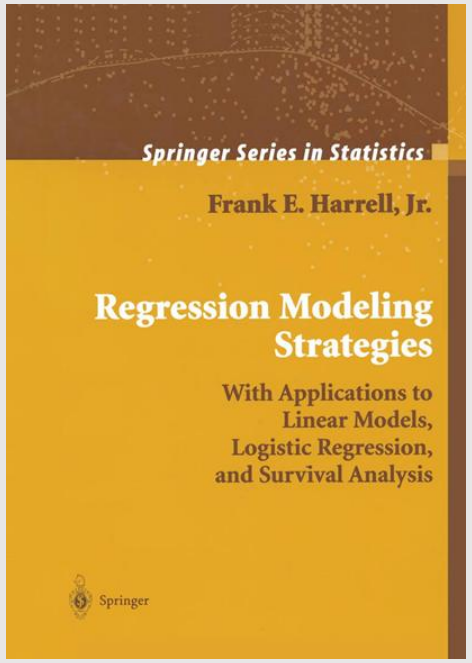
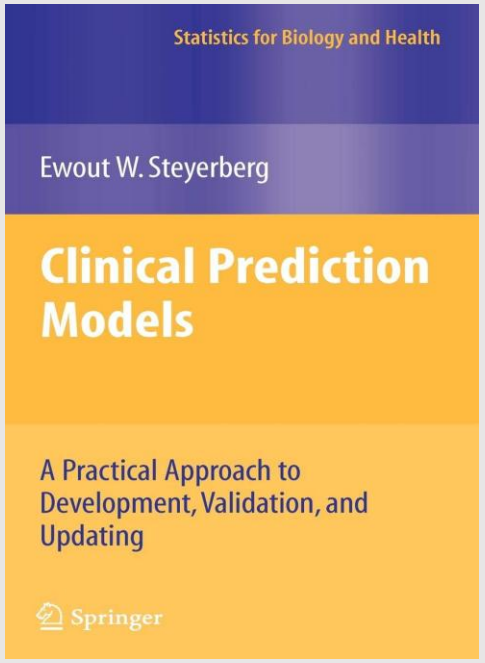
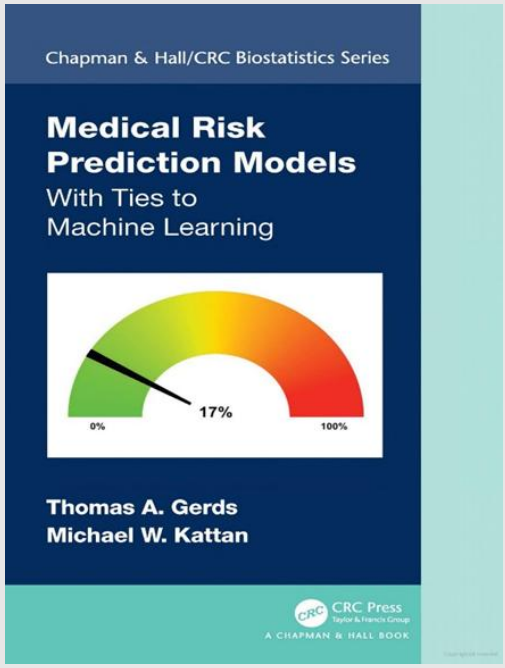
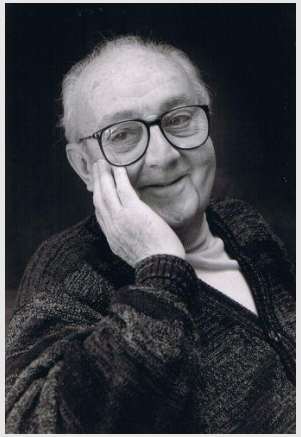
Prediction Models (II): performance & sample size





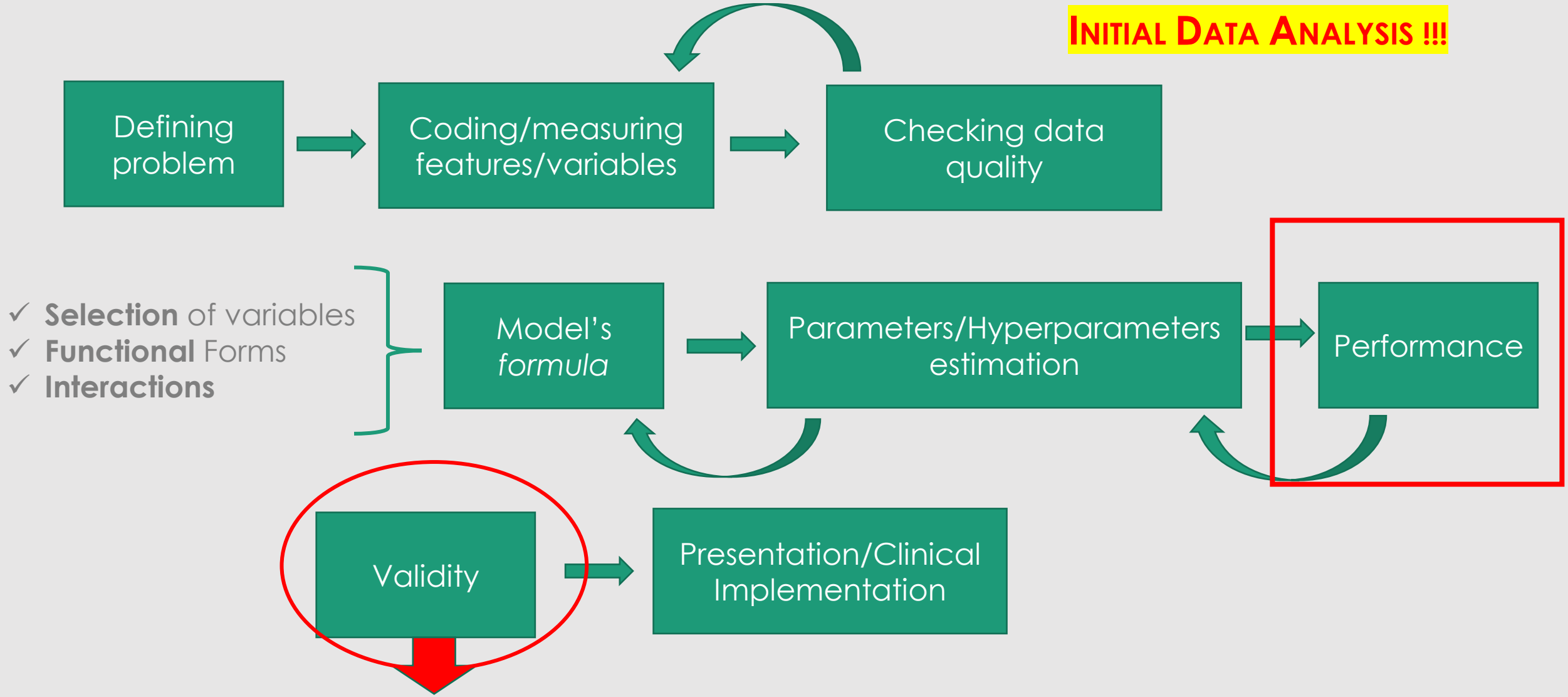




All models are wrong but some are useful.

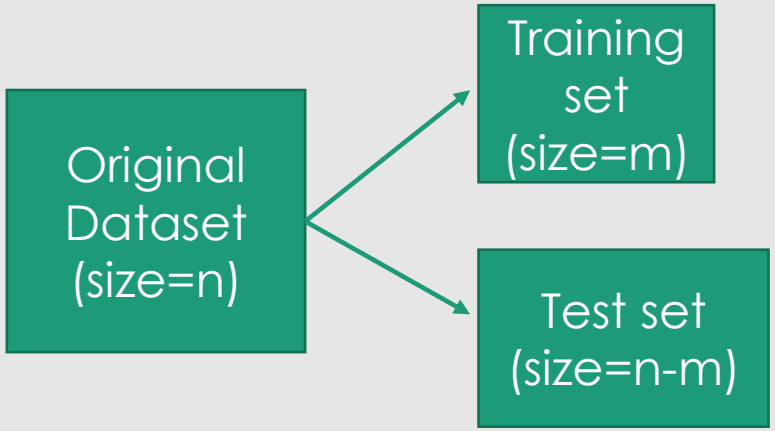
George E.P. Box
(1919 – 2013)

Some steps should be considered in developing prediction models:

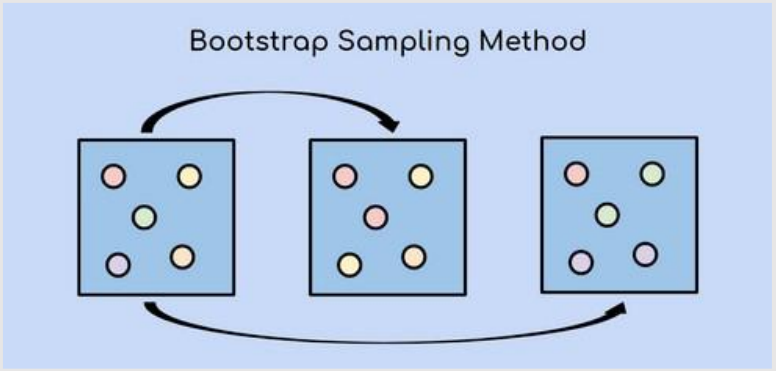


HOW TO EVALUATE MODEL PERFORMANCE* (internal validation)

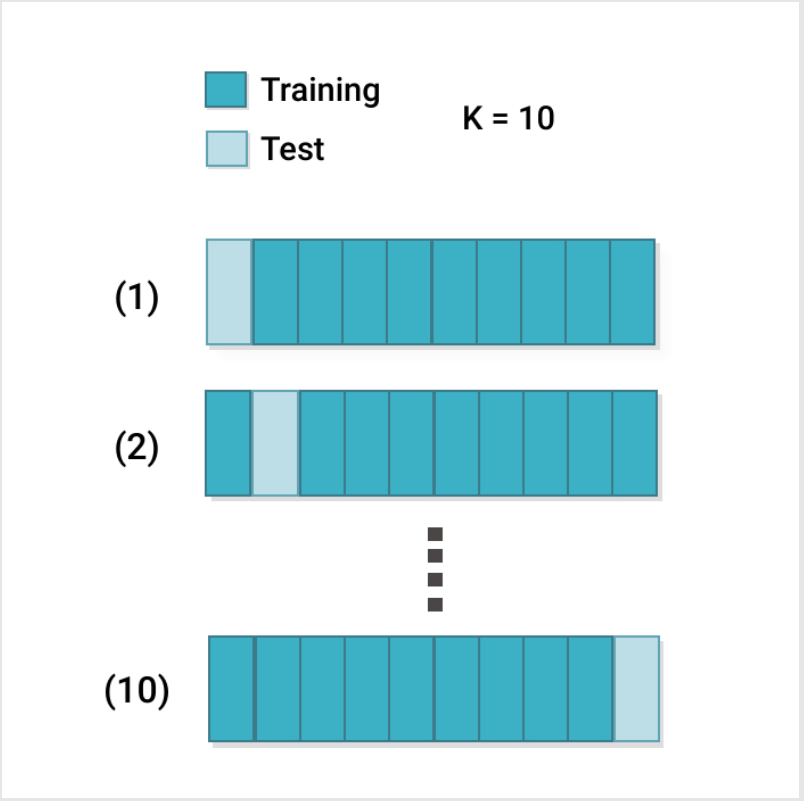
«Single» Split



Bootstrap validation



K-fold Cross-validation



*Note that [*ideally*] we should validate the *entire* model building process…

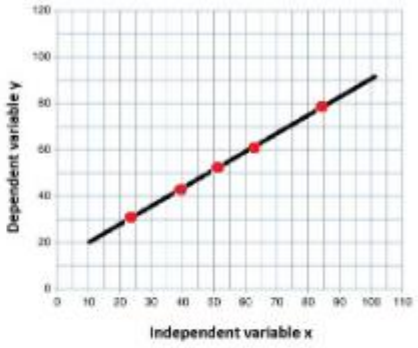
** ML has a further distinction in **validation set** for hyperparameters and **test set** for performance

Classical measure of overall performance: R squared

R^2 Values	Interpretation	Graph

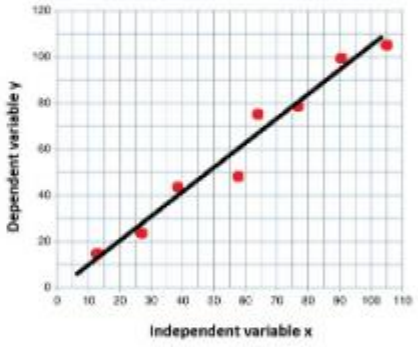
$$y = f(x) + \varepsilon$$

$R^2 = 1$ All the variation in the y values is accounted for by the x values



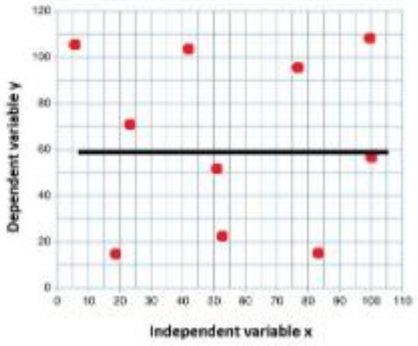
$$\bar{y} = \frac{1}{n}\sum_i y_i \qquad e_i = y_i - f_i$$

$R^2 = 0.83$ 83% of the variation in the y values is accounted for by the x values



$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \qquad SS_{res} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

$R^2 = 0$ None of the variation in the y values is accounted for by the x values



R^2 (coefficient of determination) is the proportion of the variance for a dependent variable that's explained by an independent variable in a regression model.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

fraction of variance **unexplained**

$$SS_{reg} = \sum_i (f_i - \bar{y})^2$$

R squared for multivariable (*generalized*) models

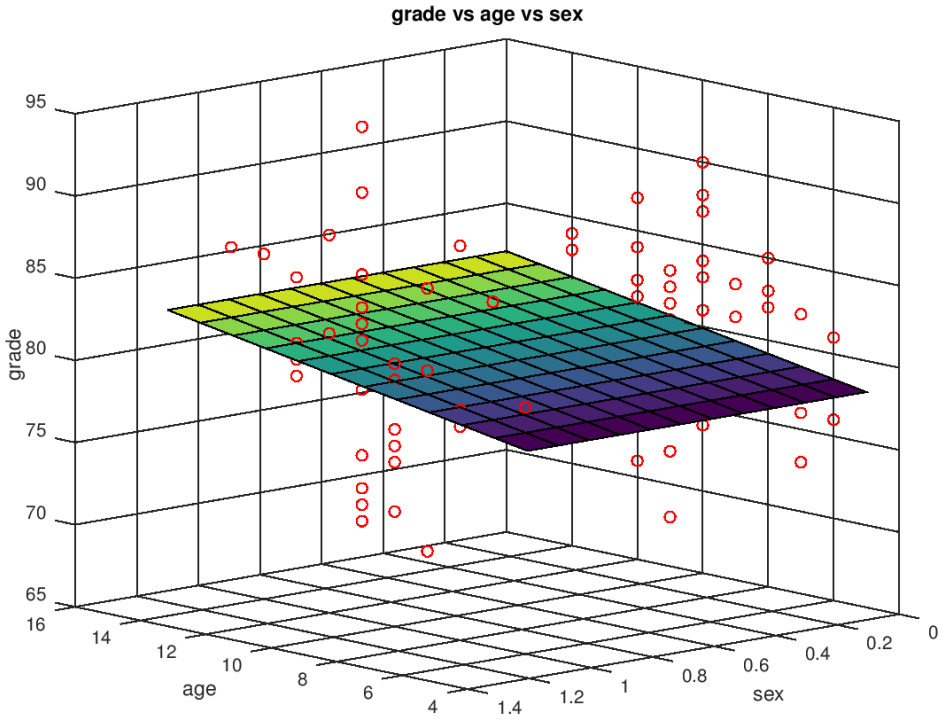
R^2 : % of variation in Y explained by the model
[**adjusted for *p*=#covariates, *n*=sample size**]

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

Binary/[time-to-event] models:

- Cox and Snell R^2
- Nagelkerke's R^2

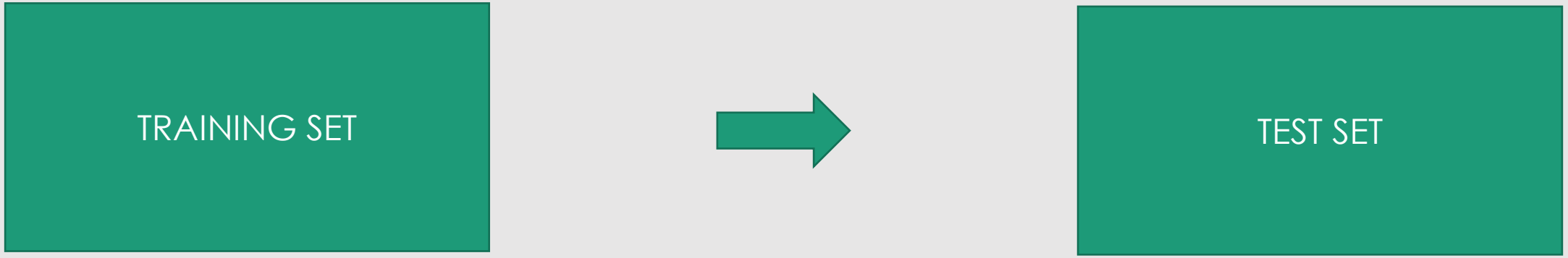
$$R^2_{CS} = 1 - \exp\left[\frac{2}{n}\left(ln(Lik_{Null}) - ln(Lik_{Model})\right)\right]$$



likelihood of the null model with only the intercept *vs* a given set of parameters

Measures of the accuracy of predictions

Are our predictions **reliable**?



Calibration: does the model predict *accurately*?
calibration **slope**, 1 : perfect calibration

Discrimination: does the model *discriminate* well?
C statistic (**AUC**ROC), 1: perfect discrimination, 0.5 : flipping a coin

Calibration (binary outcome/logistic regression)

For given values of the model covariates, we can obtain the predicted probability:

$$P(Y = 1|X_1, \dots X_p) = \frac{odds}{1 + odds} = \frac{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}$$

The model is said to be **well calibrated** if the observed risk **matches** the predicted risk (probability).

That is, if we were to take a large group of observations which are assigned a value P(Y=1)=0.2 the **proportion** of these observations with Y=1 ought to be close to 20%.

If instead the observed proportion was 80%, we would probably agree that the model is not performing well - it is under-estimating risk for these observations.

The comparison between predicted probabilities and observed proportions is the basis for the **Hosmer-Lemeshow (HL) test**.

Based on the estimated parameter values $\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_p$, for each observation in the sample the probability that Y=1 is calculated, depending on each observation's covariate values:

$$\hat{\pi} = \frac{exp(\widehat{\beta_0} + \widehat{\beta_1} x_1 + \cdots + \widehat{\beta_p} x_p)}{1 + exp(\widehat{\beta_0} + \widehat{\beta_1} x_1 + \cdots + \widehat{\beta_p} x_p)}$$

We divide the sample in groups up according to their predicted probabilities, or risks.

The observations in the sample are then split into ***g* groups** according to their predicted probabilities.

Suppose (as is commonly done) that *g*=10.

Then the first group consists of the observations with the lowest 10% predicted probabilities. The second group consists of the 10% of the sample whose predicted probabilities are next smallest, etc etc…

Suppose for the moment, artificially, that all of the observations in the first group had a predicted probability of 0.1.

Then, if our model is correctly specified, we would expect the proportion of these observations who have Y=1 to be 10%.

Of course, even if the model is correctly specified, the observed proportion will deviate *to some extent* from 10%, but not by too much (random variability…).

If the proportion of observations with Y=1 in the group were instead 90%, this is suggestive that our model is not accurately predicting probability (risk), i.e. an indication that our model is not fitting the data well.

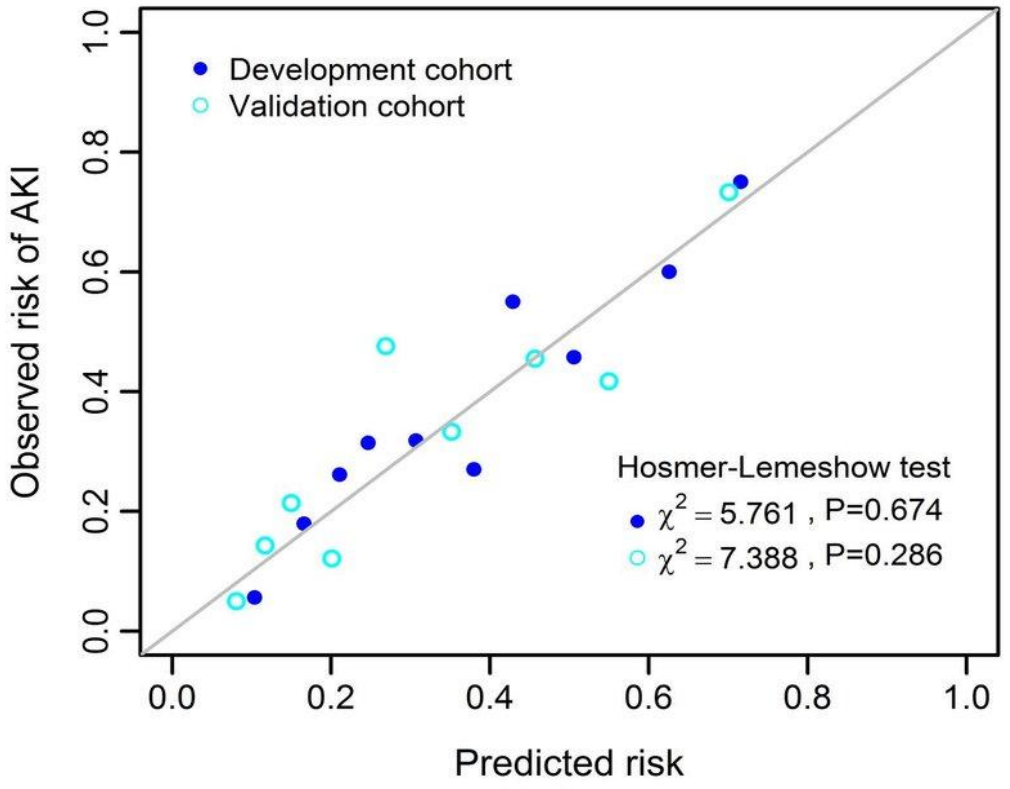
To calculate how many "Y=1" observations we would expect, the Hosmer-Lemeshow test takes the *average* of the predicted probabilities in the i-th group, and multiplies this by the number of observations in the group.

This calculation is then stratified with respect to the observed relative frequency of the outcomes in the groups.

Provided **p+1<g** (p=#covariates) the test statistic approximately follows a chi-squared distribution with g-2 degrees of freedom. Differences are computed for the "event" (k=1)and for the "non-event" (k=0).

If the p-value is small, this is indicative of poor fit.

$$\chi_{g-2} = \sum_{k=0}^{1} \sum_{l=1}^{g} \frac{(o_{kl} - e_{kl})^2}{e_{kl}}$$



But….a large p-value ***does not mean*** the model fits well, since lack of evidence against a null hypothesis is not equivalent to evidence in favour of the alternative hypothesis…

For example: if our sample size is small, do not reject H_0 may simply be a consequence of the test having lower power to detect misspecification, rather than being indicative of good fit.

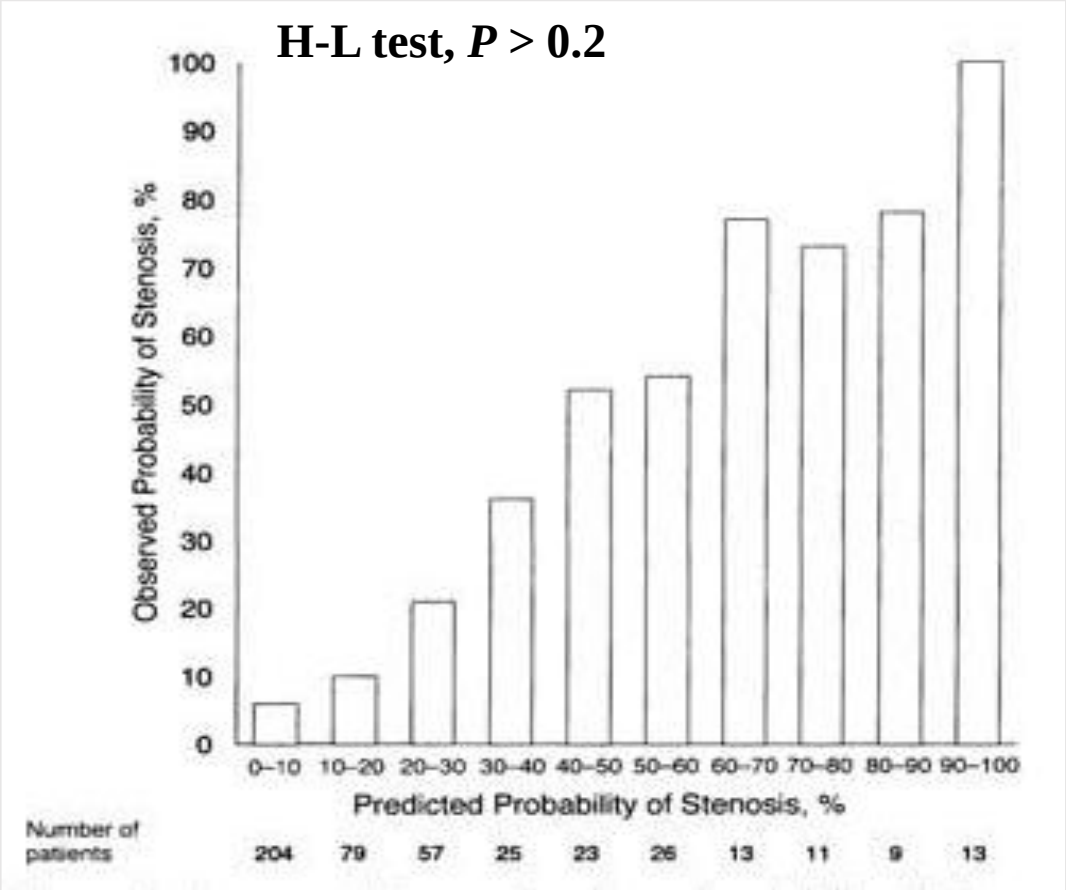
Block 3.3

Derivation of Scores in the Prediction Rule

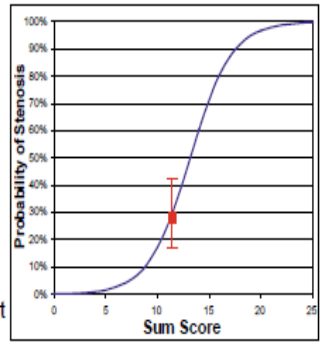
The multivariable logistic regression model can be written as:

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

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



	A	B	C	D	E	F	G	H
1	**Prediction rule for renal artery stenosis**							
2								
3	**Predictor**			Value	Score			
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	*Sumscore*				*11*			
18				Formula	Score chart			
19	*Predicted probability of renal artery stenosis*			28%	*25%*			
20	*Confidence interval*		17%	-	43%			
21	* femoral or carotid bruit, angina pectoris, claudication, myocardial infarction, CVA, or vascular surgery							
22	** or cholesterol lowering therapy							

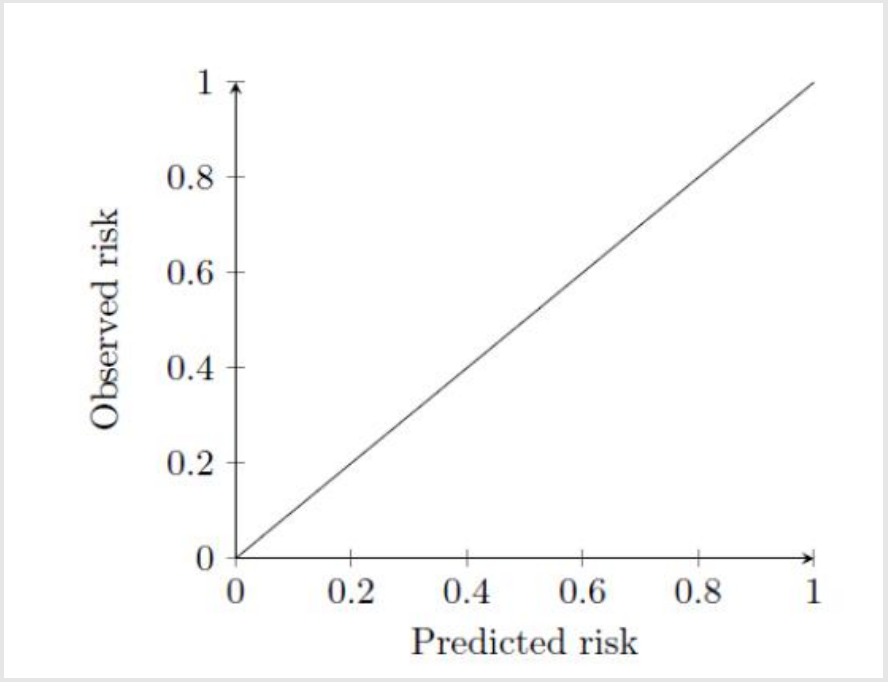


See figure for graphical illustration

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

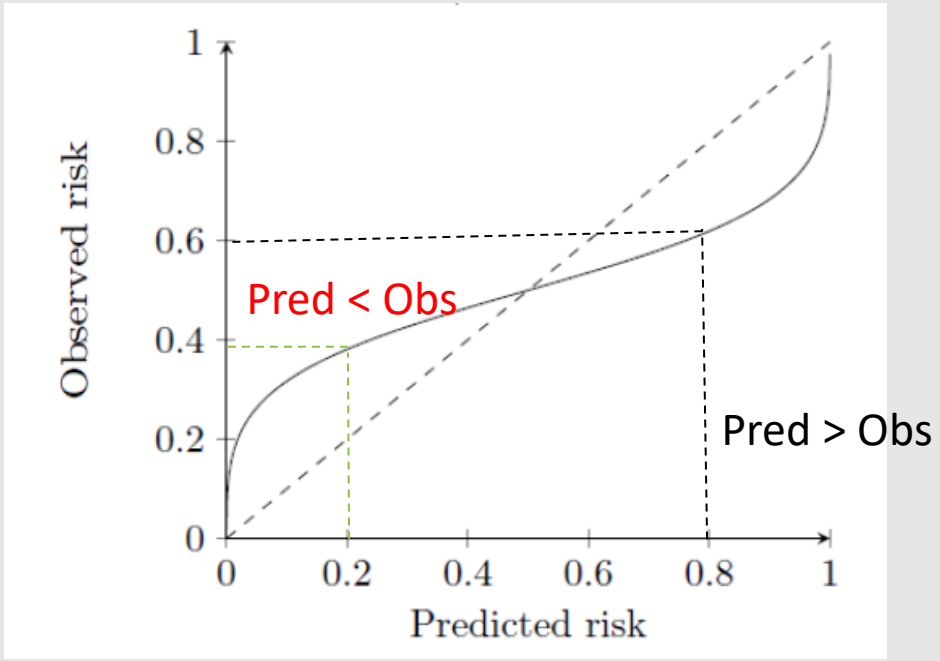
The sum score was 11, the estimated probability or renal artery stenosis was 28% [95% confidence interval 17–43%].

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



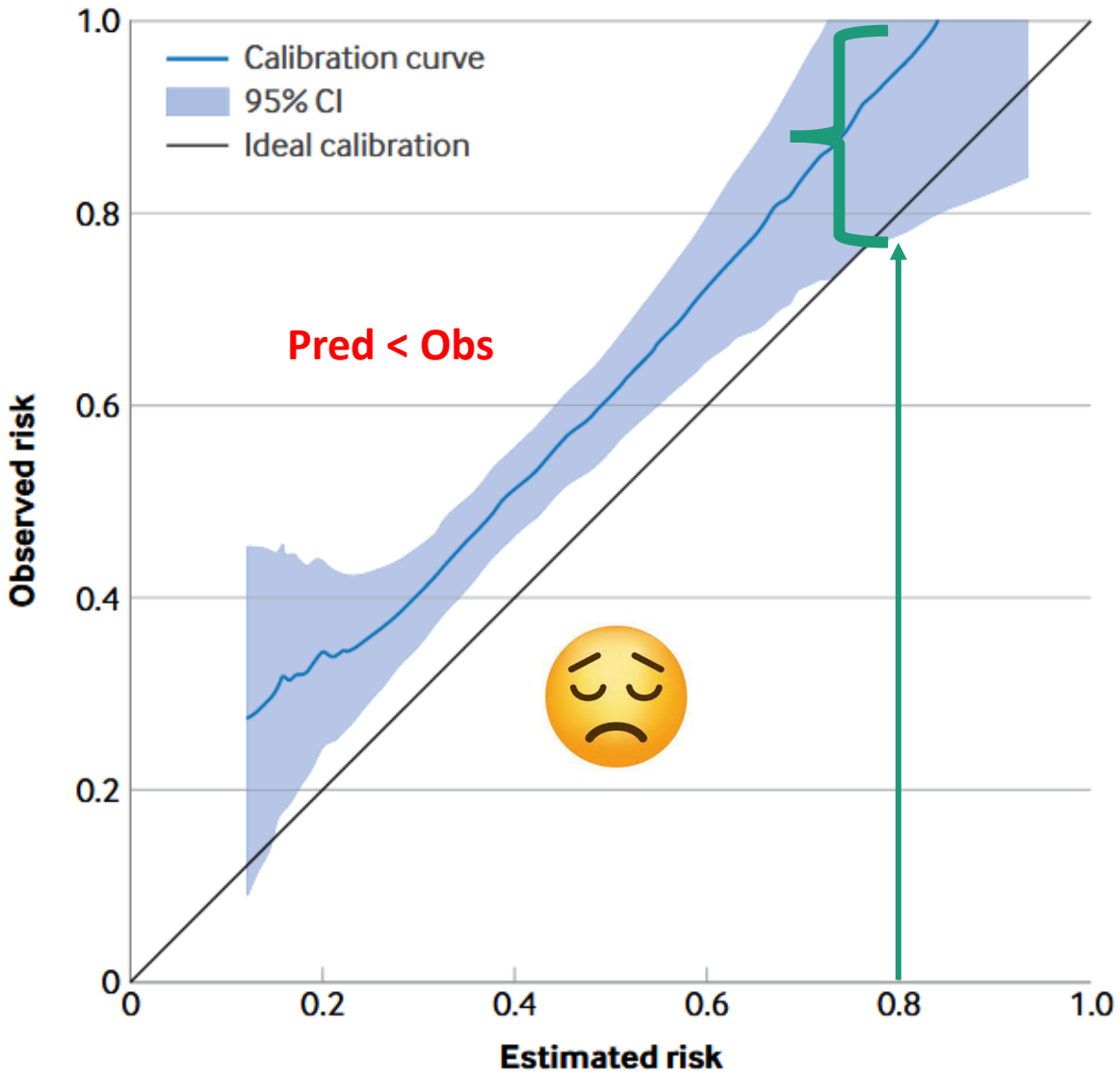
IDEAL MODEL:

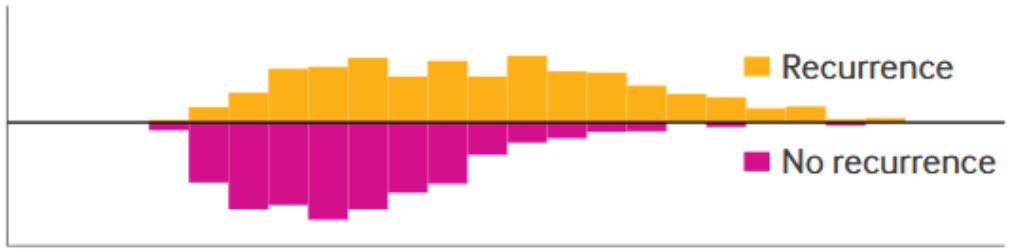
- perfect calibration
- calibration slope = 1



Uncalibrated MODEL:

- predicted risks too extreme
- calibration slope < 1
- may lead to harm!





Calibration plot and its 95% confidence interval, from an ***external validation*** of a model to estimate five year recurrence risk after a primary breast cancer diagnosis.

If a new individual is estimated a risk of **0.8** by the model, we could say:

In a group of 100 individuals with the same estimated risk as you, the model suggests that between **78** and **100** will have a recurrence by five years.

Distribution of **estimated risks** for those with and with no recurrence by five years.

Discrimination of a regression model [binary outcome] : AUC of the ROC curve

Should we be content to use a model so long as it is well calibrated? Unfortunately not.

To see why, suppose we fit a logistic model for our outcome Y but without any covariates, i.e. the model:

$$P(Y = 1) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

This (null) model assigns every observation **the same predicted probability** : it does not use any covariates.

Therefore β_0 will be the *observed overall log odds of a positive outcome*, such that the predicted value of P(Y=1) will be identical to the proportion of Y=1 observations in the dataset.

This (rather useless) model assigns every observation the same predicted probability. It will have good calibration ! - in future samples the observed proportion will be close to our estimated probability.

However, **the model isn't really useful** because it doesn't **discriminate** between those at high risk and those at low risk. The situation is analogous to a weather forecaster who, every day, says the chance of rain tomorrow is 10%. This prediction might be well calibrated (over a long period), but it doesn't tell people whether it is more or less likely to rain on a given day, and so isn't really a helpful forecast!

As well as being well calibrated, we would therefore like our model to have high **discrimination** ability.

In the binary outcome context, this means that observations with Y=1 ought to be predicted **high probabilities**, and those with Y=0 ought to be assigned **low probabilities**.

Such a model allows us to discriminate between low and high risk observations.

Recall the important notions of **sensitivity** and **specificity** of a test or prediction rule (from block 1!):

Sensitivity: probability of the model predicting an observation as '*positive*' given that is true (Y=1).

In words, the sensitivity is the proportion of truly positive observations which is classified as such by the model or test.

Specificity: probability of the model predicting '*negative*' given that the observation is 'negative' (Y=0).

Our model or prediction rule is perfect at classifying observations if it has 100% sensitivity and 100% specificity. In practice this is (usually) not attainable.

So how can we summarize the **discrimination ability** of our logistic regression model?

For each observation, our fitted model can be used to calculate the fitted probabilities $P(Y = 1|X_1, \ldots X_p)$

On their own, these don't tell us how to classify observations as positive or negative.

One way to create such a classification rule is **to choose a cut-point c**, and classify those observations with a fitted **probability > c as positive** and **those <= c as negative**.

For this specific cut-off, the sensitivity is the proportion of observations with Y=1 which have a predicted probability > c, and similarly the specificity is the proportion of Y=0 observations with a predicted probability <= c:

Predicted Probability		**Outcome**		
		Y=1	**Y=0**	**Tot**
cutoff	**> c**	a	b	a+b
	<=c	c	d	c+d
	Tot	a+c	b+d	n

Sensibility=a/a+c

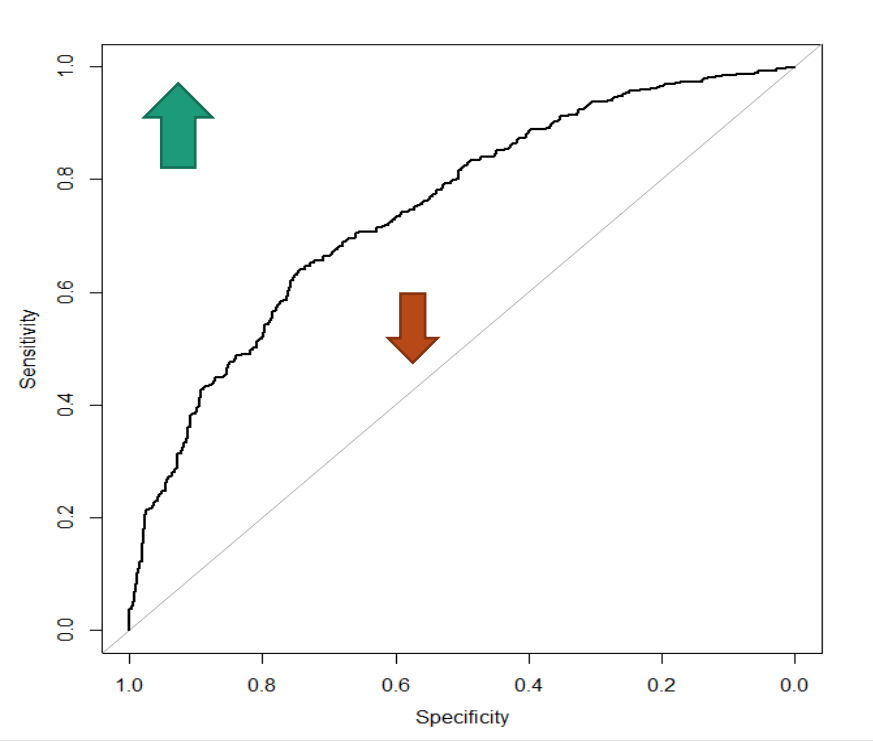
Specificity=d/b+d

If we increase the cut-point c, fewer observations will be predicted as positive.

This will mean that fewer of the Y=1 observations will be predicted as positive (reduced sensitivity), but more of the Y=0 observations will be predicted as negative (increased specificity).

In picking the cut-point, there is thus an intrinsic **trade-off** between sensitivity and specificity.

Now we come to the ROC curve: we plot all the values of sensitivity against (1-specificity), as the value of the cut-point c is increased from 0 through to 1:



A model with **high discrimination ability** will have high sensitivity and specificity simultaneously, leading to a ROC curve which goes close to the top left corner of the plot.

A model with **no discrimination ability** will have an ROC curve which is the 45 degree diagonal line.

Area under the ROC curve:

To **summarize** the discrimination ability of a model we can report the area under the ROC curve (with corresponding 95% CI).

A model with high discrimination ability has an ROC curve which goes closer to the top left hand corner of the plot, whereas a model with low discrimination ability has an ROC curve close to a 45 degree line.

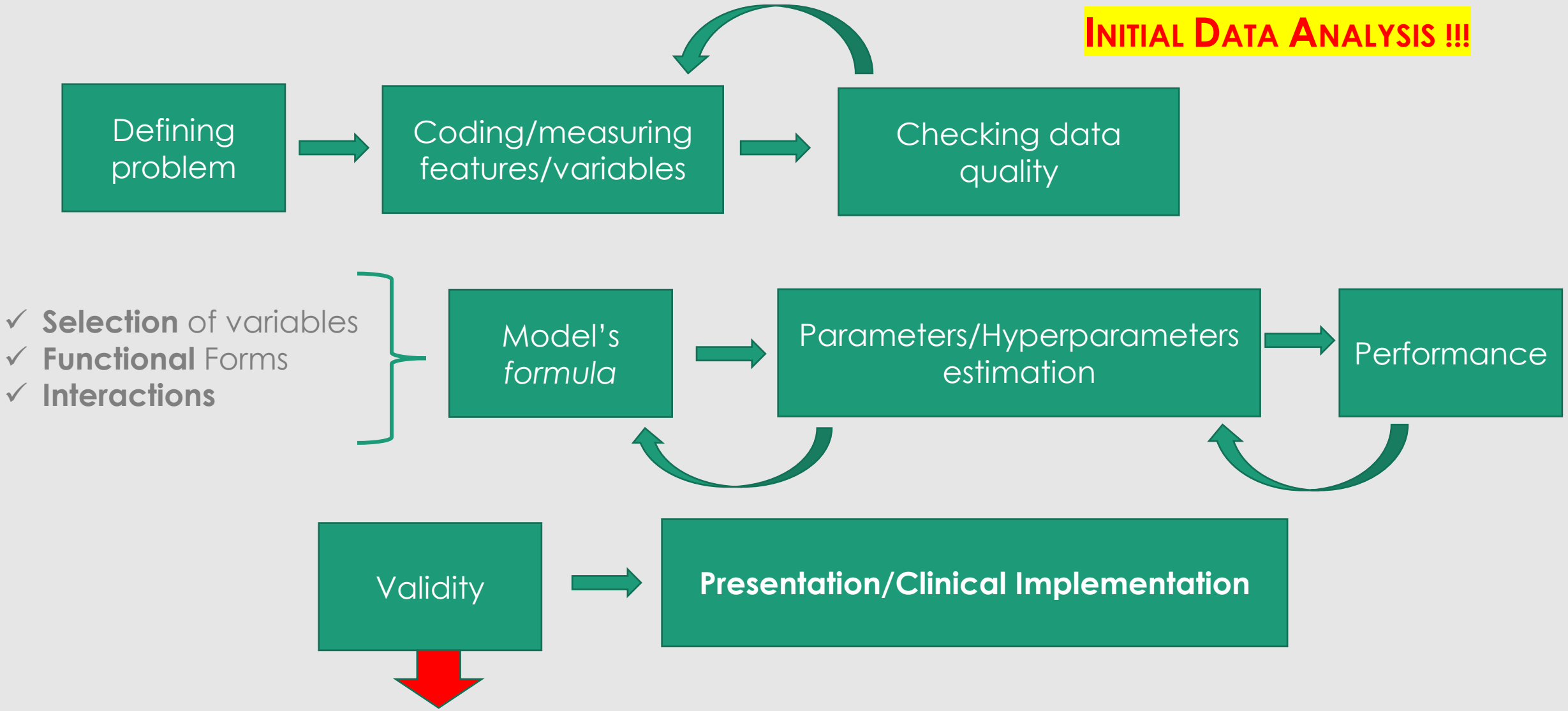
Thus AUC ranges from 1, corresponding to perfect discrimination, to 0.5, corresponding to a model with no discrimination ability.

The area under the ROC curve is also sometimes referred to as the c-statistic (c for concordance).

The AUC has a somewhat appealing interpretation:

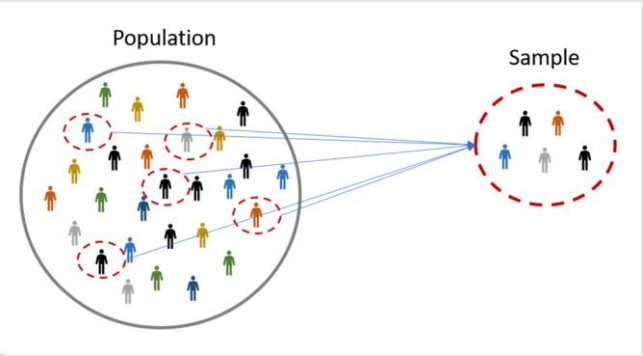
The AUC is the probability that if you were to take *a random pair of observations*, one with Y=1 and one with Y=0, the observation with Y=1 has a **higher predicted probability** than the other. The AUC thus gives the probability that the model **correctly ranks the risk** of such pairs of observations.

To summarize again… it is a quite long journey …!



*On **external** dataset !!!*

Last but not Least !
Sample Size



Model development phase:

- We should have a ***large enough sample size*** to develop a model that provides *accurate* risk predictions in ***new*** individuals from target population

- Many (most?) models do not perform well in new data

Why?

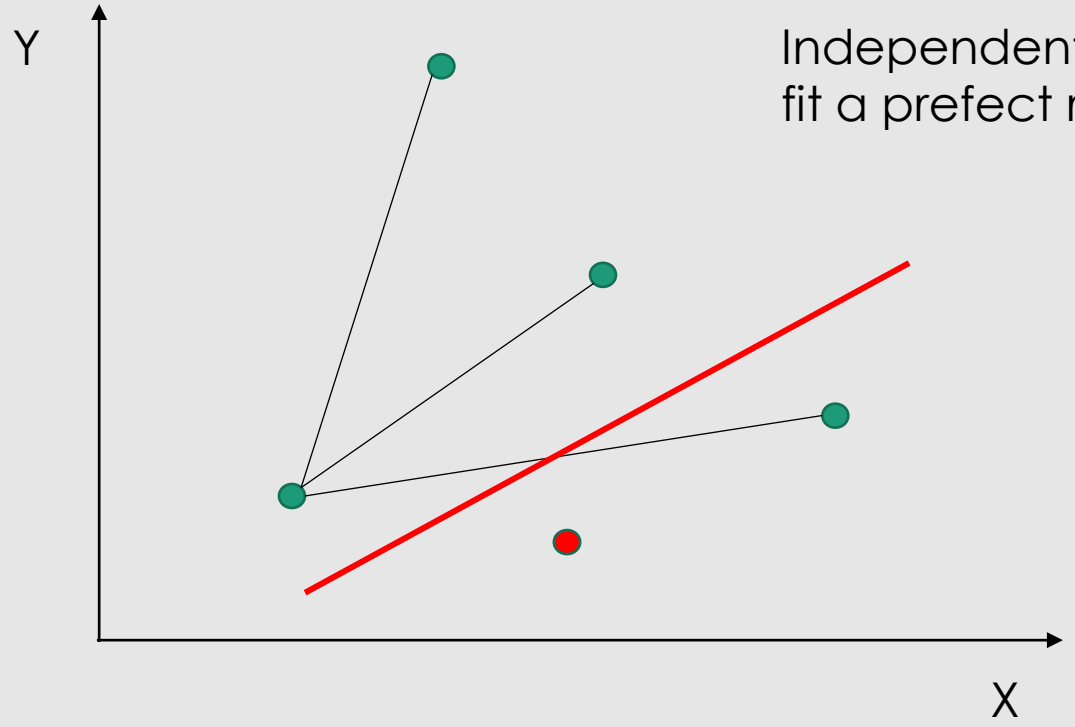
Often:

- **small** sample sizes (large imprecision)
- **overfitting** (poor generalization)

The well-known concept of *degrees of freedom*

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

What is the *minimum number* of observations required to estimate this simple linear regression model?

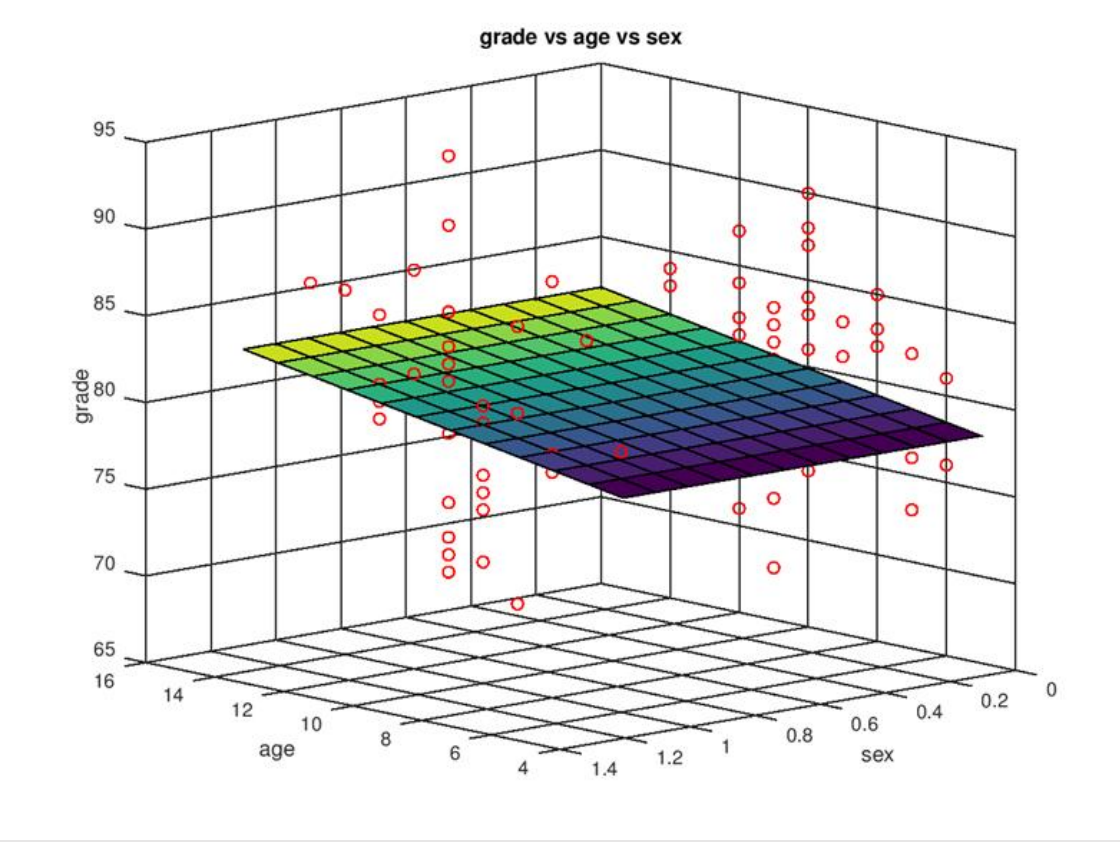


Independently from *where* the second point is, we will always fit a prefect regression line with two points ($R^2 = 1$)…

Only when a third point is included the model gain some *freedom* to assess the *best* fitting line…(df=1)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

What is the *minimum number* of observations required to estimate this multivariable linear regression model?



Any 3 points in a 3-dimensional space can identify a *perfect* plane

We need at least 4 points to gain 1 degree of freedom !

Given k parameters (regression coeff) to be estimated:

$$df = n - k - 1$$

Rule-of-thumb: a fitted regression model is likely to be *reliable* when the number of predictors *p* is less than *m*/10 or *m*/20, where *m* is the limiting sample size

Type of Outcome	Limiting sample Size *m*
Continuous	*n* (total sample size)
Binary	min(n1, n2)
Time-to-event	Number of *failures*

Linear regression
For 3 predictors we need > 3*10 = 30 individuals

Survival model
For 3 predictors we need more than 3*10=30 failure events (eg. deaths)

Logistic regression
Assuming that cases is the rarer category,
for 3 predictors we need more than 3*10=30 **cases**

Recent **guidelines** have been proposed:

Sample size for model development

RESEARCH ARTICLE 2018

WILEY Statistics in Medicine

Minimum sample size for developing a multivariable prediction model: Part I – Continuous outcomes

Richard D. Riley[1] | Kym I.E. Snell[1] | Joie Ensor[1] | Danielle L. Burke[1] | Frank E. Harrell Jr[2] | Karel G.M. Moons[3] | Gary S. Collins[4]

RESEARCH ARTICLE 2018

WILEY Statistics in Medicine

Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes

Richard D Riley[1] | Kym IE Snell[1] | Joie Ensor[1] | Danielle L Burke[1] | Frank E Harrell Jr[2] | Karel GM Moons[3] | Gary S Collins[4]

Sample size for model validation

RESEARCH ARTICLE 2020

Statistics in Medicine WILEY

Minimum sample size for external validation of a clinical prediction model with a continuous outcome

Lucinda Archer[1] | Kym I. E. Snell[1] | Joie Ensor[1] | Mohammed T. Hudda[2] | Gary S. Collins[3] | Richard D. Riley[1]

RESEARCH ARTICLE 2020

Statistics in Medicine WILEY

Minimum sample size for external validation of a clinical prediction model with a binary outcome

Richard D. Riley[1] | Thomas P. A. Debray[2] | Gary S. Collins[3,4] | Lucinda Archer[1] | Joie Ensor[1] | Maarten van Smeden[2] | Kym I. E. Snell[1]

RESEARCH ARTICLE 2021

Statistics in Medicine WILEY

Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome

Richard D. Riley[1] | Gary S. Collins[2,3] | Joie Ensor[1] | Lucinda Archer[1] | Sarah Booth[4] | Sarwar I. Mozumder[4] | Mark J. Rutherford[4] | Maarten van Smeden[5] | Paul C. Lambert[4,6] | Kym I. E. Snell[1]

Calculate sample size that is needed to:

- minimise potential ***overfitting***
- estimate *overall* risk ***precisely***

Requires calculations for **multiple** criterion

Example: sample size for binary [& time-to-event] outcomes

Example of criteria to meet:

- $S \geq 0.9$ (calibration slope)
  [< 10% overfitting]
- R^2 *apparent* – R^2 *adjusted* < 0.05
- AUC ≥ 0.80

Closed-form formulae for each criterion:

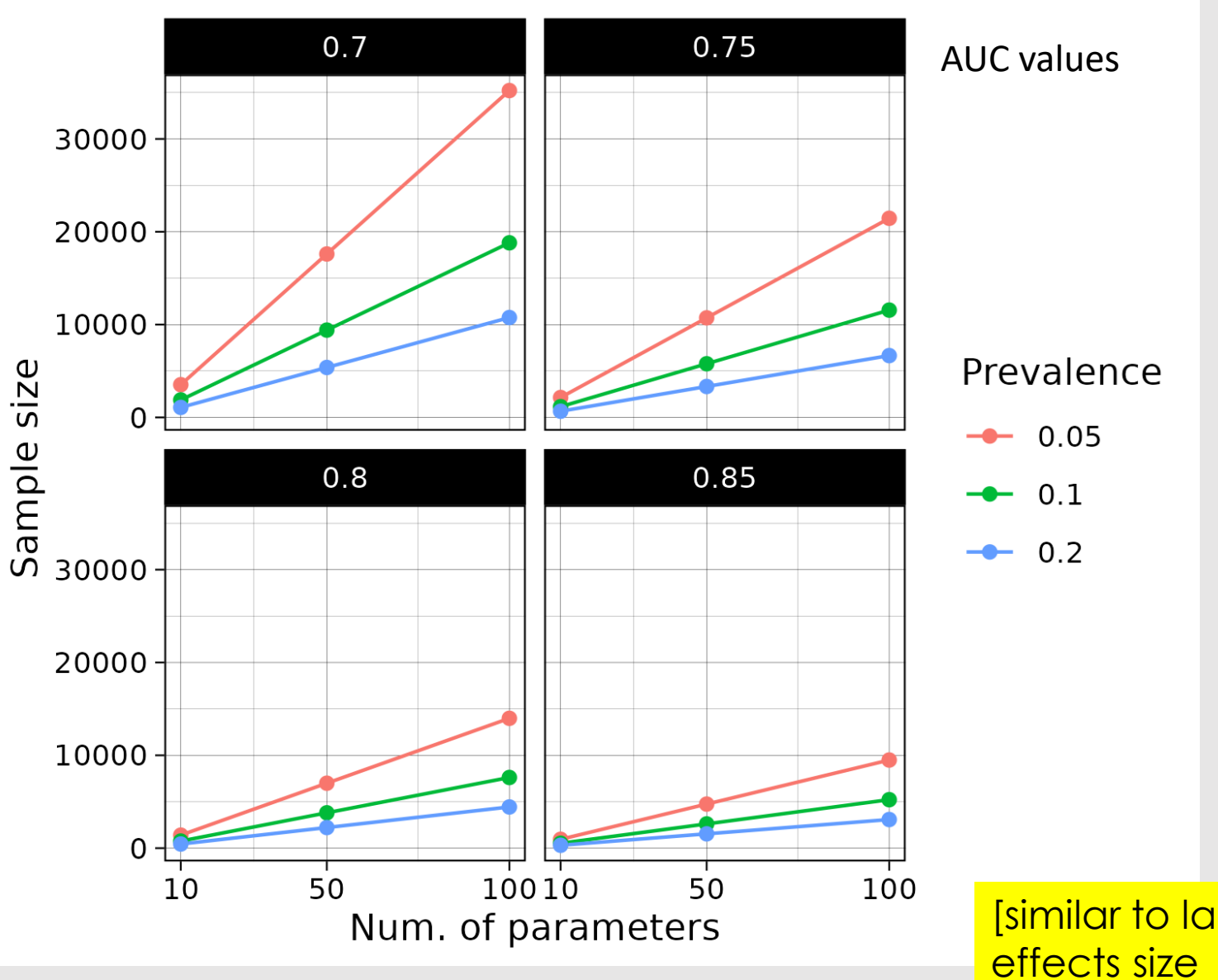
You should pre-specify:

- # predictors parameters
- desired S (calibration)
- Overall «risk» in the target population
- Model's anticipated R^2 / AUC

Stata module by Ensor et al: **PMSAMPSIZE**

pmsampsize

Binary outcome



Prevalence/Incidence ↑

Sample Size ↓

parameters ↑

Sample Size ↑

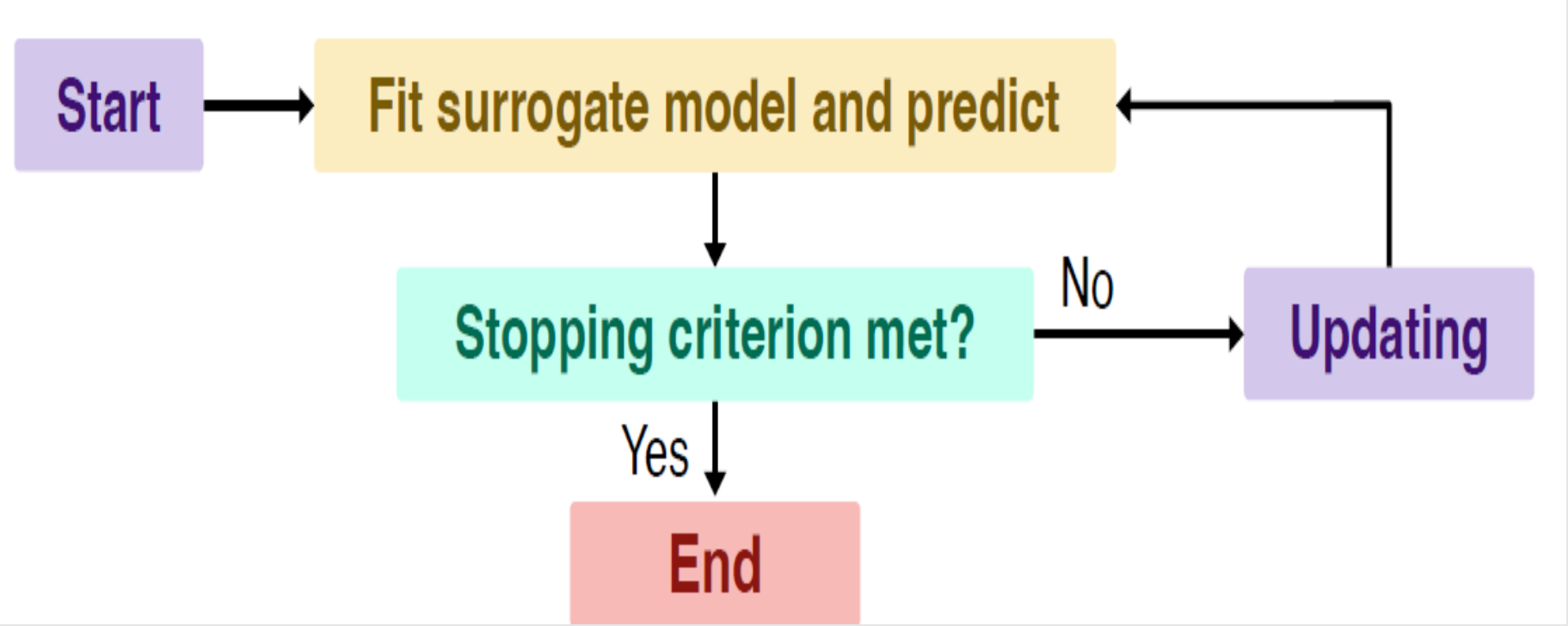
AUC/R^2 ↑

Sample Size ↓

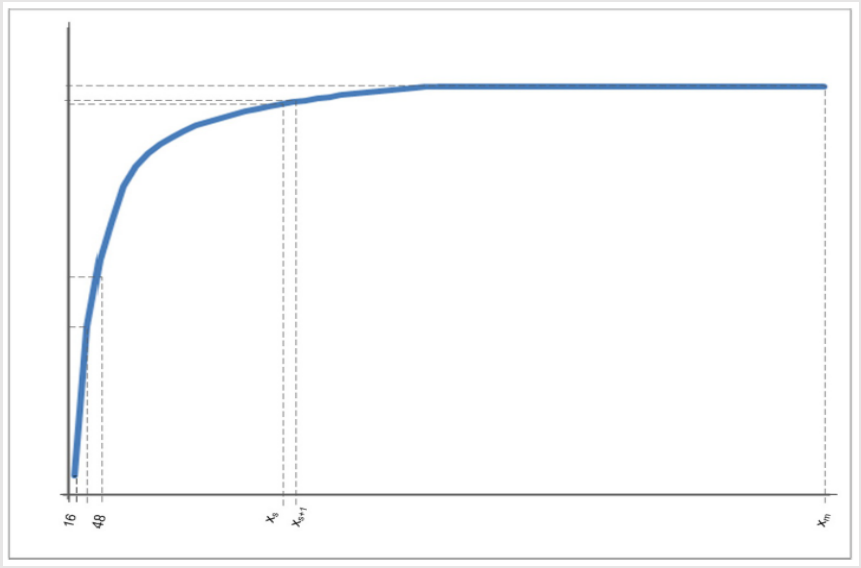
[similar to larger effects size in hypothesis tests]

We can estimate ***minimum sample size*** for continuous, binary, and survival outcomes [*standard* regression approaches]

For ML algorithms, work is in progress in developing ***simulation-based* approaches**



Basic idea : estimation of the **shape of the learning curve** for increasing sample size, to reach a *pre-determined performance* (with real or *simulated* data):



Figueroa *et al. BMC Medical Informatics and Decision Making* 2012, **12**:8
http://www.biomedcentral.com/1472-6947/12/8

BMC Medical Informatics & Decision Making

RESEARCH ARTICLE Open Access

Predicting sample size required for classification performance

Rosa L Figueroa[1†], Qing Zeng-Treitler[2*†], Sasikiran Kandula[2†] and Long H Ngo[3†]

Learning Curves: Asymptotic Values and Rate of Convergence

Corinna Cortes, L. D. Jackel, Sara A. Solla, Vladimir Vapnik, and John S. Denker
AT&T Bell Laboratories
Holmdel, NJ 07733

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE

The Shape of Learning Curves: a Review

Tom Viering, Marco Loog

✓ Flexible and applicable to *any* model and data format

× *Require **pre-hoc** model specification (hyperparameters/# of layers..)*

× *Require available [or syntethic?] data*

Supplementary materials

Last but not least (II) …clinical usefulness !

npj | precision oncology

Comment

Published in partnership with The Hormel Institute, University of Minnesota

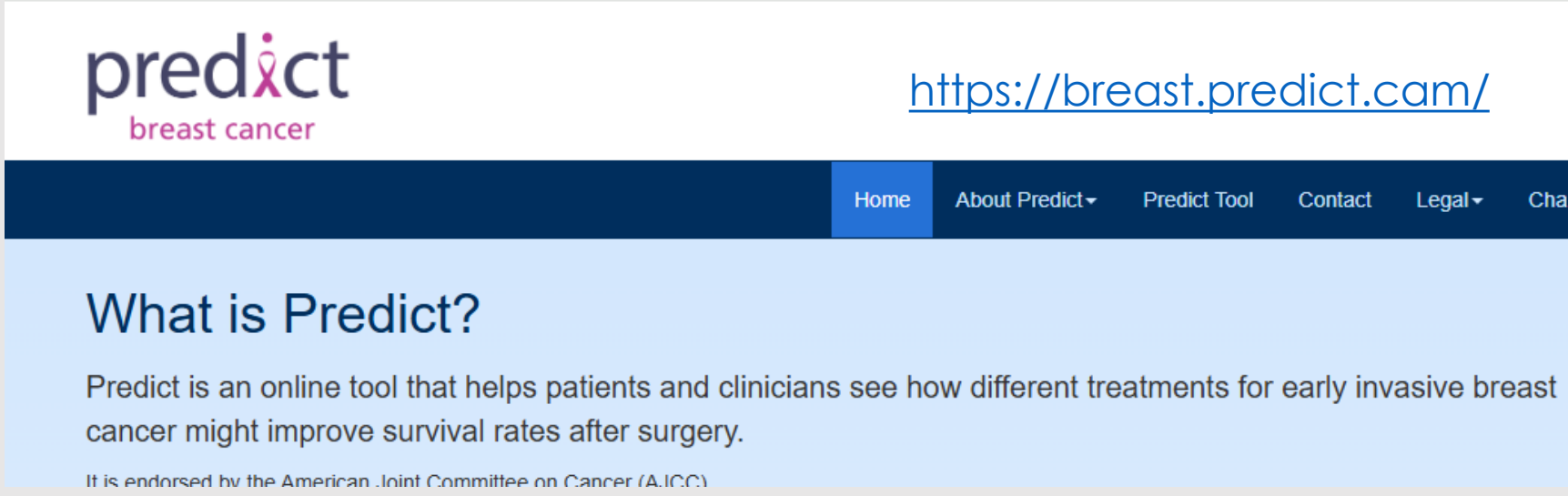
https://doi.org/10.1038/s41698-024-00553-6

All models are wrong and yours are useless: making clinical prediction models impactful for patients

1. Success in *academia* is not the same as success in the **clinic**
2. Successful models use data that are available in **routine practice**
3. Successful models are linked to **actions**

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

How might **different treatments** for early invasive breast cancer **improve survival** rates after surgery ?



Clinical Usefulness: the Net Benefit

$\hat{p}$ predicted probability from the model

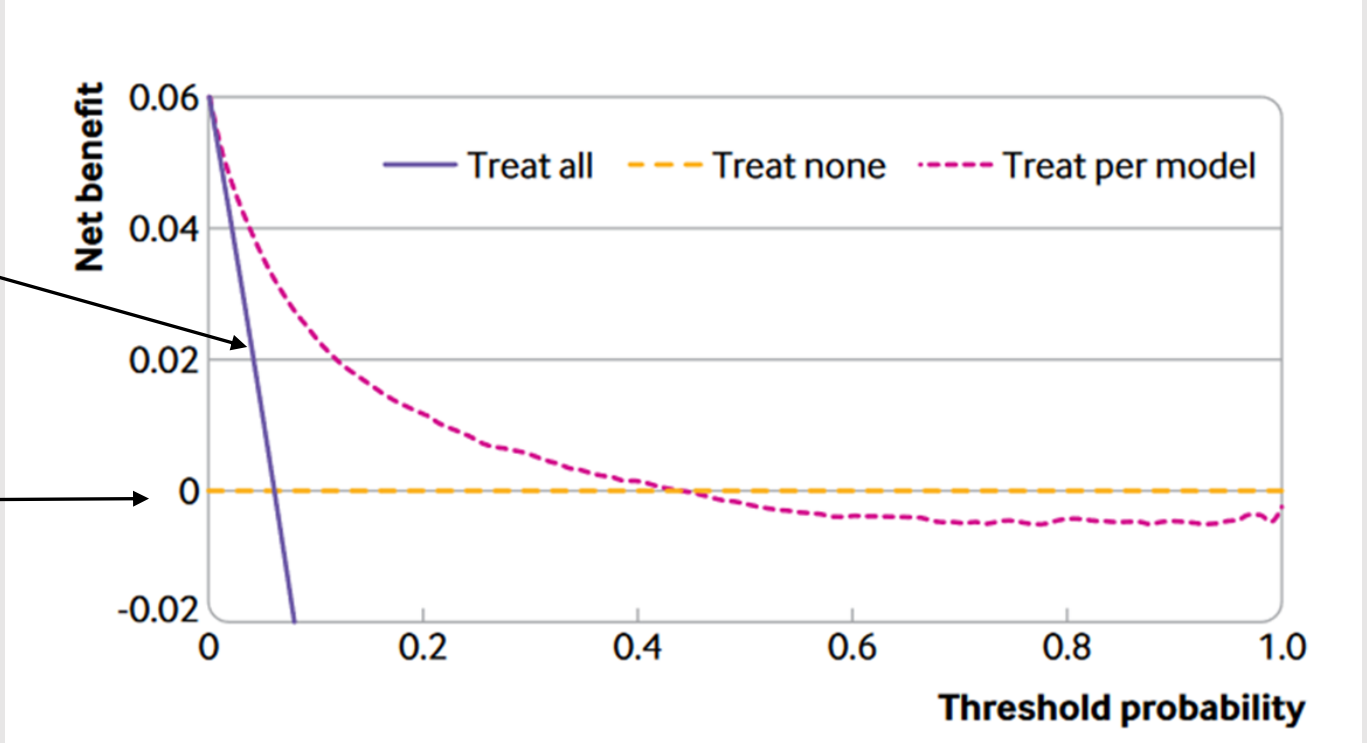
p_t **threshold probability**

$$NB = \frac{True\ Positive}{n} - \frac{False\ Positive}{n}\left(\frac{p_t}{1-p_t}\right)$$

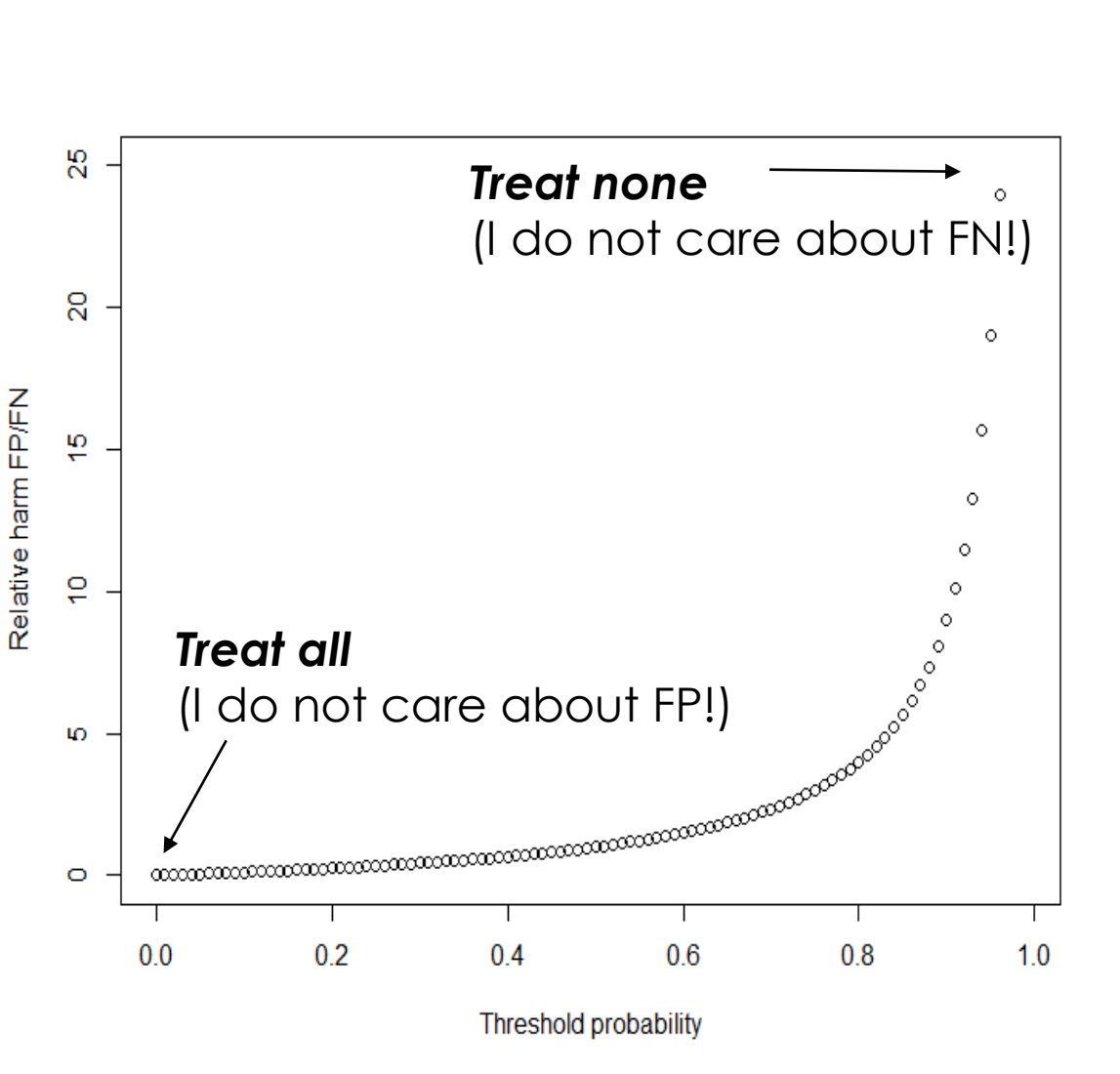
$\hat{p} \geq p_t$ clinical «action»

TP and FP fixed but the relative harm is increasing over p_t

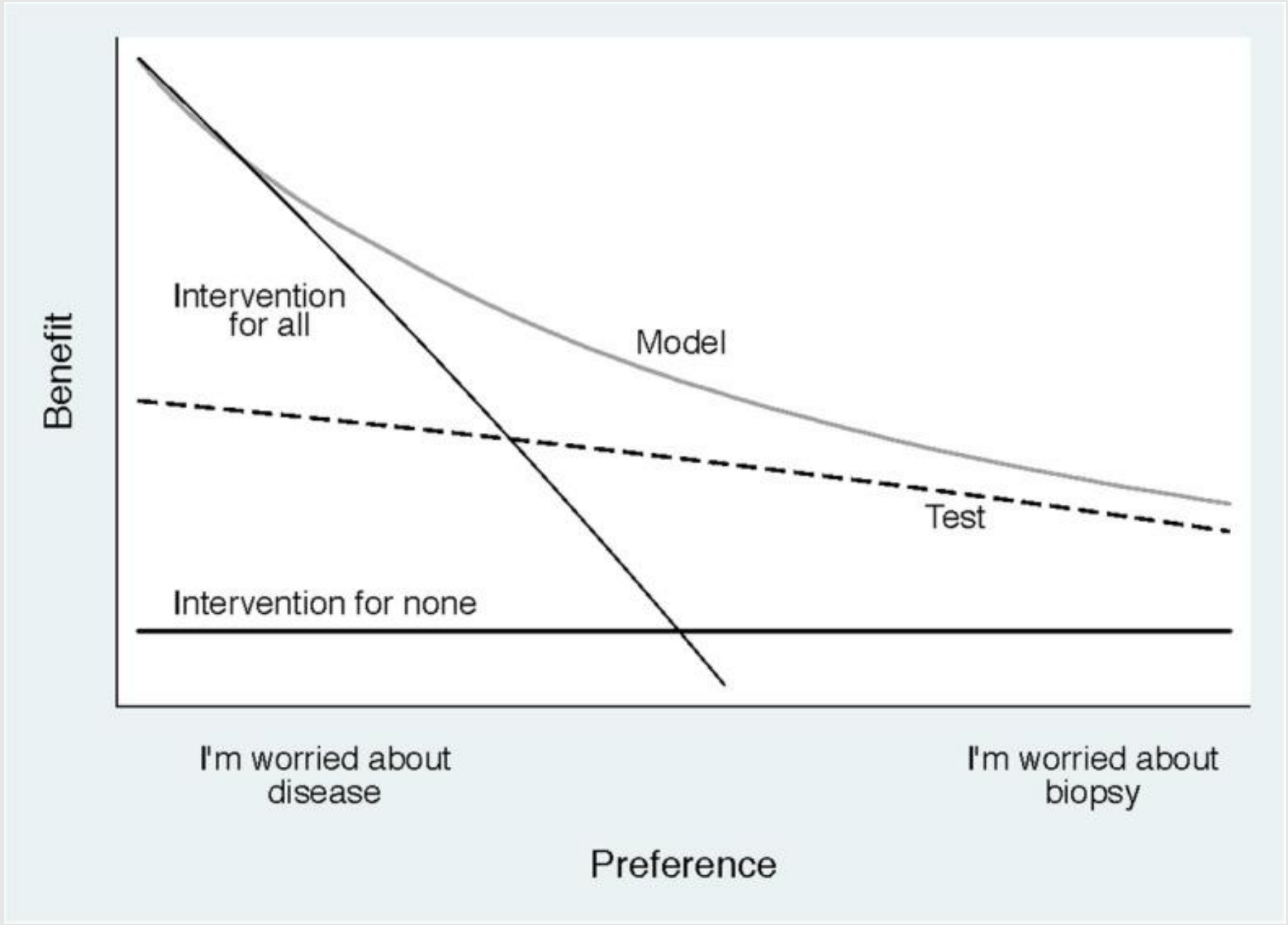
The true- and false-positives when considering **all patients as negative** are both 0, and NB is 0.



https://mskcc-epi-bio.github.io/decisioncurveanalysis/index.html



1. Chose a value for p_t.
2. Calculate TP and FP using p_t as the cut-point.
3. Calculate NB of the prediction model.
4. Vary p_t over an appropriate range and repeat steps 2 – 3.
5. Plot NB on the y axis against p_t on the x axis.
6. Repeat steps 1 – 5 *for each model* under consideration.
7. Repeat steps 1 – 5 for the strategy of assuming all patients are positive **(TP and FP fixed).**
8. Draw a straight line at y=0 : NB assuming that all patients are negative


Benefit
Intervention for all
Model
Test
Intervention for none
I'm worried about disease
I'm worried about biopsy
Preference