Píndoles estadístiques UEB-VHIR

Busqueu la fama, i aquí és on aneu a començar a pagar: Estratègies per a la construcció de models i biomarcadors

Santiago Pérez-Hoyos

Unitat d'Estadística i Bioinformàtica

Divendres 28 de febrer de 12:30 a 13:30
Sala d'Actes de Traumatologia i Rehabilitació

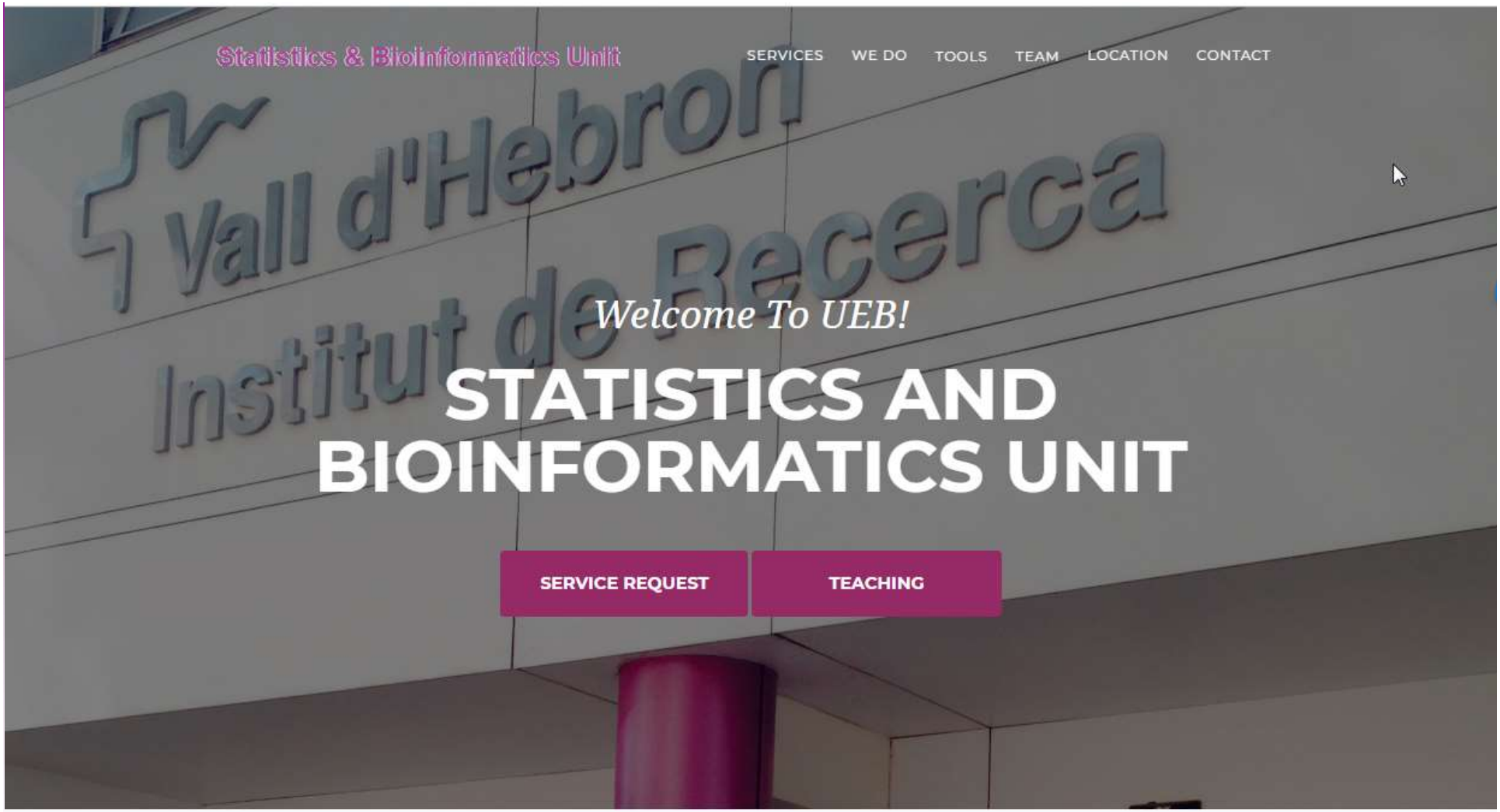
Les píndoles estadístiques son sessions divulgatives, organitzades per la Unitat d'Estadística i Bioinformàtica (UEB) del VHIR, on es presenten problemes i solucions estadístiques dirigides als professionals interessats del Campus Vall d'Hebron

Vall d'Hebron Institut de Recerca VHIR

ueb.vhir.org

GRBIO

UNIVERSITAT DE BARCELONA

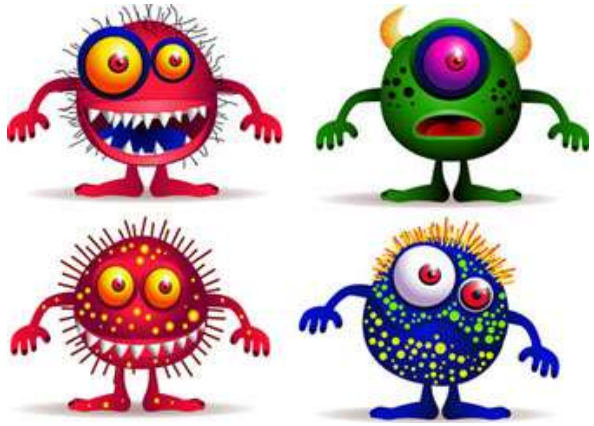

Statistics & Bioinformatics Unit
SERVICES
WE DO
TOOLS
TEAM
LOCATION
CONTACT
Vall d'Hebron
Institut de Recerca
Welcome To UEB!
STATISTICS AND
BIOINFORMATICS UNIT
SERVICE REQUEST
TEACHING

ueb.vhir.org

Busqueu la fama, i aquí és on aneu a començar a pagar, amb suor

Estratègies per a la construcció de models i biomarcadors

28 de Febrer . 12:30 h



Biomarkers are everywhere and almost everybody looks for one

Feb '20
20

Vall d'Hebron obté més d'1M€ de benefici per l'explotació de la llicència d'una patent

Jan '20
29

La càrrega microbiana és un marcador de resposta al trasplantament de microbiota fecal en pacients amb malaltia de Crohn

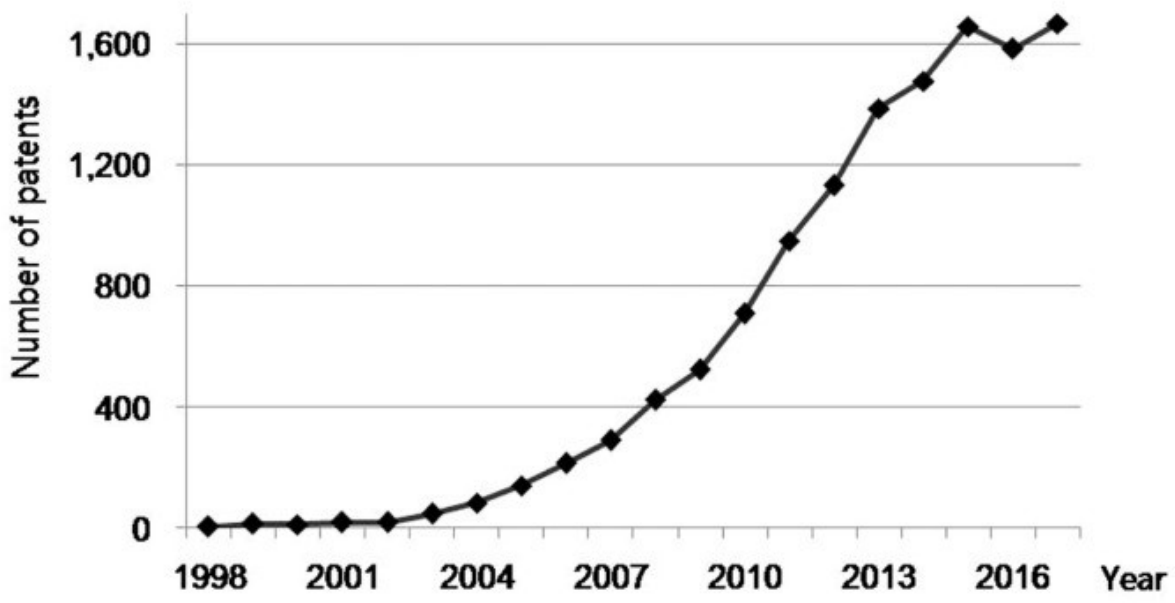
Jan '20
22

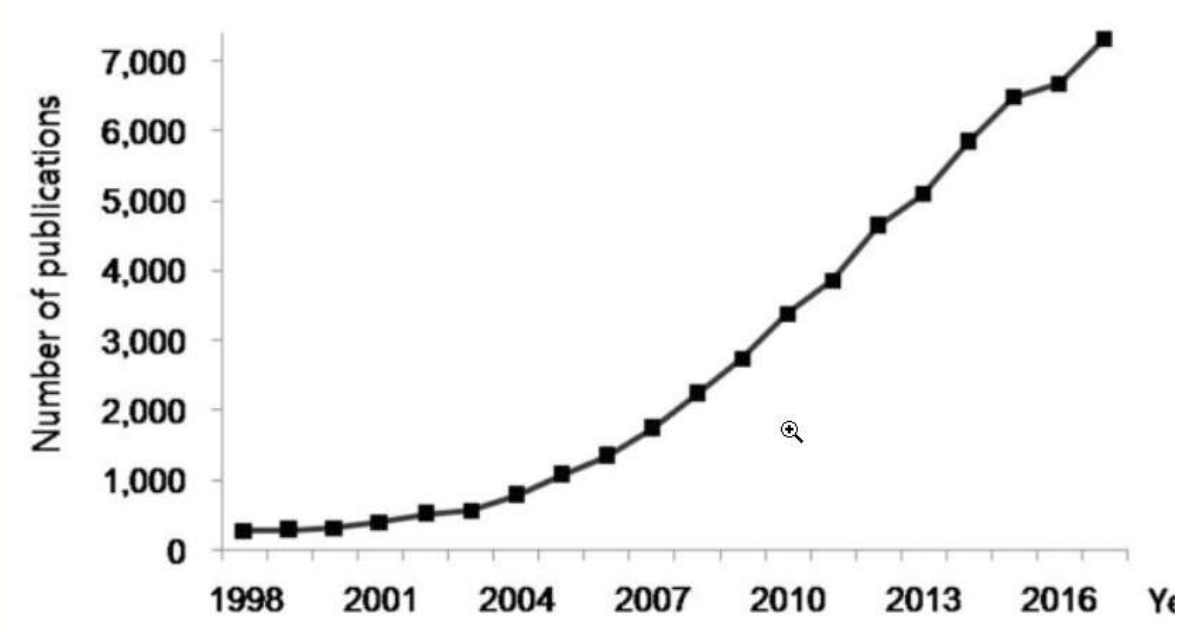
Millora en la identificació de biomarcadors mitjançant l'estandardització de l'ús de mostres de teixit humà

Dec '19
23

Researchers identify a new tumor biomarker in endometrial, lung, and colorectal cancer

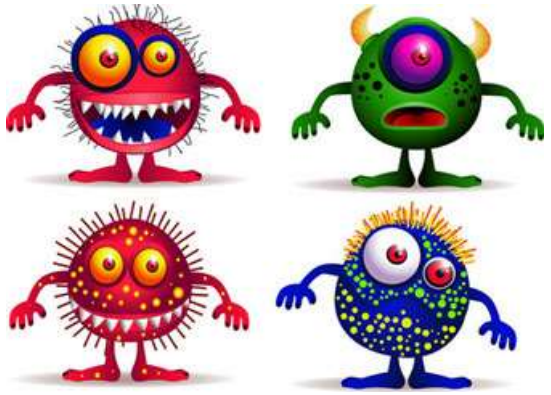
Publications with biomarker in the title and Biomarker patents by year





Biomarkers and Disease

- Natural history
- Risk prediction
- Phenotype definition
- Clinical and biological heterogeneity
- Diagnostic or screening tests
- Response to treatment
- Prognosis



Biomarker Discovery and Validation Steps

- **Correlation**: a biomarker vs a disease or status of a disease
- **Calculation**: Build statistical model to express relation
- **Reproducibility**: Test-retest, Concordance index
- **Discrimination**: Sensitivity and Specificity, Diagnostic Measures AUC
- **Validation**: **Internal** (Goodness of fit, performance Scores, Bootstrap, Cross-Validation)
- **Validation**: **External** (Training/Test, performance Scores, Prognosis, AUC)

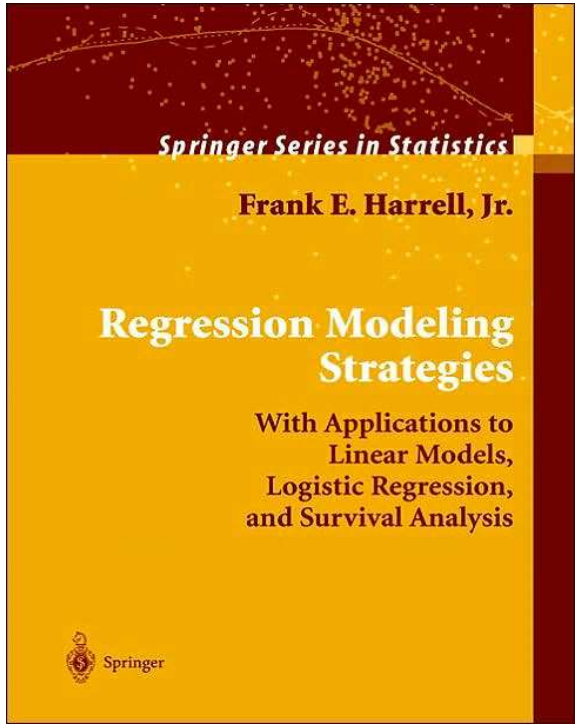
Some bibliography

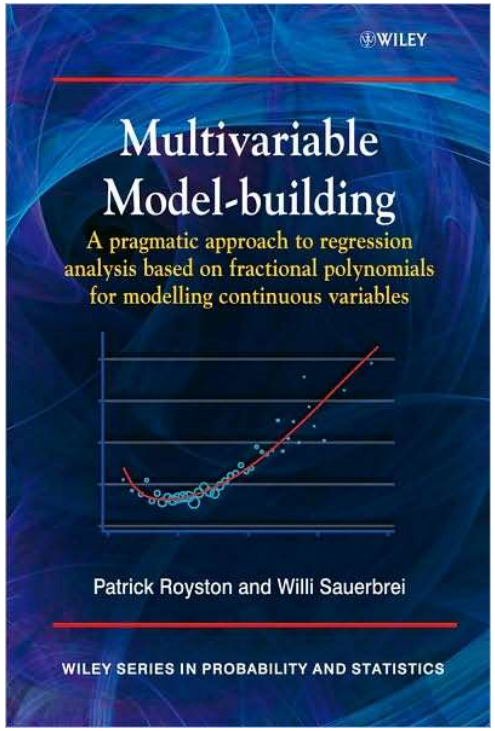
Statistics in Medicine 1996

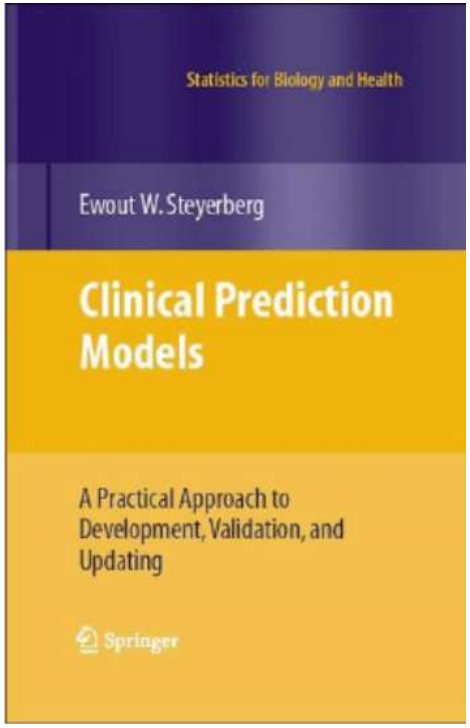
TUTORIAL IN BIOSTATISTICS

MULTIVARIABLE PROGNOSTIC MODELS: ISSUES IN DEVELOPING MODELS, EVALUATING ASSUMPTIONS AND ADEQUACY, AND MEASURING AND REDUCING ERRORS

FRANK E. HARRELL Jr., KERRY L. LEE AND DANIEL B. MARK







Some papers to read

Rev Esp Cardiol. 2011;64(6):501–507

Focus on: Contemporary Methods in Biostatistics (I)

Regression Modeling Strategies

Eduardo Núñez,[a,b,*] Ewout W. Steyerberg,[c] and Julio Núñez[a]

[a] Servicio de Cardiología, Hospital Clínico Universitario, INCLIVA, Universitat de Valencia, Spain
[b] Cuore International, Reading, Pennsylvania, United States
[c] Department of Public Health, Erasmus MC, Rotterdam, The Netherlands

Received: 18 April 2017 | Revised: 13 November 2017 | Accepted: 17 November 2017
DOI: 10.1002/bimj.201700067

REVIEW ARTICLE

Biometrical Journal

Variable selection – A review and recommendations for the practicing statistician

Georg Heinze | Christine Wallisch | Daniela Dunkler

BIOSTATISTICS & EPIDEMIOLOGY
https://doi.org/10.1080/24709360.2019.1618653

Taylor & Fra
Taylor & Francis Grou

Check for

Statistical modeling methods: challenges and strategies

Steven S. Henley[a,b,c], Richard M. Golden[d] and T. Michael Kashner[a,b,e]

Smith J Big Data (2018) 5:32
https://doi.org/10.1186/s40537-018-0143-6

Journal of Big Data

SHORT REPORT Open Access

CrossMark

Step away from stepwise

Gary Smith*

Transplant International

REVIEW

Five myths about variable selection

Georg Heinze & Daniela Dunkler

Transplant International 2017; 30: 6–10

Some guidelines and recommendations

Topic Group 2: Selection of variables and functional forms in multivariable analysis

Chairs: **Georg Heinze, Aris Perperoglou, Willi Sauerbrei**
Members: **Michal Abrahamowicz, Heiko Becher, Harald Binder, Daniela Dunkler, Rolf Groenwold, Frank Harrell, Nadja Klein, Geraldine Rauch, Patrick Royston, Matthias Schmid**

equator network

Enhancing the QUAlity and Transparency Of health Research

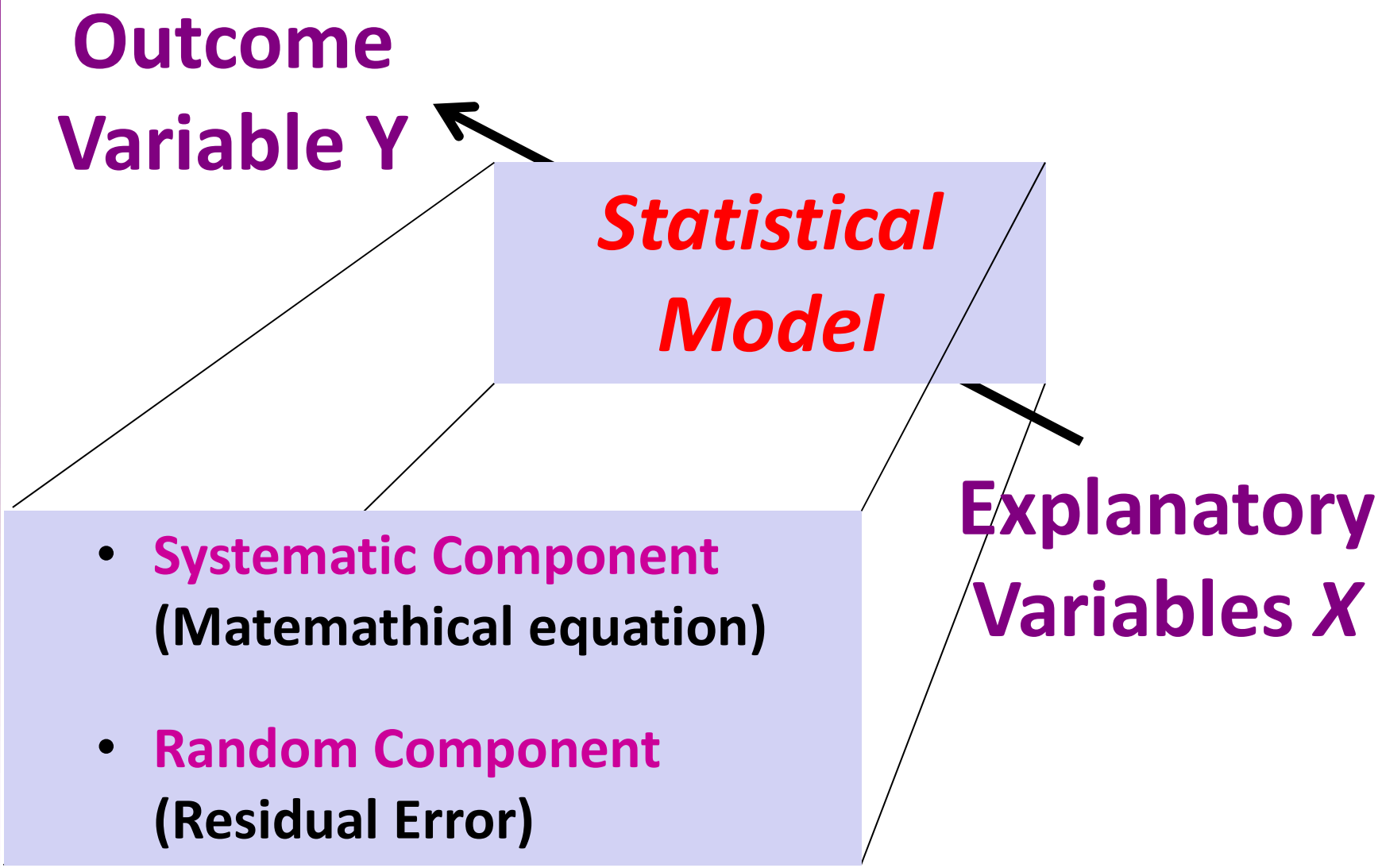
Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement

Reporting guideline provided for? (i.e. exactly what the authors state in the paper)	Reporting of studies developing, validating, or updating a prediction model, whether for diagnostic or prognostic purposes. TRIPOD Checklist for Prediction Model Development: Word | PDF TRIPOD Checklist for Prediction Model Validation: Word | PDF TRIPOD Checklist for Prediction Model Development and Validation: Word | PDF

REporting recommendations for tumour MARKer prognostic studies (REMARK)

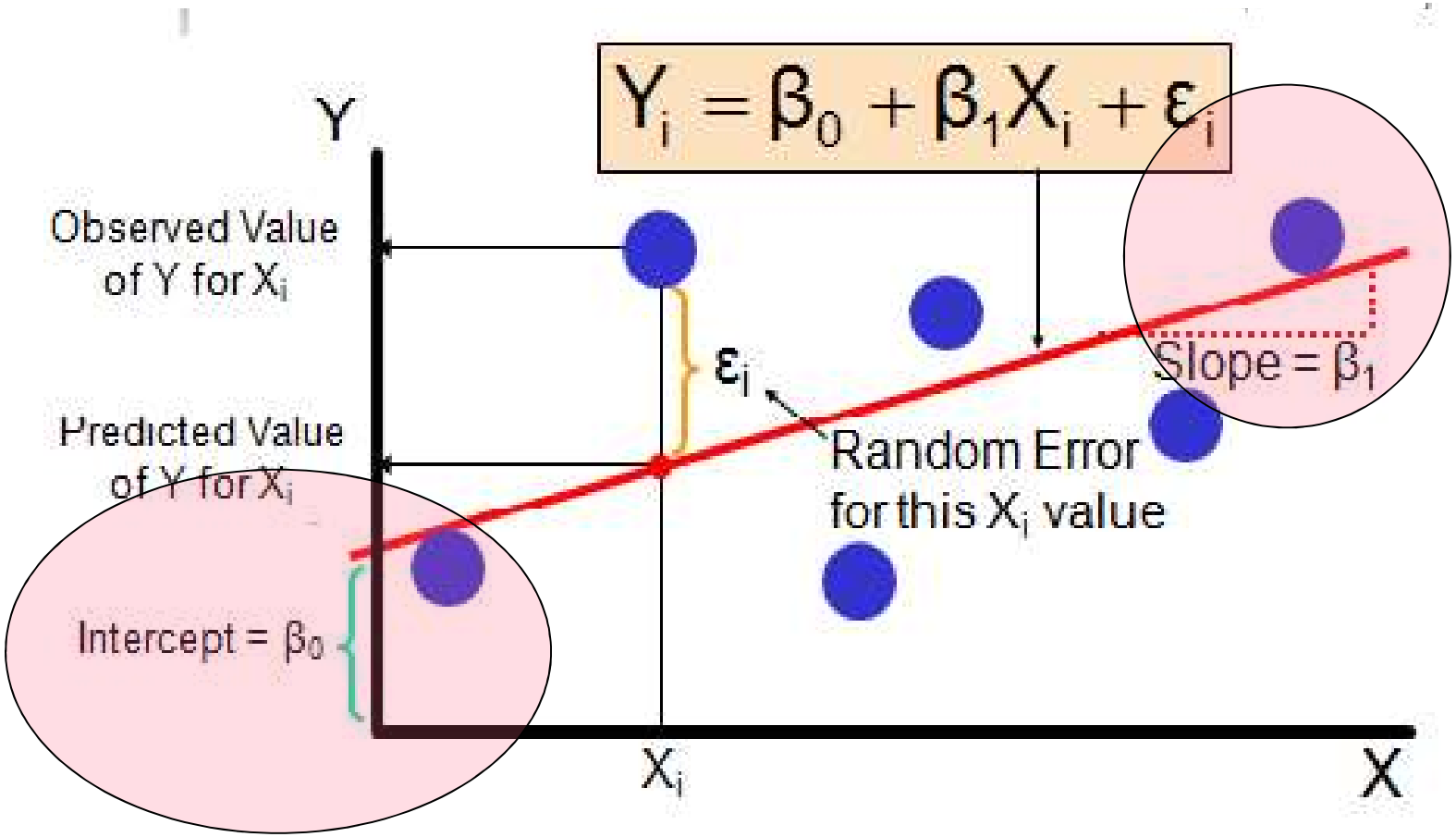
Reporting guideline provided for? (i.e. exactly what the authors state in the paper)	Tumour marker prognostic studies
Full bibliographic reference	McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM. REporting recommendations for tumour MARKer prognostic studies (REMARK). Br J Cancer. 2005;93(4):387-391.
Language	English

Statistical Modellling



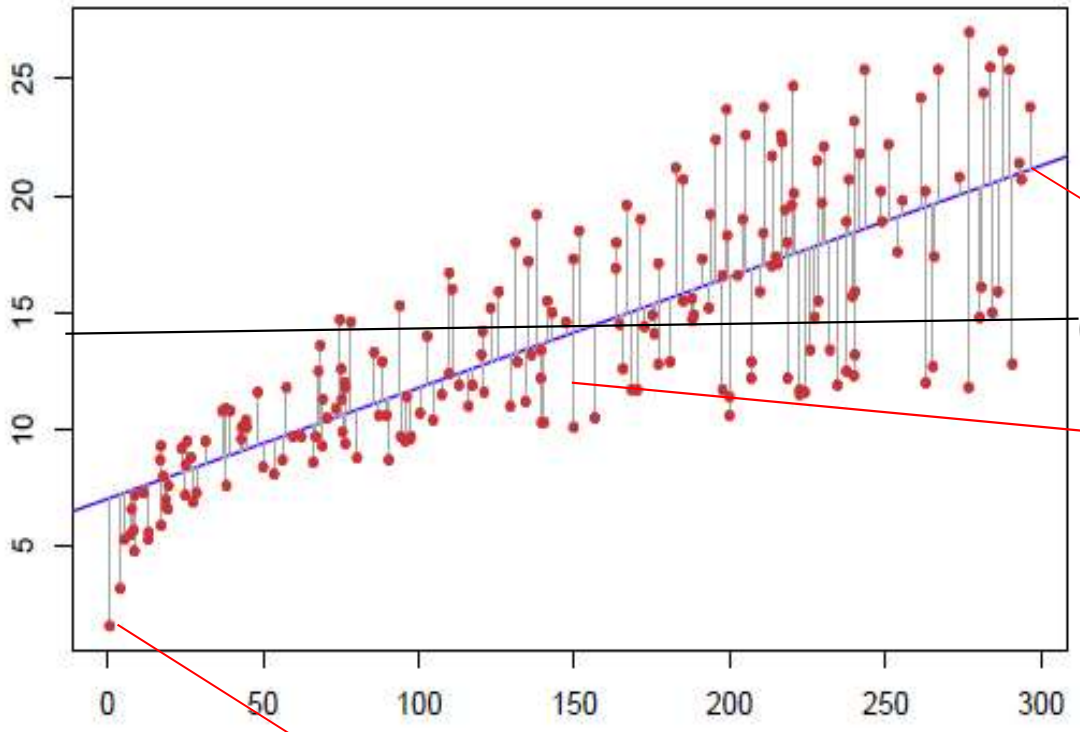
Simple linear model

Response variable=explanatory variables +error



Estimation of the parameters by least squares

R^2= Explained variability (“Performance mesure”)



Predicted value

($\bar{y}$)

Residuals(e)

observed value (Y)

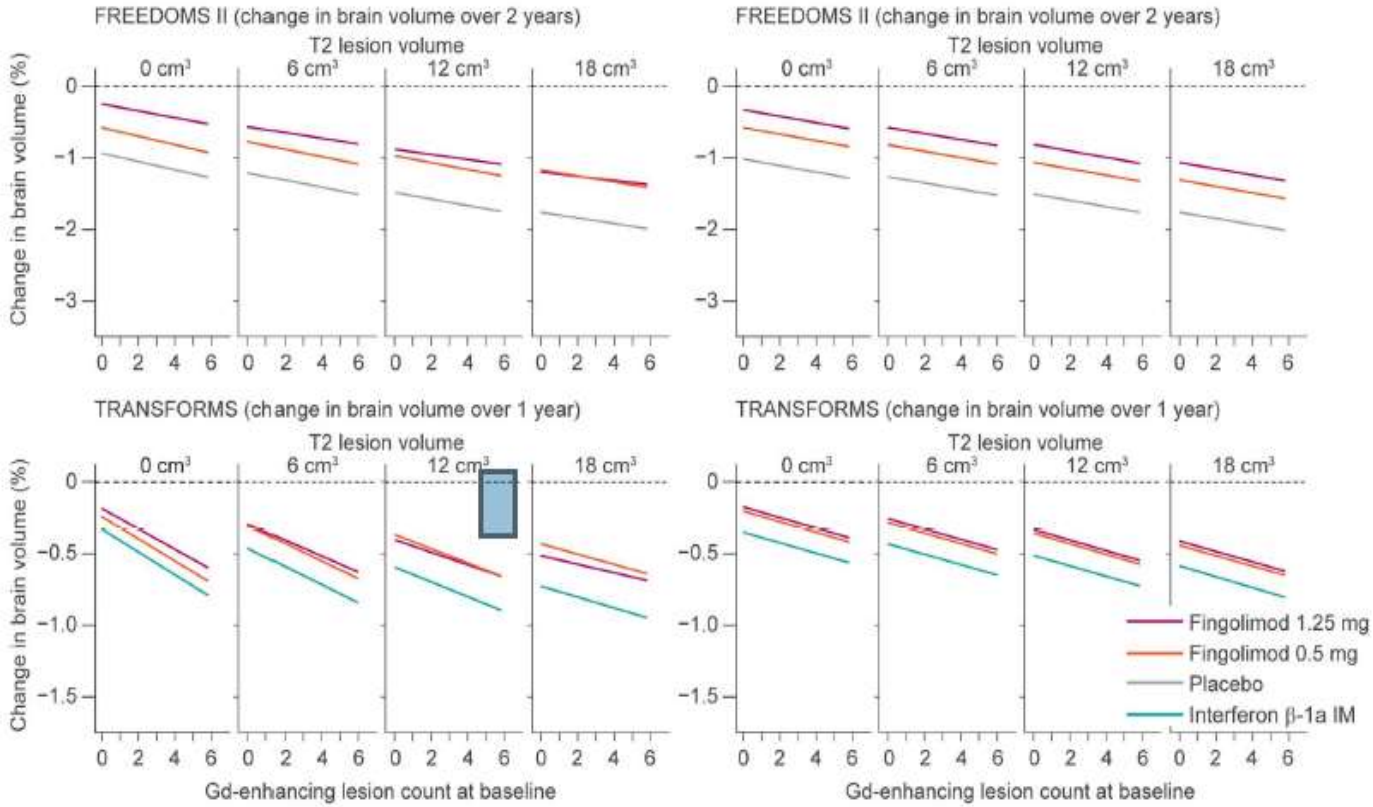
Example

Correlation between brain volume loss and clinical and MRI outcomes in multiple sclerosis

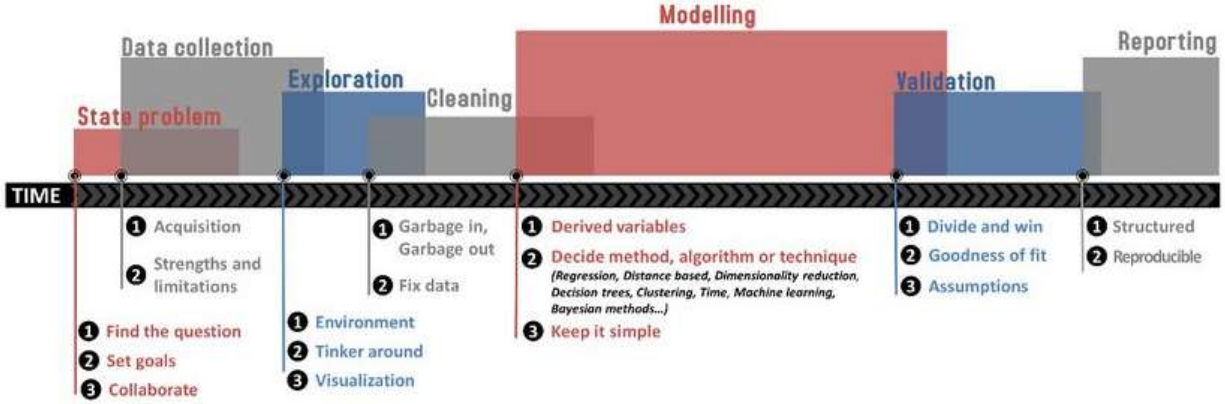
Neurology®

February 24, 2015; 84 (8)

Ernst-Wilhelm Radue, Frederik Barkhof, Ludwig Kappos, Till Sprenger, Dieter A. Häring, Ana de Vera, Philipp von Rosenstiel, Jeremy R. Bright, Gordon Francis, Jeffrey A. Cohen

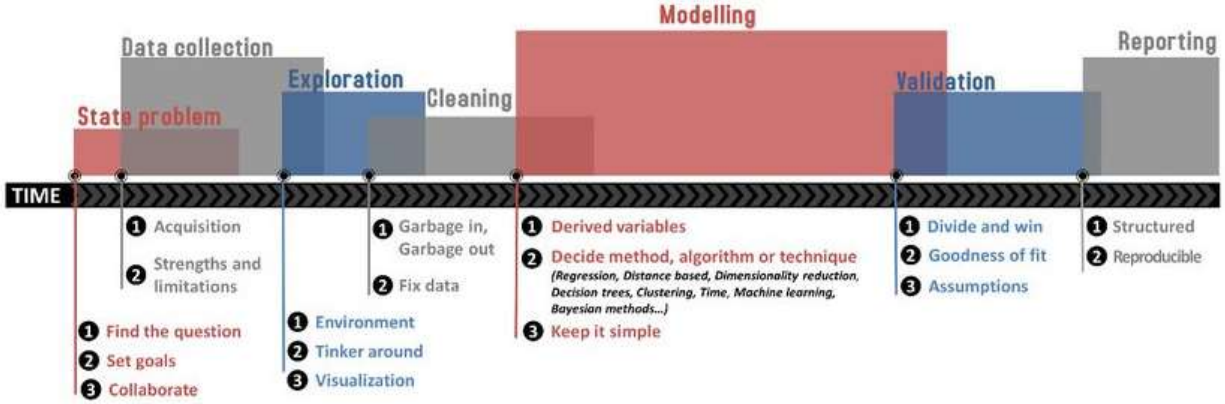


Modelling Steps



1. **Determining the aim of the model** (Prediction or explanatory)
2. **Ascertainment of the outcome** (binary, continuous,survival, repeated measure, etc.) and data
3. Choosing the **appropriate statistical model** based on design and outcome
4. **Model Estimation** (Missing values,Functional shape, Selection variables,shrinkage , avoid overfitting)
5. **Assessing model performance**
6. **Model Validation** (internal and external)
7. **Presenting Results**

Modelling Steps



1. **Determining the aim of the model** (Prediction or explanatory)
2. **Ascertainment of the outcome** (binary, continuous,survival, repeated measure, etc.) and data
3. Choosing the **appropriate statistical model** based on design and outcome
4. **Model Estimation** (Missing values,Functional shape, Selection variables,shrinkage , avoid overfitting)
5. **Assessing model performance**
6. **Model Validation** (internal and external)
7. **Presenting Results**

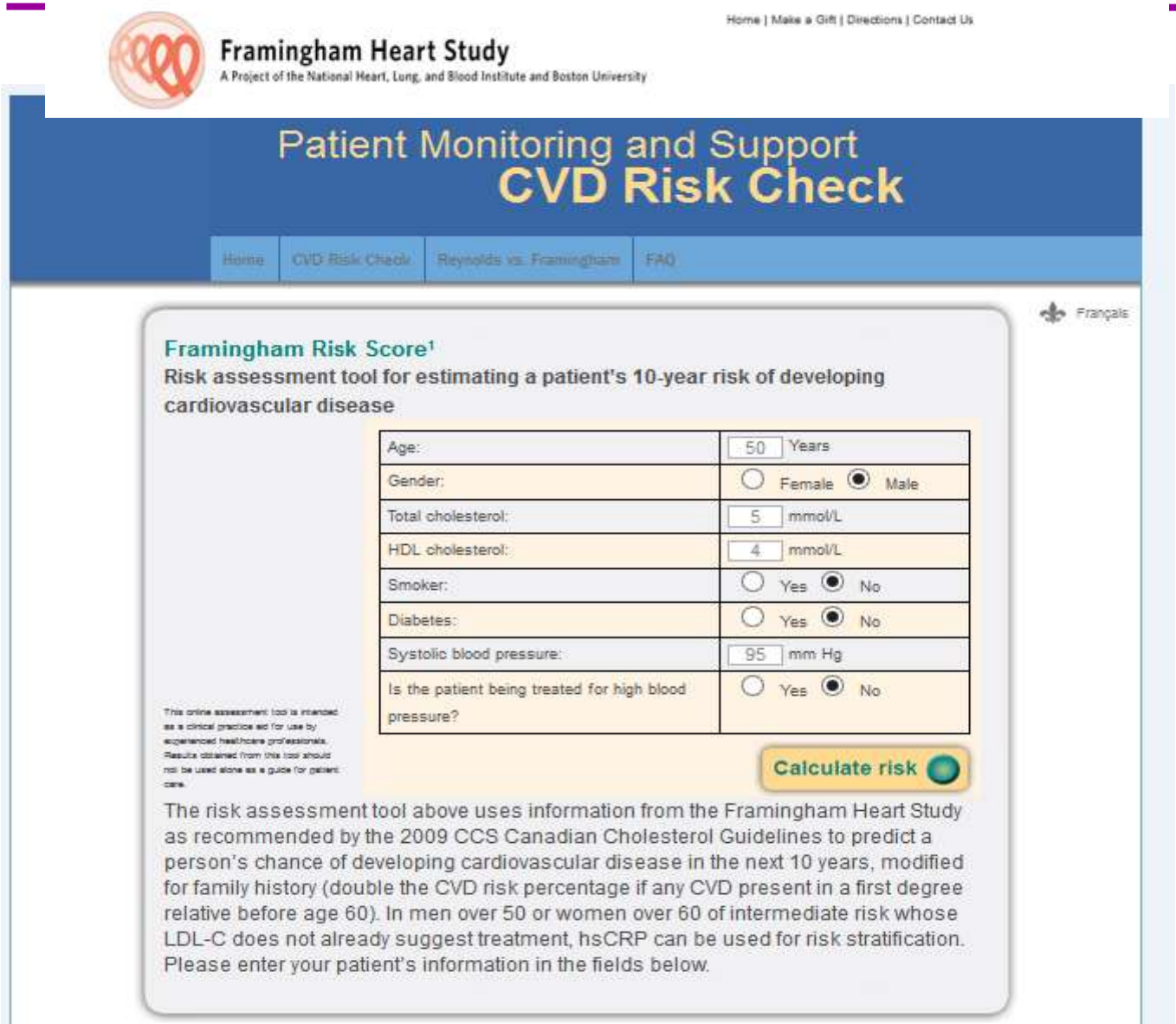
Aim of the model (I)

Predictive models

To calculate the **probability of the outcome** in each subject, often beyond the the data from which was calculated (eg Framingham calculator)

- Balance **complexity** versus **parsimony**
- Incoporate as much **accurate data** as possible
- **Impute data** if missing as sample size is important
- Specifify the complexity of the variables
- **Limit the number of interactions** (only prestablished)
- For binary endpoints at least **10-14 events per variable**
- Aware of problems with selection strategies
- **Use prior knowledge**
- **Validate the final model** for calibration and discrimination

Example



https://www.cvdriskchecksecure.com/FraminghamRiskScore.aspx

Aim of the model (I)

Model for Effect Estimation

Tools for **effect estimation** (risk factors, epidemiology, prognosis) , understanding the effect of predictors or as basis for hypothesis contrast

- Balance between **accurate estimation** and **complex modelling**
- **Minimize categorization** of continous variables.
- Keep in mind the number of events per variable
- **Group predictors** or reduce covariates (PCA)
- Model **non linear** relationships
- Test for interaccions
- **Multiple imputation** if missingness
- Use of propensity scores,
- Internal validation is a plus

Intensive care admission and hospital mortality in the elderly after non-cardiac surgery

Med Intensiva. 2018;42(8):463–472

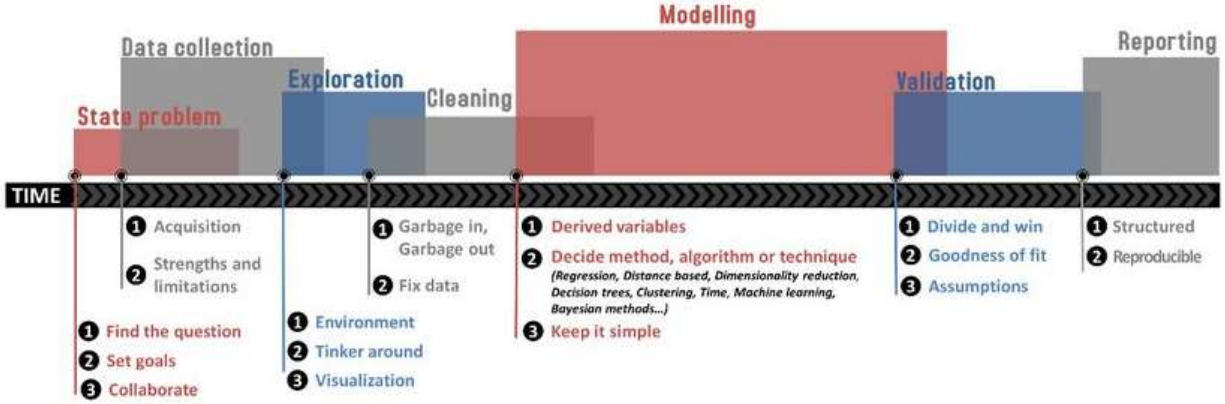
M. de Nadal[a,*], S. Pérez-Hoyos[b], J.C. Montejo-González[c], R. Pearse[d], C. Aldecoa[e], on behalf of the European Surgical Outcomes Study (EuSOS) in Spain[1]

Table 4 Adjusted logistic regression for hospital mortality.

	Hospital mortality	Crude OR (95% CI)	Adjusted OR (95% CI)	*P* value
ASA score				
I	45 (4.0%)	1	1	<0.001
II	70 (2.6%)	0.61 (0.42–0.90)	0.59 (0.39–0.90)	
III	53 (3.8%)	0.93 (0.62–1.40)	0.71 (0.42–1.20)	
IV	32 (14.0%)	3.86 (2.39–6.23)	2.20 (1.14–4.26)	
V	8 (72.7%)	54.91 (13.75–219.36)	17.81 (3.80–83.59)	
Urgency of surgery				
Elective	139 (3.1%)	1	1	<0.001
Urgent	48 (5.5%)	1.85 (1.32–2.59)	1.44 (0.99–2.09)	
Emergency	21 (18.6%)	6.82 (4.09–11.39)	3.31 (1.77–6.19)	
Surgical speciality				
Orthopaedics	36 (2.5%)	1	1	<0.001
Breast	4 (2.4%)	0.97 (0.34–2.76)	1.30 (0.45–3.75)	
Gynaecology	26 (5.9%)	2.41 (1.44–4.03)	3.29 (1.88–5.76)	
Vascular	20 (7.0%)	2.91 (1.66–5.11)	1.78 (0.95–3.33)	
Upper gastrointestinal	20 (8.4%)	3.53 (2.00–6.20)	3.45 (1.88–6.33)	
Lower gastrointestinal	33 (5%)	2.02 (1.25–3.26)	1.57 (0.94–2.64)	
Hepato-biliary	14 (4.5%)	1.81 (0.96–3.39)	1.91 (0.96–3.80)	
Plastic	5 (2.2%)	0.85 (0.33–2.20)	0.96 (0.37–2.51)	
Urology	20 (3.2%)	1.28 (0.74–2.23)	1.46 (0.81–2.62)	
Kidney	1 (2%)	0.79 (0.11–5.87)	0.70 (0.09–5.51)	
Head and neck	17 (2.6%)	1.04 (0.58–1.87)	1.24 (0.68–2.27)	
Other	9 (2.8%)	1.10 (0.52–2.30)	1.04 (0.48–2.26)	
Diabetes insulin-dependent	22 (8.7%)	1.02 (0.64–1.61)	2.08 (1.22–3.53)	0.007

Data are number (percentage column) and OR (odds ratio) and CI (confidence interval). ASA = American Society of Anaesthesiologists.

Modelling Steps



1. **Determining the aim of the model** (Prediction or explanatory)
2. **Ascertainment of the outcome** (binary, continuous,survival, repeated measure, etc.) and dat
3. Choosing the **appropriate statistical model** based on design and outcome
4. **Model Estimation** (Missing values,Functional shape, Selection variables,shrinkage , avoid overfitting)
5. **Assessing model performance**
6. **Model Validation** (internal and external)
7. **Presenting Results**

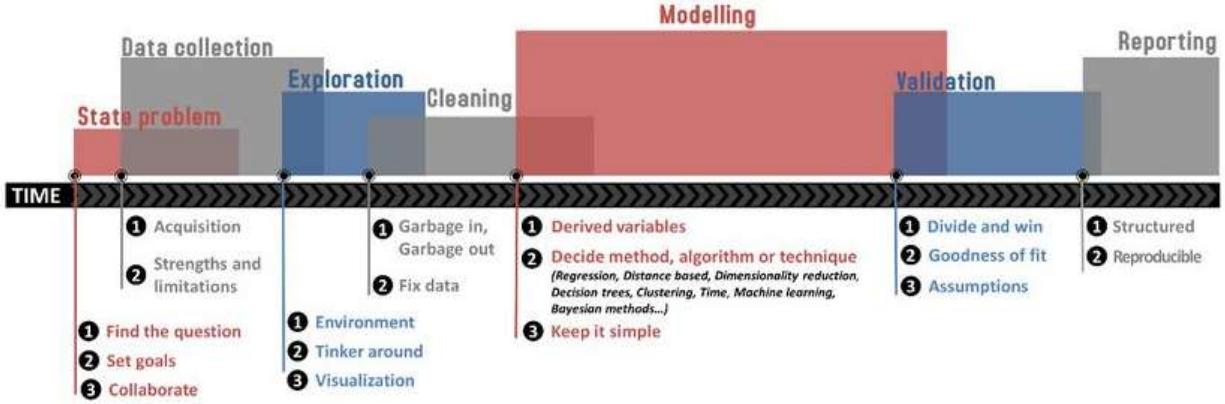
Ascertainment of the outcome

- Clinically relevant
- Missing outcomes
- Enough Follow up
- Repeated measures
- Minimize endpoint misclassification error
- Prefer hard outcomes for prognostic models
- If combined endpoint, ensure the direction of the effect is the same for both components
- Consider using new outcomes such as days alive and out of hospital , Quality of Life, etc.

Ascertainment of the data

- Most important predictor or response variables not collected
- Subjects in the dataset are not representative of the population to which inferences are needed
- Data collection sites do not represent the population of sites
- Key variables missing in large numbers of subjects
- Data not missing at random
- No operational definitions for key variables and/or measurement errors severe
- No observer variability studies done

Modelling Steps



1. **Determining the aim of the model** (Prediction or explanatory)
2. **Ascertainment of the outcome** (binary, continuous,survival, repeated measure, etc.) and data
3. Choosing the **appropriate statistical model** based on design and outcome
4. **Model Estimation** (Missing values,Functional shape, Selection variables,shrinkage , avoid overfitting)
5. **Assessing model performance**
6. **Model Validation** (internal and external)
7. **Presenting Results**

Statistical Models

Outcome Y	Explanatory X	Statistical Model
Continuous	Continuous	**Linear Regression**
Continuous	Categorical	**Anova,Linear Regression**
Dichotomous (Yes/No)	Continuous/Categorical	**Logistic Regression**
Categorical(multinomial)	Continuous/Categorical	**Polytomous or Multinomial Regression**
Count data	Continuous/Categorical	**Poisson Regression**
Time to Event	Continuous/Categorical	**Cox or Parametric Survival Regression**
Continuous	Repeated measures(time/subjects)	**Repeated, mixed or random effects models.**
Variables subset	Continuous/Categorical	**Multivariate analysis (PCA, Cluster,Machine Learning)**

Effect Measures after regression fit

Linear Regression

Intercept β_0	Mean level at baseline
Slope β_i	Change of response by increasing 1 unit explanatory i

Logistic Regression

Intercept $e^{\beta 0}$	Ratio cases/no cases (Odds) at baseline
Odds Ratio $e^{\beta i}$	Relative Odds by increasing 1 unit explanatory i

Cox Regression

No intercept Estimate $h_0(t)$	Instantaneous hazard function at baseline
Hazard Ratio $e^{\beta i}$	Relative Hazard by increasing 1 unit explanatory i

Some preliminary considerations

Linear Regression

- Assumes linearity of relation
- Categorize predictor is not usually a good option

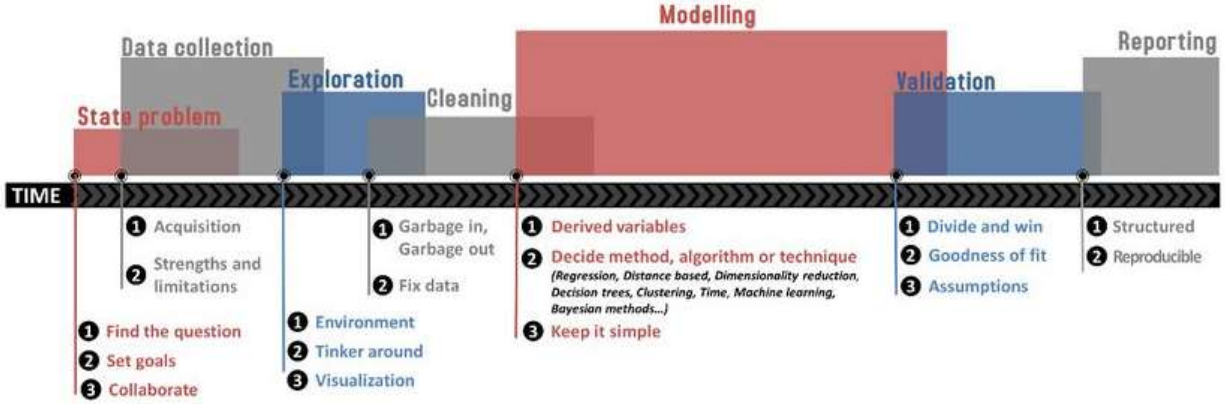
Logistic Regression

- Only need to know if outcome is present/absent
- The outcome should be clear and not product of classification of continuous variables.
- Time of follow-up should be the same and not lost of follow up
- May require 20 events per candidate predictor
- Odds Ratio is not Relative Risk

Cox Regression

- Relative hazard must be proportional
- All subjects must have a mínimum potential follow-up
- Competing Risk should be taken into account

Modelling Steps



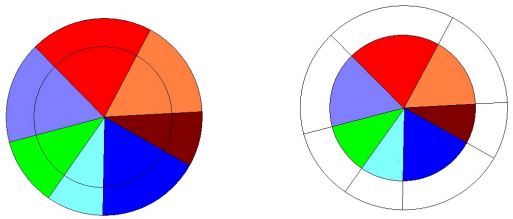
1. **Determining the aim of the model** (Prediction or explanatory)
2. **Ascertainment of the outcome** (binary, continuous,survival, repeated measure, etc.) and data
3. Choosing the **appropriate statistical model** based on design and outcome
4. **Model Estimation** (Missing values,Functional shape, Selection variables,shrinkage , avoid overfitting)
5. **Assessing model performance**
6. **Model Validation** (internal and external)
7. **Presenting Results**

Points to take into account when estimating a model

- Problems with data (Missing Data, categorization, influential observations)
- Functional shape of the relation (linearity, non linearity,etc.)
- Varible selection strategy(Automatic selection)
- Overfitting/Underfitting

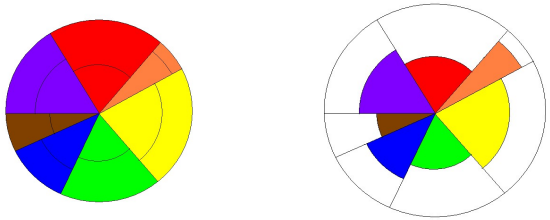
Missing Data

MCAR
(missing completely at random)



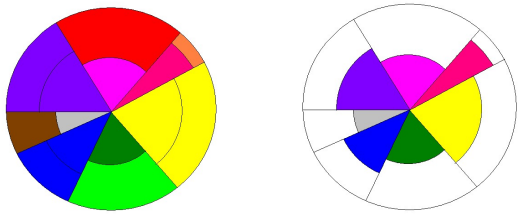
Lost of statistical power. No bias on estimated parameters

MAR
(missing at random)



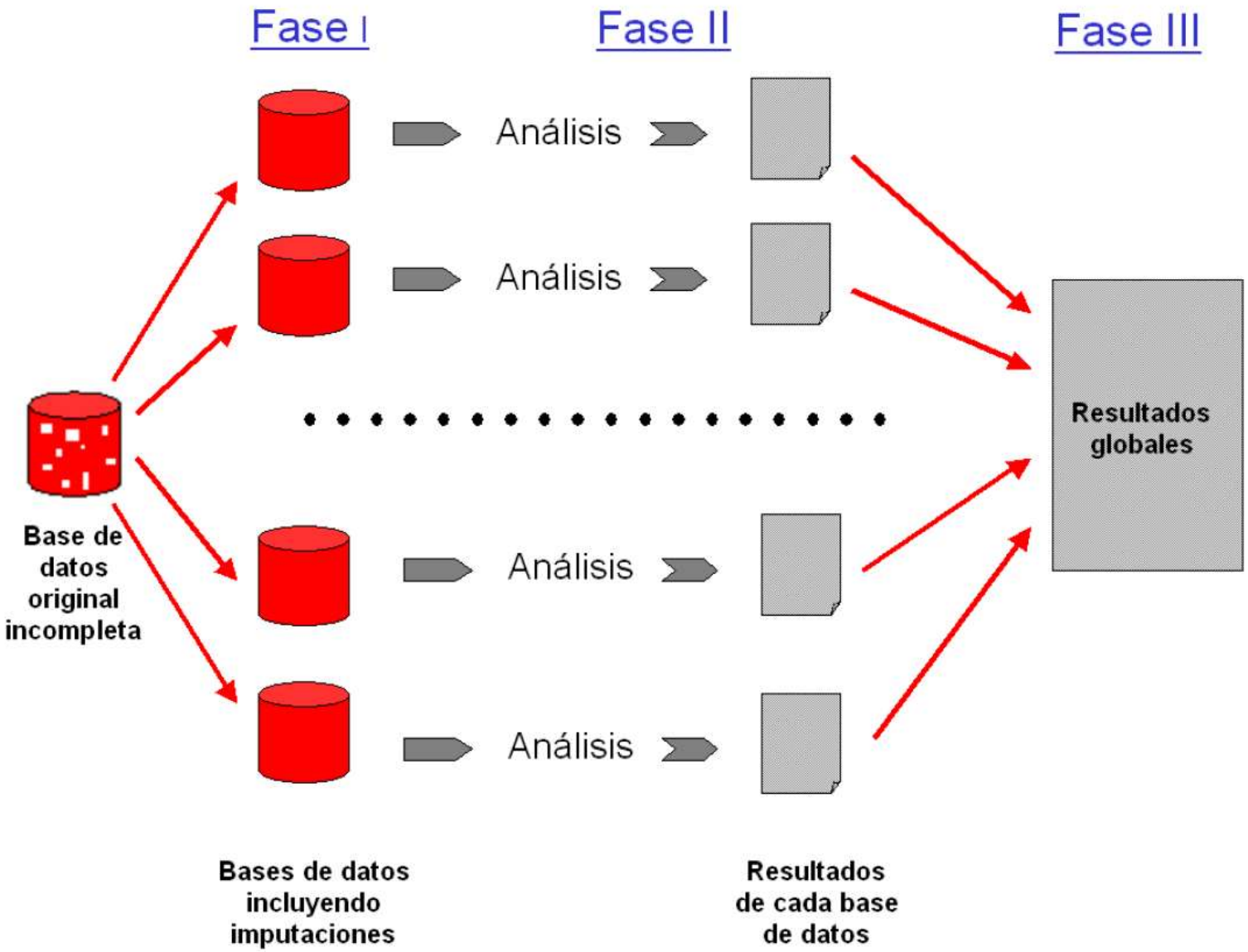
Lost of statistical power and bias with >25%th missingness. Use múltiple imputation

MNAR
(missing not at random)



Lost of statistical power. Bias cannot be reduced. Sensitivity analysis must be conducted

Multiple Imputation(MICE)



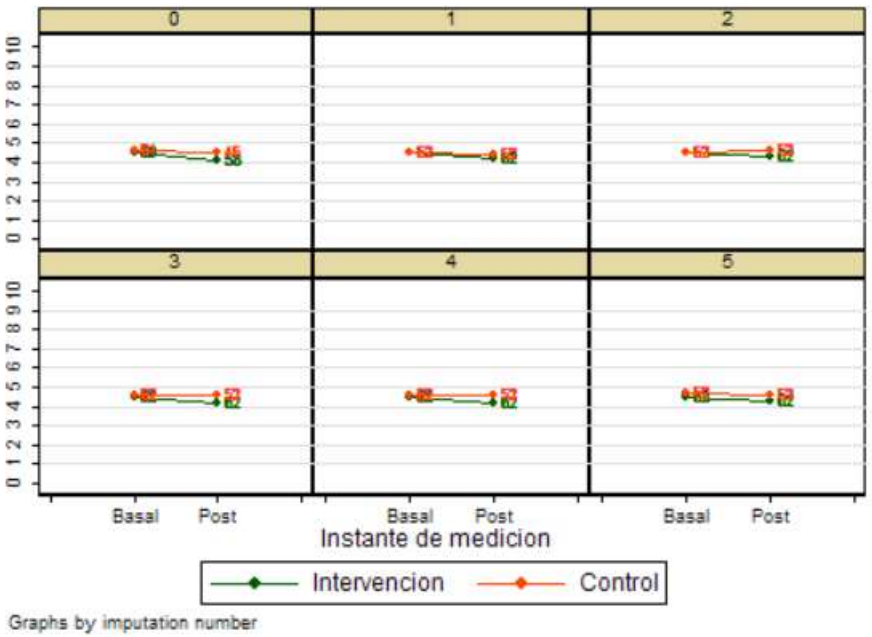
Clinical Effectiveness of Online Training in Palliative Care of Primary Care Physicians

, PhD,[2] and Yolanda Agra-Varela, PhD[3]

TABLE 2. GROUP COMPARISONS OF PAIN INTENSITY AND PAIN IMPACT OF THE BRIEF PAIN INVENTORY

Original data *Imputed data*	*Value*	*95% CI*	P	*Difference with reference group*
Pain intensity				
Time 1 in intervention (*reference*)	4.40	[3.87-4.93]	——	0
	4.49	*[3.96-5.02]*		*0*
Effect in time 2	−0.22	[−0.59-0.15]	0.246	−0.22
	−0.24	*[−0.65-0.17]*	*0.247*	*−0.24*
Effect in control	0.13	[−0.65-0.91]	0.740	0.13
	0.10	*[−0.69-0.89]*	*0.808*	*0.14*
Interaction between time 2 and control	0.28	[−0.27-0.83]	0.319	0.19
	0.28	*[−0.39-0.96]*	*0.403*	*0.16*
Difference between time 2 and time 1 in control	0.06	[−0.35; 0.47]	0.768	0.06
	0.04	*[−0.50-0.59]*	*0.352*	*0.04*
Difference of intervention and control in time 2	0.41	[−0.38-1.20]	0.307	0.41
	0.38	*[−0.43-1.20]*	*0.356*	*0.38*
Pain impact				
Time 1 in intervention (*reference*)	6.32	[5.56-7.07]		
	6.41	*[5.68-7.15]*		
Effect in time 2	−0.18	[−0.67-0.30]		
	−0.49	*[−1.04-0.07]*		
Effect in control	0.32	[−0.79-1.42]		
	0.28	*[−0.82-1.37]*		
Interaction between time 2 and control	0.23	[−0.49-0.94]		
	0.46	*[−0.34-1.26]*		
Difference between time 2 and time 1 in control	0.05	[−0.48-0.56]		
	-0.03	*[−0.71-0.66]*		
Difference of intervention and control in time 2	0.55	[−0.58-1.67]		
	0.38	*[−0.36-1.84]*		

Brief Pain Inventory scale is 1 to 11.
CI, confidence interval.



Data Categorization??

Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents

Caroline Bennette[1] and Andrew Vickers[2*]

BMC Medical Research Methodology 2012, **12**:21

Statistics Notes

The cost of dichotomising continuous variables

BMJ VOLUME 332 6 MAY 2006

STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:127–141
Published online 11 October 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2331

Dichotomizing continuous predictors in multiple regression: a bad idea

Patrick Royston[1,*,†], Douglas G. Altman[2] and Willi Sauerbrei[3]

Dichotomizing Continuous Variables in Statistical Analysis: A Practice to Avoid

Neal V. Dawson, MD, and Robert Weiss, PhD

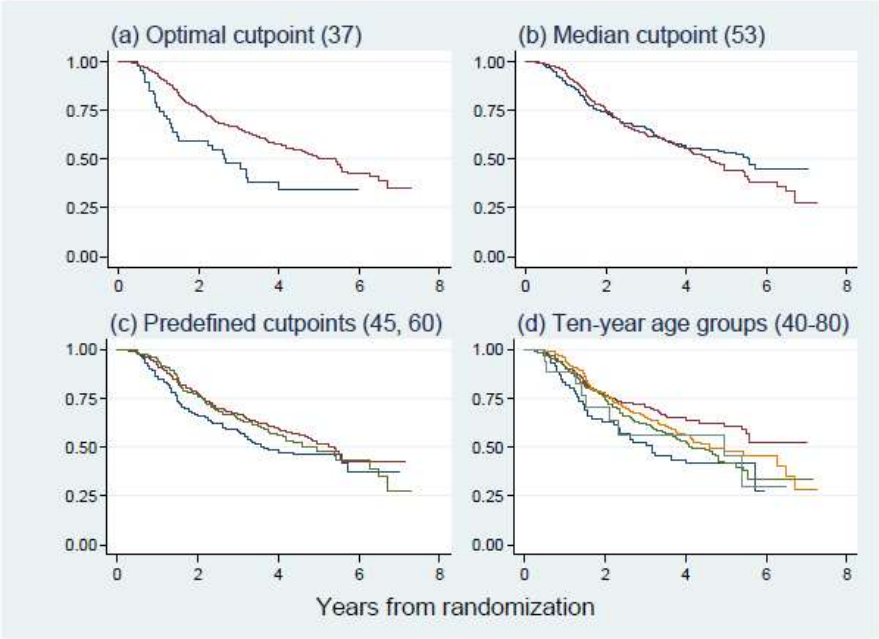
MEDICAL DECISION MAKING/MAR–APR 2012

Problems with Data Categorization

- Loss of Information and Power
- Assumes **homogeneity of risk within groups**
- Choice of the **cutpoint arbitrarely**
- Categorization assumes that there is **a discontinuity in response** as interval boundaries are crossed
- Optimal cut point runs of a high risk of spurious significant result and difference overestimation
- When using dichotomized variable for adjustment **residual confounding remains**
- Preferible keep it continuous and use log transformation or non linear functional relationships

Age Categorization Example (Sauerbrei)

Age as prognostic factor – cutpoint analyses



The youngest group is always in blue.
(a)'Optimal' (37 years); HR (older vs younger) 0.54, p= 0.004
(b) median (53 years); HR (older vs younger) 1.1, p= 0.4
(c) predefined from earlier analyses (45, 60years);
(d) popular (10-year groups)

Example bad categorization by quartiles

Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents

BMC Medical Research Methodology 2012, **12**:21

Caroline Bennette[1] and Andrew Vickers[2*]

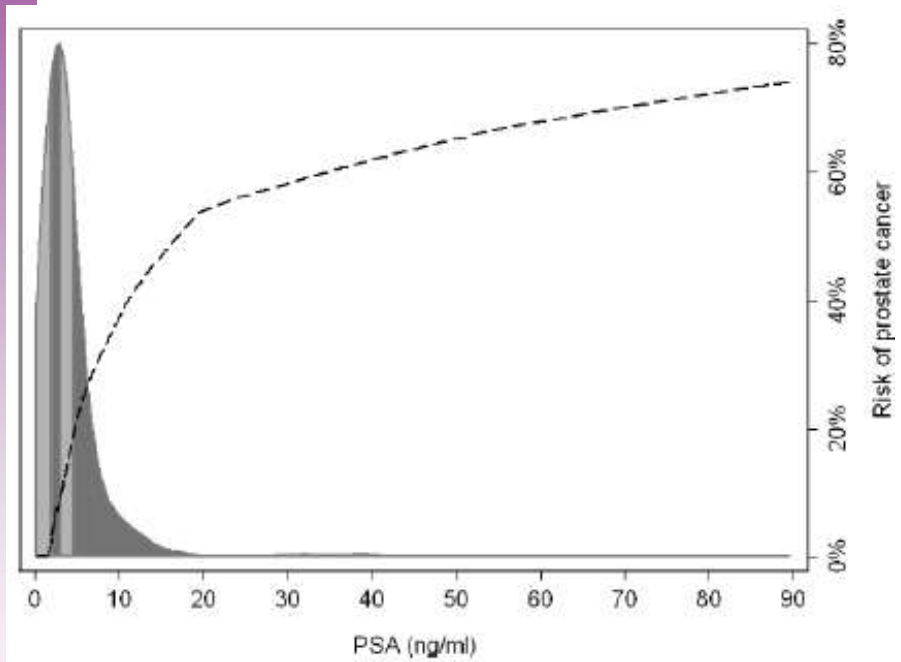


Figure 1 The risk of prostate cancer by level of PSA, with the distribution of PSA levels (ng/ml) among a population-based sample. The shading in the distribution gives the quartiles of PSA.

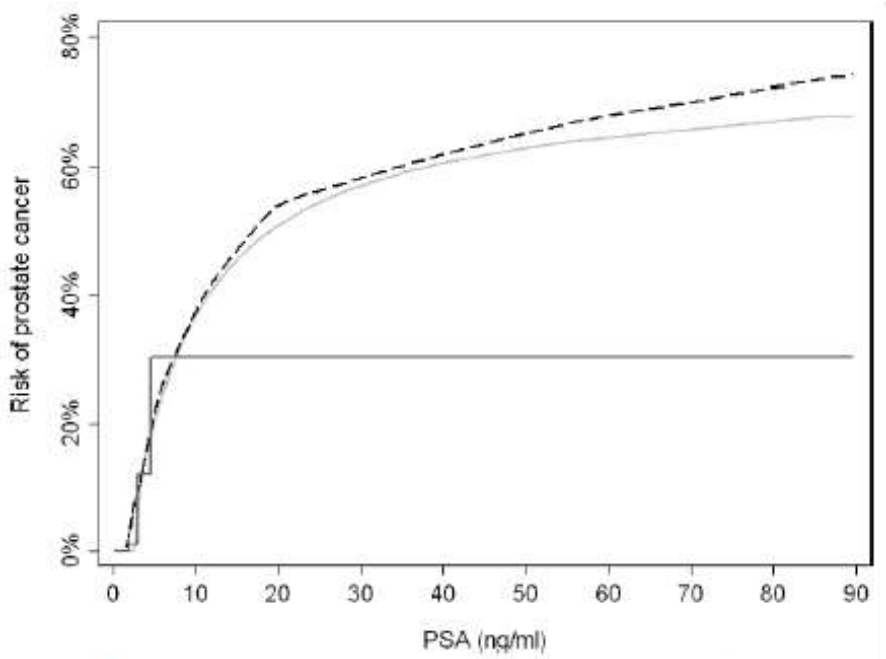
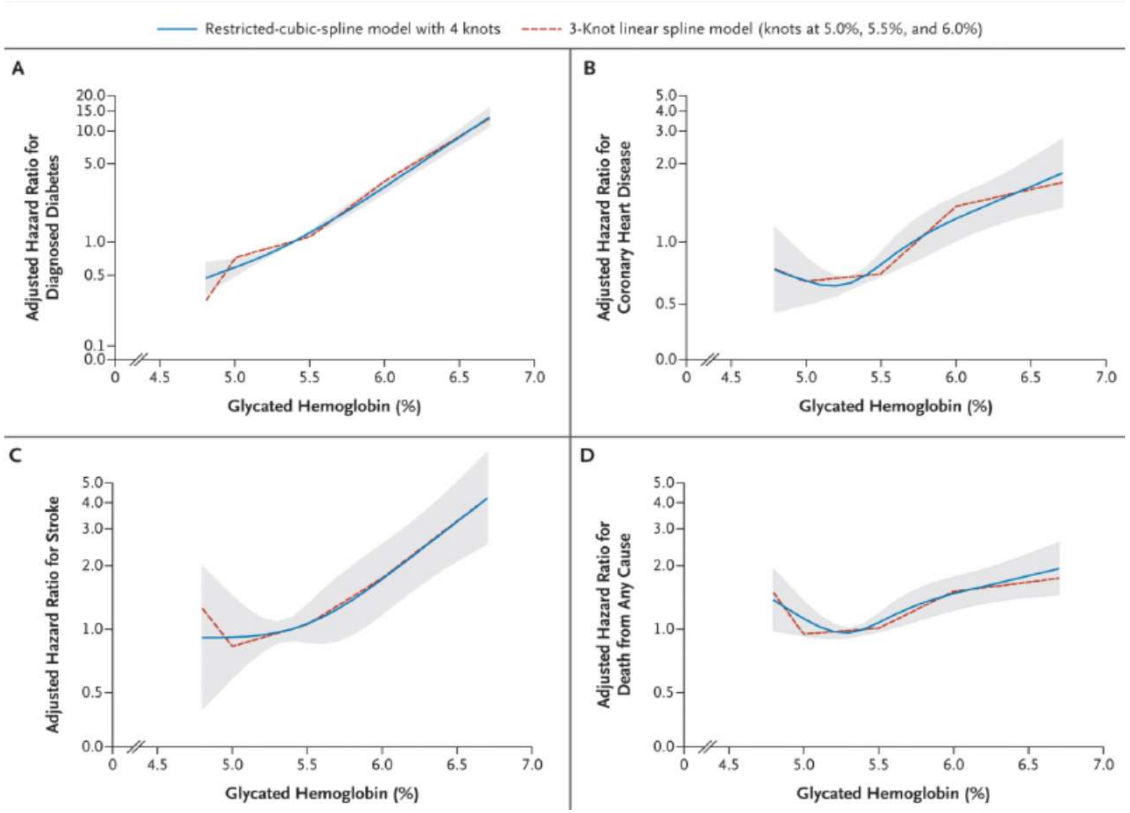


Figure 2 Prostate cancer risk by PSA (black dashed line), with predicted risks using either cubic splines (light gray solid line) or quartiles (dark gray solid line).

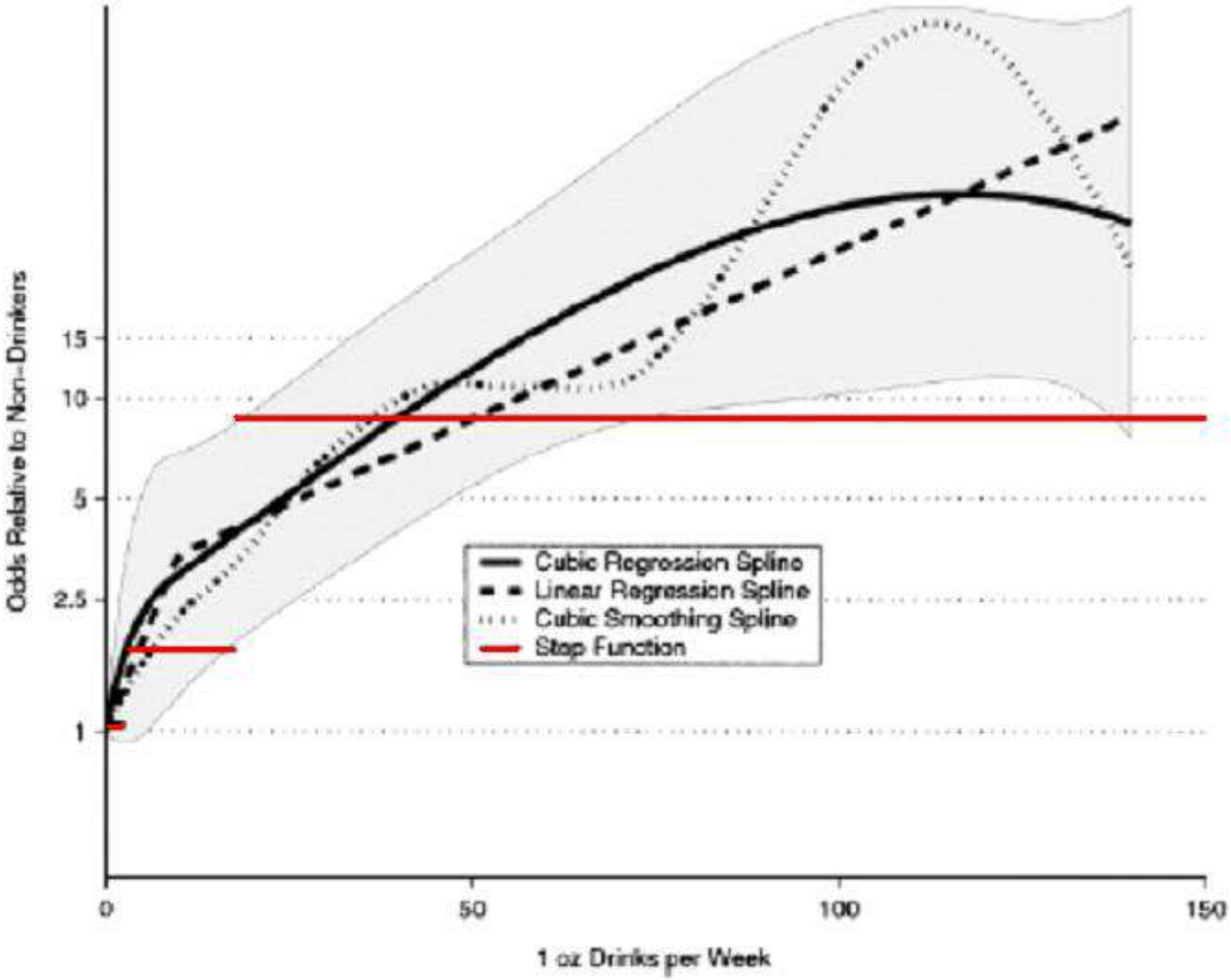
Non Linearity Solutions

- Simple Polynomial (quadratic, cubic, etc.) $Y=\alpha+\beta X+BX^2$
- Spline linear function
- Cubic Splines
- Restricted Cubic Splines



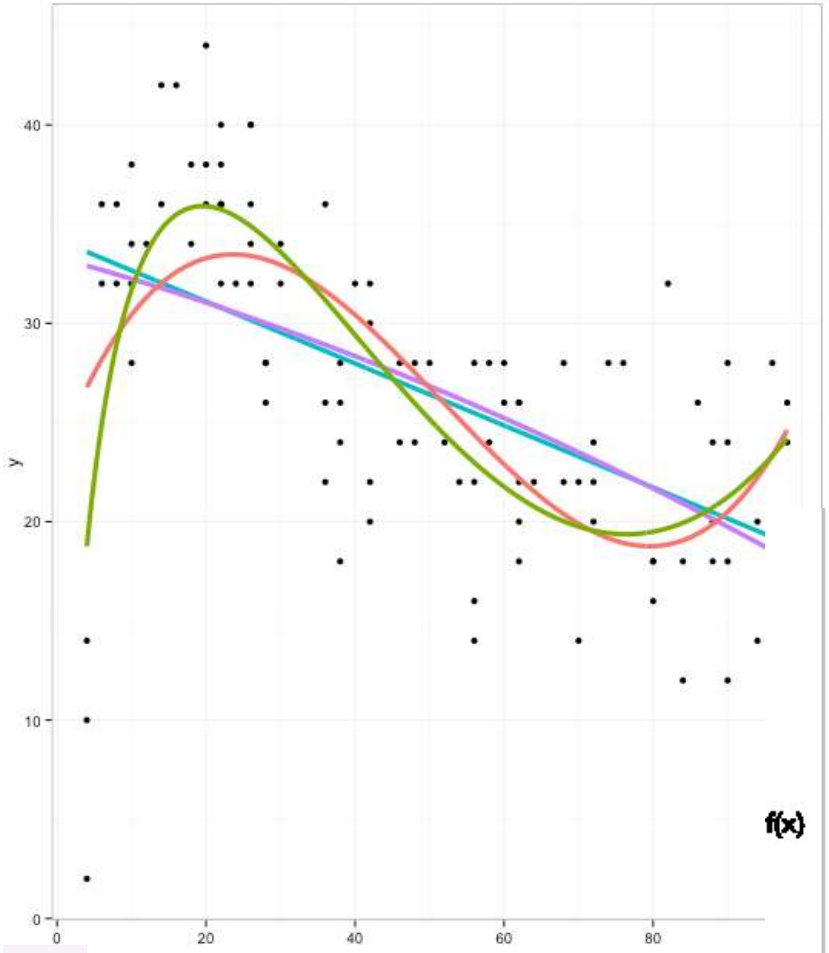
E. Selvin et al. \Glycated Hemoglobin, Diabetes, and Cardiovascular Risk in Nondiabetic Adults". In: NEJM 362.9 (Mar. 2010), pp. 800{

Alcohol consumption as risk factor for oral cancer

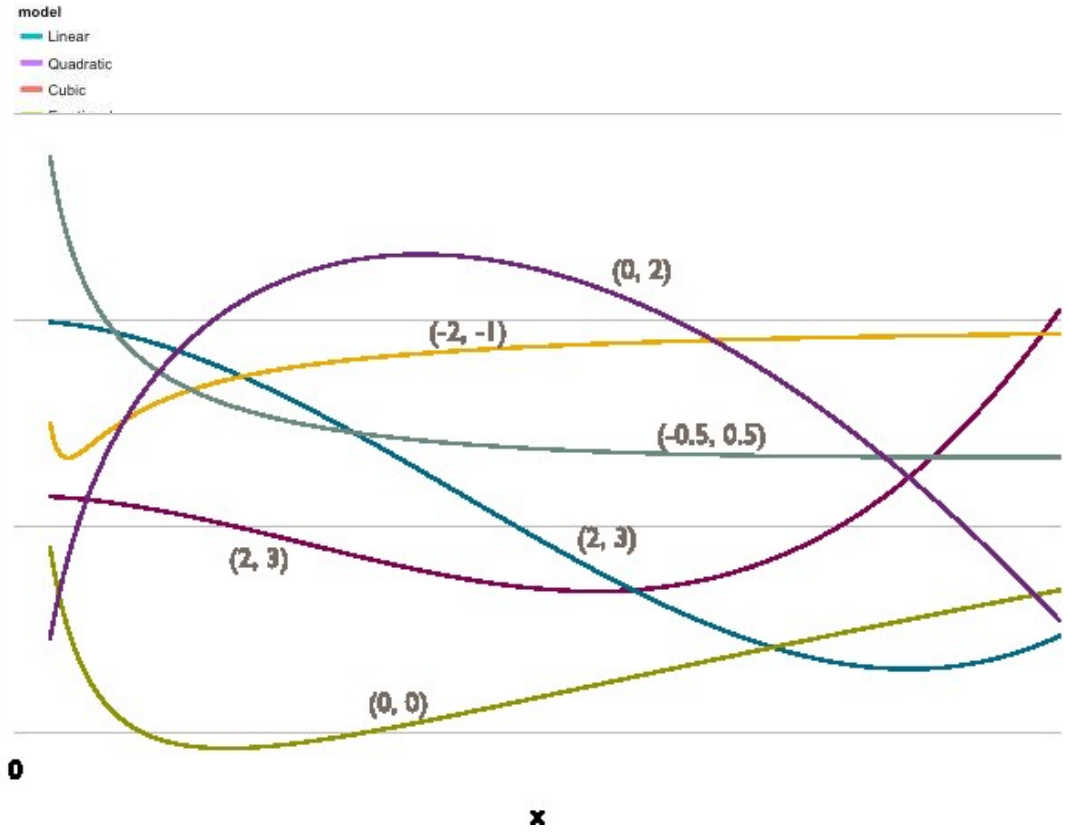


Rosenberg et al, StatMed 2003

Fractional polynomials



- Mathematical equations can defined several shapes



Example

Intensive care admission and hospital mortality in the elderly after non-cardiac surgery

Med Intensiva. 2018;42(8):463–472

M. de Nadal[a,*], S. Pérez-Hoyos[b], J.C. Montejo-González[c], R. Pearse[d], C. Aldecoa[e], on behalf of the European Surgical Outcomes Study (EuSOS) in Spain[1]

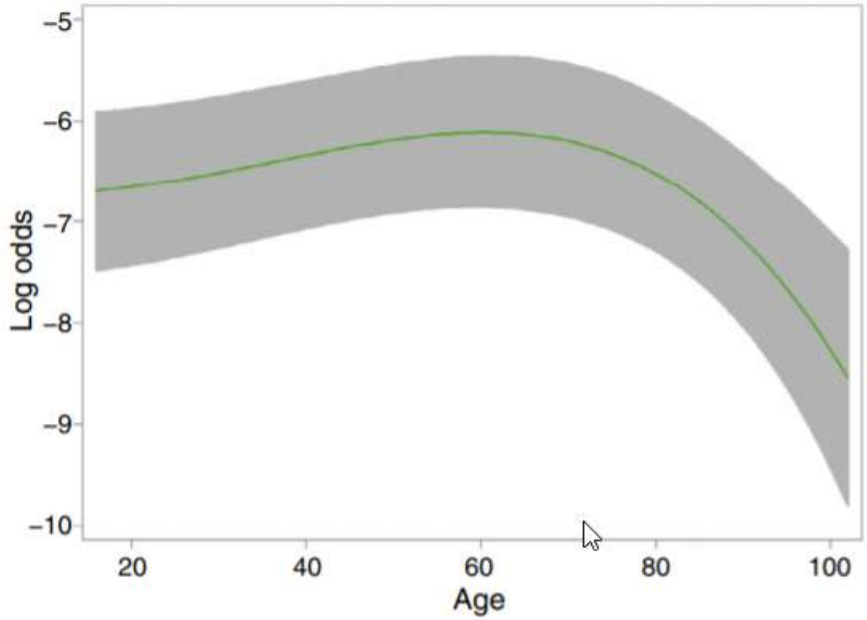


Figure 2 Relationship between age and intensive care unit (ICU) admission after fitting a fractional polynomial model (−3,3).

Table 2 Adjusted logistic regression for intensive care unit (ICU) admission.

	ICU admission	Crude OR (95% CI)	Adjusted OR (95% CI)	*P* value
Age				
<65 years	339 (10.6)	1	1	0.004
65–74 years	181 (16.8)	1.72 (1.41–2.09)	1.08 (0.84–1.39)	
75–85 years	140 (14.5)	1.41 (1.14–1.75)	0.73 (0.55–0.97)	
>85 years	17 (9.0)	0.84 (0.50–1.40)	0.41 (0.22–0.77)	

Variable selection

- Including right variables is a balance between **parsimony and complexity**
- Predictive models should include the variables that reflect the pattern in the population the sample has been drawn
- Effect estimation-The fitted model should **represent the theoretical framework** and estimated parameters should be **corrected for overfitting**
- Overfitting ocurres when a model include too many predictors that may exits in the sample but not in the population . I will not be replicated in other sample
- A **mininimum number of events** is recomended.

Five myths about variable selection

Georg Heinze & Daniela Dunkler *Transplant International* 2017; 30: 6–10

- Myth 1: "The number of variables in a model should be reduced until there are 10 events per variables."
- Myth 2: "Only variables with proven univariable-model significance should be included in a model."
- Myth 3: "Insignificant effects should be eliminated from a model."
- Myth 4: "The reported P-value quantifies the type I error of a variable being falsely selected."
- Myth 5: "Variable selection simplifies analysis."

Often there is no scientific reason to perform variable selection.

Variable selection should always be accompanied by sensitivity analyses to avoid wrong conclusions

Which variables should be included?

Effect of underfitting and overfitting

Illustration by simp le example in linear regression models (mean of 3 runs)

3 predictors $r_{1,2} = 0.5, r_{1,3} = 0, r_{2,3} = 0.7, N = 400, \sigma^2 = 1$

Correct model M_1 $y = 1 \cdot x_1 + 2 \cdot x_2 + \varepsilon$

	M_1 (true)	M_2 (overfitting)	M_3 (underfitting)
$\hat{\beta}_1$	1.050 (0.059)	1.04 (0.073)	-
$\hat{\beta}_2$	1.950 (0.060)	1.98 (0.105)	2.53 (0.068)
β_3	-	-0.03 (0.091)	-
$\hat{\sigma}^2$	1.060	1.060	1.90
R^2	0.875	0.875	0.77

M_2 overfitting $y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$

Standard errors larger (variance inflation)

M_3 underfitting $y = \beta_2 x_2 + \varepsilon$

$\hat{\beta}$ ‚biased', different interpretation, R^2 smaller, stand. error (VIF↓, $\hat{\sigma}^2$↑)?

Patric Royston, Willi Sauerbrei.. Multivariable Model - Building: A Pragmatic Approach to Regression Anaylsis based on Fractional Polynomials for Modelling Continuous Variables. Willey 2008

Avoid overfitting

- Include and exclude variables **according to the literature** and topic expertise
- Do **not** considered variables with **very narrow variability**
- Eliminate or **impute** values of predictors with a lot of **missing data**
- Apply shrinkage and **penalization techniques** on the regression coeficients
- Try to group or select variables that represent the same (collinearity problems)

Basic methods of selection

Full Model

- Variance inflation in the case of multicollinearity
- Correlation less than 0.7 not produces collinearity problems

Stepwise procedures

- Based on prespecified α_{in},α_{out} and p values of test comparing models with and without the predictor
- Forward selection (FS) Adding from Null Model
- Backward selection (BS) Dropping from Full Model
- Stepwise selection Adding-Dropping in each Step

All subset selection

- Cp Mallows
- AIC
- BIC

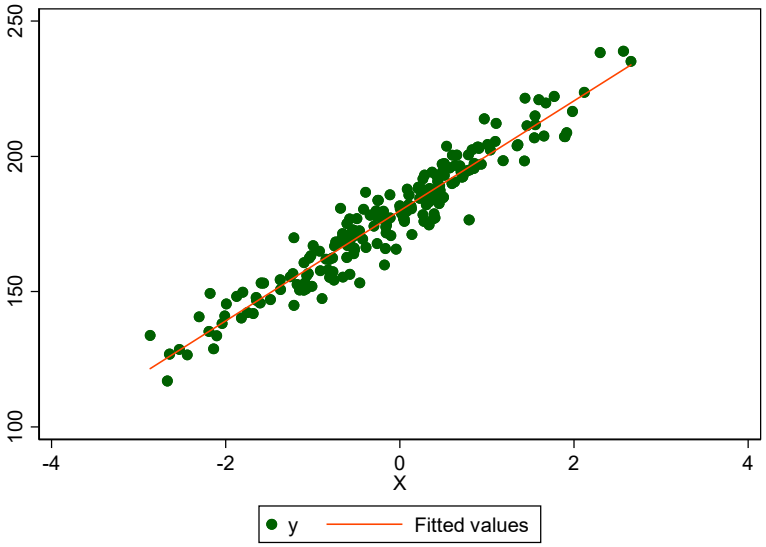
I must not use Stepwise Regression
I must not use Stepwise Regression
I must not use Stepwise Regression
I must not use Stepwise Regression
I must not use Stepwise Regression
I must not use Stepwise Regression
I must not use Stepwise Regression
I must not use Stepwise Regression
I must not use Stepwise Regression
I must not use Stepwise Regression
I must not use Stepwise Regression
I must not use Stepwise Regression
I must not use Stepwise Regression
I must not use Stepwise Regression
I must not use Stepwise Regression
I must not use Stepwise Regression
I must not use Stepwise Regression
I must not use Stepwise Regression
I must not use Stepwise Regression
I must not use Stepwise Regression
I must not use Stepwise Regression
I must not use Stepwise Regression
I must not use Stepwise Regression

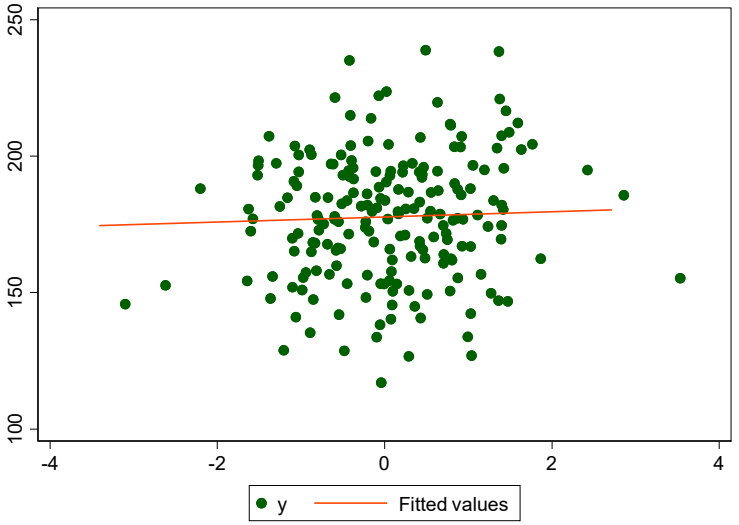
Criticism to Stepwise Methods

- Not include all the variables that influence on the response or include some that are not influent
- Variables must be included by chance (Multiplicity of tests)
- Bias in results specially in smaller data sets. Relaxing p value for select variables a value over 0.05.
- Small changes in data change the final selection
- A reason for rejecting in some journals
- Overfitting problems

Example

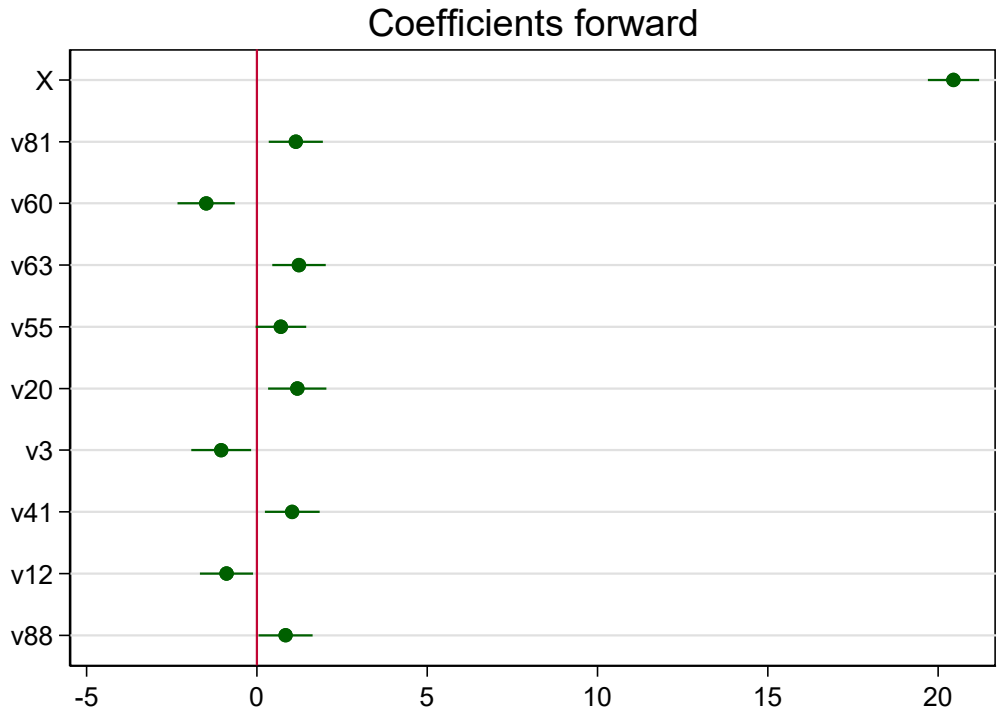
- We generate **90** variables that are random noise **v1-v90** ~ normal(0,1)
- We generate variable **X** as random noise normal(0,1)
- We generate response variable **Y** as Y= 180+20* X+rnormal(0,6)
- Sample Size=**200**





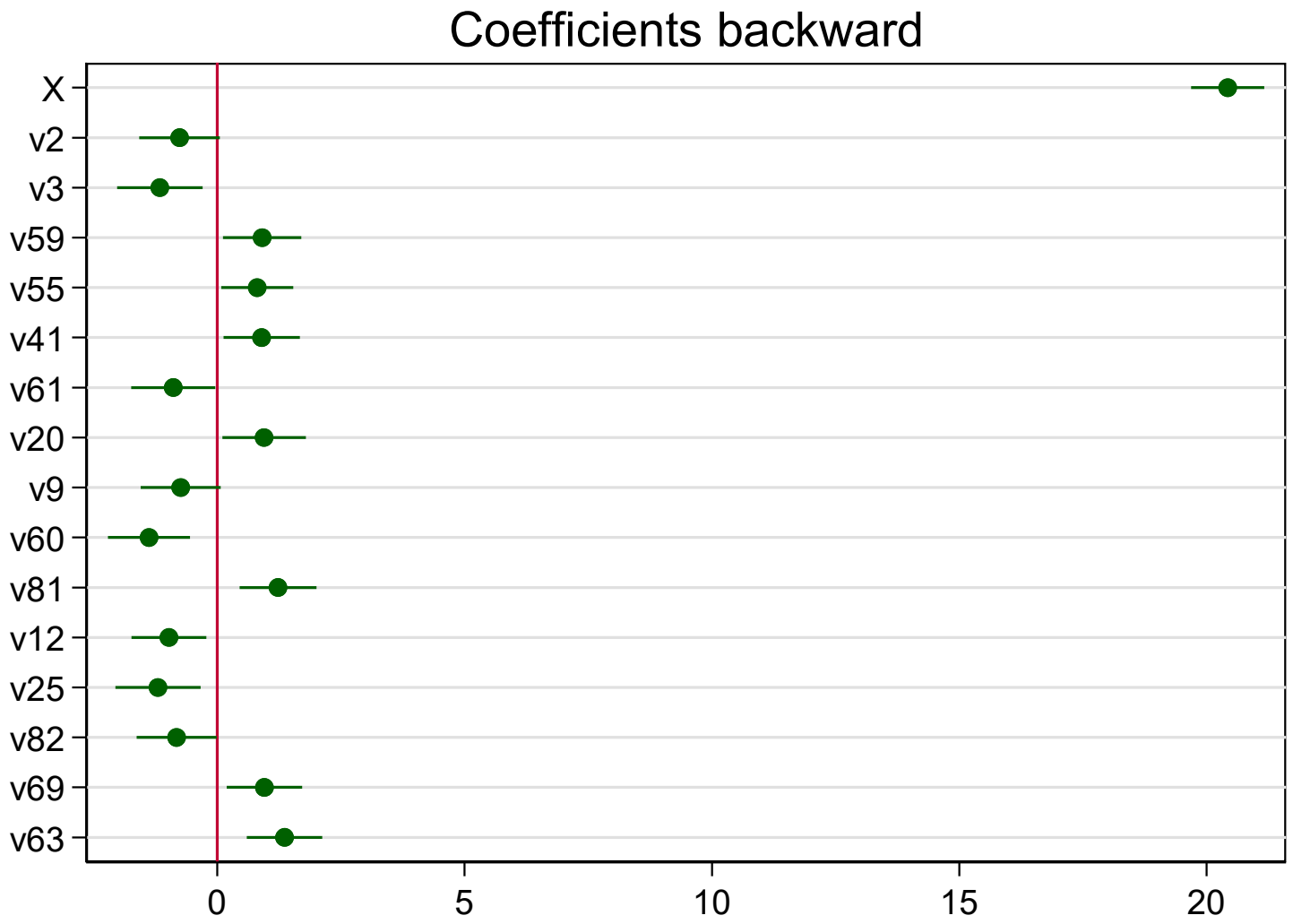
Forward Stepwise (Prob entry 0.05)

10 variables were “over” included

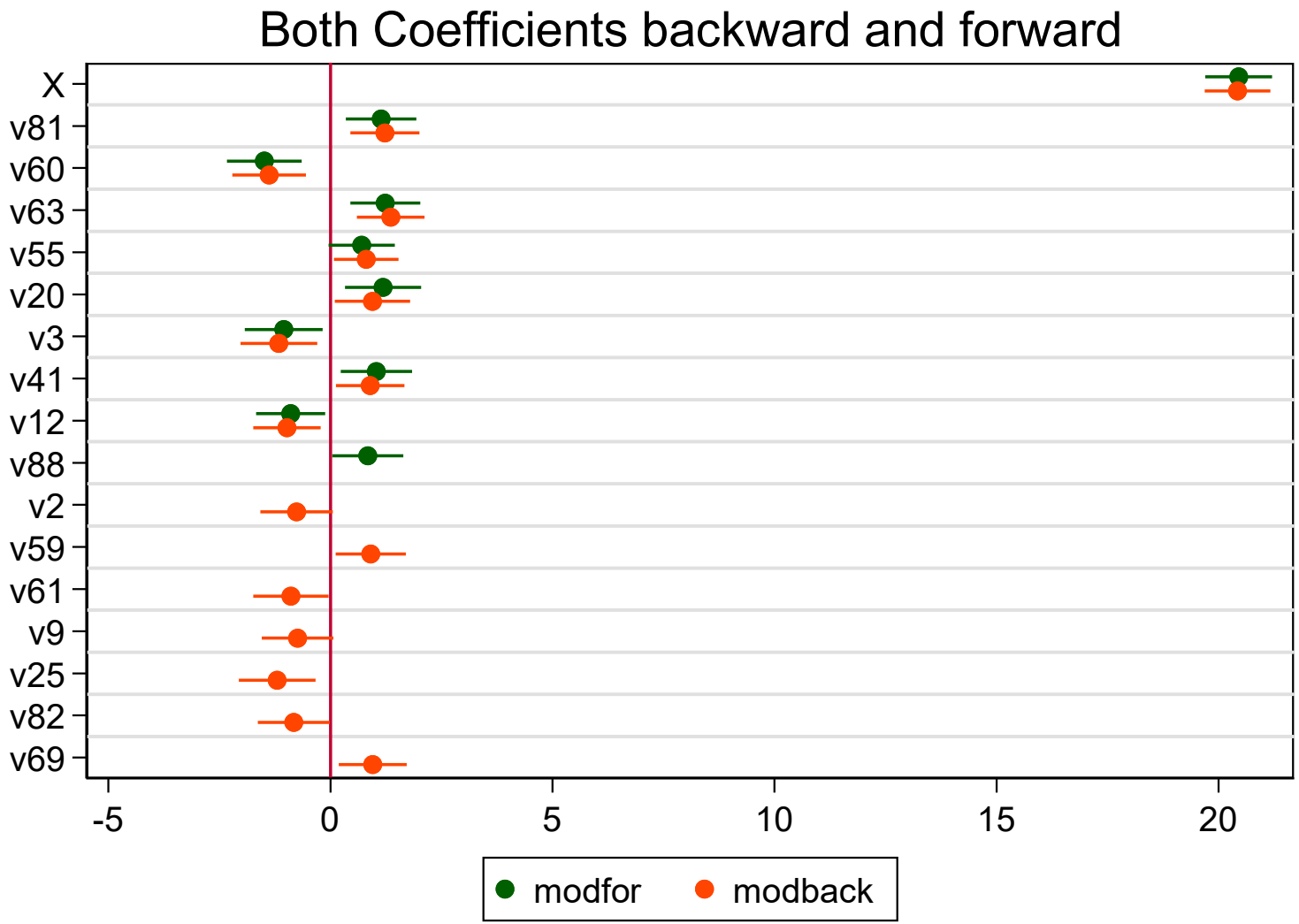


Backward Stepwise (Prob remove 0.1)

15 variables were “over” included



Both Methods Common coefficients



Some solutions to automatic variable selection

- Multivariable fractional polynomial (Royston)
  - Two algorithms for fp model selection for each continuous predictor.
  - Combine backwards elimination with fp selection procedures
- Trade off between bias and variance methods

$$L_{OLS}(\hat{\beta}) = \sum_{i-1}^{n}(y_i - x_i'\hat{\beta})^2 = ||y - X\hat{\beta}||^2$$

- Ridge regression
  - Use the shrinkage estimator by penalized each covariate

$$L_{ridge}(\hat{\beta}) = \sum_{i-1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda\sum_{j-1}^{m}\hat{\beta}_j^2 = ||y - X\hat{\beta}||^2 + \lambda||\hat{\beta}||^2.$$

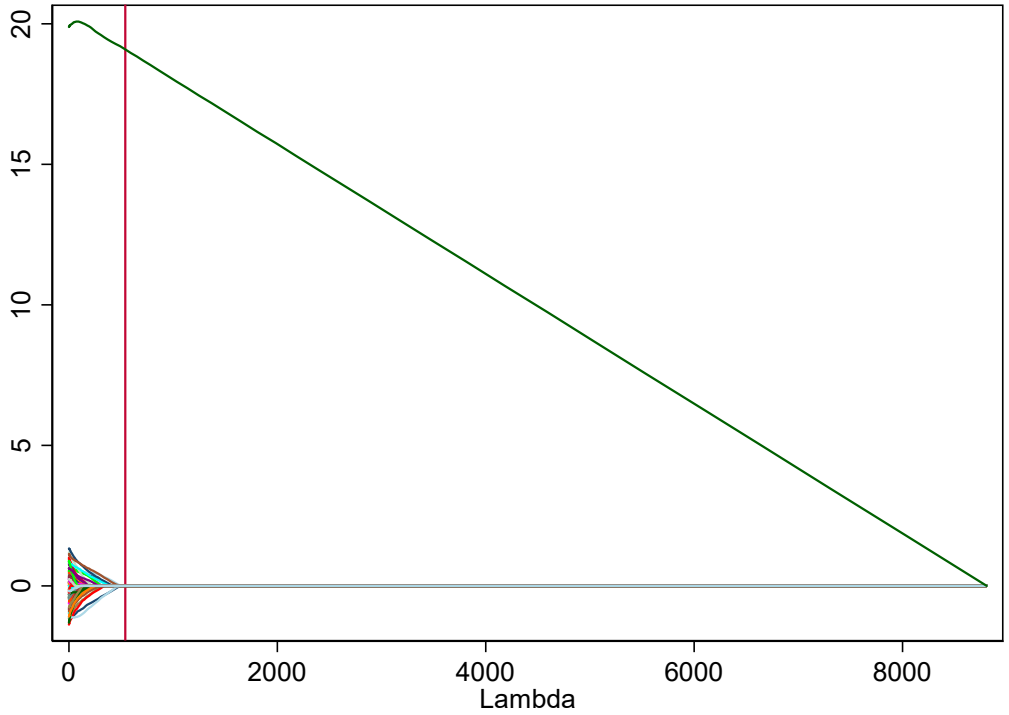
- Lasso
  - The penalty is with the sum of the absolute value of coefficients.
  - Set the coefficients at 0 and tune them by varying lambda

$$L_{lasso}(\hat{\beta}) = \sum_{i-1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda\sum_{j-1}^{m}|\hat{\beta}_j|.$$

- Elastic Net
  - Combine Ridge and Lasso penalties

$$L_{enet}(\hat{\beta}) = \frac{\sum_{i-1}^{n}(y_i - x_i'\hat{\beta})^2}{2n} + \lambda(\frac{1-\alpha}{2}\sum_{j-1}^{m}\hat{\beta}_j^2 + \alpha\sum_{j-1}^{m}|\hat{\beta}_j|),$$

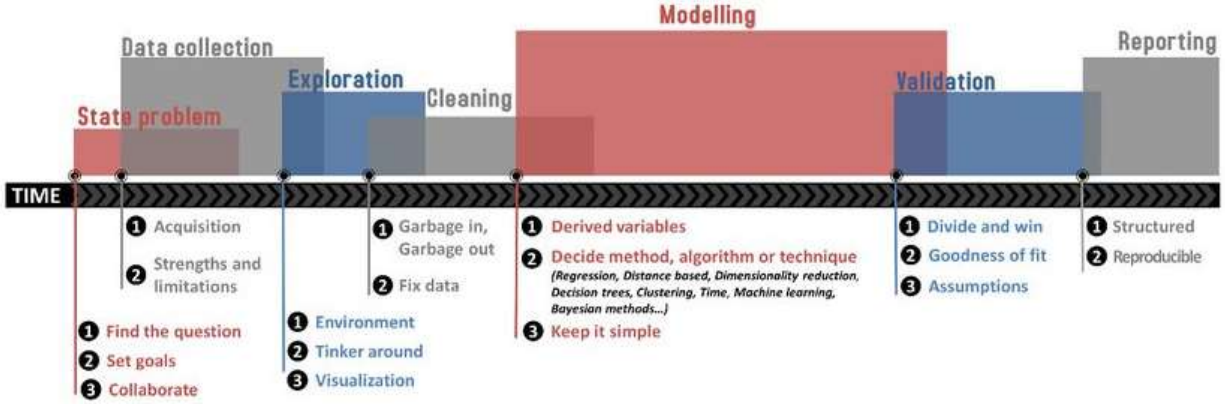
Lasso selection



Use **lambda=540**.4546507102212 (selected by EBIC).

Selected	**Lasso**	**Post-est OLS**
X	**19.0921801**	**20.3402383**
Partialled-out*		
_cons	179.7451205	179.8799902

Modelling Steps



1. **Determining the aim of the model** (Prediction or explanatory)
2. **Ascertainment of the outcome** (binary, continuous,survival, repeated measure, etc.) and data
3. Choosing the **appropriate statistical model** based on design and outcome
4. **Model Estimation** (Missing values,Functional shape, Selection variables,shrinkage , avoid overfitting)
5. **Assessing model performance**
6. **Model Validation** (internal and external)
7. **Presenting Results**

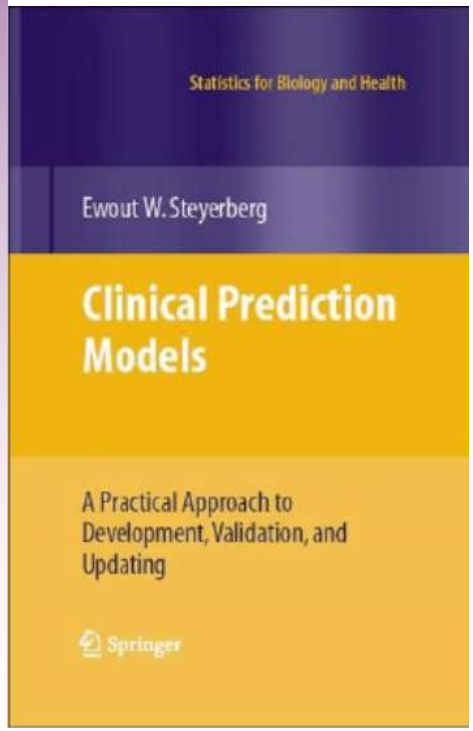
Performance measures

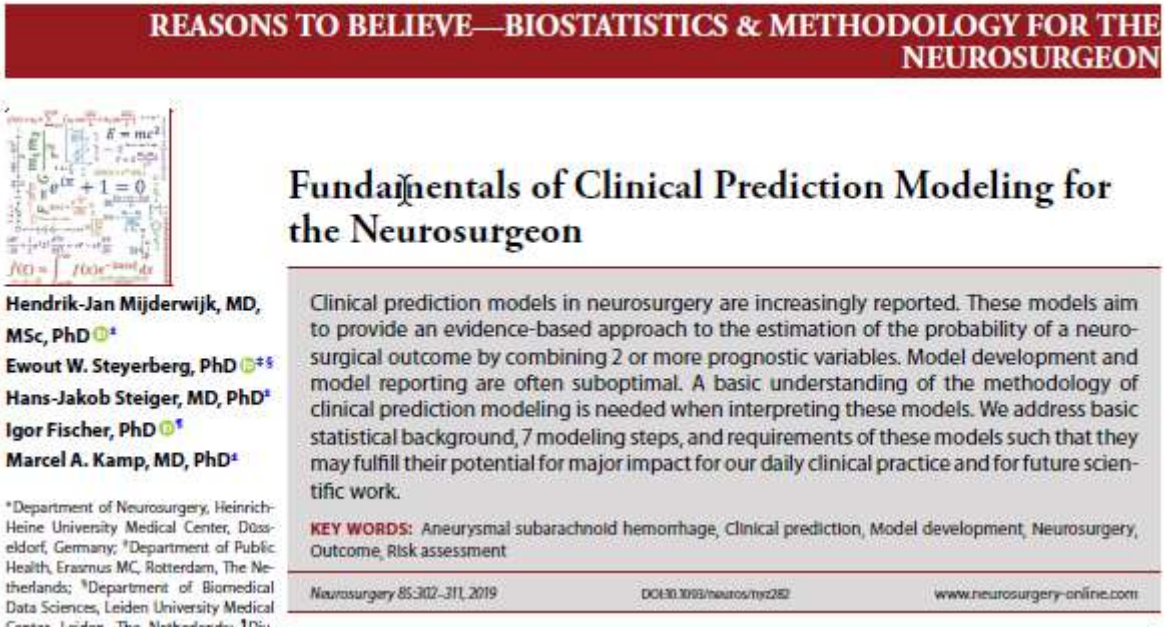
Rev Esp Cardiol. 2011;64(9):788–794

Enfoque: Métodos contemporáneos en bioestadística (IV)

Medidas del rendimiento de modelos de predicción y marcadores pronósticos: evaluación de las predicciones y clasificaciones

Ewout W. Steyerberg[a,*], Ben Van Calster[a,b] y Michael J. Pencina[c]





Performance measures

Overall Measurements

- Distance between the predicted outcome ŷ and actual outcome y
- **R^2** % of variability described by the model (continuous outcome)
- **D2** % Deviance "explained" by the model (generalized linear models)
- Brier Score (distance between prediction and event status)

  Brier score: Y*(1 – *p*)^2 + (1 – Y)**p*^2
  $Brier_{scaled}$ = 1 – Brier / $Brier_{max}$
  $Brier_{max}$ = mean(*p*) x (1 – mean(*p*))^2 + (1 – mean(*p*)) x mean(*p*)^2
  $Brier_{scaled}$ very similar to Pearson R^2 for binary outcome

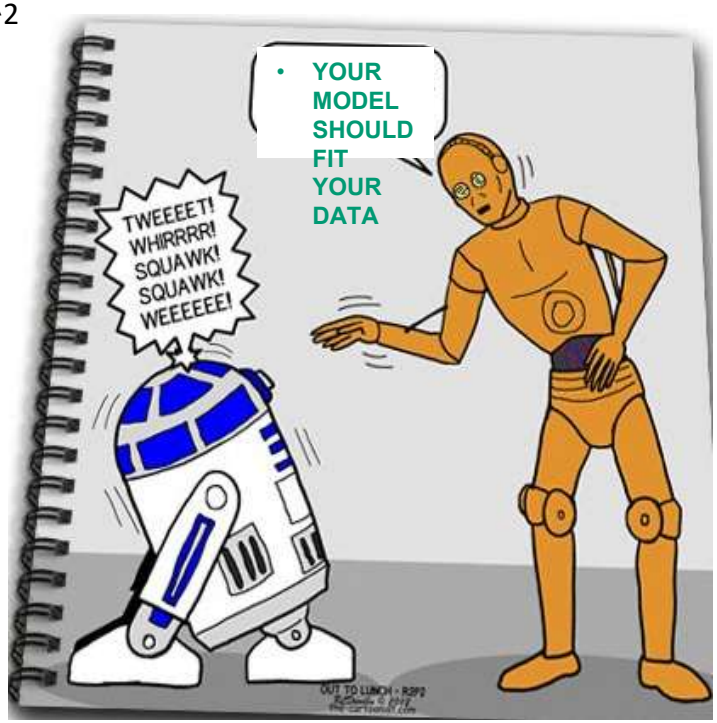
- Goodness of fit Test (Hosmer-Lemeshow)

```
Residual standard error: 2.126 on 998 degrees of freedom
Multiple R-squared:  0.7971,    Adjusted R-squared:  0.7969
F-statistic:  3920 on 1 and 998 DF,  p-value: < 2.2e-16
```

The model explain near the 80% of the variability

```
    Null deviance: 1226.77  on 999  degrees of freedom
Residual deviance:  692.73  on 995  degrees of freedom
AIC: 702.73
```

$$D^2 = \frac{\text{Null deviance} - \text{Residual deviance}}{\text{Null deviance}} * 100 = 43.53\%$$



Performance measures

BIOSTATISTICS & EPIDEMIOLOGY
https://doi.org/10.1080/24709360.2019.1618653

Taylor & Francis
Taylor & Francis Group

Statistical modeling methods: challenges and strategies

Steven S. Henley[a,b,c], Richard M. Golden[d] and T. Michael Kashner[a,b,e]

Table 1. Examples of summary level model fit measures.

No.	Model fit measures	Description
1	Sum of Squared Errors (SSE)	Sum of squared differences (residuals) between predicted and observed values. Measures deviation from actual values [5,6,39].
2	R^2, adjusted R^2, Pseudo-R^2 Statistics	Coefficient of determination (R^2) compares the predictive performance of the model to a constrained version of the model [5,6,10,12,33,36,82,83].
3	Log-Likelihood (LL)	Kullback-Leibler based measure of model fit to observed data. Selects the model that makes the in-sample data (training data) most likely [12,38,84].
4	Akaike Information Criterion (AIC)	AIC allows comparison between nested, overlapping, or nonnested models having different numbers of parameters. Selects the model that makes the out-of-sample data most likely. Assumes all models are correctly specified [18,61,80,81,85–87].
5	Akaike Information Criterion with finite sample correction (AICc)	AICc allows comparison between nested, overlapping, or nonnested models having different numbers of parameters, with small sample size correction. Selects the model that makes the out-of-sample data most likely. Assumes all models are correctly specified [61,87–89].
6	Bayesian Information Criterion (BIC), also know as the Schwarz Criterion (SIC)	BIC/SIC allows comparison between nested, overlapping, or nonnested models having different numbers of parameters. Selects the most probable model given the data. Applicable for both correctly specified and misspecified models [61,86,87,90].
7	Bayesian predictive information criterion (BPIC)	Hierarchical modeling generalization of the AIC and BIC [91,92].
8	Generalized Akaike Information Criterion (GAIC)	GAIC allows comparison between nested, overlapping, or nonnested models having different numbers of parameters. Selects the model that makes the out-of-sample data most likely. Applicable for both correctly specified and misspecified models [61,86,87,93,94].
9	Generalized Bayesian Information Criterion (GBIC)	GBIC allows comparison between nested, overlapping, or nonnested models having different numbers of parameters, with small sample correction superior to BIC. Selects the most probable model given the data. Applicable for both correctly specified and misspecified models [95–97].
10	Kullback Information Criterion (KIC)	KIC is an asymptotically unbiased estimator of Kullback's symmetric divergence that allows comparison between nested, overlapping, or nonnested models having different numbers of parameters [98,99]. Assumes all models are correctly specified.

Performance measures

Discrimination Measurements

- Capacity to discriminate between those **with** and those **without** the outcome
- Concordance statistic c , Harrell's c, Royston's D statistic
- **ROC Curve** (Sensitivity, Specificity)
- **c=AUC** in binary results
- **Box-plot** of predicted probability

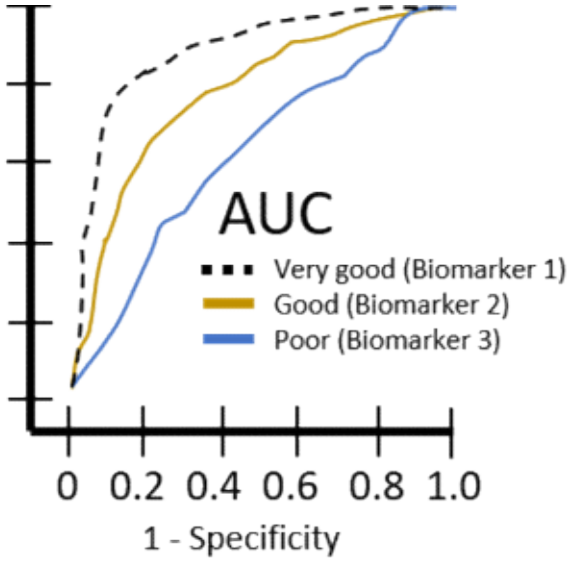
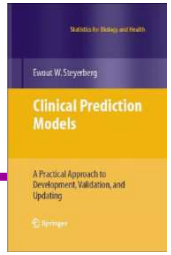
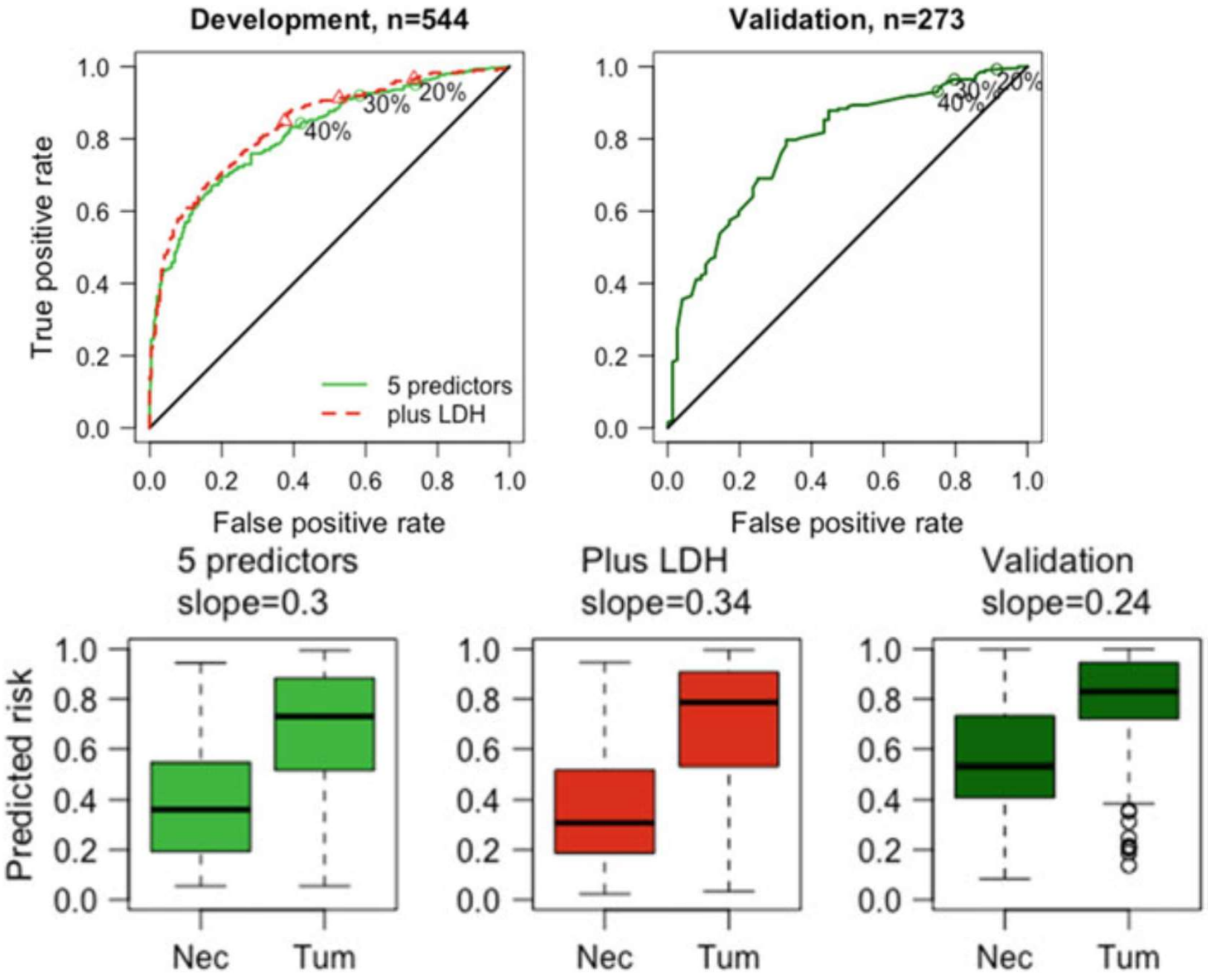


Figure 3. Comparison of ROC curves across three potential biomarkers. The higher the AUC value, the higher predictive value of the biomarker. Biomarker 3 has very poor predictive power (AUC ~0.5) as it cannot differentiate between healthy and diseased patients at all.

Example Outcome: testicular cancer model



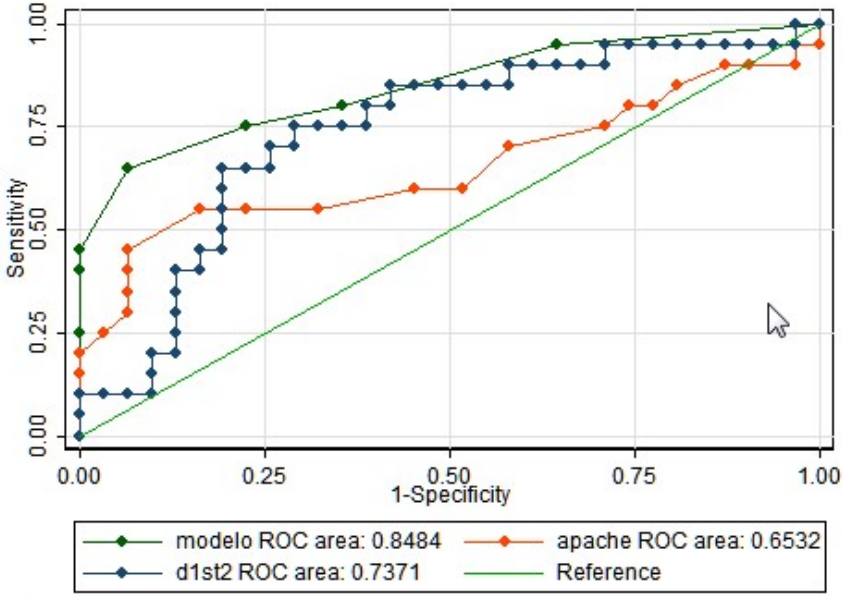


Regressio Multivariant logit per a mort_60

Number of obs = 51

VARIABLE		OR	(95%CI)	p-value
Día 1 ST-2 >4153.60	No(<4153.60)	1		0.0095
	Si(>4153.60)	8.12	(1.67;39.60)	
APACHE II >26.00	No(<26.00)	1		0.0062
	Si(>26.00)	15.88	(2.19;114.85)	
Inmunosupresión previa	No	1		0.0254
	Sí	7.90	(1.29;48.43)	

LL model= -22.01 ; AIC model= 52.03 ; BIC model= 59.75



Detailed report of Sensitivity and Specificity

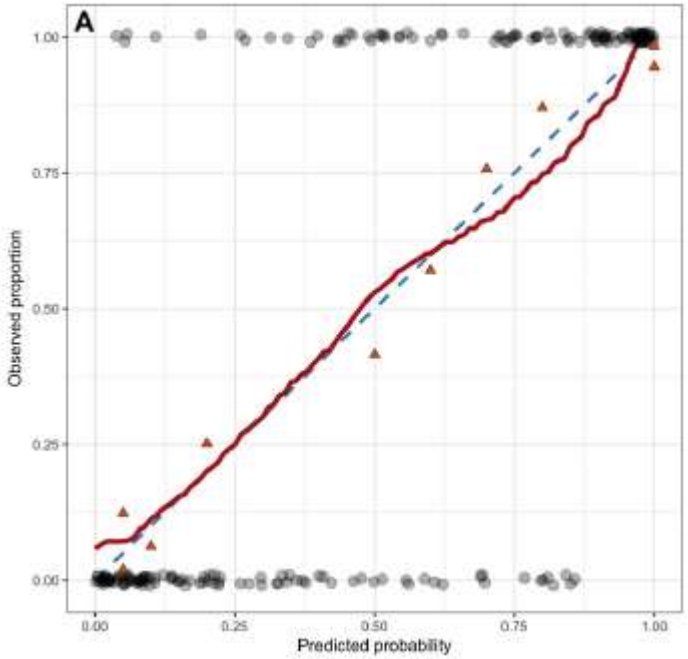
CUTPOINT	SENSITIVITY	SPECIFICITY	CORRECTED CLASSIFIED	LR+	LR-
(>= .0393311)	100.00%	0.00%	39.22%	1.0000	
(>= .244451)	95.00%	35.48%	58.82%	1.4725	0.1409
(>= .2496142)	80.00%	64.52%	70.59%	2.2545	0.3100
(>= .3939386)	75.00%	77.42%	76.47%	3.3214	0.3229
(>= .7244239)	65.00%	93.55%	82.35%	10.0750	0.3741
(>= .8370442)	45.00%	100.00%	78.43%		0.5500
(>= .8407952)	40.00%	100.00%	76.47%		0.6000
(>= .9766)	25.00%	100.00%	70.59%		0.7500
(> .9766)	0.00%	100.00%	60.78%		1.0000

Observaciones	ROC AREA	Std. Error	Asymptotic Normal [95% Conf. Interval]	Indice Youden	Punto corte según Youden	Mejor Punto corte AUC
51	0.8484	0.0585	0.7338 ; 0.9630	0.5855	0.72	0.39

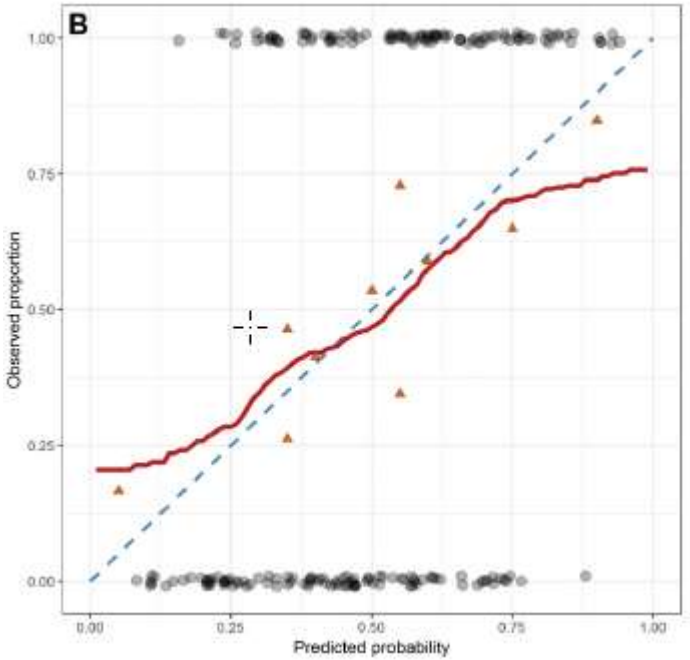
Performance measures

Calibration Measurements

- **Agreement** between the **predictions** of the model with the **observed** outcome
- **Calibration Plot** (QQ plot in linear regression)ç
- **Calibration test** (slope and intercept)



Well Calibrated

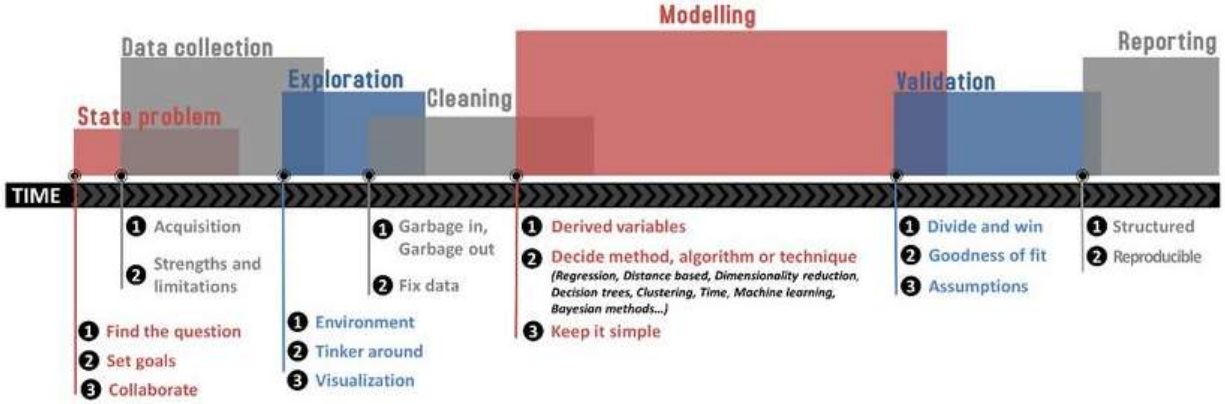


Poor Calibrated

Some calibration and goodness-of-fit tests

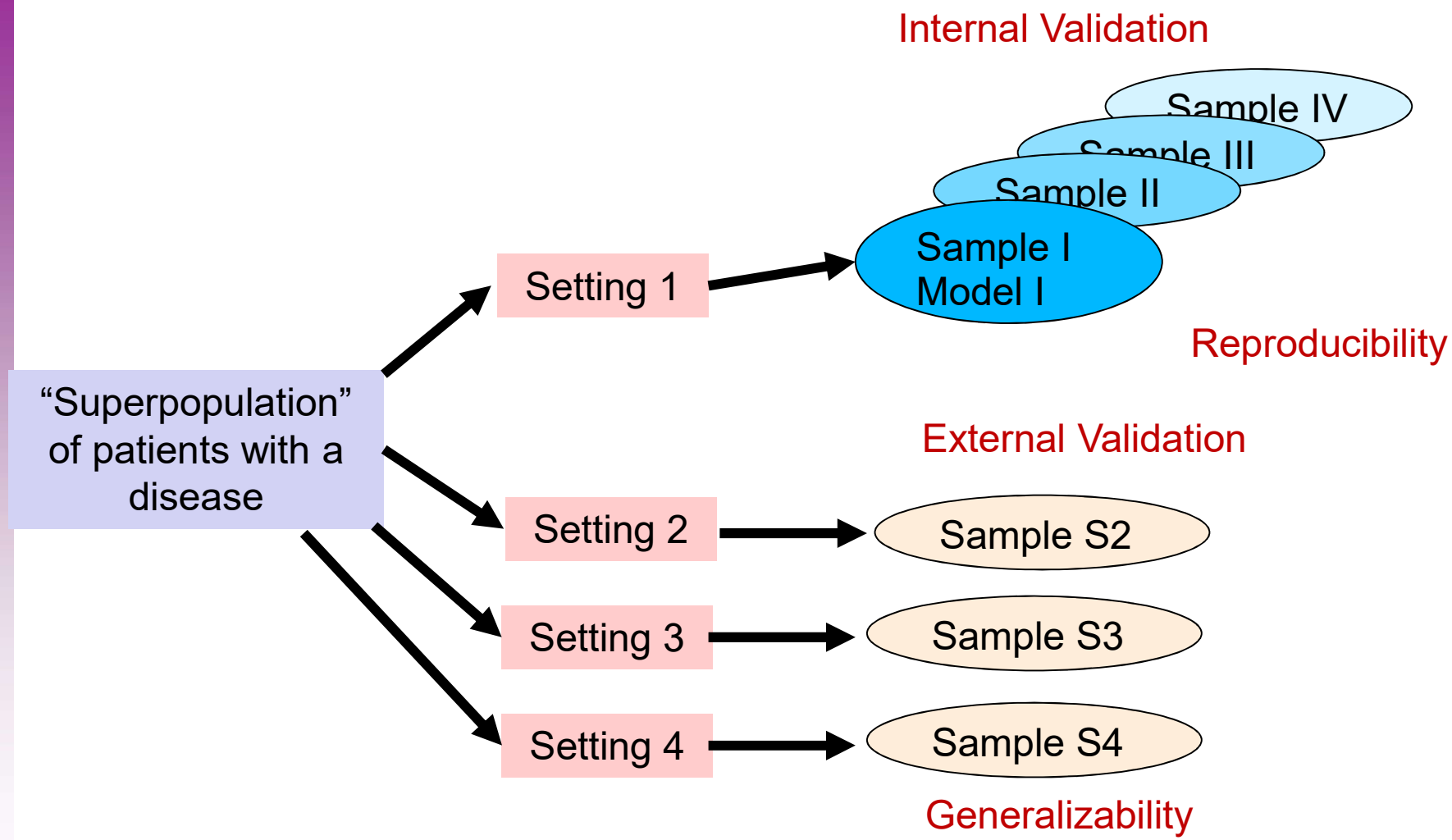
Performance aspect	Calculation	Visualization	Pros	Cons
Calibration-in-the-large	Compare mean(y) versus mean(ŷ)	Calibration graph	Key issue in validation; statistical testing possible	By definition OK in model development setting
Calibration slope	Regression slope of linear predictor	Calibration graph	Key issue in validation; statistical testing possible	By definition OK in model development setting
Calibration test	Joint test of calibration-in-the-large and calibration slope	Calibration graph	Efficient test of 2 key issues in calibration	Insensitive to more subtle miscalibration
Harrell's E statistic	Absolute difference between smoothed y versus ŷ	Calibration graph	Conceptually easy, summarizes miscalibration over whole curve	Depends on smoothing algorithm
Hosmer-Lemeshow test	Compare observed versus predicted in grouped patients	Calibration graph or table	Conceptually easy	Interpretation difficult; low power in small samples
Goeman – Le Cessie test	Consider correlation between residuals	-	Overall statistical test; supplementary to calibration graph	Very general
Subgroup calibration	Compare observed versus predicted in subgroups	Table	Conceptually easy	Not sensitive to various miscalibration patterns

Modelling Steps



1. **Determining the aim of the model** (Prediction or explanatory)
2. **Ascertainment of the outcome** (binary, continuous,survival, repeated measure, etc.) and data
3. Choosing the **appropriate statistical model** based on design and outcome
4. **Model Estimation** (Missing values,Functional shape, Selection variables,shrinkage , avoid overfitting)
5. **Assessing model performance**
6. **Model Validation** (internal and external)
7. **Presenting Results**

Model Validation

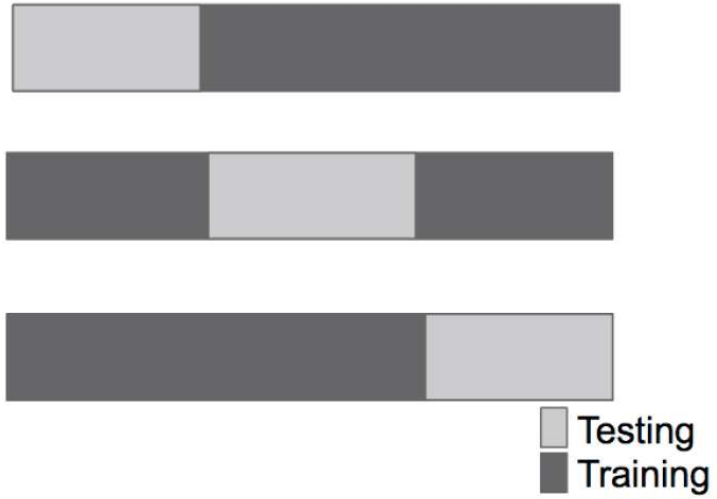


Internal Validation

- With another sample performance is worse
- Optimism difference between model and true performance
- Internal validation quantify optimism

Cross Validation

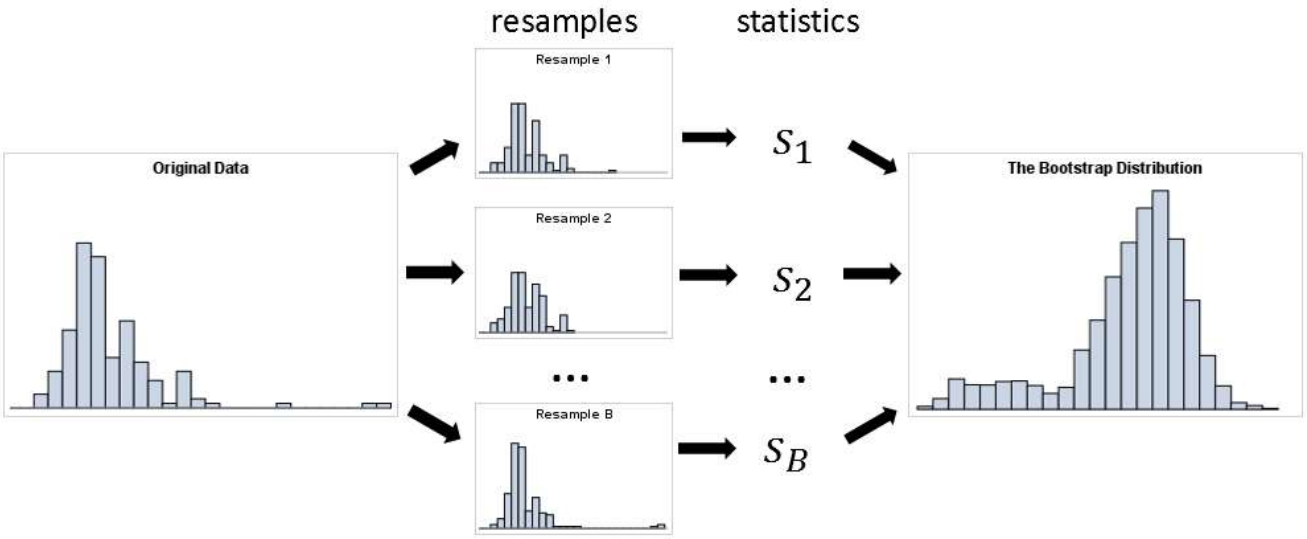
K-Fold Validation



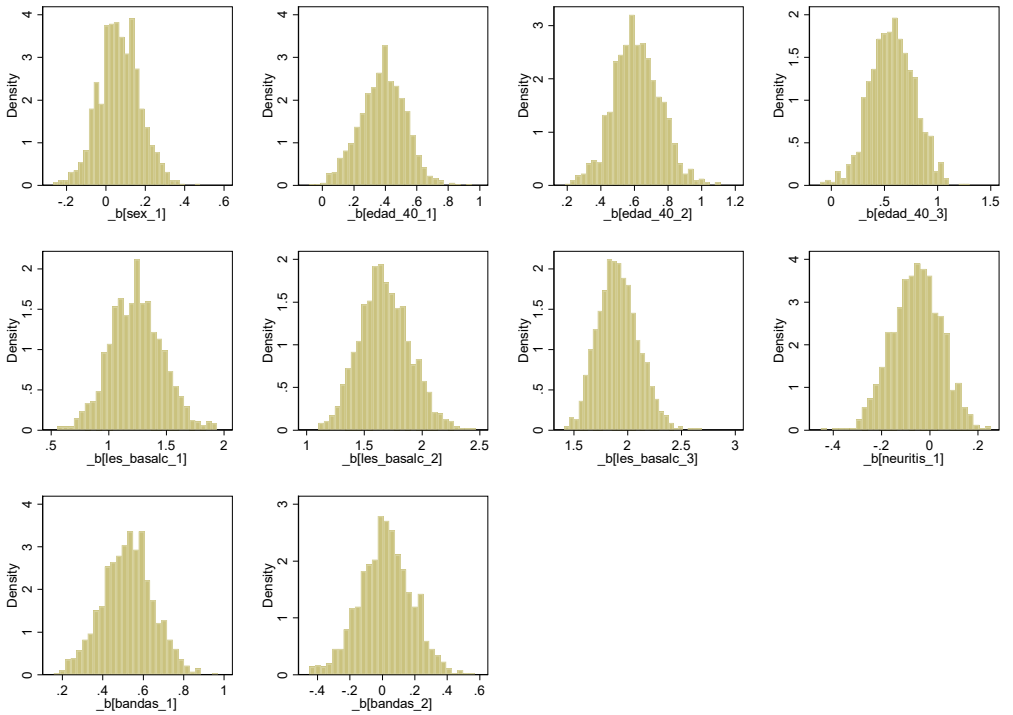
Internal Validation

Leave one-out

Bootstrap



Bootstrap Example



Parametric Weibull
Survival of time to EDSS 2
1000 bootstrap samples

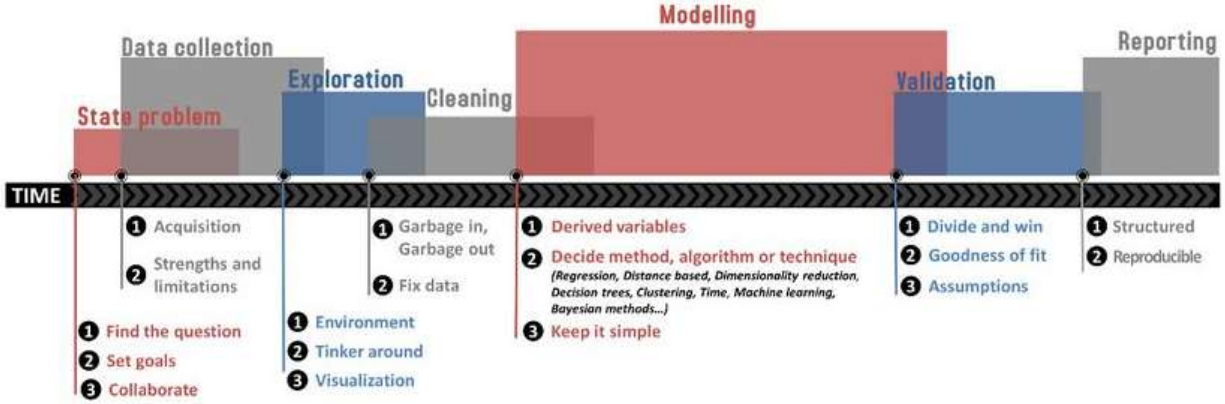
Baseline coefficients after Bootstrap

	Variable	Observados	P05	P25	P50	P75	P95	Mean	Bias	Dif Obs-p50	% Dif
1	Sexe=Female	.4132637	.1630866	.3170085	.4136112	.5124905	.6555218	.4130368	-.0002269	-.0003475	-.0840867
2	Age=30-39 years	.161318	-.1720701	.0296619	.1589391	.3047683	.5007014	.1632275	.0019095	.0023789	1.474665
3	Age=20-29 years	.3775422	.0567566	.2517222	.3823829	.5165071	.6970804	.3847625	.0072202	-.0048407	-1.282161
4	Age=0-19 years	.2654378	-.2432458	.0410203	.2497428	.4776609	.8561057	.2736707	.0082329	.015695	5.912873
5	Baseline T2 Lesions=1-3	.5783235	-.1554955	.2835953	.5926914	.9432228	1.494569	.6760066	.0976831	-.0143679	-2.484405
6	Baseline T2 Lesions=4-9	-.3210706	-.8739628	-.5540106	-.3356344	-.1192428	.2007287	-.3399132	-.0188427	.0145638	-4.536012
7	Baseline T2 Lesions=10+	-.8886151	-1.329505	-1.05744	-.901857	-.7458786	-.5288672	-.908949	-.0203338	.0132419	-1.490173
8	Optic Neuritis=Yes	.2310295	-.0417112	.1188567	.2340286	.3436311	.5018564	.2324119	.0013823	-.0029991	-1.298146
9	Oligoclonal bands=Yes	-.3369213	-.6973094	-.4829128	-.3391455	-.2131469	-.0226448	-.3485017	-.0115804	.0022242	-.6601542
10	Oligoclonal bands=Unknown	-.0158983	-.4162021	-.1977348	-.0224718	.1504095	.4213704	-.0190344	-.0031361	.0065735	-41.34719
11	Intercept	6.471763	6.019392	6.299173	6.490042	6.695776	7.040196	6.503956	.0321932	-.018279	-.2824423
12	Sigma	.9029559	.8109448	.8592739	.8967155	.9334115	.9851352	.8971602	.9918995	.0062404	.6911079
13	1/ sigma	1.107474	1.015089	1.071339	1.115181	1.163773	1.233131	1.118415	1.008167	-.007707	-.695908

External Validation

- The model is aplicable to other Setting?
- It is not easy to find a sample with same measurements from other setting
- Predict values and calculate performance measurements.

Modelling Steps

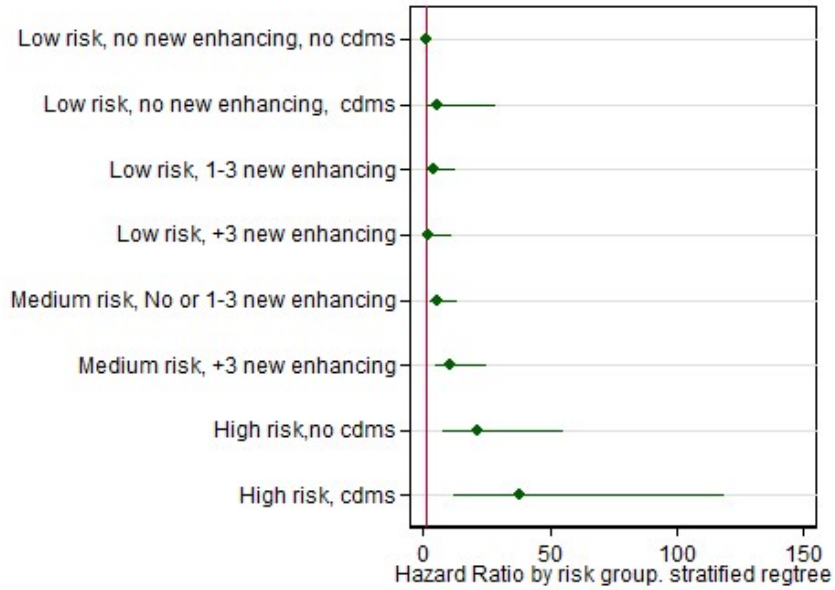


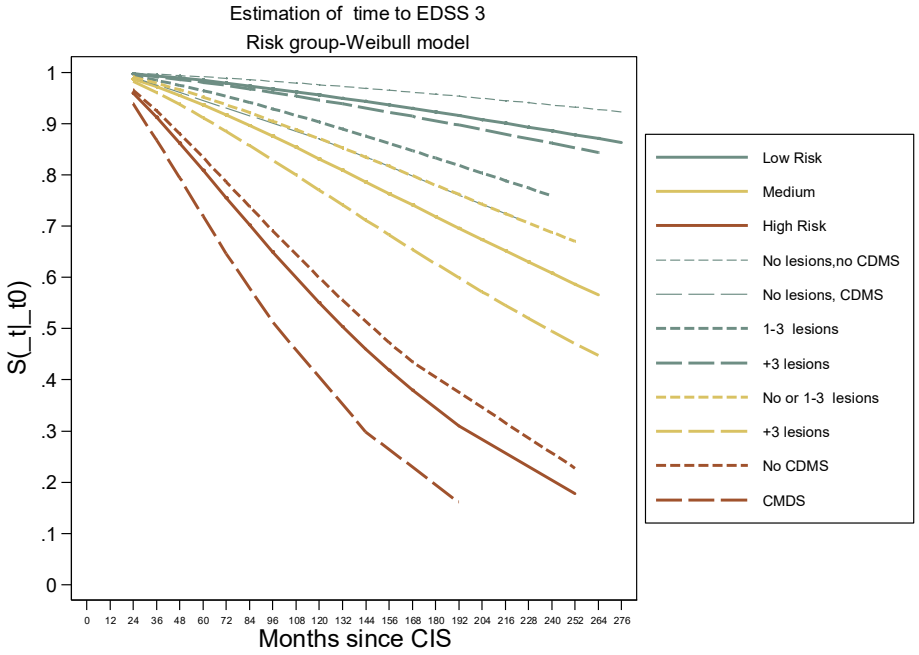
1. **Determining the aim of the model** (Prediction or explanatory)
2. **Ascertainment of the outcome** (binary, continuous,survival, repeated measure, etc.) and data
3. Choosing the **appropriate statistical model** based on design and outcome
4. **Model Estimation** (Missing values,Functional shape, Selection variables,shrinkage , avoid overfitting)
5. **Assessing model performance**
6. **Model Validation** (internal and external)
7. **Presenting Results**

Presenting Results: Some considerations

- Present effect measures(Difference, RR, OR, HR) and there confidence intervals as table or forest plot
- Use absolute measures as NNT
- Include performance measures
- Use Tripod guidelines

Example First Year update risk to EDSS 3 Model



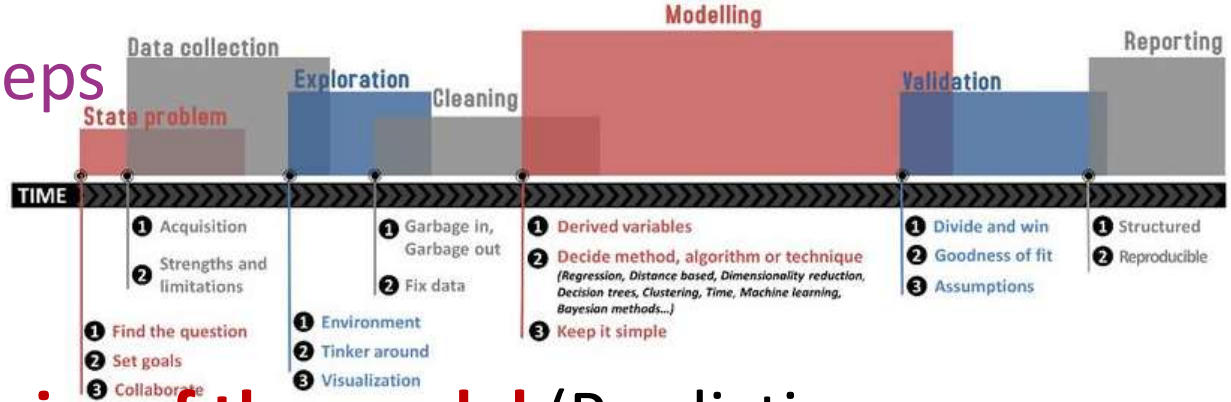


Number of obs = 886

VARIABLE		HR	(95%CI)	p-value
Updated year risk score	Low risk, no new enhancing, no cdms	1		0.0000
	Low risk, no new enhancing, cdms	5.71	(1.15;28.31)	
	Low risk, 1-3 new enhancing	4.20	(1.35;13.02)	
	Low risk, +3 new enhancing	2.27	(0.46;11.25)	
	Medium risk, No or 1-3 new enhancing	5.70	(2.41;13.49)	
	Medium risk, +3 new enhancing	10.73	(4.57;25.22)	
	High risk,no cdms	21.08	(8.00;55.58)	
	High risk, cdms	38.21	(12.28;118.92)	

Options= "d(weibull) "
Sigma = **0.7174** ; Intercept = **-10.3491**
Harell C = 0.7142 ; AIC model= 697.71 ; BIC model= 740.79

Modelling Steps



1. **Determining the aim of the model** (Prediction or explanatory)
2. **Ascertainment of the outcome** (binary, continuous,survival, repeated measure, etc.) and data
3. Choosing the **appropriate statistical model** based on design and outcome
4. **Model Estimation** (Missing values,Functional shape, Selection variables,shrinkage , avoid overfitting)
5. **Assessing model performance (**Discrimation/Calibration)
6. **Model Validation** (internal and external)
7. **Presenting Results**

"Prediction is very difficult, especially if it's about the future."

-- Niels Bohr
Physics Nobel prize 1922

Baby, remember my name

All models are wrong but some are useful

George E.P. Box

Gràcies