Statistical Considerations for Studies with a Performance Goal

Heng Li[1], Yunling Xu[1], Xu Yan[1]
[1]Food and Drug Administration, 10903 New Hampshire Ave., Silver Spring, MD 20993

Abstract
In the safety and effectiveness evaluation of medical devices, randomized controlled trials (RCTs) provide the highest level of evidence. However, it is not always feasible to conduct an RCT due to practical or ethical reasons. As such, a single arm study in which comparison against a pre-specified performance goal is made plays a substantial role in pre-market device evaluation. While this type of study design has potential benefits, such as savings in cost or time, statistical challenges arise regarding the validity of study design and the interpretability of study results. In this paper, we will discuss these challenges, focusing on issues with weighted performance goal.

Key Words: Performance goal, medical device, single arm study

1. Introduction

A single-arm study plays an important role in the pre-market evaluation of safety and effectiveness performance of an investigational medical device. In such a study design, usually the primary endpoint is compared with an objective performance criterion (OPC) or a performance goal (PG). In this short note, we will first briefly discuss the differences between PG and OPC in Section 2, and then discuss when a single-arm study with a performance goal can be considered and when it is suboptimal in Section 3 and Section 4. Section 5 will focus the discussion on the weighted performance goal. The last Section will summarize the pros and cons of a single-arm study with a performance goal.

2. OPC vs. PG

Per FDA Guidance (2013), an Objective Performance Criterion (OPC) refers to a numerical target value derived from historical data from clinical studies and/or registries and may be used in a dichotomous (pass/fail) manner by FDA for the review and comparison of safety or effectiveness endpoints. While a Performance Goal (PG) refers to a numerical value (point estimate) that is considered sufficient by FDA for use as a comparison for a safety and/or effectiveness endpoint. From the definition, the difference between these two terms is not apparent. But an OPC is usually developed when

- device technology has sufficiently matured, and
- can be based on publicly available information or on information pooled from all available studies on a particular kind of device, and
- a subject-level meta-analysis is preferred

Whereas a PG does not provide a level of evidence that is as rigorous as an OPC. The device technology is not as well-developed or mature for use of a PG as for an OPC, and

the data used to generate a PG is not considered as robust as that used to develop an OPC. Furthermore, an OPC typically cannot be developed by a single company, or unilaterally by FDA. While a PG in current practice is developed by a single company for a particular submission, even though this is not recommended by the FDA Guidance (2013).

Both OPC and PG will tend to have greater validity if it is commissioned or adopted by a medical/ scientific society or a standards organization or is described in an FDA guidance document. Both may become obsolete over time as technology improves and as additional knowledge accrues.

Since no control group is involved, comparison to an OPC or PG cannot demonstrate either superiority or non-inferiority. For example, if the null and alternative hypotheses are defined as follows: H_0: $p \geq 8\%$, H_1: $p < 8\%$, then when the null hypothesis is rejected an appropriate claim would be that the performance goal of 8% is met.

3. When a PG study can be considered

There are several scenarios in which a study with a PG can be considered for planning a medical device study. For example, when the target patient population lies in the grey area between patients who are suitable for medical therapy and patients who are suitable for open surgery, it is challenging to specify an optimal treatment for the control arm in a randomized controlled trial (RCT). Therefore, a study with a PG may be considered in this scenario. Another scenario is that the current standard of care is the off-label use of a device: since in a regulatory setting for a pre-market study it is inappropriate to use a device off-label as the control, there is no clinical equipoise for control in planning an RCT. Therefore, a study with a PG may be considered under this scenario. While the two examples above are cases where an RCT may not be feasible, the following example describes the scenario when RCT can be planned but may not be the least burdensome approach. Sometimes if the device technology is mature and its performance is well understood but an OPC has not been developed, then a study with a PG can be considered.

When a PG study is planned, several factors need to be taken into consideration in developing the PG, which include but is not limited to: whether the data used to develop the PG is from the same target patient population; whether the data used to develop the PG uses the same measurement and same definition of the primary endpoints; and whether the data reflects the current medical practice. Clinical/engineering input is of paramount importance when developing a PG. The value of PG should not be dictated by the sample size, nor should it be developed based on the investigational device's own previous data.

4. When a PG study is not optimal

A PG study may not be optimal when the performance of the current standard of care is not well understood/established, or when the treatment effect is expected to vary significantly across different subgroups. For example, in the treatment of a certain disease, the success rate may be affected by many important covariates, and there may be interactions among these covariates. In such a case, it is difficult to develop a one-size-fits-all PG.

When a PG study is not an optimal approach, alternative approaches can be considered. Depending on the feasibility and other design factors, one may consider

- Designing a randomized controlled trial, or
- Narrowing down the target patient population to the subgroup that is clinically most important, or
- Establishing for each subgroup its own PG, if each subgroup can be clearly defined.

If the target patient population can be divided into two mutually exclusive subgroups, we have seen submissions in which a study with weighted performance goals is proposed. However, we do not recommend this approach due to difficulty in interpretation. We will discuss our concerns on this approach in the next section.

5. Weighted PG

For illustrative purposes, let us consider a hypothetical proposal for a clinical study for device X, intended for treating subjects at high surgical risk. The subjects can be classified into type A or B (mutually exclusive). The null and alternative hypotheses as formulated by the sponsor are:

H_0: $P_{MAE} \geq PG$, H_1: $P_{MAE} < PG$,

where P_{MAE} = proportion of subjects experiencing one or more MAE,

$PG = w_1 * 22\% + w_2 * 28\%$,

with 22% and 28% being the performance goals for subjects of type A and B, respectively,

w_1 = observed proportion of subjects of type A
w_2 = observed proportion of subjects of type B
$w_1 + w_2 = 1$,

and the test statistic is expressed as: $$z = \frac{\widehat{P_{MAE}} - PG}{\sqrt{\widehat{P_{MAE}}(1 - \widehat{P_{MAE}})/n}}$$

The above construction is evidently inappropriate from a statistical perspective. Since w_1 and w_2 are not pre-specified as fixed values, but will be determined by the final enrollment, they should be considered as random variables and hence cannot appear in a hypothesis. Moreover, in the test statistic, w_1 and w_2 are treated as constants, which is self-contradictory.

To address this issue, one possible solution is to treat w_1 and w_2 (=1-w_1) as sample proportions and rewrite the hypotheses as follows:

H_0: $W_1(P_a - 22\%)+(1 - W_1)(P_b - 28\%) \geq 0$
H_1: $W_1(P_a - 22\%)+(1 - W_1)(P_b - 28\%) < 0$,

where W_1 represents the true proportion of subgroup A in the target population (and (1-W_1) represents the true proportion of subgroup B in the target population).

Because w_1 and w_2 (=1-w_1) . are not specified as a constant, the actual enrolled proportion of each subgroup may not be close to its true proportion in the target population. As such, the observed value $w_1(p_a - 22\%)+(1 - w_1)(p_b - 28\%)$ may not represent the comparison between the true MAE rate and the PG of the target population, especially when the enrolled proportion deviates significantly from the true proportion in the real world. Therefore, rejecting H_0 does not necessarily indicating the true MAE rate of the target population is below the PG.

An alternative solution is to fix the enrollment proportions based on the population proportion. Suppose the sponsor can pre-specify the values of the true proportions in the target population to be 0.35 and 0.65, respectively. Then, the value of PG can be calculated as:

PG = 0.35 * 22% + 0.65 * 28% = 25.9%

The null and alternative hypotheses become:

H_0: $P_{MAE} \geq 25.9\%$, H_1: $P_{MAE} < 25.9\%$

However, it may be difficult to enroll each subgroup with the same proportion as the pre-planned one without delaying study completion. So the sponsor proposes the minimum and maximum enrollment proportions for Type A subjects as 0.2 and 0.5. As such, the observed MAE rate in the study can still be a biased estimate of the true MAE rate in the target patient population, if the actual enrolled proportions of subgroups A and B deviate from 0.35 and 0.65.

Observed MAE rate in the study: $\widehat{P_{MAE}} = \frac{n_a}{n} p_a + \frac{n_b}{n} p_b$

True MAE rate in the target patient population: $0.35*P_a + 0.65*P_b$.

Since the PG is derived for the population based on the weights of 0.35 and 0.65 for each subgroup, the comparison between the observed MAE rate and the PG may not be appropriate. We can further illustrate this concern using Figure 1. If the enrolled proportion of subgroup A and subgroup B is 0.2 and 0.8, respectively. Then the upper bound of 95% confidence interval of the estimated MAE rate should be the area under the blue line. While if the enrolled proportion for each subgroup is 0.35 and 0.65 respectively, then the upper bound of 95% confidence interval of the estimated MAE rate should be the area under the orange line. Similarly, if the enrolled proportion for each subgroup is 0.5 and 0.5 respectively, then the upper bound of 95% confidence interval of the estimated MAE rate should be the area under the purple line. From Figure 1 one can see that if the enrolled proportion of each subgroup deviates from 0.35 and 0.65, then the hypothesis reject region is impacted.

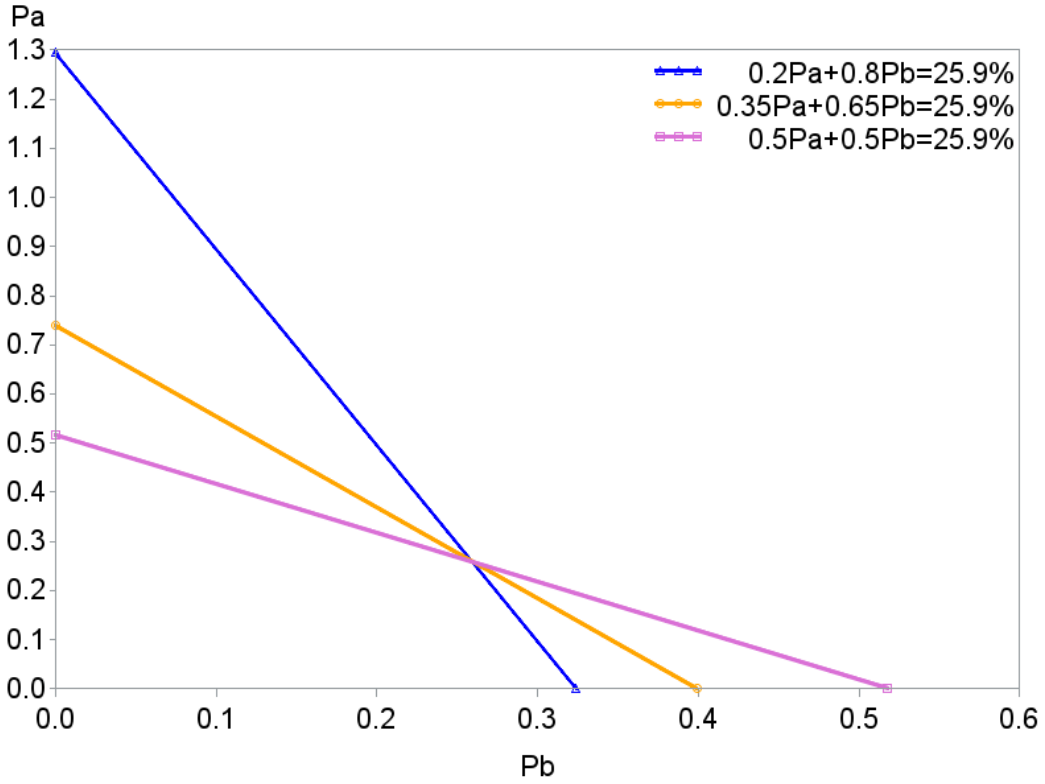

Figure 1: Observed MAE Rate

An alternative solution can be to redefine the hypotheses as follows regardless of what the actual enrolled proportion is for each subgroup:

H_0: $0.35 * P_a + 0.65 * P_b \geq 25.9\%$
H_1: $0.35 * P_a + 0.65 * P_b < 25.9\%$

Or equivalently

H_0: $0.35 * (P_a - 22\%) + 0.65 * (P_b - 28\%) \geq 0$
H_1: $0.35 * (P_a - 22\%) + 0.65 * (P_b - 28\%) < 0$

The test statistic can be: $$z = \frac{0.35p_a + 0.65p_b - 0.259}{\sqrt{0.35^2 \times \frac{p_a(1-p_a)}{n_a} + 0.65^2 \times \frac{p_b(1-p_b)}{n_b}}}$$

This approach compares the unbiased estimate of MAE rate for the overall target patient population with its performance goal. However, under this approach, rejecting H_0 does not necessarily indicate $P_a < 22\%$ and $P_b < 28\%$. For example, Figure 2 gives two scenarios when the observed MAE rate in subgroup A (or B) is very high but the null hypothesis is still rejected. As such, it may be difficult to interpret the study results for the overall target population.

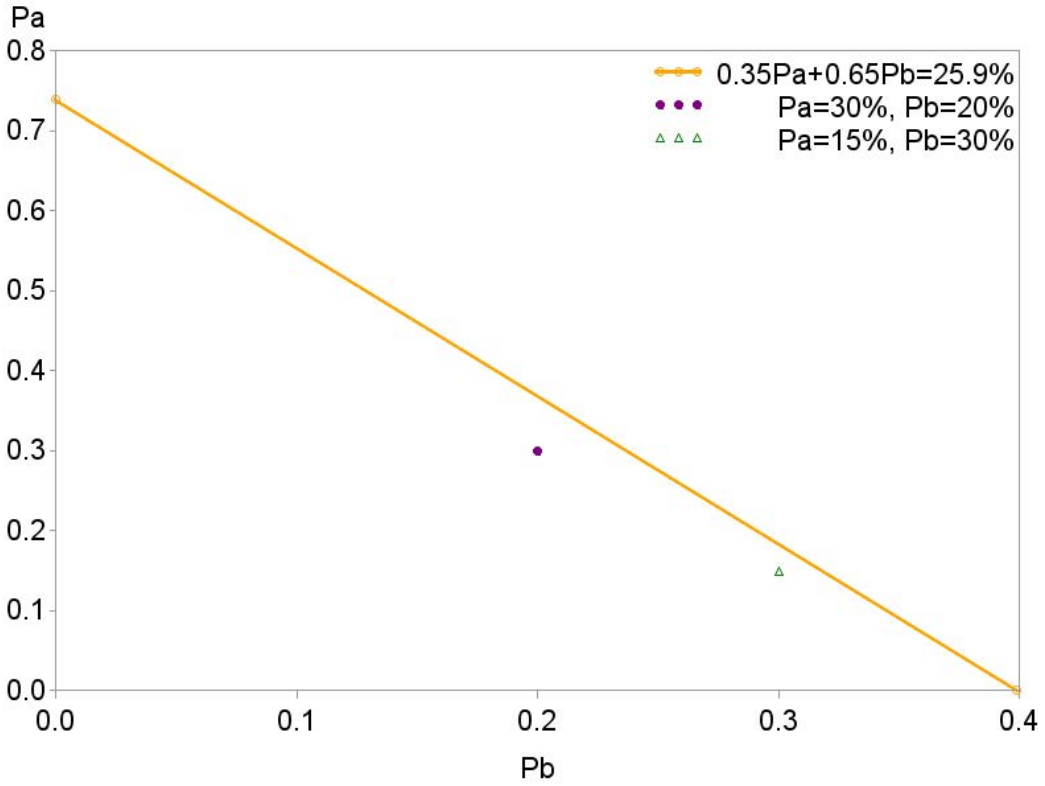

Figure 2: Observed MAE Rate

6. Discussion

In summary, there are pros and cons when planning a single-arm study with a PG. This type of study is considered cost efficient, simple, and easy to be conducted. However, since it lacks a comparison group, the investigational device cannot be compared with a treatment group. Furthermore, due to the nature of its non-blinded study design, selection bias and evaluation bias may be introduced. Sometimes the PG is not well developed, so the PG may not be consistent for the same indication for use among the same types of devices. For a study with weighted PG, caution needs to be exercised in interpreting the results.

Acknowledgements

Thanks to Dr. Ram Tiwari, Dr. Lilly Yue, and other colleagues from CDRH/Division of Biostatistics for their valuable suggestions.

References

1. Food and Drug Administration (FDA). 2013. Guidance for Industry, Clinical Investigators, and Food and Drug Administration Staff: Design Considerations for Pivotal Clinical Investigations for Medical Devices (released November 7, 2013). http://www.fda.gov/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm373750.htm (accessed April 15, 2015).