Defining information fractions in group sequential clinical trials with multiple endpoints

Tu Xu∗, Qin Qin, Xin Wang

AbbVie Inc., North Chicago, IL 60064, United States

1. Introduction

In late phase clinical trials, group sequential designs [1] have been popularly used to enable potentially early stopping for efficacy or futility to make promising drugs available to patients at an earlier time or allow ineffective drugs to be terminated earlier to save development cost and allow patients to take other promising treatments. The alpha spending approaches [2,3], such as The Lan-DeMets approach with Pocock or O’Brien-Fleming type [4,5], are commonly utilized for group sequential designs to determine decision boundaries at the planned interim and final analyses such that the overall type I error is strongly controlled at a pre-specified level (e.g. one-sided 0.025). The information fraction (IF), defined as the fraction of total information expected at the scheduled end of the trial [3], is an important parameter of the alpha spending function. In addition, it is generally assumed [6] that the correlation of standardized test statistics at two interim analyses equals \( \sqrt{t_1/t_2} \), where \( t_1 \) and \( t_2 \) are the corresponding IFs. Calendar time and information time are two types of popularly methods to define the information fraction. Lan and DeMets [7] compared the performance of calendar time and information time on the type I error and power, and pointed out that the type I error rate would be inflated when the information fraction is overestimated. They showed that if O’Brien-Fleming alpha spending function with one-sided significance level of 0.025 is used in a trial with 4 interim analyses, the one-sided type I error rate is inflated to 0.034 when the interim analyses are conducted at the estimated IFs \( t_{ik} \) = 0.3, 0.6, 0.8, 0.9 while the actual IFs are. \( t_{ik} = 0.1, 0.2, 0.3, 0.6, k \in [1,2,3,4] \).

Here we conduct a simple numerical evaluation for better illustration. Assume that for an endpoint \( E \), one interim analysis is planned to be conducted at the estimated IF \( t_{i} \), where \( 0 < t_i < 1 \). Let the standardized test statistics at the interim and final analyses be \( Z_{i0} \) and \( Z_{i1} \), where \( Z_i \) is the actual IF at the interim analysis. Denote the corresponding efficacy boundary calculated based on the IF \( t_i \) as \( c_i \) and \( c_1 \). Then the probability of crossing efficacy boundary throughout the trial is expressed as

\[
Pr(Z_{i0} \geq c_i \cup Z_{i1} \geq c_1) = 1 - Pr(Z_{i0} \leq c_i \cap Z_{i1} \leq c_1).
\]

Under the null hypothesis, the joint distribution of \( (Z_{i0}, Z_{i1}) \) is known to asymptotically follow a bivariate normal distribution [8].

\[
\begin{bmatrix}
Z_{i0} \\
Z_{i1}
\end{bmatrix} \sim N\left(\begin{bmatrix}0 \\
0\end{bmatrix}, \begin{bmatrix}1 & \sqrt{t_1} \\
\sqrt{t_1} & 1\end{bmatrix}\right).
\]

It was shown in Ref. [9] that \( Pr(Z_{i0} < c_i \cap Z_{i1} < c_1) \) is increasing w.r.t. \( t_i \). Under the null hypothesis, there is

\[
Pr(Z_{i0} < c_i \cap Z_{i1} < c_1) = 1 - \alpha \quad \text{if} \quad Z_i = t_{i}, \quad \text{where} \quad \alpha \quad \text{is the pre-specified significance level. Therefore, type I error rate would be inflated if} \quad t_i < t_{i}\;
\]

under the null hypothesis.

We further examine the magnitude of type I error inflation in two ways. Fig. 1 (A) depicts the change of type I error rate w.r.t. the actual IF \( t_i \), given that the efficacy boundary is calculated using O’Brien-Fleming boundary with the estimated IF \( t_i = 0.5 \) and a one-sided significance level 0.025. Fig. 1 (B) depicts the change of type I error rate w.r.t. the estimated IF \( t_i \), given that the actual IF \( t_i = 0.5 \) and efficacy boundary calculated using O’Brien-Fleming boundary with the estimated IF \( t_i \) and a one-sided significance level 0.025. Fig. 1 (B) shows that the type I error could be inflated as large as 0.03 if the IF is significantly overestimated.

∗ Corresponding author.
E-mail address: tu.xu@abbvie.com (T. Xu).

https://doi.org/10.1016/j.conctc.2018.03.005
Received 17 January 2018; Received in revised form 9 March 2018; Accepted 14 March 2018
Available online 17 March 2018
2. A case study

In Phase 3 confirmatory oncology clinical trials, the hierarchical testing procedure is frequently used for testing multiple endpoints. The strong control of family-wise error rate (FWER) is required for the statistical testing for multiple endpoints to be considered valid, where

![Fig. 1. Plot of type 1 error rate.](image)

![Fig. 2. Hierarchical testing procedure of primary and key secondary endpoints.](image)

**Table 1**
The summary of time to CR and ORR.

|                | 25th Pctl | Median | 75th Pctl | 90th Pctl | Min | Max |
|----------------|-----------|--------|-----------|-----------|-----|-----|
| Time to ORR (months) | 1.5       | 2.0    | 3.5       | 6.0       | 1.0 | 12.0 |
| Time to CR (months)  | 2.5       | 4.0    | 6.0       | 9.0       | 1.0 | 15.0 |

**Table 2**
The definition of information fraction.

| Endpoint                          | IF Definition                                                                 | Estimated IF |
|----------------------------------|------------------------------------------------------------------------------|--------------|
| OS                               | Fraction of total OS events accrued                                         | 50%          |
| ORR                              | Fraction of total patients with at least 6 months follow up since enrollment | 250/400 = 52.5% |
| CR rate                          | Fraction of total patients with at least 9 months follow up since enrollment | 150/400 = 37.5% |
| ORR rate by the initiation of Cycle 2 | Fraction of total patients who started Cycle 2 Day 1 treatment or discontinued by the end of Cycle 1 | 100%         |
the FWER is defined as the probability of falsely rejecting at least one null hypothesis. Glimm et al. [8] and Tamhane et al. [10] showed that the type I error rate for each ranked point must be strongly controlled and thus the alpha spending function needs to be pre-specified for each ranked endpoint under the group sequential setting. However, there was no further discussion on type I error rate inflation for each ranked endpoint that could be caused by the overestimation of IFs.

In practice, due to the operational convenience for regulatory submissions with the Phase 3 data, it is always preferred that the interim analysis for all key endpoints be conducted at the same time, where the timing is driven by the data maturity of the primary endpoint(s). Furthermore, for the same reason, the same alpha spending function applied for primary endpoint(s) is often used for other secondary endpoints, including the specification of IFs. However, due to heterogeneity among the nature of data maturity, such as those for overall survival (OS), progression free survival (PFS), and overall response rate (ORR), applying the same IFs for all endpoints could cause significant under-estimation/overestimation of IFs. It would potentially lead to the FWER inflation as discussed in Section 1. A commonly used alternative approach is to allocate conservative α to all secondary endpoints at interim analyses [7]. Although the strategy remedies the concern of type I error inflation, it sacrifices the opportunity of early stopping for all key secondary endpoints, and thus could cause a waste of development costs.

To overcome the aforementioned drawbacks, we propose to pre-specify the definition of IFs for each secondary endpoint by utilizing the preliminary results from Phase 1/2 clinical trials. For illustrative purpose, we consider a case study on developing the specification of IFs for each endpoint.

In a Phase 3 oncology trial, the primary endpoint is OS, and all key secondary endpoints are 1) overall response rate (ORR), 2) complete response (CR) rate, and 3) ORR rate by the initiation of Cycle 2. One interim analysis is planned to occur when 50% of total OS events are accrued. To strongly control the FWER, the primary endpoint and all key secondary endpoints are to be tested in a hierarchical testing procedure as illustrated in Fig. 2. The total sample size is set as 400.

It is known that the data maturity of response rate endpoints (CR rate and ORR) [11], which is the proportion of subjects with matured ORR or CR data, is different from time-to-event endpoints (OS). Therefore, we specify the IFs for CR rate and ORR at the interim analysis by considering the preliminary results of time to CR and time to ORR from Phase 2 trials as summarized in Table 1.

To better understand the data maturity of CR rate and ORR, a discussion is held within cross-functional study team to review the preliminary results aforementioned. It turns out that the 90th percentile is considered to be a clinically reasonable cutoff. Therefore, the definitions for IFs for CR rate and ORR are formed as outlined in Table 2.

For instance, given that at the time of interim analysis, there are 250 subjects with 6 month follow-up, 150 subjects with 9 month follow-up, and all 400 subjects have either started the cycle 2 treatment or discontinued treatment by the end of Cycle 1, an estimation of IFs is presented in Table 2.

As demonstrated in Table 2, our proposed framework provides more objective estimates of the IFs for each endpoint according to its corresponding data maturity mechanism. If the same IF as OS (50%) is applied to all endpoints, there is high risk that the IF of CR rate would be overestimated and the IFs of ORR and ORR by initiation of Cycle 2 would be underestimated. Correspondingly, the FWER could be inflated by testing CR rate and there is also potential power loss if ORR or ORR by initiation of Cycle 2 is tested at the interim analysis.

3. Discussion

Although the operating characteristics of the group sequential design has been well understood for each individual endpoint, attention still needs to be paid on the specification of IFs for group sequential designs with multiple endpoints. The numerical evaluation illustrates that the type I error could be inflated if the IFs are overestimated. On the other hand, allocating a conservative alpha for secondary endpoints at interim analysis would cause a waste of development cost in case that the trial could be stopped early if adequate α is allocated. We propose a framework that the IFs of each secondary endpoints are defined based on the outcomes from early phase trials. The framework provides a more objective way to estimate IFs such that the risk of severely overestimating/underestimating the IFs is mitigated. It thus preserves the FWER and allocates adequate α to each individual endpoint at the interim analysis. Especially, for those endpoints that is fully matured at the interim analysis, the full α could be utilized for testing.

Disclosure

This manuscript was sponsored by AbbVie, Inc. AbbVie contributed to the design, research, and interpretation of data, writing, reviewing, and approving the content. Tu Xu, Qin Qin, and Xin Wang are employees of AbbVie, Inc. The authors thank the Associate Editor and referees for their constructive comments and suggestions.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.conctc.2018.03.005.

References

[1] S. Emerson, T. Fleming, Symmetric group sequential test designs, Biometrics 45 (1989) 905–923.
[2] S. Wang, A. Tsilim, Approximately optimal one-parameter boundaries for group sequential trials, Biometrics 43 (1987) 193–199.
[3] D. DeMets, Interim analysis: the alpha spending function approach, Stat. Med. 13 (1994) 1341–1352.
[4] G. Lan, D. DeMets, Discrete sequential boundaries for clinical trials, Biometrika 70 (1983) 659–663.
[5] G. Lan, Information and information fractions for design and sequential monitoring of clinical trials, Commun. Stat. Theor. Meth. 23 (1994) 403–420.
[6] G. Lan, J. Witte, The B-value: a tool for monitoring data, Biometrics 44 (1988) 579–585.
[7] G. Lan, D. DeMets, Group sequential procedures: calendar versus information time, Stat. Med. 8 (1989) 1191–1198.
[8] E. Glimm, W. Maurer, F. Bretz, Hierarchical testing of multiple endpoints in group sequential trials, Stat. Med. 29 (2010) 219–228.
[9] S. Kotz, N. Balakrishnan, N. Johnson, Continuous multivariate distributions, volume 1: models and applications, Wiley, 2000.
[10] A. Tamhane, C. Mehta, L. Liu, Testing a primary and a secondary endpoint in a group sequential design, Biometrics 66 (2010) 1174–1184.
[11] K. Kelly, S. Halabi, Oncology Clinical Trials: Successful Design, Conduct and Analysis, Demos Medical Publishing, 2009.