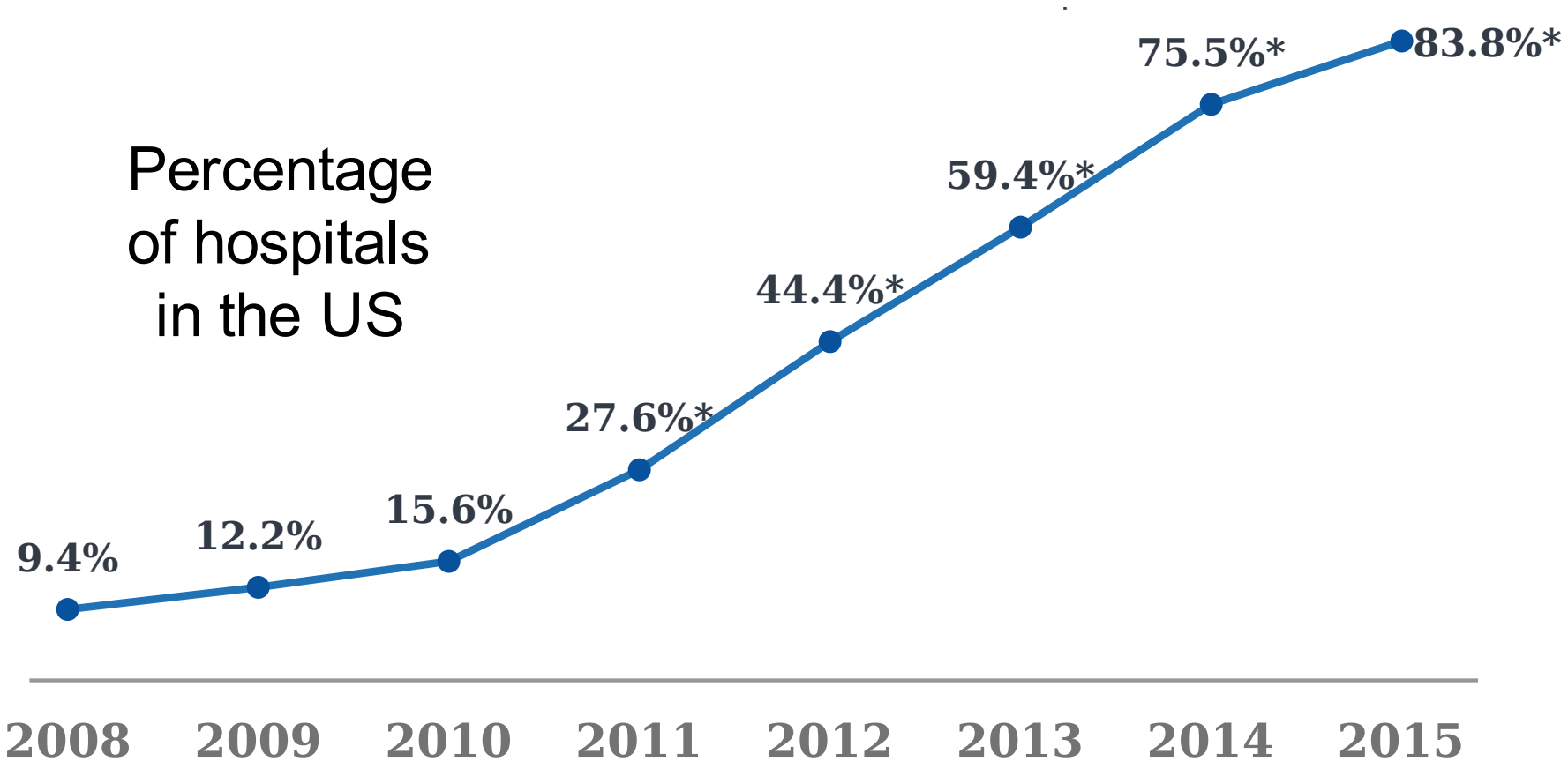
Why Does Interpretability Matter in Health Care?

David Sontag
Clinical Machine Learning Group, MIT

Department of Electrical Engineering and Computer Science (EECS),
Computer Science and Artificial Intelligence Laboratory (CSAIL)
Institute for Medical Engineering & Science (IMES)

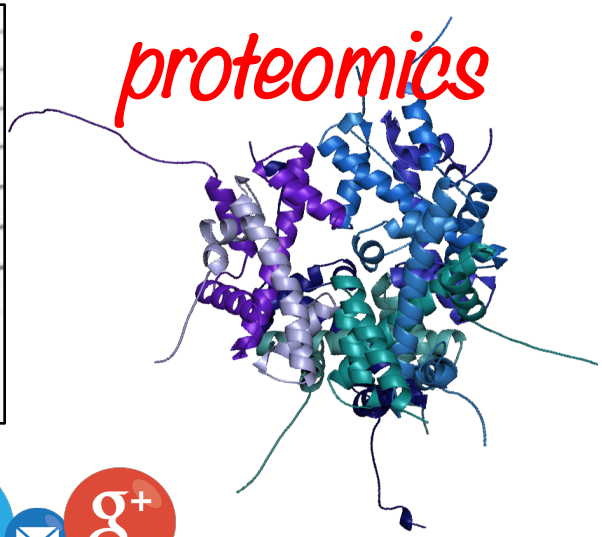
Massachusetts Institute of Technology

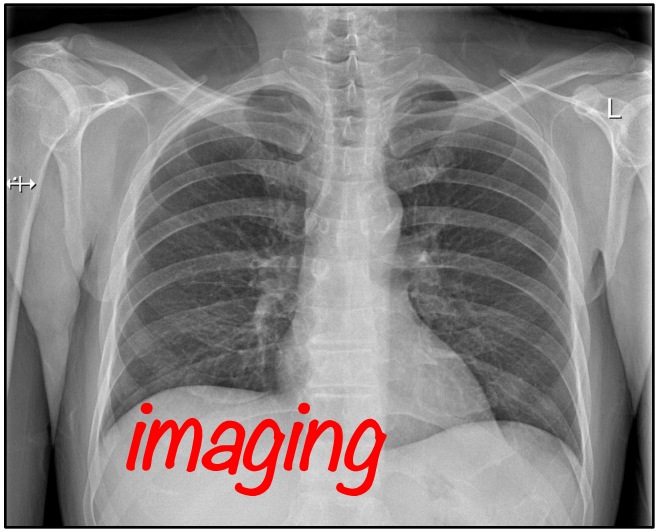
Adoption of Electronic Health Records (EHR) has increased 9x since 2008



[Henry et al., ONC Data Brief, May 2016]

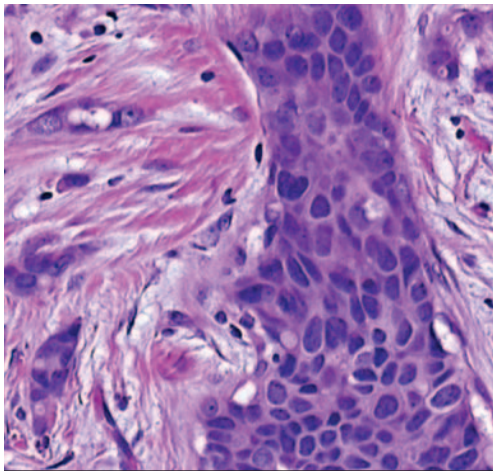
Wealth of digital health data available





Where can machine learning help?

Computational pathologist
(Beck et al., *Sci Transl Med*, 2011)



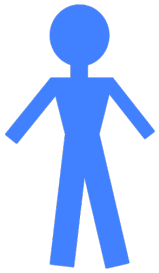
Differential diagnosis
(INTERNIST-1/Quick Medical Reference, 1980's)

Symptoms	Differential diagnosis
Cough	1. Common cold
Fever	2. Flu
Headache	3. Strep throat
Sore throat	4. Meningitis

Finding undiagnosed Type 2 diabetics
(my lab: Razavian et al., *Big Data* 2016)

Improving EHR documentation
(my lab: Jernite et al., 2013)

In-silico models for precision medicine

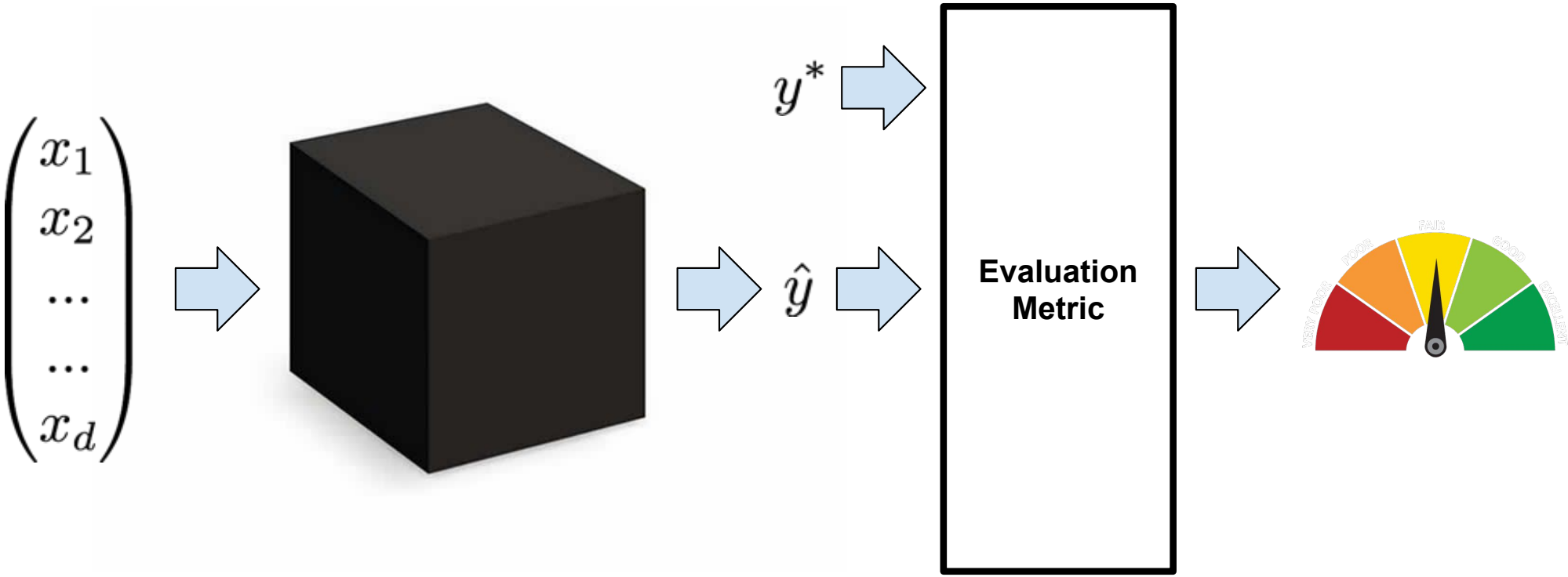


For this specific individual, which medication is better, A or B?

Outline for today's talk

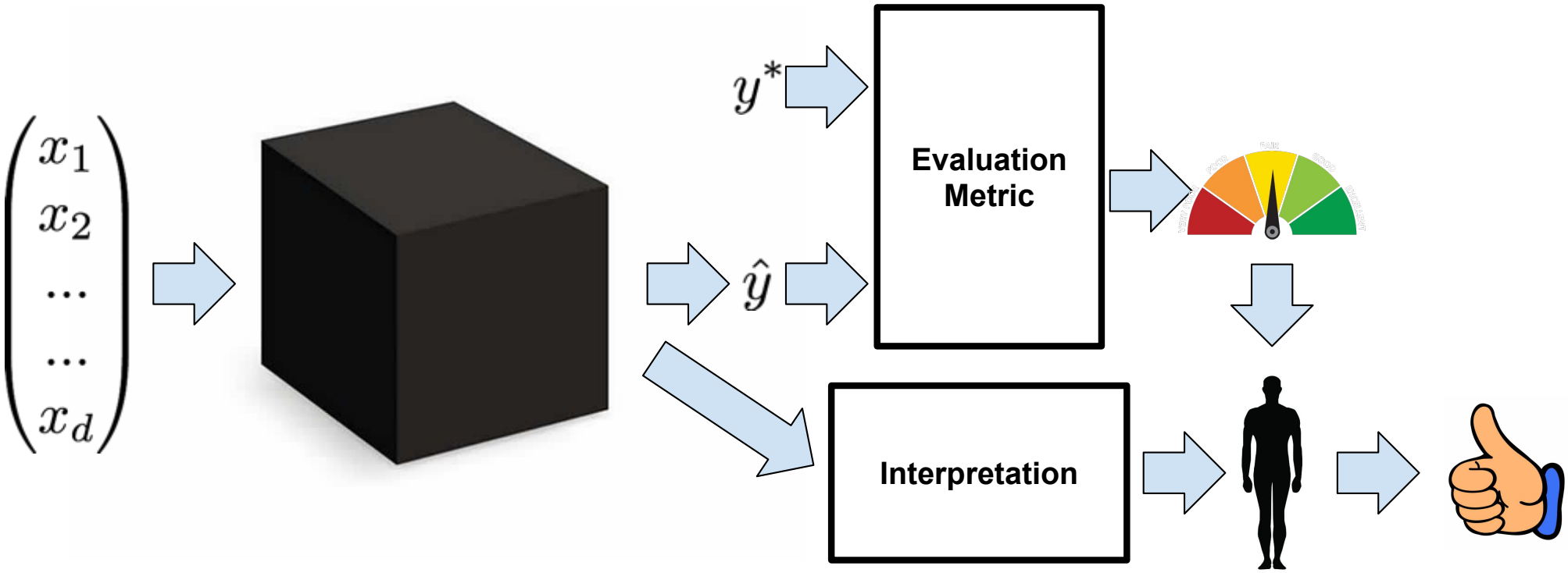
- **Reasons to want interpretable models:** trust, causality, transferability, informativeness
- **Case studies:**
  - Early detection of Type 2 diabetes
  - Framingham Coronary Heart Disease Risk Score
  - Improving clinical documentation with "auto-complete"

Typical ICML paper: we get SOTA on ___ benchmark



(Slide credit: Zachary Lipton)

Real world ML: “it’s complicated”



When humans are consumer of ML, often we want something the metric doesn’t capture. But, what?

(Slide credit: Zachary Lipton)

Trust

- Does the model *know* when it's uncertain?
- Does the model make same mistakes as human? (e.g., would we be happy delegating decision making authority?)
- Are we *comfortable* with the model?

(Slide credit: Zachary Lipton)

Causality

- We may want models to tell us something about the natural world
- Supervised models are trained simply to make predictions, but often used to take actions
- Naïve interpretations can be misleading

(Slide credit: Zachary Lipton)

Transferability

- The idealized training setups often differ from the real world
  - E.g., data leakage, errors in outcome definition from observational data

- Real problem may be non-stationary, noisier, etc.

- Want sanity-checks that the model doesn't depend on weaknesses in setup

(Slide credit: Zachary Lipton)

Informativeness

- We may train a model to make a decision

- But it's real purpose is usually to aid a person in making a decision

- Thus an interpretation may be valuable for the extra bits it carries

  I.e., ability to integrate model output with human prior beliefs

(Slide credit: Zachary Lipton)

CASE STUDY 1

Early detection of Type 2 diabetes and its complications

Work led by Narges Razavian

Early Detection of Type 2 Diabetes

- Global prevalence will go from 171 million in 2000 to 366 million in 2030
- 25% of people in the US with diabetes are undiagnosed
- Leads to complications of cardiovascular, cerebrovascular, renal, and vision systems

Traditional risk assessment

- Use small number of risk factors (e.g. ~20)
- Easy to ask/measure in the office
- 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

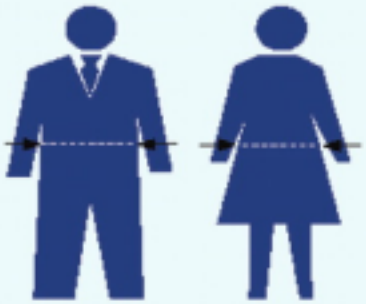
2. Body-mass index

(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

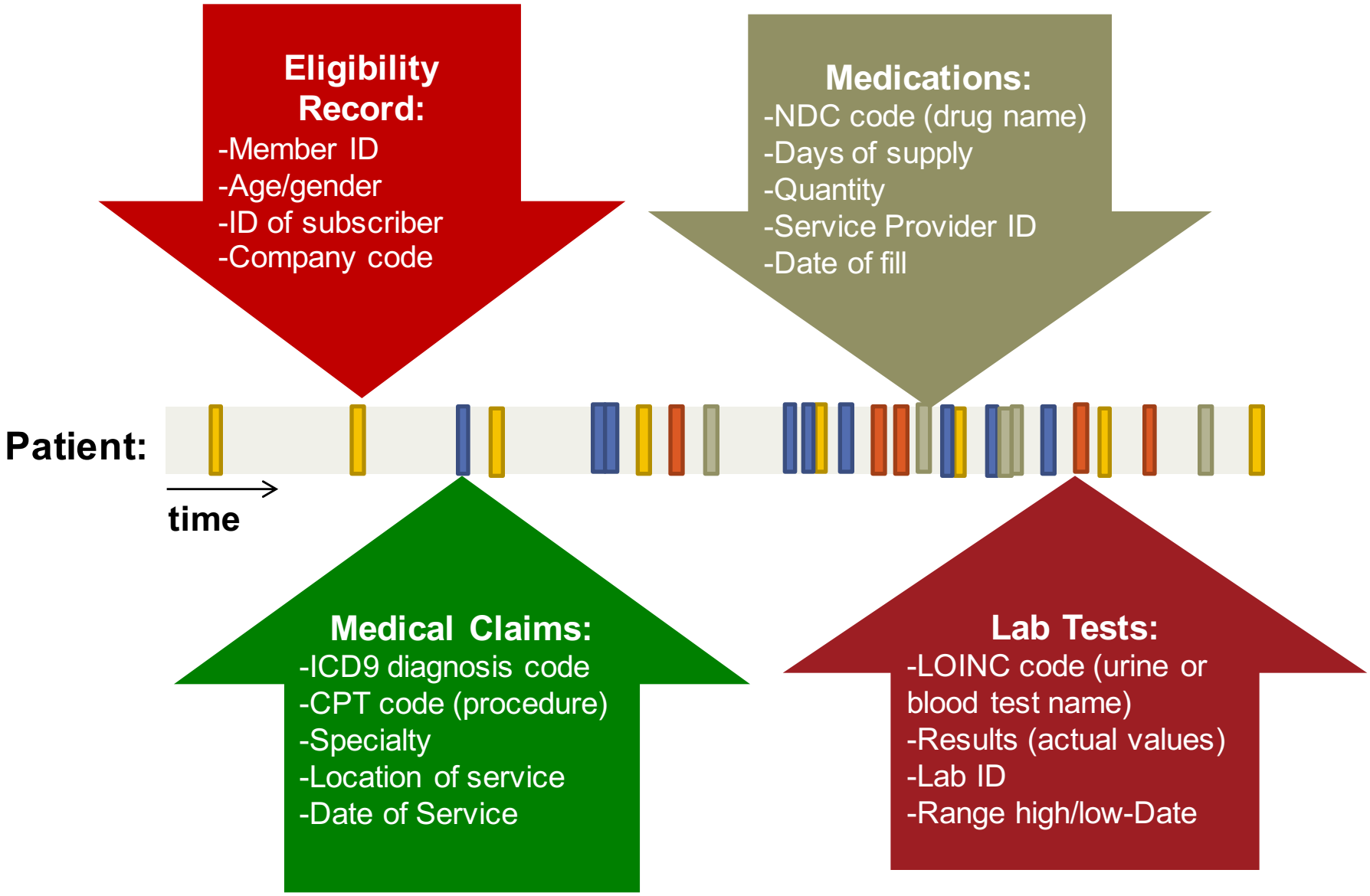
Please turn over

Population-Level Risk Stratification

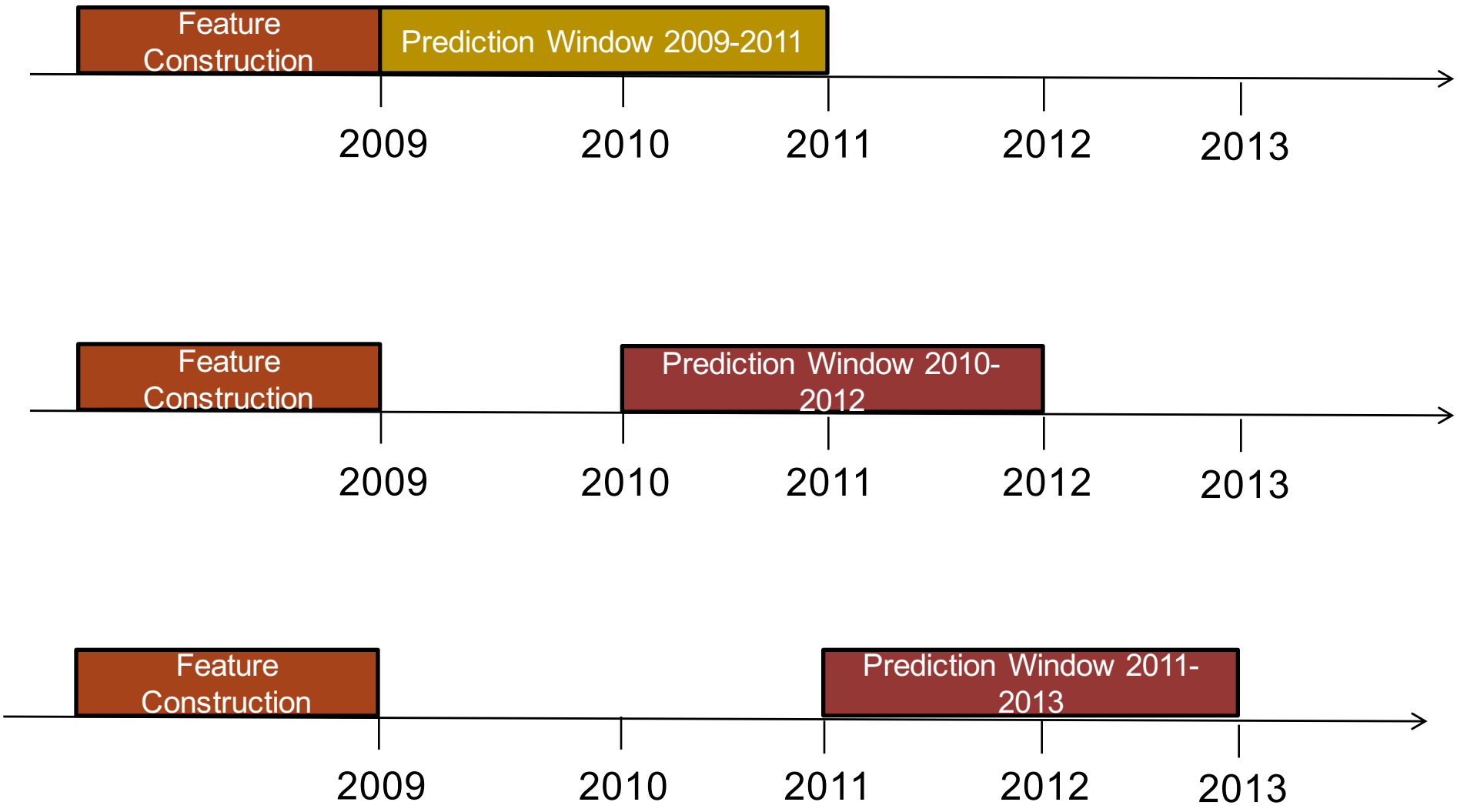
- Key idea: Use automatically collected administrative, utilization, and clinical data
- Machine learning will find surrogates for risk factors that would otherwise be missing
- Enables risk stratification at the population level – millions of patients

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

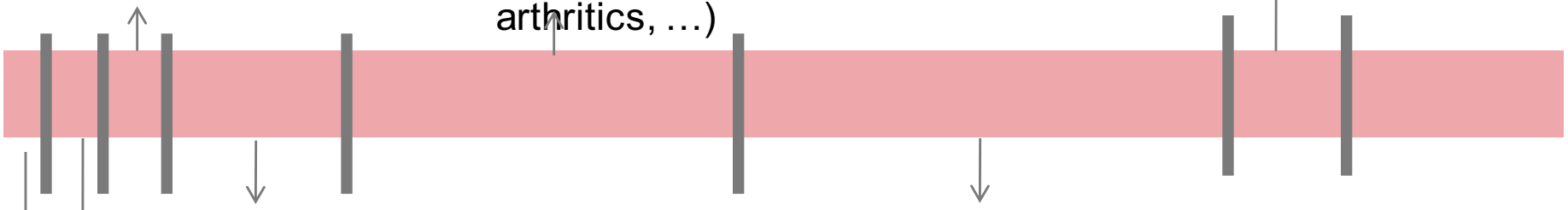
Administrative & Clinical Data



ML task formulation



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)

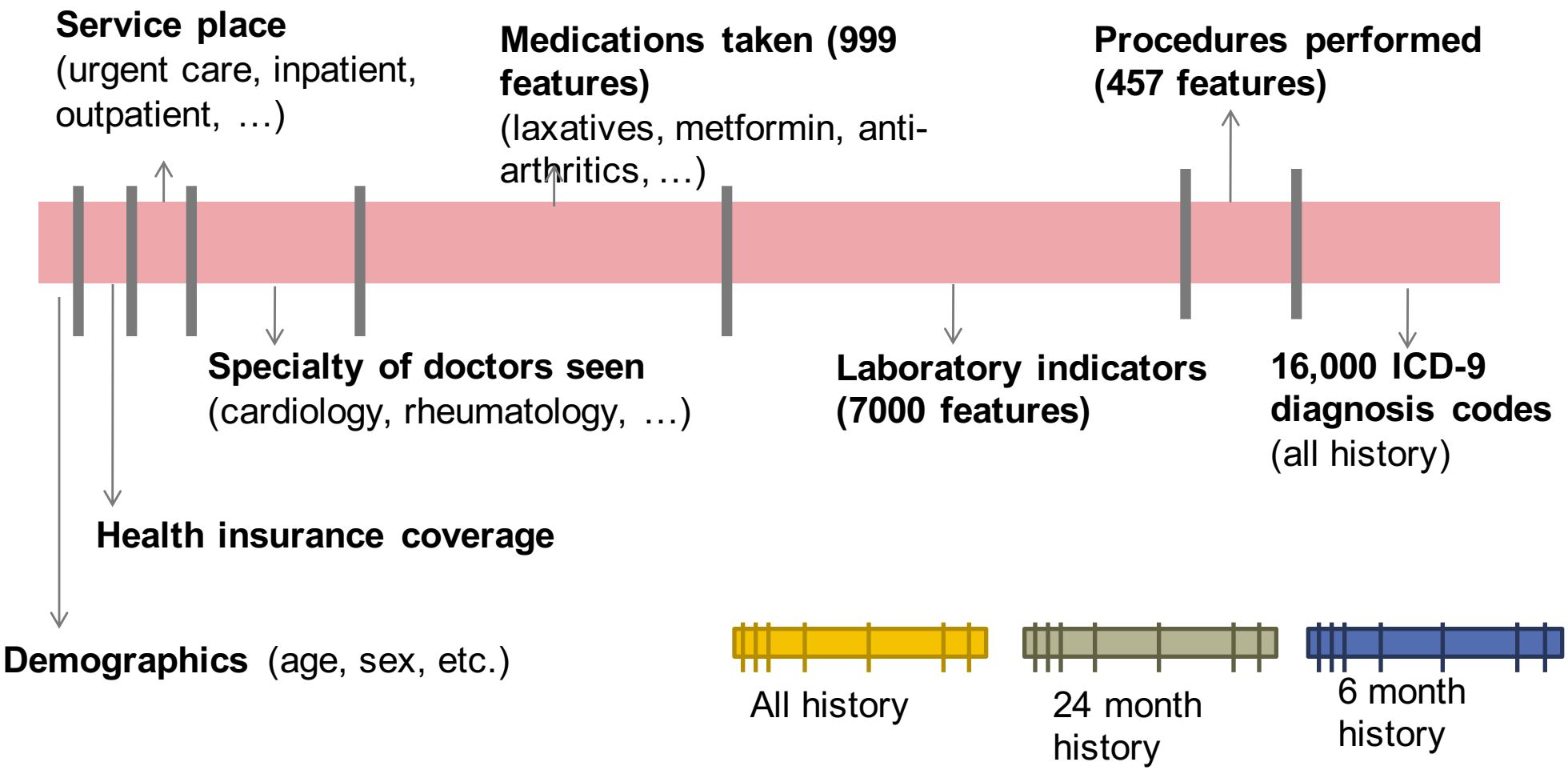
Health insurance coverage

Demographics (age, sex, etc.)

For the 1000 most frequent lab tests:

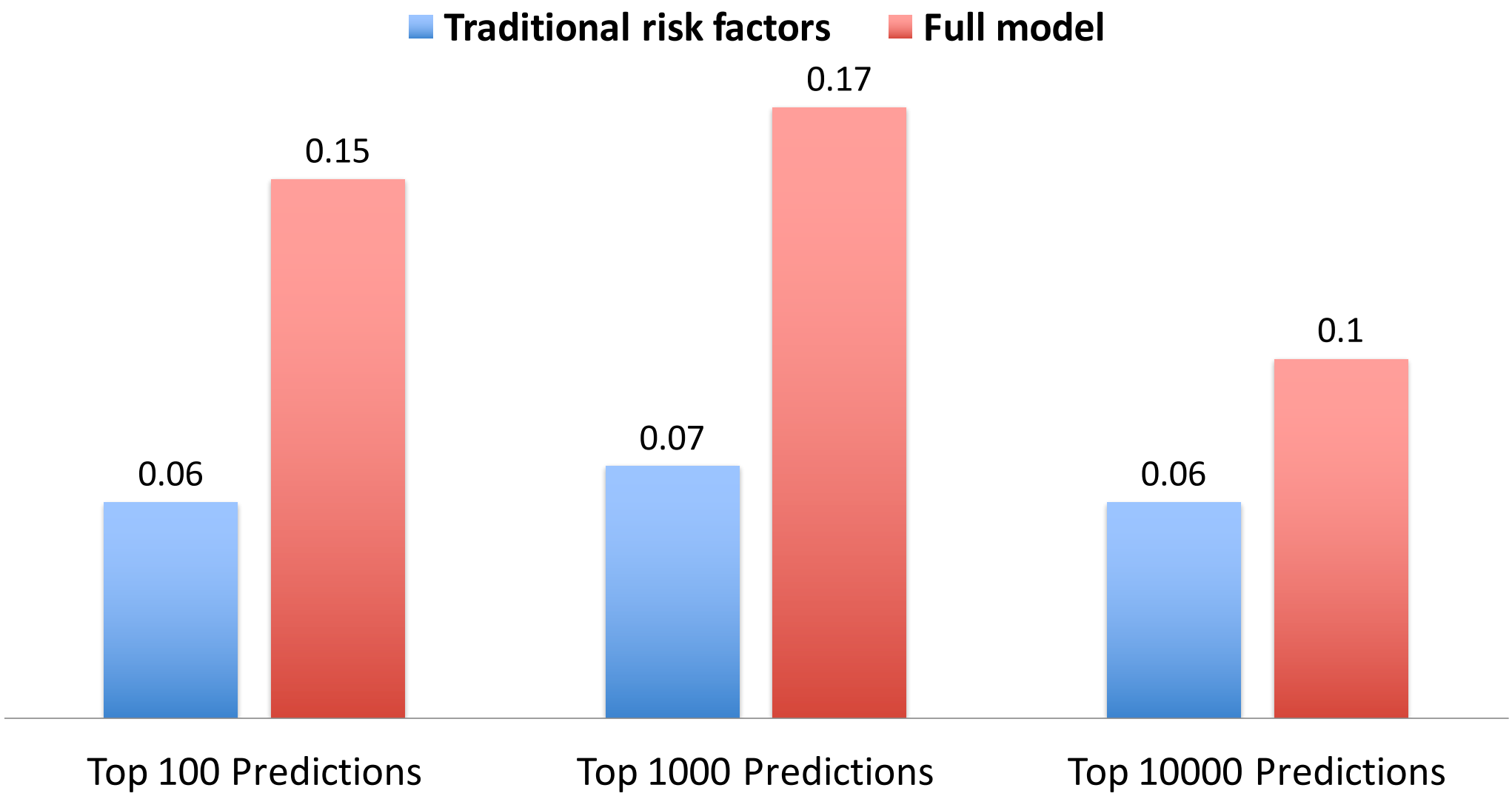
- Was the test ever administered?
- Was the result ever low?
- Was the result ever high?
- Was the result ever normal?
- Is the value increasing?
- Is the value decreasing?
- Is the value fluctuating?

Features used in models



Total features per patient: 42,000

Positive predictive value (PPV)



Diabetes 1-year gap

Questions

1. Did we set up the prediction task correctly so that it mimics how we would apply it prospectively?

   Transferability

2. Do our models suggest any causal hypotheses of the mechanism in which patients become diabetic?

   Causality

3. Is anything fundamentally new discovered?

   Trust, Informativeness

4. Is the model likely to be useful in new settings?

   Transferability

What are the Discovered Risk Factors?

- 769 variables have non-zero weight
- No time to look at all 769. Instead, we do a regression with much higher regularization *just for visualization purposes*

What are the Discovered Risk Factors?

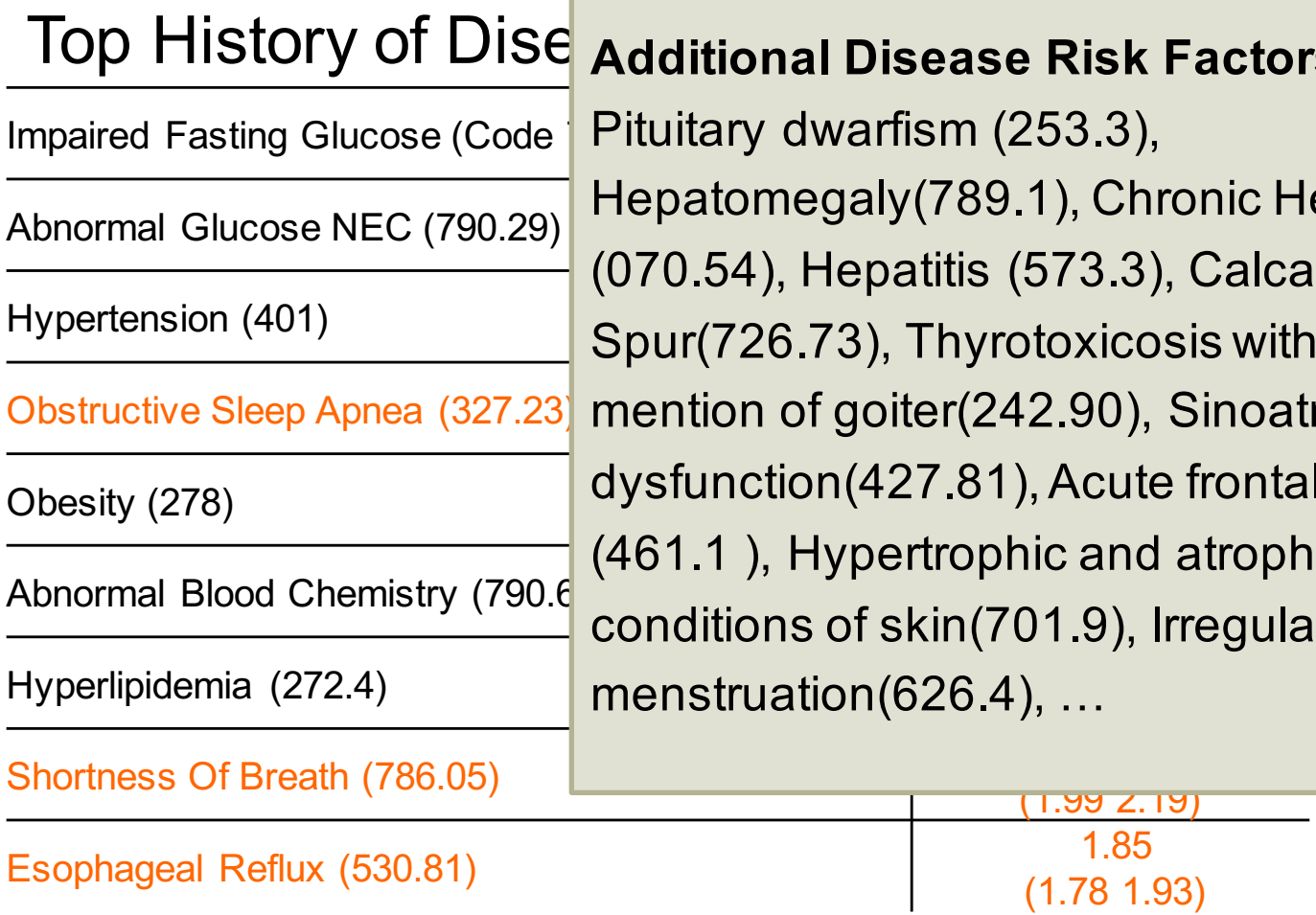
- 769 variables have non-zero weight

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



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

Questions

1. Did we set up the prediction task correctly so that it mimics how we would apply it prospectively?

   Transferability

2. Do our models suggest any causal hypotheses of the mechanism in which patients become diabetic?

   Causality

3. Is anything fundamentally new discovered?

   Trust, Informativeness

4. Is the model likely to be useful in new settings?

   Transferability

1. Did we set up the prediction task correctly so that it mimics how we would apply it prospectively?

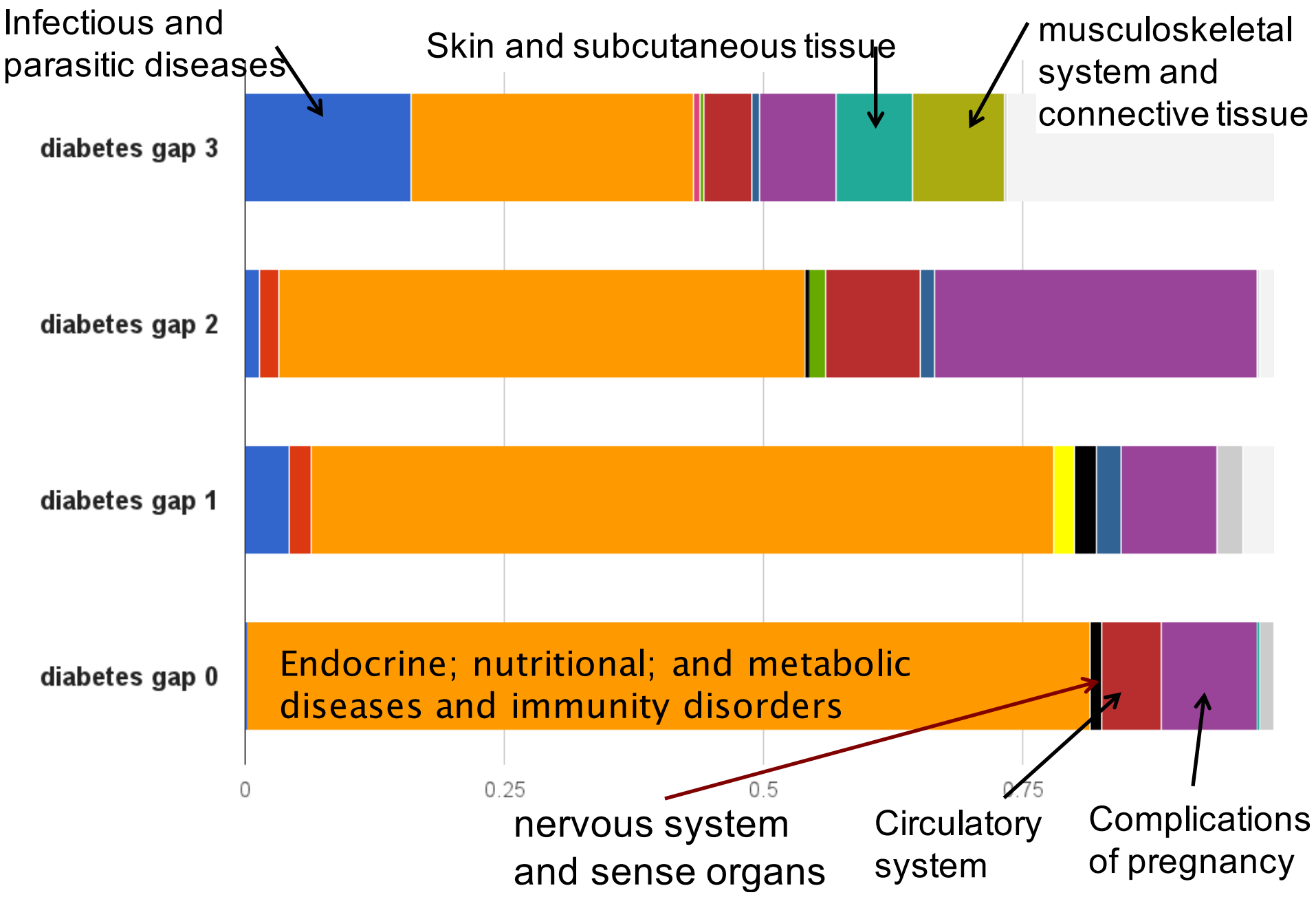
For gap=0, possibly not!

We see that the diabetic medication ***Metformin*** *(a first-line diabetic treatment) is predictive*

Table 3. Top predictive variables for type 2 diabetes onset within 2009–2010 (gap = 0), using patient

Variable type	*Variable evaluation period*[a]	*Variable description*	*Number with diabetes*	*Number without diabetes*	*OR (95% CI)*
Laboratory test	Past 2 years	Hemoglobin A1c/hemoglobin.total—high (LOINC-4548-4)	1845	8710	9.28 (8.81 9.78)
	Past 2 years	Glucose—high (LOINC-2345-7)	5274	58,736	4.58 (4.43 4.73)
	Past 2 years	Hemoglobin A1c/hemoglobin.total—request for test (LOINC-4548-4)	3908	45,519	4.06 (3.92 4.21)
	Entire history	Cholesterol.in HDL—low (LOINC-2085-9)	3233	49,524	2.94 (2.83 3.06)
	Entire history	Triglyceride—high (LOINC-2571-8)	6056	106,818	2.85 (2.77 2.94)
	Entire history	Cholesterol.total/cholesterol.in HDL—high (LOINC-9830-1)	3114	56,032	2.46 (2.37 2.56)
	Entire history	Alanine aminotransferase—high (LOINC-1742-6)	1208	22,205	2.26 (2.13 2.40)
	Entire history	Cholesterol.in VLDL—request for test (LOINC-13458-5)	3029	63,166	2.09 (2.01 2.18)
	Entire history	Cholesterol.total/cholesterol.in HDL—decreasing (LOINC-9830-1)	3277	75,701	1.89 (1.81 1.96)
	Past 2 years	Carbon dioxide—request for test (LOINC-2028-9)	6044	158,472	1.77 (1.72 1.83)
ICD9 history	Entire history	Abnormal glucose (ICD9 790.29)	1198	10,099	5.00 (4.70 5.32)
	Entire history	Impaired fasting glucose (ICD9 790.21)	1285	11,521	4.72 (4.45 5.01)
	Entire history	Hypertension (ICD9 401)	12,175	227,759	4.09 (3.97 4.22)
	Entire history	Chronic liver disease (ICD9 571.8)	619	6845	3.71 (3.41 4.03)
	Entire history	Obesity (ICD9 278)	3104	48,000	2.90 (2.78 3.01)
	Entire history	Obstructive sleep apnea (ICD9 327.23)	1178	17,302	2.84 (2.67 3.02)
	Entire history	Hypersomnia with sleep apnea (ICD9 780.53)	1138	16,965	2.79 (2.63 2.97)
	Entire history	Abnormal blood chemistry (ICD9 790.6)	2388	38,726	2.68 (2.56 2.80)
	Entire history	Hyperlipidemia (ICD9 272.4	8745	186,016	2.62 (2.54 2.69)
	Entire history	Anemia (ICD9 285.9)	3421	75,500	1.99 (1.92 2.07)
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	Entire history	Acute bronchitis (ICD9 466.0)	3229	93,559	1.46 (1.41 1.52)
NDC medication history	Entire history	Medication group: Metformin	286	1142	10.17 (8.93 11.59)
	Entire history	Medication group: antiarthritics	3055	88,506	1.46 (1.40 1.51)
	Entire history	Medication group: nonsteroidal anti-inflammatory drugs	3216	94,531	1.44 (1.38 1.49)
Healthcare utilization	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	Entire history	Service place code: home	4386	113,223	1.72 (1.66 1.77)
	Entire history	Dental coverage = yes	4142	119,108	1.50 (1.45 1.55)
	Entire history	Specialty code: internal medicine	7246	227,156	1.45 (1.40 1.49)
	Entire history	Procedure group: ophthalmologic and otologic diagnosis and	6681	247,300	1.13 (1.09 1.16)

Mechanism? Risk Factors per Body System



Preventing diabetic onset or progression

- Our goal ultimately is to use the models' predictions to prioritize *interventions* to prevent progression of diabetes

- What should these interventions be?
  1. **Approach 1:** Look for existing interventions in the data that might show promise (but perhaps are inconsistently applied)

     "Gastric bypass surgery" is highest negative weight (9th most predictive feature)

     Does *gastric bypass surgery* prevent *onset of diabetes*?

  2. **Approach 2**: Characterize the patient population that we can predict well for, and use clinical expertise

CASE STUDY 2

Framingham Coronary Heart Disease (CHD) Risk Score

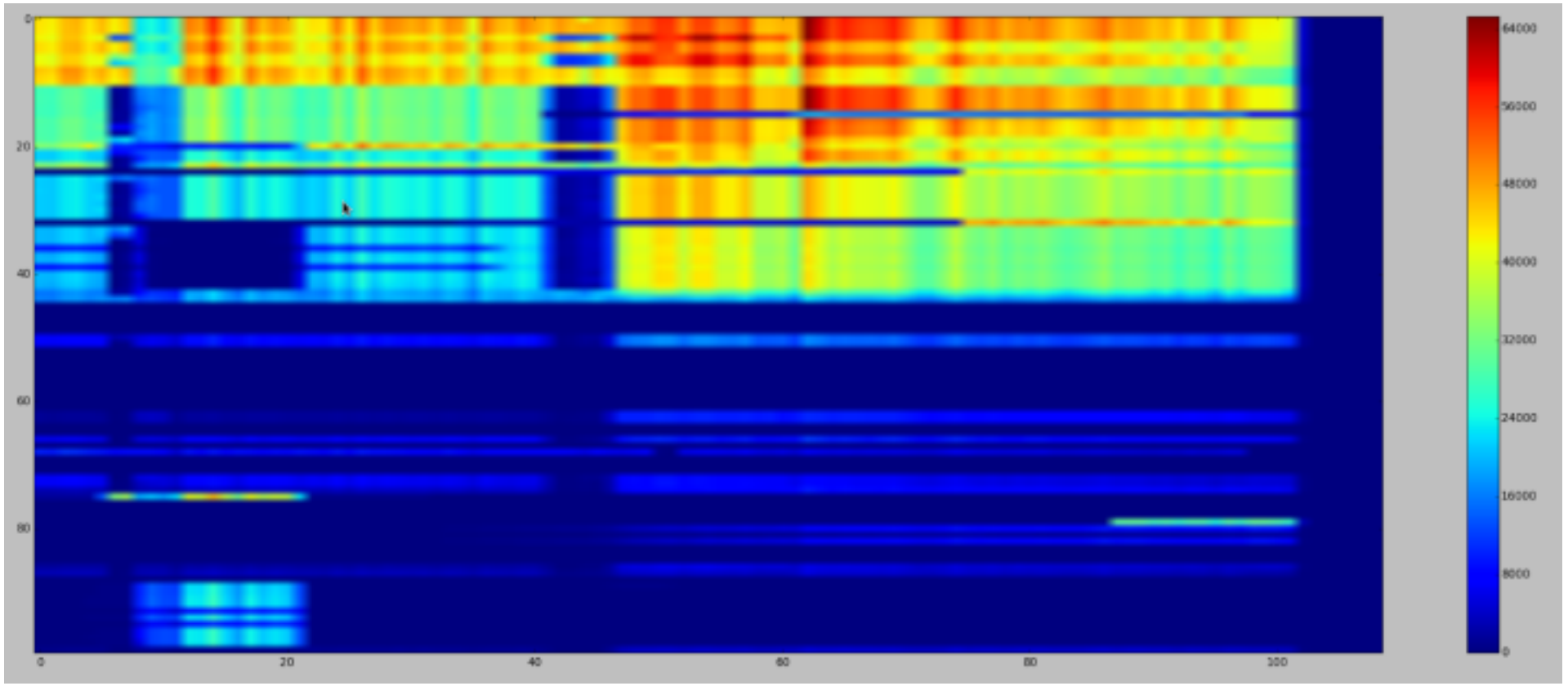
Transferability: non-stationary

- Data created during health care is from a non-stationary process due to changes in:
  - Medical science
  - Incentives & regulations
  - Business processes

(Slide credit: Ken Jung)

Transferability: non-stationary

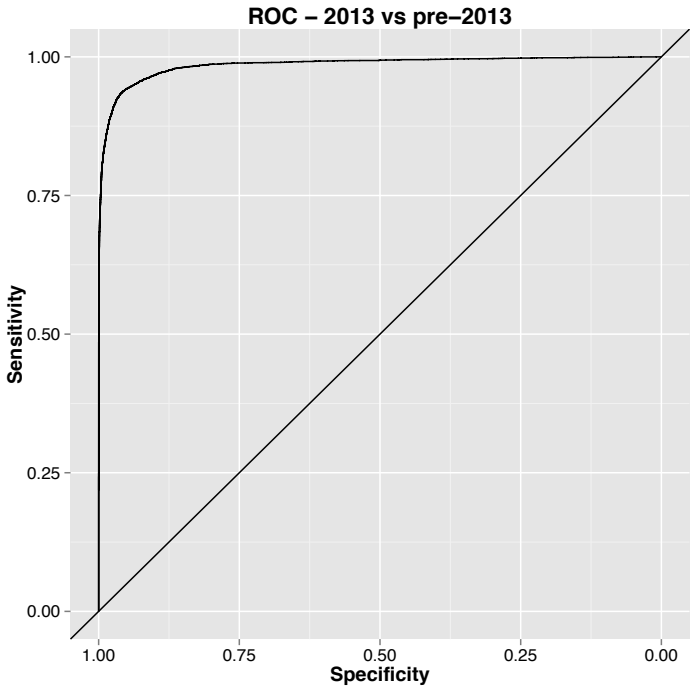
Top 100 lab measurements over time



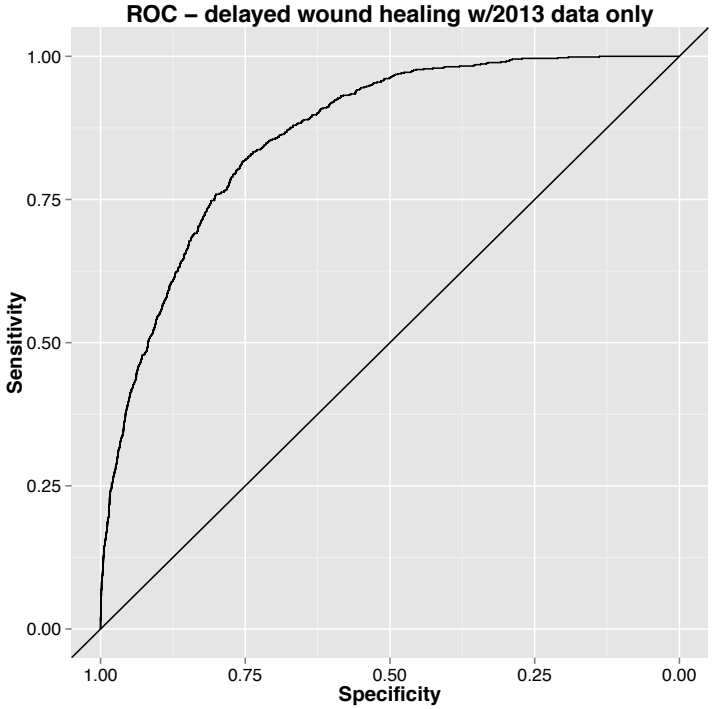
Time (in months, from 1/2005 up to 1/2014)

Transferability: non-stationary

- Testing for covariate shift (wound healing):



Distinguish 2013 from pre-2013



Distinguish first 2/3 of 2013 from last 1/3 of 2013

(Slide credit: Ken Jung)

Case study on transferability: Framingham CHD risk score

- Many ML models are trained in one place and deployed more broadly
- **Example:** Framingham coronary heart disease (CHD) risk score
  - Model based on 6 major risk factors: age, BP, smoking, diabetes, total cholesterol (TC), and high-density lipoprotein cholesterol (HDL-C)

[Wilson et al., Circulation, 1998]

CHD score sheet for men using TC or LDL-C categories.

Step 1

Age		
Years	LDL Pts	Chol Pts
30-34	-1	[-1]
35-39	0	[0]
40-44	1	[1]
45-49	2	[2]
50-54	3	[3]
55-59	4	[4]
60-64	5	[5]
65-69	6	[6]
70-74	7	[7]

Step 2

LDL - C		
(mg/dl)	(mmol/L)	LDL Pts
<100	<2.59	-3
100-129	2.60-3.36	0
130-159	3.37-4.14	0
160-190	4.15-4.92	1
≥190	≥4.92	2

Cholesterol		
(mg/dl)	(mmol/L)	Chol Pts
<160	<4.14	[-3]
160-199	4.15-5.17	[0]
200-239	5.18-6.21	[1]
240-279	6.22-7.24	[2]
≥280	≥7.25	[3]

Step 3

HDL - C			
(mg/dl)	(mmol/L)	LDL Pts	Chol Pts
<35	<0.90	2	[2]
35-44	0.91-1.16	1	[1]
45-49	1.17-1.29	0	[0]
50-59	1.30-1.55	0	[0]
≥60	≥1.56	-1	[-2]

Step 4

Blood Pressure					
Systolic (mm Hg)	Diastolic (mm Hg) <80	80-84	85-89	90-99	≥100
<120	0 [0] pts				
120-129		0 [0] pts			
130-139			1 [1] pts		
140-159				2 [2] pts	
≥160					3 [3] pts

Note: When systolic and diastolic pressures provide different estimates for point scores, use the higher number

Step 5

Diabetes		
	LDL Pts	Chol Pts
No	0	[0]
Yes	2	[2]

Step 6

Smoker		
	LDL Pts	Chol Pts
No	0	[0]
Yes	2	[2]

(sum from steps 1-6)

Step 7

Adding up the points	
Age	_____
LDL-C or Chol	_____
HDL - C	
Blood Pressure	_____
Diabetes	_____
Smoker	_____
Point total	_____

Key	
Color	Relative Risk
green	Very low
white	Low
yellow	Moderate
rose	High
red	Very high

(determine CHD risk from point total)

Step 8

CHD Risk			
LDL Pts Total	10 Yr CHD Risk	Chol Pts Total	10 Yr CHD Risk
<-3	1%		
-2	2%		
-1	2%	[<-1]	[2%]
0	3%	[0]	[3%]
1	4%	[1]	[3%]
2	4%	[2]	[4%]
3	6%	[3]	[5%]
4	7%	[4]	[7%]
5	9%	[5]	[8%]
6	11%	[6]	[10%]
7	14%	[7]	[13%]
8	18%	[8]	[16%]
9	22%	[9]	[20%]
10	27%	[10]	[25%]
11	33%	[11]	[31%]
12	40%	[12]	[37%]
13	47%	[13]	[45%]
≥14	≥56%	[≥14]	[≥53%]

(compare to average person your age)

Step 9

Comparative Risk			
Age (years)	Average 10 Yr CHD Risk	Average 10 Yr Hard* CHD Risk	Low** 10 Yr CHD Risk
30-34	3%	1%	2%
35-39	5%	4%	3%
40-44	7%	4%	4%
45-49	11%	8%	4%
50-54	14%	10%	6%
55-59	16%	13%	7%
60-64	21%	20%	9%
65-69	25%	22%	11%
70-74	30%	25%	14%

* Hard CHD events exclude angina pectoris

** Low risk was calculated for a person the same age, optimal blood pressure, LDL-C 100-129 mg/dL or cholesterol 160-199 mg/dl, HDL-C 45 mg/dL for men or 55 mg/dL for women, non-smoker, no diabetes

Risk estimates were derived from the experience of the Framingham Heart Study, a predominantly Caucasian population in Massachusetts, USA

Peter W. F. Wilson et al. Circulation. 1998;97:1837-1847

American Heart Association.

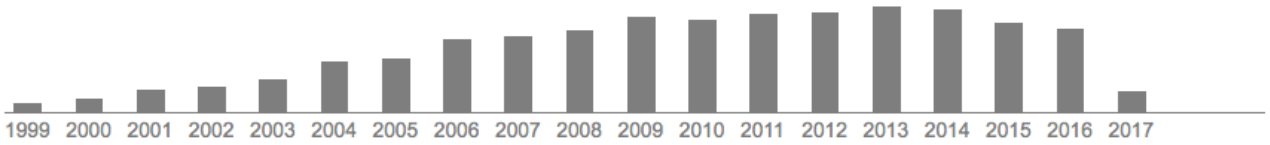
Case study on transferability: Framingham CHD risk score

- Many ML models are trained in one place and deployed more broadly
- **Example:** Framingham coronary heart disease (CHD) risk score

Prediction of coronary heart disease using risk factor categories

[HTML] from ahajournals.org
Full text - MIT Libraries

Authors	Peter WF Wilson, Ralph B D'Agostino, Daniel Levy, Albert M Belanger, Halit Silbershatz, William B Kannel
Publication date	1998/5/1
Journal	Circulation
Volume	97
Issue	18
Pages	1837-1847
Publisher	Lippincott Williams & Wilkins
Description	Background—The objective of this study was to examine the association of Joint National Committee (JNC-V) blood pressure and National Cholesterol Education Program (NCEP) cholesterol categories with coronary heart disease (CHD) risk, to incorporate them into coronary prediction algorithms, and to compare the discrimination properties of this approach with other noncategorical prediction functions. Methods and Results—This work was designed as a prospective, single-center study in the setting of a community-based ...
Total citations	Cited by 8422



Case study on transferability: Framingham CHD risk score

- Many ML models are trained in one place and deployed more broadly
- **Example:** Framingham coronary heart disease (CHD) risk score
  - 99% of Framingham participants are of European descent
  - How well does it generalize to a Chinese population?

- C-statistic (=AUC on censored data) 0.705/0.742 (M/F)
- Re-fit using local data only slightly improves C-statistic (=AUC on censored data), to 0.736/0.759 (M/F)

[Liu et al., JAMA '04]

Could we say the same about our more complex machine learning models?

What would we need to look at to get confidence that they would transfer as well?

CASE STUDY 3

Improving clinical documentation with "auto-complete"

Improving Quality of Structured Data

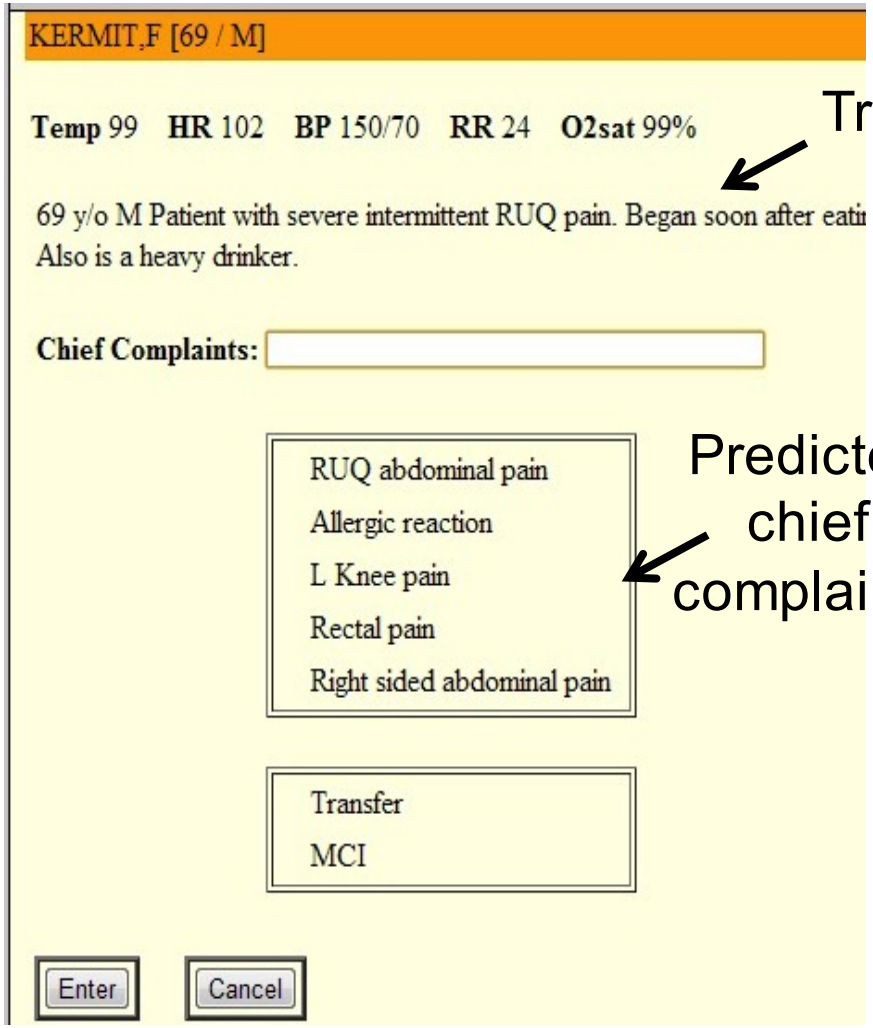
- Much of the valuable data in EHRs is in the form of free text notes
- Collecting structured data is slow and error-prone
- We can make the process faster and more accurate:

  1. Automatically keeping problem lists up to date
  2. Assigning diagnosis codes and other documentation
  3. Assigning chief complaints

  by leveraging the text data in a patient's EHR

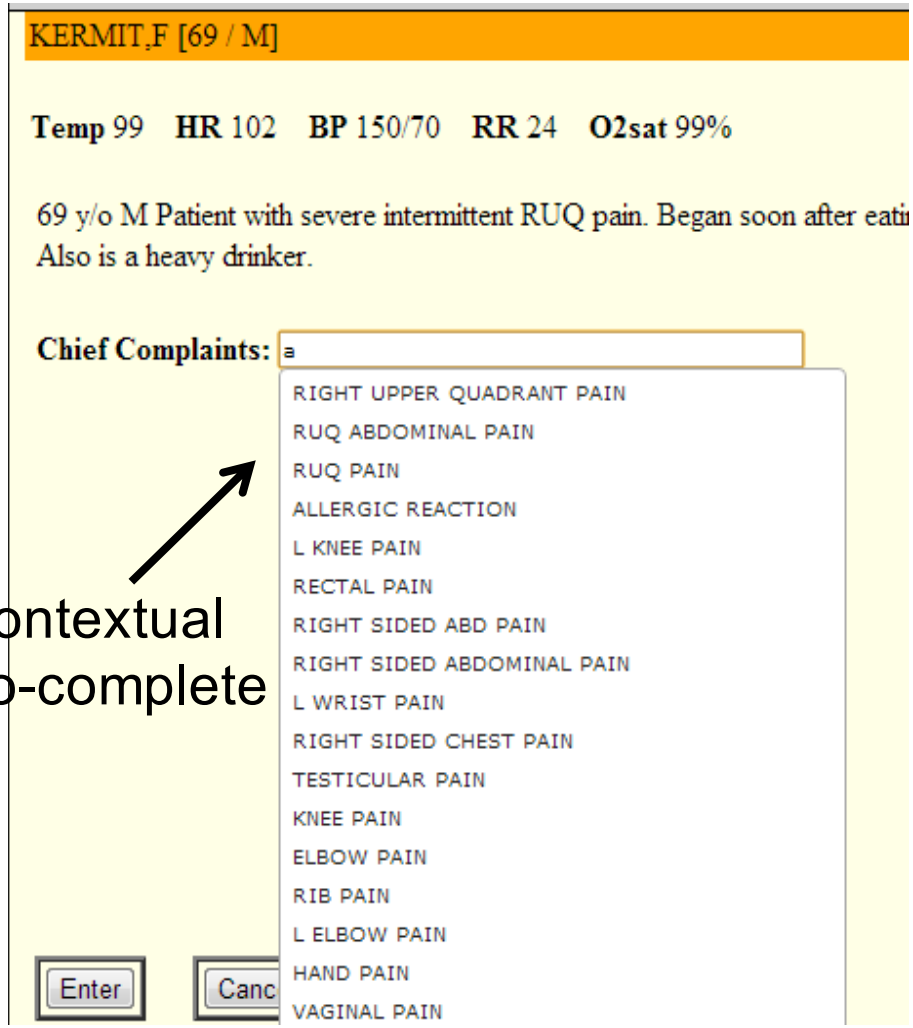
Example: Chief complaints

Changed workflow to have chief complaints assigned *last*. Predict them.



Triage note

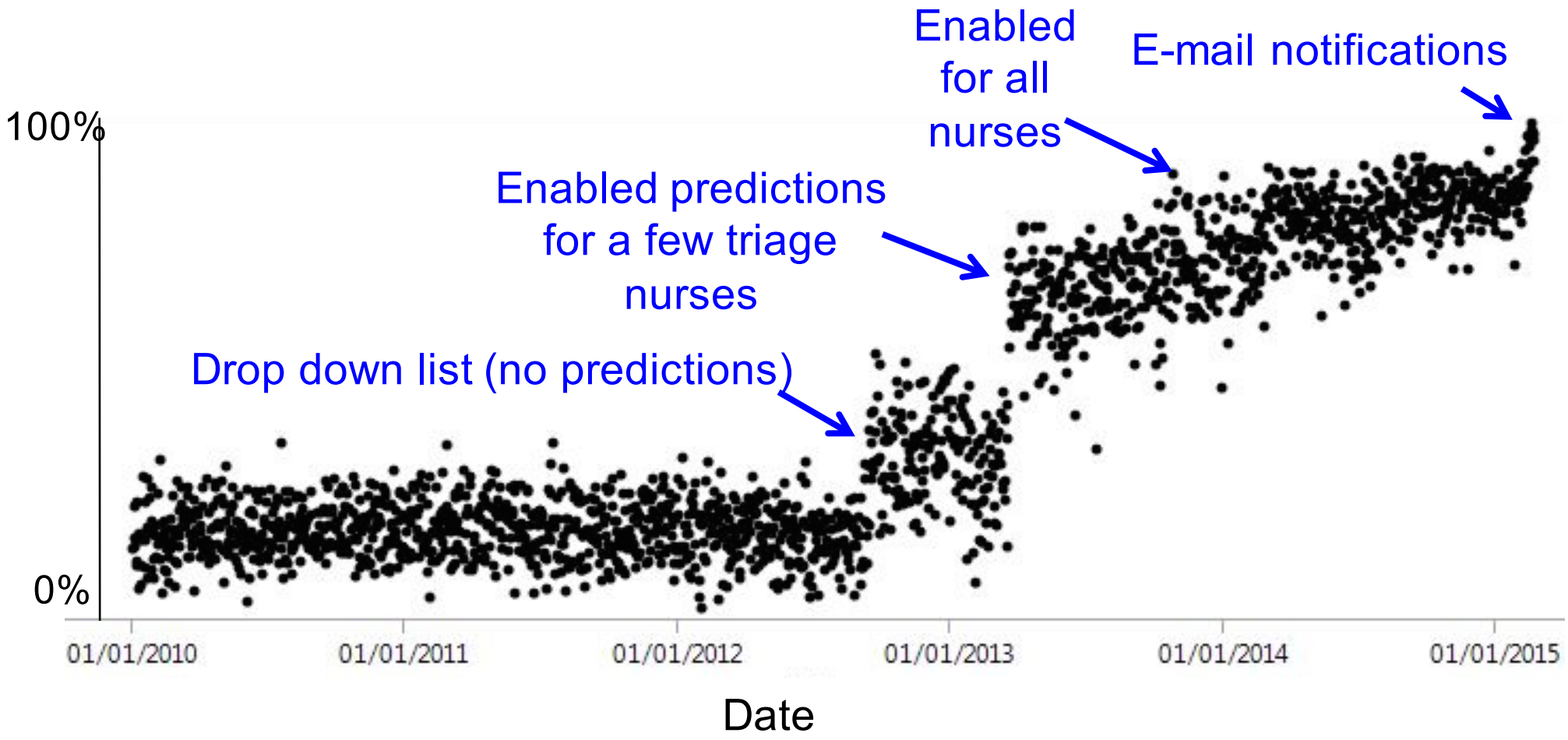
Predicted chief complaints



Contextual auto-complete

Using for all 55,000 patients/year that present at BIDMC ED

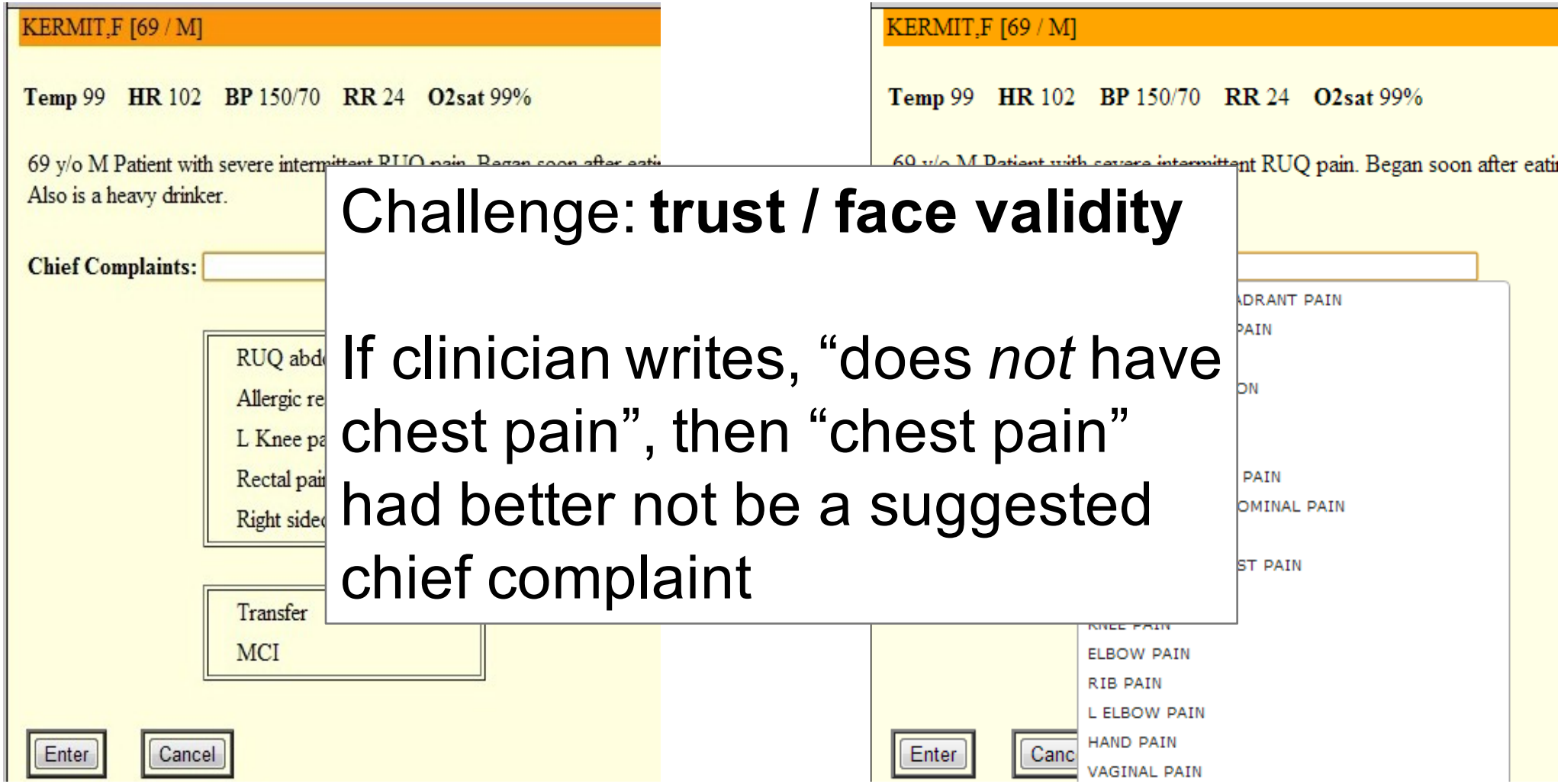
Example: Chief complaints



Percentage of *standardized* chief complaints (per week)

Example: Chief complaints

Changed workflow to have chief complaints assigned *last*. Predict them.



Challenge: **trust / face validity**

If clinician writes, “does *not* have chest pain”, then “chest pain” had better not be a suggested chief complaint

Using for all 55,000 patients/year that present at BIDMC ED

Conclusions

- Model introspection seems to be essential for using machine learning in health care
- Fertile terrain to develop new ML methods for directly tackling these issues where "interpretability" arises:
  - Building trust
  - Checking that prediction task is set up properly
  - Identifying causal hypotheses
  - Getting the gist of what is predictive
  - Assessing transferability