Biostatistics

Applications in Medicine

Nuno Sepúlveda, 20.11.2023

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

Weibull regression model

Log-linear formulation (similar to linear regression)

$$\log T_i = \beta_0 + \sum_j \beta_j x_{ij} + \sigma_0 \epsilon_i \qquad \epsilon_i \mid \rightsquigarrow \text{Gumbel}(\mu = 0, \sigma = 1)$$

$$\log T_i \rightsquigarrow \text{Gumbel}\left(\mu = \beta_0 + \sum_j \beta_j x_j, \sigma = \sigma_0\right)$$

(see slide 17)

$$T_i \rightsquigarrow \text{Weibull}\left(\lambda = \exp\left\{\frac{\beta_0 + \sum_j \beta_j x_j}{\sigma}\right\}, \gamma = \frac{1}{\sigma}\right)$$

Weibull regression model as a proportional hazard model

$$T_i \rightsquigarrow \text{Weibull}\left(\lambda = \exp\left\{\frac{\beta_0 + \sum_j \beta_j x_j}{\sigma}\right\}, \gamma = \frac{1}{\sigma}\right)$$

$$h_{\gamma,\lambda}(t) = \frac{\gamma}{\lambda}\left(\frac{t}{\lambda}\right)^{\gamma-1}, \; t > 0$$

$$h_{\gamma,\left\{\beta_j\right\}}(t) = \frac{1}{\sigma e^{\frac{\beta_0 + \sum_j \beta_j x_j}{\sigma}}}\left(\frac{t}{e^{\frac{\beta_0 + \sum_j \beta_j x_j}{\sigma}}}\right)^{\frac{1}{\sigma}-1}$$

Weibull regression model as a proportional hazard model

$$h_{\gamma,\left\{\beta_j\right\}}(t) = \frac{1}{\sigma e^{\frac{\beta_0 + \sum_j \beta_j x_j}{\sigma}}} \left(\frac{t}{e^{\frac{\beta_0 + \sum_j \beta_j x_j}{\sigma}}} \right)^{\frac{1}{\sigma}-1}$$

$$= \frac{1}{\sigma e^{\frac{\beta_0}{\sigma}}} \left(\frac{t}{e^{\frac{\beta_0}{\sigma}}} \right)^{\frac{1}{\sigma}-1} \left(\frac{1}{e^{\frac{\sum_j \beta_j x_j}{\sigma}}} \right)^{\frac{1}{\sigma}}$$

$$= \underbrace{\frac{1}{\sigma e^{\frac{\beta_0}{\sigma}}} \left(\frac{t}{e^{\frac{\beta_0}{\sigma}}} \right)^{\frac{1}{\sigma}-1}}_{h_0(t)} \times \underbrace{e^{-\frac{\sum_j \beta_j x_j}{\sigma^2}}}_{\text{effect of covariates}}$$

Estimation and statistical validation

Maximum likelihood estimation using numerical methods (e.g., Newton-Raphson)

$$\left\{\hat{\beta}_j, j = 0,\ldots,p\right\}, \hat{\sigma}$$

Validation of the model

Standardized residuals: $\hat{e}_i = \dfrac{\log t_i - \widehat{\log} t_i}{\hat{\sigma}}$

they should follow a Gumbel distribution with mu=0 and sigma=1

Cox-Snel residuals: $\tilde{e}_i = \left(t_i e^{-\widehat{\log} t_i}\right)^{1/\hat{\sigma}}$

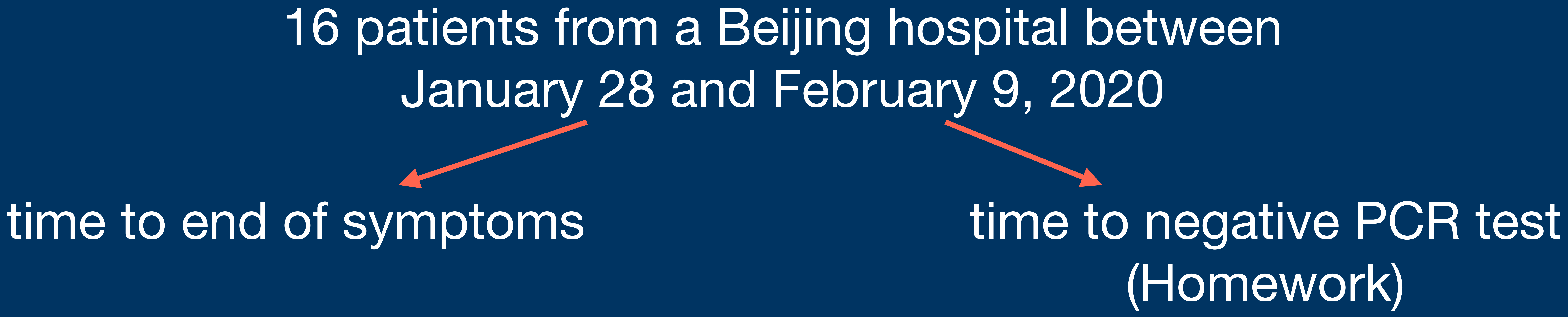
they should follow a Exponential distribution with parameter 1 (see slide 16 of previous lecture)

Weibull regression model is not a generalized linear model

Weibull distribution is not a member of the exponential family.

Homework!

Exercise 1: data about recovery from a SARS-CoV-2 infection



16 patients from a Beijing hospital between
January 28 and February 9, 2020

time to end of symptoms

time to negative PCR test
(Homework)

Package survival (survreg function)

Fit a Weibull regression model with time to end of symptoms as the outcome and age and gender as the covariate. Draw your conclusions.

Check the model validity by testing a Gumbel distribution in the standardized residuals and exponential distribution in the Cox-Snel residuals

Endpoint:
time to event

Which the study design should be used?

What are the practical problems of this study design?

Truncated versus censored data

Truncated data

The last time observed in an individual is exactly the end of follow-up

Truncation is defined by study design

Censored data

The last time observed in an individual is within the period of follow-up

Censoring is likely “uncontrolled” truncation caused by external uncontrolled factors to the study)

Truncated data

Prospective/longitudinal studies

$t_i \in (\tau, +\infty)$

Retrospective/cross-sectional studies

$t_i \in (0,\tau)$

τ is the length of the study

Types of censored data

Censored data

$$t_i \in \left(t_i^*, +\infty\right)$$

Right censoring
(when some individuals drop out from the study)

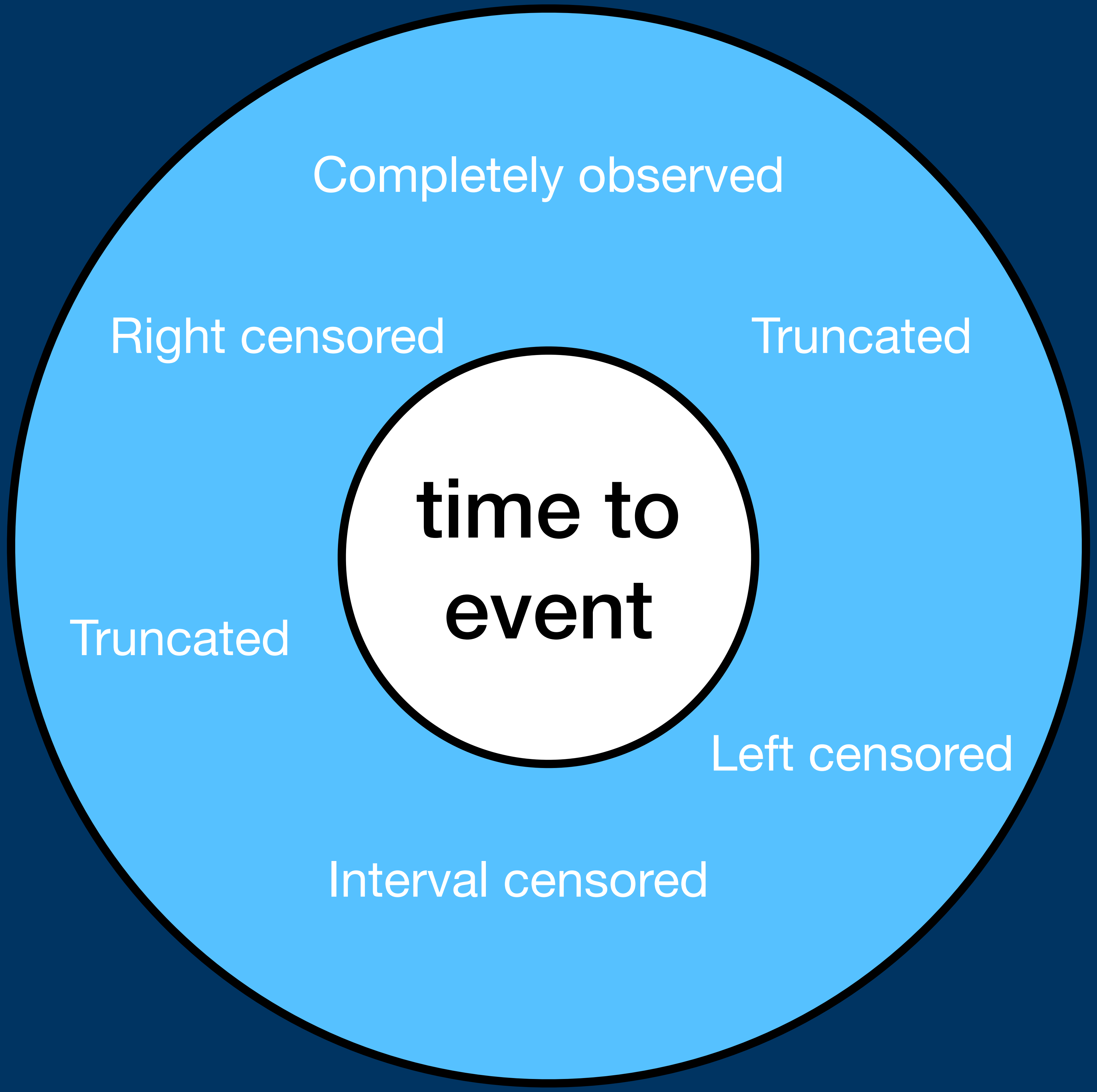
$$t_i \in \left(0, t_i^+\right)$$

Left censoring
(when the event of interest already occurred in some of the individuals)

$$t_i \in \left(t_i^*, t_i^+\right)$$

Intervalar censoring
(when the monitoring of the event is done in intervals)

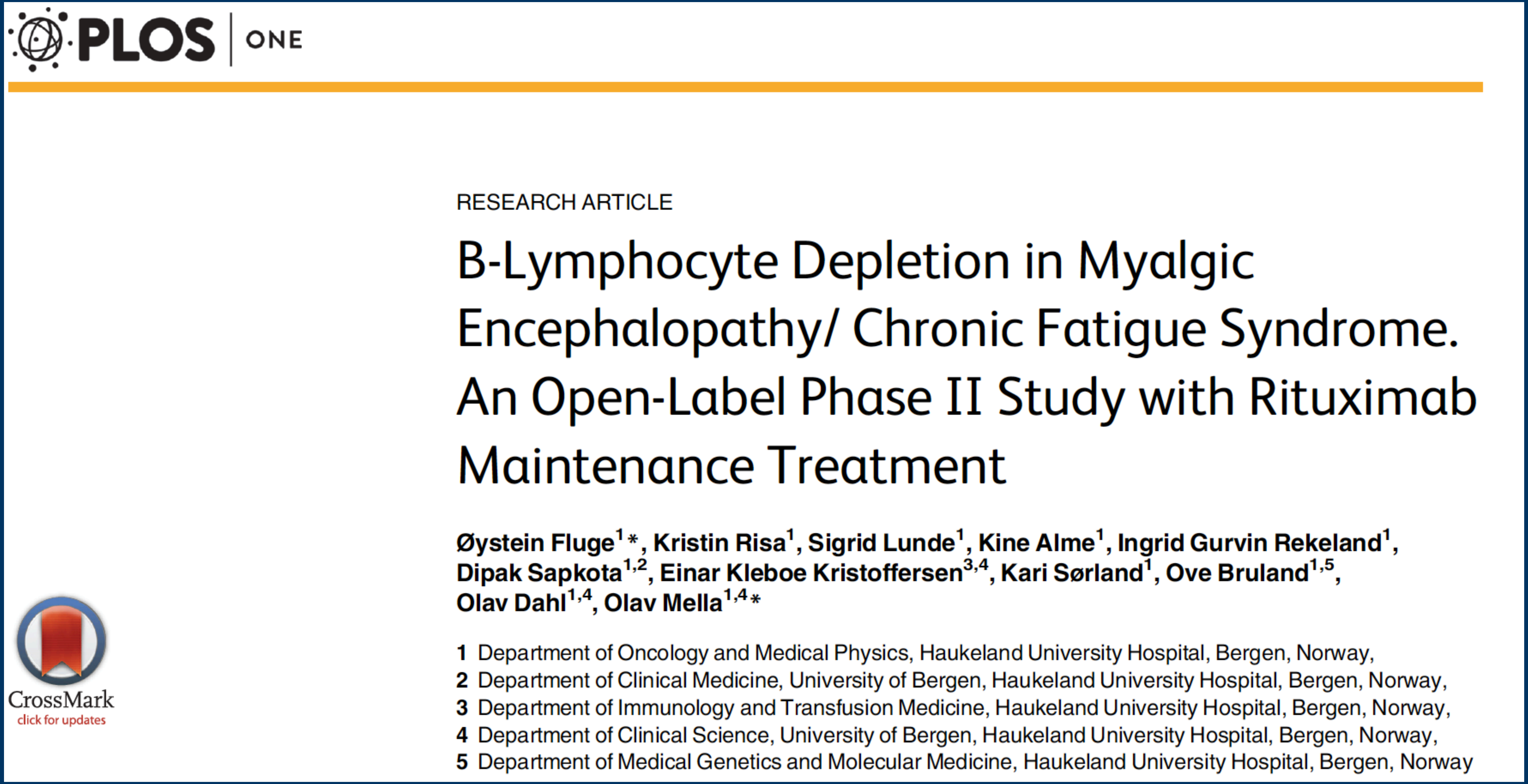
Real-world survival studies included incomplete data of multiple sources



Basic assumption

The occurrence of censoring is independent of the process leading to the observation of event of interest

Example: Rituximab clinical trial



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]*

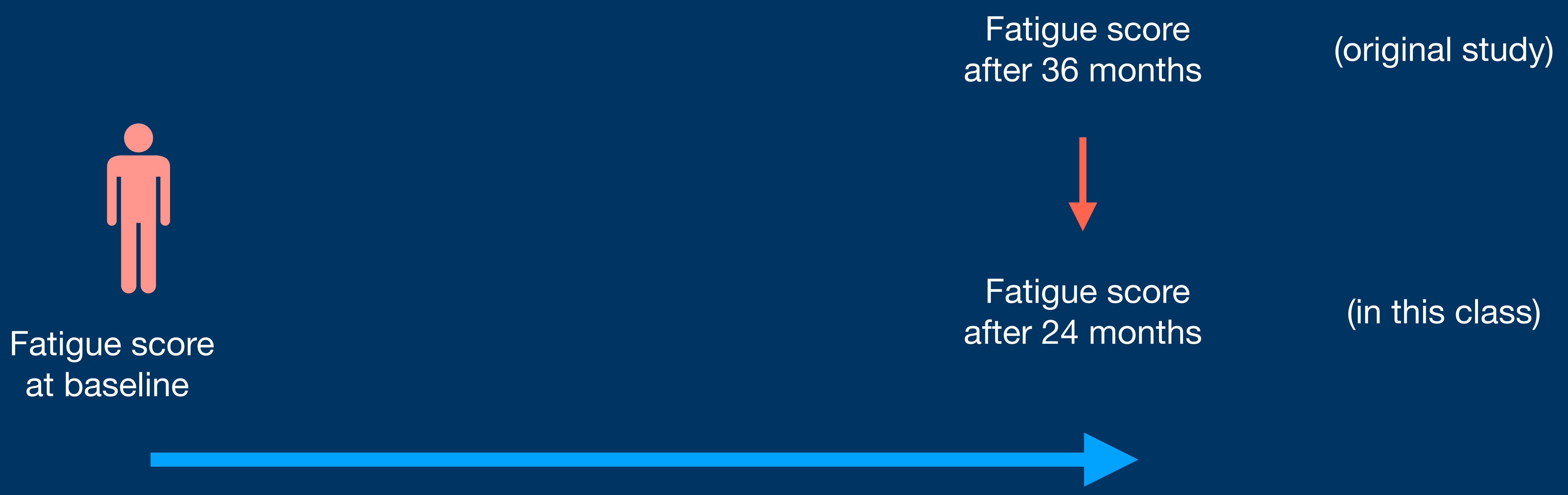
1 Department of Oncology and Medical Physics, Haukeland University Hospital, Bergen, Norway,
2 Department of Clinical Medicine, University of Bergen, Haukeland University Hospital, Bergen, Norway,
3 Department of Immunology and Transfusion Medicine, Haukeland University Hospital, Bergen, Norway,
4 Department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

Study design and follow-up:

Rituximab (n=29)

Biomarker
Fatigue score



Fatigue score
after 36 months

(original study)

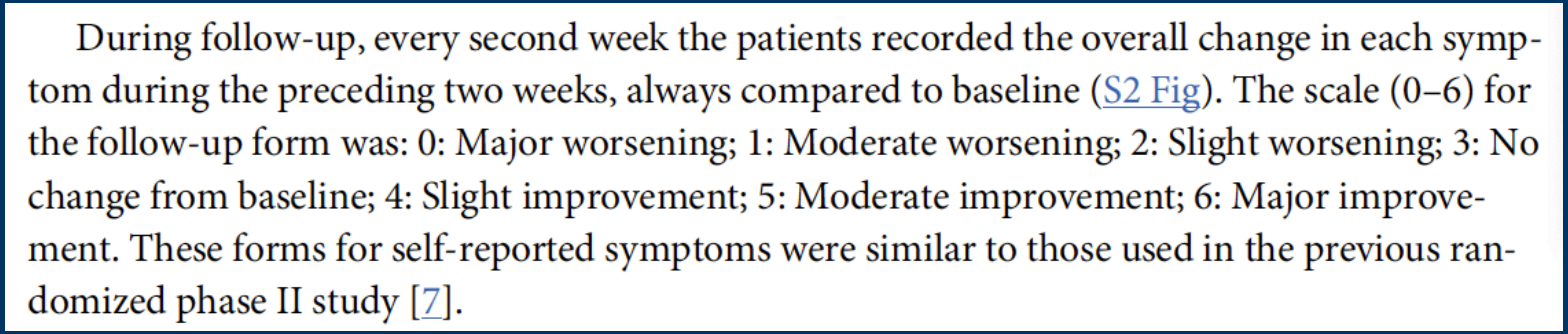
Fatigue score
after 24 months

(in this class)

Fatigue score
at baseline

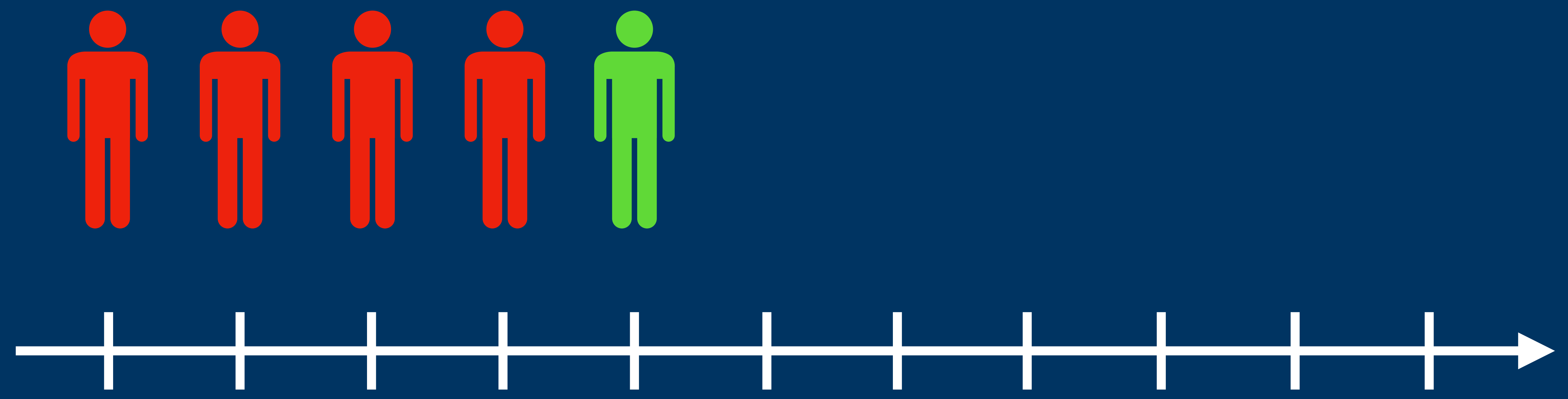
Treatment response: Fatigue Score > 4.

Endpoint:
Time to first positive response
to treatment



During follow-up, every second week the patients recorded the overall change in each symptom during the preceding two weeks, always compared to baseline (S2 Fig). The scale (0–6) for the follow-up form was: 0: Major worsening; 1: Moderate worsening; 2: Slight worsening; 3: No change from baseline; 4: Slight improvement; 5: Moderate improvement; 6: Major improvement. These forms for self-reported symptoms were similar to those used in the previous randomized phase II study [7].

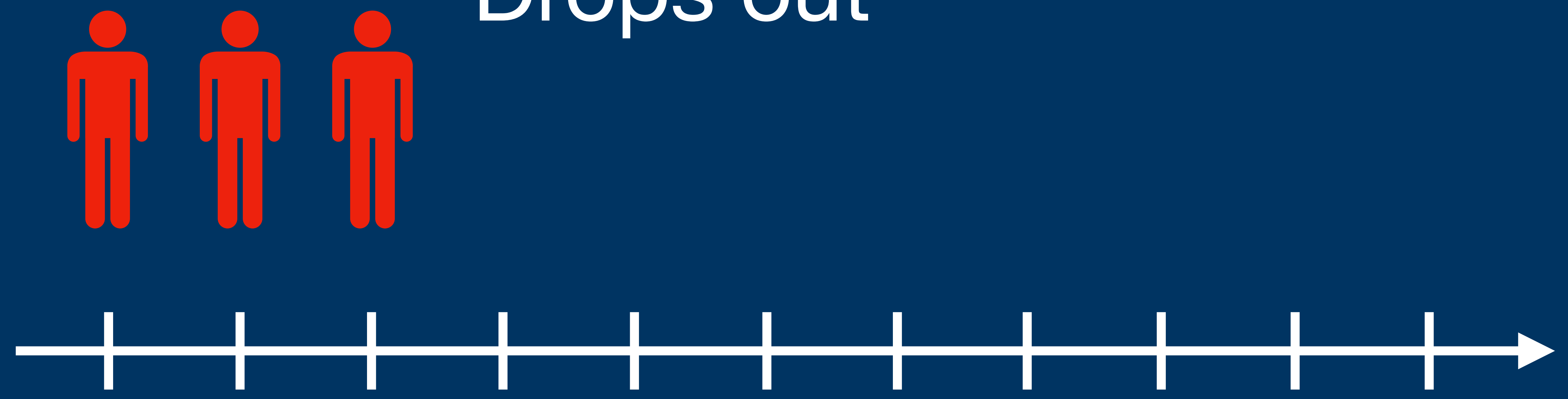
What type of incomplete data?



Begin of study

End of study

What type of incomplete data?

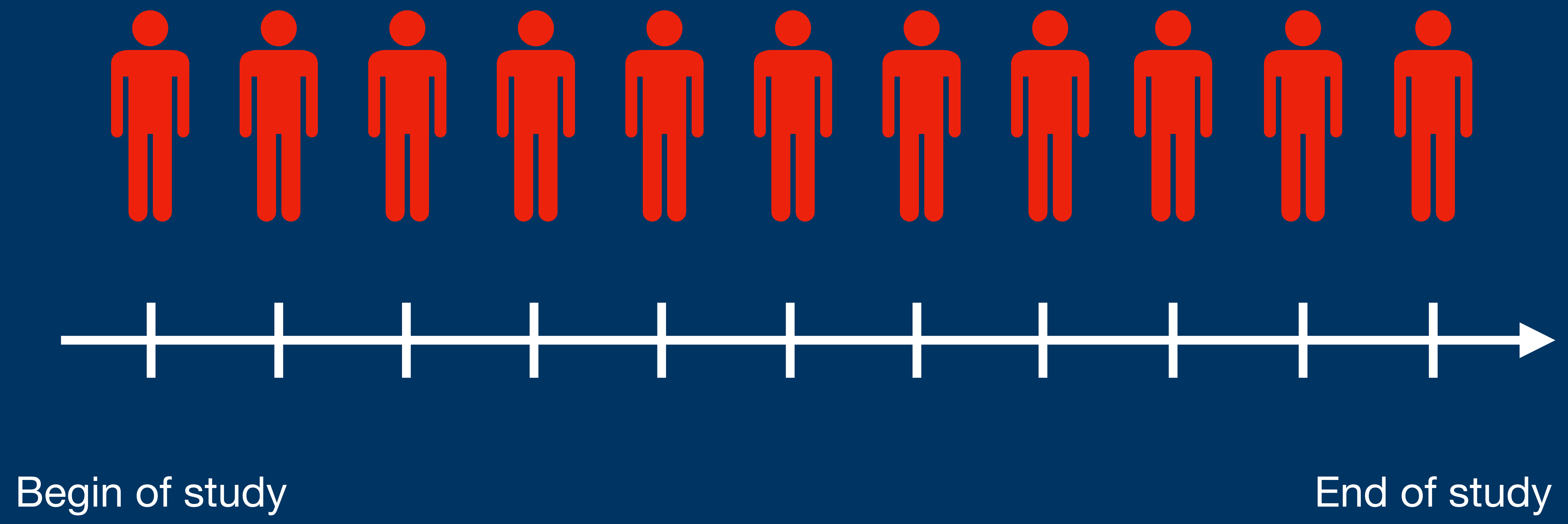


Drops out

Begin of study

End of study

What type of incomplete data?



Exercise 2: rituximab clinical trial data

Identify and quantify censored and truncated data concerning time to treatment response

Should you consider interval censoring in this case?

Basic mathematical formulation of the problem

Right/left censored data

$$\{\mathrm{t}_i, d_i\}, i = 1,\ldots,n$$

$$\mathrm{t}_i = \begin{cases} t_i^*, & \text{if } t_i \text{ is right censored} \\ t_i, & \text{if } t_i \text{ is completely observed} \end{cases}$$

$$\mathrm{t}_i = \begin{cases} t_i^+, & \text{if } t_i \text{ is left censored} \\ t_i, & \text{if } t_i \text{ is completely observed} \end{cases}$$

Interval censored data

$$\{\mathrm{a}_i, b_i, d_i\}, i = 1,\ldots,n$$

$$\mathrm{a}_i = \begin{cases} t_i^*, & \text{if } t_i \text{ is interval censored} \\ t_i, & \text{if } t_i \text{ is completely observed} \end{cases}$$

$$\mathrm{b}_i = \begin{cases} t_i^+, & \text{if } t_i \text{ is interval censored} \\ t_i, & \text{if } t_i \text{ is completely observed} \end{cases}$$

$$d_i = \begin{cases} 0, & \text{if } t_i \text{ is censored} \\ 1, & \text{if } t_i \text{ is completely observed} \end{cases}$$

In practice

Package survival

Survival time

$$\text{time}_i = \begin{cases} t_i^*, & \text{if } t_i \text{ is right or interval censored} \\ t_i, & \text{if } t_i \text{ is completely observed} \\ t_i^+, & \text{if } t_i \text{ is left censored} \end{cases}$$

Event indicator

$$d_i = \begin{cases} 0, & \text{if } t_i \text{ is right censored} \\ 1, & \text{if } t_i \text{ is completely observed} \\ 2, & \text{if } t_i \text{ is left censored} \\ 3, & \text{if } t_i \text{ is interval censored} \end{cases}$$

$$t_i \in \left(t_i^*, t_i^+\right)$$

$$\text{time2}_i = \begin{cases} t_i^+, & \text{if } t_i \text{ is interval censored} \\ 0, & \text{otherwise} \end{cases}$$

Likelihood function of a parametric model under different censoring mechanisms

$$T_i \,|\, \theta \rightsquigarrow F(\theta)$$

Weibull, Gamma, Lognormal, Log-logistic, etc

Right censored data

$$L\left(\theta \,|\, \{t_i, d_i\}\right) \equiv \prod_{i=1}^{n} f_\theta(t_i)^{d_i} S_\theta(t_i)^{1-d_i}$$

Left censored data

$$L\left(\theta \,|\, \{t_i, d_i\}\right) \equiv \prod_{i=1}^{n} f_\theta(t_i)^{d_i} F_\theta(t_i)^{1-d_i}$$

Interval censored data

$$L\left(\theta \,|\, \{a_i, b_i, d_i\}\right) \equiv \prod_{i=1}^{n} f_\theta(a_i)^{d_i} \left(F_\theta(b_i) - F_\theta(a_i)\right)^{1-d_i}$$

Parametric estimation

$$\hat{\theta} = \underset{\theta}{\text{argmax}} L\left(\theta \mid \left\{t_i, d_i\right\}\right)$$

No closed-form expressions

Numerical solutions for the maximum likelihood equations

Exercise 3: rituximab clinical trial data

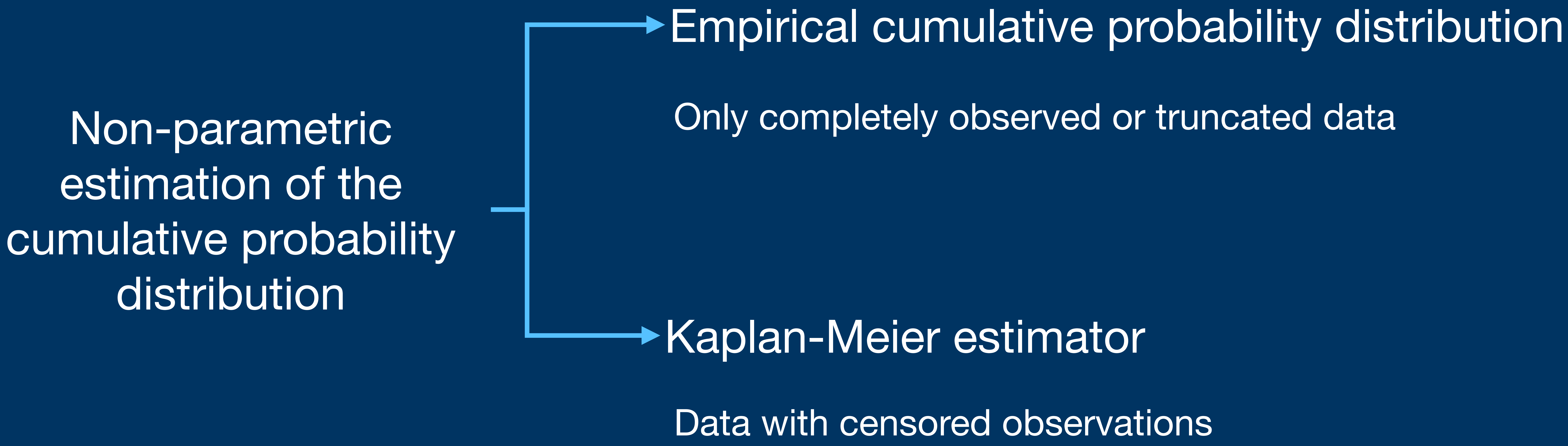
Estimate lognormal, weibull and log-logistic model to data on time to treatment response using the “survreg” function of package survival.

What is the best model for the data?

Checking the adequacy of a parametric model

Can we use the Kolmogorov-Smirnov test directly to data?

Kaplan-Meier estimator for the survival function



Non-parametric estimation of the cumulative probability distribution

Empirical cumulative probability distribution

Only completely observed or truncated data

Kaplan-Meier estimator

Data with censored observations

Kaplan-Meier estimator for the survival function

$$\hat{S}(t) = \prod_{i:t_{(}i)\leq t} \left(1 - \frac{d_i}{n_i}\right) \qquad t \in (0, t_{\max})$$

$d_i =$ number of individuals in which the event was observed at $t_{(i)}$

$n_i =$ number of individuals without the event of interest at $t_{(i-1)}$

$\left\{ t_{(i)}, i = 1, \ldots, r \right\} =$ unique times when the event of interest was observed

Kaplan-Meier estimator for the survival function

$$\hat{S}\left(t_{(1)}\right) = 1 - \frac{d_1}{n_1}$$

$n_1 =$ number of individuals without the event of interest at time $0 = n$

$$\hat{S}\left(t_{(i)}\right) = \hat{S}\left(t_{(i-1)}\right)\left(1 - \frac{d_i}{n_i}\right)$$

Exercise 3: rituximab clinical trial data

Estimate the survival curve of time to treatment response using the Kaplan-Meier estimator

Compare the Kaplan-Meier estimated survival curve to the survival curve predicted by the best parametric model from Exercise 2.