

# **Biostatistics**

## **Applications in Medicine**

**Nuno Sepúlveda, 27.10.2025**

# Syllabus

## 1. General review

- a. What is Biostatistics?
- b. Population/Sample/Sample size
- c. Type of Data – quantitative and qualitative variables
- d. Common probability distributions
- e. Work example – Malaria in Tanzania

## 2. Applications in Medicine

- a. Construction and analysis of diagnostic tools – Binomial distribution, sensitivity, specificity, ROC curve, Rogal-Gladen estimator
- b. Estimation of treatment effects - generalized linear models
- c. Survival analysis - Kaplan-Meier curve, log-rank test, Cox's proportional hazards model

## 3. Applications in Genetics, Genomics, and other 'omics data

- a. Genetic association studies – Hardy-Weinberg test, homozygosity, minor allele frequencies, additive model, multiple testing correction
- b. Methylation association studies – M versus beta values, estimation of biological age
- c. Gene expression studies based on RNA-seq experiments – Tests based on Poisson and Negative-Binomial

## 4. Other Topics

- a. Estimation of Species diversity – Diversity indexes, Poisson mixture models
- b. Serological analysis – Gaussian (skew-normal) mixture models
- c. Advanced sample size and power calculations

Prevent

Diagnose

Medicine

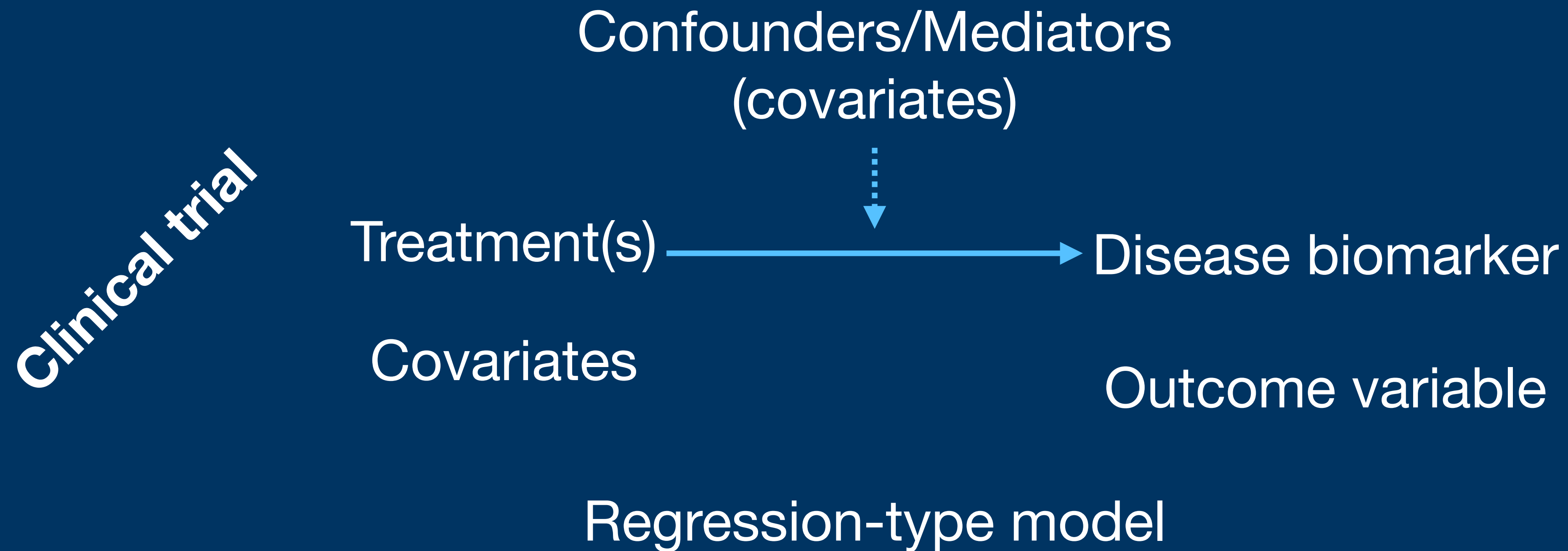
Treat

Improve

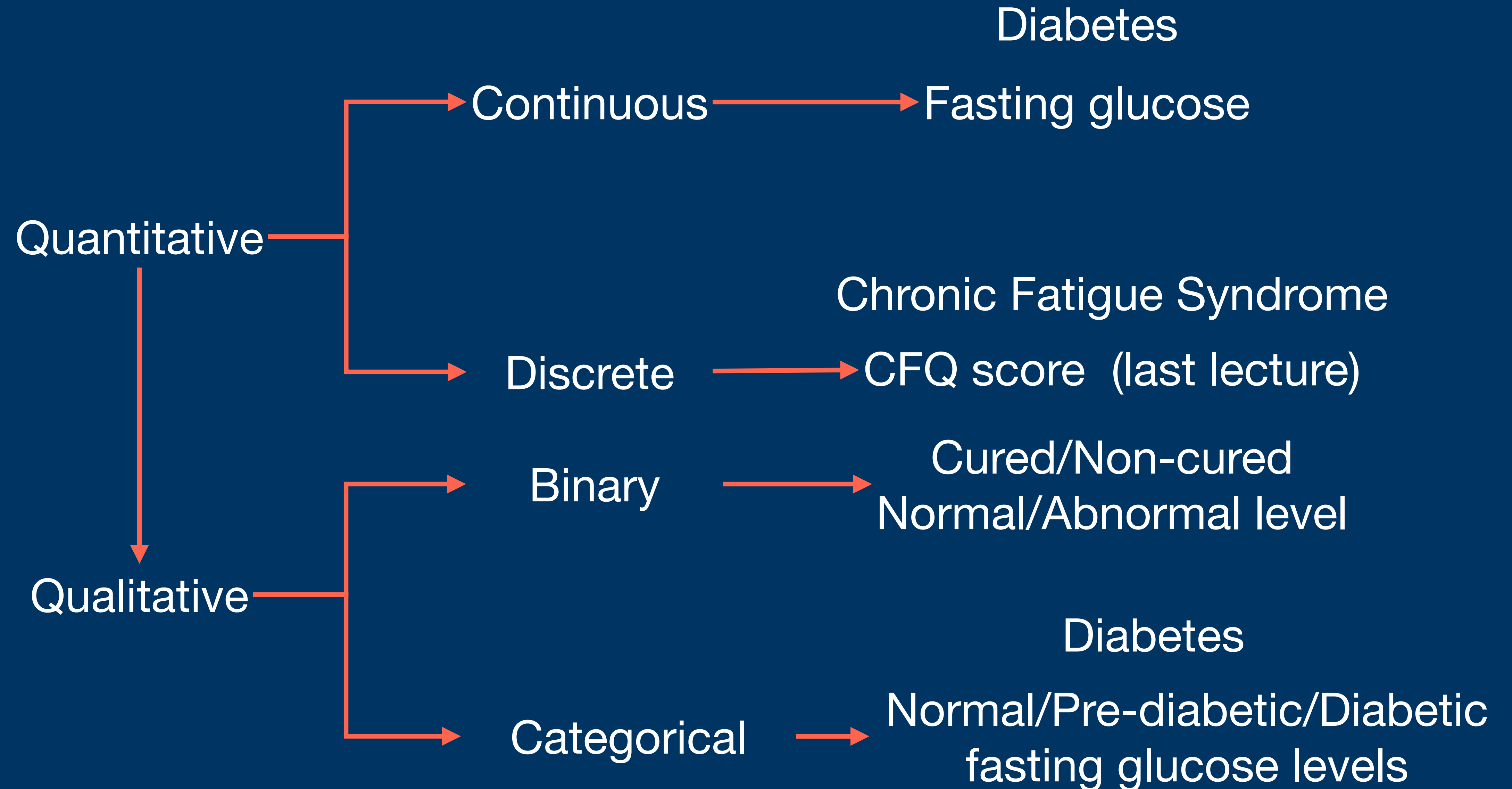
Develop

# Basic question

What are the treatment effects on a disease biomarker?



# Disease biomarker



# Generalised linear models



# Generalised linear models

$$Y | \theta \rightsquigarrow F(\theta)$$

Random component

# Generalised linear models

$Y_1, \dots, Y_n$  Outcomes

$Y_i$  = random variable representing the biomarker value of individual  $i$

$x_{11}, \dots, x_{1p}$

$\vdots \quad \vdots$  Covariates

$x_{n1}, \dots, x_{np}$

$x_{ij}$  = value of covariate  $j$  of individual  $i$



# Generalised linear models

$$Y_i | \theta_i \rightsquigarrow F(\theta_i)$$

Random component

$$g(\theta_i) = \alpha + \sum_{j=1}^p \beta_j x_{ij}$$

Systematic component

$g(\cdot)$  = link function

# Generalised linear models

$$Y_i | \theta_i \rightsquigarrow F(\theta_i)$$

Random component

$$g(\theta_i) = \alpha + \sum_{j=1}^p \beta_j x_{ij}$$

Systematic component

$g(\cdot)$  = link function

$F(\theta)$  should belong to the exponential family of distributions

# Exponential family of distributions

$$f_{X_i}(x | \theta) = h(x) e^{\eta(\theta)T(x) - A(\theta)}$$

The support of the distribution does not depend on the parameter

$\eta(\cdot)$  = canonical link function

# Exercise: Is Bernoulli distribution a member of exponential family?

$$f_{X_i}(x | \pi_i) = \pi_i^x (1 - \pi_i)^{1-x}$$

$$= e^{x \log \pi + (1-x) \log(1-\pi)}$$

$$= e^{x \log \frac{\pi}{1-\pi} + \log(1-\pi)}$$

$$h(x) = 1$$

$$T(x) = x$$

$$\eta(\pi) = \log \frac{\pi}{1-\pi}$$

$$A(\pi) = -\log(1-\pi)$$

# Generalised linear models

What are the main advantages of using these models?

# Popular GLMs: linear regression

$$Y_i | \mu_i, \sigma \rightsquigarrow \text{Normal}(\mu_i, \sigma)$$

Random component

+

$$\mu_i = \alpha + \sum_{j=1}^p \beta_j x_{ij}$$

Systematic component

$$g(\mu_i) = \mu_i$$

Canonical link function

# Popular GLMs: **logistic regression**

$$Y_i | \pi_i \rightsquigarrow \text{Bernoulli}(\pi_i)$$

Random component

+

$$g(\pi_i) = \alpha + \sum_{j=1}^p \beta_j x_{ij}$$

Systematic component

$$g(\pi_i) = \log \frac{\pi_i}{1 - \pi_i}$$

canonical link function

logit

# Popular GLMs: **probit regression**

$$Y_i | \pi_i \rightsquigarrow \text{Bernoulli}(\pi_i)$$

Random component

+

$$g(\pi_i) = \alpha + \sum_{i=1}^p \beta_i x_{ij}$$

Systematic component

$$g(\pi_i) = \Phi^{-1}(\pi_i)$$

Probit link function

where  $\Phi^{-1}(\cdot)$  is the quantile function of a standard Normal distribution



# Popular GLMs: complementary log-log

$$Y | \pi \rightsquigarrow \text{Bernoulli}(\pi)$$

Random component

+

$$g(\pi) = \alpha + \sum_{i=1}^p \beta_i x_i$$

Systematic component

$$g(\pi) = \log(-\log(1 - \pi))$$

Complementary log-log link function

# A theoretical note on the link functions for the Bernoulli model

$$g(\pi_i) = \log \frac{\pi_i}{1 - \pi_i}$$

$$g(\pi_i) = \Phi^{-1}(\pi_i)$$

$$g(\pi) = \log(-\log(1 - \pi))$$

# A theoretical note on the link functions for the Bernoulli model

$$\eta_i = \log \frac{\pi_i}{1 - \pi_i} \Leftrightarrow \pi_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$

$$\eta_i = \Phi^{-1}(\pi_i) \Leftrightarrow \pi_i = \Phi(\eta_i)$$

$$\eta_i = \log(-\log(1 - \pi_i)) \Leftrightarrow \pi_i = 1 - e^{-e^{\eta_i}}$$

# A theoretical note on the link functions for the Bernoulli model

$$\eta_i = \log \frac{\pi_i}{1 - \pi_i} \Leftrightarrow \pi_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$

Cumulative distribution of a standard  
Logistic distribution

$$\eta_i = \Phi^{-1}(\pi_i) \Leftrightarrow \pi_i = \Phi(\eta_i)$$

Cumulative distribution of a Standard  
Normal distribution

$$\eta_i = \log(-\log(1 - \pi_i)) \Leftrightarrow \pi_i = 1 - e^{-e^{\eta_i}}$$

1- Cumulative distribution of an Extreme  
Value distribution

# A theoretical note on the link functions for the Bernoulli model

$$\eta_i = \log \frac{\pi_i}{1 - \pi_i} \Leftrightarrow \pi_i = \frac{1}{1 + e^{-\eta_i}}$$

Cumulative distribution of a standard  
Logistic distribution

$$\eta_i = \Phi^{-1}(\pi_i) \Leftrightarrow \pi_i = \Phi(\eta_i)$$

Cumulative distribution of a Standard  
Normal distribution

$$\eta_i = \log(-\log(1 - \pi_i)) \Leftrightarrow \pi_i = 1 - e^{-e^{\eta_i}}$$

1- Cumulative distribution of an Extreme  
Value distribution

**Practical Implication:** The inverse of any cumulative probability distribution can be used as a link function

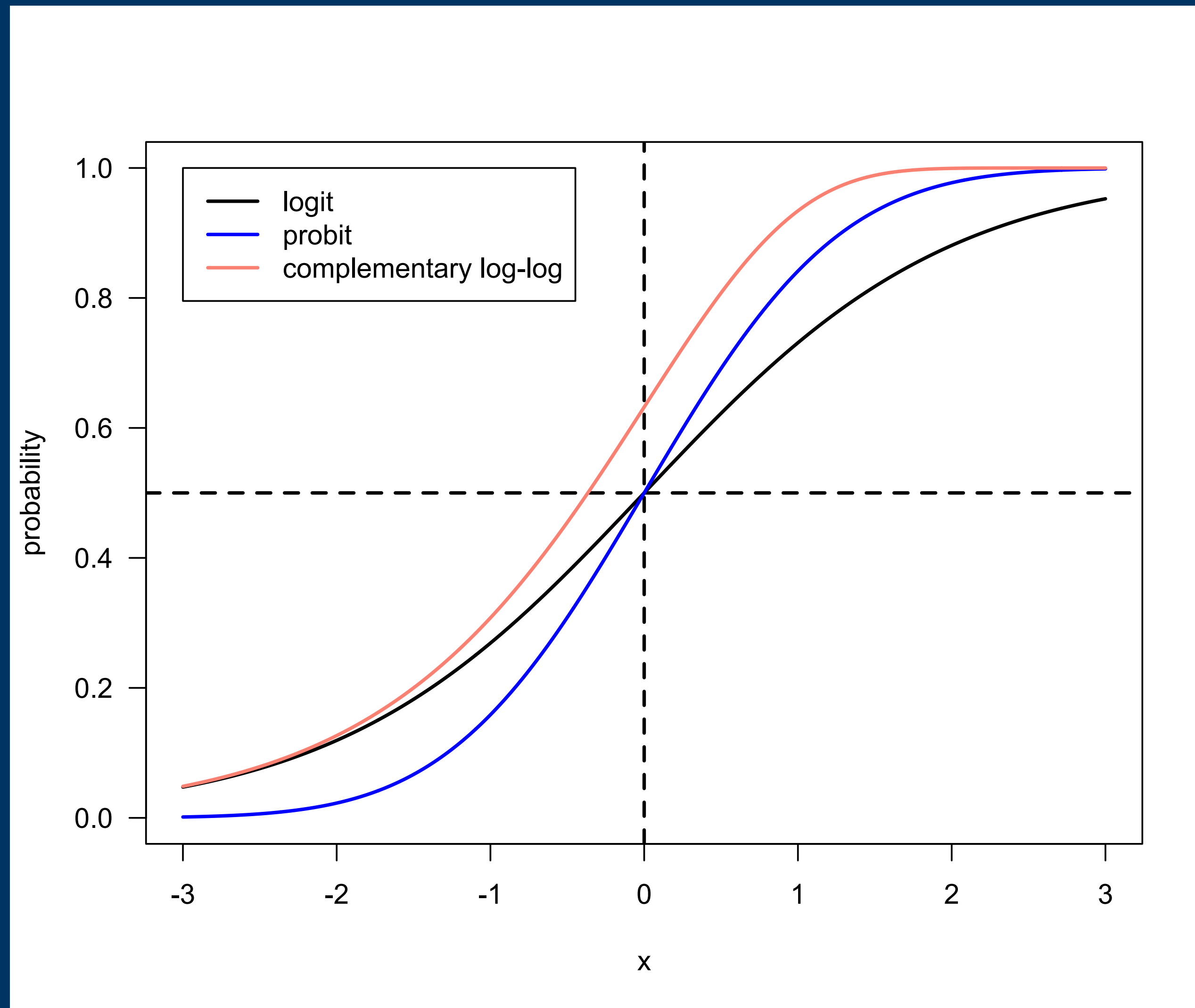
## Exercise:

Construct a link function based on the cumulative probability function of an Exponential distribution and a Weibull distribution.

$$f_{\lambda}(x) = 1 - e^{-\lambda x}$$

$$f_{\lambda,k}(x) = 1 - e^{-\left(\frac{x}{\lambda}\right)^k}$$

# A practical note on the link functions for the Bernoulli model



# Exercise:

RESEARCH ARTICLE

## B-Lymphocyte Depletion in Myalgic Encephalopathy/ Chronic Fatigue Syndrome. An Open-Label Phase II Study with Rituximab Maintenance Treatment

Øystein Fluge<sup>1\*</sup>, Kristin Risa<sup>1</sup>, Sigrid Lunde<sup>1</sup>, Kine Alme<sup>1</sup>, Ingrid Gurvin Rekeland<sup>1</sup>, Dipak Sapkota<sup>1,2</sup>, Einar Kleboe Kristoffersen<sup>3,4</sup>, Kari Sørland<sup>1</sup>, Ove Bruland<sup>1,5</sup>, Olav Dahl<sup>1,4</sup>, Olav Mella<sup>1,4\*</sup>

- 1 Department of Oncology and Medical Physics, Haukeland University Hospital, Bergen, Norway,
- 2 Department of Clinical Medicine, University of Bergen, Haukeland University Hospital, Bergen, Norway,
- 3 Department of Immunology and Transfusion Medicine, Haukeland University Hospital, Bergen, Norway,
- 4 Department of Clinical Science, University of Bergen, Haukeland University Hospital, Bergen, Norway,
- 5 Department of Medical Genetics and Molecular Medicine, Haukeland University Hospital, Bergen, Norway





# Exercise:

## Abstract

### Background

Myalgic Encephalopathy/Chronic Fatigue Syndrome (ME/CFS) is a disease of unknown etiology. We previously reported a pilot case series followed by a small, randomized, placebo-controlled phase II study, suggesting that B-cell depletion using the monoclonal anti-CD20 antibody rituximab can yield clinical benefit in ME/CFS.

### Methods

In this single-center, open-label, one-armed phase II study (NCT01156909), 29 patients were included for treatment with rituximab (500 mg/m<sup>2</sup>) two infusions two weeks apart, followed by maintenance rituximab infusions after 3, 6, 10 and 15 months, and with follow-up for 36 months.

### Findings

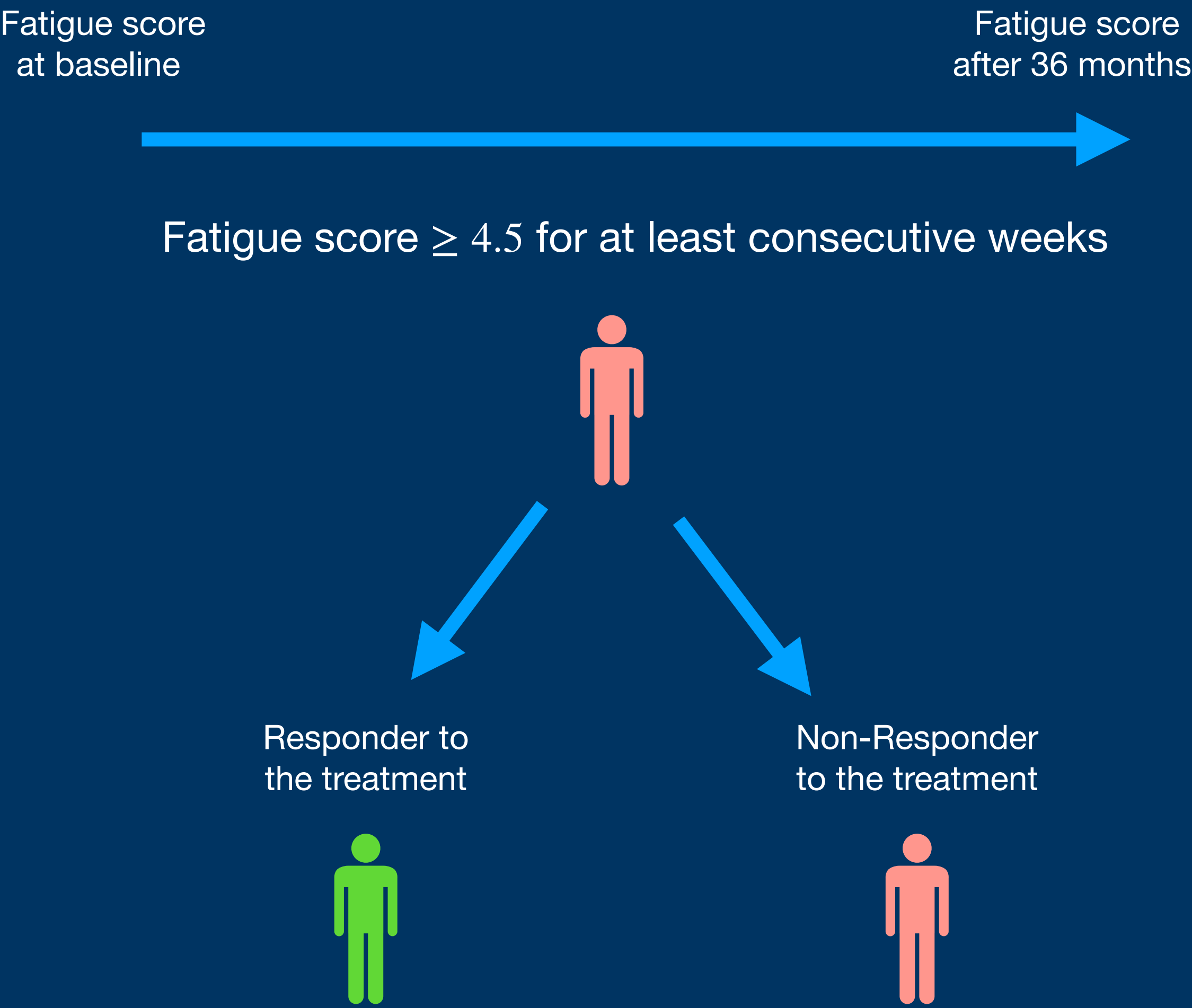
Major or moderate responses, predefined as lasting improvements in self-reported *Fatigue score*, were detected in 18 out of 29 patients (intention to treat). Clinically significant responses were seen in 18 out of 28 patients (64%) receiving rituximab maintenance treatment. For these 18 patients, the mean response durations within the 156 weeks study period were 105 weeks in 14 major responders, and 69 weeks in four moderate responders. At end of follow-up (36 months), 11 out of 18 responding patients were still in ongoing clinical remission. For major responders, the mean lag time from first rituximab infusion until start of clinical response was 23 weeks (range 8–66). Among the nine patients from the placebo group in the previous randomized study with no significant improvement during 12

# Exercise:

Rituximab (n=29)



Biomarker  
Fatigue score



## Let's analyse the data

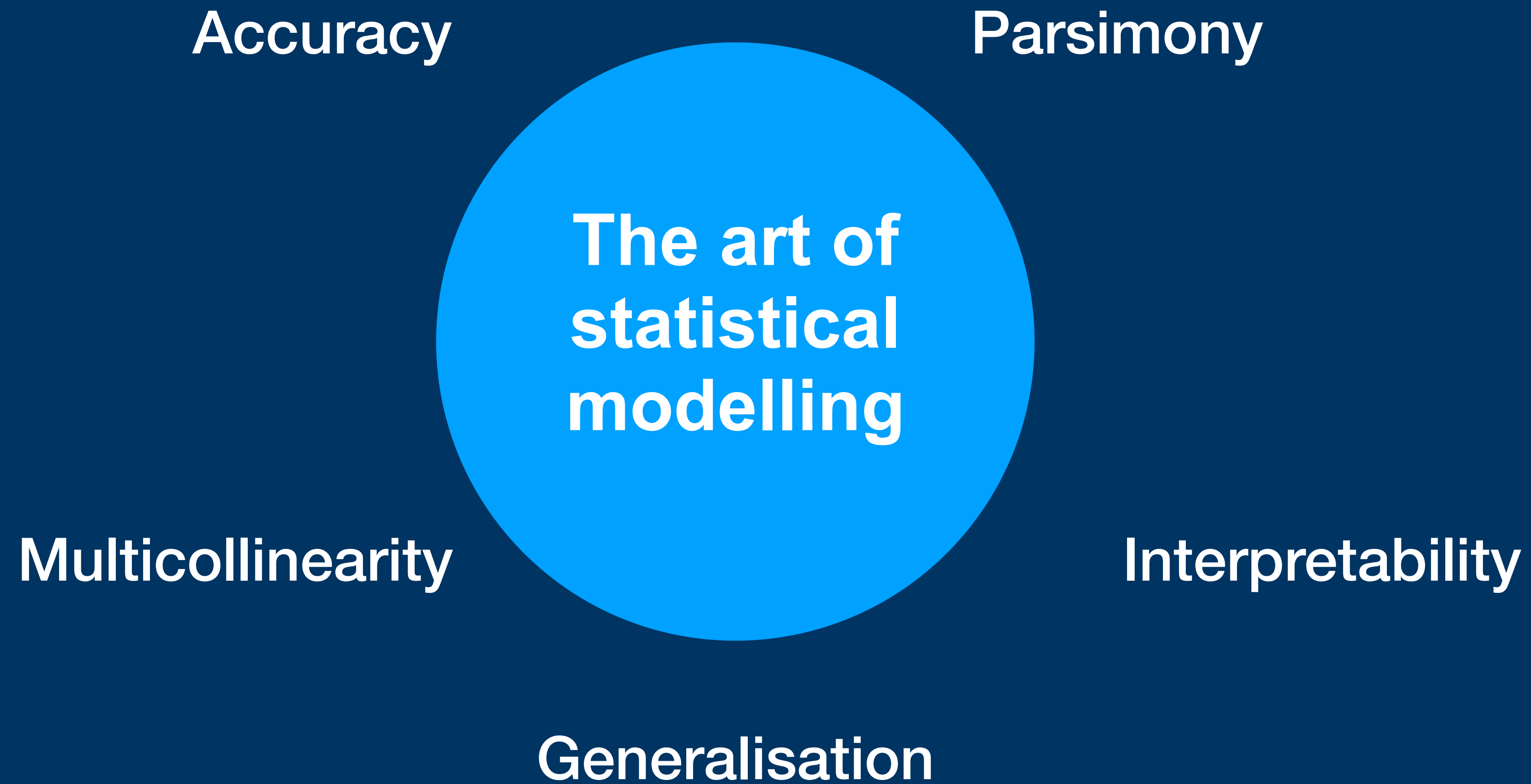
dataset: data\_mecfs\_rituximab.csv:

Estimate the probability of treatment response using statistical inference methods for the binomial distribution?

Use `binom.test` or `prop.test` functions

Construct an appropriate regression model to understand whether age, gender, disease duration affect the treatment success?

Use `glm` function



# The art of constructing a model

## Select the best link function

- Fit models with different link functions and compare them

- Fit models with flexible link functions (e.g., Aranda-Ordaz link function for Bernoulli models)

## Select the best subset of covariates (feature selection)

- Forward/Backward/Stepwise Regression

- Penalised regression (LASSO or Elastic-Net)

# The art of constructing a model

## Select the best link function

- Fit models with different link functions and compare them

- Fit models with flexible link functions (e.g., Aranda-Ordaz link function for Bernoulli models)

## Select the best subset of covariates (feature selection)

- Forward/Backward/Stepwise Regression

- Penalised regression (LASSO or Elastic-Net)

# Model comparison and selection

**AIC - Akaike's Information Criterion**

$$\text{AIC}(M) = (-2)\log\text{-L}(\hat{\theta} | M, \mathbf{x}) + 2p$$

**BIC - Bayesian Information Criterion**

$$\text{BIC}(M) = (-2)\log\text{-L}(\hat{\theta} | M, \mathbf{x}) + p \log(n)$$

$\log\text{-L}(\hat{\theta} | M, \mathbf{x})$  is the log-likelihood function evaluated on the parameter estimates

$p$  is the number of parameters of model  $M$

$n$  is the sample size

**Choose the model with the lowest values of one of these measures**



# Forward selection

“Empty” Model

Add covariate

Add covariate

Add covariate

⋮

Stop procedure

Increased accuracy **compensates**  
increased model complexity

Increased accuracy **does not compensate**  
increased model complexity



# Backward elimination

“All covariates” Model

Remove covariate

Remove covariate

Remove covariate

⋮

Stop procedure

Decreased model complexity **does not have** an impact on model accuracy

Decreased model complexity **has an impact** on model accuracy

# Stepwise regression

“Empty” Model

Add covariate 1

Add covariate 2

Remove covariate 1

Add covariate 3

Remove covariates 1, 2

⋮

Stop procedure



Increased accuracy **compensates**  
increased model complexity

Increased accuracy **does not compensate**  
increased model complexity

# Stepwise regression

## Advantages

Remove multicollinearity

Easy automation

Speed

## Disadvantages

Overestimation of the number of predictors

Inflated type I errors

Unstable to slight changes in the data

# Model comparison and selection

**AIC - Akaike's Information Criterion**

$$\text{AIC}(M) = (-2)\log\text{-L}(\hat{\theta} | M, \mathbf{x}) + 2p$$

**BIC - Bayesian Information Criterion**

$$\text{BIC}(M) = (-2)\log\text{-L}(\hat{\theta} | M, \mathbf{x}) + p \log(n)$$

$\log\text{-L}(\hat{\theta} | M, \mathbf{x})$  is the log-likelihood function evaluated on the parameter estimates

$p$  is the number of parameters of model  $M$

$n$  is the sample size

**Choose the model with the lowest values of one of these measures**