Machine Learning for Healthcare

6.S897, HST.S53

Lecture 2: Risk stratification

Prof. David Sontag

MIT EECS, CSAIL, IMES

(Thanks to Narges Razavian for some of the slides)

Massachusetts
Institute of
Technology

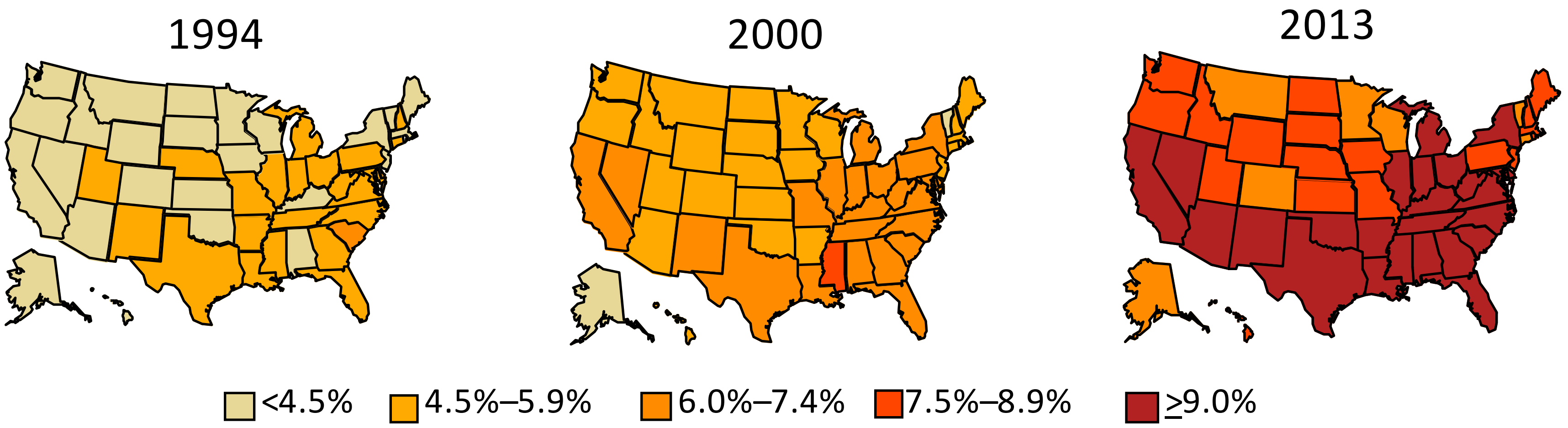
Outline for today’s class

1. Case study for risk stratification: Early detection of Type 2 diabetes
2. Framing as supervised learning problem
3. Deriving labels
4. Evaluating risk stratification algorithms
5. Non-stationarity

Outline for today’s class

1. **Case study for risk stratification: Early detection of Type 2 diabetes**
2. Framing as supervised learning problem
3. Deriving labels
4. Evaluating risk stratification algorithms
5. Non-stationarity

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.

Type 2 Diabetes Can Be Prevented *

Requirement for successful large scale prevention program

1. Detect/reach truly at risk population

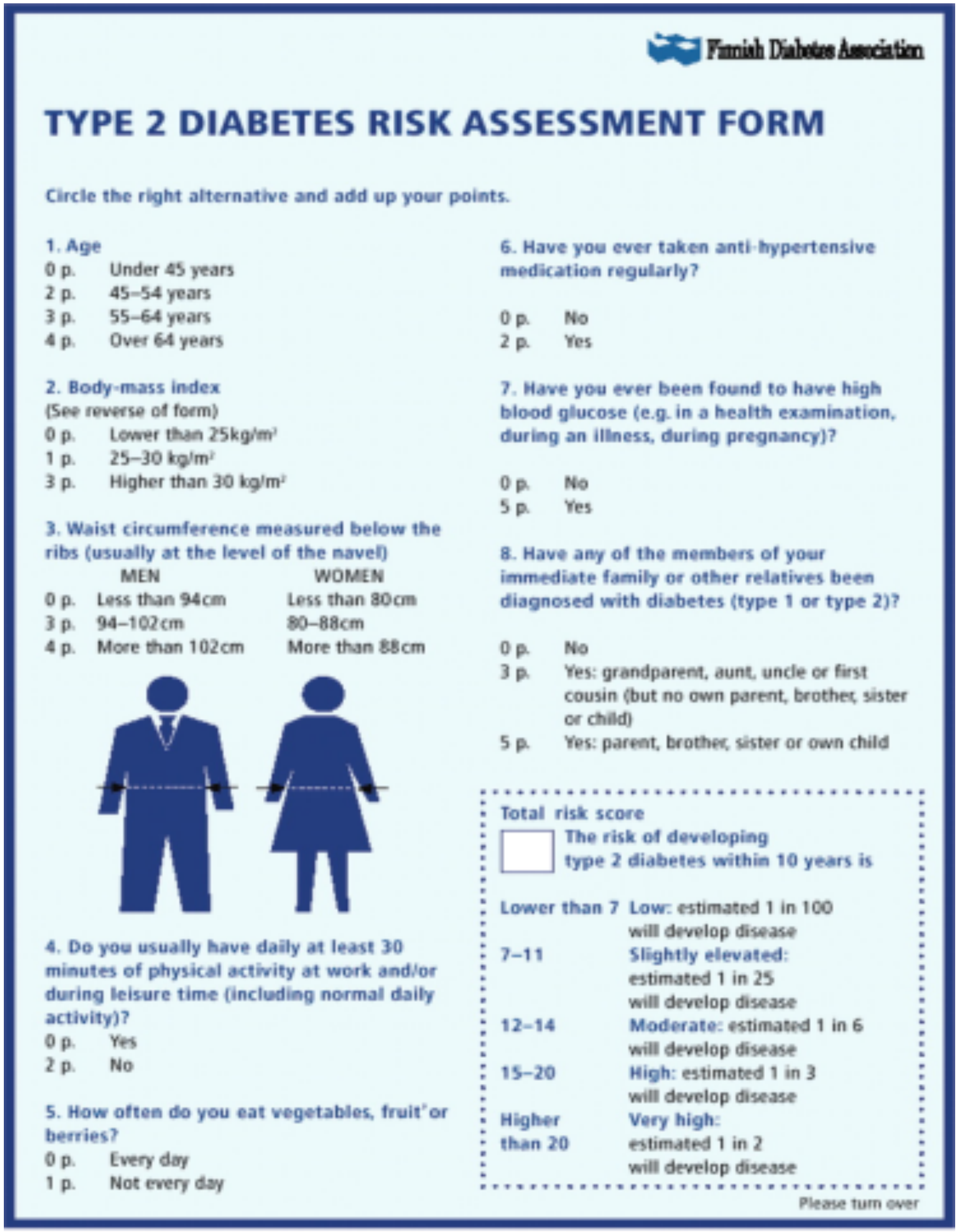
2. Improve the interventions

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/m²
1 p.	25–30 kg/m²
3 p.	Higher than 30 kg/m²

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

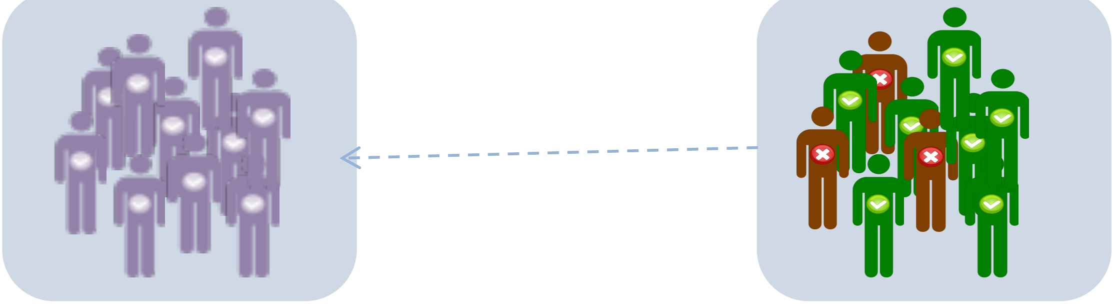
Population-Level Risk Stratification

- Key idea: Use readily available administrative, utilization, and clinical data
- Machine learning will find surrogates for risk factors that would otherwise be missing
- Perform risk stratification at the population level – millions of patients

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

A Data-Driven approach on Longitudinal Data

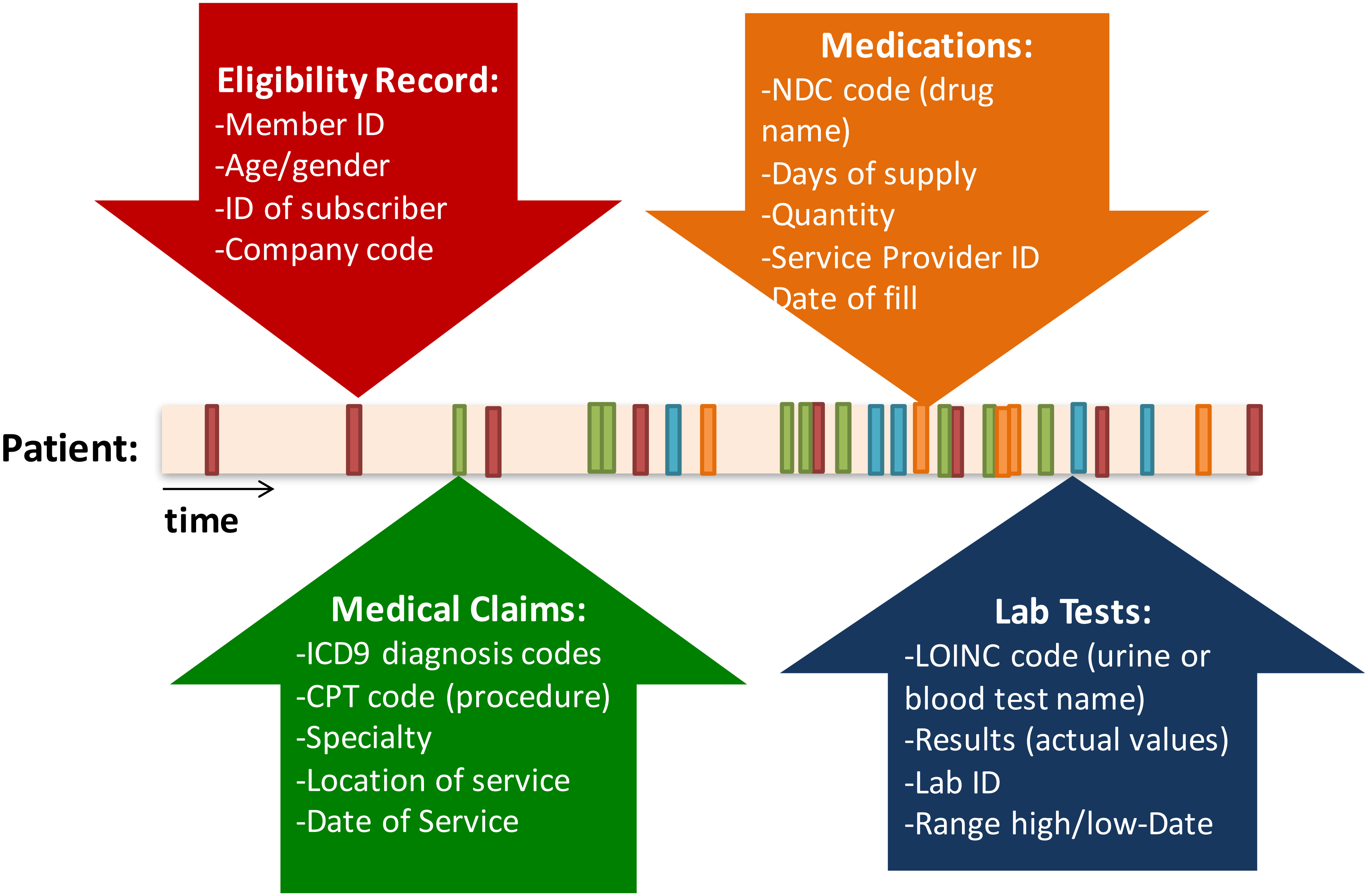
- Looking at individuals who got diabetes *today,* (compared to those who didn't)
  - Can we infer which variables in their record could have predicted their health outcome?



A Few Years Ago

Today

Reminder: 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

Patient:

time

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
4011 Benign hypertension	447017
2724 Hyperlipidemia NEC/NOS	382030
4019 Hypertension NOS	372477
25000 DMII wo cmp nt st uncntr	339522
2720 Pure hypercholesterolem	232671
2722 Mixed hyperlipidemia	180015
V7231 Routine gyn examination	178709
2449 Hypothyroidism NOS	169829
78079 Malaise and fatigue NEC	149797
V0481 Vaccin for influenza	147858
7242 Lumbago	137345
V7612 Screen mammogram NEC	129445
V700 Routine medical exam	127848

Disease	count
53081 Esophageal reflux	121064
42731 Atrial fibrillation	113798
7295 Pain in limb	112449
41401 Crnry athrscl natve vssl	104478
2859 Anemia NOS	103351
78650 Chest pain NOS	91999
5990 Urin tract infection NOS	87982
V5869 Long-term use meds NEC	85544
496 Chr airway obstruct NEC	78585
4779 Allergic rhinitis NOS	77963
41400 Cor ath unsp vsl ntv/gft	75519

Disease	count
71947 Joint pain-ankle	28648
3004 Dysthymic disorder	28530
2689 Vitamin D deficiency NOS	28455
V7281 Preop cardiovsclr exam	27897
7243 Sciatica	27604
78791 Diarrhea	27424
V221 Supervis oth normal preg	27320
36501 Opn angl brderln lo risk	26033
37921 Vitreous degeneration	25592
4241 Aortic valve disorder	25425
61610 Vaginitis NOS	24736
70219 Other sborheic keratosis	24453
3804 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

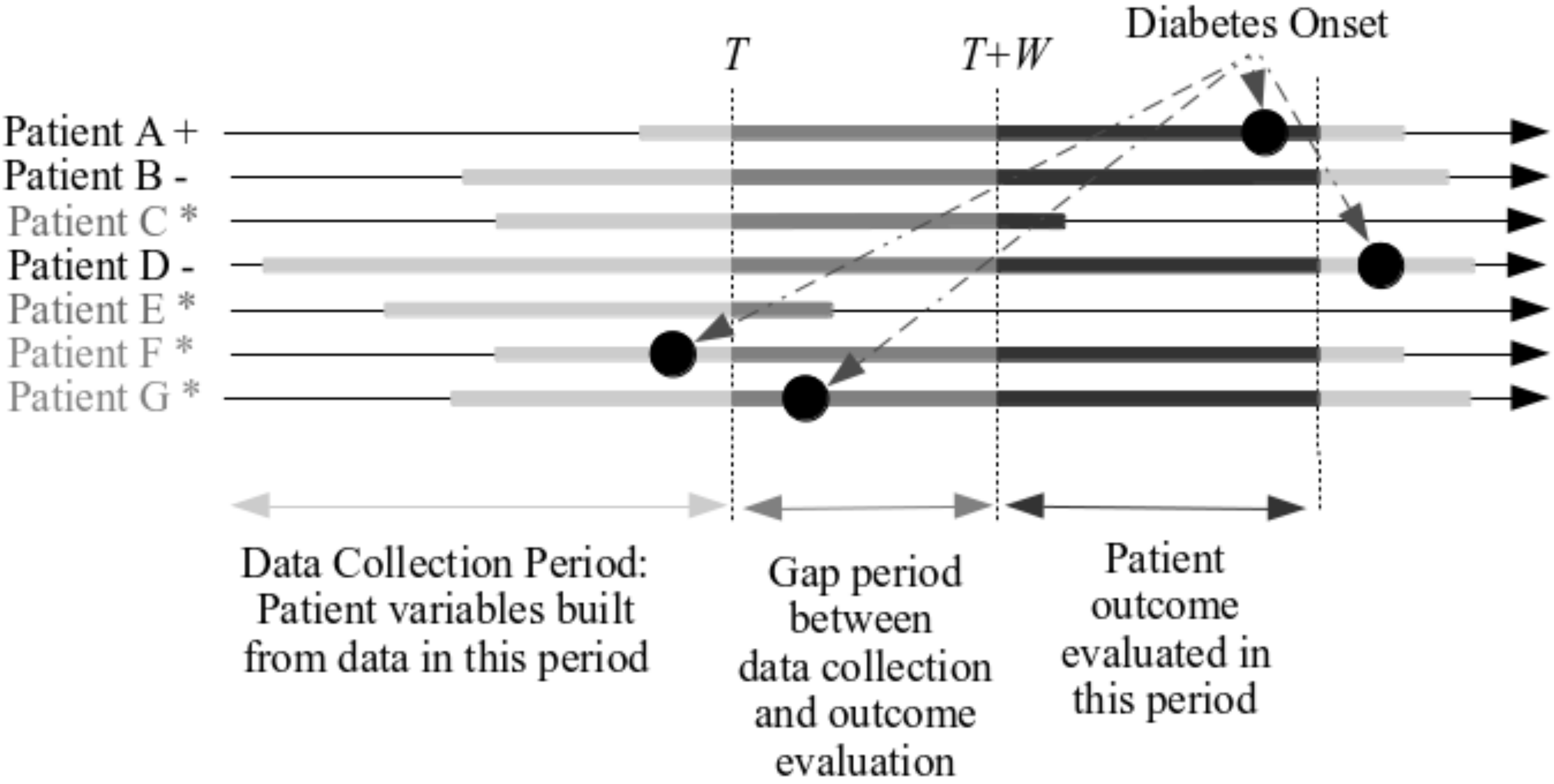
Count of people who have the test result (ever)

Outline for today’s class

1. Case study for risk stratification: Early detection of Type 2 diabetes
2. **Framing as supervised learning problem**
3. Deriving labels
4. Evaluating risk stratification algorithms
5. Non-stationarity

Framing for supervised machine learning

Align by absolute time


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

Gap is important to prevent label leakage

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

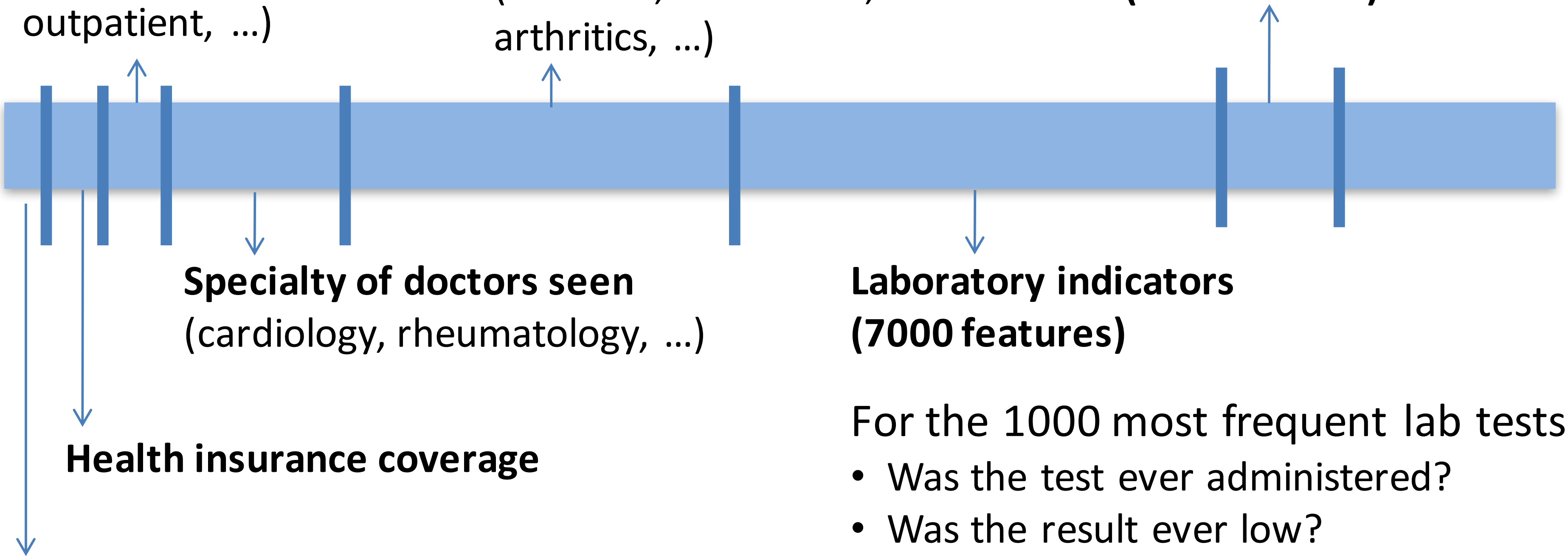
NOTE:

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

Methods

- L1 Regularized Logistic Regression
  - Simultaneously optimizes predictive performance *and*
  - Performs feature selection, choosing the subset of the features that are most predictive
- This prevents overfitting to the training data

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)

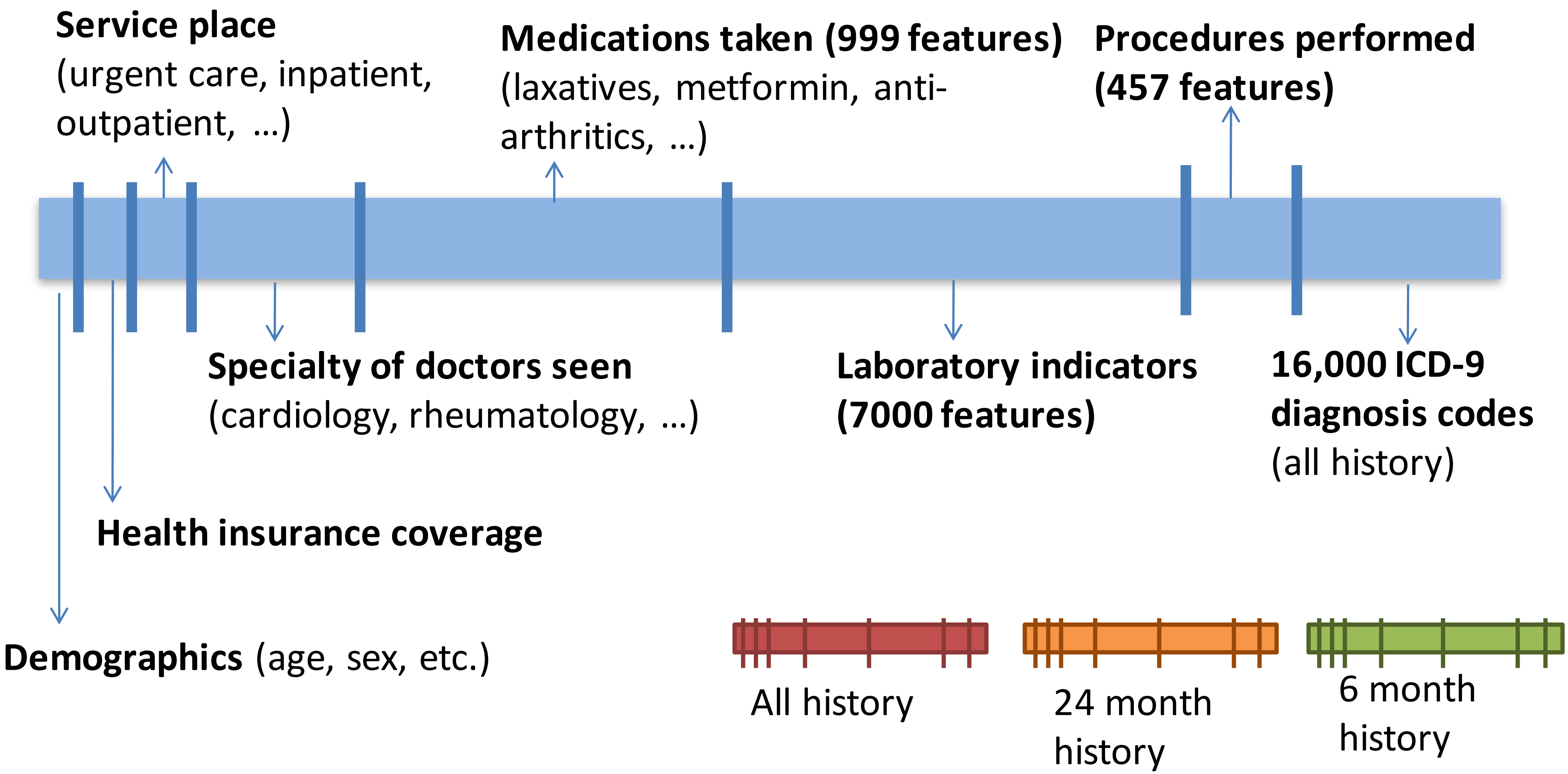
Health insurance coverage

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

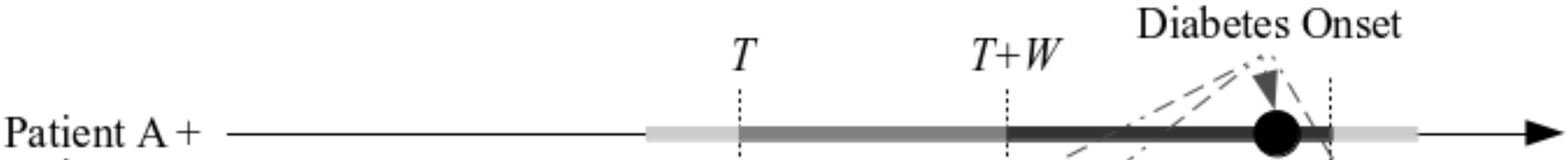
6 month history

Total features per patient: 42,000

Outline for today's class

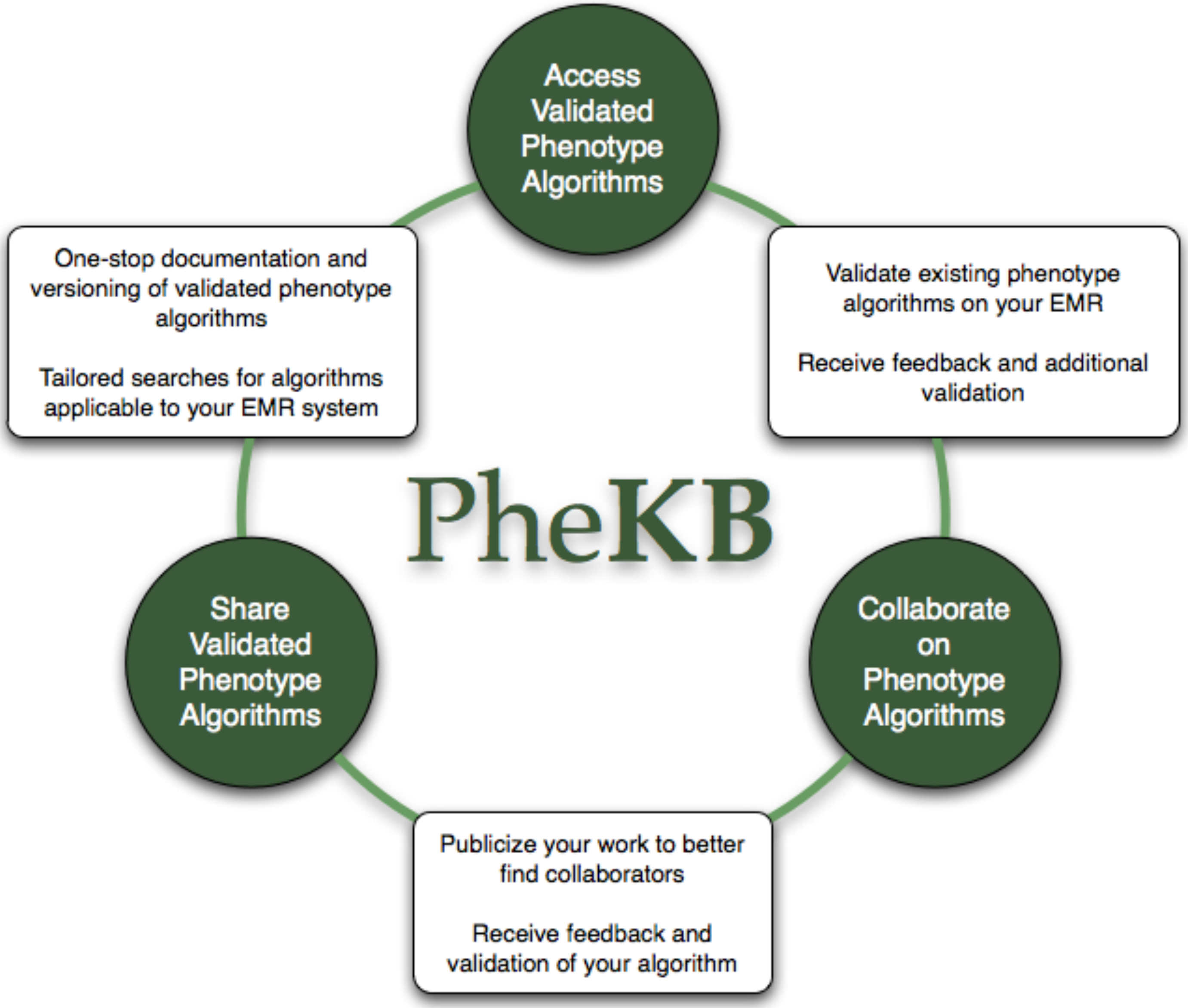
1. Case study for risk stratification: Early detection of Type 2 diabetes
2. Framing as supervised learning problem
3. **Deriving labels**
4. Evaluating risk stratification algorithms
5. Non-stationarity

Where do the labels come from?



1. Manually label data by chart review
2. Electronic phenotyping from medical records
3. Use machine learning to get the labels themselves

Electronic phenotyping



Access Validated Phenotype Algorithms

Validate existing phenotype algorithms on your EMR

Receive feedback and additional validation

Collaborate on Phenotype Algorithms

Publicize your work to better find collaborators

Receive feedback and validation of your algorithm

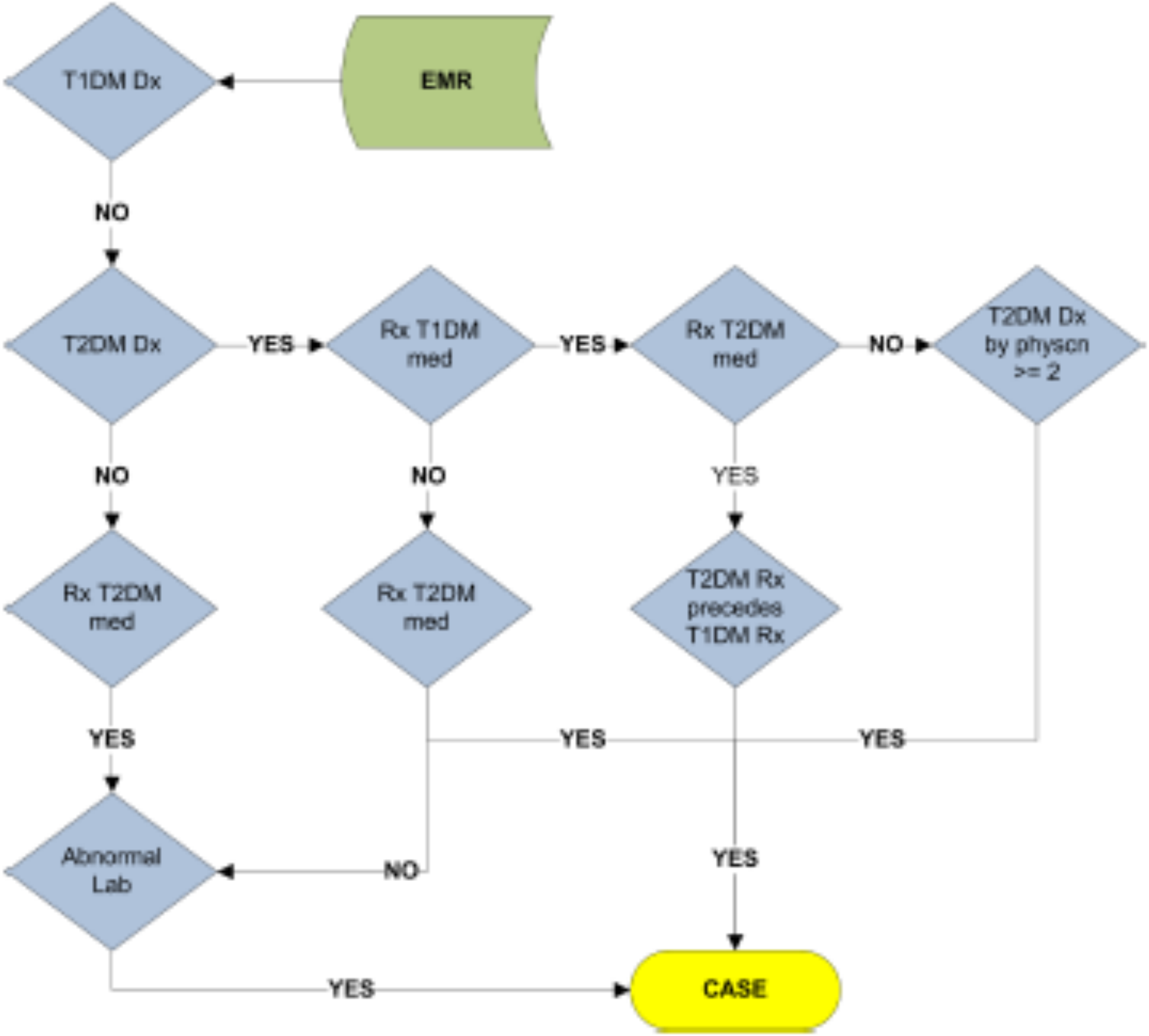
Share Validated Phenotype Algorithms

One-stop documentation and versioning of validated phenotype algorithms

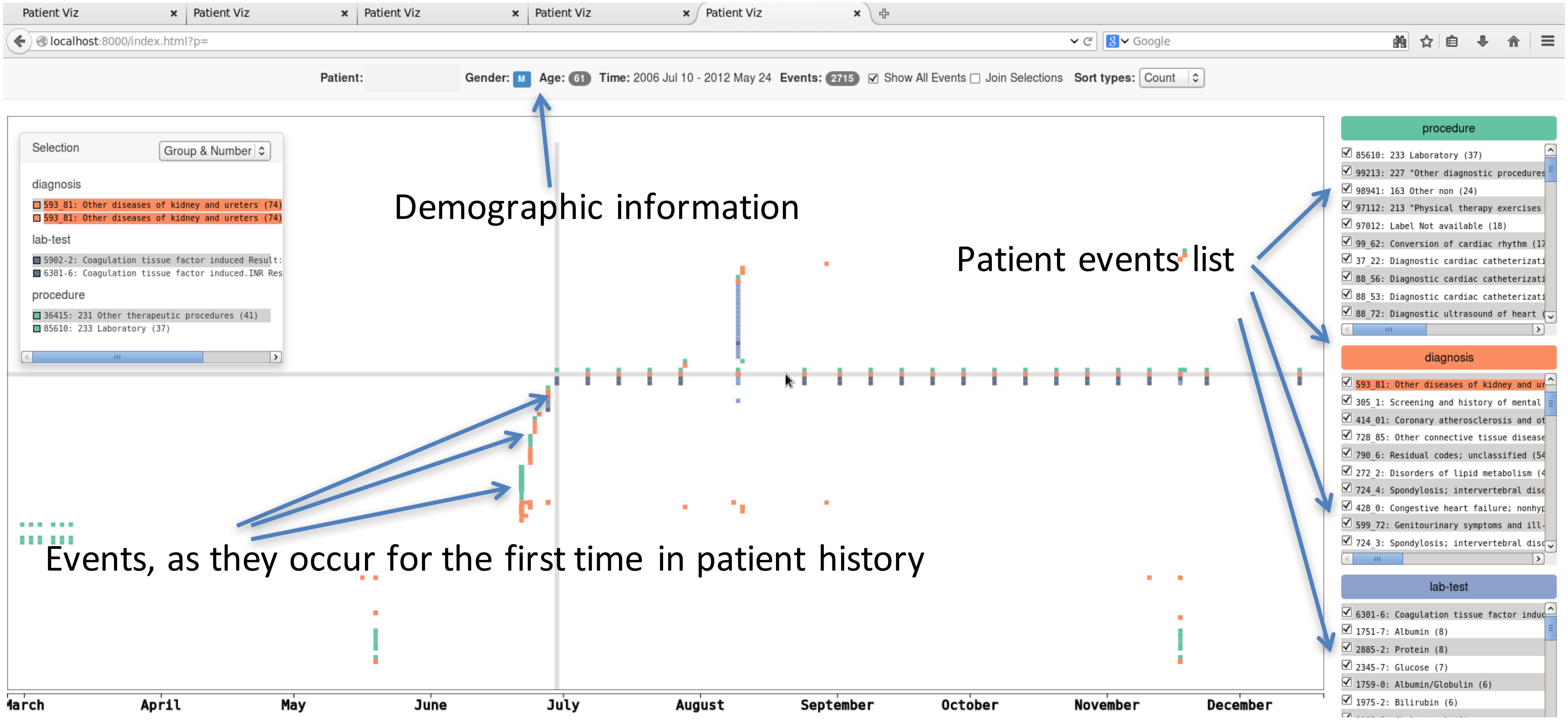
Tailored searches for algorithms applicable to your EMR system

Electronic phenotyping

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



Visualization (looking at individual patients) is important to sanity check labeling method



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

Getting the labels using the Anchor & Learn Framework

- Use a combination of domain expertise (simple rules) and vast amounts of data (machine learning)
- Method does not require any manual labeling
- Anchors are highly transferable between institutions

[Halpern et al., AMIA 2014]

What are anchors?

- Rather than provide gold-standard labels, construct a simple rule that can catch some positive cases.
- Examples:

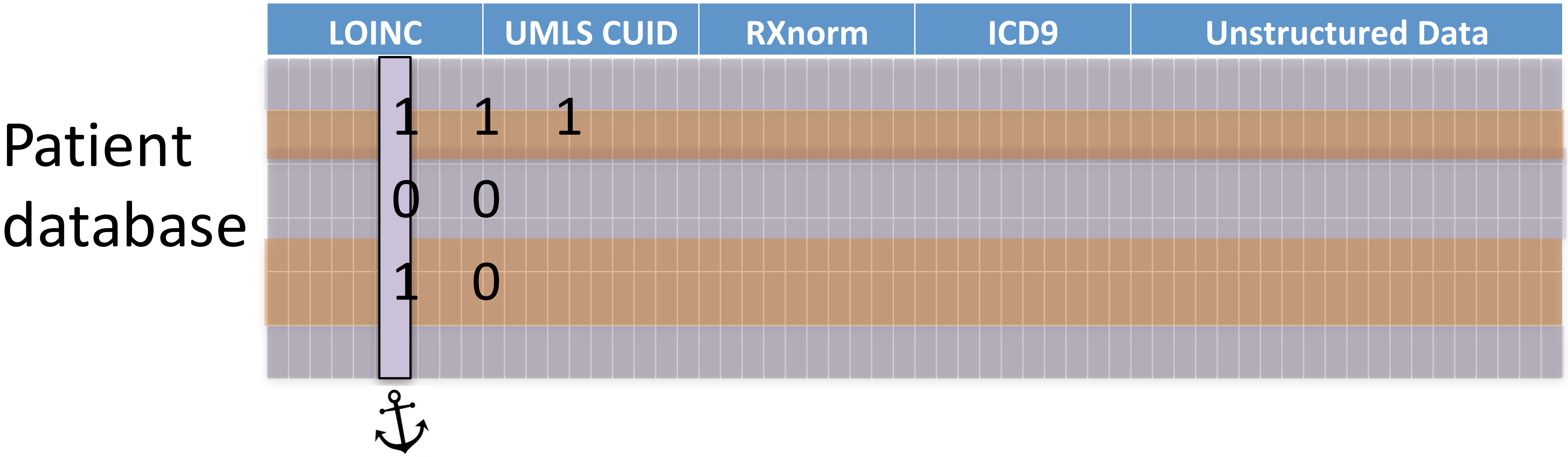
Clin. state var	Possible Anchor
Diabetic	gsn:016313 (insulin) in Medications
Cardiac	ICD9:428.X (heart failure) in Diagnoses
Nursing home	“from nursing home” in text
Social work	“social work consulted” in text

What are anchors?

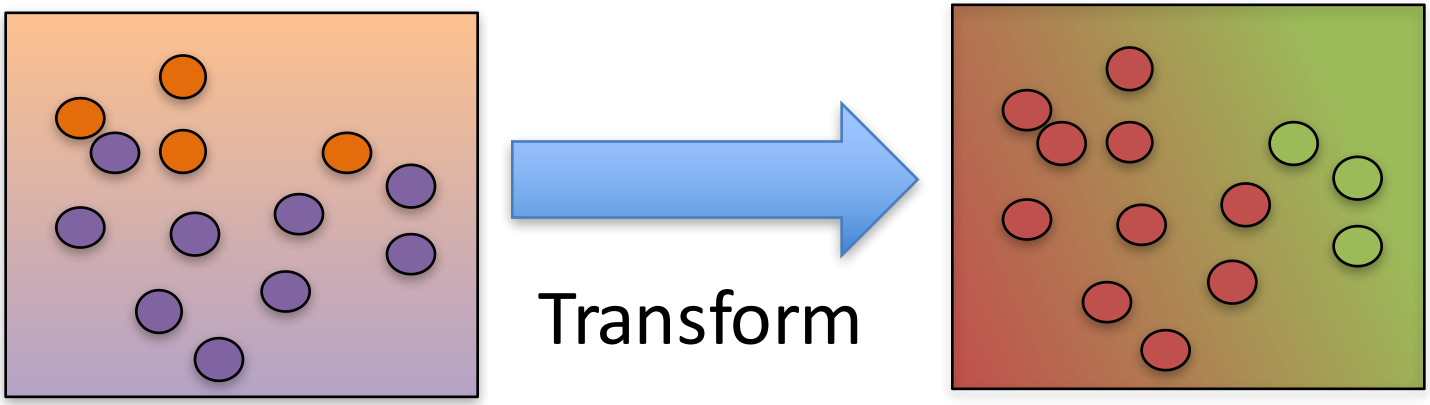
- Rather than provide gold-standard labels, construct a simple rule that can catch some positive cases. **Low sensitivity here is ok!**
- Examples:

Clin. state var	Possible Anchor
Diabetic	gsn:016313 (insulin) in Medications
Cardiac	ICD9:428.X (heart failure) in Diagnoses
Nursing home	"from nursing home" in text
Social work	"social work consulted" in text

Learning with Anchors



- Identify anchors
- Learn to predict the anchors (anchor as pseudo-labels)
- Account for the difference between anchors and labels



Predict anchor

Predict label

Theoretical basis for anchors

- Unobserved variable: Y, Observation: A
- *A* is an **anchor** for *Y* if conditioning on *A*=1 gives uniform samples from the set of *positive cases*.

Theoretical basis for anchors

- Unobserved variable: Y, Observation: A
- *A* is an **anchor** for *Y* if conditioning on *A*=1 gives uniform samples from the set of *positive cases*.
- Alternative formulation – two necessary conditions:

$$P(Y = 1|A = 1) = 1 \quad \text{AND} \quad A \perp \mathcal{X}|Y$$

Positive condition **Conditional independence**

$\mathcal{X}$ represents all *other* observations.

Theoretical basis for anchors

- Unobserved variable: Y, Observation: A
- *A* is an **anchor** for *Y* if conditioning on *A*=1 gives uniform samples from the set of *positive cases*.
- Alternative formulation – two necessary conditions:

$$P(Y = 1 | A = 1) = 1 \quad \text{AND} \quad A \perp \mathcal{X} | Y$$

Positive condition

e.g. If patient is taking *insulin*, the patient is surely **diabetic**.

Conditional independence

e.g. If we know the patient had **heart failure**, knowing whether the *diagnosis code* appears does not inform us about the rest of the record.

$\mathcal{X}$ repre s.

Theoretical basis for anchors

- Unobserved variable: Y, Observation: A
- *A* is an **anchor** for *Y* if conditioning on *A*=1 gives uniform samples from the set of *positive cases*.
- Theorem [Elkan & Noto 2008]:

In the above setting, a function to predict A

can be transformed to predict Y

- Can also use more recent advances on ***learning with noisy labels*** (e.g., Natarajan et al., NIPS '13)

Learning with anchors

[Elkan & Noto 2008]

Input: anchor A

unlabeled patients

Output: prediction rule

1. Learn a calibrated classifier (e.g. logistic regression) to predict:

$$\Pr(A = 1 \mid \mathcal{X})$$

2. Using a validate set, let *P* be the patients with A=1. Compute:

$$C = \frac{1}{|\mathcal{P}|} \sum_{k \in \mathcal{P}} \Pr(A = 1 \mid \mathcal{X}^{(k)})$$

3. For a previously unseen patient *t*, predict:

$$\begin{aligned} \frac{1}{C}\Pr(A = 1|\mathcal{X}^{(t)}) &\quad \text{if } A^{(t)} = 0 \\ 1 &\quad \text{if } A^{(t)} = 1 \end{aligned}$$

Learning

Learn to predict *A* from the other variables.

Calibration

C is the average model prediction for patients with anchors.

Transformation

If no anchor present, according to a scaled version of the anchor-prediction model.

Outline for today’s class

1. Case study for risk stratification: Early detection of Type 2 diabetes
2. Framing as supervised learning problem
3. Deriving labels
4. **Evaluating risk stratification algorithms**
5. Non-stationarity

What are the Discovered Risk Factors?

- 769 variables have non-zero weight

Top History of Disease	Odds Ratio
Impaired Fasting Glucose (Code 790.21)	4.17 (3.87 4.49)
Abnormal Glucose NEC (790.29)	4.07 (3.76 4.41)
Hypertension (401)	3.28 (3.17 3.39)
Obstructive Sleep Apnea (327.23)	2.98 (2.78 3.20)
Obesity (278)	2.88 (2.75 3.02)
Abnormal Blood Chemistry (790.6)	2.49 (2.36 2.62)
Hyperlipidemia (272.4)	2.45 (2.37 2.53)
Shortness Of Breath (786.05)	2.09 (1.99 2.19)
Esophageal Reflux (530.81)	1.85 (1.78 1.93)

Diabetes
1-year gap

What are the Discovered Risk Factors?

- 769 variables have non-zero weight

Top History of Diseas	
Impaired Fasting Glucose (Code	
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)	(1.99 2.19)
Esophageal Reflux (530.81)	1.85 (1.78 1.93)

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

Top Lab Factors	Odds Ratio
Hemoglobin A1c /Hemoglobin.Total (High - past 2 years)	5.75 (5.42 6.10)
Glucose (High- Past 6 months)	4.05 (3.89 4.21)
Cholesterol.In VLDL (Increasing - Past 2 years)	3.88 (3.53 4.27)
Potassium (Low - Entire History)	2.58 (2.24 2.98)
Cholesterol.Total/Cholesterol.In HDL (High - Entire History)	2.29 (2.19 2.40)
Erythrocyte mean corpuscular hemoglobin concentration -(Low - Entire History)	2.25 (1.92 2.64)
Eosinophils (High - Entire History)	2.11 (1.82 2.44)
Glomerular filtration rate/1.73 sq M.Predicted (Low -Entire History)	2.07 (1.92 2.24)
Alanine aminotransferase (High Entire History)	2.04 (1.89 2.19)

Diabetes

1-year gap

What are the Discovered Risk Factors?

- 769 variables have non-zero weight

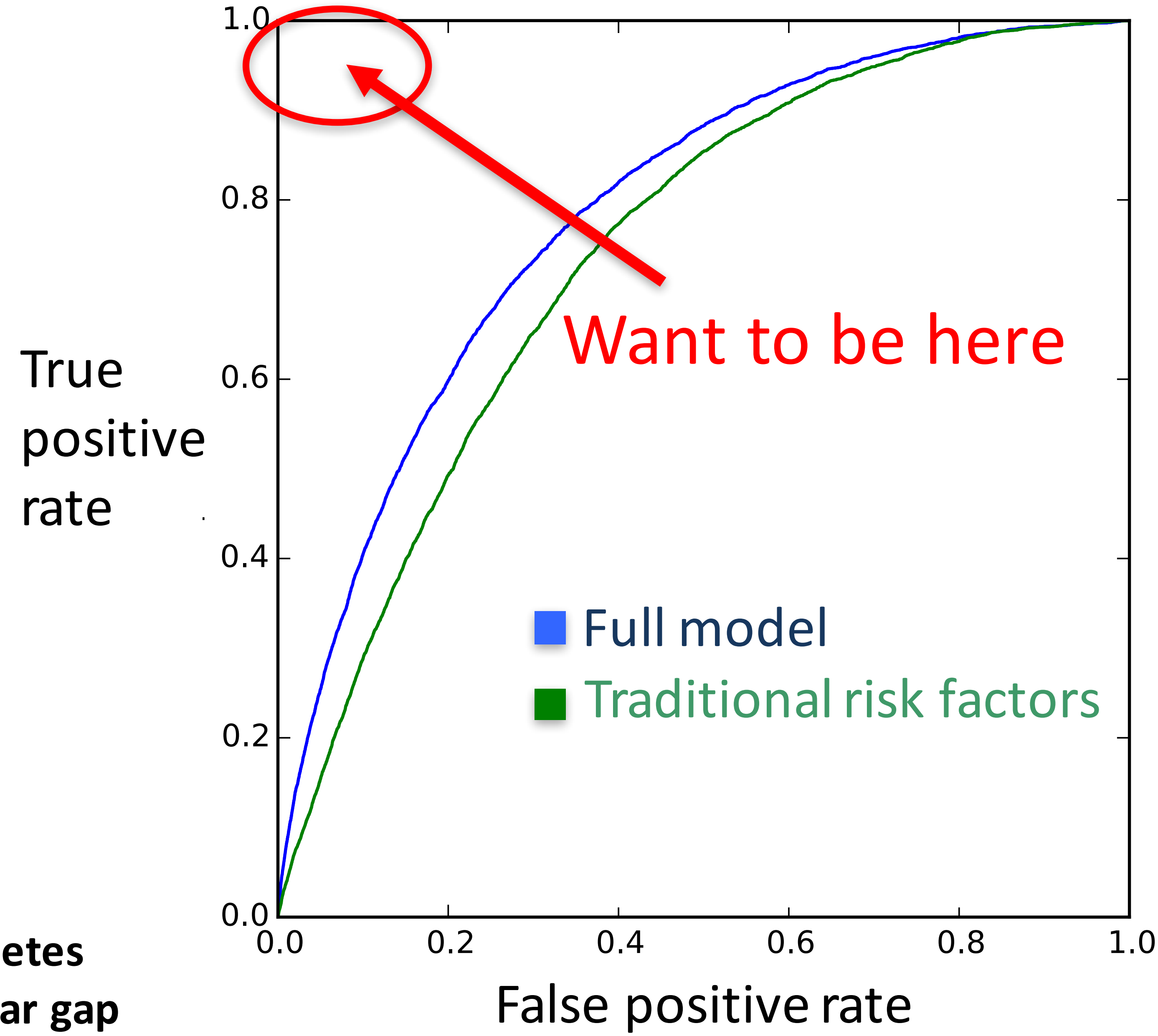
Top Lab Factors

Hemoglobin A1c /Hemoglobin.Total (High	
Glucose (High- Past 6 months)	
Cholesterol.In VLDL (Increasing - Past 2	
Potassium (Low - Entire History)	
Cholesterol.Total/Cholesterol.In HDL (Hig	(2.10 2.40)
Erythrocyte mean corpuscular hemoglobin concentration -(Low - Entire History)	2.25 (1.92 2.64)
Eosinophils (High - Entire History)	2.11 (1.82 2.44)
Glomerular filtration rate/1.73 sq M.Predicted (Low -Entire History)	2.07 (1.92 2.24)
Alanine aminotransferase (High Entire History)	2.04 (1.89 2.19)

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

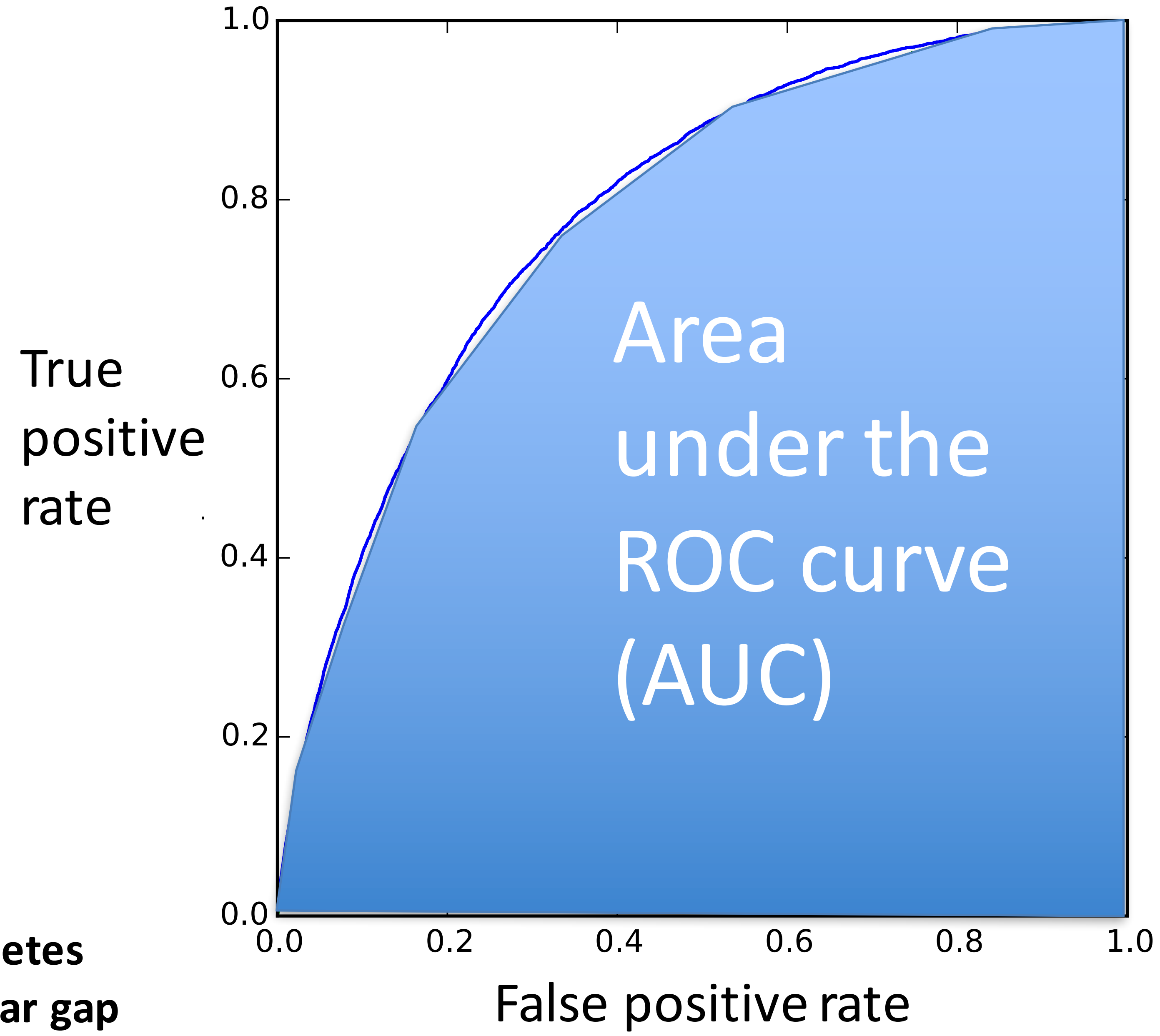


Want to be here

Obtained by varying prediction threshold

Diabetes 1-year gap

Receiver-operator characteristic curve

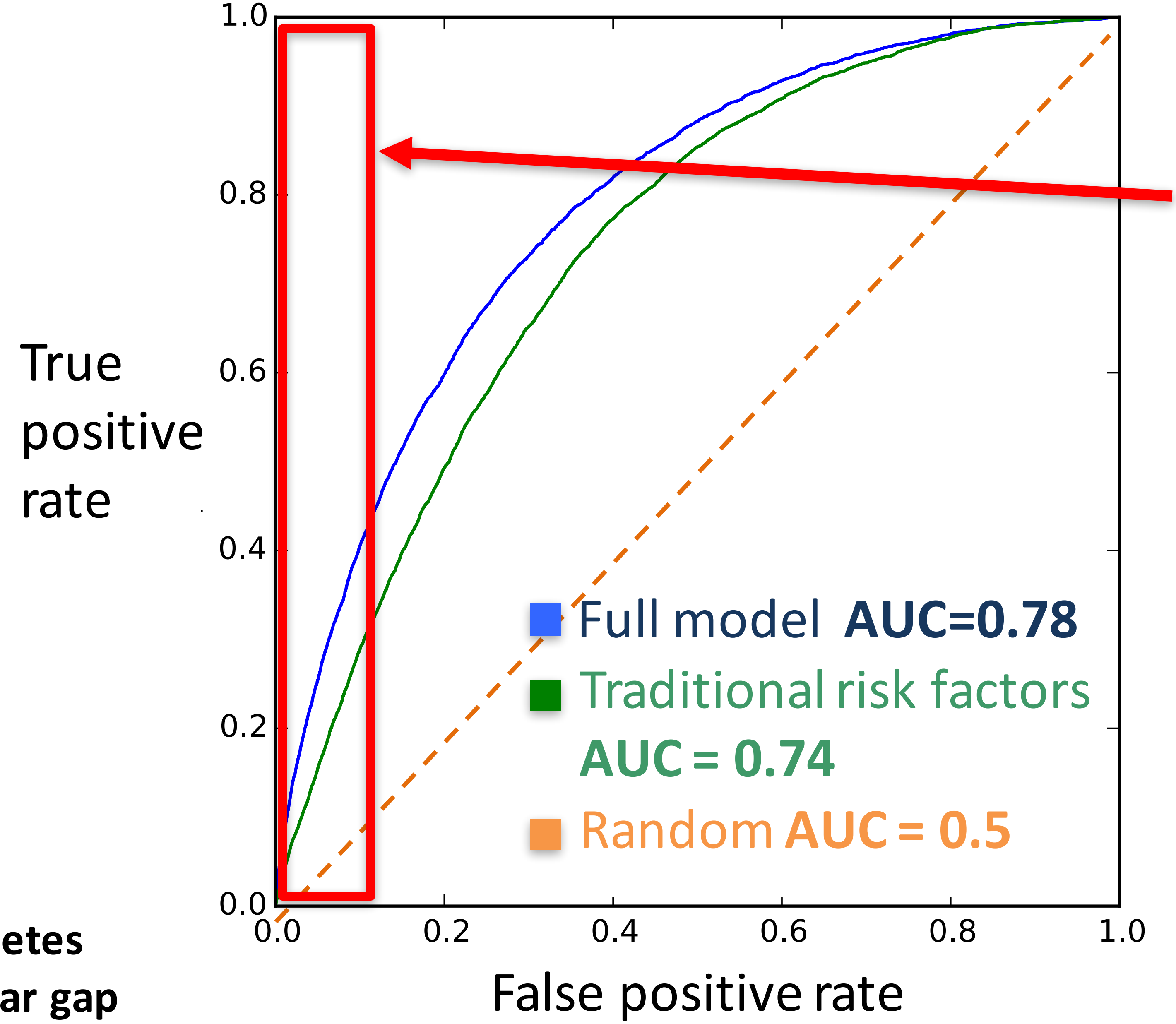


AUC = Probability that algorithm ranks a positive patient over a negative patient

Invariant to amount of class imbalance

Diabetes 1-year gap

Receiver-operator characteristic curve

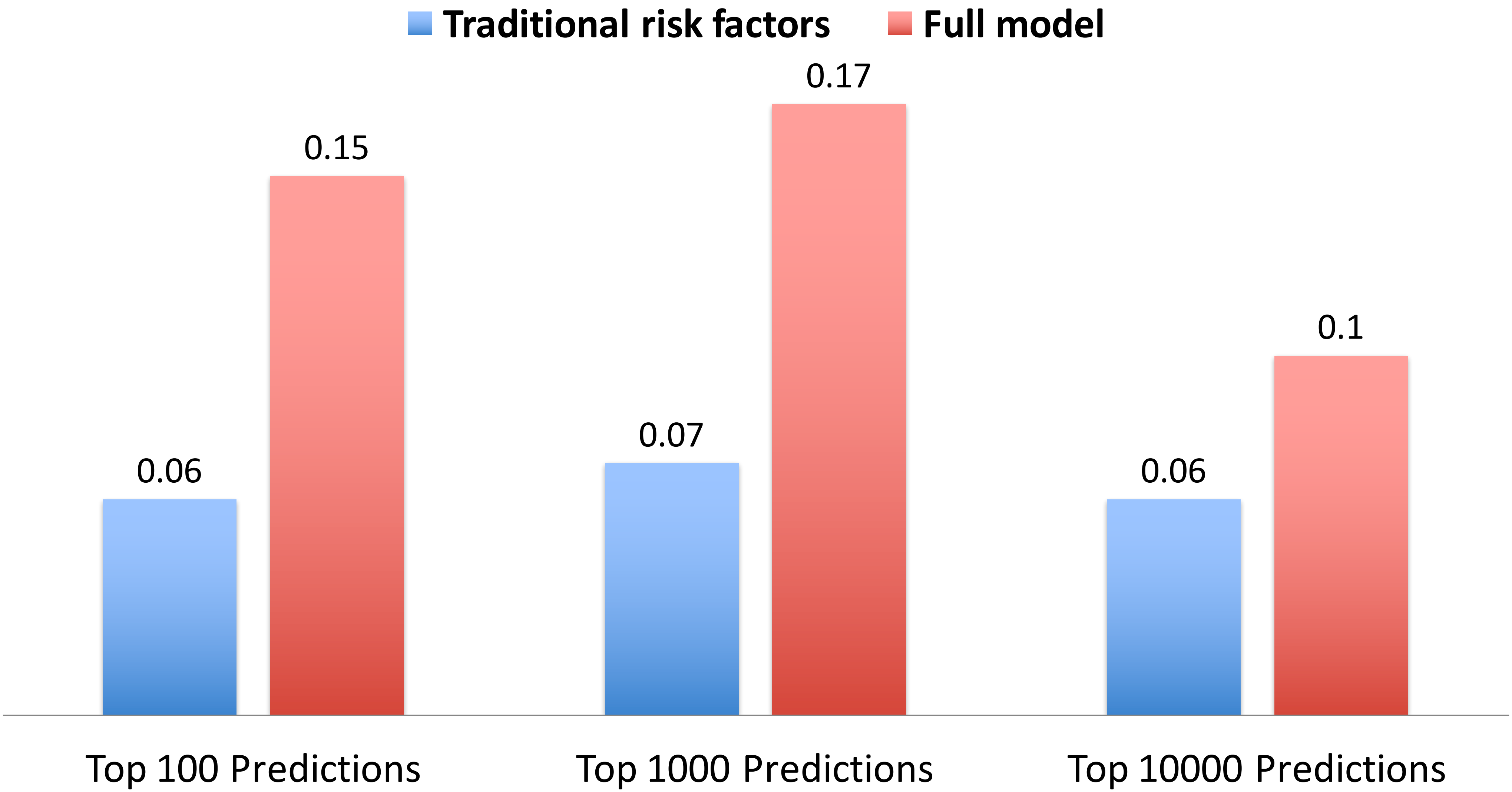


Risk stratification usually focuses on just this region

(because of the cost of interventions)

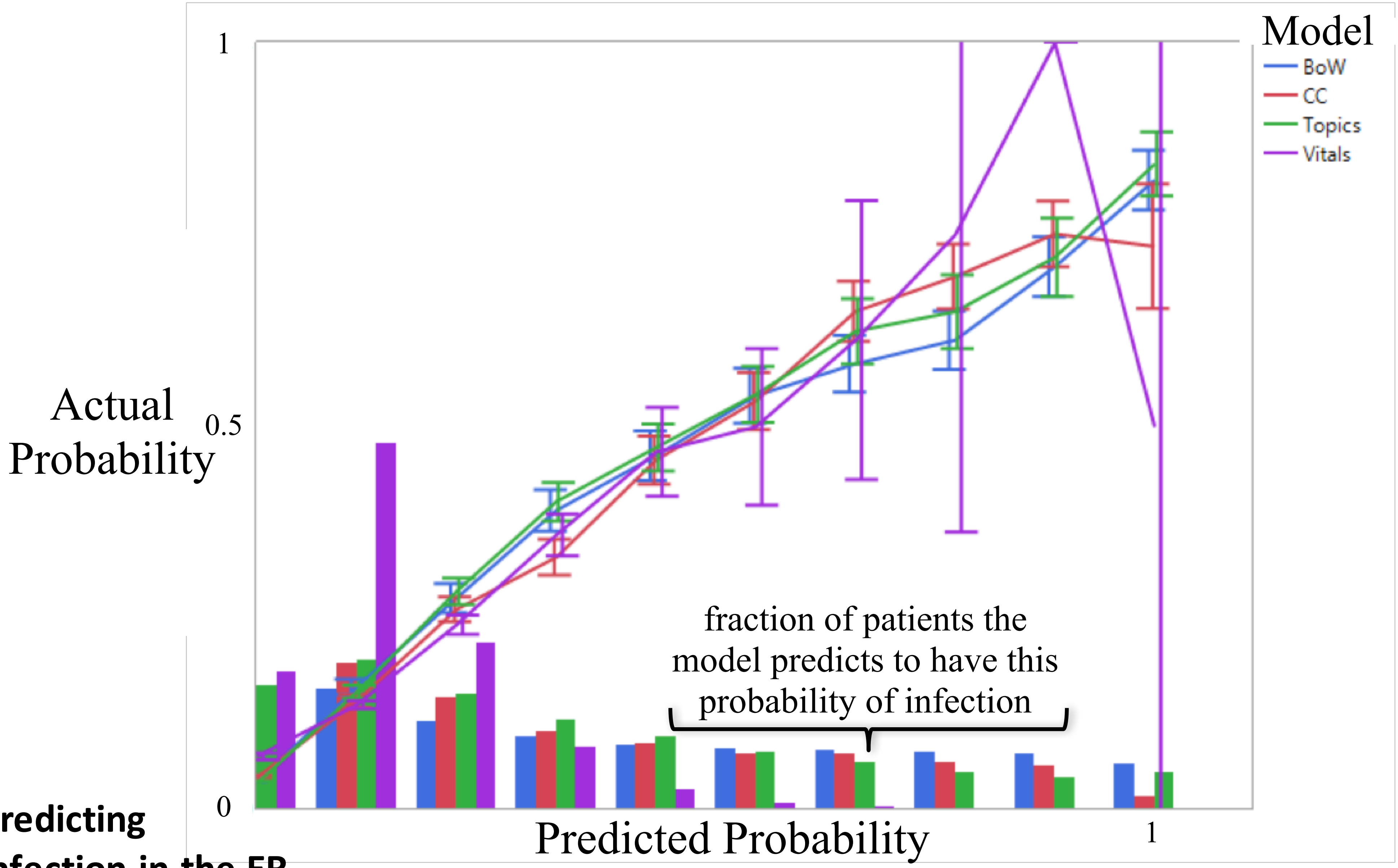
Diabetes 1-year gap

Positive predictive value (PPV)



Diabetes 1-year gap

Calibration (*note: different dataset*)



Predicting infection in the ER

Outline for today’s class

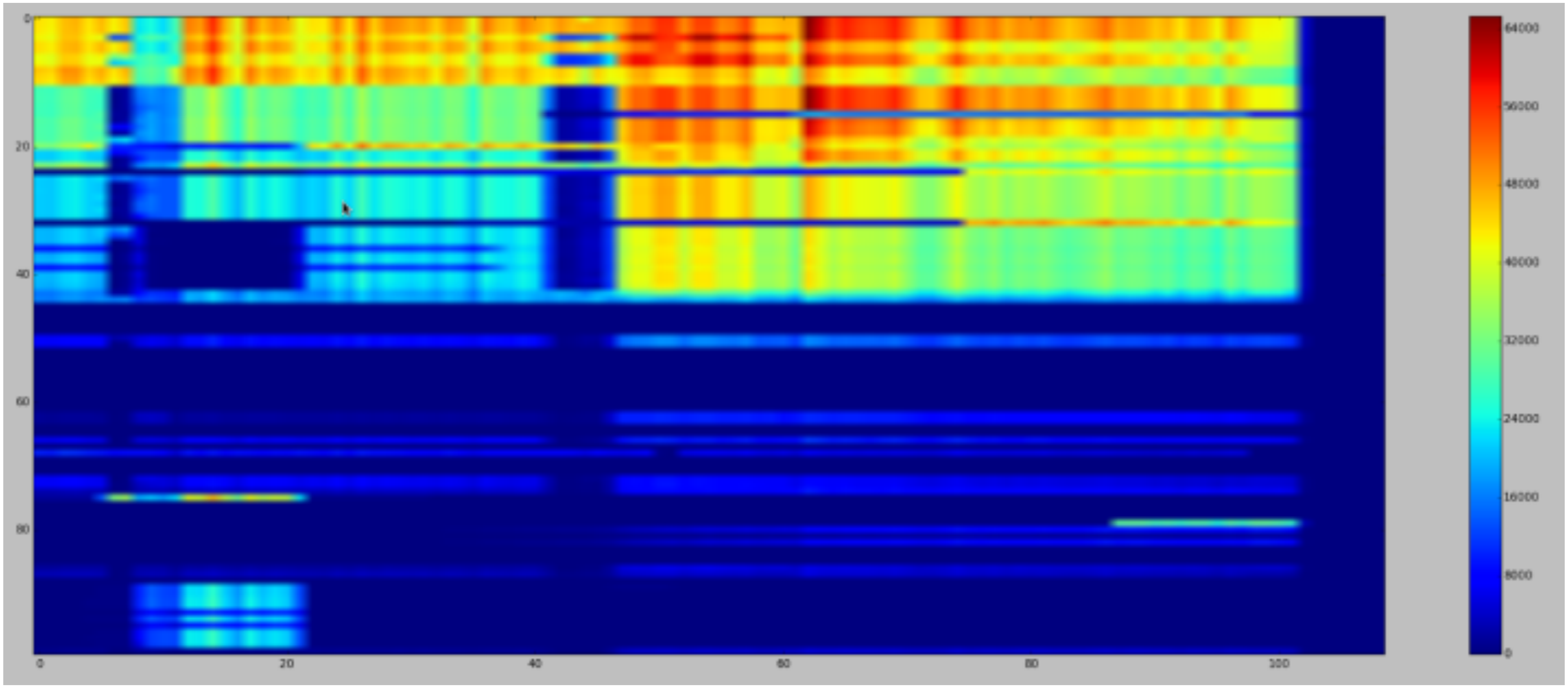
1. Case study for risk stratification: Early detection of Type 2 diabetes
2. Framing as supervised learning problem
3. Deriving labels
4. Evaluating risk stratification algorithms
5. **Non-stationarity**

Major challenge: non-stationarity

- ICD10 rolled out in 2015: predictive models learned using ICD9 features are no longer useful!
- Logistical issues => some features may not be available!
- Prevalence and significance of features may change over time
- Automatically derived labels may change meaning

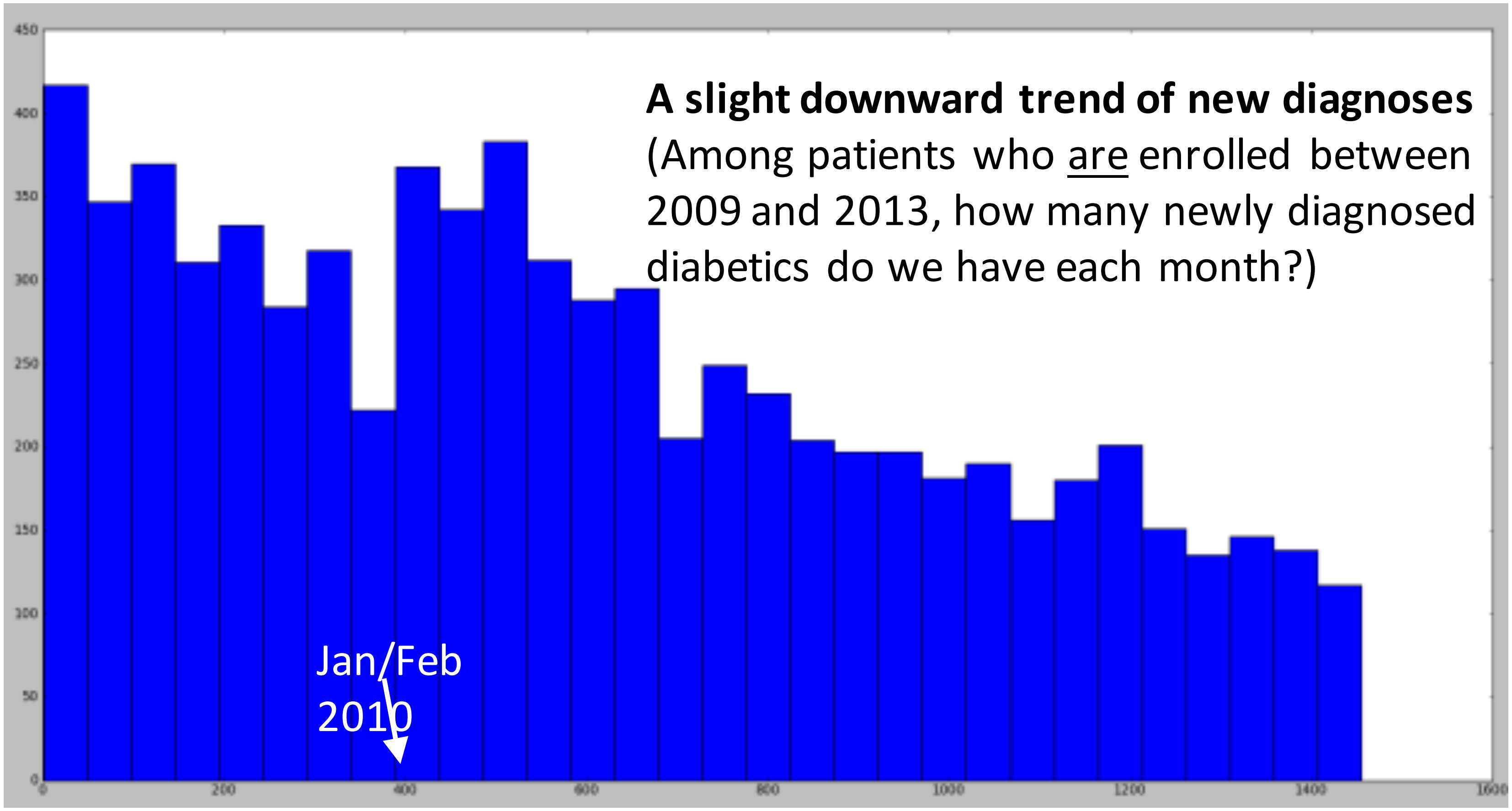
Top 100 lab measurements over time

Labs



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

Diabetes Onset after 2009



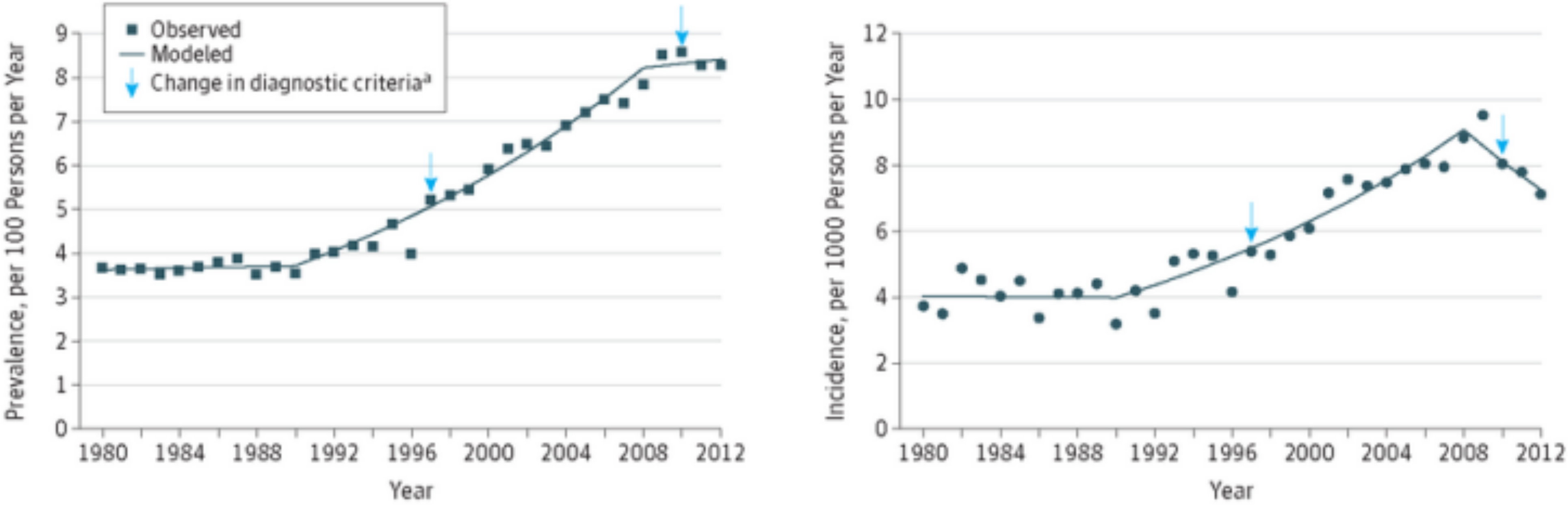
A slight downward trend of new diagnoses
(Among patients who are enrolled between 2009 and 2013, how many newly diagnosed diabetics do we have each month?)

Jan/Feb 2010

Number of people who had diabetes onset

Months, after 2009/01/01

Diabetes Onset after 2009



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;312(12):1218-1226.

External validity

- Motivates multi-institution evaluations
- Good practice is to let the test data be from a future year