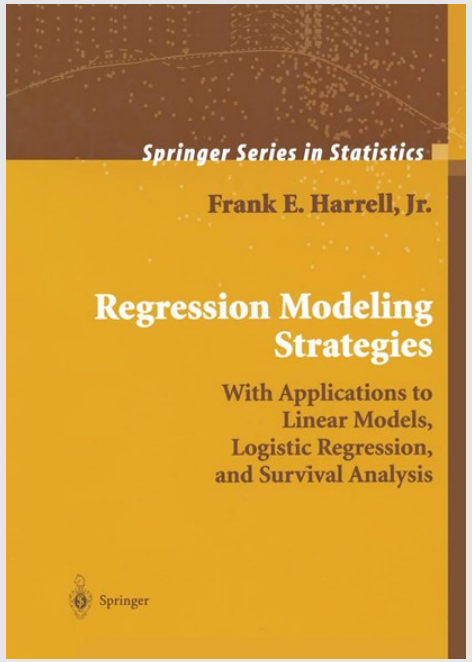
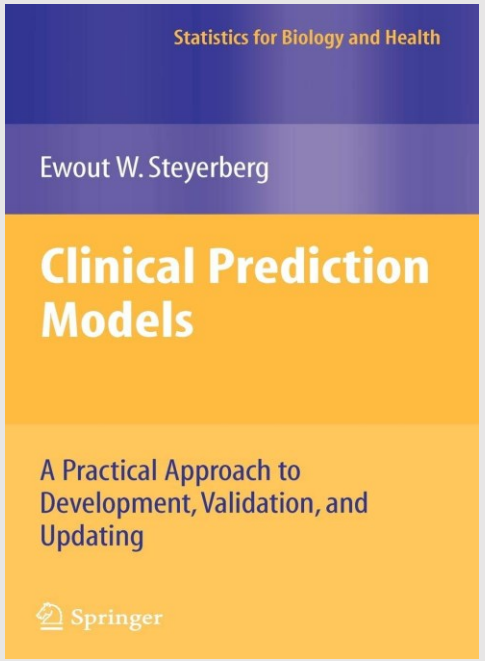
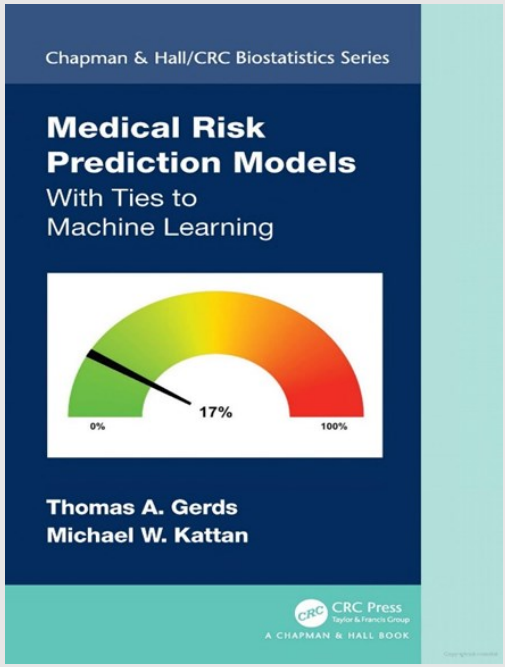
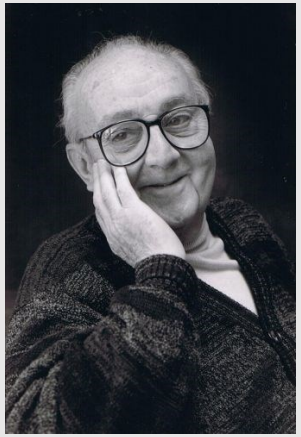
Prediction Models in Epidemiological & Clinical Research: performance & sample size







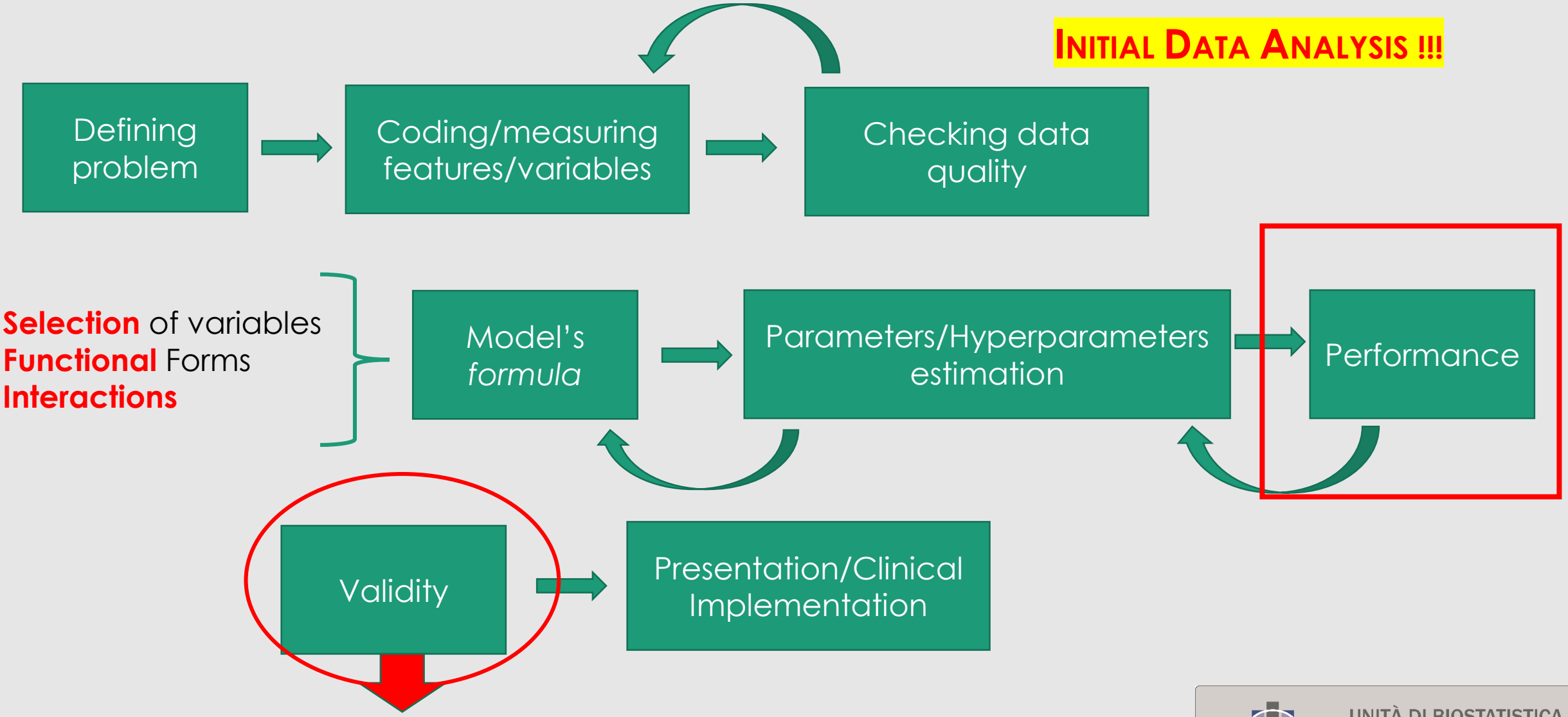


All models are wrong but some are useful.

George E.P. Box
(1919 – 2013)

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Some steps should be considered in developing prediction models:



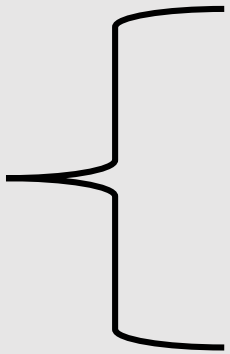
UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Measuring performance

While working on the development/validation of a prediction model, evaluating the *performance* is a crucial step.

1. R^2-type measures or % of the *explained variation* of the outcome

2. Are our predictions ***reliable*** ?

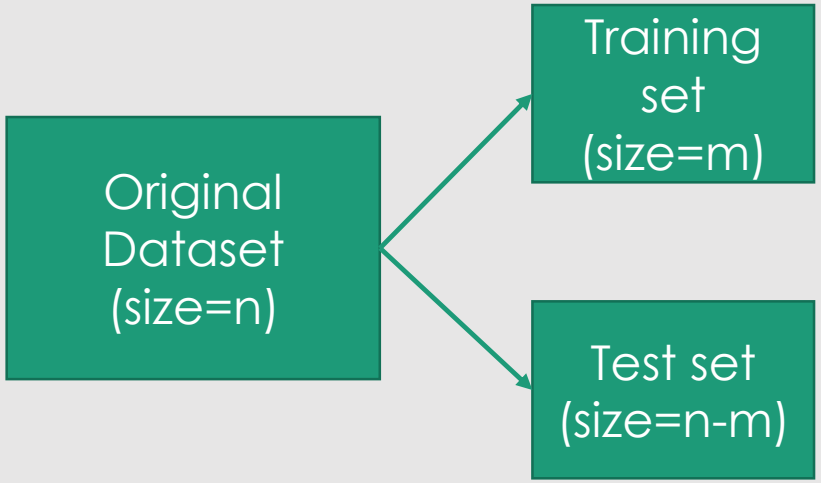


1.1 **Calibration**: does the model predict *accurately*? [calibration slope, 1 : perfect calibration]

1.2 **Discrimination**: does the model *discriminate* well? [C statistic (AUCROC), 1: perfect discrimination, 0.5 : flipping a coin]

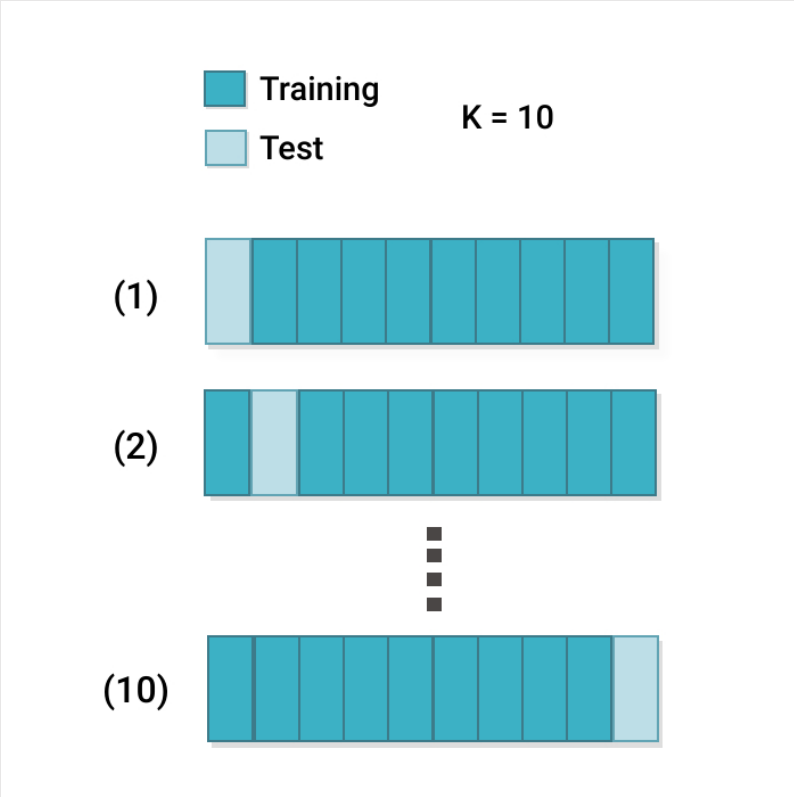
How to use the data in building the model* and perfomance evaluations ?

Single Split



- *Conditional* performance
- *Dependence* on the single split
- Waste of data

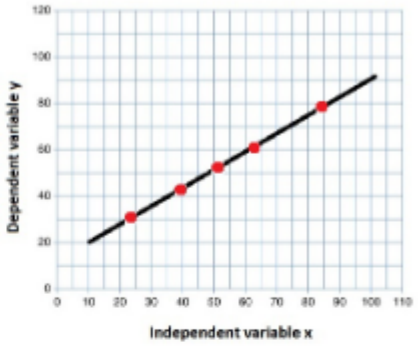
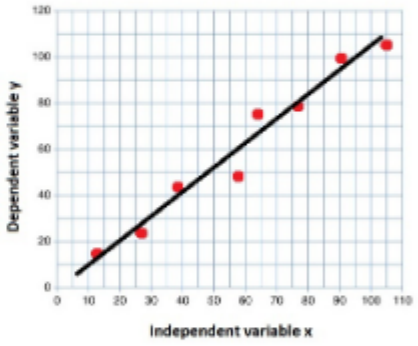
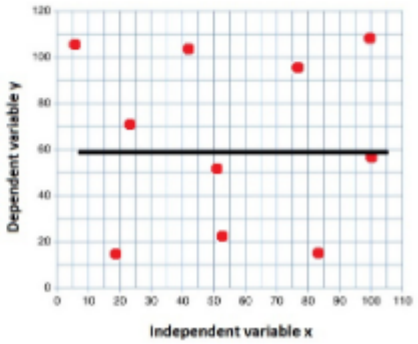
Cross-validation (boostrap)



- «average» performance

* different scenarios: 1. Evaluating performance of a given model vs 2. Comparing alternative models …

Overall performance: R squared

R^2 Values	Interpretation	Graph
$R^2 = 1$	All the variation in the y values is accounted for by the x values	
$R^2 = 0.83$	83% of the variation in the y values is accounted for by the x values	
$R^2 = 0$	None of the variation in the y values is accounted for by the x values	

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

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

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

R^2 (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**

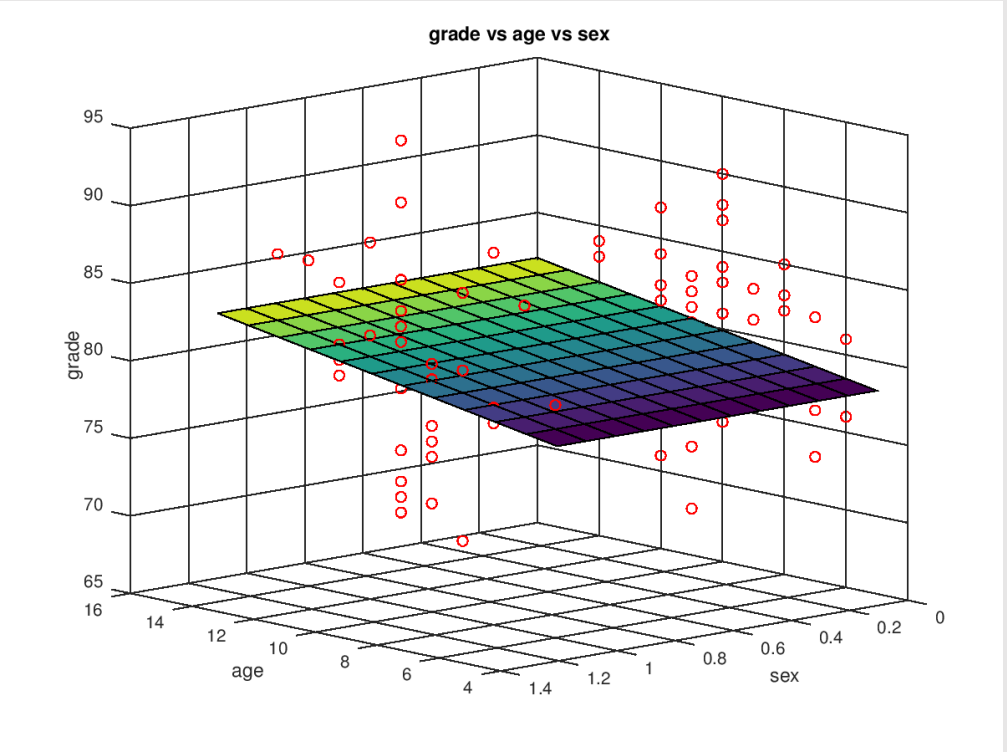
$$SS_{reg} = \sum_i (f_i - \bar{y})^2$$

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

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

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.

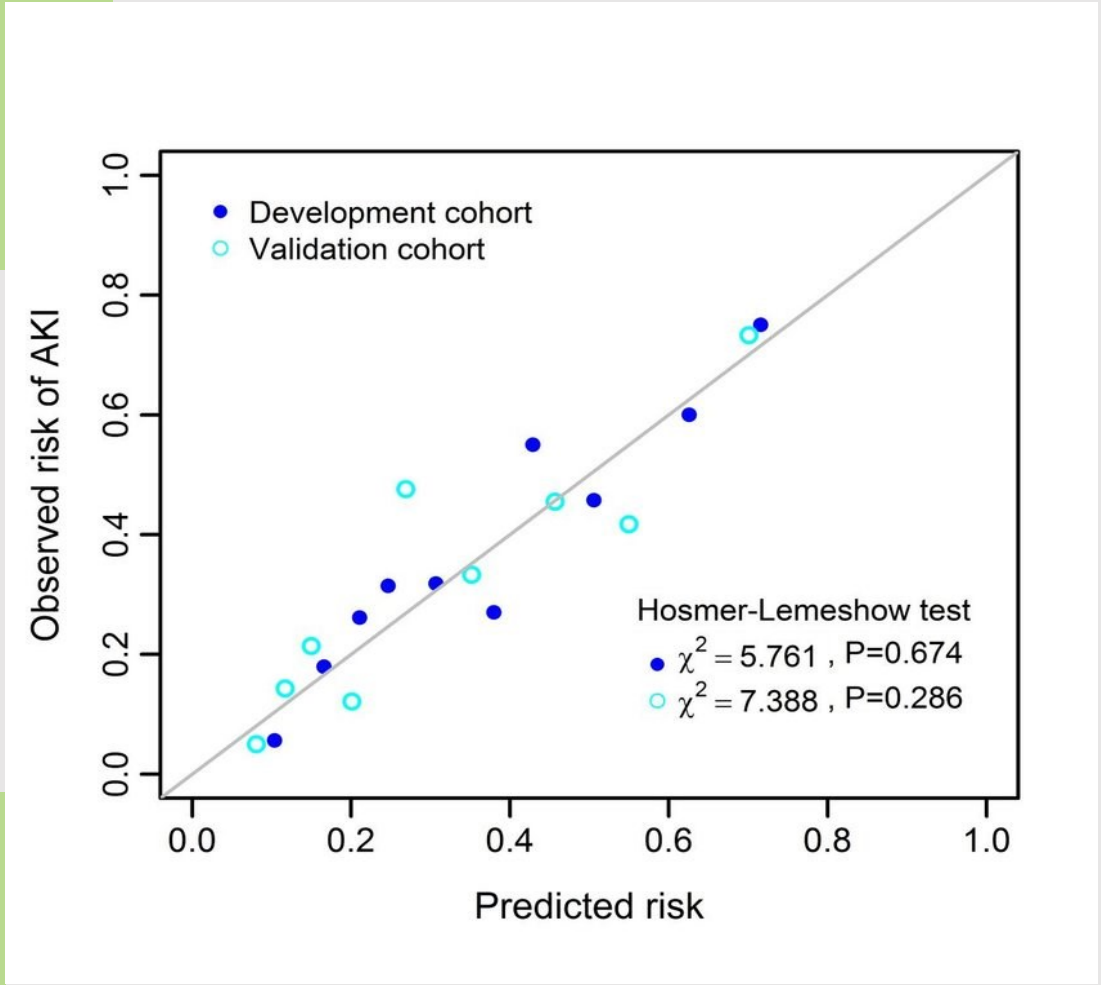
To calculate how many "Y=1" observations we would expect, the Hosmer-Lemeshow test takes the *average* of the predicted probabilities in the i-th group, and multiplies this by the number of observations in the group.

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.

Calibration *in the large*:

Level 1: ***Mean* calibration** (calibration in the large)

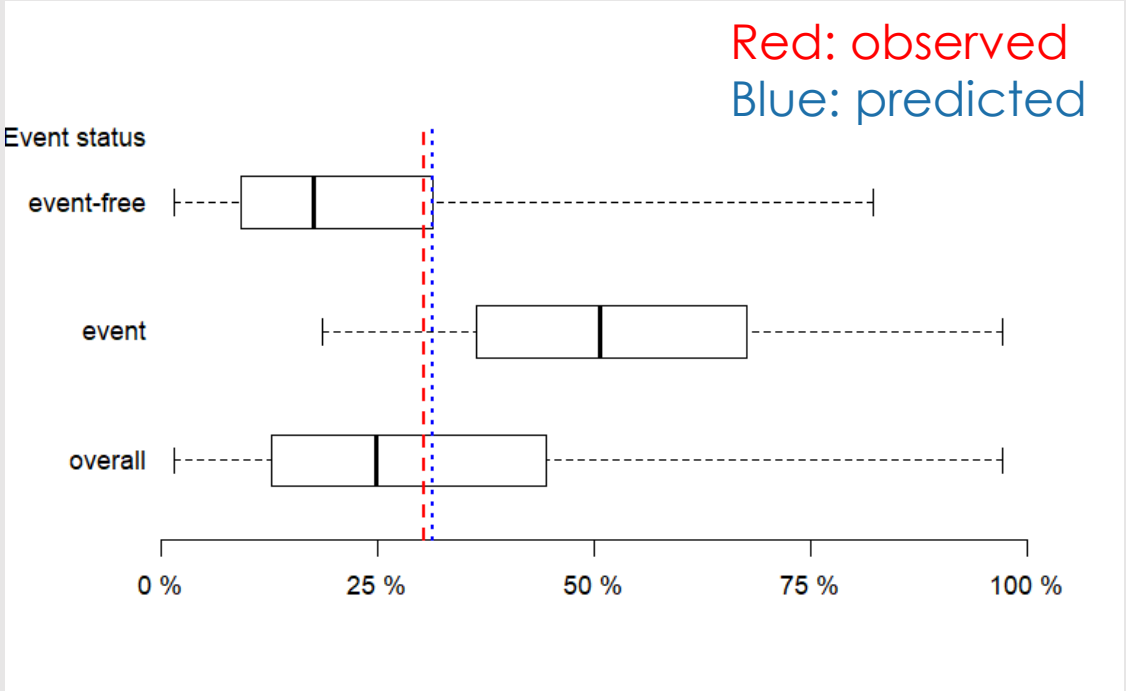
Mean estimated risk = observed proportion of event

"*On average, risks are not over-or underestimated.*"

Compare event rate with **average predicted** risk.

O:E ratio of observed events / expected events = 1

If violated *adjust* the intercept of the model.



Logistic calibration model: $\log\left(\frac{\pi}{1-\pi}\right) = a + b * LP$

LP: Linear Predictor

Taking fixet ad 1 the slope: $b = 1$

Estimate the calibration intercept: $[a|b = 1]$ideally ≈ 0 $\log\left(\frac{\pi}{1-\pi}\right) = \hat{a} + offset(LP)$

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

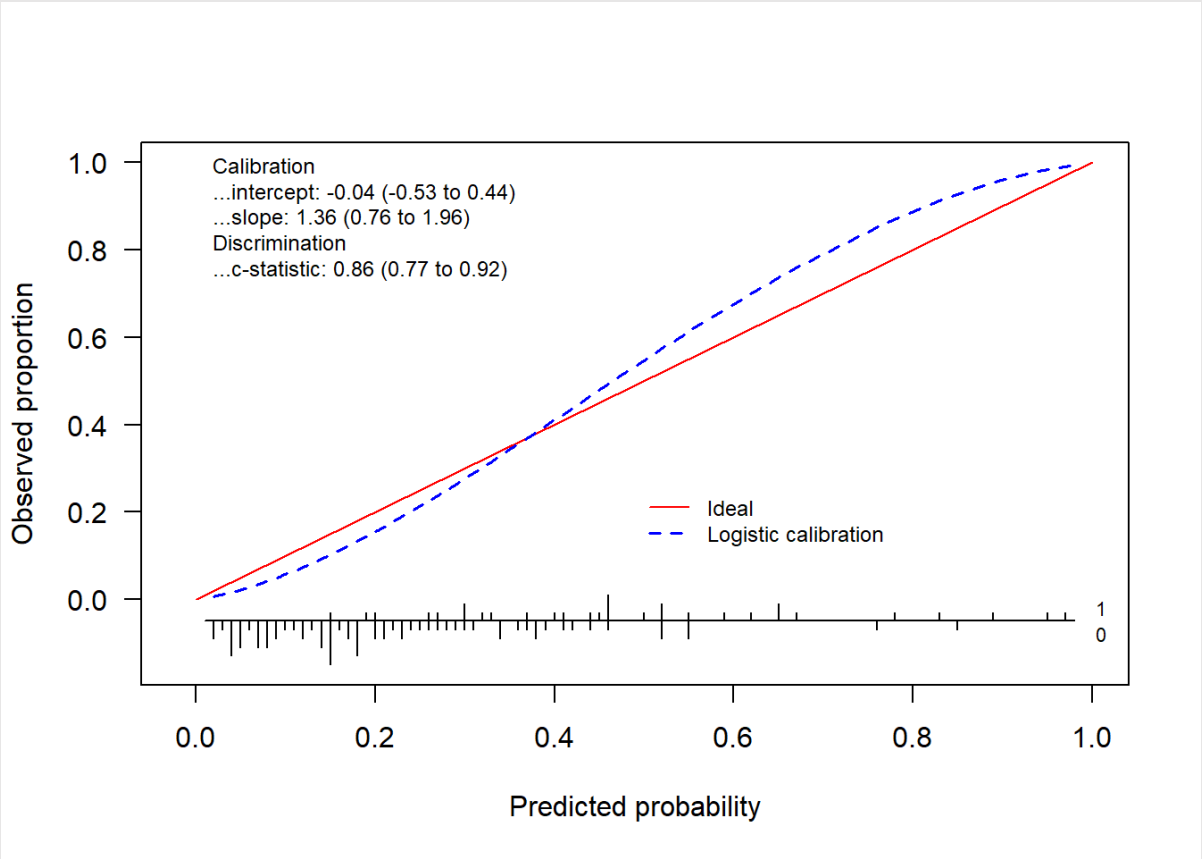
Level 2: *Weak* calibration

*"On **average**, risks are not over-or underestimated, nor too extreme/modest."*

No **systematic** over- or under-fitting

Logistic calibration model: $\log\left(\frac{\pi}{1-\pi}\right) = a + b * LP$

Linear Predictor



Estimate the calibration slope: $\hat{b}$

Estimate the calibration intercept: $\hat{a}$

Then *adjust* estimated probabilities using: $\hat{a} + \hat{b} * LP$

(the slope is called the **shrinkage** factor)

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

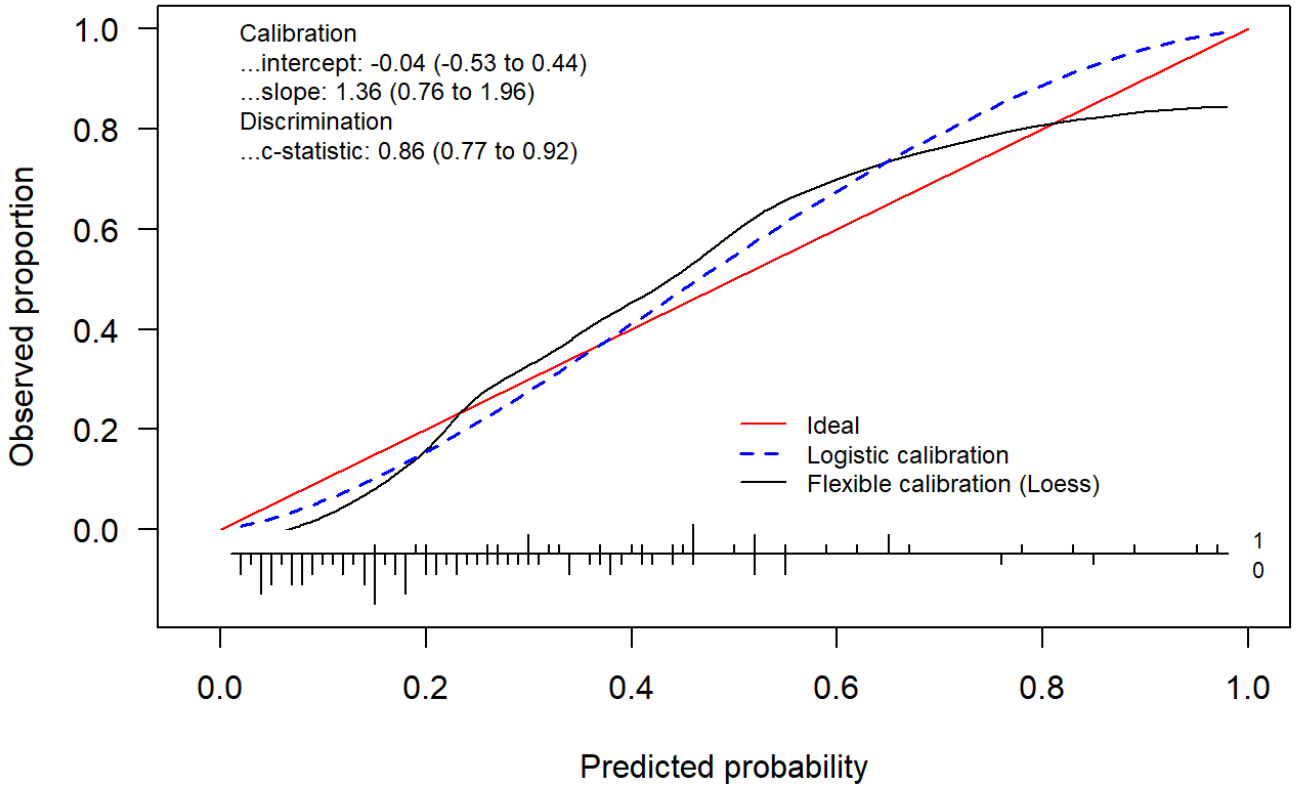
Level 3: ***Moderate*** **calibration**

"Among patients with estimated risk xx, the proportion of events is xx."

Use *calibration plots* (density/loess/splines…)

Note that the **flexible** calibration curve is more sensible to deviations, with respect to the logistic regression approach, especially at the ***extremes*** of the distribution.

But, it does not give us a ***numerical summary*** of calibration, it is sensible to the smoothing method used and it does not take into account the number of subjects in each *bin* of the smoothing function.



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

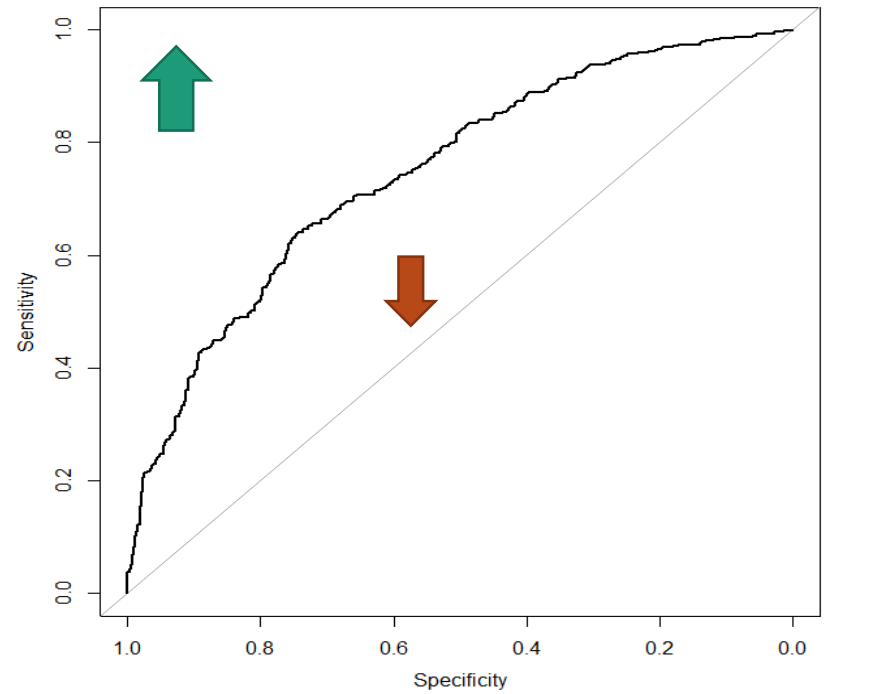
Specificity=d/b+d

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.

Assessing the performance of prediction models: a framework for some traditional and novel measures
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3575184/

Sample Sizes for Various Response Variables (basic indications): event per variable [EPV]

Type of Response Variable	Limiting Sample Size m
Continuous	n (total sample size)
Binary	$\min(n_1, n_2)$ [h]
Failure (survival) time	number of failures [j]

A fitted regression model is likely to be **reliable** when the **number of predictors** (or candidate predictors if using variable selection) ***p*** is less than *m*/10 or *m*/20, where *m* is the "limiting sample size".

A good average requirement is $p < m/15$

When a model is fitted that is ***too complex*** (i.e. **too many parameters** to estimate for the *amount of information* in the data), the goodness of fit of the model will be exaggerated and future observed values will not agree with predicted values.

In this situation, **overfitting** is said to be present, and some of the findings of the analysis come from fitting noise and not just a signal, or finding **spurious** associations between X (independent variables) and Y (outcome).

h: n1 and n2 are the marginal frequencies of the two response levels.

j: failures: events in the survival jargon

Of note: the number of non-intercept parameters in the model is usually > number of variables

Categorical variables, nonlinear terms require >1 parameters to be estimated and included in the model

1 categorical variable with 4 categories : 3 parameters

$$EPV \equiv EPP = \frac{\#Events}{candidate\ predictors\ parameters}$$

…but… why one rule ? Sample size should be tailored to the problem!

Here we focus on a more complex approach than EPV, based on minimizing the **expected overfitting** and ensuring precise parameter estimation.

Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement

G S Collins*,[1], J B Reitsma[2], D G Altman[1] and K G M Moons[2]

Sample Size

Item 8. Explain how the study size was arrived at. [D;V]

What do we want?

Development

We want to have a large enough sample size to develop a model that predicst as accurately as we can.

Validation

We want to have a large enough sample size to accurately and precisely estimate model performance. This is to be intended as **external validation**.

Of note:

- Use as much data as possible to develop your model… [cross-validation/boostrap to ***internally*** evaluate optimism]

- Avoid (randomly) ***single-splitting*** your data ***to develop and then validate*** your model*
  - Reduces development sample size (overfitting)
  - Reduces validation sample size (inadequate to evaluate model performance)

Much better **external** validation (different place/time…)

Medical data : often low-moderate sample size!

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Recent **guidelines** have been proposed in the biostatistical community:

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]

Summary:

Calculate sample size that is needed to:

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

Requires calculations for **multiple** criterion

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Development:

Calculate sample size that is needed to:

- Minimize potential overfitting
- Estimate parameters precisely

A series of **closed form solutions** compute the required sample size to precisely estimate key performance measures:

Continuous outcomes

- A ***shrinkage*** factor >=0.9 (***calibration slope***)
- A small difference (<=0.05) in $\boldsymbol{R^2}$ apparent vs adjusted
- Precise estimation of the residual standard deviation
- Precise estimation of the average outcome

Binary/Time to event outcomes

- A **shrinkage** factor >=0.9 (***calibration slope***)
- A small difference (<=0.05) in Nagelkerke's $\boldsymbol{R^2}$ apparent vs adjusted
- A margin of error <=0.05 in *overall* risk estimate
- A certain level of the AUC(>= 0.80)

Parameters required in input

Cox-Snell R^2 may be small…

For example, for a logistic regression model with an outcome proportion of:

- 50% the max Cox-Snell R^2 is 0.75
- 5% the max Cox-Snell R^2 is 0.33
- 1% the max Cox-Snell R^2 is 0.11

What about *No-existing-model* thing?

When there is no existing model for a particular research question (**rare!**) take into account that healthcare outcomes are generally **low** signal:noise ratio.

Assume a **low** R^2 [i.e. : between 15% and 20%]

Last but not least: **timing** of data collection vs sample size calculation

BEFORE (primary data source) :

- If you can anticipate the expected sample size and proportion of events, then
  You can ***limit*** the number of variables you will collect

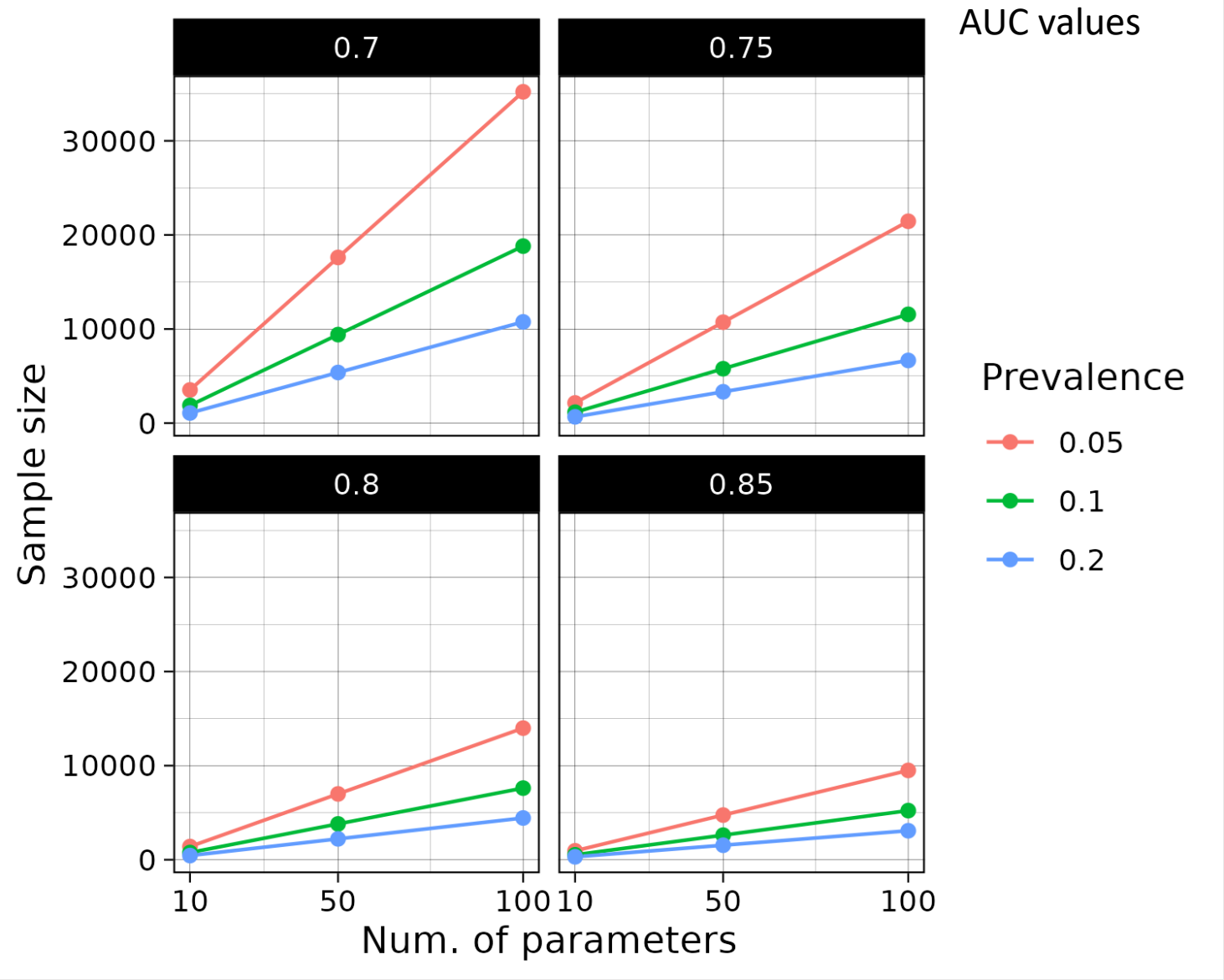
- If you know a priori how many predictors you wanto to examine, then
  You will need to collect a ***suitably sized*** sample

AFTER (secondary data source) :

- Your sample size and number of events are fixed

- You can then ***restrict*** the number of variables (and complexity) you will include in the modelling

Binary outcome



Prevalence/Incidence ↓ Sample Size

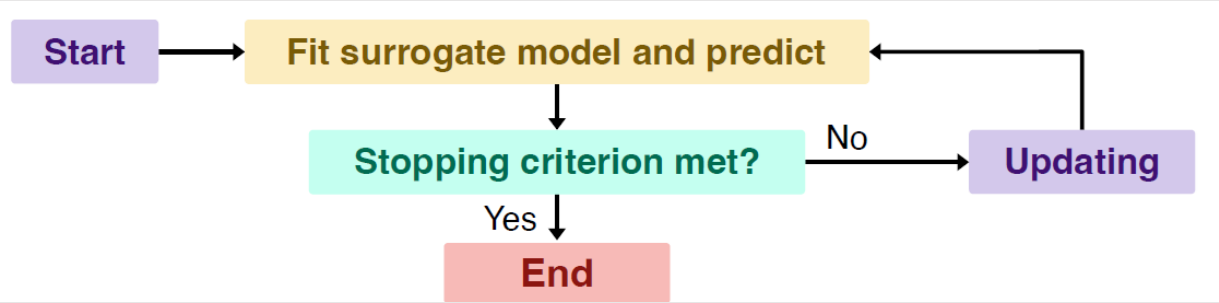
parameters to be estimated ↑ Sample Size

AUC/R^2 [like larger effects size in hypothesis tests] ↓ Sample Size

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Existing tools can estimate minimum samples for continuous, binary, and survival outcomes [“***standard***” statistical tools]

Work is in progress in developing *simulation-based* approaches that works with *any* outcome or method [!ML algorithms!].



The pmsims package for R

Flexible	Any model or data type
User-friendly	Defaults for common scenarios
Efficient	Estimation via surrogate modelling

Ewan Carr, Gordon Forbes, Diana Shamsutdinova, Daniel Stahl & Felix Zimmer

Department of Biostatistics & Health Informatics
King’s College London

https://github.com/ewancarr/pmsims-iscb

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute