The hazards of hazard ratios

L.J. Wei, Harvard University

Many thanks to H. Uno, Lu Tian, B. Claggett, Dae Kim, Lihui Zhao,
Bo Huang, T. Cai, Zack McCaw, Ray Sun

I did not create this provocative title

- MA Hernan (Epidemiology, 2010)
- Alexander et al (NEJM, 2018)

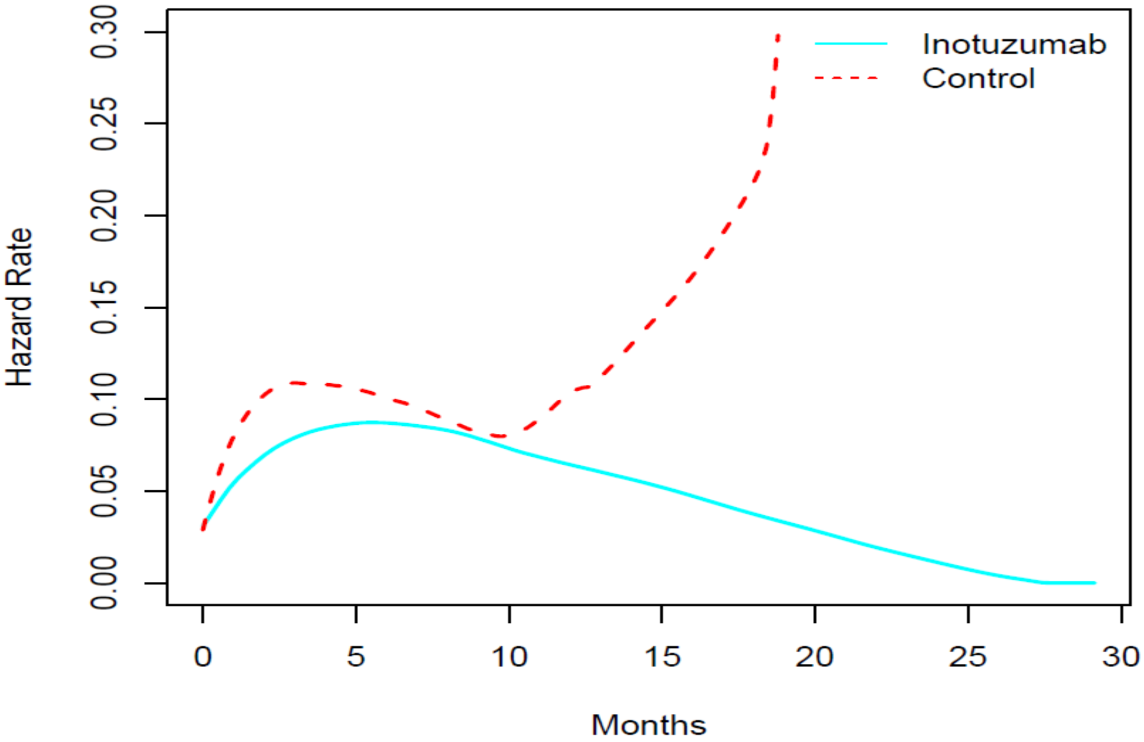
Time to a clinical event as the endpoint for a single group

- How to empirically summarize the “survival” (event-free) time profile for each treatment group?
- Kaplan-Meier (cumulative incidence curve)
- Event rate (at a specific time point)
- Median “survival” time (may not be observable)
- Hazard curve (hard to estimate well nonparametrically)

- Restricted mean survival time (or t-year mean survival time), RMST

An example of hazard rate curve or function

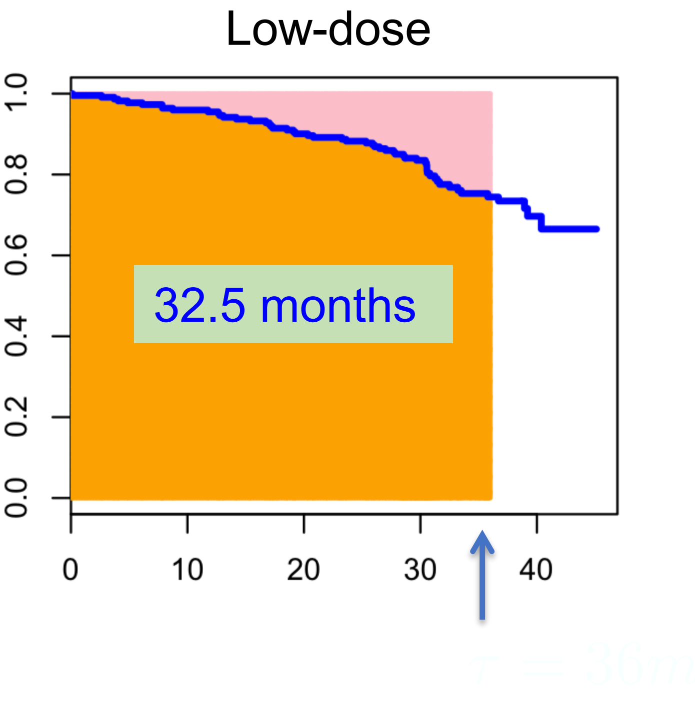
Study in acute lymphoblastic leukemia comparing inotuzumab with chemotherapy (NEJM, 2016)

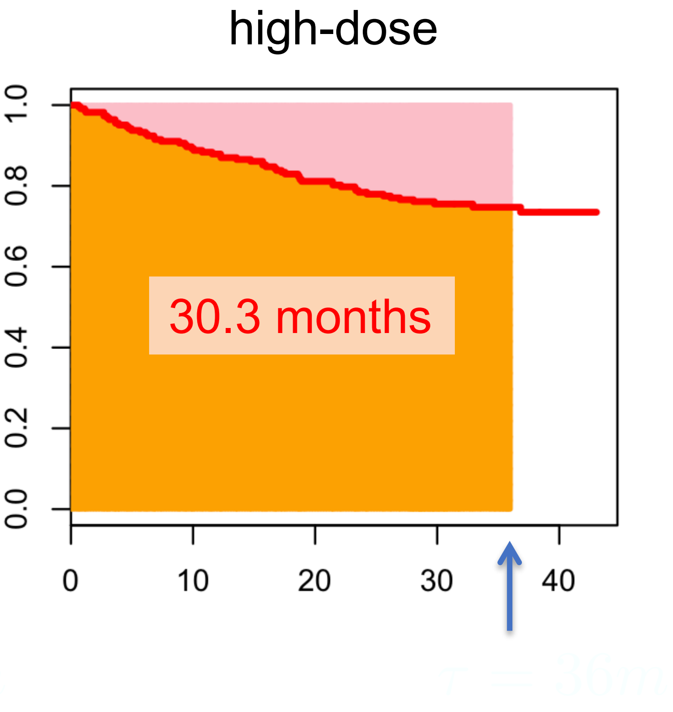


An example of t-year mean survival time or restricted mean survival time (RMST)

Restricted mean survival time (RMST):

2.2 months; Conf interval (0.5, 4.0), p=0.014





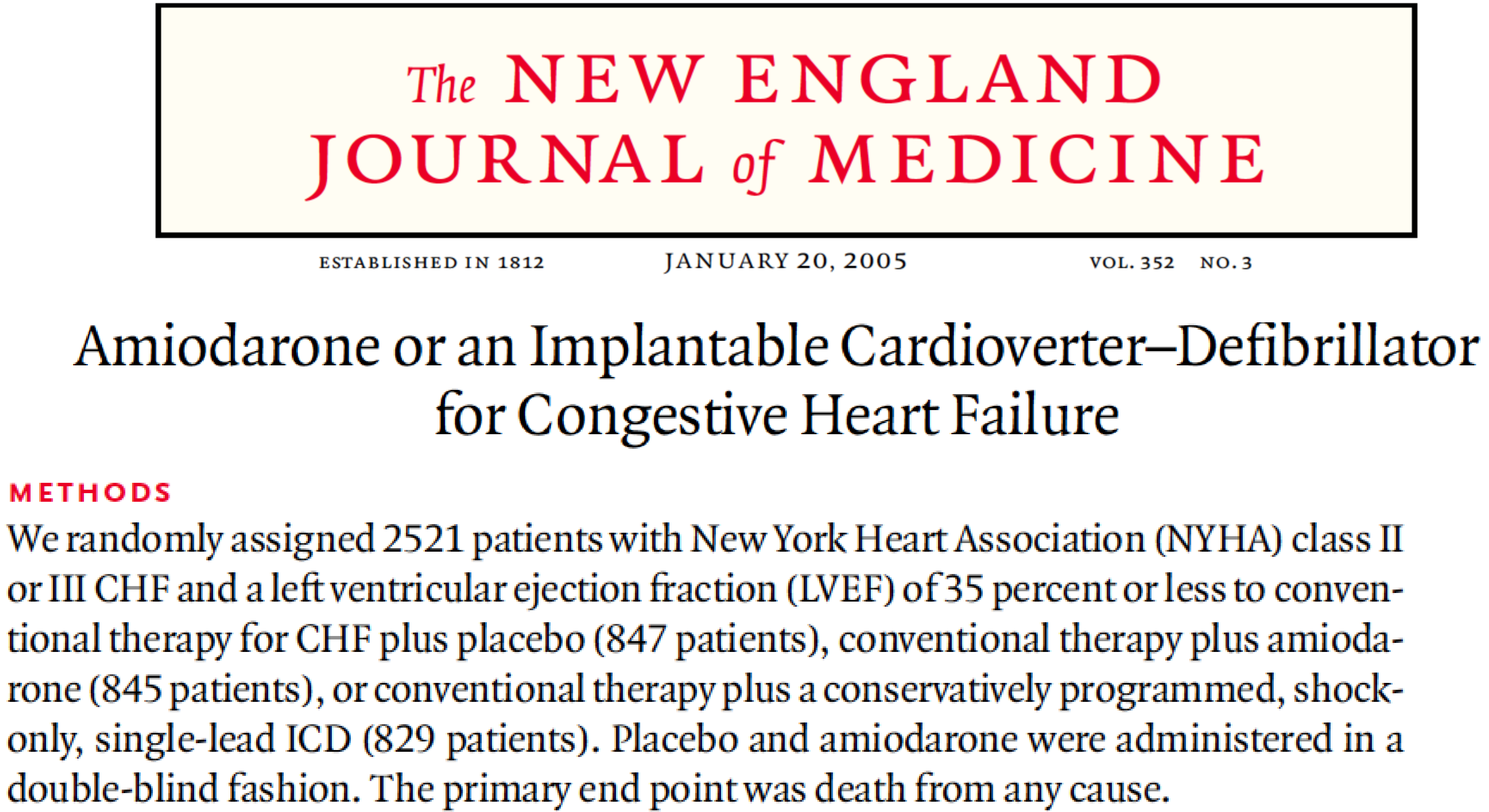
Metrics for quantifying the group difference

- Event rate difference (or ratio)
- Difference of two median failure times
- Hazard ratio (routinely used in practice)
- Difference (ratio) between two RMSTs.

- (Moving beyond p-value, consider an “estimand”)
- (Ideally using estimate to do testing too, such as logrank test and HR)

How to communicate with patients via various summaries for treatment effect?

- Which summary can be comprehended easily by clinical practitioners and patients?



The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 JANUARY 20, 2005 VOL. 352 NO. 3

Amiodarone or an Implantable Cardioverter–Defibrillator for Congestive Heart Failure

METHODS

We randomly assigned 2521 patients with New York Heart Association (NYHA) class II or III CHF and a left ventricular ejection fraction (LVEF) of 35 percent or less to conventional therapy for CHF plus placebo (847 patients), conventional therapy plus amiodarone (845 patients), or conventional therapy plus a conservatively programmed, shock-only, single-lead ICD (829 patients). Placebo and amiodarone were administered in a double-blind fashion. The primary end point was death from any cause.

Benefit vs harm of ICD?

Risks associated with ICD implantation are uncommon but may include (from Mayo Website):

- Infection at the implant site
- Allergic reaction to the medications used during the procedure
- Swelling, bleeding or bruising where your ICD was implanted
- Damage to the vein where your ICD leads are placed
- Bleeding around your heart, which can be life-threatening
- Blood leaking through the heart valve where the ICD lead is placed
- Collapsed lung (pneumothorax)

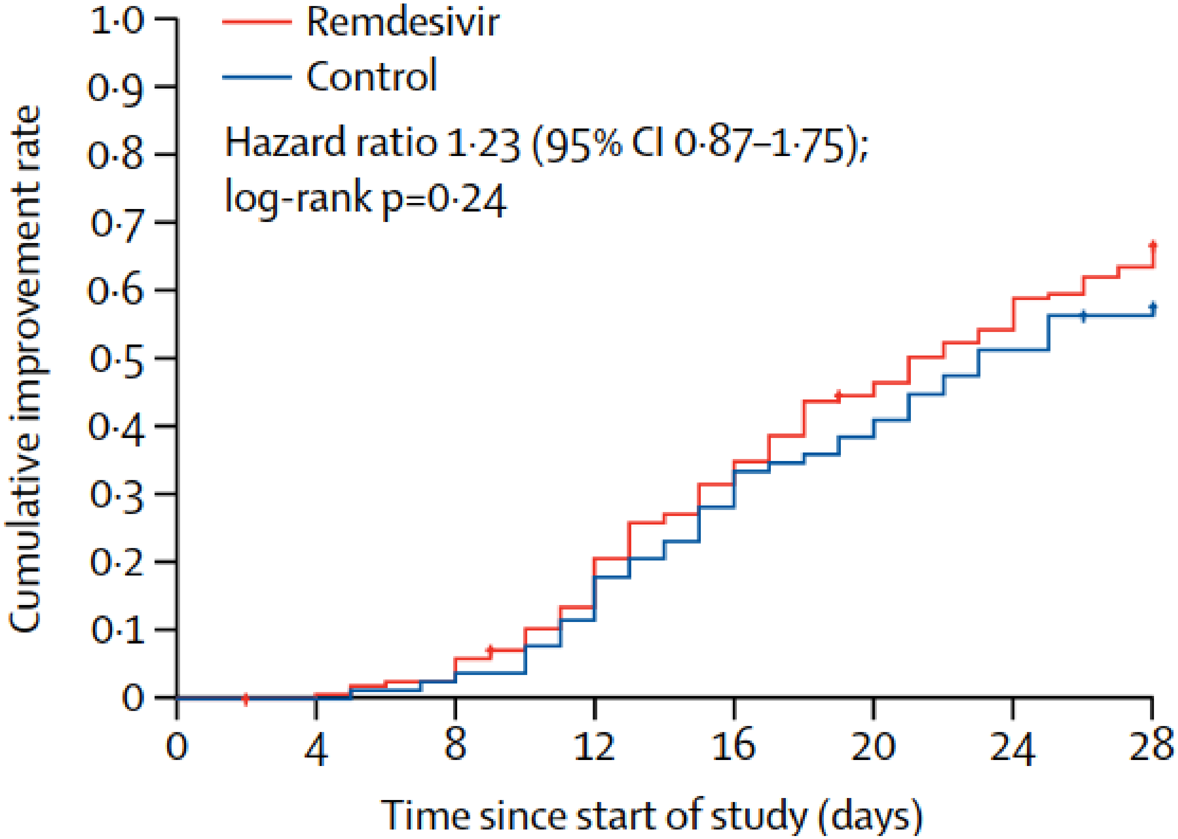
Shared decision making between patients and clinicians

GJ is a 79-year-old woman with hypertension, diabetes, osteoporosis, depression, and New York Heart Association class II heart failure with a left ventricular ejection fraction of 30%. She is a potential candidate for an implantable cardioverter-defibrillator (ICD), and you would like to discuss this with her using evidence from a clinical trial. Which of the following statistics would be most helpful in explaining the possible survival benefit of an ICD?

- *p-value comparing mortality of ICD and placebo groups was 0.007.*
- *hazard ratio (HR) for mortality was 0.77.*
- *absolute risk reduction was 7%, from 36% to 29%, over 5 years.*
- *number-needed-to-treat (NNT) was 15 over 5 years.*
- *ICD will prolong life from 49.1 to 51.4 months, an average of 2.3 months, over 5 years.*

Another example of the difficulty of interpret hazard ratio

Remdesivir in adults with severe COVID-19 (Lancet online April 29, 2020)



- The primary endpoint was the time to a clinical improvement within 28-days of follow-up via a WHO six-point ordinal scale outcome. The shorter, the better.
- The observed HR was 1·23 (95% CI [0·87,1·75]). A 23% of hazard increase over placebo is difficult to interpret clinically. Moreover, a clinical improvement event might not be observed due to death.
- The area under the curve (AUC) up to 28-days would be a reasonable summary of the treatment efficacy.
- the AUCs are 7·6 and 6·7 days for remdesivir and control, respectively. On average, patients with remdesivir enjoyed 7·6 days of improvement with 28-days follow-up.
- Patients who died by 28-days without improvement contributed zero days to this average value. The difference is 0·9 days (95% CI [-1·1,2·8]) numerically favoring remdesivir.

Annals of Internal Medicine®

Information for Authors - General Statistical Guidance

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

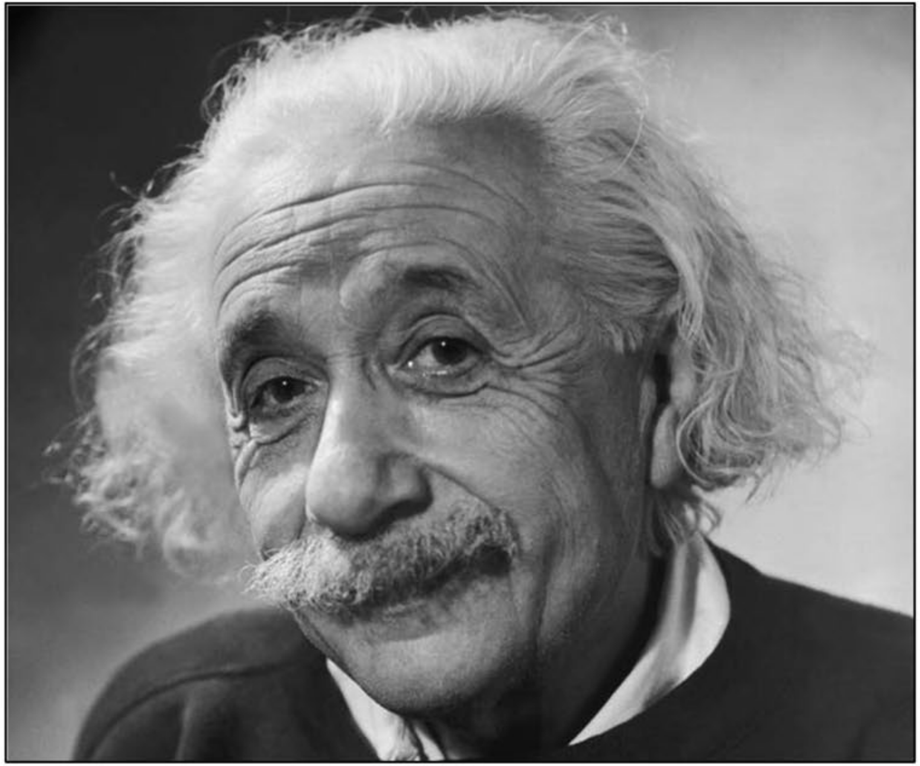
Hazard Ratios and Standardized Cumulative Incidence

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.

Let us see what Sir David told us..

In an interview, Professor David R. Cox, the creator of the proportional hazards model, stated, “Of course, another issue is the physical or substantive basis for the proportional hazards model. I think that’s one of its weaknesses…”

> "If you can't explain it simply, you just don't understand it well enough."
>
> — Albert Einstein



Beyond “translational” what are other advantages of RMST analysis?

*HR does not have a causal treatment effect interpretation.

*When proportional hazards assumption is not met, HR is difficult to interpret, which is not a simple average of hazard ratios over time. The parameter HR estimated depending on the censoring distributions.

*RMST based statistics can be more powerful than HR under non-PH.

*When HR gives significant results, so does RMST.

*For equivalence or non-inferiority studies, RMST does not require a large study like HR (event driven).

*RMST uses more data than HR

t-year mean survival time

Uno et al. (2014, JCO)

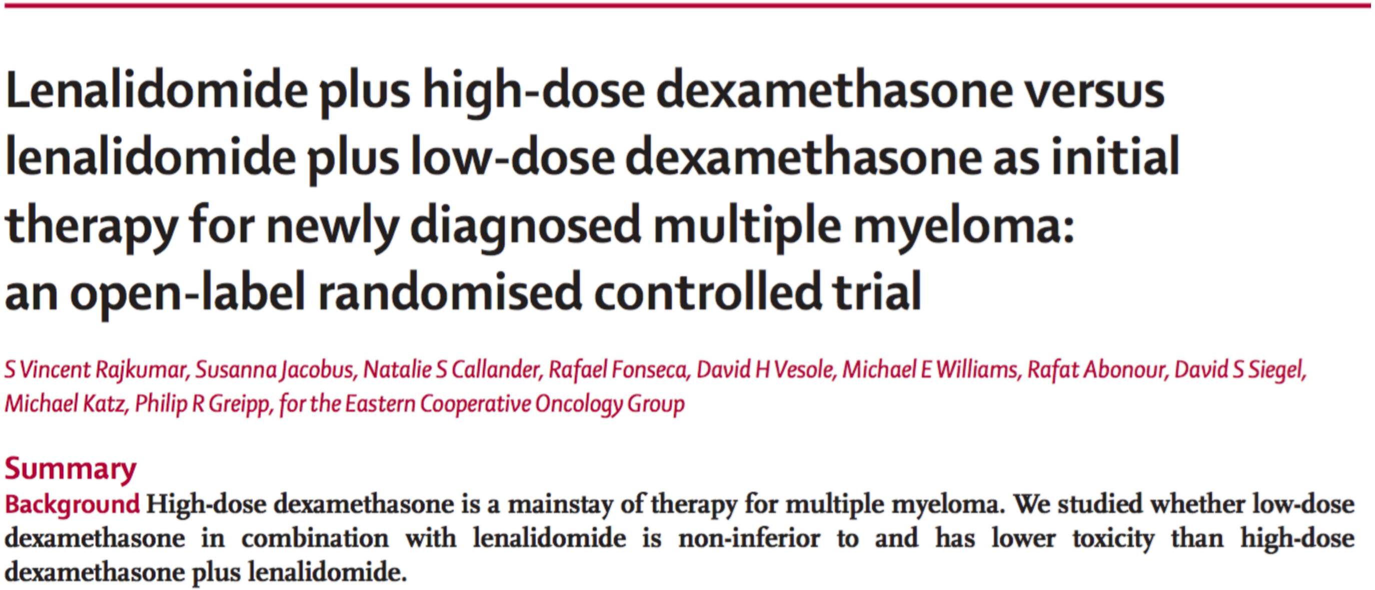
Pak et al. (2017, JAMA-Oncology)

Uno et al. (2015, Annals of Internal Medicine)

Z. McCaw, G. Yin and L.J. Wei (Circulation, 2019)

For non-proportional hazards, RMST can be more powerful and interpretable

Example:
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

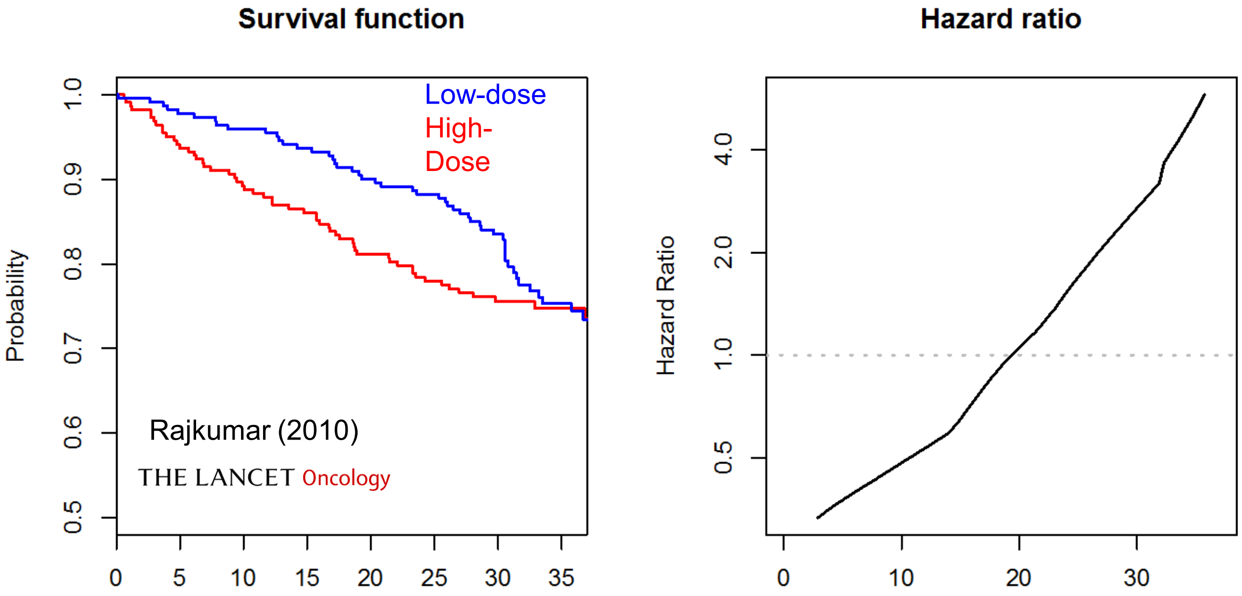
S Vincent Rajkumar, Susanna Jacobus, Natalie S Callander, Rafael Fonseca, David H Vesole, Michael E Williams, Rafat Abonour, David S Siegel, Michael Katz, Philip R Greipp, for the Eastern Cooperative Oncology Group

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)

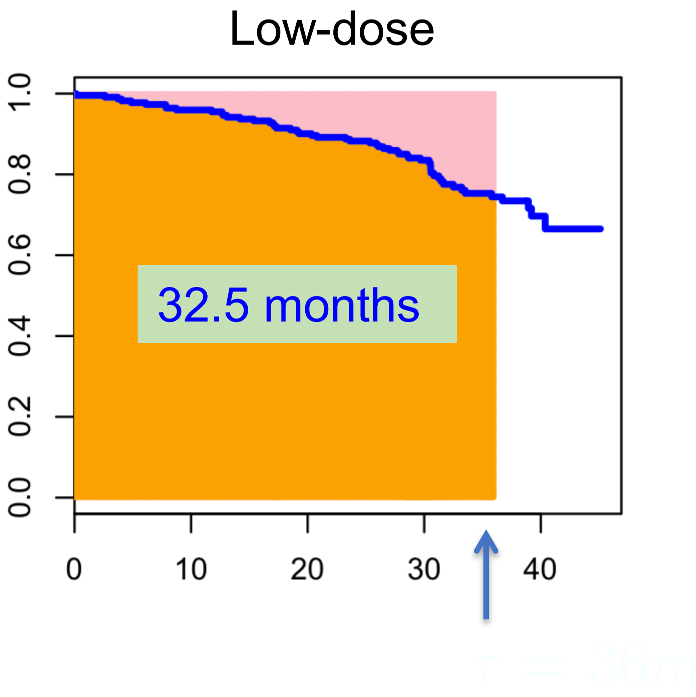
ECOG Myeloma study (OS, low Dex vs. High Dex)



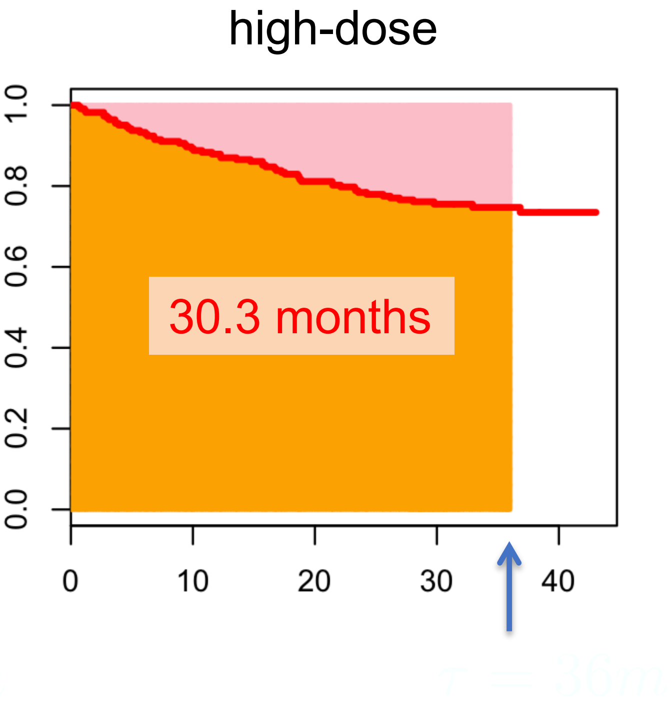
HR= 0.87 (0.95CI: 0.60 to 1.27), p=0.46

Restricted mean survival time (RMST) Difference:

2.2 months; CI: 0.5 to 4.0, p=0.014



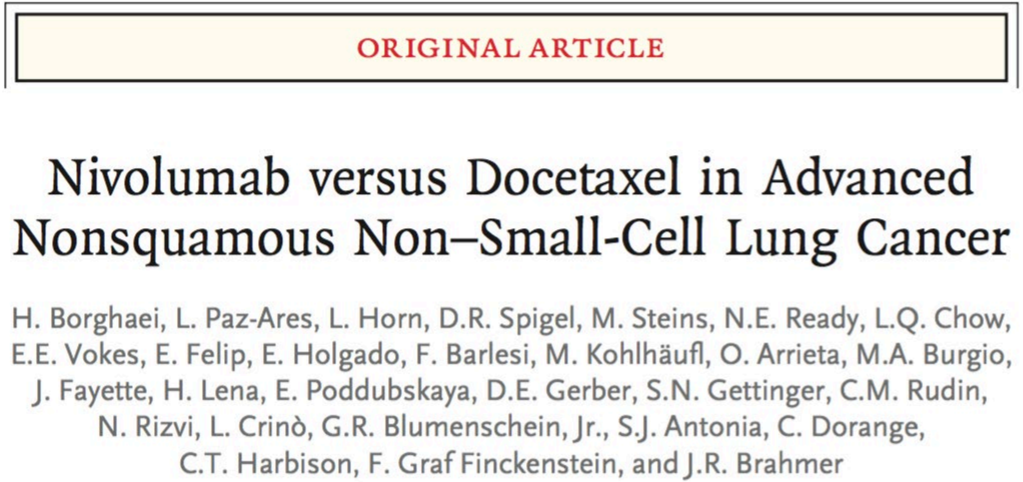
Low-dose



high-dose

Another example, RMST procedure can be more powerful than the hazard ratio’s

CheckMate 057 Study

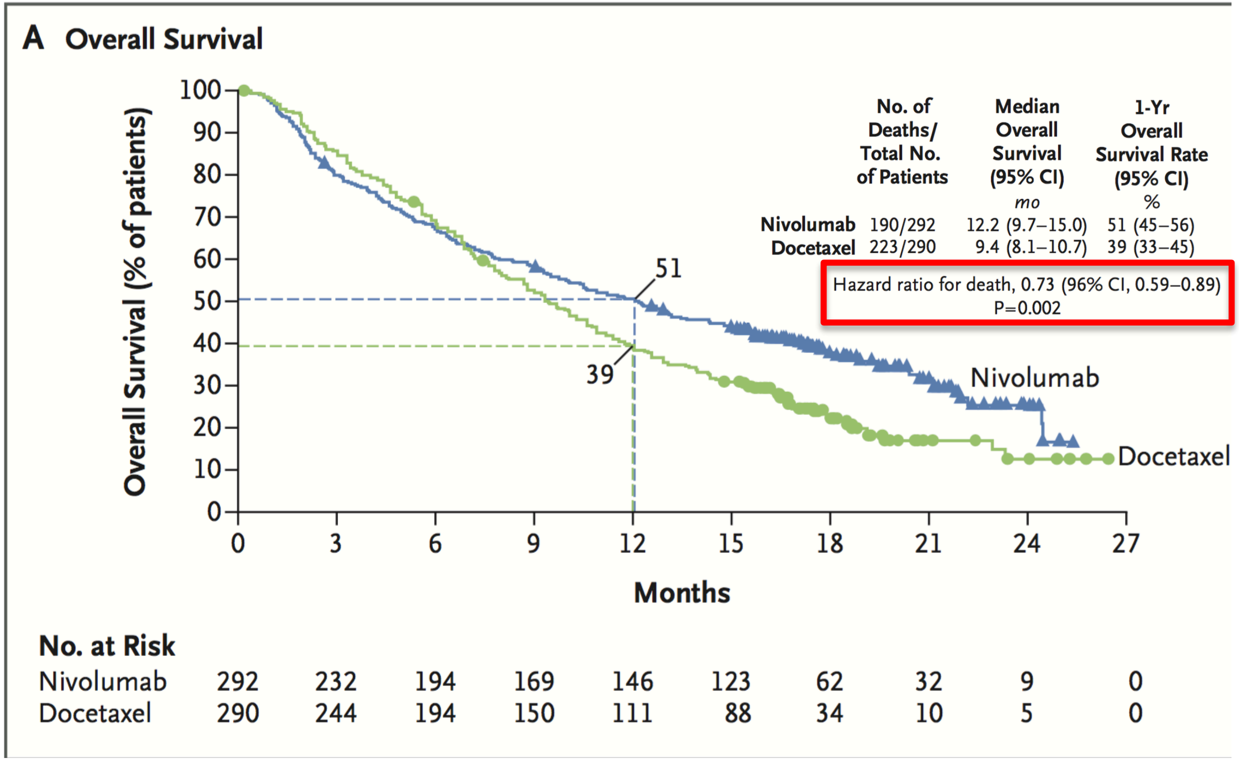


ORIGINAL ARTICLE

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)



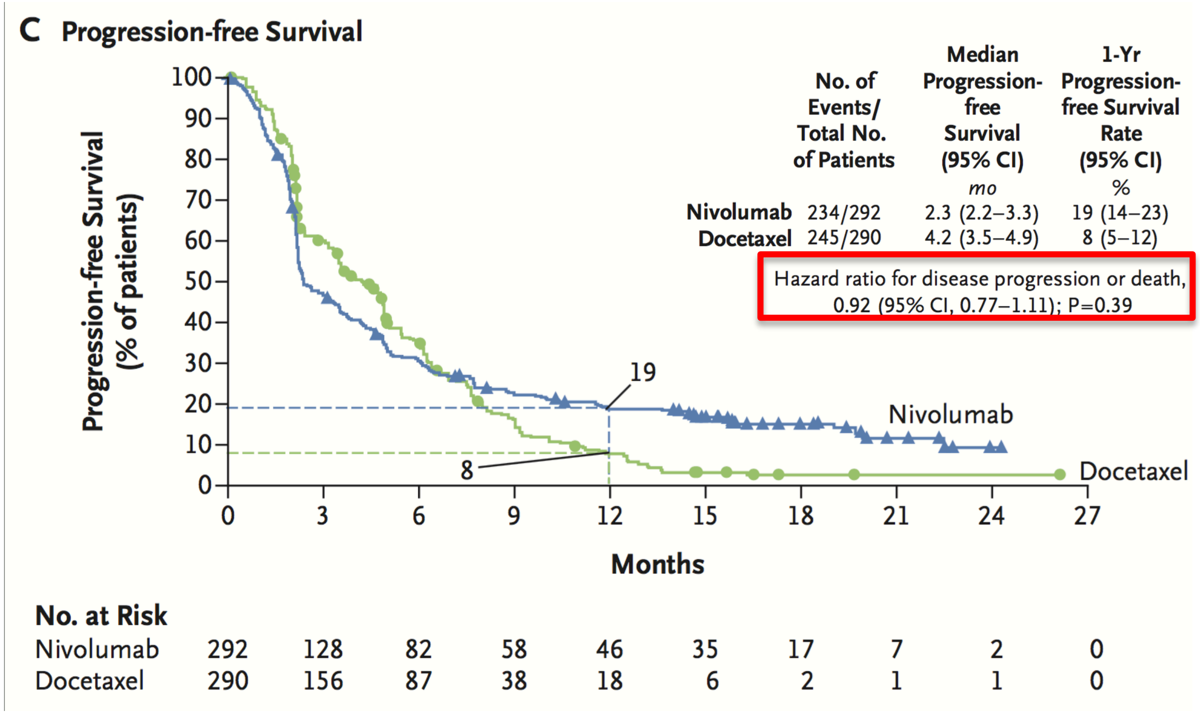
A Overall Survival

	No. of Deaths/ Total No. of Patients	Median Overall Survival (95% CI)	1-Yr Overall Survival Rate (95% CI)
		mo	%
Nivolumab	190/292	12.2 (9.7–15.0)	51 (45–56)
Docetaxel	223/290	9.4 (8.1–10.7)	39 (33–45)

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

No. at Risk										
Nivolumab	292	232	194	169	146	123	62	32	9	0
Docetaxel	290	244	194	150	111	88	34	10	5	0

Borghaei et al. (2015, NEJM)



	No. of Events/ Total No. of Patients	Median Progression-free Survival (95% CI)	1-Yr Progression-free Survival Rate (95% CI)
		mo	%
Nivolumab	234/292	2.3 (2.2–3.3)	19 (14–23)
Docetaxel	245/290	4.2 (3.5–4.9)	8 (5–12)

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

No. at Risk

Nivolumab	292	128	82	58	46	35	17	7	2	0
Docetaxel	290	156	87	38	18	6	2	1	1	0

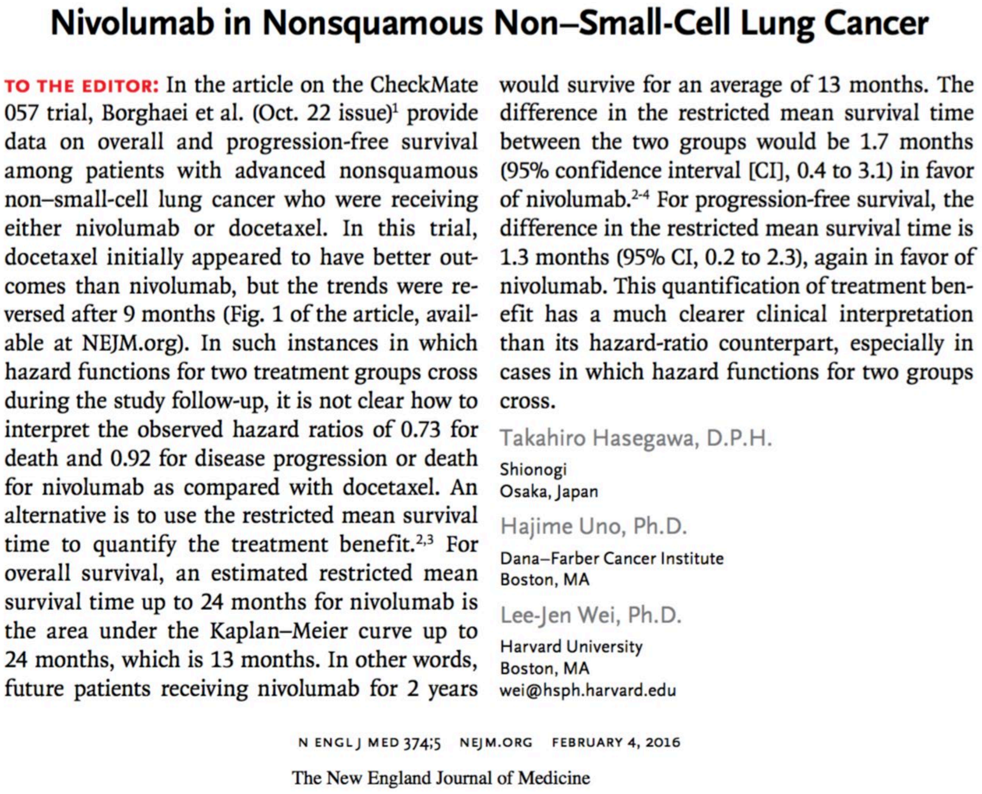
Borghaei et al. (2015, NEJM)

RMST analysis for PFS

The difference between two arms was 1.3 months with

95% CI: (0.2, 2.3), **statistically significant**!

Our letter to the editor



Nivolumab in Nonsquamous Non–Small-Cell Lung Cancer

TO THE EDITOR: In the article on the CheckMate 057 trial, Borghaei et al. (Oct. 22 issue)[1] provide data on overall and progression-free survival among patients with advanced nonsquamous non–small-cell lung cancer who were receiving either nivolumab or docetaxel. In this trial, docetaxel initially appeared to have better outcomes than nivolumab, but the trends were reversed after 9 months (Fig. 1 of the article, available at NEJM.org). In such instances in which hazard functions for two treatment groups cross during the study follow-up, it is not clear how to interpret the observed hazard ratios of 0.73 for death and 0.92 for disease progression or death for nivolumab as compared with docetaxel. An alternative is to use the restricted mean survival time to quantify the treatment benefit.[2,3] For overall survival, an estimated restricted mean survival time up to 24 months for nivolumab is the area under the Kaplan–Meier curve up to 24 months, which is 13 months. In other words, future patients receiving nivolumab for 2 years would survive for an average of 13 months. The difference in the restricted mean survival time between the two groups would be 1.7 months (95% confidence interval [CI], 0.4 to 3.1) in favor of nivolumab.[2-4] For progression-free survival, the difference in the restricted mean survival time is 1.3 months (95% CI, 0.2 to 2.3), again in favor of nivolumab. This quantification of treatment benefit has a much clearer clinical interpretation than its hazard-ratio counterpart, especially in cases in which hazard functions for two groups cross.

Takahiro Hasegawa, D.P.H.
Shionogi
Osaka, Japan

Hajime Uno, Ph.D.
Dana–Farber Cancer Institute
Boston, MA

Lee-Jen Wei, Ph.D.
Harvard University
Boston, MA
wei@hsph.harvard.edu

N ENGL J MED 374;5 NEJM.ORG FEBRUARY 4, 2016

The New England Journal of Medicine

HR gives significant result, so does RMST

Analysis of 7 clinical studies for heart failure

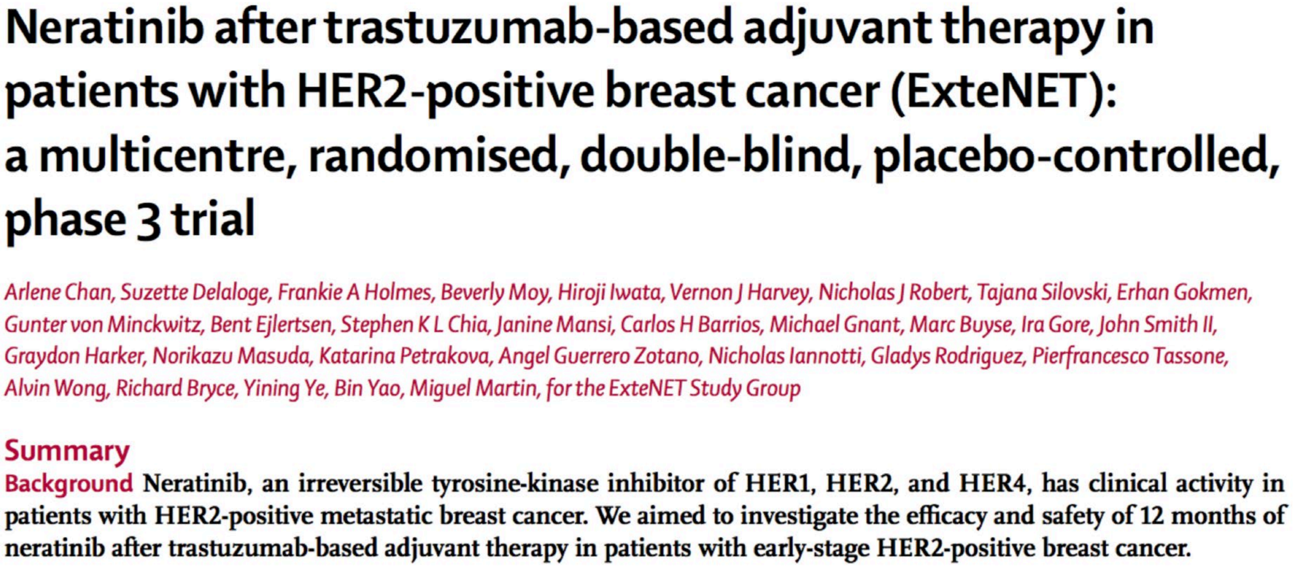
- C. Perego; M. Sbolli; C. Specchia; C. Oriecuia; G. Peveri; M.Metra; M. Fiuzat; L.J. Wei; C.M. O’Connor;M.A. Psotka

Inova Heart Vascular Inst. Falls Church, US; Univ of Bressia, Italy; Univ of Milan, Italy; Duke Univ; SPEDALI CIVILI Hosp, Italy, Harvard Univ.

Trial	Treatment(s)	Outcome	Hazard Ratio (HR)		RMST (months)		TIME (months)
			HR (95% CI)	p-value	RMST Difference (95% CI)	p-value	
CONSENSUS	Enalapril vs Placebo	All-cause death	0.73[b]	0.003	2.2 (1 – 3.4)	< 0.001	12
RALES	Spironolactone vs Placebo	All-cause death	0.70 (0.60-0.82)	0.001	2.2 (1.1 – 3.4)	< 0.001	34
COPERNICUS	Carvedilol vs Placebo	Death or cardiovascular hospitalization	0.73 (0.63-0.84)	<0.001	1.7 (1.1 – 2.4)	<0.001	21
MERIT-HF	Metoprolol CR/XL vs Placebo	All-cause death	0.66 (0.53-0.81)	< 0.001	0.4 (0.2 – 0.7)	<0.001	18
SHIFT	Ivabradine vs Placebo	Cardiovascular death or HF hospitalization	0.82 (0.75-0.90)	< 0.001	1.0 (0.5 – 1.5)	< 0.001	30
PARADIGM-HF	Sacubitril/valsartan vs Enalapril	Cardiovascular death or HF hospitalization	0.80 (0.73-0.87)	< 0.001	1.5 (0.9 – 2.0)	< 0.001	41
DAPA-HF	Dapagliflozin vs Placebo	Cardiovascular death or worsening HF	0.74 (0.65-0.85)	< 0.001	0.9 (0.5 – 1.2)	< 0.001	24

Highly statistical significance may not be clinically significant

ExteNET Study



Neratinib after trastuzumab-based adjuvant therapy in patients with HER2-positive breast cancer (ExteNET): a multicentre, randomised, double-blind, placebo-controlled, phase 3 trial

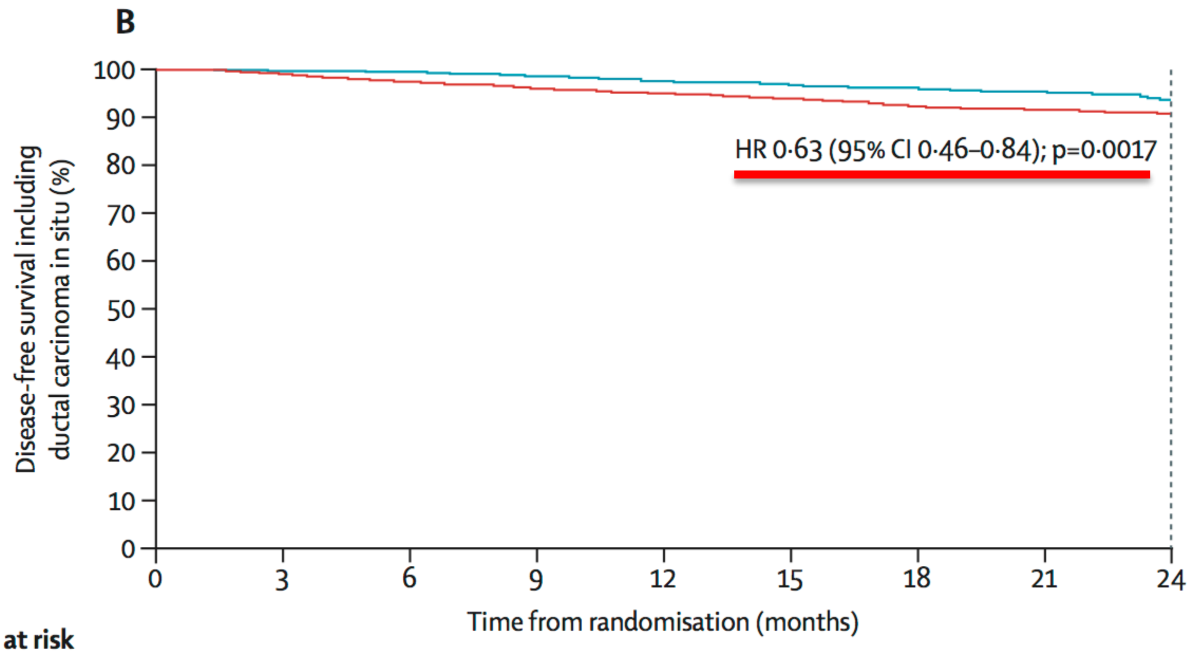
Arlene Chan, Suzette Delaloge, Frankie A Holmes, Beverly Moy, Hiroji Iwata, Vernon J Harvey, Nicholas J Robert, Tajana Silovski, Erhan Gokmen, Gunter von Minckwitz, Bent Ejlertsen, Stephen K L Chia, Janine Mansi, Carlos H Barrios, Michael Gnant, Marc Buyse, Ira Gore, John Smith II, Graydon Harker, Norikazu Masuda, Katarina Petrakova, Angel Guerrero Zotano, Nicholas Iannotti, Gladys Rodriguez, Pierfrancesco Tassone, Alvin Wong, Richard Bryce, Yining Ye, Bin Yao, Miguel Martin, for the ExteNET Study Group

Summary

Background Neratinib, an irreversible tyrosine-kinase inhibitor of HER1, HER2, and HER4, has clinical activity in patients with HER2-positive metastatic breast cancer. We aimed to investigate the efficacy and safety of 12 months of neratinib after trastuzumab-based adjuvant therapy in patients with early-stage HER2-positive breast cancer.

Chan et al. (2016, Lancet Onc)

Disease-free survival including ductal carcinoma in situ (DCIS)



HR 0·63 (95% CI 0·46–0·84); p=0·0017

Number at risk									
Neratinib group	1420	1291	1260	1229	1189	1150	1108	1033	662
Placebo group	1420	1366	1324	1290	1241	1206	1159	1086	701

Chan et al. (2016, Lancet Onc)

Issues and concerns

- The PH may be ok, but the hazard ratio is difficult to explain with a short-term follow-up

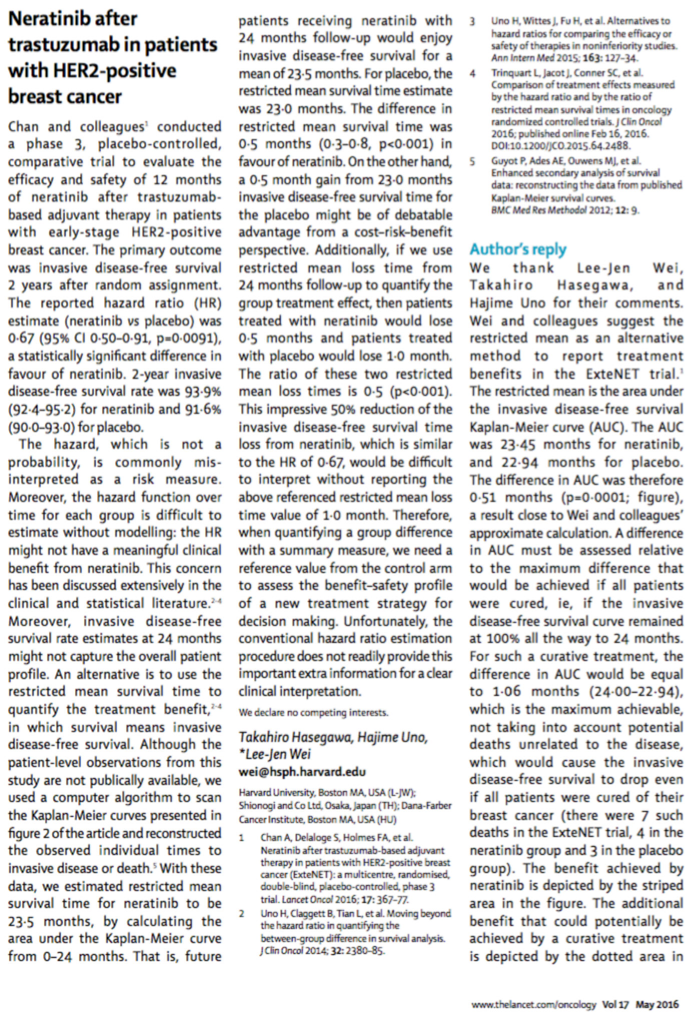
- What is the gain from the extra treatment clinically?

- No median survival time estimate

Our analysis results for a clear clinical interpretation

Disease-free survival including DCIS

Up to 24 months		Estimate	95% CI	P-value
RMST	Neratinib	23.43	(23.28, 23.58)	
	Placebo	22.84	(22.62, 23.06)	
	Difference	0.59	(0.33, 0.86)	<0.001
	Ratio	1.03	(1.01, 1.04)	<0.001



Neratinib after trastuzumab in patients with HER2-positive breast cancer

Chan and colleagues[1] conducted a phase 3, placebo-controlled, comparative trial to evaluate the efficacy and safety of 12 months of neratinib after trastuzumab-based adjuvant therapy in patients with early-stage HER2-positive breast cancer. The primary outcome was invasive disease-free survival 2 years after random assignment. The reported hazard ratio (HR) estimate (neratinib vs placebo) was 0·67 (95% CI 0·50–0·91, p=0·0091), a statistically significant difference in favour of neratinib. 2-year invasive disease-free survival rate was 93·9% (92·4–95·2) for neratinib and 91·6% (90·0–93·0) for placebo.

The hazard, which is not a probability, is commonly misinterpreted as a risk measure. Moreover, the hazard function over time for each group is difficult to estimate without modelling: the HR might not have a meaningful clinical benefit from neratinib. This concern has been discussed extensively in the clinical and statistical literature.[2-4] Moreover, invasive disease-free survival rate estimates at 24 months might not capture the overall patient profile. An alternative is to use the restricted mean survival time to quantify the treatment benefit,[2-4] in which survival means invasive disease-free survival. Although the patient-level observations from this study are not publically available, we used a computer algorithm to scan the Kaplan-Meier curves presented in figure 2 of the article and reconstructed the observed individual times to invasive disease or death.[5] With these data, we estimated restricted mean survival time for neratinib to be 23·5 months, by calculating the area under the Kaplan-Meier curve from 0–24 months. That is, future patients receiving neratinib with 24 months follow-up would enjoy invasive disease-free survival for a mean of 23·5 months. For placebo, the restricted mean survival time estimate was 23·0 months. The difference in restricted mean survival time was 0·5 months (0·3–0·8, p<0·001) in favour of neratinib. On the other hand, a 0·5 month gain from 23·0 months invasive disease-free survival time for the placebo might be of debatable advantage from a cost–risk–benefit perspective. Additionally, if we use restricted mean loss time from 24 months follow-up to quantify the group treatment effect, then patients treated with neratinib would lose 0·5 months and patients treated with placebo would lose 1·0 month. The ratio of these two restricted mean loss times is 0·5 (p<0·001). This impressive 50% reduction of the invasive disease-free survival time loss from neratinib, which is similar to the HR of 0·67, would be difficult to interpret without reporting the above referenced restricted mean loss time value of 1·0 month. Therefore, when quantifying a group difference with a summary measure, we need a reference value from the control arm to assess the benefit–safety profile of a new treatment strategy for decision making. Unfortunately, the conventional hazard ratio estimation procedure does not readily provide this important extra information for a clear clinical interpretation.

We declare no competing interests.

*Takahiro Hasegawa, Hajime Uno, *Lee-Jen Wei*
wei@hsph.harvard.edu

Harvard University, Boston MA, USA (L-JW); Shionogi and Co Ltd, Osaka, Japan (TH); Dana-Farber Cancer Institute, Boston MA, USA (HU)

1 Chan A, Delaloge S, Holmes FA, et al. Neratinib after trastuzumab-based adjuvant therapy in patients with HER2-positive breast cancer (ExteNET): a multicentre, randomised, double-blind, placebo-controlled, phase 3 trial. *Lancet Oncol* 2016; **17:** 367–77.

2 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:** 2380–85.

3 Uno H, Wittes J, Fu H, et al. Alternatives to hazard ratios for comparing the efficacy or safety of therapies in noninferiority studies. *Ann Intern Med* 2015; **163:** 127–34.

4 Trinquart L, Jacot J, Conner SC, et al. Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. *J Clin Oncol* 2016; published online Feb 16, 2016. DOI:10.1200/JCO.2015.64.2488.

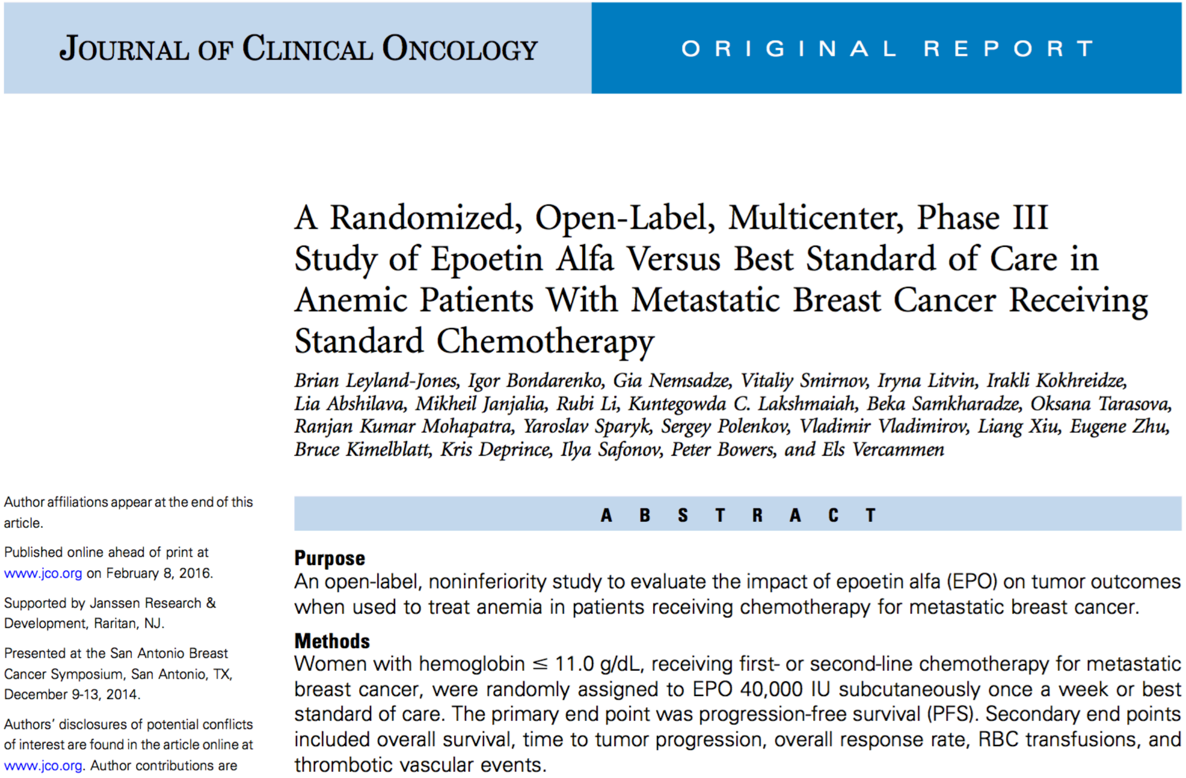
5 Guyot P, Ades AE, Ouwens MJ, et al. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol* 2012; **12:** 9.

Author's reply

We thank Lee-Jen Wei, Takahiro Hasegawa, and Hajime Uno for their comments. Wei and colleagues suggest the restricted mean as an alternative method to report treatment benefits in the ExteNET trial.[1] The restricted mean is the area under the invasive disease-free survival Kaplan-Meier curve (AUC). The AUC was 23·45 months for neratinib, and 22·94 months for placebo. The difference in AUC was therefore 0·51 months (p=0·0001; figure), a result close to Wei and colleagues' approximate calculation. A difference in AUC must be assessed relative to the maximum difference that would be achieved if all patients were cured, ie, if the invasive disease-free survival curve remained at 100% all the way to 24 months. For such a curative treatment, the difference in AUC would be equal to 1·06 months (24·00–22·94), which is the maximum achievable, not taking into account potential deaths unrelated to the disease, which would cause the invasive disease-free survival to drop even if all patients were cured of their breast cancer (there were 7 such deaths in the ExteNET trial, 4 in the neratinib group and 3 in the placebo group). The benefit achieved by neratinib is depicted by the striped area in the figure. The additional benefit that could potentially be achieved by a curative treatment is depicted by the dotted area in

Non-inferiority studies

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

ABSTRACT

Purpose

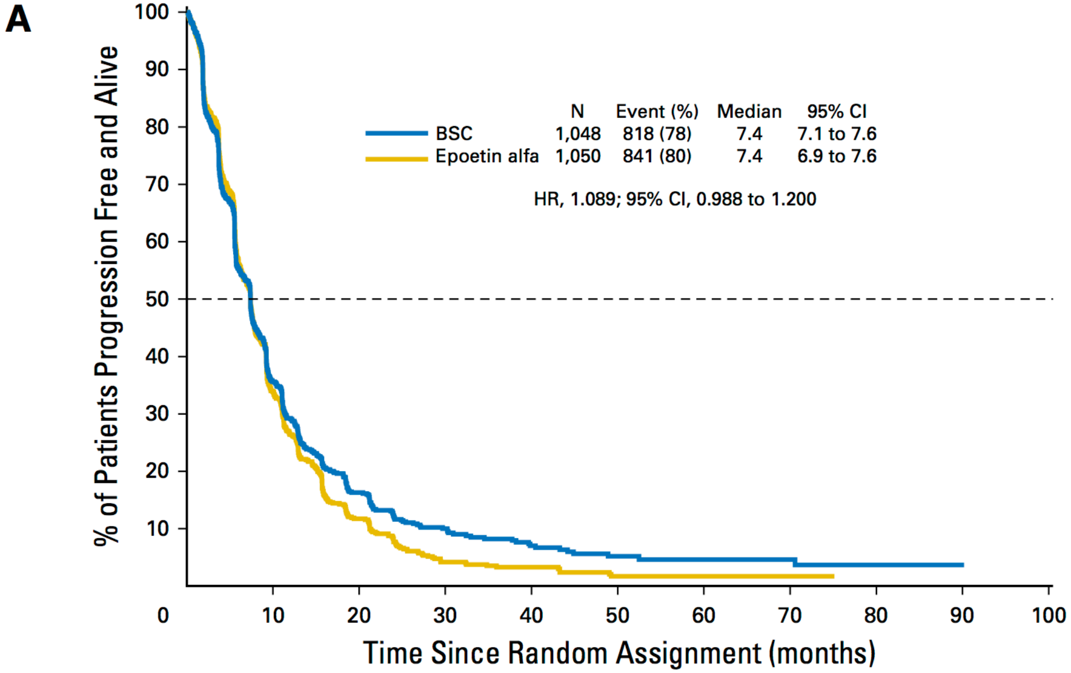
An open-label, noninferiority study to evaluate the impact of epoetin alfa (EPO) on tumor outcomes when used to treat anemia in patients receiving chemotherapy for metastatic breast cancer.

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)

PFS by Investigator



	N	Event (%)	Median	95% CI
BSC	1,048	818 (78)	7.4	7.1 to 7.6
Epoetin alfa	1,050	841 (80)	7.4	6.9 to 7.6

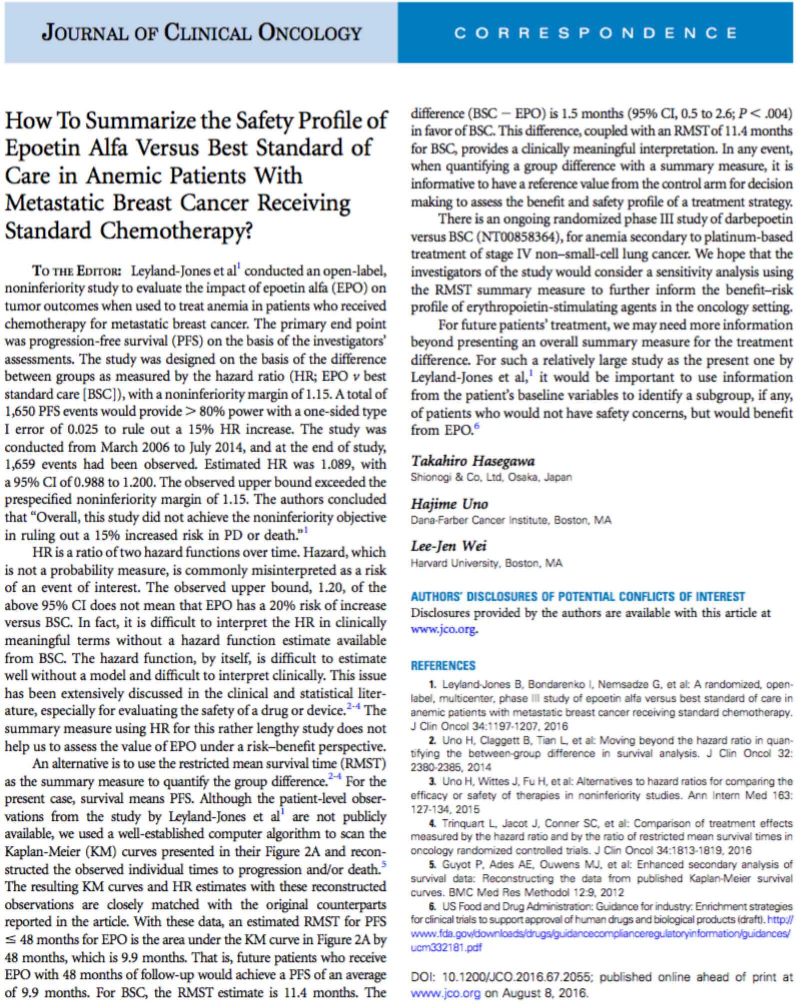
HR, 1.089; 95% CI, 0.988 to 1.200

No. at risk											
BSC	1,048	299	106	42	22	12	7	5	2	1	0
Epoetin alfa	1,050	283	72	21	13	5	4	2	0	0	0

Leyland-Janes et al. (2016, JCO)

Our analysis results

	RMST up to		Estimate	95% CI	P-value
PFS	48 months	BSC	11.40	(10.56, 12.23)	
		Epoetin alfa	9.87	(9.23, 10.51)	
		Difference	-1.53	(-2.58, -0.47)	0.004



Journal of Clinical Oncology CORRESPONDENCE

How To Summarize the Safety Profile of Epoetin Alfa Versus Best Standard of Care in Anemic Patients With Metastatic Breast Cancer Receiving Standard Chemotherapy?

To the Editor: Leyland-Jones et al[1] conducted an open-label, noninferiority study to evaluate the impact of epoetin alfa (EPO) on tumor outcomes when used to treat anemia in patients who received chemotherapy for metastatic breast cancer. The primary end point was progression-free survival (PFS) on the basis of the investigators' assessments. The study was designed on the basis of the difference between groups as measured by the hazard ratio (HR; EPO *v* best standard care [BSC]), with a noninferiority margin of 1.15. A total of 1,650 PFS events would provide > 80% power with a one-sided type I error of 0.025 to rule out a 15% HR increase. The study was conducted from March 2006 to July 2014, and at the end of study, 1,659 events had been observed. Estimated HR was 1.089, with a 95% CI of 0.988 to 1.200. The observed upper bound exceeded the prespecified noninferiority margin of 1.15. The authors concluded that "Overall, this study did not achieve the noninferiority objective in ruling out a 15% increased risk in PD or death."[1]

HR is a ratio of two hazard functions over time. Hazard, which is not a probability measure, is commonly misinterpreted as a risk of an event of interest. The observed upper bound, 1.20, of the above 95% CI does not mean that EPO has a 20% risk of increase versus BSC. In fact, it is difficult to interpret the HR in clinically meaningful terms without a hazard function estimate available from BSC. The hazard function, by itself, is difficult to estimate well without a model and difficult to interpret clinically. This issue has been extensively discussed in the clinical and statistical literature, especially for evaluating the safety of a drug or device.[2-4] The summary measure using HR for this rather lengthy study does not help us to assess the value of EPO under a risk–benefit perspective.

An alternative is to use the restricted mean survival time (RMST) as the summary measure to quantify the group difference.[2-4] For the present case, survival means PFS. Although the patient-level observations from the study by Leyland-Jones et al[1] are not publicly available, we used a well-established computer algorithm to scan the Kaplan-Meier (KM) curves presented in their Figure 2A and reconstructed the observed individual times to progression and/or death.[5] The resulting KM curves and HR estimates with these reconstructed observations are closely matched with the original counterparts reported in the article. With these data, an estimated RMST for PFS $\le$ 48 months for EPO is the area under the KM curve in Figure 2A by 48 months, which is 9.9 months. That is, future patients who receive EPO with 48 months of follow-up would achieve a PFS of an average of 9.9 months. For BSC, the RMST estimate is 11.4 months. The difference (BSC − EPO) is 1.5 months (95% CI, 0.5 to 2.6; $P < .004$) in favor of BSC. This difference, coupled with an RMST of 11.4 months for BSC, provides a clinically meaningful interpretation. In any event, when quantifying a group difference with a summary measure, it is informative to have a reference value from the control arm for decision making to assess the benefit and safety profile of a treatment strategy.

There is an ongoing randomized phase III study of darbepoetin versus BSC (NT00858364), for anemia secondary to platinum-based treatment of stage IV non–small-cell lung cancer. We hope that the investigators of the study would consider a sensitivity analysis using the RMST summary measure to further inform the benefit–risk profile of erythropoietin-stimulating agents in the oncology setting.

For future patients' treatment, we may need more information beyond presenting an overall summary measure for the treatment difference. For such a relatively large study as the present one by Leyland-Jones et al,[1] it would be important to use information from the patient's baseline variables to identify a subgroup, if any, of patients who would not have safety concerns, but would benefit from EPO.[6]

Takahiro Hasegawa
Shionogi & Co, Ltd, Osaka, Japan

Hajime Uno
Dana-Farber Cancer Institute, Boston, MA

Lee-Jen Wei
Harvard University, Boston, MA

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

Disclosures provided by the authors are available with this article at www.jco.org.

REFERENCES

1. Leyland-Jones B, Bondarenko I, Nemsadze G, et al: 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. J Clin Oncol 34:1197-1207, 2016

2. 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 32: 2380-2385, 2014

3. Uno H, Wittes J, Fu H, et al: Alternatives to hazard ratios for comparing the efficacy or safety of therapies in noninferiority studies. Ann Intern Med 163: 127-134, 2015

4. Trinquart L, Jacot J, Conner SC, et al: Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. J Clin Oncol 34:1813-1819, 2016

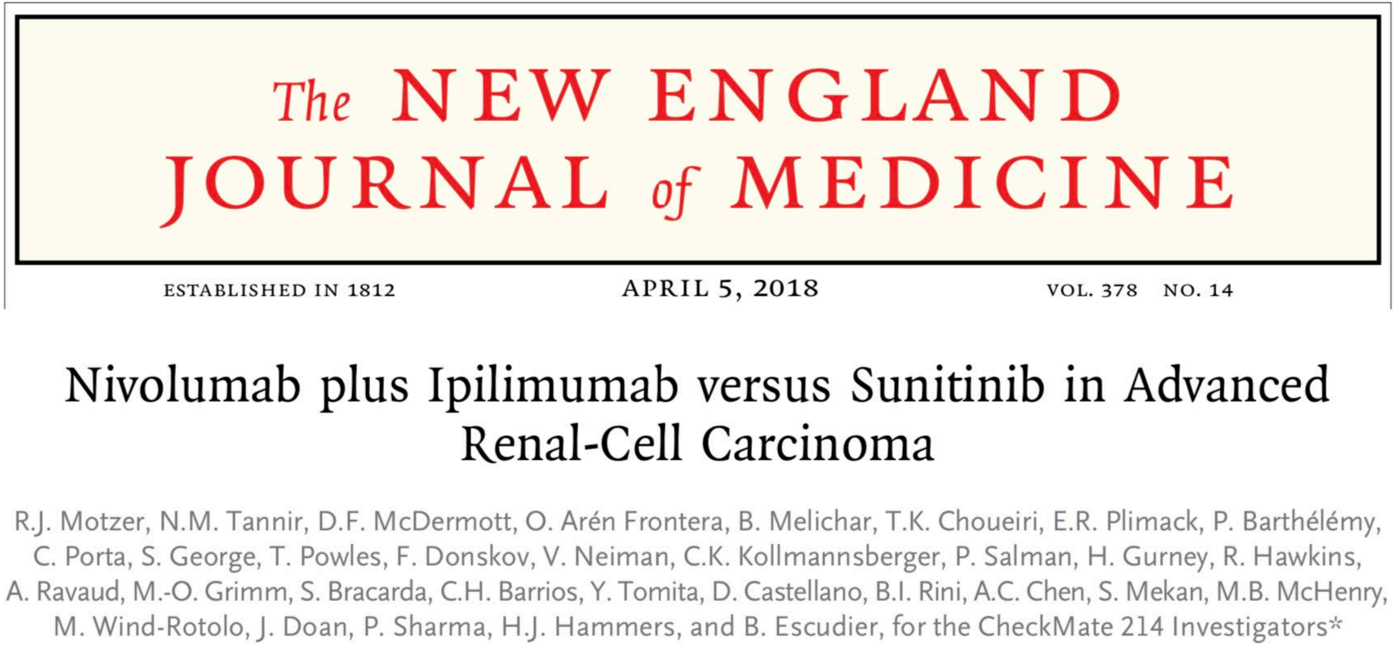
5. Guyot P, Ades AE, Ouwens MJ, et al: Enhanced secondary analysis of survival data: Reconstructing the data from published Kaplan-Meier survival curves. BMC Med Res Methodol 12:9, 2012

6. US Food and Drug Administration: Guidance for industry: Enrichment strategies for clinical trials to support approval of human drugs and biological products (draft). http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm332181.pdf

DOI: 10.1200/JCO.2016.67.2055; published online ahead of print at www.jco.org on August 8, 2016.

Quantifying long term survival

Example of immunotherapy trial (CheckMate 214)



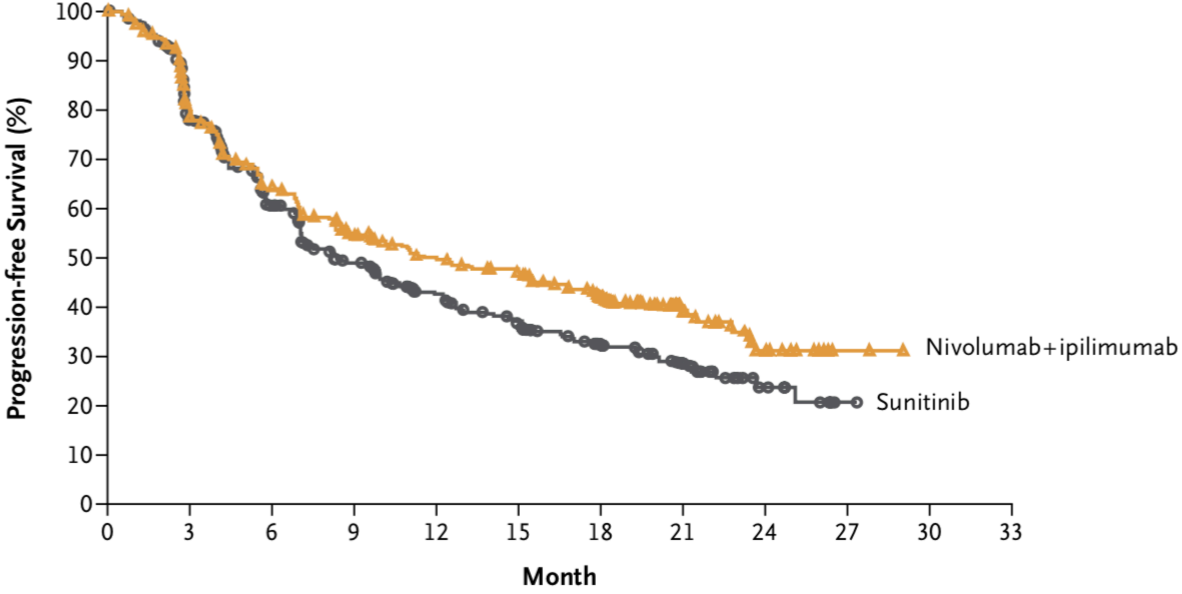
The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 APRIL 5, 2018 VOL. 378 NO. 14

Nivolumab plus Ipilimumab versus Sunitinib in Advanced Renal-Cell Carcinoma

R.J. Motzer, N.M. Tannir, D.F. McDermott, O. Arén Frontera, B. Melichar, T.K. Choueiri, E.R. Plimack, P. Barthélémy, C. Porta, S. George, T. Powles, F. Donskov, V. Neiman, C.K. Kollmannsberger, P. Salman, H. Gurney, R. Hawkins, A. Ravaud, M.-O. Grimm, S. Bracarda, C.H. Barrios, Y. Tomita, D. Castellano, B.I. Rini, A.C. Chen, S. Mekan, M.B. McHenry, M. Wind-Rotolo, J. Doan, P. Sharma, H.J. Hammers, and B. Escudier, for the CheckMate 214 Investigators*

CheckMate 214: Progression-Free Survival



No. at Risk

Nivolumab+ipilimumab	425	304	233	187	163	149	118	46	17	3	0
Sunitinib	422	282	191	139	107	86	57	33	11	1	0

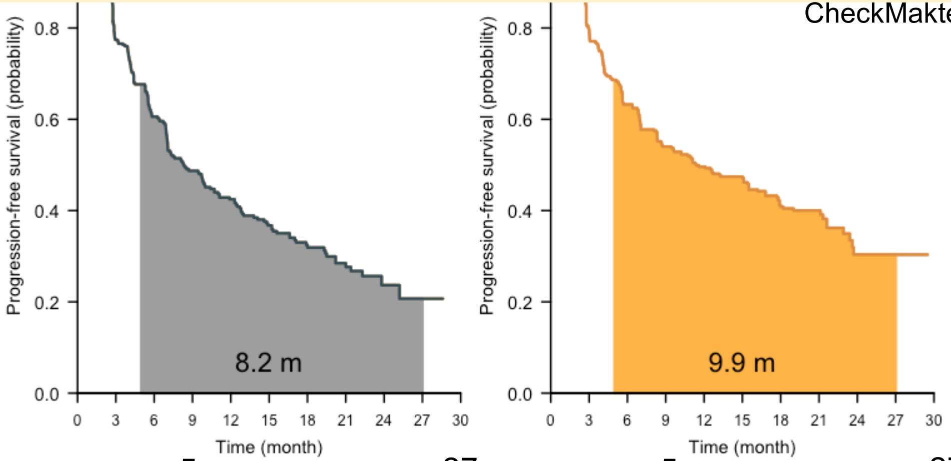
Motzer et al. (2018, *NEJM*)

Long-term RMST-based analysis

(Horiguchi, Tian, Uno, Cheng et al. 2018, JAMA Onc)

Consider area under the curve only on $[t_1, \tau]$

CheckMakte214: PFS



t_1=5 τ=27 t_1=5 τ=27

Difference in RMST [5 – 27m]
1.7m (95%CI: 0.3m to 3.2m)

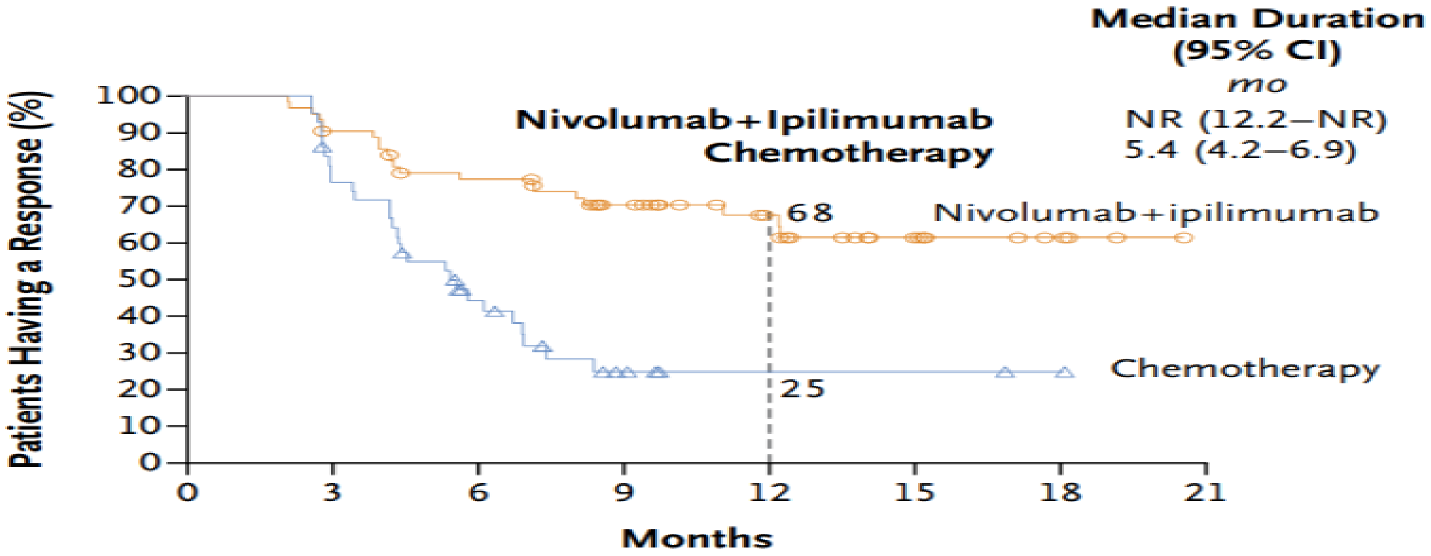
Traditional way to analyze duration of response (DOR) data

DOR among responders

- Construct KM curve of DOR from responders only
- No statistical inference for comparing two treatment groups
- Response is an outcome after randomization
- Under-estimating the treatment effect if there are more responders in the treated group

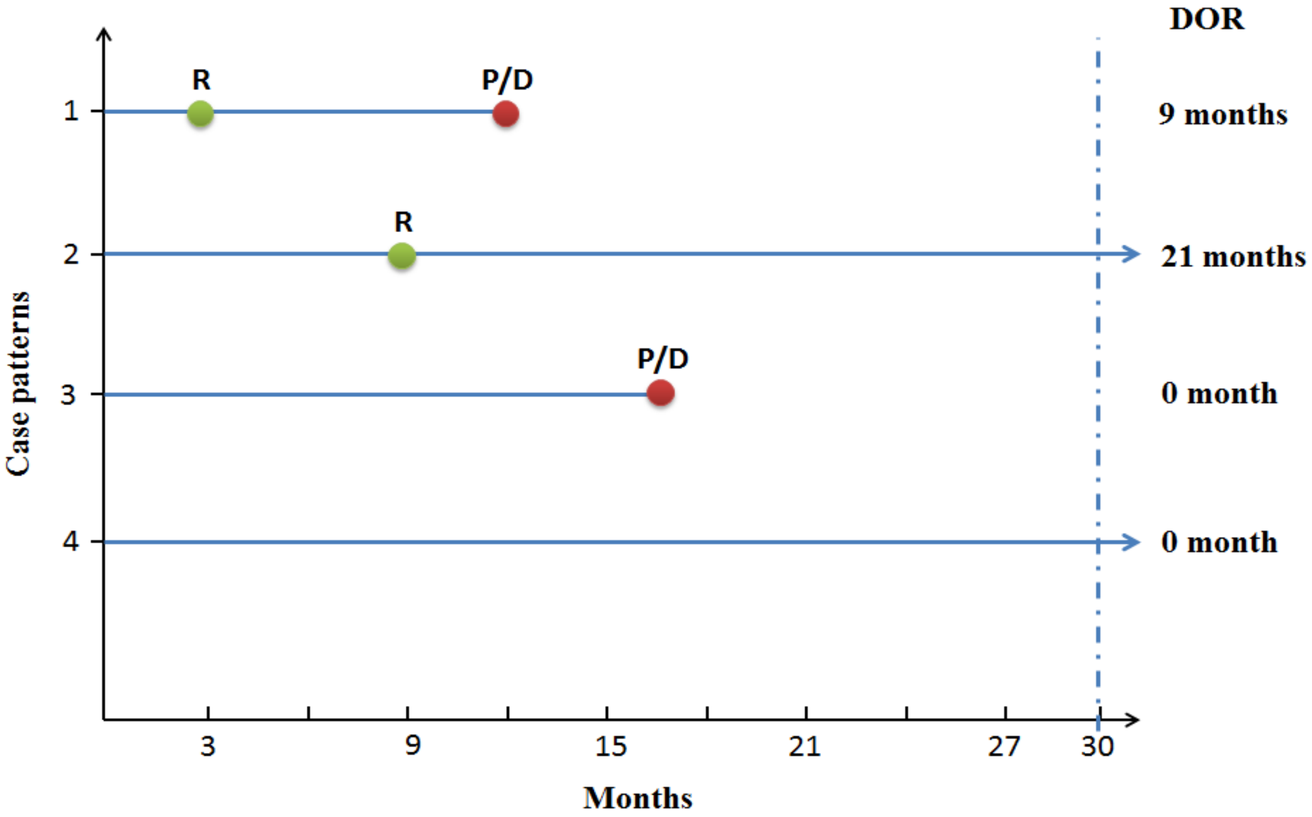
Nivolumab plus Ipilimumab in Lung Cancer with a High Tumor Mutational Burden, NEJM, 2018

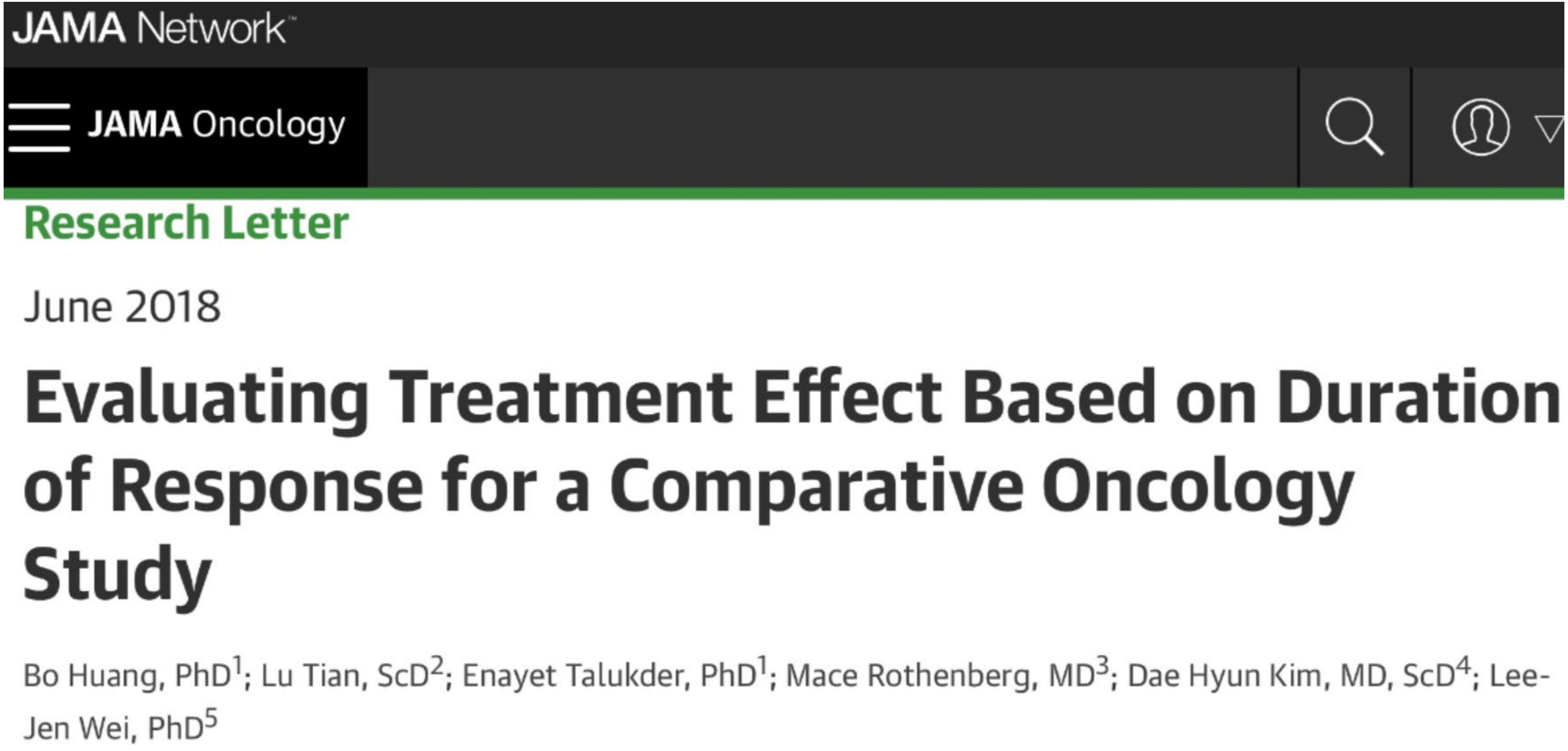
B Duration of Response



No. at Risk

Nivolumab + ipilimumab	63	56	46	32	22	10	5	0
Chemotherapy	43	32	15	5	2	2	1	0


DOR
R
P/D
1
9 months
R
2
21 months
P/D
3
0 month
4
0 month
Case patterns
3
9
15
21
27
30
Months



JAMA Network

JAMA Oncology

Research Letter

June 2018

Evaluating Treatment Effect Based on Duration of Response for a Comparative Oncology Study

Bo Huang, PhD[1]; Lu Tian, ScD[2]; Enayet Talukder, PhD[1]; Mace Rothenberg, MD[3]; Dae Hyun Kim, MD, ScD[4]; Lee-Jen Wei, PhD[5]

Cox model with baseline covariate adjustment (or stratified Cox)

- When two sample PH assumption is ok, the Cox ANCOVA is not valid (incoherent)
- Augmentation procedures (Tsiatis et al.; Tian et al.)
- For stratified analysis, a simple and coherent procedure is available (Tian et al. 2019, Statistics in Med)

Identifying a high value subgroup of patients who benefit from treatment

- How to use patient baseline information to identify a high value subgroup?

How to use the real-world observational study data?

- How to integrate clinical trial data with observational data to evaluate treatment effect/toxicity?

Totality of evidence on the treatment effect/toxicity?

- For each patient, we have response, progression, death, toxicity information, how can be integrate them to create a clinically interpretable study endpoint?

- RMST can be used for designing the study (JAMA-Oncology, Pak et al. 2017)
- Regression analysis for RMST
- R package: survRM2, and SAS: PROC RMSTREG

There is an R package for designing studies with RMST (SSRMST).

- https://cran.r-project.org/web/packages/SSRMST/index.html

Maybe we need to move out of box for design and analysis of clinical studies?

- The crisis of COVID-19 is a great lesson for clinical trialists, regulatory agencies, pharmaceutical industry, academicians
- Need transparency, efficiency, unbiasedness, robustness for studies