Methods for Accommodating Nonproportional Hazards in Clinical Trials: Ready for the Primary Analysis?

Boris Freidlin, PhD[1] and Edward L. Korn, PhD[1]

INTRODUCTION

Evaluation of new anticancer therapies in randomized clinical trials (RCTs) is typically based on comparing a new treatment with a standard one, using a time-to-event end point such as overall survival or progression-free survival (PFS). Although the statistical framework underlying the design of these RCTs is centered on formal testing of a treatment effect, methods for estimation (quantification) of the treatment benefit are also specified. Currently, log-rank statistical tests and/or proportional hazards models are commonly used for the trial design and primary analysis. These methods are optimized for treatment effects that do not change substantially over time (the proportional hazard assumption).

Introduction of immunotherapeutic agents with potentially delayed treatment effects has renewed interest in statistical methods that can better accommodate general departures from proportional hazards and, particularly, a delayed treatment effect. This has led to considerable attention in, and some controversy about, appropriate statistical methodology for comparing survival curves, as demonstrated by the comments and replies on trial reports[1-24] and at a Duke–US Food and Drug Administration workshop[25] that offered alternatives to the standard log-rank/hazard-ratio methodology. While these new methods could be useful, as outlined in comprehensive reviews,[26-30] we offer a caution about some of these methods' limitations in translating statistical evidence into clinical evidence, both for formal treatment-effect hypothesis testing and for estimation (when used for the primary analysis).

TESTING TREATMENT EFFECTS

The two most commonly discussed treatment-effect testing approaches developed to be sensitive to departures from the proportional hazards assumption are based on weighted log-rank tests and restricted mean survival times (RMSTs). Weighted log-rank tests allow one to give more weight to emphasize a particular part of the survival curve, in contrast to the standard log-rank test, which weights all parts of the survival curves equally. For example, one of the first weighted log-rank tests, the generalized Wilcoxon test,[31,32] gives more weight to the early portions of the survival curve and thus is sometimes recommended for situations in which the treatment effect may dissipate over time.[33] In addition to the early-emphasis Wilcoxon test ($G^{1,0}$), the general family of weighted log-rank tests[34] includes the standard log-rank test and a late-emphasis test ($G^{0,1}$), which gives more weight to the later portions of the survival curves. Accordingly, the late-emphasis test has been suggested for situations in which the treatment effect potentially may be delayed.[26,35]

As an example, consider the PFS curves from the KEYNOTE-042 trial[36] (Fig 1), which compared pembrolizumab with chemotherapy in first-line, metastatic non–small-cell lung cancer. This is a good example for evaluating alternative methods, because the observed survival curves cross, implying nonproportional hazards. The standard log-rank test is not significant with a hazard ratio (HR) estimate of 1.07. Using the late-emphasis test in this immunotherapy setting allows one to focus the comparison of the separation in the tails of the PFS curves, rejecting the null hypothesis in *favor* of pembrolizumab with a one-sided $P < .0001$ (individual patient data are reconstructed[37] from Fig 1). The drawback with this approach, however, is that if the treatment effect is not delayed, then using the late-emphasis test will result in a considerable loss of power as compared with the standard log-rank test. Moreover, because it down-weights the early events, the late-emphasis test does not properly account for existence of early harm, potentially leading to clinically incorrect conclusions. For example, it is possible to have the experimental-arm survival curve always below the control-arm curve but with the late-emphasis test rejecting the null hypothesis in *favor* of the experimental-treatment arm (Fig A1).

To avoid the potential loss of power with using the late-emphasis test, numerous versatile testing procedures have been developed that involve multiple weighted log-rank tests.[38-40] For example, Karrison[40] suggested using the maximum of the log-rank, early-emphasis, and late-emphasis tests as the test statistic. However, as noted by Karrison,[40] these tests can reject the null hypothesis both in favor of the experimental treatment and in favor of control treatment on the same data.

ASSOCIATED CONTENT

Appendix

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on October 3, 2019 and published at jco.org on October 24, 2019: DOI https://doi.org/10.1200/JCO.19.01681

Published by American Society of Clinical Oncology

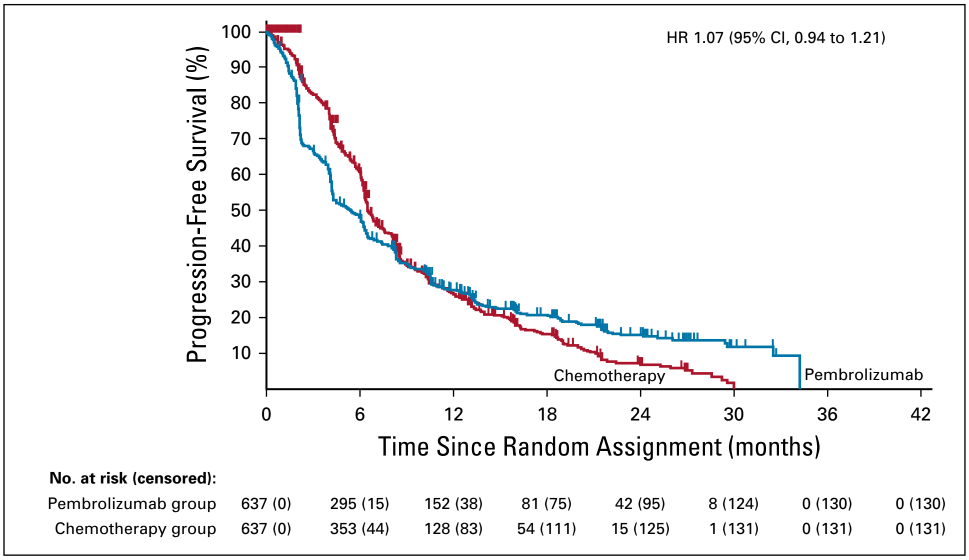

FIG 1. Progression-free survival curves for patients with non–small-cell lung cancer in the KEYNOTE-042 trial for the PD-L1 TPS 1% or greater population. This is Figure 3C from Mok et al.[36] The individual patient data were reconstructed by digitizing the progression-free survival curves using WebPlotDigitizer (https://automeris.io/WebPlotDigitizer/) and then using the algorithm from Guyot et al.[37] All statistical tests presented in this paper were done using these reconstructed data. HR, hazard ratio.

Indeed, when applied to KEYNOTE-042[36,37] to test the superiority of pembrolizumab, the maximum test rejects the null hypothesis in favor of pembrolizumab (one-sided $P <$ .0001); at the same time, when applied to the same data to test the superiority of chemotherapy, the maximum test rejects the null hypothesis in favor of chemotherapy (one-sided $P <$ .0001). The same result is obtained from the Max-Combo test,[25] which additionally includes a test ($G^{1,1}$) that gives more weight to the middle portion of the survival curves. This is an unfortunate situation, given that there would not appear to be any clinically meaningful *overall* advantage established for either arm. (Note that although the curves suggest a potential subpopulation that may benefit from pembrolizumab, this subpopulation needs to be prospectively identified to improve treatment.)

The other commonly discussed approach to accommodate nonproportional hazards is the RMST, which graphically corresponds to the area under the Kaplan-Meier curve over a specified time.τ.[41-44] An RMST test is based on the area between the experimental arm and the control arm Kaplan-Meier curves up to time τ: the larger the area, the greater the treatment effect. Although there is no proportional hazards assumption required for this analysis, it does have the major limitation that the area between the curves is only calculated up to a specified time τ, and the statistical significance of the results depends on the chosen τ.[16,22] For example, in comments[1,6,8,12,21] on trial reports published in *Journal of Clinical Oncology* that suggested using the RMST, the chosen τs were 15, 21, 24, 45, 48 and 108 months. Because the selection of τ in these comments appears to be based on the observed Kaplan-Meier curves, the statistical significance quoted for the RMST is potentially exaggerated if the τs were chosen to maximize the statistical significance. On the other hand, prospectively selecting τ before the study starts can be challenging because of the uncertainties about what the survival curves will look like, and a poor choice could result in dramatically reduced power.

To address the difficulty of choosing τ for the RMST methodology, versatile RMST methods have been developed to allow choosing from a range of τ values to maximize the observed treatment-arm difference while accounting for this data-driven choice of τ in the calculation of the P value.[29,45] Although these methods will produce a valid P value in terms of type 1 error, they have the same flaw as the versatile weighted log-rank tests: They can sometimes yield a statistically significant result that is clinically meaningless by focusing exclusively on a particular part of the survival curves. For example, application of a versatile RMST procedure[45] to KEYNOTE-042 data (Fig 1)[36,37] rejects the null hypothesis that the curves are equal in favor of chemotherapy, with a one-sided $P <$ .0001; this is because the procedure intentionally selects the chemotherapy-arm advantage in the first 8 months as the most statistically relevant portion of the curves.

Modern definitive (ie, phase III) RCTs follow a set of design and conduct practices (eg, prespecification of a primary outcome, formal interim analysis plans, sample size with sufficient power against clinically meaningful alternatives) that ensure a statistically significant result will correspond to a clinically significant result (with few exceptions).[46] Thus, one should be cautious about abandoning the log-rank test in favor of tests that can squeeze out more statistically significant results from observations that have no clinical significance. In specific situations where the clinical interest focuses on a specific aspect of the survival curves, an appropriate test should be used regardless of whether there are proportional hazards. For example, in RTOG-0534,[47] for patients with prostate cancer with an increasing prostate-specific antigen level after prostatectomy, the primary end point was the freedom from biochemical progression at 5 years, because it was thought that short-term delays in progression were not as clinically relevant as long-term freedom from progression. In other situations in which it is known that the intervention cannot affect early events, a weighted log-rank test that down-weights early events may also be appropriate. Examples include some screening trials (eg, National Lung Screening Trial[48]) and prevention trials (eg, Women's Heath Trial).[49]

ESTIMATION OF TREATMENT EFFECTS

Even with proportional hazards, the HR is not a particularly intuitive measure of treatment effect. (The exception is when the survival curves are approximately exponential in shape, in which case the HR is the ratio of the control and experimental arm medians.) RMST-based approaches are among several possible summary measures that may complement the estimated HR.[30] However, contrary to its proponents,[28,50,51] we find it lacking as a particularly insightful summary of the clinical benefit of an experimental treatment. First, although mathematically well defined, the notion of a restricted mean has no common-sense interpretation: What is the clinical interpretation of "the mean survival time to some prespecified time point [τ]"[50]? Second, use of RMST for estimation requires prespecification of τ, which can dramatically change the value of the estimated treatment effect. For example, consider an RCT in which 70% of patients in the control arm and 90% of those in the experimental arm are cured at 2 years (and the curves are approximately exponential up to 2 years): the RMST difference is 2.6, 5, and 7.4 months when τ is 2, 3, and 4 years, respectively. Given these changing values, it is not clear how the approach elucidates the clinical impact of the 20% increase in cure rates.

As with other summary measures, the RMST value without the context of the survival curves could be misleading about clinical significance of the experimental treatment. For example, consider a trial in which metronomic chemotherapy was compared with placebo in progressive pediatric malignant solid tumors.[52] No significant improvement was reported for the metronomic therapy, with median PFS of 49 and 46 days in the experimental and control arms, respectively (HR, 0.69; 95% CI, 0.47 to 1.03; log-rank $P = .07$). An RMST reanalysis[10] reported a 0.8-month difference in PFS RMST (2.4 *v* 1.6 months for the metronomic and placebo arms, respectively), which was statistically significant ($P = .02$). Fang et al[10] concluded that their RMST analysis "provided a more clinically meaningful interpretation of the treatment effect." Given absence of any clinically meaningful differences in the observed PFS curves,[52] one would have to agree with Pramanik et al[11] that regardless of the statistical significance, "this meager difference between mean survivals may remain clinically unimportant."

In another example, a trial evaluating the efficacy of neratinib after trastuzumab-based adjuvant therapy in early-stage HER2-positive breast cancer[53] reported a disease-free survival HR of 0.67 (log-rank $P = .0091$). In their reanalysis, Hasegawa et al[4] noted that the observed HR of 0.67 corresponds to just a 0.5-month improvement in RMST disease-free survival (from 23.0 to 23.5 months) and suggested that this improvement was of "debatable advantage." In response, Chan et al[5] noted that the RMST analysis used a τ of 24 months. This implies that the maximum possible RMST is 24 months (which would be if all experimental patients were cured), so the maximum possible improvement in RMST is 1.0 month (from the 23-month RMST observed in the control arm). Therefore, the observed 0.5-month improvement in the neratinib arm represents 50% of the maximum possible effect (ie, everybody is cured)—an improvement that would seem of clinical significance.

Kaplan-Meier curves (with CIs) for the experimental and control arms of an RCT offer a comprehensive display of the experimental-arm effectiveness. Trying to reduce these curves to a single summary of treatment benefit is challenging. Although the clinical utility of an estimator is somewhat in the eyes of the beholder, we are unimpressed by the RMST ability to consistently capture the magnitude of clinical benefit across clinical settings.

To our knowledge, the use of HRs and log-rank tests as primary analysis tools has not impeded the development, testing, and acceptance of effective oncologic therapies (eg, the checkpoint inhibitors). When there are concerns about delayed treatment effects and/or long-term cures, log-rank–based designs with slightly inflated sample size (10%),[54] additional follow-up,[27] and modified interim futility analyses[27,55] can provide robust power. Methods for accommodating nonproportional hazards such as RMST, weighted log-rank tests, and others[56-59] can be useful secondary analyses because it is often difficult to have a single summary measure to accurately reflect the totality of clinical effect. However, before abandoning log-rank test–based primary analyses of definitive RCTs, we will need to see more convincing evidence of how these alternative methods can improve development of effective cancer therapies.

AFFILIATION

[1]National Cancer Institute, Bethesda, MD

CORRESPONDING AUTHOR

Boris Freidlin, PhD, Biometric Research Program, National Cancer Institute, Bethesda, MD 20892; e-mail: freidlinb@ctep.nci.nih.gov.

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST AND DATA AVAILABILITY STATEMENT

Disclosures provided by the authors and data availability statement (if applicable) are available with this article at DOI https://doi.org/10.1200/JCO.19.01681.

AUTHOR CONTRIBUTIONS

Conception and design: All authors
Financial support: All authors
Provision of study material or patients: All authors
Collection and assembly of data: All authors
Data analysis and interpretation: All authors
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

REFERENCES

1. Hasegawa T, Uno H, Wei LJ: How to summarize the safety profile of epoetin alfa versus best standard of care in anemic patients with metastatic breast cancer receiving standard chemotherapy? J Clin Oncol 34:3818, 2016
2. Hasegawa T, Uno H, Wei LJ: Nivolumab in nonsquamous non-small-cell lung cancer. N Engl J Med 374:492-493, 2016
3. Borghaei H, Brahmer J: Nivolumab in nonsquamous non-small-cell lung cancer. N Engl J Med 374:493-494, 2016
4. Hasegawa T, Uno H, Wei LJ: Neratinib after trastuzumab in patients with HER2-positive breast cancer. Lancet Oncol 17:e176, 2016
5. Chan A, Buyse M, Yao B: Neratinib after trastuzumab in patients with HER2-positive breast cancer - Author's reply. Lancet Oncol 17:e176-e177, 2016
6. Horiguchi M, Uno H, Wei LJ: Patients with advanced melanoma who discontinued treatment with nivolumab and ipilimumab as a result of adverse events lived significantly longer than patients who continued treatment. J Clin Oncol 36:720-721, 2018
7. Lo SN, Sandhu S: Reply to M. Horiguchi et al. J Clin Oncol 36:722-723, 2018
8. Horiguchi M, Uno H, Wei LJ: Evaluating noninferiority with clinically interpretable statistics for the PROSELICA study to assess treatment efficacy of a reduced dose of cabazitaxel for treating metastatic prostate cancer. J Clin Oncol 36:825-826, 2018
9. Eisenberger MA, Zhang W, Shun Z, et al: Reply to M. Horiguchi et al. J Clin Oncol 36:826-827, 2018
10. Fang X, Uno H, Wei LJ: Assessing metronomic chemotherapy for progressive pediatric solid malignant tumors. JAMA Oncol 4:743, 2018
11. Pramanik R, Vishnubhatla S, Bakhshi S: Assessing metronomic chemotherapy for progressive pediatric solid malignant tumors—reply. JAMA Oncol 4:744, 2018
12. Sun R, Horiguchi M, Wei LJ: Interpreting the benefit of trifluridine/tipiracil in metastatic colorectal cancer with respect to progression-free survival and overall survival. J Clin Oncol 36:1378-1379, 2018
13. Uno H, Kim DH, Wei LJ: Interpreting the association of first-in-class immune checkpoint inhibition and targeted therapy with survival in patients with stage IV melanoma. JAMA Oncol 4:1135-1136, 2018
14. Sinnamon AJ, Gimotty PA, Karakousis GC: Interpreting the association of first-in-class immune checkpoint inhibition and targeted therapy with survival in patients with stage IV melanoma–reply. JAMA Oncol 4:1136-1137, 2018
15. Uno H, Tian L, Wei LJ: Estimating and interpreting the overall survival benefit of checkpoint inhibitors via meta-analysis. JAMA Oncol 4:1137-1138, 2018
16. Gebski V, Yang JC, Lee CK: Estimating and interpreting the overall survival benefit of checkpoint inhibitors via meta-analysis–reply. JAMA Oncol 4:1138-1139, 2018
17. Sun R, Wei LJ: Regional hyperthermia with neoadjuvant chemotherapy for treatment of soft tissue sarcoma. JAMA Oncol 5:112-113, 2019
18. Mansmann U, Lindner LH, Issels R: Regional hyperthermia with neoadjuvant chemotherapy for treatment of soft tissue sarcoma–reply. JAMA Oncol 5:114, 2019
19. McCaw ZR, Vassy JL, Wei LJ: Palbociclib and fulvestrant in breast cancer. N Engl J Med 380:796-797, 2019
20. Turner NC, Huang X, Cristofanilli M: Palbociclib and fulvestrant in breast cancer. Reply. [Reply] N Engl J Med 380:797, 2019
21. McCaw ZR, Wei LJ: Interpreting the survival benefit from neoadjuvant chemoradiotherapy before surgery for locally advanced squamous cell carcinoma of the esophagus. J Clin Oncol 37:1032-1033, 2019
22. Li J, Liu Q, Yang H, et al: Reply to Z. McCaw and L-J. Wei. J Clin Oncol 37:1034, 2019
23. McCaw ZR, Jiang F, Wei LJ: Trastuzumab therapy for 9 weeks vs 1 year for human epidermal growth factor receptor 2-positive breast cancer. JAMA Oncol 5:117-118, 2019
24. Gebski V, Huttunen T, Joensuu H: Trastuzumab therapy for 9 weeks vs 1 year for human epidermal growth factor 2-positive breast cancer–reply. JAMA Oncol 5:118, 2019
25. Duke University, US Food and Drug Administration: Public workshop: Oncology clinical trials in the presence of non-proportional hazards. https://healthpolicy.duke.edu/events/public-workshop-oncology-clinical-trials-presence-non-proportional-hazards
26. Fine GD: Consequences of delayed treatment effects on analysis of time-to-event endpoints. Drug Inf J 41:535-539, 2007
27. Chen TT: Statistical issues and challenges in immuno-oncology. J Immunother Cancer 1:18, 2013
28. 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
29. Royston P, Parmar MK: 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 16:16, 2016
30. Saad ED, Zalcberg JR, Péron J, et al: Understanding and communicating measures of treatment effect on survival: Can we do better? J Natl Cancer Inst 110:232-240, 2018
31. Gehan EA: A generalized two-sample Wilcoxon test for doubly censored data. Biometrika 52:650-653, 1965
32. Prentice RL: Linear rank tests with right censored data. Biometrika 65:167-179, 1978
33. Jatoi I, Anderson WF, Jeong JH, et al: Breast cancer adjuvant therapy: Time to consider its time-dependent effects. J Clin Oncol 29:2301-2304, 2011

34. Fleming TR, Harrington DP: Counting Processes and Survival Analysis. New York, NY, Wiley, 1981
35. Hasegawa T: Sample size determination for the weighted log-rank test with the Fleming-Harrington class of weights in cancer vaccine studies. Pharm Stat 13: 128-135, 2014
36. Mok TSK, Wu YL, Kudaba I, et al: Pembrolizumab versus chemotherapy for previously untreated, PD-L1-expressing, locally advanced or metastatic non-small-cell lung cancer (KEYNOTE-042): A randomised, open-label, controlled, phase 3 trial. Lancet 393:1819-1830, 2019
37. 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
38. Tarone RE: On the distribution of the maximum of the logrank statistic and the modified Wilcoxon statistic. Biometrics 37:79-85, 1981
39. Lee JW: Some versatile tests based on simultaneous use of weighted log-rank statistics. Biometrics 52:721-725, 1996
40. Karrison TG: Versatile tests for comparing survival curves based on weighted log-rank statistics. Stata J 16:678-690, 2016
41. Karrison TG: Restricted mean life with adjustment for covariates. J Am Stat Assoc 82:1169-1176, 1987
42. Pepe MS, Fleming TR: Weighted Kaplan-Meier statistics: A class of distance tests for censored survival data. Biometrics 45:497-507, 1989
43. 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 30:2409-2421, 2011
44. Seruga B, Pond GR, Hertz PC, et al: Comparison of absolute benefits of anticancer therapies determined by snapshot and area methods. Ann Oncol 23: 2977-2982, 2012
45. Horiguchi M, Cronin AM, Takeuchi M, et al: A flexible and coherent test/estimation procedure based on restricted mean survival times for censored time-to-event data in randomized clinical trials. Stat Med 37:2307-2320, 2018
46. Cook JA, Fergusson DA, Ford I, et al: There is still a place for significance testing in clinical trials. Clin Trials 16:223-224, 2019
47. Pollack A, Karrison TG, Balogh AG, et al: Short term androgen deprivation therapy without or with pelvic lymph node treatment added to prostate bed only salvage radiotherapy: The NRG Oncology/RTOG 0534 SPPORT trial. Int J Radiat Oncol Biol Phys 102:1605, 2018
48. Aberle DR, Adams AM, Berg CD, et al: Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 365:395-409, 2011
49. Self S, Prentice R, Iverson D, et al: Statistical design of the Women's Health Trial. Control Clin Trials 9:119-136, 1988
50. 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
51. A'Hern RP: Restricted mean survival time: An obligatory end point for time-to-event analysis in cancer trials? J Clin Oncol 34:3474-3476, 2016
52. Pramanik R, Agarwala S, Gupta YK, et al: Metronomic chemotherapy vs best supportive care in progressive pediatric solid malignant tumors: A randomized clinical trial. JAMA Oncol 3:1222-1227, 2017
53. 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 17:367-377, 2016
54. Hoering A, Durie B, Wang H, et al: End points and statistical considerations in immuno-oncology trials: Impact on multiple myeloma. Future Oncol 13: 1181-1193, 2017
55. Korn EL, Freidlin B: Interim futility monitoring assessing immune therapies with a potentially delayed treatment effect. J Clin Oncol 36:2444-2449, 2018
56. Chen TT: Milestone survival: A potential intermediate endpoint for immune checkpoint inhibitors. J Natl Cancer Inst 107:djv156, 2015
57. Péron J, Lambert A, Munier S, et al: Assessing long-term survival benefits of immune checkpoint inhibitors using the net survival benefit. J Natl Cancer Inst [epub ahead of print on March 5, 2019]
58. Sposto R: Cure model analysis in cancer: An application to data from the Children's Cancer Group. Stat Med 21:293-312, 2002
59. Kim HT, Gray R: Three-component cure rate model for nonproportional hazards alternative in the design of randomized clinical trials. Clin Trials 9:155-163, 2012

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

Methods for Accommodating Nonproportional Hazards in Clinical Trials: Ready for the Primary Analysis?

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/jco/site/ifc.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

No potential conflicts of interest were reported.

APPENDIX

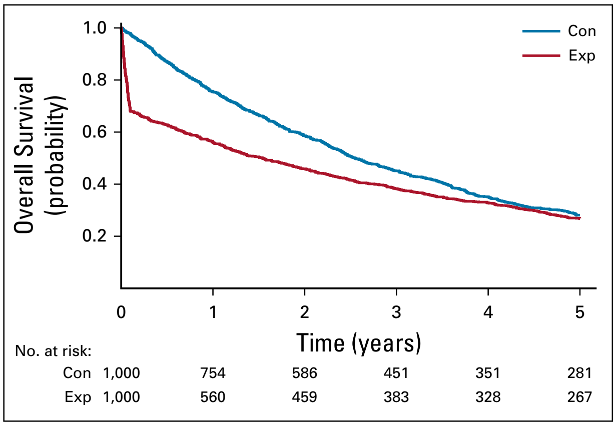

FIG A1. An example of a hypothetical randomized clinical trial where the survival curve for an experimental arm (red line) is always below the control arm curve (blue line) yet the late-emphasis test ($G^{0,1}$) rejects the null hypothesis in favor of the experimental arm with one-sided p-value of 0.0046. The trial data were generated assuming 1000 patients per arm with instant accrual and 5 years of follow-up; in the control arm survival was assumed to follow an exponential distribution with constant hazard of 0.25, in the experimental arm survival was assumed to follow a piecewise exponential distribution with a hazard of 4 in the first 1.2 months and a hazard of 0.19 after the first 1.2 months.