Biostatistics

Applications in Medicine

Nuno Sepúlveda, 26.10.2023

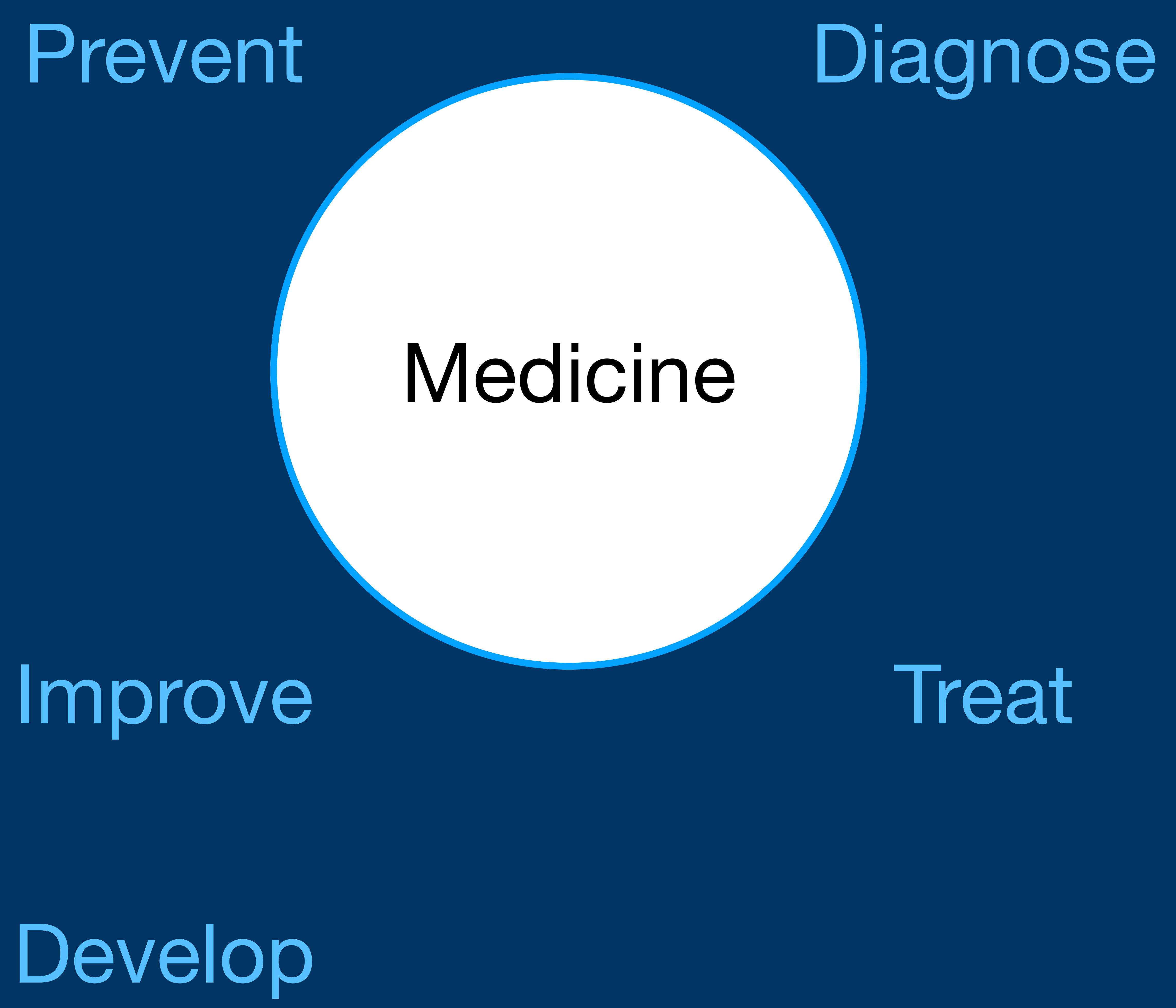
Syllabus

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
Improve
Treat
Develop

Basic question

What are the treatment effects on a disease biomarker?

Confounders/Mediators
(covariates)

Clinical trial

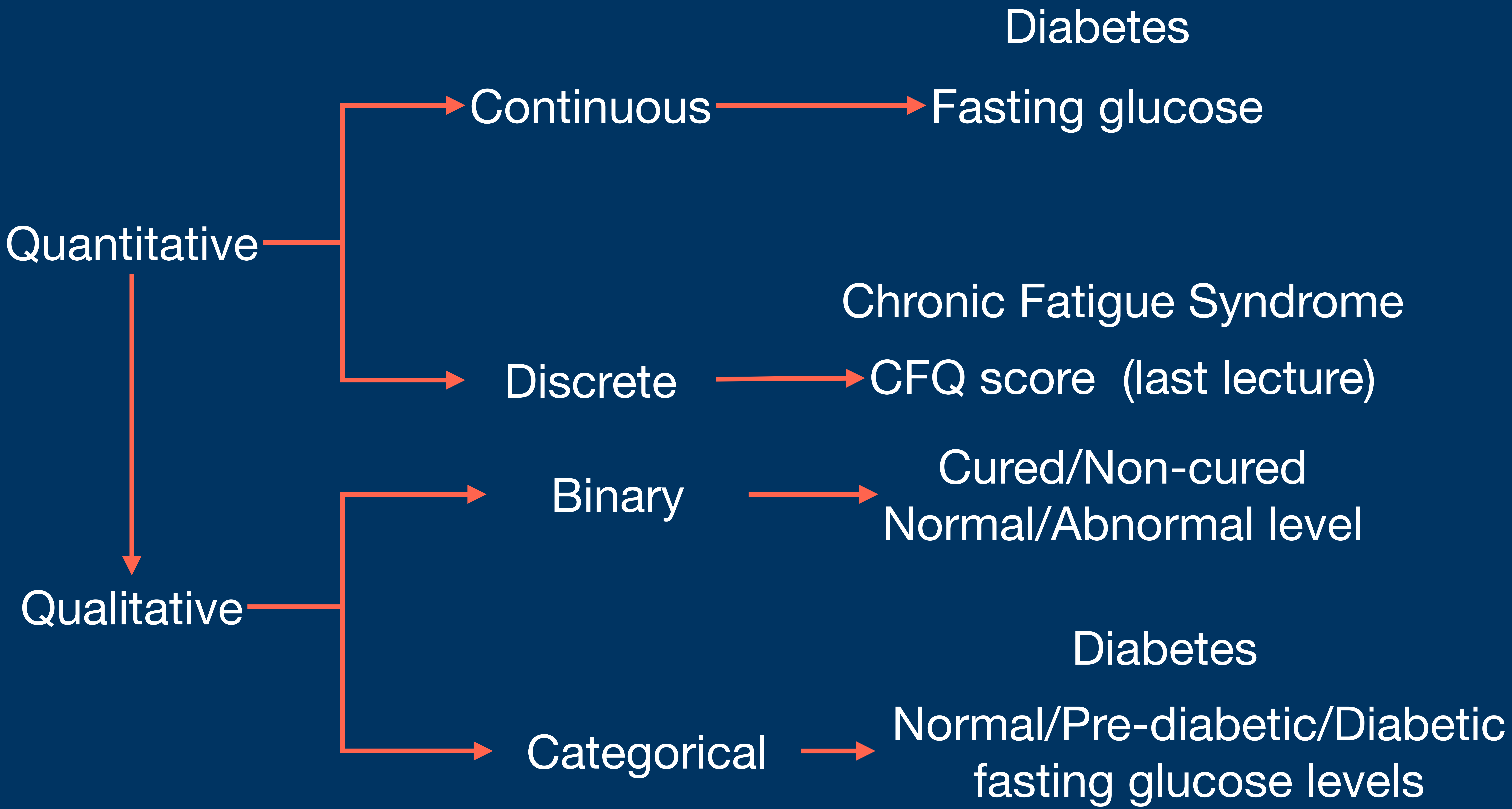
Treatment(s) → Disease biomarker

Covariates

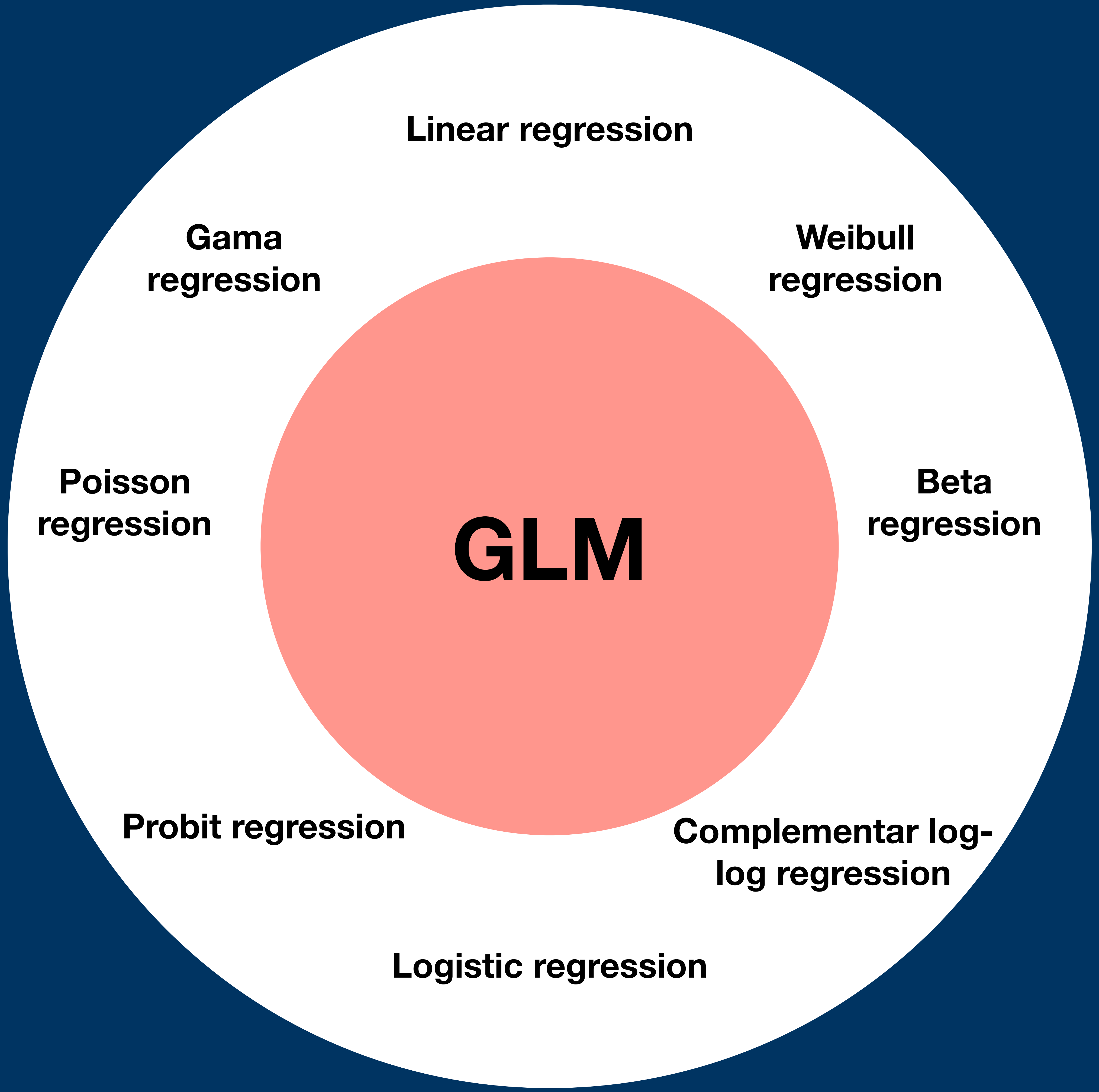
Outcome variable

Regression-type model

Disease biomarker



Generalised linear models



Generalised linear models

$Y \mid \theta \rightsquigarrow F(\theta)$ Random component

Generalised linear models

$Y_1, \ldots, Y_n$ Outcomes

Y_i = random variable representing the biomarker value of individual i

$$\begin{matrix} x_{11}, \ldots, x_{1p} \\ \vdots \qquad \vdots \\ x_{n1}, \ldots, x_{np} \end{matrix}$$

Covariates

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

Generalised linear models

$$Y_i \mid \theta_i \rightsquigarrow F\left(\theta_i\right)$$

Random component

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

Systematic component

$g\left(\,\cdot\,\right) =$ link function

Generalised linear models

$Y_i \,|\, \theta_i \rightsquigarrow F\left(\theta_i\right)$ Random component

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

$g(\,\cdot\,) =$ link function

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

Exponential family of distributions

$$f_{X_i}\left(x \,|\, \theta_i\right) = h\left(x\right) e^{\eta(\theta_i)T(x) - A(\theta_i))}$$

The support of the distribution does not depend on the parameter

$\eta\left(\,\cdot\,\right) =$ canonical link function

Exercise: Is Bernoulli distribution a member of exponential family?

$$f_{X_i}\left(x \mid \pi_i\right) = \pi_i^x (1 - \pi_i)^{1-x}$$

Generalised linear models

What are the main advantages of using these models?

Popular GLMs: linear regression

$Y_i \mid \mu_i, \sigma \rightsquigarrow \text{Normal}(\mu_i, \sigma)$ Random component

+

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

$g\left(\mu_i\right) = \mu_i$ Canonical link function

Popular GLMs: logistic regression

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

Random component

+

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

Systematic component

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

canonical link function

logit

Popular GLMs: probit regression

$Y_i \mid \pi_i \rightsquigarrow \text{Bernoulli}(\pi_i)$ Random component

+

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

$g\left(\pi_i\right) = \Phi^{-1}(\pi_i)$ Probit link function

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

Popular GLMs: complementary log-log

$Y \mid \pi \rightsquigarrow \text{Bernoulli}(\pi)$ Random component

+

$g(\pi) = \alpha + \sum_{i=1}^{p} \beta_1 x_i$ Systematic component

$g(\pi) = \log(-\log(1 - \pi))$ Complementary log-log link function

A theoretical note on the link functions for the Bernoulli model

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

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

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

A theoretical note on the link functions for the Bernoulli model

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

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

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

A theoretical note on the link functions for the Bernoulli model

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

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

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

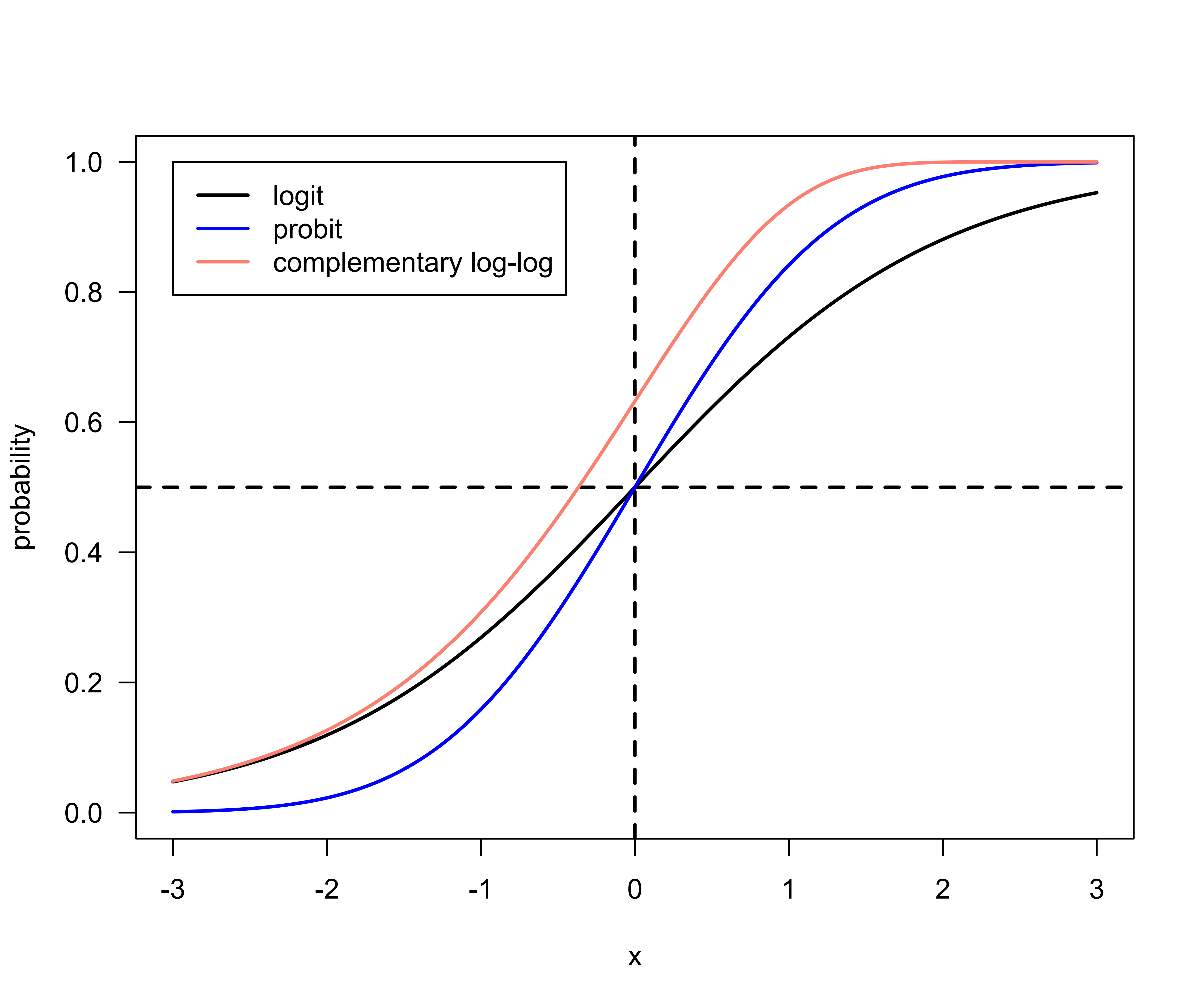
Cumulative distribution of a Standard Normal distribution

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

1- Cumulative distribution of an Extreme Value distribution

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

A practical note on the link functions for the Bernoulli model



Exercise:

PLOS | ONE

RESEARCH ARTICLE

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

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

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

CrossMark
click for updates

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

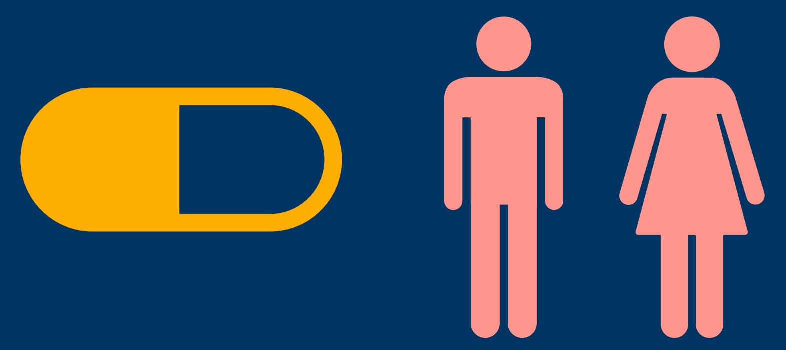
In this single-center, open-label, one-armed phase II study (NCT01156909), 29 patients were included for treatment with rituximab (500 mg/m^2) two infusions two weeks apart, followed by maintenance rituximab infusions after 3, 6, 10 and 15 months, and with follow-up for 36 months.

Findings

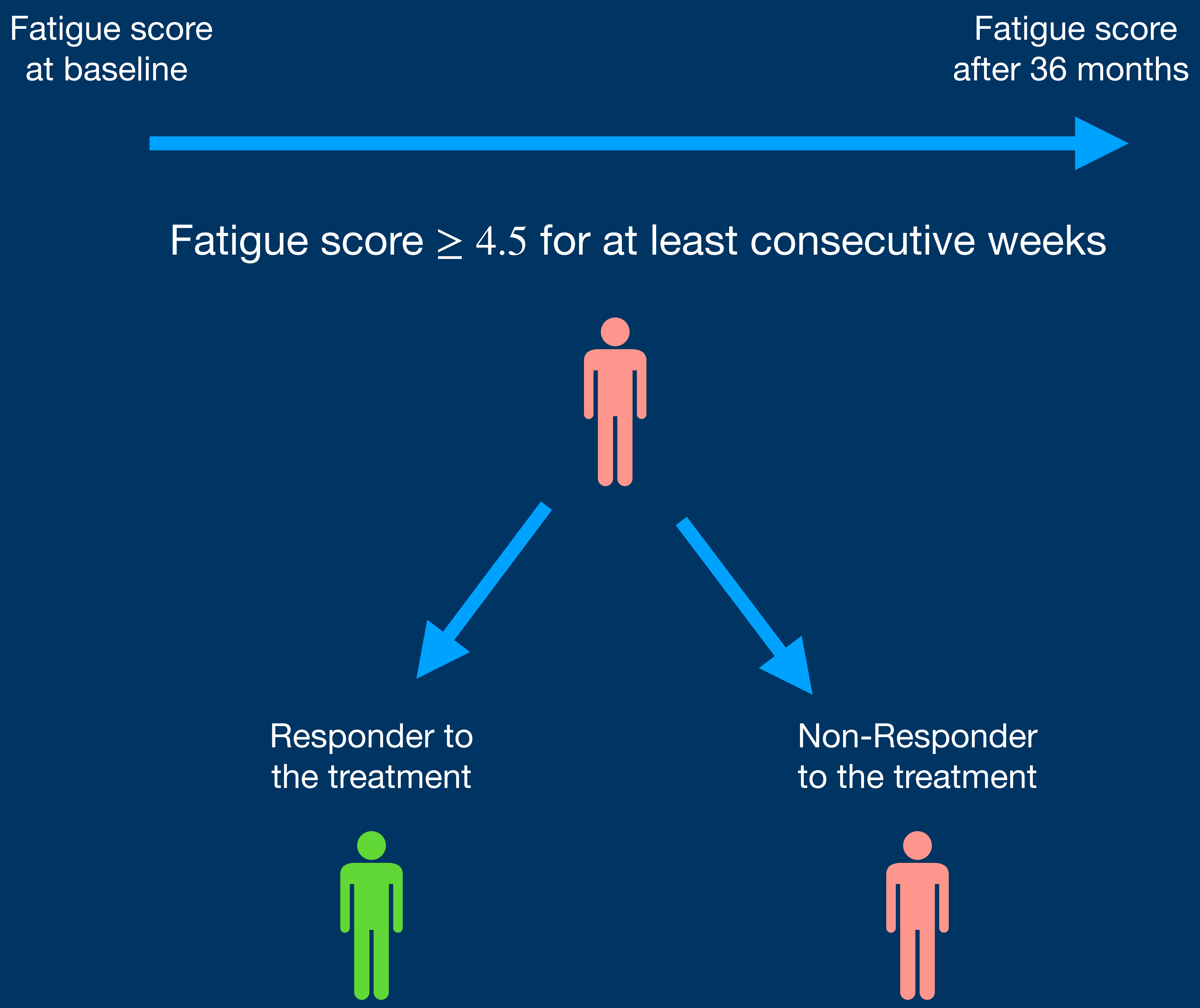
Major or moderate responses, predefined as lasting improvements in self-reported *Fatigue score*, were detected in 18 out of 29 patients (intention to treat). Clinically significant responses were seen in 18 out of 28 patients (64%) receiving rituximab maintenance treatment. For these 18 patients, the 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 a logistic regression model to understand whether age, gender, disease duration affect the treatment success?

Use glm function