Dana-Farber Cancer Institute

HARVARD MEDICAL SCHOOL

Restricted Mean Survival Time in Practice: An Easy-to-Understand Approach

Angel Cronin, Xiang Meng, and Hajime Uno

Department of Data Science

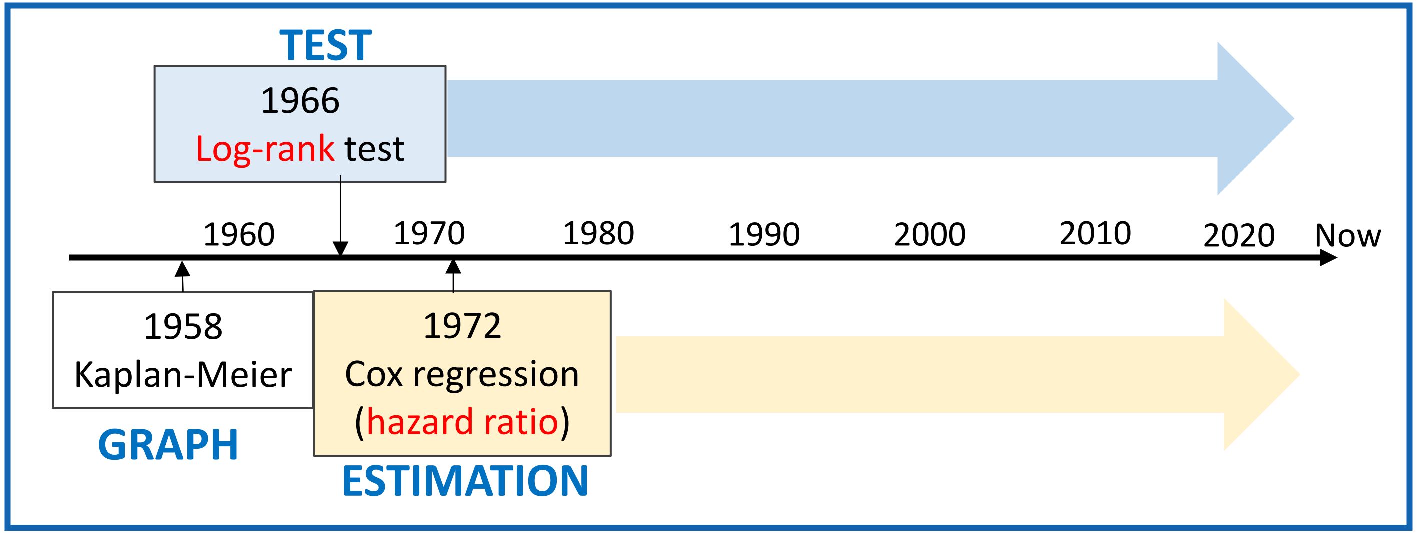
Dana-Farber Cancer Institute

November 5 and 12, 2025

Data Science Training 2025

Welcome!

Methods for time-to-event data

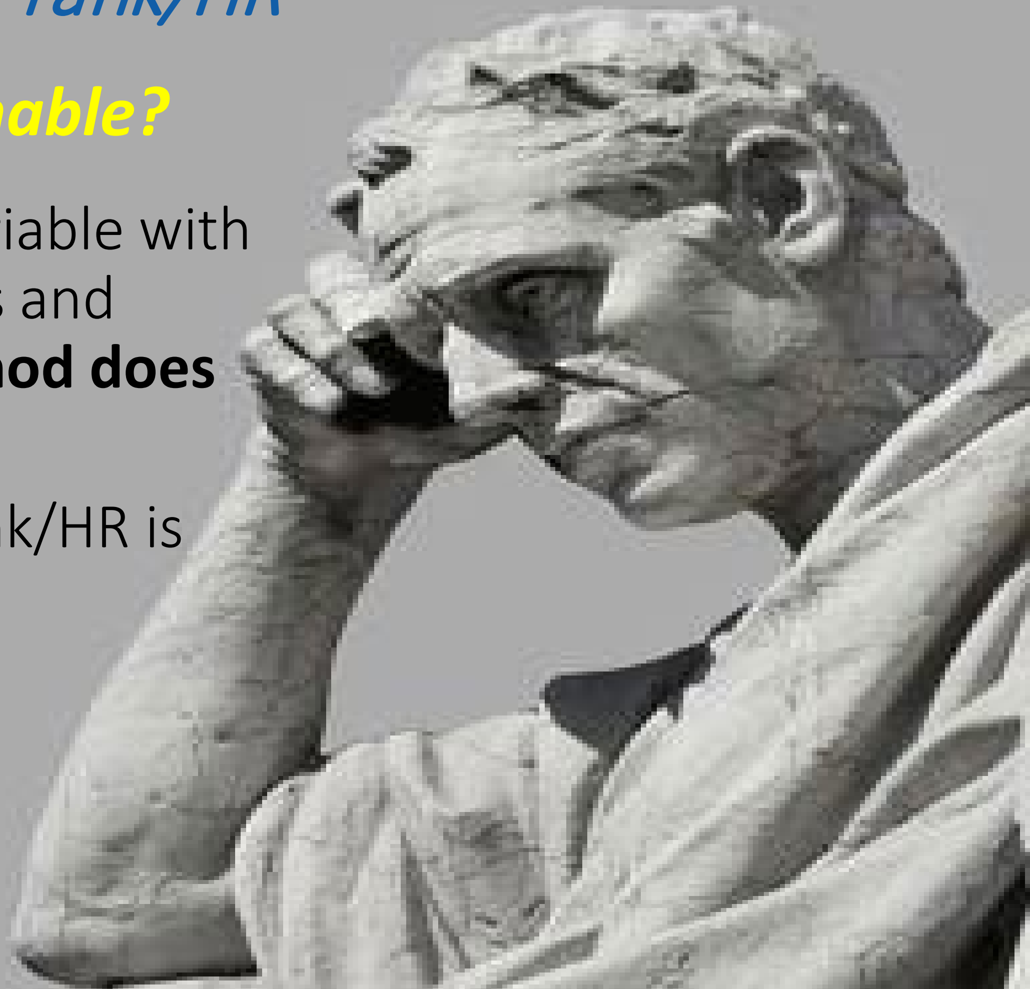


>95% of cancer RCTs (ph3) were using this Test/Estimation method (Uno et al. 2020, Oncologist)

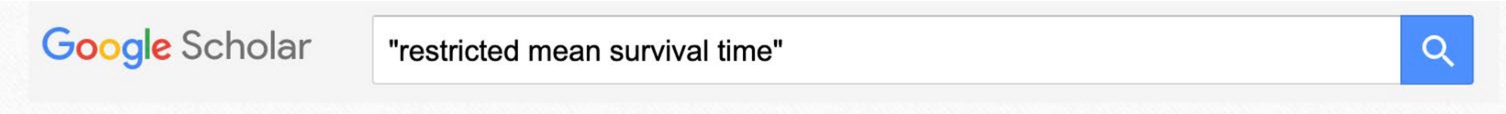
Near universal use of log-rank/HR

Is this tradition reasonable?

- Clinical research are highly variable with respect to their characteristics and research questions. **One method does not fit all.**
- ***No method is perfect.*** Log-rank/HR is not an exception.

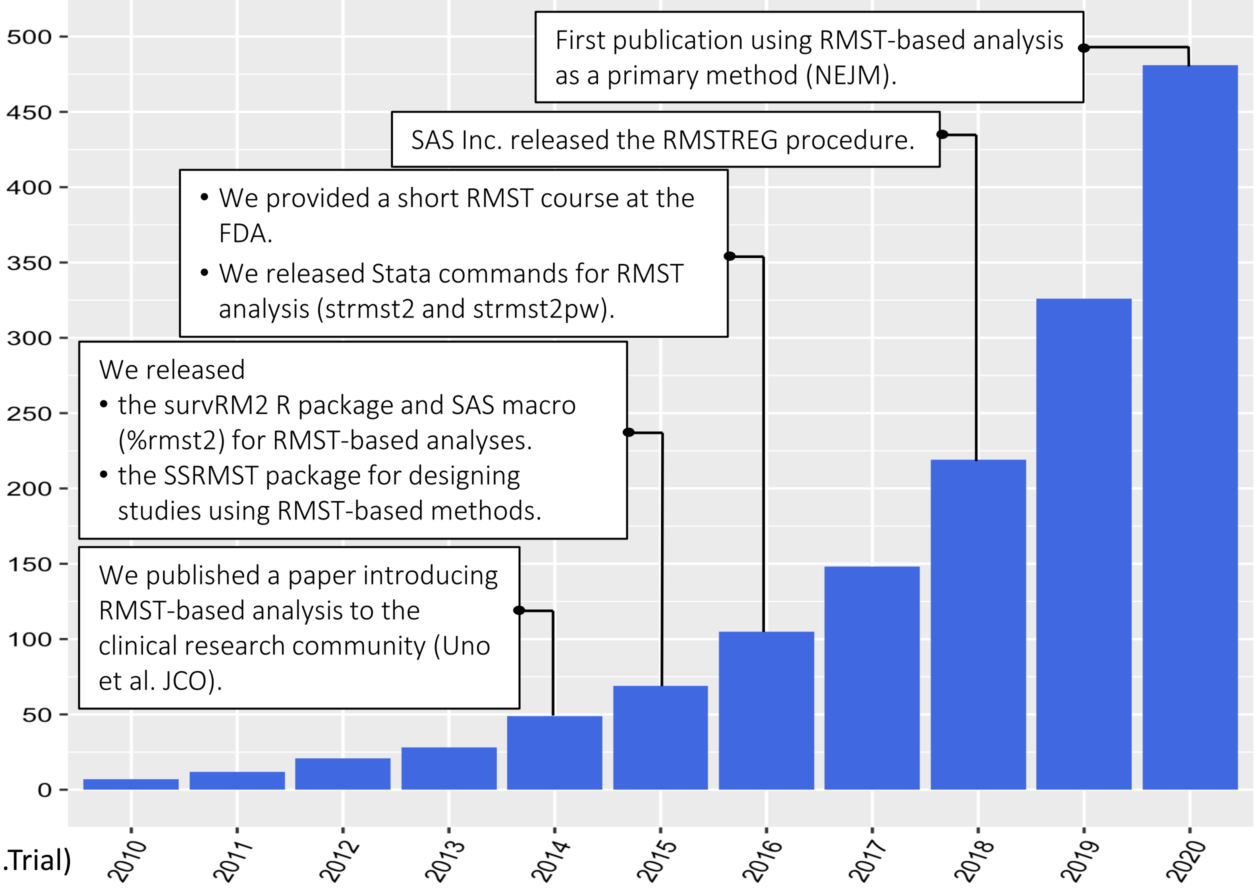


Signs of change



"restricted mean survival time"

RMST is getting popular…



First publication using RMST-based analysis as a primary method (NEJM).

SAS Inc. released the RMSTREG procedure.

- We provided a short RMST course at the FDA.
- We released Stata commands for RMST analysis (strmst2 and strmst2pw).

We released

- the survRM2 R package and SAS macro (%rmst2) for RMST-based analyses.
- the SSRMST package for designing studies using RMST-based methods.

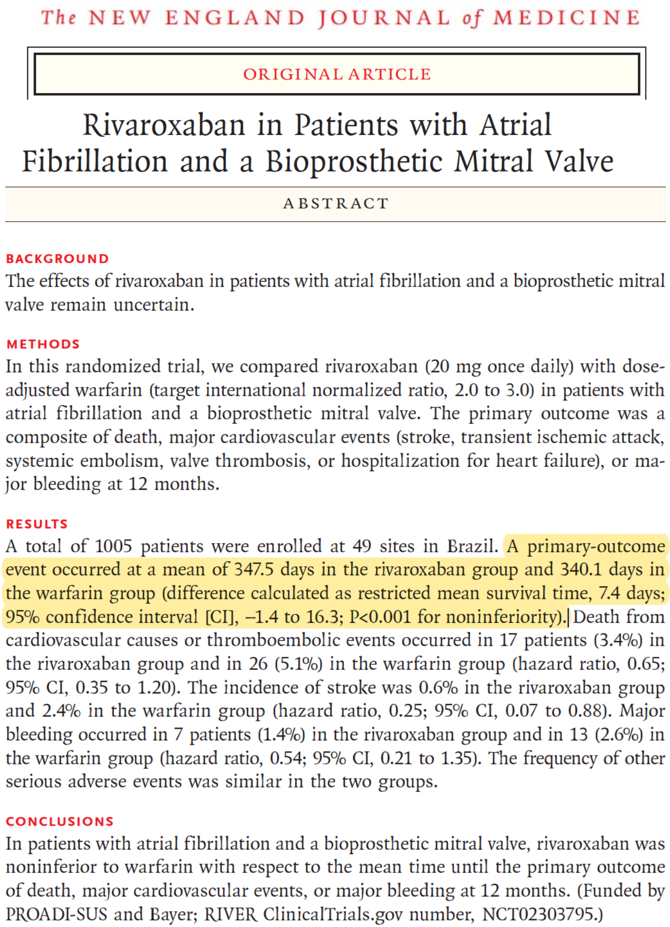
We published a paper introducing RMST-based analysis to the clinical research community (Uno et al. JCO).

Glasziou, Simes, Gelber (**1990**)
Partitioned survival curve
(Quality Adjusted Survival)

RMST was proposed
by Irwin (**1947**)

Karrison
(**1997**, Cont.Clin.Trial)

Guimarães et al. (November 14, 2020)



The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Rivaroxaban in Patients with Atrial Fibrillation and a Bioprosthetic Mitral Valve

ABSTRACT

BACKGROUND

The effects of rivaroxaban in patients with atrial fibrillation and a bioprosthetic mitral valve remain uncertain.

METHODS

In this randomized trial, we compared rivaroxaban (20 mg once daily) with dose-adjusted warfarin (target international normalized ratio, 2.0 to 3.0) in patients with atrial fibrillation and a bioprosthetic mitral valve. The primary outcome was a composite of death, major cardiovascular events (stroke, transient ischemic attack, systemic embolism, valve thrombosis, or hospitalization for heart failure), or major bleeding at 12 months.

RESULTS

A total of 1005 patients were enrolled at 49 sites in Brazil. A primary-outcome event occurred at a mean of 347.5 days in the rivaroxaban group and 340.1 days in the warfarin group (difference calculated as restricted mean survival time, 7.4 days; 95% confidence interval [CI], –1.4 to 16.3; P<0.001 for noninferiority). Death from cardiovascular causes or thromboembolic events occurred in 17 patients (3.4%) in the rivaroxaban group and in 26 (5.1%) in the warfarin group (hazard ratio, 0.65; 95% CI, 0.35 to 1.20). The incidence of stroke was 0.6% in the rivaroxaban group and 2.4% in the warfarin group (hazard ratio, 0.25; 95% CI, 0.07 to 0.88). Major bleeding occurred in 7 patients (1.4%) in the rivaroxaban group and in 13 (2.6%) in the warfarin group (hazard ratio, 0.54; 95% CI, 0.21 to 1.35). The frequency of other serious adverse events was similar in the two groups.

CONCLUSIONS

In patients with atrial fibrillation and a bioprosthetic mitral valve, rivaroxaban was noninferior to warfarin with respect to the mean time until the primary outcome of death, major cardiovascular events, or major bleeding at 12 months. (Funded by PROADI-SUS and Bayer; RIVER ClinicalTrials.gov number, NCT02303795.)

RMST was used as the primary analysis

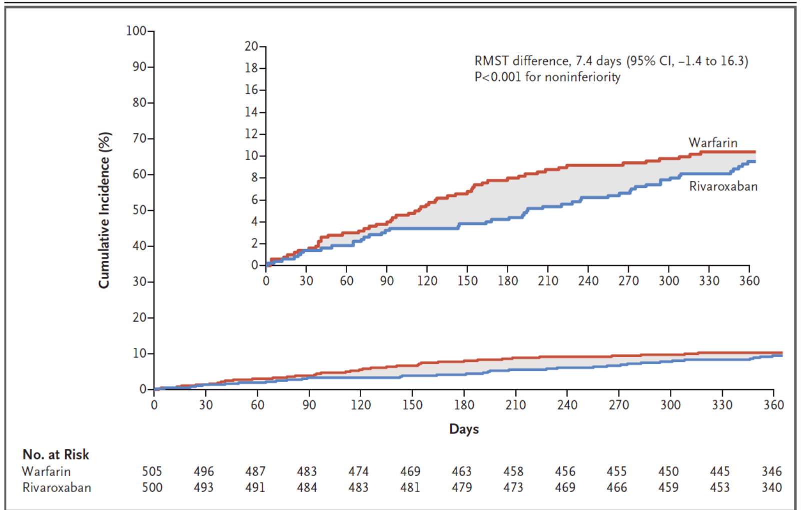
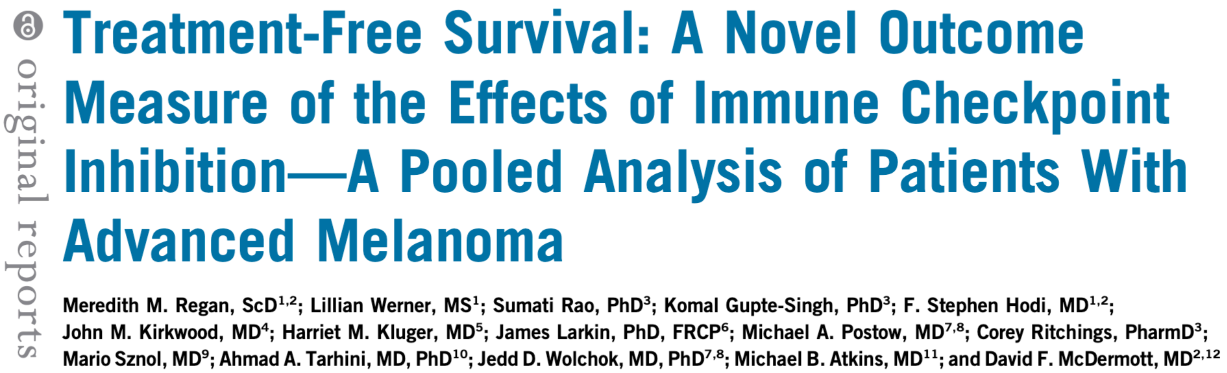


Figure 1. Kaplan–Meier Analysis of the Primary Outcome.

Shown is the primary outcome (death, major cardiovascular events, or major bleeding) in the rivaroxaban group and the warfarin group, as calculated according to the restricted mean survival time (RMST) method. The inset shows the same data on an expanded y axis.

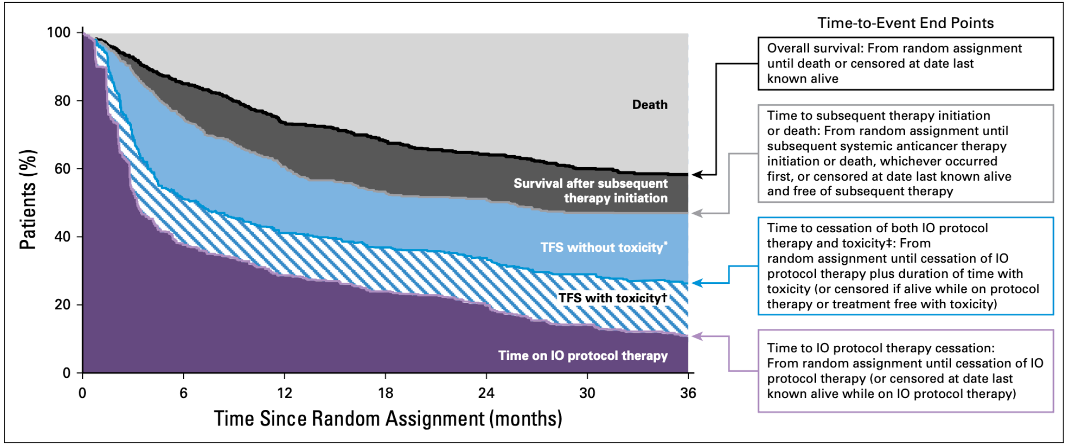
Regan et al. 2019 JCO



original reports

Treatment-Free Survival: A Novel Outcome Measure of the Effects of Immune Checkpoint Inhibition—A Pooled Analysis of Patients With Advanced Melanoma

Meredith M. Regan, ScD[1,2]; Lillian Werner, MS[1]; Sumati Rao, PhD[3]; Komal Gupte-Singh, PhD[3]; F. Stephen Hodi, MD[1,2]; John M. Kirkwood, MD[4]; Harriet M. Kluger, MD[5]; James Larkin, PhD, FRCP[6]; Michael A. Postow, MD[7,8]; Corey Ritchings, PharmD[3]; Mario Sznol, MD[9]; Ahmad A. Tarhini, MD, PhD[10]; Jedd D. Wolchok, MD, PhD[7,8]; Michael B. Atkins, MD[11]; and David F. McDermott, MD[2,12]



Overall survival: From random assignment until death or censored at date last known alive

Time to subsequent therapy initiation or death: From random assignment until subsequent systemic anticancer therapy initiation or death, whichever occurred first, or censored at date last known alive and free of subsequent therapy

Time to cessation of both IO protocol therapy and toxicity‡: From random assignment until cessation of IO protocol therapy plus duration of time with toxicity (or censored if alive while on protocol therapy or treatment free with toxicity)

Time to IO protocol therapy cessation: From random assignment until cessation of IO protocol therapy (or censored at date last known alive while on IO protocol therapy)

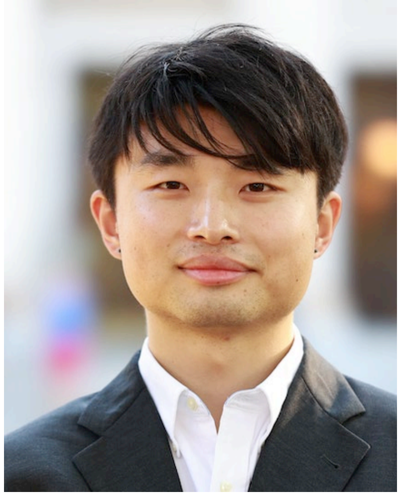
FIG 1. Illustration of the end points that partition the area under the overall survival curve into treatment-free survival (TFS) and other resulting health states. (*) Time after cessation of immuno-oncology (IO) protocol therapy without toxicity before initiation of subsequent systemic anticancer therapy or death. (†) Time after cessation of IO protocol therapy with toxicity while treatment free. (‡) Includes toxicity that persisted since protocol therapy and toxicity that newly presented after protocol therapy cessation.

Overview of the course

Goal: Participants can immediately apply the RMST methods in practice when appropriate.

Instructors:

Angel Cronin



Xiang Meng

Hajime Uno

Schedule

Date	Contents
2025-11-05 (120 min)	Part 1: Estimation of between-group difference - Limitation of Hazard Ratio - RSMT definition - Examples
2025-11-12 (120 min)	Part 2: More on RMST - Power considerations - Study design - Regression analysis - Stratified analysis - Other applications of RMST

SCAN ME

Each includes

- 10 min break
- Demo and exercise using R (all materials are on the DS Training webpage)
- Feedback and Q&A

Hazard Ratio and its limitations

Hazard function plays an important role in survival analysis

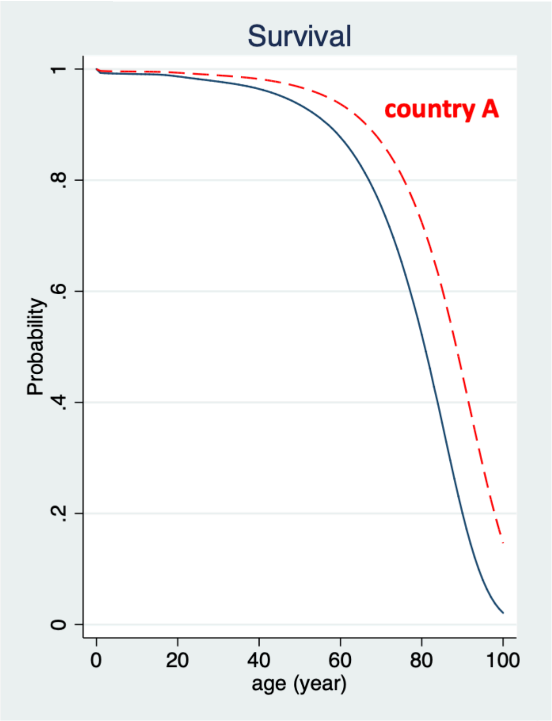
- Hazard function

$$\lambda(t) = \lim_{\Delta \to 0+} \frac{\Pr\{t \leq T < t + \Delta \mid T \geq t\}}{\Delta}$$

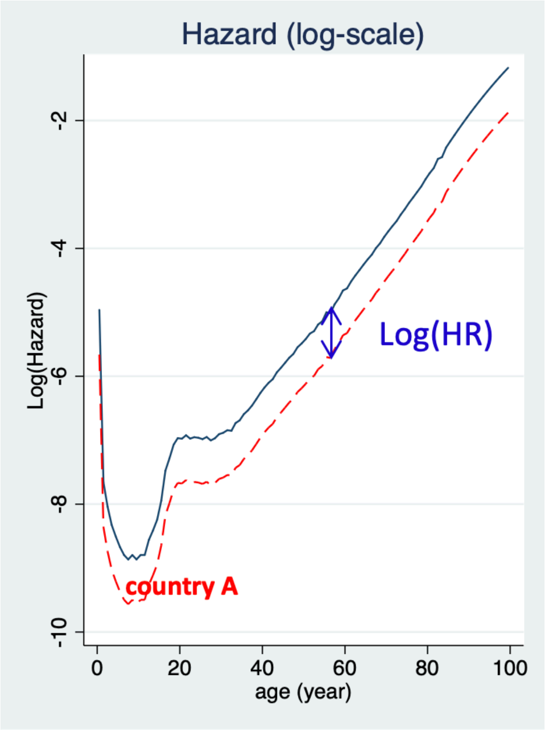
- Instantaneous rate for a patient who survived right before at time t, but died right after t.
- Difficult to estimate well without assumptions.
  - A typical assumption: the proportional hazards (PH) assumption

Data where proportional hazards (PH) assumption **holds**

US National Vital Statistics (2002)



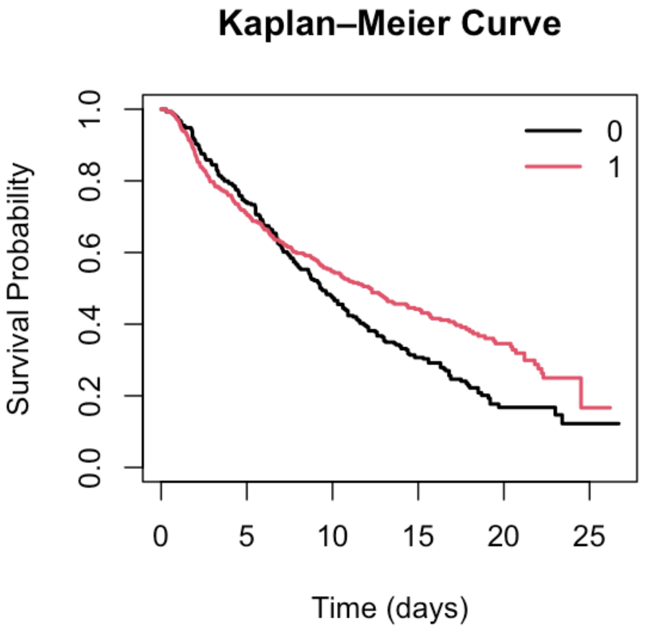
$$S(t) = \exp(-\int_0^t \lambda(u)du)$$



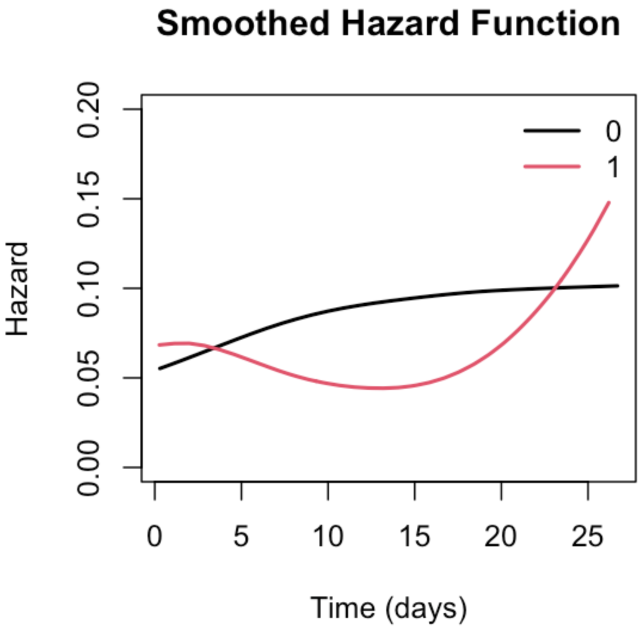
$$\log(\lambda(t))$$

Data where proportional hazards (PH) assumption **does not hold**

Data: CheckMate 057 (Advanced Nonsquamous NSCLC), Borghaei, et al. (2015, NEJM)


Kaplan–Meier Curve

$$S(t) = \exp(-\int_0^t \lambda(u)du)$$


Smoothed Hazard Function

$$\log(\lambda(t))$$

Significant advantages of Sir David Cox’s HR approach

- Easy to use: quantifies difference between two survival curves (i.e., HR).
- Supported by rigorous and elegant theories.*

- Great experience (history) in clinical research community.
- Accessibility (software, instructions)

* including handling time-varying covariates

Limitations of the hazard ratio

1. Traditional hazard ratio is a model-based estimand

- Proportional Hazards Model:

$$\lambda(t \mid Z) = \lambda_0(t)\exp(\beta Z) \rightarrow \log(\mathrm{HR}) = \beta$$

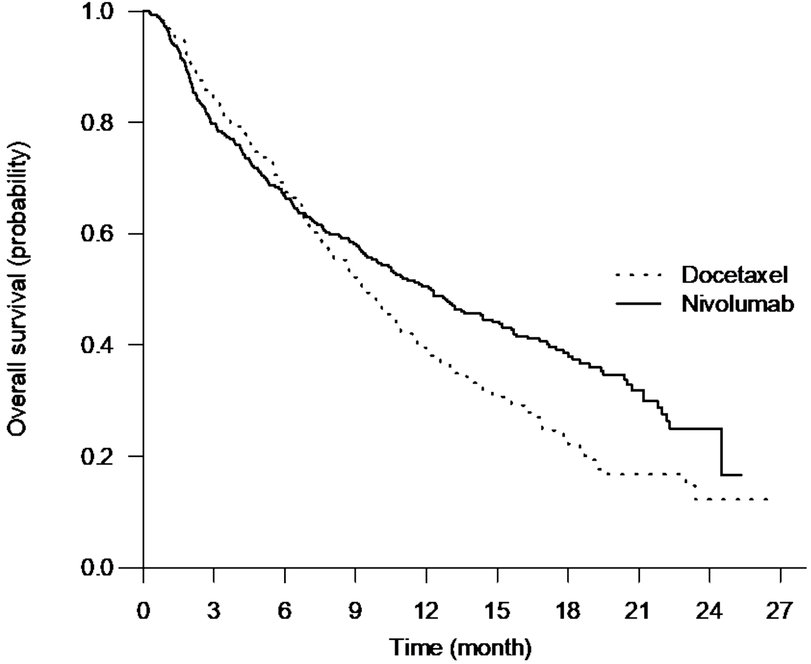
where Z is the treatment indicator (1: Treat, 0: Control).

- When the PH assumption fails, e.g., $\lambda(t \mid Z) = \lambda_0(-tZ)\exp(\beta Z)$
  - the estimand is not well-defined*.
  - with the same KM curve, HR can be different.

* More precisely, the estimator converges to some non-interpretable quantity

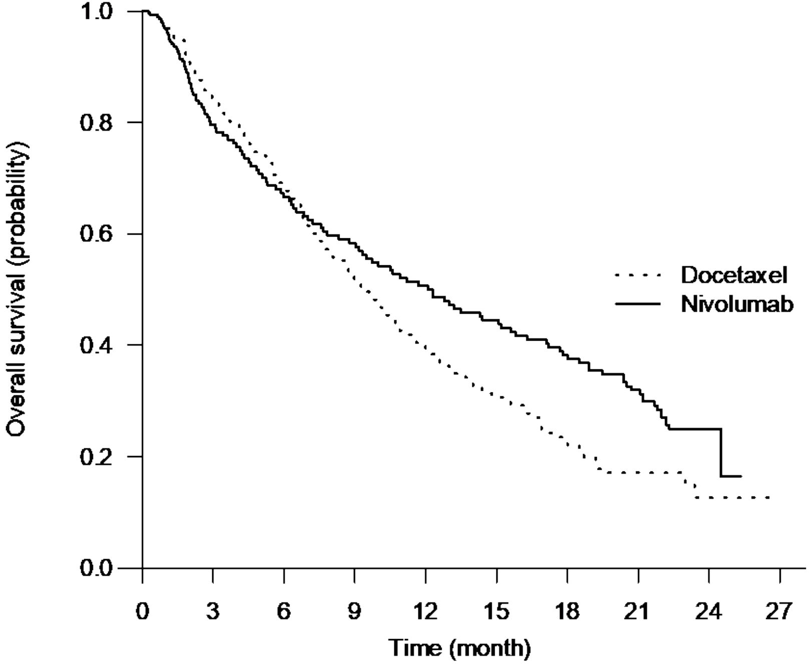
CheckMate 057 data: PH does not hold

Original data from the paper



Nivolumab 292	232	194	169	146	123	64	32	9	0
Docetaxel 290	244	194	150	112	86	36	10	5	0

Same time distribution, recreated censoring

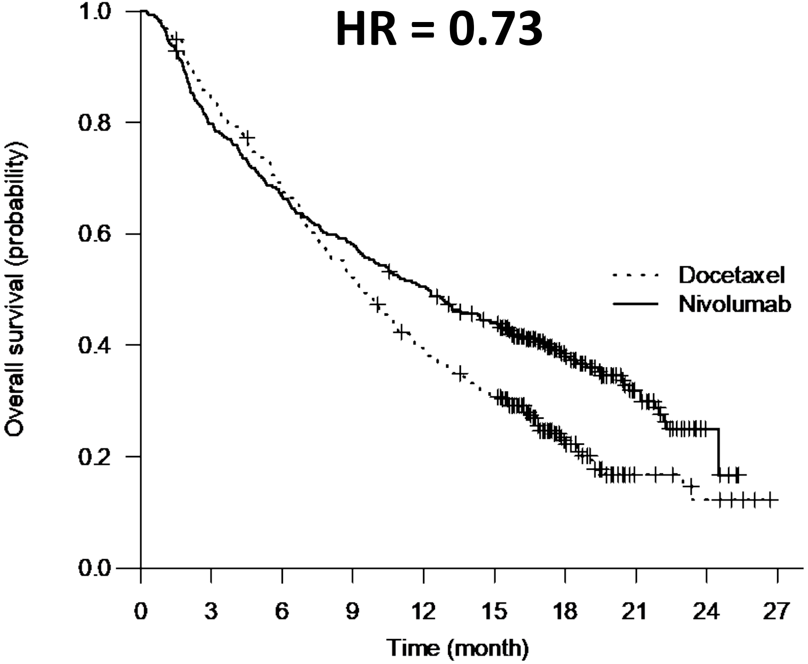


Nivolumab 292	116	97	84	73	64	55	46	35	0
Docetaxel 290	122	97	74	57	43	32	24	17	0

Numbers in the table: number of alive patients by the time

CheckMate 057 data: PH does not hold

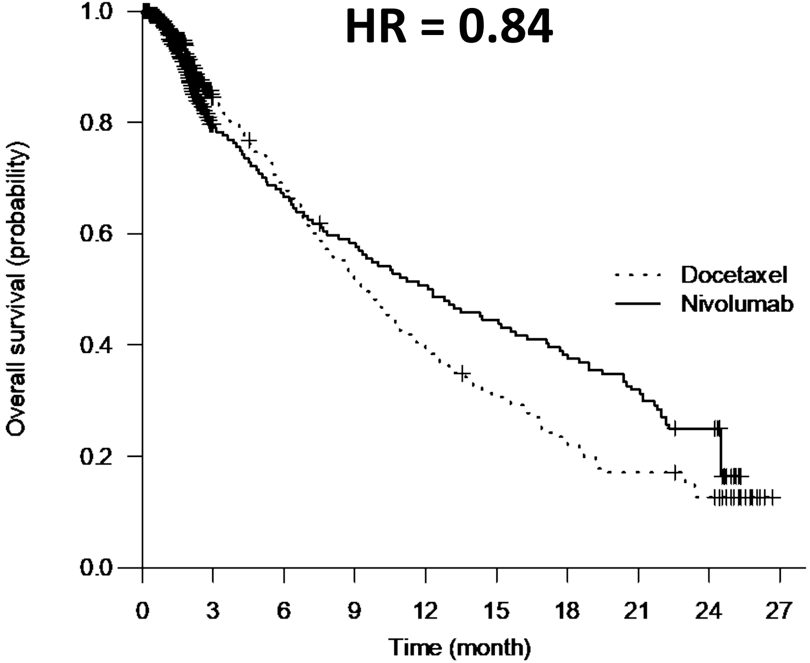
Original data from the paper



HR = 0.73

Nivolumab 292	232	194	169	146	123	64	32	9	0
Docetaxel 290	244	194	150	112	86	36	10	5	0

Same time distribution, recreated censoring



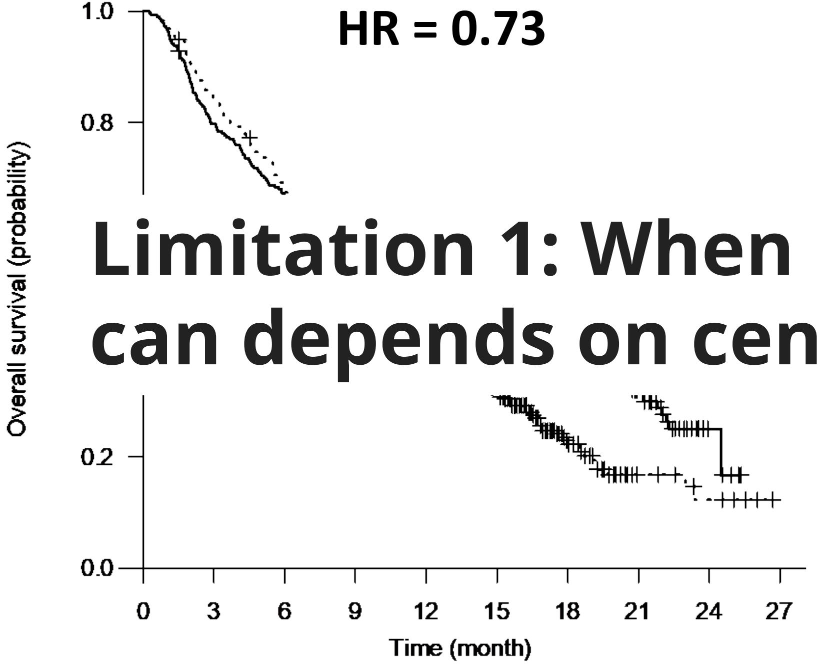
HR = 0.84

Nivolumab 292	116	97	84	73	64	55	46	35	0
Docetaxel 290	122	97	74	57	43	32	24	17	0

Numbers in the table: number of alive patients by the time

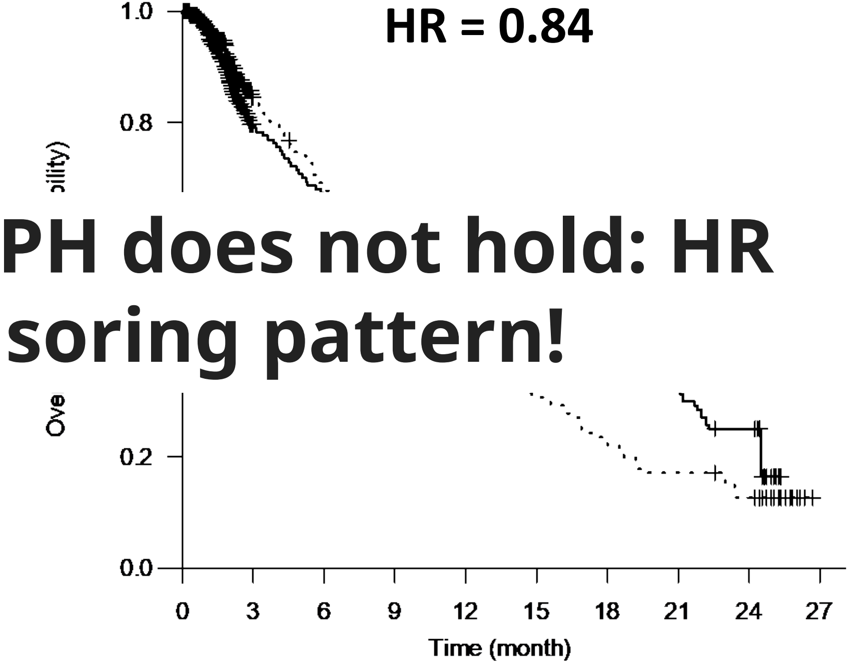
CheckMate 057 data: PH does not hold

Original data from the paper



Nivolumab 292	232	194	169	146	123	64	32	9	0
Docetaxel 290	244	194	150	112	86	36	10	5	0

Same time distribution, recreated censoring



Nivolumab 292	116	97	84	73	64	55	46	35	0
Docetaxel 290	122	97	74	57	43	32	24	17	0

Limitation 1: When PH does not hold: HR can depends on censoring pattern!

Numbers in the table: number of alive patients by the time

Limitation 2: HR alone cannot address a practically important recommendation

(Draw) A numerical example…

N=10,000 in New treatment group

N=10,000 in Placebo group

Followed everybody for 10 years

Study 1: Observed ~1 adverse event around 5 years in each group

Study 2: Observed ~4000 adverse event around 5 years in each group

Limitation 2: HR alone cannot address a practically important recommendation

*Does **HR = 0.8** indicate the new treatment gives a clinically meaningful large effect?*

	Overall Survival (Study 1)	**Overall Survival (Study 2)**
Event Rate	**1/10000**	**4000/1000**
Hazard Ratio		
Hazard in Treatment group		
Hazard in Control group		

Limitation 2: HR alone cannot address a practically important recommendation

*Does **HR = 0.8** indicate the new treatment gives a clinically meaningful large effect?*

	Overall Survival (Study 1)	**Overall Survival (Study 2)**
Event Rate	**1/10000**	**4000/1000**
Hazard Ratio	**0.8**	**0.8**
Hazard in Treatment group		
Hazard in Control group		

Limitation 2: HR alone cannot address a practically important recommendation

*Does **HR = 0.8** indicate the new treatment gives a clinically meaningful large effect?*

	Overall Survival (Study 1)	**Overall Survival (Study 2)**
Event Rate	**1/10000**	**4000/1000**
Hazard Ratio	**0.8**	**0.8**
Hazard in Treatment group	0.00001	0.0511
Hazard in Control group		

Limitation 2: HR alone cannot address a practically important recommendation

*Does **HR = 0.8** indicate the new treatment gives a clinically meaningful large effect?*

	Overall Survival (Study 1)	**Overall Survival (Study 2)**
Event Rate	**1/10000**	**4000/1000**
Hazard Ratio	**0.8**	**0.8**
Hazard in Treatment group	0.00001	0.0511
Hazard in Control group	0.000008	0.0409

Baseline matters

Alaska temperature

10% increase: 32°F -> 35.2°F

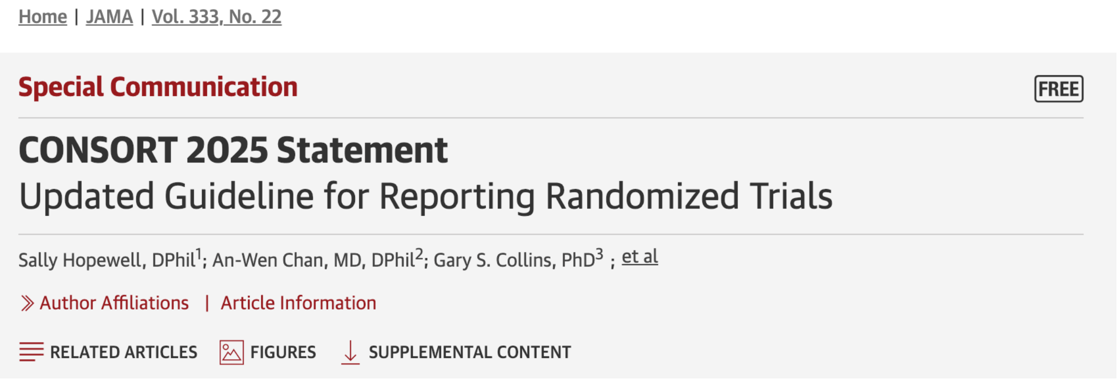
No clear difference: still cold

Florida temperature

10% increase: 75°F -> 82.5°F

More comfortable

Several resources have acknowledged the importance of the baseline hazard



Home | JAMA | Vol. 333, No. 22

Special Communication

FREE

CONSORT 2025 Statement

Updated Guideline for Reporting Randomized Trials

Sally Hopewell, DPhil[1]; An-Wen Chan, MD, DPhil[2]; Gary S. Collins, PhD[3] ; et al

» Author Affiliations | Article Information

RELATED ARTICLES FIGURES SUPPLEMENTAL CONTENT

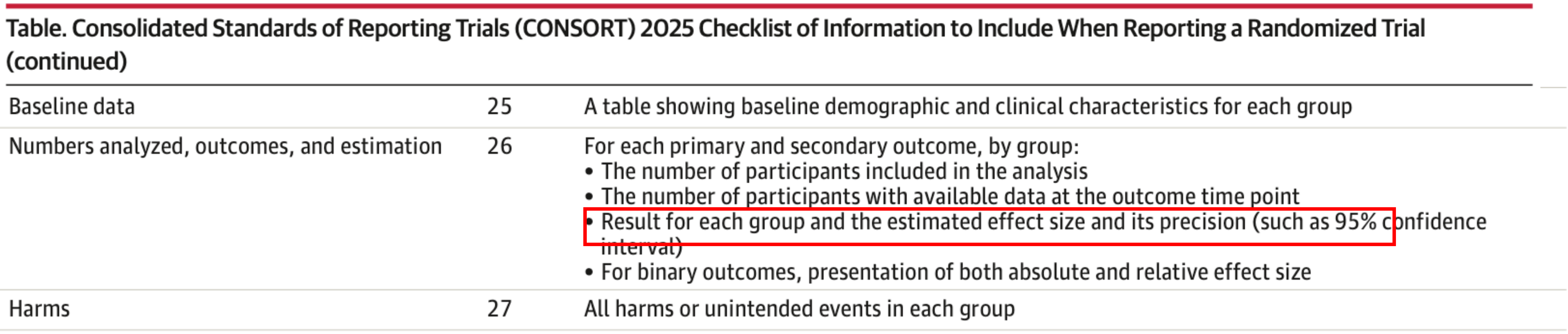


Table. Consolidated Standards of Reporting Trials (CONSORT) 2025 Checklist of Information to Include When Reporting a Randomized Trial (continued)

Baseline data	25	A table showing baseline demographic and clinical characteristics for each group
Numbers analyzed, outcomes, and estimation	26	For each primary and secondary outcome, by group: • The number of participants included in the analysis • The number of participants with available data at the outcome time point • Result for each group and the estimated effect size and its precision (such as 95% confidence interval) • For binary outcomes, presentation of both absolute and relative effect size
Harms	27	All harms or unintended events in each group

Annals of Internal Medicine

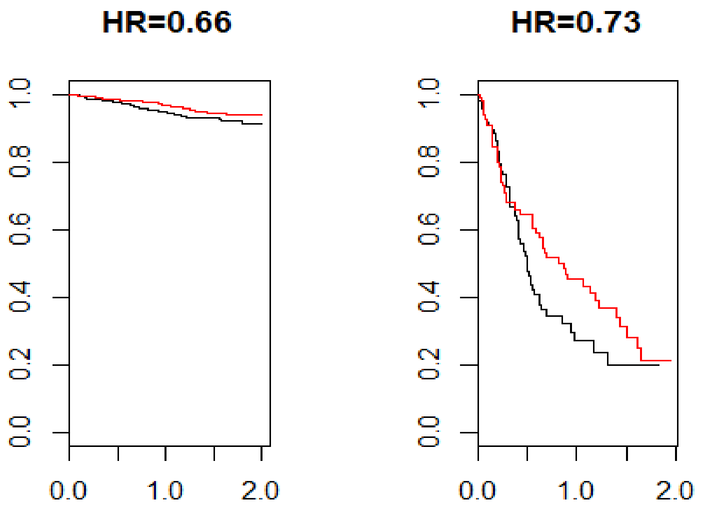
Information for Authors: General Statistical Guidance

https://annals.org/aim/pages/author-information-statistics-only

Hazard Ratios and Standardized Cumulative Incidence

Authors often report results from analysis of survival or time-to-event data using hazard ratios estimated from proportional hazards Cox models. **Hazard ratios are notoriously difficult to interpret clinically, may be sensitive to the length of follow-up, and rely on model assumptions, such as proportional hazards. In addition, presenting estimates of effect in both absolute and relative terms increases the likelihood that results will be correctly interpreted.** For all of these reasons, we recommend that authors present cumulative incidence curves (inverted Kaplan-Meier plots) along with tabular summaries of absolute differences in cumulative incidence, with 95% confidence bounds, at meaningful times, when reporting results from survival analyses. When such an analysis requires covariate adjustment, authors can estimate and present covariate-standardized (weighted) cumulative incidence curves with differences in adjusted cumulative incidence at meaningful times.

Limitation 3: The HR may not align well with the graphic presentation of survival curves



Result of baseline difference + PH assumption violation

Limitation 4. Wide CI when the event rate is low

For a safety study: When the number of events is small, the hazard ratio estimate is very unstable and the confidence interval is very wide, implying that there is not enough information to make a decision

… even if the PH assumption is correct

Going back to our numerical example…

N=10,000 in New treatment group

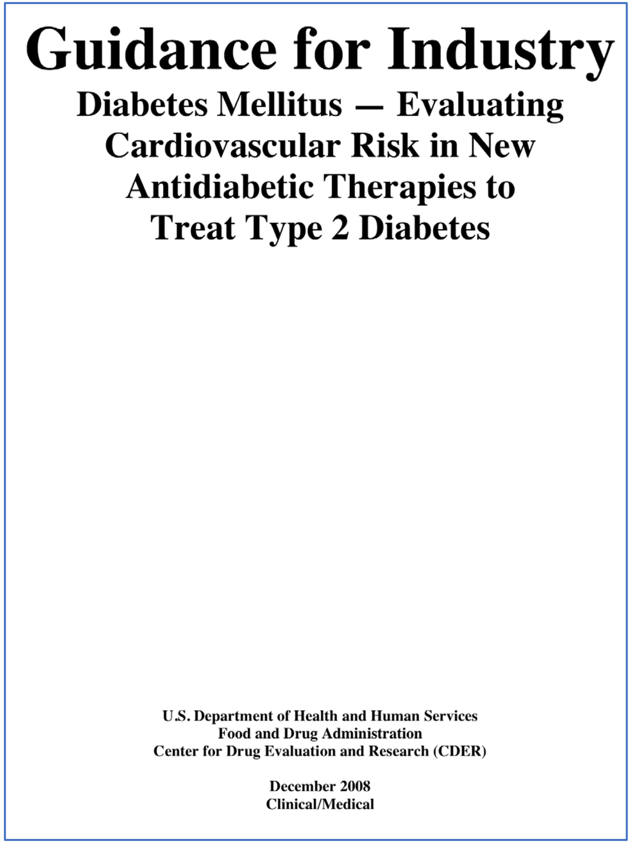
N=10,000 in Placebo group

Followed everybody for 10 years

Study 1: Observed only 1 adverse event around 5 years in each group

95% Confidence Interval of HR

(0.1 to 16)



Guidance for Industry

Diabetes Mellitus — Evaluating Cardiovascular Risk in New Antidiabetic Therapies to Treat Type 2 Diabetes

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)

December 2008
Clinical/Medical

1. If the modeling assumption does not hold *(usually it does not hold in practice)*, the HR estimate depends on study-specific censoring distribution.
2. Baseline matters.
3. Censoring pattern matters: Precision of HR depends on the # of events, not exposure times.

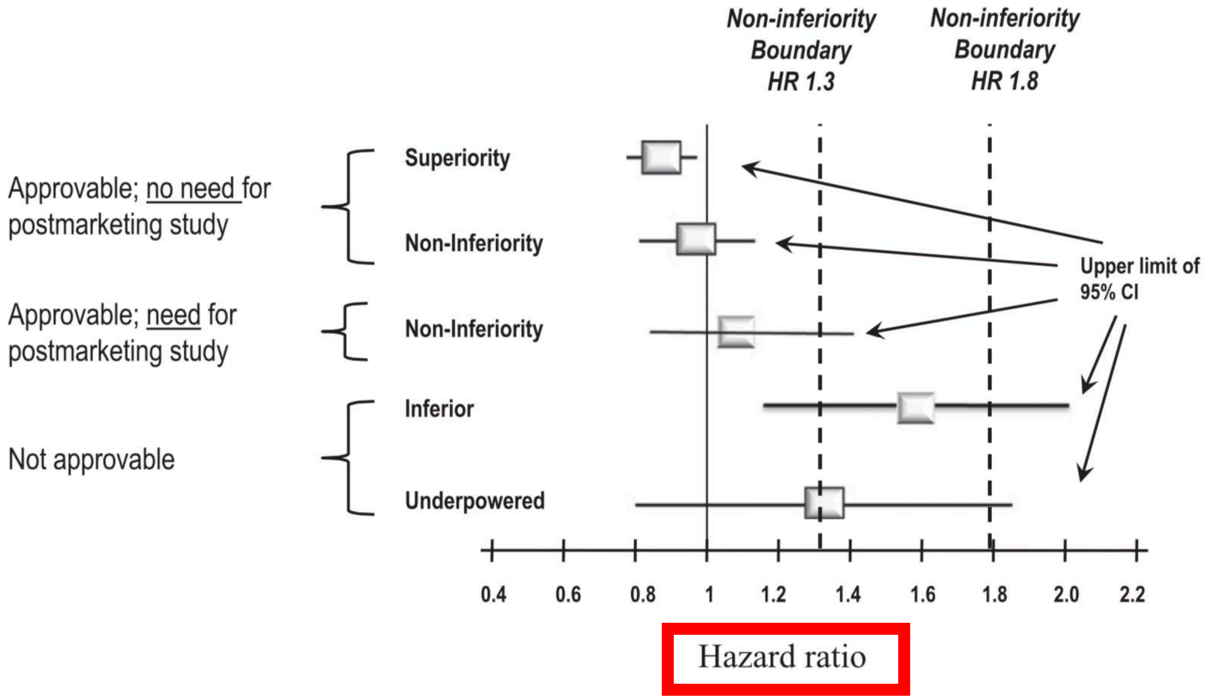
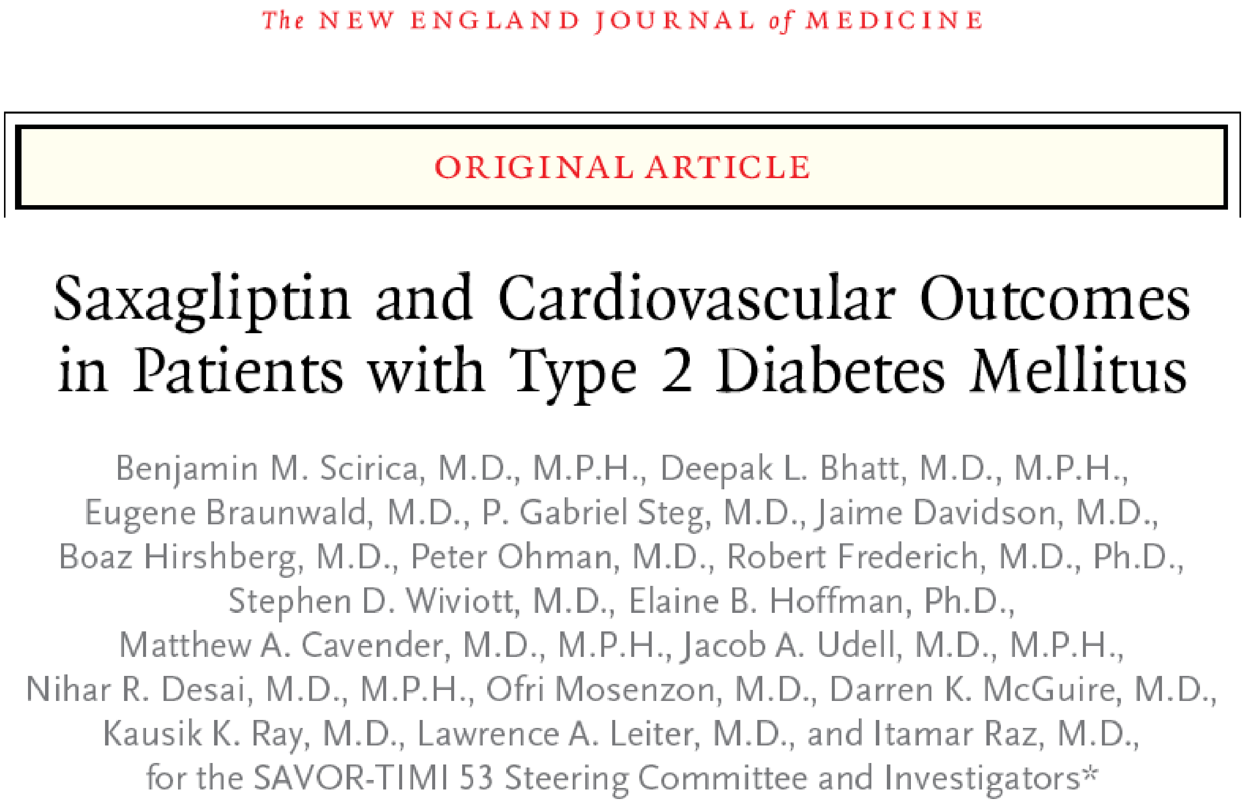


Figure 1—*FDA CV safety: CI bars. The FDA guidelines provide statistical hurdles for approval. Five hypothetical examples of possible hazard ratios and the upper limit of the 95% CI of a development plan are shown as well as the regulatory consequences of each outcome.*

Hirshberg & Raz (2011, Diabetes Care)

Example: SAVOR-TIMI 53 trial



The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Saxagliptin and Cardiovascular Outcomes in Patients with Type 2 Diabetes Mellitus

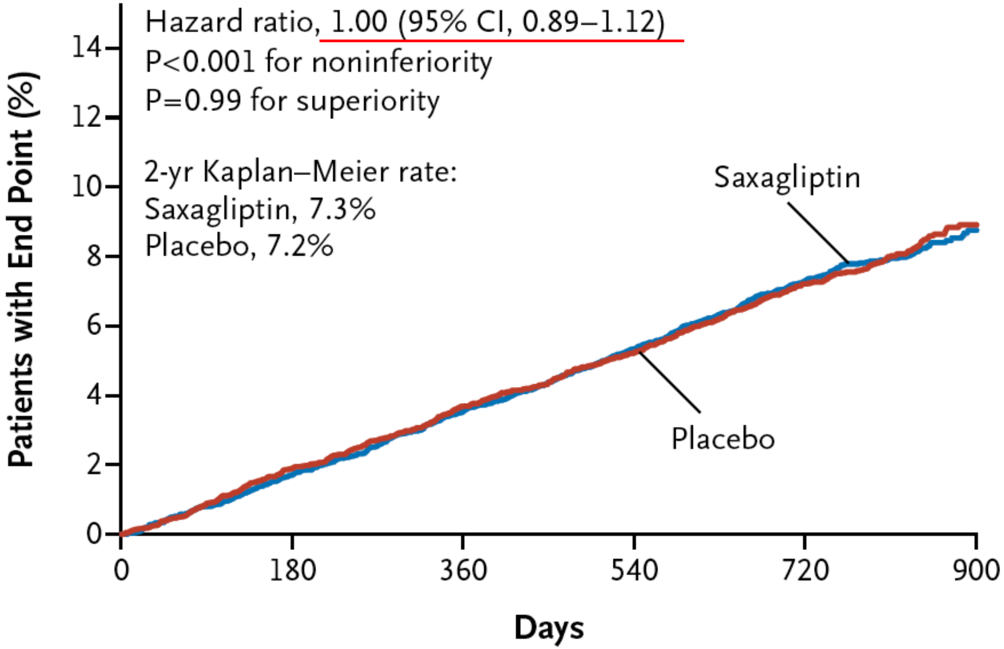
Benjamin M. Scirica, M.D., M.P.H., Deepak L. Bhatt, M.D., M.P.H.,
Eugene Braunwald, M.D., P. Gabriel Steg, M.D., Jaime Davidson, M.D.,
Boaz Hirshberg, M.D., Peter Ohman, M.D., Robert Frederich, M.D., Ph.D.,
Stephen D. Wiviott, M.D., Elaine B. Hoffman, Ph.D.,
Matthew A. Cavender, M.D., M.P.H., Jacob A. Udell, M.D., M.P.H.,
Nihar R. Desai, M.D., M.P.H., Ofri Mosenzon, M.D., Darren K. McGuire, M.D.,
Kausik K. Ray, M.D., Lawrence A. Leiter, M.D., and Itamar Raz, M.D.,
for the SAVOR-TIMI 53 Steering Committee and Investigators*

NEJM, October 3, 2013

SAVOR-TIMI 53 (saxagliptin vs. placebo)

- Primary endpoint: CV death, nonfatal MI, or nonfatal ischemic stroke
- Primary analysis: time to event
- 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)

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:
Saxagliptin, 7.3%
Placebo, 7.2%

No. at Risk

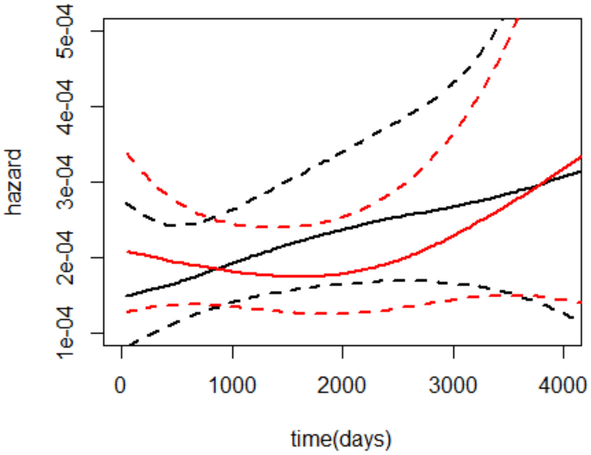
Placebo	8212	7983	7761	7267	4855	851
Saxagliptin	8280	8071	7836	7313	4920	847

Ref: A way to estimate the hazard function

- The lack of reference hazard function hinders the effective interpretation of an estimated HR.
- The hazard function is difficult to estimate non-parametrically

$$\int_0^\tau \frac{1}{h} K\left(\frac{t - t_0}{h}\right) d\widehat{\Lambda}(t),$$

where $\widehat{\Lambda}(t)$ is the NA estimator of the cumulative hazards function



Ref: R code for estimating the hazard function

```
library(bshazard)
fit1=bshazard(Surv(time, status==2)~1, data=pbc[pbc$trt==1,])
fit2=bshazard(Surv(time, status==2)~1, data=pbc[pbc$trt==2,])
plot(c(0, 4000), c(0.0001, 0.0005), xlab="time(days)", ylab="hazard", type="n")
lines(fit1, col=1, lwd=2)
lines(fit2, col=2, lwd=2)
```

Ref: Check proportional hazards assumptions

- In the two-group comparisons, one may plot

$$t \ vs \ \log\left(-\log\left(S_j(t)\right)\right), j = 0, 1$$

to examine if the two curves are approximately parallel.

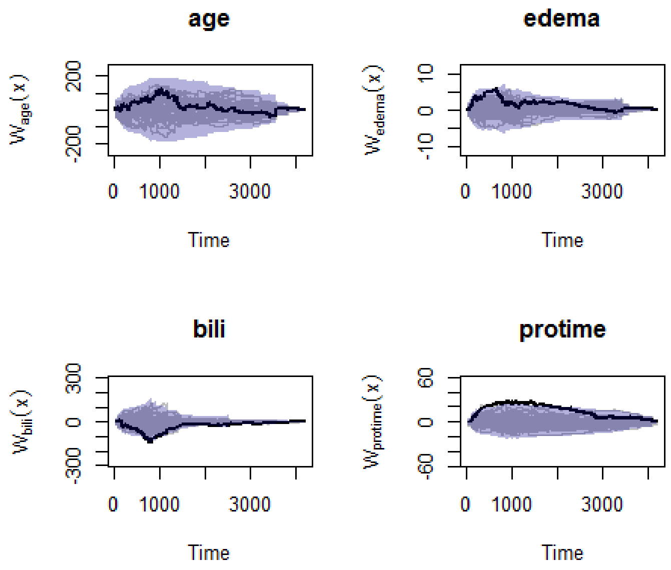
- In the regression setting, one may use the cumulative martingale-based residuals

$$\sum_{i=1}^{n} \int_0^t Z_i d\widehat{M}_i(u),$$

which are supposed to be approximated by a mean zero Gaussian process, where

$$\widehat{M}_i(t) = I(Y_i \le t)\delta_i - \widehat{\Lambda}_0(\min(t, Y_i))e^{\widehat{\beta}' Z_i}$$

Potential violation of the PH assumption



Ref: R code for checking the PH assumption

```
library(survival)
library(gof)
data(pbc)
fit.cox <- coxph(Surv(time,status==2) ~ age + edema + bili + protime, data=pbc)
system.time(pbc.gof <- cumres(fit.cox,R=2000))
par(mfrow=c(2,2))
plot(pbc.gof, ci=TRUE, legend=NULL)
```

Ref: The problems of goodness of fit tests

- All the statistical models including the Cox regression are merely approximation to the truth.
  - If the sample size is sufficiently large, one would always reject the PH assumption
  - If the sample size is small, one may not be able to reject the PH assumption even if the violation is nontrivial.
- PH assumptions with different sets of covariates are not compatible:
  - $\lambda(t|Z) = \lambda_0(t)\exp(\beta Z)$
  - $\lambda(t|Z, X) = \lambda_0(t)\exp(\beta Z + \gamma' X)$

  normally can not hold at the same time
  - This cast doubts on the normal practice in analyzing clinical trial data: first estimate the HR using treatment indicator only and then estimate the adjusted HR using a multivariate Cox regression.