Machine Learning for Healthcare
6.871Jx

Risk Stratification Part II

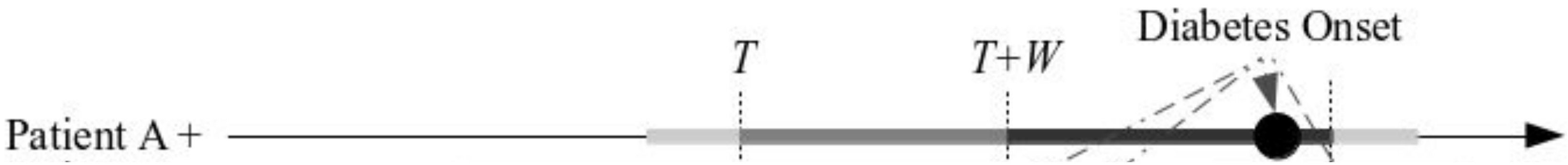
David Sontag

CSAIL

INSTITUTE FOR MEDICAL ENGINEERING & SCIENCE

HEALTH SCIENCES & TECHNOLOGY

Where do the labels come from?

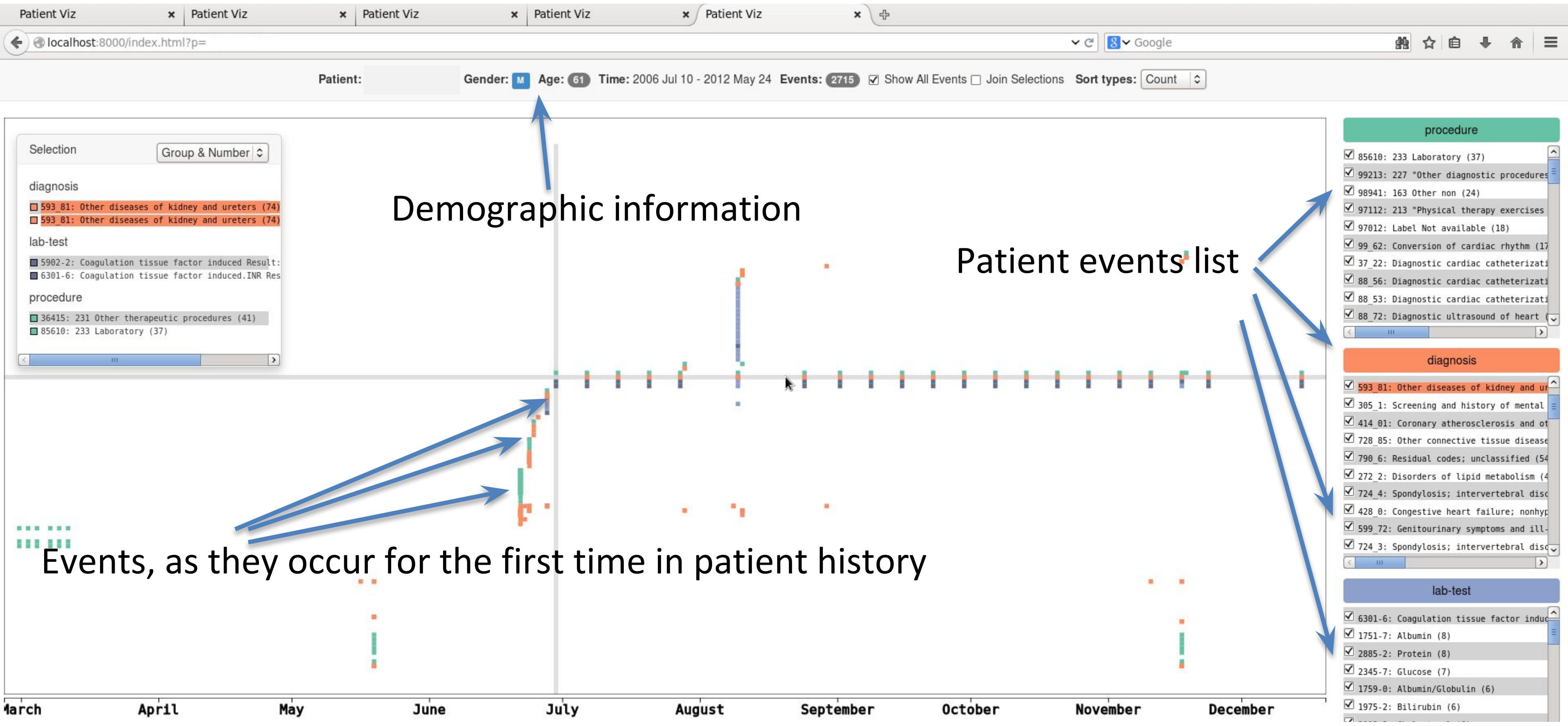


Typical pipeline:

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

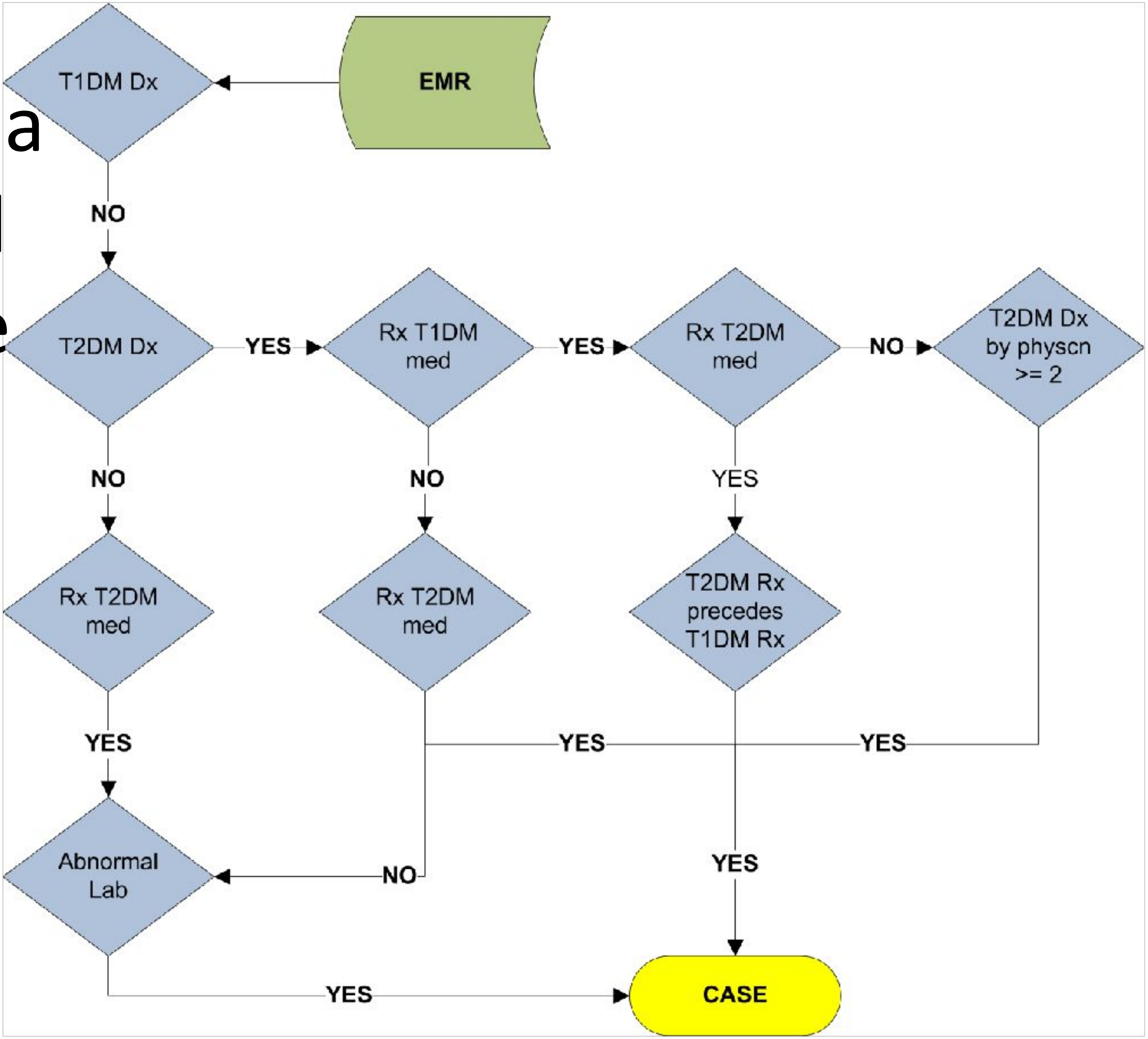
Step 1: Visualization of individual patient data is an important part of chart review



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

Step 2: Example of a rule-based phenotype

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



Source: https://phekb.org/sites/phenotype/files/T2DM-algorithm.pdf

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

Search

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 Disease or Syndrome

Owner Phenotyping Groups

View Phenotyping Groups

Data Model - Any -

Apply

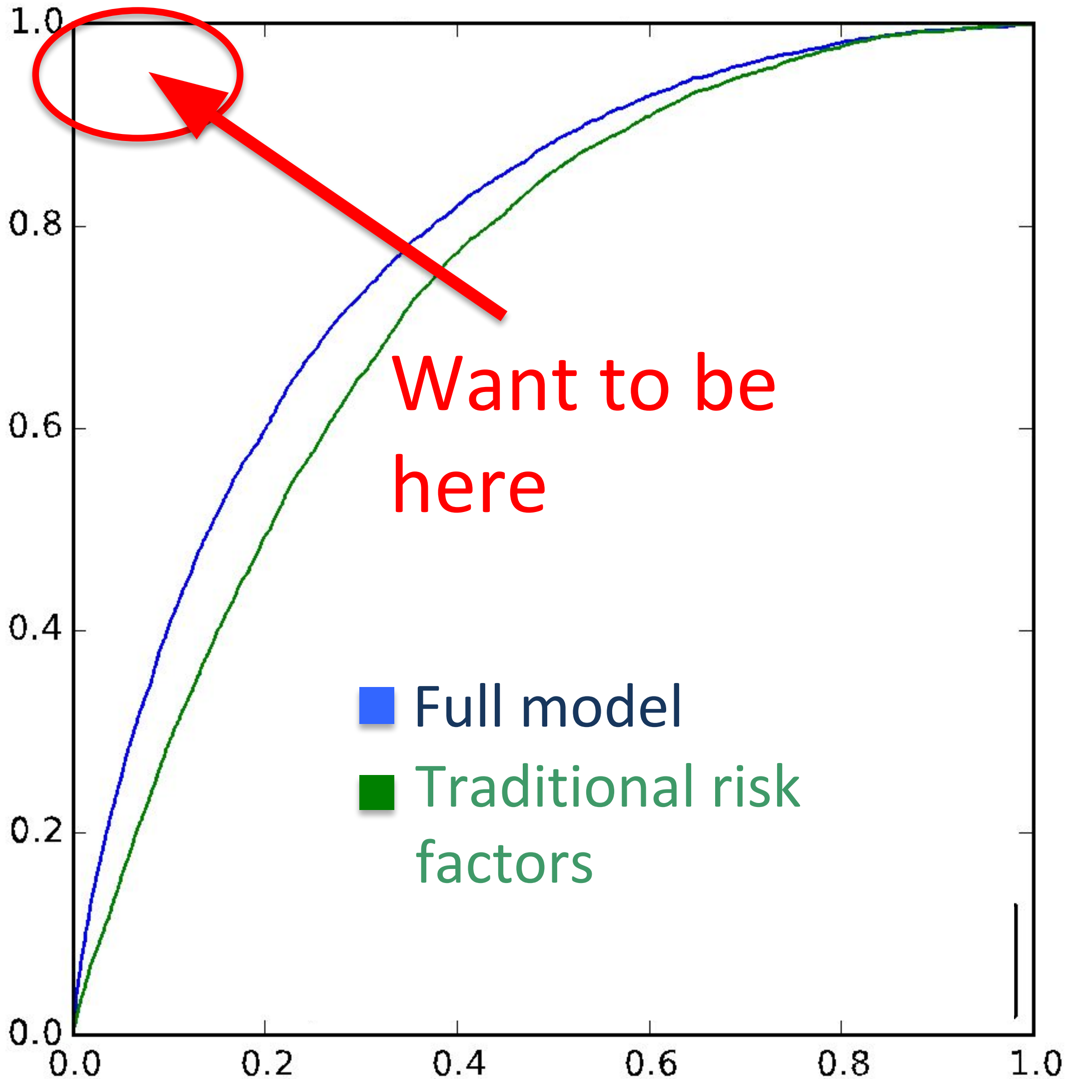
Title	Institution	Data Modalities and Methods Used	Owner Phenotyping Groups	View Groups	Has new content	Status	Type
Abdominal Aortic Aneurysm (AAA)	Geisinger	CPT Codes, ICD 9 Codes, Vital Signs	eMERGE Geisinger Group	eMERGE Geisinger Group, eMERGE Phenotype WG		Final	Disease or Syndrome
ADHD phenotype 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 -		ICD 9 Codes, Medications,	Vanderbilt -	Vanderbilt -			Disease

Outline for today's class

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

Receiver-operator characteristic curve

True positive rate



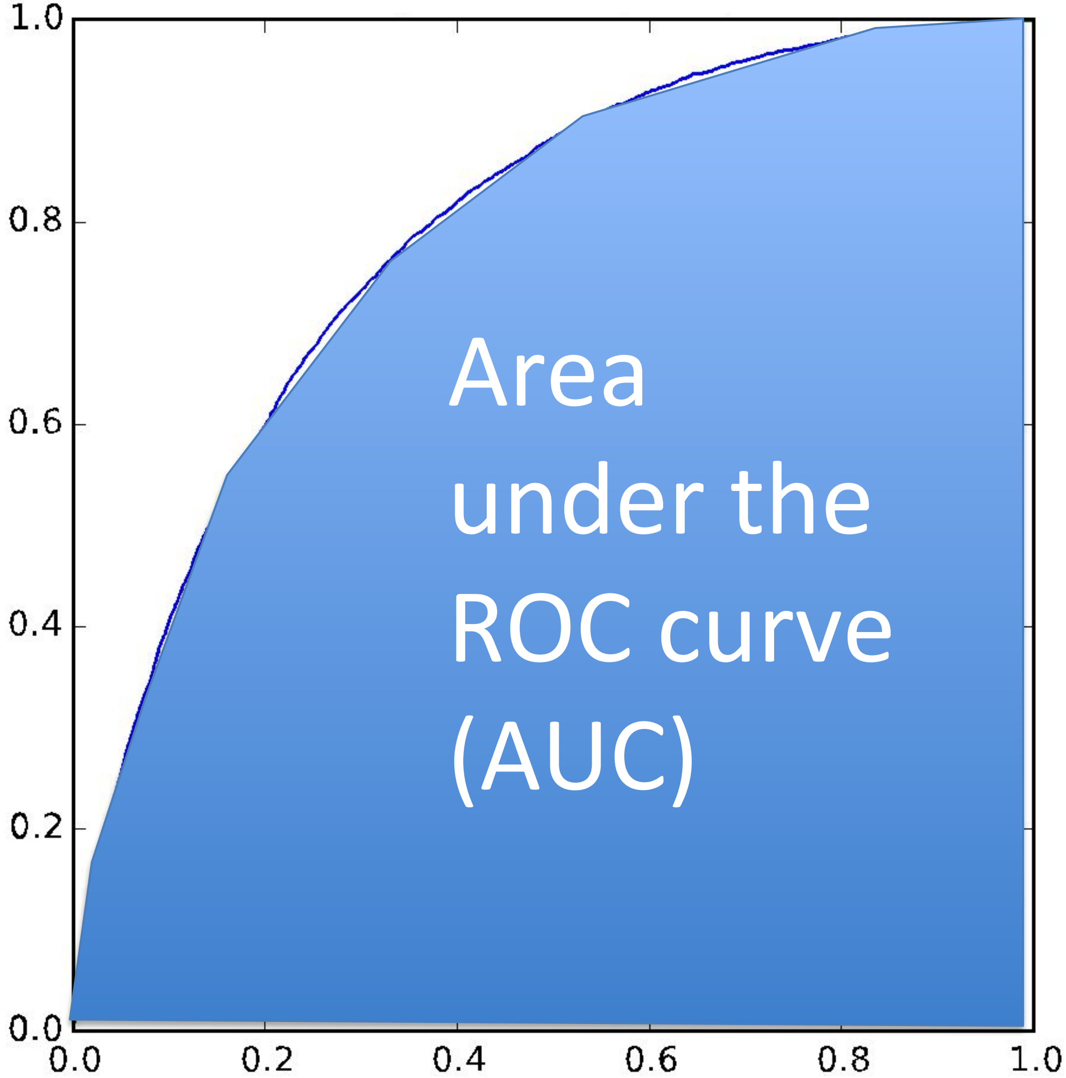
Obtained by varying prediction threshold

False positive rate

Diabetes 1-year gap

Receiver-operator characteristic curve

True positive rate



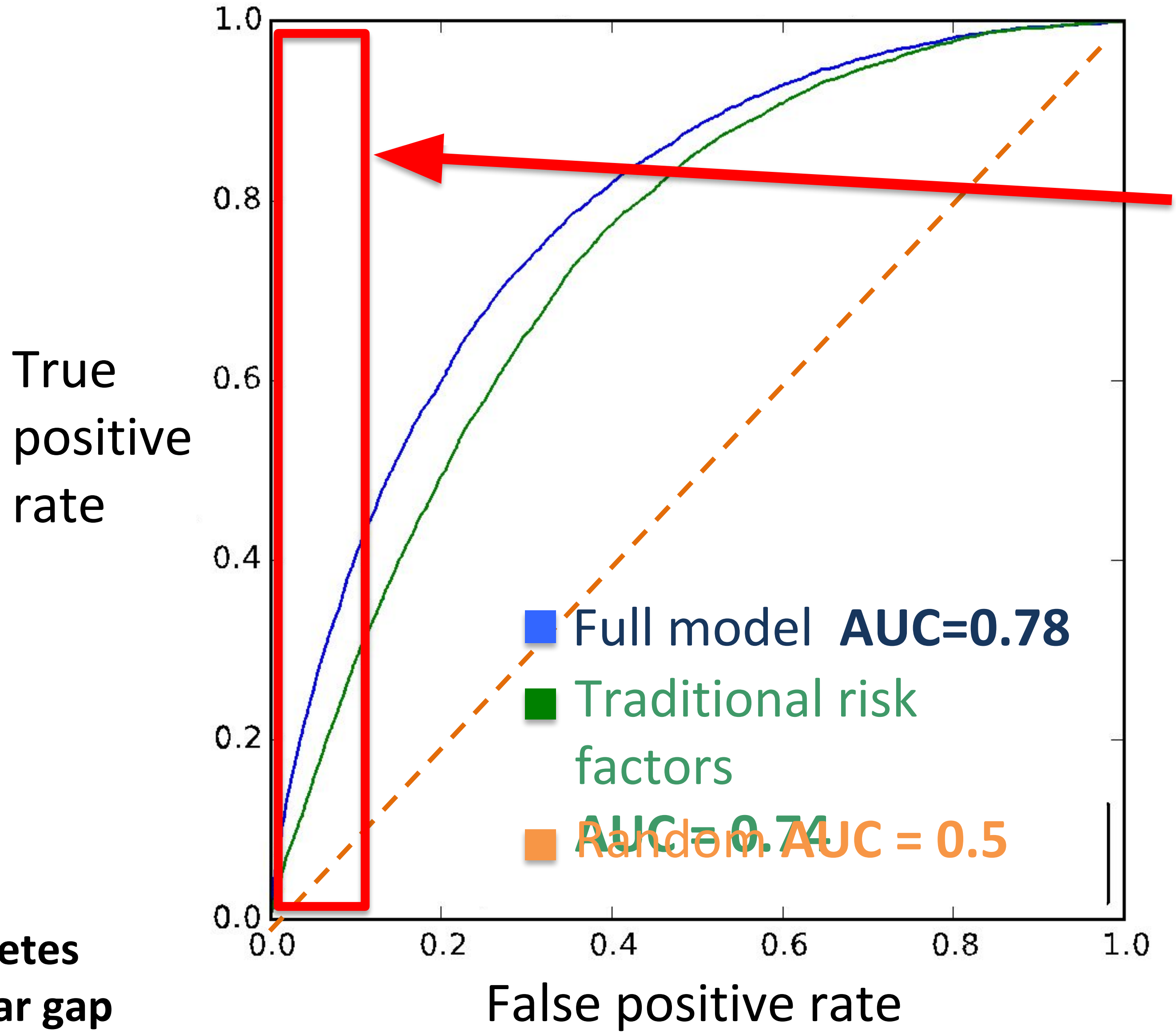
False positive rate

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

Invariant to amount of class imbalance

Diabetes 1-year gap

Receiver-operator characteristic curve

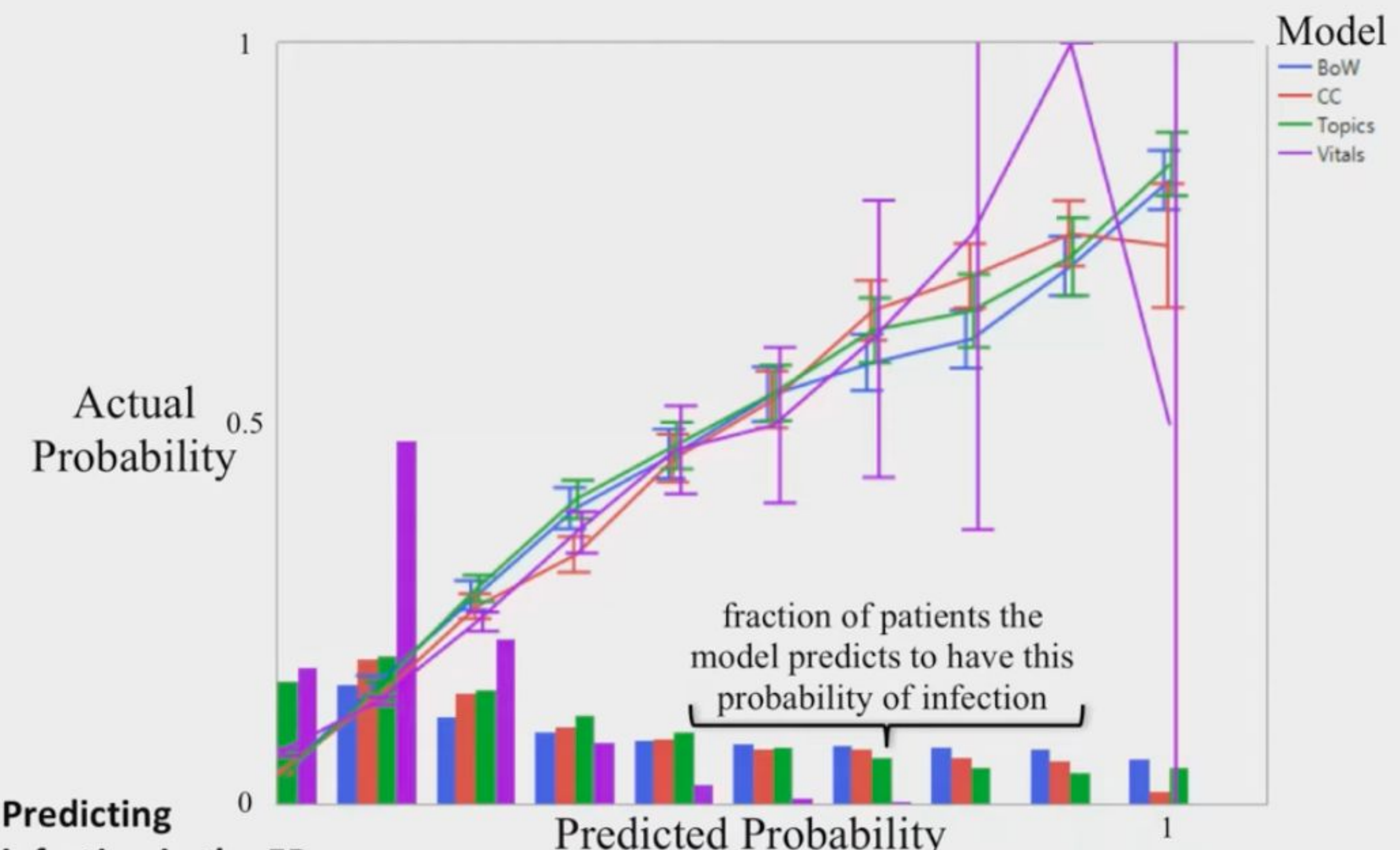


Risk stratification usually focuses on just this region

(because of the cost of interventions)

Diabetes
1-year gap

Calibration (*note: different dataset*)



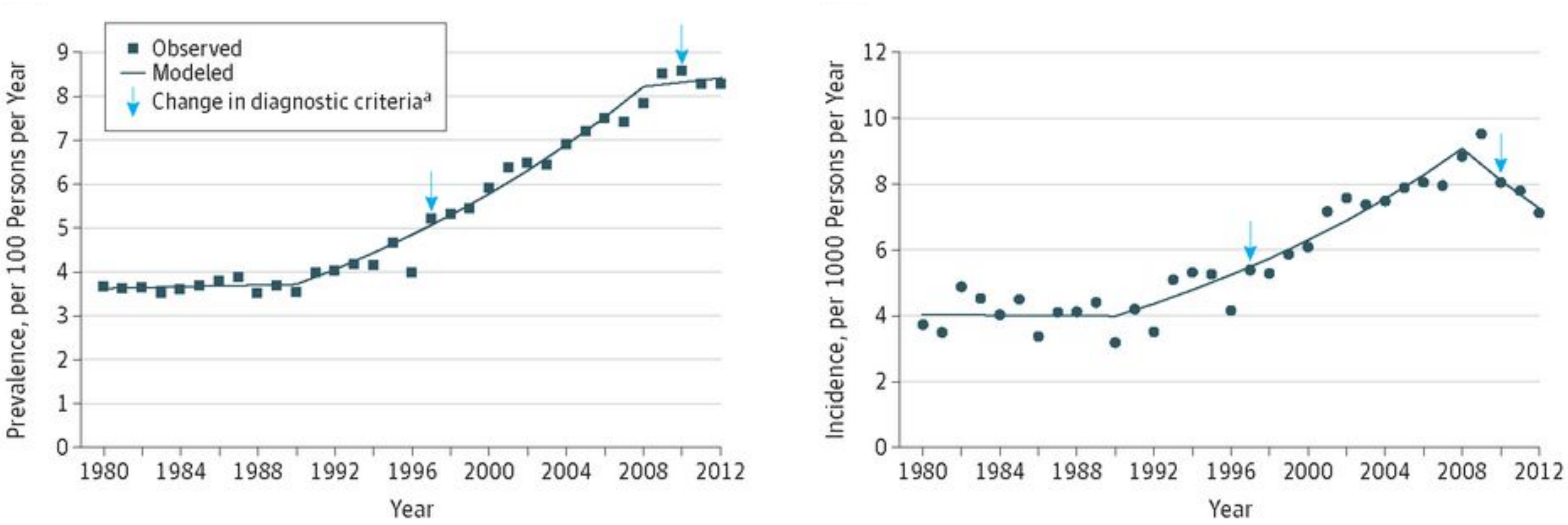
Predicting infection in the ER

Outline for today's class

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

Non-stationarity:

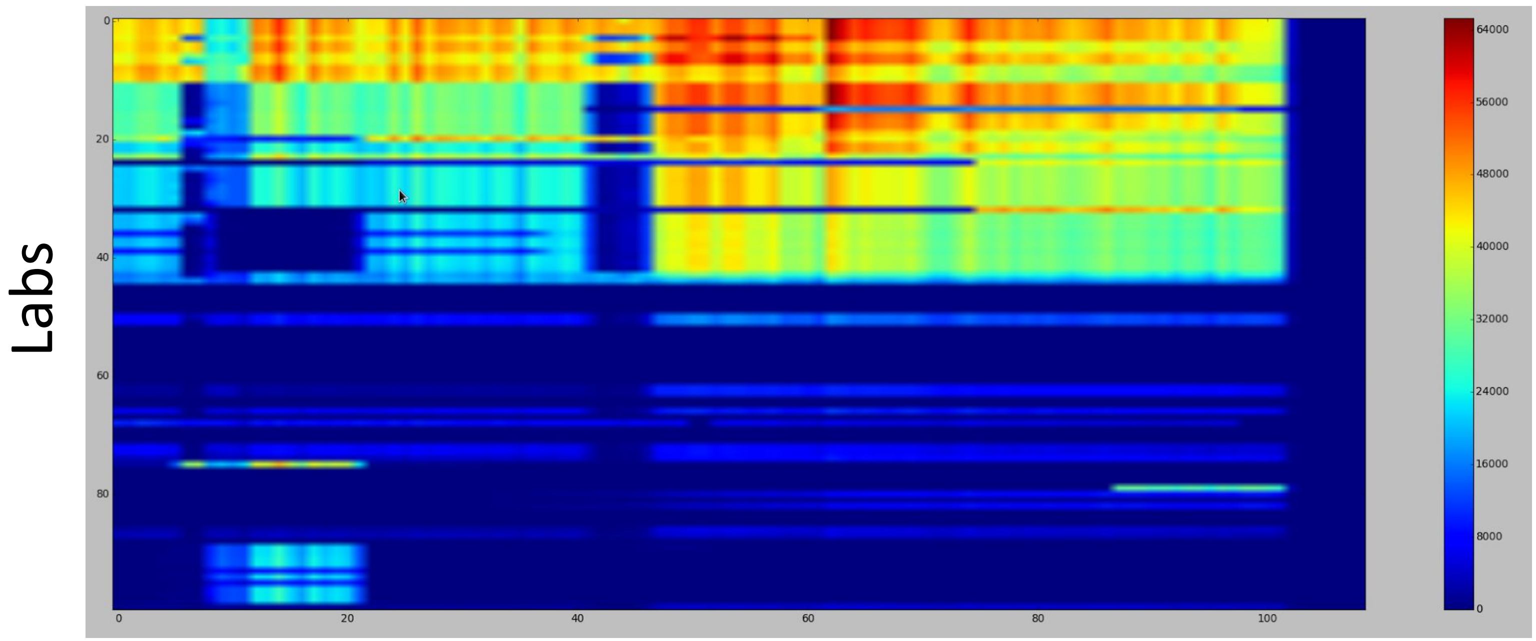
Diabetes Onset After 2009



→ 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.]

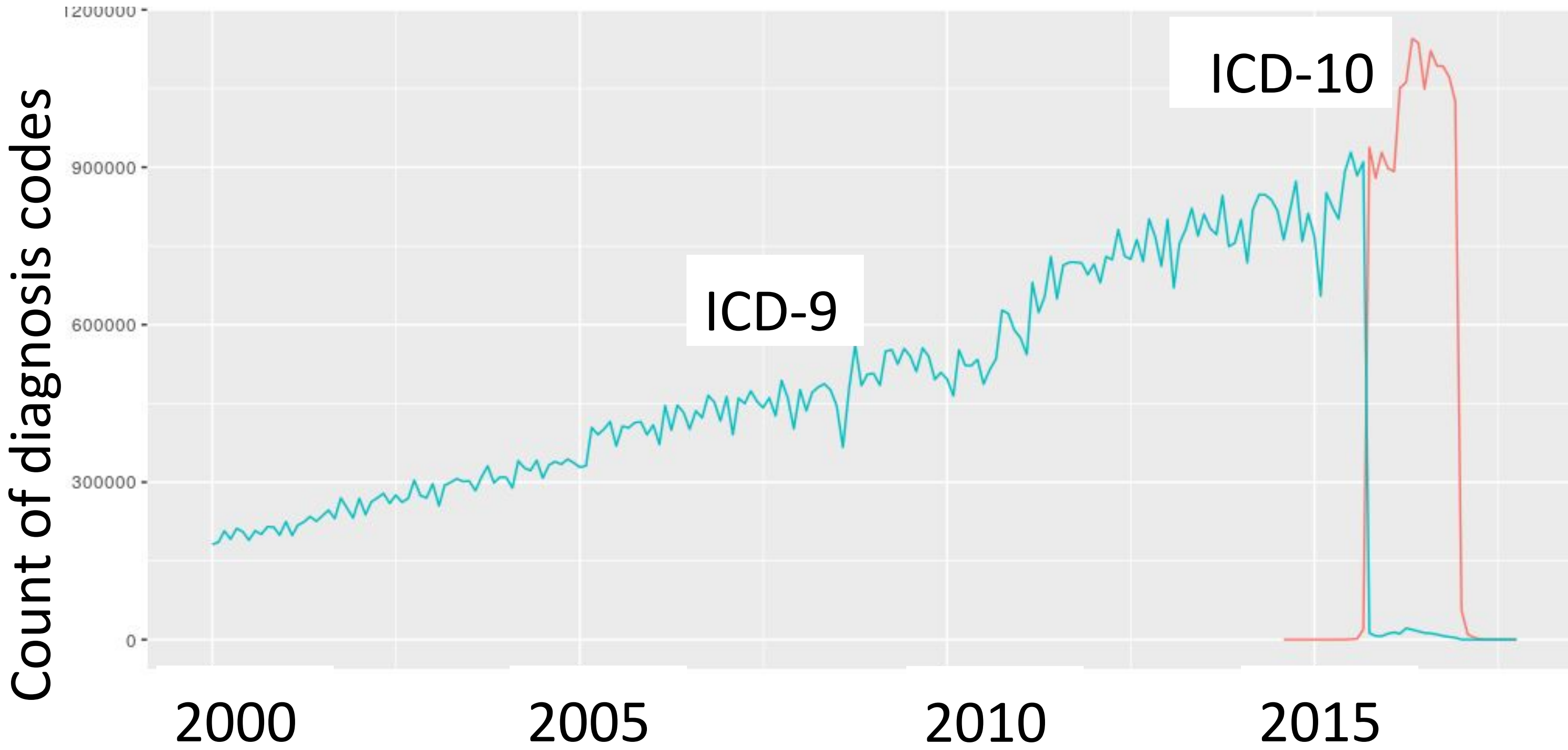
Non-stationarity:
Top 100 lab measurements over time



→ Significance of features may change over time

[Figure credit: Narges Razavian]

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

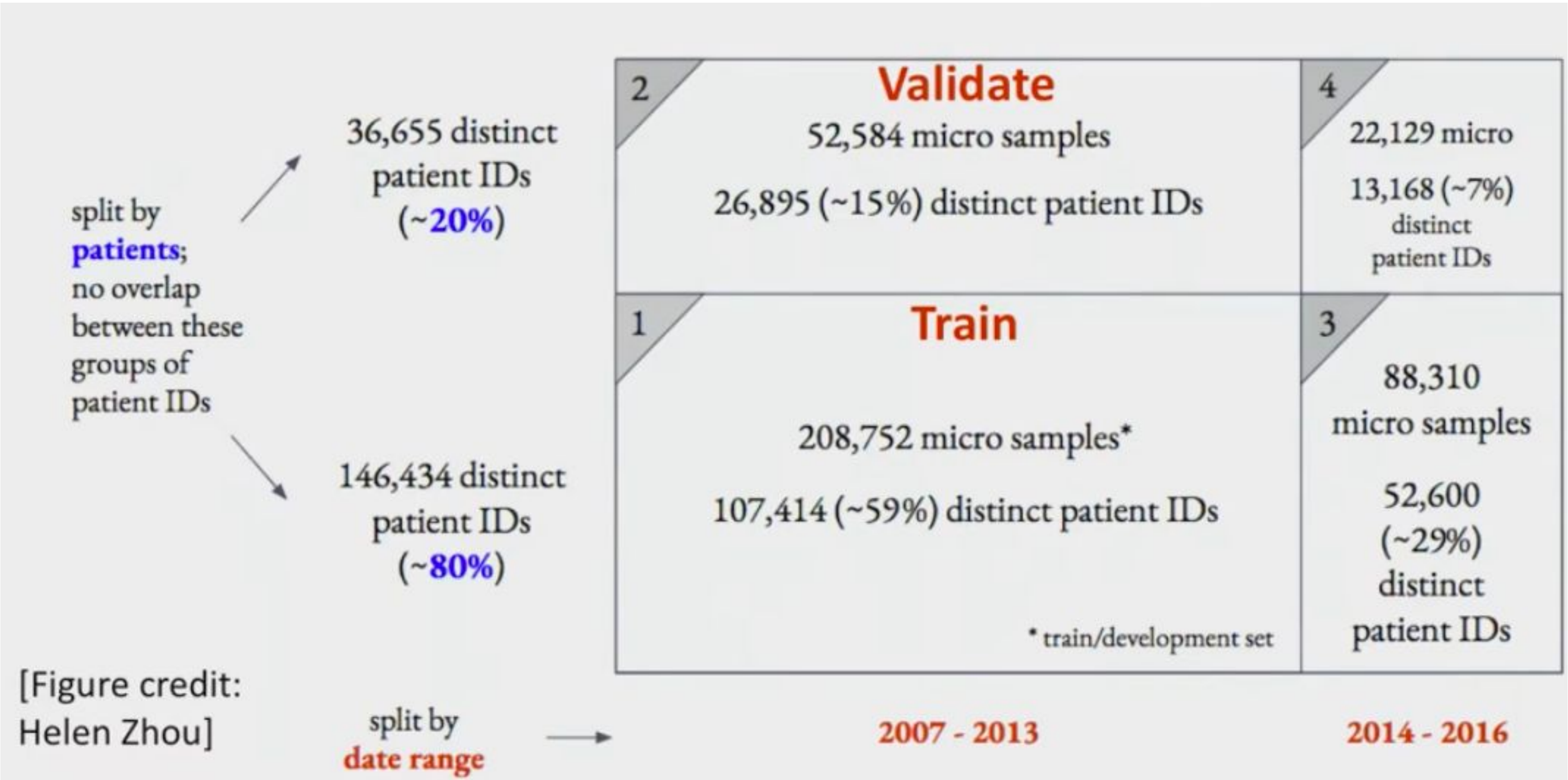


→ Significance of features may change over time

[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:



[Figure credit: Helen Zhou]

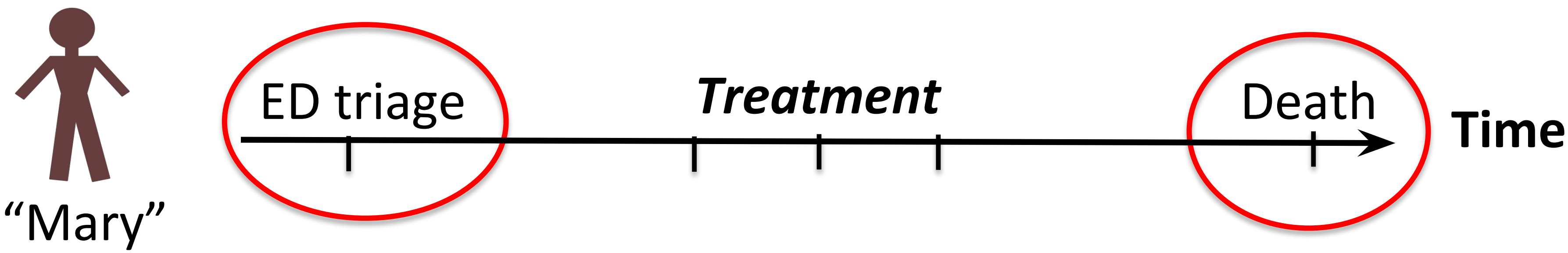
Intervention-tainted outcomes

- 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

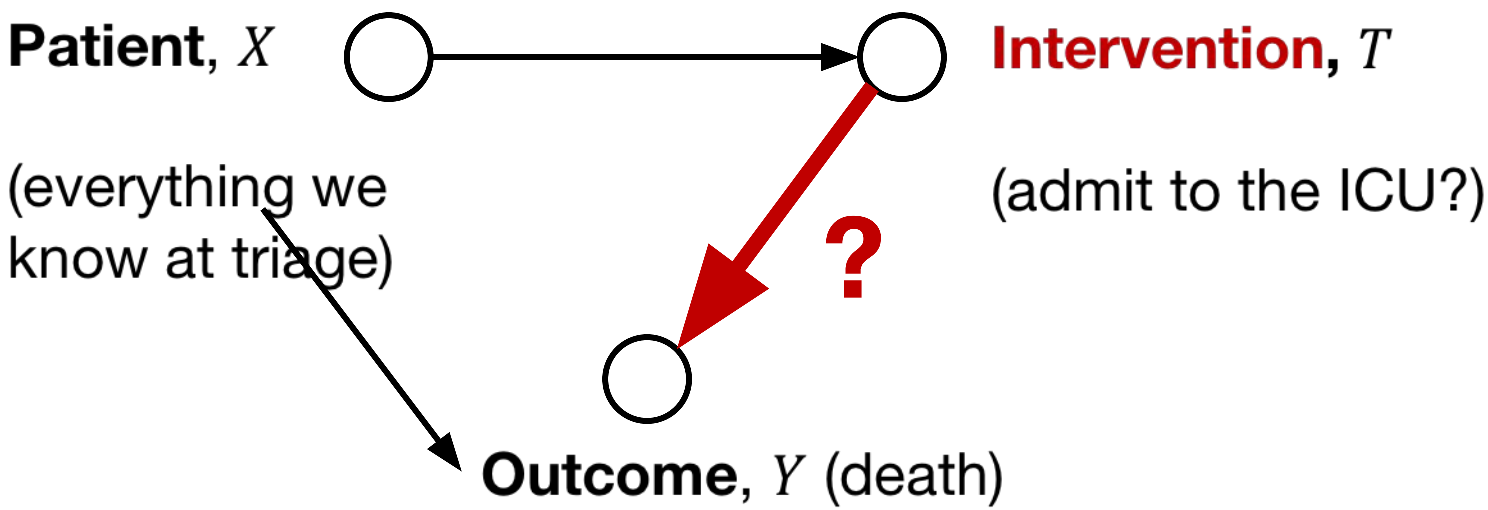
Intervention-tainted outcomes

- 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

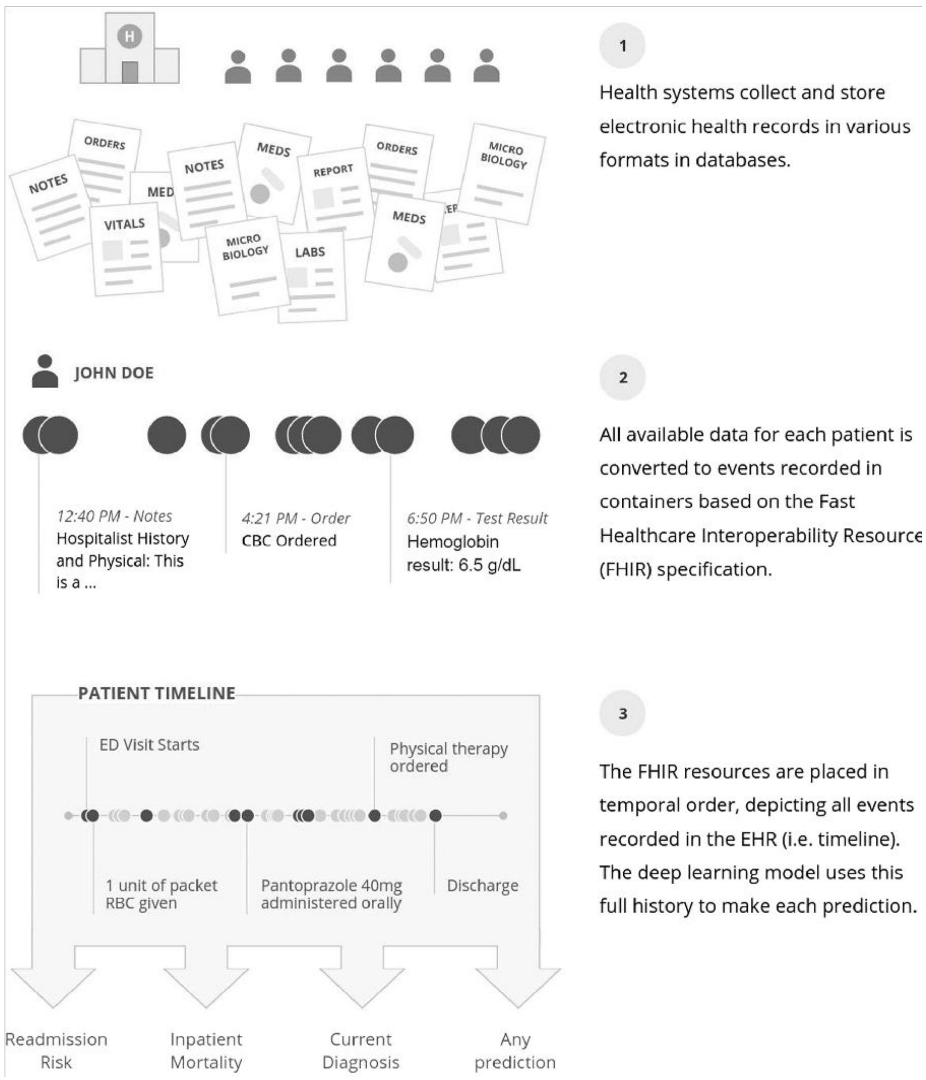
- The rigorous way to address this problem is through the language of **causality:**



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



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, 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

[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, 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 fe[illegible]	[illegible]	[illegible]
Baseli[illegible]	[illegible]	[illegible]
30-da[illegible]		
Deep l[illegible]	[illegible]	[illegible]
Full fe[illegible]	[illegible]	[illegible]
Full fe[illegible]	[illegible]	[illegible]
Baseli[illegible]	[illegible]	[illegible]
Length [illegible]		
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 [illegible] '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]

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