Imperial College London

AI in Medicine I

Risk scores and stratification

Julia Schnabel I32 – Chair for Computational Imaging and AI in Medicine
School of Computation, Information & Technology

Slides courtesy of Daniel Rueckert –
Some of the slides adopted from David Sontag, MIT

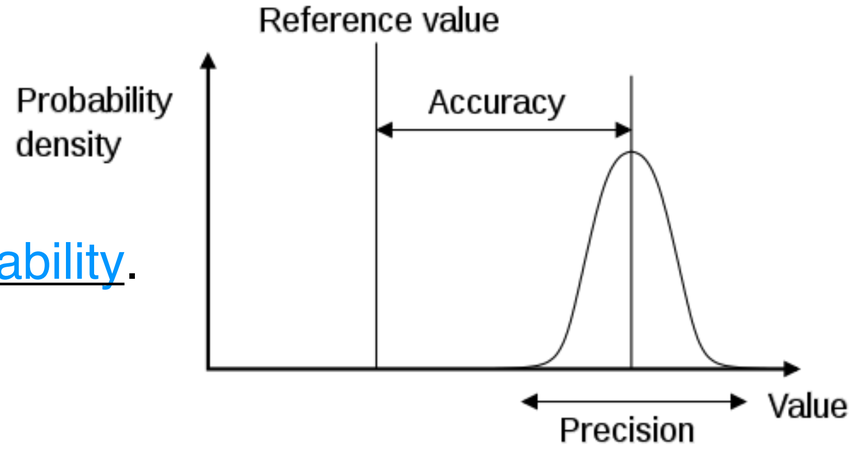
Outline

- Evaluation of diagnostic tests:
  - Accuracy, sensitivity, specificity
  - Positive and negative predictive values
  - Relative risk
  - Odds ratio
- Risk scores and stratification
- Case study: Early detection of Type 2 diabetes
  - Framing as supervised learning problem
  - Evaluating risk stratification algorithms

Outline

- **Evaluation of diagnostic tests:**
  - **Accuracy, sensitivity, specificity**
  - **Positive and negative predictive values**
  - **Relative risk**
  - **Odds ratio**
- Risk scores and stratification
- Case study: Early detection of Type 2 diabetes
  - Framing as supervised learning problem
  - Evaluating risk stratification algorithms

How to assess diagnostic tests?



- **Precision**
  - is a description of random errors, a measure of statistical variability.
  - the repeatability, or reproducibility of the measurement
- **Accuracy** (two definitions)
  - Description of systematic errors, a measure of statistical bias; as these cause a difference between a result and a "true" value, ISO calls this trueness.
  - Alternatively, ISO defines accuracy as describing a combination of both types of observational error above (random and systematic), so high accuracy requires both high precision and high trueness
- **Robustness**
  - refers to the degradation in performance with respect to varying noise levels or other measurement artefacts

How to assess diagnostic tests: Confusion Matrix

- True positive (TP)
  – eqv. with hit
- True negative (TN)
  – eqv. with correct rejection
- False positive (FP)
  – eqv. with false alarm, Type I error
- False negative (FN)
  – eqv. with miss, Type II error

		Predicted condition	
	Total population = P + N	**Positive (PP)**	**Negative (PN)**
Actual condition	**Positive (P)**	**True positive (TP),** hit	**False negative (FN),** type II error, miss, underestimation
	Negative (N)	**False positive (FP),** type I error, false alarm, overestimation	**True negative (TN),** correct rejection

Accuracy, Precision, Recall, ...

Accuracy

$$ACC = \frac{TP+TN}{P+N},\ P = TP + FN,\ N = TN + FP$$

Precision or **positive predictive value**

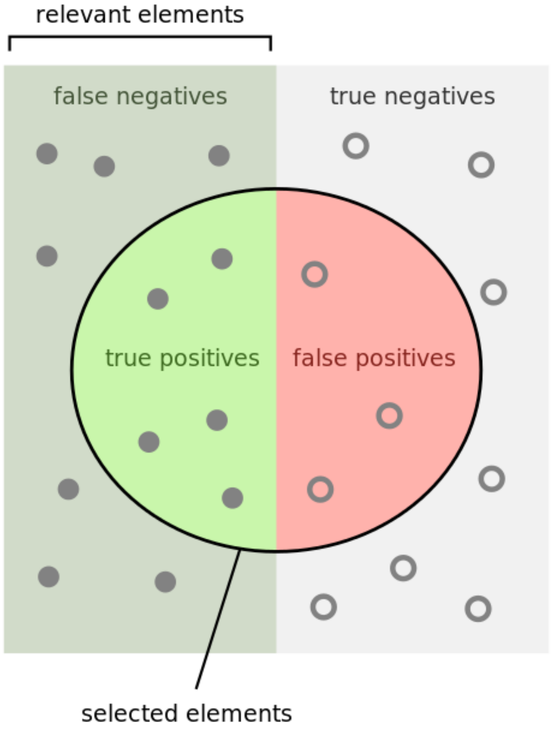
$$PPV = \frac{TP}{TP + FP}$$

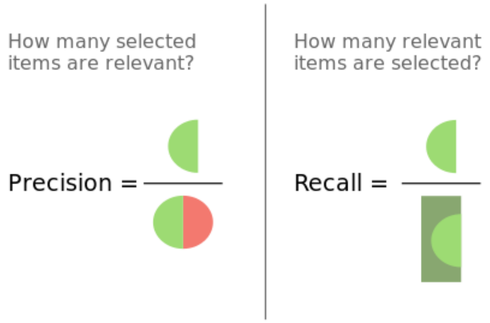
Recall, sensitivity, **hit rate** or **true positive rate**

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$$

Specificity or **true negative rate**

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP}$$





Positive and negative predictive values

- Positive and negative predictive values (PPV and NPV respectively) are the proportions of positive and negative results in statistics
- They describe the diagnostic tests that are true positive and true negative results, respectively.
- PPV and NPV describe the performance of a diagnostic test or other statistical measure.
  - A high result can be interpreted as indicating the accuracy of such a statistic.
  - PPV and NPV are not intrinsic to the test; they depend also on the **prevalence**.

Positive predictive value

$$\mathrm{PPV} = \frac{\text{Number of true positives}}{\text{Number of true positives} + \text{Number of false positives}} = \frac{\text{Number of true positives}}{\text{Number of positive calls}} = \frac{TP}{TP+FP}$$

$$\mathrm{PPV} = \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + (1 - \text{specificity}) \times (1 - \text{prevalence})}$$

Since: Prevalence: $\frac{P}{P+N}$ Sensitivity: $\frac{TP}{P}$ Specificity: $\frac{TN}{N}$

$$\mathrm{FDR} = 1 - \mathrm{PPV} = \frac{\text{Number of false positives}}{\text{Number of true positives} + \text{Number of false positives}} = \frac{\text{Number of false positives}}{\text{Number of positive calls}}$$

Complement of PPV: False discovery rate

Negative predictive value

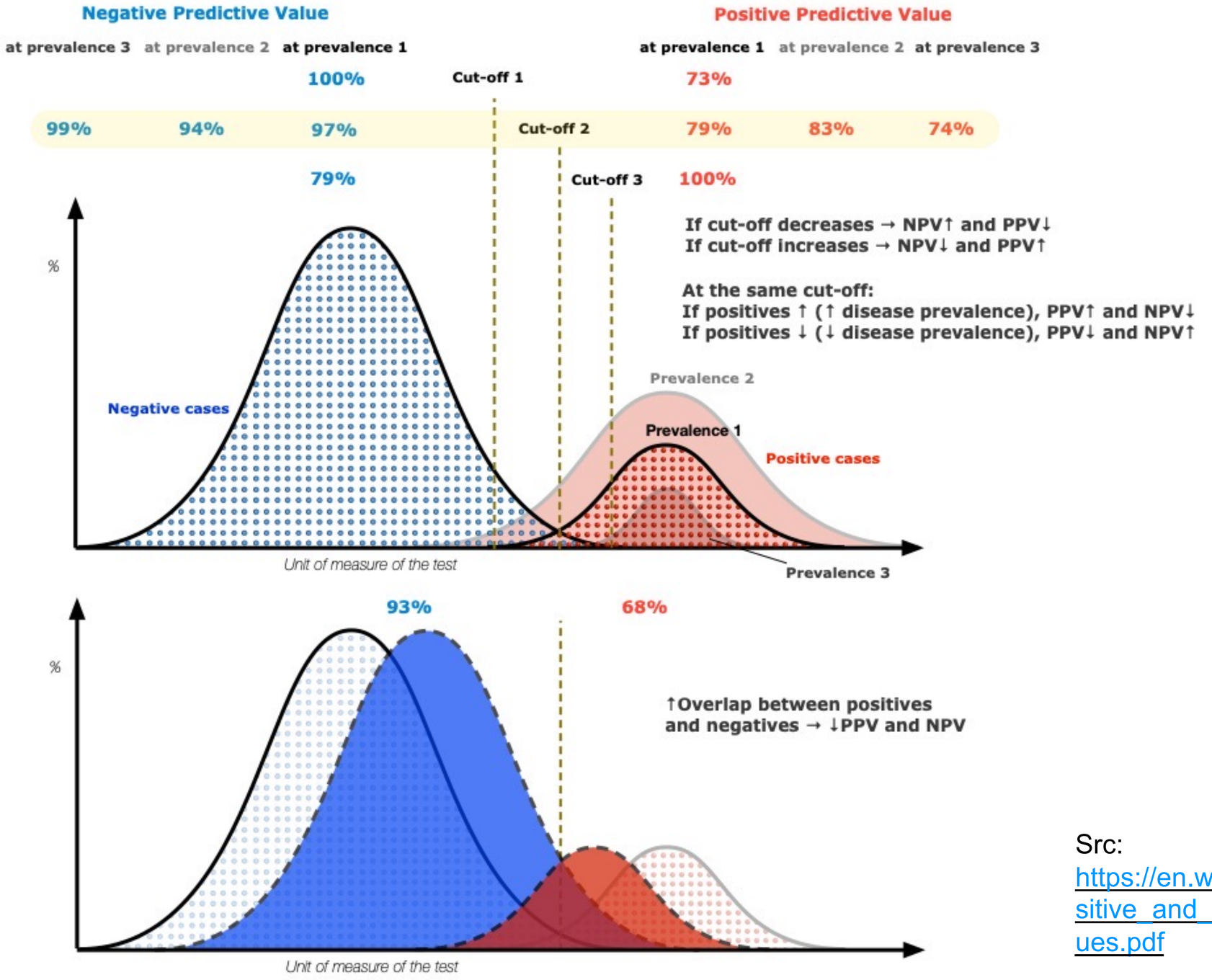
$$\mathrm{NPV} = \frac{\text{Number of true negatives}}{\text{Number of true negatives} + \text{Number of false negatives}} = \frac{\text{Number of true negatives}}{\text{Number of negative calls}}$$

$$\mathrm{NPV} = \frac{\text{specificity} \times (1 - \text{prevalence})}{\text{specificity} \times (1 - \text{prevalence}) + (1 - \text{sensitivity}) \times \text{prevalence}}$$

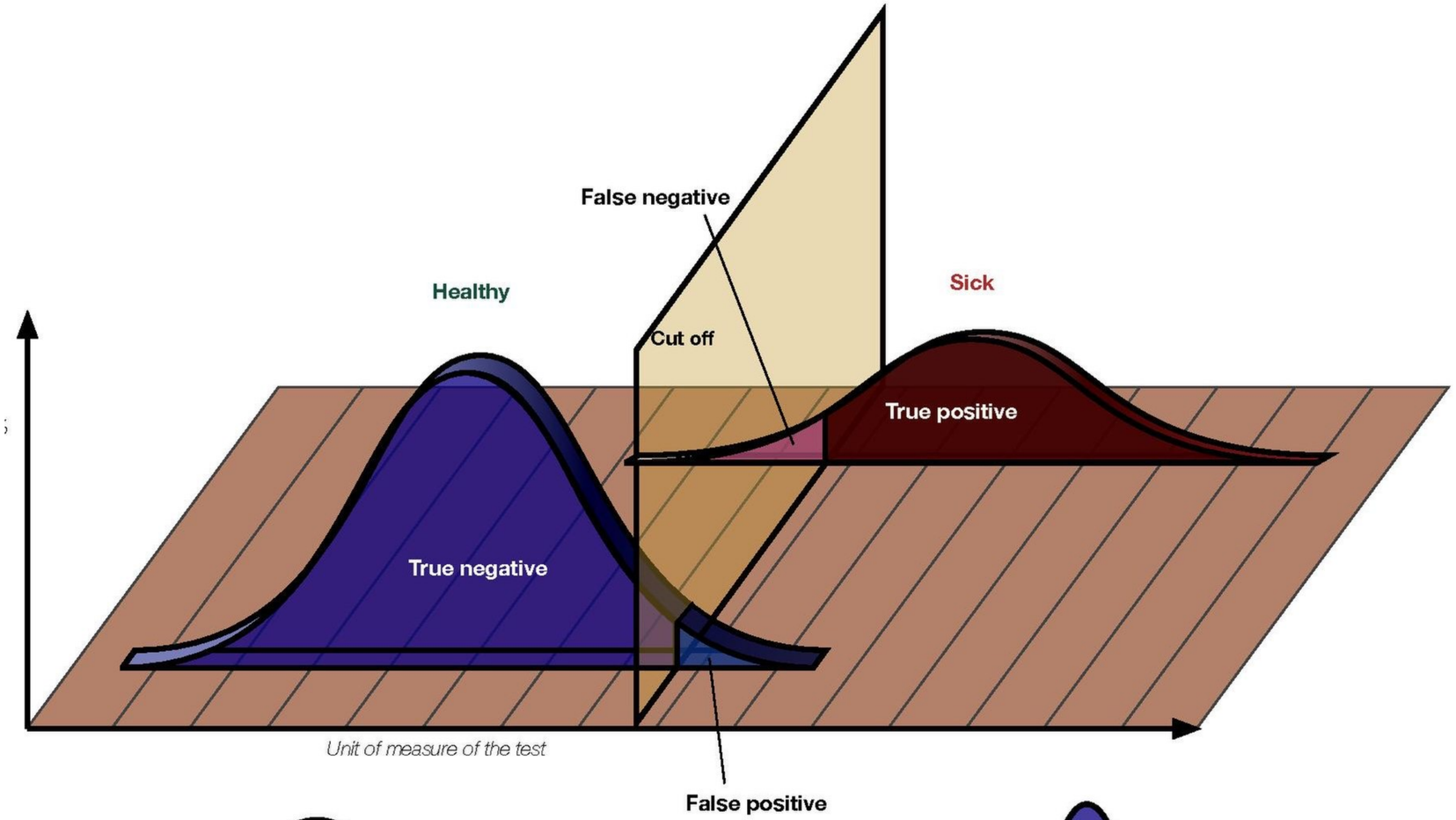
Since: Prevalence: $\frac{P}{P+N}$ Sensitivity: $\frac{TP}{P}$ Specificity: $\frac{TN}{N}$

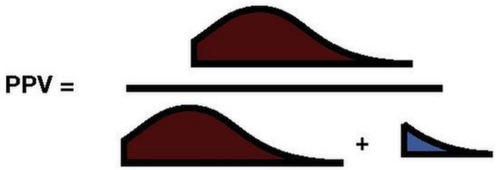
$$\mathrm{FOR} = 1 - \mathrm{NPV} = \frac{\text{Number of false negatives}}{\text{Number of true negatives} + \text{Number of false negatives}} = \frac{\text{Number of false negatives}}{\text{Number of negative calls}}$$

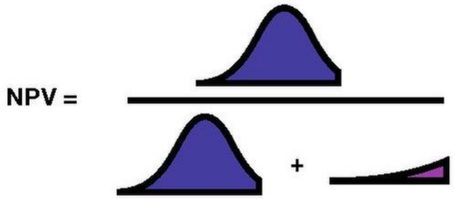
Complement of NPV: False omission rate

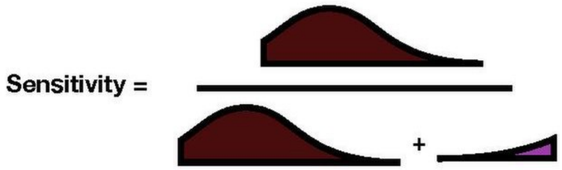


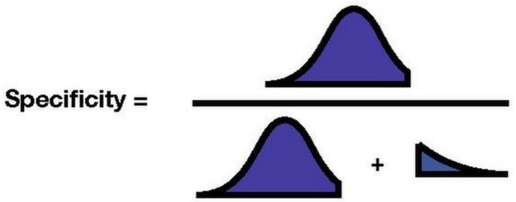
Src: https://en.wikipedia.org/wiki/File:Positive_and_negative_predictive_values.pdf











Src: https://en.wikipedia.org/wiki/Positive_and_negative_predictive_values#/media/File:PPV,_NPV,_Sensitivity_and_Specificity.svg

Source: Wikipedia, see https://en.wikipedia.org/wiki/Positive_and_negative_predictive_values

		Predicted condition			
	Total population $= P + N$	**Positive (PP)**	**Negative (PN)**	Informedness, bookmaker informedness (BM) $= TPR + TNR - 1$	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Positive (P)**	**True positive (TP),** hit	**False negative (FN),** type II error, miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
	Negative (N)	**False positive (FP),** type I error, false alarm, overestimation	**True negative (TN),** correct rejection	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN}$ $= 1 - FOR$	Markedness (MK), deltaP (Δp) $= PPV + NPV - 1$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
	Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$	F_1 score $= \frac{2PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP + FN + FP}$

Source: Wikipedia, see https://en.wikipedia.org/wiki/Positive_and_negative_predictive_values

		Fecal occult blood screen test outcome			
	Total population (pop.) = 2030	Test outcome **positive**	Test outcome **negative**	Accuracy (ACC) = (TP + TN) / pop. = (20 + 1820) / 2030 ≈ **90.64%**	F_1 score $= 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$ ≈ **0.174**
Patients with bowel cancer (as confirmed on endoscopy)	Actual condition positive	**True positive** (TP) = 20 (2030 × 1.48% × 67%)	**False negative** (FN) = 10 (2030 × 1.48% × (100% – 67%))	True positive rate (TPR), recall, sensitivity = TP / (TP + FN) = 20 / (20 + 10) ≈ **66.7%**	False negative rate (FNR), miss rate = FN / (TP + FN) = 10 / (20 + 10) ≈ **33.3%**
	Actual condition negative	**False positive** (FP) = 180 (2030 × (100% – 1.48%) × (100% – 91%))	**True negative** (TN) = 1820 (2030 × (100% – 1.48%) × 91%)	False positive rate (FPR), fall-out, probability of false alarm = FP / (FP + TN) = 180 / (180 + 1820) = **9.0%**	Specificity, selectivity, true negative rate (TNR) = TN / (FP + TN) = 1820 / (180 + 1820) = **91%**
	Prevalence = (TP + FN) / pop. = (20 + 10) / 2030 ≈ **1.48%**	Positive predictive value (PPV), precision = TP / (TP + FP) = 20 / (20 + 180) = **10%**	False omission rate (FOR) = FN / (FN + TN) = 10 / (10 + 1820) ≈ **0.55%**	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$ = (20 / 30) / (180 / 2000) ≈ **7.41**	Negative likelihood ratio (LR–) $= \frac{\text{FNR}}{\text{TNR}}$ = (10 / 30) / (1820 / 2000) ≈ **0.366**
		False discovery rate (FDR) = FP / (TP + FP) = 180 / (20 + 180) = **90.0%**	Negative predictive value (NPV) = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.45%**	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR–}}$ ≈ **20.2**	

Relative risk

- The **relative risk** (RR) or **risk ratio** is the ratio of the probabilities

$$RR = \frac{P(outcome|exposure)}{P(outcome|no\ exposure)}$$

- Here exposure is used to mean
  - intervention or treatment
- Assuming the causal effect between the exposure and the outcome, values of relative risk can be interpreted as follows:
  - RR = 1 means that exposure does not affect the outcome
  - RR < 1 means that the risk of the outcome is decreased by the exposure, which is a "protective factor"
  - RR > 1 means that the risk of the outcome is increased by the exposure, which is a "risk factor"

Relative risk

- Suppose a radiation leak in a village* of 1,000 people increased the incidence of a rare disease.
  - Total number of people exposed to the radiation was $V_E = 400$, out of which $D_E = 20$ developed the disease, and $H_E = 380$ stayed healthy.
  - The total number of people not exposed was $V_N = 600$ out of which $D_N = 6$ developed the disease and $H_N = 594$ stayed healthy.

	Diseased	Healthy
Exposed	20	380
Not exposed	6	594

$$RR = \frac{D_E/V_E}{D_N/V_N} = \frac{20/400}{6/600} = \frac{0.05}{0.01} = 5$$

The *risk* of developing the disease given exposure is $D_E/V_E = 20/400 = .05$ and of developing the disease given non-exposure is $D_N/V_N = 6/600 = .01$

Odds ratio

- Relative risk is requires knowledge of the **prevalence** which requires knowledge about the whole population, not just a random sample.

- The **odds ratio** looks at the ratio of the odds of getting the disease if exposed and the odds if not exposed:

$$OR = \frac{D_E/H_E}{D_N/H_N} = \frac{20/380}{6/594} \approx \frac{0.052}{0.010} = 5.2$$

- In a rare-disease case like this, the **relative risk** and the **odds ratio** are almost the same.

Outline

- Evaluation of diagnostic tests:
  - Accuracy, sensitivity, specificity
  - Positive and negative predictive values
  - Relative risk
  - Odds ratio

- **Risk scores and stratification**

- Case study: Early detection of Type 2 diabetes
  - Framing as supervised learning problem
  - Evaluating risk stratification algorithms

What is a risk score?

- At its fundamental level, a risk score is a standardised metric for the likelihood that an individual will experience a particular outcome.

- In healthcare, these outcomes can include
  - events, such as hospital admissions and emergency department visits, or
  - development of a certain clinical state, such as heart disease, diabetes, cancer, or sepsis.

What is a risk score?

- Traditionally, risk stratification was based on simple scores using simple measurements

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

Mnemonic for the test is
"**H**ow **R**eady **I**s **T**his **C**hild":
Heart rate, **R**espiratory effort,
Irritability, **T**one, and **C**olour

Virginia Apgar, American physician who introduced the Apgar Score

Conventional vs. ML-based risk scores

- Traditionally, risk stratification was based on simple scores using simple measurements

- Now, based on machine learning it can be on complex, high-dimensional measurements
  - Fits more easily into workflow
  - Higher accuracy
  - Quicker to derive (special case)

- *But, new dangers may be introduced with AI-based approaches*

What is a risk score?

Epidemiology

General Cardiovascular Risk Profile for Use in Primary Care
The Framingham Heart Study

Ralph B. D'Agostino, Sr, PhD; Ramachandran S. Vasan, MD; Michael J. Pencina, PhD; Philip A. Wolf, MD; Mark Cobain, PhD; Joseph M. Massaro, PhD; William B. Kannel, MD

Background—Separate multivariable risk algorithms are commonly used to assess risk of specific atherosclerotic cardiovascular disease (CVD) events, ie, coronary heart disease, cerebrovascular disease, peripheral vascular disease, and heart failure. The present report presents a single multivariable risk function that predicts risk of developing all CVD and of its constituents.

Methods and Results—We used Cox proportional-hazards regression to evaluate the risk of developing a first CVD event in 8491 Framingham study participants (mean age, 49 years; 4522 women) who attended a routine examination between 30 and 74 years of age and were free of CVD. Sex-specific multivariable risk functions ("general CVD" algorithms) were derived that incorporated age, total and high-density lipoprotein cholesterol, systolic blood pressure, treatment for hypertension, smoking, and diabetes status. We assessed the performance of the general CVD algorithms for predicting individual CVD events (coronary heart disease, stroke, peripheral artery disease, or heart failure). Over 12 years of follow-up, 1174 participants (456 women) developed a first CVD event. All traditional risk factors evaluated predicted CVD risk (multivariable-adjusted $P<0.0001$). The general CVD algorithm demonstrated good discrimination (C statistic, 0.763 [men] and 0.793 [women]) and calibration. Simple adjustments to the general CVD risk algorithms allowed estimation of the risks of each CVD component. Two simple risk scores are presented, 1 based on all traditional risk factors and the other based on non–laboratory-based predictors.

Conclusions—A sex-specific multivariable risk factor algorithm can be conveniently used to assess general CVD risk and risk of individual CVD events (coronary, cerebrovascular, and peripheral arterial disease and heart failure). The estimated absolute CVD event rates can be used to quantify risk and to guide preventive care. **(*Circulation*. 2008;117: 743-753.)**

Key Words: cardiovascular diseases ■ coronary disease ■ heart failure ■ risk factors ■ stroke

It is widely accepted that age, sex, high blood pressure, smoking, dyslipidemia, and diabetes are the major risk factors for developing cardiovascular disease (CVD).[1] It also is recognized that CVD risk factors cluster and interact multiplicatively to promote vascular risk.[2] This knowledge led to the development of multivariable risk prediction algorithms incorporating these risk factors that can be used by primary care physicians to assess in individual patients the risk of developing all atherosclerotic CVD[3–12] or specific components of CVD, ie, coronary heart disease,[9,13–17] stroke,[18] peripheral vascular disease,[19] or heart failure.[20] Multivariable assessment has been advocated to estimate absolute CVD risk and to guide treatment of risk factors.[2,6] For instance, the Framingham formulation for predicting coronary heart disease (CHD) was incorporated into the Third Report of the Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III).[9] The Framingham CHD risk assessment tool has been validated in whites and blacks in the United States[9,10,21] and are transportable (with calibration) to culturally diverse populations in Europe, the Mediterranean region, and Asia.[9,10,22,23] Similar CHD risk prediction algorithms have been developed by other investigators worldwide and have been demonstrated to perform well.[14,15,17]

Clinical Perspective p 753

Despite the availability of several validated risk prediction algorithms, their use has lagged in primary care.[24] One potential reason for physician inertia in using risk prediction instruments is the multiplicity of such algorithms, each for

Received February 27, 2007; accepted November 30, 2007.

From Boston University, Department of Mathematics and Statistics (R.B.D., M.J.P.), School of Medicine (R.S.V., P.A.W., W.B.K.), and Department of Biostatistics (J.M.M.), Boston, Mass; Framingham Heart Study, Framingham, Mass (R.B.D., R.S.V., M.J.P., P.A.W., J.M.M., W.B.K.); and Unilever Research, Corporate Biology, Colworth Park, UK (M.C.).

Guest Editor for this article was Eric B. Rimm, ScD.

The online Data Supplement can be found with this article at http://circ.ahajournals.org/cgi/content/full/CIRCULATIONAHA.107.699579/DC1.

Correspondence to R.B. D'Agostino, PhD, Chairman, Professor of Mathematics/Statistics and Public Health, Boston University, Department of Mathematics and Statistics, 111 Cummington St, Boston, MA 02215.

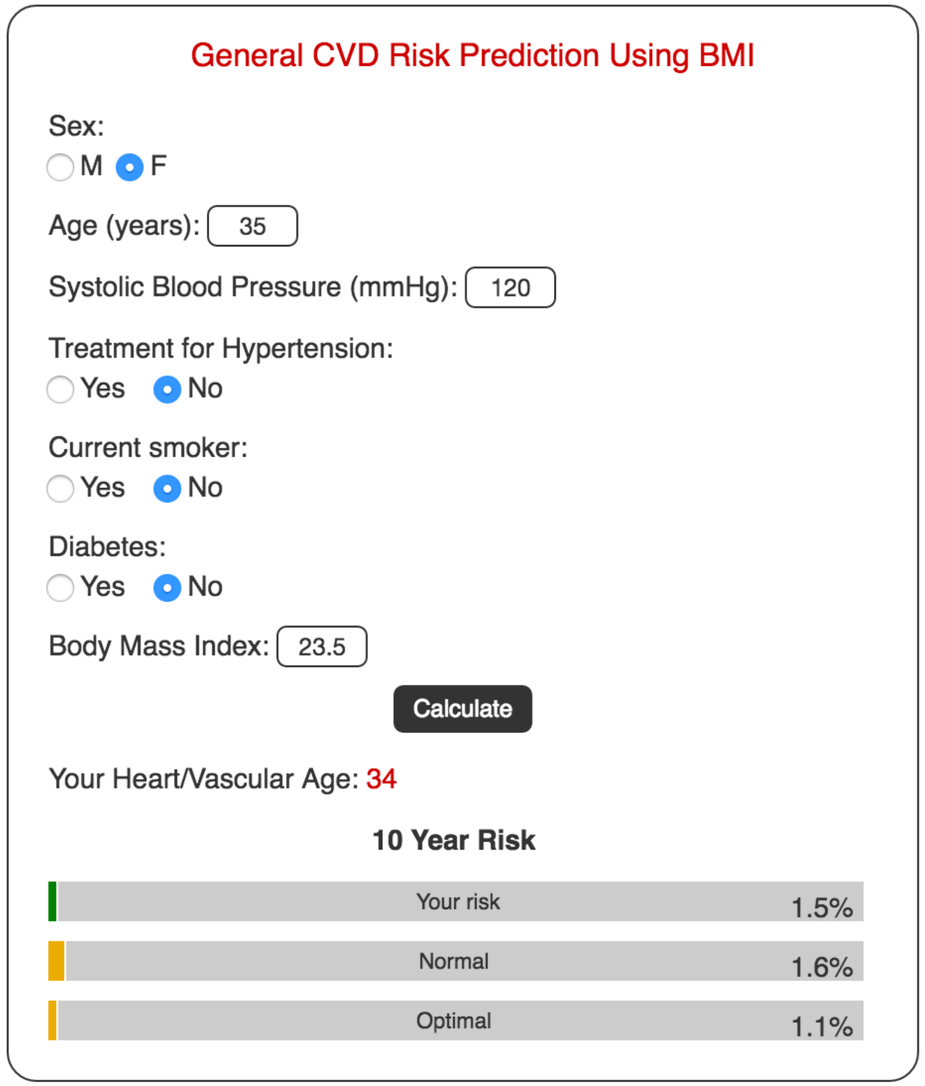
***Circulation* is available at http://circ.ahajournals.org** **DOI: 10.1161/CIRCULATIONAHA.107.699579**

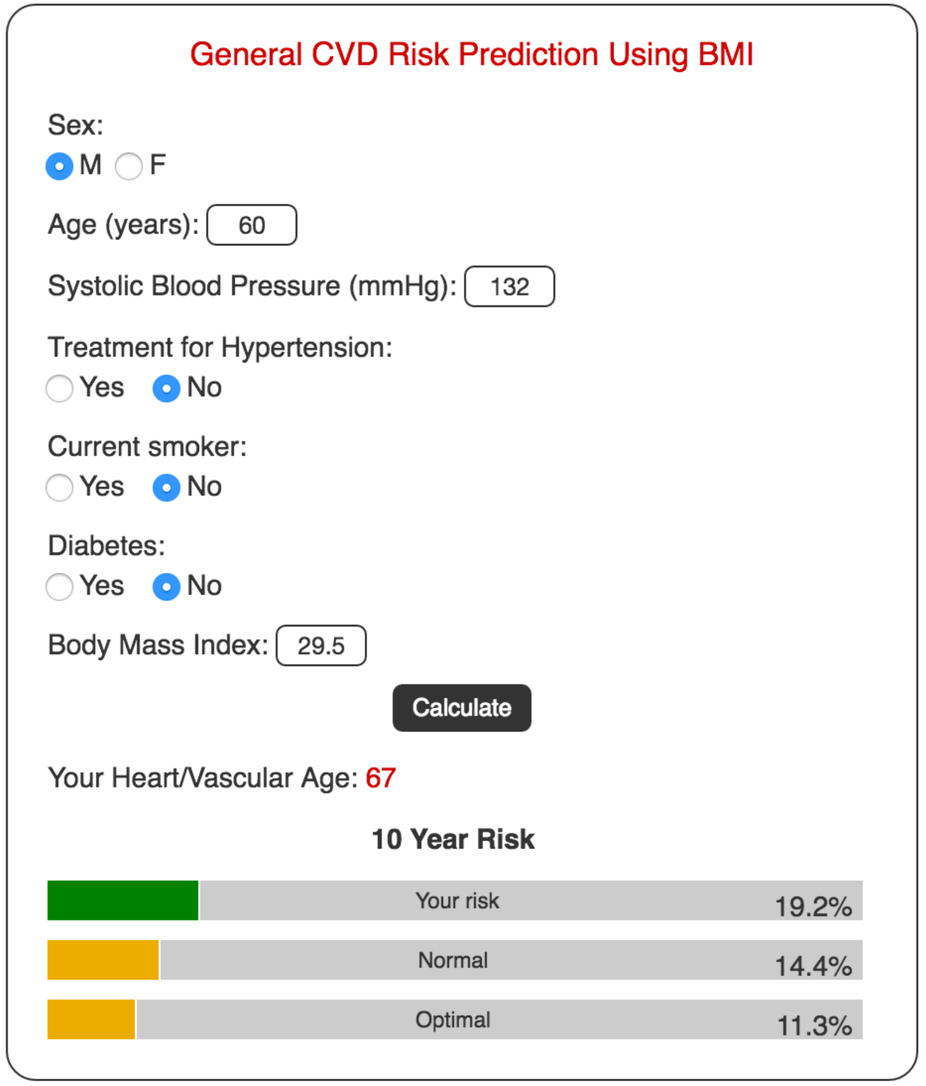
Example: Framingham CVD Risk Score

Characteristics	Women (n=4522, 28% FOC)	Men (n=3969, 22% FOC)
Age, mean (SD), y	49.1 (11.1)	48.5 (10.8)
Total-C, mean (SD), mg/dL	215.1 (44.1)	212.5 (39.3)
HDL-C, mean (SD), mg/dL	57.6 (15.3)	44.9 (12.2)
Systolic BP, mean (SD), mm Hg	125.8 (20.0)	129.7 (17.6)
BP treatment, n (%)	532 (11.76)	402 (10.13)
Smoking, n (%)	1548 (34.23)	1398 (35.22)
Diabetes, n (%)	170 (3.76)	258 (6.50)
Incident CVD events, n (%)	456 (10.08)	718 (18.09)

FOC indicates Framingham original cohort; Total-C, total cholesterol; HDL-C, HDL cholesterol; and BP, blood pressure.

What is a risk score?





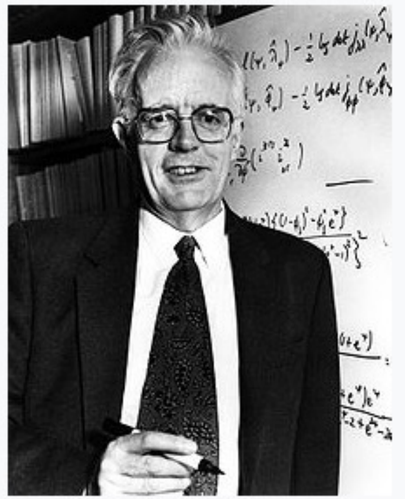
Example: Framingham CVD Risk Score

- Uses a Cox proportional-hazards regression model

$$h(t) = 1 - \lambda_0(t)^{\exp(\sum_{i=1}^{p} \beta_i x_i - \sum_{i=1}^{p} \beta_i \bar{x}_i)}$$

  - $\lambda_0(t)$ is baseline survival at follow-up time t
  - β_i is the **estimated regression coefficient** of the i-th risk factor
  - x_i is the **log-transformed value** of the i-th risk factor

Sir
David Cox
FRS FBA FRSE FRSC



Cox in 1980

Regression Models and Life-Tables

By D. R. Cox

Imperial College, London

[Read before the Royal Statistical Society, at a meeting organized by the Research Section, on Wednesday, March 8th, 1972, Mr M. J. R. Healy in the Chair]

Summary

The analysis of censored failure times is considered. It is assumed that on each individual are available values of one or more explanatory variables. The hazard function (age-specific failure rate) is taken to be a function of the explanatory variables and unknown regression coefficients multiplied by an arbitrary and unknown function of time. A conditional likelihood is obtained, leading to inferences about the unknown regression coefficients. Some generalizations are outlined.

Keywords: LIFE TABLE; HAZARD FUNCTION; AGE-SPECIFIC FAILURE RATE; PRODUCT LIMIT ESTIMATE; REGRESSION; CONDITIONAL INFERENCE; ASYMPTOTIC THEORY; CENSORED DATA; TWO-SAMPLE RANK TESTS; MEDICAL APPLICATIONS; RELIABILITY THEORY; ACCELERATED LIFE TESTS.

Example: Framingham CVD Risk Score

- Regression coefficients and hazard ratios

Variable	β*	P	Hazard Ratio	95% CI
Women [So(10)=0.95012]				
Log of age	2.32888	<0.0001	10.27	(5.65–18.64)
Log of total cholesterol	1.20904	<0.0001	3.35	(2.00–5.62)
Log of HDL cholesterol	−0.70833	<0.0001	0.49	(0.35–0.69)
Log of SBP if not treated	2.76157	<0.0001	15.82	(7.86–31.87)
Log of SBP if treated	2.82263	<0.0001	16.82	(8.46–33.46)
Smoking	0.52873	<0.0001	1.70	(1.40–2.06)
Diabetes	0.69154	<0.0001	2.00	(1.49–2.67)
Men [So(10)=0.88936]				
Log of age	3.06117	<0.0001	21.35	(14.03–32.48)
Log of total cholesterol	1.12370	<0.0001	3.08	(2.05–4.62)
Log of HDL cholesterol	−0.93263	<0.0001	0.39	(0.30–0.52)
Log of SBP if not treated	1.93303	<0.0001	6.91	(3.91–12.20)
Log of SBP if treated	1.99881	<0.0001	7.38	(4.22–12.92)
Smoking	0.65451	<0.0001	1.92	(1.65–2.24)
Diabetes	0.57367	<0.0001	1.78	(1.43–2.20)

So(10) indicates 10-year baseline survival; SBP, systolic blood pressure.
*Estimated regression coefficient

Example: Framingham CVD Risk Score

Points	Age, y	HDL	Total Cholesterol	SBP Not Treated	SBP Treated	Smoker	Diabetic	
−3				<120				
−2		60+						
−1		50–59			<120			
0	30–34	45–49	<160	120–129		No	No	
1		35–44	160–199	130–139				
2	35–39	<35		140–149	120–129			
3			200–239		130–139	Yes		
4	40–44		240–279	150–159			Yes	
5	45–49		280+	160+	140–149			
6					150–159			
7	50–54				160+			
8	55–59							
9	60–64							
10	65–69							
11	70–74							
12	75+							
Points allotted								Total

SBP indicates systolic blood pressure.

Example: Framingham CVD Risk Score

Points	Risk, %
≤−2	<1
−1	1.0
0	1.2
1	1.5
2	1.7
3	2.0
4	2.4
5	2.8
6	3.3
7	3.9
8	4.5
9	5.3
10	6.3
11	7.3
12	8.6
13	10.0
14	11.7
15	13.7
16	15.9
17	18.5
18	21.5
19	24.8
20	28.5
21+	>30

FRAMINGHAM RISK SCORE (FRS)

Estimation of 10-year Cardiovascular Disease (CVD) Risk

Date: _____

Patient's Name: _____

Step 1[1]

In the "points" column enter the appropriate value according to the patient's age, HDL-C, total cholesterol, systolic blood pressure, and if they smoke or have diabetes. Calculate the total points.

Risk Factor		Risk Points		Points
		Men	Women	
Age				
30-34		0	0	
35-39		2	2	
40-44		5	4	
45-49		7	5	
50-54		8	7	
55-59		10	8	
60-64		11	9	
65-69		12	10	
70-74		14	11	
75+		15	12	
HDL-C (mmol/L)				
>1.6		-2	-2	
1.3-1.6		-1	-1	
1.2-1.29		0	0	
0.9-1.19		1	1	
<0.9		2	2	
Total Cholesterol				
<4.1		0	0	
4.1-5.19		1	1	
5.2-6.19		2	3	
6.2-7.2		3	4	
>7.2		4	5	
Systolic Blood Pressure (mmHg)		Not Treated / Treated	Not Treated / Treated	
<120		-2 / 0	-3 / -1	
120-129		0 / 2	0 / 2	
130-139		1 / 3	1 / 3	
140-149		2 / 4	2 / 5	
150-159		2 / 4	4 / 6	
160+		3 / 5	5 / 7	
Smoker	Yes	4	3	
	No	0	0	
Diabetes	Yes	statin-indicated condition		
	No	0	0	
Total Points				

1 Adapted from: D'Agostino RB et al.(i). General cardiovascular risk profile for use in primary care. The Framingham Heart Study. Circ 2008;117:743-53.
2 Adapted from: Genest J et al.(i). 2009 Canadian Cardiovascular Society/Canadian guidelines for the diagnosis and treatment of dyslipidemia and prevention of cardiovascular disease in the adult. Can J Cardiol. 2009;25(10):567-579.
3 Adapted from: Anderson T et al.(i). 2012 Update of the Canadian Cardiovascular Society guidelines for the diagnosis and treatment of dyslipidemia for the prevention of cardiovascular disease in the adult. Can J Cardiol. 2013;29(2):151-167.

‡ apoB: apolipoprotein B stat, CVD: cardiovascular disease, FRS: Framingham Risk Score, HDL-C: high-density lipoprotein cholesterol, LDL-C: low-density lipoprotein cholesterol.
* Statins indicated as initial therapy
** Consider LDL-C < 1.8 mmol/L for subjects with acute coronary syndrome (ACS) within past 3 months

Provided courtesy of **Canadian Cardiovascular Society**
Leadership. Knowledge. Community.

Step 2[1]

Using the total points from Step 1, determine the 10-year CVD risk (%).*

Total Points	10-Year CVD Risk (%)*	
	Men	Women
-3 or less	<1	<1
-2	1.1	<1
-1	1.4	1.0
0	1.6	1.2
1	1.9	1.5
2	2.3	1.7
3	2.8	2.0
4	3.3	2.4
5	3.9	2.8
6	4.7	3.3
7	5.6	3.9
8	6.7	4.5
9	7.9	5.3
10	9.4	6.3
11	11.2	7.3
12	13.3	8.6
13	15.6	10.0
14	18.4	11.7
15	21.6	13.7
16	25.3	15.9
17	29.4	18.51
18	>30	21.5
19	>30	24.8
20	>30	27.5
21+	>30	>30

Step 3[1]

Using the total points from Step 1, determine heart age (in years).

Heart Age, *y*	Men	Women
<30	<0	<1
30	0	
31		1
32	1	
34	2	2
36	3	3
38	4	
39		4
40	5	
42	6	5
45	7	6
48	8	7
51	9	8
54	10	
55		9
57	11	
59		10
60	12	
64	13	11
68	14	12
72	15	
73		13
76	16	
79		14
>80	≥17	15+

* Double cardiovascular disease risk percentage for individuals between the ages of 30 and 59 without diabetes if the presence of a positive history of premature cardiovascular disease is present in a first-degree relative before 55 years of age for men and before 65 years of age for women. This is known as the modified Framingham Risk Score.[3]

Step 4[2,3]

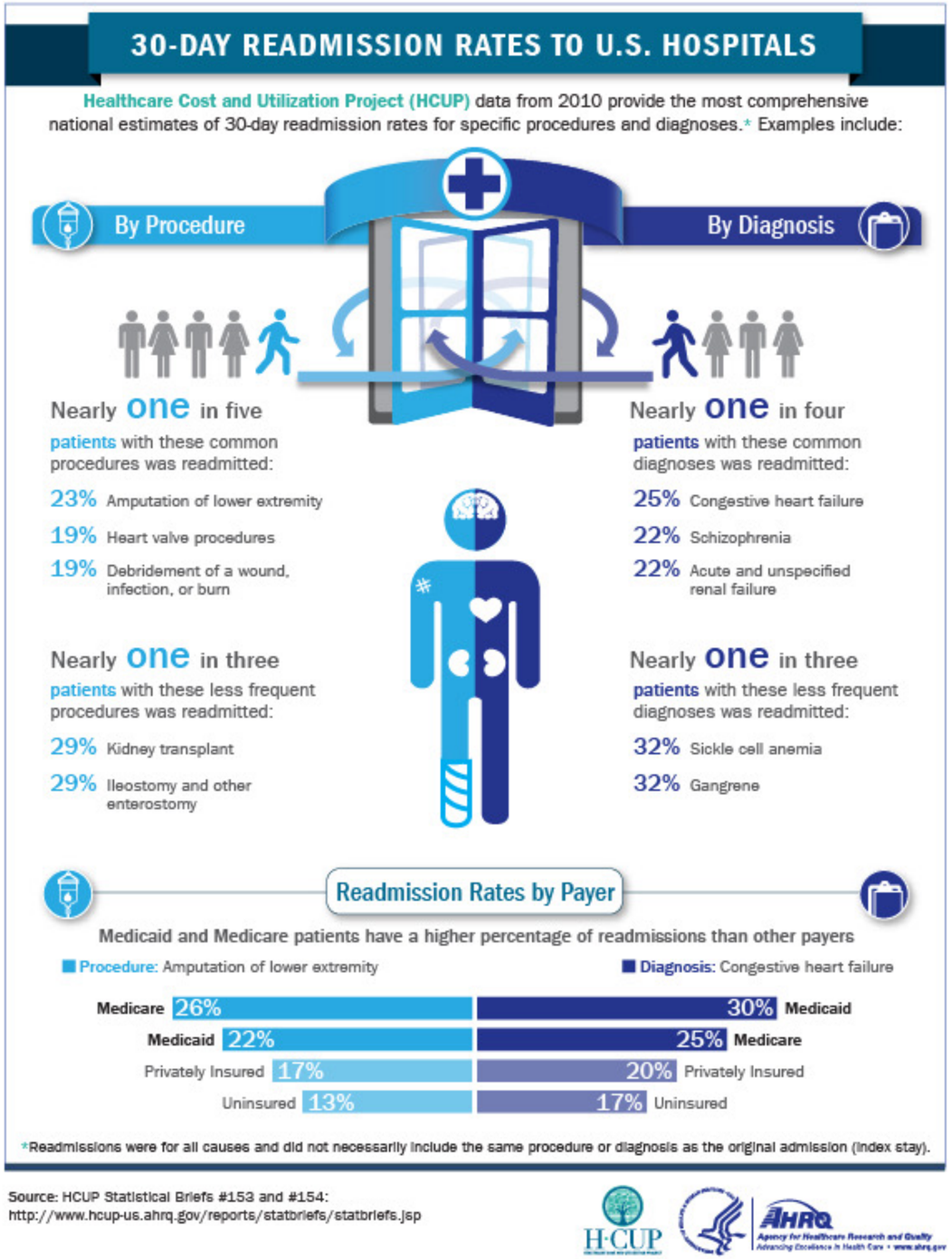
Using 10-year CVD risk from Step 2, determine if patient is Low, Moderate or High risk.† Indicate Lipid and/or Apo B targets

Risk Level†	Initiate Treatment If:	Primary Target (LDL-C)	Alternate Target
High FRS ≥20%	• Consider treatment in all *(Strong, High)*	• ≤2 mmol/L or ≥50% decrease in LDL-C *(Strong, Moderate)*	• Apo B ≤0.8 g/L or • Non-HDL-C ≤2.6 mmol/L *(Strong, High)*
Intermediate FRS 10-19%	• LDL-C ≥3.5 mmol/L *(Strong, Moderate)* • For LDL-C <3.5 mmol/L consider if: • Apo B ≥1.2 g/L • OR Non-HDL-C ≥4.3 mmol/L *(Strong, Moderate)* • *Men ≥50 and women ≥60 with 1 risk factor: low HDL-C, impaired fasting glucose, high waist circumference, smoker, hypertension*	• ≤2 mmol/L or ≥50% decrease in LDL-C *(Strong, Moderate)*	• Apo B ≤0.8 g/L or • Non-HDL-C ≤2.6 mmol/L *(Strong, Moderate)*
Low FRS <10%	• statins generally not indicated	• statins generally not indicated	• statins generally not indicated
Statin-indicated conditions**	• *Clinical atherosclerosis** • *Abdominal aortic aneurysm* • *Diabetes mellitus Age ≥ 40 years 15-Year duration for age ≥ 30 years (DM1) Microvascular disease* • *Chronic kidney disease (age ≥ 50 years) eGFR <60 mL/min/1.73 m2 or ACR > 3 mg/mmol*		

Lipid targets LDL-C: _____ or Apo B: _____

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



Likelihood of hospital readmission?

Examples of risk stratification

- Preterm infant's risk of severe morbidity?

RESEARCH ARTICLE

PREMATURE INFANTS

Integration of Early Physiological Responses Predicts Later Illness Severity in Preterm Infants

Suchi Saria,[1] Anand K. Rajani,[2] Jeffrey Gould,[2] Daphne Koller,[1]* Anna A. Penn[2]*

(Published 8 September 2010; Volume 2 Issue 48 48ra65)

Physiological data are routinely recorded in intensive care, but their use for rapid assessment of illness severity or long-term morbidity prediction has been limited. We developed a physiological assessment score for preterm newborns, akin to an electronic Apgar score, based on standard signals recorded noninvasively on admission to a neonatal intensive care unit. We were able to accurately and reliably estimate the probability of an individual preterm infant's risk of severe morbidity on the basis of noninvasive measurements. This prediction algorithm was developed with electronically captured physiological time series data from the first 3 hours of life in preterm infants (≤34 weeks gestation, birth weight ≤2000 g). Extraction and integration of the data with state-of-the-art machine learning methods produced a probability score for illness severity, the PhysiScore. PhysiScore was validated on 138 infants with the leave-one-out method to prospectively identify infants at risk of short- and long-term morbidity. PhysiScore provided higher accuracy prediction of overall morbidity (86% sensitive at 96% specificity) than other neonatal scoring systems, including the standard Apgar score. PhysiScore was particularly accurate at identifying infants with high morbidity related to specific complications (infection: 90% at 100%; cardiopulmonary: 96% at 100%). Physiological parameters, particularly short-term variability in respiratory and heart rates, contributed more to morbidity prediction than invasive laboratory studies. Our flexible methodology of individual risk prediction based on automated, rapid, noninvasive measurements can be easily applied to a range of prediction tasks to improve patient care and resource allocation.

INTRODUCTION

Early, accurate prediction of a neonate's morbidity risk is of significant clinical value because it allows for customized medical management. The standard Apgar score has been used for more than 50 years to assess neonatal well-being and the need for further medical management. We aimed to develop a modern tool akin to an "electronic" Apgar assessment that reflects a newborn's physiological status and is predictive of future illness severity. Such an improvement in neonatal risk stratification may better inform decisions regarding aggressive use of intensive care, need for transport to tertiary centers, and resource allocation, thus potentially reducing the estimated $26 billion per year in U.S. health care costs resulting from preterm birth (*1*). Gestational age and birth weight are highly predictive of death or disability (*2*) but do not estimate individual illness severity or morbidity risk (*3*). These perinatal risk factors, in addition to laboratory measurements, have been incorporated into currently used algorithms for mortality risk assessment of preterm infants (*4*–*6*). These algorithms, however, predict mortality rather than morbidity (*3*). They also rely on invasive testing and require extraction of data from multiple sources to make a risk assessment.

Although it has been recognized that changes in heart rate characteristics (*7*) or variability (*8*) can suggest impending illness and death in a range of clinical scenarios, from sepsis (*9*) in intensive care patients to fetal intolerance of labor (*10*), the predictive accuracy of a single parameter is limited. Intensive care providers observe multiple physiological signals in real time to assess health, but certain informative patterns may be subtle. To achieve improved accuracy and speed of individual morbidity prediction for preterm neonates, we developed a new probability score (PhysiScore) based on physiological data obtained noninvasively after birth along with gestational age and birth weight. Two recent advances enabled the use of multiple complex physiological signals for this purpose: the digitization of medical records, which allows linking of real-time physiological signals with later outcomes, and the increasing sophistication of machine learning and pattern recognition algorithms, which allows optimization of PhysiScore in an automated, unbiased manner. We evaluated PhysiScore's use for predicting overall morbidity and mortality, specific risk for infants with infection or cardiovascular and pulmonary complications, and a combination of complications associated with poor long-term neurodevelopment and compared its performance to standard scoring systems in a preterm neonatal cohort.

RESULTS

PhysiScore development based on patient characteristics and morbidities

To develop our prediction tool, we studied a total of 138 preterm neonates that were 34 weeks gestational age or less and <2000 g in weight without major congenital malformations and with baseline characteristics and morbidities as shown in Table 1. Mean birth weight was 1367 g at an estimated mean gestational age of 29.8 weeks, placing these infants at significant risk of both short- and long-term complications.

Patients were then classified as high morbidity (HM) or low morbidity (LM) on the basis of their illnesses. The HM group was defined as any patient with major complications associated with short- or long-term morbidity. Short-term morbidity complications included culture-positive sepsis, pulmonary hemorrhage, pulmonary hypertension, and

[1]Department of Computer Science, Stanford University, Stanford, CA 94305, USA. [2]Department of Pediatrics, Division of Neonatal and Developmental Medicine, Stanford University School of Medicine, Palo Alto, CA 94305, USA.
*To whom correspondence should be addressed. E-mail: apenn@stanford.edu (A.A.P.); koller@cs.stanford.edu (D.K.)

- Will this COVID19 patient need to be admitted to the ICU?

ORIGINAL CLINICAL REPORT

OPEN

Stratifying Deterioration Risk by Acuity at Admission Offers Triage Insights for Coronavirus Disease 2019 Patients

Joseph Beals IV, PhD[1]
Jaime J. Barnes, DO[2]
Daniel J. Durand, MD[3]
Joan M. Rimar, DNSc[4]
Thomas J. Donohue, MD[4]
S. Mahfuz Hoq, MD[5]
Kathy W. Belk, BA[1]
Alpesh N. Amin, MD[6]
Michael J. Rothman, PhD[3]

OBJECTIVES: Triaging patients at admission to determine subsequent deterioration risk can be difficult. This is especially true of coronavirus disease 2019 patients, some of whom experience significant physiologic deterioration due to dysregulated immune response following admission. A well-established acuity measure, the Rothman Index, is evaluated for stratification of patients at admission into high or low risk of subsequent deterioration.

DESIGN: Multicenter retrospective study.

SETTING: One academic medical center in Connecticut, and three community hospitals in Connecticut and Maryland.

PATIENTS: Three thousand four hundred ninety-nine coronavirus disease 2019 and 14,658 noncoronavirus disease 2019 adult patients admitted to a medical service between January 1, 2020, and September 15, 2020.

INTERVENTIONS: None.

MEASUREMENTS AND MAIN RESULTS: Performance of the Rothman Index at admission to predict in-hospital mortality or ICU utilization for both general medical and coronavirus disease 2019 populations was evaluated using the area under the curve. Precision and recall for mortality prediction were calculated, high- and low-risk thresholds were determined, and patients meeting threshold criteria were characterized. The Rothman Index at admission has good to excellent discriminatory performance for in-hospital mortality in the coronavirus disease 2019 (area under the curve, 0.81–0.84) and noncoronavirus disease 2019 (area under the curve, 0.90–0.92) populations. We show that for a given admission acuity, the risk of deterioration for coronavirus disease 2019 patients is significantly higher than for noncoronavirus disease 2019 patients. At admission, Rothman Index–based thresholds segregate the majority of patients into either high- or low-risk groups; high-risk groups have mortality rates of 34–45% (coronavirus disease 2019) and 17–25% (noncoronavirus disease 2019), whereas low-risk groups have mortality rates of 2–5% (coronavirus disease 2019) and 0.2–0.4% (noncoronavirus disease 2019). Similarly large differences in ICU utilization are also found.

CONCLUSIONS: Acuity level at admission may support rapid and effective risk triage. Notably, in-hospital mortality risk associated with a given acuity at admission is significantly higher for coronavirus disease 2019 patients than for noncoronavirus disease 2019 patients. This insight may help physicians more effectively triage coronavirus disease 2019 patients, guiding level of care decisions and resource allocation.

DOI: 10.1097/CCE.0000000000000400

Outline

- Evaluation of diagnostic tests:
  - Accuracy, sensitivity, specificity
  - Positive and negative predictive values
  - Relative risk
  - Odds ratio
- Risk scores and stratification
- **Case study: Early detection of Type 2 diabetes**
  - **Framing as supervised learning problem**
  - **Evaluating risk stratification algorithms**

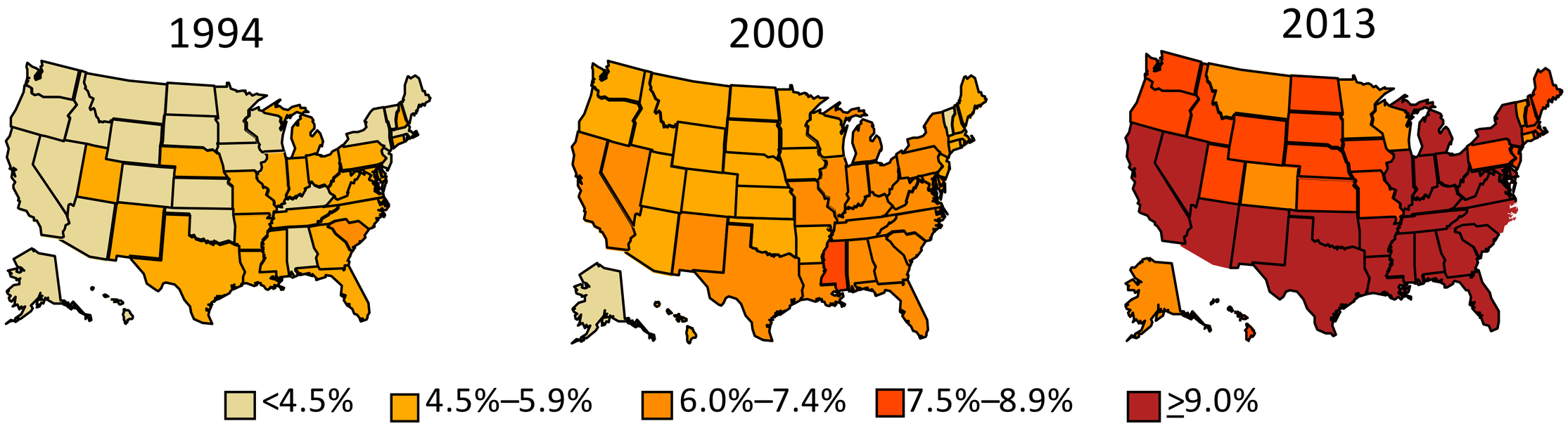
Big Data
Volume 3 Number 4, 2015
Mary Ann Liebert, Inc.
DOI: 10.1089/big.2015.0020

ORIGINAL ARTICLE

Population-Level Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors

Narges Razavian,[1] Saul Blecker,[2] Ann Marie Schmidt,[3] Aaron Smith-McLallen,[4] Somesh Nigam,[4] and David Sontag[1,*]

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

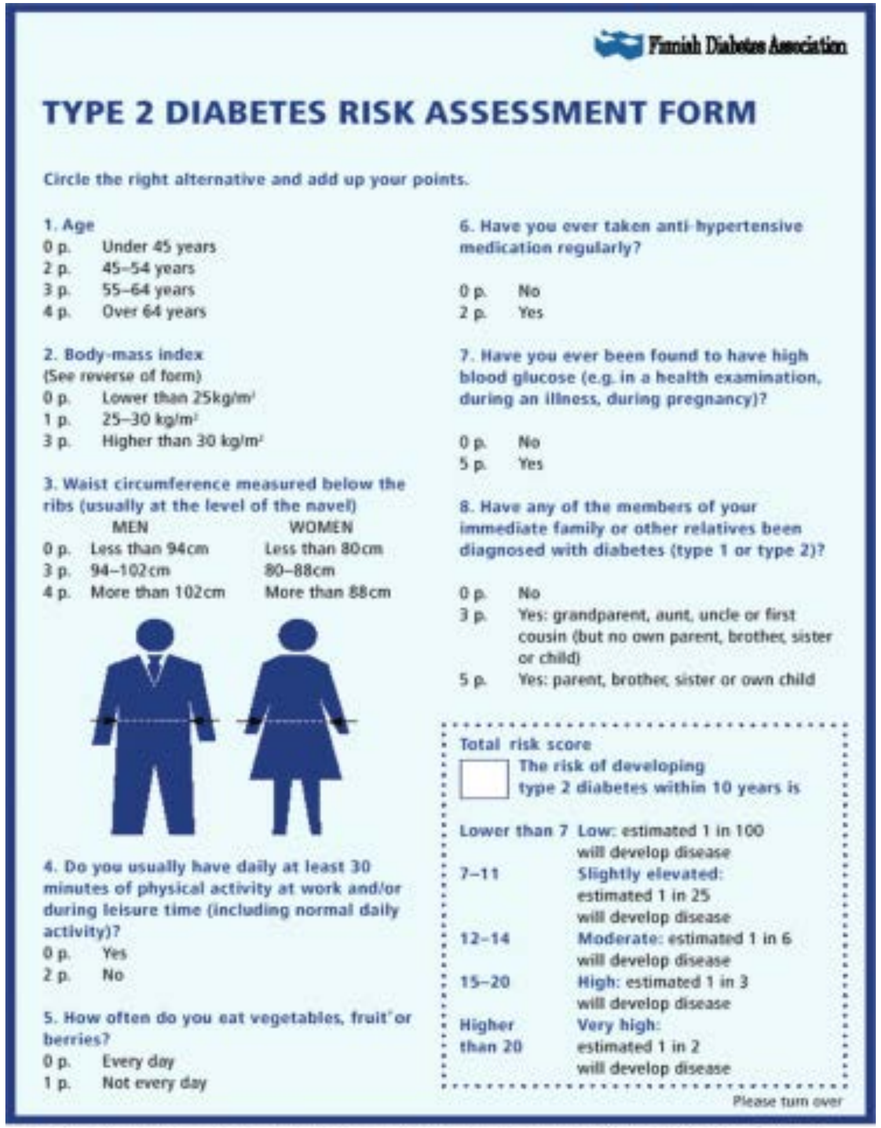
- Requirement for successful large scale prevention programme

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.

Individual 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

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

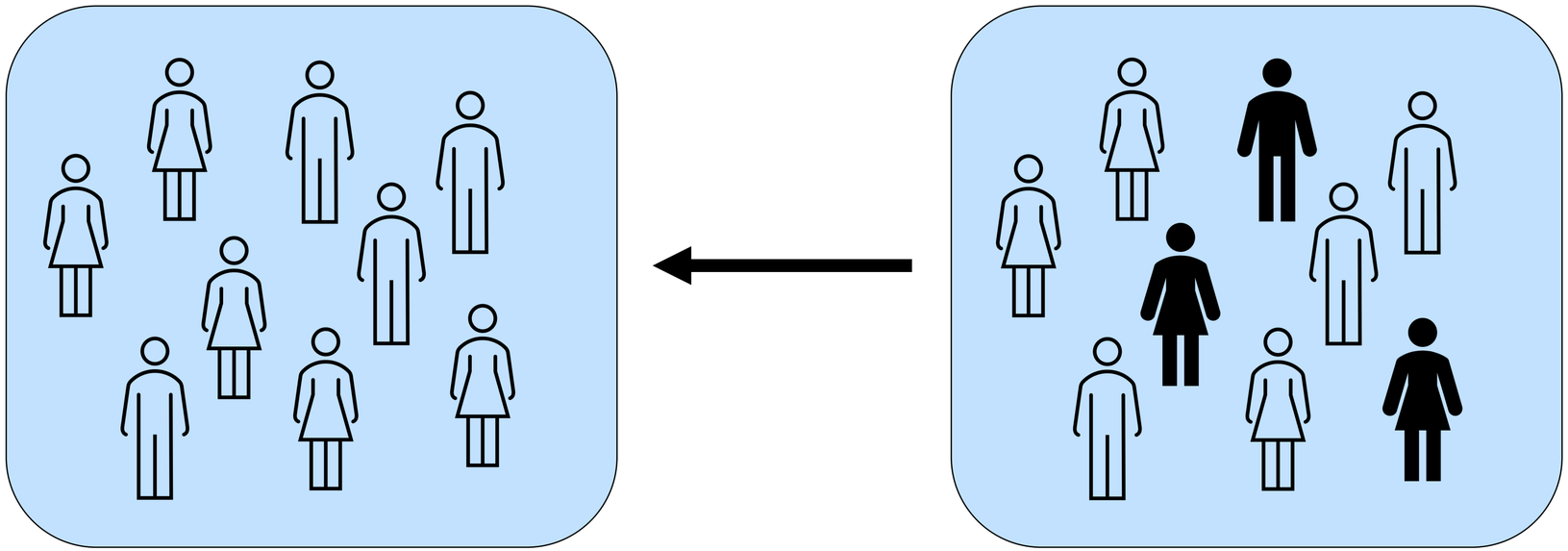
Population-Level Risk Stratification

Key ideas:

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

Using longitudinal data

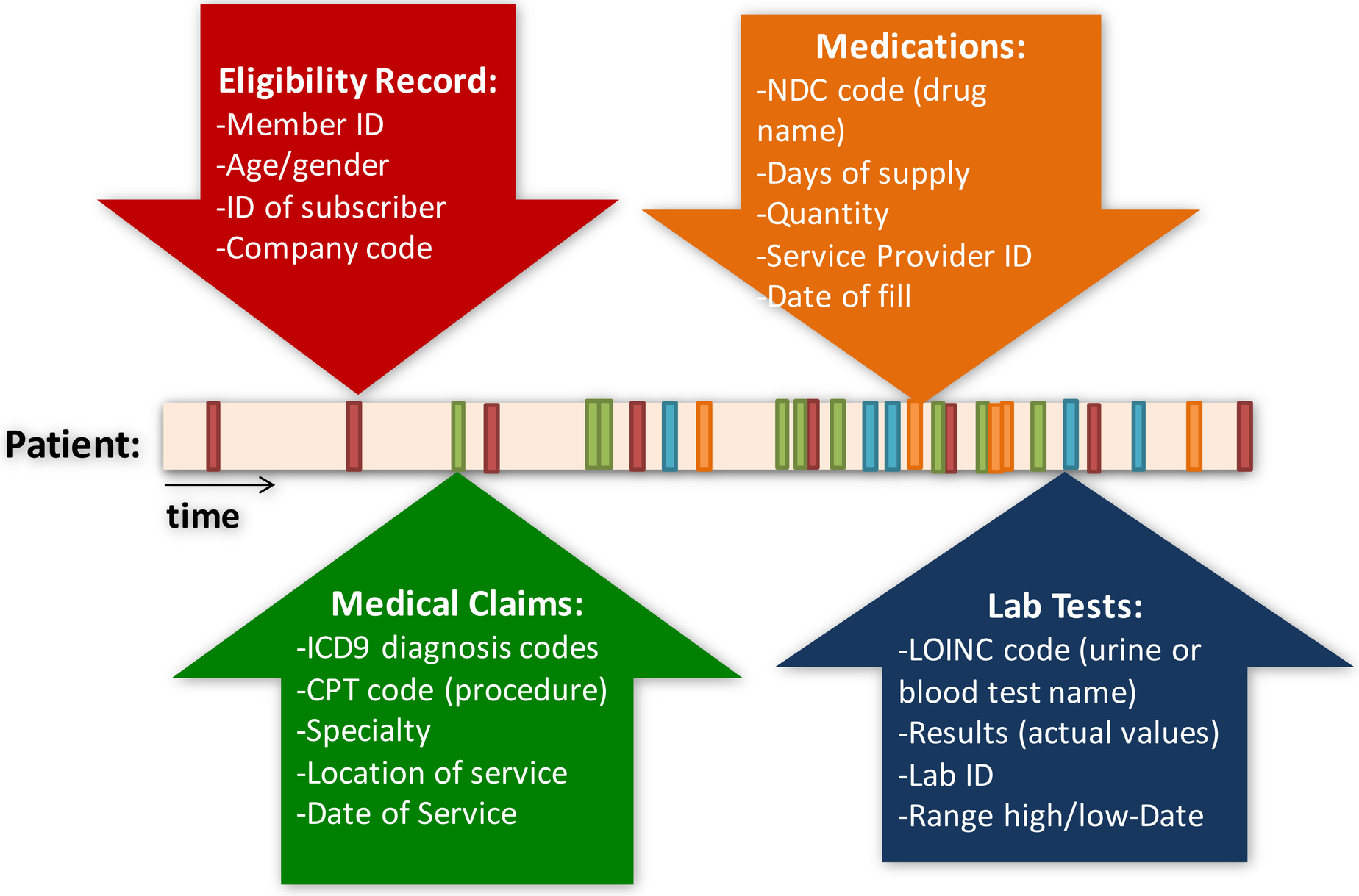
- Looking at individuals who got diabetes today (compared to those who did not)
  - Can we infer which variables in their record could have predicted their health outcome?



years ago

today

Using administrative & clinical data



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

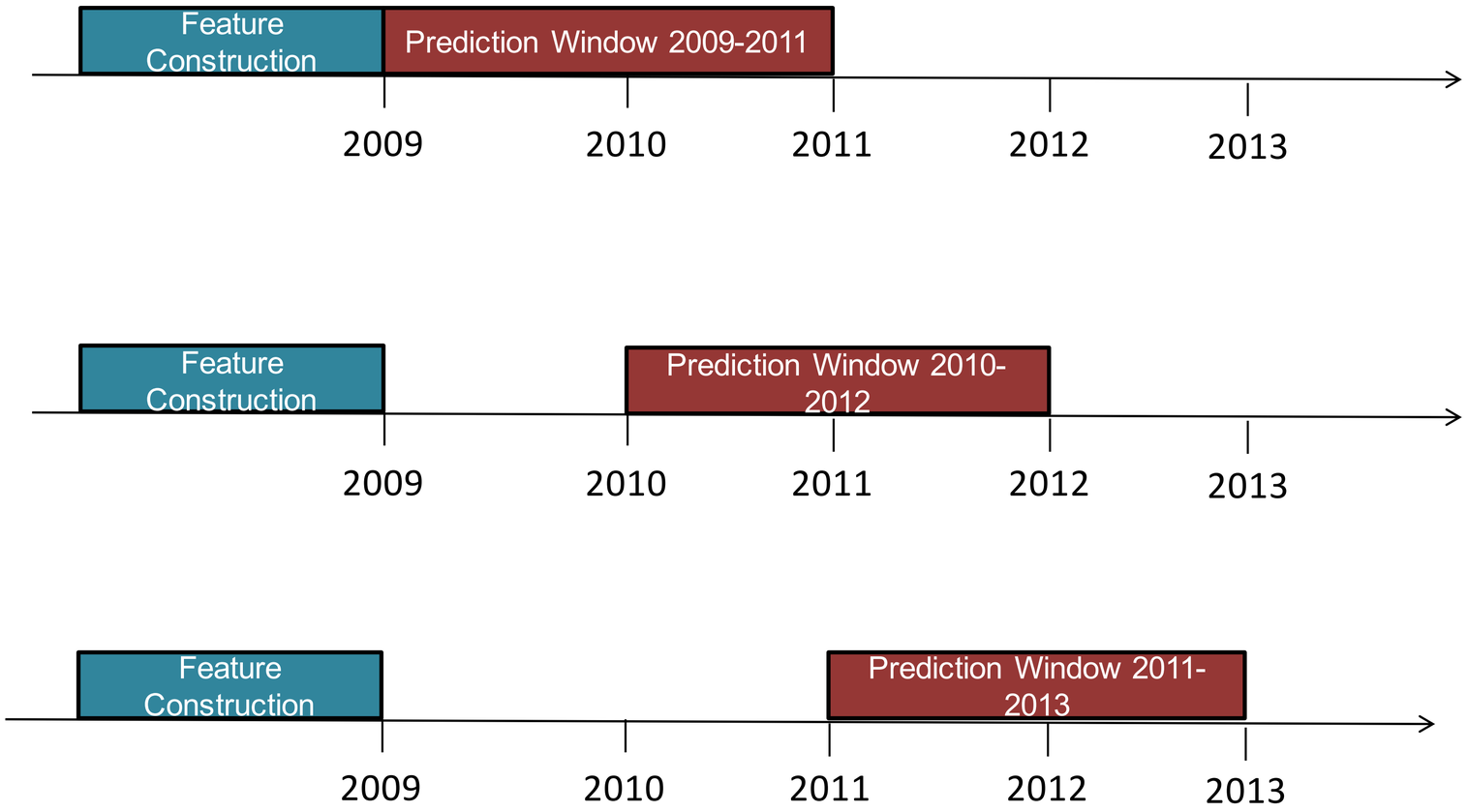
Lab test	
2085-9 Cholesterol.in HDL	1155666
718-7 Hemoglobin	1152726
4544-3 Hematocrit	1147893
9830-1 Cholesterol.total/Cholesterol.in HDL	1037730
33914-3 Glomerular filtration rate/1.73 sq M.predicted	561309
785-6 Erythrocyte mean corpuscular hemoglobin	1070832
6690-2 Leukocytes	1062980
789-8 Erythrocytes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

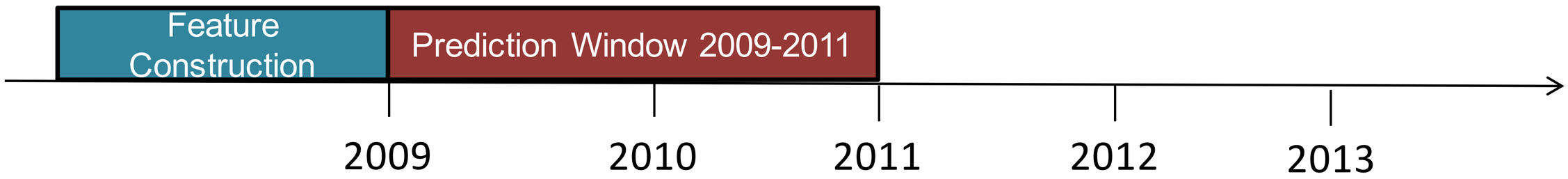
Framing as a supervised ML problem

- Binary classification:
  - Given that information about the past at a given time point
  - Predict a future outcome in a specified time window, e.g. developing diabetes



Prevent label leakage

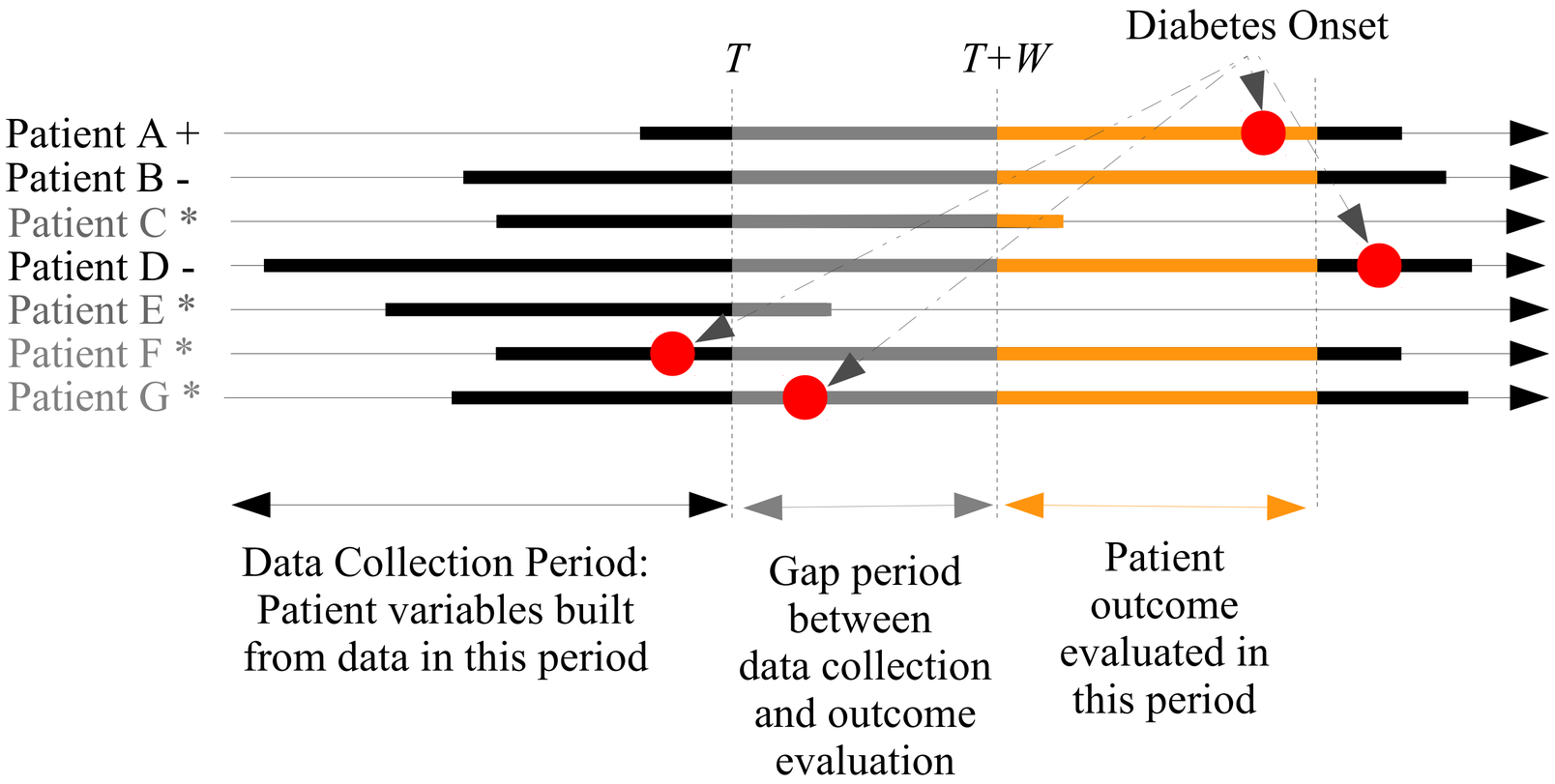
Framing as a supervised ML problem



- **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

Framing as a supervised ML problem

- Exclude patients that are left- and right-censored.



This is an example of alignment by ***absolute time***

Alternative alignment strategies

- Align by ***relative time***, e.g.
  - 2 hours into patient stay in ER
  - Every time patient sees Primary Care Physican (PCP)
  - When individual turns 40 years old

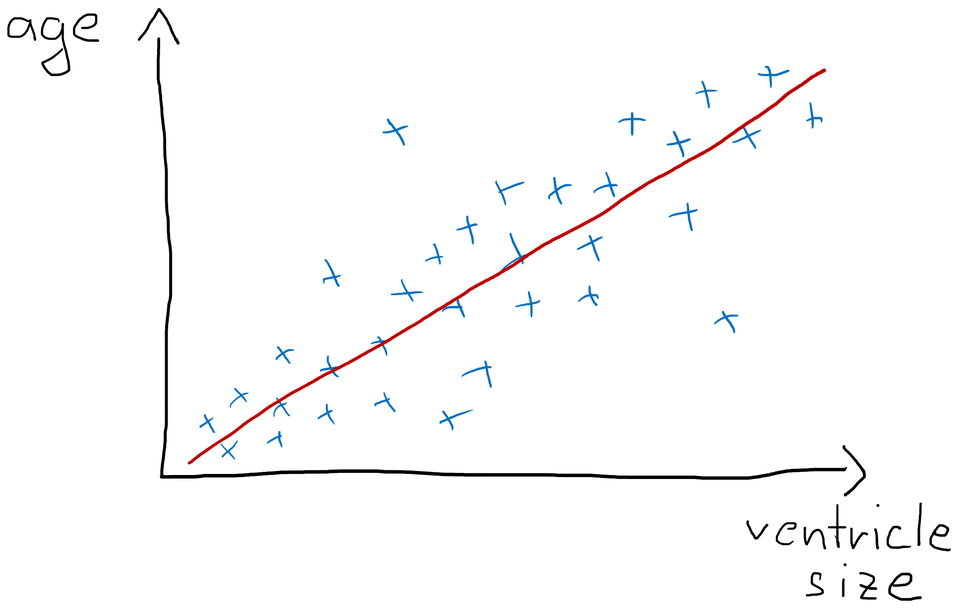
- Align by ***data availability***

- **NOTE:** If multiple data points exist per patient, make sure each patient is in *only* train, validate, *or* test set

Methods

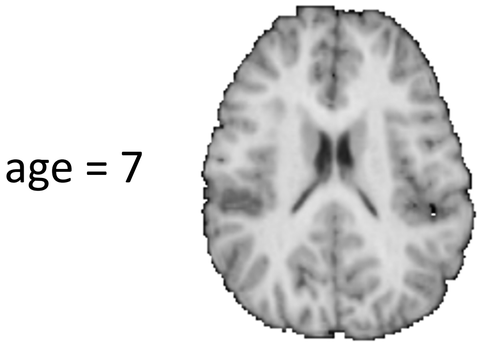
- L1 regularized logistic regression
  - Simultaneously optimises predictive performance *and*
  - Performs feature selection, choosing the subset of the features that are most predictive
- This prevents overfitting to the training data

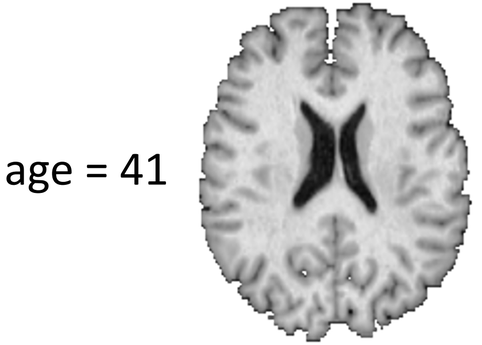
Recap: Linear Regression

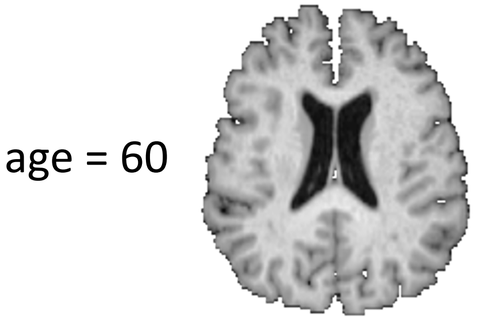


$$h_\Theta(x) = \Theta_0 + \Theta_1 x$$

bias

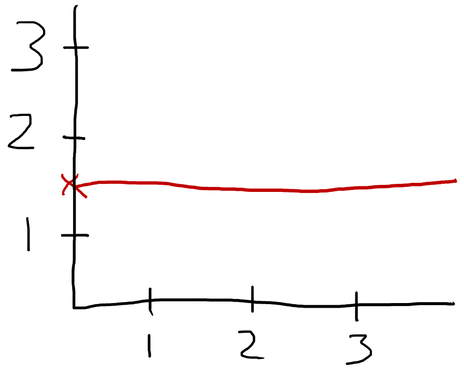






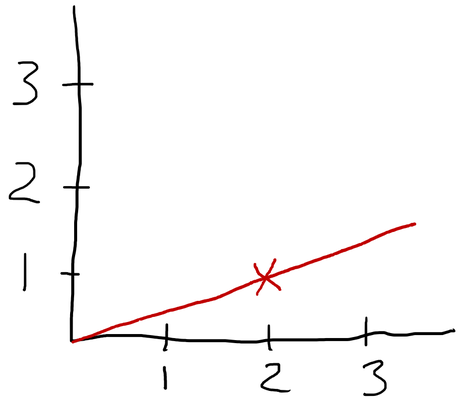
Recap: Linear Regression

$$h_\Theta(x) = \Theta_0 + \Theta_1 x$$



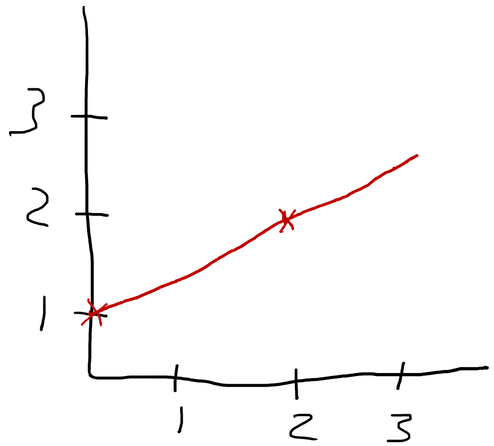
$\Theta_0 = 1.5$

$\Theta_1 = 0$



$\Theta_0 = 0$

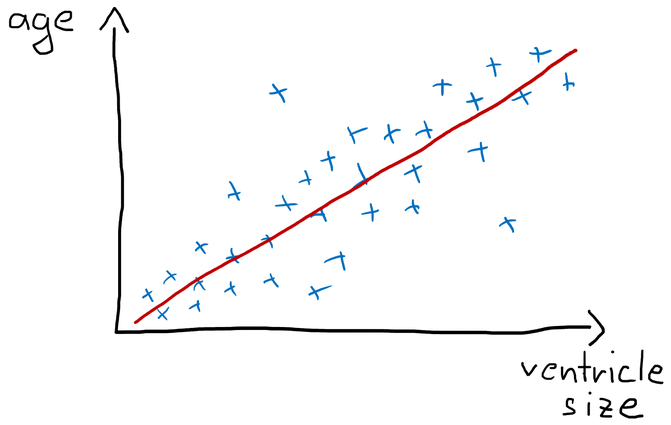
$\Theta_1 = 0.5$



$\Theta_0 = 1$

$\Theta_1 = 0.5$

Recap: Linear Regression



$$h_\Theta(x) = \Theta_0 + \Theta_1 x$$

$$\min_\Theta \frac{1}{m} \sum_{i=1}^{m} \left(h_\Theta\left(x^{(i)}\right) - y^{(i)} \right)^2$$

cost function $J(\theta)$

Recap: Linear Regression for Multiple Variables

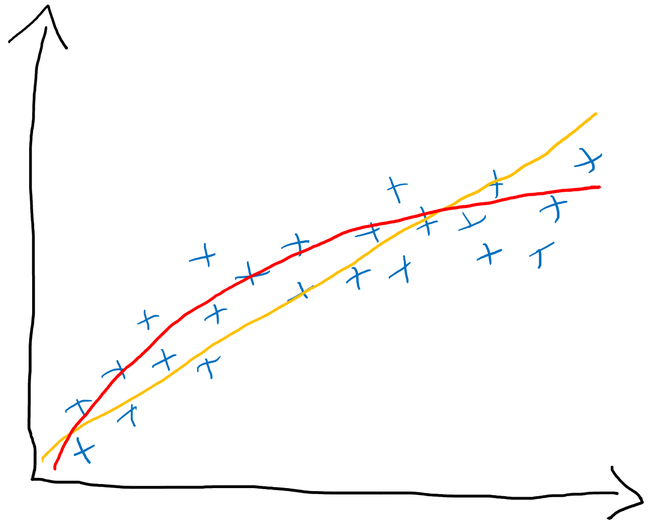
Given input vector $\mathbf{x}$ and output y

$$h_\Theta(\mathbf{x}) = \Theta^\top \mathbf{x}$$
$$= \Theta_0 x_0 + \Theta_1 x_1 + \Theta_2 x_2 + \cdots + \Theta_n x_n$$

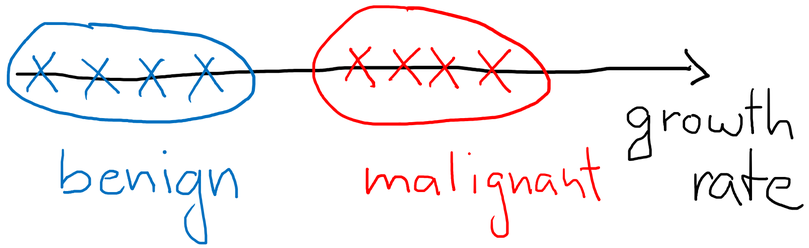
$x_0 = 1$

Polynomial Regression

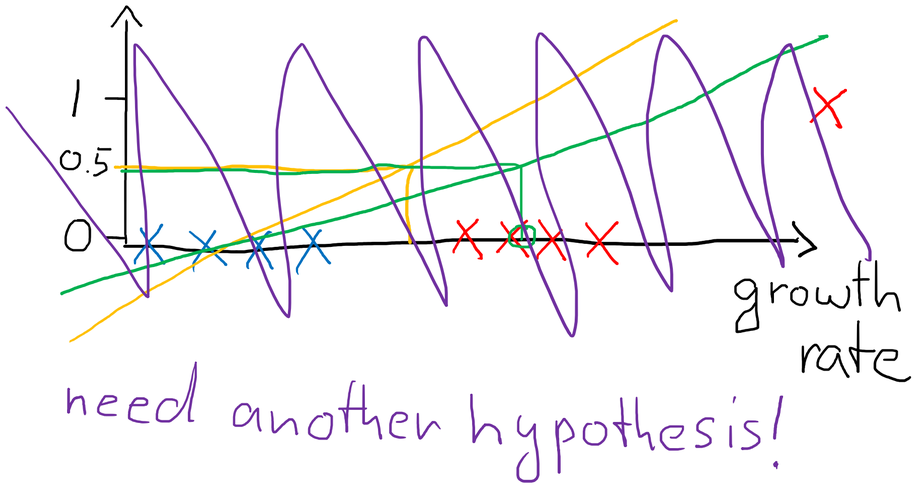
$$h_\Theta(x) = \Theta_0 + \Theta_1 x + \Theta_2 x^2$$



Recap: Logistic Regression



$$h_\Theta(\mathbf{x}) = \Theta^T\mathbf{x}$$



$$h_\Theta(\mathbf{x}) = g(\Theta^T\mathbf{x})$$

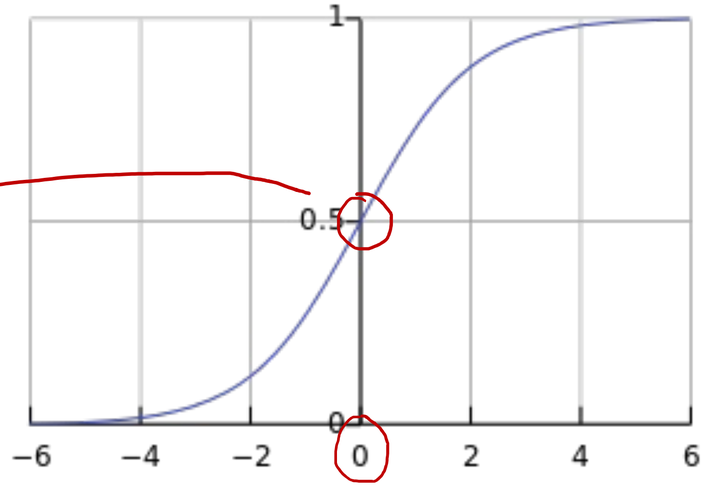
$$g(z) = \frac{1}{1+e^{-z}}$$

sigmoid function
logistic function

Recap: Logistic Regression

Hypothesis

$$h_\Theta(\mathbf{x}) = \frac{1}{1 + e^{-\Theta^T \mathbf{x}}} \in [0,1]$$



Prediction

$$h_\Theta(\mathbf{x}) \geq 0.5 \rightarrow y = 1$$

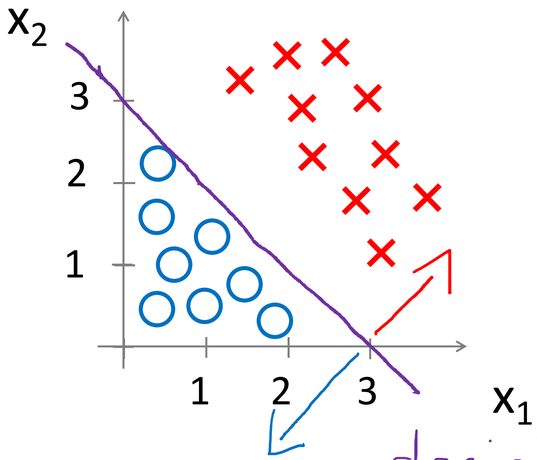
$$h_\Theta(\mathbf{x}) < 0.5 \rightarrow y = 0$$

if $\theta^T x \geq 0$

$$g(z) = \frac{1}{1 + e^{-z}}$$

$g(z) \geq 0.5$ if $z \geq 0$

Recap: Decision Boundary

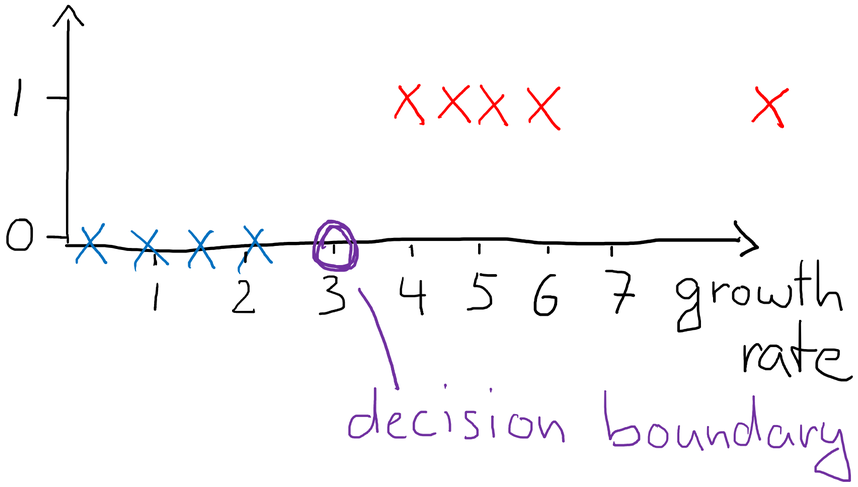


$$h_\Theta(\mathbf{x}) = g(\Theta_0 + \Theta_1 x_1 + \Theta_2 x_2)$$

$\Theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix} \rightarrow -3 + x_1 + x_2 \geq 0$

$x_1 + x_2 \geq 3$

decision boundary $x_1 + x_2 = 3$



$$h_\Theta(\mathbf{x}) = g(\Theta_0 + \Theta_1 x_1)$$

$\Theta = \begin{bmatrix} -3 \\ 1 \end{bmatrix} \rightarrow -3 + x_1 \geq 0$

$x_1 \geq 3$

$x_1 = 3$

L1 regularization

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

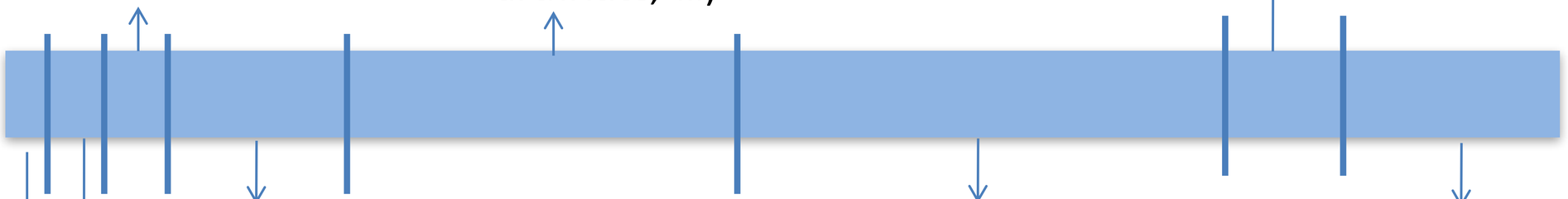
$$\min_w \sum_i \ell(x_i, y_i; w) + \quad ||w||_1 \qquad\qquad ||\vec{w}||_1 = \sum_d |w_d|$$

instead of

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

- Why?

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)

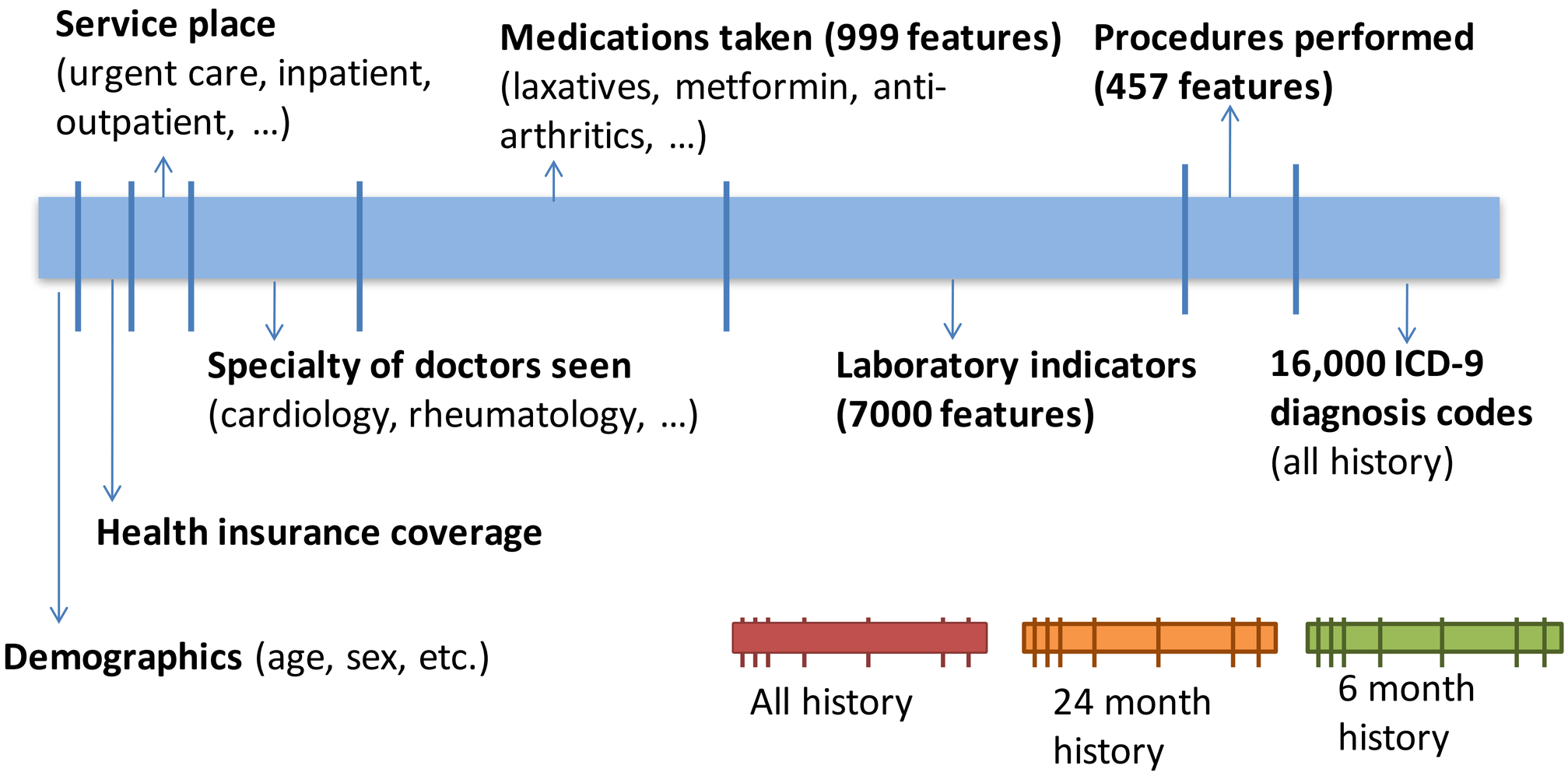
Health insurance coverage

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



Total features per patient: 42,000

What are the Discovered Risk Factors?

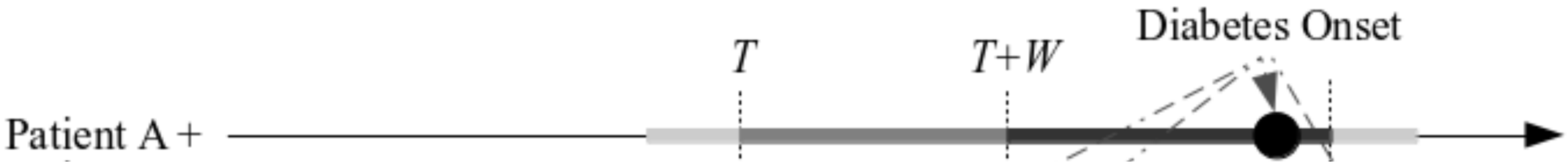
- 769 variables have non-zero weight

Recall: Odds ratio quantifies the strength of association between two events

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

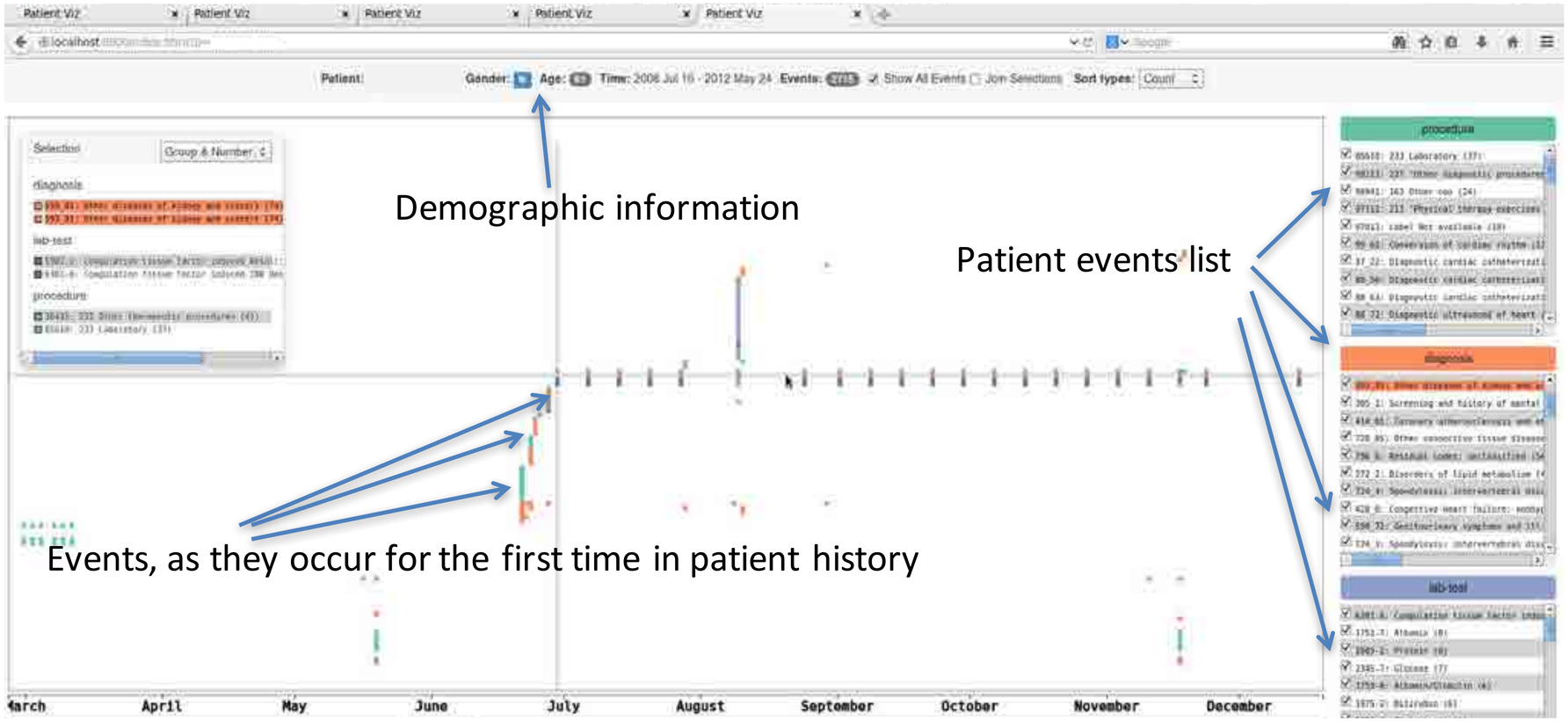
Where do the labels come from?



Typical pipeline:

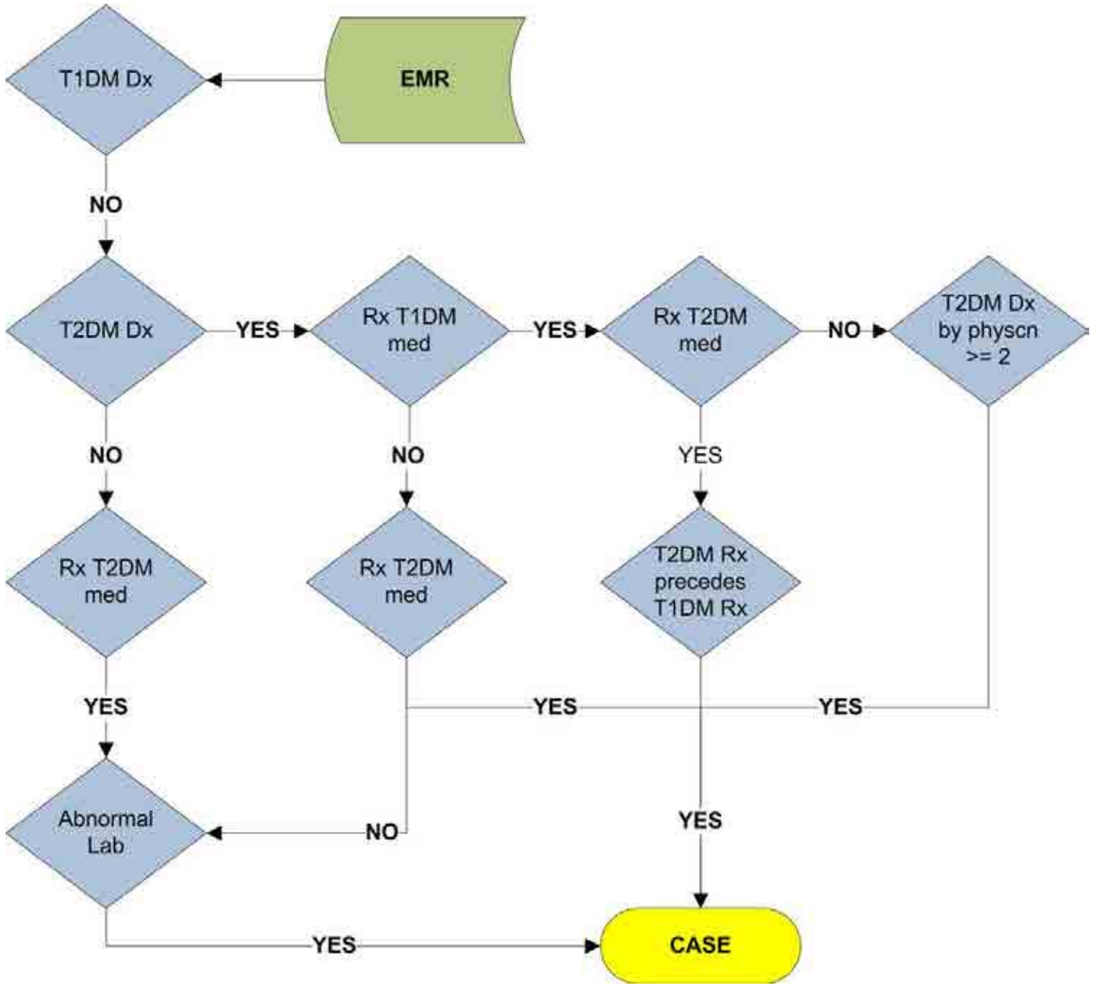
- Step 1:
  - Manually label several patients' data by "chart review"
- Step 2:
  - Come up with a simple rule to automatically derive label for all patients, **or**
  - Use **machine learning** to get the labels themselves

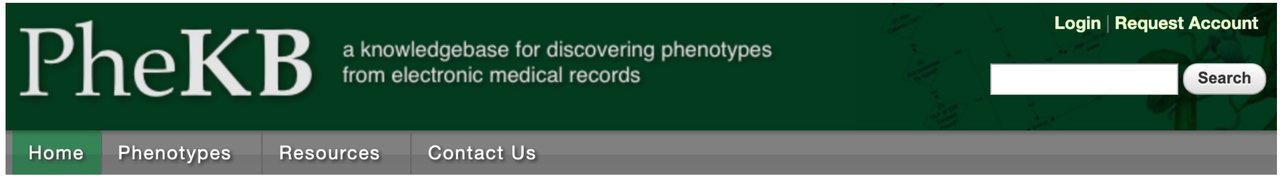
Where do the labels come from?
Step 1



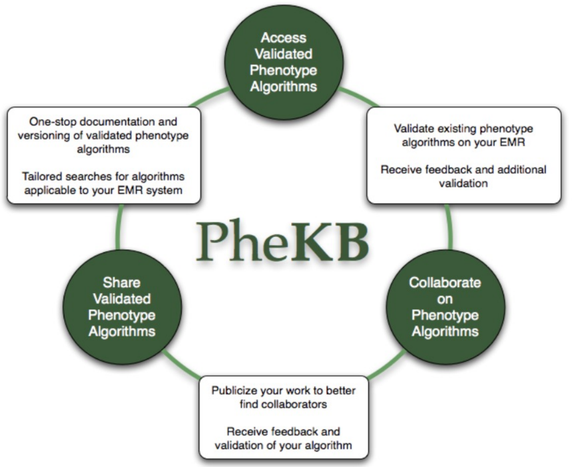
https://github.com/nyuvis/patient-viz

Where do the labels come from?
Step 2





What is the Phenotype KnowledgeBase?



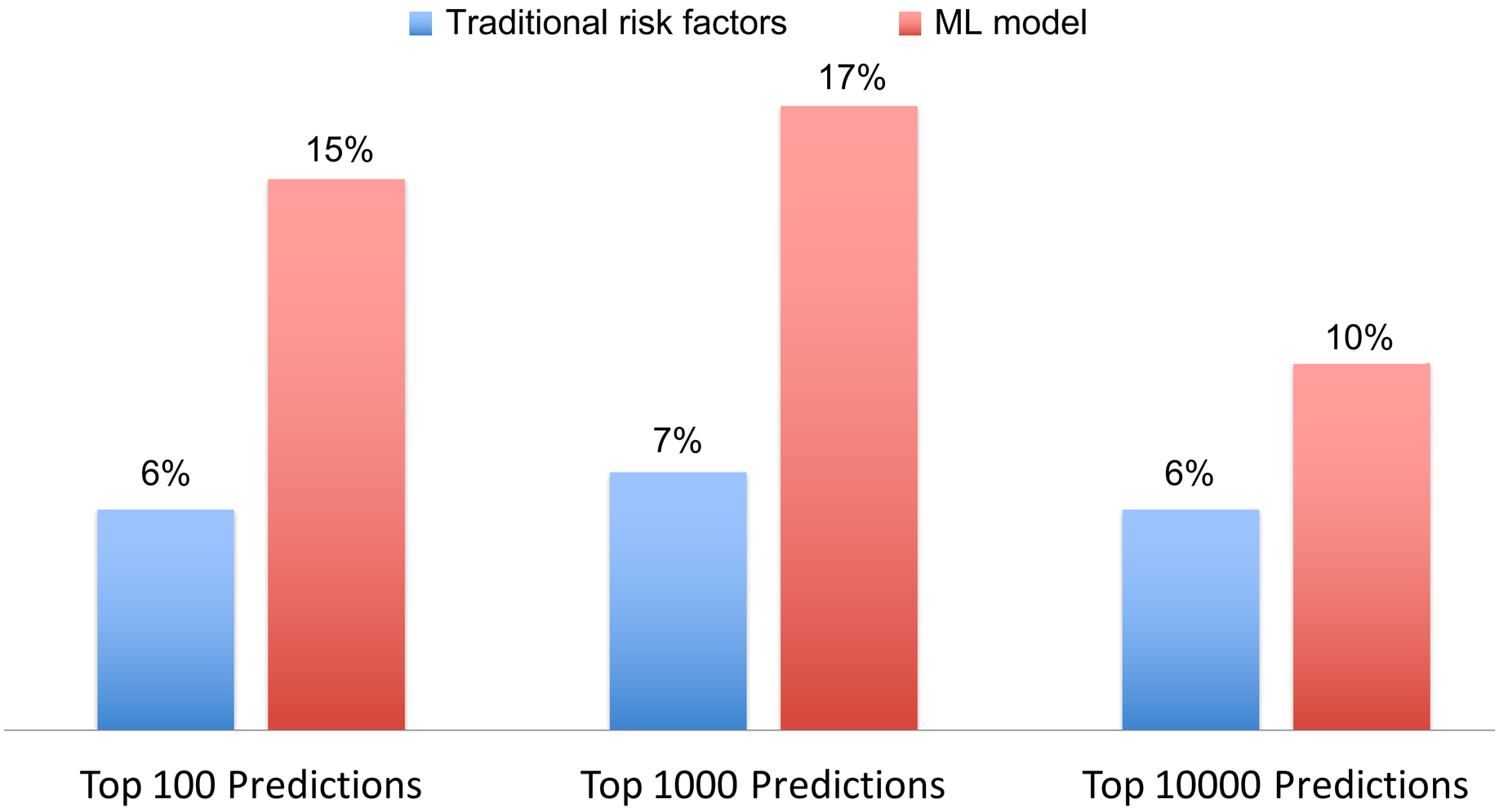
Health Data is becoming an increasing important source for clinical and genomic research. Researchers create and iteratively refine algorithms using structured and unstructured data to better identify cohorts of subjects within the health data.

The Phenotype Knowledgebase website, PheKB, is a collaborative environment to building and validating electronic algorithms to identify characteristics of patients within health data. PheKB was functionally designed to enable

Most Recent Phenotypes

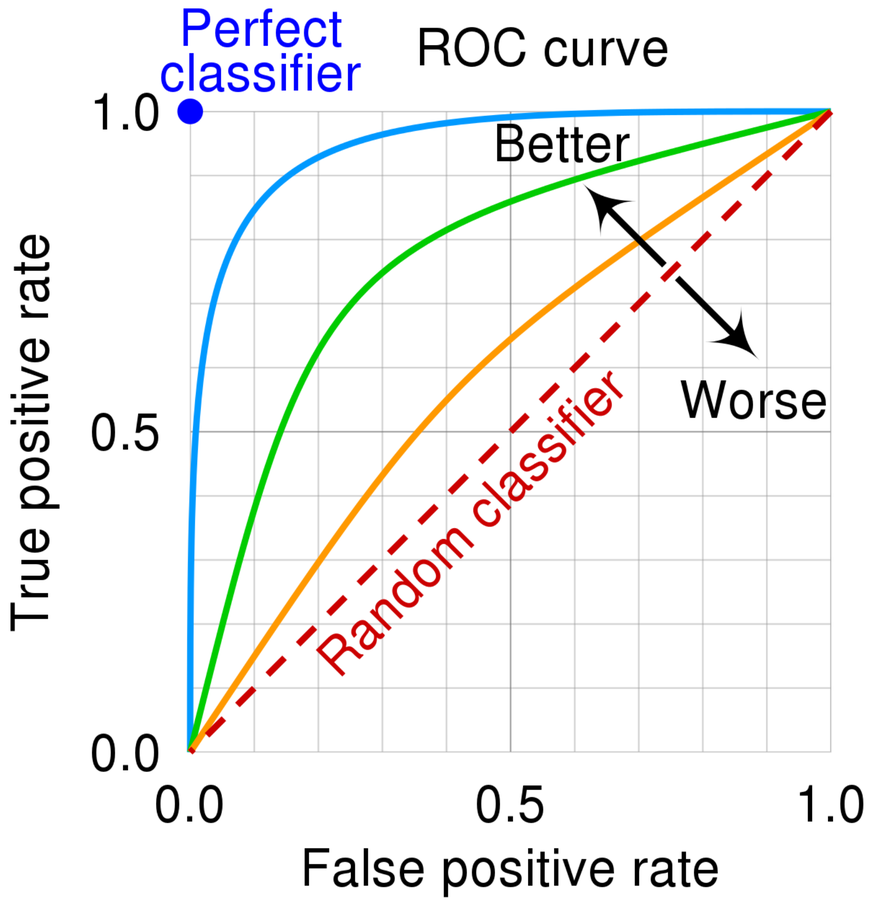
- HIV
- Functional seizures
- RxNorm RxCUI codes for Cancer Therapies
- Type 1 Diabetes
- Body Mass Index (BMI)

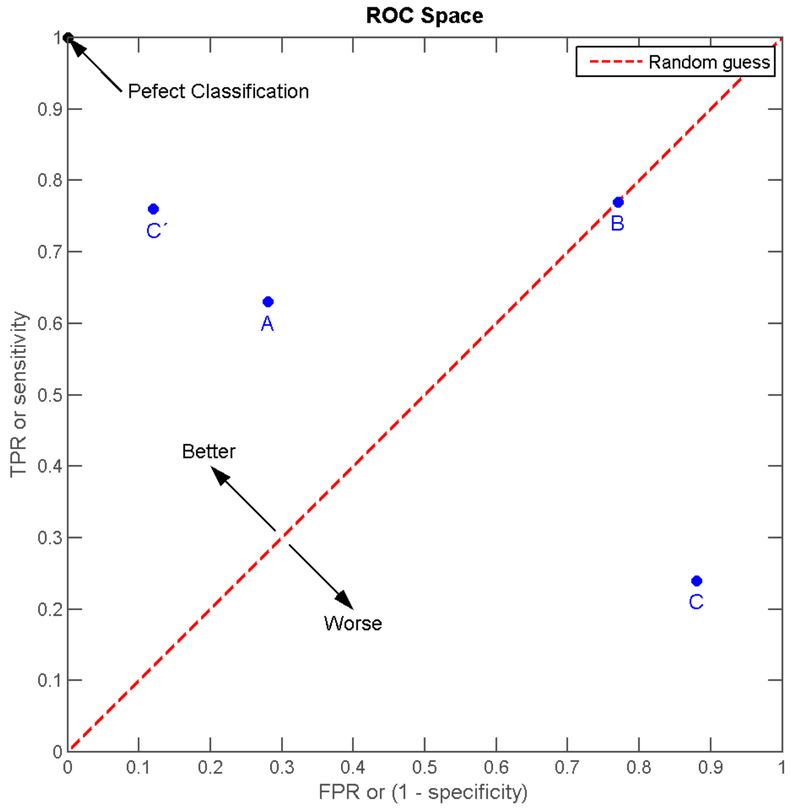
How to evaluate risk stratification models: Positive predictive value (PPV)



Diabetes 1-year gap

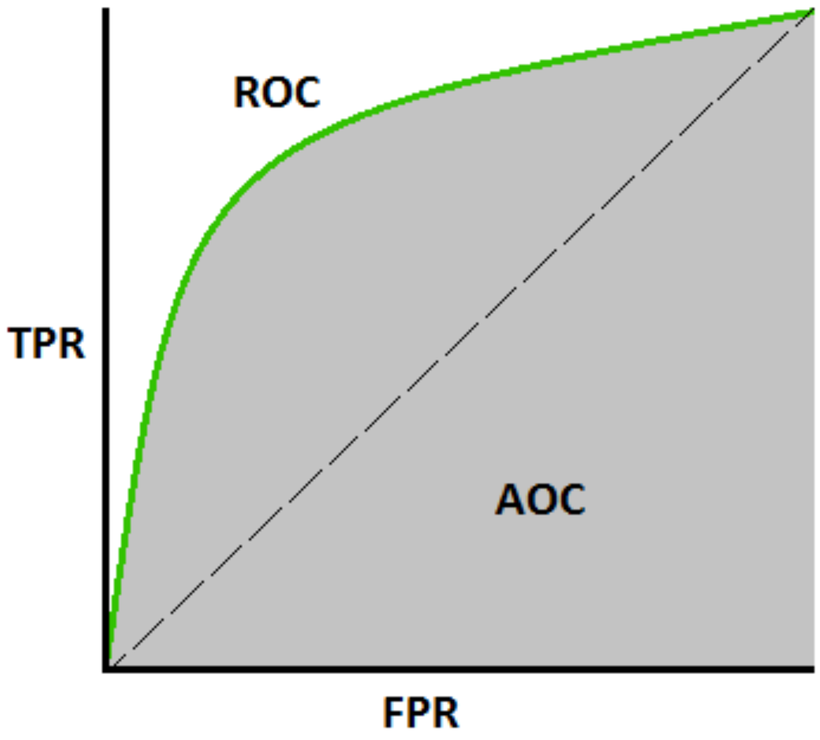
Recap: Receiver-operator characteristic curve





Recap: Receiver-operator characteristic curve

- Area under the curve

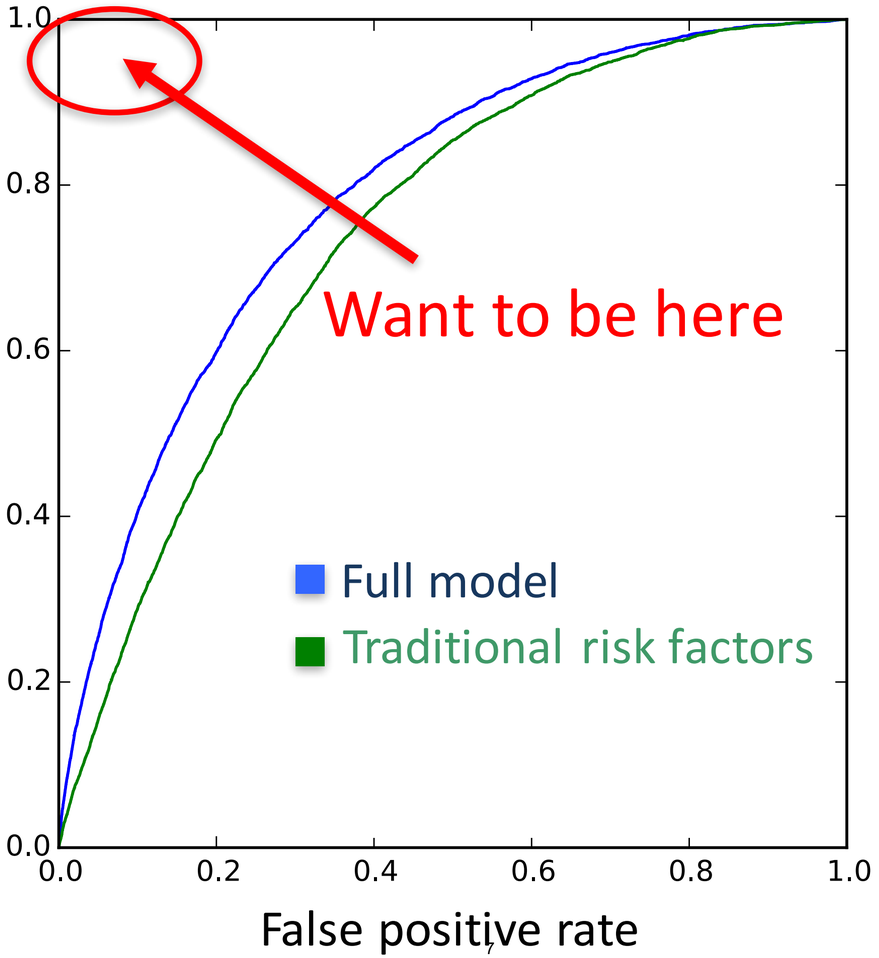


AUC =
Probability that classifier ranks a positive patient over a negative patient

Invariant to amount of class imbalance

$$AUC(f) = \frac{\sum_{t_0 \in \mathcal{D}^0} \sum_{t_1 \in \mathcal{D}^1} \mathbf{1}[f(t_0) < f(t_1)]}{|\mathcal{D}^0| \cdot |\mathcal{D}^1|},$$

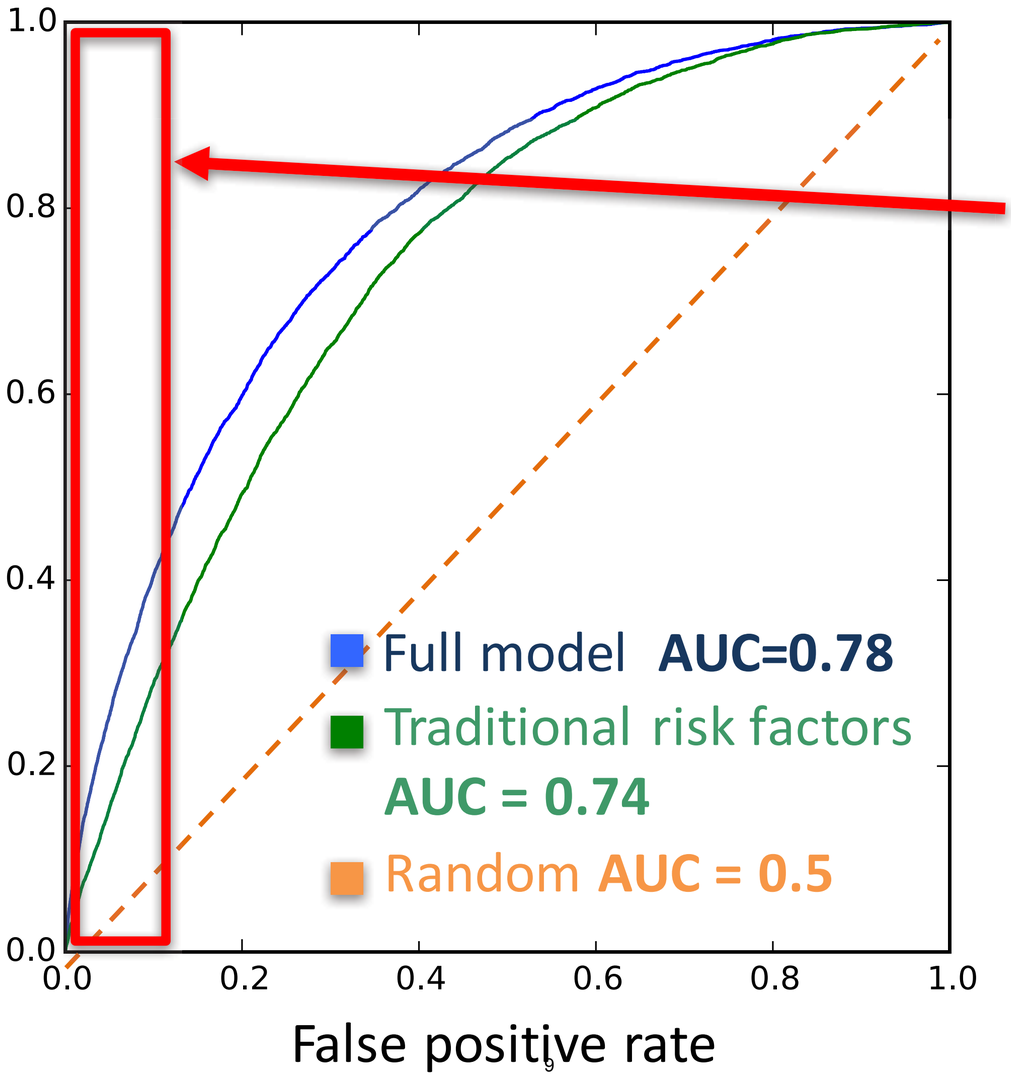
How to evaluate risk stratification models: Receiver-operator characteristic curve



Obtained by varying prediction threshold

Diabetes 1-year gap

How to evaluate risk stratification models: Receiver-operator characteristic curve

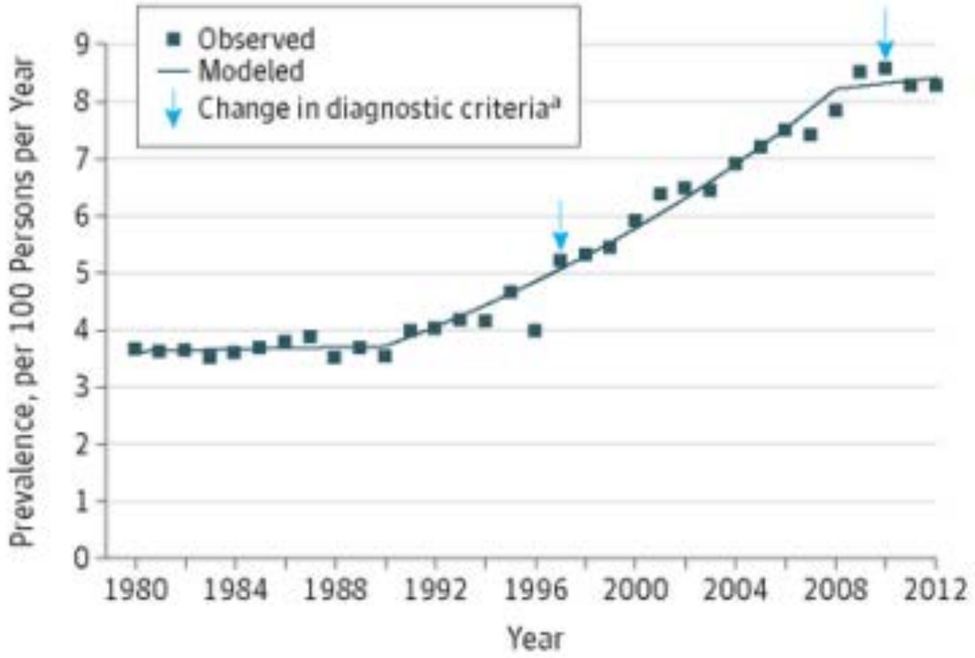


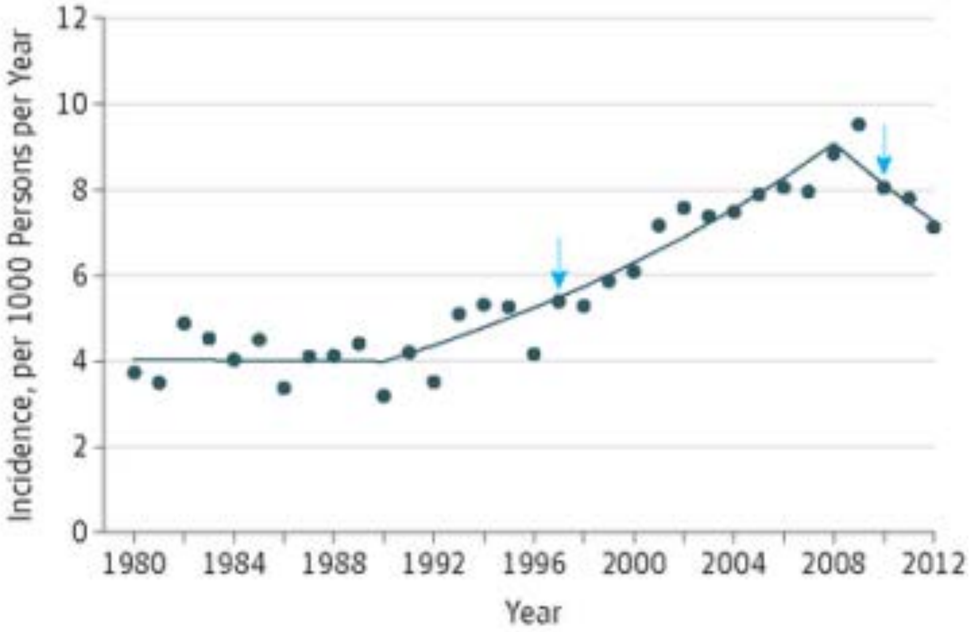
Risk stratification usually focuses on just this region

(because of the cost of interventions)

Diabetes
1-year gap

Pitfalls: Non-stationarity



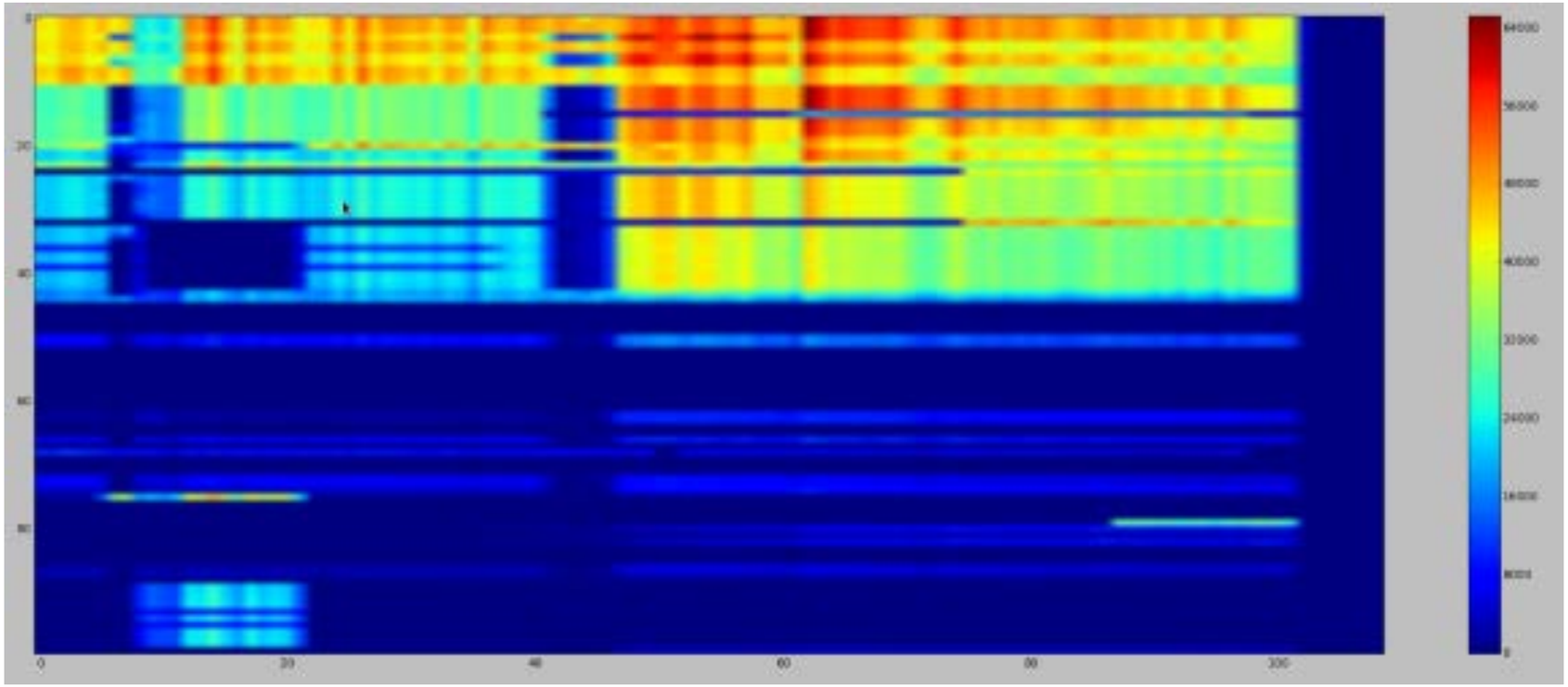


→ Automatically derived labels may change meaning

Pitfalls: Non-stationarity

Top 100 lab measurements over time

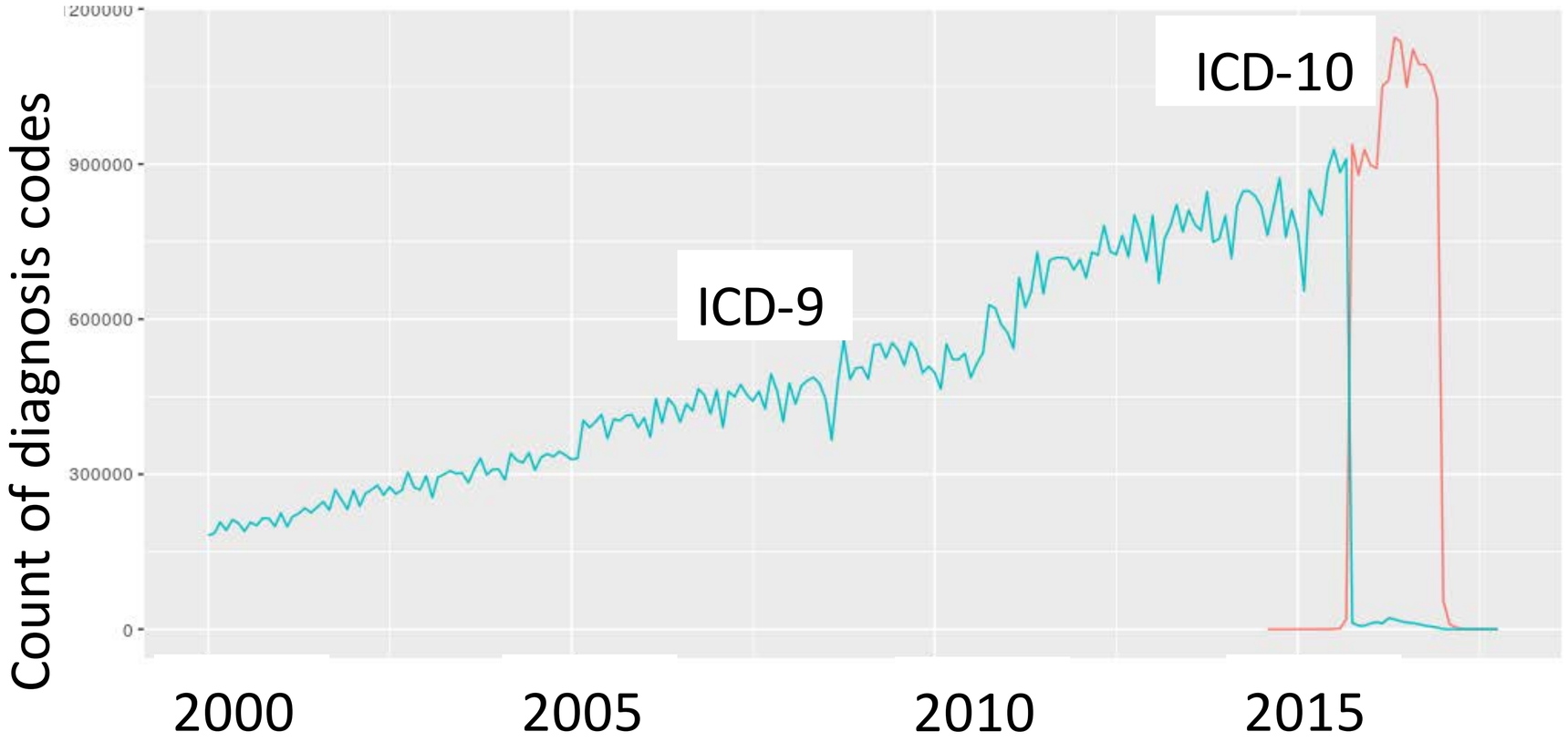
Labs



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

→ Significance of features may change over time

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

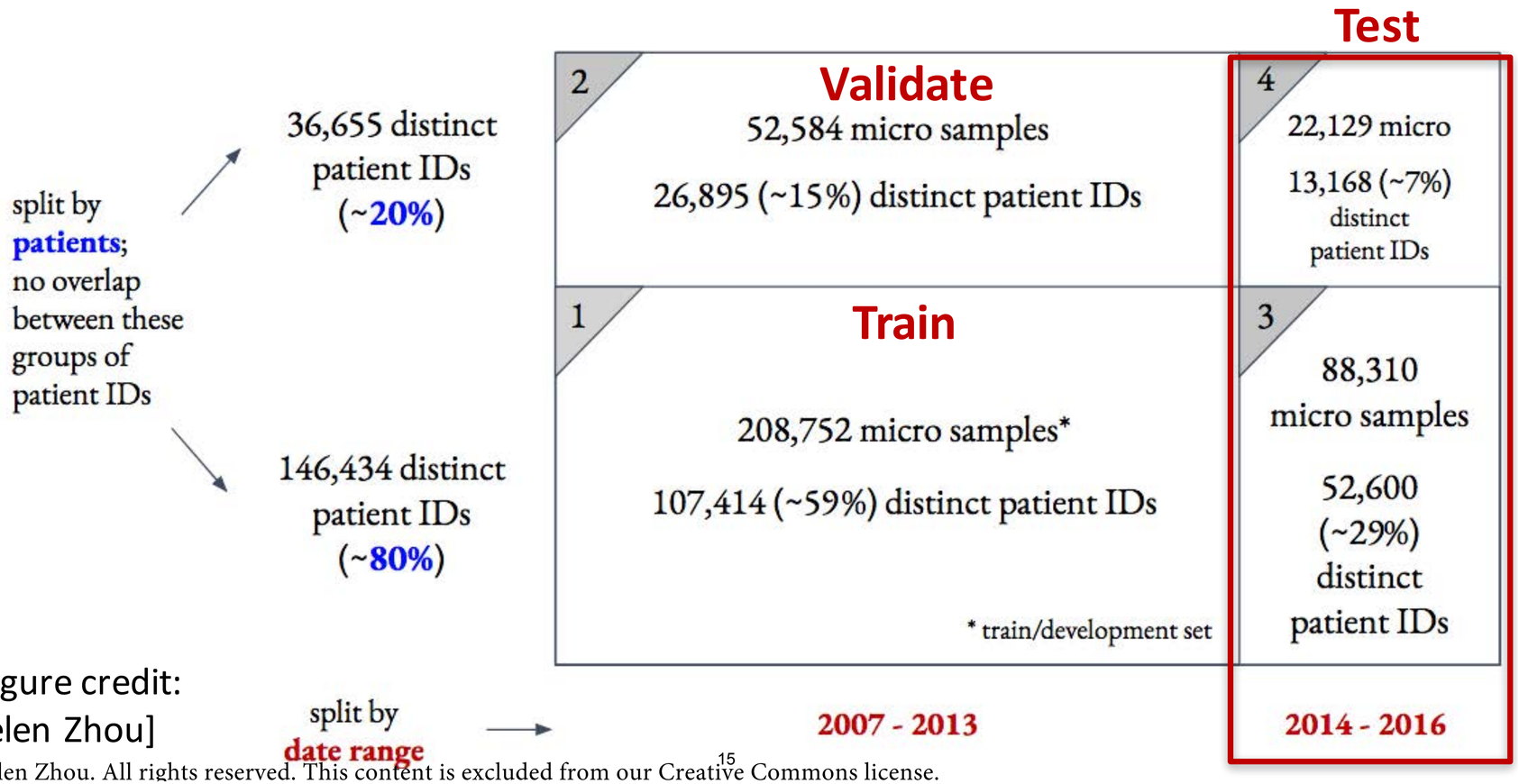


→ Significance of features may change over time

*ICD: International Classification of Diseases Code

Re-thinking evaluation in the face of non-stationarity

- **Question:** How is model evaluation flawed?
- **Answer:** Use test data from a future time period



[Figure credit: Helen Zhou]

Intervention-tainted outcomes

- Example:
  - Patients with pneumonia who have a history of asthma have lower risk of dying from pneumonia
  - Thus, **HasAsthma(x) => LowerRisk(x)**
- What's wrong with the learnt model?
  - Risk stratification drives **interventions**
  - If low risk, might not admit to ICU. But this was precisely what prevented patients from dying!

Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission

Rich Caruana
Microsoft Research
rcaruana@microsoft.com

Yin Lou
LinkedIn Corporation
ylou@linkedin.com

Johannes Gehrke
Microsoft
johannes@microsoft.com

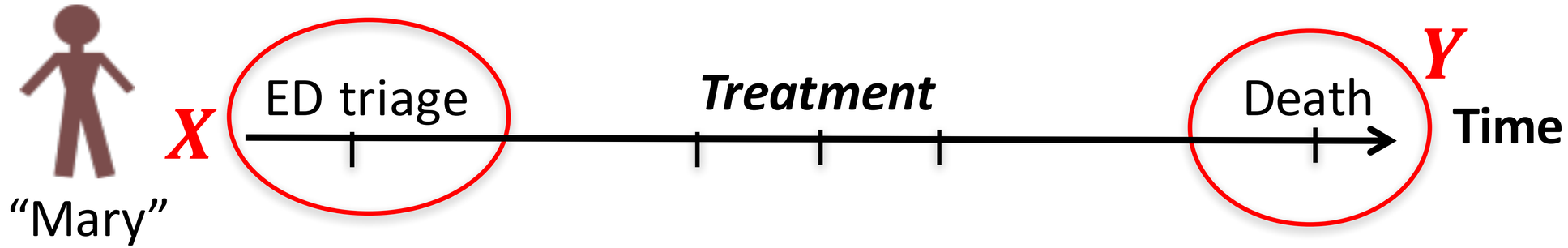
Paul Koch
Microsoft Research
paulkoch@microsoft.com

Marc Sturm
NewYork-Presbyterian Hospital
mas9161@nyp.org

Noémie Elhadad
Columbia University
noemie.elhadad@columbia.edu

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, we must recognize it is happening
  - Interpretable models help with this

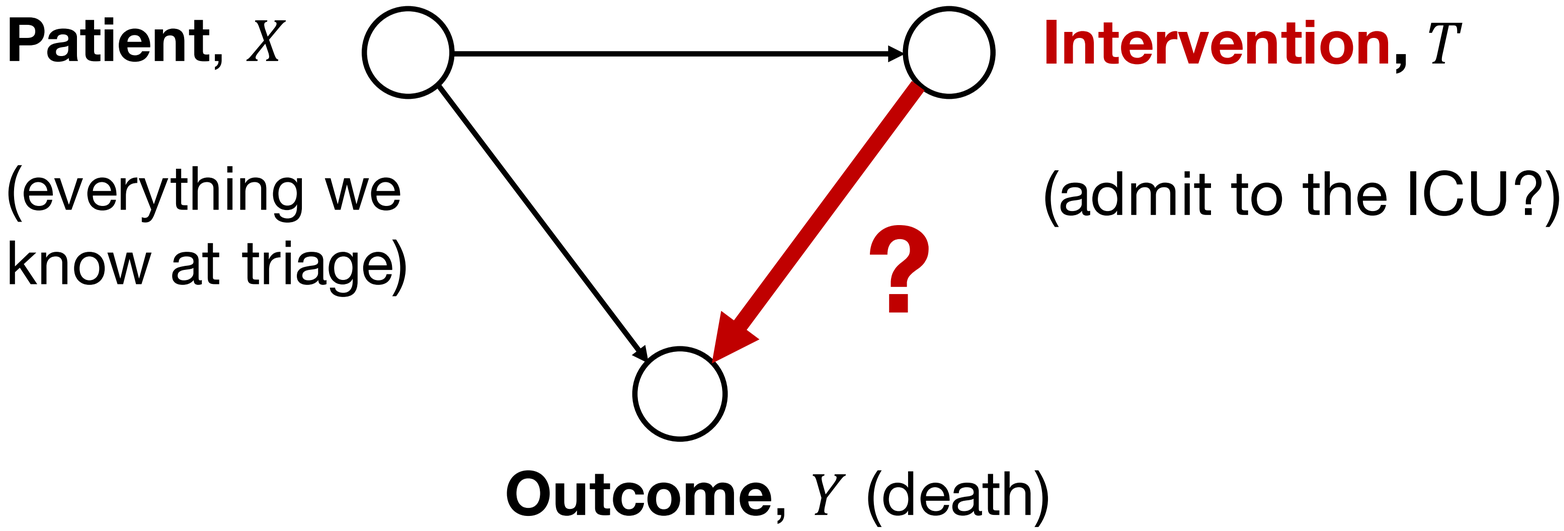
Intervention-tainted outcomes

Solutions:

- Modify model, e.g. by removing the HasAsthma(x) => LowerRisk(x) rule (will not work with high-dimensional data)
- Re-define outcome by finding a pre-treatment surrogate (e.g., lactate levels)
- Consider treated patients as right-censored by treatment

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?

What about deep learning?

npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Scalable and accurate deep learning with electronic health records

Alvin Rajkomar[1,2], Eyal Oren[1], Kai Chen[1], Andrew M. Dai[1], Nissan Hajaj[1], Michaela Hardt[1], Peter J. Liu[1], Xiaobing Liu[1], Jake Marcus[1], Mimi Sun[1], Patrik Sundberg[1], Hector Yee[1], Kun Zhang[1], Yi Zhang[1], Gerardo Flores[1], Gavin E. Duggan[1], Jamie Irvine[1], Quoc Le[1], Kurt Litsch[1], Alexander Mossin[1], Justin Tansuwan[1], De Wang[1], James Wexler[1], Jimbo Wilson[1], Dana Ludwig[2], Samuel L. Volchenboum[3], Katherine Chou[1], Michael Pearson[1], Srinivasan Madabushi[1], Nigam H. Shah[4], Atul J. Butte[2], Michael D. Howell[1], Claire Cui[1], Greg S. Corrado[1] and Jeffrey Dean[1]

	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)

Some final thoughts….

How does risk stratification differ from differential diagnosis?

Differential diagnosis	Risk stratification
Usually iterative/active	Usually passive
Often considers a large set of conditions	Often just one condition
Has to consider rare conditions (needs hybrid knowledge/ML approaches)	Often focuses on settings where there is enough training data

What is left/right censoring?

LC: occurs when we may not have enough data/time for a subject to derive a complete set of features

RC: occurs when a subject leaves the study before an event occurs, or the study ends before the event has occurred.

What is the difference between relative risk (=risk ratio) and odds ratio?

RR: Ratio of risk of an event in one group (e.g., exposed group) versus the risk of the event in the other group (e.g., nonexposed group).

OR: Ratio of odds of an event in one group versus the odds of the event in the other group. Indicates association between exposure and outcome.

Literature

- https://www.liebertpub.com/doi/epdf/10.1089/big.2015.0020

Big Data
Volume 3 Number 4, 2015
Mary Ann Liebert, Inc.
DOI: 10.1089/big.2015.0020

ORIGINAL ARTICLE

Population-Level Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors

Narges Razavian,[1] Saul Blecker,[2] Ann Marie Schmidt,[3] Aaron Smith-McLallen,[4] Somesh Nigam,[4] and David Sontag[1,*]

- https://dl.acm.org/doi/abs/10.1145/2783258.2788613

RESEARCH-ARTICLE

Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission

Authors: Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, Noemie Elhadad Authors Info & Claims

KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining • August 2015 • Pages 1721–1730 • https://doi.org/10.1145/2783258.2788613

https://www.nature.com/articles/s41746-018-0029-1

npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Scalable and accurate deep learning with electronic health records

Alvin Rajkomar[1,2], Eyal Oren[1], Kai Chen[1], Andrew M. Dai[1], Nissan Hajaj[1], Michaela Hardt[1], Peter J. Liu[1], Xiaobing Liu[1], Jake Marcus[1], Mimi Sun[1], Patrik Sundberg[1], Hector Yee[1], Kun Zhang[1], Yi Zhang[1], Gerardo Flores[1], Gavin E. Duggan[1], Jamie Irvine[1], Quoc Le[1], Kurt Litsch[1], Alexander Mossin[1], Justin Tansuwan[1], De Wang[1], James Wexler[1], Jimbo Wilson[1], Dana Ludwig[2], Samuel L. Volchenboum[3], Katherine Chou[1], Michael Pearson[1], Srinivasan Madabushi[1], Nigam H. Shah[4], Atul J. Butte[2], Michael D. Howell[1], Claire Cui[1], Greg S. Corrado[1] and Jeffrey Dean[1]

Imperial College London

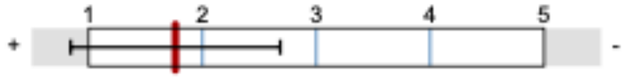
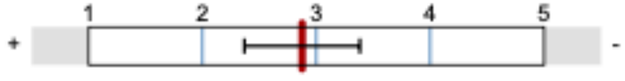
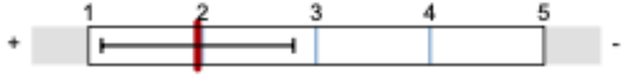
That’s all for this lecture

- now for the lecture evaluation (so far)

Univ.-Prof. Dr. Daniel Rückert
Prof. Dr. Julia Schnabel

0000000960 Künstliche Intelligenz in der Medizin I (IN2403) (WS 22/23)
Erfasste Fragebögen = 14

Globalwerte

Vorlesungskonzept	= Lecture concept		mw=1,8 s=0,8
Vermittlung der Inhalte	= Content delivery		mw=1,8 s=0,9
Kompetenzerwerb	= Competency acquisition		mw=1,8 s=0,7
Umfang und Schwierigkeitsgrad (3 best)	= Scope and level		mw=2,9 s=0,5
Eindruck Dozent*in	= Impression of lecturer(s)		mw=1,6 s=0,7
Insgesamt finde ich die Vorlesung...	= Overall I find the lecture...		mw=1,8 s=0,7
Zusatzangebot (Skalenbreite: 5)	= Additional offers		mw=1,9 s=0,9
Organisation	= ditto		mw=2 s=0,8

Individual (anonymised) comments

- Generally very positive (Thank you!)
- **Lectures**
  - Content is interesting, only in person option leads to better lecture environment
  - Everyone seems very competent, motivated and always happy to help.
  - Presents the topic very clearly
  - *The lecturers are really helpful and eager to answer questions from the students.*
  - The style of teaching is good. Especially I am happy to see that lecturers seem to care about delivering some practical experience, supporting all the theory.
  - *Slides are good and comprehensive*
  - *Some of the visual explanations were particularly good.*
  - The slides and the assignments are always uploaded on time.
  - *Themas gut vorbereitet und verlinkt zu den heutigen Anwendungen*

Individual (anonymised) comments

- Room for improvement (grouped):
- **Lectures:**
  - The lectures could perhaps be recorded (several)
  - *More detailed slides (ie more like lecture notes); more visuals, animations in the presentations would make it easier to understand, or lecture slides that have the notes of what the lecturer said*
  - Guest lectures are too fast and includes too much content, impossible to learn and follow
  - Sometimes there's no microphone and the lecturer can't be understood past the first few rows. / We had microphone issues many times. / Often we had microphone issues.
  - *It would be good to make a 5-10 second pause more often and ask if we have questions*
  - *Seeing actual clinical use of some content*

Individual (anonymised) comments

- Room for improvement (grouped):
- **Tutorials / Practicals:**
  - The given code files for assignments are very badly written. […]
  - Upload solutions of the exercises (both practicals and theoreticals) / Feedback about the practical exercise session - answers for the practical exercise […] (several)
- **Online forum:**
  - Hardly interactive and informative
- **General:**
  - [dislike] concept that of having to work with 4 [others] on projects that will directly affect final grade
  - It was not specified […] how many points you need in each project to get a specific grade
  - […] related to exam / which parts of the lecture is crucial from profs perspective might be defined clearer