Utility of Restricted Mean Survival Time in Oncology Clinical Trials
An Incomplete, Gentle Review of Current Developments and Methods

Chen Hu

Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins
Baltimore, MD, USA
NRG Oncology Statistics & Data Management Center
Philadelphia, PA, USA

Joint UCL-IDDI Workshop
Louvain-la-Neuve, Belgium
Dec. 11, 2018
Treatment Effect on Survival

Original Investigation

The Net Chance of a Longer Survival as a Patient-Oriented Measure of Treatment Benefit in Randomized Clinical Trials

Julien Péron, MD, PhD; Pascal Roy, MD, PhD; Brice Ozenne, PhD; Laurent Roche, PhD; Marc Buyse, ScD

Invited Commentary

Describing Differences in Survival Curves

Rick Chappell, PhD; Xiaotian Zhu, PhD

JAMA Oncology  July 2016  Volume 2, Number 7
Acknowledgment

- Lee-Jen Wei
  - Hajime Uno
  - Lu Tian
  - Brian Claggett
  - Lihui Zhao

- Rick Chappell
- Jim Dignam
- Ted Karrison
- Chiung-Yu Huang
- Yifei Sun
- Ravi Varadhan
Why is the Cox model so popular

- Traditionally, we use hazard ratio as a measurement of between treatment difference for event driven studies
  - and logrank test for hypothesis testing
- It is semi-parametric
- Allow time-dependent covariate (internal and external)
- Justification for the large sample theory
- Efficiency of the hazard ratio estimate
- Commercial software available
- No other alternatives to catch the profile of the difference between two groups over time
Cox model for association

• Define hazard (risk) level as a dependent variable which is being explained by the time-related component (so called baseline hazard) and covariates-related component

• Exploring the association between a covariate (independent variable) and survival time

• Like other regression models, it is an approximation to the true model

• It is difficult to validate an association?

• Model is based on several restrictive assumptions which need to be carefully verified before interpretation of parameters estimates
  • One assumption of proportional hazard which results directly from the model formula and means that hazard ratio needs to be constant over time
### Table 1. Advantages and disadvantages of different measures of treatment effect

<table>
<thead>
<tr>
<th>Measure</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hazard ratio</td>
<td>Almost always reported</td>
<td>Not practical for patient communication</td>
</tr>
<tr>
<td></td>
<td>Clear interpretation</td>
<td>Difficult to interpret for nonproportional hazards</td>
</tr>
<tr>
<td></td>
<td>Takes entire survival curve into account</td>
<td>Depends on choice(s) of t</td>
</tr>
<tr>
<td></td>
<td>Easy to read off survival curves</td>
<td>Loses information</td>
</tr>
<tr>
<td>Difference between survival probabilities at different time points (t)</td>
<td>Easy to read off survival curves</td>
<td>Not directly patient-relevant</td>
</tr>
<tr>
<td></td>
<td>Easy to remember</td>
<td>Not always reached</td>
</tr>
<tr>
<td>Difference between medians</td>
<td>Takes entire survival curve (until chosen time t) into account</td>
<td>Affected by schedule of assessment for end points other than overall survival</td>
</tr>
<tr>
<td></td>
<td>Does not depend on proportional hazards assumption</td>
<td>Loses information</td>
</tr>
<tr>
<td></td>
<td>Intuitive interpretation as difference between areas under the survival curves</td>
<td>Statistically unstable</td>
</tr>
<tr>
<td></td>
<td>Easy to remember</td>
<td>Almost never reported</td>
</tr>
<tr>
<td>Difference between restricted means</td>
<td>Takes entire survival curve into account</td>
<td>Difficult interpretation if survival curves are far from 0 at the largest follow-up time t</td>
</tr>
<tr>
<td></td>
<td>Does not depend on proportional hazards assumption</td>
<td>Potential for misunderstanding the key role of truncation time in its computation</td>
</tr>
<tr>
<td></td>
<td>Intuitive interpretation</td>
<td>Almost never reported</td>
</tr>
<tr>
<td>Net benefit</td>
<td>Can be readily interpreted as a net probability of benefit</td>
<td>Estimation requires a parametric distribution assumption if survival curves do not reach 0</td>
</tr>
<tr>
<td></td>
<td>Can express benefit in terms of absolute gains in survival time</td>
<td>Imprecise estimation if data are not mature (survival curves far from 0 at the largest follow-up time t)</td>
</tr>
<tr>
<td></td>
<td>Takes entire survival curve into account</td>
<td>Recently proposed, hence little experience</td>
</tr>
<tr>
<td></td>
<td>Does not depend on proportional hazards assumption</td>
<td></td>
</tr>
</tbody>
</table>
|                                 | Prioritizes the more relevant component of a composite end point | |}

Understanding and Communicating Measures of Treatment Effect on Survival: Can We Do Better? (Saad et al. (2018))
Examples when PH fails: CheckMate 057

Nivolumab versus Docetaxel in Advanced Nonsquamous NonSmall-Cell Lung Cancer (Borghaei et al. (2015))
Challenges (and Issues) of Cox Model

- **Interpretation**
  - Critique: why methodological convenience should dictate the nature of scientific question
  - HR is NOT a simple average of the hazard ratio over time
  - HR depends on underlying study-specific censoring distributions (or follow-up time)

- **Model Inconsistency**
  - When the PH is correct in each subgroup, the PH does NOT hold in the pooled sample except some special cases
  - When the PH is correct the pooled sample, the PH does not hold in all subgroups except some special cases
  - Adjusted and unadjusted analyses are estimating different quantities each other
Challenges (and Issues) of Cox Model, cont.

- Challenges in non-inferiority trials
  - The scale of null hypothesis is especially important
    - for example, consider HR=1.3 as a "tolerable" margin?
    - If the event rates are low → HR=1.3 may be clinically meaningless from the absolute risk interpretation
    - If the event rates are high → we may need a NI margin of 1.1?
  - The usual null of "no effect" for superiority trials is invariant to scale
- Causality
  - While log-rank test is asymptotically consistent, the corresponding estimated HR cannot be considered as a casual estimate
  - Likelihood contributions beyond the first are conditioned on survival past larger and larger times (Aalen et al. (2015))
Restricted Mean Survival Time

- Area under the survival curve before (restricted to) a landmark time $\tau$.
- Originally proposed by Irwin in 1948 but recently publicized by Uno et al. (2014).
- Interpreted as the mean of the minimum of the event and landmark times, or the mean life before the landmark.
  - Let $S(t)$ be the survival function for a random variable $T > 0$
    \[
    \mu_\tau = \int_0^\tau S(t) dt = E[\min(T, \tau)]
    \]
- Not the mean conditional on event occurring before that time.
  - E.g., the mean life of children in developed countries restricted to 5 years is nearly 5.
  - Life expectancy conditional on death before 5 is close to 0.
Treatment Effect based on RMST

- Difference: $\mu_{\tau,1} - \mu_{\tau,2}$
- Ratio: $\mu_{\tau,1}/\mu_{\tau,2}$
- Proportion of potential life years achieved: $\mu_{\tau,1}/\tau$
- Restricted mean time lost (RMTL), i.e., $\text{RMTL} = \tau - \text{RMST}$
  - Difference of RMTL
  - Ratio of RMTL

  When the event rate is low and the event time distribution is exponential, the ratio of RMTL will be close to the HR

$$\frac{\int_0^\tau 1 - \exp^{-\lambda_1 t} dt}{\int_0^\tau 1 - \exp^{-\lambda_2 t} dt} \approx \frac{\int_0^\tau \lambda_1 t dt}{\int_0^\tau \lambda_2 t dt} = \frac{\lambda_1}{\lambda_2}$$
CheckMate 057 Revisit

A

HR=0.73 (96% CI, 0.59-0.89), \( p=0.002 \)
\( \Delta \text{RMST}=1.7 \)m (95% CI, 0.4-3.1); \( P = .01 \)

B

Pak et al. (2017)
Example 2: ECOG E4A03 Trial

- E4A03 trial to compare low- and high-dose dexamethasone for patients with newly diagnosed multiple myeloma
- One of the endpoints is overall survival, $n = 445$.
- The trial stopped early at the second interim analysis; the low dose was superior.

Uno et al. (2014)
- **Cox PH analysis**
  - The proportional hazards assumption is not valid.
  - The PH estimator is estimating a quantity which cannot be interpreted and, worse, depends on the study-specific censoring distributions.
  - The logrank test is not powerful.
  - In conventional analysis, we have log-rank test: $p = 0.47$ and hazard ratio: $HR=0.87$ (0.60, 1.27).

- **RMST analysis**
  - Restricted mean (up to 40 months).
  - 35.4 months vs. 33.3 months.
  - Difference $= 2.1$ (0.1, 4.2) months; $p = 0.04$.
  - Ratio of RMST $= 35.4/33.3 = 1.06$ (1.00, 1.13).
  - Ratio of RMTL $= 6.7/4.6 = 1.46$ (1.02, 2.13).
Nonparametric Estimation and Inference of RMST

- **Notations**
  - \( T_i \): failure time; \( C_i \): (independent) censoring time
  - \( Y_i = T_i \wedge C_i, \Delta_i = I(T_i \leq C_i) \)
  - \((Y_i, \Delta_i, X_i)\): observed data
  - \( Y_i^\tau = Y_i \wedge \tau, \Delta_i^\tau = I(T_i \wedge \tau \leq C_i) \)
  - \((Y_i^\tau, \Delta_i^\tau, X_i)\): derived data based on \( \tau \)

- **KM-based estimator**, where \( \hat{S} \) is a KM estimator of \( T \)
  \[
  \tilde{\mu}_\tau = \int_0^\tau \hat{S}(t)dt
  \]

- **Inference**: Based on the martingale approach (Andersen et al. (2012)), we have a variance estimator of \( \tilde{\mu}_\tau \)
  \[
  \hat{V}(\tilde{\mu}_\tau) = \sum_{i=1}^D \left\{ \int_{t_i}^\tau \hat{S}(t)dt \right\} \frac{d_i}{R(t_i)[R(t_i) - d_i]}
  \]
  where \( d_i \) and \( R(t_i) \) is the number of events and risk set at \( t_i \), for \( t_1 < t_2 < \cdots < t_D \)
Inverse probability censoring weighting (IPCW) approach

\[ \hat{\mu}_\tau = n^{-1} \sum_{i=1}^{n} \frac{\Delta_i^\tau}{\hat{G}(Y_i^\tau)} Y_i^\tau \]

where \( \hat{G}(\cdot) \) is the KM estimator of censoring time \( C \)

Based on results from Satten and Datta (2001), we also have

\[ \hat{S}(t) = n^{-1} \sum_{i=1}^{n} I(Y_i > t) \frac{\Delta_i^\tau}{\hat{G}(Y_i^\tau)} + O_p(n^{-1/2}) \]

with some algebra, we can show \( \hat{\mu}_\tau - \tilde{\mu}_\tau = O_p(n^{-1/2}) \), e.g., \( \hat{\mu}_\tau \) and \( \tilde{\mu}_\tau \) are asymptotically equivalent at \( n^{-1/2} \) rate.

Provide a natural connection for building an ANCOVA-type regression model (Tian et al. (2014))
Two-sample Testing

- Logrank test
  - Robust
  - The most powerful under PH alternatives
  - Various weighted versions exist

- RMST-based testing
  - Convert the estimated treatment effect into a coherent test, e.g., Uno et al. (2015), Tian et al. (2018)
  - Power *depends* on the pattern of difference, $\tau$, etc.,
  - For a fixed $\tau$, should *NOT* assume it will be better than log-rank test even under non-PH
    - if KM curves separate early $\rightarrow$ likely more powerful
    - if KM curves separate late $\rightarrow$ possibly less powerful
Saxagliptin and Cardiovascular Outcomes in Patients with Type 2 Diabetes Mellitus


Scirica et al. (2013)
SAVOR-TIMI 53 (saxagliptin vs. placebo)

- Primary endpoint: time to CV death, nonfatal MI, or nonfatal ischemic stroke
- 1040 primary events needed to show superiority (efficacy)
- 457 primary events needed to show non-inferiority
  - upper bound of HR<1.3, for safety
  - no matter what the underlying event rates are
- A total of 16,492 patients were enrolled
- Median follow-up time: 2.1 years
- Observed events: 613 (Saxagliptin) vs. 609 (Placebo)
SAVOR-TIMI 53 trial: Primary Endpoint

A Primary End Point

Hazard ratio, 1.00 (95% CI, 0.89–1.12)
P<0.001 for noninferiority
P=0.99 for superiority

2-yr Kaplan–Meier rate:
Saxaglaptin, 7.3%
Placebo, 7.2%

No. at Risk
Placebo  8212  7983  7761  7267  4855  851
Saxaglaptin  8280  8071  7836  7313  4920  847

Uno et al. (2015)
SAVOR-TIMI 53 trial: RMST

A. Area above the cumulative incidence (Saxagliptin)

B. Area above the cumulative incidence (Placebo)

RMST: 860 days
SAVOR-TIMI 53 trial: Impacts to Sample Size

- Three methods compared
  - Hazard Ratio (HR)
  - Difference in event rate at Day 900 ($\Delta \hat{S}(900)$)
  - Difference in RMST at Day 900 ($\Delta \hat{\mu}(900)$)

<table>
<thead>
<tr>
<th>Estimate</th>
<th>All Data</th>
<th>25%</th>
<th>20%</th>
<th>15%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N=16,492</td>
<td>N=4123</td>
<td>N=3298</td>
<td>N=2427</td>
</tr>
<tr>
<td>HR</td>
<td>1.00</td>
<td>(0.89, 1.12)</td>
<td>(0.80, 1.26)</td>
<td>(0.78, 1.28)</td>
</tr>
<tr>
<td>$\Delta \hat{S}(900)$</td>
<td>0%</td>
<td>(-1.2, 0.9)</td>
<td>(-2.3, 2.0)</td>
<td>(-2.6, 2.2)</td>
</tr>
<tr>
<td>$\Delta \hat{\mu}(900)$</td>
<td>0 day</td>
<td>(-5, 4)</td>
<td>(-9, 9)</td>
<td>(-11, 10)</td>
</tr>
</tbody>
</table>

Uno et al. (2015)
Design NI or Safety Trial with RMST

- The standard approach using HR requires a practically infeasible size of a safety study when the event rate is very low (e.g., annual event rate 1% - 1.5%)
- Difference of RMST provides a CI tight enough to make a decision about safety of the new therapy with much smaller study
- The clinical interpretation is crucial for a safety or superiority study
Construction of RMST

- Restricted Mean Survival Time (RMST), for $0 \leq t \leq \tau$
  - $m(t) = E\{\min(D, t)\} = \int_0^t S_D(u)du$
  - $m(t) = E\{\int_0^t I(D \geq u)du\}$
  - $\int_0^t I(D \geq u)du$: cumulative at-risk process, e.g., area under at-risk process

![Graphs of Patient 1, Patient 2, and Patient 3 showing Y(t) against Time.]
Analysis of Duration of Response (DOR)

- Duration of response: time from response (R) to progression/death (P/D)
- Clinically meaningful, yet challenging to summarize due to its post-randomization nature
- Kaplan-Meier is often used for descriptive summary
- Huang et al. (2018) reported the marginal RMST of DOR
Analysis of Duration of Response (DOR), cont.

- $\mu_{\tau,R}$: RMST for min$(R, P, D)$
- $\mu_{\tau,PD}$: RMST for min$(P, D)$
- $\mu_{\tau,DOR} = \mu_{\tau,PD} - \mu_{\tau,R}$
Analysis of Duration of Response (DOR), cont.

- $\Delta \mu(30) = 7.4$ months (95% CI, 6.0-8.8 months; $P < .001$).

Huang et al. (2018)
Glasziou et al. (1990) analyzed the Ludwig III randomized clinical trial of adjuvant Pt-based chemotherapy vs. observation in nodal-invasive breast cancer using restricted mean life.

The landmark time reflected patient followup: 7 years.

Survival states: Toxicity, Health (Time Without Symptoms and Toxicity, TWiST), Relapse, Death.
Survival Partition for Quality-Adjusted Survival, cont.

Glasziou et al. (1990)
Survival Partition for Quality-Adjusted Survival, cont.

Glasziou et al. (1990)
Competing risks are often of interest when localized therapy, e.g., surgery or radiotherapy is under investigation.

- **Failure time** $T^*$
- **Right censoring time** $C$
- **Observed data** $T = \min(T^*, C), \delta = I(T^* < C), D = I(\delta = 1)D^*$

For example, in lung cancer trial:
- $T$: time to first failure (subject to censoring)
- $\delta$: event indicator, 1 = event, 0 = censoring
- $D = 1$: in-field failure;
  $D = 2$: out-field failure;
  $D = 3$: death without cancer recurrence
Cumulative Incidence Function (CIF)

- Area under CIF is Restricted Mean Time Loss (RMTL) to the specific failure cause.
Zhao et al. (2018) argued analysis using RMTL should be included when reporting competing risks.
• Treatment effects on different endpoints may point to different directions, creating challenges and difficulties to clinicians

• Allogeneic bone marrow transplantation (BMT) is widely used to re-establish the damaged hematopoietic function in treating acute and chronic leukemia and other hematological malignancies

• Multiple types of events can occur in Post-BMT
  • Relapse
  • Graft-versus-host disease (GVHD)
  • Death

• Evaluating different treatment options is challenging, especially when treatment can have heterogeneous effects or even qualitatively differing impacts on different events
Example in BMT: ATG Trial

A  Incidence of Clinical Extensive Chronic GVHD

B  Relapse

No. at Risk
ATG  Non-ATG

No. at Risk
ATG  Non-ATG
Other Results from ATG Trial

C Relapse-free Survival

D Overall Survival

E Nonrelapse-Related Death

F Chronic GVHD–free+Relapse-free Survival

No. at Risk
ATG 83 76 61 58 55 52 49 47 33
Non-ATG 72 67 61 60 58 56 54 54 35

No. at Risk
ATG 83 78 70 63 62 58 54 53 36
Non-ATG 72 68 64 63 61 60 59 56 35

No. at Risk
ATG 83 76 47 42 37 35 34 34 22
Non-ATG 72 67 32 21 19 17 16 15 8
Interpretations of ATG Trial

- ATG substantially decreased incidence of chronic GVHD (panel A)
- ATG also increased the incidence of relapse (panel B)?
- RFS (panel C) and OS (panel D) are slightly better for non-ATG?
- Confusions remain among clinicians
Reverse Counting Process

- Proposed by Prof. L.J. Wei (Claggett et al. (2018))

- $D$: time to terminal event

- $Y(t) = \sum_{k=1}^{K} I(T_k \geq t) + I(D \geq t)$, reverse counting process with $K$ distinct morbidity events
  - $T_i$ time to morbidity $i$, such as GVHD or relapse
  - reflects individual’s disease burden and health condition over time
  - $Y(\cdot)$ after $D$ is not defined
Reverse Counting Process Illustration

- **Patient 1**: Censored, Death, GVHD, Relapse
- **Patient 2**: Censored, GVHD, Death
- **Patient 3**: GVHD, Relapse

Frequency plots for each patient:

- **Patient 1**: Y(t)
  - Y(t) = 1 at t = 10
  - Y(t) = 2 at t = 20

- **Patient 2**: Y(t)
  - Y(t) = 1 at t = 10
  - Y(t) = 2 at t = 15

- **Patient 3**: Y(t)
  - Y(t) = 1 at t = 10
  - Y(t) = 2 at t = 15

Time axis (0-20) with intervals (0, 5, 10, 15, 20)
Cumulative Marker Process in Presence of Terminal Event

- $M(t) = \int_0^t Y^*(u)I(D \geq u)\,du$, cumulative marker process
  - $Y^*(u)I(D \geq u)$ takes 0 after terminal event occurs
  - Area under marker trajectory
- Cumulative Mean $\mu(t) = E\{M(t)\} = \int_0^t E\{Y^*(u)I(D \geq u)\}\,du$
  - Ideal treatment: prolong survival and maintain high marker value
Standardized Summary Metric: “Morbidity”-Adjusted RMST
For the non-standardized case, \( Y^*(t) = \sum_{k=1}^{K} I(T_k \geq t) + 1 \), Claggett et al. (2018) considered

\[
\mu(t) = \int_0^t E\{Y^*(u)I(D \geq u)du\}
\]

\[
= \int_0^t E\{Y(u)du\}
\]

\[
= E\{\sum_{k=1}^{K} \min(T_k, D, t) + \min(D, t)\}
\]

• sum of (restricted) mean event-free survival times up to \( t \)
Nonparametric Estimation

- Induced informative censoring of $M(\cdot)$
- Consider the framework proposed by Sun et al. (2017) for benefit-risk assessment in general
- Notice
  \[
  \mu(t) = \int_0^t E\{Y^*(u)I(D \geq u)\} \, du = \int_0^t S_D(u) E\{Y^*(u)|D \geq u\} \, du
  \]
- Moment-type estimator
  \[
  \hat{\mu}(t) = \int_0^t \hat{S}_D(u) \frac{\sum_{i=1}^n Y_i^*(u)I(X_i \geq u)}{\sum_{i=1}^n I(X_i \geq u)} \, du
  \]
  where $\hat{S}_D(u)$ is K-M estimator of $S_D(u)$, the survivor function of $D$

- Theorems (Sun et al. (2017))
  - Consistency of $\hat{\mu}$
  - Weak convergence of $\sqrt{n}\{\hat{\mu}(t) - \mu(t)\}(0 \leq t \leq \tau)$
- Weights can/should be flexibly incorporated to reflect individual preferences for morbidities
  - Different weights can be used as personalized decision making tools
Bone Marrow Transplant Data Analysis

- A multi-center, non-comparative trial of patients prepared for allogeneic marrow transplants with a radiation-free conditioning regimen for patients with acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) (Copelan et al. (1991))

- Here we analyze and “compare” ALL and High-risk AML patients for illustration purpose solely
Regular and Standardized RMST

RMST (yrs): ALL

RMST (yrs): AML High Risks
Regular and Standardized RMST Differences

RMST (yrs): ALL vs. AML High Risks

RMST Difference p-value
Covariate Adjustment

- Linear models
  - Adjusting for covariates associated with outcome $Y$ increases the precision of the treatment effect estimate

- Logistic regression and Cox PH regression
  - SE of the treatment effect increases (or at best does not decrease)
  - Maybe still want to adjust (for prognostic covariates) because unadjusted estimate is biased towards the null
  - Such “bias” is inherent due to the fact that the unadjusted and adjusted model estimates different measures of treatment effect

- Covariate-adjusted RMST?
  - Gains in precision is not granted
Covariate-adjusted RMST Difference

- Karrison (1987) incorporated covariates $Z$ by fitting a piecewise exponential model, e.g.,
  \[ \lambda_g(t|Z) = \lambda_{gk} \exp(Z\beta), \quad t \in (t_{k-1}, t_k], \quad g = 1, 2 \]

- hazard functions are piecewise constant on $t \in (t_{k-1}, t_k]$
- non-PH is allowed for treatment effect; PH is assumed for covariates $Z$
- Zucker (1998) considered
  \[ \lambda_g(t|Z) = \lambda_{0g}(t) \exp(Z\beta), \quad g = 1, 2 \]

- Baseline hazard $\lambda_{0g}$ completely unspecified, no need to specify the intervals $(t_{k-1}, t_k]$
- $\hat{\mu}_{\tau,g}$ can be estimated by integrating $\hat{S}_g(t|Z) = \exp[-e^{Z\beta}\hat{\Lambda}_{0g}(t)]$
- Average $S_g(t)$ over the entire covariate distribution in both arms
- Covariate-adjusted RMST has similar results as linear model (Karrison and Kocherginsky (2018)), e.g., covariate adjustment offers unbiased and more precision estimation of treatment effect
Tian et al. (2014) considered a regression model with link function
\[ \eta(\mu_{\tau}(X)) = \alpha + \beta X \]

With logic link, \( \beta \) of the treatment indicator becomes
\[
\log\left\{ \frac{\mu_{\tau,1}(\tau - \mu_{\tau,2})}{\mu_{\tau,2}(\tau - \mu_{\tau,1})} \right\}
\]

an odds-ratio like summary for the group contrast

From the IPCW expression, we have the estimating equation
\[
S_n(\beta) = \sum_{i=1}^{n} \frac{\Delta^T_i}{\hat{G}(Y^T_i)} X_i \left\{ Y^T_i - \eta^{-1}(\beta X_i) \right\}
\]

True parameter \( \beta_0 \) can be estimated by solving \( S_n(\beta) = 0 \). It can be shown \( \hat{\beta} \) is consistent

Inference can be obtained through perturbation-resampling method (Tian et al. (2005))
Choice of the truncation time point, $\tau$

- In a confirmatory study, $\tau$ should be pre-specified
  - Often if not always difficult
- The choice would depend on
  - clinical motivation or interest (short-term? Long-term?)
  - Follow-up time of the study
  - Precision at the tail part of the KM curves
- When choosing $\tau$ a posteriori, objective rules like “effective sample size” (Karrison (1987)) can be useful
  - e.g., choose the largest $t$ such that $\hat{N}_{Eff}(t) > \frac{2}{3}N$, where
    $$\hat{N}_{Eff}(t) = \frac{\hat{S}(t)(1-\hat{S}(t))}{\hat{V}\{\hat{S}(t)\}}$$
Final Notes

- Collective efforts to change the culture of reporting and interpreting HR alone
- Model-free and clinically interpretable metrics like RMST should be appreciated and better interpreted
- A lot of potential (and fun) to exploit the additivity of RMST


Thank You!