Evaluating the impact of delayed effects in confirmatory trials

José L. Jiménez

Biostatistical Sciences and Pharmacometrics
Novartis Pharma A.G.
jose_luis.jimenez@novartis.com

PSI One-Day Meeting: Non-proportional hazards and applications in immuno-oncology
April 29th, 2021

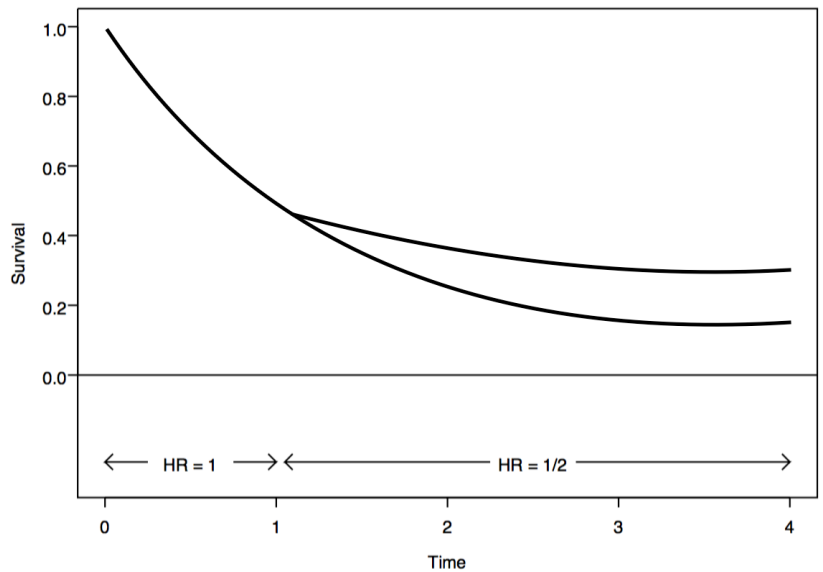
Acknowledgments

- Byron Jones – Novartis (co-author)
- Viktoriya Stalbovskaya – Merus (formerly at Novartis and co-author).
- Ekkehard Glimm – Novartis.
- Thomas Jaki – Lancaster University.
- Franz König – Medical University of Vienna.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 633567

Introduction (I)

- Immuno-oncology (IO) is a rapidly evolving area in the development of anti-cancer drugs.
- The effect of an IO agent is not typically directed to the tumor itself; it instead boosts the patient's immune system, and this effect may not be observed immediately.
- This may translate to inferior or equal overall survival (OS) compared to control treatment in the first months of therapy, and superior OS thereafter leading to non-proportionality of hazards.



Introduction (II)

What to expect from this presentation?

We study the behavior of the weighted log-rank test as an alternative to the log-rank test with and without delayed effects answering questions such us:

- How much does the power drop with delayed effects?
- How much power can we gain using weighted log-rank vs. long-rank in a study with delayed effects?
- What are the optimal ρ and γ (the weighted log-rank parameters) values?
- Can we guarantee a certain power level?
- What if we think know the delay, but it turns out to be wrong?

Introduction (III)

Simulated example setting

- Group sequential design with 1 interim analysis for efficacy at 75% of the information fraction (no interim analysis for futility)
- $\alpha = 0.025$ (1-sided test) and $1 - \beta = 0.9$
- O'Brien - Fleming alpha spending function
- Control group's median survival: 6 months
- Experimental groups's median survival: 9 months (group with IO agent)
- Study duration: 25 months
- Recruitment duration: 17.5 month
- Randomization ratio: 1:1

Log-rank vs. weighted log-rank

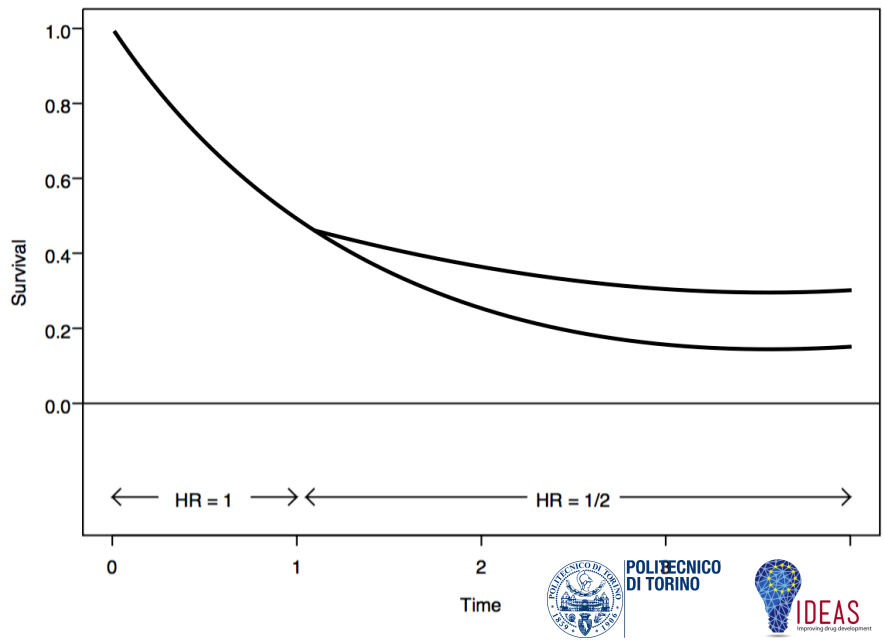
- Weighted log-rank statistic:

$$T_w = \frac{\left[\sum_{t=1}^{T} w_t(O_{1t} - E_{1t})\right]^2}{\sum_{t=1}^{T} w_t^2 V_t},$$

where $w_t = \hat{S}(t)^\rho(1-\hat{S}(t))^\gamma$ and $\hat{S}$ is the estimated pooled survival function.

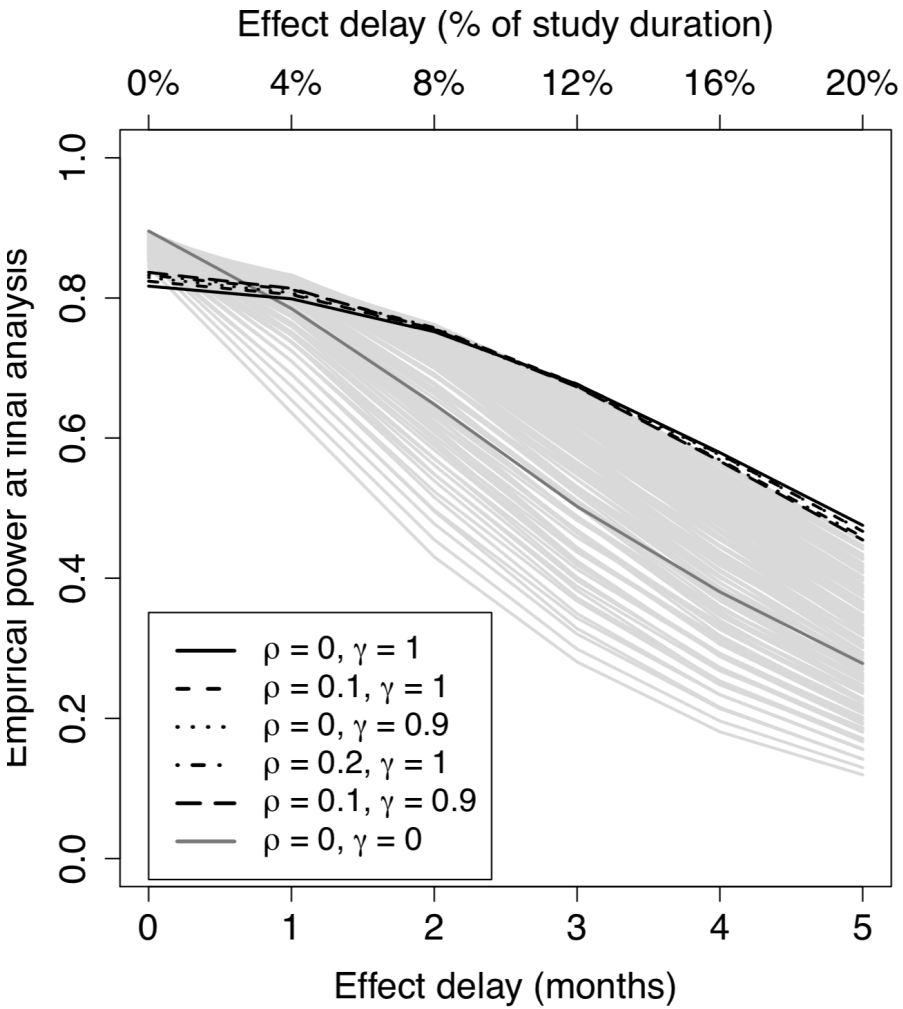
- Potential power "gain" with respect to log-rank if we use $(\rho = 0, \gamma = 1)$.

- Important parameter value combinations:
  - $(\rho = 0, \gamma = 0)$ = equal weights (log-rank)
  - $(\rho = 1, \gamma = 0)$ = weight early differences
  - $(\rho = 0, \gamma = 1)$ = weight late differences



Log-rank vs. weighted log-rank (III)

- We consider a study where the sample size has been calculated assuming proportional hazards, and calculate the empirical power for each combination of ρ and γ for different delay times.



- For $(\rho = 0, \gamma = 1)$, the study is **not** sufficiently powered.
- We cannot guarantee (at least) 80% of power for medium - large delays.

POLITECNICO DI TORINO

IDEAS

Sample size calculation methods

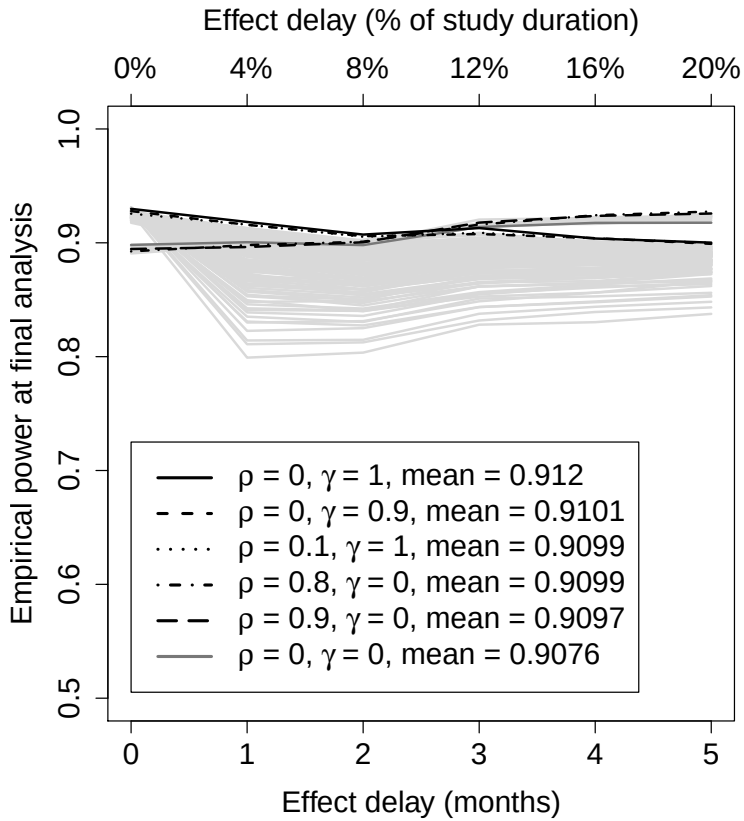
Most common approach:

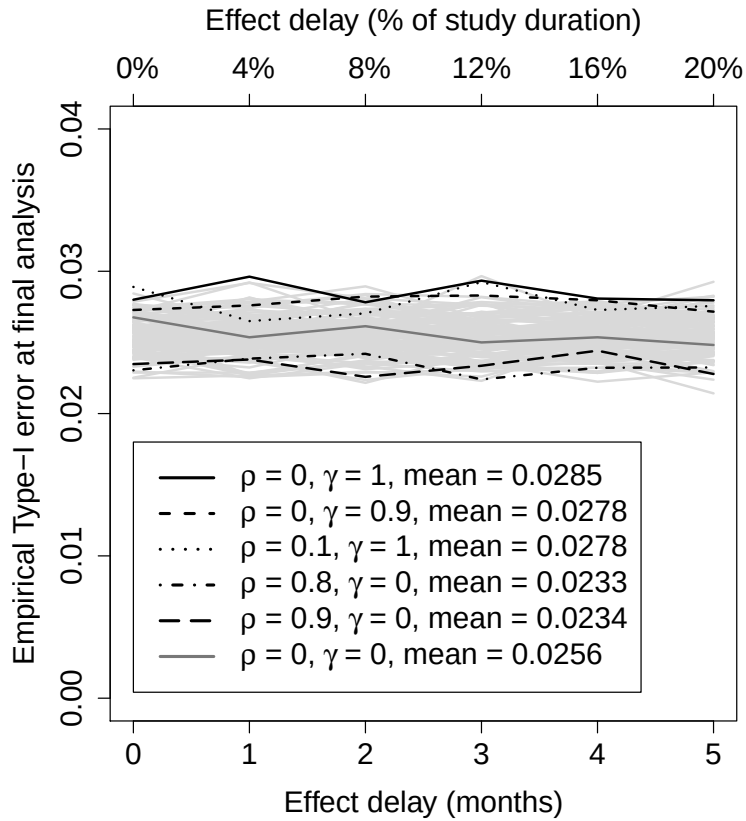
- Schoenfeld, D. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika*, 68(1), 316-319.

Alternative method that incorporates delayed effects and the Fleming and Harrington class of weights in the formulation:

- Lakatos, E. (1988). Sample sizes based on the log-rank statistic in complex clinical trials. *Biometrics*, 229-241.
- Hasegawa, T. (2014). Sample size determination for the weighted log-rank test with the Fleming–Harrington class of weights in cancer vaccine studies. *Pharmaceutical statistics*, 13(2), 128-135.

Sample size calculation following Hasegawa (2014)





	Delay (months)	0	1	2	3	4	5
$(\rho = 0, \gamma = 0)$	# of events	258	359	492	686	986	1436
	# of patients	330	456	621	860	1228	1777
$(\rho = 0, \gamma = 1)$	# of events	369	376	406	468	578	741
	# of patients	472	478	512	587	719	917

Controlling type-I error - The combination test statistic

- To protect type-I error rate inflation we use the "weighted inverse normal" combination test statistic:

$$z^* = \sqrt{\frac{n_1}{n_2}} z_1 + \sqrt{\frac{n_2 - n_1}{n_2}} z_2, \tag{1}$$

where

  - z_1 is the weighted log-rank statistic at the interim analysis
  - z_2 is the weighted log-rank statistic at the final analysis using the data from second stage alone.
  - $z_1 \sim N(0,1)$, $z_2 \sim N(0,1)$ and $z^* \sim N(0,1)$

- With the combination test approach we guarantee type-I error rate control.

So far...

- We have seen its impact in a group sequential setting.
- We gain some power using weighted-log rank, but this may not be enough.
- Following Hasegawa (2014), we can guarantee a certain power level as long as we know the delay.
- We can guarantee type-I error control.

So far...

- We have seen its impact in a group sequential setting.
- We gain some power using weighted-log rank, but this may not be enough.
- Following Hasegawa (2014), we can guarantee a certain power level as long as we know the delay.
- We can guarantee type-I error control.

What if we underestimated (or simply don't know) the delay?

So far...

- We have seen its impact in a group sequential setting.
- We gain some power using weighted-log rank, but this may not be enough.
- Following Hasegawa (2014), we can guarantee a certain power level as long as we know the delay.
- We can guarantee type-I error control.

What if we underestimated (or simply don't know) the delay?

Sample size reassessment

Mehta and Pocock's "promising zone" approach (I)

In case ϵ is out of the range we compute the conditional power:

$$\mathrm{CP}_{\hat{\delta}_1}(z_1, n_2) = 1 - \Phi\left(\frac{z_\alpha\sqrt{n_2} - z_1\sqrt{n_1}}{\sqrt{n_2 - n_1}} - \frac{z_1\sqrt{n_2 - n_1}}{\sqrt{n_1}}\right). \qquad (2)$$

We divide the conditional power results in 3 zones:

- Favorable: If $\mathrm{CP}_{\hat{\delta}_1}(z_1, n_2) \geq 0.8$.
- Promising: If $0.8 > \mathrm{CP}_{\hat{\delta}_1}(z_1, n_2) \geq \mathrm{CP}_{\mathrm{min}}$.
  - Sample size re-estimation.
- Unfavorable: If $\mathrm{CP}_{\hat{\delta}_1}(z_1, n_2) < \mathrm{CP}_{\mathrm{min}}$.

Type-I error is protected as long as $\mathrm{CP}_{\mathrm{min}} > 0.5$ (see Chen et al., (2004)).

Mehta and Pocock's "promising zone" approach (II)

Some issues with this methodology:

- The results of Chen et al., (2004) don't allow to increase the sample size in situations when the greatest benefits might accrue.
- Jennison and Turnbull (2015) showed that it is possible to obtain an optimal sample size reassessment rule that yields a lower expected sample size for the same power curve.

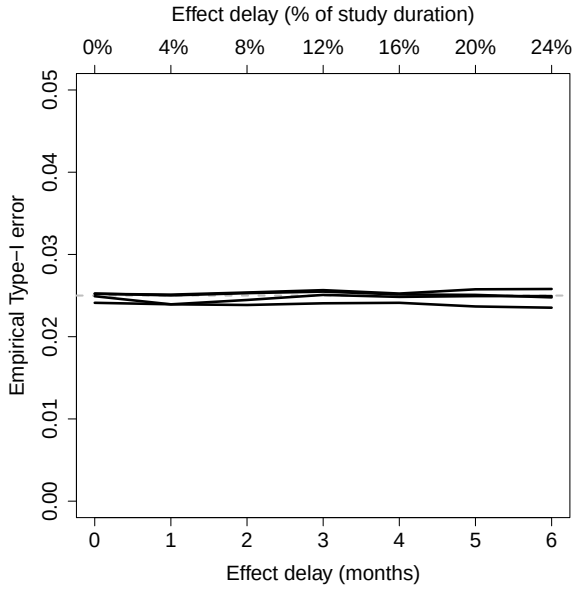
Jennison and Turnbull's "start small then ask for more" approach (I)

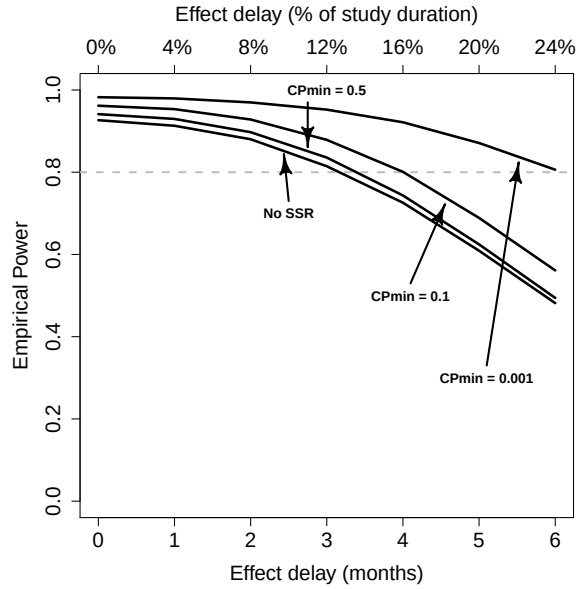
To obtain the optimal number of events at the final analysis (n_2^*), we need to maximize

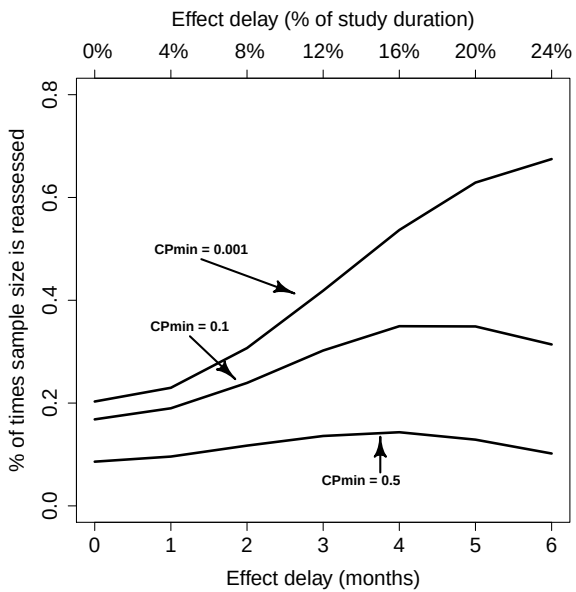
$$\mathrm{CP}_{\hat{\delta}_1}(z_1, n_2^*) - \eta(n_2^* - n_2), \tag{3}$$

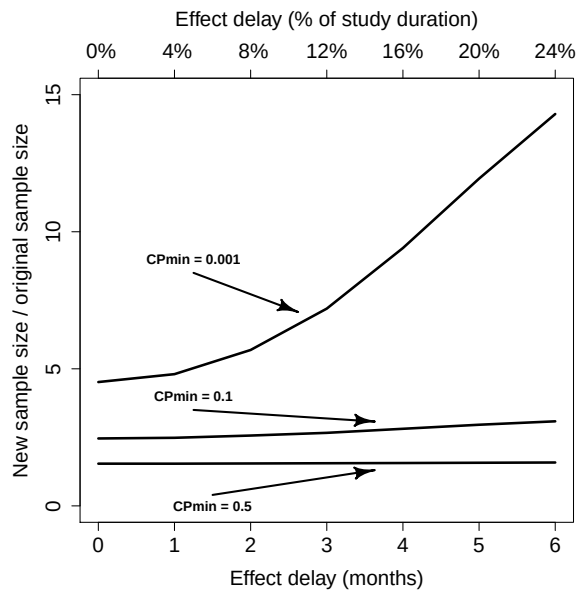
where η can be considered as "a tuning parameter that controls the degree to which the sample size may be increased when interim data are promising but not overwhelming".

Simulation results (I)





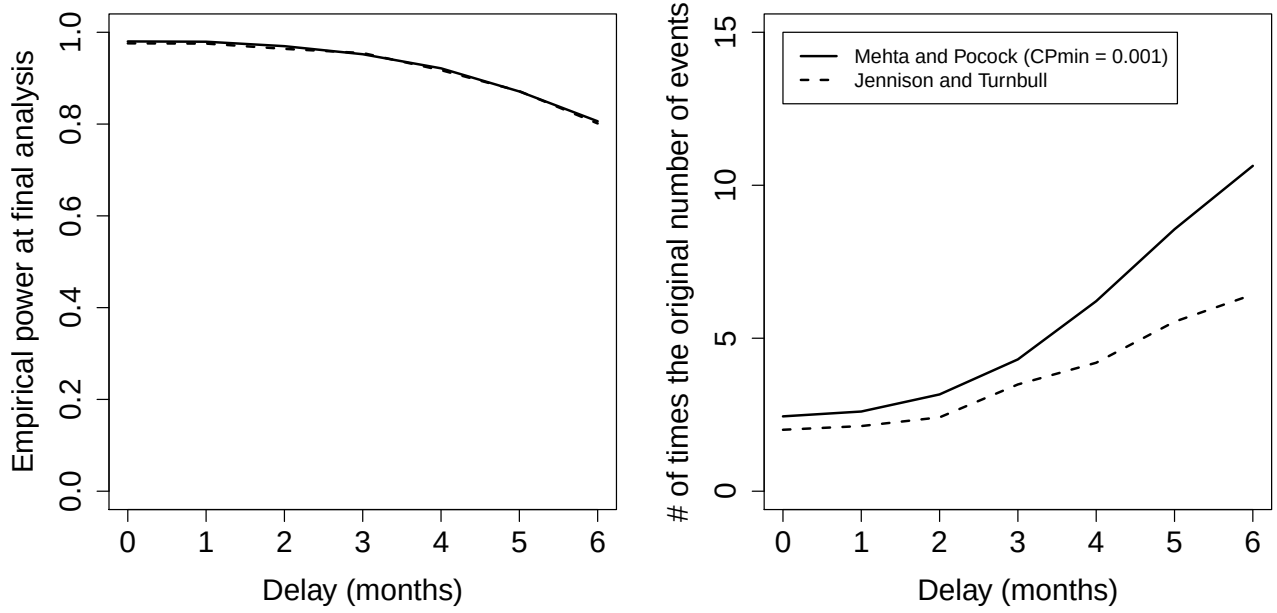




POLITECNICO DI TORINO

IDEAS

Simulation results (II)



Jennison and Turnbull's approach requires less patients for the same power curve

Conclusions

- The weighted log-rank with parameter combination $(\rho = 0, \gamma = 1)$ outperforms the log-rank test.
- Low values of ρ and high values of γ are hence appropriate.
- However, the difference may not enough when dealing with larger delayed effects.
- We can guarantee a certain power if we follow the methods proposed by Lakatos (1988) and Hasegawa (2014).
  - Need for type-I error control.
- In case the delay is underestimated (or unknown) at the design stage, through sample size reassessment we can achieve enough power at the end of the trial.
  - We need to avoid any back-calculation of the conditional power.

Where to find this work?

MAIN PAPER

Properties of the weighted log-rank test in the design of confirmatory studies with delayed effects

José L. Jiménez ✉, Viktoriya Stalbovskaya, Byron Jones

First published: 27 December 2018 | https://doi.org/10.1002/pst.1923

POLITECNICO DI TORINO

IDEAS

Thank you for your attention!

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 633567

POLITECNICO DI TORINO

IDEAS