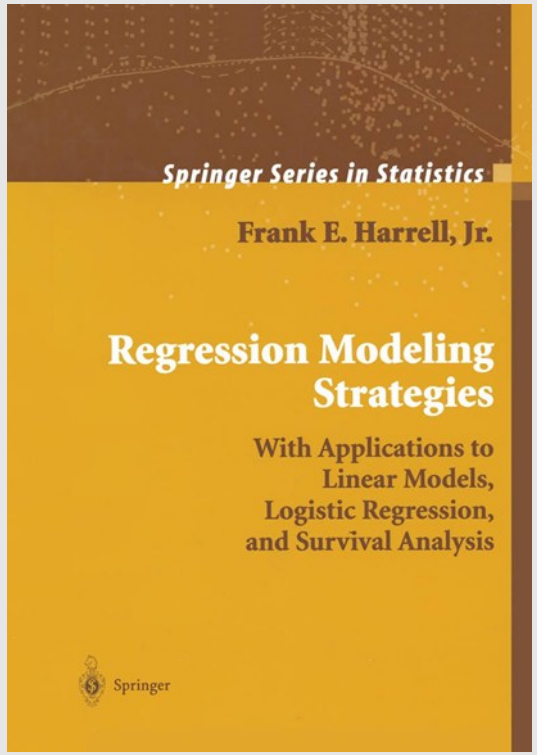
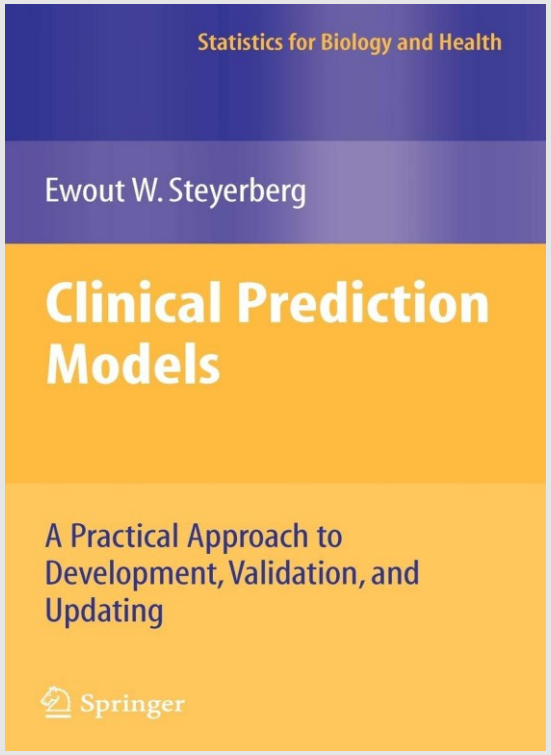
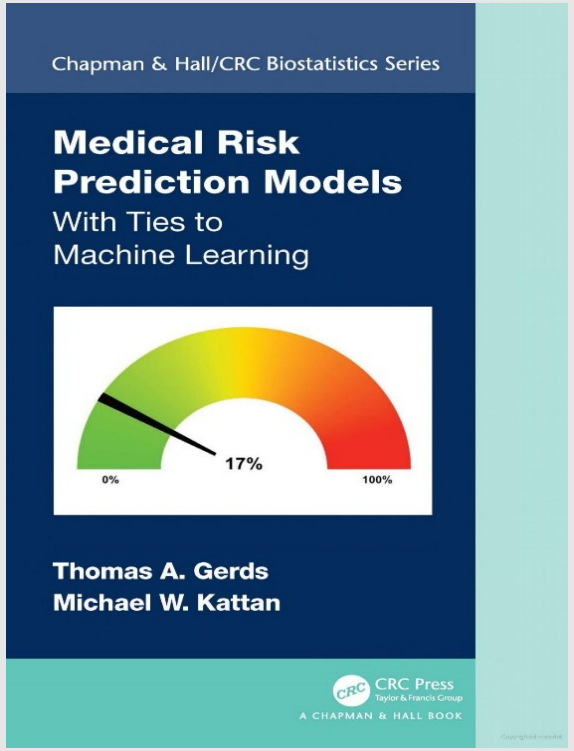
Prediction Models in Epidemiological & Clinical Research (I)







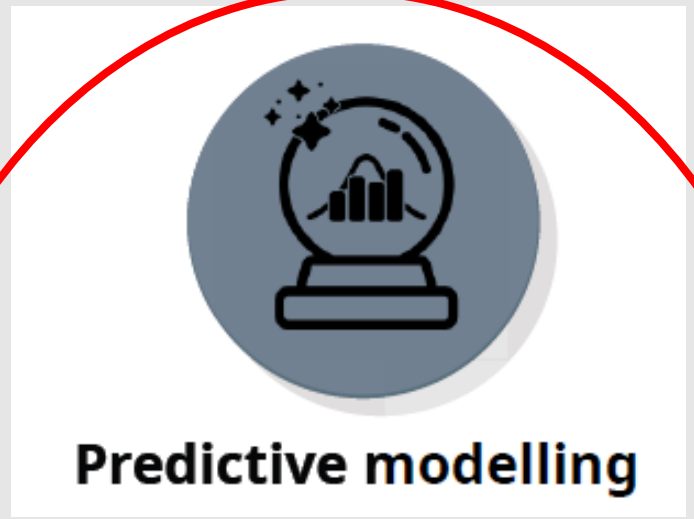
All models are wrong, but some are useful
Box & Draper, 1987

Statisticians, like artists, have the bad habit of *falling in love* with their models.
G.E.P. Box

Different scientific aims

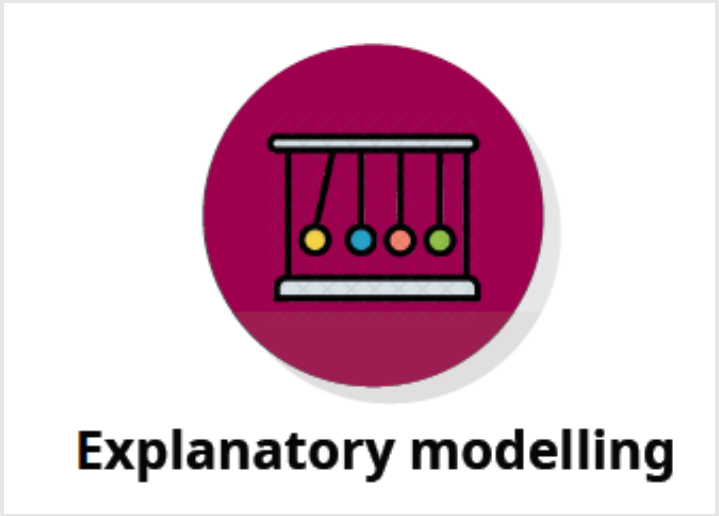
Describe the outcome of interest:
which factors *affect* it and how?

Estimate a *prevalence* in function of age and sex



Accurate predictions of future observations.
No concern about causality and confounding (**association**)

Risk of developing CVD in the next x years



Testing and comparing existing causal theories.

Effect of LDL on CVD risk

Policy recommendations (statins)

Independently from the primary scientific aim, a **basic statistical tool** is the REGRESSION model approach.

The key difference is in **the scope of** the REGRESSION model according to the main aim...

The **Descriptive** Aim (The "What")

Regression as a tool for **parsimonious** summarization.

Describing the Population: How regression describes the *average* individual.

Trend Identification: Using splines or polynomials to describe non-linear patterns

Data Reduction: Using regression to simplify complex datasets into understandable trends.

The Predictive Aim (The "Who")

Focus on accuracy and generalizability of predictions.

The **Bias-Variance** Tradeoff: Why "more variables" isn't always better.

Feature Selection: we don't care about the specific role of a variable (if it is a "confounder" or "independent predictor"..) only if *it adds signal*.

Validation: The necessity of Cross-Validation and External Validation.

The Causal Aim (The "Why")

Focus on unbiased estimation of an **effect** of an intervention/exposure.

Counterfactual Framework: Potential Outcomes.

Directed Acyclic Graphs (DAGs): How to choose variables based on theory, not p-values...

Identifiability: assumptions of exchangeability, positivity, and consistency.

We are moving to an era of ***personalized*** evidence-based medicine that asks for an ***individualized*** approach to shared medical decision-making.

In **evidence-based medicine** a central place is reserved to results from RCTs (***average*** effect) - sometimes grouped in meta-analyses.

Effect of <u>one specific</u> treatment/exposure of interest.

Prediction* models** summarize the effects of **multiple** predictors to provide more ***individualized predictions of the **risk** of a diagnostic or prognostic outcome.

"Personalized" predictions central to many domains of medicine:

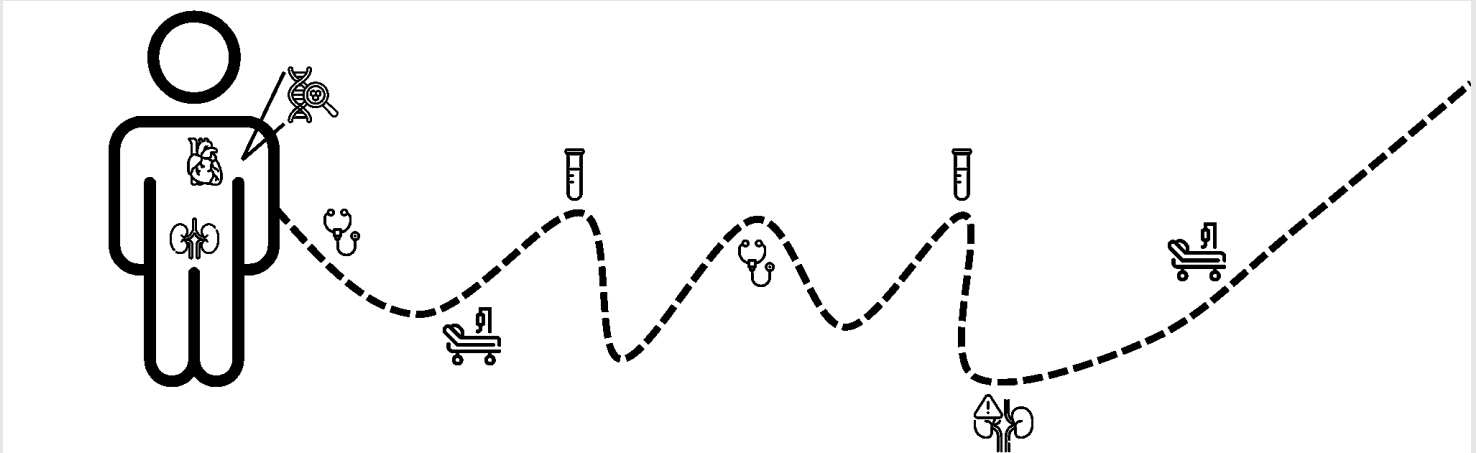
- **Screening**: find diseases early and treat better. Whether screening is useful depends on the improvement in prognosis compared to a *no screening* strategy.

- **Diagnosis**: Estimate the probability of having a disease without invasive tools, based on patient's characteristics.

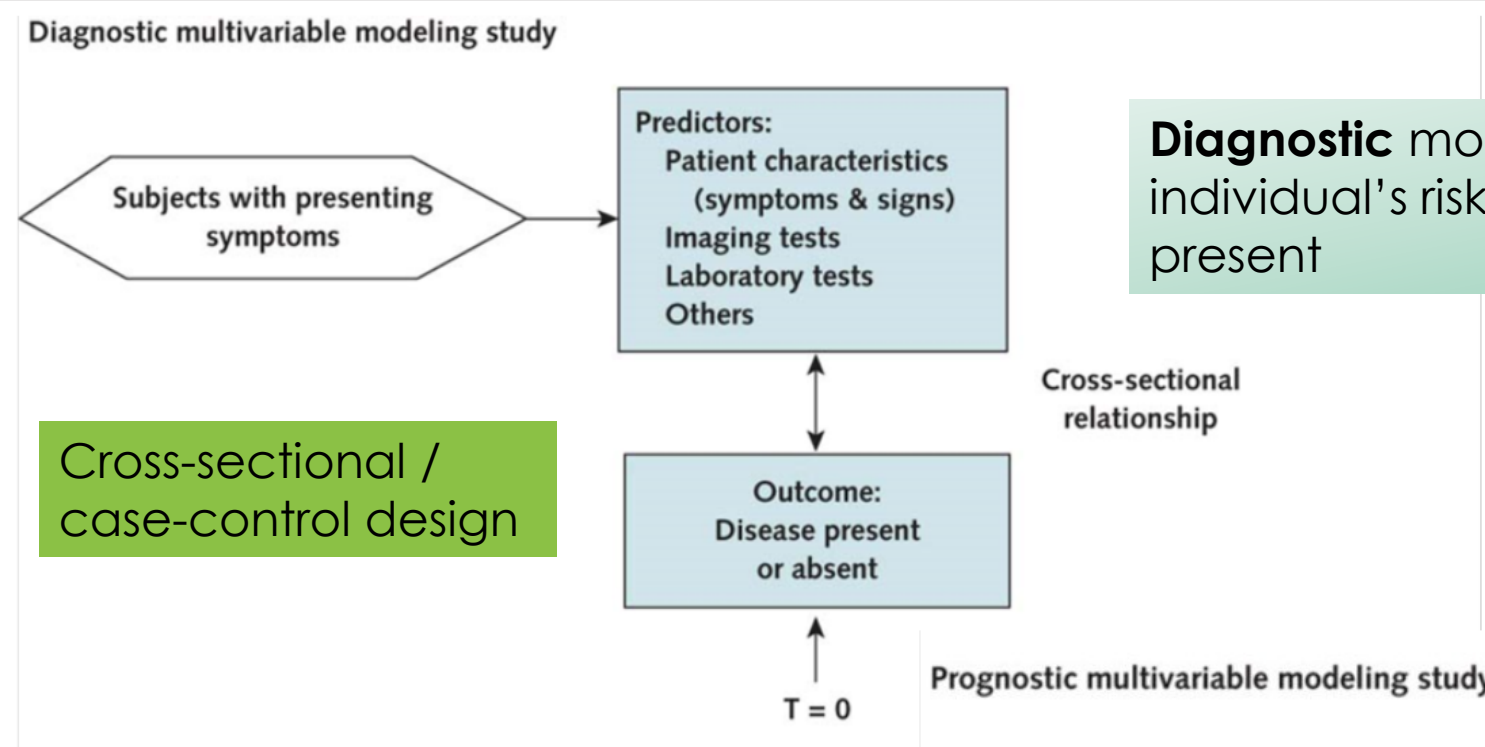
- **Prognosis**: The "individual" trajectory given a specific disease (i.e. risk of unfavourable outcomes)

Prediction model

[the term **prediction** is more general than **prognostic**, since we could also take into account possible ***interactions*** *between patient's features and treatments]*



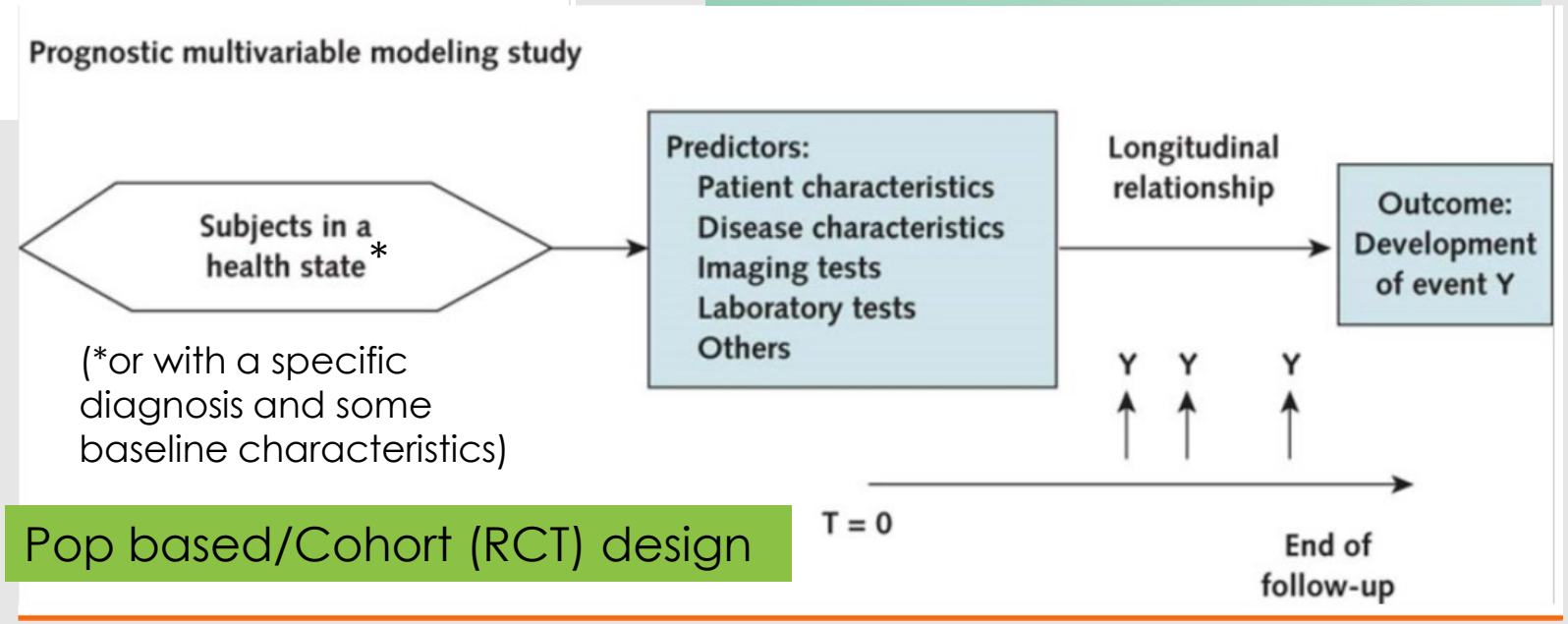
Diagnostic / Prognostic models



Diagnostic models aim to estimate an individual's risk that a disease is ***already*** present

Cross-sectional / case-control design

Estimate the risk of particular health state ***occurring in the future***



(*or with a specific diagnosis and some baseline characteristics)

Key difference : **temporal relationship** between the moment of prediction and the outcome

Pop based/Cohort (RCT) design

Prognosis/Prediction

1. Overall prognosis is the ***average risk*** of an outcome or the expected value of an outcome (e.g. pain score) among people with the health condition of interest in a particular healthcare setting

2. Prognostic factor whose values (levels) ***are associated with*** changes in the outcome's risk or expected value

3. Prognostic model individual's outcome risk or expected outcome value using **combinations** of **prognostic** factors.

4. *Prediction* model how to ***tailor treatment decisions*** for individual patients according to **whether they are likely to benefit** from particular treatments (overlap with **causal** models).

Prediction is about getting a **probability/risk** of the outcome of interest (e.g., what is my risk of developing CVD over the next 10 years) specifically ***IF I*** **do some therapy/change** in lifestyle vs not.

https://www.prognosisresearch.com/

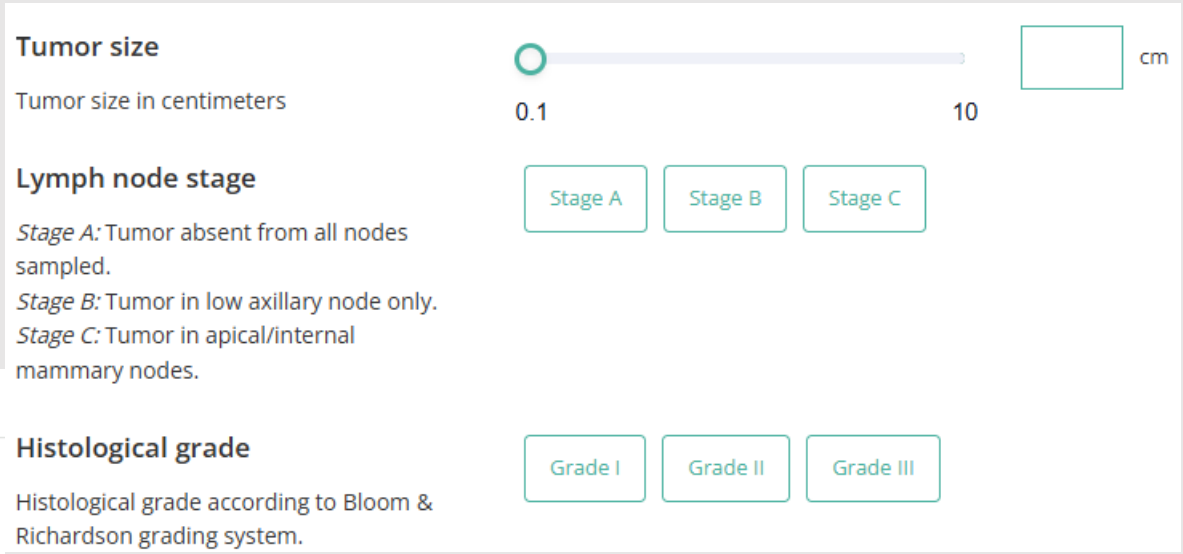
Examples

1. **Overall prognosis** 5 out of 6 women diagnosed with breast cancer in the UK in 2020 will be alive in 2025

2. **Prognostic factor** among women with breast cancer, *social isolation* is associated with higher risks of recurrences (RR=1.43, 95% CI 1.15-1.77)

3. **Prognostic model** : Nottingham Prognostic Index

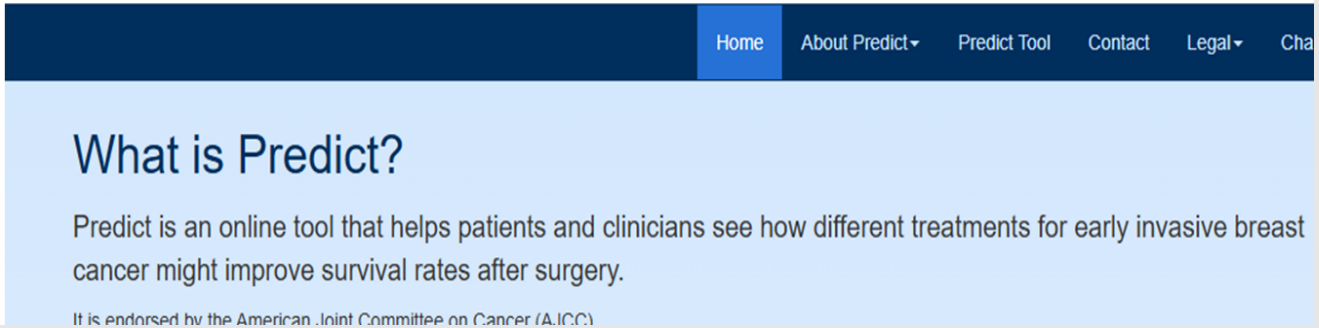
4. **Prediction model**



https://github.com/WintonCentre/predict-v21-main

predict breast cancer

https://breast.predict.cam/



General aim: combine **multiple** patient characteristics to predict the ***probability*** of a health outcome

Diagnostic / Prognostic models:

- Increasingly **recommended in Clinical Guidelines**

E.g. **QRISK** (CV diseases), **FRAX** (risk of developing osteoporotic & hip fracture), **SAPS** and **APACHE** (ICU scoring systems)….

- Typically **developed using standard regression** approaches (logistic, Cox…)

- Widely **available, easy-to-use** (to both the public and healthcare professionals) on websites, and smartphone apps

For reporting guidance, or risk of bias assessments and checklists for diagnostic and prognostic model studies: TRIPOD and PROBAST **(+ TRIPOD-AI)**

https://www.tripod-statement.org/

https://www.probast.org/

Prognostic/predictive models combine several ***characteristics/features*** (related to the patient, the disease, or treatment) to predict outcome.

Typically, a ***limited*** number of predictors are considered (relative to the # of subjects).

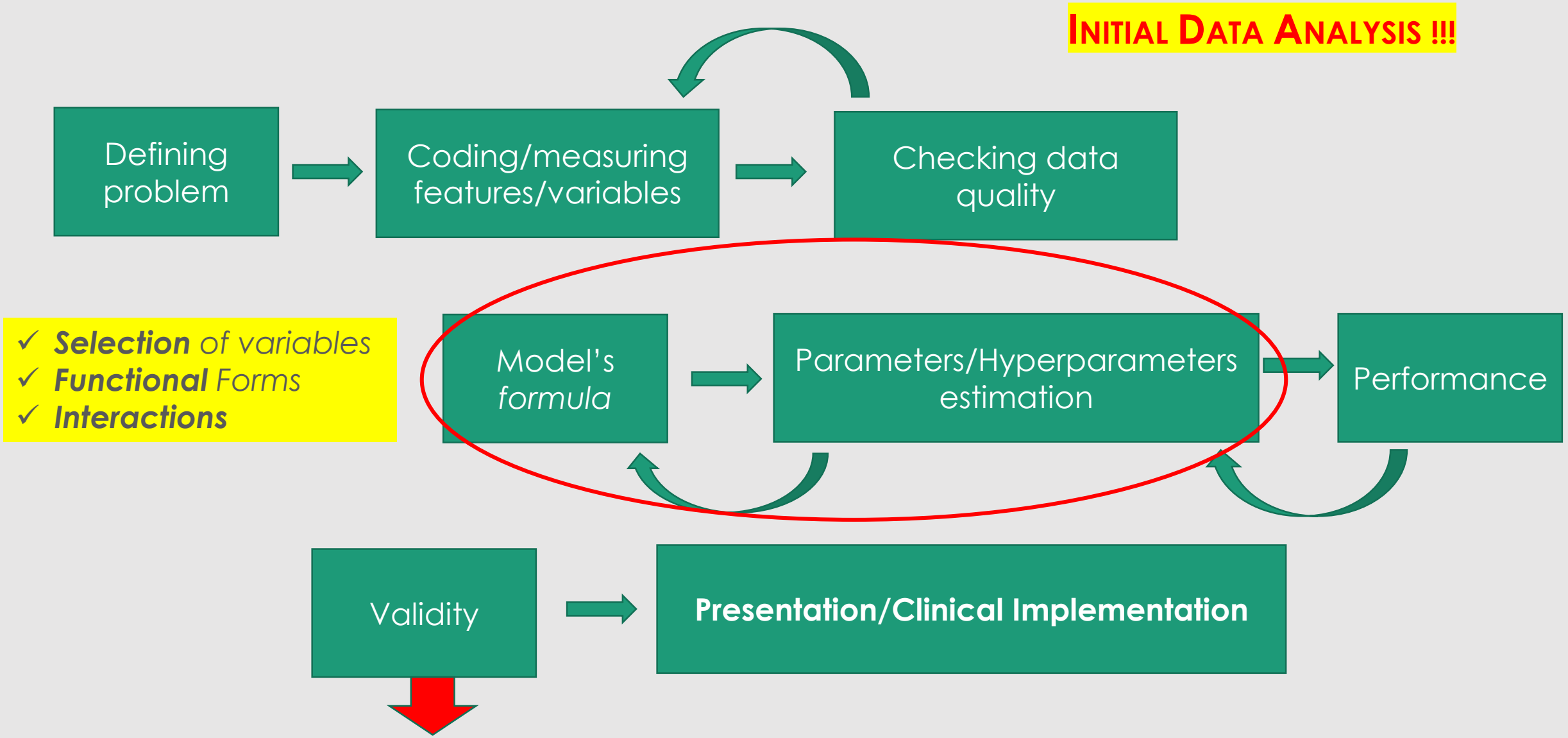
Our **focus** here is on the models which are **the most widely used** in the clinical field. We will consider situations where the ***initial number*** of candidate predictors is ***limited***.

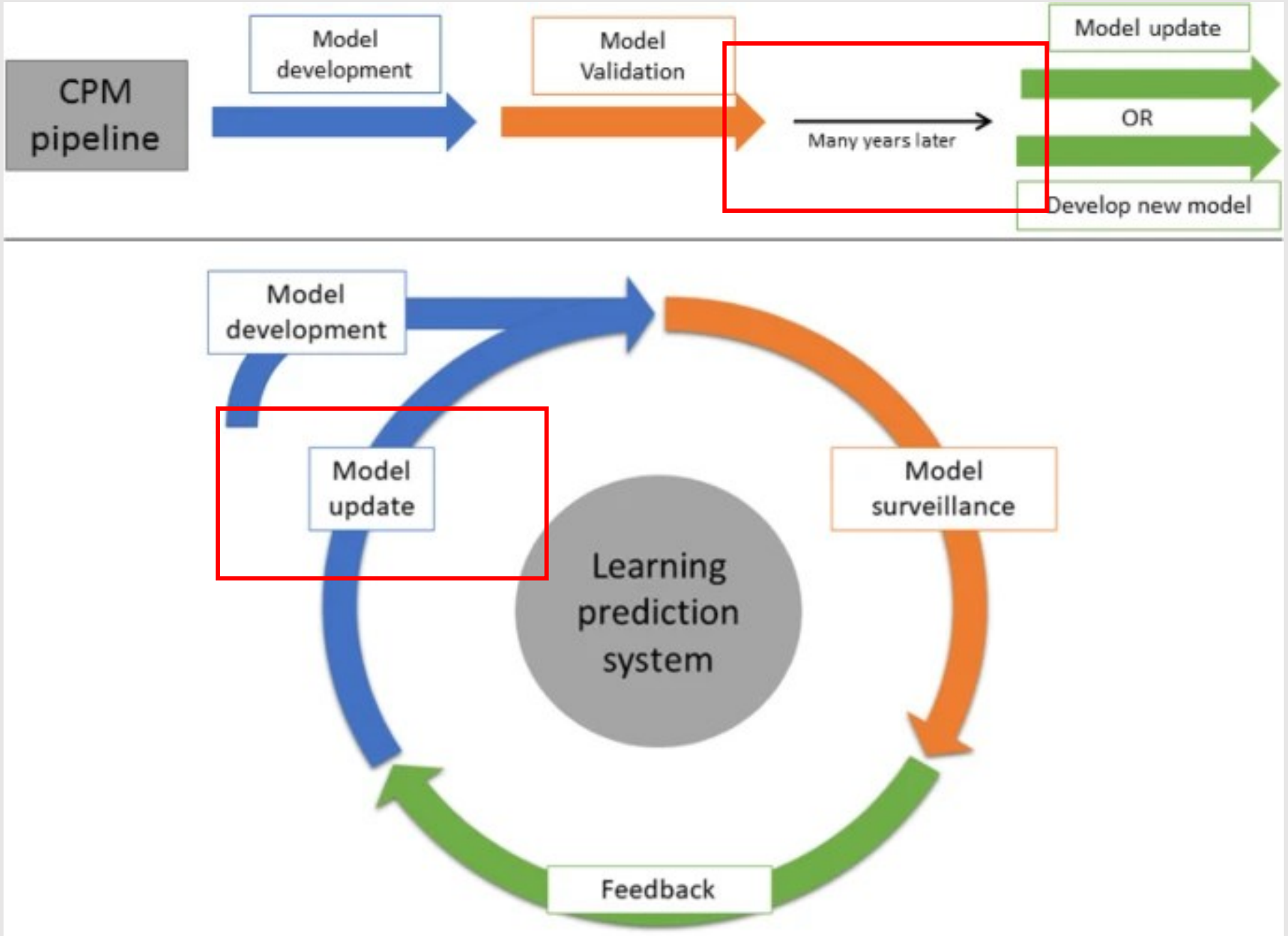
This is in contrast to areas such as bioinformatics, genomics, proteomics, or metabolomics… there is more complex data, larger numbers of candidate predictors (often >10,000, or even >1 M).

!! Topic not covered !!

We assume that **subject knowledge** is available, from previous studies and experts (e.g., medical doctors)

Basic steps:


INITIAL DATA ANALYSIS !!!
Defining problem
Coding/measuring features/variables
Checking data quality
✓ Selection of variables
✓ Functional Forms
✓ Interactions
Model's formula
Parameters/Hyperparameters estimation
Performance
Validity
Presentation/Clinical Implementation
On external dataset !!!


CPM pipeline
Model development
Model Validation
Many years later
Model update
OR
Develop new model
Model development
Model update
Learning prediction system
Model surveillance
Feedback

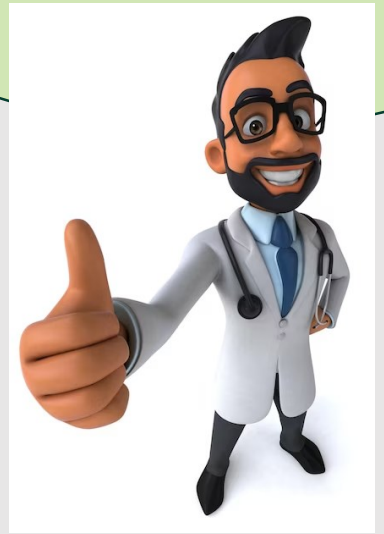
Defining the framework: initial checklist

- **Target population** who would be eligible to use the model and whatever inclusion/exclusion criteria
- **Time origin** baseline *time zero* ***(if time is involved)***
- **Target of prediction** event of interest (scale of measure?)
- ***Competing risks*** *: events after which the event of interest cannot occur or is not of interest any longer* ***[survival setting]***
- **Prediction time horizon** how far in time from the baseline the prediction is projected
- **Predictor/Prognostic variables** list of the predictors/features [*measured at baseline] (how they were measured)*

How we **select** variables in the model ?

Depend on the scientific aim !

Theory-driven



Data-driven

Predictive modelling

3 basic steps in the [*classical*] model building process:

1. **Missing data**

2. Variables **selection** ***(functional forms/interactions)***

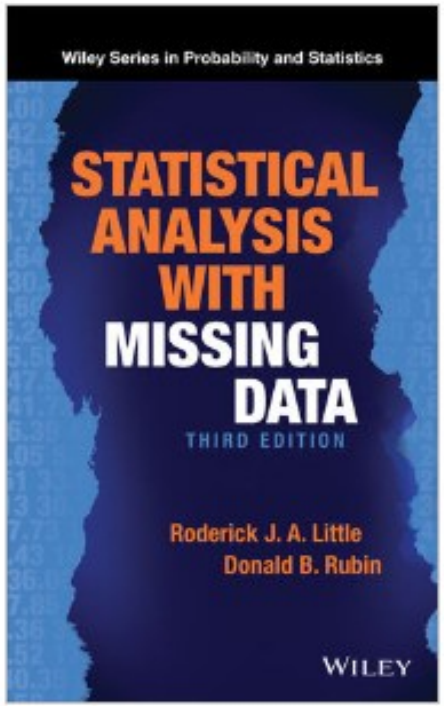
3. **Evaluating performance**

missing data… something more!

MCAR
Missing completely at random

the fact that data are missing is ***independent*** of the observed and *unobserved* data

no **systematic** differences between participants with missing data and those with **complete** data



MAR
Missing at random

the fact that the data are missing is **systematically** related to the observed but not the *unobserved* data

Complete case analyses may or may not result in bias. Proper **accounting** for the known factors can produce unbiased results in analysis

MNAR
Missing not at random

the fact that the data are missing is **systematically** related to the unobserved data…

if the complete case analysis is biased this issue **cannot be** addressed…

Why Not Just Use "**Complete Case Analysis**"?

The Problem… Deleting rows with missing values (*listwise deletion*) leads to

1. Reduced Power: You lose data.

2. **Bias**: If the "missingness" isn't random (e.g., sicker patients don't show up for follow-ups), your results will be biased.

The goal of **MI** (Multiple Imputation): To preserve the sample size and account for the *uncertainty* caused by missing data, rather than just "making up" a single number…

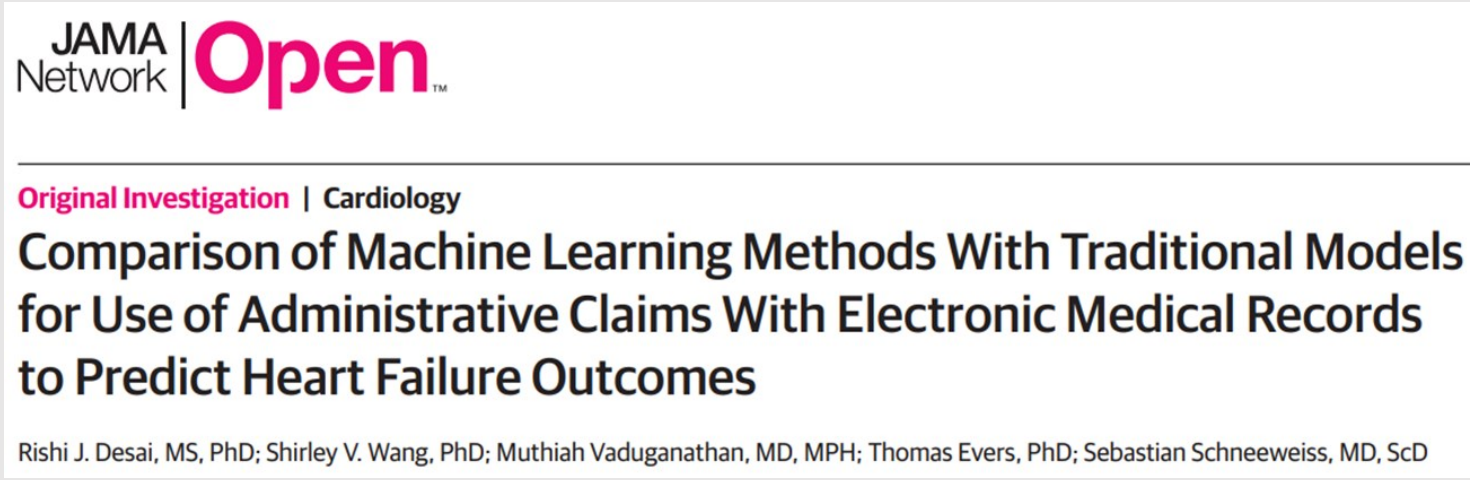
Missing data imputation

For laboratory results (serum sodium, serum potassium, serum urea nitrogen, and serum creatinine levels), missing values were observed in the range of 5% to 25%. Therefore, we used the ***multiple imputation procedure***

Multiple imputation in a nutshell

- predict ***M*** different values for each missing value, leading to ***M*** imputed datasets
- perform the statistical analysis *on each imputed datasets* to estimate the parameter of interest $\boldsymbol{\theta}$ and then combine the results to provide a unique estimation for $\boldsymbol{\theta}$ and for its associated variability **(Rubin's Rules)**

*In Desai et al. there is no information on the specific imputation algorithm that was used.


JAMA Network | Open

Original Investigation | Cardiology

Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes

Rishi J. Desai, MS, PhD; Shirley V. Wang, PhD; Muthiah Vaduganathan, MD, MPH; Thomas Evers, PhD; Sebastian Schneeweiss, MD, ScD

Understanding Rubin's Rules...

To get the final result, we don't just *average* the results. We must account for two types of variance:

Within-imputation variance: The standard error of the estimate *within* one dataset.

Between-imputation variance: How much the estimates vary across the ***M*** datasets.

The Result: A wider (and more honest) *confidence interval* that reflects our uncertainty about the missing values imputation process.

Block 3.2

The Pooled Estimate ($\bar{Q}$): the arithmetic mean of the estimates (e.g., the regression coefficients $\hat{\beta}$) from each imputed dataset:

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m} \hat{Q}_i$$

m is the number of imputations.

Within-Imputation Variance (W): the average of the squared standard errors from each model.

$$W = \frac{1}{m}\sum_{i=1}^{m} S\,E_i^2$$

Between-Imputation Variance (B): how much the individual estimates $\hat{Q}_i$ vary around the pooled average $\bar{Q}$.

$$B = \frac{1}{m-1}\sum_{i=1}^{m} \left(\hat{Q}_i - \bar{Q}\right)^2$$

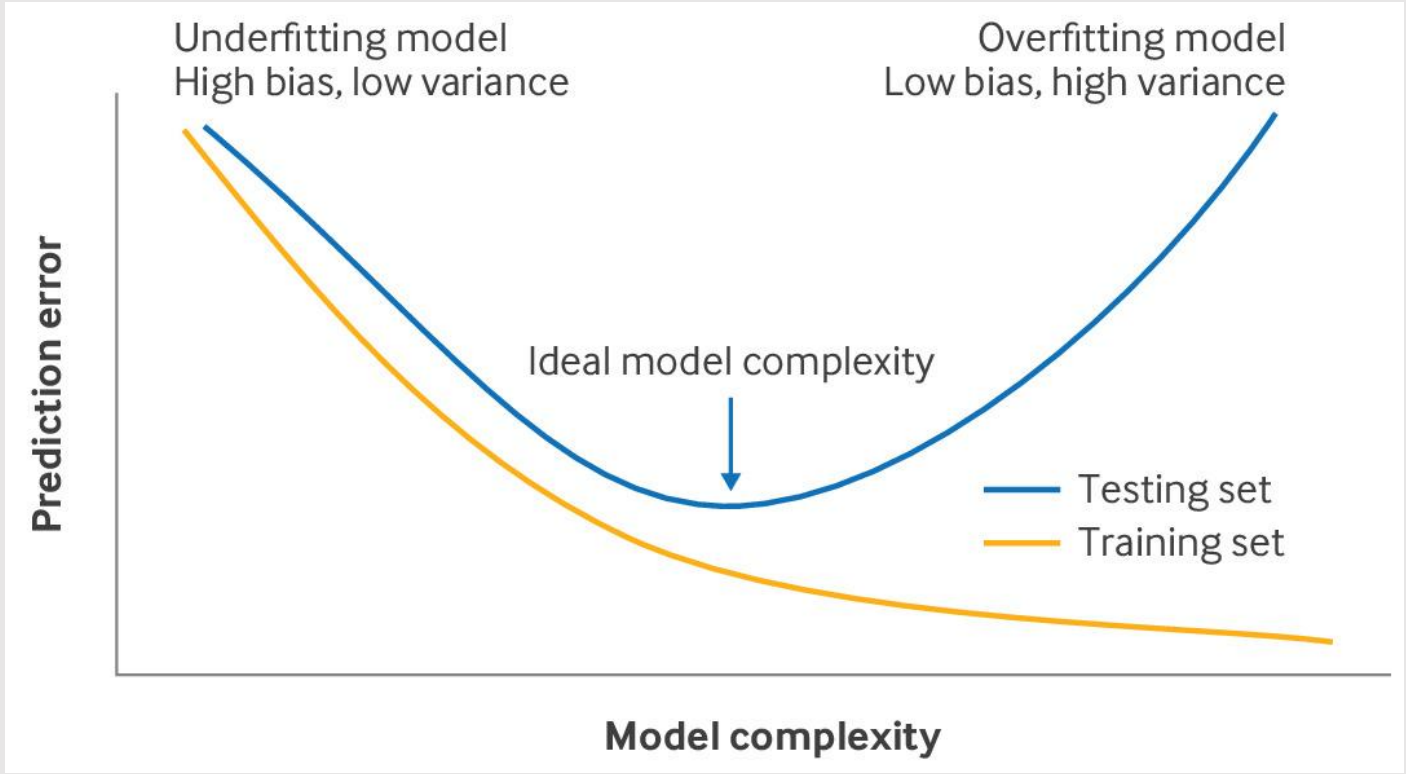
Total Variance (T): the value used to calculate your final p-values and 95% confidence intervals:

$$T = W + B + \frac{B}{m}$$

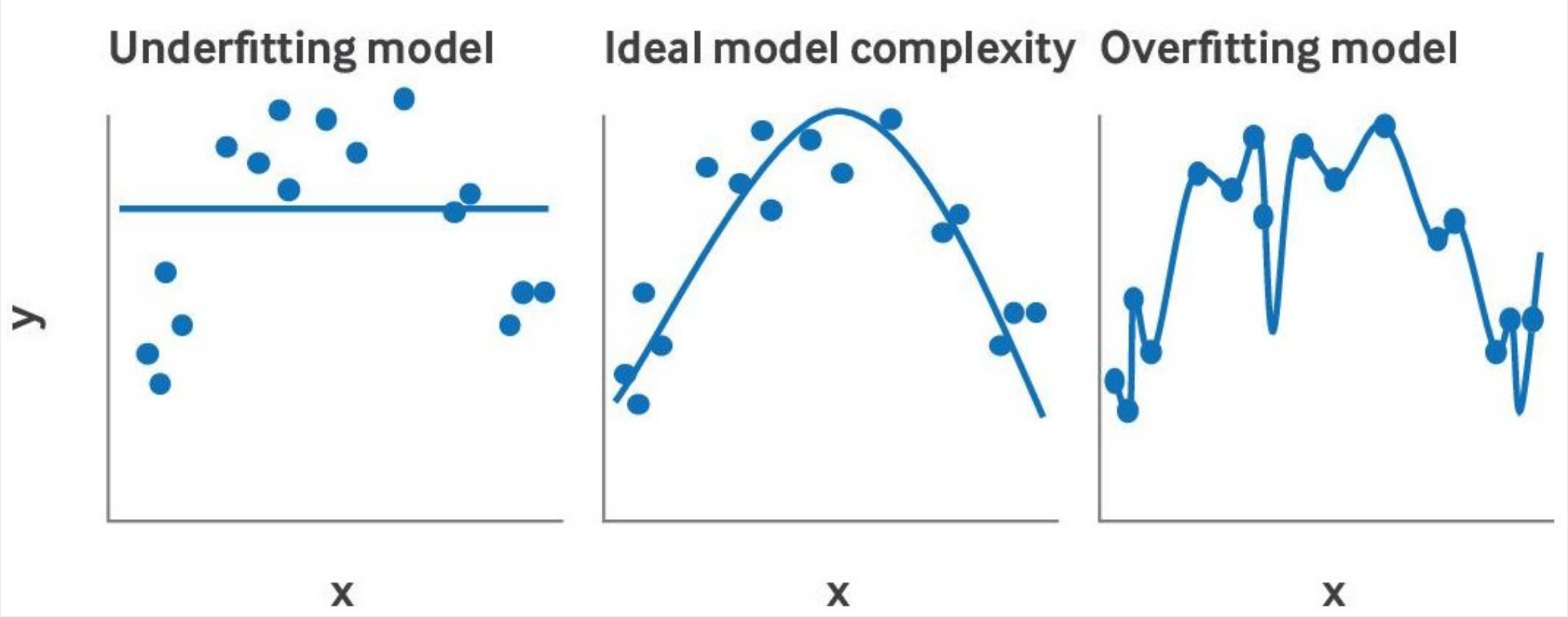
Note: The $\frac{B}{m}$ term is a "penalty" for using a finite number of imputations. As m (the number of datasets) increases toward infinity, this penalty disappears.

2. Variables/Features **selection**

Best subset selection	Estimate all 2^k possible models. Choose the best model according to an information criterion, for example AIC, BIC.	No subset of variables attains a better information criterion.
LASSO	Imposes a penalty on the sum of squares or log likelihood that is equal to the absolute sum of regression coefficients.	Relative weight of penalty is optimized by cross-validated sum of squares or deviance.



The well-known *bias-variance* trade-off



AIC/BIC based rules

The focus of **information criteria (IC)** is on selecting a model from a set of *plausible* models. Since including more variables in a model will slightly increase the *apparent* model fit (i.e. the model ***likelihood***), IC ***penalize*** the apparent model fit for model complexity (*more* variables, *k*=number of variables).

$$AIC = -2logL + 2k$$

penalty

goodness of fit

"smaller is better"

Log-likelihood is a measure of how likely one is to see their observed data, *given* a model.

The Bayesian IC (BIC) roughly speaking is more *parsimonius* (as *n* become large, AIC could select an *unnecessarily* complex model).

$$BIC = -2logL + \log(n) * 2k$$

n =sample size/number of events

Penalization (RIDGE/LASSO)

Penalized estimation is a technique used to address the challenges of overfitting and multicollinearity, especially when dealing with a **large** number of candidate predictor variables (and in absence of external knowledge).

The core idea is to add a **penalty term** to the loss function that the model tries to minimize during the training process.

This penalty term *discourages* overly complex models by imposing a cost on the magnitude or number of the model's coefficients.

$$X_i \in \mathbb{R}^d$$

$$Y_i = \beta X_i + \varepsilon_i$$

$$\hat{\beta}_{LSE} = argmin_\beta \frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta X_i)^2$$

$$\hat{\beta}_{RIDGE} = argmin_\beta \frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta X_i)^2 + \lambda\|\beta\|_2^2$$

$$\hat{\beta}_{LASSO} = argmin_\beta \frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta X_i)^2 + \lambda\|\beta\|_1$$

Just a further note:

We're seeing in the recent years **an overemphasis on prognostic algorithms** due to the ML/DL explosion.

In clinical research **prediction** is about getting an individualised ***probability/risk*** of the outcome of interest (e.g., not only what is my **general risk (=prognosis)** of developing CVD **over the next 10 years**, but **IF I DO** something ***(=> counterfactual prediction)***…)

Typically we are ***interested*** in prediction especially when:

- We **can act** on an the predicted risk : e.g., send a patient for further testing or monitoring of some specific risk factors

- We can **intervene to modify** that risk (e.g., stop smoking, giving a treatment…)

- It is useful **to communicate** this risk to the patient

Explainability of the prediction algorithm is **crucial**

Some (*old but gold*) examples of diagnostic/prognostic models in health research

https://www.bmj.com/content/386/bmj-2023-078276

Research Methods & Reporting

Developing clinical prediction models: a step-by-step guide

BMJ 2024 ; 386 doi: https://doi.org/10.1136/bmj-2023-078276 (Published 03 September 2024)
Cite this as: *BMJ* 2024;386:e078276

Article | Related content | Metrics | Responses | Peer review

Orestis Efthimiou [1 2], *senior lecturer*; *Michael Seo* [2], *doctoral student*; *Konstantina Chalkou* [3], *senior statistician*; *Thomas Debray* [4], *senior statistician*; *Matthias Egger* [2 5], *professor*; *Georgia Salanti* [2], *professor*

Diagnostic workup example

Diagnostic models may be useful to estimate the probability of an underlying disease, so that we can decide on further testing.

When a diagnosis is very *unlikely*, no further testing is indicated, while more tests may be indicated when the diagnosis *is not yet sufficiently certain* for decision-making on therapy.

Further testing usually involves one or more imperfect [**possibly invasive**] tests (sensitivity <100%, specificity<100%)

Many reference tests are not truly "gold standard", while they are used as definitive in determining whether a subject has the disease. The reference test may **not be suitable to apply in all subjects** suspected of the disease because it is burdensome (e.g., invasive) or costly.

	Disorder	No Disorder
Positive Test Result	True Positive (TP)	False Positive (FP)
Negative Test Result	False Negative (FN)	True Negative (TN)

Sensitivity = TP/(TP+FN)
Specificity = TN/(TN+FP)
PPV = TP/(TP+FP)
NPV = TN/(FN+TN)

Example of logistic regression as a diagnostic model (an old one)

Renal artery stenosis is a rare cause of hypertension.

The reference standard for **diagnosing** renal artery stenosis, renal angiography, is **invasive** and **costly**.

Aim: develop a *prediction rule* for renal artery stenosis from clinical characteristics.

The rule might then be used **to select patients** for renal angiography.

Logistic regression analysis performed with data from **477** hypertensive patients who underwent renal angiography. A simplified **prediction rule** was derived from the regression model for use in clinical practice.

Age, sex, atherosclerotic vascular disease, recent onset of hypertension, smoking history, body mass index, presence of an abdominal bruit, serum creatinine concentration, and serum cholesterol level were selected as predictors.

Diagnostic accuracy of the regression model was similar to that of renal scintigraphy. The conclusion was that this clinical prediction model **can help to pre-select patients** for renal angiography in an efficient manner by reducing the number of angiographic procedures without the risk of missing many renal artery stenosis.

Krijnen et al., A clinical prediction rule for renal artery stenosis. Annals of Internal Medicine(1998)

Block 3.2

Derivation of Scores in the Prediction Rule

The multivariable logistic regression model can be written as:

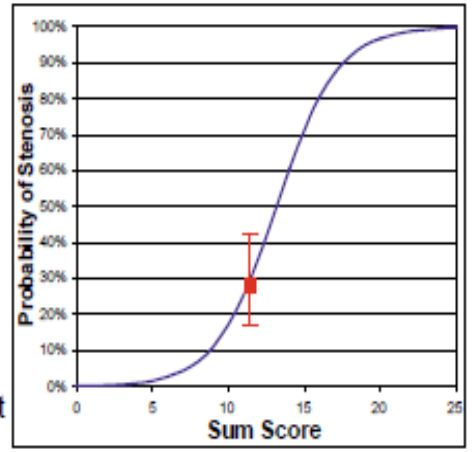
predicted probability of stenosis = $1/1 + e^{-(LP)}$,

where linear predictor $LP = -7.859 + 0.059 \times \text{age} + 0.033 \times (75 - \text{age}) \times \text{ever smoked} - 0.996 \times \text{sex} + 0.585 \times \text{atherosclerotic vascular disease} + 0.642 \times \text{recent onset} - 1.027 \times \text{obesity} + 1.693 \times \text{abdominal bruit} + 0.502 \times \text{hypercholesterolemia} + 0.032 \times \text{serum creatinine concentration}$.

Prediction rule for renal artery stenosis

Predictor		Value	Score
Smoking	former or current =1	1	-
Current age	years	45	4.4
Gender	male = 1	1	0
Atherosclerotic vascular disease*	yes = 1	0	0
Onset of hypertension within 2 years	yes = 1	1	1
Body mass index >= 25 kg/m2	yes = 1	0	2
Presence of abdominal bruit	yes = 1	0	0
Serum creatinine concentration	µmol/L	112	4.1
Serum cholesterol level > 6.5 mmol/L**	yes = 1	0	0
Sumscore			*11*

		Formula	Score chart
Predicted probability of renal artery stenosis		28%	*25%*
Confidence interval	17%	-	43%



See figure for graphical illustration

* femoral or carotid bruit, angina pectoris, claudication, myocardial infarction, CVA, or vascular surgery

** or cholesterol lowering therapy

45-year-old male with recent onset of hypertension.

According to a score chart, the sum score was 11, corresponding to a probability of stenosis of 25%. According to exact logistic regression calculations, the probability was 28% [95% confidence interval 17–43%].

Example of a prognostic model for Public Health [another quite classical one]

Various models have been developed to predict the future occurrence of disease in asymptomatic subjects in the population.

Well-known examples include the Framingham risk functions for cardiovascular disease

The Framingham risk functions (estimated by a regression model suitable for survival data) underpin several current policies for preventive interventions.

For example, statin therapy is only considered for those with relatively high risk of cardiovascular disease.

Appendix

Risk Estimation From Cox Model and From Score Sheet

The following examples illustrate the direct application of the Cox model and the use of the score sheet to estimate CVD risk in women and men.

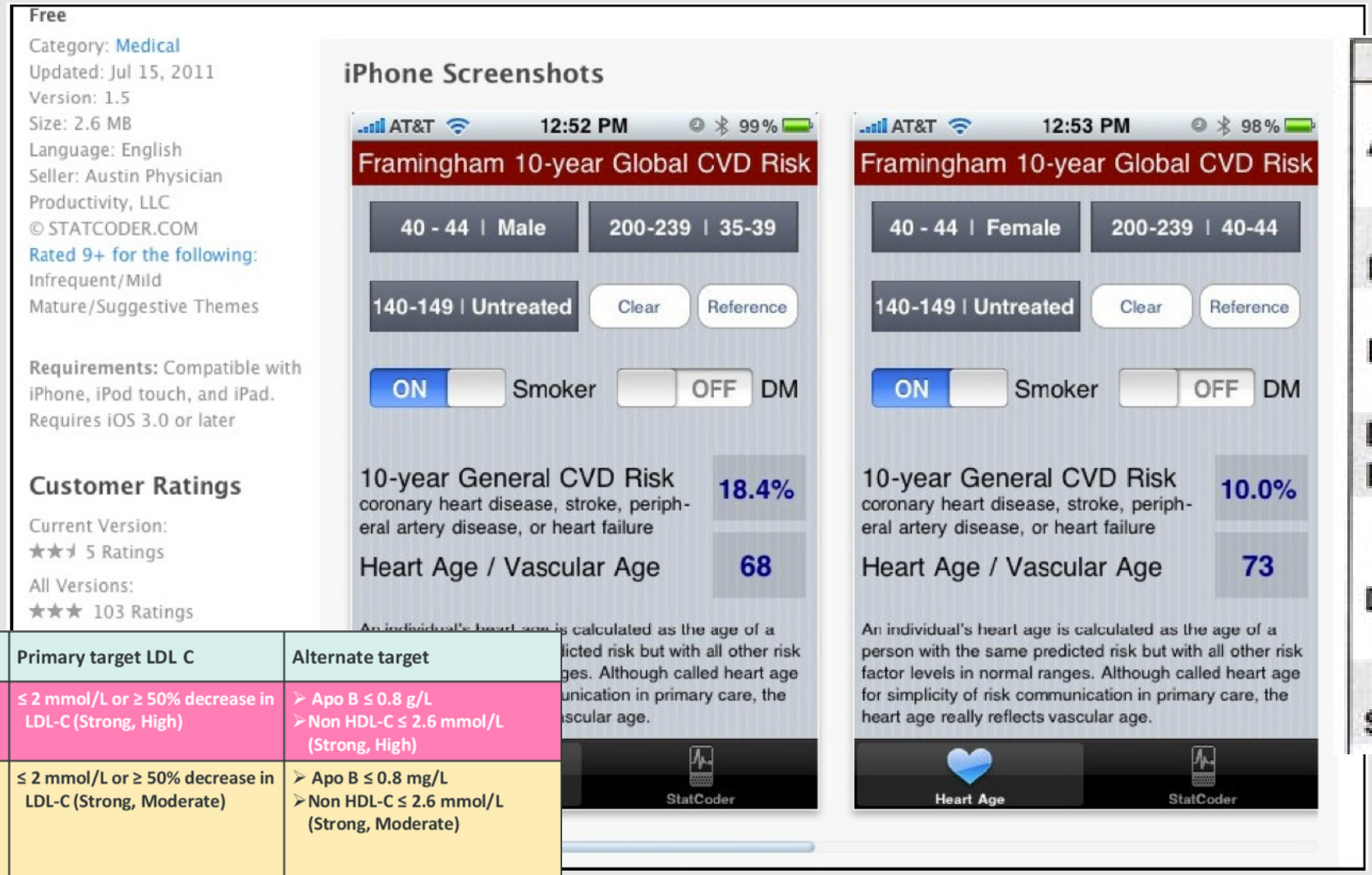
General formula: equation

$$\hat{p} = 1 - S_0(t)^{\exp(\Sigma_{i=1}^{p} \beta_i X_i - \Sigma_{i=1}^{p} \beta_i \bar{X}_i)},$$

https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.107.699579?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed#FD1

https://www.framinghamheartstudy.org/fhs-about/

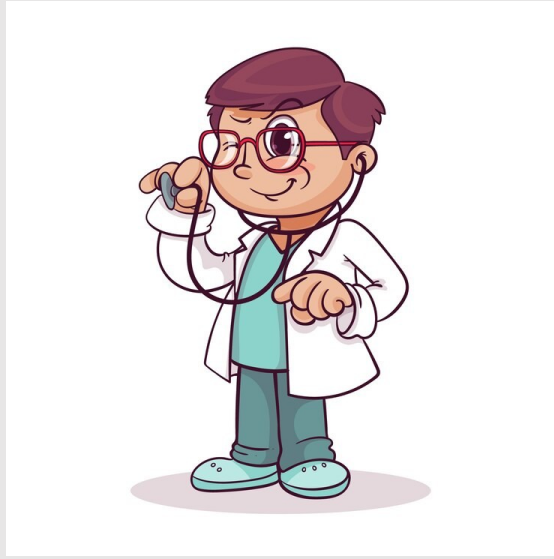
The Framingham Risk Score (**derived by a Cox model**) is used to estimate the 10-year cardiovascular risk of an individual:



Risk level	Initiate therapy 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, High)	➢ Apo B ≤ 0.8 g/L ➢ 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 consider if: Apo B ≥ 1.2 g/L or Non-HDL-C ≥ 4.3 mmol/L (Strong, Moderate)	≤ 2 mmol/L or ≥ 50% decrease in LDL-C (Strong, Moderate)	➢ Apo B ≤ 0.8 mg/L ➢ Non HDL-C ≤ 2.6 mmol/L (Strong, Moderate)
Low FRS < 10%	➢ LDL-C ≥ 5.0 mmol/L ➢ Familial hypercholesterolemia (Strong, Moderate)	≥ 50% reduction in LDL-C (Strong, Moderate)	

It's tough to make predictions, especially about the future*.

Team work !!



* A quote attributed to many people, from the Nobel prize-winning Quantum physicist Niels Bohr to legendary baseball player (and philosopher) Yogi Berra.