A martingale framework for survival tests

Bigstatistics Seminar 20 May 2026

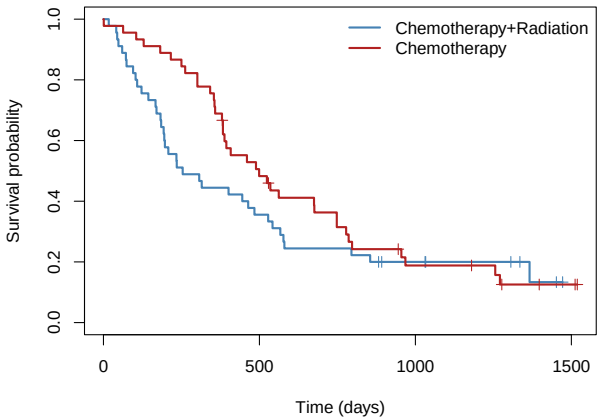
Tijn Jacobs (VU Amsterdam)

joint work with Dante de Roos (CWI)

CWI
Centrum Wiskunde & Informatica

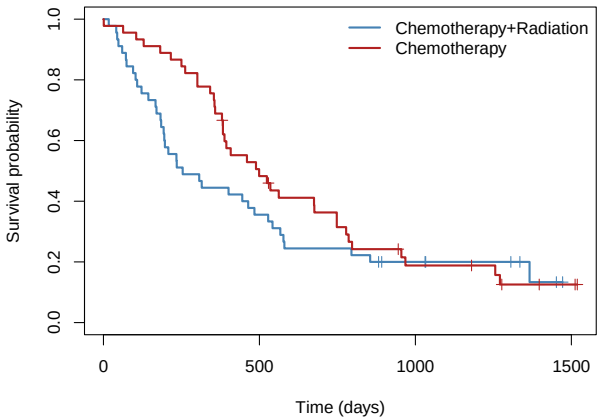
VU VRIJE UNIVERSITEIT AMSTERDAM

Is there a difference?



Is the survival in these two groups different?

Is there a difference?



Is the survival in these two groups different?

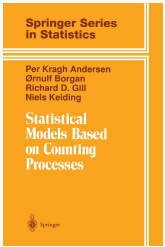
$p = 0.25$

What we promise

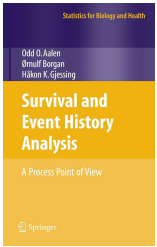
- Finite-sample validity.
  Type-I error controlled exactly at every sample size — no large-sample approximation.
- Continuous monitoring.
  Inspect the evidence as the data accumulate; stop whenever it suffices.
- Meta-analysis by multiplication.
  Evidence from independent studies combines into a single valid test.
- Asymptotic power one.
  Power tends to 1 under mild conditions, with no proportional-hazards assumption.

Two theoretical frameworks

Counting-process survival analysis

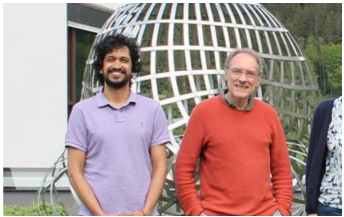


Andersen, Borgan, Gill, Keiding (1993)



Aalen, Borgan, Gjessing (2008)

Test martingales & e-processes



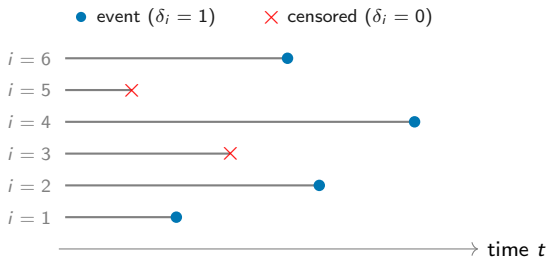
Ville (1939)

Shafer & Vovk (2019)

Ramdas, Grünwald, Vovk, Shafer (2023)

The counting-process language

Survival data



For $i = 1, \ldots, n$ we observe a pair (t_i, δ_i):

- t_i — the observed time.
- $\delta_i = 1$ if the event happened, 0 if censored.

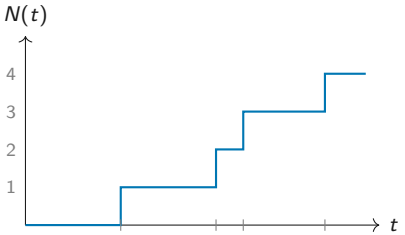
We summarise the data by two processes that change in time:

- $N(t)$: events observed by time t.
- $Y(t)$: subjects still at risk at time t.

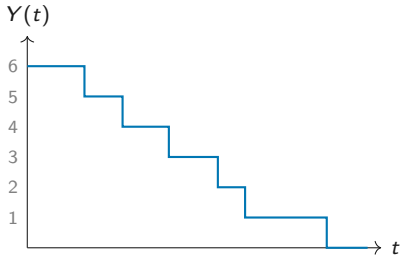
Counting process and at-risk process

$$N(t) = \sum_{i=1}^{n} \mathbf{1}\{t_i \leq t,\ \delta_i = 1\}$$

$$Y(t) = \sum_{i=1}^{n} \mathbf{1}\{t_i \geq t\}$$



Jumps up by 1 at each observed event.



Drops by 1 at each event *or* censoring.

Hazard and cumulative hazard

The hazard rate:

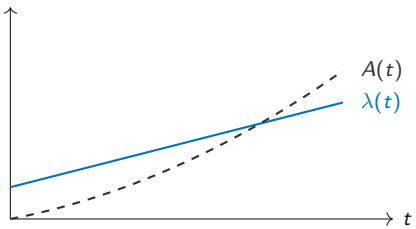
$$\lambda(t)\,dt \;\approx\; \mathbb{P}\big(T \in [t,\, t+dt] \,\big|\, T \geq t\big)\,.$$

The local rate of failure among survivors.

The cumulative hazard:

$$A(t) = \int_0^t \lambda(s)\,ds.$$

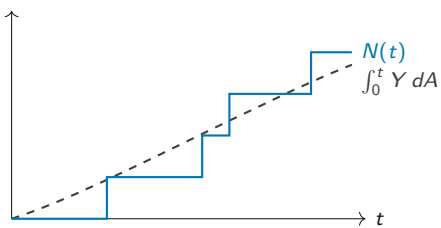
Accumulates risk over time.



A is the area under λ up to time t.

The fundamental decomposition

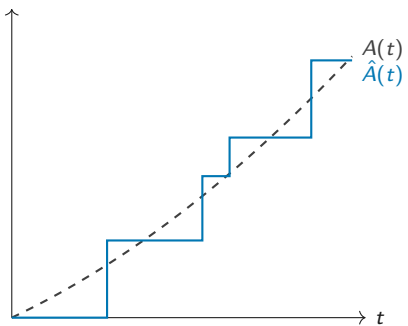
$$N(t) = \underbrace{\int_0^t Y(s)\, dA(s)}_{\text{expected events (predictable)}} + \underbrace{M(t)}_{\text{surprise}}$$



- $Y(s)\, dA(s)$: expected events in $[s, s + ds]$ given the past.
- $M(t) = N(t) - \int_0^t Y\, dA$ is the gap.
- M is a mean-zero martingale.
- Observed minus expected, accumulated.

Nelson–Aalen: estimating A from the data

$$\hat{A}(t) \;=\; \int_0^t \frac{dN(s)}{Y(s)} \;=\; \sum_{t_j \le t,\ \delta_j = 1} \frac{1}{Y(t_j)}.$$



- Step function, jumps only at observed events.
- $\hat{A} - A$ is itself a martingale.
- Consistent estimator of A under independent censoring.

Constructing the test

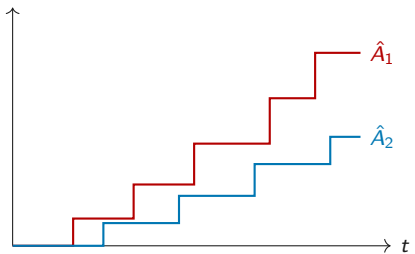
Two samples

Two independent counting processes:

- $N_1(t)$, $Y_1(t)$ with hazard λ_1.
- $N_2(t)$, $Y_2(t)$ with hazard λ_2.

We test:

$$H_0 : A_1(t) = A_2(t) \quad \forall\, t \geq 0,$$
$$H_1 : A_1(t) \neq A_2(t) \ \text{ for some } t.$$



Nelson–Aalen curves diverging under H_1.

The Nelson–Aalen difference

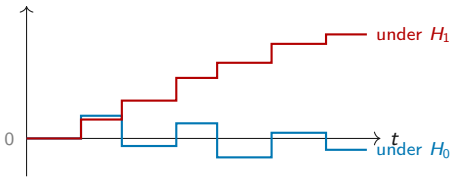
Consider the running difference:

$$X(t) \;=\; \hat{A}_1(t) - \hat{A}_2(t).$$

Under H_0, with $A_1 = A_2$:

$$X(t) \;=\; \underbrace{\left(\hat{A}_1 - A_1\right)}_{\text{martingale}} \;-\; \underbrace{\left(\hat{A}_2 - A_2\right)}_{\text{martingale}}.$$

X is a mean-zero martingale.



Sample paths of $X(t)$.

The Nelson–Aalen difference

Consider the running difference:

$$X(t) \;=\; \hat{A}_1(t) - \hat{A}_2(t).$$

Under H_0, with $A_1 = A_2$:

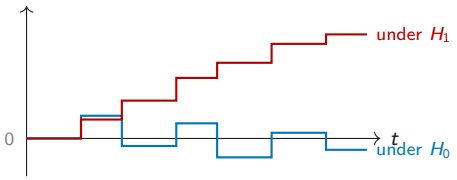
$$X(t) \;=\; \underbrace{\left(\hat{A}_1 - A_1\right)}_{\text{martingale}} \;-\; \underbrace{\left(\hat{A}_2 - A_2\right)}_{\text{martingale}}.$$

X is a mean-zero martingale.

We bet against it with a predictable weight W:

$$Z(t) \;=\; \int_0^t W(s)\, dX(s).$$

Z is still a mean-zero martingale.



Sample paths of $X(t)$.

From martingale to running evidence

Wrap Z multiplicatively — the Doléans–Dade exponential:

$$\mathcal{E}(Z)_t \;=\; \prod_{s \le t} \bigl(1 + W(s)\,\Delta X(s)\bigr).$$

Only event times contribute: each event multiplies in one factor.

Pick W so that every factor stays nonnegative:

$$-Y_1(s) \;\le\; W(s) \;\le\; Y_2(s).$$

Then under H_0 the process $\mathcal{E}(Z)$ is:

- $\mathcal{E}(Z)_0 = 1$;
- nonnegative for all t;
- a supermartingale with: $\mathbb{E}\bigl[\mathcal{E}(Z)_t\bigr] \le 1$.

Read it as a running evidence score: starts at one, grows under H_1, fair under H_0.

Choosing the weight W

W must be:

- predictable — $W(s)$ uses only data strictly before s;
- admissible — $-Y_1(s) \le W(s) \le Y_2(s)$.

Beyond that, we design W to target a specific alternative.

Some familiar choices:

- *Logrank*: $W(s) = \dfrac{Y_1(s)\, Y_2(s)}{Y_1(s) + Y_2(s)}$.
- *Anytime-valid logrank* of Ter Schure et al. (2024) is the special case

$$W(s) = \frac{(\theta - 1)\, Y_1(s)\, Y_2(s)}{\theta\, Y_1(s) + Y_2(s)},$$

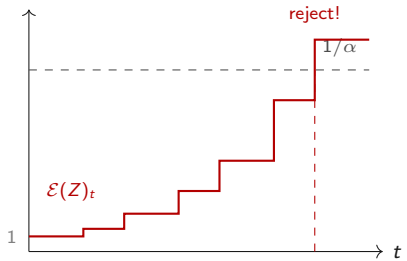
  for a chosen hazard ratio θ.
- *Adaptive*: $W(s) = Q_\gamma(s) \cdot \dfrac{Y_1(s)\, Y_2(s)}{Y_1(s) + Y_2(s)}$ — coming up.

Validity and composability

Theorem
Fix $\alpha \in (0,1)$ and reject H_0 the first time $\mathcal{E}(Z)_t \geq 1/\alpha$. This defines a level-$\alpha$ test:

$$\mathbb{P}_{H_0}\left(\sup_{t \geq 0} \mathcal{E}(Z)_t \;\geq\; 1/\alpha\right) \;\leq\; \alpha.$$

The bound holds at every sample size, and at any data-dependent stopping time.

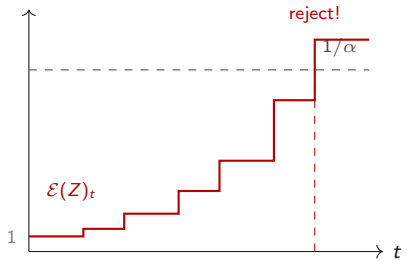


Validity and composability

Theorem
Fix $\alpha \in (0,1)$ and reject H_0 the first time $\mathcal{E}(Z)_t \geq 1/\alpha$. This defines a level-$\alpha$ test:

$$\mathbb{P}_{H_0}\left(\sup_{t \geq 0} \mathcal{E}(Z)_t \;\geq\; 1/\alpha\right) \;\leq\; \alpha.$$

The bound holds at every sample size, and at any data-dependent stopping time.



Composability. For independent studies, the terminal values $\mathcal{E}_1, \mathcal{E}_2, \ldots$ of separate e-processes multiply into a valid combined e-value:

$$\mathbb{P}_{H_0}(\mathcal{E}_1 \cdot \mathcal{E}_2 \cdots \;\geq\; 1/\alpha) \leq \alpha.$$

Other results

Consistency*

Under mild conditions on W:

$$\mathbb{P}_{H_1}\left(\sup_{t\geq 0} \mathcal{E}(Z)_t \;\geq\; 1/\alpha\right) \;\longrightarrow\; 1.$$

Other results

Consistency*

Under mild conditions on W:

$$\mathbb{P}_{H_1}\left(\sup_{t\geq 0} \mathcal{E}(Z)_t \;\geq\; 1/\alpha\right) \;\longrightarrow\; 1.$$

Competing risks

Cause-specific counting processes $N_{j,k}$, $k = 1, \ldots, K$. Global null:

$$H_0 :\; \lambda_{1,k}(t) = \lambda_{2,k}(t) \quad \forall\, t,\; k.$$

Product e-process over causes:

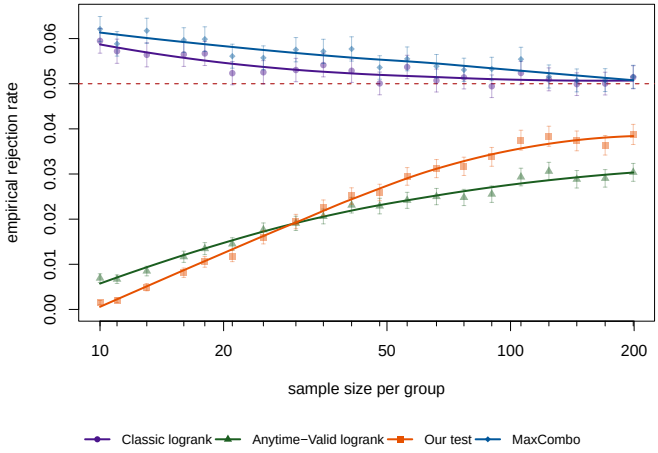
$$\mathcal{E}_t^{\mathrm{cr}} \;=\; \prod_{k=1}^{K} \mathcal{E}(Z_k)_t.$$

Admissibility per cause:

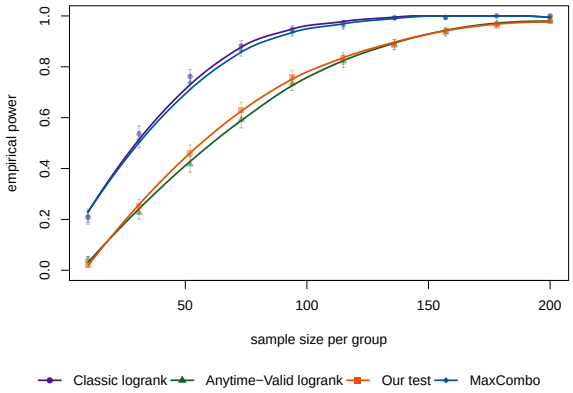
$$-Y_1(s) \;\leq\; W_k(s) \;\leq\; Y_2(s) \quad \forall\, s, \forall\, k.$$

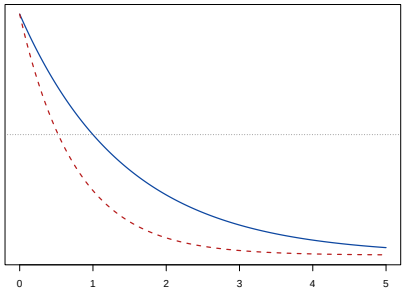
Empirical evaluation

Simulation under H_0



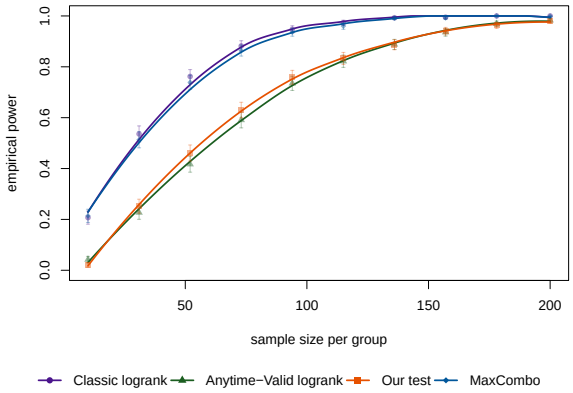
Simulation under H_1: proportional hazards





DGP: exponential survival with HR $= 1.9$.

Simulation under H_1: proportional hazards

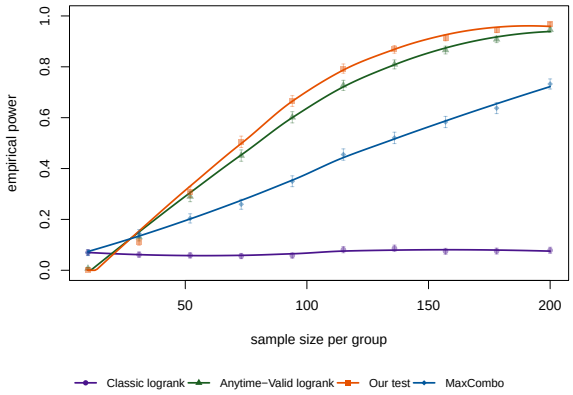


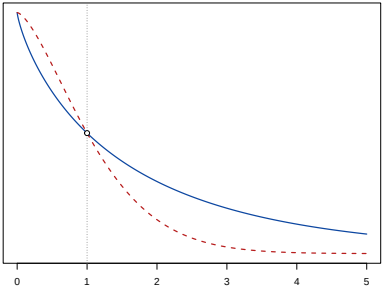
Question

How many extra observations are needed on average to match the logrank's power?

Test	Extra n	SD
Anytime-Valid logrank	37.8	5.4
Our test	33.7	5.2

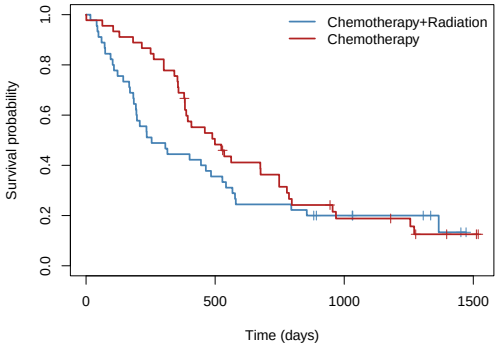
Simulation under H_1: crossing hazards





DGP: both groups have median $= 1$; curves cross at the median.

Back to the GTSG example

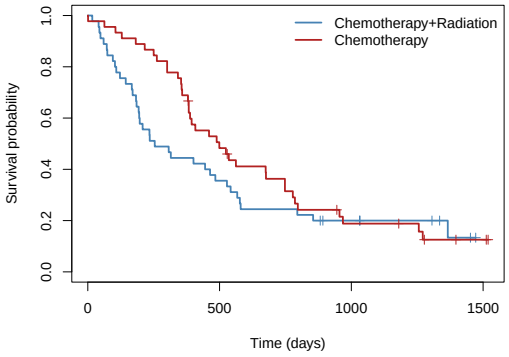


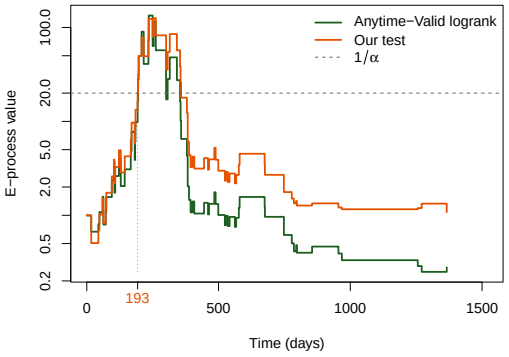
Test	p	$\sup_t \mathcal{E}_t$	$\mathcal{E}_\infty$
Classic logrank	0.251	—	—
MaxCombo	0.061	—	—
Anytime-Valid logrank	—	134.05	0.28
Our test	—	126.14	1.08

Reject at $\alpha = 0.05$ iff $p < 0.05$ or $\sup_t \mathcal{E}_t \geq 20$.

$\mathcal{E}_\infty$ is the e-process at the final event time — multiplies across studies.

Back to the GTSG example





Thank you — questions & input?

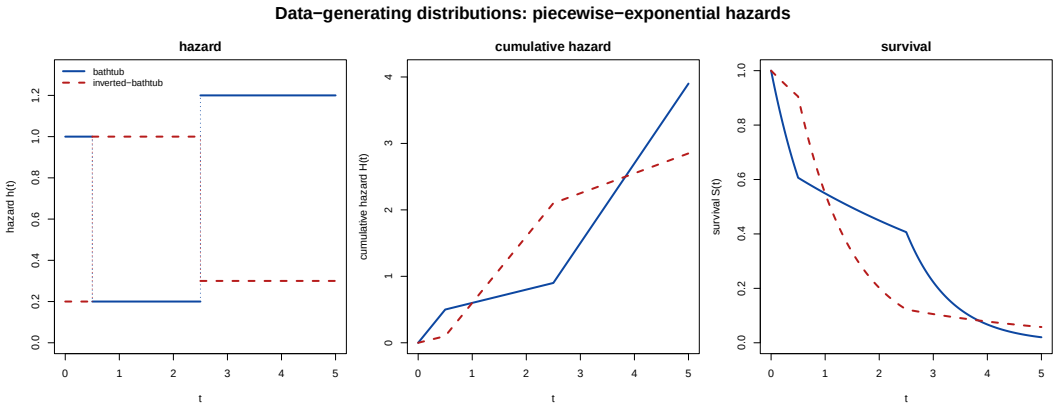
Thank you!

Questions, comments, input most welcome.

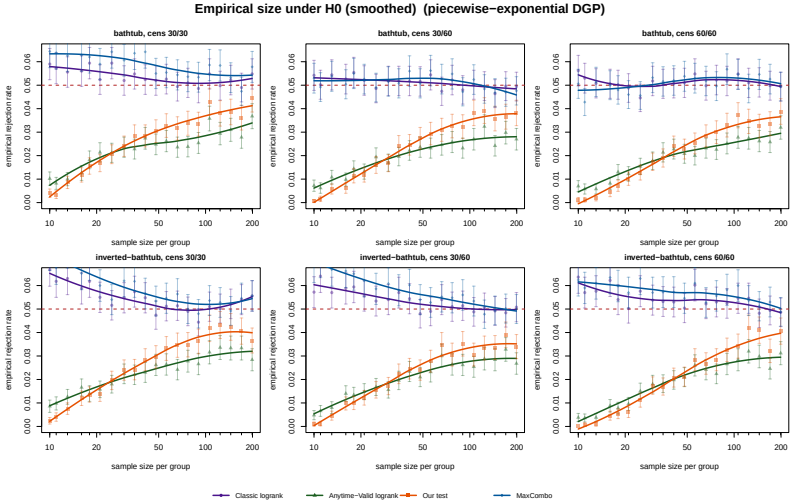
Funded by
the European Union

This project has received funding from the European Research Council (ERC) under the European Union's Horizon Europe program under Grant agreements No. 101074082 and No. 101142168. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Data-generating processes under the null



Type-I error: six scenarios



Non-proportional-hazards methods: a landscape

Hypothesis-testing approaches ($n = 98$)	
Log-rank tests	63 (64.3%)
Kaplan–Meier-based tests	26 (26.5%)
Combination tests	20 (20.4%)
Other tests	12 (12.2%)

Bardo, Huber, Benda, Brugger, Fellinger, Galaune, Heinz, Heinzl, Hooker, Klinglmüller, König, Mathes, Mittlböck, Posch, Ristl, Friede (2024). *Methods for Non-Proportional Hazards in Clinical Trials: A Systematic Review.* Statistical Methods in Medical Research, 33(6), 1069–1092. doi:10.1177/09622802241242325