Issues and concerns of some conventional analytic methods
--- Moving beyond the comfort zone ---

Hajime Uno, PhD
Department of Biostatistics and Computational Biology
Department of Medical Oncology
Dana-Farber Cancer Institute / Harvard Medical School

DANA-FARBER CANCER INSTITUTE

March 1, 2017 (Tokyo)

HARVARD MEDICAL SCHOOL

Acknowledgements

- LJ Wei (Harvard)
- Lu Tian (Stanford)
- Brian Claggett (Harvard/Brigham)
- Takahiro Hasegawa (Shionogi)


Prof. LJ Wei

Statistical analyses for clinical studies

We tend to follow a pattern….

- Take the previously used approach
- Use methods frequently used in medical literature

Why?

- Hardly getting criticized?
- Avoiding delay of review processes?
- No need to explain the methods?

Some methods have become almost routine!!

What is the primary goal of clinical studies?

Obtaining robust, clinically interpretable risk-benefit information for patients in a well-defined target population

Contents

- Issues of conventional analytic methods in survival data analysis
  - Issues of hazard ratio
  - Alternatives to hazard ratio
- Issues of some other routinely used methods…
- Conclusions

Issues of conventional analytic methods in survival data analysis

A standard practice…

Description

Test

Estimation of treatment effect

Kaplan-Meier

Log-rank test

Estimate HR by Cox reg

Why are we almost routinely using HR?

Several desirable properties

A lot of experiences in practice

Elegant, rigorous theories

Everyone knows

No other choice

It is already in our comfort zone?

What is hazard function?

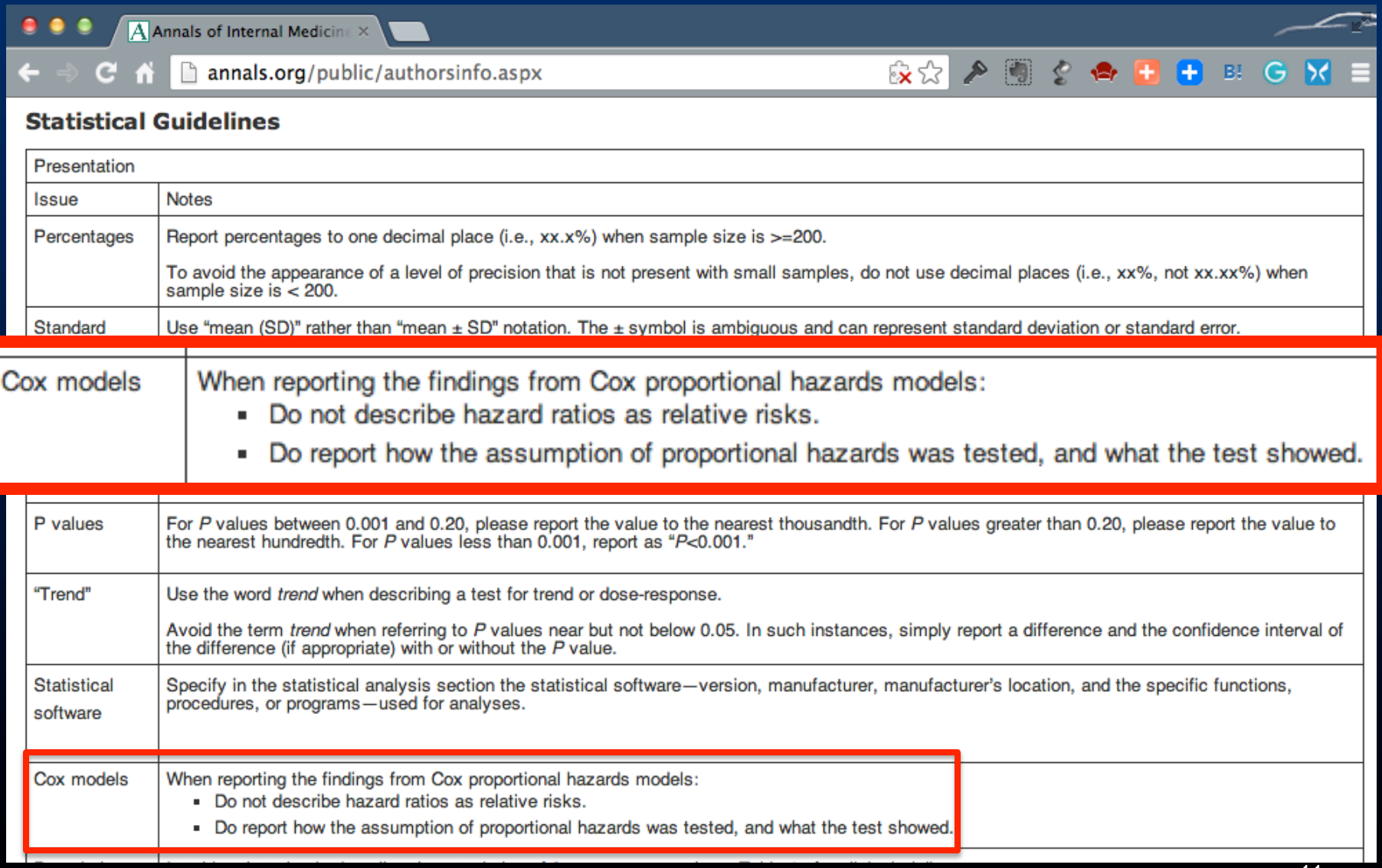
$$h(t)\Delta t \approx \Pr[t < T \leq t + \Delta t \mid T \geq t]$$

Closely related to *conditional probability*, which may be more interesting metric for some subjects (e.g., cancer survivors)

Note the hazard function is ***NOT*** in the probability scale

Ref. *Annals of Internal Medicine*, Guideline for Authors

http://annals.org/public/authorsinfo.aspx

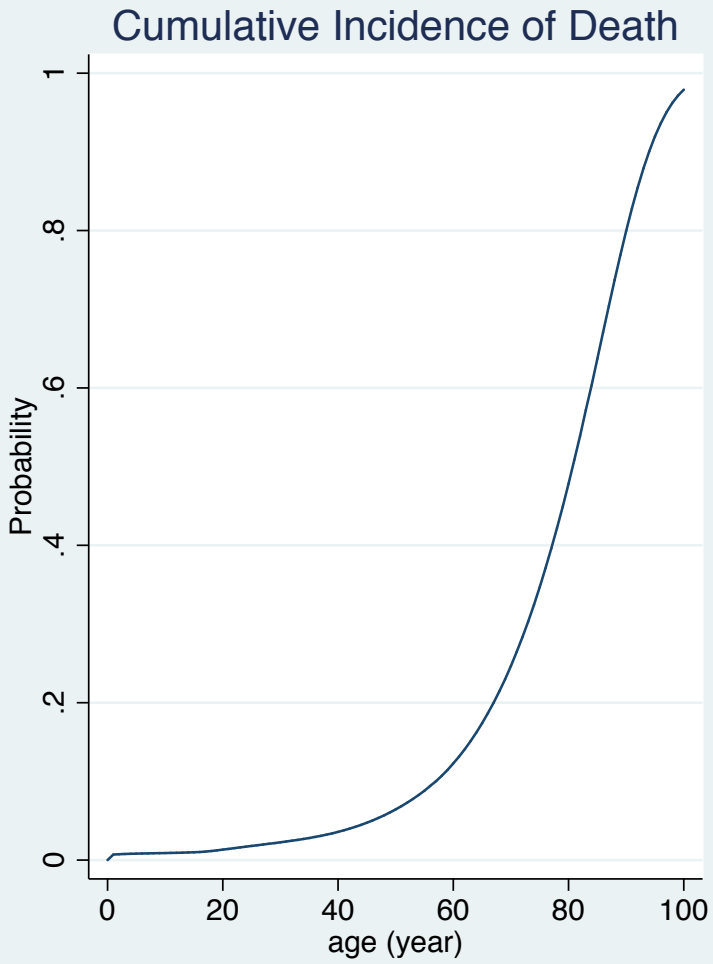


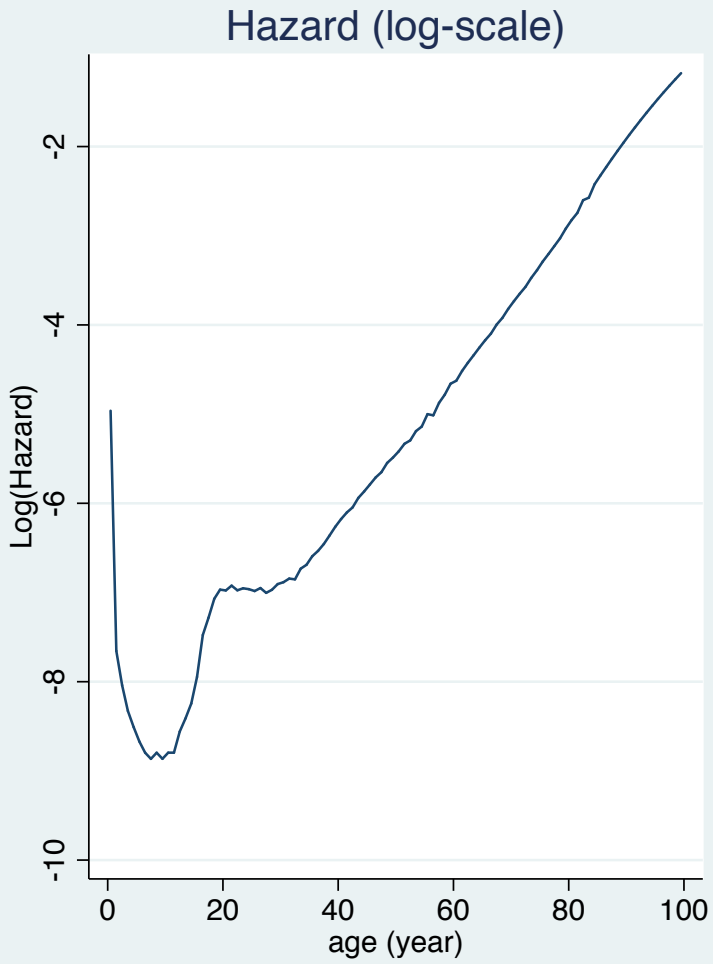
Statistical Guidelines

Presentation	
Issue	Notes
Percentages	Report percentages to one decimal place (i.e., xx.x%) when sample size is >=200. To avoid the appearance of a level of precision that is not present with small samples, do not use decimal places (i.e., xx%, not xx.xx%) when sample size is < 200.
Standard	Use "mean (SD)" rather than "mean ± SD" notation. The ± symbol is ambiguous and can represent standard deviation or standard error.
Cox models	When reporting the findings from Cox proportional hazards models: ▪ Do not describe hazard ratios as relative risks. ▪ Do report how the assumption of proportional hazards was tested, and what the test showed.
P values	For *P* values between 0.001 and 0.20, please report the value to the nearest thousandth. For *P* values greater than 0.20, please report the value to the nearest hundredth. For *P* values less than 0.001, report as "*P*<0.001."
"Trend"	Use the word *trend* when describing a test for trend or dose-response. Avoid the term *trend* when referring to *P* values near but not below 0.05. In such instances, simply report a difference and the confidence interval of the difference (if appropriate) with or without the *P* value.
Statistical software	Specify in the statistical analysis section the statistical software—version, manufacturer, manufacturer's location, and the specific functions, procedures, or programs—used for analyses.
Cox models	When reporting the findings from Cox proportional hazards models: ▪ Do not describe hazard ratios as relative risks. ▪ Do report how the assumption of proportional hazards was tested, and what the test showed.

What is hazard function?

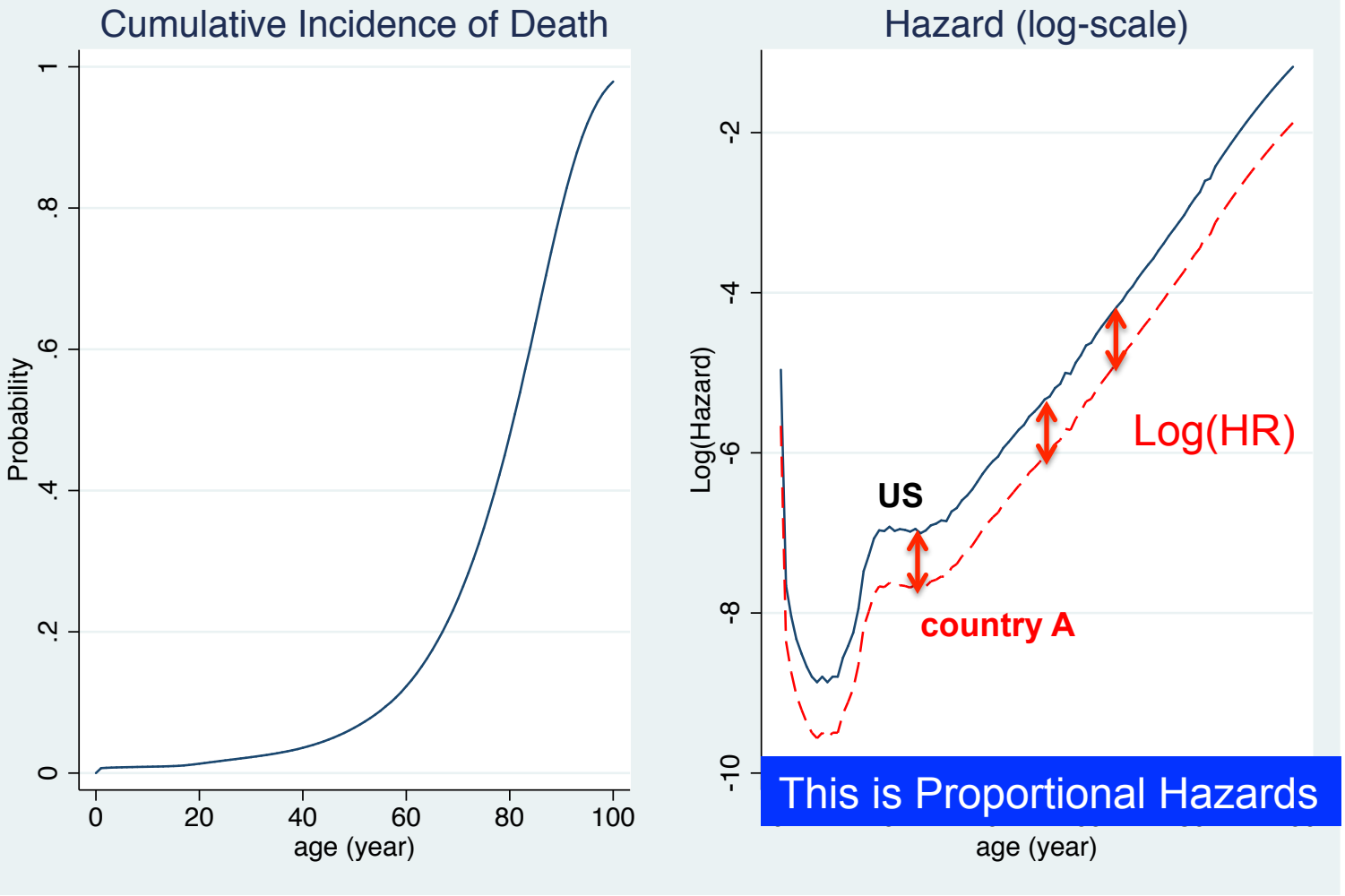
US National Vital Statistics (2002)





What is Hazard Ratio (HR)?

US National Vital Statistics (2002)



Proportional hazard (PH) assumption

- Ratio of two hazard functions (i.e.,hazard ratio) is constant overtime
  (Log-hazard functions are parallel)

- Basis of the valid inference of hazard ratio

Issues of HR

Issues and concerns about hazard ratio estimate (1)

… if the PH assumption is violated

- HR is NOT a simple average of the hazard ratio over time
  - HR depends on underlying study-specific censoring distributions (or follow-up time…)

Ex.

- If follow-up time are different, HR will be different
- If the rates of dropout is different, HR will be different

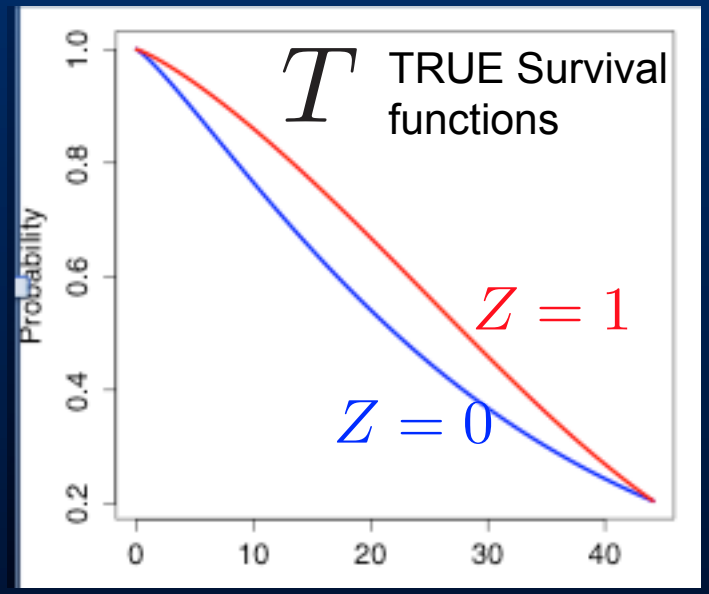
What are we estimating?

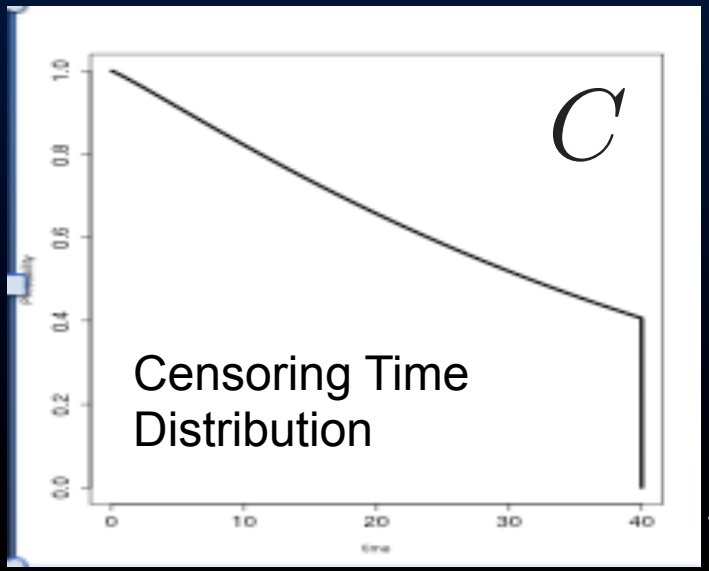
A numerical study for illustration

- Consider two groups and their true survival functions
- Consider a censoring time distribution
- Generate 10 millions of (T, C, Z)
- Then, observable data (X, Δ, Z)

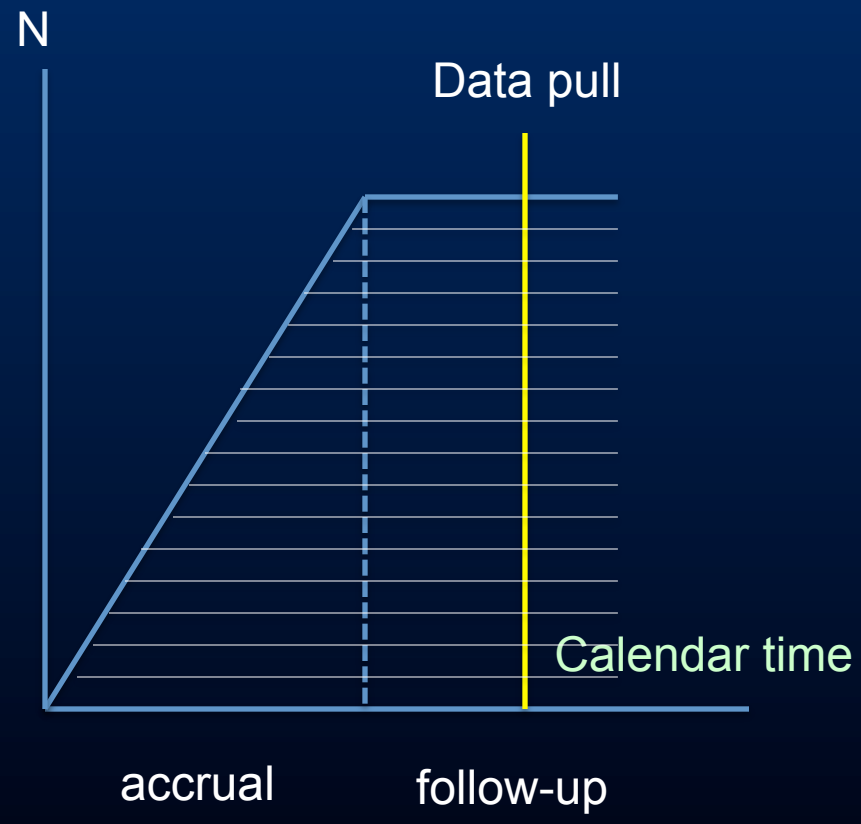
$$X = \min(T, C)$$
$$\Delta = 1 \text{ if } T \leq C$$
$$0 \text{ otherwise}$$

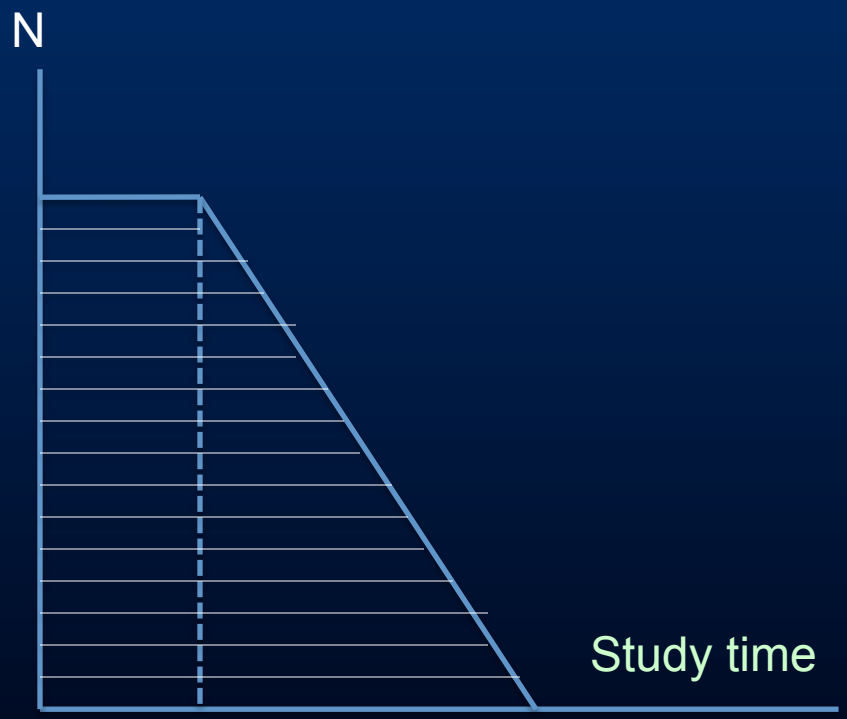
- Calculate HR with the observable data

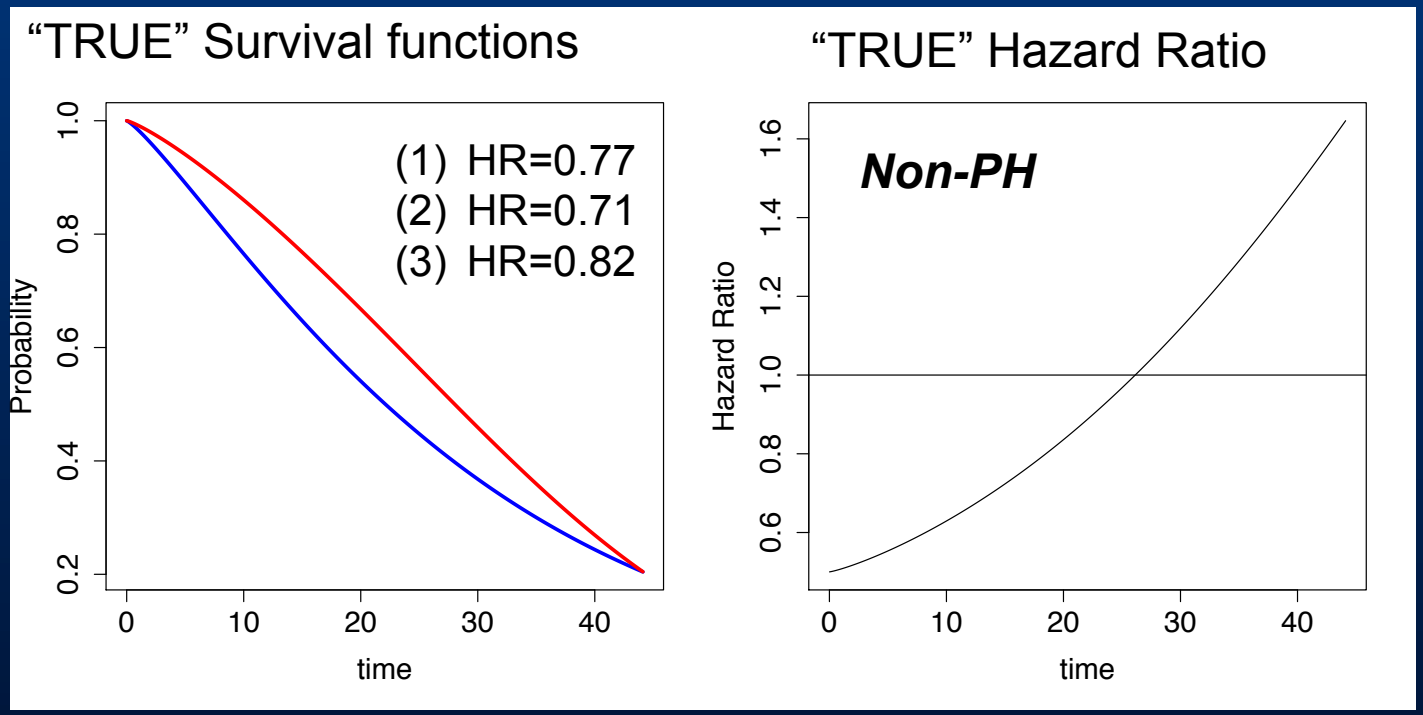




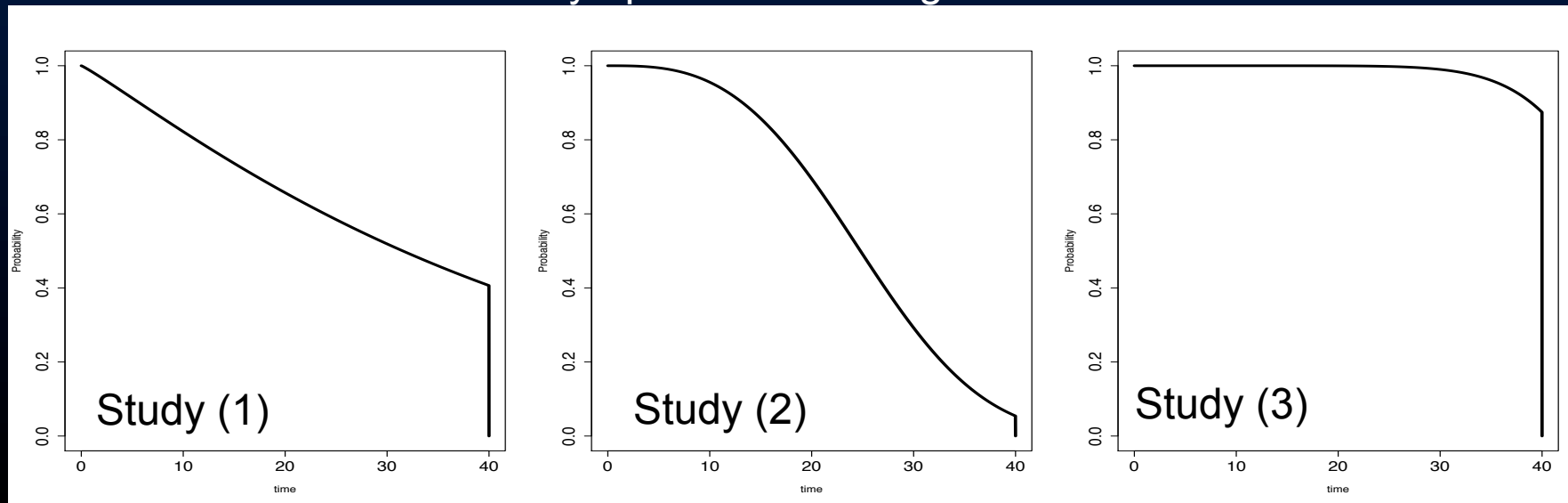
Typical censoring pattern in event-driven trials



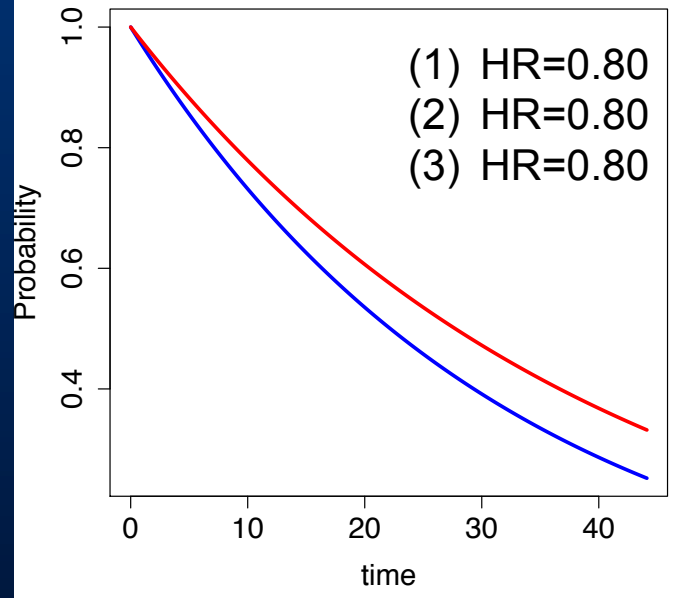




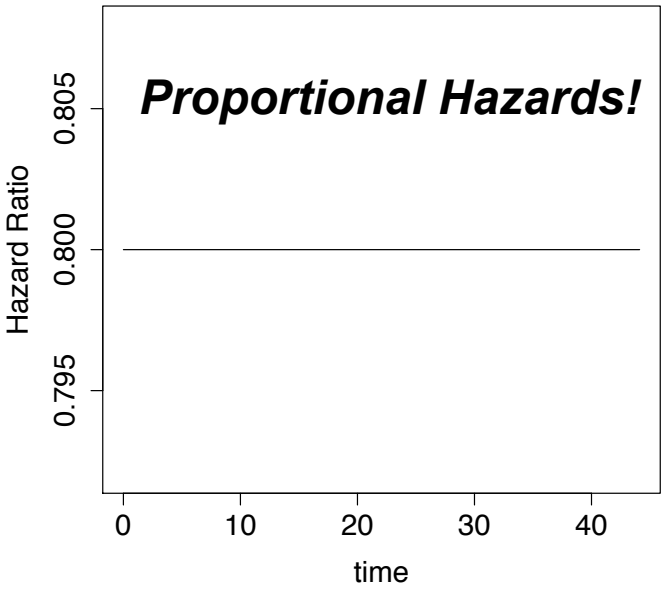
Study-specific censoring distribution



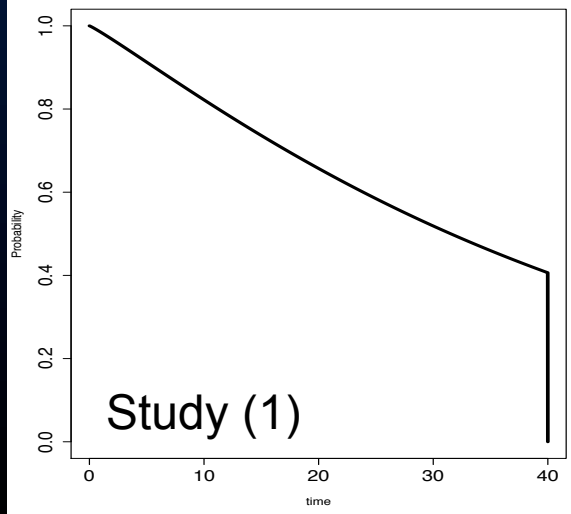
"TRUE" Survival functions

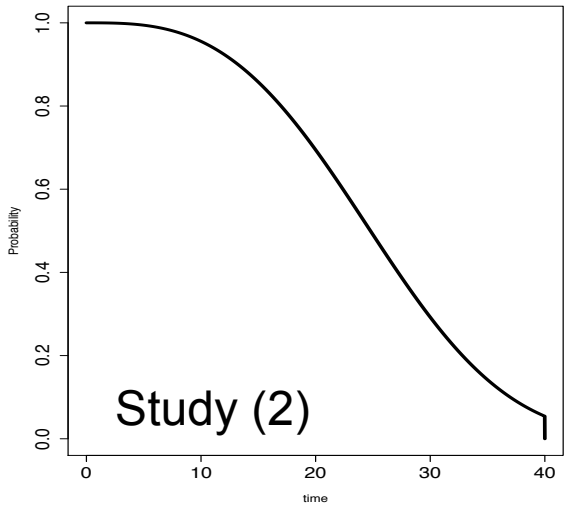


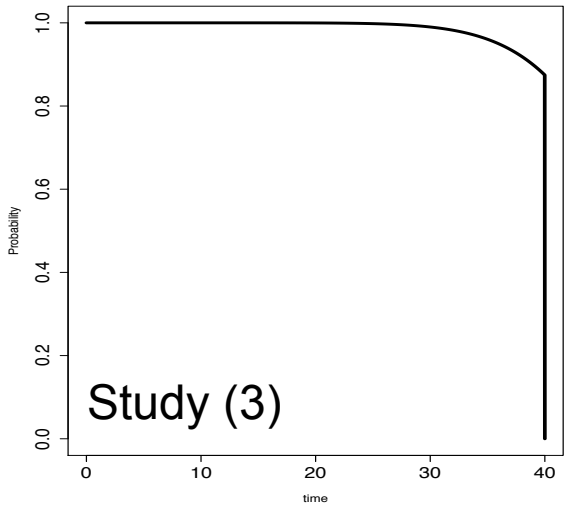
"TRUE" Hazard Ratio



Study-specific censoring distribution







ECOG myeloma study

Lenalidomide plus high-dose dexamethasone versus lenalidomide plus low-dose dexamethasone as initial therapy for newly diagnosed multiple myeloma: an open-label randomised controlled trial

S Vincent Rajkumar, Susanna Jacobus, Natalie S Callander, Rafael Fonseca, David H Vesole, Michael E Williams, Rafat Abonour, David S Siegel, Michael Katz, Philip R Greipp, for the Eastern Cooperative Oncology Group

Summary

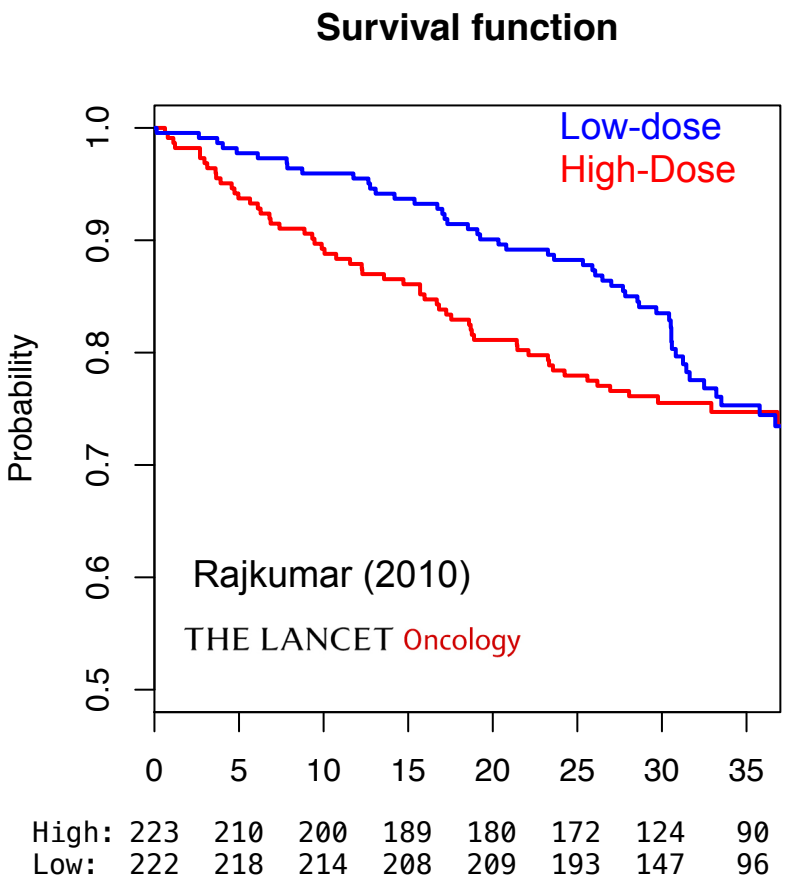
Background High-dose dexamethasone is a mainstay of therapy for multiple myeloma. We studied whether low-dose dexamethasone in combination with lenalidomide is non-inferior to and has lower toxicity than high-dose dexamethasone plus lenalidomide.

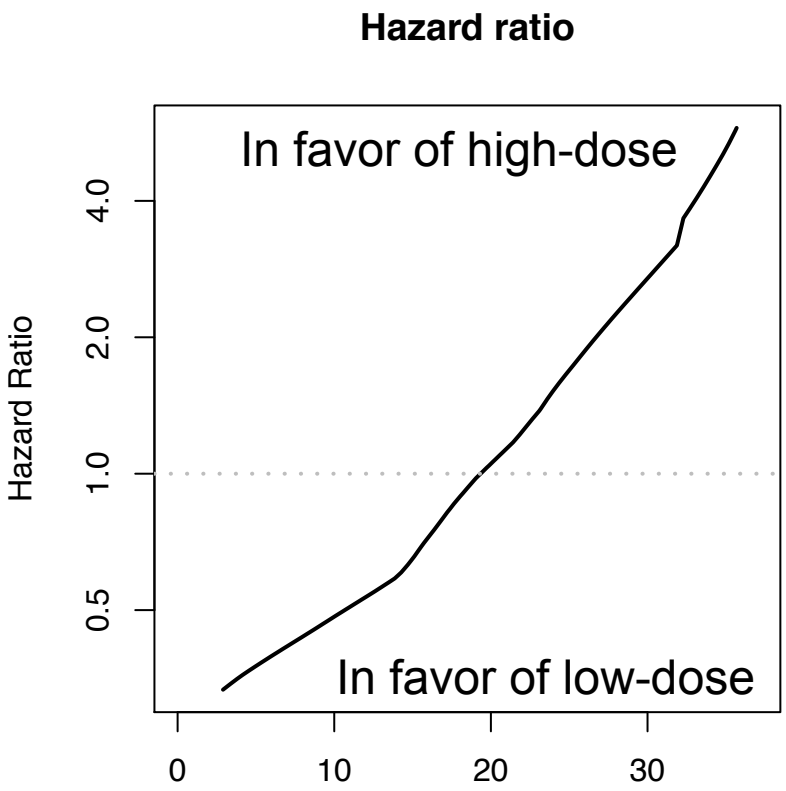
Rajkumar et al. (2010, Lancet Oncology)

ECOG myeloma study (low- vs. high dose)

- A phase III randomized trial to compare low- and high-dose dexamethasone for newly diagnosed multiple myeloma
- N=445 (223 on high-dose, 222 on low-dose)
- One of the endpoints was overall survival

MM Example (ECOG E4A03 OS, low Dex vs. High Dex)





HR= 0.87 (0.95CI: 0.60 to 1.27), p=0.46

How do we interpret 0.87 ??

If the PH assumption is correct, then is the HR ok?

Issues and concerns about hazard ratio estimate [2]

… even if the PH assumption is correct

No reference number

A HR is difficult to interpret clinically without any absolute hazard to serve as reference

Numerical examples:

- 3-year event rate

  50% (Cont) → 40% (Treat) (the ratio is 0.8)

  This is 20% risk reduction from 50%

  1% (Cont) → 0.8% (Treat) (the ratio is 0.8)

  This is 20% risk reduction from 1%

- Median survival time

  10 months (Cont) → 12 months (Treat)

  This is 20% improvement from 10 months

- Now, Hazard Ratio = 0.8?

  This is 20% hazard reduction. But ***from what?***

Issues and concerns about hazard ratio estimate [3]

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

A numerical example…

N=10,000 in New treatment group
N=10,000 in Placebo group

Followed everybody for 10 years

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

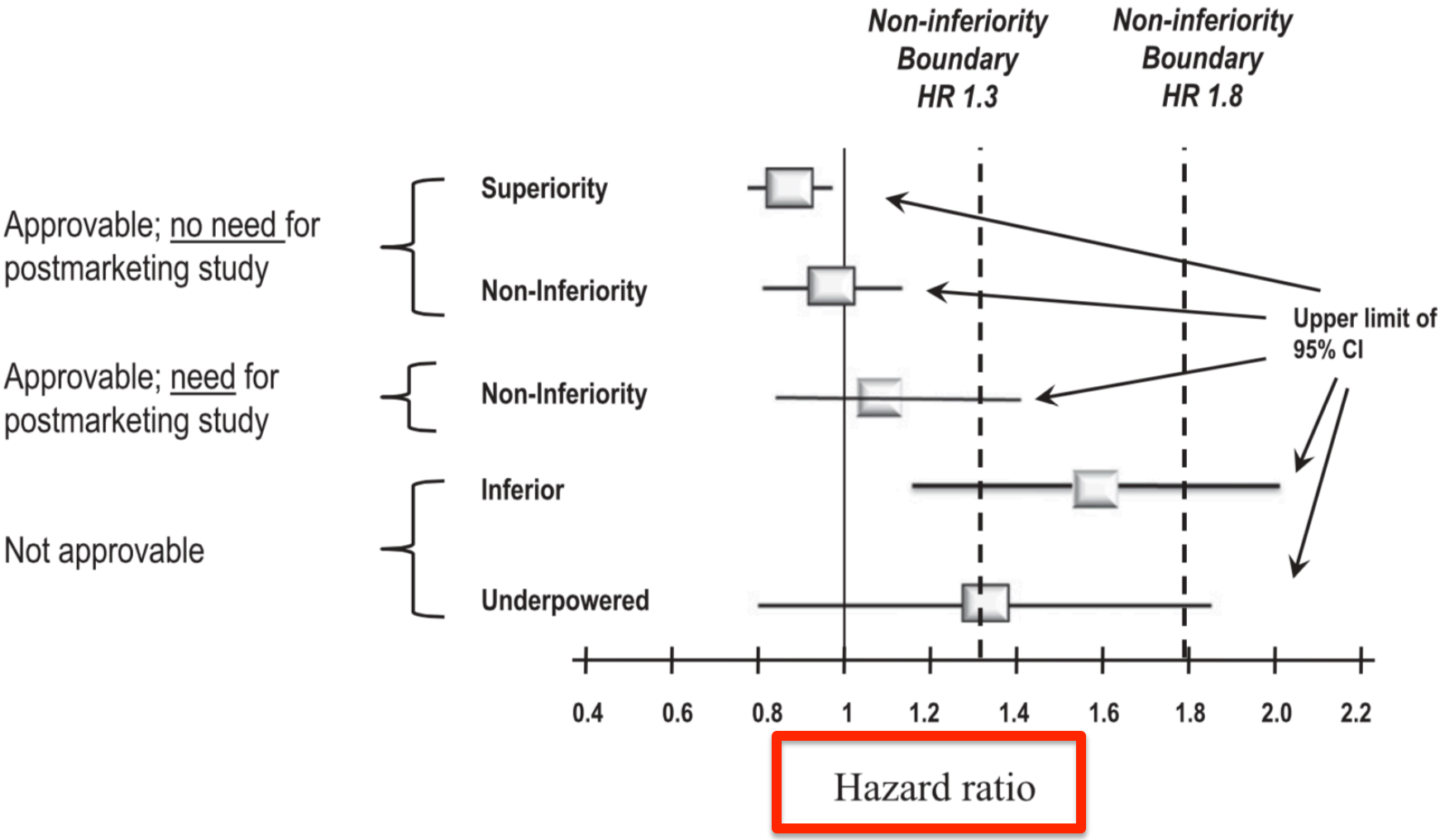


Figure 1—*FDA CV safety: CI bars. The FDA guidelines provide statistical hurdles for approval. Five hypothetical examples of possible hazard ratios and the upper limit of the 95% CI of a development plan are shown as well as the regulatory consequences of each outcome.*

Hirshberg & Raz (2011, Diabetes Care)

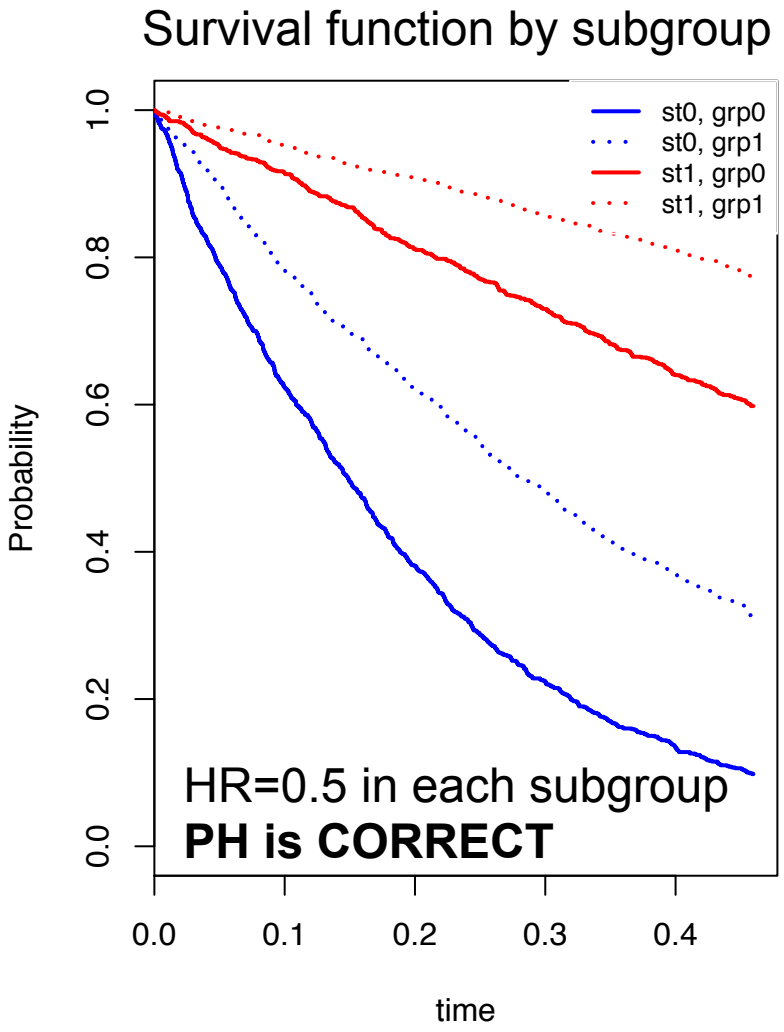
Issues and concerns about hazard ratio estimate (4)

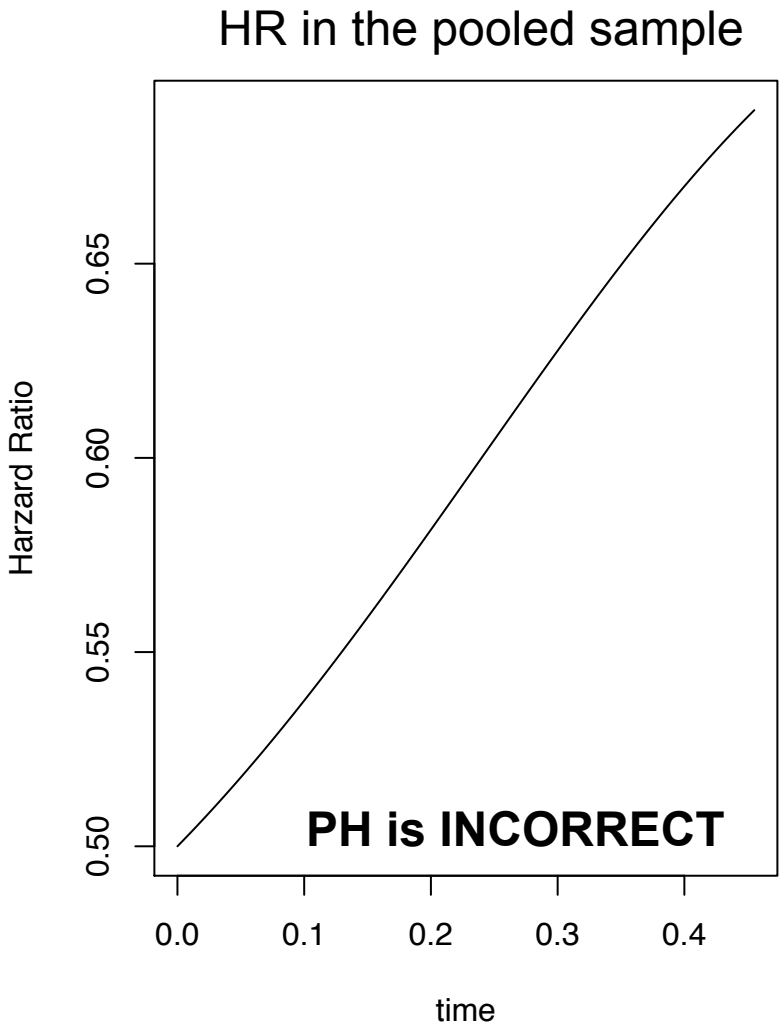
Generally it is incoherent…

- 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

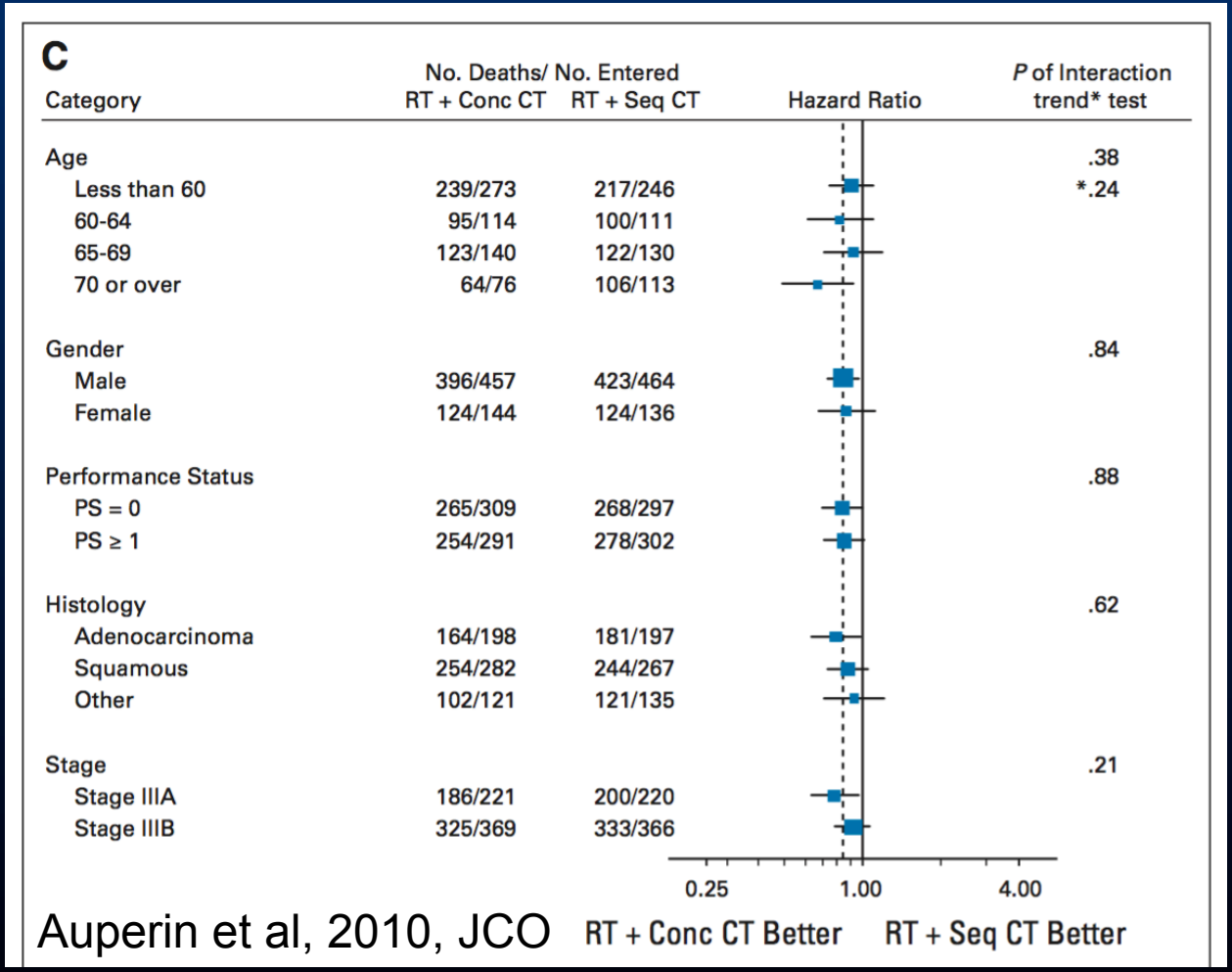
Adjusted and unadjusted analyses are estimating different quantities each other

Numerical example





What is the implication?



Auperin et al, 2010, JCO

Often, we see HR forest plots like this in journals…. but

If the PH is CORRECT with the pooled sample, the PH is INCORRECT in these subgroups, vice versa….

PH assumption cannot be correct in all of these

Summary of the issues of HR

- If the modeling assumption does not hold *(usually it does not hold in practice)*, we do not even know what we are estimating
- No reference number for interpretation
- Does not help much if rare event case
- Incoherent

Checking the PH assumption may be critical in practice…

- Check by your eye ball – (subjective…)
  - Log(-log(S(t))) vs. t
- Statistical tests
  - Include time-varying covariates in Cox's model
  - Goodness of fit tests
    - Schoenfeld residuals (Schoenfeld, 1982)
    - Weighted residuals (Grambsh & Therneau, 1994)
    - Cumulative residuals (Lin & Wei, 2002)

However, can we actually rule out non-PH cases by statistical tests?

Testing the PH assumption

- Null hypothesis: "PH is correct"
- Alternative hypothesis: "PH is NOT correct"
  - If SIG → reject the null & take "PH is NOT correct"
  - If NS → *retain* the null hypothesis

- N.S. does not necessary means the null hypothesis ("PH is correct") is true
- Also, if sample size is huge, the test will be sig. anyway

Examples from recent publications

1. EPOETIN Safety Study
   Leyland-Janes et al. (2016, JCO)
   ← Letter to the editor (JCO 2016; 34(3)1:3818)

2. LABAs Safety Study
   Stempel et al. (2016, NEJM)
   ← Letter to the editor (NEJM 2016;375(11):1097)

Example 1: EPOETIN safety study

JOURNAL OF CLINICAL ONCOLOGY

ORIGINAL REPORT

A Randomized, Open-Label, Multicenter, Phase III Study of Epoetin Alfa Versus Best Standard of Care in Anemic Patients With Metastatic Breast Cancer Receiving Standard Chemotherapy

Brian Leyland-Jones, Igor Bondarenko, Gia Nemsadze, Vitaliy Smirnov, Iryna Litvin, Irakli Kokhreidze, Lia Abshilava, Mikheil Janjalia, Rubi Li, Kuntegowda C. Lakshmaiah, Beka Samkharadze, Oksana Tarasova, Ranjan Kumar Mohapatra, Yaroslav Sparyk, Sergey Polenkov, Vladimir Vladimirov, Liang Xiu, Eugene Zhu, Bruce Kimelblatt, Kris Deprince, Ilya Safonov, Peter Bowers, and Els Vercammen

Author affiliations appear at the end of this article.

Published online ahead of print at www.jco.org on February 8, 2016.

Supported by Janssen Research & Development, Raritan, NJ.

Presented at the San Antonio Breast Cancer Symposium, San Antonio, TX, December 9-13, 2014.

Authors' disclosures of potential conflicts of interest are found in the article online at www.jco.org. Author contributions are

A B S T R A C T

Purpose

An open-label, noninferiority study to evaluate the impact of epoetin alfa (EPO) on tumor outcomes when used to treat anemia in patients receiving chemotherapy for metastatic breast cancer.

Methods

Women with hemoglobin ≤ 11.0 g/dL, receiving first- or second-line chemotherapy for metastatic breast cancer, were randomly assigned to EPO 40,000 IU subcutaneously once a week or best standard of care. The primary end point was progression-free survival (PFS). Secondary end points included overall survival, time to tumor progression, overall response rate, RBC transfusions, and thrombotic vascular events.

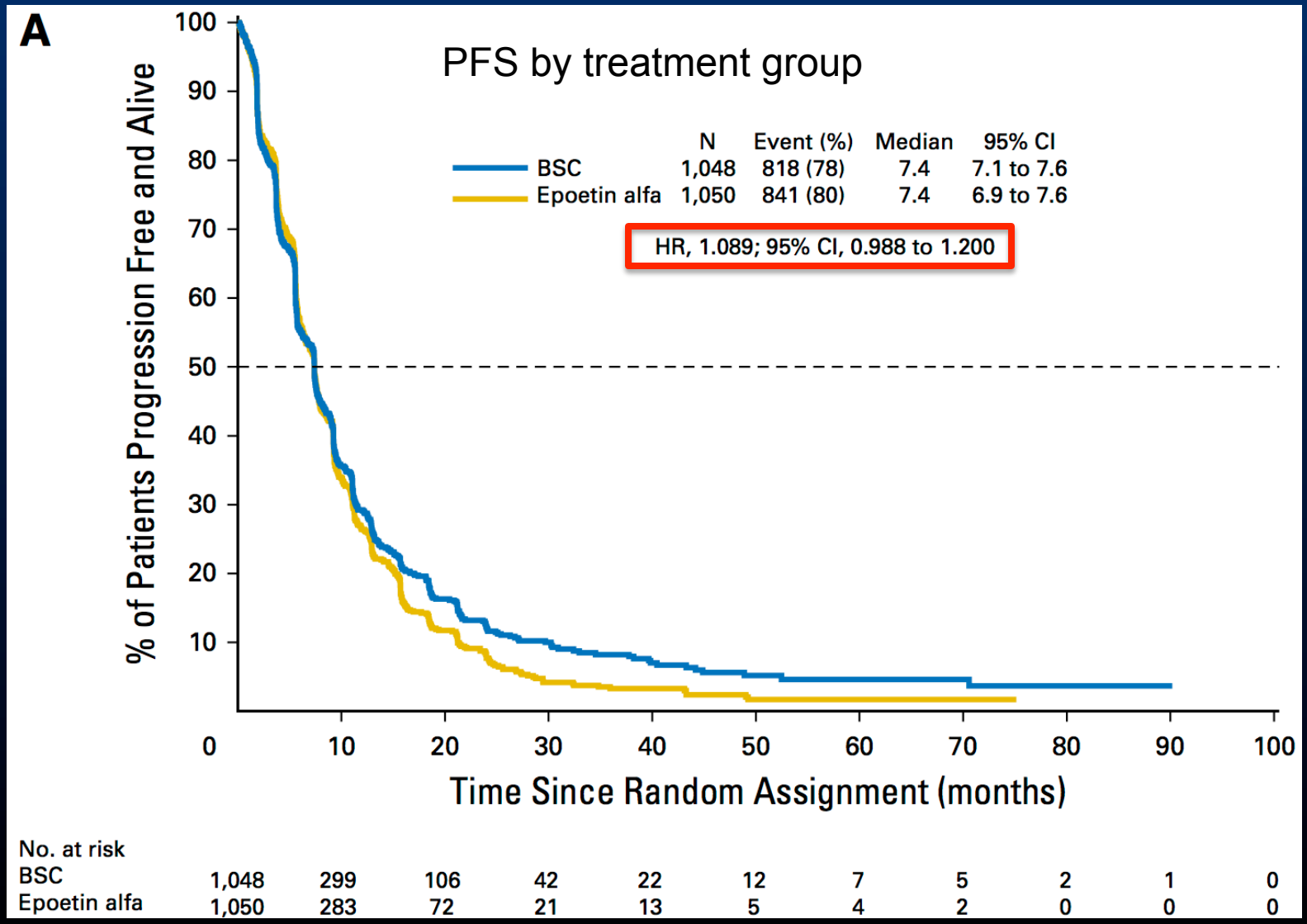
Leyland-Janes et al. (2016, JCO)

Example 1: EPOETIN safety study

- Noninferiority study (EPOETIN vs. BSC)
- Primary endpoint: Progression-Free Survival (PFS)
- **NI criteria: Upper 95%CI of HR < 1.15**
- Planned total PFS events: **1,650** to achieve 80% of power with a 0.025 one-sided alpha level

- **Results: HR estimate [EPO/BSC] was 1.089 (95%CI, 0.988 to 1.200)**

Example 1: EPOETIN safety study



Leyland-Janes et al. (2016, JCO)

Issues of the HR here

- The PH assumption is violated (Cum. residual test: $p<0.001$)

- **HR estimate [EPO/BSC]**
  1.089 (95%CI, 0.988 to 1.200) is not interpretable

  → Need some other measures that provide clinically interpretable and meaningful information of the value of EPOETIN

Example 2: Long-acting beta-agonists (LABAs) safety study for asthma patients

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Serious Asthma Events with Fluticasone plus Salmeterol versus Fluticasone Alone

David A. Stempel, M.D., Ibrahim H. Raphiou, Ph.D., Kenneth M. Kral, M.S., Anne M. Yeakey, M.D., Amanda H. Emmett, M.S., Charlene M. Prazma, Ph.D., Kathleen S. Buaron, B.S.N., and Steven J. Pascoe, M.B., B.S., for the AUSTRI Investigators*

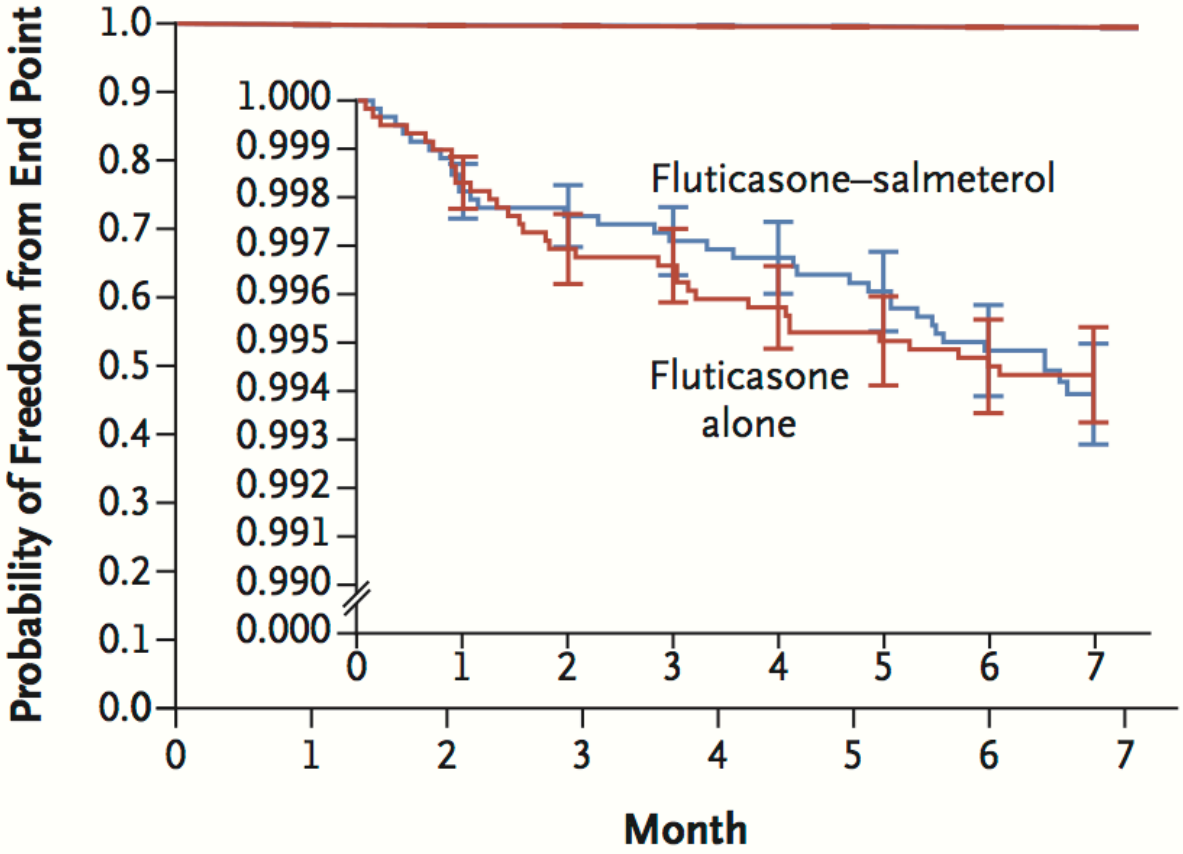
Stempel et al. (2016, NEJM)

LABAs safety study for asthma

- Noninferiority study (fluticasone-salmeterol vs. fluticasone alone)
- Primary endpoint: 1st serious asthma-related event
- **NI criteria: Upper 95% CI of HR < 2.0**
- Planned total events: **87** to achieve 90% of power with a 0.025 one-sided alpha level (Total planned sample size: **11,644**)

- **Results: 67 events (N=11,679)**
  HR estimate was 1.03 (95%CI, 0.64 to 1.66)

LABAs safety study for asthma



No. at Risk

Fluticasone–salmeterol	5834	5798	5761	5731	5707	5671	5625	527
Fluticasone alone	5845	5811	5770	5726	5695	5669	5621	529

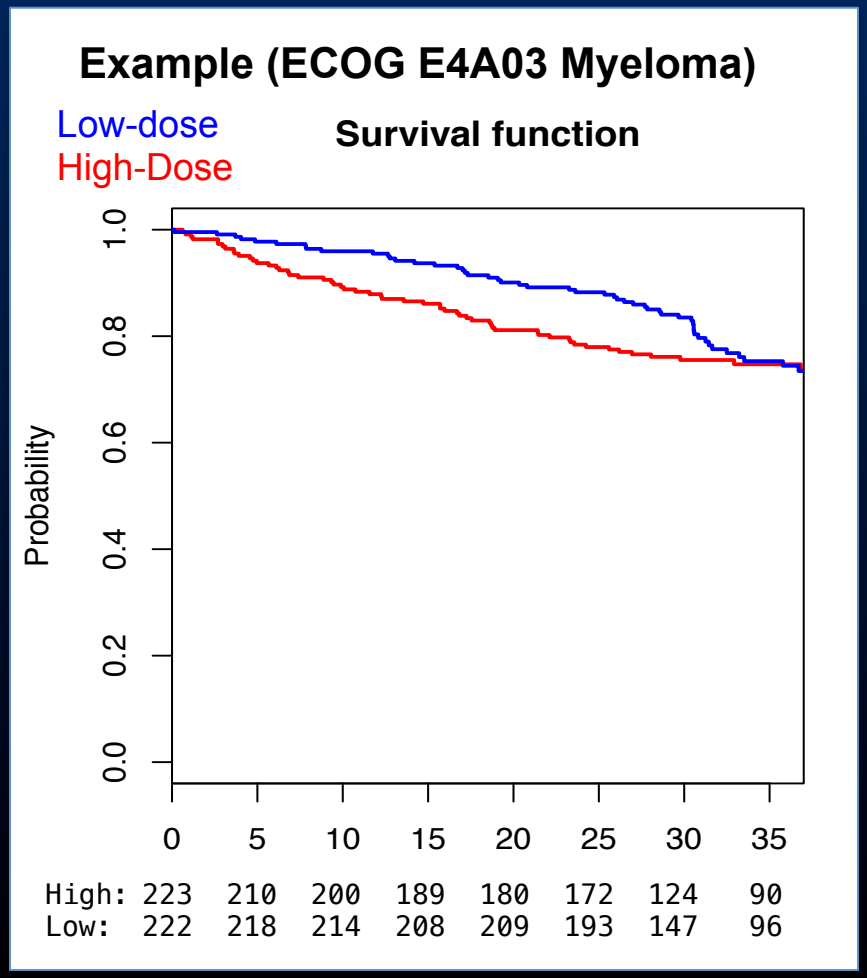
Stempel et al. (2016, NEJM)

Issues of the HR here

- No clear reference value to assess if the observed increase of hazard indicates a clinically important difference
- 95%CI of HR (0.64 to 1.66) met the pre-specified NI criterion. However, not clear that a possible **66%** increase of hazard would be acceptable clinically to make such a claim because of no reference value

Alternatives to HR

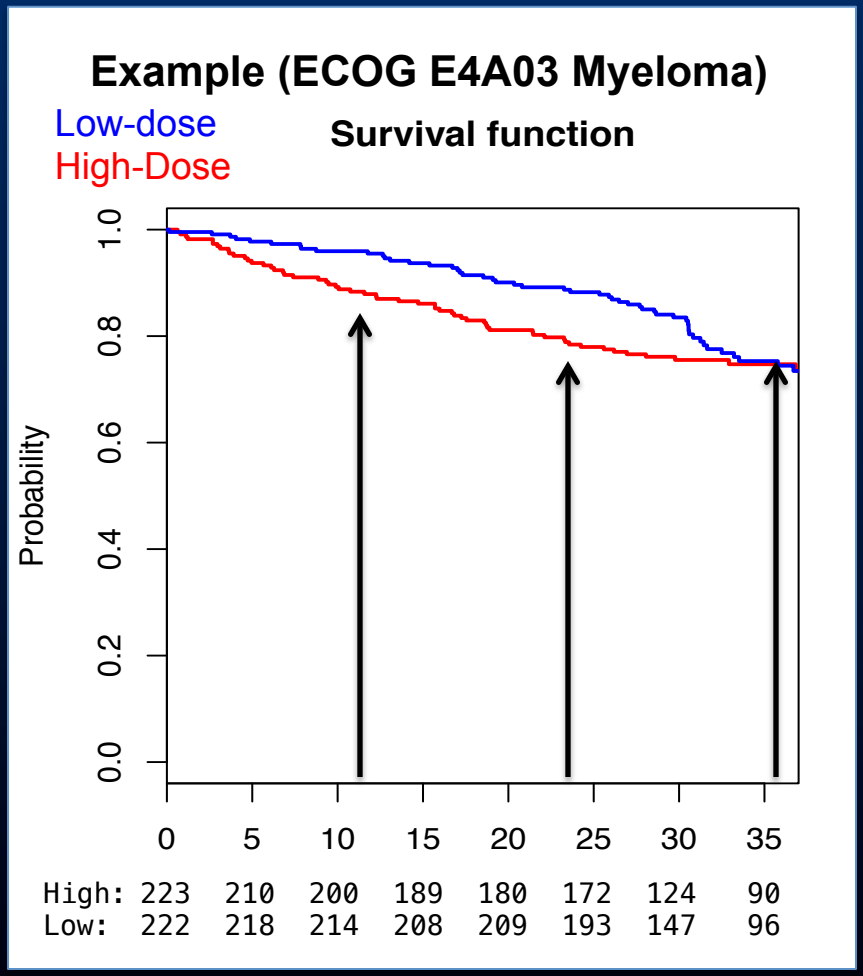
What are alternative measures that do not have shortcomings of the HR?



Desirable measures…

- have no strong assumption on the relationship between two survival time distributions
- have a clear baseline reference

(1) t-year survival probability

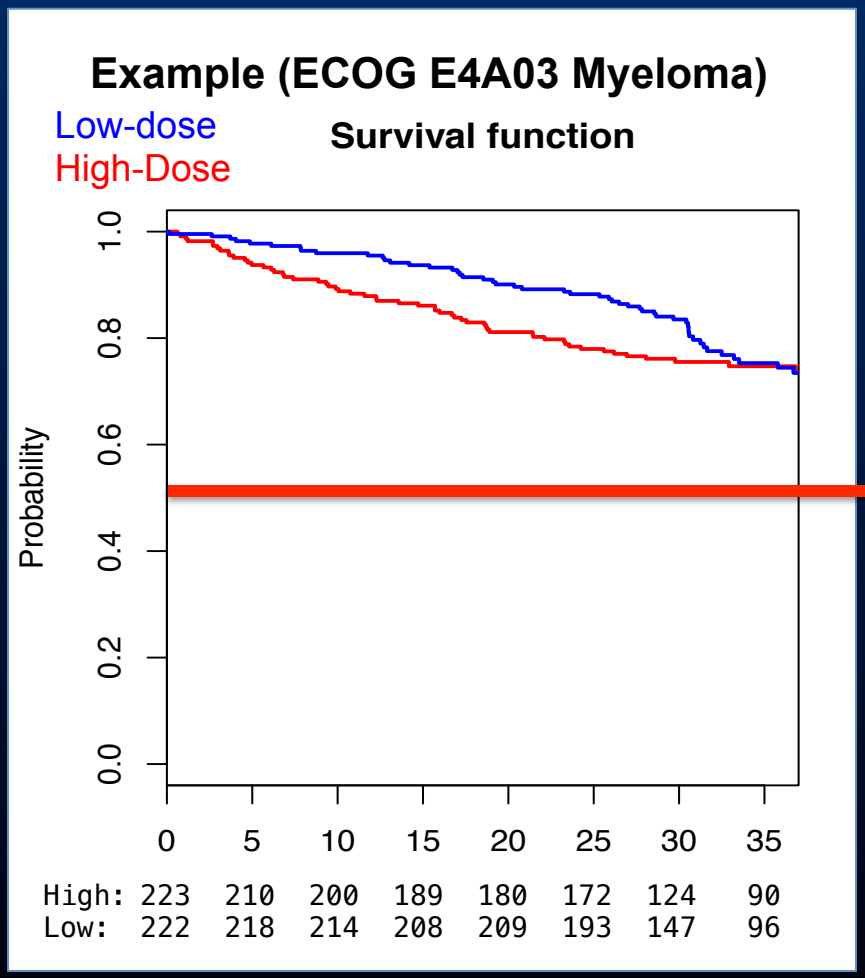


$$S_1(t) - S_0(t)$$

$$S_1(t)/S_0(t)$$

1-year?
2-year?
3-year?

(2) Median survival time



$$S_1^{-1}(0.5) - S_0^{-1}(0.5)$$

$$S_1^{-1}(0.5) \;/\; S_0^{-1}(0.5)$$

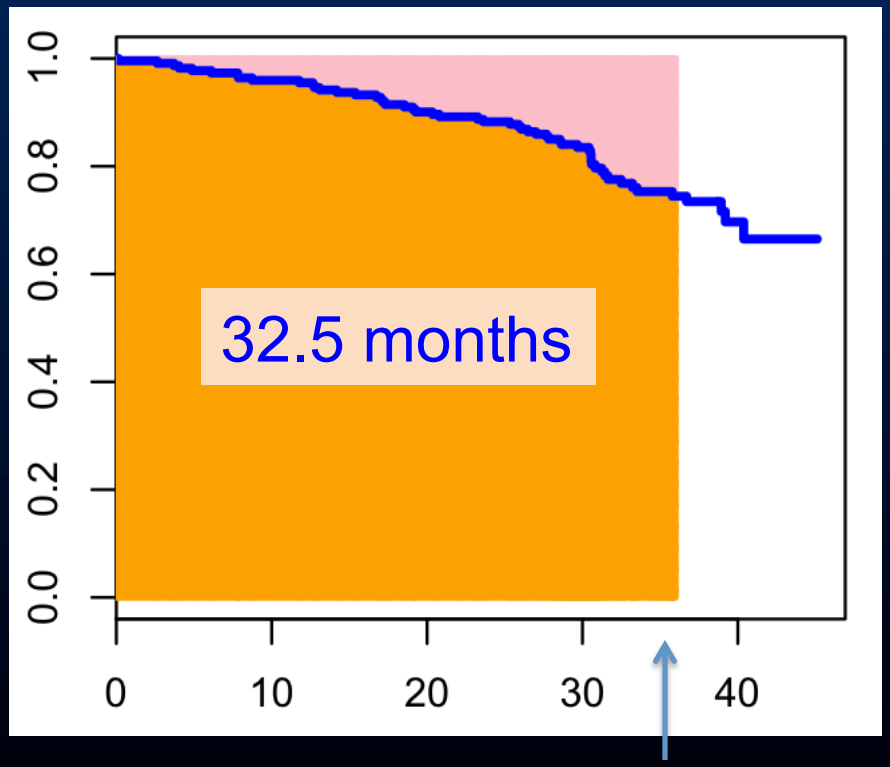
???

Sometimes this is inestimable…

(3) Restricted mean survival time (RMST)

$$\int_0^{\tau} S_1(u)du$$

Low-dose



$\tau = 36m$

Interpretation:
If you follow-up patients on low-dose for 36m, patients will survive 32.5 months *on average*
$(\tau-\text{year life expectancy})$

Note:
RMST is estimable even when median survival time is inestimable

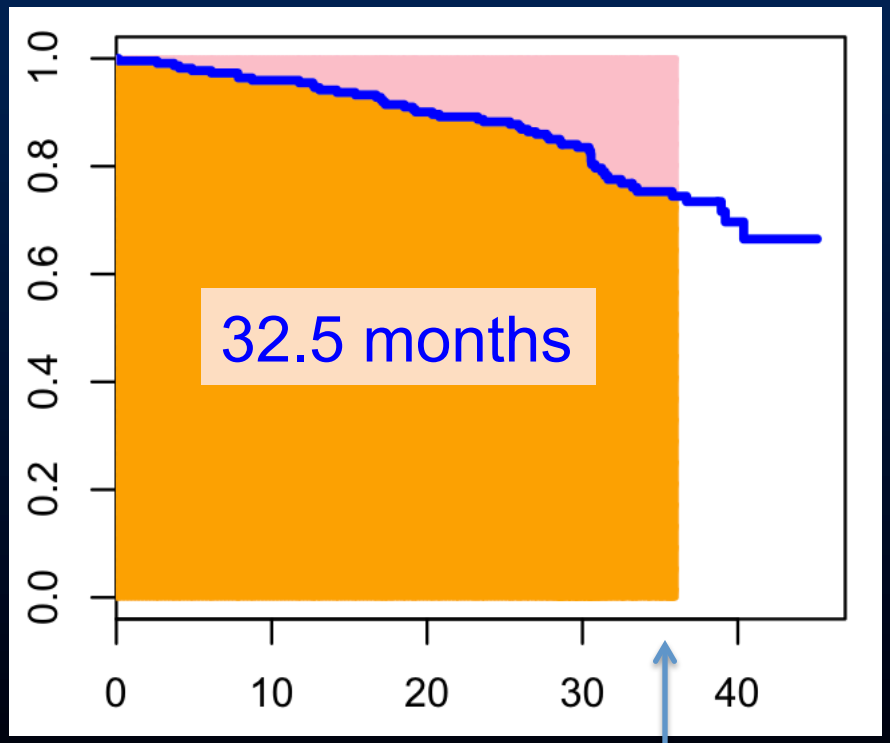
(3) Restricted mean survival time (RMST)

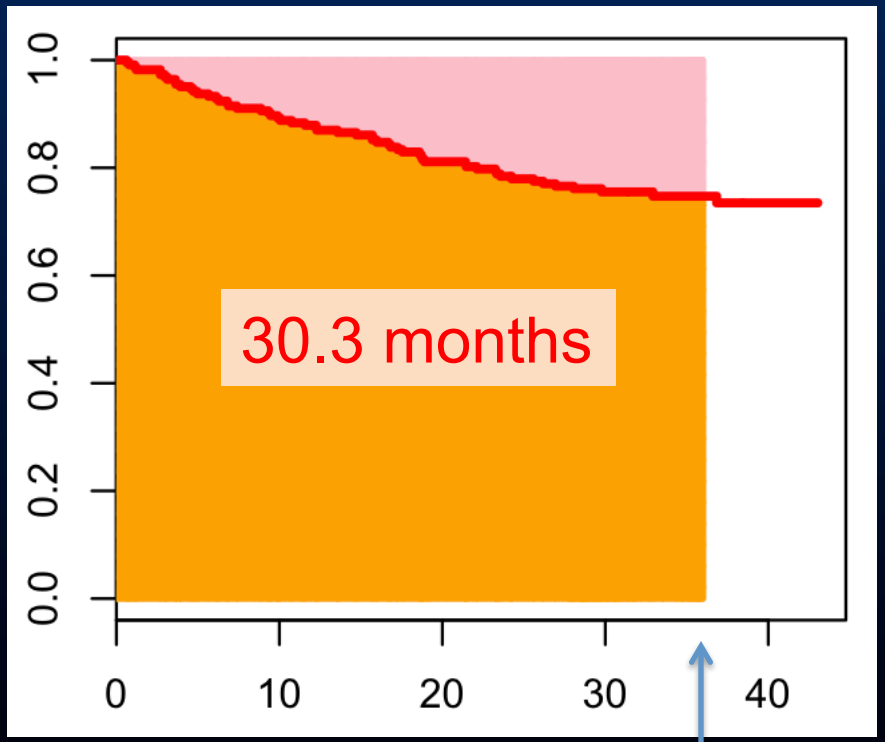
$$\int_0^{\tau} S_1(u)du$$

Low-dose

$$\int_0^{\tau} S_0(u)du$$

high-dose





$\tau = 36m$

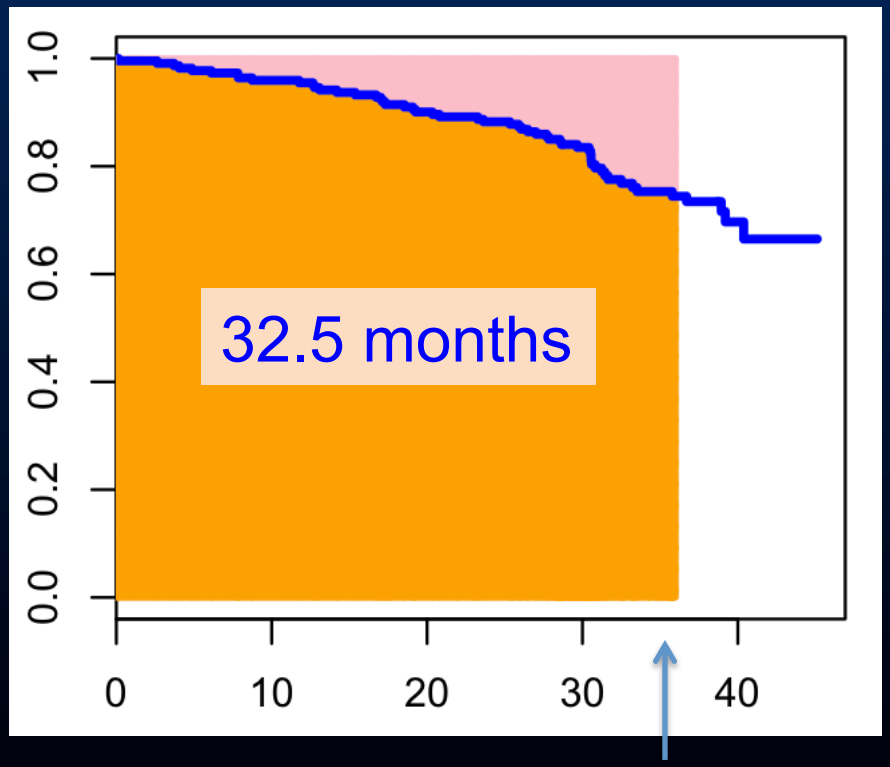
$\tau = 36m$

(3) Restricted mean survival time (RMST)

Difference in RMST: 2.2 months (0.95CI: 0.5 to 4.0, p=0.014)

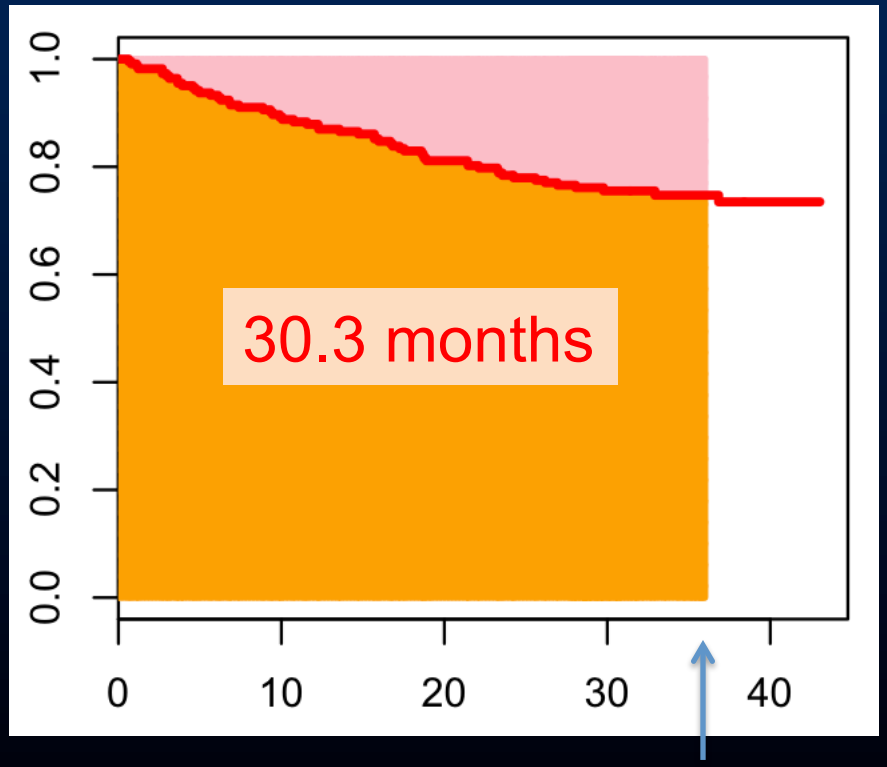
Recall: HR= 0.87 (0.95CI: 0.60 - 1.27), NS

Low-dose



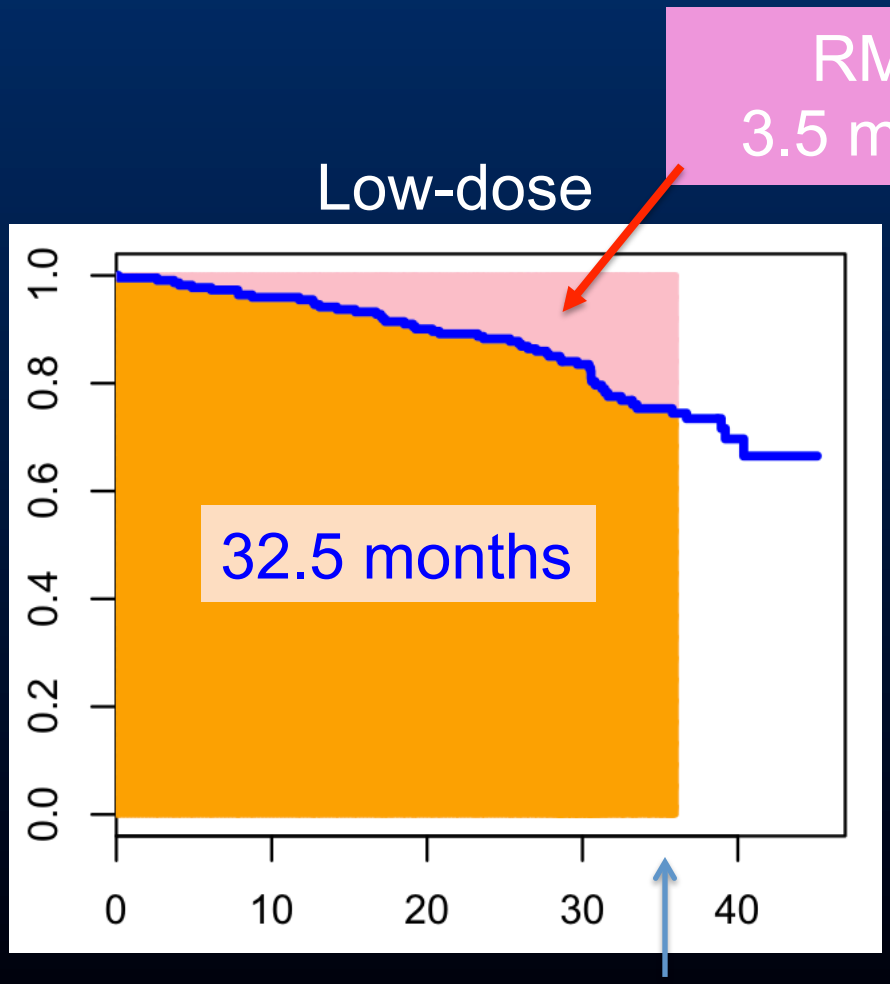
$\tau = 36m$

high-dose



$\tau = 36m$

(4) Restricted mean time lost (RMTL)



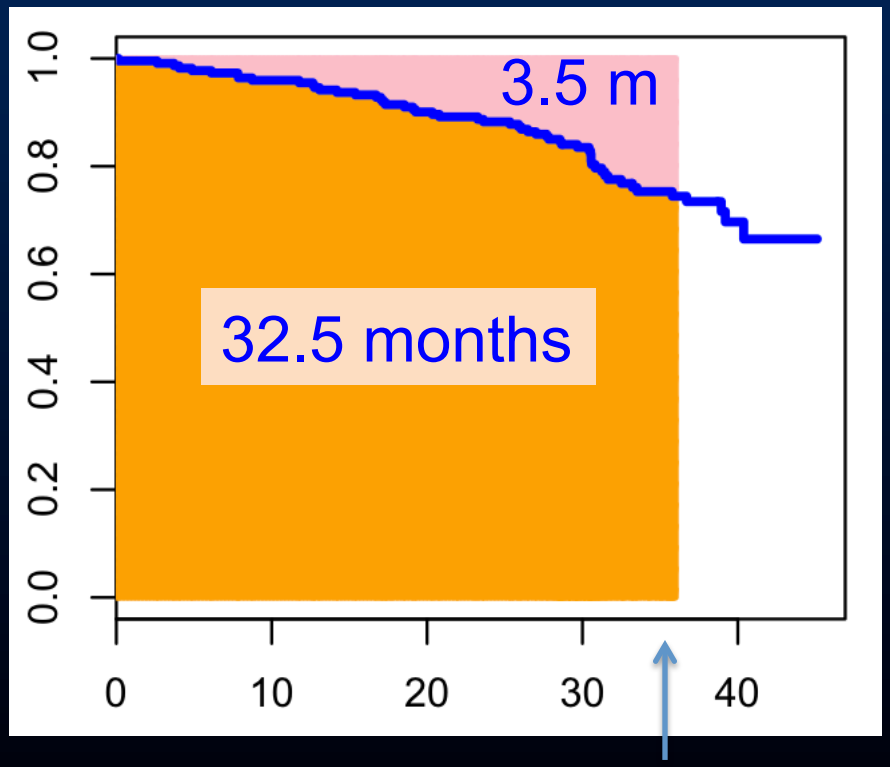
RMTL = τ - RMST

Area above the survival curve

(4) Restricted mean time lost (RMTL)

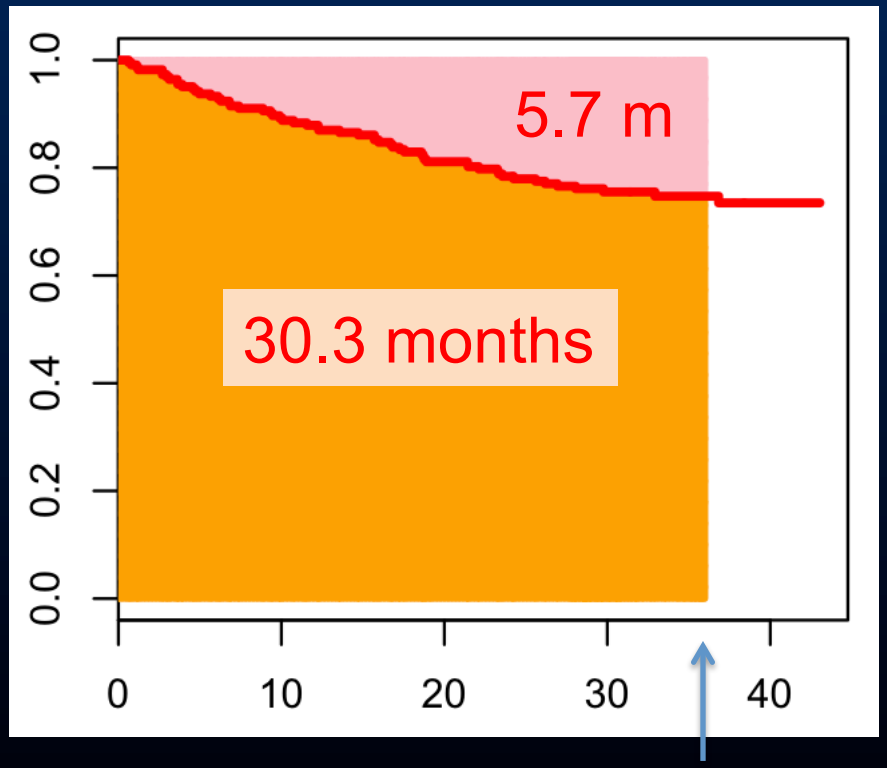
Ratio of RMTL: 0.61 (0.95CI: 0.42 to 0.90, p=0.013)

Low-dose



$\tau = 36m$

high-dose



$\tau = 36m$

Ratio of RMTL and HR

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

$$\frac{\int_0^\tau 1 - e^{-\lambda_1 t} dt}{\int_0^\tau 1 - e^{-\lambda_0 t} dt} \approx \frac{\int_0^\tau \lambda_1 t dt}{\int_0^\tau \lambda_0 t dt} = \frac{\lambda_1}{\lambda_0}$$

Results of the RMST (RMTL) analysis with the ECOG Myeloma data

Metric	Low dose	High dose	Difference (0.95 CI)	Ratio (0.95 CI)
RMST (36 months)	32.5	30.3	2.2 (0.5, 4.0) P=0.014	1.07 (1.01, 1.14) P=0.015
RMTL (36 months)	3.5	5.7	---	0.61 (0.42, 0.90) P=0.013

Go back to the two examples

1. EPOETIN Safety Study
   Leyland-Janes et al. (2016, JCO)
   ← Letter to the editor (JCO 2016; 34(3)1:3818)

2. LABAs Safety Study
   Stempel et al. (2016, NEJM)
   ← Letter to the editor (NEJM 2016;375(11):1097)

Revisit EPOETIN safety study

JOURNAL OF CLINICAL ONCOLOGY

ORIGINAL REPORT

A Randomized, Open-Label, Multicenter, Phase III Study of Epoetin Alfa Versus Best Standard of Care in Anemic Patients With Metastatic Breast Cancer Receiving Standard Chemotherapy

Brian Leyland-Jones, Igor Bondarenko, Gia Nemsadze, Vitaliy Smirnov, Iryna Litvin, Irakli Kokhreidze, Lia Abshilava, Mikheil Janjalia, Rubi Li, Kuntegowda C. Lakshmaiah, Beka Samkharadze, Oksana Tarasova, Ranjan Kumar Mohapatra, Yaroslav Sparyk, Sergey Polenkov, Vladimir Vladimirov, Liang Xiu, Eugene Zhu, Bruce Kimelblatt, Kris Deprince, Ilya Safonov, Peter Bowers, and Els Vercammen

Author affiliations appear at the end of this article.

Published online ahead of print at www.jco.org on February 8, 2016.

Supported by Janssen Research & Development, Raritan, NJ.

Presented at the San Antonio Breast Cancer Symposium, San Antonio, TX, December 9-13, 2014.

Authors' disclosures of potential conflicts of interest are found in the article online at www.jco.org. Author contributions are

A B S T R A C T

Purpose

An open-label, noninferiority study to evaluate the impact of epoetin alfa (EPO) on tumor outcomes when used to treat anemia in patients receiving chemotherapy for metastatic breast cancer.

Methods

Women with hemoglobin ≤ 11.0 g/dL, receiving first- or second-line chemotherapy for metastatic breast cancer, were randomly assigned to EPO 40,000 IU subcutaneously once a week or best standard of care. The primary end point was progression-free survival (PFS). Secondary end points included overall survival, time to tumor progression, overall response rate, RBC transfusions, and thrombotic vascular events.

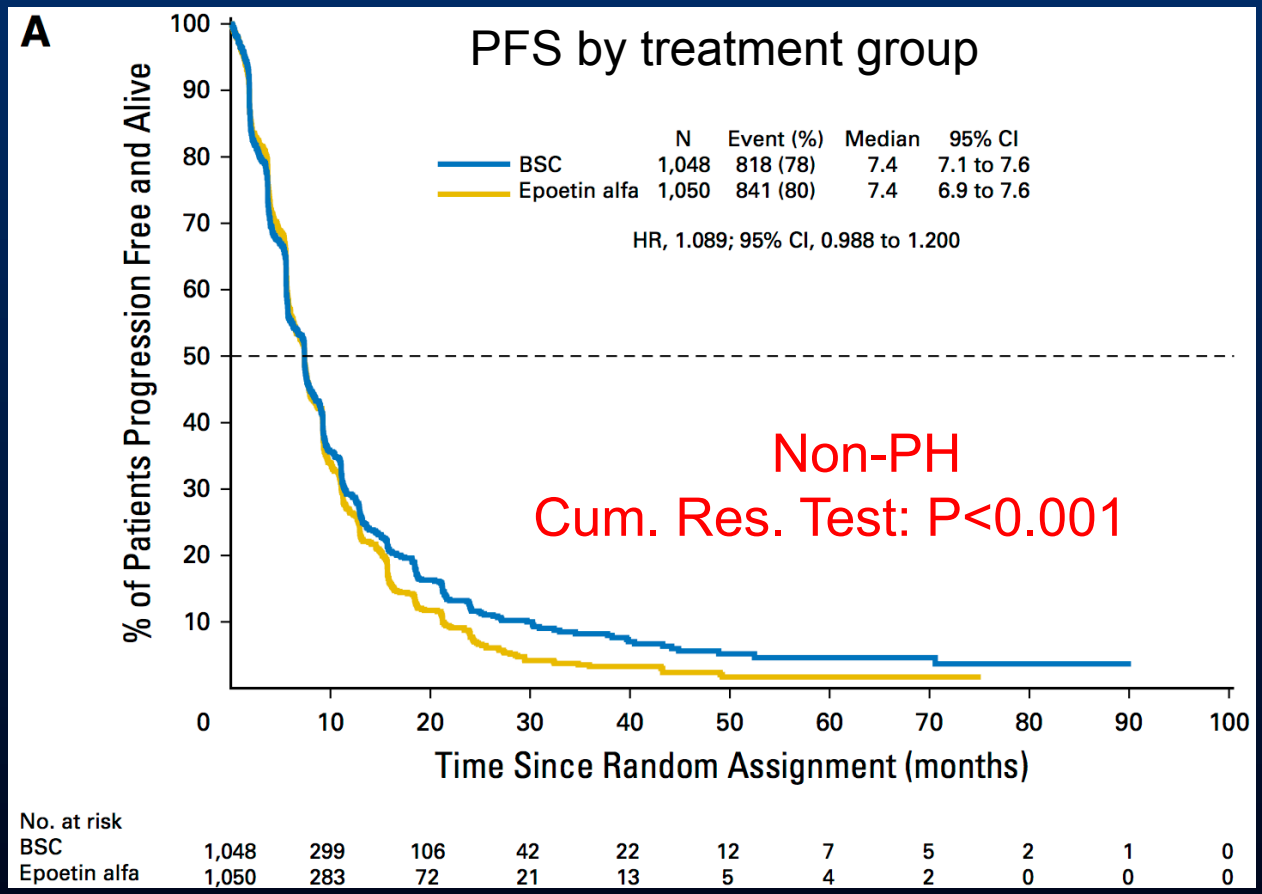
Leyland-Janes et al. (2016, JCO)

Revisit EPOETIN safety study

- Noninferiority study (EPO vs. BSC)
- Primary endpoint: PFS
- NI margin: HR 1.15
- Planned total PFS events: 1650 to achieve 80% of power with a 0.025 one-sided alpha level

- Based on their analysis results, the authors concluded that *"Overall, this study did not achieve the noninferiority objective in ruling out a 15% increased risk in PD or death."*

Revisit EPOETIN safety study



Leyland-Janes et al. (2016, JCO)

HR: 1.089 (95%CI, 0.988 to 1.200)

What if we used RMST?

EPOETIN safety study

Metric	EPO (months)	BSC (months)	Difference (0.95 CI)	Ratio (0.95 CI)
RMST (48 months)	9.9	11.4	-1.5 (-2.6, -0.5) P=0.004	1.04 (1.01, 1.07) P=0.004
RMTL (48 months)	38.1	36.6	---	0.87 (0.79, 0.96) P=0.004

These are clinically interpretable and meaningful information of the value of EPO

Ref: HR: 1.089 (95%CI, 0.988 to 1.200)

Revisit LABAs safety study

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Serious Asthma Events with Fluticasone plus Salmeterol versus Fluticasone Alone

David A. Stempel, M.D., Ibrahim H. Raphiou, Ph.D., Kenneth M. Kral, M.S., Anne M. Yeakey, M.D., Amanda H. Emmett, M.S., Charlene M. Prazma, Ph.D., Kathleen S. Buaron, B.S.N., and Steven J. Pascoe, M.B., B.S., for the AUSTRI Investigators*

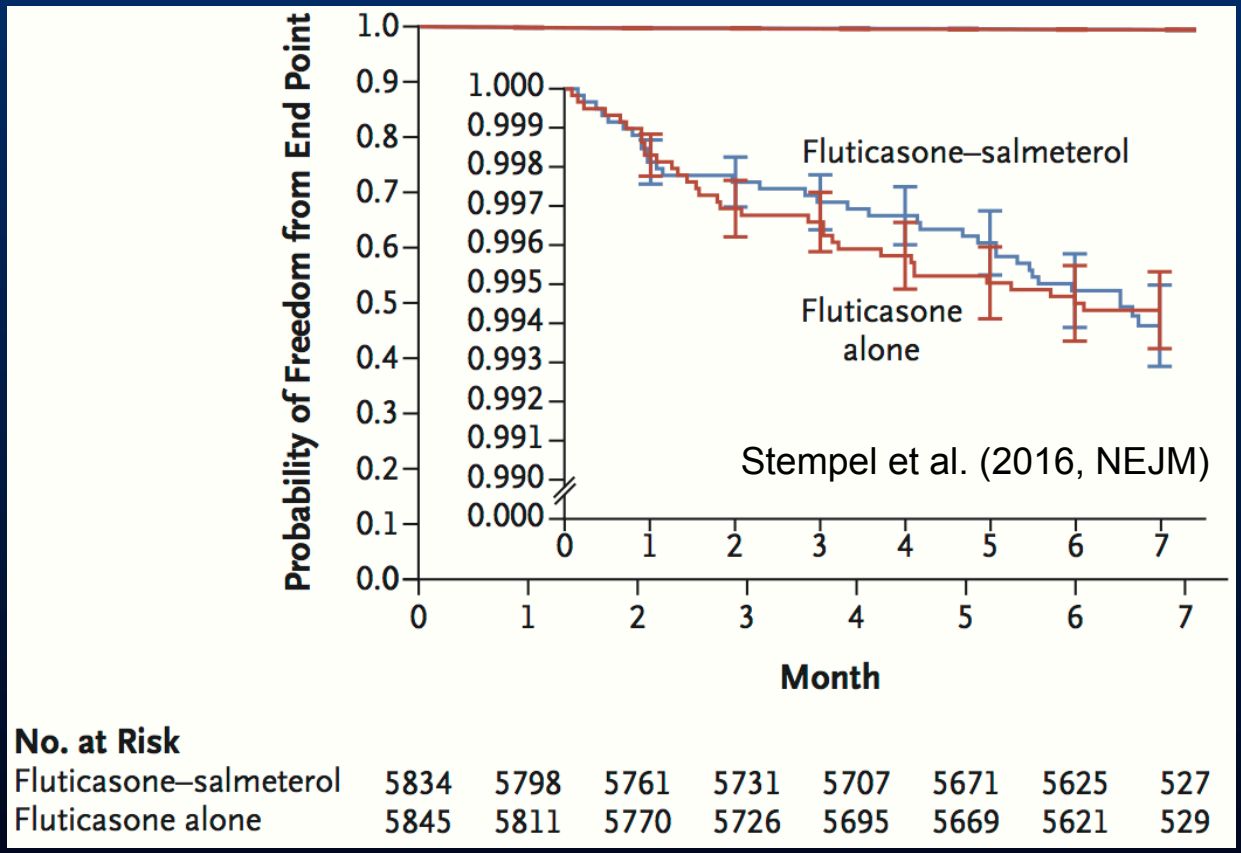
Stempel et al. (2016, NEJM)

Revisit LABAs safety study

- Noninferiority study (fluticasone-salmeterol vs. fluticasone alone)
- Primary endpoint: 1st serious asthma-related event
- NI criteria: Upper 95% CI of HR < 2.0
- Planned total events: 87 to achieve 90% of power with a 0.025 one-sided alpha level (Total planned sample size: 11644)

- **Results: HR in the fluticasone–salmeterol group was 1.03 (95%CI, 0.64 to 1.66)**

Revisit LABAs safety study



HR: 1.03 (95%CI, 0.64 to 1.66)

What if we used RMST?

Results of the RMST (RMTL) analysis LABAs safety study

Metric	Flu–sal (days)	Flu alone (days)	Difference (0.95 CI)	Ratio (0.95 CI)
RMST (210 days)	209.3	209.2	0.1 (-0.3, 0.5)	1.001 (0.999, 1.002)
RMTL (210 days)	0.7	0.8	---	0.858 (0.504, 1.461)

These numbers would help us understand the treatment difference much better
(ref. 95%CI of HR was 0.64 to 1.66)

Message of our letter

- HR 1.03 (95%CI, 0.64 to 1.66) met the pre-specified NI criterion. However, not clear that a possible 66% increase of hazard would be acceptable clinically to make such a claim.
- For a safety study, using HR may not be appropriate.
- For example, RMST difference, 0.1day
- (95%CI: -0.3 to 0.5 days) has a much clearer clinical interpretation than the HR.
- With this measure, a much smaller study size could be sufficient for a non-inferiority claim.

What if a smaller study?

95% confidence intervals for various measures

	All data	**50%**	**25%**	**20%**
	N=11,679	N=5840	N=2920	N=2336
Hazard Ratio	**(0.64, 1.66)**	(0.51, 2.00)	(0.38, 2.78)	(0.00, $1.14*10^{11}$)
Difference in RMST at Day 210 [days]	**(-0.3, 0.5)**	(-0.5, 0.7)	(-0.7, 0.9)	(-0.7, 1.0)

Maybe already tight enough to make a NI decision

A standard practice…

Description

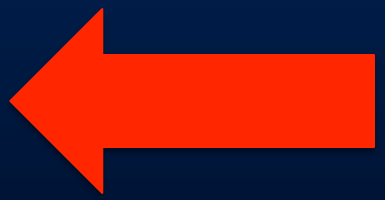
Test

Estimation of treatment effect

Kaplan-Meier

Log-rank test

Estimate HR by Cox reg



What about testing?

- Logrank test
  - Robust
  - Most powerful under PH alternatives
  - PL score test in Cox's PH model
- Model-free measures can be used for testing (conversion to test)
  - Robust
  - RMST-based tests are comparable under PH and outperform under some non-PH cases
  - Estimation (or a clinically interpretable metric) and testing will be coherent

Problems of some other routinely used methods

- Meta-analysis
  - fixed-effect, random-effects models
  - censored data

- Are unadjusted and adjusted analyses estimating the same thing?

Conclusions

- Almost routinely, the HR is used to summarize the treatment effect, but the HR estimates may not provide clinically interpretable information with respect to risk-benefit perspectives
- Robust alternative measures (e.g., RMST difference) would be useful

Remarks

Move beyond the comfort zone

- Some methods have become routine, but some of them have significant limitations regarding robustness and clinical interpretability
- It seems that investigators tend to choose routine methods without giving adequate considerations to these issues in practice
- Continuation of such trends may ultimately slow the advancement of clinical research on public health

References (1)

Papers regarding issues of the HR

- Uno et al. Moving Beyond the Hazard Ratio in Quantifying the Between-Group Difference in Survival Analysis. *JCO 2014*, *32*(22), 2380–2385

- Uno et al. Alternatives to Hazard Ratios for Comparing the Efficacy or Safety of Therapies in Noninferiority Studies. *Annals of Internal Medicine 2015*, *163*(2), 127–12.

References (2)

Letter to the editor

1. Hasegawa et al. How To Summarize the Safety Profile of Epoetin Alfa Versus Best Standard of Care in Anemic Patients With Metastatic Breast Cancer Receiving Standard Chemotherapy? *JCO. 2016 34(31):3818*

2. Hasegawa et al. Safety Study of Salmeterol in Asthma in Adults. *NEJM. 2016*; 375(11):1097.

Some papers re RMST in medical journals

- Trinquart et al. (2016). Comparison of Treatment Effects Measured by the Hazard Ratio and by the Ratio of Restricted Mean Survival Times in Oncology Randomized Controlled Trials. *JCO*
- Chappel & Zhu (2016). Describing Differences in Survival Curves. *JAMA Oncology*.
- Péron et al. The Net Chance of a Longer Survival as a Patient-Oriented Measure of Treatment Benefit in Randomized Clinical Trials. *JAMA Oncology*, *2*(7), 901–5.
- Ahern RP. (2016). Restricted Mean Survival Time: An Obligatory End Point for Time-to-Event Analysis in Cancer Trials? *JCO*, 1–4.

Computer codes

Implementation of RMST analysis

Computer programs are available on 3 major platforms

- **R:** *survRM2* package
- **Stata:** *strmst2* command
  (Cronin, Tian, and Uno, Stata Journal, in press)
- **SAS:** SAS macro *%rmst2*

visit my website: http://bcb.dfci.harvard.edu/~huno/

What you can do with these packages:

- Two-sample comparison based on the RMST
  (difference in RMST, ratio of RMST, and ratio of RMTL)
- RMST regression


LIFE BEGINS AT
THE END
OF YOUR
COMFORT ZONE
– NEALE DONALD WALSCH
BLOG.ZERO DEAN.COM

Backup

Choice of τ

- In a confirmatory study, τ should be pre-specified
- 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

Note: Others measures (incl. HR) also have a τ ***implicitly***. Extrapolation beyond the end of the study followup is always a challenge

Choice of τ (ad-hoc)

When choosing τ a posteriori, we will need some objective rule….

For example,

- Based on "effective sample size" (Karrison, 1987)

$$\text{Choose the largest } t$$
$$\text{s.t. } \hat{N}_{EFF}(t) > \tfrac{2}{3}N, \text{ where}$$

$$\hat{N}_{EFF}(t) = \hat{S}(t)(1-\hat{S}(t))/\hat{V}\{\hat{S}(t)\}$$

Adjusting for covariates

- Standard stratified analyses will work with RMST

- ANCOVA-type regression for RMST (Tian et al. 2014, Biostatistics)

Study Design with RMST

Key elements of designing of safety (noninferiority) study

1. metric/ parameter used to compare groups
2. the noninferiority (NI) margin w.r.t. the above metric
3. statistical inference procedure (e.g., 95% CI for the group contrast measure) for assessing NI
4. parametric distribution for the outcome variable
5. expected patient accrual over time
6. timing for the end of the study or patient's potential follow-up time
7. number of patients expected to enroll

Numerical example:
How to Design a CV Outcome Trial via RMST?

The standard approach is to use the HR

- Show the upper bound of the 95% CI for HR < 1.5
- To make this happen with a probability 90%, we need to observe 256 MACE events

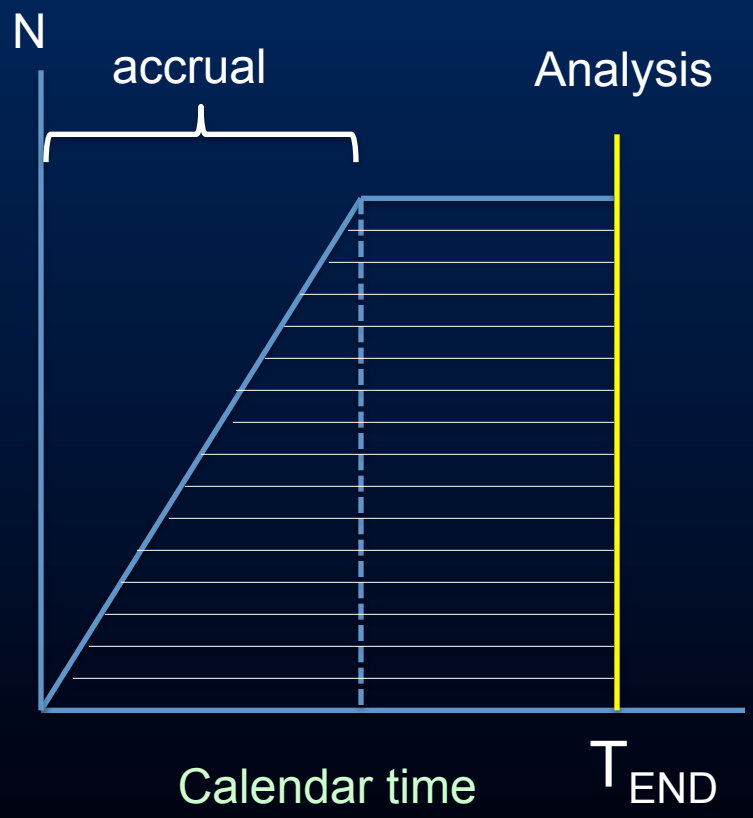
For the treatment, assuming

- the MACE annual event rate is 1% or 1.5%

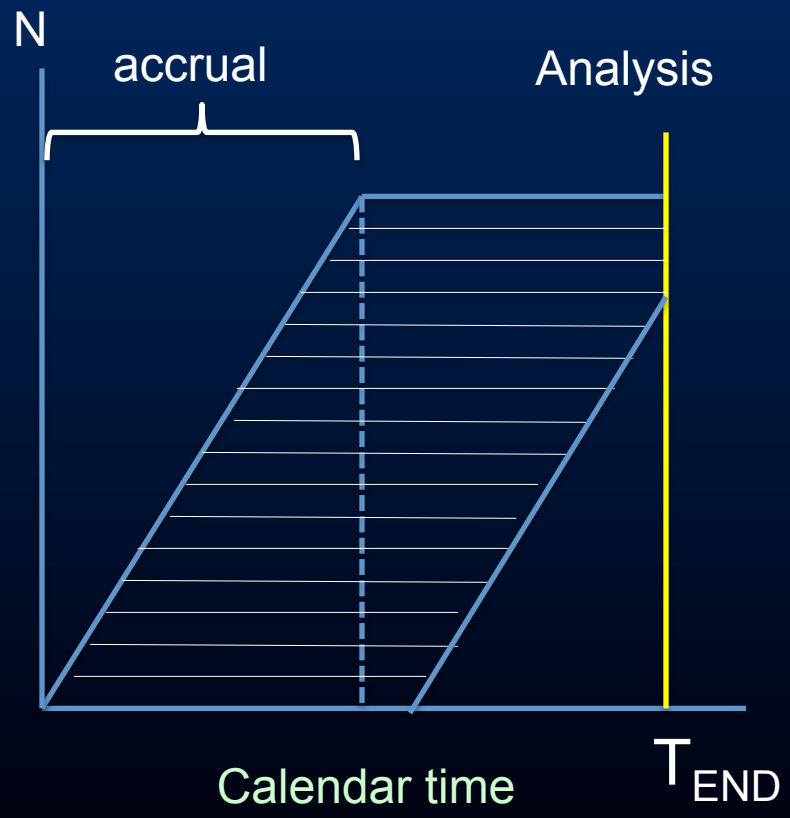
How large and how long a study is needed taking this standard HR interval approach for 2 year followup?

Two kinds of follow-up scheme

Pattern 1: no maximum FU for each patient



Pattern 2: With a maximum FU for each patient



T_{END} denotes the study duration

Consider Follow-up pattern 1

Event Rate annually (%)	Accrual Rate per (yr)	Accrual Period (yr)	Total N*	T_{END} (yr)	# of events	Upper Bound of CI
1	1000	5	**5000**	**7.7**	256	1.5
1	500	10	**5000**	**10.3**	256	1.5
1.5	1000	5	**5000**	**6.0**	256	1.5
1.5	500	7	**3500**	**8.5**	256	1.5

* If we assume 20% of enrolled patients will not be used for the analysis, divide this number by 0.8

Follow-up pattern 2
(max. FU of each pat. 2.1Y)

Event Rate yearly(%)	Accrual Rate per (yr)	Accrual Period (yr)	Total N* enrolled	T_{END} (yr)	# of events	Upper Bound of CI for HR
1	2000	**6.1**	**12196**	**8.1**	256	1.5
1	1500	**8.1**	**12196**	**10.1**	256	1.5
1.5	1500	**5.4**	**8133**	**7.4**	256	1.5
1.5	1000	**8.1**	**8133**	**10.1**	256	1.5

* If we assume 20% of enrolled patients will not be used for the analysis, divide this number by 0.8

What if we use RMST?

(with 15 days as the NI margin, 2% of 2 years)

Questions

- Would it be possible to make a decision much earlier using difference of RMST?
- How tight would the 95%CI for Diff. of RMST be?

Follow-up pattern 2
(max. FU of each pat. 2.1Y)

Event Rate yearly (%)	Accrual Rate per (yr)	Accrual Period (yr)	Total N*	T_{END} (yr)	# of events	Upper Bound of CI for HR	Upper Bound of CI for RMST diff (days)
1	2000	**6.1**	**12196**	**8.1**	256	1.5	**3**
		3.0	**6000**	**3.0**	**82**	**2.1**	**5**
		2.5	**5000**	**2.5**	**61**	**2.3**	**6**
1	1500	**8.1**	**12196**	**10.1**	256	1.5	
		3.0	**4500**	**3.0**	**61**	**2.3**	**6**
		2.5	**3750**	**2.5**	**46**	**2.6**	**7**
1.5	1500	**5.4**	**8133**	**7.4**	256	1.5	**5**
		3.0	**4500**	**3.0**	**92**	**1.9**	**8**
		2.5	**3750**	**2.5**	**69**	**2.2**	**9**
1.5	1000	**8.1**	**8133**	**10.1**	256	1.5	
		3.0	**3000**	**3.0**	61	2.4	**9**
		2.5	**2500**	**2.5**	46	2.7	**11**

Follow-up pattern 2
(max. FU of each pat. 2.1Y)

How about using NI margin of 15 days?

Event Rate at Yearly(%)	Accrual Rate per (yr)	Accrual Period (yr)	Total N*	T_{END} (yr)	# of events	Upper Bound of CI for HR	Upper Bound of CI for RMST diff (days)
1	500	2	1000	4	21	4.6	12
				3	18	5.3	12
1.5	500	2	1000	4	31	3.4	14
				3	27	3.7	15

Follow-up pattern 2 (max. FU of each pat. 2.1Y)

A higher event rate case

Cum. Event Rate at Year 2	Accrual Rate per (yr)	Accrual Period (yr)	Total N* enrolled	T_{END} (yr)	# of events at T_{END}	Upper Bound of 95% CI for HR 90% Percentile	Upper Bound of 95% CI for RMST diff. (2yr) 90% Percentile
10%	500	**5.12**	**2560**	**7.12**	256	1.50	
				5	**208**	**1.57**	**17.4 days**
				4	**155**	**1.69**	**19.4 days**
				3	**103**	**1.92**	**23.6 days**

* If we assume 20% of enrolled patients will not be used for the analysis, divide this number by 0.8

Summary of numerical studies for designing safety studies

- 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

R package for study design using RMST

SSRMST (sample size calculation using RMST)
https://cran.r-project.org/web/packages/SSRMST/index.html

Description: The package calculates the study sample size and power in designing clinical trials using the differences in restricted mean survival times (RMST).

- Superiority test
- Non-inferiority test

Estimating long-term treatment effects

- RMST difference provide clinically useful information but it is restricted to the truncation point of τ
- Estimating long-term treatment effects is often difficult

Estimating long-term treatment effects

Several attempts

1. Using flexible parametric model and project the future survival curves

- Relies on the adequacy of the fitted model
- It is hard to validate the future projection is good

2. Life-table analysis (Claggett et al, 2016, NEJM)

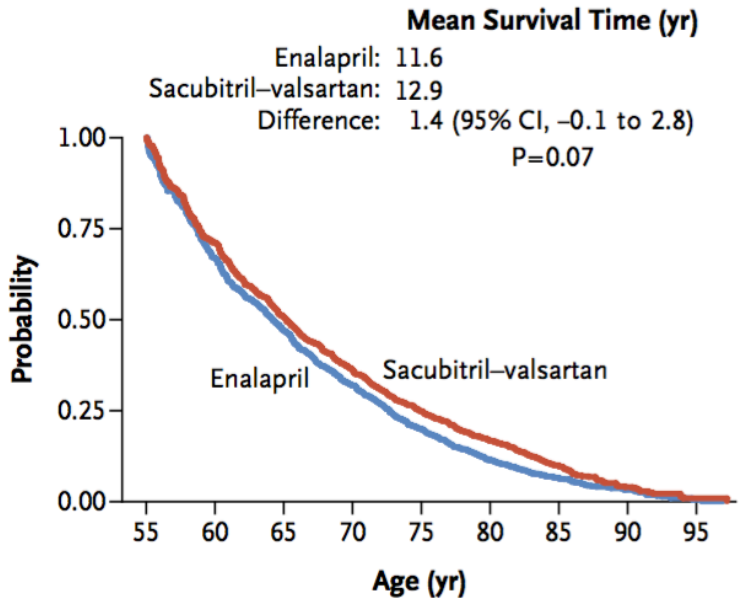
- Traditional analyses estimate the risk of death (event) in each treatment arm as a function of “time since randomization”
- Life-table analyses focus on “time since birth” (i.e. age)
- Estimate age-specific death rates (event rates) for each treatment group
- Create Kaplan-Meier curves using “Age” rather than “Time from Randomization” as the time scale
- Use the area between two KM curves to estimate average delay in onset of death (event) over remaining patient lifetime

2. Life-table analysis (Claggett et al, 2016, NEJM)

Estimating the Long-Term Treatment Benefits of Sacubitril–Valsartan

TO THE EDITOR: Although data from clinical trials can be used to estimate the effectiveness of a new therapy as compared with a control during the study follow-up, estimating long-term treatment effects is often difficult. The PARADIGM-HF trial showed that sacubitril–valsartan was superior

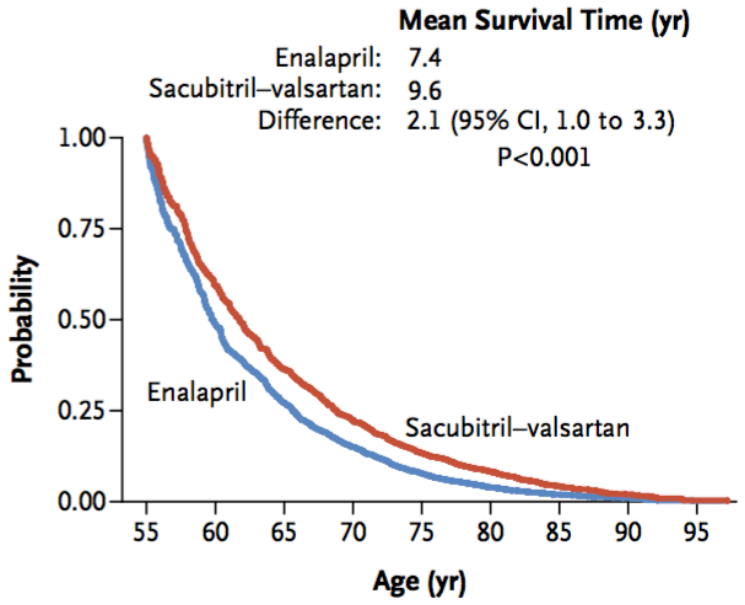
A Survival after Age 55 Yr



No. at Risk

Enalapril	158	280	388	274	284	210	80	14	1
Sacubitril–valsartan	178	276	343	267	270	208	73	15	0

B Freedom from Primary End Point after Age 55 Yr



No. at Risk

Enalapril	145	249	352	253	260	190	73	13	1
Sacubitril–valsartan	171	258	323	246	244	198	68	15	0

Interim Analysis with RMST

- Murray and Tsiatis (Biometrics, 1999) showed the independent increment structure of RMST when a common τ is used for all interim and final analyses

- When the τ changes across the planned analyses, simulation methods will be used to maintain the type I error

- Again, τ's should be pre-specified and should be clinically motivated

Reconstruction of individual-level data

- Guyot, P., Ades, A. E., Ouwens, M. J., & Welton, N. J. (2012). Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Medical Research Methodology*, *12*(1), 9.

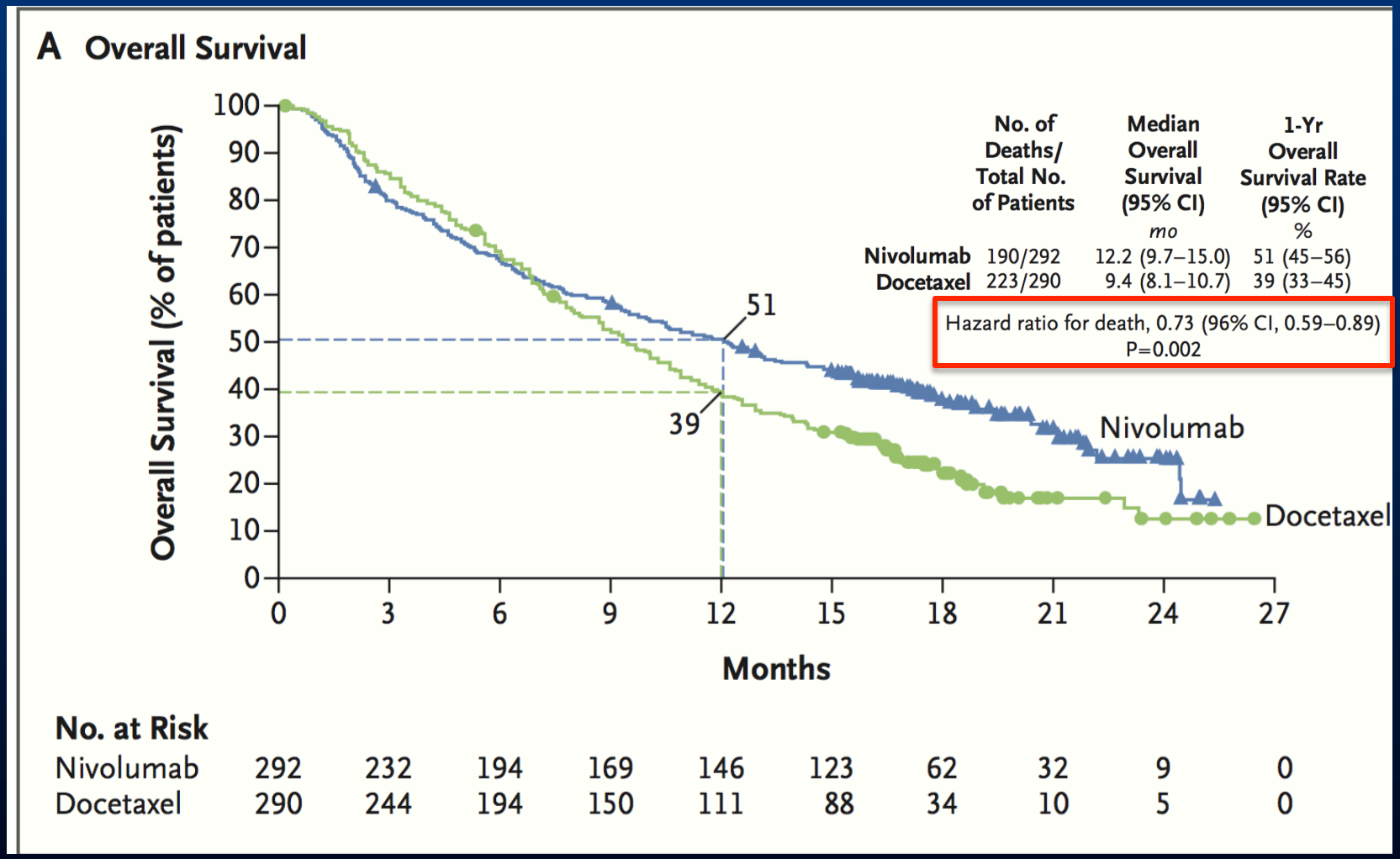
CheckMate 057 Study of Nivolumab

ORIGINAL ARTICLE

Nivolumab versus Docetaxel in Advanced Nonsquamous Non–Small-Cell Lung Cancer

H. Borghaei, L. Paz-Ares, L. Horn, D.R. Spigel, M. Steins, N.E. Ready, L.Q. Chow, E.E. Vokes, E. Felip, E. Holgado, F. Barlesi, M. Kohlhäufl, O. Arrieta, M.A. Burgio, J. Fayette, H. Lena, E. Poddubskaya, D.E. Gerber, S.N. Gettinger, C.M. Rudin, N. Rizvi, L. Crinò, G.R. Blumenschein, Jr., S.J. Antonia, C. Dorange, C.T. Harbison, F. Graf Finckenstein, and J.R. Brahmer

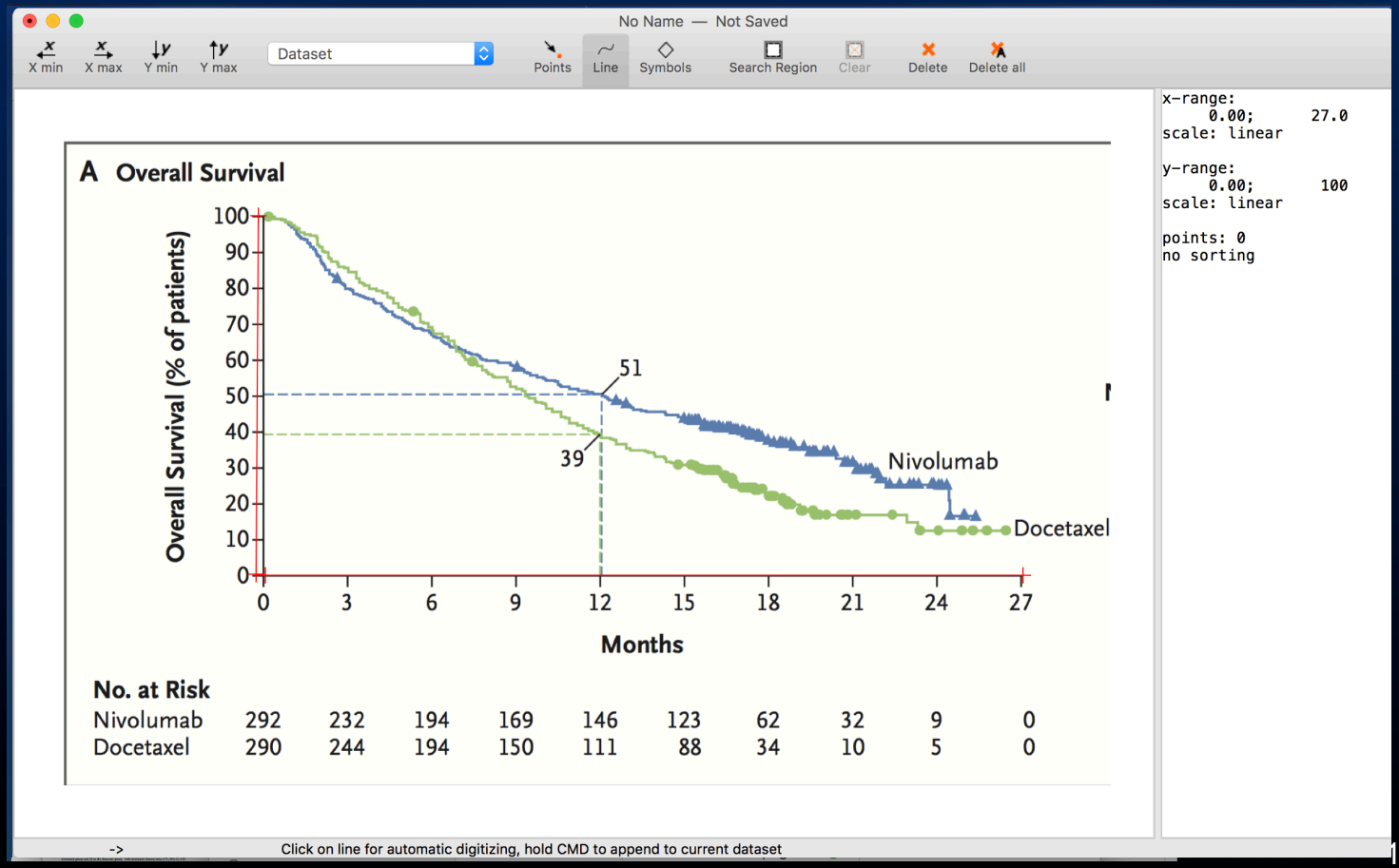
Borghaei et al. (2015, NEJM)



Borghaei et al. (2015, NEJM)

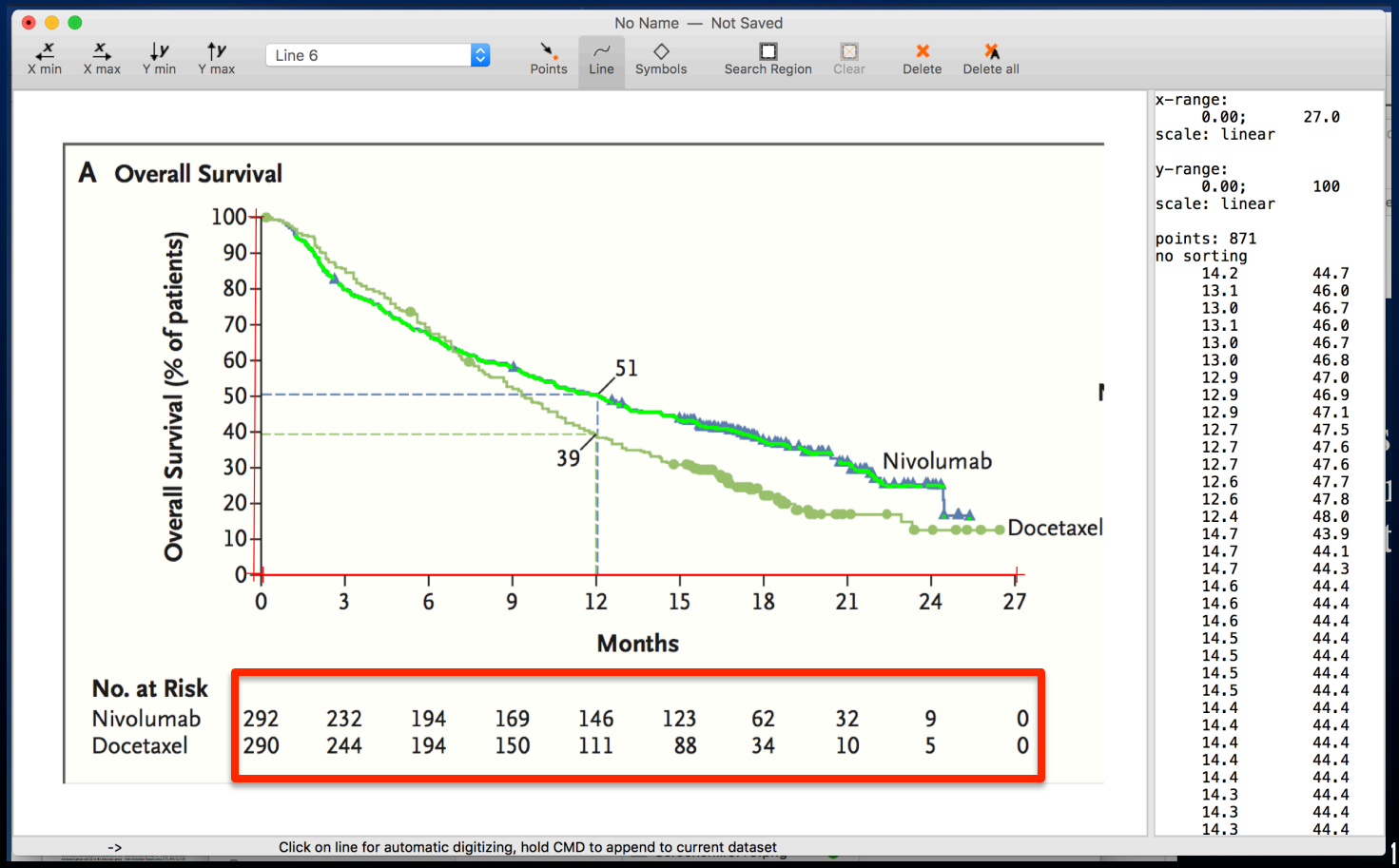
Ref.) How to reconstruct individual-level data from published KM curves

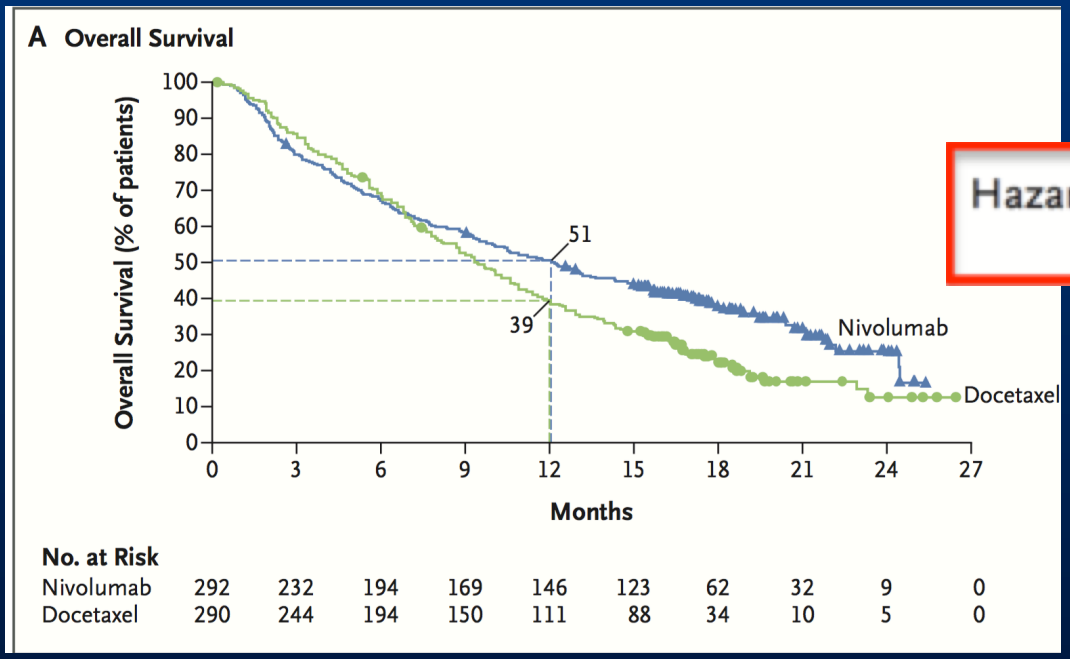
1. Scan the KM plot and digitize it using a digitizer software (eg. *DigitizeIt*)
2. Input the information of # of risk set
3. Run an algorithm to generate the individual-level data from these information



Ref.) How to reconstruct individual-level data from published KM curves

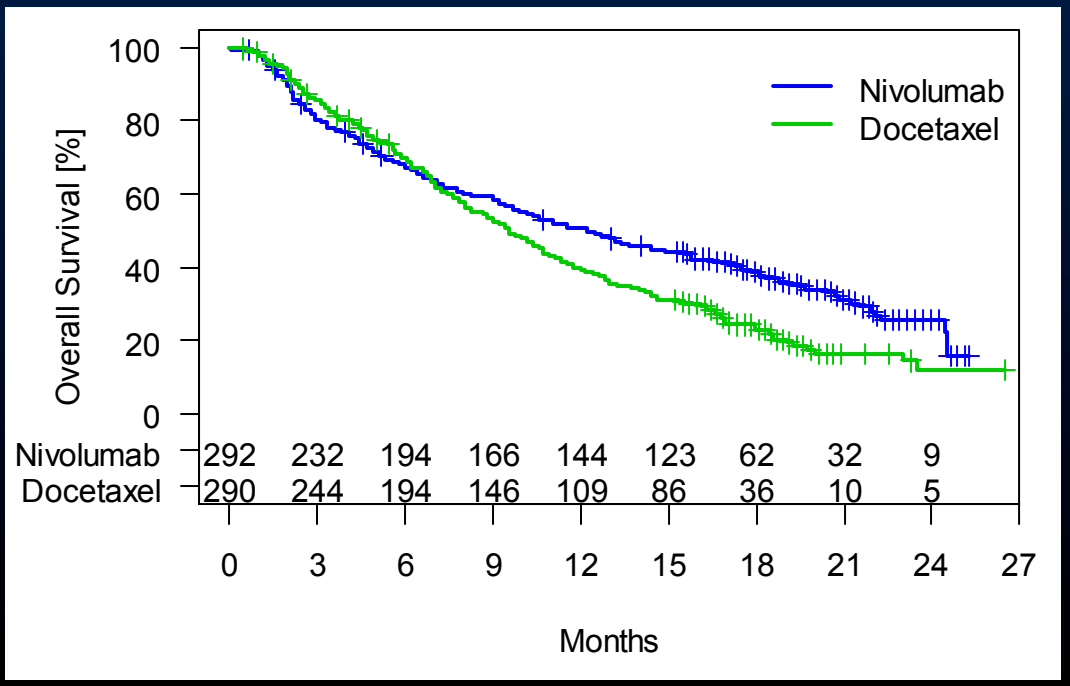
1. Scan the KM plot and digitize it using a digitizer software (eg. *DigitizeIt*)
2. Input the information of # of risk set
3. Run an algorithm to generate the individual-level data from these information





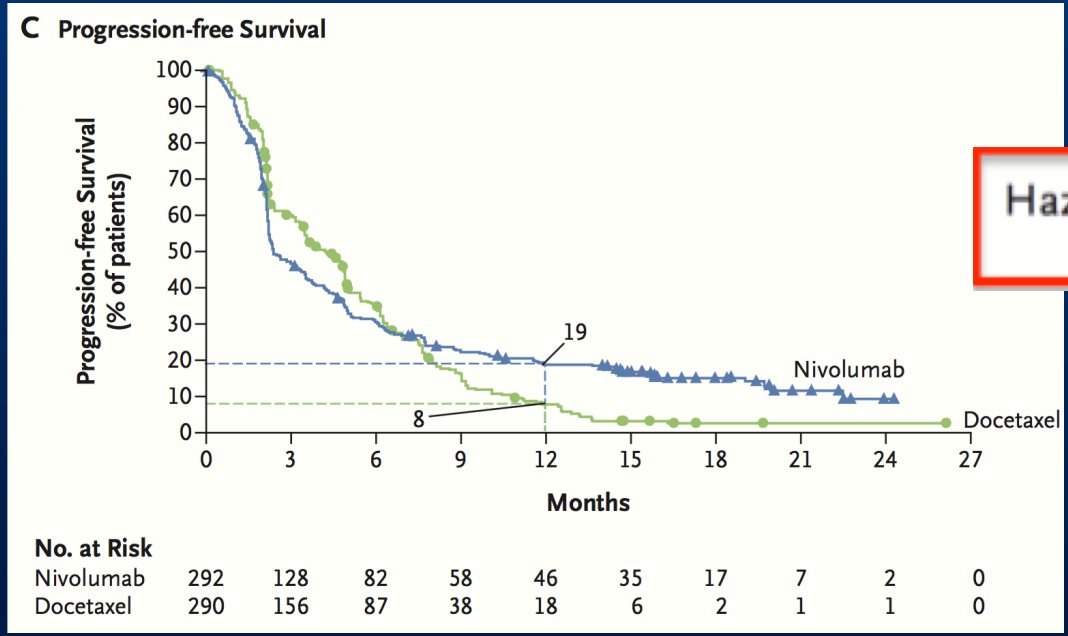
The original KM in the NEJM paper

Hazard ratio for death, 0.73 (96% CI, 0.59–0.89)
P=0.002



The KM drawn with the reconstructed data

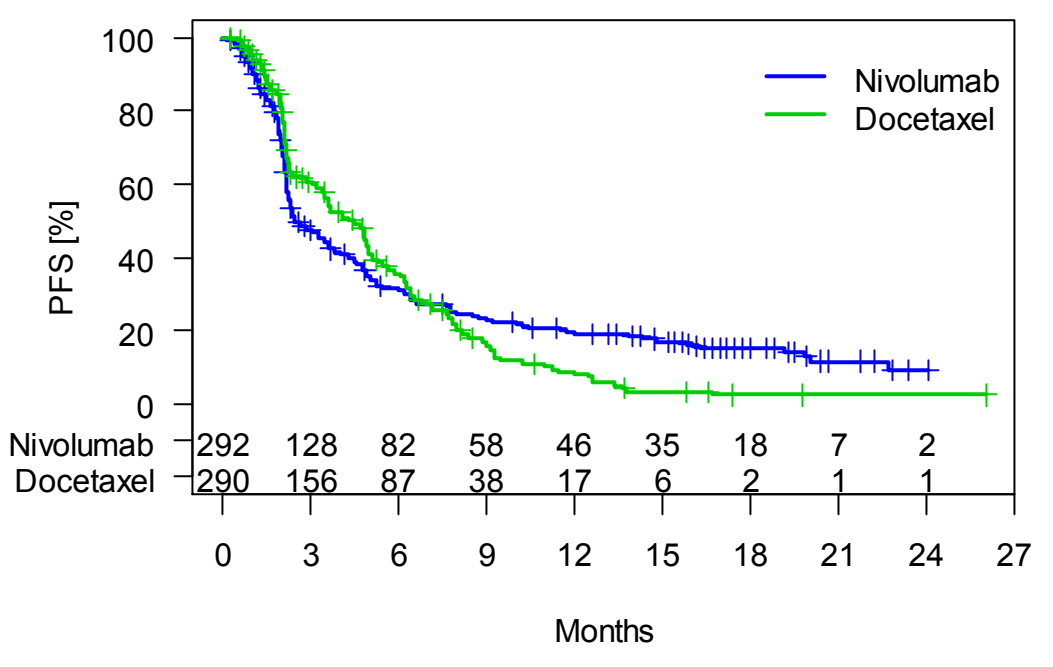
RMST 24m (0.95CI) [months]	
Nivolumab	13.0 (12.0 to 14.0)
Docetaxel	11.3 (10.9 to 12.2)
Difference	1.7 (0.4 to 3.1) **P=0.012**



The original KM in the NEJM paper

Hazard ratio for disease progression or death, 0.92 (95% CI, 0.77–1.11); P=0.39

NS



The KM drawn with the reconstructed data

RMST 24m (0.95CI) [months]	
Nivolumab	6.7 (5.8 to 7.6)
Docetaxel	5.4 (4.8 to 6.0)
Difference	1.3 (0.2 to 2.3) **P=0.021**

SIGNIFICANT!!

Alternatives to the hazard ratio in survival analysis – moving beyond the comfort zone (FDA, Aug 28, 2016)

H. Uno and L. J. Wei

Short Course Explores Survival Analysis | Department of Biostatistics | Harvard T.H. Chan School of Public Health

Short Course Explores Survival Analysis The primary goal for conducting clinical studies is to obtain robust, clinically interpretable results with respect to...

HSPH.HARVARD.EDU

Statistical Test & Estimation

Statistical test

- Gives an answer

SIGNIFICANT ($p<0.05$) or NS ($p>0.05$)

- It nicely fits situations where we need to make a decision

ASA News
AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®
732 North Washington Street, Alexandria, VA 22314 • (703) 684-1221 • Toll Free: (888) 231-3473 • www.amstat.org • www.twitter.com/AmstatNews

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

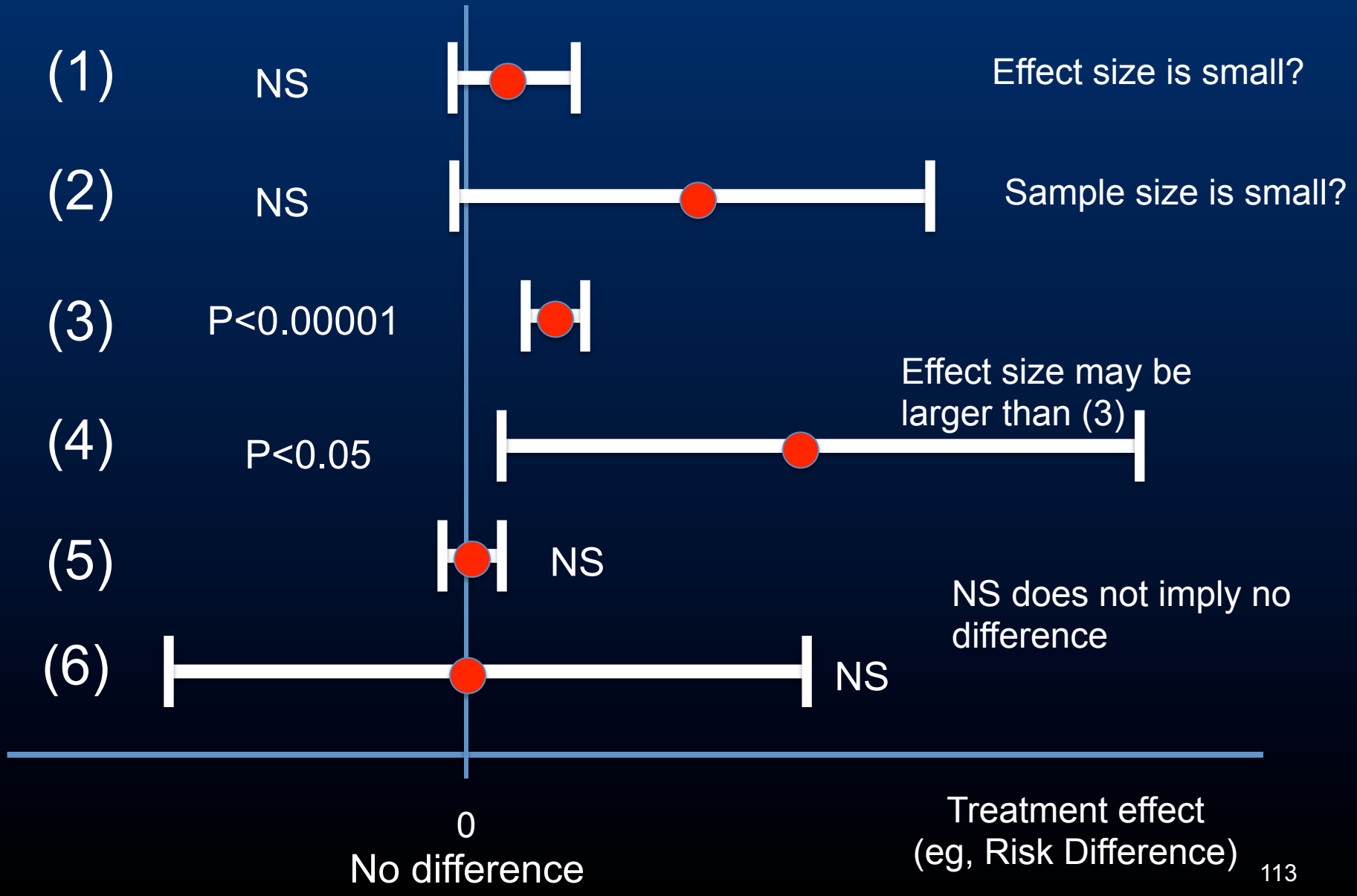
March 7, 2016

The American Statistical Association (ASA) has released a "Statement on Statistical Significance and *P*-Values" with six principles underlying the proper use and interpretation of the *p*-value [http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XIOaE2MN]. The ASA releases this guidance on *p*-values to improve the conduct and interpretation of quantitative science and inform the growing emphasis on reproducibility of science research. The statement also notes that the increased quantification of scientific research and a proliferation of large, complex data sets has expanded the scope for statistics and the importance of appropriately chosen techniques, properly conducted analyses, and correct interpretation.

The statement's six principles

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Test results and 0.95CI



Test vs. Estimation

- Clearly, CI (i.e., estimation) is more informative than p-value (i.e., testing)

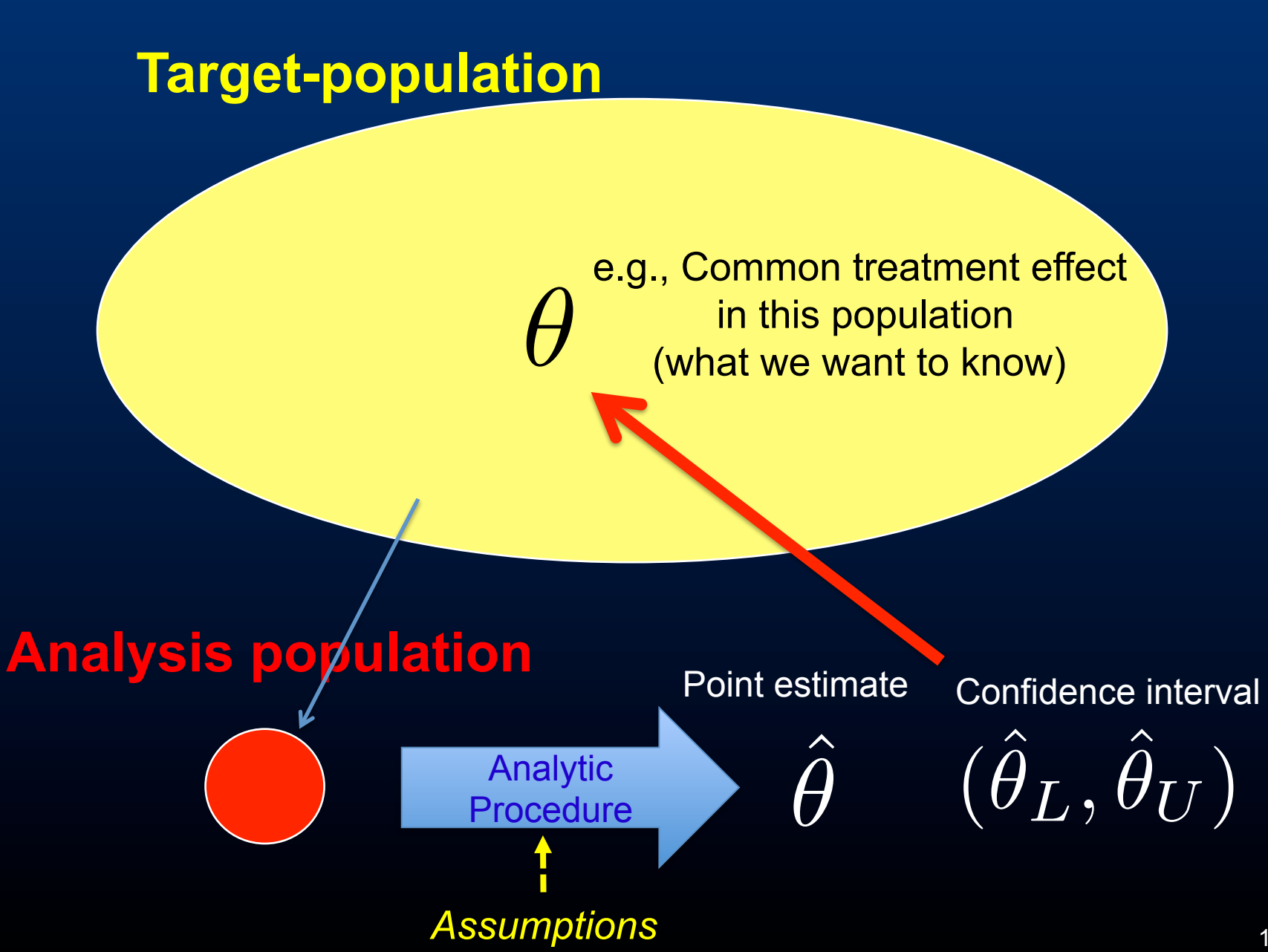
- However, generally, estimation is more challenging than testing…

Example: Hazard ratio (HR)

Test: If the treatment assignment is randomized, testing the null hypothesis (no treatment effect; HR = 1) is **valid** even if the modeling assumption is wrong

- Non proportional hazard
- Covariate omission in the Cox model
- Wrong functional form of covariates

Estimation: The correct model specification is the basis of the valid estimation of HR


Target-population
θ
e.g., Common treatment effect
in this population
(what we want to know)
Analysis population
Analytic
Procedure
Assumptions
Point estimate
$\hat{\theta}$
Confidence interval
$(\hat{\theta}_L, \hat{\theta}_U)$

What would be important for estimates to be clinically useful information?

1. **Generalizability:** the target population should be clear, and the estimate should be generalizable to the target population

2. **Validity:** i.e., the estimate should be estimating what we want to estimate

3. **Interpretability:** i.e., we should be able to assess if the estimated treatment effect is a clinically meaningful difference or not