On Weighted Performance Goals in Medical Device Single-Arm Clinical Studies

Nelson Lu and Yunling Xu
Center for Devices and Radiological Health, US Food and Drug Administration, 10903 New Hampshire Ave., Silver Spring, MD 20993

Abstract

Under some circumstances, a pre-market application of medical devices may be supported by a single arm study where clinical endpoint is compared with a performance goal (PG). Occasionally, performance is expected to be different between two clinically defined subgroups of patients, and an approach based on a weighted PG is adopted. In this article, we discuss considerations of current practice where weights need to be pre-specified and fixed. We also propose an approach which relaxes the fix weights requirement.

Key Words: Performance goal, medical device, subgroup

1. Introduction

The randomized controlled trials (RCT) are considered as a gold standard for causal inference. The data collected from a well-designed and well-conducted RCT may provide the strongest evidence in evaluating the effectiveness and safety of an investigational device in the premarket setting. However, due to feasible or ethical reasons, conducting a RCT is not always practical.

Alternatively, observational (nonrandomized) studies may be utilized in device evaluation under some circumstances. If a comparative conclusion is not critical in the regulatory decision making or such a claim is not sought after by a sponsor, it is common that the primary endpoint results from such an open label, single arm study are compared against a numerical value. A common type of a numerical value is a Performance Goal (PG). A PG can be considered sufficient for use as a comparison for a safety and/or effectiveness endpoint for some kinds of medical devices. [1]

Sometimes, in the design stage, the performance is expected to be greatly different between two clinically defined subgroups. When a PG approach is appropriate, separate PGs may need to be specified. A proper way is to compare results from subjects of each subgroup to the corresponding PG. However, in some medical device fields, it is not uncommon to compare results of combined subjects against a weighted PG for sake of saving cost.

Our discussion is mainly focused on the binary endpoint as it is the most common type in medical device studies. In this paper, random variable Y denotes the clinical outcome. $Y = 1$ indicates the favorable clinical outcome and $Y = 0$ otherwise. Let S be the random

variable denoting subgroup; $S = i$, $i = 1, 0$. For each i ($i = 1, 0$), PG_i, w_i, π_i represents the PG, weight, and success rate in the Subgroup i, respectively.

2. Current Practice

In this section, we describe two approaches that are commonly proposed in the submissions.

Approach 1
Letting $PG = w_1 PG_1 + w_0 PG_0$, the null and alternative hypotheses are listed below:

$$H_0: \pi \leq PG$$
$$H_1: \pi > PG\ ,$$

where π denotes the success rate. This is often proposed to be tested using a one-sample binomial test.

In this approach, the weights w_1 and w_0 are not meant to be parameters. Instead, they are required to be pre-specified and fixed constants, and the proportions of the enrolled subjects from both groups are controlled at w_1 and w_0. The hypothesis is tested using a one-sample binomial test.

Note that, such a test is inappropriate from the theoretical perspective. One assumption of the test is that each observation follows the Bernoulli distribution with the success rate π. This assumption may be violated as the success rates are different between the two subgroups.

Approach 2
Letting $PG = w_1 PG_1 + w_0 PG_0$, the null and alternative hypotheses are listed below:

$$H_0: \pi_1 w_1 + \pi_0 w_0 \leq PG$$
$$H_1: \pi_1 w_1 + \pi_0 w_0 > PG.$$

In this approach, the weights w_1 and w_0 arepre-specified constants. However, unlike the enrollment condition required for adopting Approach 1, there are no constraints on the proportions of the enrolled subjects for each group. A common way to test the hypothesis is to use a Z-test expressed below:

$$\mathrm{Z} = (w_1 \hat{\pi}_1 + w_2 \hat{\pi}_2 - PG)/\sqrt{\mathrm{var}(w_1 \hat{\pi}_1 + w_2 \hat{\pi}_2)},$$

where $\mathrm{var}(w_1 \hat{\pi}_1 + w_2 \hat{\pi}_2) = w_1^2\, PG_1(1 - PG_1)/n_1 + w_2^2\, PG_2(1 - PG_2)/n_2$.

It can be observed that Z is mathematically valid for any proportion of actual enrollment. However, if actual enrollment of a subgroup is greatly below the pre-specified weight, Z may be greatly influenced by these subjects. To avoid such a situation, a cap is usually put on the enrollment for each subgroup.

3. Proposed Approach

The natural and reasonable interpretation for the weight is that it represents the proportion of subjects in the target population. This target population may reflect the population in intended use of the investigational device.

Knowledge in the proportion of subgroups in the overall patient population may not lead to a correct idea in proportion of subgroups in the target population. For example, asymptomatic patients may be less likely to implant a device than symptomatic patients. Therefore, even if there are more asymptomatic patients, the population to receive the device may consist of higher proportion of symptomatic patients.

Both approaches described in Section 2 require weights to be pre-specified constants. However, sometimes it may be challenging to have them specified correctly in the design stage. Issues do arise when the weights are wrongly specified.

First, wrongly specified weights make it difficult to interpret the estimated quantity. Observed that

$$\pi = \Pr(Y = 1) = \Pr(S = 1)\Pr(Y = 1|S = 1) + \Pr(S = 2)\Pr(Y = 1|S = 2)$$
$$= \Pr(S = 1)\,\pi_1 + \Pr(S = 2)\pi_2.$$

When weights are wrongly specified (i.e.$\Pr(S = 1) \neq w_1$), π does not always equal to $w_1\pi_1 + w_2\pi_2$. Consequently, the estimation of $w_1\pi_1 + w_2\pi_2$ does not match the actual goal of estimating π, $\Pr(Y = 1)$.

Second, enrollment time may be significantly prolonged with off-target ore-specified weights. This can be illustrated by the following example, which mimics a true case. In the design phase, 300 subjects were planned to be enrolled with w_1 to be set to be 0.7. The enrollment cap was set at from 65% to 75% for Subgroup 1, or 195 to 225 subjects. The enrollment cap for Subgroup 2 was thus 25% to 35%, or 75 to 105 subjects. After one year, a total of 150 subjects (half of the sample size) had been enrolled. Of those 150 subjects, 50 were in Subgroup 1 and 100 in Subgroup 2. Clearly, the enrollment of Subgroup 2 almost reached the cap, while only about a quarter of the Subgroup 1 subjects were enrolled. With a similar enrollment rate, another three years were needed to finish the enrollment. Note that without the enrollment caps, the enrollment might be finished in about one more year.

Our approach is proposed to avoid the issues caused by incorrectly specified weights. If weights cannot be reasonably specified at the design stage, they need to be estimated. In doing so, let ω, a parameter, denote the true proportion of Subgroup 1: $\Pr(S = 1) = \omega$, $\Pr(S = 0) = 1 - \omega$. Therefore, $\pi = \pi_1\omega + \pi_0(1 - \omega)$.

The covariance of *Y* and *S* can be derived. As
$\mathrm{E}(YS) = \Pr(Y = 1, S = 1) = \Pr(S = 1)\Pr(Y = 1|S = 1) = \omega\pi_1$,
$\mathrm{Cov}(Y, S) = \mathrm{E}(YS) - \mathrm{E}(Y)\mathrm{E}(S) = \pi_1\omega - \pi\omega = \omega(\pi_1 - \pi) = \omega(1 - \omega)(\pi_1 - \pi_0)$.

Writing $\boldsymbol{Y} = (Y \quad S)'$, which is distributed in a bivariate Bernoulli with correlation $\rho = \omega(\pi_1 - \pi)/\sqrt{\pi(1 - \pi)\omega(1 - \omega)}$.

With a random sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$,
$\begin{pmatrix}\bar{Y}\\ \bar{S}\end{pmatrix} \dot{\sim} N\left(\begin{pmatrix}\pi\\ \omega\end{pmatrix}, \boldsymbol{\Sigma}/n\right)$, where $\boldsymbol{\Sigma} = \begin{pmatrix}\pi(1-\pi) & \omega(\pi_1-\pi)\\ \omega(\pi_1-\pi) & \omega(1-\omega)\end{pmatrix}$, by central limit theorem.

The hypothesis can be formulated in a similar fashion adopted as the current practice:

$$H_0\colon \pi_1\omega + \pi_0(1-\omega) \le PG_1\omega + PG_0(1-\omega)$$
$$H_1\colon \pi_1\omega + \pi_0(1-\omega) > PG_1\omega + PG_0(1-\omega).$$

This can be expressed as

$$H_0\colon (1 \quad PG_0 - PG_1)(\pi \quad \omega)^t - PG_0 \le 0$$
$$H_1\colon (1 \quad PG_0 - PG_1)(\pi \quad \omega)^t - PG_0 > 0.$$

The hypothesis may be tested using a Z-test as expressed below:

$$Z = n\,(c^t\bar{\boldsymbol{Y}} - PG_0)/\sqrt{c^t\hat{\boldsymbol{\Sigma}}c}$$

where $c^t = (1 \quad PG_0 - PG_1)$.

In order for this method to work, $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ needs to be a random sample. The method may be broken if the sampling method is inappropriate. An example of inappropriate sampling scheme is that a cap enrollment is in place for the subgroup.

4. Discussions

A couple of points need to be mentioned regarding the paradigm including the formulation of the hypothesis.

In the setup, there are essentially two parameters, π_1 and π_0. However, the hypothesis is on a linear combination of these two parameters. Therefore, the two-dimension parameter space is reduced to the one-dimension space. The statement in the alternative hypothesis does not necessarily imply that both $\pi_1 > \mathrm{PG}_1$ and $\pi_0 > \mathrm{PG}_0$. The one-dimensional hypothesis only works perfectly if π_1 and π_0 have some deterministic relationship.

In order for the hypotheses formulation working fairly well, certain assumptions on the parameters, π_1 and π_0, are needed. Although in theory each of these parameters takes on the space of (0, 1), in practice there should be some relationship between them as the same device is applied to two populations of patients. It is reasonable to assume that the performances between two groups are somewhat positively correlated. To elaborate this, consider a case where subgroup 1 represents a more sever condition and poorer clinical results are almost guaranteed for this type of patients. There are probably no reasons to believe that $\pi_1 > \pi_0$.

It may be necessary to evaluate $\Pr(\text{Rejecting } H_0 | \pi_i \le \mathrm{PG}_i, H_1)$ for various reasonable scenarios of π_1 and π_0 to assess the risk. This probability is above the significance level since it is under the alternative hypothesis. If some of the evaluated probabilities are unacceptable, some actions need to be taken. One possibility is to lower the significance level of the test. Another is to put additional success criteria for each subgroup. One

simple case of such criteria is to sample average for each subgroup needs to be greater than the associated performance goal.

5. An Example

In this section, a hypothetical example is provided to illustrate the use of the proposed approach.

A certain type of disease is currently treated with a class of devices under off-label use. An investigational device has been newly developed to aim the treatment of such disease. Based on the clinical experience and rationale, the clinical performance of the investigational device is expected to be different between asymptomatic ($S = 1$) and symptomatic patients ($S = 0$) regarding the primary effectiveness endpoint, which is 30-day success. The performance goals are set at PG_1= 0.8 and PG_0= 0.6 based on the clinical experience.

Say that the one arm study is proposed and accepted. To avoid a situation that the proportions of symptomatic and asymptomatic patients in the target population are wrongly specified, the proposed approach is adopted. A total of 300 subjects are planned to be enrolled. The success rules are listed in the following:

(1) $Z > 1.96$
(2) $\hat{\pi}_1 > PG_1$ and $\hat{\pi}_0 > PG_0$

Rule (1) indicates that the test outlined in Section 3 is conducted at one-sided significance level of 0.025. Rule (2) is in place as an attempt to address issues discussed in Section 5.

A simulation study has been conducted to elaluate the operating characteristics of this procedure. For a particular set of (π_1, π_2, ω), a random sample of bivariate Bernoulli $\boldsymbol{Y} = (Y \quad S)'$ can be generated using, for example, the method proposed by Park, Park, and Shin [2]. To implement this method, the following quantities were computed first:

$\alpha_{11} = -\log(\pi)$,
$\alpha_{12} = \log\left(1 + \rho\sqrt{(1-\pi)(1-\omega)/\pi\omega}\right) = \log(1 + (\pi_1 - \pi_2)(1-\omega)/\pi)$, and
$\alpha_{22} = -\log(\omega)$.

For each simulated data set, $Z_1 \sim Poisson(\alpha_{11} - \alpha_{12})$, $Z_2 \sim Poisson(\alpha_{22} - \alpha_{12})$, and $Z_3 \sim Poisson(\alpha_{12})$ were drawn. The observations *Y* and *S* were then derived following the expression below:

$Y = Z_1 + Z_3$, and
$S = Z_2 + Z_3$.

The success rules could be examined based on this simulated data set. A total of 5000 data sets were simulated. The probability of meeting the success criteria was estimated by the proportion of data sets satisfying the success rules.

The following table presentes the simulation results. The top three cases are associated with scenarios under the H_0, the middle three cases under H_1, and the bottom three under Under H_1, but $\pi_0 < PG_0$.

ω	π_1	π_0	$\omega\pi_1 + (1-\omega)\pi_0$	$\omega PG_1 + (1-\omega)PG_0$	Probability meeting success criteria
			Under H_0		
0.25	0.8	0.6	0.65	0.65	0.023
0.50	0.8	0.6	0.70	0.70	0.029
0.75	0.8	0.6	0.75	0.75	0.029
			Under H_1		
0.25	0.88	0.68	0.76	0.65	0.87
0.50	0.88	0.68	0.78	0.70	0.92
0.75	0.88	0.68	0.83	0.75	0.90
			Under H_1, but $\pi_0 < PG_0$		
0.25	0.88	0.58	0.655	0.65	0.05
0.50	0.88	0.58	0.73	0.70	0.19
0.75	0.88	0.58	0.805	0.75	0.34

Additional reasonable or concerned scenarios under H_1 and $\pi_i < PG_i, i = 0, 1$ may need to be assessed. If some of these probabilities are not acceptable, stricter success criteria may be imposed.

6. Concluding Remarks

Regarding the study design under a situation where the performance of a medical device is expected to be greatly different between two subgroups, the randomized controlled trial where randomization is carried out by subgroups is considered to provide the highest level of evidence in pre-market evaluation. If a comparative claim is not pursued, a one-arm study where results from each subgroup are compared to the associated PG, using separate hypothesis testing by subgroups, may be considered.

Many sponsors utilize the weighted PG approach for this situation. While such a design and the associated statistical analysis method have been accepted in some medical device applications, their potential limitations may not have been addressed extensively. The fundamental issue is that the hypothesis testing is only conducted on a linear combination of the two performance parameters that are associated with two subgroups.

If the concerns can be adequately addressed, the current practice (Approach 1 and 2 presented in Section 2) can be adopted if weights can be reasonably correctly identified. Otherwise, our proposed approach may be considered.

References

[1] US Food and Drug Administration. (2013) Design considerations for pivotal clinical investigations for medical devices - guidance for industry, clinical investigators, institutional review boards and Food and Drug Administration staff. November 7, 2013. Available at: http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM373766.pdf. Accessed on September 22, 2016

[2] Park, G.P., Park, T., and Shin, D.W. (1996) A simple method for generating correlated binary variates. *The American Statistician* 50(4): 306-310