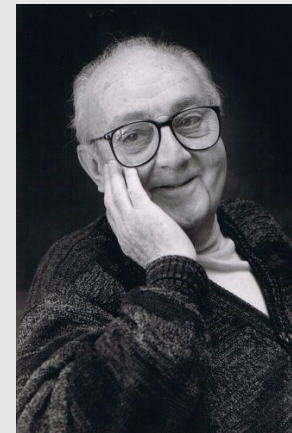
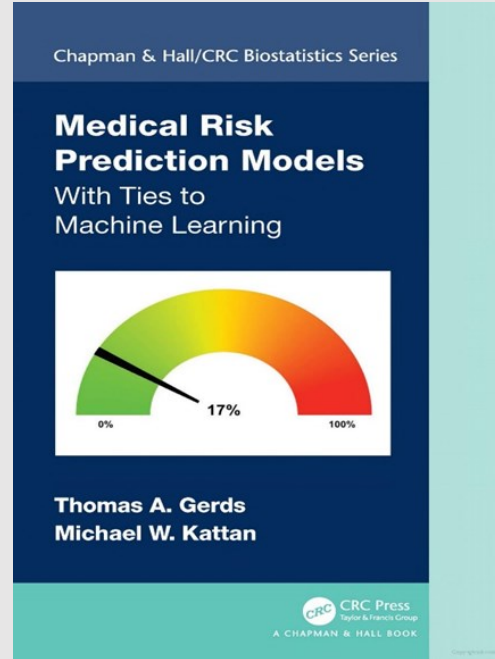
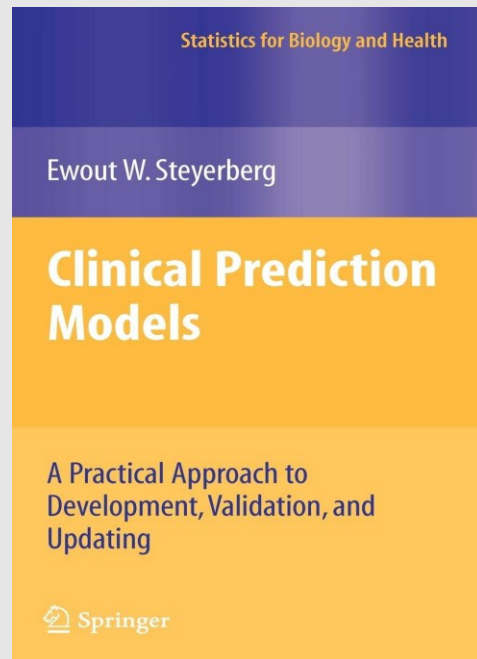
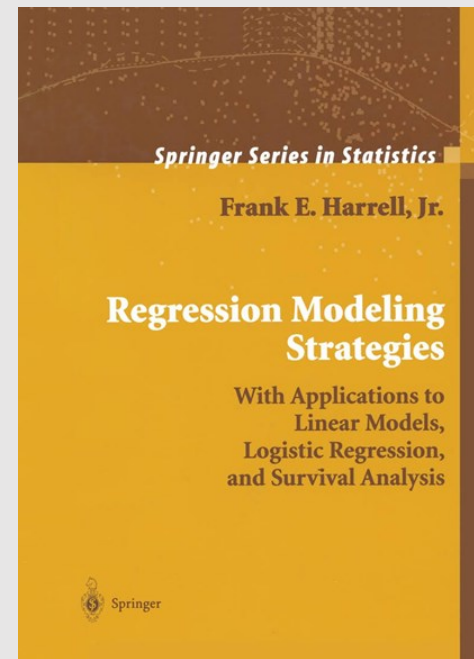


# Prediction Models in Epidemiological & Clinical Research



All models are wrong but some are useful.

**George E.P. Box**  
(1919 – 2013)

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

**G.E.P. Box**

# Summary

- **Prediction** regression models : Diagnostic/Prognostic
- **Steps** in building a prediction regression model
- The Basic Ones: Covariates Selection/Functional form/Interactions

“Models aren’t made to be unquestioned oracles. Instead of “follow the models” let’s “incorporate the models” in our decision making process.

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.

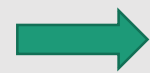
**One specific** treatment/exposure of interest.

**Observational** studies are increasingly used to enhance our knowledge of the real world. **Efficacy  $\neq$  Effectiveness**

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

“**Personalized**” predictions are central to many domains of medicine:

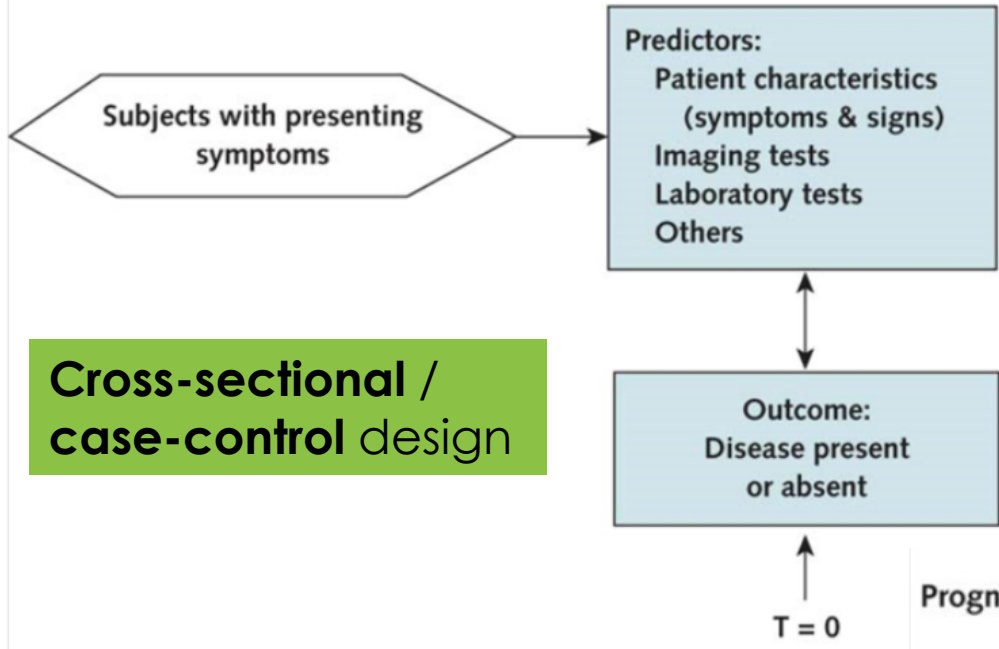
- **Screening:** find diseases early and treat better. Whether screening is useful depends on the **improvement in prognosis** compared to a *no screening* strategy. Selecting patients more at risk of developing a disease could be a useful *pre-screening* step.
- **Diagnosis:** Estimate the probability of a diagnosis without invasive tools, based on patient's characteristics.
- **Therapy:** New treatments appear nearly every day, but **their impact on prognosis** is often rather limited. Treatment effects could be *small* relative to the effects of determinants of the *natural history* of a disease. The “individual” benefits need to be considered and exceed any side effects and harms.



**Prediction** model that take into account possible treatment\*covariates *interactions*

# Diagnostic / Prognostic models

## Diagnostic multivariable modeling study



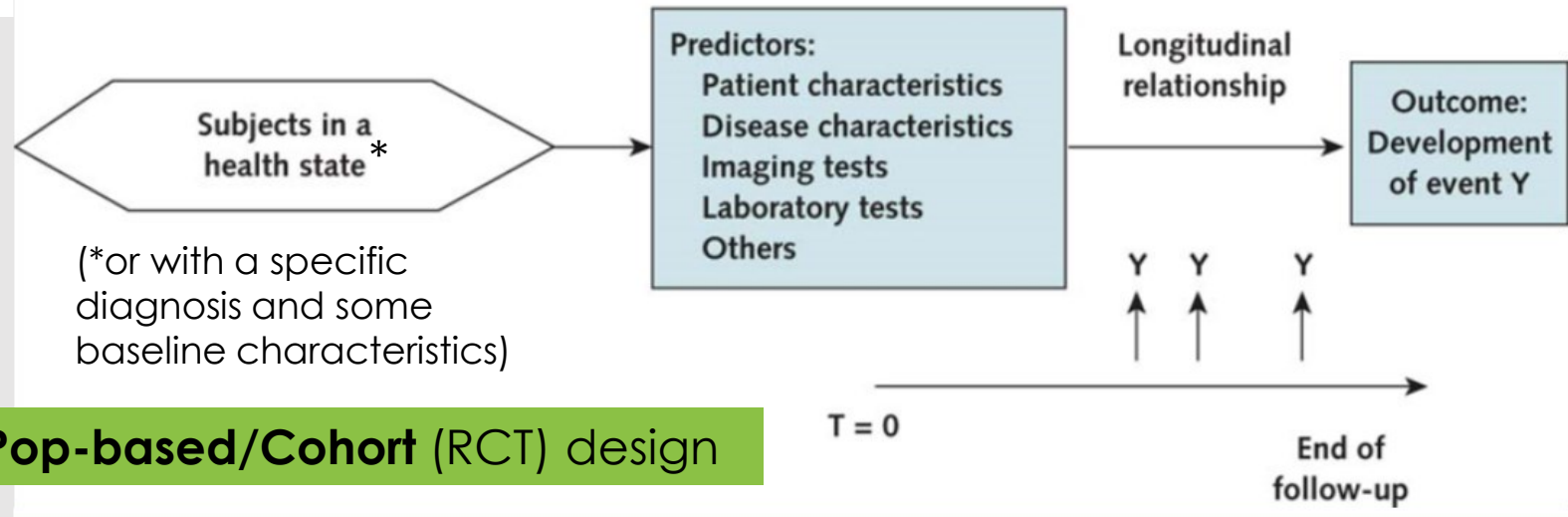
Cross-sectional / case-control design

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

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

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

## Prognostic multivariable modeling study



Pop-based/Cohort (RCT) design

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

# Prognosis/Prediction

- 1. Overall prognosis** Estimate the **average risk** of an outcome (e.g. death) 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** Identify factors whose values (levels) **are associated with** changes in the outcome's risk or expected value
- 3. Prognostic model** Predict an *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.

In this context “**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) **IF I do some therapy/change** in lifestyle vs not (in causal inference “*counterfactual prediction*” is also used).

<https://www.prognosisresearch.com/>

# Examples

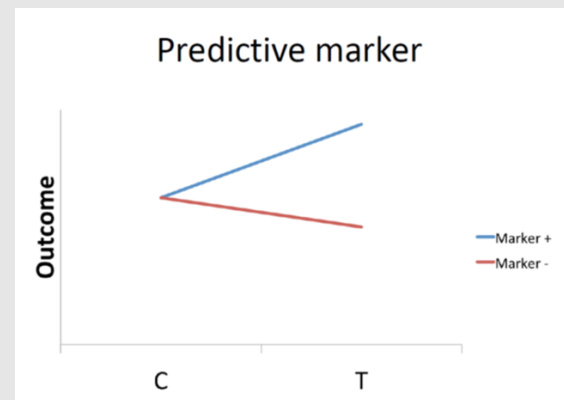
**1.Overall prognosis:** 5 out of 6 women diagnosed with breast cancer in the UK in 2019 will be alive in 2024

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

**3.Prognostic model:** “PREDICT” is an online tool that clinicians can use to estimate 5-years survival probability for a woman after breast cancer surgery

[https://breast.predict.nhs.uk/predict\\_v2.0.html](https://breast.predict.nhs.uk/predict_v2.0.html)

**4.Prediction model:** women with breast cancer estrogen receptor (ER) positive have reduced 10-yr recurrence and mortality **IF treated with** a drug (tamoxifen), whilst in women with ER-negative, this drug had little or no effect.





## Block 3.1

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 is under way !!)**

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

<https://www.probast.org/>



Clinical prediction models combine a certain number of **characteristics/features** (related to the patient, the disease, or treatment) to predict a diagnostic or prognostic/predictive outcome.

Typically, a **limited** number of predictors are considered.

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**, say below 20 - 30.

**This is in contrast to areas such as bioinformatics, genomics, proteomics, or metabolomics... more complex data and high-dimensional # candidate predictors (often >10,000, or even >1 M).**

**!! Data mining or reduction techniques not covered !!**

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

# Initial checklist

- **Target population:** who would be eligible to use the model and whatever inclusion/exclusion criteria
- **Time origin:** baseline *time zero* (if there is time involved!)
- **Target of prediction:** event/parameter of interest
- **Competing risks** events *after which* the event of interest cannot occur or is not of interest any longer **[survival setting, block 4]**
- **Prediction time horizon:** how far in time from the baseline the prediction is projected (if there is time involved!)
- **Predictor/Prognostic variables:** list of the predictors/features [*measured at baseline*] (*how they were measured / context !*)

# Type of Data and Choice of Model (*classical ones* !)

## Type of Data [outcomes]

- Continuous measurement
- Count data
- Binary data
- **Censored lifetimes**

## Possible Model

- **Linear** regression model (normal outcomes)
- **Poisson** regression
- **Logistic** regression
- **Proportional hazards** regression (Cox)

# What does a (classical) model look like?

Binary outcome, logistic regression model:  $p$  = Probability of CV hospitalization

$$\text{logit}(\hat{p}_i) = \alpha + \beta_1 \text{Age}_i + \beta_2 \text{Sex}_i + \cdots + \beta_n \text{Hypertension}_i$$



$LP_i$  Linear Predictor

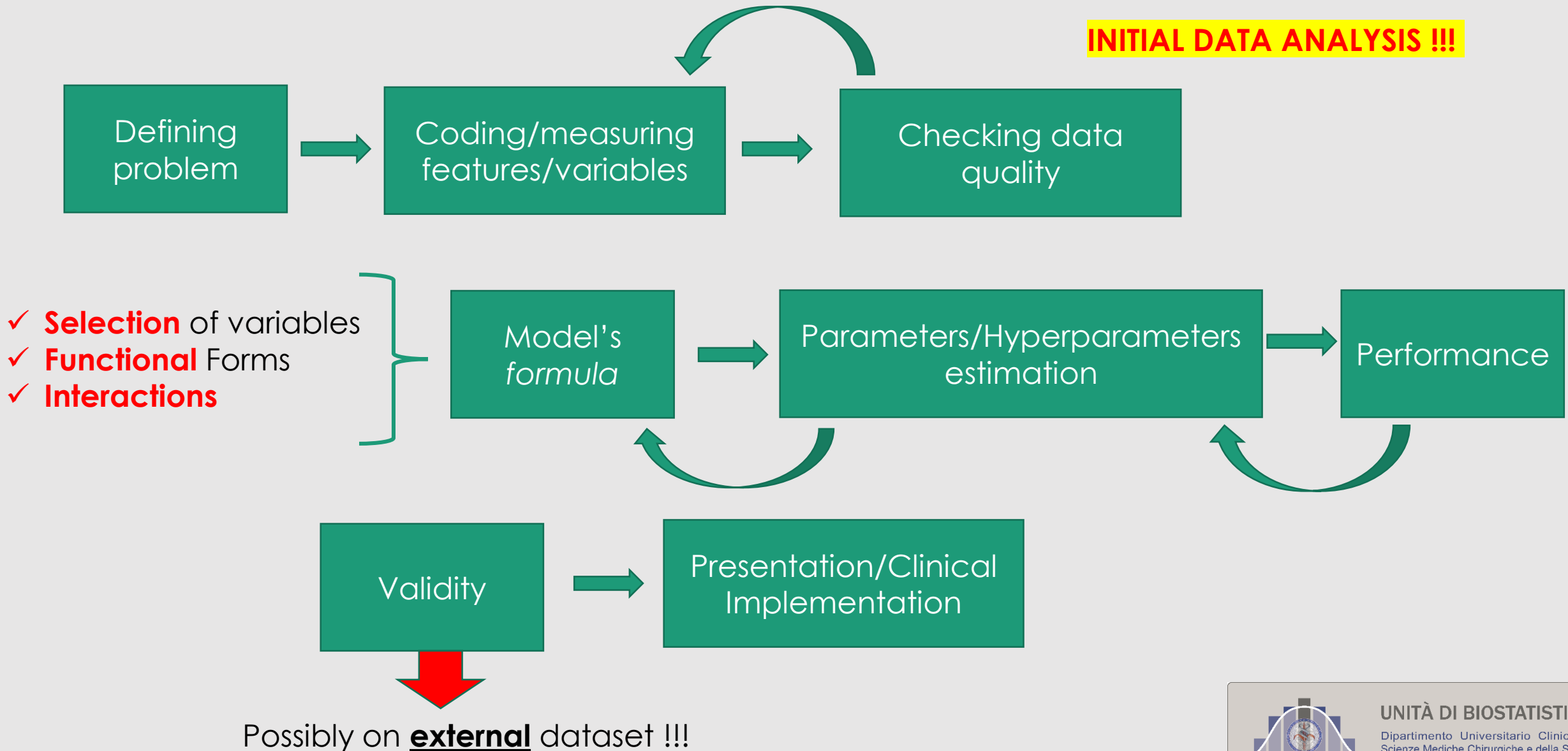
Continuous outcome, linear regression model:  $Y$  = Heart rate

$$E(Y_i) = \alpha + \beta_1 \text{Age}_i + \beta_2 \text{Sex}_i + \cdots + \beta_n \text{Hypertension}_i$$



$LP_i$  Linear Predictor

**Some** basic (also iterative!) **steps** should be considered in developing prediction models:



# 1. How to **select** variables in the model ?

**Theory-driven**



**Data-driven**

Something in-between...



# 1. Selection of variables

- Subject matter knowledge
- Chronology
- Costs of collecting measurements
- Availability at time of model use
- ...

• **DAG's like criteria [causal]  
(directed acyclic graphs)**

- Availability in data set (missing values)
- Variability (rare categories)



[**IDA** considerations]

Discussion with experts





# 1. Variable selection methods\*

Basic\*\* algorithms:

There is no Universal Solution !!!

- **Full model**
- *Univariable filtering*
- *Forward selection*
- *Backward elimination*
- **AIC/BIC based rules**

Stepwise-like methods

- **Directed acyclic graph (DAG) based selection (causal)**

\*in general (low-dimensional) modeling problems

## Full Model:

1. Do not perform **any variable selection** [except for highly correlated features]
2. Select for each variable a suitable *functional form*
3. Explore *biologically plausible* interactions

The initial list is usually *pre-selected* by expertise



If sample size permits...

## Univariable filtering:

Still by far the most often applied method in medical literature

1. Select a significance level (e.g.,  $\alpha=0.20$  or  $\alpha=0.10$ )
2. Estimate univariable models
3. Use all variables in multivariable model with univariable p-value  $< \alpha$



Univariable selection work only with *perfectly* uncorrelated variables....

## Stepwise methods

## Forward selection

Select a significance level  $\alpha_1$ .

- Estimate a **null** model
- For  $j=1, \dots, p$  consider **adding**  $x_j$  [find the most significant]
- Repeat:  
While the most significant excluded term has  $p < \alpha_1$  add it and re-estimate.

## Backward elimination

Select a significance level  $\alpha_2$

- Estimate the **full** model
- For  $j=1, \dots, p$  consider **dropping**  $x_j$  [find the least significant]
- Repeat:  
While least significant term has  $p \geq \alpha_2$  remove it and re-estimate.

## Variant: Stepwise forward



- Estimate a **null** model.
- Repeat:

While the most significant excluded term has  $p < \alpha_1$  add it and re-estimate.

If least significant included term has  $p \geq \alpha_2$  remove it and re-estimate.

Select  $\alpha_1$  and  $\alpha_2$

## Variant: Stepwise backward



- Estimate **the full** model.
- Repeat:

While least significant term has  $p \geq \alpha_2$  remove it and re-estimate.

If most significant excluded term has  $p < \alpha_1$  add it and re-estimate.



## AIC/BIC based rules

The focus of information criteria 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*), information criteria were developed to *penalize* the apparent model fit for model complexity (*more* variables,  $k$ =number of variables).

$$AIC = -2\log L + 2k$$

↓
↑

goodness of fit
 penalty

“smaller is better”

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

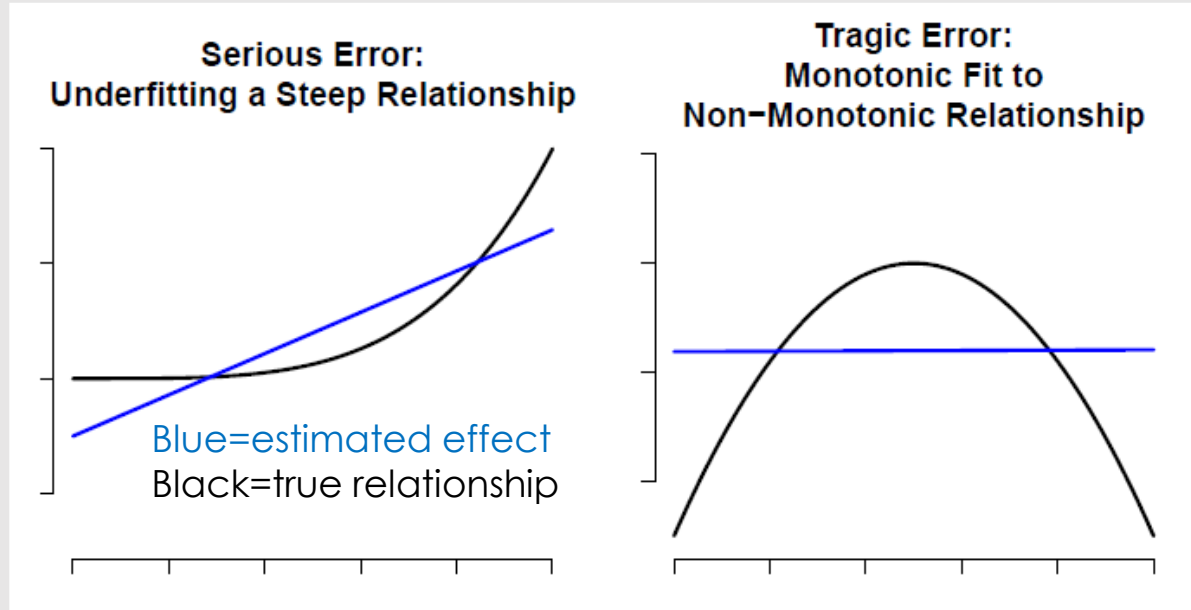
The Bayesian information criterion (BIC) roughly speaking is more *parsimonious* (as  $n$  become large, AIC could select an unnecessarily complex model).

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

$n$  =sample size/number of events

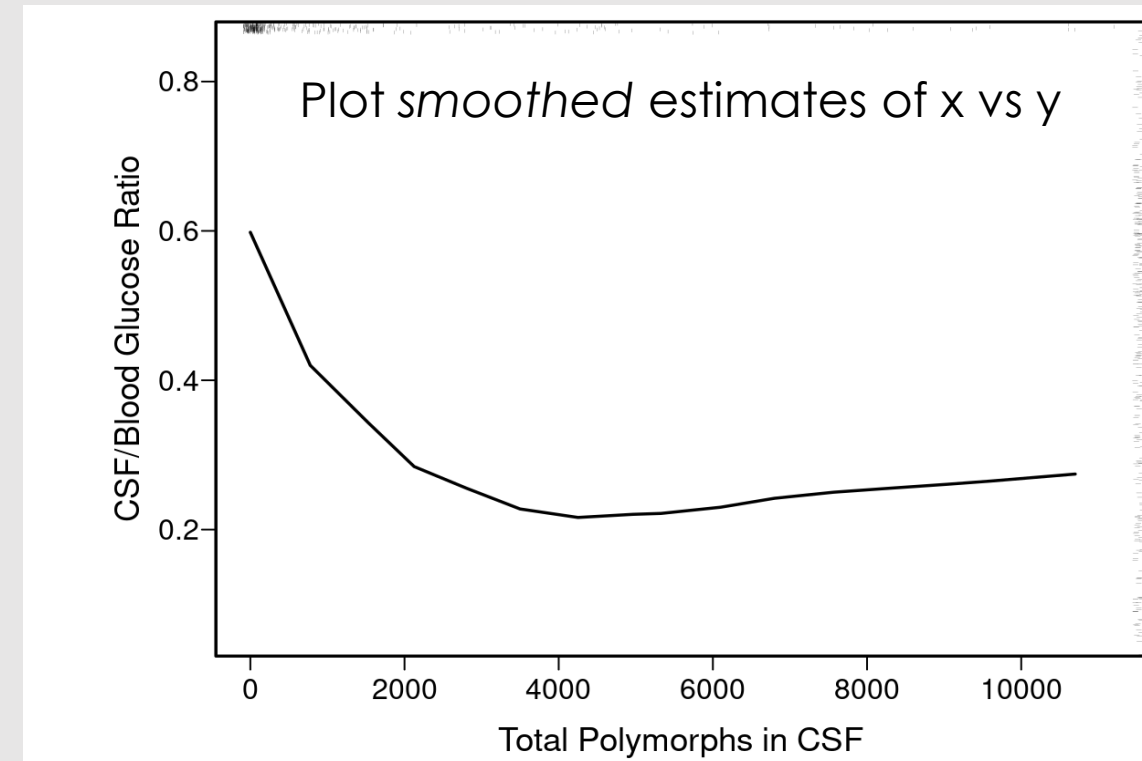
## 2. Functional forms

### Numerical variables:



Rarely expect linearity.

curve fitting could help

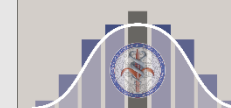


### Nominal variables:

- choose an *appropriate reference* (frequent, standard group, etc.)
- collapse *rare* groups if possible

### Ordinal variables:

- ordinal coding
- collapse *rare* adjacent groups if reasonable



## Block 3.1

All regression models **should** make assumptions about the **shape** of the relationship between predictor X and response variable Y.

Many analysts assume **linear** relationships *by default*.


**Splines** (piecewise polynomials) are natural nonlinear generalizations.

In epidemiology many practitioners analyze continuous data using **percentiling/classes**, but this is nearly always a bad idea.

REVIEW

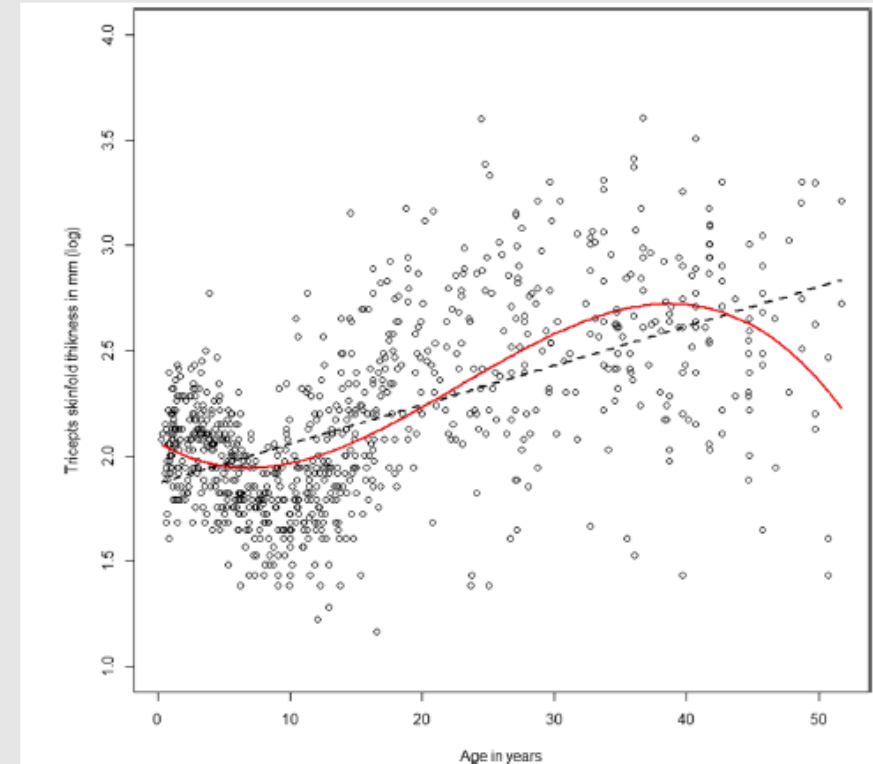
Open Access

### A review of spline function procedures in R

Aris Perperoglou<sup>1\*</sup> , Willi Sauerbrei<sup>2</sup>, Michal Abrahamowicz<sup>3</sup>, Matthias Schmid<sup>4</sup> on behalf of TG2 of the STRATOS initiative



**Splines** are useful to reproduce flexible shapes. **Knots** are placed at several places within the data range, to identify the points where adjacent functional pieces join each other. **Smooth** functional pieces (usually low-order polynomials) are chosen to fit the data between two consecutive knots. The **type** of polynomial and the **number and placement** of **knots** is what then defines the type of spline.

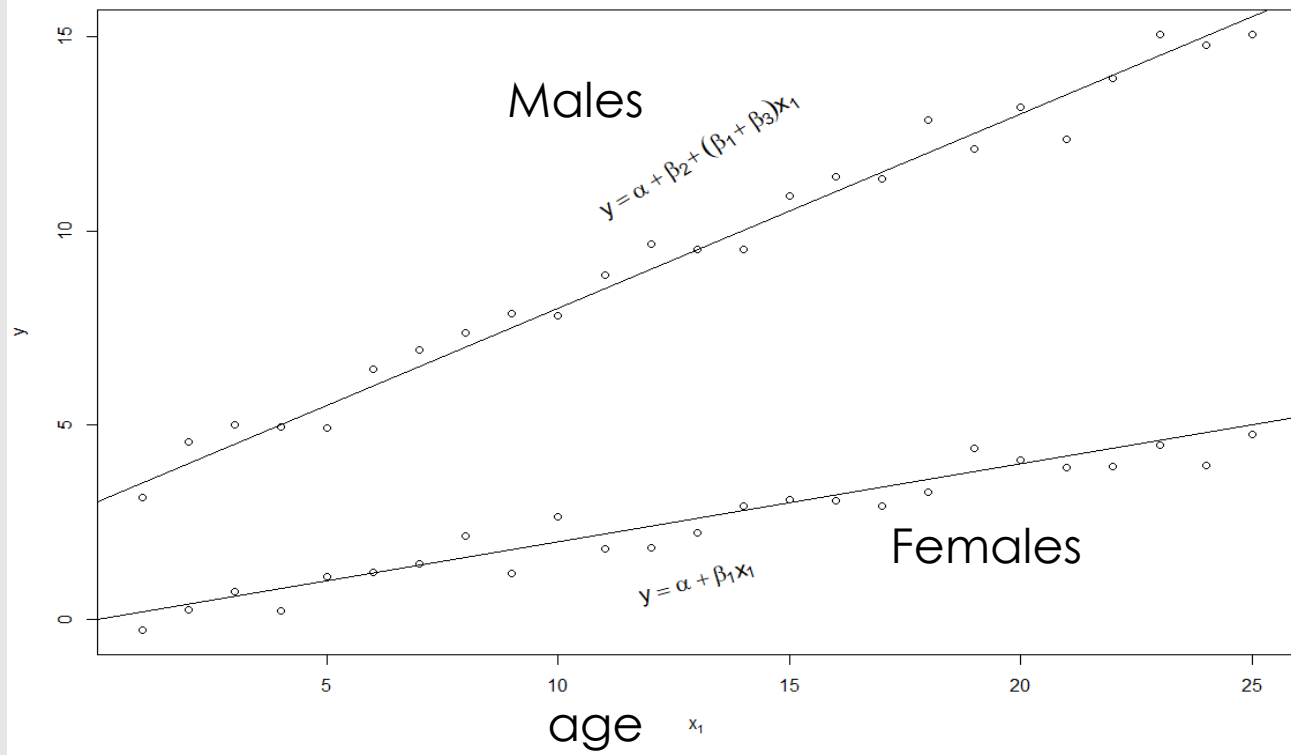


### 3. Interactions (basic example)

This is how one allows the slope of a predictor **to vary** by categories of another variable.

Example: separate slope for males and females:

$$E(y|x) = \alpha + \beta_1 * age + \beta_2 * [sex = m] + \beta_3 * age * [sex = m]$$



$$E(y|age, sex = m) = \alpha + \beta_1 * age + \beta_2 + \beta_3 * age \\ = (\alpha + \beta_2) + (\beta_1 + \beta_3) * age$$

$\alpha$  : mean  $y$  for 0-year-old female

$\beta_1$ : slope of age for females

$\beta_2$  : mean  $y$  for males - mean  $y$  for females, (0-year-olds)

$\beta_3$ : increment in slope in going from females to males

$$E(y|age, sex = f) = \alpha + \beta_1 * age$$

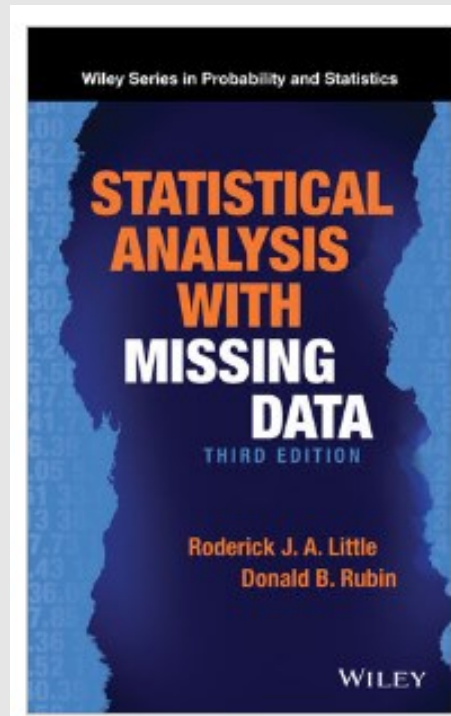


# Just a note about missing data...

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...

## The great power of regression models (if the assumption holds..)

We face primarily an **estimation** problem (prognosis/prediction):

- What is the *probability* that a male patient of 45 years with hypertension has a renal artery stenosis ?
- What is the *risk* of dying within 30 days after an acute myocardial infarction for a 67-years old female patient with diabetes ?
- What is the *expected 2-year survival probability* for a male patient of 54 years old with esophageal cancer ?

but *simultaneously* we can also “**test**” and somehow “**quantify**” associations [**estimation of causal effects is another topic!!!**]:

- Is the risk of renal artery stenosis *increasing* with values of a *specific* biomarker?
- Is age a *significant predictor* of 30-day mortality after an acute myocardial infarction?
- *How important* is nutritional status for survival with esophageal cancer?

More general:

- What are the *most relevant prognostic/predictors* in a certain disease ?
- Which is the *direction/intensity* of these associations ?

Statistical models may serve **simultaneously** to address **both** estimation **and** hypothesis testing.

Statistical models **summarize** patterns of the data available for analysis. In doing so, it is *inevitable* that **assumptions** have to be made (additivity?, linearity? *normal* distribution of the residuals ?....)

Some of these assumptions can be checked on data, for example, whether variable's effect work in an **additive** way or if continuous variables have reasonably **linear** effects.


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 of developing CVD **over the next 10 years**, but **IF I DO** something...??)

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 algorithm is crucial

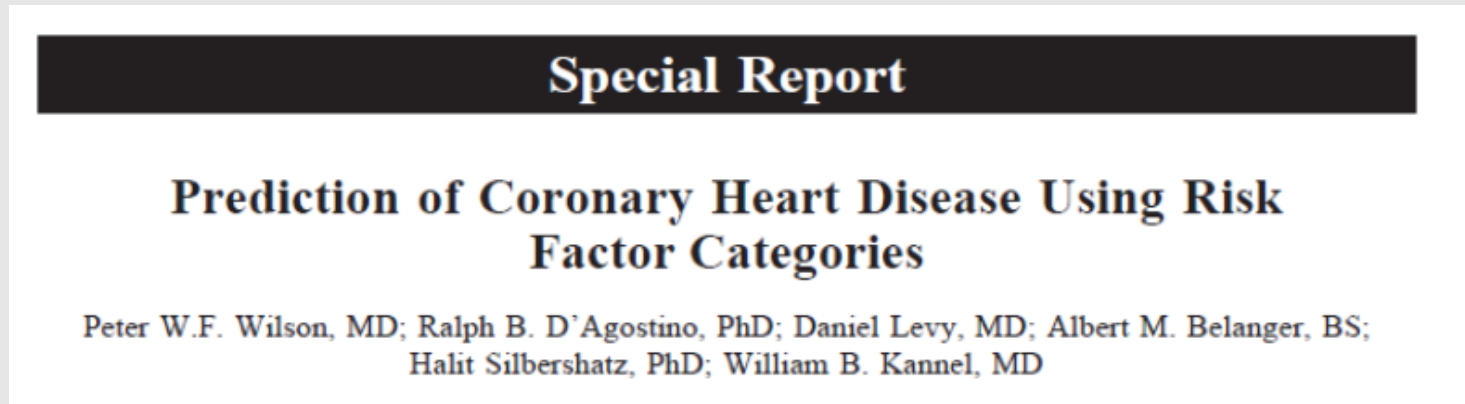
## Example of a prognostic model for Public Health [a classical one]

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.



**Special Report**

**Prediction of Coronary Heart Disease Using Risk Factor Categories**

Peter W.F. Wilson, MD; Ralph B. D'Agostino, PhD; Daniel Levy, MD; Albert M. Belanger, BS; Halit Silbershatz, PhD; William B. Kannel, MD

<https://framingham.com/heart/profile.htm>

## Block 3.1

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

**Free**  
 Category: Medical  
 Updated: Jul 15, 2011  
 Version: 1.5  
 Size: 2.6 MB  
 Language: English  
 Seller: Austin Physician Productivity, LLC  
 © STATCODER.COM  
 Rated 9+ for the following:  
 Infrequent/Mild Mature/Suggestive Themes

**Requirements:** Compatible with iPhone, iPod touch, and iPad. Requires iOS 3.0 or later

**Customer Ratings**  
 Current Version: ★★½ 5 Ratings  
 All Versions: ★★★ 103 Ratings

**iPhone Screenshots**

Heart Age / Vascular Age is calculated as the age of a person with the same predicted risk but with all other risk factor levels in normal ranges. Although called heart age for simplicity of risk communication in primary care, the heart age really reflects vascular age.

Adding up	
Age	
LDL-C or Chol	
HDL - C	
Blood Pressure	
Diabetes	
Smoker	

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)	<ul style="list-style-type: none"> <li>&gt; Apo B ≤ 0.8 g/L</li> <li>&gt; Non HDL-C ≤ 2.6 mmol/L (Strong, High)</li> </ul>
Intermediate FRS 10%-19%	<ul style="list-style-type: none"> <li>&gt; LDL-C ≥ 3.5 mmol/L (Strong, Moderate)</li> <li>&gt; For LDL-C &lt; 3.5 consider if: Apo B ≥ 1.2 g/L or Non-HDL-C ≥ 4.3 mmol/L (Strong, Moderate)</li> </ul>	≤ 2 mmol/L or ≥ 50% decrease in LDL-C (Strong, Moderate)	<ul style="list-style-type: none"> <li>&gt; Apo B ≤ 0.8 mg/L</li> <li>&gt; Non HDL-C ≤ 2.6 mmol/L (Strong, Moderate)</li> </ul>
Low FRS < 10%	<ul style="list-style-type: none"> <li>&gt; LDL-C ≥ 5.0 mmol/L</li> <li>&gt; Familial hypercholesterolemia (Strong, Moderate)</li> </ul>	≥ 50% reduction in LDL-C (Strong, Moderate)	



## 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)

$$\text{Sensitivity} = \text{TP}/(\text{TP}+\text{FN})$$

$$\text{Specificity} = \text{TN}/(\text{TN}+\text{FP})$$

$$\text{PPV} = \text{TP}/(\text{TP}+\text{FP})$$

$$\text{NPV} = \text{TN}/(\text{FN}+\text{TN})$$



## Block 3.1

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.

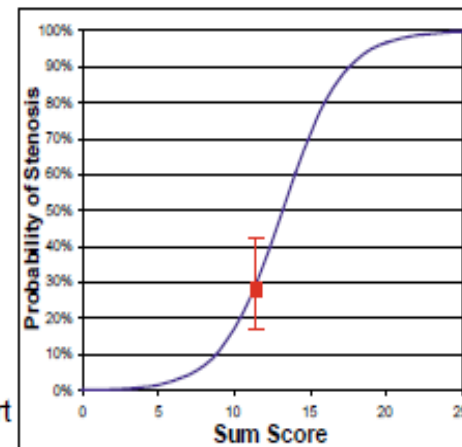
Age, sex, atherosclerotic vascular disease, recent onset of hypertension, smoking history, body mass index, presence of an abdominal bruit, serum creatinine concentration, and serum cholesterol level were selected as predictors.

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 3.1

	A	B	C	D	E	F	G	H
1	<b>Prediction rule for renal artery stenosis</b>							
2								
3	<b>Predictor</b>			<b>Value</b>	<b>Score</b>			
4	Smoking	former or current =1		1	-			
5	Current age	years		45	4.4			
6	Gender	male = 1		1	0			
7	Atherosclerotic vascular disease*	yes = 1		0	0			
8	Onset of hypertension within 2 years	yes = 1		1	1			
9	Body mass index >= 25 kg/m2	yes = 1		0	2			
10	Presence of abdominal bruit	yes = 1		0	0			
11	Serum creatinine concentration	µmol/L		112	4.1			
12	Serum cholesterol level > 6.5 mmol/L**	yes = 1		0	0			
17	<i>Sumscore</i>				11			
18				<b>Formula</b>	<b>Score chart</b>			
19	<i>Predicted probability of renal artery stenosis</i>			28%	25%			
20	<i>Confidence interval</i>			17%	-	43%	See figure for graphical illustration	
21	* femoral or carotid bruit, angina pectoris, claudication, myocardial infarction, CVA, or vascular surgery							
22	** 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%].