Risk Stratification

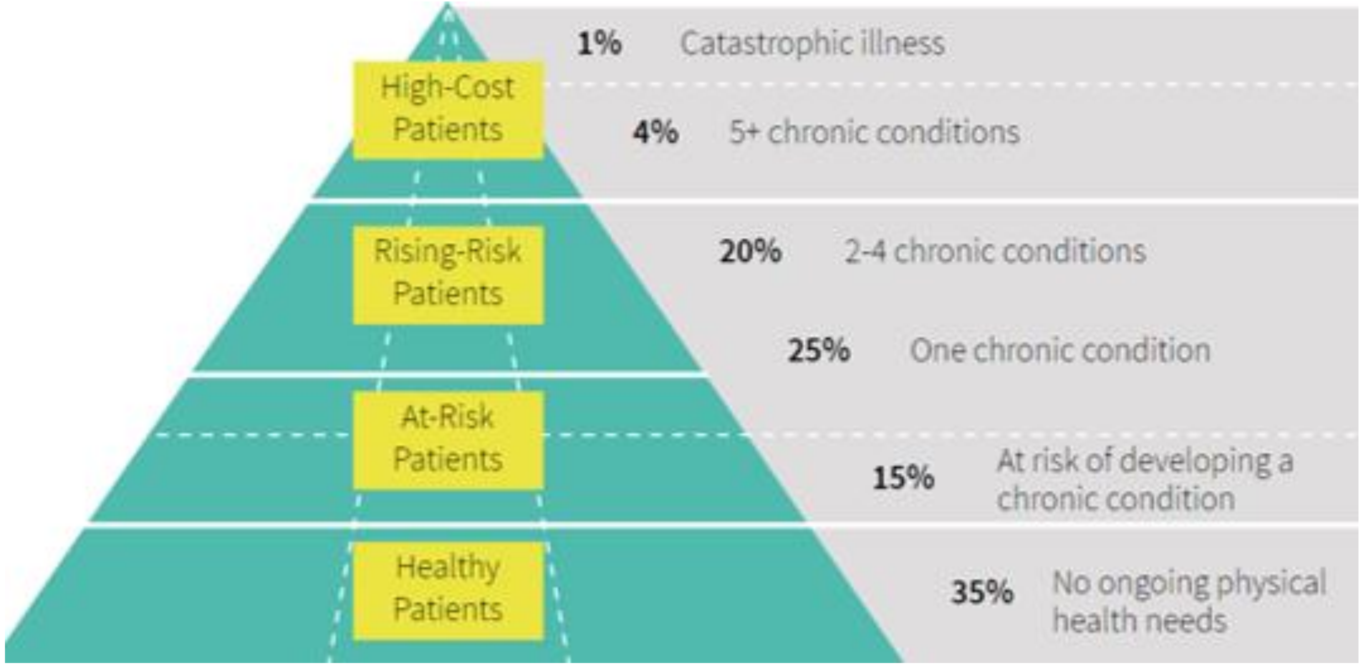
Artificial Intelligence for Digital Health (AID)
M.Sc. in Digital Health – University of Pisa
Davide Bacciu (davide.bacciu@unipi.it)

Lecture Outline

- Population health management and the need for subject stratification
- Risk scoring and stratification
- Machine learning for risk stratification
    - A simple scoring model
    - Identifying risk factors
    - Assessment and validation of risk predictors
- Data challenges in risk scoring
    - Censoring
    - Competing risks

Risk stratification fundamentals

Stratification/Segmentation



Source: English NHS

Tool designed to enhance comprehension of population needs by segmenting it into smaller, more manageable groups, with each group exhibiting comparable healthcare requirements and priorities

Population

A subgroup we want to know something about

→

Example

- Very preterm newborns
- Men aged 75+ who require coronary bypass surgery
- Women who smoke 20+ cigarettes per-day

Risk

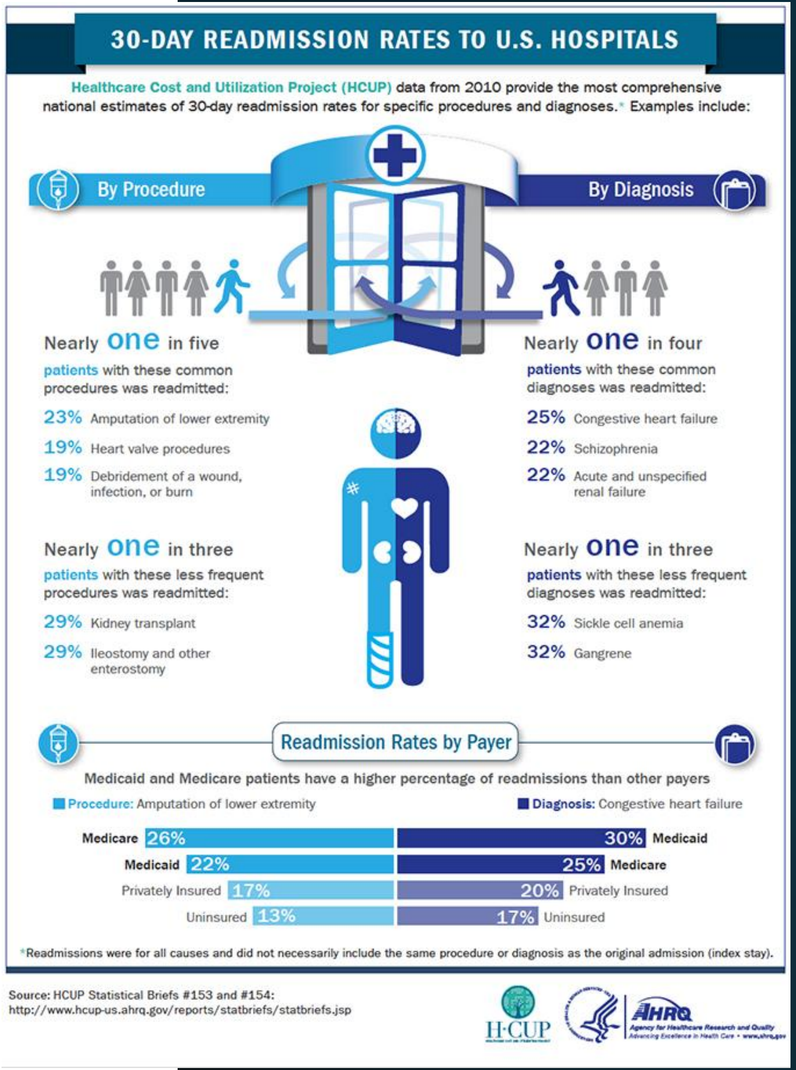
The likelihood of an adverse clinical outcome, such as a worsening of symptoms, declining quality of life due to illness or injury, or death

Risk Stratification

- Separate a patient population into multiple segments (at least 2) based on the risk of incurring into an outcome
  - A typical stratification separates high-risk individuals from low-risk individuals
- It is associated with interventions that target high-risk subjects
- Goal is population health management
  - Consider patients as both individuals and members of distinct subgroups based on their conditions and medical needs
  - A "one-size-fits-all" model, where the same level of resources is offered to every patient, is clinically ineffective and prohibitively expensive
    - Improve patient outcomes
    - Optimize healthcare systems operations

Population Health Management

- Select patients who would benefit from working with a specialist
- Identify patients for coordinated care solutions
- Improve scheduling (i.e. schedule longer appointments for higher-risk patients)
- Prioritize existing resources and identify needs gaps



Source: US AHRQ Agency

Risk Stratification Example

Source: womensagenda.com.au/tag/cardiovascular-disease/

Does a patient need to be admitted to the intensive coronary care unit?

Risk Stratification Example

Source: edition.cnn.com

What is the risk of severe morbidity as adults in early preterms?

Traditional Risk Stratification

Simple scores computed on data entered by a human (expert or patient) using questionnaire-like methods

Apgar Scoring System

	Indicator	0 Points	1 Point	2 Points
A	Activity (muscle tone)	Absent	Flexed limbs	Active
P	Pulse	Absent	< 100 BPM	> 100 BPM
G	Grimace (reflex irritability)	Floppy	Minimal response to stimulation	Prompt response to stimulation
A	Appearance (skin color)	Blue Pale	Pink body Blue extremeties	Pink
R	Respiration	Absent	Slow and irregular	Vigorous cry

HIE Help Center
hiehelpcenter.org

Apgar score measures status of newborns immediately after birth (7–10 normal, 4–6 moderately abnormal, 0–3 critical)

Apgar score for risk stratification

Newborns with lower Apgar scores are more likely than babies with higher scores to need resuscitation

Apgar Score

Gestational age______________weeks

Sign	0	1	2	1 minute	5 minute	10 minute	15 minute	20 minute
Color	Blue or Pale	Acrocyanotic	Completely Pink					
Heart rate	Absent	<100 minute	>100 minute					
Reflex irritability	No Response	Grimace	Cry or Active Withdrawal					
Muscle tone	Limp	Some Flexion	Active Motion					
Respiration	Absent	Weak Cry; Hypoventilation	Good, Crying					
			Total					

Comments:

Resuscitation

Minutes	1	5	10	15	20
Oxygen					
PPV/NCPAP					
ETT					
Chest Compressions					
Epinephrine					

Diabetes risk stratification

- Allows self assessment
- Computes 5 risk classes

Finnish Diabetes Association

Type 2 diabetes risk assessment form

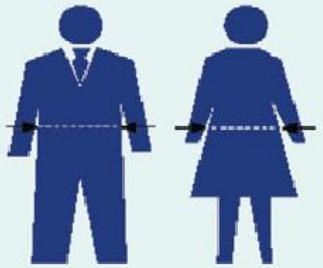
Circle the right alternative and add up your points.

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 25 kg/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 94 cm	Less than 80 cm
3 p.	94–102 cm	80–88 cm
4 p.	More than 102 cm	More than 88 cm



4. Do you usually have daily at least 30 min 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 antihypertensive 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 one in 100 will develop disease
7–11	Slightly elevated: estimated one in 25 will develop disease
12–14	Moderate: estimated one in 6 will develop disease
15–20	High: estimated one in three will develop disease
Higher than 20	Very high: estimated one in 2 two will develop disease

Please turn over

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

Limitations of traditional risk assessment

- A manual screening step needs to be done for every subject
  - Either in the physician's office or as surveys
  - Costly and time-consuming
  - Infeasible for regular screening for millions of individuals
- Not easy to adapt to multiple surrogates, e.g. when variables are missing
  - Discovery of surrogates not straightforward

Both can be attacked by a data-driven approach leveraging machine learning

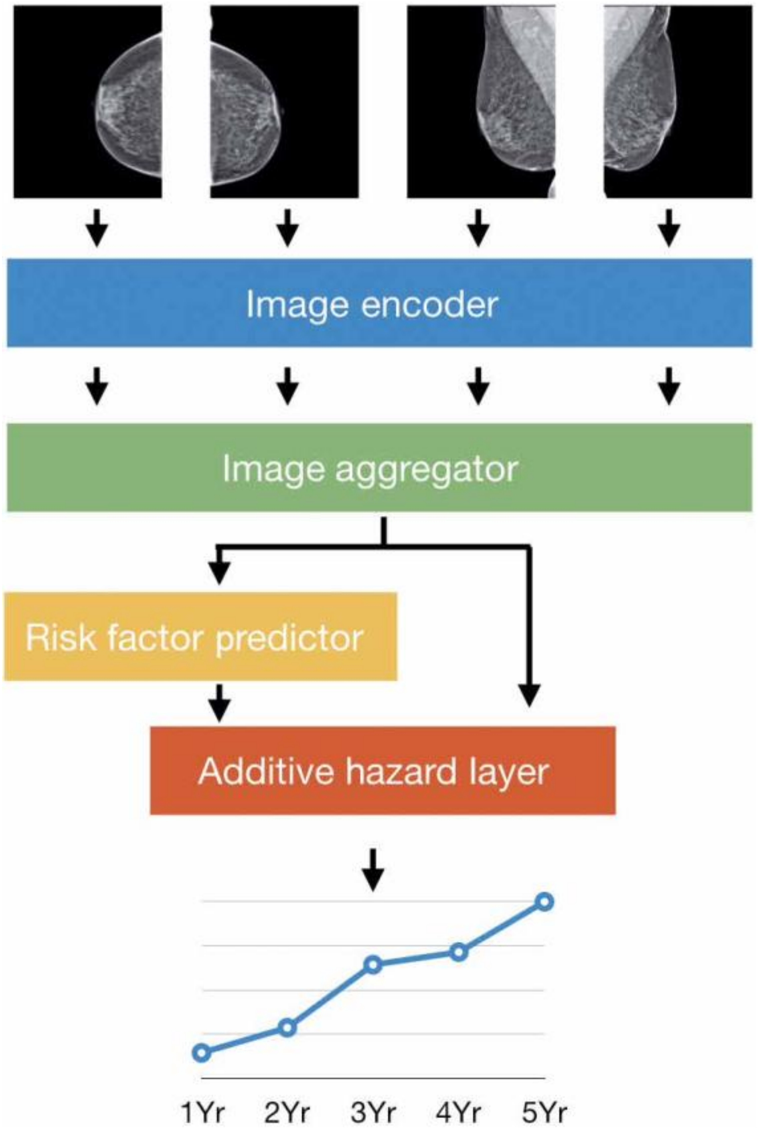
Machine learning for risk stratification

ML for risk stratification

Data and problem characterized by critical temporal aspects

- Prediction time in the future
- Availability of data from a starting indexing point

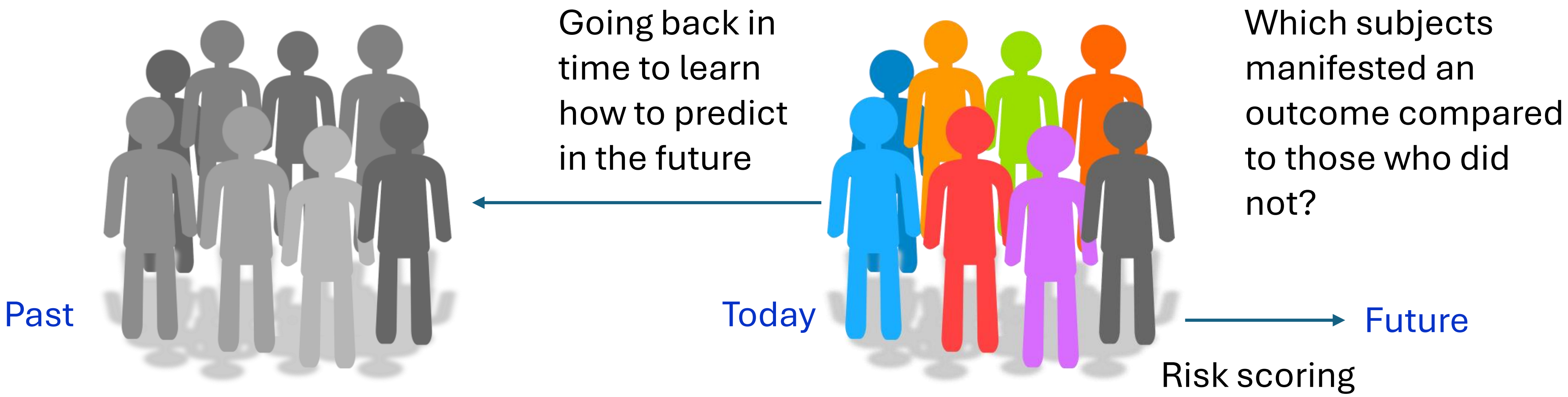
Risk stratification often coupled with the problem of identifying the relevant risk factors



Source: Yala et al 2021, Science Translational

Longitudinal Data

- Measurements are taken for the same subjects on multiple occasions
  - Population-level predictions
  - Individual-level predictions in the future
- Different from time series data, where measurements are taken for a single subject (or few subjects) for a long period of time and inference is conducted within the subject



Risk scoring

- Risk score is a numerical representation of the likelihood that a patient will experience a particular health event
- Predicted using ML models incorporating various patient-specific factors

$$RiskScore = w_1x_1 + w_2x_2 + \cdots + w_Dx_D$$

Linear risk modeling

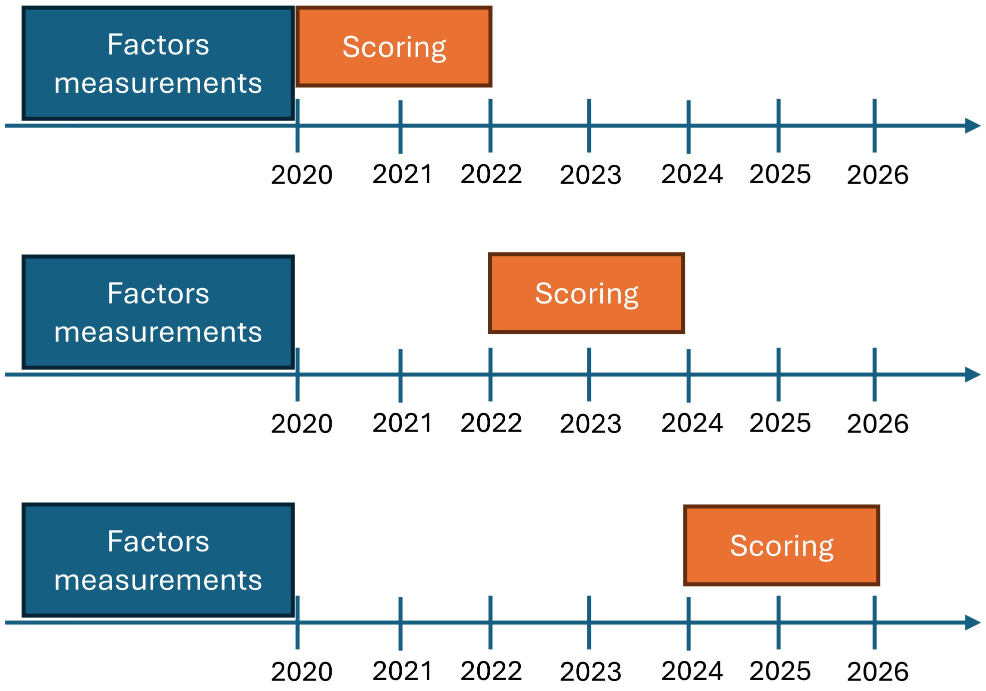
x_i are input variables or features related to the patient's health, e.g. age, blood pressure, cholesterol levels, lifestyle, genetic markers

w_i are coefficients assigned to each feature and determine how much each factor contributes to the final risk score (risk factors)

A **higher risk score** indicates a **greater likelihood** of developing the outcome

Risk Scoring as Supervised Learning

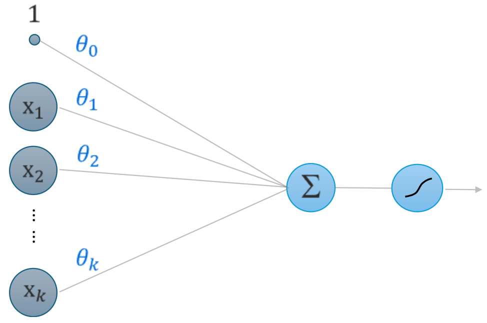
Consider the problem of stratifying high-risk vs low-risk subjects w.r.t developing an outcome at different time periods in the future



- Can be casted as three binary classification problems
- Building on the linear risk modelling
- Biasing the model to focus on few relevant factors

Any suggestion?

The Return of Logistic Regression



$$\mathrm{P}(\overset{\text{outcome}}{y} = 1|\overset{\text{factors}}{\boldsymbol{x}}) = \sigma\left(\sum_{k=1}^{D} \theta_k x_k\right) = \sigma(\boldsymbol{\theta x})$$

$$Loss = \sum_{n=1}^{N} \underbrace{-\mathrm{y}\log\big(\sigma(\boldsymbol{\theta x}_n)\big) - (1-\mathrm{y})\log(1-\sigma(\boldsymbol{\theta x}_n))}_{\text{Score fitting to data}} + \underbrace{\lambda \parallel \boldsymbol{\theta} \parallel_1}_{\text{Bias towards few factors}}$$

Logistic regression is a widely used model in risk scoring because of a nice property of its coefficients

Odds Ratio (OR)

- The ratio of the odds of an outcome O occurring in a group characterized by a certain condition/exposition (C) versus those of a group without the condition (N)

$$OR = \frac{P(O|C)/(1-P(O|C)) \longrightarrow \boldsymbol{ODD_C}}{P(O|N)/(1-P(O|N)) \longrightarrow \boldsymbol{ODD_N}}$$

- OR > 1: Higher odds of the outcome in the exposed group C
- OR = 1: No difference in odds
- OR < 1: Lower odds of the outcome in the exposed group C

Quantifies the strength of the association between an outcome and a condition/exposition, providing hints on risk factors

Odds Ratio - Example

- Outcome: lung cancer
- Condition: smoking/not-smoking
- Given a case study:
  - Smokers: 30 have lung cancer, 70 do not.
  - Non-Smokers: 10 have lung cancer, 90 do not.
- Odds of lung cancer in smokers: 30/70 = 0.43
- Odds of lung cancer in non-smokers: 10/90 = 0.11

$$OR = \frac{0.43}{0.11} = 3.9$$

- Suggests that smokers have 3.9 times higher odds of developing lung cancer than non-smokers.

OR and Logistic Regression

- The coefficients of the logistic regression is associated to the odds ratio between the outcome variable and the corresponding free variables
- Given

$$\mathrm{P}(y = 1|\boldsymbol{x}) = \sigma\left(\sum_{k=1}^{D} \theta_k x_k\right)$$

Then

$$OR\left(y \middle| x_k, \boldsymbol{x}_{1/k}\right) \approx e^{\theta_k}$$

That is the OR between outcome y and risk factor x_k when the other independent variables $\boldsymbol{x}_{1/k}$ are fixed

Predicting Risk in Time

- We may want to capture (and predict) how a subject risk evolves in time
  - E.g. a daily risk assessment of hospitalized patients
- We introduce time marked patient features and outcomes (e.g. for each day of patient admission)
  - $\boldsymbol{x}_t \rightarrow$ patient features at time t (e.g. lab results or medications taken on the day)
  - $y_t \rightarrow$ patient risk at time t (e.g. patient infected with hospitalization-related disease)
- A naïve solution would consider a straightforward adaptation of logistic regression

$$\mathrm{P}(y_t = 1|\boldsymbol{x}_t) = \sigma\left(\sum_{k=1}^{D}\sum_{t}\theta_k x_{tk}\right) = \sigma(\boldsymbol{\theta}\boldsymbol{x}_t)$$

Accounting for Time Varying Effects

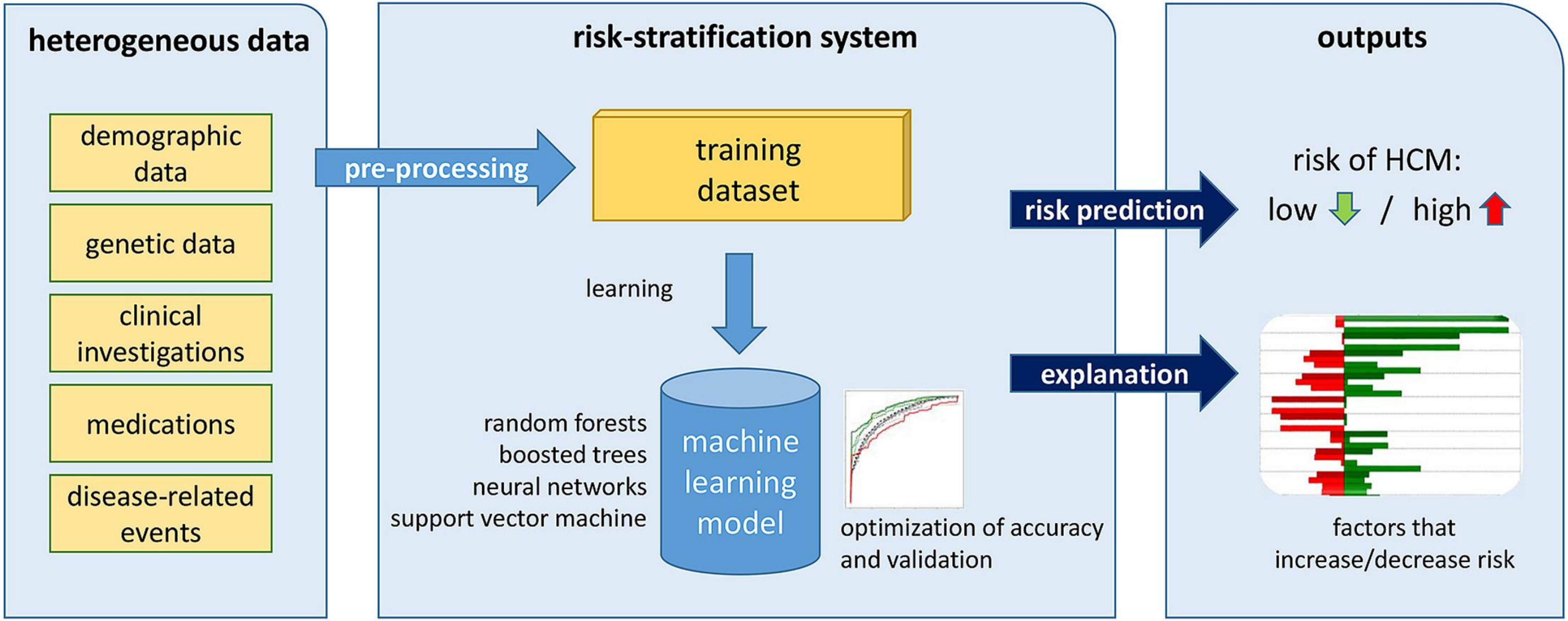
- Previous formulation pools all training examples together and learns a single model
  - Model parameters $\boldsymbol{\theta}$ are independent of time
- We may want a model that changes as the patient spends more time in the hospital as we expect the factors contributing to patient risk to change with time
  - Partition days into periods $j = 1 \ldots T$ with τ_j being the set of days of the j -th period
  - Re-parameterize the logistic regression such that the subject risk on day $t \in \tau_j$ is proportional to $(\theta_0 + \theta_j)\boldsymbol{x}_t$

Shared time-invariant knowledge

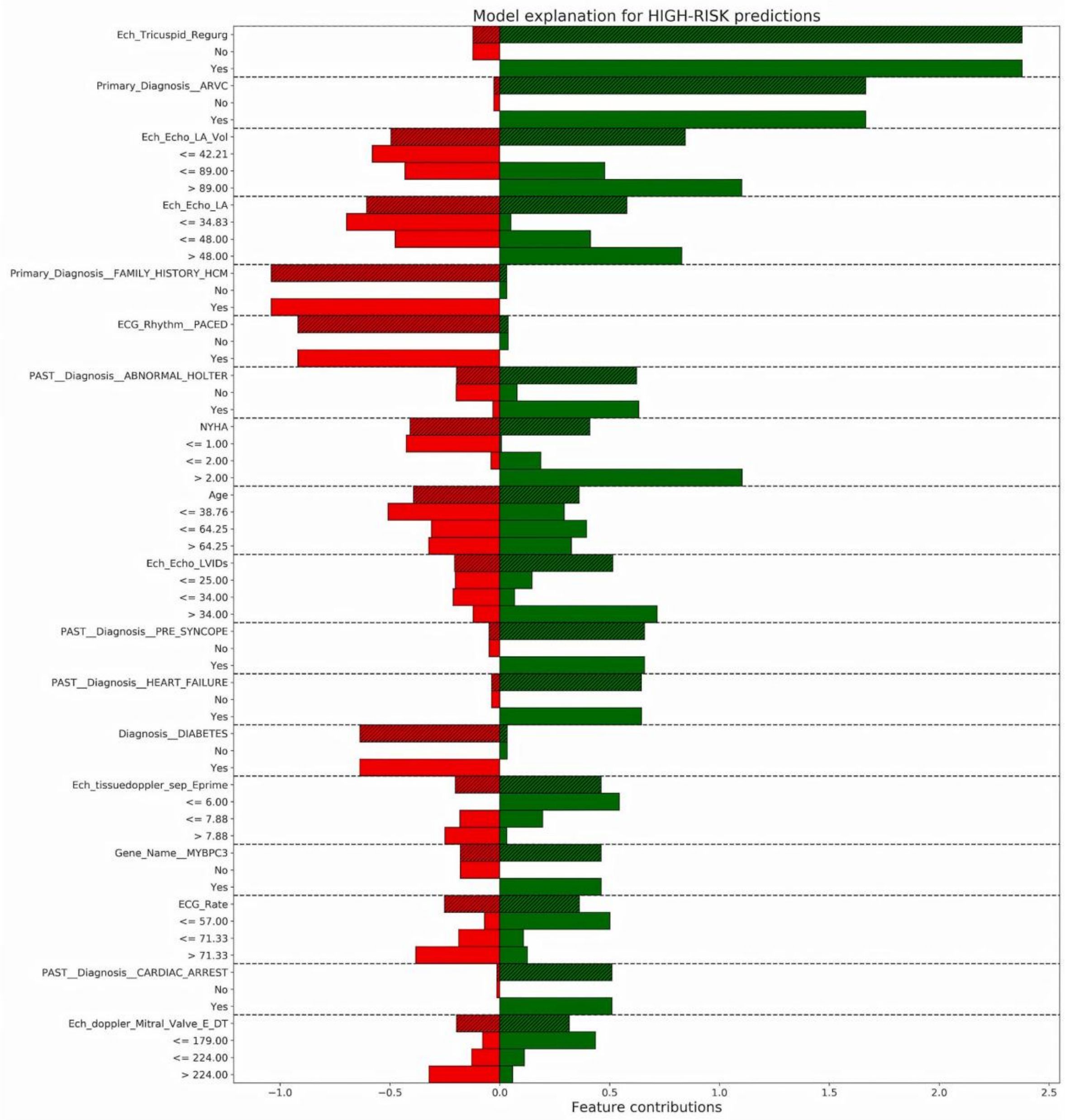
Changing time-specific knowledge

Do we need to use only logistic regression?

Source: Smole et al, CIBM journal, 2021



Source: Smole et al, CIBM journal, 2021



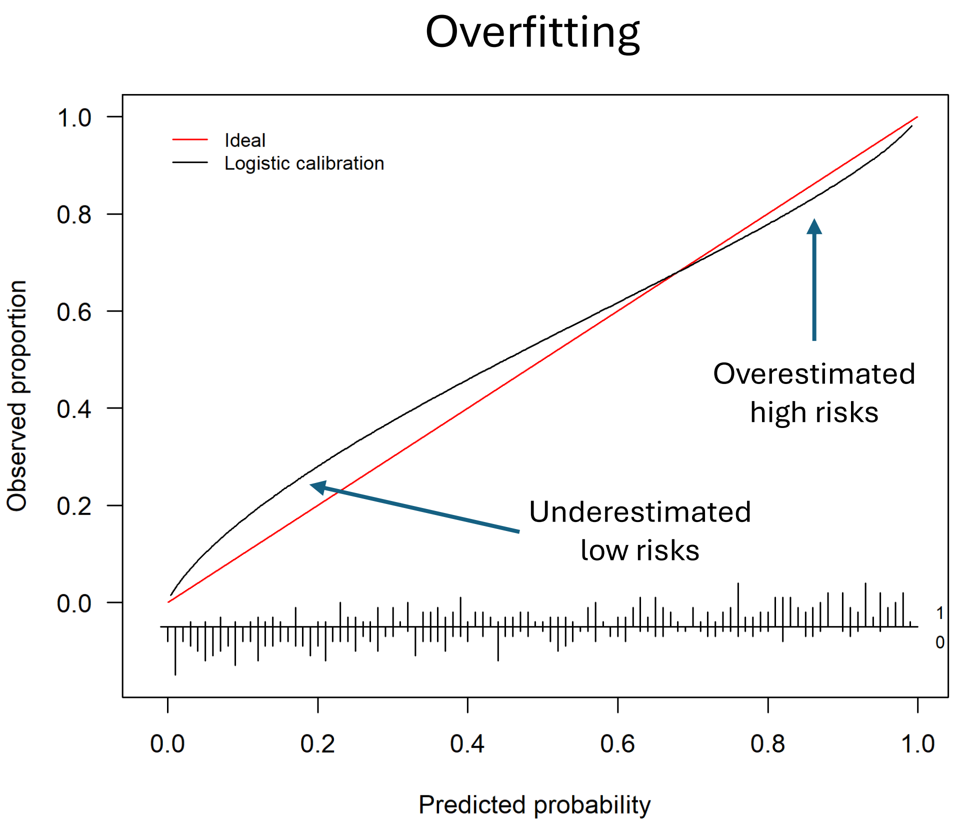
Of course not, but you may want to retain the ability to assess risk factors

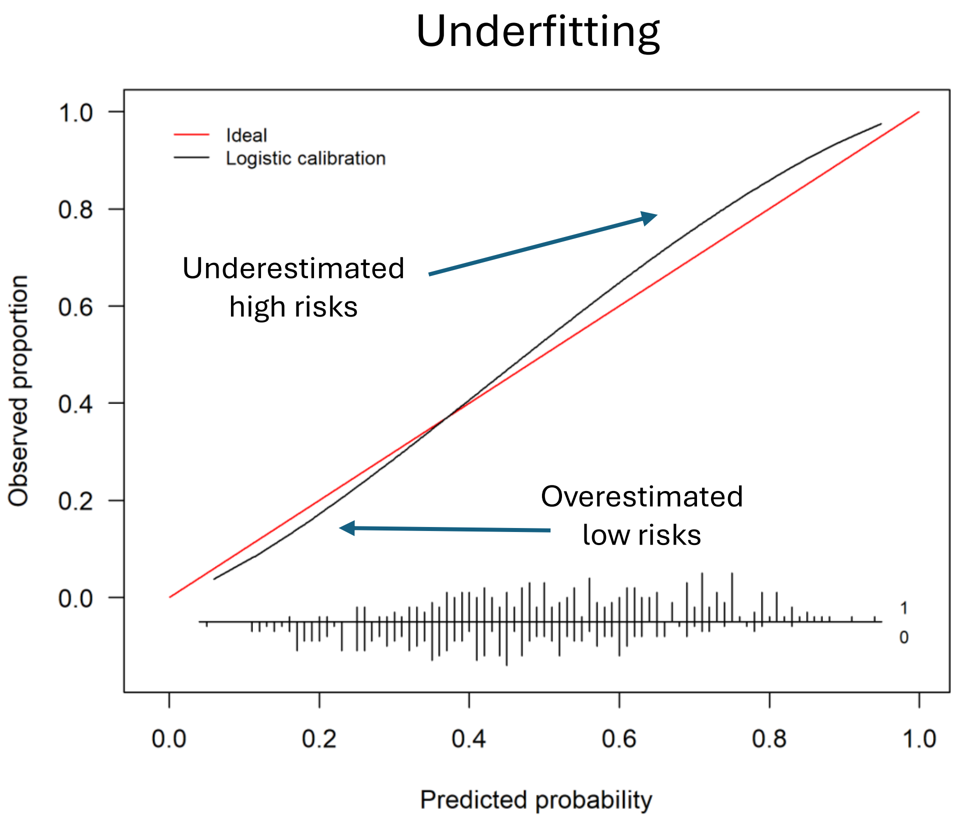
Key caveats in risk prediction models

Performance assessment in risk prediction

- Calibration - The ability to accurately predict the absolute risk level
  - In binary outcomes it is assessed by looking at the difference between the mean observed risk and the mean predicted risk
  - To provide an assessment that generalizes beyond the training set, it needs to be computed in cross-validation settings (calibration set)
- Discrimination - The ability to accurately separate individuals into low and high risk
  - A popular discrimination measure for binary outcomes is the area under the receiver operating characteristic curve (AUC)

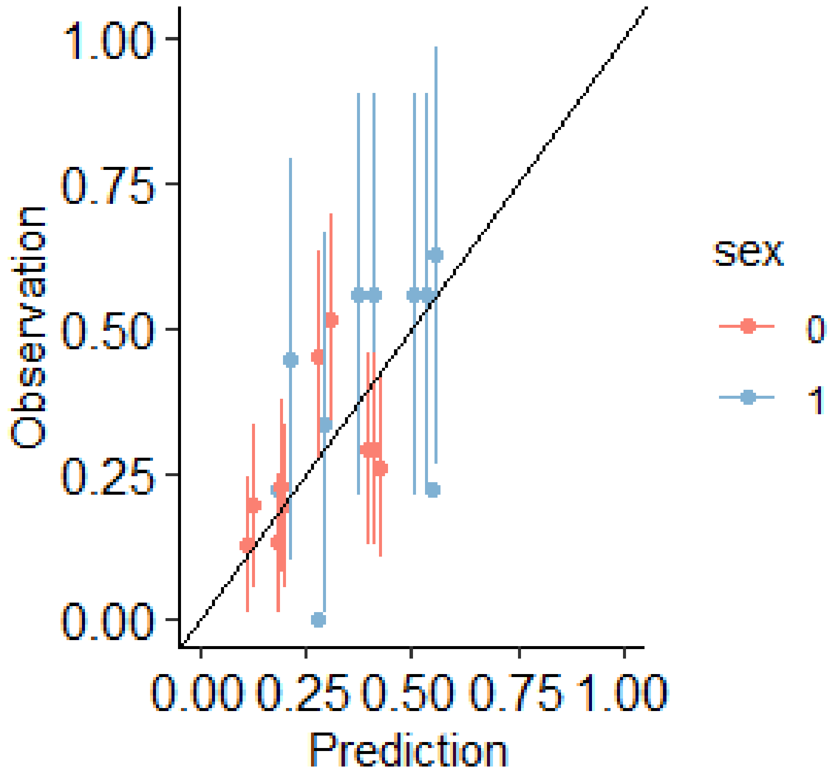
Calibration Plot





Source: cran.r-project.org/web/packages/CalibrationCurves

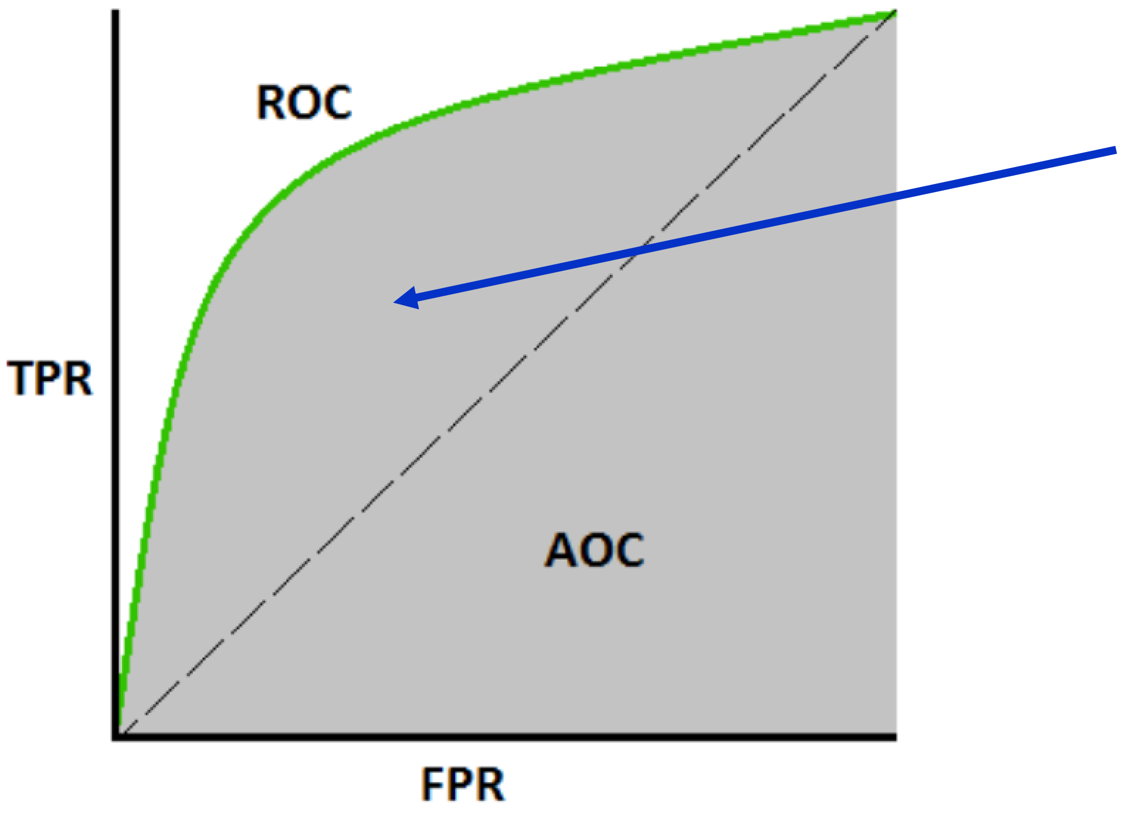
Calibration Plot on Observations



Calibration can be evaluated also according to a grouping factor

Source: cran.r-project.org/web/packages/predtools

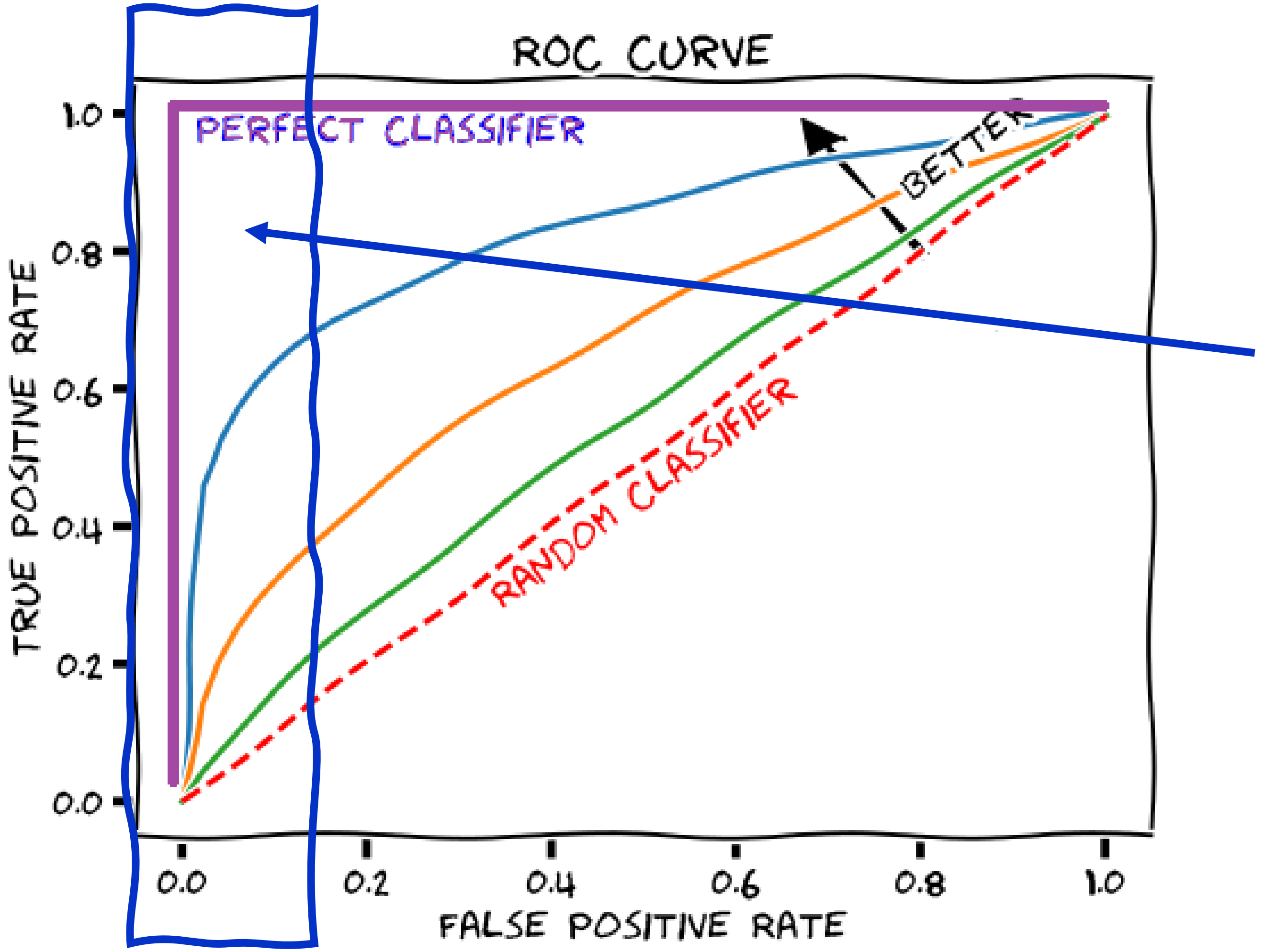
AUC and Risk Stratification



AUC = Probability that model ranks a positive outcome subject over a negative one

Invariant to class imbalance

AUC and Risk Stratification

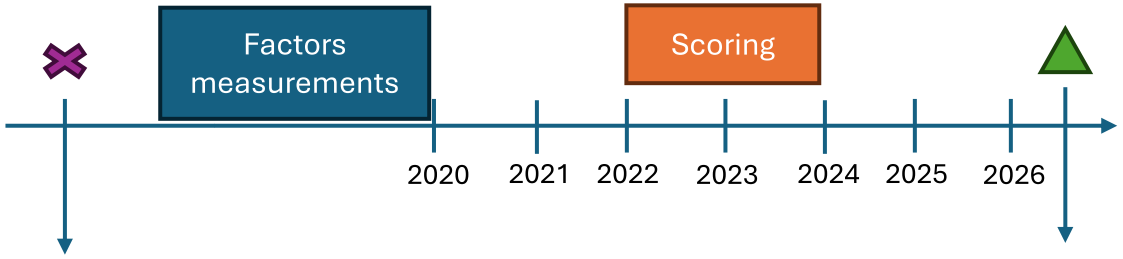


Risk stratification usually focuses on just this region (because of costs of interventions in false positives)

Can tune the classifier to operate in this region by adjusting the decision threshold

Censoring

A.k.a. why risk stratification may not be your standard classification problem



The observations of some factors for some subjects may not be available (e.g. it was collected in a different hospital)

Left censoring

Outcomes for some patient may not have materialized yet (so labelling of the outcome is not available)

Right censoring

Exclude patients that are either left- or right-censored may substantially deplete or bias your dataset

Dealing With Left-Censoring

- Imputation is an approach to fill-in the voids when observations are not available for some features
  - Typically, does not work well with left-censoring due to the large amount of missing values
- Instead, we engineer features to include many binary variables indicating whether a particular observation is available for an individual

Level of LDL cholesterol

$x_1 = 152$

Instead of representing this we expand it into multiple factors

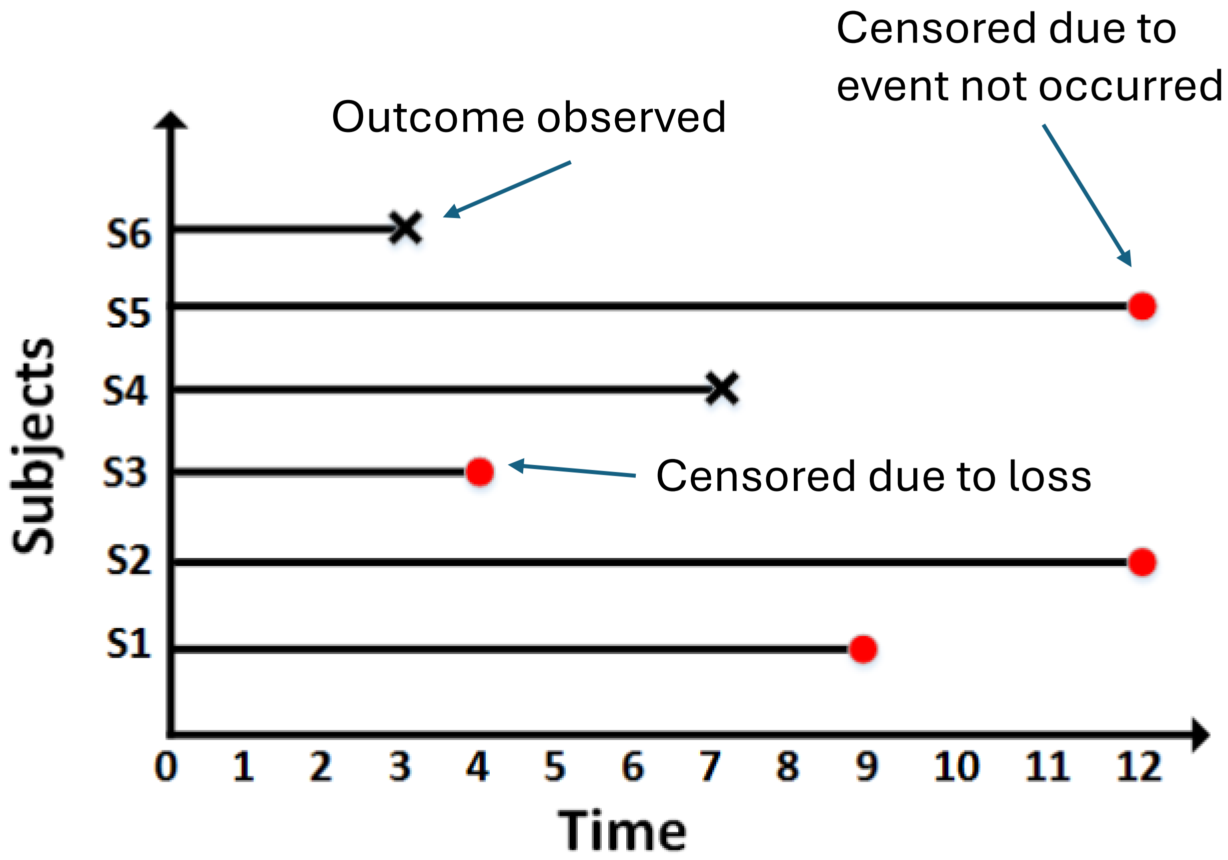
$$x_1 = \begin{cases} 1 \text{ if LDL test available} \\ 0 \text{ o.w.} \end{cases}$$

$$x_2 = \begin{cases} 0 \;\text{ test result normal} \\ 1 \;\text{ test result abnormal} \\ 2 \text{ test result high} \end{cases}$$

$$x_3 = \begin{cases} 0 \text{ result decreasing} \\ 1 \text{ result increasing} \end{cases}$$

Note that these features are typically computed for different time windows

Dealing with Right-Censoring



- With right-censored data we need to steer away from a classification-based approach
- Filtering censored data typically leaves with too few subjects for training
    - Introduces bias creating pessimistic models due to focus on non-censored data
- Classification is too coarsely grained: we will be interested in predicting the time-to-event (rather than event occurrence)

Need a new approach: survival modelling (next lecture)

Competing Risks

- Competing risks arises when we have multiple outcomes of interest and the occurrence of one outcome prevents us from observing another outcome.
- Consider heart disease and mortality as two outcomes of interest
  - If a patient dies from a cause other than heart disease, we can no longer observe the heart disease outcome
  - We have full observation of death outcome but only partial about hearth disease (we only know that occurrence time is greater than time-of-death)

Wrap-up

Take home lessons

- Risk stratification is a powerful tool to categorize patients into segments based on their risk of incurring an outcome
- Stratification helps in targeting interventions for high-risk individuals and optimizing healthcare system operations (Population Health Management)
- Risk stratification can be addressed as a classification tasks in machine learning, with some caveats
  - There are critical factors concerning temporal aspects of the data and task
  - The ability to single out and measure the contribution of relevant risk factors is key for the biomedical application
- Logistic regression as an effective tool for risk scoring
  - Odds ratio associated with feature parameters quantify the strength of the association between an outcome and a condition
  - Can be extended to consider time-varying behaviors
- Performance assessment in risk prediction models concerns
  - Calibration - how well the predicted probabilities of an event match the actual observed outcomes
  - Discrimination - ability to separate individuals into low and high-risk categories
- Risk modeling is characterized by censored data: we do not have complete information about a subject or their outcome within the study period

Next Lecture

- Survival analysis
  - From event prediction to time-to-event regression
- Survival function estimation with baseline statistical models
  - Kaplan-Meier
  - Cox regression
- A broader view into machine learning for survival analysis