# Machine Learning for Healthcare

6.871, HST.956

Lecture 4: Risk stratification

David Sontag

# Course announcements

- Recitation Friday at 2pm (1-390) – optional
- Office hours Mon 12:30-2pm in 32-G9 lounge
  - *Except for next week! Weds 4-6pm*
- No class Tuesday

- Reflection questions due Tuesday 5pm
- Problem set 1 due Mon Feb 24th 11:59pm
- Sign up for lecture scribing

- All course communication through Piazza

# Roadmap

- **Module 1: Overview of clinical care & data** (3 lectures)

- **Module 2: Using ML for risk stratification and diagnosis** (9 lectures)
  - Supervised learning with noisy, biased, or censored labels
  - Interpretability; Methods for detecting dataset shift; Fairness; Uncertainty

- **Module 3: Suggesting treatments** (4 lectures)
  - Causal inference; Off-policy reinforcement learning

QUIZ

- **Module 4: Understanding disease and its progression** (3 lectures)
  - Unsupervised learning on censored time series with substantial missing data
  - Discovery of disease subtypes; Precision medicine

- **Module 5: Human factors** (3 lectures)
  - Differential diagnosis; Utility-theoretic trade-offs
  - Automating clinical workflows
  - Translating technology into the clinic

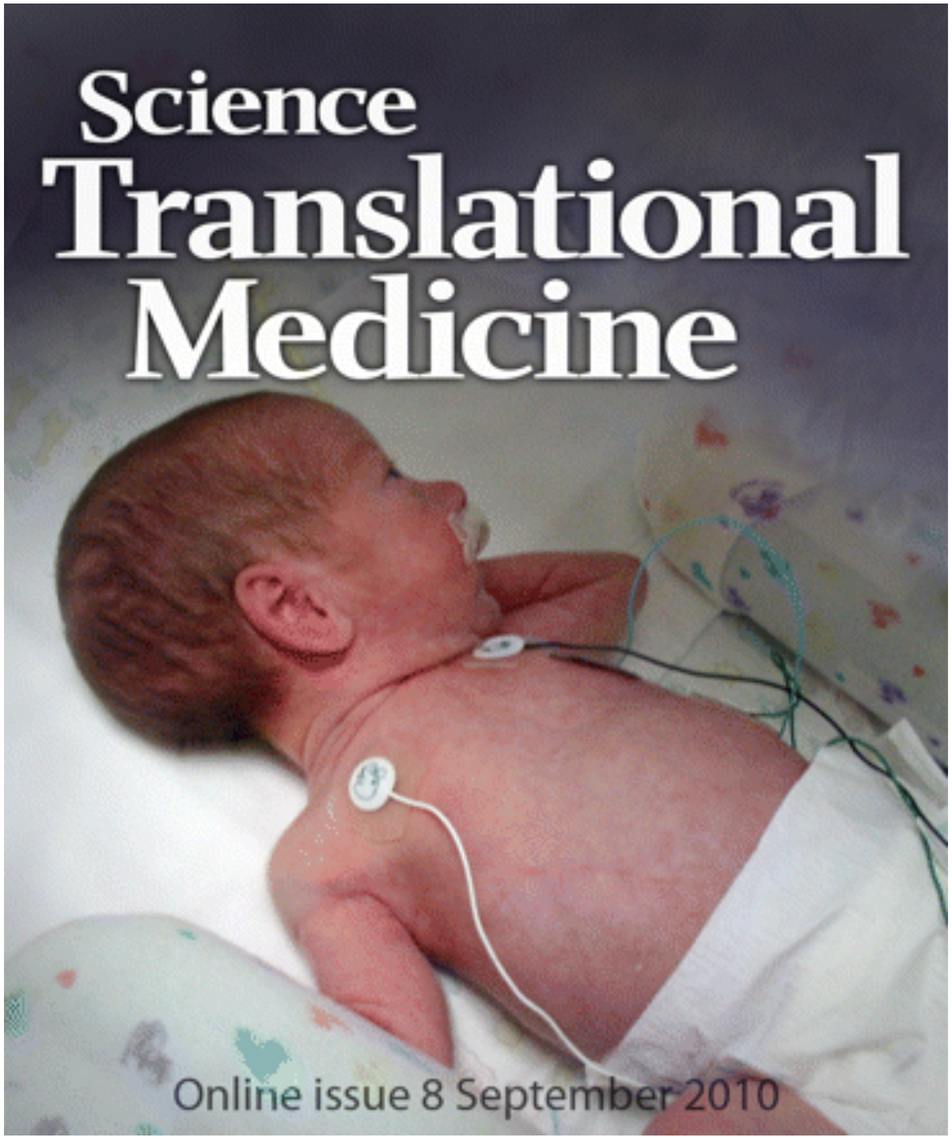
# Outline for today’s class

1. **Risk stratification**
2. Case study: Early detection of Type 2 diabetes
   – Framing as supervised learning problem
   – Deriving labels
   – Evaluating risk stratification algorithms
3. Subtleties with ML-based risk stratification

# What *is* risk stratification?

- Separate a patient population into **high-risk** and **low-risk** of having an outcome
  - Predicting something in the future
  - Goal is different from diagnosis, with distinct performance metrics
- Coupled with **interventions** that target high-risk patients
- Goal is typically to reduce cost and improve patient outcomes

# Examples of risk stratification



Preterm infant’s risk of severe morbidity?

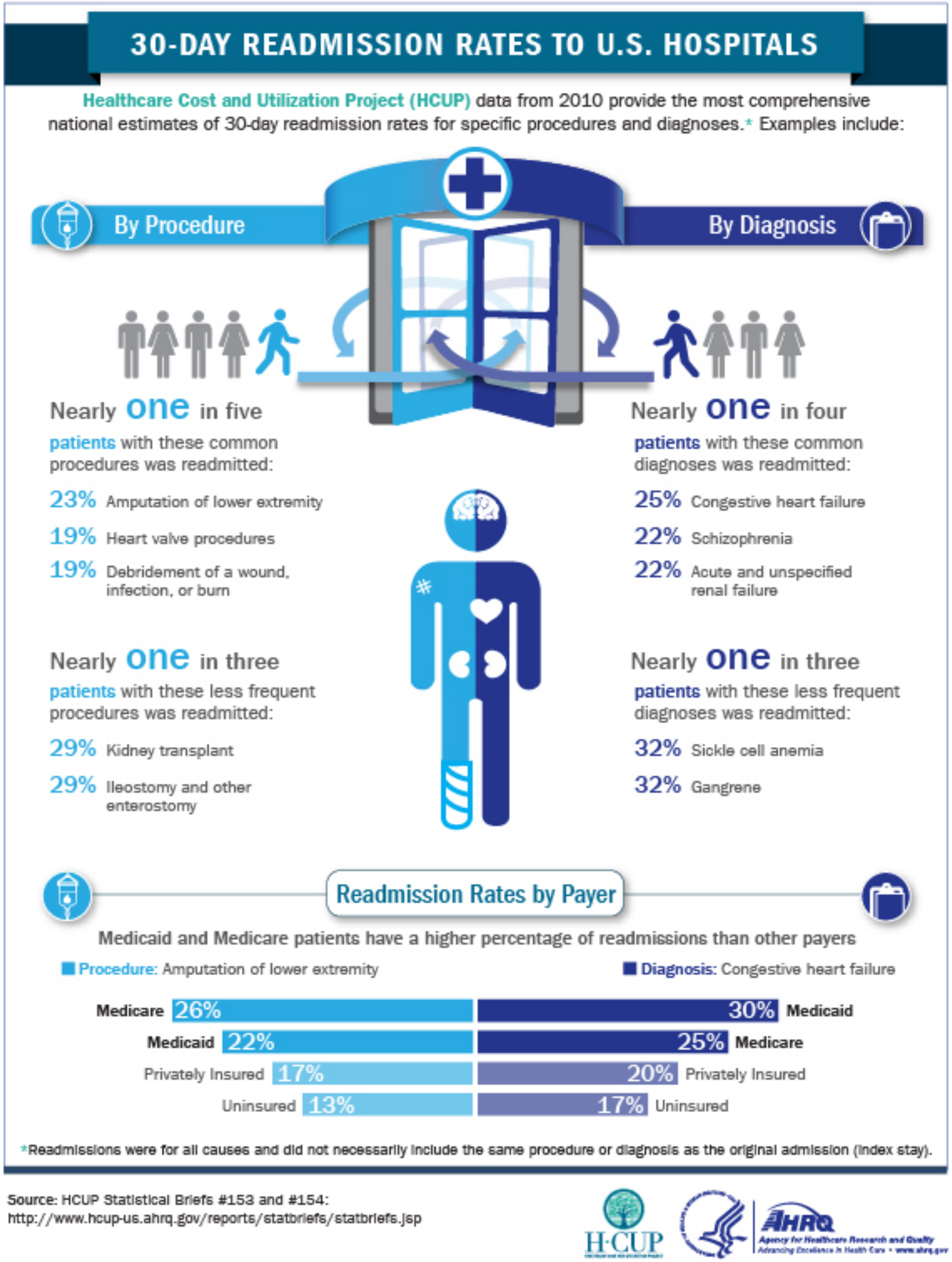
(Saria et al., Science Translational Medicine 2010)

# Examples of risk stratification

Figure source: https://www.drmani.com/heart-attack/

Does this patient need to be admitted to the coronary-care unit?

(Pozen et al., NEJM 1984)



## 30-DAY READMISSION RATES TO U.S. HOSPITALS

Healthcare Cost and Utilization Project (HCUP) data from 2010 provide the most comprehensive national estimates of 30-day readmission rates for specific procedures and diagnoses.* Examples include:

### By Procedure

Nearly **one** in five patients with these common procedures was readmitted:

| | |
|---|---|
| 23% | Amputation of lower extremity |
| 19% | Heart valve procedures |
| 19% | Debridement of a wound, infection, or burn |

Nearly **one** in three patients with these less frequent procedures was readmitted:

| | |
|---|---|
| 29% | Kidney transplant |
| 29% | Ileostomy and other enterostomy |

### By Diagnosis

Nearly **one** in four patients with these common diagnoses was readmitted:

| | |
|---|---|
| 25% | Congestive heart failure |
| 22% | Schizophrenia |
| 22% | Acute and unspecified renal failure |

Nearly **one** in three patients with these less frequent diagnoses was readmitted:

| | |
|---|---|
| 32% | Sickle cell anemia |
| 32% | Gangrene |

### Readmission Rates by Payer

Medicaid and Medicare patients have a higher percentage of readmissions than other payers

| Procedure: Amputation of lower extremity | | Diagnosis: Congestive heart failure | |
|---|---|---|---|
| Medicare | 26% | 30% | Medicaid |
| Medicaid | 22% | 25% | Medicare |
| Privately Insured | 17% | 20% | Privately Insured |
| Uninsured | 13% | 17% | Uninsured |

*Readmissions were for all causes and did not necessarily include the same procedure or diagnosis as the original admission (index stay).

Source: HCUP Statistical Briefs #153 and #154:
http://www.hcup-us.ahrq.gov/reports/statbriefs/statbriefs.jsp

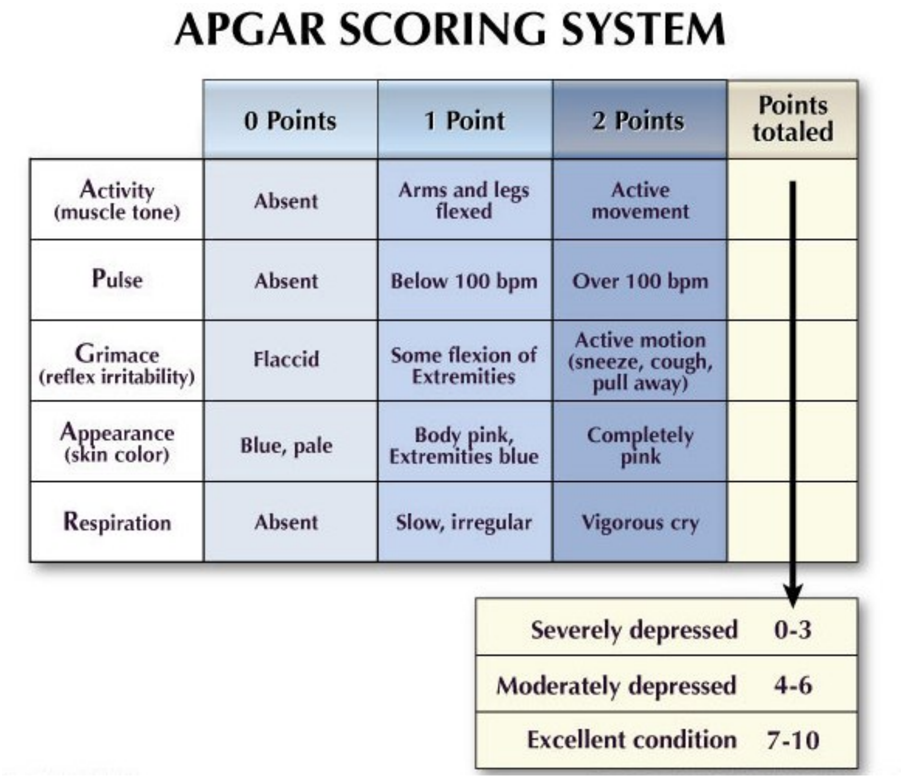
H·CUP
AHRQ
Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • www.ahrq.gov

# Likelihood of hospital readmission?

Figure source: https://www.air.org/project/revolving-door-u-s-hospital-readmissions-diagnosis-and-procedure

# Old vs. New

- Traditionally, risk stratification was based on simple scores using human-entered data



APGAR SCORING SYSTEM

| | 0 Points | 1 Point | 2 Points | Points totaled |
|---|---|---|---|---|
| Activity (muscle tone) | Absent | Arms and legs flexed | Active movement | |
| Pulse | Absent | Below 100 bpm | Over 100 bpm | |
| Grimace (reflex irritability) | Flaccid | Some flexion of Extremities | Active motion (sneeze, cough, pull away) | |
| Appearance (skin color) | Blue, pale | Body pink, Extremities blue | Completely pink | |
| Respiration | Absent | Slow, irregular | Vigorous cry | |

| | |
|---|---|
| Severely depressed | 0-3 |
| Moderately depressed | 4-6 |
| Excellent condition | 7-10 |

# Old vs. New

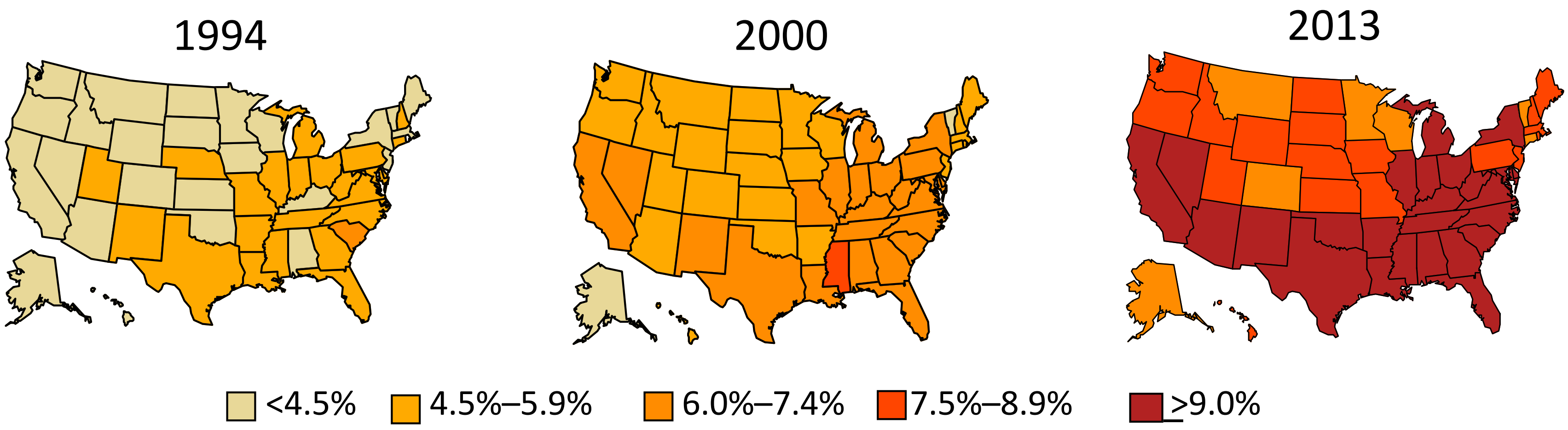
- Traditionally, risk stratification was based on simple scores using human-entered data
- Now, based on machine learning on high-dimensional data
  - Fits more easily into workflow
  - Higher accuracy
  - Quicker to derive (can special case)
- **But, ML approach comes with new challenges – to be discussed**

# Outline for today's class

1. Risk stratification
2. **Case study: Early detection of Type 2 diabetes**
   - Framing as supervised learning problem
   - Deriving labels
   - Evaluating risk stratification algorithms
3. Subtleties with ML-based risk stratification

[Razavian, Blecker, Schmidt, Smith-McLallen, Nigam, Sontag. *Big Data.* '16]

# Type 2 Diabetes: A Major public health challenge



$245 billion: Total costs of diagnosed diabetes in the United States in 2012
$831 billion: Total fiscal year federal budget for healthcare in the United States in 2014

# Type 2 Diabetes Can Be Prevented *

Requirement for successful large scale prevention program

1. Detect/reach truly at risk population
2. Improve the interventions
3. Lower the cost of intervention

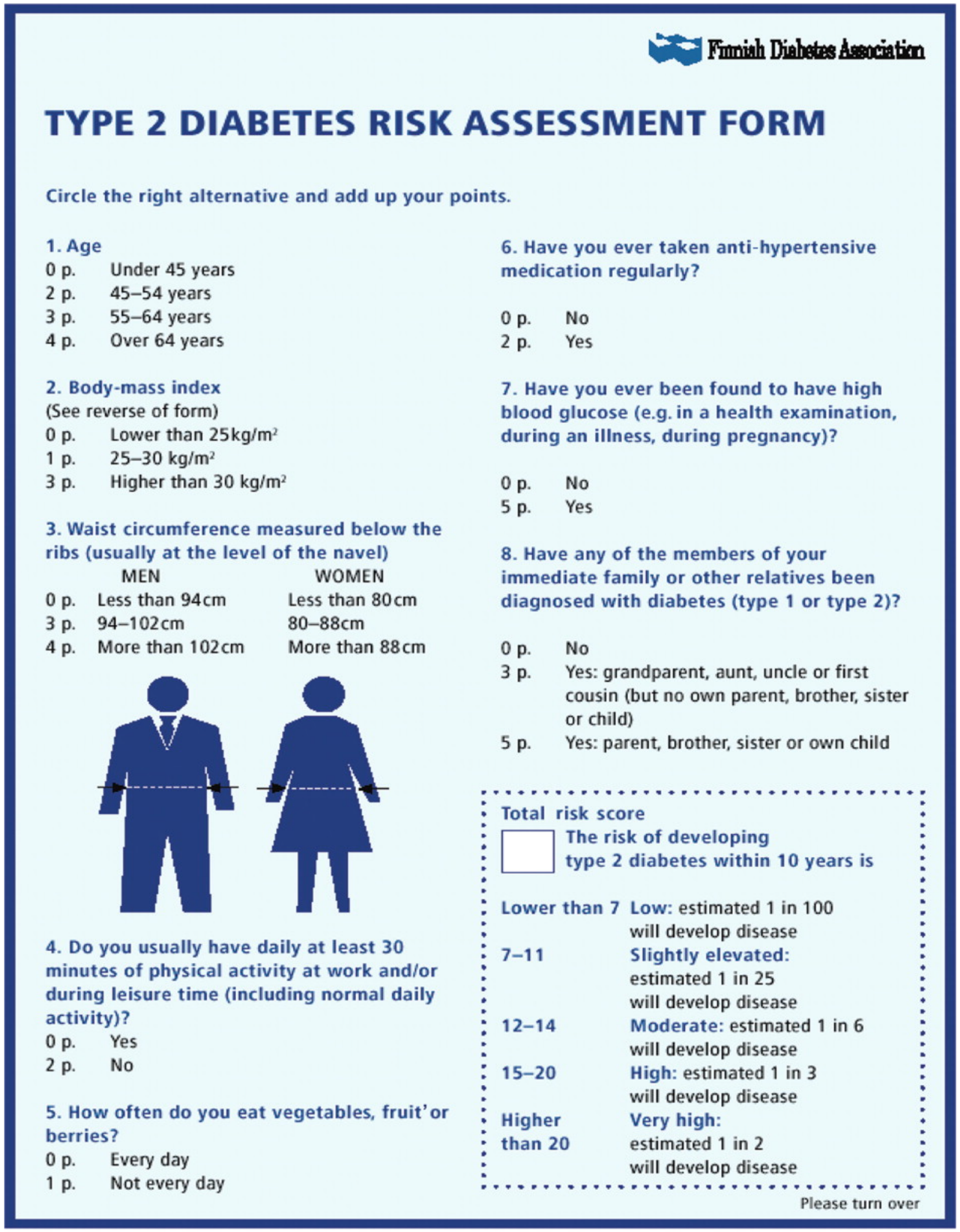
* Diabetes Prevention Program Research Group. "Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin." The New England journal of medicine 346.6 (2002): 393.

# Traditional Risk Prediction Models

- Successful Examples
  - ARIC
  - KORA
  - FRAMINGHAM
  - AUSDRISC
  - FINDRISC
  - San Antonio Model

- Easy to ask/measure in the office, or for patients to do online

- Simple model:
  can calculate scores by hand



Finnish Diabetes Association

## TYPE 2 DIABETES RISK ASSESSMENT FORM

Circle the right alternative and add up your points.

**1. Age**

| | |
|---|---|
| 0 p. | Under 45 years |
| 2 p. | 45–54 years |
| 3 p. | 55–64 years |
| 4 p. | Over 64 years |

**2. Body-mass index**
(See reverse of form)

| | |
|---|---|
| 0 p. | Lower than 25kg/$m^2$ |
| 1 p. | 25–30 kg/$m^2$ |
| 3 p. | Higher than 30 kg/$m^2$ |

**3. Waist circumference measured below the ribs (usually at the level of the navel)**

| | MEN | WOMEN |
|---|---|---|
| 0 p. | Less than 94cm | Less than 80cm |
| 3 p. | 94–102cm | 80–88cm |
| 4 p. | More than 102cm | More than 88cm |

**4. Do you usually have daily at least 30 minutes of physical activity at work and/or during leisure time (including normal daily activity)?**

| | |
|---|---|
| 0 p. | Yes |
| 2 p. | No |

**5. How often do you eat vegetables, fruit'or berries?**

| | |
|---|---|
| 0 p. | Every day |
| 1 p. | Not every day |

**6. Have you ever taken anti-hypertensive medication regularly?**

| | |
|---|---|
| 0 p. | No |
| 2 p. | Yes |

**7. Have you ever been found to have high blood glucose (e.g. in a health examination, during an illness, during pregnancy)?**

| | |
|---|---|
| 0 p. | No |
| 5 p. | Yes |

**8. Have any of the members of your immediate family or other relatives been diagnosed with diabetes (type 1 or type 2)?**

| | |
|---|---|
| 0 p. | No |
| 3 p. | Yes: grandparent, aunt, uncle or first cousin (but no own parent, brother, sister or child) |
| 5 p. | Yes: parent, brother, sister or own child |

**Total risk score**

□ The risk of developing type 2 diabetes within 10 years is

| | |
|---|---|
| Lower than 7 | **Low:** estimated 1 in 100 will develop disease |
| 7–11 | **Slightly elevated:** estimated 1 in 25 will develop disease |
| 12–14 | **Moderate:** estimated 1 in 6 will develop disease |
| 15–20 | **High:** estimated 1 in 3 will develop disease |
| Higher than 20 | **Very high:** estimated 1 in 2 will develop disease |

Please turn over

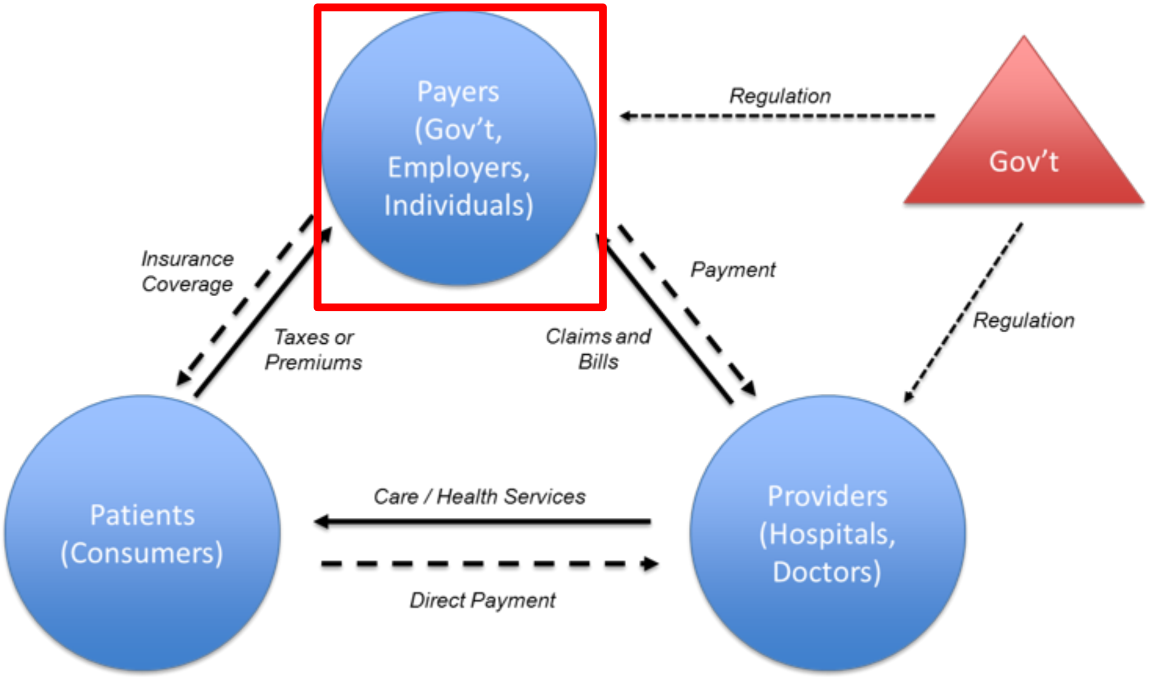
Test designed by Professor Jaakko Tuomilehto, Department of Public Health, University of Helsinki, and Jaana Lindström, MFS, National Public Health Institute.

# Challenges of Traditional Risk Prediction Models

- A screening step needs to be done for every member in the population
  - Either in the physician’s office or as surveys
  - Costly and time-consuming
  - Infeasible for regular screening for millions of individuals
- Models not easy to adapt to multiple surrogates, when a variable is missing
  - Discovery of surrogates not straightforward

# Population-Level Risk Stratification

- Key idea: Use readily available administrative, utilization, and clinical data



Source for figure: http://www.mahesh-vc.com/blog/understanding-whos-paying-for-what-in-the-healthcare-industry

# Population-Level Risk Stratification

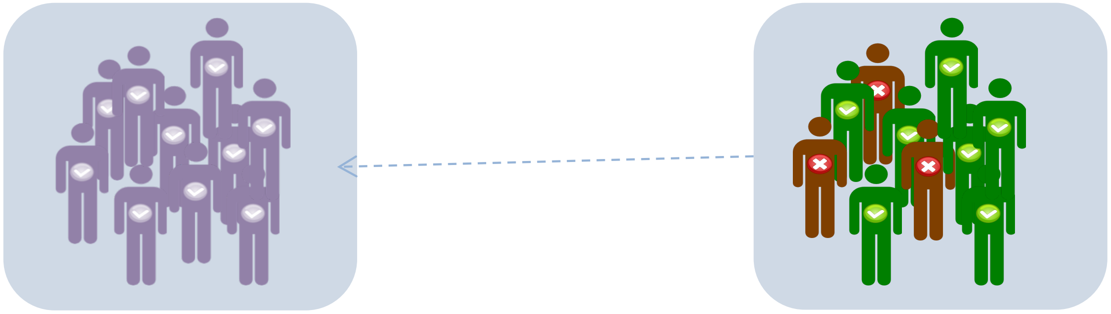
- Key idea: Use readily available administrative, utilization, and clinical data

- Machine learning will find surrogates for risk factors that would otherwise be missing

- Perform risk stratification at the population level – millions of patients

# A Data-Driven approach on Longitudinal Data

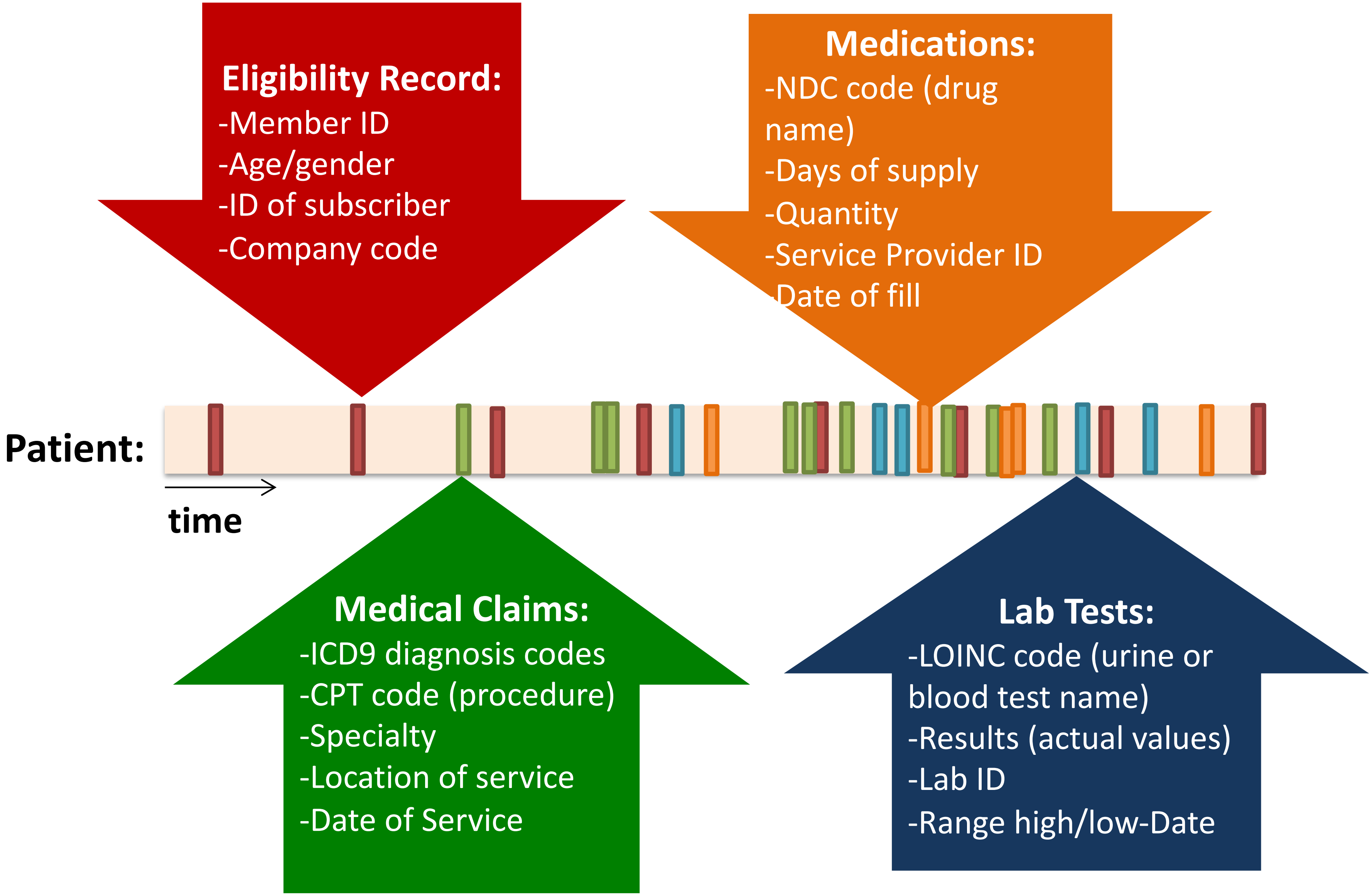
- Looking at individuals who got diabetes *today,* (compared to those who didn’t)
  - Can we infer which variables in their record could have predicted their health outcome?



A Few Years Ago

Today

# Administrative & Clinical Data



**Eligibility Record:**

-Member ID

-Age/gender

-ID of subscriber

-Company code

**Medications:**

-NDC code (drug name)

-Days of supply

-Quantity

-Service Provider ID

-Date of fill

**Medical Claims:**

-ICD9 diagnosis codes

-CPT code (procedure)

-Specialty

-Location of service

-Date of Service

**Lab Tests:**

-LOINC code (urine or blood test name)

-Results (actual values)

-Lab ID

-Range high/low-Date

# Top diagnosis codes

| Disease | count |
|---|---|
| **401.1 Benign hypertension** | 447017 |
| 272.4 Hyperlipidemia NEC/NOS | 382030 |
| 401.9 Hypertension NOS | 372477 |
| **250.00 DMII wo cmp nt st uncntr** | 339522 |
| 272.0 Pure hypercholesterolem | 232671 |
| 272.2 Mixed hyperlipidemia | 180015 |
| V72.31 Routine gyn examination | 178709 |
| 244.9 Hypothyroidism NOS | 169829 |
| **780.79 Malaise and fatigue NEC** | 149797 |
| **V04.81 Vaccin for influenza** | 147858 |
| **724.2 Lumbago** | 137345 |
| **V76.12 Screen mammogram NEC** | 129445 |
| **V70.0 Routine medical exam** | 127848 |

| Disease | count |
|---|---|
| **530.81 Esophageal reflux** | 121064 |
| 427.31 Atrial fibrillation | 113798 |
| **729.5 Pain in limb** | 112449 |
| 414.01 Crnry athrscl natve vssl | 104478 |
| 285.9 Anemia NOS | 103351 |
| **786.50 Chest pain NOS** | 91999 |
| **599.0 Urin tract infection NOS** | 87982 |
| V58.69 Long-term use meds NEC | 85544 |
| **496 Chr airway obstruct NEC** | 78585 |
| 477.9 Allergic rhinitis NOS | 77963 |
| 414.00 Cor ath unsp vsl ntv/gft | 75519 |

| Disease | count |
|---|---|
| 719.47 Joint pain-ankle | 28648 |
| 300.4 Dysthymic disorder | 28530 |
| 268.9 Vitamin D deficiency NOS | 28455 |
| V72.81 Preop cardiovsclr exam | 27897 |
| **724.3 Sciatica** | 27604 |
| **787.91 Diarrhea** | 27424 |
| **V2.21 Supervis oth normal preg** | 27320 |
| 365.01 Opn angl brderln lo risk | 26033 |
| 379.21 Vitreous degeneration | 25592 |
| 424.1 Aortic valve disorder | 25425 |
| 616.10 Vaginitis NOS | 24736 |
| 702.19 Other sborheic keratosis | 24453 |
| 380.4 Impacted cerumen | 24046 |

**Out of 135K patients who had laboratory data**

# Top lab test results

| Lab test | |
|---|---|
| 2160-0 Creatinine | 1284737 |
| 3094-0 Urea nitrogen | 1282344 |
| 2823-3 Potassium | 1280812 |
| 2345-7 Glucose | 1299897 |
| 1742-6 Alanine aminotransferase | 1187809 |
| 1920-8 Aspartate aminotransferase | 1187965 |
| 2885-2 Protein | 1277338 |
| 1751-7 Albumin | 1274166 |
| 2093-3 Cholesterol | 1268269 |
| 2571-8 Triglyceride | 1257751 |
| 13457-7 Cholesterol.in LDL | 1241208 |
| 17861-6 Calcium | 1165370 |
| 2951-2 Sodium | 1167675 |

| Lab test | |
|---|---|
| 2085-9 Cholesterol.in HDL | 1155666 |
| 718-7 Hemoglobin | 1152726 |
| 4544-3 Hematocrit | 1147893 |
| 9830-1 Cholesterol.total/Cholesterol.in HDL | 1037730 |
| 33914-3 Glomerular filtration rate/1.73 sq M.predicted | 561309 |
| 785-6 Erythrocyte mean corpuscular hemoglobin | 1070832 |
| 6690-2 Leukocytes | 1062980 |
| 789-8 Erythrocytes | 1062445 |
| 787-2 Erythrocyte mean corpuscular volume | 1063665 |

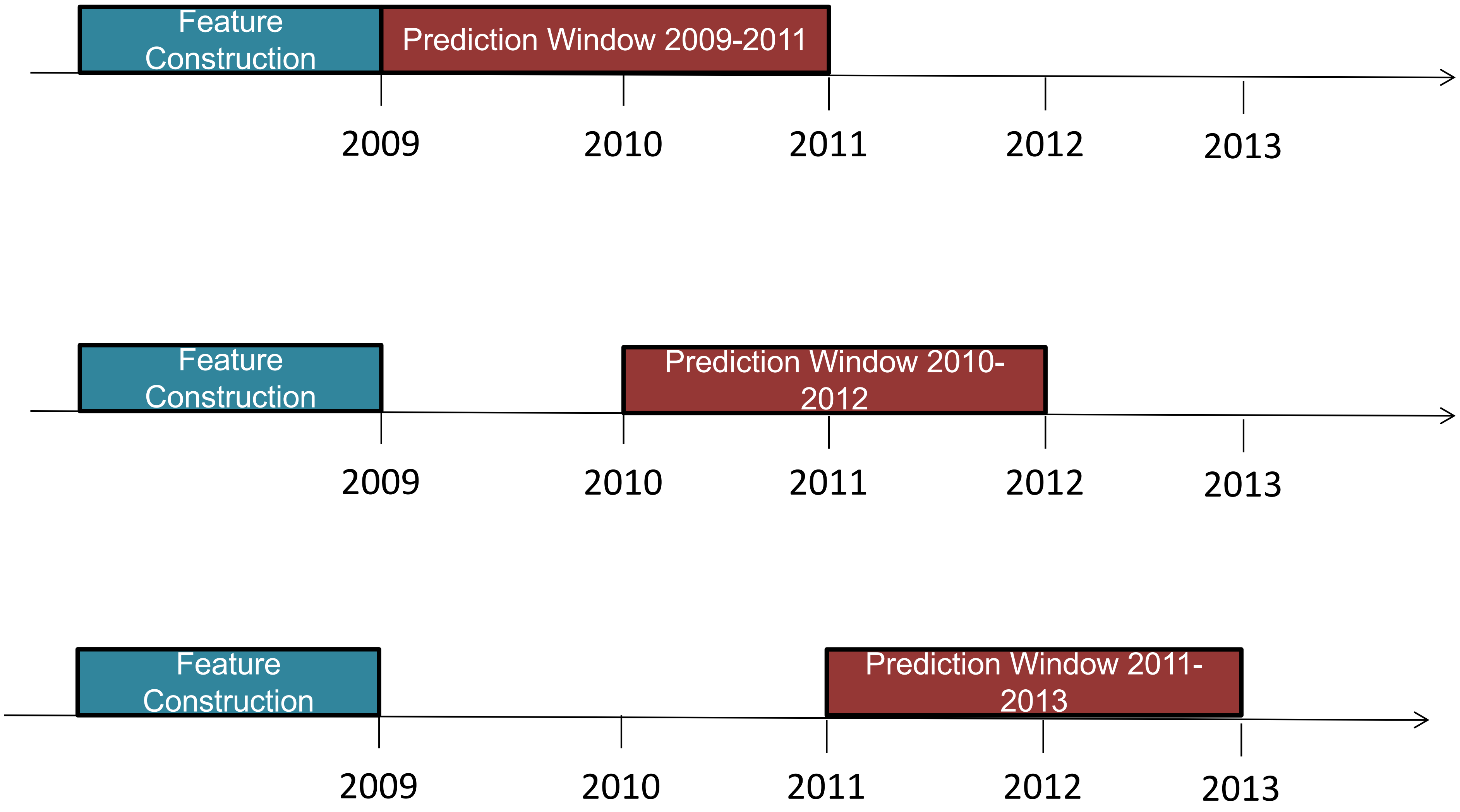
| Lab test | |
|---|---|
| 770-8 Neutrophils/100 leukocytes | 952089 |
| 731-0 Lymphocytes | 943918 |
| 704-7 Basophils | 863448 |
| 711-2 Eosinophils | 935710 |
| 5905-5 Monocytes/100 leukocytes | 943764 |
| 706-2 Basophils/100 leukocytes | 863435 |
| 751-8 Neutrophils | 943232 |
| 742-7 Monocytes | 942978 |
| 713-8 Eosinophils/100 leukocytes | 933929 |
| 3016-3 Thyrotropin | 891807 |
| 4548-4 Hemoglobin A1c/Hemoglobin.total | 527062 |

**Count of people who have the test result (ever)**

# Outline for today’s class

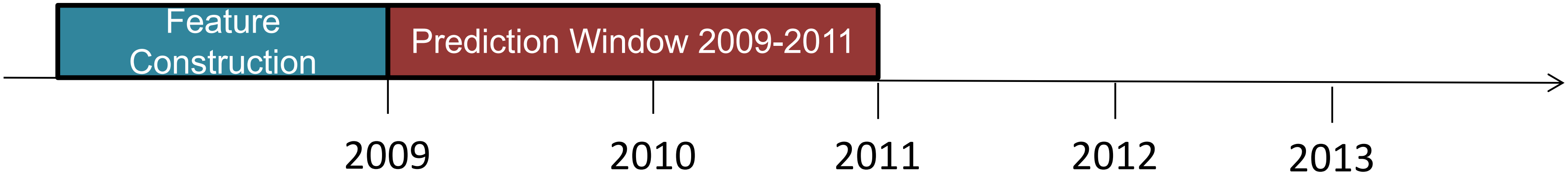
1. Risk stratification
2. Case study: Early detection of Type 2 diabetes
   - **Framing as supervised learning problem**
   - Deriving labels
   - Evaluating risk stratification algorithms
3. Subtleties with ML-based risk stratification

# Framing for supervised machine learning



**Gap is important to prevent label leakage**

# Framing for supervised machine learning

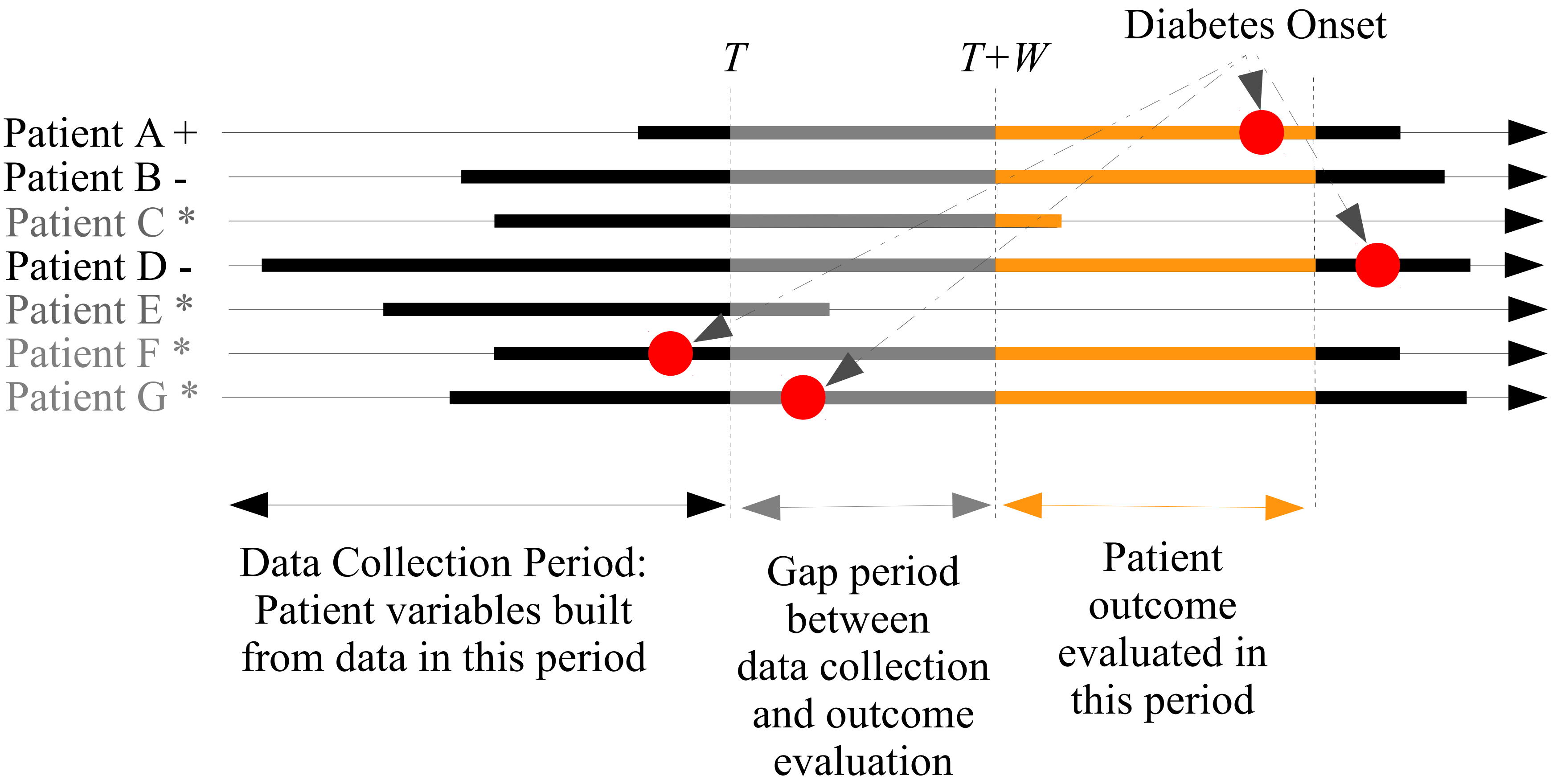


**Problem: Data is censored!**

- Patients change health insurers frequently, but data doesn’t follow them
- *Left censored*: may not have enough data to derive features
- *Right censored*: may not know label

# Reduction to binary classification

Exclude patients that are left- and right-censored.


Diabetes Onset
T
T+W
Patient A +
Patient B -
Patient C *
Patient D -
Patient E *
Patient F *
Patient G *
Data Collection Period: Patient variables built from data in this period
Gap period between data collection and outcome evaluation
Patient outcome evaluated in this period

This is an example of alignment by *absolute time*

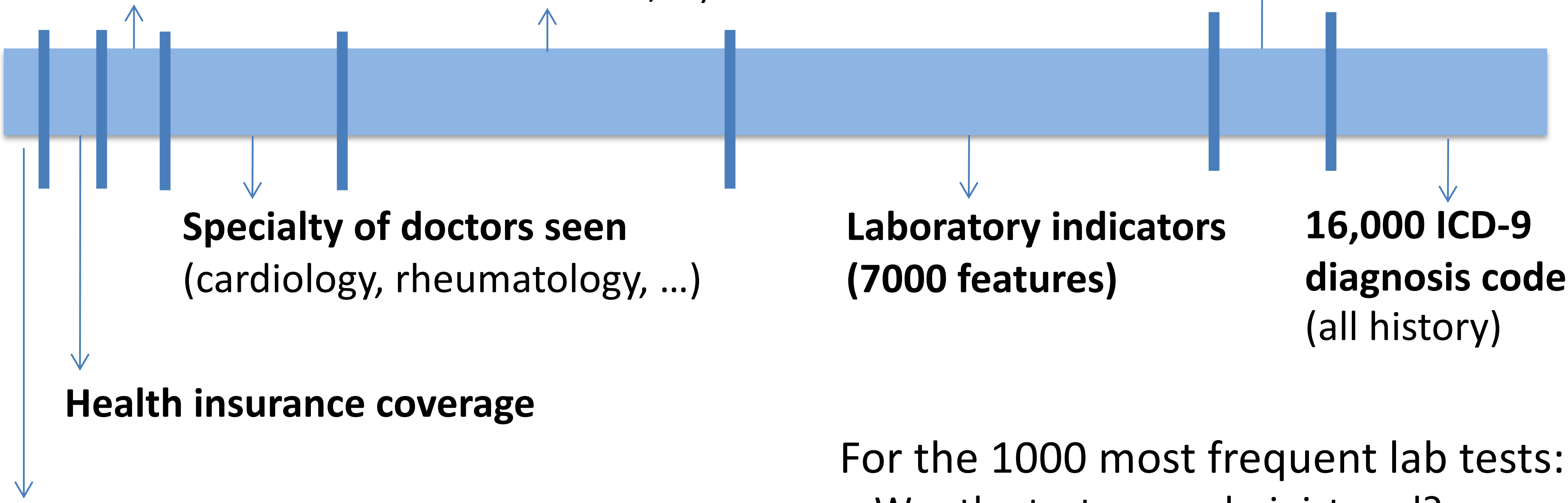
# Alternative framings

- Align by relative time, e.g.
  - 2 hours into patient stay in ER
  - Every time patient sees PCP
  - When individual turns 40 yrs old
- Align by data availability

**NOTE:**

- If multiple data points per patient, make sure each patient in *only* train, validate, or test

# Features used in models



**Service place** (urgent care, inpatient, outpatient, …)

**Medications taken (999 features)** (laxatives, metformin, anti-arthritics, …)

**Procedures performed (457 features)**

**Specialty of doctors seen** (cardiology, rheumatology, …)

**Laboratory indicators (7000 features)**

**16,000 ICD-9 diagnosis codes** (all history)

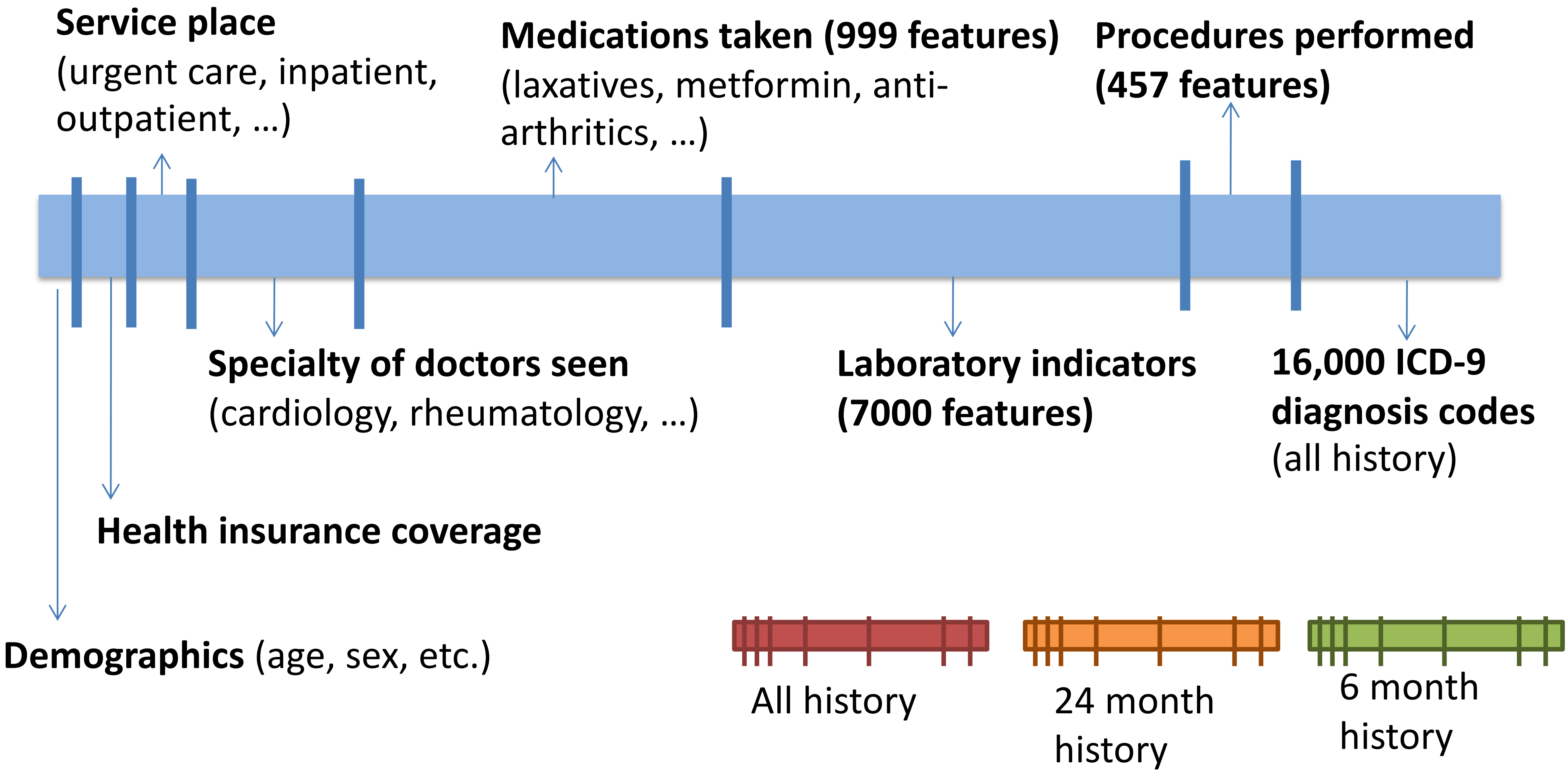
**Health insurance coverage**

**Demographics** (age, sex, etc.)

For the 1000 most frequent lab tests:

- Was the test ever administered?
- Was the result ever low?
- Was the result ever high?
- Was the result ever normal?
- Is the value increasing?
- Is the value decreasing?
- Is the value fluctuating?

# Features used in models



**Service place**
(urgent care, inpatient, outpatient, …)

**Medications taken (999 features)**
(laxatives, metformin, anti-arthritics, …)

**Procedures performed (457 features)**

**Specialty of doctors seen**
(cardiology, rheumatology, …)

**Laboratory indicators (7000 features)**

**16,000 ICD-9 diagnosis codes**
(all history)

**Health insurance coverage**

**Demographics** (age, sex, etc.)

All history

24 month history

6 month history

## 10s-100s of thousands of features

# Logistic regression with L1 regularization

- Penalizing the L1 norm of the weight vector leads to *sparse* (read: many 0's) solutions for *w.*

$$\min_w \sum_i \ell(x_i, y_i; w) + \lambda ||w||_1 \qquad ||\vec{w}||_1 = \sum_d |w_d|$$

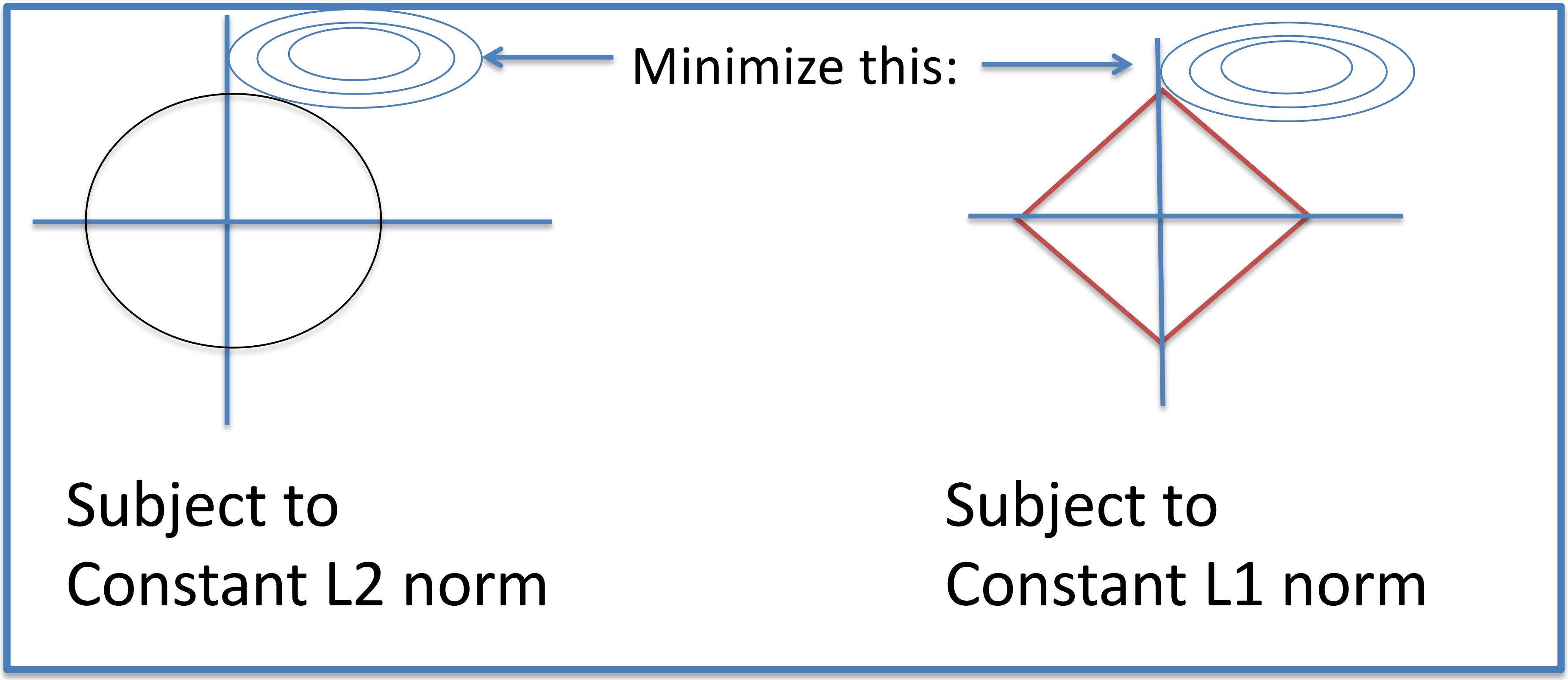
instead of

$$\min_w \sum_i \ell(x_i, y_i; w) + \lambda ||w||_2^2 \qquad ||\vec{w}||_2^2 = \sum_d w_d^2$$

- Why?

# Logistic regression with L1 regularization

- Penalizing the L1 norm of the weight vector leads to *sparse* (read: many 0’s) solutions for *w.*



Minimize this:

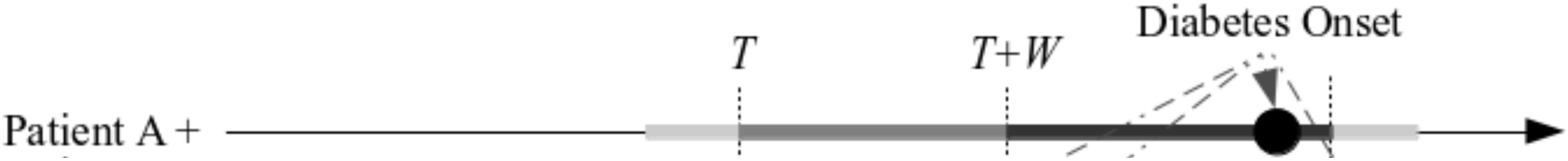
Subject to
Constant L2 norm

Subject to
Constant L1 norm

# Outline for today’s class

1. Risk stratification
2. Case study: Early detection of Type 2 diabetes
   - Framing as supervised learning problem
   - **Deriving labels**
   - Evaluating risk stratification algorithms
3. Subtleties with ML-based risk stratification

# Where do the labels come from?

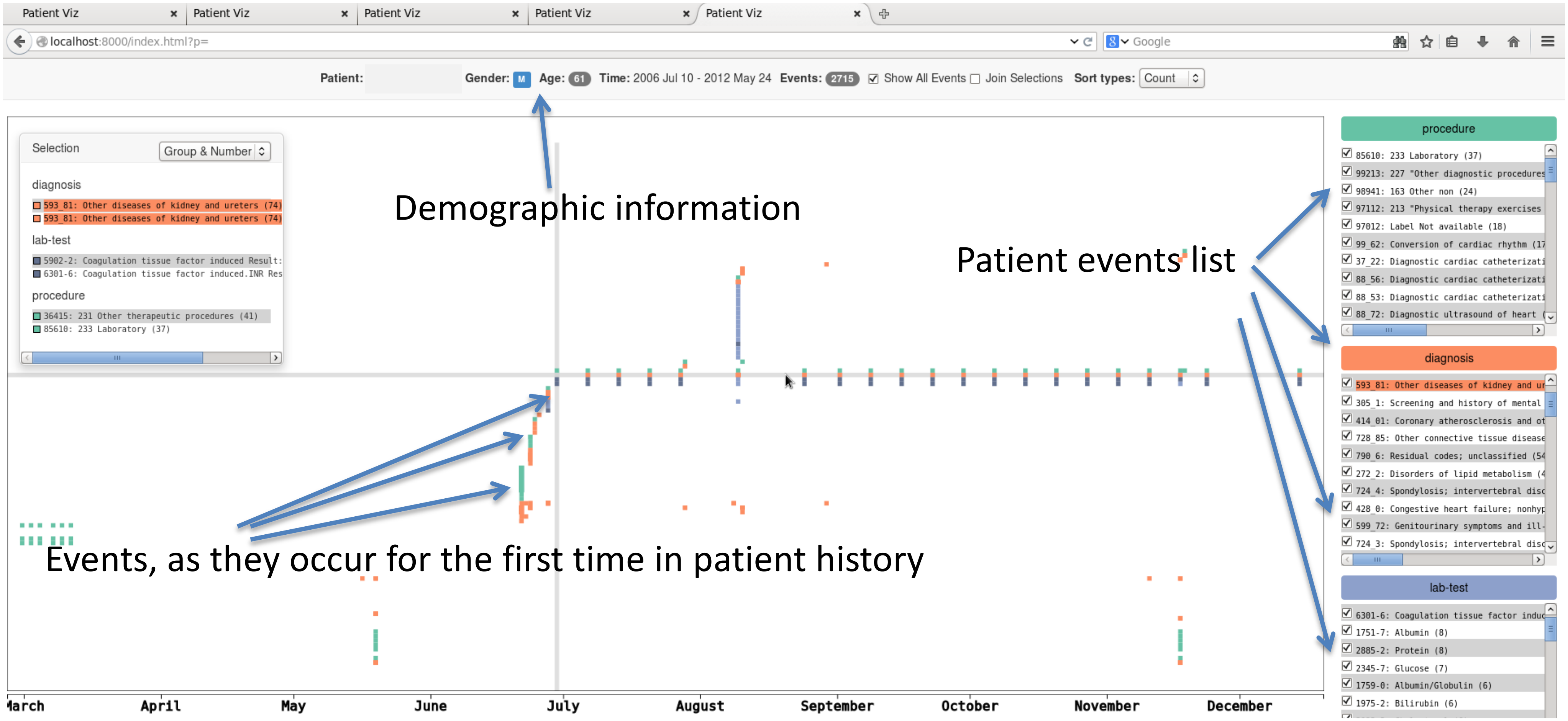


Typical pipeline:

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



**Patient:** **Gender:** M **Age:** 61 **Time:** 2006 Jul 10 - 2012 May 24 **Events:** 2715 ☑ Show All Events □ Join Selections **Sort types:** Count

Selection: Group & Number

diagnosis
- 🟧 593_81: Other diseases of kidney and ureters (74)
- 🟧 593_81: Other diseases of kidney and ureters (74)

lab-test
- ■ 5902-2: Coagulation tissue factor induced Result:
- ■ 6301-6: Coagulation tissue factor induced.INR Res

procedure
- 🟩 36415: 231 Other therapeutic procedures (41)
- 🟩 85610: 233 Laboratory (37)

procedure
- ☑ 85610: 233 Laboratory (37)
- ☑ 99213: 227 "Other diagnostic procedures
- ☑ 98941: 163 Other non (24)
- ☑ 97112: 213 "Physical therapy exercises
- ☑ 97012: Label Not available (18)
- ☑ 99_62: Conversion of cardiac rhythm (17
- ☑ 37_22: Diagnostic cardiac catheterizati
- ☑ 88_56: Diagnostic cardiac catheterizati
- ☑ 88_53: Diagnostic cardiac catheterizati
- ☑ 88_72: Diagnostic ultrasound of heart (

diagnosis
- ☑ 593_81: Other diseases of kidney and ur
- ☑ 305_1: Screening and history of mental
- ☑ 414_01: Coronary atherosclerosis and ot
- ☑ 728_85: Other connective tissue disease
- ☑ 790_6: Residual codes; unclassified (54
- ☑ 272_2: Disorders of lipid metabolism (4
- ☑ 724_4: Spondylosis; intervertebral disc
- ☑ 428_0: Congestive heart failure; nonhyp
- ☑ 599_72: Genitourinary symptoms and ill-
- ☑ 724_3: Spondylosis; intervertebral disc

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

Demographic information

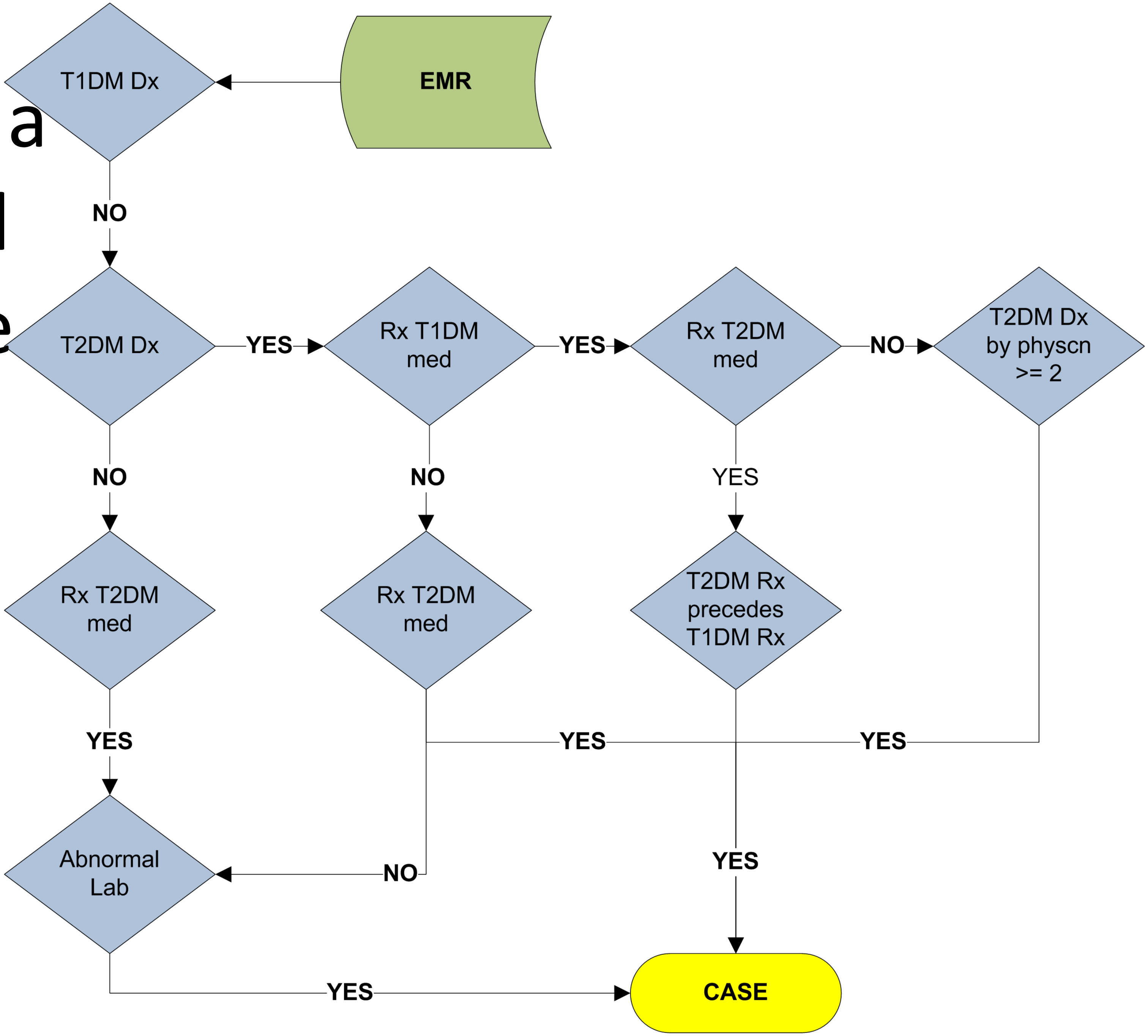
Patient events list

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.



EMR

T1DM Dx

NO

T2DM Dx

YES

Rx T1DM med

YES

Rx T2DM med

NO

T2DM Dx by physcn >= 2

NO

Rx T2DM med

YES

Abnormal Lab

NO

Rx T2DM med

YES

NO

YES

T2DM Rx precedes T1DM Rx

YES

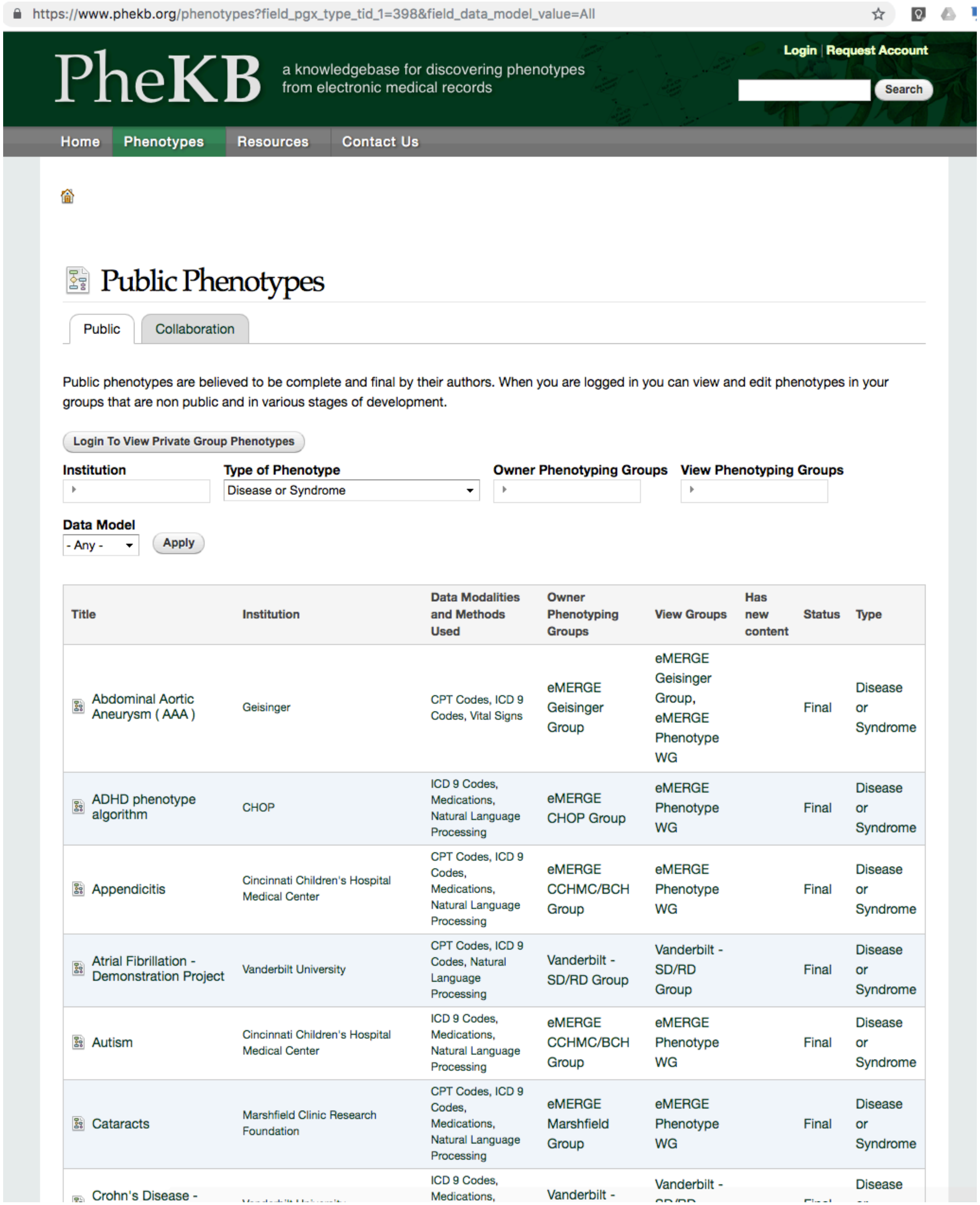
YES

YES

CASE

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

## Public Phenotypes

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.

Institution:

Type of Phenotype: Disease or Syndrome

Owner Phenotyping Groups:

View Phenotyping Groups:

Data Model: - Any -

| 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 - SD/RD | | Final | Disease or |

If the derived label is noisy, how does it affect learning?

# Outline for today’s class

1. Risk stratification
2. Case study: Early detection of Type 2 diabetes
   - Framing as supervised learning problem
   - Deriving labels
   - **Evaluating risk stratification algorithms**
3. Subtleties with ML-based risk stratification

# What are the Discovered Risk Factors?

- 769 variables have non-zero weight
- Highly weighted diagnosis codes:

| History of Disease |
|---|
| Impaired Fasting Glucose (Code 790.21) |
| Abnormal Glucose NEC (790.29) |
| Hypertension (401) |
| Obstructive Sleep Apnea (327.23) |
| Obesity (278) |
| Abnormal Blood Chemistry (790.6) |
| Hyperlipidemia (272.4) |
| Shortness Of Breath (786.05) |
| Esophageal Reflux (530.81) |

**Additional Disease Risk Factors Include:** Pituitary dwarfism (253.3), Hepatomegaly(789.1), Chronic Hepatitis C (070.54), Hepatitis (573.3), Calcaneal Spur(726.73), Thyrotoxicosis without mention of goiter(242.90), Sinoatrial Node dysfunction(427.81), Acute frontal sinusitis (461.1 ), Hypertrophic and atrophic conditions of skin(701.9), Irregular menstruation(626.4), …

**Diabetes**
**1-year gap**

# What are the Discovered Risk Factors?

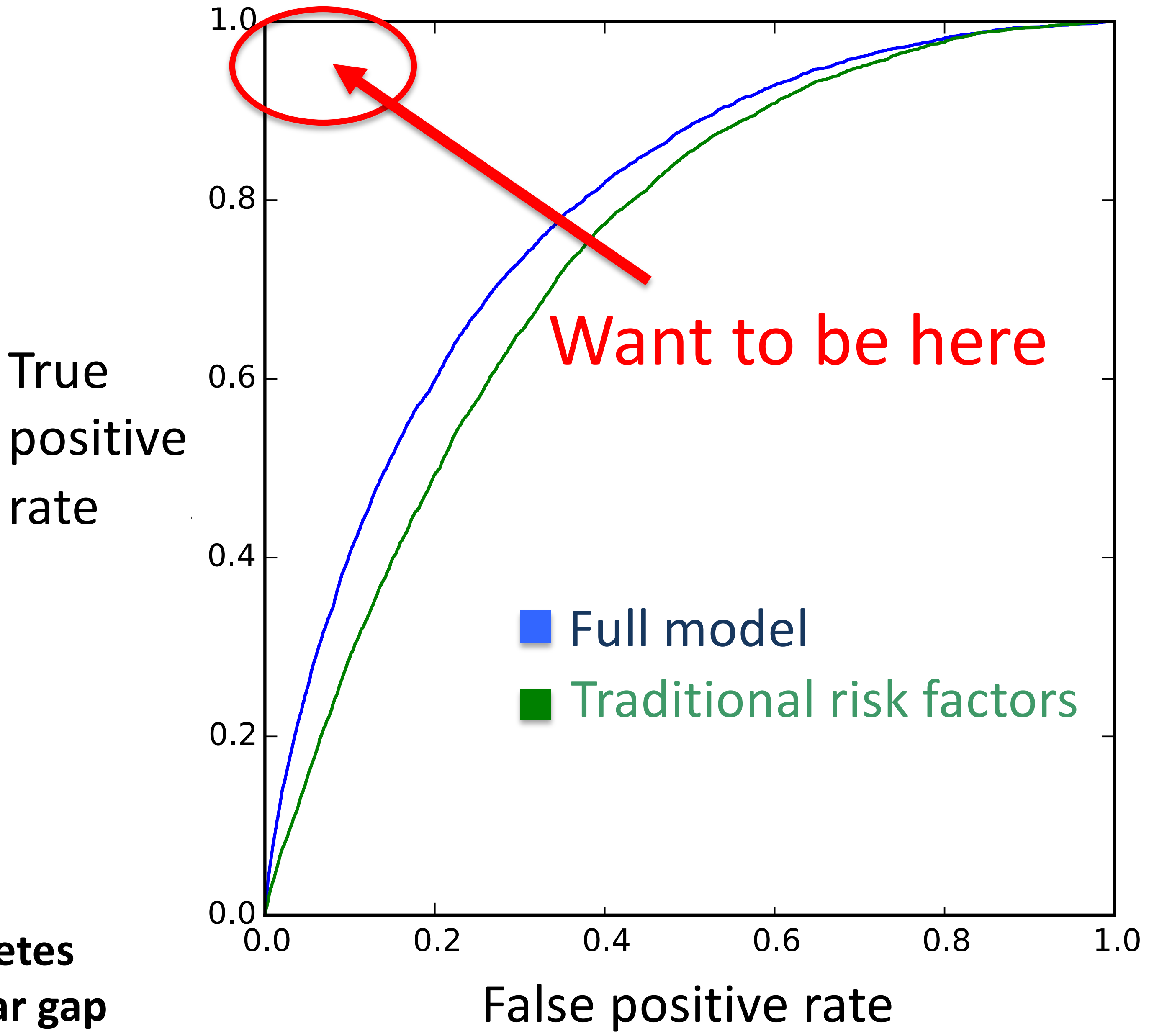
- 769 variables have non-zero weight
- Highly weighted laboratory features:

| Top Lab Factors |
|---|
| Hemoglobin A1c /Hemoglobin.Total (H |
| Glucose (High- Past 6 months) |
| Cholesterol.In VLDL (Increasing - Pas |
| Potassium (Low - Entire History) |
| Cholesterol.Total/Cholesterol.In HDL ( |
| Erythrocyte mean corpuscular hemoglobin concentration -(Low - Entire History) |
| Eosinophils (High - Entire History) |
| Glomerular filtration rate/1.73 sq M.Predicted (Low -Entire History) |
| Alanine aminotransferase (High Entire History) |

**Additional Lab Test Risk Factors Include:**
Albumin/Globulin (Increasing -Entire history), Urea nitrogen/Creatinine -(high -Entire History), Specific gravity (Increasing, Past 2 years), Bilirubin (high -Past 2 years),…

**Diabetes**
**1-year gap**

# Receiver-operator characteristic curve

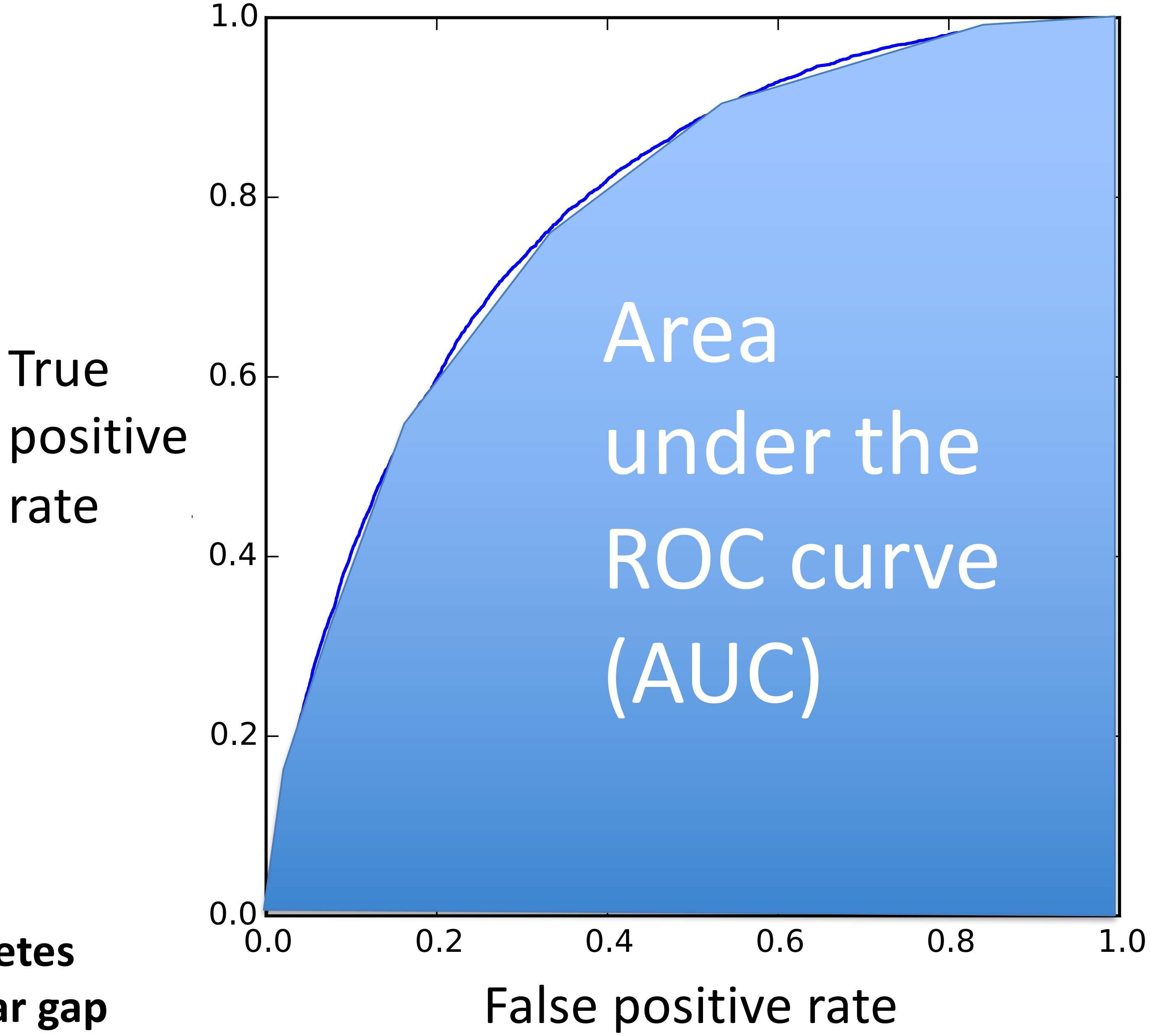


Want to be here

Obtained by varying prediction threshold

**Diabetes 1-year gap**

# Receiver-operator characteristic curve

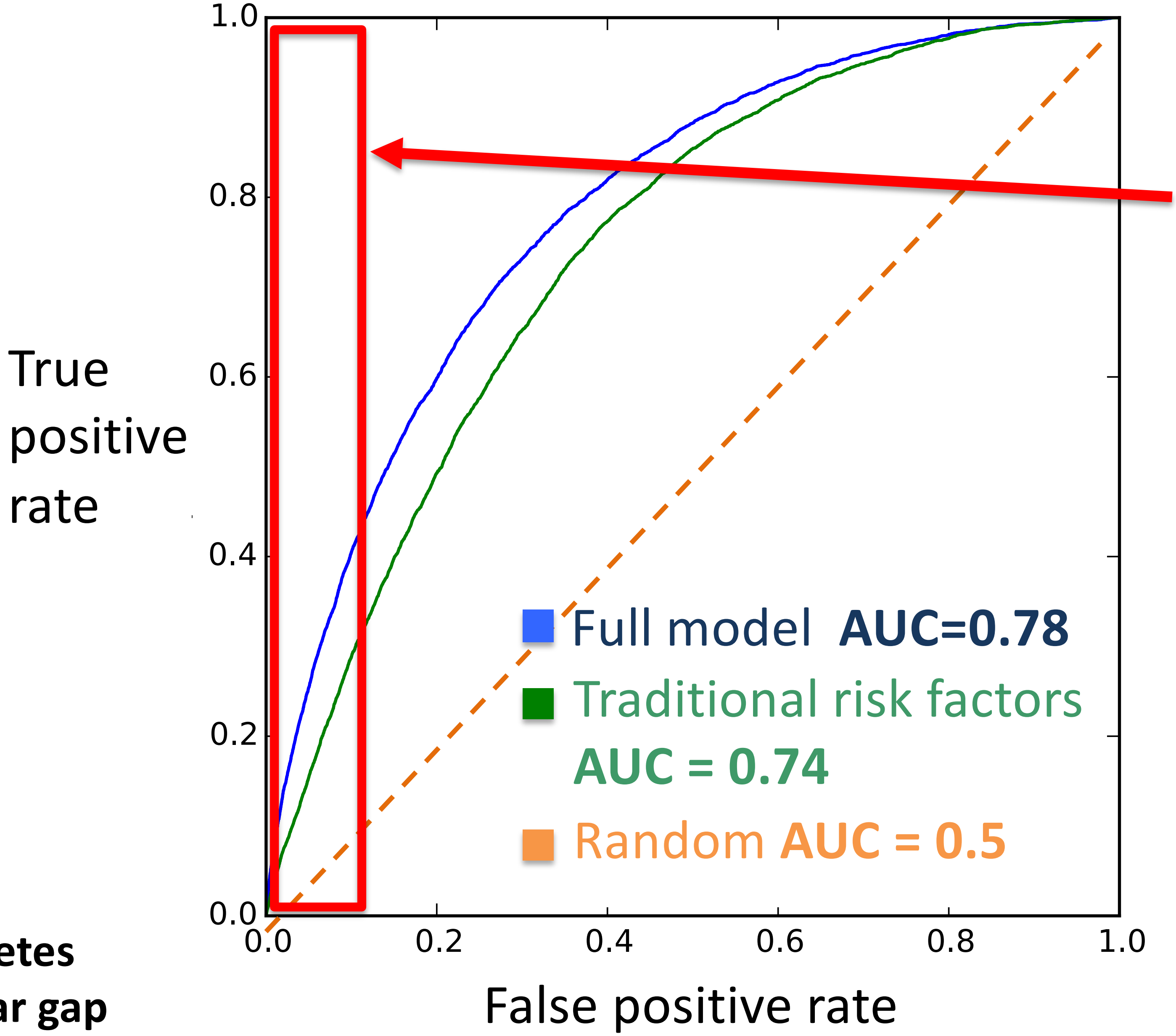


AUC = Probability that algorithm ranks a positive patient over a negative patient

Invariant to amount of class imbalance

**Diabetes 1-year gap**

# Receiver-operator characteristic curve

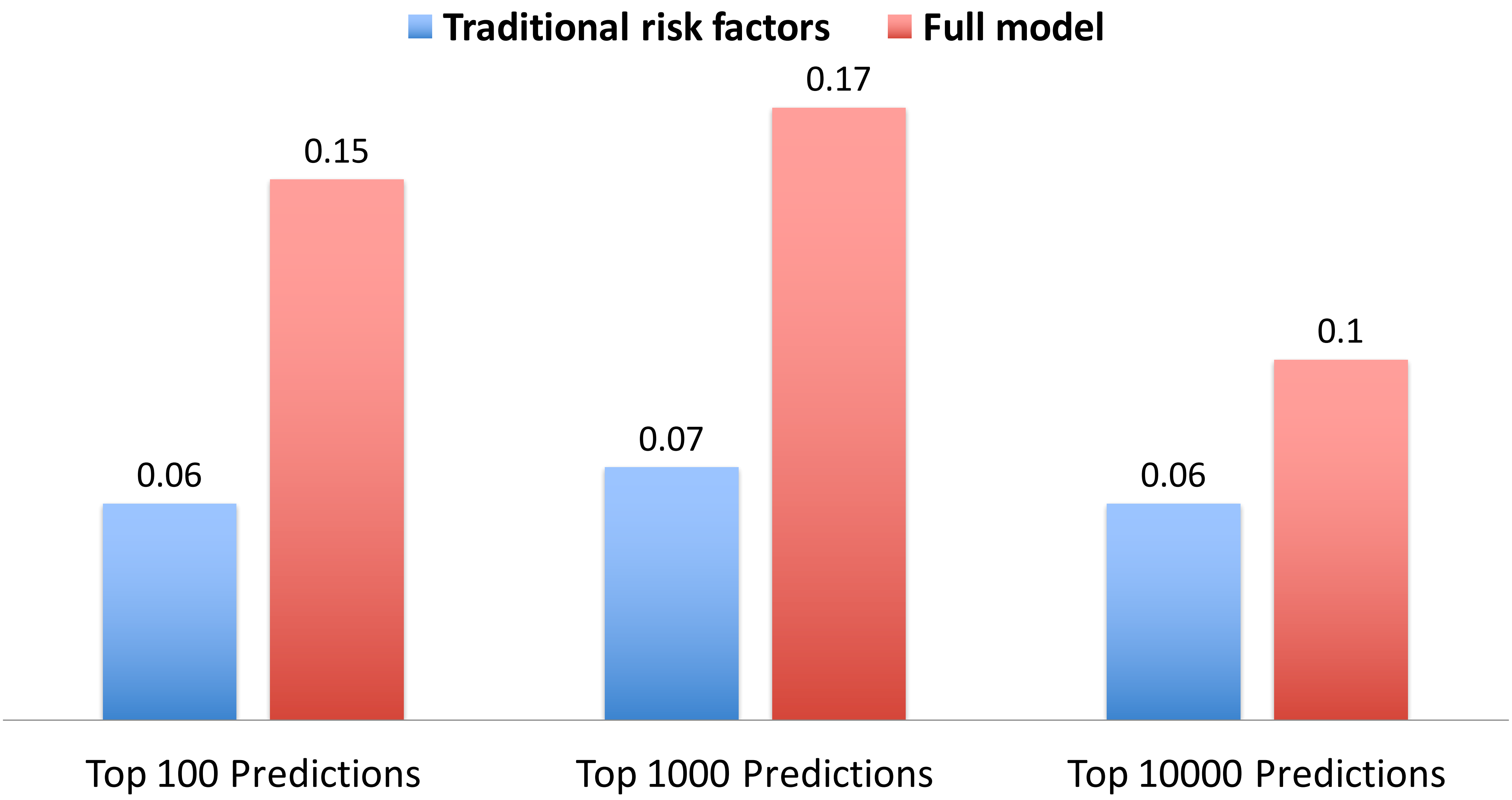


*Risk stratification* usually focuses on just this region

(because of the cost of interventions)

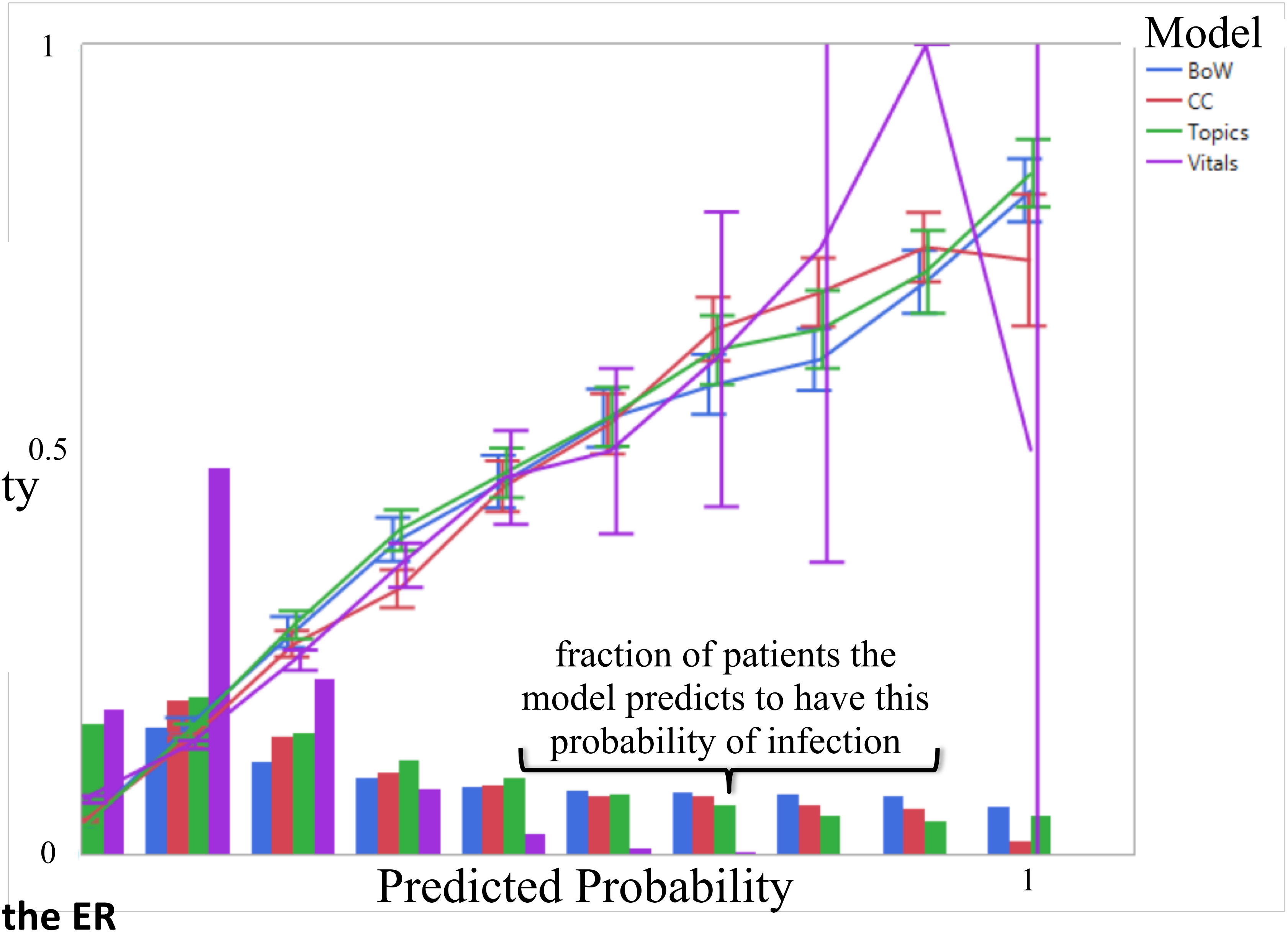
**Diabetes**
**1-year gap**

# Positive predictive value (PPV)



**Diabetes 1-year gap**

# Calibration (*note: different dataset*)

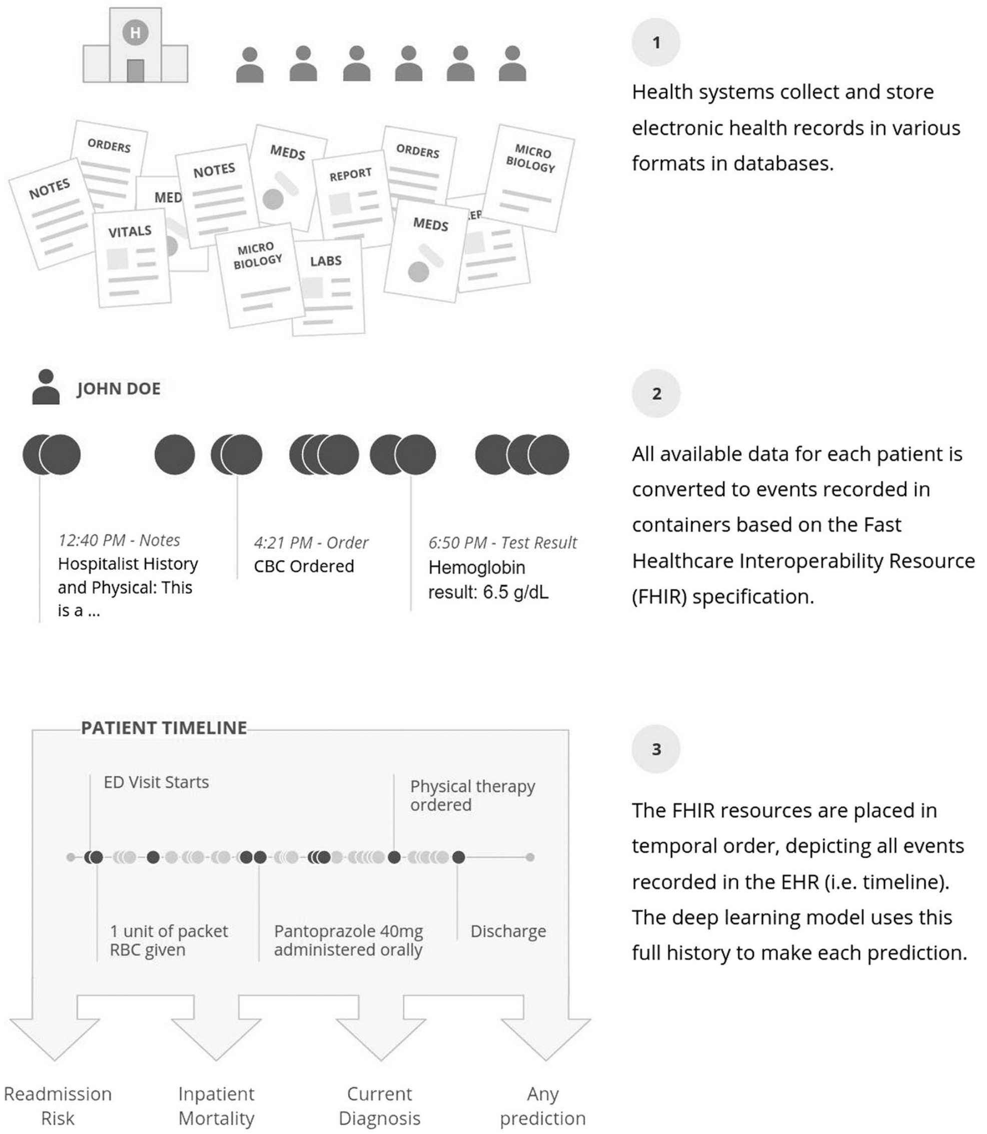


**Predicting infection in the ER**

# Outline for today’s class

1. Risk stratification
2. Case study: Early detection of Type 2 diabetes
   - Framing as supervised learning problem
   - Deriving labels
   - Evaluating risk stratification algorithms
3. **Subtleties with ML-based risk stratification**

# No big wins from deep models on structured data/text



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

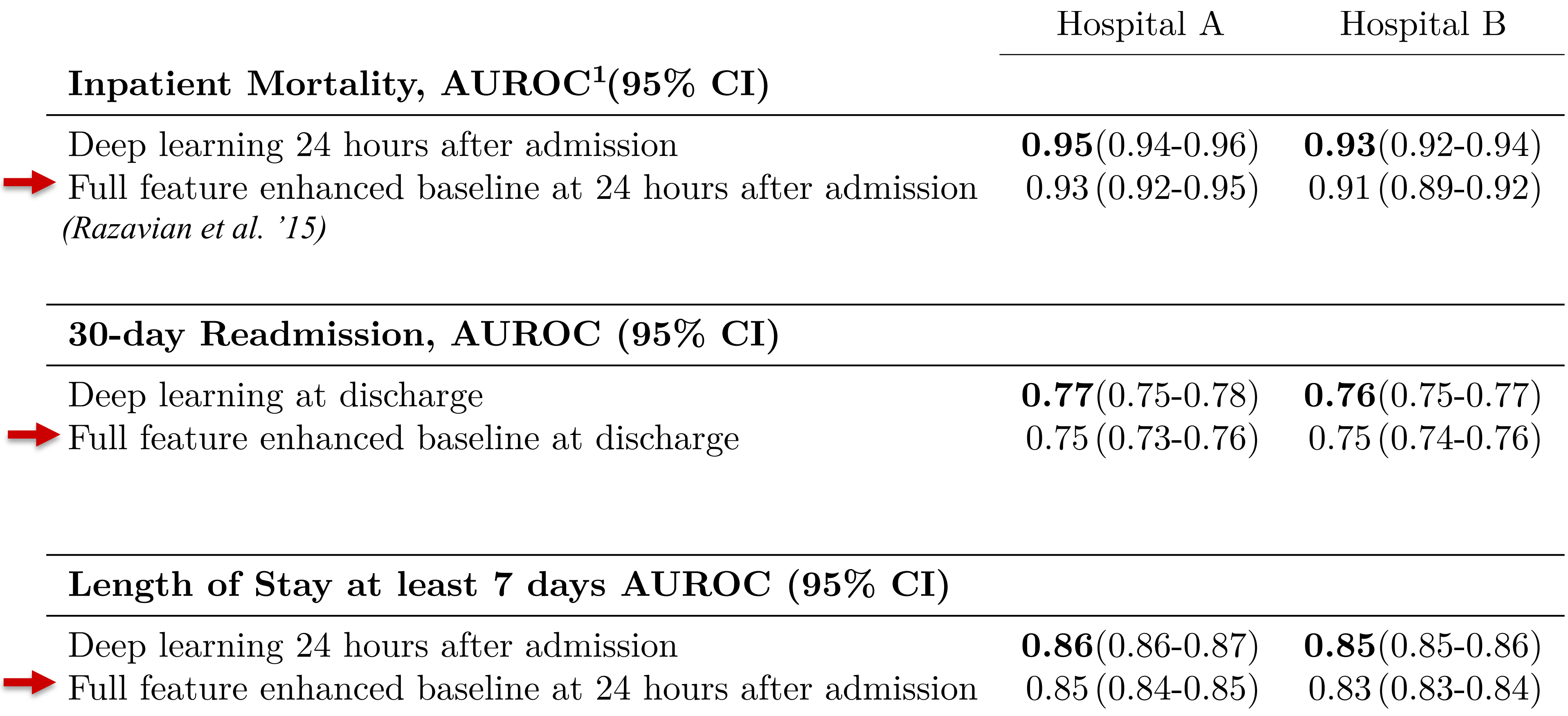
2 All available data for each patient is converted to events recorded in containers based on the Fast Healthcare Interoperability Resource (FHIR) specification.

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

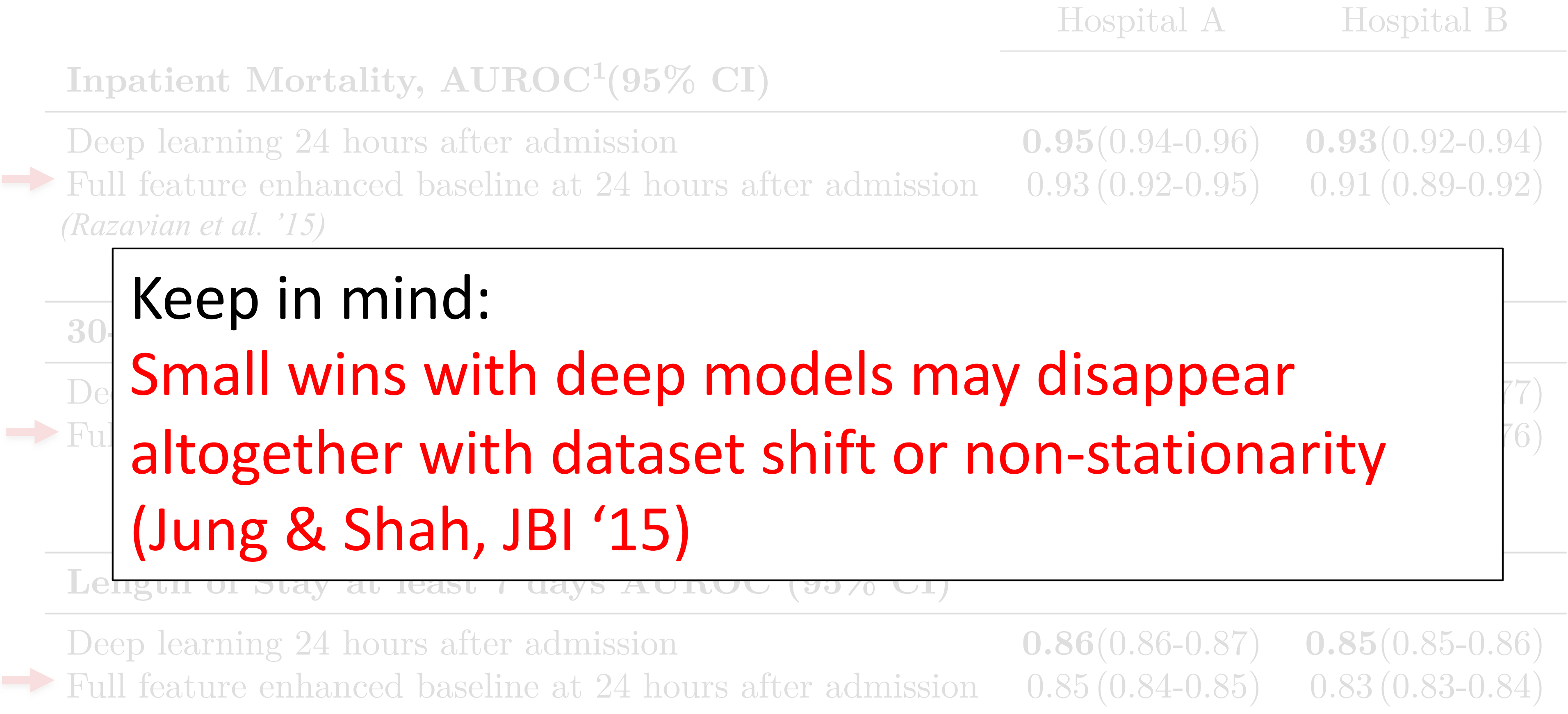
# No big wins from deep models on structured data/text



| | 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) |
| *(Razavian et al. '15)* | | |
| **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) |
| **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) |

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



| | 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) |
| *(Razavian et al. '15)* | | |
| **30** | | |
| De | | 7) |
| → Fu | | 6) |
| **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) |

Keep in mind:

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

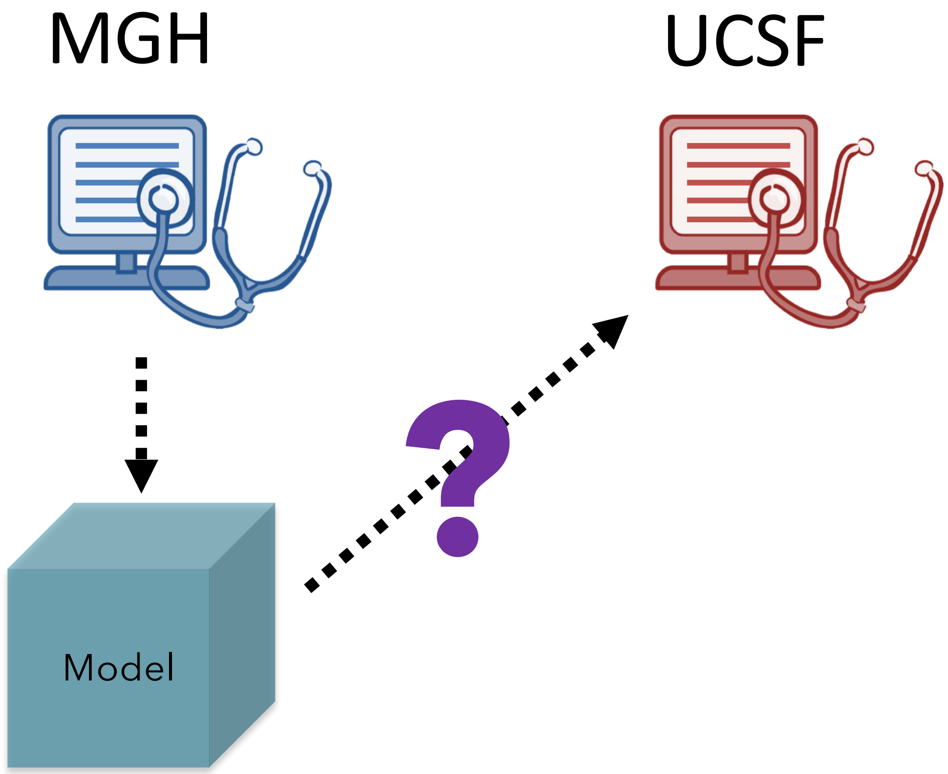
[Rajkomar et al., Scalable and accurate deep learning with electronic health records. *Nature Digital Medicine*, 2018. **electronic supplementary material**: https://static-content.springer.com/esm/art%3A10.1038%2Fs41746-018-0029-1/MediaObjects/41746_2018_29_MOESM1_ESM.pdf]

# 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

# Dataset shift / non-stationarity:

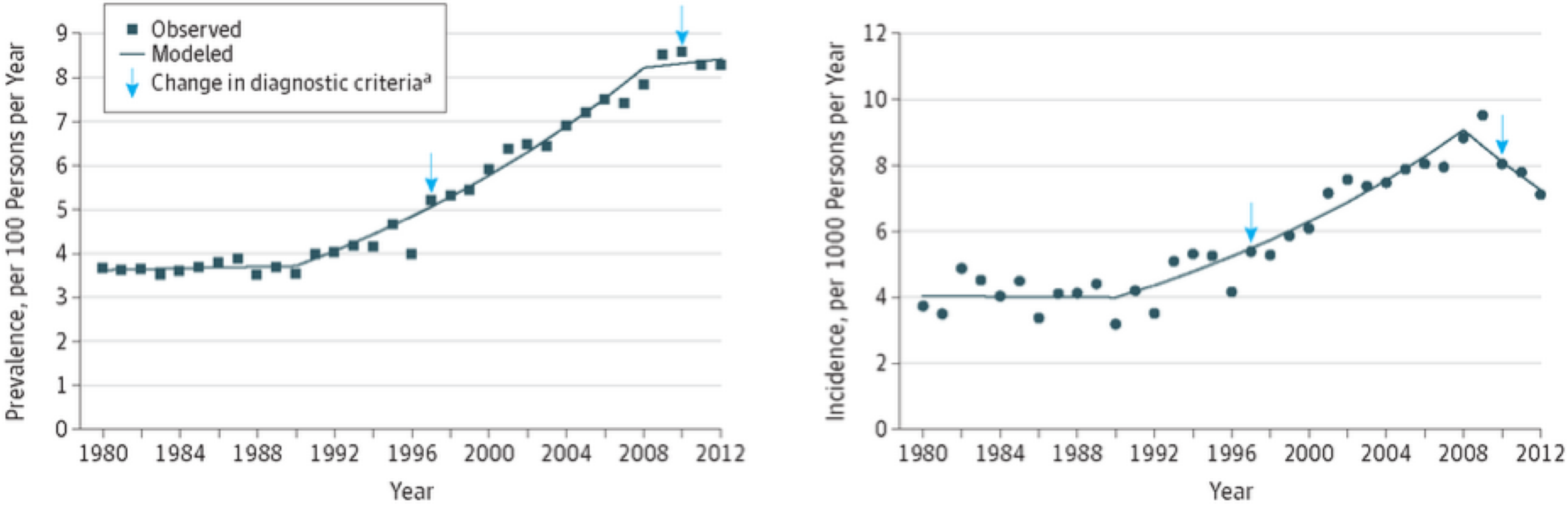
*Models often do not generalize*



[Figure adopted from Jen Gong and Tristan Naumann]

# Dataset shift / non-stationarity:

## *Diabetes Onset After 2009*

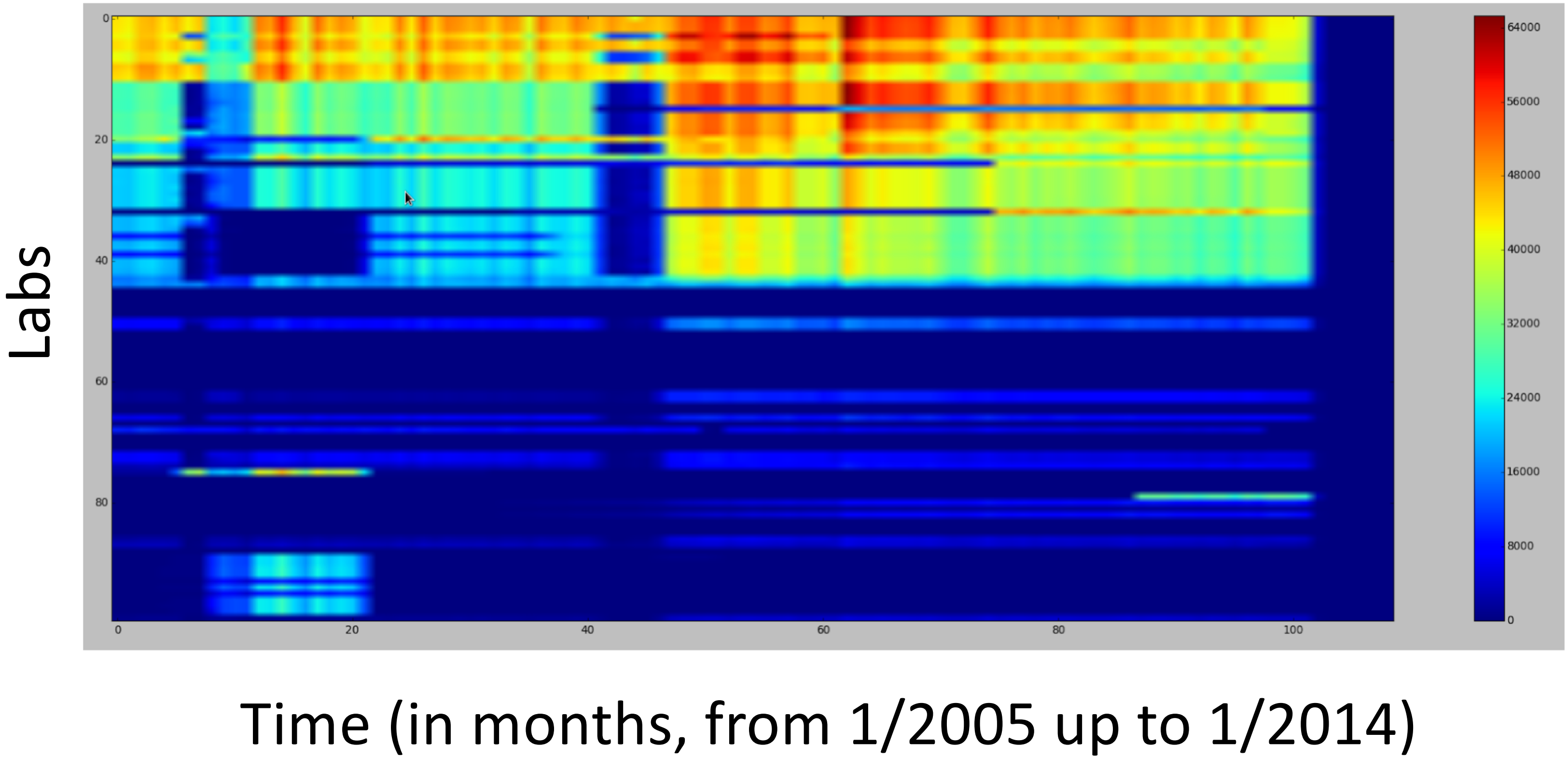


→ Automatically derived labels may change meaning

[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.]

# Dataset shift / non-stationarity:

*Top 100 lab measurements over time*

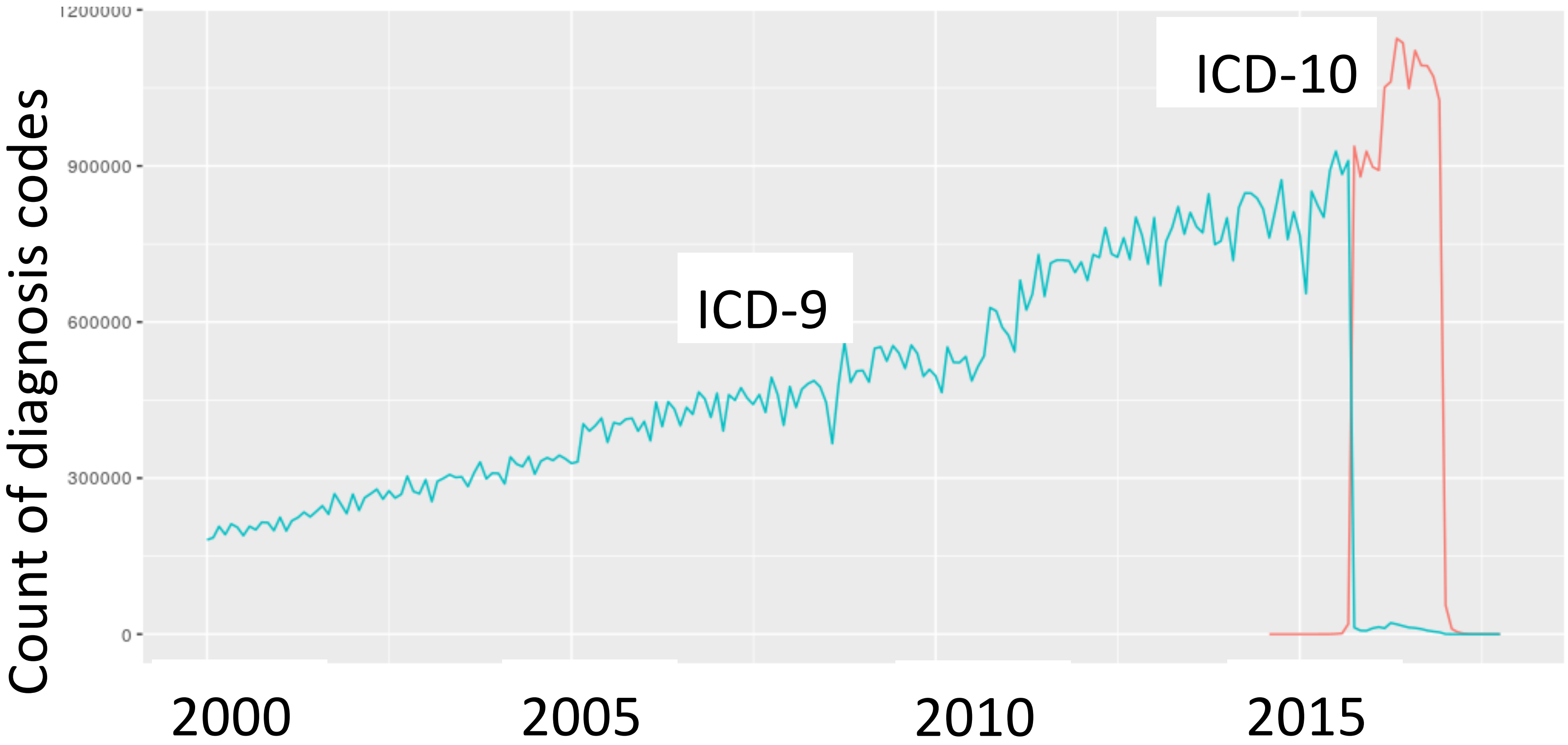


→ Significance of features may change over time

[Figure credit: Narges Razavian]

# Dataset shift / non-stationarity:

## *ICD-9 to ICD-10 shift*

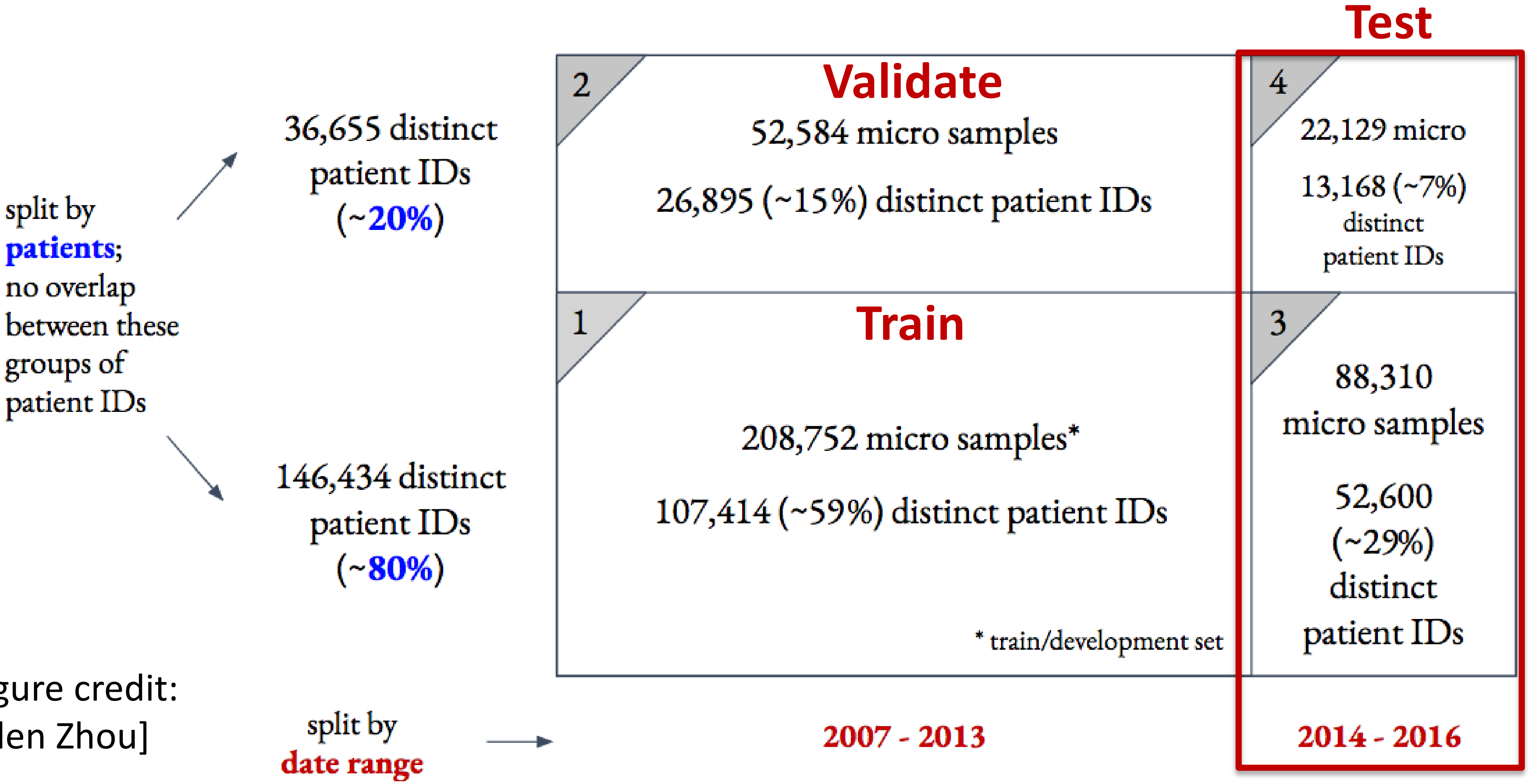


→ Significance of features may change over time

[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

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

* train/development set

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

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

split by **date range**

**2007 - 2013**

**2014 - 2016**

[Figure credit: Helen Zhou]

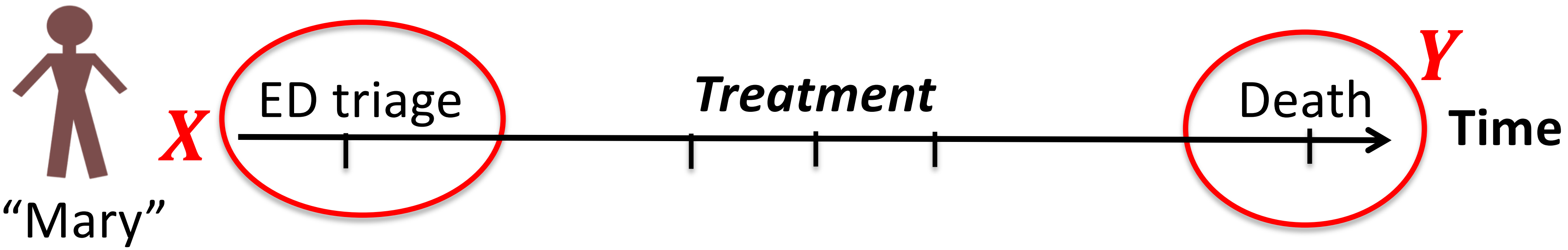
# Intervention-tainted outcomes

- Example from Caruana et al.:
  - 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

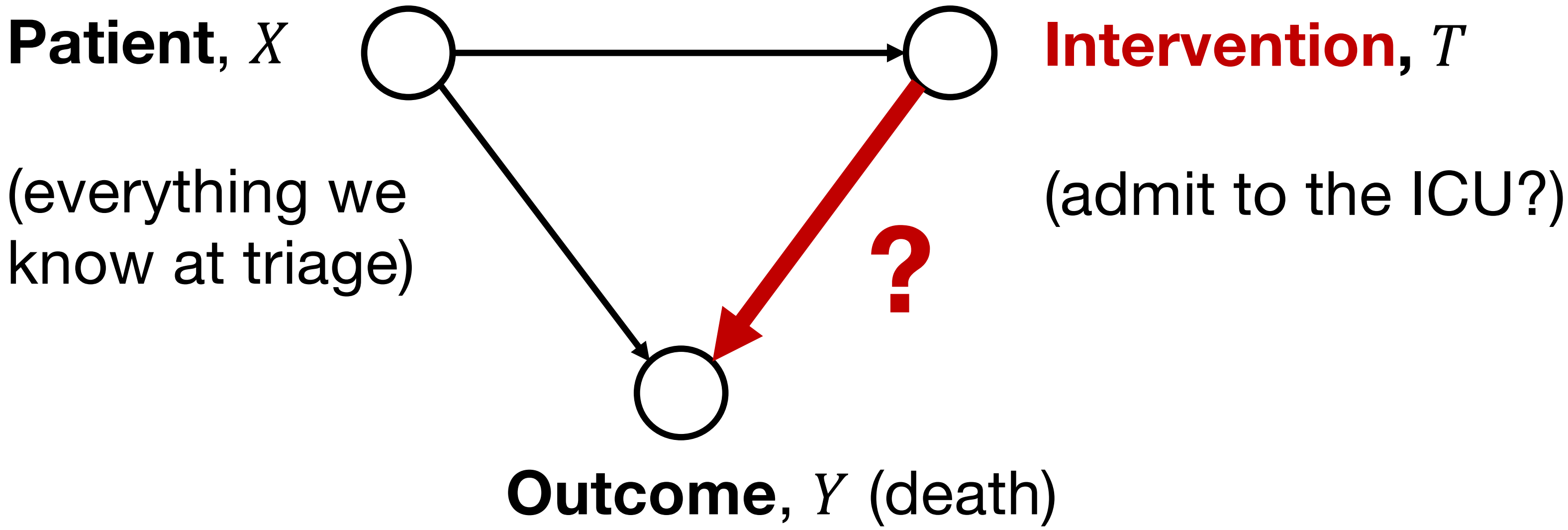
# Intervention-tainted outcomes

- 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?)

?

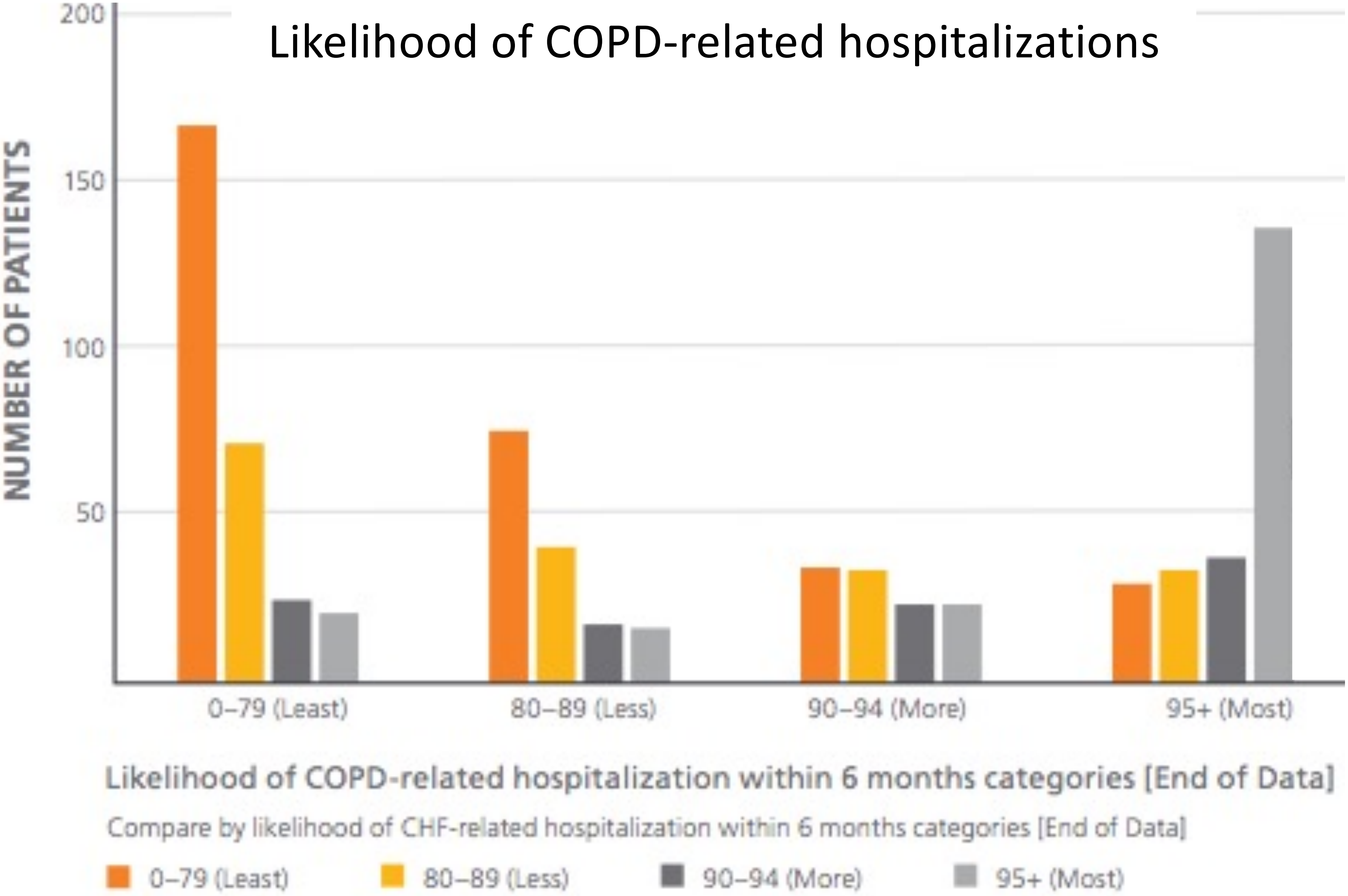
**Outcome**, $Y$ (death)

Will admission to ICU lower likelihood of death for patient?

- We return to this in Lecture 14

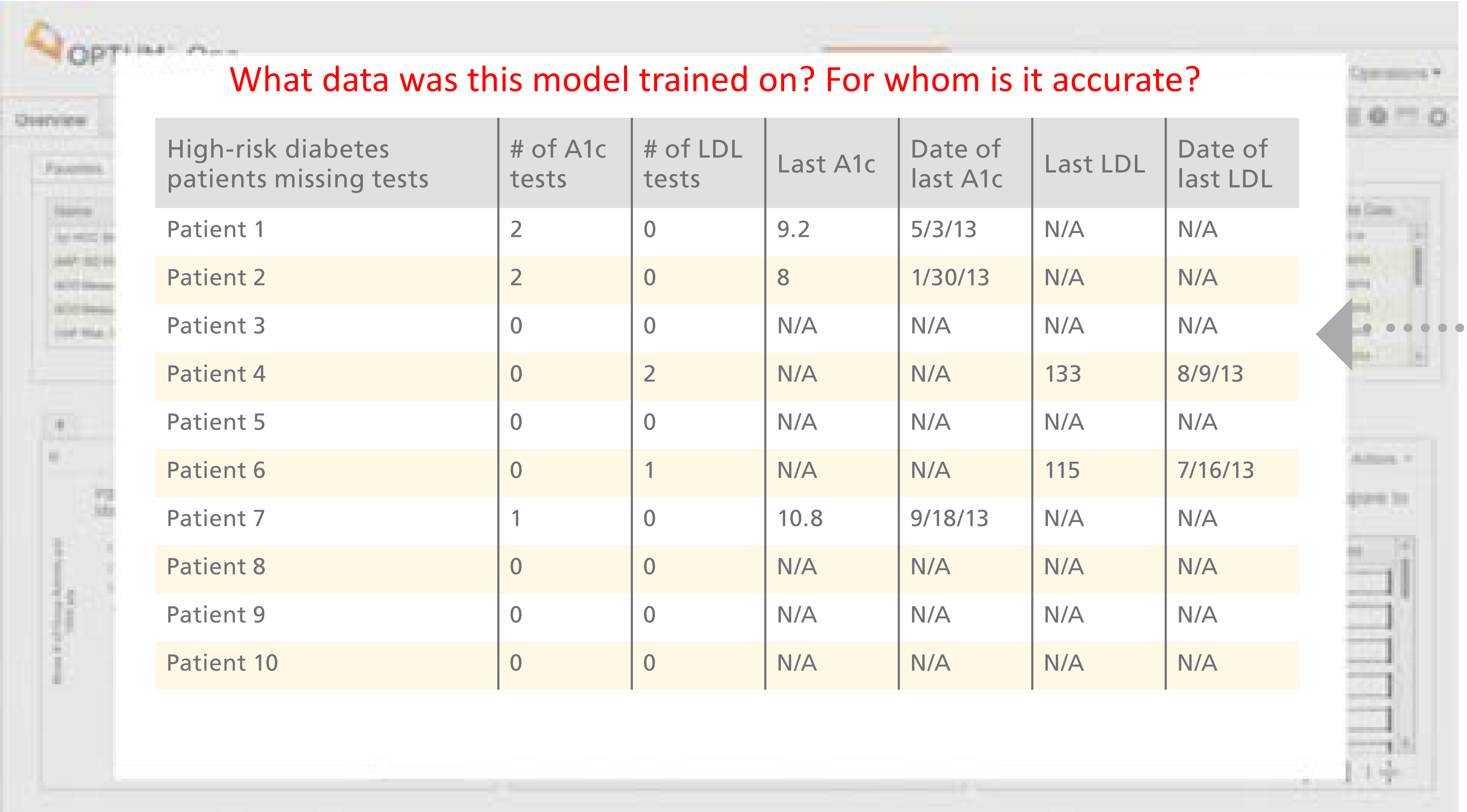
# Example commercial product

Likelihood of COPD-related hospitalizations



Optum Whitepaper, “Predictive analytics: Poised to drive population health"

# Example commercial product



What data was this model trained on? For whom is it accurate?

| High-risk diabetes patients missing tests | # of A1c tests | # of LDL tests | Last A1c | Date of last A1c | Last LDL | Date of last LDL |
|---|---|---|---|---|---|---|
| Patient 1 | 2 | 0 | 9.2 | 5/3/13 | N/A | N/A |
| Patient 2 | 2 | 0 | 8 | 1/30/13 | N/A | N/A |
| Patient 3 | 0 | 0 | N/A | N/A | N/A | N/A |
| Patient 4 | 0 | 2 | N/A | N/A | 133 | 8/9/13 |
| Patient 5 | 0 | 0 | N/A | N/A | N/A | N/A |
| Patient 6 | 0 | 1 | N/A | N/A | 115 | 7/16/13 |
| Patient 7 | 1 | 0 | 10.8 | 9/18/13 | N/A | N/A |
| Patient 8 | 0 | 0 | N/A | N/A | N/A | N/A |
| Patient 9 | 0 | 0 | N/A | N/A | N/A | N/A |
| Patient 10 | 0 | 0 | N/A | N/A | N/A | N/A |

Optum Whitepaper, “Predictive analytics: Poised to drive population health"

# Summary and next steps

- Risk stratification is being used to drive clinical decisions and resource allocation
  - *Are the models fair?*
- It can be very difficult to derive high-quality labels for supervised ML in healthcare
  - *Can one learn from noisy, biased, or censored labels?*
- Interpretability of models important for assessing whether retrospective evaluation is representative of future deployment
  - Identifying errors in label/outcome derivation
  - Assessing robustness to dataset shift
- To achieve scalability, we need ML algorithms that can detect and be robust to dataset shift
- Often the right question is not one of prediction but causal inference (counterfactual estimation)