On testing based on restricted mean survival time for time-to-event outcomes

Hajime Uno, PhD
Dana-Farber Cancer Institute
Harvard Medical School

Joint work with
Miki Horiguchi
(Kitasato University, Tokyo)

Annual Conference of the ISCB at Vigo, Spain (Jul 10, 2017)

Setting

Consider

- Randomized clinical trial to compare two groups
- time-to-event outcome is the primary endpoint

What we need to do…

- Testing equality of the two event time distributions (Test)

  ex. "The survival benefit of this drug is highly significant!!"

- Estimating magnitude of the treatment effect (0.95CI for a primary summary measure)

Important information for decision-making

Test-estimation coherency

Testing		Estimation
Test result is SIG	⟺	CI of the treatment effect excludes the null value (e.g, HR=1)
Test result is NS	⟺	Includes the null value

Example of test-estimation coherency

Test: Logrank test

Estimation of the treatment effect:

Cox's estimator (Hazard ratio)

Outline

- Issues of the conventional practices
- Alternatives and challenge
- Restricted mean survival time (RMST) based versatile test
- Application
- Numerical studies
- Conclusions

A conventional practice

Description ➡ Test ➡ Estimation of treatment effect

Kaplan-Meier ➡ Logrank test ➡ HR by Cox PH

A conventional practice

"logrank-HR" test-estimation practice

Logrank test is

- the most powerful nonparametric test in PH
- equivalent to the score test for testing HR equals 1 via the Cox's model (test-estimation coherency)

However, the proportional hazards (PH) assumption does not always hold in practice, which raises some problems

What if non-PH?

Test: Logrank is not optimal anymore (may perform rather poorly, e.g., cross-hazard cases)

Estimation of treatment effect:

- Interpretation of HR is not obvious as a quantitative summary of the treatment effect
  - HR is not a simple average of the hazard ratio over time
  - HR depends on underlying study-specific censoring distributions (or follow-up time)
- Thus, it may not be useful as the primary summary of the treatment effect for decision making

A conventional practice when non-PH

Test

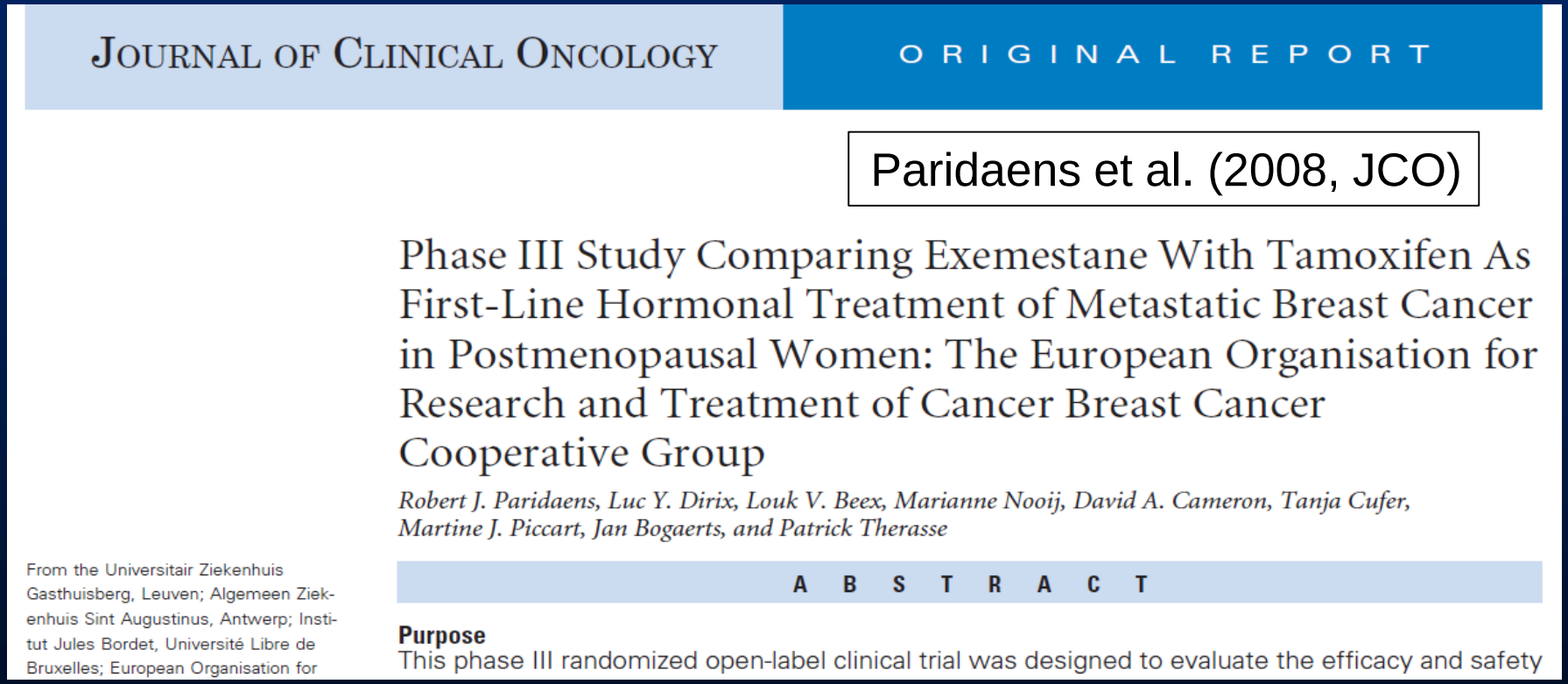
- Another member of the weighted logrank tests (e.g., Wilcoxon test) is often chosen

Estimation of treatment effect

- Difference in median survival time (without CI) is often reported as the quantitative information of treatment effect

This conventional practice can also be problematic sometimes!

Metastatic breast cancer example



JOURNAL OF CLINICAL ONCOLOGY

ORIGINAL REPORT

Paridaens et al. (2008, JCO)

Phase III Study Comparing Exemestane With Tamoxifen As First-Line Hormonal Treatment of Metastatic Breast Cancer in Postmenopausal Women: The European Organisation for Research and Treatment of Cancer Breast Cancer Cooperative Group

Robert J. Paridaens, Luc Y. Dirix, Louk V. Beex, Marianne Nooij, David A. Cameron, Tanja Cufer, Martine J. Piccart, Jan Bogaerts, and Patrick Therasse

From the Universitair Ziekenhuis Gasthuisberg, Leuven; Algemeen Ziekenhuis Sint Augustinus, Antwerp; Institut Jules Bordet, Université Libre de Bruxelles; European Organisation for

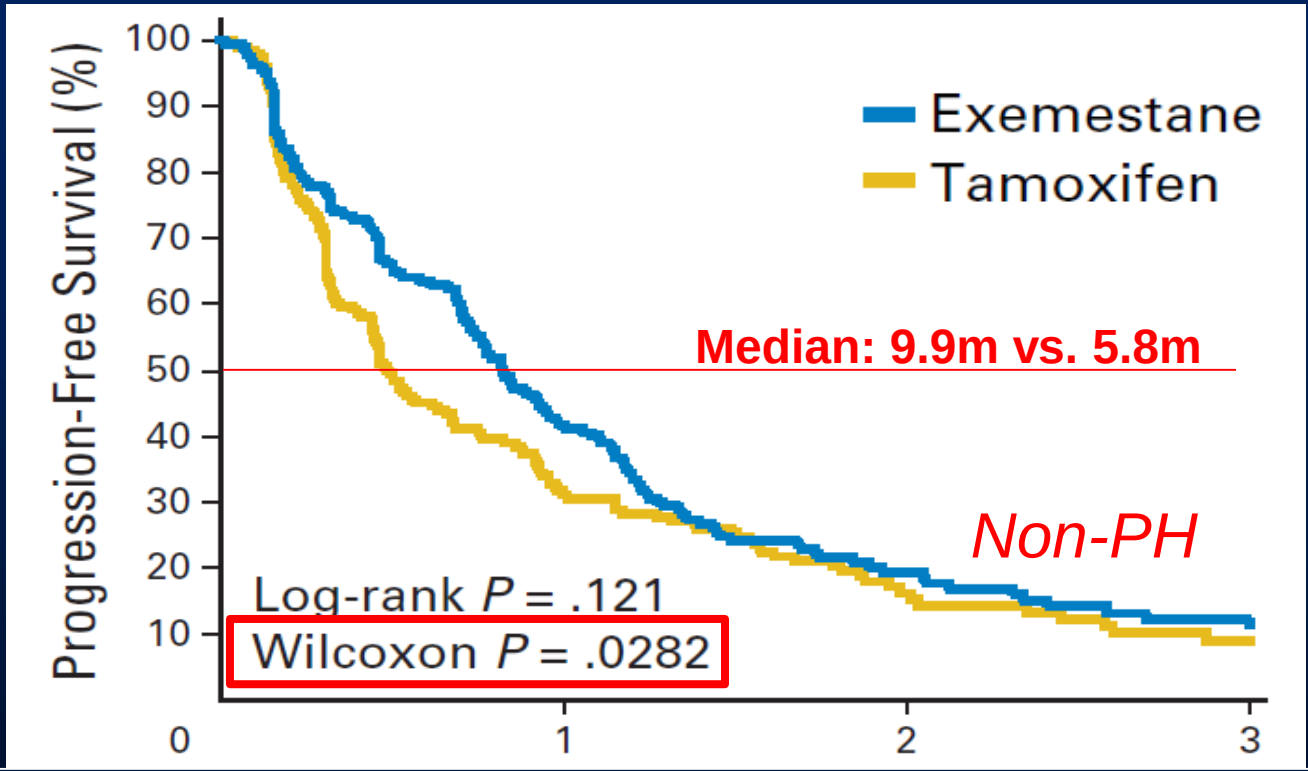
A B S T R A C T

Purpose

This phase III randomized open-label clinical trial was designed to evaluate the efficacy and safety

- N=371 (182 on exemestane, 189 on tamoxifen)
- The primary endpoint was PFS

Metastatic breast cancer example



"The 4.1-month difference in median PFS between treatment arms was statistically significant using the Wilcoxon test (P = .028)" Paridaens et al. (2008, JCO)

This is not a correct statement statistically

How does the conventional design and analysis fail?

Metastatic breast cancer example

Paridaens et al. (2008, JCO)

	Design stage
median PFS	10.0 mon 7.14 mon
Pattern of difference	PH
HR	0.78
Number of events	278
sample size	342
The primary test	logrank

Metastatic breast cancer example

Paridaens et al. (2008, JCO)

	Design stage	**It turned out**
median PFS	10.0 mon 7.14 mon	9.9 mon 5.8 mon
Pattern of difference	PH	**Non-PH**
HR	0.78	**0.84 ?**
Number of events	278	319
sample size	342	371
The primary test	logrank	**p=0.121**

Why did this happen?

Failed to guess the pattern of the difference

- Power depends on the underlying true difference between the two survival functions
- No or little information is usually available at design stage

This is a common challenge in many clinical trials

Solution and problem

- Versatile tests can capture various patterns of difference between two survival curves

  Examples:
  - Max/liner comb of weighted logrank tests (e.g., Tarone (1981))
  - Adaptively weighted logrank test (Yang and Prentice, 2010)
  - Adaptively weighted KM-based test (Uno et al. 2015)
  - Min of logrank p-val and RMST permutation test p-val (Royston and Parmer 2016)

Solution and problem

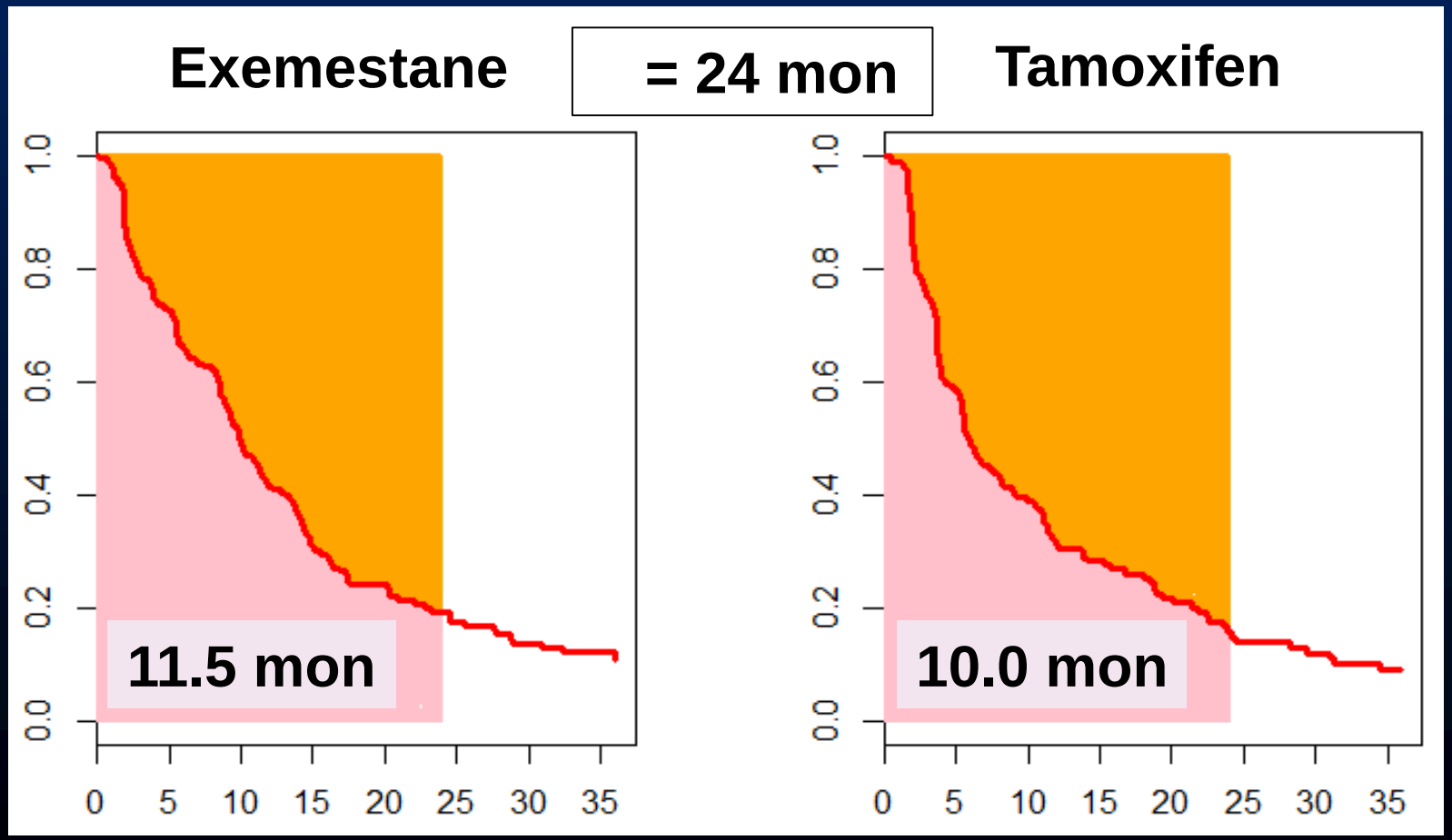
- However, a problem with most of the versatile tests is that test-estimation coherency will be a challenge

Proposal

RMST-based versatile test

Restricted Mean Survival Time (RMST)

$= E_x min_{E,Fz} =$



Difference in RMST

$$= \quad -$$

robust, clinically interpretable, and model-free summary measure

Standard RMST-based test

- using a pre-specified fixed
- based on the test statistic, $( ) = ( )/ ( )$

$( )$: standard error estimate of $( )$

Proposed RMST-based versatile test

- Uses data-dependent
  can detect various patterns of the difference

- Has companion quantification procedures to provide a corresponding, robust, clinically interpretable summary of the treatment effect
  test-estimation coherency

Details of the proposed test

i. Instead of choosing a fixed , we consider a set of , = { , , …, }

ii. For each , we can calculate the test statistic, E F = ( )/ ( )

iii. The test statistic is then obtained as

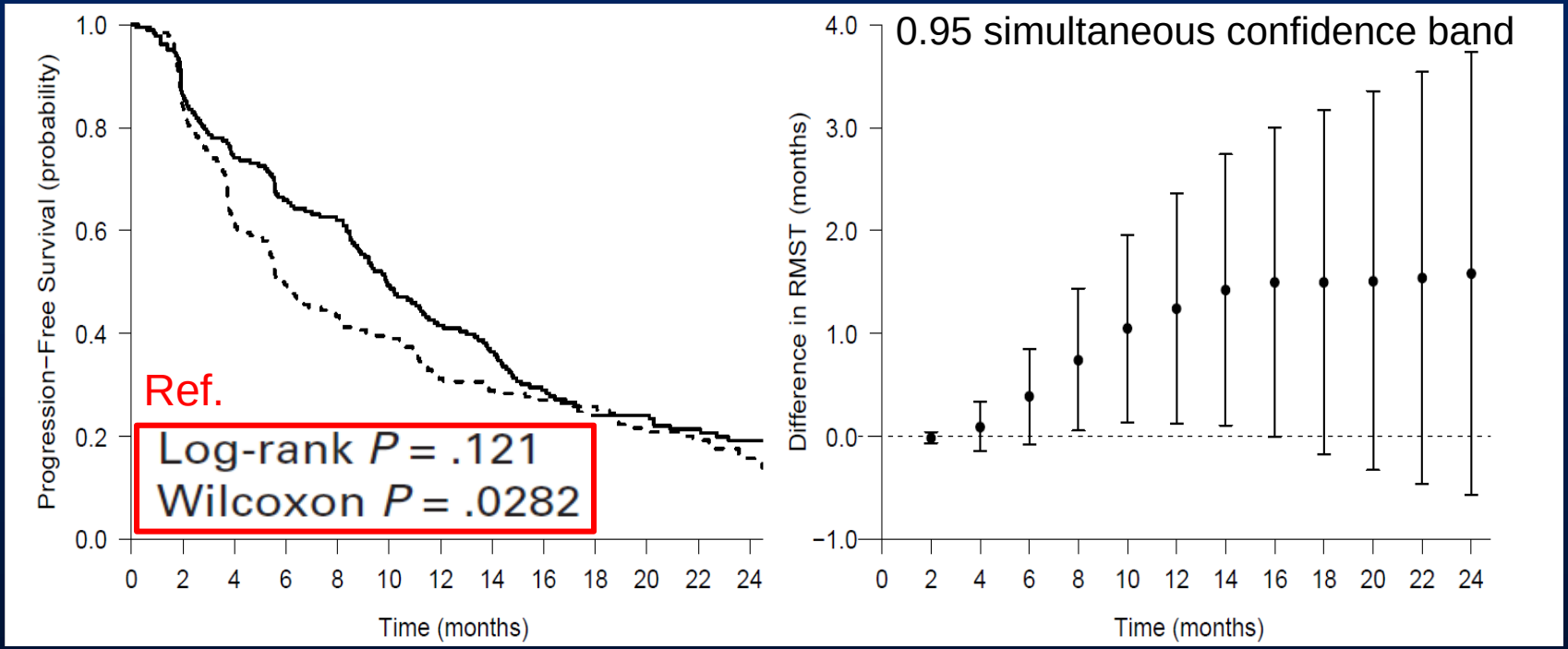
= max ( ) (one-sided)

= max| E F| (two-sided)

iv. The null distribution of can be derived via a perturbation-resampling procedure

Application

Metastatic breast cancer example



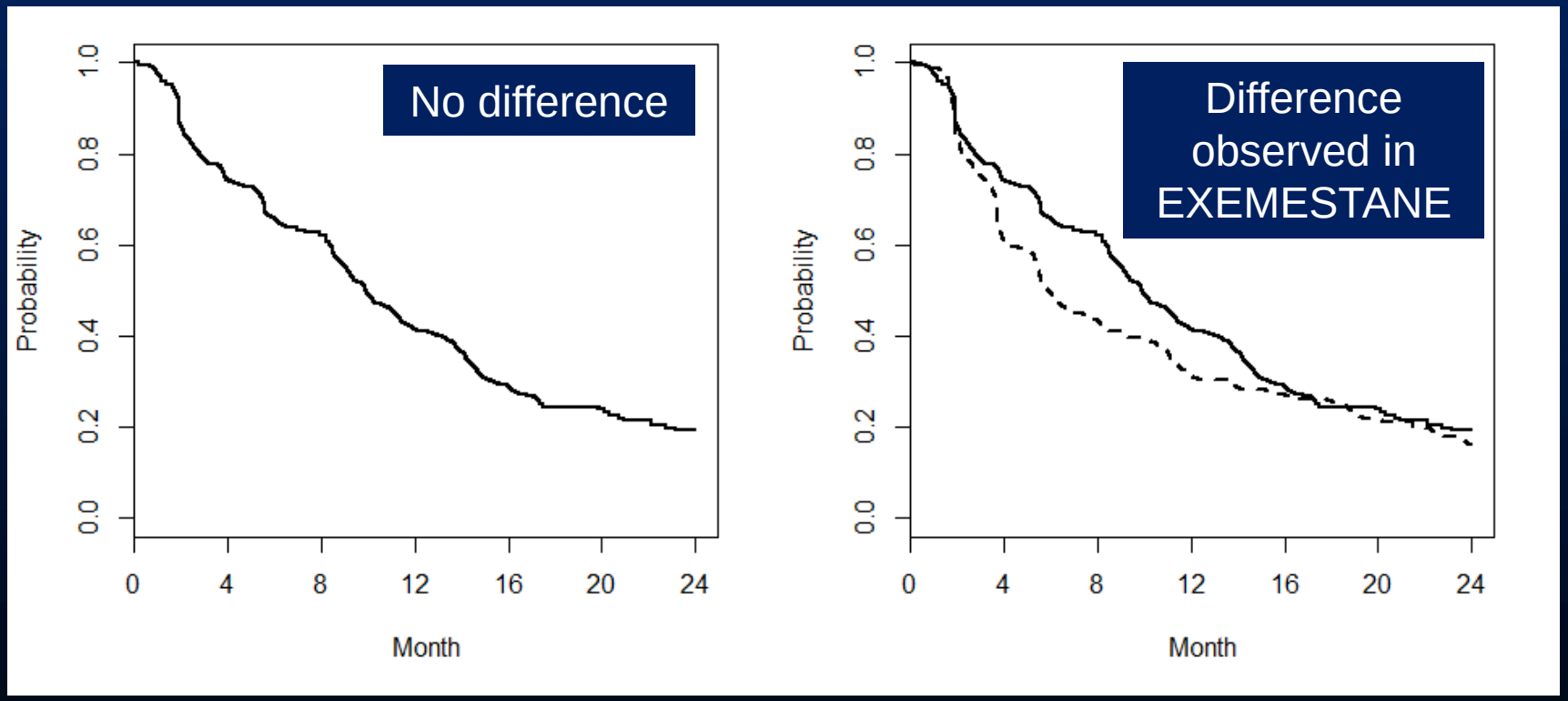
RMST-based versatile test	
	{2, 4, 6, …, 24} mon
Selected	10 mon
P-value	**P=0.010**
Difference (10 mon)	1.05 (0.95CI, 0.14 to 1.96) mon

Numerical studies

1. Metastatic breast cancer example
2. Another cancer example (RAINBOW study)

Metastatic breast cancer example

Event time distribution



Metastatic breast cancer example

Other parameters

- Test: **two-sided**
- Total study time: **24 months**
- Time points for : = *{ , , , ... , }* (months)
- Sample size: **N = 300 (per arm)**
- The number of the perturbation-resampling: **5000**
- The number of iterations: **2000**
- Comparisons:
  - **Log-rank test**
  - **Peto-Prentice-Wilcoxon test**
  - **Standard RMST-based test (= months)**

Results

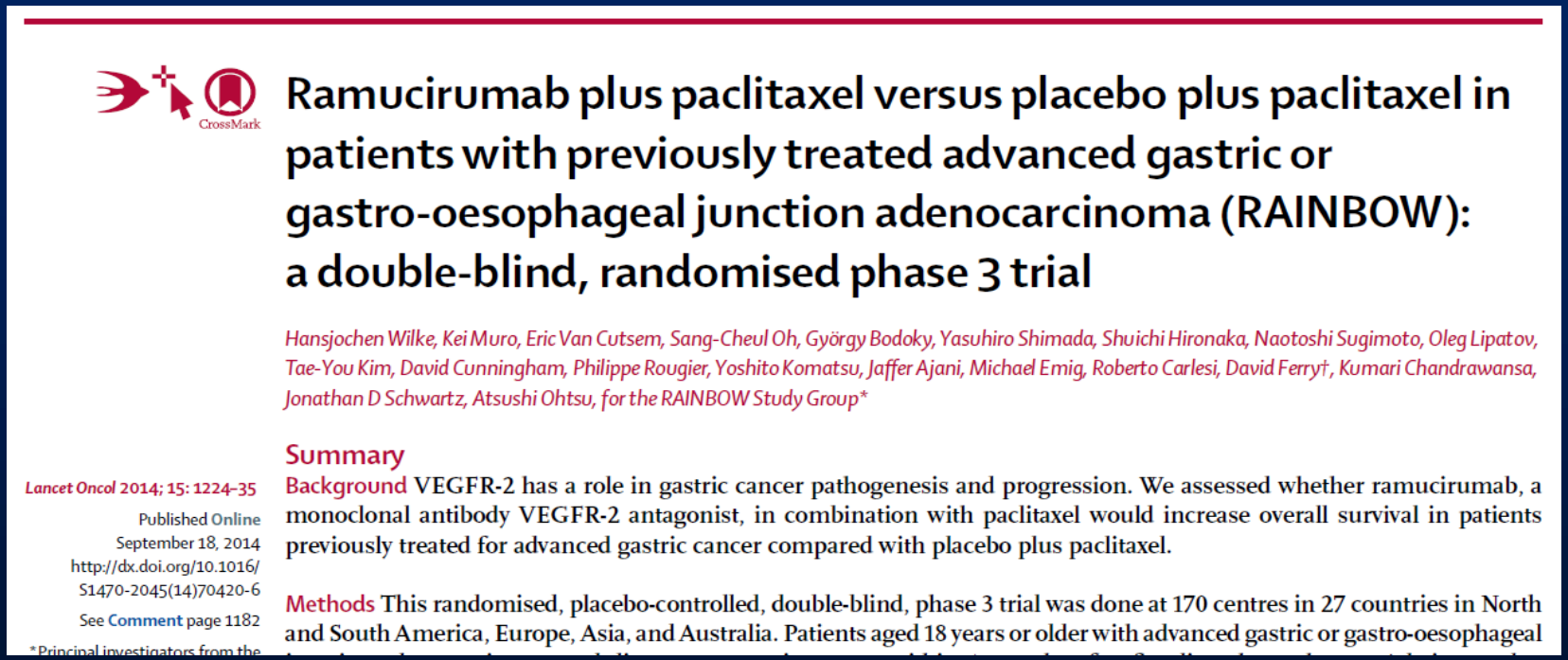
Size (nominal 0.05)

Test	No censoring	Light censoring	Moderate censoring	EXEM. censoring
Logrank	0.048	0.052	0.055	0.054
Peto-Prentice Wilcoxon	0.054	0.050	0.052	0.048
Standard RMST	0.053	0.053	0.051	0.055
Versatile RMST	0.056	0.048	0.047	0.051

Power

Test	No censoring	Light censoring	Moderate censoring	EXEM. censoring
Logrank	0.568	0.572	0.587	0.560
Peto-Prentice Wilcoxon	0.809	0.815	0.812	0.820
Standard RMST	0.649	0.644	0.624	0.652
Versatile RMST	0.932	0.930	0.926	0.934

Another cancer example (RAINBOW study)



Ramucirumab plus paclitaxel versus placebo plus paclitaxel in patients with previously treated advanced gastric or gastro-oesophageal junction adenocarcinoma (RAINBOW): a double-blind, randomised phase 3 trial

*Hansjochen Wilke, Kei Muro, Eric Van Cutsem, Sang-Cheul Oh, György Bodoky, Yasuhiro Shimada, Shuichi Hironaka, Naotoshi Sugimoto, Oleg Lipatov, Tae-You Kim, David Cunningham, Philippe Rougier, Yoshito Komatsu, Jaffer Ajani, Michael Emig, Roberto Carlesi, David Ferry†, Kumari Chandrawansa, Jonathan D Schwartz, Atsushi Ohtsu, for the RAINBOW Study Group**

Lancet Oncol 2014; 15: 1224–35
Published Online
September 18, 2014
http://dx.doi.org/10.1016/S1470-2045(14)70420-6
See Comment page 1182

Summary

Background VEGFR-2 has a role in gastric cancer pathogenesis and progression. We assessed whether ramucirumab, a monoclonal antibody VEGFR-2 antagonist, in combination with paclitaxel would increase overall survival in patients previously treated for advanced gastric cancer compared with placebo plus paclitaxel.

Methods This randomised, placebo-controlled, double-blind, phase 3 trial was done at 170 centres in 27 countries in North and South America, Europe, Asia, and Australia. Patients aged 18 years or older with advanced gastric or gastro-oesophageal

Wilke et al. (2014, Lancet)

- A phase III randomized trial to compare ramucirumab + paclitaxel and placebo + paclitaxel for advanced gastric cancer.
- N=665 (330 on ramucirumab, 335 on placebo)

Overall Survival

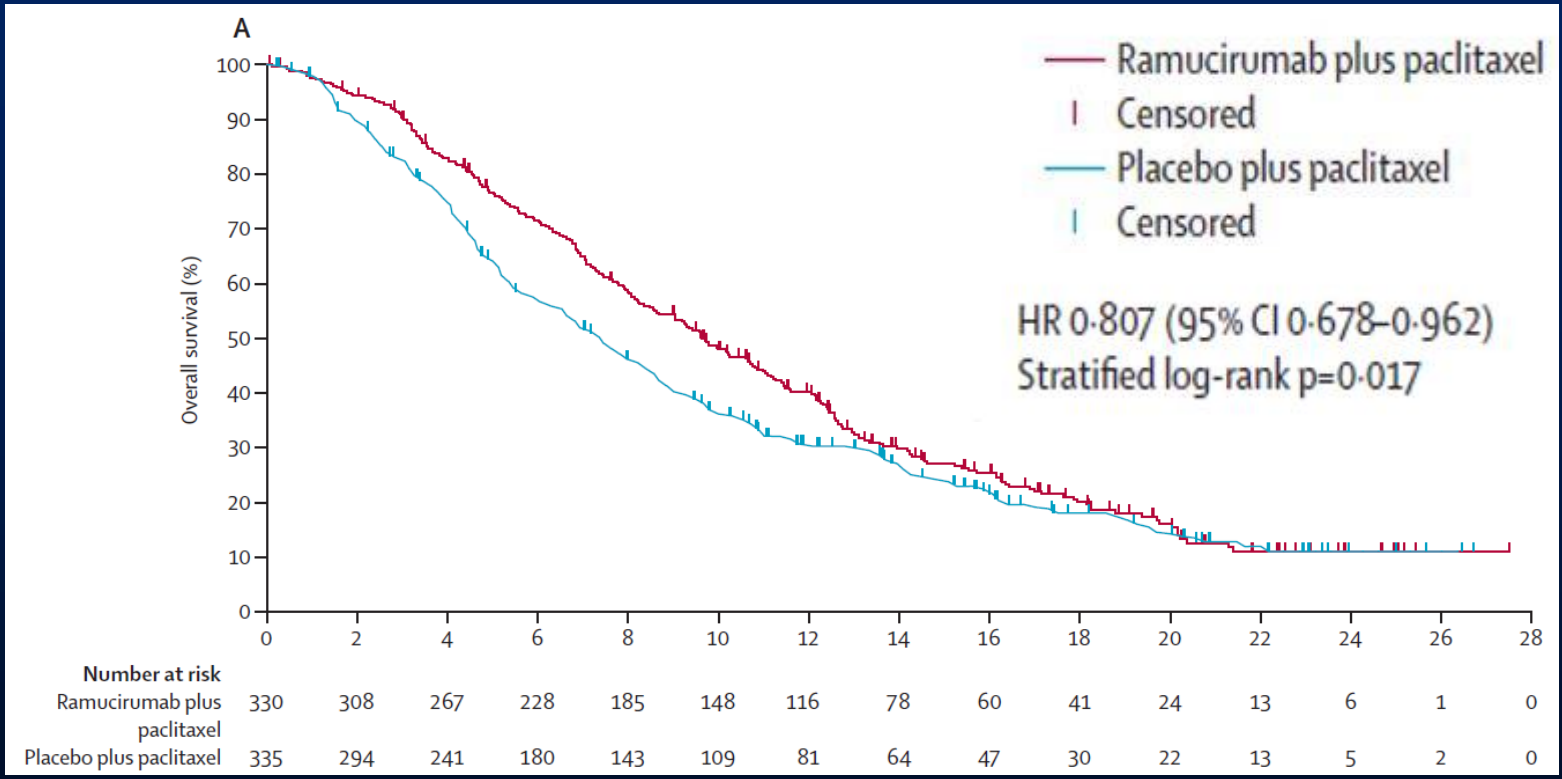


Figure 2, Wilke et al. (2014, Lancet)

The KM curves suggest that the PH assumption does not hold (the cumulative residual test, p=0.002)

RAINBOW study

Other parameters

- Test: **two-sided**
- Total study time: **21 months**
- Time points for : = *{ , , , , , , }* (months)
- Sample size: **N = 300 (per arm)**
- The number of the perturbation-resampling: **5000**
- The number of iterations: **2000**
- Comparisons:
  - **Log-rank test**
  - **Peto-Prentice-Wilcoxon test**
  - **Standard RMST-based test (= months)**

RAINBOW study
Results

Power

Test	No censoring	Light censoring	Moderate censoring	RAINBOW censoring
Logrank	0.435	0.467	0.497	0.586
Peto-Prentice Wilcoxon	0.870	0.875	0.872	0.893
Standard RMST	0.746	0.746	0.722	0.736
Versatile RMST	0.940	0.945	0.940	0.945

Summary of numerical studies

- The empirical significance levels of all tests are close to their nominal value of 0.05

- Although the proposed test is inferior to the logrank test under PH alternatives by theory, the proposed test is dramatically powerful for the pattern of the difference seen in those cancer clinical trials

Conclusions

Conclusions

- Several practical issues arise from the conventional design and analysis using the "logrank-HR" test-estimation practice
- The RMST-based versatile test is
  - a model-free, pre-specified test, and
  - dramatically powerful for patterns of difference seen in some recent cancer clinical trials
- It also provides corresponding robust and interpretable quantitative information of the treatment effect (test-estimation coherency)

END

References

Uno H, Claggett B, Tian L, et al. Moving Beyond the Hazard Ratio in Quantifying the Between-Group Difference in Survival Analysis. J Clin Oncol 2014; 32: 2380-2385.

Uno H, Wittes J, Fu H, et al. Alternatives to Hazard Ratio for Comparing the Efficacy or Safety of Therapies in Noninferiority Studies. Ann Intern Med Jul 2015; 163(2): 127-34.

Miller. Survival Analysis. Wiley 1981.

Tian et al. Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. Biostatistics 2014; 15: 222-233.

Uno et al. A versatile test for equality of two survival functions based on weighted differences of Kaplan–Meier curves. Statist. Med. 2015**;** 34: 3680-3695.

Royston P, Parmar MKB. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. Stat Med 2011; 30: 2409-2421.

Royston P, Parmar MKB. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Med Res Methodol 2013; 13: 152.

Royston P, Parmar MKB. Augmenting the logrank test in the design of clinical trials in which non-proportional hazards of the treatment effect may be anticipated. BMC Med Res Methodol 2016; 16: 16.

References

Guyot P, Ades AE, Ouwens MJ, et al. Enhanced secondary analysis of survival data: Reconstructing the data from published Kaplan-Meier survival curves. BMC Med Res Methodol 2012; 12: 9.

Rajkumar et al. 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. Lancet Oncol 2010; 11: 29–37.

Paridaens et al. Phase III Study Comparing Exemestane With Tamoxifen As First-Line Hormonal Treatment of Metastatic Breast Cancer in Postmenopausal Women: The European Organisation for Research and Treatment of Cancer Breast Cancer Cooperative Group. JCO 2008; 26: 4883-4890.

Wilke et al. Ramucirumab plus paclitaxel versus placebo plus paclitaxel in patients with previously treated advanced gastric or gastro-oesophageal junction adenocarcinoma (RAINBOW): a double-blind, randomised phase 3 trial. *Lancet Oncol* 2014; 15: 1224–35.

Zhao et al. On the Restricted Mean Survival Time Curve in Survival Analysis. Biometrics 2016; 72: 215–221.