Assessment of Hazard Ratios in Oncology Clinical Trials Terminated Early for Superiority
A Systematic Review

Masashi Shimura, PhD; Shogo Nomura, PhD; Masashi Wakabayashi, ME; Kazushi Maruo, PhD; Masahiko Gosho, PhD

Abstract

IMPORTANCE Group sequential designs allow potential early trial termination at the interim analysis, before study completion. Traditional maximum likelihood estimate is commonly used to quantify the treatment effect in group sequential design trials; however, in published clinical trials, a bias-adjusted estimator has rarely been reported.

OBJECTIVE To emphasize the need for considering overestimation of treatment effect by applying 2 bias-adjusted estimators to previously published, early-terminated oncology clinical trials.

EVIDENCE REVIEW Trials published from 2013 to 2017 were identified by searching MEDLINE and Embase on February 23, 2018. This review was restricted to oncology clinical trials using group sequential designs with a single preplanned interim analysis as well as 2-arm randomized clinical trials that were subsequently stopped for efficacy reasons. Each article was independently reviewed by 3 biostatisticians during text screening, and differences in opinion were resolved by discussion. This report presents the unadjusted hazard ratio (HR) of an experimental arm to a reference arm and 2 bias-adjusted HRs calculated by using the conditional mean-adjusted estimator (CMAE) and weighted CMAE (WCMAE).

FINDINGS In total, 198 abstracts were screened for eligibility, of which, 19 eligible clinical trials were identified as applicable to the bias-adjusted estimators. Unadjusted HRs ranged from 0.203 (95% CI, 0.150-0.276) to 0.71 (95% CI, 0.60-0.84), number of events at the interim analysis from 58 to 540, and information time from 48% to 82%. In each study, the HRs adjusted by CMAE and WCMAE were higher than the unadjusted HR. Bias-adjusted estimates in large trials (243 and 414 events at the interim analysis) were similar to the unadjusted HR. However, in small trials (eg, with 58 events at the interim analysis), bias-adjusted estimates were highly disparate from the unadjusted HR. In trials with large treatment effects (eg, HRs of 0.20 and 0.22), the difference between unadjusted and bias-adjusted HRs was small even though the number of events at the interim analysis was small; larger differences were observed when the unadjusted HR was greater than 0.5.

CONCLUSIONS AND RELEVANCE In this systematic review of oncology clinical trials that were stopped for efficacy at the interim analysis, relatively large differences were noted between the unadjusted and adjusted HRs when the number of events at the interim analysis was small or when the unadjusted HR was close to the boundaries. These findings suggest presenting the 2 bias-adjusted HRs along with the unadjusted HR in the data monitoring committee meeting.

Key Points

Question Do bias-adjusted hazard ratios differ from unadjusted hazard ratios when oncology clinical trials are stopped for efficacy at the interim analysis?

Findings In this systematic review of 19 clinical trials, 2 bias-adjusted hazard ratios—calculated using conditional mean-adjusted estimator and weighted conditional mean-adjusted estimator—were distinct from the unadjusted hazard ratio in small trials. Larger differences between the unadjusted and bias-adjusted values were observed when the estimated hazard ratio was greater than 0.5.

Meaning These findings suggest presenting the bias-adjusted hazard ratios, along with the unadjusted hazard ratio, in the data monitoring committee meeting because bias-adjusted estimators may play an important role in the committee's decision.

Author affiliations and article information are listed at the end of this article.
Introduction

In clinical trials with long follow-up times (eg, oncology), group sequential designs (GSDs) allow early trial termination based on the interim analysis. Group sequential design is adaptive and has boundaries to stop a trial when there is sufficient evidence of efficacy. In a clinical trial, the boundaries of the GSD control the overall type I error rate (the probability of erroneously concluding a beneficial effect when there is no effect on the experimental treatment) by suppressing the significance level at each stage. Group sequential design is widely used to assess whether a trial should be terminated early for efficacy or lack thereof, especially in oncology clinical trials.

The maximum likelihood method is commonly used to quantify treatment effects, such as hazard ratio (HR) and risk ratio, in GSD trials. It is an intuitive and standard statistical method used in both non-GSDs and GSDs. Controlling the overall type I error rate is based on a maximum likelihood estimate (MLE), derived from the maximum likelihood method, by adjusting efficacy stopping boundaries. However, even controlling type I errors does not address the bias of MLE, and the MLE may overestimate treatment effects, especially when a trial is terminated early for efficacy reasons. Magnitude of the overestimations of treatment effects generally reported in medical journals may not be negligible. For example, an estimate of efficacy (eg, HR) allows a data monitoring committee (DMC) to assess the risks and benefits of treatment in addition to providing safety information. The US Food and Drug Administration has issued draft guidance on adaptive designs, including GSD, stating that a bias-adjusted estimate informing the stopping rule of a GSD should be prospectively planned and used when reporting study results. Several researchers have proposed bias-adjusted estimators, such as the conditional mean-adjusted estimator (CMAE) and weighted CMAE (WCMAE), to address the issue. The WCMAE is a modified CMAE, which is calculated by weighting the MLE obtained at the interim analysis and the effect size prespecified when calculating the sample size. The statistical detail of the CMAE and WCMAE is shown in eAppendix 1 in the Supplement.

However, in published clinical trials, a bias-adjusted estimate of the HR is rarely reported. This suggests that an unadjusted estimate obtained using the maximum likelihood method may be nonchalantly presented to the DMC at the interim analysis. It might be critical to display bias-adjusted estimators in addition to the MLE when the DMC, sponsor, and investigators interpret the results of the trial. Therefore, we applied the CMAE and the WCMAE to early-terminated oncology clinical trials published in major journals to emphasize the need for adjusting HR overestimation.

Methods

Literature Search

We conducted this systematic review according to the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) reporting guideline. The procedures for systematic review are provided in eAppendix 2 in the Supplement. This review was restricted to oncology clinical trials using GSD with a single preplanned interim analysis. We included 2-arm randomized clinical trials that were stopped on the basis of efficacy. We identified trials published in each year from 2013 to 2017 by searching MEDLINE and Embase on February 23, 2018. The search was restricted to 11 scientific journals: Annals of Internal Medicine, CA: A Cancer Journal for Clinicians, Cancer Discovery, JAMA, JAMA Oncology, Journal of Clinical Oncology, Lancet Oncology, Nature Reviews Clinical Oncology, The BMJ, The Lancet, and The New England Journal of Medicine. We used a free-text search with relevant keywords, including hazard ratio and at least 1 of the following keywords: interim analysis((i/e)s), group sequential, two stage, stop, stopping, terminate, termination, halt, close, continue, continuation, prematurely, independent data monitoring, data and safety monitoring board, DSMB, Brien-Fleming, Pocock, Lan-DeMets, Fisher information, boundary(ies), stop, terminate, or independent data monitoring.
Eligibility Criteria and Quality Control
We included articles that met all the following inclusion criteria: oncology clinical trials, 2-arm randomized trials, trials that were stopped at an interim analysis on the basis of efficacy, and clinical trials with a preplanned interim analysis. We excluded articles that met any of the following exclusion criteria: protocol articles, articles on statistical methodology, retrospective studies, meta-analyses, articles on integrated database analyses, follow-up analyses, subgroup analyses, noninferiority clinical trials, and clinical trials that were stopped because of safety concerns. We defined articles as eligible if they met all the inclusion criteria and did not meet any exclusion criteria. The definition of eligibility was the same in abstract screening and full-text screening.

We used abstract screening to reduce the number of articles examined with full-text screening. We excluded articles that obviously did not meet the eligibility criteria, including articles that were not based on oncology, not clinical trials, not randomized, or were negative studies. Then, we retrieved the full texts of the remaining articles and determined their eligibility for full-text screening. Each article was independently reviewed by 3 biostatisticians (M.S., S.N., and M.W.) during abstract and full-text screening, and differences in opinion were resolved by discussion.

Results
Eligibility Screening and Data Extraction
A total of 198 abstracts were screened for eligibility (Figure 1). Ninety-seven articles that obviously were not eligible were excluded, including duplicate articles that were eliminated by comparing registration numbers from 3 registration databases: clinicaltrial.gov, UMIN, and ISRCTN Registry. The full texts of the remaining 101 articles were assessed, and 19 eligible clinical trials were identified for application of the bias-adjusted estimators.7-25

Of the 19 eligible trials, two of them (Lonial et al7 and Dimopoulos et al15) included 2 efficacy end points and a single planned interim analysis. However, these 2 trials met the termination criteria for only 1 end point. No trials used any bias-adjusted estimators. If the information needed to calculate the bias-adjusted estimates (alpha spending function, number of events, and planned HR) was missing, the information was gathered by referring to the protocols, statistical analysis plan, and related articles. All data extracted are shown in the eTable in the Supplement.

Comparison of Unadjusted HRs and Bias-Adjusted HRs
Figure 2 shows the unadjusted HRs based on the standard maximum likelihood method, HRs adjusted by the CMAE and WCMAE, study number, end points, sample sizes, events at the interim analysis, and planned events at the final analysis for each trial. The experimental treatment was more efficacious than the reference treatment if the HR was less than 1. The most common type of end point.

Figure 1. Results of Abstract and Full-text Screening
point for the interim analysis was progression-free survival time (79% of the trials), followed by overall survival time (16%) and disease-free survival time (5%). The unadjusted HRs ranged from 0.203 (95% CI, 0.150-0.276)\textsuperscript{19} to 0.71 (95% CI, 0.60-0.84),\textsuperscript{22} number of events at the interim analysis from 58 to 540, and information time from 48% to 82%. The information time is a proportion of the number of events at the interim analysis and that at the final analysis. The planned information time (designated as IT[P] in Figure 2) was extracted from the main text of the 19 articles. Two of them could not be shown as a percentage because the interim analysis of the two studies was not based on the number of events. The actual information time (designated as IT[A] in Figure 2) was calculated by the actual number of events at the interim analysis dividing the planned number of events at the final analysis. Progression-free survival was the end point in all trials with a unadjusted HR of 0.6 or less. In each study, the HRs adjusted by the CMAE and WCMAE were higher than the associated unadjusted HR.

The differences between the unadjusted HRs and adjusted HRs was associated with the number of events at the interim analysis and the unadjusted HR itself (Figure 2). The bias-adjusted estimators in the large trials, such as the study by Hortobagyl et al\textsuperscript{11} (243 events at the interim analysis) and the one by Dimopoulos et al\textsuperscript{16} (414 events at the interim analysis), were similar to the unadjusted HR. However, the bias-adjusted estimators in the small trials, such as the study by Oza et al\textsuperscript{14} (58 events at the interim analysis), were highly disparate from the unadjusted HRs. In trials with large treatment effects (eg, Dimopoulos et al\textsuperscript{19} and Byrd et al\textsuperscript{20}), when the number of events at the interim analysis was small, the difference between unadjusted and bias-adjusted HRs was also small. For treatments with extremely positive effects, such as 0.20 (Chanan-Kahn et al\textsuperscript{21}) and 0.22 (Byrd et al\textsuperscript{20}), the risks of overestimating unadjusted HR would be minimal. Conversely, larger differences between the unadjusted and bias-adjusted HRs were generally observed when the estimated HR was larger than 0.5.

![Figure 2](https://jamanetwork.com/)

**Figure 2. Unadjusted and Bias-Adjusted Hazard Ratios (HRs) With End Points (EPs) in 19 Eligible Oncology Clinical Trials**

<table>
<thead>
<tr>
<th>Study No.</th>
<th>Source</th>
<th>IA, No.</th>
<th>SS, No.</th>
<th>FA, No.</th>
<th>IT(P/A)</th>
<th>EP</th>
<th>Unadjusted HR</th>
<th>Adjusted HR (CMAE)</th>
<th>Adjusted HR (WCMAE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Lomtal et al,\textsuperscript{17} 2015</td>
<td>384</td>
<td>646</td>
<td>466</td>
<td>70/82</td>
<td>PFS</td>
<td>0.70</td>
<td>0.71</td>
<td>0.72</td>
</tr>
<tr>
<td>2</td>
<td>Stewart et al,\textsuperscript{2} 2015</td>
<td>431</td>
<td>792</td>
<td>526</td>
<td>80/82</td>
<td>PFS</td>
<td>0.69</td>
<td>0.70</td>
<td>0.71</td>
</tr>
<tr>
<td>3</td>
<td>Marcus et al,\textsuperscript{17} 2017</td>
<td>245</td>
<td>1203</td>
<td>370</td>
<td>67/66</td>
<td>PFS</td>
<td>0.66</td>
<td>0.69</td>
<td>0.72</td>
</tr>
<tr>
<td>4</td>
<td>Stupp et al,\textsuperscript{19} 2015</td>
<td>187</td>
<td>695</td>
<td>374</td>
<td>NA/50</td>
<td>PFS</td>
<td>0.62</td>
<td>0.68</td>
<td>0.71</td>
</tr>
<tr>
<td>5</td>
<td>Hortobagyl et al,\textsuperscript{11} 2016</td>
<td>243</td>
<td>668</td>
<td>302</td>
<td>70/80</td>
<td>PFS</td>
<td>0.56</td>
<td>0.57</td>
<td>0.58</td>
</tr>
<tr>
<td>6</td>
<td>Sehn et al,\textsuperscript{12} 2016</td>
<td>175</td>
<td>396</td>
<td>260</td>
<td>67/67</td>
<td>PFS</td>
<td>0.55</td>
<td>0.56</td>
<td>0.57</td>
</tr>
<tr>
<td>7</td>
<td>Goetz et al,\textsuperscript{13} 2017</td>
<td>194</td>
<td>493</td>
<td>240</td>
<td>79/81</td>
<td>PFS</td>
<td>0.54</td>
<td>0.56</td>
<td>0.57</td>
</tr>
<tr>
<td>8</td>
<td>Oza et al,\textsuperscript{14} 2015</td>
<td>58</td>
<td>130</td>
<td>121</td>
<td>48/48</td>
<td>PFS</td>
<td>0.53</td>
<td>0.73</td>
<td>0.81</td>
</tr>
<tr>
<td>9</td>
<td>Dimopoulos et al,\textsuperscript{16} 2015</td>
<td>414</td>
<td>929</td>
<td>526</td>
<td>75/79</td>
<td>PFS</td>
<td>0.53</td>
<td>0.53</td>
<td>0.53</td>
</tr>
<tr>
<td>10</td>
<td>van Oers et al,\textsuperscript{15} 2015</td>
<td>187</td>
<td>474</td>
<td>280</td>
<td>67/67</td>
<td>PFS</td>
<td>0.50</td>
<td>0.51</td>
<td>0.51</td>
</tr>
<tr>
<td>11</td>
<td>Turner et al,\textsuperscript{17} 2015</td>
<td>195</td>
<td>521</td>
<td>238</td>
<td>60/82</td>
<td>PFS</td>
<td>0.42</td>
<td>0.42</td>
<td>0.42</td>
</tr>
<tr>
<td>12</td>
<td>Palumbo et al,\textsuperscript{18} 2016</td>
<td>189</td>
<td>498</td>
<td>295</td>
<td>60/64</td>
<td>PFS</td>
<td>0.39</td>
<td>0.39</td>
<td>0.39</td>
</tr>
<tr>
<td>13</td>
<td>Dimopoulos et al,\textsuperscript{19} 2016</td>
<td>169</td>
<td>569</td>
<td>295</td>
<td>60/57</td>
<td>PFS</td>
<td>0.37</td>
<td>0.37</td>
<td>0.37</td>
</tr>
<tr>
<td>14</td>
<td>Byrd et al,\textsuperscript{20} 2014</td>
<td>117</td>
<td>391</td>
<td>234</td>
<td>50/50</td>
<td>PFS</td>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
</tr>
<tr>
<td>15</td>
<td>Chanan-Kahn et al,\textsuperscript{11} 2016</td>
<td>239</td>
<td>578</td>
<td>342</td>
<td>50/70</td>
<td>PFS</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
</tr>
<tr>
<td>16</td>
<td>Beer et al,\textsuperscript{21} 2014</td>
<td>540</td>
<td>1717</td>
<td>765</td>
<td>67/71</td>
<td>OS</td>
<td>0.71</td>
<td>0.72</td>
<td>0.73</td>
</tr>
<tr>
<td>17</td>
<td>Parker et al,\textsuperscript{13} 2013</td>
<td>314</td>
<td>921</td>
<td>640</td>
<td>50/49</td>
<td>OS</td>
<td>0.70</td>
<td>0.76</td>
<td>0.78</td>
</tr>
<tr>
<td>18</td>
<td>Robert et al,\textsuperscript{24} 2015</td>
<td>222</td>
<td>704</td>
<td>288</td>
<td>70/77</td>
<td>OS</td>
<td>0.69</td>
<td>0.74</td>
<td>0.73</td>
</tr>
<tr>
<td>19</td>
<td>Masuda et al,\textsuperscript{25} 2017</td>
<td>202</td>
<td>910</td>
<td>352</td>
<td>NA/57</td>
<td>DFS</td>
<td>0.70</td>
<td>0.80</td>
<td>0.83</td>
</tr>
</tbody>
</table>

The trials were sorted by EPs (progression-free survival [PFS], overall survival [OS], and disease-free survival [DFS]) and unadjusted HRs. The third and fourth columns present the number of events at the interim analysis (IA) and the sample size (SS), respectively. The fifth column presents the planned number of events at the final analysis (FA). The information time (IT) is a proportion of the planned (IT[P]) and the actual (IT[A]) number of events at the interim analysis. The IT[P] was extracted from the main text of the 19 articles. Two of them could not be shown as a percentage because the interim analysis of the 2 studies was not based on the number of events. The IT[A] was calculated by the actual number of events at the interim analysis dividing the planned number of events at the final analysis. CMAE indicates conditional mean-adjusted estimator; WCMAE, weighted CMAE.
Discussion

If the estimated treatment effect is not large or the number of events at the interim analysis is small, the unadjusted HR should be interpreted with skepticism. The effect of other factors on bias adjustment, such as information time, was explained by these 2 factors. Even if the total number of events is small, large information time reduced the difference between unadjusted HR and bias-adjusted HRs. It is difficult to control the bias of the unadjusted HR before starting the trial, but researchers can reduce the magnitude of overestimation by increasing the number of events at the interim analysis in advance. The performance of the CMAE and WCMAE depends on the unmeasured true treatment effect. Thus, we recommend presenting the 2 bias-adjusted estimators with the MLE, especially when the trial is terminated early, although most standard statistical software (eg, SAS; SAS Institute Inc) cannot directly produce bias-adjusted estimators.

The importance of the number of events is easy to interpret; data about treatment effects from small trials are generally limited, and the amount of data at the interim analysis is smaller than that at the final analysis. Thus, a positive result at the interim analysis tends to exaggerate the treatment effect when trials are halted for efficacy reasons. The DMC members and stakeholders of clinical trial sponsors should carefully interpret and report the results from such trials when the number of events at the interim analysis is small.

Since the true HR is unknown, operating characteristics of CMAE and WCMAE should be evaluated statistically under the hypothetical situation where the true HR is known. Shimura et al4 had analytically compared the unadjusted HR, CMAE, and WCMAE along with computer simulations and found that the CMAE and WCMAE reduced the bias that was defined as the difference of HR between estimates and the true value. Although the WCMAE underestimated the difference in treatment effect when the true HR was relatively low (a true HR of 0.5), the magnitude of underestimation was not serious.

Study end points might also be critical when interpreting study results. In Figure 2, the treatment effect of progression-free survival tended to be larger than that of overall survival. This caused differences in the estimated treatment effect between the unadjusted and bias-adjusted estimators because, as discussed in the Results section, the adjustment becomes larger when the unadjusted HR is close to 1.

In GSDs, the conditional bias-adjusted estimators (the CMAE and WCMAE) do not affect the boundaries and significance levels of early stopping of a trial at the interim analysis. Therefore, researchers can use the same boundaries and significance levels used in standard GSDs. Another advantage of the CMAE and WCMAE is that they do not require patient-level data. To calculate the bias-adjusted estimators, we require the unadjusted HR, boundary value, number of events at the interim analysis, and planned HR (only for WCMAE). The information is generally provided in the article or in the protocol. Thus, researchers can calculate the HR for the CMAE and WCMAE using the existing literature.

The bias-adjusted estimators may play an important role in the decision of the DMC. In many cases, the estimate from CMAE and WCMAE becomes more conservative than the unadjusted HR. Thus, presenting them in the DMC meeting might reduce the probability of early termination. Data monitoring committees are generally wary of interim results with few events, since 1 more or 1 less event in either group could change the conclusion of the interim analysis. Therefore, it is recommended to present both the bias-adjusted HRs as well as the unadjusted HR.

Limitations

One of the major limitations of our study is that the bias-adjusted estimators we considered could only be applied to superiority trials with a single interim analysis. As a result, the literature search was restricted to trials in which only 1 interim analysis was planned. Therefore, we have no information on the bias in HR estimation for long-term trials with more than 1 interim analysis.
Conclusions

In actual clinical trials, this systematic review found relatively large differences between the unadjusted and adjusted HRs when the number of events at the interim analysis was small or unadjusted HR was close to the boundaries. Therefore, we recommend presenting the 2 bias-adjusted estimators with the unadjusted HR in the DMC meeting to assess whether early termination for efficacy reasons is recommended at the interim analysis.

ARTICLE INFORMATION

Accepted for Publication: April 16, 2020.


Open Access: This is an open access article distributed under the terms of the CC-BY-NC-ND License. © 2020 Shimura M et al. JAMA Network Open.

Corresponding Author: Masashi Shimura, PhD, Data Science Department, Taiho Pharmaceutical Co, Ltd, 1-2-4, Uchikanda, Chiyoda-ku, Tokyo, Japan (m-shimura@taiho.co.jp).

Author Affiliations: Data Science Department, Taiho Pharmaceutical Co, Ltd, Tokyo, Japan (Shimura); Center for Research and Administration and Support, Biostatistics Division, National Cancer Center, Chiba, Japan (Nomura, Wakabayashi); Department of Biostatistics, Faculty of Medicine, University of Tsukuba, Tsukuba, Japan (Maruo, Gosho).

Author Contributions: Dr Shimura 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.

Concept and design: Shimura, Wakabayashi, Maruo, Gosho.

Acquisition, analysis, or interpretation of data: All authors.

Drafting of the manuscript: Shimura, Nomura, Wakabayashi.

Critical revision of the manuscript for important intellectual content: Shimura, Nomura, Maruo, Gosho.

Statistical analysis: Shimura, Wakabayashi.

Administrative, technical, or material support: Shimura, Nomura.

Supervision: Shimura, Maruo, Gosho.

Conflict of Interest Disclosures: Dr Nomura reported receiving personal fees from AstraZeneca, Chugai Pharmaceutical Co, Ltd, Taiho Pharmaceutical, and Pfizer outside the submitted work. Dr Wakabayashi reported receiving personal fees from Chugai Pharmaceutical Co, Ltd and Johnson & K. K. Medical Company outside the submitted work. No other disclosures were reported.

REFERENCES


SUPPLEMENT.
eAppendix 1. Bias-Adjusted Estimators
eReference.
eAppendix 2. Procedures for Systematic Review
eTable. All Extracted Data


June 23, 2020 7/7