JAMA Oncology | Original Investigation

Log-Rank Test vs MaxCombo and Difference in Restricted Mean Survival Time Tests for Comparing Survival Under Nonproportional Hazards in Immuno-oncology Trials
A Systematic Review and Meta-analysis

Pralay Mukhopadhyay, PhD; Jiabu Ye, PhD; Keaven M. Anderson, PhD; Satrajit Roychoudhury, PhD; Eric H. Rubin, MD; Susan Halabi, PhD; Richard J. Chappell, PhD

IMPORTANCE The log-rank test is considered the criterion standard for comparing 2 survival curves in pivotal registrational trials. However, with novel immunotherapies that often violate the proportional hazards assumptions over time, log-rank can lose power and may fail to detect treatment benefit. The MaxCombo test, a combination of weighted log-rank tests, retains power under different types of nonproportional hazards. The difference in restricted mean survival time (dRMST) test is frequently proposed as an alternative to the log-rank under nonproportional hazard scenarios.

OBJECTIVE To compare the log-rank with the MaxCombo and dRMST in immuno-oncology trials to evaluate their performance in practice.

DATA SOURCES Comprehensive literature review using Google Scholar, PubMed, and other sources for randomized clinical trials published in peer-reviewed journals or presented at major clinical conferences before December 2019 assessing efficacy of anti-programmed cell death protein-1 or anti-programmed death/ligand 1 monoclonal antibodies.

STUDY SELECTION Pivotal studies with overall survival or progression-free survival as the primary or key secondary end point with a planned statistical comparison in the protocol. Sixty-three studies on anti-programmed cell death protein-1 or anti-programmed death/ligand 1 monoclonal antibodies used as monotherapy or in combination with other agents in 35 902 patients across multiple solid tumor types were identified.

DATA EXTRACTION AND SYNTHESIS Statistical comparisons (n = 150) were made between the 3 tests using the analysis populations as defined in the original protocol of each trial.

MAIN OUTCOMES AND MEASURES Nominal significance based on a 2-sided .05-level test was used to evaluate concordance. Case studies featuring different types of nonproportional hazards were used to discuss more robust ways of characterizing treatment benefit instead of sole reliance on hazard ratios.

RESULTS In this systematic review and meta-analysis of 63 studies including 35 902 patients, between the log-rank and MaxCombo, 135 of 150 comparisons (90%) were concordant; MaxCombo achieved nominal significance in 15 of 15 discordant cases, while log-rank did not. Several cases appeared to have clinically meaningful benefits that would not have been detected using log-rank. Between the log-rank and dRMST tests, 137 of 150 comparisons (91%) were concordant; log-rank was nominally significant in 5 of 13 cases, while dRMST was significant in 8 of 13. Among all 3 tests, 127 comparisons (85%) were concordant.

CONCLUSIONS AND RELEVANCE The findings of this review show that MaxCombo may provide a pragmatic alternative to log-rank when departure from proportional hazards is anticipated. Both tests resulted in the same statistical decision in most comparisons. Discordant studies had modest to meaningful improvements in treatment effect. The dRMST test provided no added sensitivity for detecting treatment differences over log-rank.

JAMA Oncol. doi:10.1001/jamaoncol.2022.2666
Published online July 21, 2022.

Supplemental content

Author Affiliations: Otsuka America Pharmaceutical, Inc, Rockville, Maryland (Mukhopadhyay); Merck & Co, Inc, Kenilworth, New Jersey (Ye, Anderson, Rubin); Pfizer Inc, New York, New York (Roychoudhury); Duke Cancer Institute, Duke University, Durham, North Carolina (Halabi); Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina (Halabi); Department of Statistics, University of Wisconsin Madison (Chappell); Department of Biostatistics and Medical Informatics, University of Wisconsin Madison (Chappell).

Corresponding Author: Pralay Mukhopadhyay, PhD, Otsuka America Pharmaceutical, Inc, 2440 Research Blvd, Rockville, MD 20850 (pralay.mukhopadhyay@otsuka-us.com).

Novel immunotherapies in oncology have improved patient care and treatment. Several recent trials in immuno-oncology demonstrated nonconstant treatment effect or violation of the proportional hazards assumption.[1-10] These trials showed delayed separation[9] or crossing of Kaplan-Meier curves,[10] suggesting the relative risk of progression or death between immuno-oncology drugs and control were not constant over time, reflective of the biological mechanism of these therapies.

From a statistical perspective, standard approaches, such as the log-rank test, may not be optimal for comparing 2 treatment arms when violation of the proportional hazards assumption is expected. Similarly, hazard ratios (HRs) from the proportional hazards model may be difficult to interpret as a single measure of treatment benefit.

While the log-rank test remains consistent under a variety of patterns of ordered hazards, it is most powerful when the underlying effect satisfies the proportional hazards assumption. Nonproportional hazards may have a considerable impact on the statistical power of the log-rank test to detect a treatment difference.[11] Many alternative tests have been explored, including weighted log-rank tests,[12] weighted Kaplan-Meier test,[13] combination tests,[14,15] milestone survival,[16] and the difference in restricted mean survival time (dRMST) or ratios of RMST.[17,18] For a discussion of these tests, please refer to publications by Lin and colleagues[11] and Roychoudhury and colleagues.[19] The MaxCombo test combines log-rank test (unweighted) and several weighted log-rank tests to handle different types of treatment effect (eg, delayed) and recently demonstrated robust results regarding power under different proportional hazards and nonproportional hazard assumptions.[11]

We conducted a comprehensive analysis comparing log-rank test with MaxCombo using data from randomized immuno-oncology trials of nearly 36 000 patients across different solid tumor types to help clinical trialists and statisticians understand the advantages and limitations associated with MaxCombo and facilitate informed decision-making for primary analyses in phase 3 trials. Performance of dRMST relative to log-rank test and MaxCombo was evaluated because of its interpretability and frequent use in practice. Lastly, we used case studies to examine these 3 tests under different nonproportional hazard scenarios.

Key Points

Question How do the alternative statistical tests MaxCombo and difference in restricted mean survival time tests compare with the log-rank test under nonproportional hazards using data from immuno-oncology trials?

Findings Of 150 comparisons, 90% were concordant between MaxCombo and log-rank, and in 10% of cases, MaxCombo detected nominal significance while log-rank did not. The difference in restricted mean survival time test provided no added sensitivity for detecting treatment differences over log-rank.

Meaning The MaxCombo test provides a pragmatic alternative to log-rank under nonproportional hazards.

Methods

Literature Search and Data Sources

A comprehensive literature review was conducted using criteria outlined in eFigure 1 (PRISMA diagram) and the eMethods in the Supplement. Briefly, we searched for randomized clinical trials with registrational intent published in peer-reviewed journals or presented at major clinical conferences before December 2019 that assessed the efficacy (defined as overall survival or progression-free survival) of anti-programmed cell death protein-1 or anti-programmed death/ligand 1 monoclonal antibodies in solid tumors. This study followed the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) reporting guideline.

Data Sources

Individual patient level data from the trial sponsors were not available. To overcome this limitation, a reverse engineering method was used to extract the event time and censoring information from published Kaplan-Meier curves for all the studies that made these analyses feasible.[20]

Statistical Methods

We performed 150 statistical comparisons with log-rank test, MaxCombo, and dRMST using the primary analysis populations defined in the original protocol, with comparisons made at the 2-sided .05 level. Additional information is provided in eMethods in the Supplement, including weights used for MaxCombo. The MaxCombo test uses a combination of Fleming-Harrington weighted log-rank test and adaptively selects the best test statistic based on underlying data. The primary analysis used MaxCombo 2, which accounts for proportional hazards and delayed treatment effect scenarios, and is considered most relevant for immuno-oncology trials. MaxCombo 3 additionally incorporates the scenario of diminishing effect with longer follow-up and was used as a sensitivity analysis in eFigure 2 in the Supplement. The minimum of the maximum observed event time between the 2 arms was used to select the truncation time for computing dRMST.[18] Additional summary measures besides HRs included dRMST between arms and milestone survival estimates at 6, 12, and 18 months (when estimable), chosen for their clinical relevance across tumor types, particularly in the metastatic setting. Piecewise HRs at intervals of 0 to 6 months and more than 6 months were provided to describe changes in treatment effect over time. Trial maturity was defined as the proportion of patients with events.

For each study, if both tests being compared had 2-sided *P* values ≤.05 (nominally significant) or >.05 (not nominally significant), the results were considered concordant; other outcomes were considered discordant. For studies with discordant results, where MaxCombo was significant but log-rank test was not, the proportional hazards assumption may have been violated. We tested for this using the Grambsch-Therneau test at the .05 level.[21]

Table 1. Concordance Among Log-Rank (LR), MaxCombo 2 (MC2), and Difference in Restricted Mean Survival Time (dRMST) Tests for 150 Comparisons

Test result	No. (%)
LR,[a] MC2,[a] dRMST[a]	93 (62.0)
LR,[b] MC2,[b] dRMST[b]	34 (22.7)
LR,[b] MC2,[a] dRMST[b]	10 (6.7)
LR,[a] MC2,[a] dRMST[b]	5 (3.3)
LR,[b] MC2,[a] dRMST[a]	5 (3.3)
LR,[b] MC2,[b] dRMST[a]	3 (2)
LR,[a] MC2,[b] dRMST[a]	0
LR,[a] MC2,[b] dRMST[b]	0

[a] $P \leq .05$.

[b] $P > .05$.

Table 2. Overall Concordance Between Either MaxCombo 2 (MC2) or Difference in Restricted Mean Survival Time (dRMST) Compared With Log-Rank Test (LRT)

	LRT, No. (%)	
	Significant[a]	Not significant[b]
MC2		
Significant[a]	98 (65.3)[c]	15 (10.0)
Not significant[b]	0	37 (24.7)[c]
dRMST		
Significant[a]	93 (62.0)[c]	8 (5.3)
Not significant[b]	5 (3.3)	44 (29.3)[c]

[a] $P \leq .05$.

[b] $P > .05$.

[c] Concordant results between the indicated tests using a 2-sided .05-level test.

[d] Percentages are based on 150 comparisons.

Results

This systematic review and meta-analysis of 63 studies included 35 902 patients with non–small cell lung cancer, small cell lung cancer, squamous cell carcinoma of the head and neck, melanoma, kidney cell carcinoma, and tumors of gastrointestinal, gastrourinary, breast, and ovarian origin (PRISMA diagram, eFigure 1 in the Supplement). Studies of the following anti-anti-programmed cell death protein-1 and anti-programmed death/ligand 1 therapies were considered, as monotherapy or in combination with other agents: pembrolizumab, nivolumab, atezolumab, durvalumab, avelumab, cemiplimab, tislelizumab, camrelizumab, and sintilimab.

Table 1 shows 8 possible combinations of outcomes ($P \leq .05$; $P > .05$) from the 3 tests (log-rank test, MaxCombo 2, and dRMST). In 127 comparisons (85%), all 3 tests were concordant (93 [62%] achieved nominal significance using all 3 tests, whereas 34 [23%] did not), while 23 (15%) were discordant. MaxCombo 2 and dRMST differed in 18 (12%) of the 150 comparisons.

Of the 150 statistical comparisons between log-rank test and MaxCombo 2, 135 comparisons (90%) were concordant and 15 (10.0%) were discordant (**Table 2**). In these 15 comparisons, MaxCombo 2 achieved nominal significance, while the log-rank test did not. However, these 15 represented 29% of cases where log-rank test did not reach the nominal significance level. In 12 cases (80%), the Grambsch-Therneau test was statistically significant at the .05 level, suggesting that the proportional hazards assumption was not met. Notably, the trial maturity had been reached for the 15 discordant cases, with 12 of 15 cases having 70% or more patients with events (eTable 1 in the Supplement). The median (range) HR in these studies was 0.86 (0.74-0.92) (eTable 2 in the Supplement). The conclusions were identical using MaxCombo 3.

In 13 of 150 statistical comparisons (8.7%) between dRMST and log-rank test, the outcomes were not concordant (Table 2). In 8 of the 13 cases (5.3%), dRMST was nominally significant, while the log-rank test was not. In the remaining 5 (3.3%) cases, log-rank test was nominally significant, but there was no difference in dRMST.

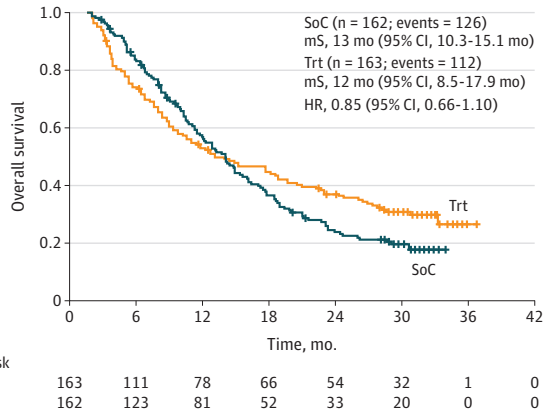
eFigure 2 in the Supplement shows the *P* values from log-rank test vs MaxCombo 2, log-rank test vs MaxCombo 3, and log-rank test vs dRMST. eFigure 3 in the Supplement shows the 15 comparisons where the outcomes from log-rank test and MaxCombo 2 were discordant. Through visual inspection, we separated them into 3 groups: severe crossing of Kaplan-Meier curves in 8 comparisons, moderate crossing of Kaplan-Meier curves in 4 comparisons, and delayed separation of Kaplan-Meier curves in 3 comparisons.

eTable 1 in the Supplement shows the *P* values for log-rank test, MaxCombo 2, dRMST, and Grambsch-Therneau tests, as well as trial maturity. eTable 2 in the Supplement shows the HRs over the entire treatment period for these 23 studies with any discordance, along with HRs in the interval of 0 to 6 months and more than 6 months; the Kaplan-Meier survival estimates at 6, 12, and 18 months (when estimable); and the dRMST between arms.

eFigure 4A in the Supplement provides the Kaplan-Meier plots for the 8 studies where dRMST but not log-rank test was statistically significant. Of these 8 studies, the median (range) HR was 0.86 (0.78-0.92), while the median (range) dRMST was 1.45 (0.85-2.06). Five of these comparisons were for progression-free survival with a median (range) HR of 0.86 (0.84-0.92), while the median (range) dRMST was 1.45 (0.85-1.92) months. These 5 cases were also significant using MaxCombo 2. The remaining 3 comparisons were for overall survival with HRs of 0.78, 0.80, and 0.85, while dRMST was 2.04, 1.39, and 2.06 months, respectively. None of these 3 cases achieved nominal significance using MaxCombo 2. Of note, the Grambsch-Therneau test that measures departure from proportional hazards assumption was also not significant (eTable 1 in the Supplement).

In 5 studies where dRMST was not nominally significant while the log-rank test and MaxCombo 2 tests were significant, the median (range) HR was 0.75 (0.58-0.83), and the median (range) dRMST was 1.81 (1.42-3.94) months (eTable 2 and eFigure 4B in the Supplement). Among the cases with concordance using a 2-sided test, there were no instances where the inference would have differed if a 1-sided test was used.

Figure 1. Overall Survival in First-Line Non–Small Cell Lung Cancer in All Randomized Patients Regardless of PD-L1 Expression



No. at risk

Trt	163	111	78	66	54	32	1	0
SoC	162	123	81	52	33	20	0	0

Data from the Phase III Open Label First Line Therapy Study of MEDI 4736 (Durvalumab) With or Without Tremelimumab Versus SOC in Non Small-Cell Lung Cancer (NSCLC) (MYSTIC) trial.[23] Hazard ratio (HR) and 95% CI computed from an unstratified Cox model. PD-L1 indicates programmed death/ligand 1; mS, median survival; SoC, standard of care; Trt, treatment.

Select Case Studies

From the 15 discordant studies between log-rank test and MaxCombo 2, we selected 3 case studies, 1 from each of the 3 types of nonproportional hazard scenarios. In studies with more than 2 arms, only the comparisons of interest where the tests differed in outcome were addressed. Two case studies are discussed below. A third case study depicting delayed separation of progression-free survival Kaplan-Meier curves in first-line gastric cancer from the Study of Pembrolizumab (MK-3475) as First-Line Monotherapy and Combination Therapy for Treatment of Advanced Gastric or Gastroesophageal Junction Adenocarcinoma (KEYNOTE-062) trial[22] is described in eResults in the Supplement and shown in eFigure 6 in the Supplement. Plots of the cumulative hazard and log(HR) over time for all 3 case studies are shown in eFigure 7 in the Supplement.

Nonproportional Hazards Scenario 1: Substantial Crossing of Overall Survival Kaplan-Meier Curves in First-Line Non–Small Cell Lung Cancer

The Phase III Open Label First Line Therapy Study of MEDI 4736 (Durvalumab) With or Without Tremelimumab Versus SOC in Non Small-Cell Lung Cancer (NSCLC) (MYSTIC) trial[23] was a study of durvalumab alone or plus tremelimumab vs chemotherapy alone in first-line non-small cell lung cancer. **Figure 1** compares the intent-to-treat population (regardless of anti-programmed death/ligand 1 expression) using durvalumab plus tremelimumab vs chemotherapy alone. For the comparison of interest, the study randomized 163 patients in the durvalumab plus tremelimumab arm and 162 in the chemotherapy arm. The analysis was conducted after 112 (68.7%) and 126 (77.8%) events were observed in the experimental and control arms, respectively. Overall HR was 0.85 (95% CI, 0.66-1.10), suggesting a possible modest survival benefit. The HR in the first 0 to 6 months was 1.55 (95% CI, 1.00-2.38) and after 6 months, 0.61 (95% CI, 0.44-0.84). Median survival was 12.0 (95% CI, 8.5-17.9) and 13.0 (95% CI, 10.3-15.1) months in the treatment and standard of care groups, respectively. This example illustrates the challenge of using either single measure (HR or median) to describe treatment benefit when there is a strong departure from proportional hazards.

Although HRs have been criticized as not having a mechanistic interpretation,[24] restricted means can be computed for distinct portions of the timeline and then added to yield a unified estimate of benefit or harm during follow-up.[25,26] In this case, dRMST up to 1 year was −0.85 months and 2.03 months beyond a year, indicating the change in treatment efficacy over time. The summed dRMST was 1.19 months (95% CI, −1.44 to −3.81; $P = .37$), indicating the possibility of a small combined benefit. However, like the overall HR, the combined benefit in dRMST may not be a useful descriptor for patients given the obvious heterogeneity in treatment effect across the population.

The overall survival estimates for treatment and control groups at 6, 12, and 18 months were 69%, 50%, and 42% compared with 78%, 52%, and 34%, respectively, suggesting higher risk of early mortality in the experimental arm with a potentially important subset of patients who benefited long-term. Despite MaxCombo tests being nominally significant, the benefit was clearly modest and likely restricted to a subgroup of patients, thereby leading to a substantial crossing. Both MaxCombo 2 and MaxCombo 3 tests were positive owing to the Fleming-Harrington (with weights, 0,0.5) test statistic, which applies more weight to late events.

Nonproportional Hazard Scenario 2: Moderate Crossing of Overall Survival Kaplan-Meier Curves in First-Line Non–Small Cell Lung Cancer

The Study of Atezolizumab Compared With a Platinum Agent (Cisplatin or Carboplatin) + (Pemetrexed or Gemcitabine) in Participants With Stage IV Non-Squamous or Squamous Non-Small Cell Lung Cancer (Impower 110) study[27] evaluated atezolizumab alone vs chemotherapy alone in anti-programmed death/ligand 1 selected first-line non-small cell lung cancer (NCT02409342). The comparison shown in **Figure 2** is for the anti-programmed death/ligand 1+ (≥1% positivity of tumor cells) population using the Ventana SP142 assay. For the comparison of interest, the study randomized 166 patients in the atezolizumab arm and 162 in the chemotherapy arm. The analysis was conducted after 71 (42.8%) and 84 (51.9%) events were observed in the experimental and control arms, respectively. Although there was a modest crossing during the first 4 months (HR, 1.04; 95% CI, 0.85-1.27), there was a clinically meaningful overall treatment benefit with HR of 0.74 (95% CI, 0.54-1.02). However, log-rank test failed to reach nominal significance in this case, while both MaxCombo 2 and MaxCombo 3 did. As in case study 1, MaxCombo 2 and MaxCombo 3 tests were positive owing to the Fleming-Harrington (with weights, 0,0.5) test statistic. dRMST was 2.99 months (95% CI, −0.17 to 6.14; $P = .06$). HR in the first 6 months was 0.89 (95% CI, 0.56-1.41) and 0.64 (95% CI, 0.41-0.98) beyond 6 months.

Discussion

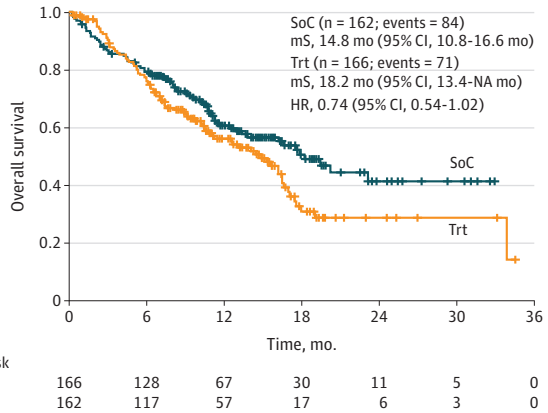
In this systematic review and meta-analysis, we evaluated the implication of using the MaxCombo test in lieu of the log-rank test in practice as a primary method for analyses. Our comprehensive analysis shows that in 90% of cases, the 2 tests came to the same conclusion at the 2-sided .05 level. However, in 10% of scenarios, primary outcomes differed; in these discordant cases, only MaxCombo was found to be nominally significant. On further review of the discordant comparisons, in many instances the benefit was deemed modest and may not have met standards for regulatory approval unless supported by additional clinical evidence. However, in some cases, the benefit was arguably strong. We suggest that such cases should often not be removed from consideration for regulatory approval based on negative log-rank test without careful evaluation of the potential risk-benefit profile. In our findings, dRMST and log-rank test were largely concordant.

Whether the use of weighted log-rank tests is clinically appropriate is an important consideration. However, immuno-oncology agents may address areas of high unmet need for patients. If a subgroup of patients might gain considerable long-term treatment benefit, arguably more emphasis should be placed on late events when looking for statistical significance, if the mechanism of the drug supports such findings (eg, immuno-oncology agents). Prespecification of the MaxCombo and any subsequent testing on subgroups of interest using appropriate type I error control are necessary to generate robust data for decision-making. We recommend MaxCombo 2 as primary and MaxCombo 3 as sensitivity analyses for immuno-oncology trials.

Not all patients derive equal benefit from new immuno-oncology therapies, resulting in a heterogeneous population. Therefore, the current practice of using a single test or single summary to compare treatment effect is not sufficiently nuanced as evidenced by the various case studies we discussed, where the use of a single HR to summarize treatment benefit was misleading. Although MaxCombo does not have a corresponding easy-to-interpret summary statistic like an HR for log-rank test, supporting MaxCombo with descriptive summary statistics may be valuable for understanding overall treatment benefits. Beyond the measures already described, the net chance of survival benefit[28,29] can be used as an intuitive measure of treatment benefit for patients and physicians. We suggest that no single number adequately describes treatment differences when there is a changing effect over time.

It is critical that a statistical test control for type I (false-positive) error. Based on a previous review,[19] the modified MaxCombo, which has been applied here, will control for type I error even in extreme scenarios[30,31] (eg, the hypothetical scenario in Freidlin and Korn,[30] where the HR was 3.53 in the first 6 months and 0.75 beyond 6 months). Notably, we did not observe such extreme cases in practice; the closest was progression-free survival in KEYNOTE-061 (eFigure 5A in the Supplement) with an HR of 1.67 in the first 6 months and 0.28 thereafter. In this case, none of the 3 tests were nominally significant (eFigure 5B in the Supplement).

Figure 2. Overall Survival in Patients With PD-L1–Positive First-Line Non–Small Cell Lung Cancer (≥1% Tumor Cell Positivity)



No. at risk

Trt	166	128	67	30	11	5	0
SoC	162	117	57	17	6	3	0

Data from the Study of Atezolizumab Compared With a Platinum Agent (Cisplatin or Carboplatin) + (Pemetrexed or Gemcitabine) in Participants With Stage IV Non-Squamous or Squamous Non-Small Cell Lung Cancer (Impower 110) study.[27] Hazard ratio (HR) and 95% CI computed from an unstratified Cox model. PD-L1 indicates programmed death/ligand 1; mS, median survival; NA, not applicable; SoC, standard of care group; Trt, treatment group.

An agent with a modest treatment benefit in the overall population may show a statistically positive outcome, providing an opportunity to explore underlying subpopulations that may benefit from the experimental therapy. In immuno-oncology trials, treatment effects may be more modest in the overall population compared with the anti-programmed death/ligand 1+ subgroup (eg, the Efficacy Study of Nivolumab Plus Ipilimumab or Nivolumab Plus Chemotherapy Against Chemotherapy in Stomach Cancer or Stomach/Esophagus Junction Cancer [CheckMate 649 study]).[32] However, when developing a novel agent, subpopulations may not be well understood. If the primary end point of a prospective phase 3 study is negative, there is minimal opportunity for further exploration in a regulatory setting without conducting another study. While this principle of rejecting a negative trial is generally sound, in some cases where overall survival benefit is considered substantial (eg, atezolizumab in first-line non-small cell lung cancer), the trial would have been deemed statistically negative using log-rank test.

Shen et al[33] recently reported that MaxCombo could not identify underlying causes of nonproportional hazards or identify populations in which the benefit exists and is therefore challenging to use for drawing inference. However, no statistical test can be completely relied on for making inference from a clinical perspective. Inference must be based on comprehensive evaluation of the data, while the statistical test is merely an initial gatekeeper. Although log-rank test is less sensitive under nonproportional hazards, with a sufficiently large sample size the log-rank test will also be statistically significant, as seen in the First Line IRESSA Versus Carboplatin/Paclitaxel in Asia (IPASS) study.[34]

Careful power considerations are needed for studies with anticipated nonproportional hazards, such as delayed

separation,[19,35] because off-base assumptions can lead to an underpowered study. In these cases, an adequate duration of follow-up and number of events planned before trial initiation are critical. Prespecification of an analysis following the final time point could further help elucidate long-term benefit. A viable multiprong approach would involve consideration of a delayed treatment effect during trial design and use of MaxCombo for primary analyses.

In our review of dRMST vs log-rank test, dRMST was not more sensitive than log-rank test. Although dRMST was concordant with log-rank test in most cases, in the 5 cases where dRMST differed from log-rank test and failed to reach statistical significance, the overall survival benefit appeared clinically meaningful with HR ranging from 0.58-0.83.

Limitations

A potential limitation of this study was using unstratified log-rank test and MaxCombo tests for each comparison and not using the stratified test or multiplicity adjustments as in the original protocols. Therefore, our conclusions may not align with the actual study results.

Conclusions

The findings of this study suggest that MaxCombo may be a pragmatic alternative to log-rank test when nonproportional hazards are anticipated. Log-rank test and MaxCombo reached the same statistical decision in most comparisons. In discordant cases, the associated clinical benefit ranged from modest to meaningful improvements in treatment effect. Notably, dRMST was not more sensitive for detecting treatment differences over log-rank test. HRs can be misleading in summarizing overall treatment benefit under nonproportional hazards and should be supplemented with additional measures (eg, milestone survival differences).

ARTICLE INFORMATION

Accepted for Publication: May 11, 2022.

Published Online: July 21, 2022. doi:10.1001/jamaoncol.2022.2666

Author Contributions: Dr Mukhopadhyay had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Co-senior authorship: Drs Halabi and Chappell.
Concept and design: Mukhopadhyay, Ye, Anderson, Roychoudhury, Chappell.
Acquisition, analysis, or interpretation of data: Mukhopadhyay, Ye, Rubin, Halabi, Chappell.
Drafting of the manuscript: Mukhopadhyay, Ye, Roychoudhury, Halabi.
Critical revision of the manuscript for important intellectual content: Mukhopadhyay, Ye, Anderson, Rubin, Halabi, Chappell.
Statistical analysis: Mukhopadhyay, Ye, Anderson, Roychoudhury, Chappell.
Administrative, technical, or material support: Ye, Roychoudhury.
Supervision: Mukhopadhyay.

Conflict of Interest Disclosures: Dr Mukhopadhyay is an employee and shareholder (or equivalent) of Otsuka America Pharmaceutical, Inc. Drs Ye and Rubin are employed by Merck Sharp & Dohme, LLC, a subsidiary of Merck & Co, Inc. Dr Halabi is a member of data monitoring committees for Sanofi, Eisai, and Ferring and reports grants from American Society of Clinical Oncology Targeted Agent and Profiling Utilization Registry, Duke University Medical Center, during the conduct of the study. No other disclosures were reported.

Funding/Support: This research and article were funded and supported by Otsuka Pharmaceutical Development & Commercialization, Inc.

Role of the Funder/Sponsor: The study sponsor was involved in all aspects of this study and manuscript preparation.

Additional Contributions: Editorial support for this manuscript was provided by Shawna Matthews, PhD, Oxford PharmaGenesis, Newtown, PA. Oxford was compensated for this work by Otsuka Pharmaceutical Development & Commercialization, Princeton, NJ.

REFERENCES

1. Kantoff PW, Higano CS, Shore ND, et al; IMPACT Study Investigators. Sipuleucel-T immunotherapy for castration-resistant prostate cancer. *N Engl J Med*. 2010;363(5):411-422. doi:10.1056/NEJMoa1001294

2. Motzer RJ, Escudier B, McDermott DF, et al; CheckMate 025 Investigators. Nivolumab versus everolimus in advanced renal-cell carcinoma. *N Engl J Med*. 2015;373(19):1803-1813. doi:10.1056/NEJMoa1510665

3. Ribas A, Kefford R, Marshall MA, et al. Phase III randomized clinical trial comparing tremelimumab with standard-of-care chemotherapy in patients with advanced melanoma. *J Clin Oncol*. 2013;31(5):616-622. doi:10.1200/JCO.2012.44.6112

4. Wolchok JD, Neyns B, Linette G, et al. Ipilimumab monotherapy in patients with pretreated advanced melanoma: a randomised, double-blind, multicentre, phase 2, dose-ranging study. *Lancet Oncol*. 2010;11(2):155-164. doi:10.1016/S1470-2045(09)70334-1

5. Larkin J, Chiarion-Sileni V, Gonzalez R, et al. Combined nivolumab and ipilimumab or monotherapy in untreated melanoma. *N Engl J Med*. 2015;373(1):23-34. doi:10.1056/NEJMoa1504030

6. Postow MA, Chesney J, Pavlick AC, et al. Nivolumab and ipilimumab versus ipilimumab in untreated melanoma. *N Engl J Med*. 2015;372(21):2006-2017. doi:10.1056/NEJMoa1414428

7. Fehrenbacher L, Spira A, Ballinger M, et al; POPLAR Study Group. Atezolizumab versus docetaxel for patients with previously treated non-small-cell lung cancer (POPLAR): a multicentre, open-label, phase 2 randomised controlled trial. *Lancet*. 2016;387(10030):1837-1846. doi:10.1016/S0140-6736(16)00587-0

8. Ferris RL, Blumenschein G Jr, Fayette J, et al. Nivolumab for recurrent squamous-cell carcinoma of the head and neck. *N Engl J Med*. 2016;375(19):1856-1867. doi:10.1056/NEJMoa1602252

9. Robert C, Long GV, Brady B, et al. Nivolumab in previously untreated melanoma without BRAF mutation. *N Engl J Med*. 2015;372(4):320-330. doi:10.1056/NEJMoa1412082

10. Borghaei H, Paz-Ares L, Horn L, et al. Nivolumab versus docetaxel in advanced nonsquamous non-small-cell lung cancer. *N Engl J Med*. 2015;373(17):1627-1639. doi:10.1056/NEJMoa1507643

11. Lin RS, Lin J, Roychoudhury S, et al. Alternative analysis methods for time to event endpoints under nonproportional hazards: a comparative analysis. *Stat Biopharm Res*. 2020;12(2):187-198. doi:10.1080/19466315.2019.1697738

12. Harrington DP, Fleming TR. A class of rank test procedures for censored survival data. *Biometrika*. 1982;69(3):553-566. doi:10.1093/biomet/69.3.553

13. Pepe MS, Fleming TR. Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. *Biometrics*. 1989;45(2):497-507. doi:10.2307/2531492

14. Breslow NE, Edler L, Berger J. A two-sample censored-data rank test for acceleration. *Biometrics*. 1984;40(4):1049-1062. doi:10.2307/2531155

15. Logan BR, Klein JP, Zhang MJ. Comparing treatments in the presence of crossing survival curves: an application to bone marrow transplantation. *Biometrics*. 2008;64(3):733-740. doi:10.1111/j.1541-0420.2007.00975.x

16. Klein JP, Logan B, Harhoff M, Andersen PK. Analyzing survival curves at a fixed point in time. *Stat Med*. 2007;26(24):4505-4519. doi:10.1002/sim.2864

17. Royston P, Parmar MK. 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. doi:10.1186/1471-2288-13-152

18. 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(22):2380-2385. doi:10.1200/JCO.2014.55.2208

19. Roychoudhury S, Anderson Kaplan-Meier, Ye J, Mukhopadhyay P. Robust design and analysis of clinical trials with nonproportional hazards: a straw man guidance from a cross-pharma working group. *Stat Biopharm Res*. Published online March 4, 2021. doi:10.1080/19466315.2021.1874507

20. Guyot P, Ades AE, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol*. 2012;12:9. doi:10.1186/1471-2288-12-9

21. Therneau TM, Grambsch PM. The Cox model. In: *Modeling survival data: extending the Cox model. Statistics for Biology and Health*. Springer; 2000. doi:10.1007/978-1-4757-3294-8_3

22. Shitara K, Van Custem E, Bang YJ, et al. Efficacy and safety of pembrolizumab or pembrolizumab plus chemotherapy vs chemotherapy alone for patients with first-line, advanced gastric cancer: the KEYNOTE-062 phase 3 randomized clinical trial. *JAMA Oncol*. 2020;6(10):1571-1580. doi:10.1001/jamaoncol.2020.3370

23. Rizvi NA, Cho BC, Reinmuth N, et al; MYSTIC Investigators. Durvalumab with or without tremelimumab vs standard chemotherapy in first-line treatment of metastatic non-small cell lung cancer: the MYSTIC phase 3 randomized clinical trial. *JAMA Oncol*. 2020;6(5):661-674. doi:10.1001/jamaoncol.2020.0237

24. Cox DR. Some remarks on the analysis of survival data. Paper presented at: Proceedings of the First Seattle Symposium in Biostatistics; 1997.

25. Paukner M, Chappell R. Window mean survival time. *Stat Med*. 2021;40(25):5521-5533. doi:10.1002/sim.9138

26. Horiguchi M, Tian L, Uno H, et al. Quantification of long-term survival benefit in a comparative oncology clinical study. *JAMA Oncol*. 2018;4(6):881-882. doi:10.1001/jamaoncol.2018.0518

27. Herbst RS, Giaccone G, de Marinis F, et al; MYSTIC Investigators. Atezolizumab for First-Line Treatment of PD-L1-Selected Patients with NSCLC. *N Engl J Med*. 2020;383(14):1328-1339. doi:10.1056/NEJMoa1917346

28. Buyse M. Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Stat Med*. 2010;29(30):3245-3257. doi:10.1002/sim.3923

29. Péron J, Roy P, Ozenne B, Roche L, Buyse M. The net chance of a longer survival as a patient-oriented measure of treatment benefit in randomized clinical trials. *JAMA Oncol*. 2016;2(7):901-905. doi:10.1001/jamaoncol.2015.6359

30. Freidlin B, Korn EL. Methods for accommodating nonproportional hazards in clinical trials: ready for the primary analysis? *J Clin Oncol*. 2019;37(35):3455-3459. doi:10.1200/JCO.19.01681

31. Magirr D, Burman CF. Modestly weighted log-rank tests. *Stat Med*. 2019;38(20):3782-3790. doi:10.1002/sim.8186

32. Moehler M, Shitara K, Garrido M, et al. LBA6_PR nivolumab (nivo) plus chemotherapy (chemo) versus chemo as first-line (1L) treatment for advanced gastric cancer/gastroesophageal junction cancer (GC/GEJC)/esophageal adenocarcinoma (EAC): first results of the CheckMate 649 study. *Ann Oncol*. 2020;31:S1191. doi:10.1016/j.annonc.2020.08.2296

33. Shen Y-L, Wang X, Mushti S, et al. Non-proportional hazards—an evaluation of the MaxCombo test in cancer clinical trials. *Stat Biopharm Res*. Published online January 4, 2022. doi:10.1080/19466315.2021.2008485

34. Mok TS, Wu YL, Thongprasert S, et al. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *N Engl J Med*. 2009;361(10):947-957. doi:10.1056/NEJMoa0810699

35. Mukhopadhyay P, Huang W, Metcalfe P, Öhrn F, Jenner M, Stone A. Statistical and practical considerations in designing of immuno-oncology trials. *J Biopharm Stat*. 2020;30(6):1130-1146. doi:10.1080/10543406.2020.1815035