Efficiency of nonparametric superiority tests based on restricted mean survival time versus the log-rank test under proportional hazards

Dominic Magirr

PSI Conference, Wembley, June 2025

NOVARTIS

Acknowledgements

Craig Wang

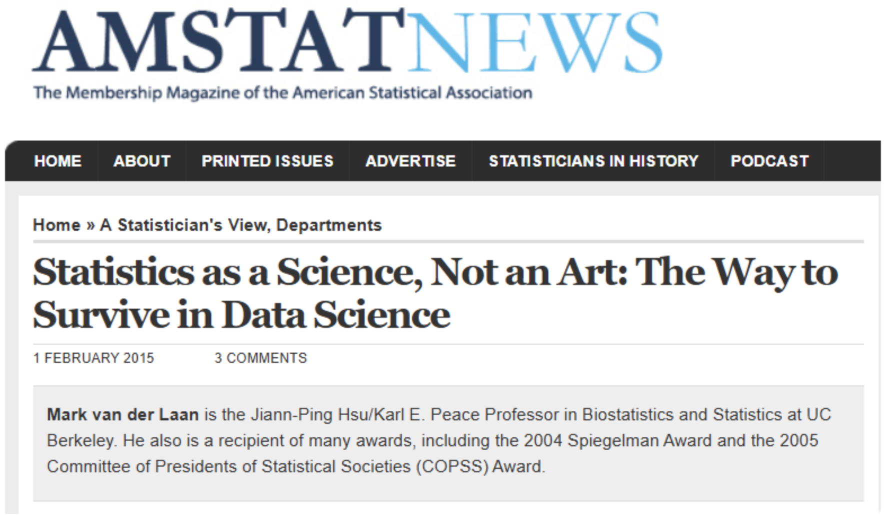
Alexander Przybylski

Xinlei Deng

Mark Baillie

Tim Morris

The Science of Statistics



Home » A Statistician's View, Departments

Statistics as a Science, Not an Art: The Way to Survive in Data Science

1 FEBRUARY 2015 3 COMMENTS

Mark van der Laan is the Jiann-Ping Hsu/Karl E. Peace Professor in Biostatistics and Statistics at UC Berkeley. He also is a recipient of many awards, including the 2004 Spiegelman Award and the 2005 Committee of Presidents of Statistical Societies (COPSS) Award.

https://magazine.amstat.org/blog/2015/02/01/statscience_feb2015/

The foundation of statistics laid down by its founders [...] could not have been to arbitrarily select a "convenient" statistical model. However, that is precisely what most statisticians blithely do, proudly referring to the quote, "All models are wrong, but some are useful."

[...]

one typically asks a few questions about the data such as: Is the outcome a survival time? Is it case-control data? And then one quickly moves on to returning output from a Cox-Ph model or a logistic regression model with some "reasonable" set of covariates

[...]

Is this mess we have created really necessary? No! As a start, we need to take the field of statistics (i.e., the science of learning from data) seriously. It is complete nonsense to state that all models are wrong, so let's stop using that quote. For example, a statistical model that makes no assumptions is always true.

Model-trusting

$$\text{logit}\, P(Y = 1 \mid A, X) = \alpha_0 + \alpha_1 A + \alpha_2 X$$

- **Direct** estimation via MLE / posterior probability
- If model is incorrect, it's unclear what α_1 means
- Compatible with Bayesian, likelihood, and frequentist (conditional and unconditional) inference
- Typically used for conditional estimands

Model-robust / assumption-lean

$$\overline{Y}_1 - \overline{Y}_0 + \frac{1}{n}\sum_{i=1}^{n}\left(A_i - \frac{1}{2}\right)h(X_i)$$

- Combines an unadjusted estimator with an "estimator of zero"
- Clever choice of h(x) to increase efficiency
- An (unconditional) frequentist approach
- Typically used for unconditional estimands

Buja et al. (2019); Vansteelandt (2021)

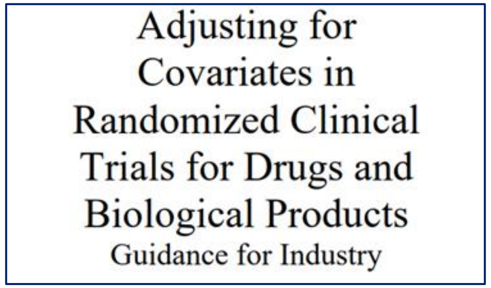
RCTs with time-to-event endpoints

Status quo

(Stratified) Cox proportional hazards models to estimate conditional hazard ratios

→

FDA Position



Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products

Guidance for Industry

→

“Model-free estimand” and “assumption-lean analysis”

Compare unconditional probability of survival (or restricted mean survival time) on the two treatment arms.

Double-robust covariate-adjusted estimators: AIPCW, TMLE

What’s in the guidance?

Should we go further?

Status quo

For all strata $k = 1, \ldots, K$:

$$S_{E,k}(t) = \{S_{C,k}(t)\}^{HR_c}$$

$$\widehat{HR_C}$$

FDA guidance

$$S_E(t) = \{S_C(t)\}^{HR_M}$$

$$\widehat{HR_M}$$

$$\widehat{HR_M} + \frac{1}{n}\sum_{i=1}^{n}\left(A_i - \frac{1}{2}\right)h(X_i)$$

Model robust

$$(T_i(1), T_i(0), X_i, A_i) \overset{\text{i.i.d.}}{\sim} f$$

$$\Delta_{RMST(\tau)} = E\{\min(T(1), \tau)\} - E\{\min(T(0), \tau)\}$$

$$\widehat{\Delta}_{RMST(\tau)}$$

$$\widehat{\Delta}_{RMST(\tau)} + \frac{1}{n}\sum_{i=1}^{n}\left(A_i - \frac{1}{2}\right)h(X_i)$$

Status quo

For all strata $k = 1, \ldots, K$:

$$S_{E,k}(t) = \{S_{C,k}(t)\}^{HR_c}$$

$$\widehat{HR_C}$$

Basel Biometric Society Presentation 25th March https://baselbiometrics.github.io/home/docs/talks/20250325/3_Magirr.pdf

FDA guidance

$$S_E(t) = \{S_C(t)\}^{HR_M}$$

$$\widehat{HR_M}$$

$$\widehat{HR_M} + \frac{1}{n}\sum_{i=1}^{n}\left(A_i - \frac{1}{2}\right)h(X_i)$$

Model robust

$$(T_i(1), T_i(0), X_i, A_i) \overset{\text{i.i.d.}}{\sim} f$$

$$\Delta_{RMST(\tau)} = E\{\min(T(1),\tau)\} - E\{\min(T(0),\tau)\}$$

$$\widehat{\Delta}_{RMST(\tau)}$$

$$\widehat{\Delta}_{RMST(\tau)} + \frac{1}{n}\sum_{i=1}^{n}\left(A_i - \frac{1}{2}\right)h(X_i)$$

NOVARTIS

Status quo

For all strata $k = 1, \dots, K$:

$$S_{E,k}(t) = \{S_{C,k}(t)\}^{HR_c}$$

$$\widehat{HR_C}$$

FDA guidance

$$S_E(t) = \{S_C(t)\}^{HR_M}$$

$$\widehat{HR_M}$$

$$\widehat{HR_M} + \frac{1}{n}\sum_{i=1}^{n}\left(A_i - \frac{1}{2}\right)h(X_i)$$

Model robust

$$(T_i(1), T_i(0), X_i, A_i) \overset{\text{i.i.d.}}{\sim} f$$

$$\Delta_{RMST(\tau)} = E\{\min(T(1), \tau)\} - E\{\min(T(0), \tau)\}$$

$$\hat{\Delta}_{RMST(\tau)}$$

$$\hat{\Delta}_{RMST(\tau)} + \frac{1}{n}\sum_{i=1}^{n}\left(A_i - \frac{1}{2}\right)h(X_i)$$

$$\widehat{HR}_M \quad \text{Vs.} \quad \hat{\Delta}_{RMST(\tau)} \qquad (1)$$

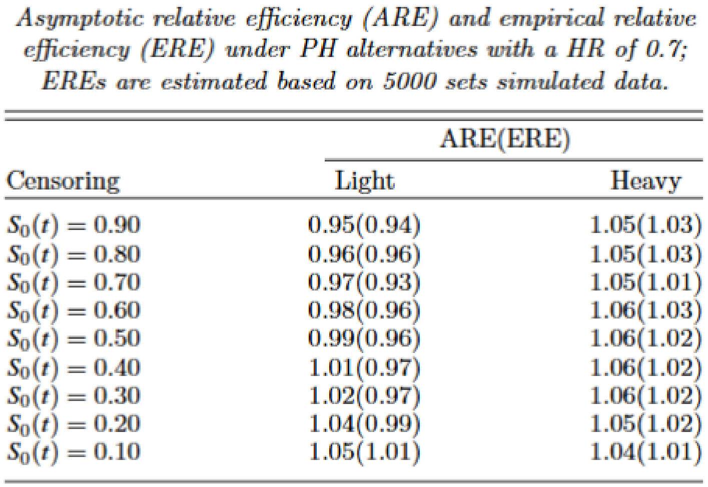
$$\widehat{HR}_M + \frac{1}{n}\sum_{i=1}^{n}\left(A_i - \frac{1}{2}\right)h(X_i) \quad \text{Vs.} \quad \hat{\Delta}_{RMST(\tau)} + \frac{1}{n}\sum_{i=1}^{n}\left(A_i - \frac{1}{2}\right)h(X_i) \qquad (2)$$

- Comparison (2) is interesting.
- Much broader than a power comparison.
- Proponents of model-robust methodology have claimed that they can improve power compared to Cox PH modelling “***even in a setting where Cox is expected to perform well.” (Chen et al., 2023).***
- But it’s difficult to compare power of (2) if we do not fully understand comparison (1).
- We do not fully understand comparison (1). → Motivation for this talk!

Efficiency of $\widehat{HR}_M$ $vs.$ $\hat{\Delta}_{RMST(\tau)}$ under proportional hazards

Tian et al., *Biometrics* 74.2 (2018): 694-702.

> *"When the PH assumption is valid, the [RMST] test performs almost as well as the PH test"*

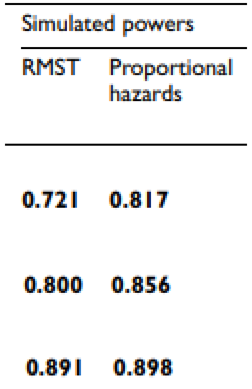


Asymptotic relative efficiency (ARE) and empirical relative efficiency (ERE) under PH alternatives with a HR of 0.7; EREs are estimated based on 5000 sets simulated data.

	ARE(ERE)	
Censoring	Light	Heavy
$S_0(t) = 0.90$	0.95(0.94)	1.05(1.03)
$S_0(t) = 0.80$	0.96(0.96)	1.05(1.03)
$S_0(t) = 0.70$	0.97(0.93)	1.05(1.01)
$S_0(t) = 0.60$	0.98(0.96)	1.06(1.03)
$S_0(t) = 0.50$	0.99(0.96)	1.06(1.02)
$S_0(t) = 0.40$	1.01(0.97)	1.06(1.02)
$S_0(t) = 0.30$	1.02(0.97)	1.06(1.02)
$S_0(t) = 0.20$	1.04(0.99)	1.05(1.02)
$S_0(t) = 0.10$	1.05(1.01)	1.04(1.01)

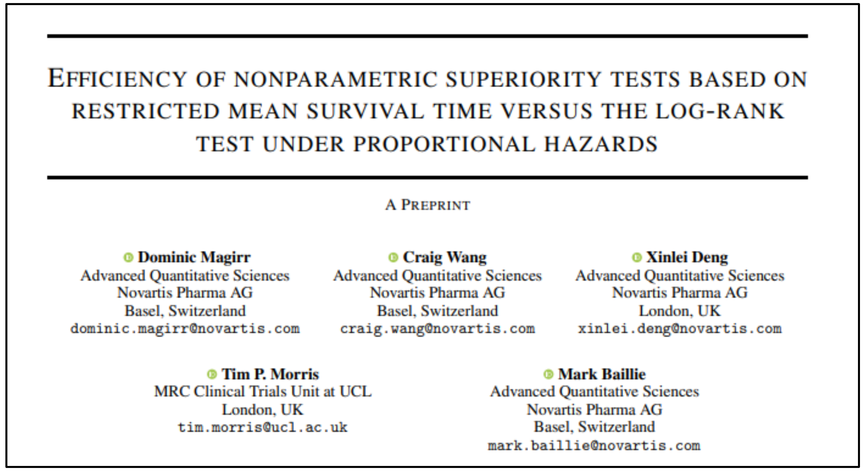
Freidlin et al., *Clinical Trials* 18.2 (2021): 188-196.

> *"For superiority testing, the proportional hazards analyses uniformly have better power than the RMST methods, although the differences are negligible in the high-event rate setting"*



	Simulated powers			
	RMST	Proportional hazards		
Low event-rate setting	**0.721**	**0.817**	→	$RE \approx 0.79$
Moderate event-rate setting	**0.800**	**0.856**	→	$RE \approx 0.86$
High event-rate setting	**0.891**	**0.898**	→	$RE \approx 0.98$

Where does this discrepancy come from?



Efficiency of nonparametric superiority tests based on restricted mean survival time versus the log-rank test under proportional hazards

A Preprint

Dominic Magirr
Advanced Quantitative Sciences
Novartis Pharma AG
Basel, Switzerland
dominic.magirr@novartis.com

Craig Wang
Advanced Quantitative Sciences
Novartis Pharma AG
Basel, Switzerland
craig.wang@novartis.com

Xinlei Deng
Advanced Quantitative Sciences
Novartis Pharma AG
London, UK
xinlei.deng@novartis.com

Tim P. Morris
MRC Clinical Trials Unit at UCL
London, UK
tim.morris@ucl.ac.uk

Mark Baillie
Advanced Quantitative Sciences
Novartis Pharma AG
Basel, Switzerland
mark.baillie@novartis.com

https://arxiv.org/pdf/2412.06442

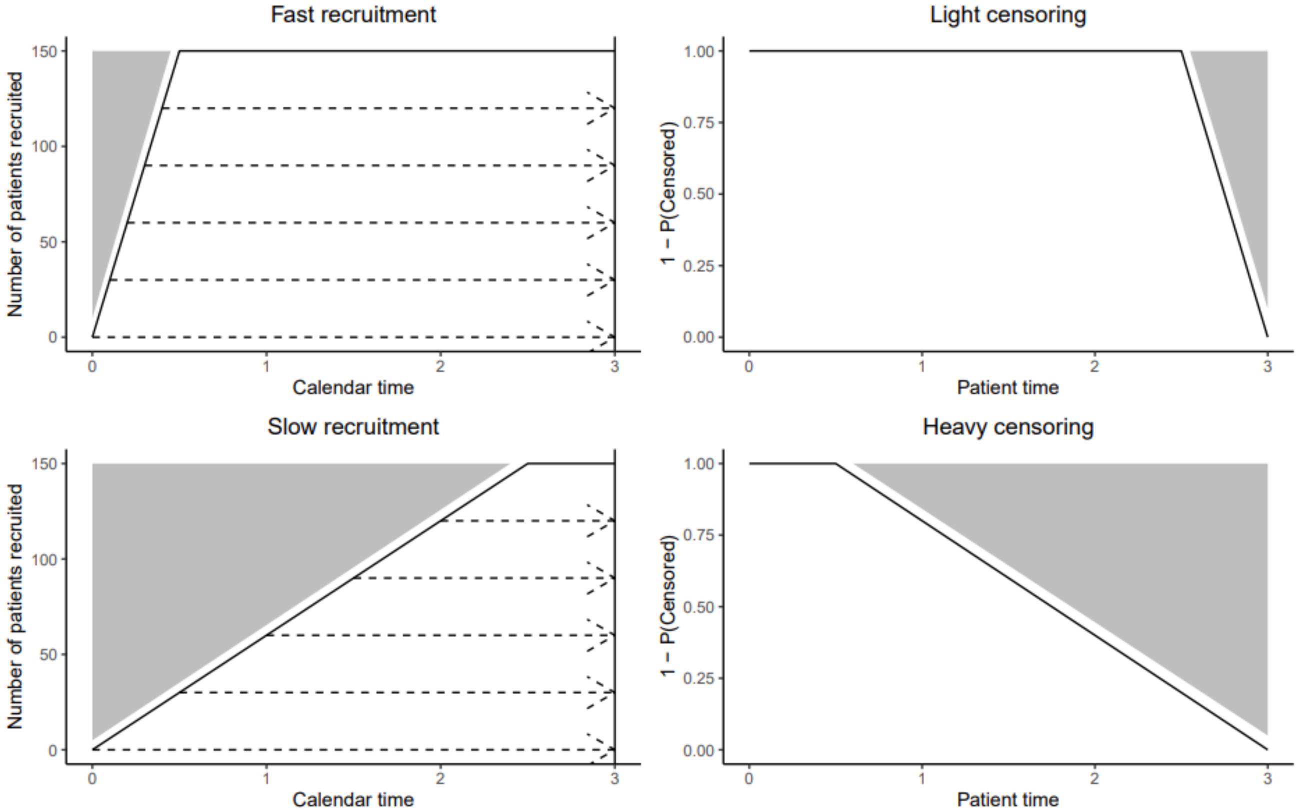


Figure 1: Illustration of how a fast recruitment rate (upper left) leads to a "light" censoring distribution (upper right), and a slow recruitment rate (lower left) leads to a "heavy" censoring distribution (lower right), when the only form of censoring is administrative censoring at the end of the study follow-up.

Asymptotics

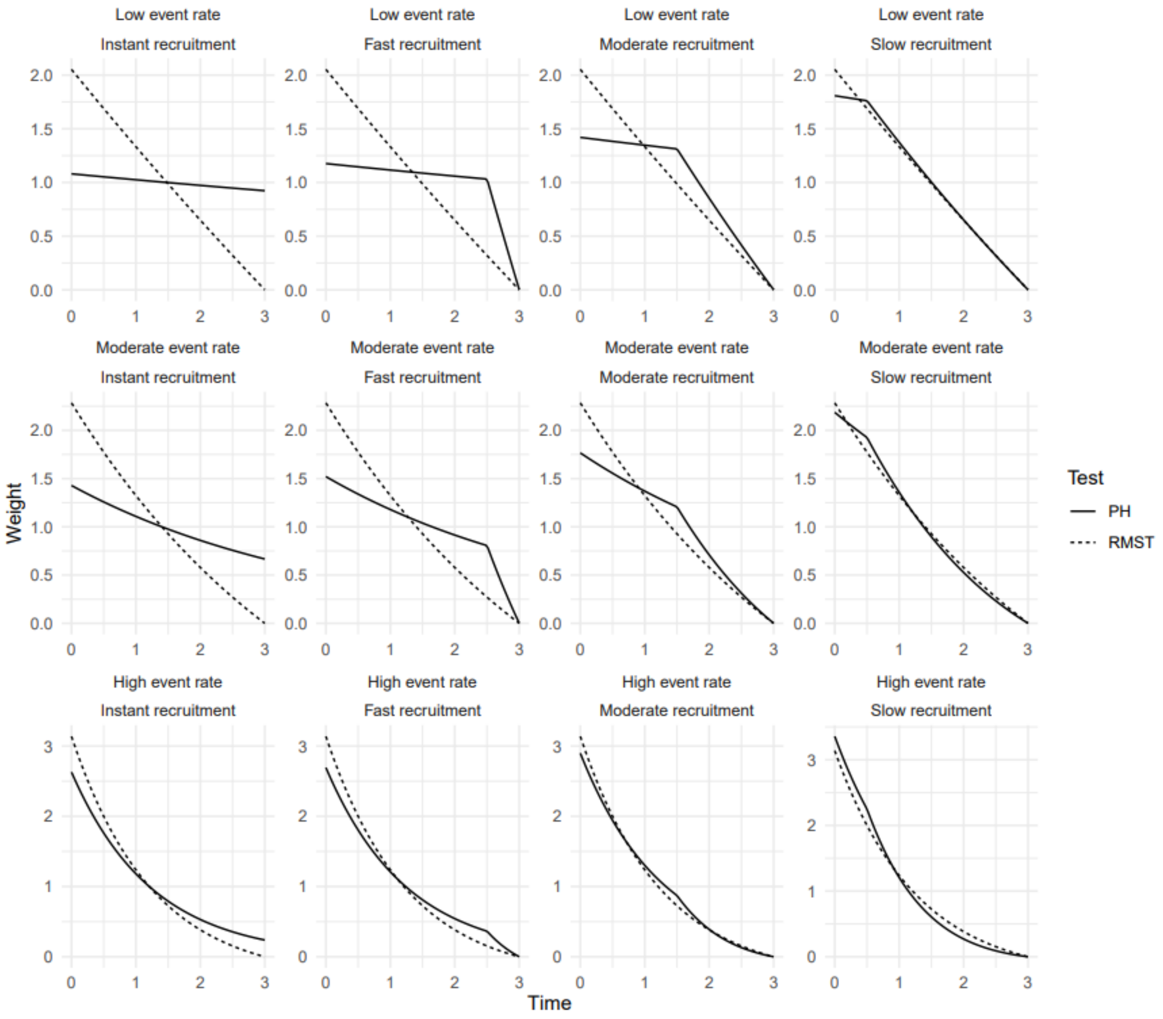
RMST test statistic:

$$\int_0^{\tau} \overset{w_{RMST}(t)}{\boxed{\frac{\int_t^{\tau} S_0(v)dv}{\int_0^{\tau} S_0(v)dv}}}\, d\left\{\widehat{\Lambda}_1(t) - \widehat{\Lambda}_0(t)\right\},$$

Log-rank statistic:

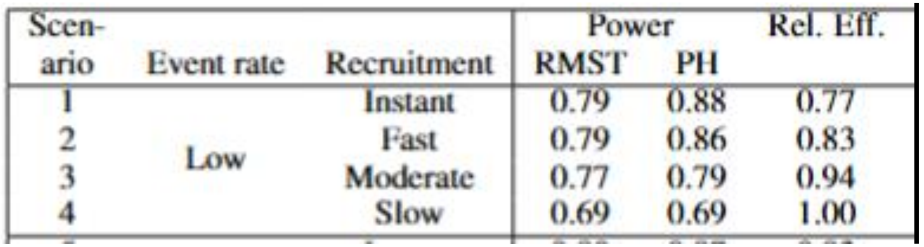
$$\int_0^{t_H} \overset{w_{PH}(t)}{\boxed{S_C(t)S_0(t)}}\, d\left\{\widehat{\Lambda}_1(t) - \widehat{\Lambda}_0(t)\right\},$$

Tian et al., (2018)



Magirr et al., (2025)

Confirmation via simulation study

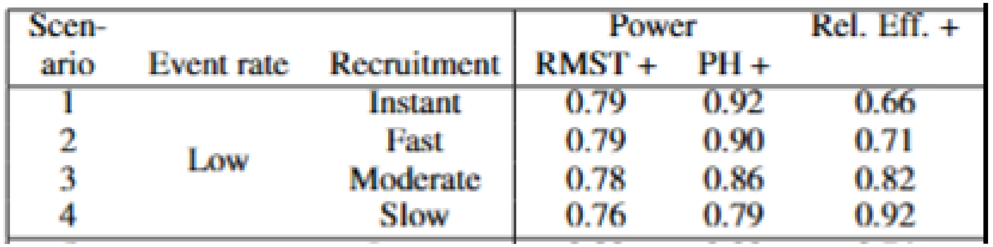


Scen- ario	Event rate	Recruitment	Power RMST	 PH	Rel. Eff.
1	Low	Instant	0.79	0.88	0.77
2		Fast	0.79	0.86	0.83
3		Moderate	0.77	0.79	0.94
4		Slow	0.69	0.69	1.00

Scenario like Freidlin et al. (2021)

Scenario like Tian et al. (2018)

Results are less favourable to RMST if restriction time τ has to be pre-specified:



Scen- ario	Event rate	Recruitment	Power RMST +	 PH +	Rel. Eff. +
1	Low	Instant	0.79	0.92	0.66
2		Fast	0.79	0.90	0.71
3		Moderate	0.78	0.86	0.82
4		Slow	0.76	0.79	0.92

Conclusions: current efficiency comparison

- There are some situations under PH where the log-rank test is substantially more efficient than a test based on a comparison of restricted mean survival time.
  - When there is a low event rate and a fast recruitment rate.
- Choice of restriction time is a difficult issue.
  - From a technical perspective, it is sometimes possible to make valid inference on RMST even when the restriction time is equal to the last observation time (Tian et al., 2020).
  - More often, it’s an easily digestible round number such as 12 or 24 months, which leads to greater efficiency loss compared to log-rank test under PH.
- Under non-PH, there is little controversy:
  - RMST is more efficient than log-rank under “early effect” scenarios.
  - Log-rank is more efficient than RMST under “late effect” scenarios.

Conclusions: wider context

- Covariate adjustment: strong arguments in favour of model-free, assumption-lean analysis methods
  - Less reliance on strong modelling assumptions.
  - Semi-parametric efficient.
- However,
  - Greater robustness to model misspecification comes at the cost of lower power under PH.
- FDA guidance on covariate adjustment (2023) has only two citations specific to time-to-event outcomes – in both cases the estimand is a marginal (unconditional) hazard ratio.
  - We should take advantage of the opportunity to adjust for **(continuous)** prognostic covariates – the methods cited in the FDA document increase precision with minimal impact on current ways of designing trials. Implementation in {RobinCar}.
  - Moving to fully model-free, assumption-lean methods would require a more radical change in study design/analysis.

References

Magirr, D., Wang, C., Deng, X., Morris, T., & Baillie, M. (2024). Efficiency of nonparametric superiority tests based on restricted mean survival time versus the log-rank test under proportional hazards. *arXiv preprint arXiv:2412.06442*.

Tian, L., Fu, H., Ruberg, S. J., Uno, H., & Wei, L. J. (2018). Efficiency of two sample tests via the restricted mean survival time for analyzing event time observations. *Biometrics*, *74*(2), 694-702.

Freidlin, B., Hu, C., & Korn, E. L. (2021). Are restricted mean survival time methods especially useful for noninferiority trials?. *Clinical Trials*, *18*(2), 188-196.

Tian, L., Jin, H., Uno, H., Lu, Y., Huang, B., Anderson, K. M., & Wei, L. J. (2020). On the empirical choice of the time window for restricted mean survival time. *Biometrics*, *76*(4), 1157-1166.

Ozenne, B. M. H., Scheike, T. H., Stærk, L., & Gerds, T. A. (2020). On the estimation of average treatment effects with right-censored time to event outcome and competing risks. *Biometrical Journal*, *62*(3), 751-763.

Ye, T., Shao, J., & Yi, Y. (2024). Covariate-adjusted log-rank test: guaranteed efficiency gain and universal applicability. *Biometrika*, *111*(2), 691-705.

Stensrud, M. J., & Hernàn, M. A. (2025). Invited Commentary: Why use methods that require proportional hazards?. *American Journal of Epidemiology*, kwae361.

Chen, D., Petersen, M. L., Rytgaard, H. C., Grøn, R., Lange, T., Rasmussen, S., ... & van der Laan, M. J. (2023). Beyond the Cox Hazard Ratio: A Targeted Learning Approach to Survival Analysis in a Cardiovascular Outcome Trial Application. *Statistics in Biopharmaceutical Research*, *15*(3), 524-539.

Tangen, C. M., & Koch, G. G. (1999). Nonparametric analysis of covariance for hypothesis testing with logrank and Wilcoxon scores and survival-rate estimation in a randomized clinical trial. *Journal of Biopharmaceutical Statistics*, *9*(2), 307-338.

Tsiatis, A. A., & Davidian, M. (2007). Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science: a review journal of the Institute of Mathematical Statistics*, *22*(4), 569.

Buja, A., Brown, L., Berk, R., George, E., Pitkin, E., Traskin, M., ... & Zhao, L. (2019). Models as approximations I. *Statistical Science*, *34*(4), 523-544.

Vansteelandt, S. (2021). Statistical modelling in the age of data science. *Observational Studies*, *7*(1), 217-228.

Bannick M, Ye T, Yi Y, Bian F (2024). *RobinCar: ROBust INference for Covariate Adjustment in Randomized clinical trials*. R package version 0.3.0.