Therapeutic Innovation & Regulatory Science (2021) 55:1001–1005
https://doi.org/10.1007/s43441-021-00299-4

ORIGINAL RESEARCH

Proof of Concept: Drug Selection? Or Dose Selection? Thoughts on Multiplicity Issues

Qian H. Li[1] · Qiqi Deng[2] · Naitee Ting[2]

Received: 5 December 2020 / Accepted: 20 April 2021 / Published online: 24 May 2021

Abstract

In new drug development process, one of the most important milestones for a drug candidate is to establish Proof of Concept (PoC) at early Phase II stage. Among many challenges in PoC clinical trial design and analysis, the application of multiplicity comparison procedures (MCP) is frequently discussed when multiple doses or drugs are included in one PoC study. In such discussion, one fundamental question of applying multiplicity adjustment is which error one should consider to control and at what level. Should it be the experiment-wise error or the compound-wise error? In this paper, the multiplicity issues in two cases of PoC studies are used as examples to discuss the concept of different types of error and the level of the error rate control. With a clear understanding of the type of error and error rate control, the debate of applications of the multiplicity adjustment procedures in the PoC studies can be reconciled.

Keywords Error rate control · Proof of concept · Multiple comparisons · Phase II clinical trials

Background

In clinical development, new drug candidates go through pre-clinical and various phases of clinical development. In the Phase II stage of clinical development, one of the most important clinical trials is the Proof of Concept (PoC) study.

During the development of drugs treating for chronic diseases except for oncology treatment, Phase I clinical trials usually recruit healthy volunteers with the objectives of studying pharmacokinetics (PK) and maximally tolerable dose. The reason is that the patient population may affect the metabolism and tangle the safety features of the drug candidate. In addition, patients tend to take other medications to control their conditions. The medication they take could interact with the study drug so that the PK property as well as adverse events could be confounded with other factors. Because drug efficacy cannot be observed from healthy volunteers, PoC studies are the first time to evaluate drug efficacy and subjects recruited are patients with the target disease.

A PoC study is designed to help the drug developer make a "Go/NoGo" decision based on the efficacy performance of a drug candidate. if the drug candidate demonstrates efficacy, the concept is considered proven Ting [1]. A "Go" decision from the PoC result implies further development of this candidate, leading to dose ranging from Phase IIb studies to long term and large-scale Phase III studies. PoC is one of the most important milestones for a drug candidate Ting et al. [2]. One of the challenges is that it may not be ethical to expose a large number of patients in a PoC study. Furthermore, the drug developer is not certain whether this candidate has a bright future and would prefer to take incremental steps. Based on both reasons, the sample size in a PoC study is somewhat limited. A classic PoC is designed with a highest dose allowable based on Phase I clinical trial results to compare with placebo. Due to the increased need of efficiency in drug development, the traditional PoC study design has been evolved with multiple dose studies or perhaps multiple drug candidates evaluated in one PoC study by sharing the same placebo control.

The scope of this paper will focus on the multiplicity issue in the PoC study design. Many statistical decision rules such as multiplicity adjustment procedures are pre-specified in PoC study design, in a hope to control the error rate of

✉ Naitee Ting
naitee.ting@boehringer-ingelheim.com

[1] Global Biometrics and Data Sciences, Bristol Myers Squibb – Global Biopharmaceutical Company, NewYork, USA

[2] Biostatistics and Data Sciences, Boehringer Ingelheim Pharmaceuticals, Inc, Ridgefield, USA

decision-making and make the decision process less ambiguous. However, many statistical decision rules are not clear on what error that is controlled and often tangled with different types of error considerations. In the frequentist setting, many different types of error under null hypotheses are all categorized as Type I error, which need to be further clarified so that the decision-makers can consciously understand what error that they are controlling and their tolerance to the error, i.e., the level the error rate controlled.

Drug Selection? Or Dose Selection?

Two cases will be introduced in this paper to elaborate the concept of error control: one is to include two drug candidates in a PoC study and the other is to include two doses of the same candidate in a PoC study. In both cases, the key question is to control the Go/NoGo decision error rate for the drug candidates. By nature of PoC studies, the focus is on the drug efficacy and one-side hypotheses is used throughout the paper. Bonferroni correction is used as an example to explain the concepts that this paper intends to clarify.

Case 1

One example of more than one drug candidates included in one PoC study comes from the pulmonary therapeutic area where both beta agonist and anticholinergic agents can reduce the severity of disease symptoms. A drug developer may consider developing one beta agonist and another anticholinergic agent with the hope that both drugs can be approved and marketed about the same time. The two drugs may be later combined to develop a combination product. Another example is that two candidates have passed the preclinical development and Phase I tests. The PoC study will include both drug candidates perhaps to select the front-runner and leave another as a back-up.

In these examples, a parallel three-group study is designed for candidate A, candidate B, and placebo P. The two sets of statistical hypotheses, corresponding to each candidate can be written as the following:

$$H_{0\mathrm{A}} : m_{\mathrm{A}} \leq m_{\mathrm{P}},\ \mathrm{vs}\ H_{1\mathrm{A}} : m_{\mathrm{A}} > m_{\mathrm{P}} \tag{1a}$$

$$H_{0\mathrm{B}} : m_{\mathrm{B}} \leq m_{\mathrm{P}},\ \mathrm{vs}\ H_{1\mathrm{B}} : m_{\mathrm{B}} > m_{\mathrm{P}}, \tag{1b}$$

where μ_{A}, μ_{B}, and μ_{P} denote mean effect of candidates A, B, and place, respectively.

With this study design, many statisticians' first instinct is to control the family-wise or experiment-wise type I error. A multiplicity control procedure, like Bonferroni correction could be considered. If the significance level is controlled at one-sided α, each hypothesis is tested at the α/2 level. Therefore, the error rate of Go/NoGo decision for each drug candidate is controlled at the level of α/2. The significance level reflects the level of acceptance to the error rate which also represents the level of confidence of decision-making.

However, if two separate trials are conducted to test the same two sets of hypotheses depicted in (1), most statisticians would not suggest using multiplicity adjustment. Therefore, each hypothesis will be tested at α level, which means that the error rate of Go/NoGo decision for each drug candidate is controlled at the level of α, rather than α/2. One cannot help to wonder for the same sets of hypotheses, why the decision would be made at different confidence levels? What exactly the Type I error rate is in this case?

Case 2

When two doses for one drug candidate are considered for a PoC study, a parallel three-group trial includes placebo (P), low dose (L) and high dose (H). Two sets of statistical hypotheses corresponding to each dose can be written as the followings:

$$\begin{aligned} &H_{0\mathrm{H}} : m_{\mathrm{H}} \leq m_{\mathrm{P}},\ \mathrm{vs}\ H_{1\mathrm{H}} : m_{\mathrm{H}} > m_{\mathrm{P}} \\ &H_{0\mathrm{L}} : m_{\mathrm{L}} \leq m_{\mathrm{P}},\ \mathrm{vs}\ H_{1\mathrm{L}} : m_{\mathrm{L}} > m_{\mathrm{P}}, \end{aligned} \tag{2}$$

where μ_{H}, μ_{L}, μ_{P} denote mean effect of high dose, low dose, and placebo, respectively.

At the stage of PoC, the objective is to use multiple doses to collectively confirm the efficacy or perhaps to gain preliminary understanding on the range of dose selection, rather than identifying the optimal dose(s) for drug labeling. The PoC objective may not be necessarily to identify the doses that will be used in the Phase III studies either as such a decision is usually based on Phase IIB dose ranging trials. Nevertheless, some statisticians may feel obligated to use MCPs due to multiple dose levels. If again the Bonferroni correction is considered, each hypothesis should be tested at the α/2 level.

It is possible that the selected dose(s) for the PoC study may not be able to reach statistical significance at the α/2 level as a limited sample size is usually used. For example, let α = 0.025 and the significance level for testing each dose will be 0.0125. If the one-sided p values are $p_{\mathrm{L}} = 0.060$ and $p_{\mathrm{H}} = 0.026$ for the low and high doses, respectively, results fail the Bonferroni correction. However, the evidence may be considered promising and providing sufficient confidence for a Go decision. It is possible that by changing the dose levels as well as sample sizes and other design characteristics such as treatment duration, the candidate may show statistical significance in the Phase III program. In this situation, should the question be how likely to observe such results if indeed the drug was ineffective? Which error rate should be controlled that could provide a high confidence for a Go/NoGo decision?

Control the Error Rate of Go/NoGo Decision

Error Rate

In hypothesis testing, one fundamental principle is to control the probability of making a false positive decision, i.e., the Type I error rate. In both cases discussed above, the experiments employ two pairwise comparisons for two sets of hypotheses, where MCPs are habitually applied. When multiple hypotheses are introduced, the error rate control has been widely discussed in classical statistical literature to control a family-wise error rate (FWER) or experiment-wise error rate (EWER) at the level of α Hochberg and Tamhane 1987 [3]. However, the EWER or FWER may carry different meaning in different study designs and for different purposes.

To address the questions raised before, it may be helpful to introduce a candidate-wise error rate for Go/NoGo decision. This candidate-wise error rate directly controls the decision error rate of the individual drug candidates. The candidate-wise Type I and Type II error rates are the error rates of wrongly moving an ineffective drug to a Go decision and wrongly moving an efficacious drug to a NoGo decision, respectively.

Case 1

The objective of the PoC study in Case 1 is to make Go/NoGo decision for each drug candidate A or B or both. The experiment-wise null hypothesis includes a hypothesis that is the intersection of the two individual null hypotheses in (1), i.e., H_0: $H_{0A} \cap H_{0B}$. H_0 implies that neither candidate is efficacious. The need of controlling the EWER is to suggest that the Go decision of one drug candidate needs to take the number drug candidates in a trial into consideration. If Bonferroni correction is applied, each candidate should be tested at the significance level of α/2 if the EWER is controlled at α. If three drug candidates are included in a trial, the decision error rate for each drug candidate needs to be controlled at α/3. This logic is in contrary to the common practice in drug development. The control of EWER may limit the use of an efficient study design and add more confusion in the decision process. As discussed earlier, if separate trials were conducted to test the same two hypotheses, most statisticians would not suggest applying multiplicity adjustment. The Go decision for one candidate should not suffer more stringent criteria because another candidate is being tested in the same trial or in a separate trial. The error rate control should fit for the purpose of the study rather than study unit. Therefore, the candidate-wise error rate should be controlled at the level of α.

There could be potential risks associated with the strategy of designing one PoC trial to evaluate two drug candidates. One of the risks is an under-performed placebo response. If a higher group mean indicates a better response, then an "under-performed placebo response" implies that the observed placebo group mean is lower than the true placebo mean. When this is the case, the observed treatment difference between candidate A and placebo as well as between candidate B and placebo could be both false positive. Bai et al. [4] provided a detailed discussion on such error rate and showed that such an error rate is not critical based on numerical results obtained from both simulations and numerical integration.

Another scenario could arise is that when limited resource is available at a time, it may be the plan that only one drug should be moved forward at a time. Then the decisions of which one to go between the two candidates can depend on the relative performance of the two drug candidates. In addition, the selection between two efficacious candidates should be done by comparing the complete efficacy and safety profile available. Multiplicity adjustment is not relevant here.

Therefore, it is recommended that the error rate in the experiment discussed in Case 1 does not need the MCP adjustment. Each hypothesis in (1) can be tested at the level of α. That is to control the drug candidate-wise error rate, instead of EWER.

Case 2

In Case 2, two doses are included for the purpose of reinforcing the Go/NoGo decision. It may not serve the purpose of determining if the individual doses will be optimal for treating patients. It is possible that a PoC study may combine both purposes of proof of concept and dose-ranging if properly designed. This topic is beyond the scope of this manuscript. Therefore, regarding Case 2, the hypothesis testing will be focused on the null hypothesis that the candidate is an ineffective drug. This null hypothesis H_0 can be written as the intersection of the two individual null hypotheses depicted in (2), i.e., H_0: $H_{0L} \cap H_{0H}$, with corresponding alternative H_1: H_{1L} U H_{1H}. The false positive implies that the Go decision is made when in fact the candidate is not efficacious, neither dose works. The FWER or EWER are the same as compound-wise error rate in this case. This error rate should be controlled, however, only needs to be controlled weakly. The weakly control usually corresponds to tests known as global tests.

For weak control of FWER, *F* tests in one-way ANOVA models can be considered. Other possible choice is to use trend tests similar to the test portion of MCP-MOD Pinheiro et al. [5]. If monotonic dose–response relationship is expected, such that $p_L = 0.060$ and $p_H = 0.026$, it is easy to see that the probability of observing such or more extreme results ($p_L \leq 0.060$ and $p_H \leq 0.026$) under null is low. That is, it is very unlikely to observe such results when the drug candidate is ineffective.

In the anti-psychotic therapeutic area, one may not see monotone increased dose–response relationship (as can be found in the label of many anti-psychotic drugs). Therefore, when the one-sided p values are $p_L = 0.015$ and $p_H = 0.800$ for the low and high doses, respectively, the concept can be considered as proven. Contrast tests might need to be modified to cover the non-increasing trends of dose–response relationship. A useful general discussion on the trend test can also be found in a paper by Li and Lagakos [6].

The Bonferroni adjustment will indeed control the FWER, even strongly Henning and Westfall [7]. However, such adjustment can be considered unnecessarily stringent as the doses studied in PoC trials may not always be able to achieve statistically significance at the level of $\alpha/2$. On the other hand, even the study results satisfy the Bonferroni adjustment, the results may not provide high confidence for a Go decision. Careful evaluation of the totality of evidence may be needed to support or against the Go decision. An extreme hypothetical example may help to illustrate the concern: suppose that ten active doses of the same drug are tested against placebo in a single clinical trial, where nine doses provide relatively large p values, say around 0.5 for one-sided tests, but only one dose (not the highest dose) is significant at $\alpha/10$. Even though such results satisfy Bonferroni correction, the totality evidence may lead hesitation for a Go decision. Therefore, Bonferroni correction may not be a sensible approach in such PoC studies either.

Discussion

In drug approval, both FDA and the drug developer bear the burden of controlling the error rate of wrongly approving /launching an ineffective drug to the market. The level of error rate control is 0.000625 ($= 0.025^2$) when two positive Phase III studies are recommended Li and Huque [8] for substantial evidence of efficacy. Similarly, PoC can be considered a confirmatory trial at a critical milestone in drug development. Hence error rate control is a key consideration in this development step. It is important to understand that the recommendation in this manuscript does not suggest letting go the error rate protection. The emphasis is to understand what the error is so that reasonable methods can be applied.

The concept of compound-wise error control can also be applied to a trial that includes multiple drugs from multiple companies and compares to a control group. A well-known example is the case of Ebola study Mulangu [9] where the test products were made by different pharmaceutical companies. The study has argued against multiplicity adjustment, stating that "The current circumstances of high mortality, intermittent outbreaks, and the need to find effective treatments as quickly as possible argue for less austere statistical penalties… Each of the primary comparisons of remdesivir, MAb114, and REGN-EB3 with ZMapp was tested at a two-sided type I error rate of 5%, without adjustment for multiplicity…" This study in fact controlled the candidate-wise error rate.

Many objectives of Phase II clinical development may be folded into the PoC studies. In addition to drug selection from multiple drug candidates and dose/regiment selection for later phases, the sponsor needs to consider selecting patient population, treatment duration, endpoints for novel disease indications, and instruments for patient reported outcomes. One major issue is the selection of primary endpoints, as the primary endpoints play important roles in understanding treatment effect, sample size considerations, study implementation, and data analysis. In addition, the preliminary assessment of benefit/risk profile, potential marketing competition, etc. also need to be considered in the objectives of a Phase II study. Many of the objectives will also play roles in the decision process in the Co/No/Go decision.

PoC is an important step in drug development. Timely Go/NoGo decision is needed to avoid delay in development program, as expiration of patent exclusivity is always a consideration in drug development. Any delay in this decision because of unclear PoC results will erode the patent life of this candidate. If a Go decision was made with low confidence, it may require cautious steps in later stages of drug development. Such as the case of a CNS drug discussed earlier, adding additional doses near the low dose and obtaining assessments of objective endpoints may further strengthen the evidence in the next steps of drug development.

References

1. Ting N. Confirm and explore, a stepwise approach to clinical trial designs. Drug Inf J. 2008;42:545–54.
2. Ting N, Chen D, Ho S, Capppelleri J. Phase II clinical development of new drugs. New York: Springer; 2017.
3. Hochberg Y, Tamhane AC. Multiple comparison procedures. New York: Wiley; 1987.
4. Bai X, Deng Q, Liu D. Multiplicity issues for platform trials with a shared control arm. J Biopharm Stat. 2020;30(6):1077–90.
5. Pinheiro JC, Bretz F, Branson M. Analysis of dose-response studies—modeling approaches, dose finding in drug development. New York: Springer; 2006.
6. Li QH, Lagakos SW. On the relationship between directional and omnibus statistical tests. Scand J Stat. 2006;33:239–46.

7. Henning KS, Westfall PH. Closed testing in pharmaceutical research: historical and recent developments. Stat Biopharm Res. 2015;7(2):126–47.
8. Li QH, Huque MF. A decision rule for evaluating several independent clinical trials collectively. J Biopharm Stat. 2003;13:621–8.
9. Mulangu S, Dodd LE, Davey RT, et al. Randomized controlled trial of ebola virus disease therapeutics. N Engl J Med. 2019;10:1056.