Overestimation of the Effect Size in Group Sequential Trials
(Running title: Overestimation in Group Sequential Trials)

Jenny J. Zhang1*, Gideon M. Blumenthal2, Kun He1, Shenghui Tang1, Patricia Cortazar2, and Rajeshwari Sridhara1
1U.S. Food and Drug Administration, CDER/OTS/OB/DBV, Silver Spring, MD 20993
2U.S. Food and Drug Administration, CDER/OND/OHOP/DOP-1, Silver Spring, MD 20993
*Corresponding author: jenny.zhang@fda.hhs.gov

Disclaimer: This article reflects the views of the authors and should not be construed to represent FDA’s views or policies.

Conflicts of interest: None.

Abstract
Group sequential designs (GSDs), which provide for interim monitoring of efficacy data and allow potential early trial termination while preserving the type I error rate, have become commonplace in oncology clinical trials. Although ethically appealing, GSDs tend to overestimate the true treatment effect size at early interim analyses. Overestimation of the treatment effect may exaggerate the benefit of a drug and provide imprecise information for physicians and their patients about a drug’s true effect. The cause and effect of such a phenomenon is generally not well-understood by many in clinical trial practice. In this paper, we provide a graphical explanation for why the phenomenon of overestimation in GSDs occurs. The potential overestimation of the magnitude of the treatment effect is of particular concern in oncology, where the more subjective endpoint of progression-free survival (PFS) has increasingly been adopted as the primary endpoint in pivotal phase III trials.

Keywords: Group sequential design; Overestimation bias; Treatment effect size; Progression-free survival
Introduction

Given the life-threatening nature and the unmet medical need for many types of cancer, there is an understandable demand within the oncology community to accelerate the drug development process. Many approaches aimed towards that goal have been proposed for phase I and II clinical trials (e.g. adaptive designs). In phase III pivotal trials, a well-accepted and ethically advantageous strategy is interim monitoring of efficacy data for potential early trial termination if one treatment is clearly superior (or inferior) to the testing alternative. However, repeated interim analyses inflate the false-positive rate or type I error. Group sequential designs (GSDs) control the overall type I error at a pre-specified level through an alpha spending function, which provides a specific statistical stopping criterion (or boundary) for each interim analysis. Thus, GSDs have become a frequently-used approach for interim monitoring in oncology phase III clinical trials to preserve the overall type I error rate (1-3).

An important, and sometimes overlooked, limitation of GSDs is that they are not aimed at describing the magnitude of the treatment effect (4). In fact, the earlier a trial is stopped for efficacy, the more likely the observed magnitude of treatment effect is an overestimate of the truth. This tendency of GSDs to overestimate the true treatment effect at an early interim analysis is well-recognized within the statistical community. However, although investigators and data monitoring committee (DMC) members may be aware of the overestimation phenomenon, few have a clear understanding of why this phenomenon occurs. Without a clear understanding of this phenomenon, the consequences of a decision to terminate a trial early may be obscured. Thus, the
objective of this paper is didactic in nature; i.e., to provide a graphical illustration, easily understood by all disciplines involved in clinical trials, of why the phenomenon of overestimation in GSDs occurs.

This potential overestimation of the magnitude of the treatment effect is of particular concern in oncology, where progression-free survival (PFS), a radiographically-assessed endpoint, has increasingly been adopted in pivotal phase III trials. Unlike overall survival (OS), the measurement and interpretation of PFS is inherently subjective. Thus, although the overestimation phenomenon is a concern for all GSDs, irrespective of the trial endpoint, the concern may be more paramount when PFS is the endpoint. Even for trials that continue to the end, statistical significance of a PFS effect may not necessarily be indicative of clinical meaningfulness (5). As a result, in trials with PFS as the primary endpoint, the magnitude of the treatment effect plays a particularly critical role in the benefit-risk assessment of a drug. In cases where that magnitude is likely overestimated, appropriate considerations and precautions should be discussed prior to decision-making.

**Random Variability**

In general, we are interested in the value of some true parameter (e.g. the treatment effect of a test drug as compared to a control, derived as a hazard ratio). However, the truth is often unknowable; our best alternative is to try to estimate it within a controlled setting, such as a clinical trial. In estimation, statisticians often emphasize the importance of bias, a measure of how far the average of the estimates is from the truth. The goal is to strive
for an unbiased estimate of the truth. Overestimation is bias in the direction that favors
the test drug; i.e. purporting that the effect of the test drug is better than it actually is.

There are many different clinical trial designs; two of the most common are fixed sample
design and group sequential design (GSD). Fixed sample design is the classical setting
where the size of the trial is fixed up-front and there are no interim analyses allowing
early trial termination for efficacy. The one and only analysis of efficacy data occurs at
the pre-defined end of the trial. In contrast, GSDs are more flexible in that they allow
repeated analyses of efficacy data over time to assess the potential for early trial
termination. However, there is a trade-off between flexibility and accuracy of estimation
(e.g. of the treatment effect) since trials that are stopped early have reduced sample size
(or number of events observed).

Regardless of the trial design, random variability always exists in estimation of the true
treatment effect and bias in estimation is often related to variability. In the setting of
fixed sample designs, as depicted in Figure 1A, the variability (solid red line) is
symmetric between the random highs (extreme positive bias) and random lows (extreme
negative bias), and stabilizes around the true treatment effect (short-dashed line) as the
trial continues to its mature end. Random highs and lows are collectively known as
random variability, which depends on factors including the study endpoint and the size of
the trial. In fixed sample designs (i.e. where there is no potential for early trial
termination), the random variability is measureable and the estimate of the treatment
effect is theoretically unbiased.
In GSDs, however, as depicted in Figure 1B where the dotted blue line denotes the statistical stopping boundary (e.g. O’Brien-Fleming), the variability favors the random highs. That is, the trial may be stopped early for efficacy when extreme positively-biased treatment effects are seen, but will continue in wait for the chance to observe a random high when negatively-biased effects are seen. This selective or differential use of variability in GSDs that allows early trial termination for efficacy leads to overestimation of the magnitude of the treatment effect. It is also intuitive that more frequent looks at the data, especially earlier on in the trial, increases the chance of observing a random high that would result in early termination and an observed effect that is an overestimate of the truth.

In a letter to JAMA (6), Ellenberg, DeMets, and Fleming gave a nice summary explanation for why overestimation of the treatment effect occurs in early-terminated trials: “...a trial terminated early for benefit will tend to overestimate [the] true effect; this happens because there always is variability in estimation of [the] true effect, and when assessing data over time, evidence of extreme benefit is more likely obtained at times when the data provide a random overestimate of [the] truth.”

It is important to note that controlling the type I error rate does not address this potential bias of overestimation. Type I error control simply assures that, under the null hypothesis (H₀) of no treatment benefit, the chance of seeing a spurious positive effect is limited to a defined level (usually 2.5% corresponding to a two-sided alpha of 0.05). If the statistical
analysis accounts for all efficacy interim analyses performed through appropriate alpha allocation, then the type I error should be controlled when there is no treatment effect (i.e. H₀ is true). However, when a treatment effect exists (i.e. H₀ is false), especially a minor one, the effect size is likely overestimated at early interim analyses, irrespective of type I error control. This is the setting we focus on in this paper.

Illustration of Overestimation in GSDs

Hypothetical example

To illustrate the overestimation in GSDs, we introduce a hypothetical example based on a randomized (1:1), placebo-controlled clinical trial in oncology. The primary endpoint is progression-free survival (PFS), defined as the time from randomization to disease progression or death from any cause, whichever occurs first. The trial is designed to detect a 50% improvement in median PFS time from 5 to 7.5 months, corresponding to a hazard ratio (HR) of 0.67. To obtain 90% power with an overall two-sided significance level of 0.05 and two planned interim analyses (at 33% and 67% information), 260 PFS events need to be observed. Given 6 months accrual, 10 months follow-up, and a 10% dropout rate, a total of 340 patients need to be enrolled. The overall type I error rate is controlled at 5% using O’Brien-Fleming boundaries following the Lan-DeMets methodology (3). Alpha is allocated as 0.00021, 0.01202, and 0.04626 for the first (87 events), second (174 events), and final (260 events) analyses, respectively.

Suppose that the trial is terminated early at the first interim analysis (i.e. after observing 87 PFS events) with an observed HR of 0.400 and a p-value of 0.00014. Note that the p-
value of 0.00014 is less than the alpha allocated for the first interim analysis of 0.00021, indicating that the observed estimate of the treatment effect has crossed the statistical stopping boundary. However, through simulation, we will show that the observed magnitudes of treatment effect at early interim analyses are overestimates of the true treatment effect.

**Simulation Study**

The main advantage of simulations is that we can specify the “true” treatment effect in a simulation, a value that is usually unknown within real clinical data. Figure 2 illustrates the results of a simulation (of 1 million trial replicates) following the hypothetical design and analysis plan described above. To provide a more impactful and didactic graphical illustration, the “true” treatment effect or HR in the simulated trials is specified at 0.60 (depicted by the dashed line). The dotted purple line denotes the stopping boundary, where all points (in dark red) falling below the dotted line have crossed their corresponding stopping boundary with an average HR denoted by the white dot; for example, at the first interim analysis, the average HR of trials that crossed the boundary at 87 events is 0.410.

Data maturity at the time of an interim analysis (i.e. % information) plays a crucial role in the reliability of the treatment effect estimate and the extent of overestimation. From Figure 2, we see that the mean HR estimate at 174 events (67% information) of 0.583 (95% CI: 0.432, 0.783) is quite close to the “true” HR of 0.60, whereas, at only 87 events (33% information), the overestimation is much greater with a mean HR of 0.410 (95%
CI: 0.271, 0.624). Note that if the trials are able to continue to the end (260 events), the average HR of trials crossing the boundary is 0.684 (95% CI: 0.535, 0.874), an underestimate of the “true” treatment effect HR of 0.60. This is due to the accumulation of random lows (negative biases) for trials that continue to the end.

**Real Data Examples**

Since most pivotal trials with interim PFS results that cross a stopping boundary are terminated early by DMCs with no further follow-up tumor assessments, illustration of the overestimation phenomenon within the same trial using real data examples is somewhat difficult. Although of notable value, an expansive evaluation of early-terminated trials is outside the scope of this paper. Bassler et al. (7) compared 91 truncated randomized clinical trials (RCTs) with 424 matching non-truncated RCTs. They found that, independent of statistical stopping rules, truncated RCTs were associated with greater effect sizes than RCTs not stopped early. However, the validity of their analyses was questioned in 4 letters to the editor (6).

We tried to find pivotal phase III trials in oncology which were terminated early due to significant interim PFS analysis results, but had reported subsequent updated results. Table 1 summarizes the trial design and results of two such trials: lapatinib plus capecitabine for HER2+ metastatic breast cancer (8-9) and sunitinib for metastatic renal cell carcinoma (10-11). Both trials were terminated early based on results from interim analyses that crossed the pre-specified stopping boundaries. The second analysis for each trial listed in the table was the update.
For both trials, the HR from the updated analysis was larger than that of the interim analysis, indicating that the observed treatment effect diminished with the subsequent analysis. As seen in Table 1, the first interim analysis of both trials was conducted around 50% information and the update occurred when greater than two-thirds of the total number of events was observed. The large differences between the HR estimates from the first interim and updated analyses suggest that the data for the first interim analysis were less mature, resulting in a likely overestimate of the treatment effect size. Both lapatinib and sunitinib were approved for serious and life-threatening conditions with great unmet medical need. In these examples, the earlier assessment allowed for earlier introduction of the drugs to U.S. patients.

Discussion

It is within the purview of DMCs to guard against early trial termination when the estimate of the treatment effect is not likely clinically meaningful. DMCs also strive to maintain ethical standards on behalf of patients. Thus, when an early stopping boundary has been crossed, DMCs often request an additional analysis at a later time point in an effort to confirm the effect. However, there are two caveats: (1) In general (e.g., when using O’Brien-Fleming boundaries), once an early boundary has been crossed, the incremental data included in the next analysis would need to be extremely negative in order to have any remarkable impact on the estimate from the new analysis. Thus, in practice, once an early boundary has been crossed, the estimated effect from the next
analysis would almost always be consistent with the previous analysis. (2) Although the incremental data included in the next analysis should somewhat ameliorate the overestimation bias, the “adjustment” would generally be very small since oftentimes the next analysis requested by the DMC is only 3-6 months later. Thus, the overestimation concern remains prominent even if the DMC waits for results of a later analysis.

There exists literature on statistical methods to adjust for the (conditional and unconditional) bias in overestimation of the treatment effect within the context of GSDs (12-16); however, they are rarely implemented in practice. Thus, only nominal estimates are usually presented at DMC meetings. This further emphasizes the need for all DMC members to have a thorough understanding of why the overestimation phenomenon occurs and its potential consequences such that they can be more appropriately accounted for in the DMC’s complex decision-making process to terminate a trial early.

Although this paper focuses on the bias conditional on the observed stopping time, we also recognize the importance of the marginal or unconditional bias. In our simulation example illustrated in Figure 2, the unconditional HR estimate was 0.575. Evaluation of the unconditional bias is particularly helpful in the trial design stage; however, there is also value in assessing the potential bias given that the trial has already stopped (conditional bias), especially based on a very early interim analysis. Fan et al. (2004) found that the conditional bias may be quite serious, even in situations in which the unconditional bias is acceptable. Most of the available adjustment methods (12-15) focus on the unconditional bias, which have little effect on the conditional bias (16).
In the clinical setting, when deciding whether to initiate a new therapy for a patient, or deciding which of several approved therapies to initiate, it is important to have a reliably precise gauge of the treatment effect. Uncertainty in the treatment effect complicates a physician’s benefit-risk assessment for an individual patient. This uncertainty should also be well-understood and taken into consideration by sponsors, investigators, DMCs and others involved in clinical trial conduct when deciding whether to terminate a trial early based on efficacy. In addition, clinicians and regulatory authorities also need to account for this potential overestimation and uncertainty in magnitude of effect when evaluating the efficacy and risk profile of a drug.
Table 1: Early-terminated phase III oncology trials with updated results reported

<table>
<thead>
<tr>
<th>Drugᵃ (total # of events)</th>
<th>Tumor Typeᵇ</th>
<th>Primary endpointᶜ</th>
<th># of events at interim analysis (%)</th>
<th>Medians (Exp / Ctrl in months)</th>
<th>HR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lap + Cap (266)</td>
<td>MBC</td>
<td>TTP</td>
<td>121 (46%)</td>
<td>8.4 / 4.4</td>
<td>0.49 (0.34, 0.71)</td>
</tr>
<tr>
<td>Sunitinib (471)</td>
<td>MRCC</td>
<td>PFS</td>
<td>250 (53%)</td>
<td>11 / 5</td>
<td>0.42 (0.32, 0.54)</td>
</tr>
</tbody>
</table>

ᵃLap = lapatinib, Cap = capecitabine;ᵇMBC = metastatic breast cancer, MRCC = metastatic renal cell carcinoma;ᶜTTP = time to progression, BC = breast cancer, PFS = progression-free survival

Captions for Figures

**Fig. 1.** Illustration of variability in estimation of true effect for fixed sample design (A) and group sequential design (B). Solid red line denotes variability; long-dashed line denotes null hazard ratio (HR) of 1.0; short-dashed line denotes true treatment effect derived as a HR; dotted blue line denotes statistical stopping boundary, where observed treatment effect estimates falling below the boundary have crossed the statistical threshold to trigger early trial termination due to efficacy. Arrows denote random highs that may coincide with interim looks at the data.

**Fig. 2.** Overestimation in GSDs. Simulation based on 1 million trial replicates; dashed line denotes “true” hazard ratio (HR) in the simulation of 0.60; dotted purple line denotes statistical stopping boundary; circles denote HR of each simulated trial replicate; dark red circles denote HR of simulated trials that crossed stopping boundary with corresponding average HR (95% CI) depicted by white dot.
Fixed sample design

A

Low

(-)

HR = 1.0

High

(+)

True treatment
effect

Time (no efficacy interim analyses)

Group sequential design

B

Low

(-)

HR = 1.0

Stopping boundary

High

(+)

Random highs

True treatment
effect

Time (efficacy interim analyses)

Statistics in CCR
Overestimation of the Effect Size in Group Sequential Trials

Jenny J Zhang, Gideon Blumenthal, Kun He, et al.

Clin Cancer Res  Published OnlineFirst July 2, 2012.

Updated version
Access the most recent version of this article at:
doi:10.1158/1078-0432.CCR-11-3118

Author Manuscript
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.