - Many time scales, significant missing data, and multi-variate observations
  - Likely *do exist* predictive nonlinear interactions, but subtle
  - Not enough data to naively deal with the above two
- Medical community has already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



Event occurrence
e.g., death, divorce, college graduation

Censoring

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

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 (**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*,

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

[Ha, Jeong, Lee. Statistical Modeling of Survival Data with Random Effects. Springer 2017]

Notation and formalization

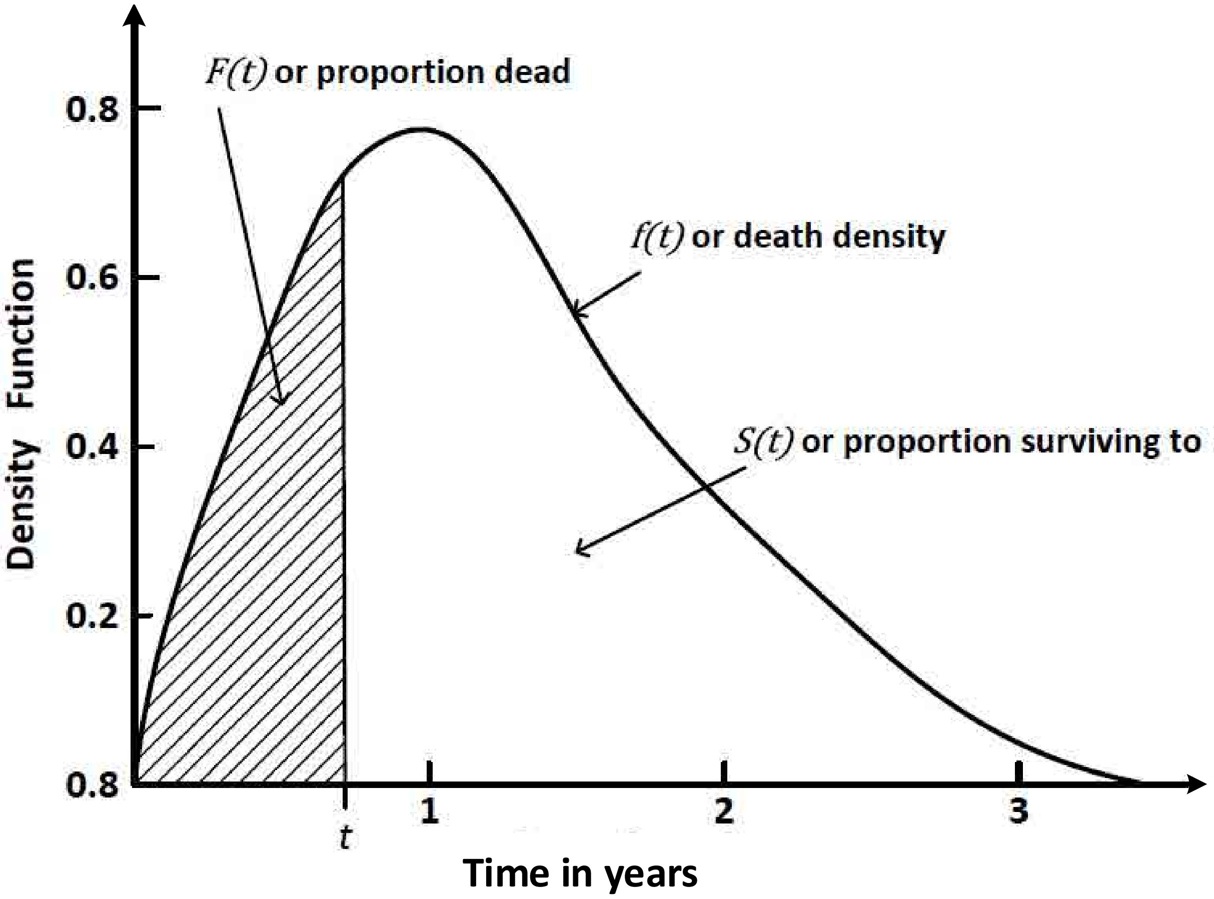
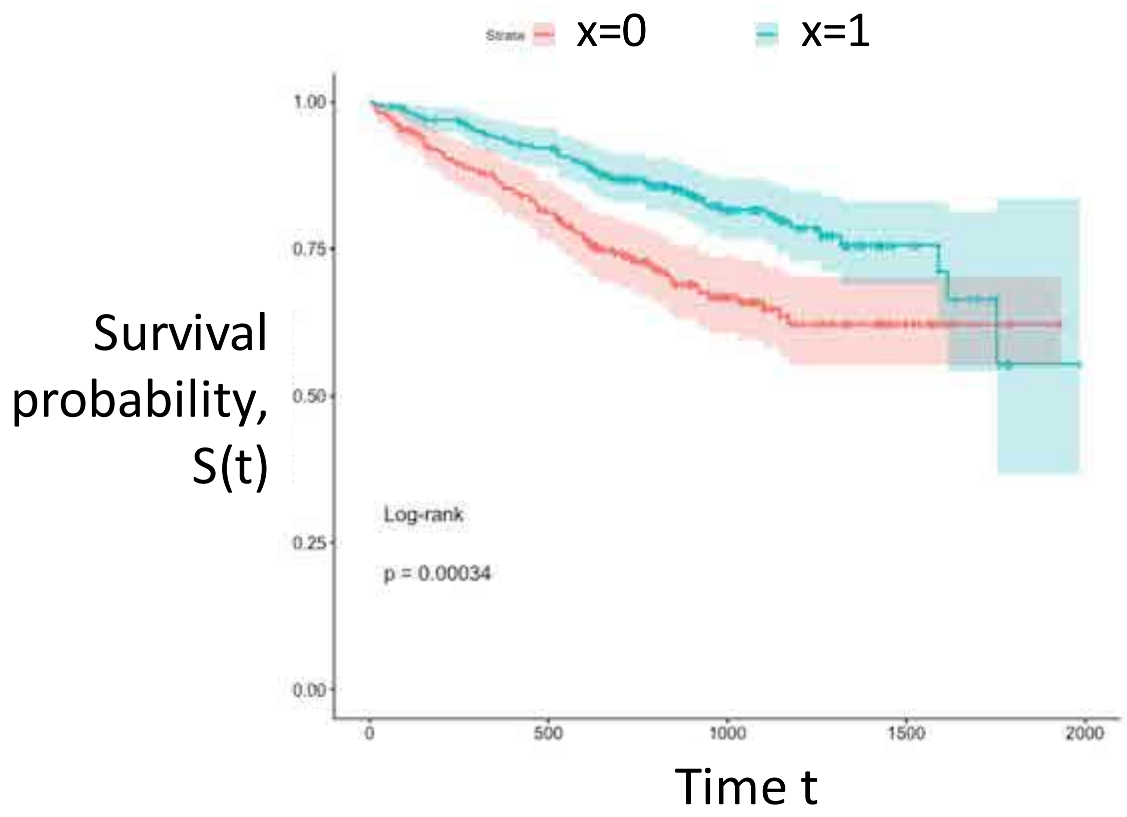


Fig. 2: Relationship among different entities $f(t)$, $F(t)$ and $S(t)$.

[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

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

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

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

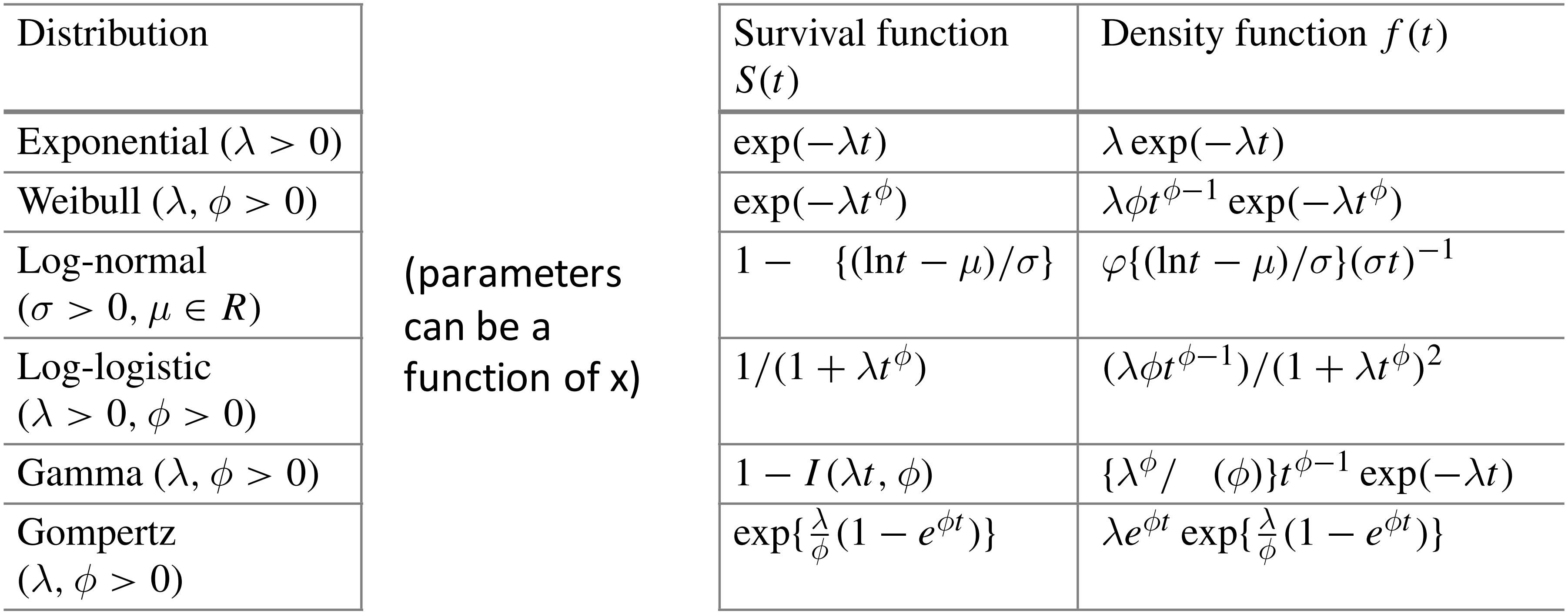
$$S_{K-M}(t) = \underset{k:y_{(k)} \leq t}{} 1 - \frac{d_{(k)}}{n_{(k)}}$$

[Figure credit: Rebecca Peyser]

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 $(\sigma > 0, \mu \in R)$	$1 - \quad \{(\ln t - \mu)/\sigma\}$	$\varphi\{(\ln t - \mu)/\sigma\}(\sigma t)^{-1}$
Log-logistic $(\lambda > 0, \phi > 0)$	$1/(1 + \lambda t^{\phi})$	$(\lambda \phi t^{\phi-1})/(1 + \lambda t^{\phi})^2$
Gamma $(\lambda, \phi > 0)$	$1 - I(\lambda t, \phi)$	$\{\lambda^{\phi}/ \quad (\phi)\} t^{\phi-1} \exp(-\lambda t)$
Gompertz $(\lambda, \phi > 0)$	$\exp\{\frac{\lambda}{\phi}(1 - e^{\phi t})\}$	$\lambda e^{\phi t} \exp\{\frac{\lambda}{\phi}(1 - e^{\phi t})\}$

(parameters can be a function of x)

[Ha, Jeong, Lee. Statistical Modeling of Survival Data with Random Effects. Springer 2017]

Maximum likelihood estimation

- Two kinds of observations: censored and uncensored

Uncensored likelihood

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

Censored likelihood

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

- Putting the two together, we get:

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

Optimize via gradient or stochastic gradient ascent!

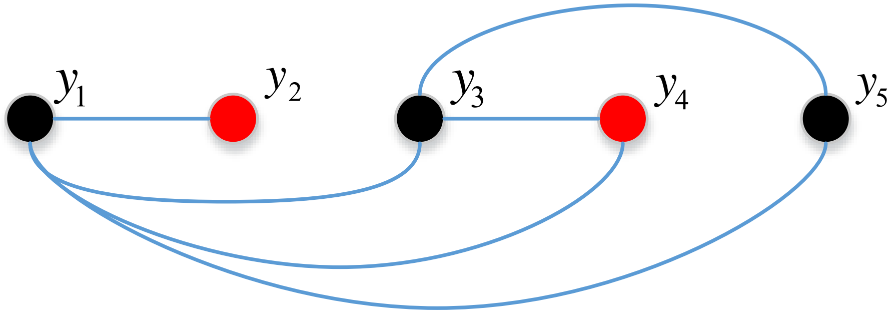
Evaluation for survival modeling

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

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

- Illustration – blue lines denote pairwise comparisons:

Black = uncensored
Red = censored



- Equivalent to AUC for binary variables and no censoring

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

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