Inflation of a Type II Error Rate in Three-Arm Non-Inferiority Trials

Nor Afzalina Azmee

Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, 35900 Tanjong Malim, Perak, MALAYSIA
E-mail: afzalina@fsmt.upsi.edu.my

ABSTRACT

The aim of a non-inferiority trial is to demonstrate that the new experimental treatment is not worse relative to the reference treatment by more than a pre-defined margin. Assuming that the inclusion of a placebo arm is properly justified, the three-arm non-inferiority trial is termed as the gold standard design and should be used whenever possible. This study focuses on the problem of sample size determination in three-arm non-inferiority trials, a crucial matter that needs to be addressed in the early stage of clinical trial. The current two-stage procedure involved in the analysis of three-arm non-inferiority trial presents a problem known as the inflation of type II error rate. In other words, the sample size obtained does not ensure enough power to reject the null hypothesis when it is false. This paper illustrates the problem via simulation study and proposes an alternative solution using assurance.

Key Words: Clinical trial, gold-standard design, sample size, power, assurance

1. Introduction

The aim of a non-inferiority trial is to demonstrate that the new experimental treatment is not worse than the reference treatment by more than a pre-defined margin. Although the fundamental idea of testing non-inferiority of a new treatment was given in the early 1980s, it was not until early 2000s that considerable interest were shown among the medical practitioners and academicians. Assay sensitivity is a problem, widely recognized as unavoidable when aiming to establish non-inferiority of the new treatment with respect to reference treatment (see Temple and Ellenberg (2000) and Pigeot et al. (2003)). For this reason, whenever and wherever possibly justified, the inclusion of an extra placebo arm is seen as a solution to solve the assay sensitivity problem. This type of clinical trial design is known as the three-arm non-inferiority trial and is termed as the gold standard design.

Determining a sample size prior to executing a clinical trial is crucial and is given an utmost attention by the practitioners (see Pigeot et al. (2003), Julious (2004), Brasher and Brant (2007)). In practice, the calculation of sample size should be made transparent in the protocol as this partly is the basis of granting the approval to run a clinical trial. The ethical committee will be concerned if the sample size is too large or too small. It is deemed unethical to enrol patients in a small trial as the desired treatment may not be demonstrated and patients may be exposed to unnecessary risk. Similarly, a sample size that is excessively large is frowned upon as the results from having additional patients

are not beneficial and will slow down the process of marketing an effective drug to the public.

A required sample size can be determined either by employing the frequentist method or the Bayesian method, although the latter approach, more often than not will be subjected to further debates among the members of the ethical committee. Nevertheless, the Bayesian approach offers a natural solution in quantifying the prior information about the unknown parameters. The approach highlighted in this paper adopts the Bayesian point of view. The method is known as assurance, defined as the unconditional probability that the trial will successfully reject the null hypothesis. This idea has been described earlier in O'Hagan and Stevens (2001) and O'Hagan et al. (2005), but not directly applicable in the setting of three-arm non-inferiority trials. Recently, Azmee et al. (2013) has developed the assurance formula in the three-arm non-inferiority trial, based on the ratio of means. Although the authors have acknowledged the problem of inflation of type II error rate seen in the simulation results, a solution has not been addressed. Therefore, the object of this paper is to illustrate a solution in tackling the inflation of type II error rate, using assurance.

2. Inflation of Type II Error Rate

The common statistical procedure for three-arm non-inferiority trials, described in Pigeot et al. (2003), is a two-stage testing, which begins with establishing superiority of reference over placebo. If only the first stage is successful, the testing will proceed to the second stage, aiming to establish non-inferiority of the experimental treatment with respect to reference treatment. Assuming that the outcome variables are normally distributed with common but unknown variances and that higher values correspond to better efficacy, the hypotheses statements in both stages are written as follows:

$$H_0 : \mu_R - \mu_P \le 0 \quad \text{versus} \quad H_1 : \mu_R - \mu_P > 0 \tag{1}$$

$$H_0 : \frac{\mu_E - \mu_P}{\mu_R - \mu_P} \le \theta \quad \text{versus} \quad H_1 : \frac{\mu_E - \mu_P}{\mu_R - \mu_P} > \theta \tag{2}$$

where μ represents the population mean, E, R and P denote the experimental, reference and placebo groups respectively and θ is the non-inferiority margin. The value of θ is positive, usually ranging from 0.5 to 0.8. Considering a linear contrast, the null hypothesis in (2) can be rejected if the following test statistic, T is shown to be greater than $t_{1-\alpha,df}$ where $t_{1-\alpha,df}$ refers to (1 - α) quantile of the central t-distribution, with significance level α and degrees of freedom $df = n_E + n_R + n_P - 3$.

$$T = \frac{\bar{X}_E - \theta\bar{X}_R + (1-\theta)\bar{X}_P}{\hat{\sigma}\sqrt{\frac{1}{n_E} + \frac{\theta^2}{n_R} + \frac{(1-\theta)^2}{n_P}}} \tag{3}$$

Note that $\bar{X}$ and n denote the sample mean and the sample size, respectively and $\hat{\sigma}$ is the unbiased estimator of the common but unknown σ. A rejection of

the null hypothesis in (2) will imply that the experimental treatment is at least $\theta \times 100$ percent as good as the reference treatment.

The procedure, as noted earlier in Pigeot et al. (2003) suffers an inflation of a type II error rate as the ratio $\rho = (\mu_E - \mu_P)/(\mu_R - \mu_P)$ increases. This is because the necessary sample size derived to achieve the main objective (which is non-inferiority) fails to maintain the type II error rate at the desired level. A frequentist solution provided by Pigeot et al. (2003), which is power adjustment for different values of ρ can only be applied if the optimal allocation of sample size is used. This paper offers a Bayesian solution, which can be applied, regardless of whether a balanced or unbalanced design is adopted by the practitioners. Details are given in the next section.

3. Assurance in Three-Arm Non-Inferiority Trials

This section demonstrates the implementation of assurance to find the required sample size, via simulation. As illustrated in O'Hagan et al. (2005) and Azmee et al. (2013), a Bayesian Clinical Trial Simulation (BCTS) is useful in avoiding complex integration and is applied in this work. Assume that the interest is to establish non-inferiority of the experimental treatment in a three-arm trial, then a sample size which tackles the inflation of a type II error rate can be determined by defining assurance as the probability of demonstrating both superiority of reference over placebo and non-inferiority of experimental relative to reference.

To illustrate the above idea, consider the following sampling distributions:

$$\begin{aligned} \overline{X}_E &\sim N(\mu_E, \sigma^2/n_E) \\ \overline{X}_R &\sim N(\mu_R, \sigma^2/n_R) \\ \overline{X}_P &\sim N(\mu_P, \sigma^2/n_P) \end{aligned} \tag{4}$$

Suppose, normal priors are assigned to the unknown parameters μ_E, μ_R and μ_P, with the following means, m_E, m_R and m_P and variances v_E, v_R and v_P. Suppose σ^2 has a log normal prior with mean a and variance b, represented as follows $\ln \sigma^2 \sim N(a,b)$. Thus, assurance via BCTS can be implemented by following the steps below.

i. Define the counters; $I = 0$ and $S = 0$, where I corresponds to a number of repetition and S corresponds to a number of successful event.
ii. Define the number of repetition, J. For example, $J = 1000$.
iii. Define the non-inferiority margin, θ. In this example, $\theta = 0.8$.
iv. Define the number of subjects in placebo arm, n_P and the allocation of sample size in the ratio of c_E:c_R:1, where c_E and c_R are the proportions of sample size in the experimental and reference groups with respect to those in the placebo groups. In this example, the sample size allocation is made to be 5:4:1 for experimental, reference and placebo groups, respectively.
v. Sample μ_E, μ_R and μ_P from the prior distributions specified. In this particular example, the prior distributions are specified as:

$$\mu_E \sim N(m_E, v_E), \quad \mu_R \sim N(m_R, v_R), \quad \mu_P \sim N(m_P, v_P)$$

vi. Sample σ^2 from the prior distribution specified, which is $\ln \sigma^2 \sim N(a,b)$.

vii. Using the results in (iv), (v) and (iv), sample the followings:

$$\bar{X}_E \sim N\left(\mu_E, \sigma^2/n_E\right), \quad \bar{X}_R \sim N\left(\mu_R, \sigma^2/n_R\right), \quad \bar{X}_P \sim N\left(\mu_P, \sigma^2/n_P\right)$$

viii. Using the results in (iv) and (vi), $\hat{\sigma}^2$ can be obtained by sampling from the chi-square distribution, with degrees of freedom, $n_E + n_R + n_P - 3$.

$$\frac{\hat{\sigma}^2\left(n_E + n_R + n_P - 3\right)}{\sigma^2} \sim \chi^2_{n_E+n_R+n_P-3}$$

ix. Using the results in (iii), (vii) and (viii), calculate the *T* statistic, given in Equation (3).

x. Using the results in (iv), (vii) and (viii), calculate statistic *U*:

$$U = \frac{\bar{X}_R - \bar{X}_P}{\hat{\sigma}\sqrt{\frac{1}{n_R} + \frac{1}{n_P}}}$$

xi. If both statistics *T* and *U* are greater than $t_{1-\alpha,df}$, where α = 0.025 and $df = n_E + n_R + n_P - 3$, update $S = S + 1$.

xii. Update $I = I + 1$.

xiii. While $I \le J$, repeat the following steps (v) – (xii).

xiv. Calculate assurance, *A*. Given the sample size in placebo arm as n_P or total sample size of $c_E n_P + c_R n_P + n_P$, assurance is $A = S/N$.

For illustration, consider plotting the assurance curves, given the following specifications; $m_E = 4.2$, $m_R = 3.8$, $m_P = 3.0$. Suppose the uncertainty regarding these values is represented with v_E, v_R and v_P set to be 0.04, although it is also possible to specify unequal values. This arrangement reflects the belief that a new treatment is thought to be better than the reference, with the ratio of the difference in means, $\rho = \left(\mu_E - \mu_P\right)/\left(\mu_R - \mu_P\right)$, possibly centred around 1.5 with some slight variation. Furthermore, suppose the uncertainty regarding σ^2 can be represented by setting $a = 0$ and $b = 0.0625$.

Figure 1 demonstrates the assurance curves based on Event 1 and Event 2. Event 1 is defined as "successfully establishing non-inferiority of the experimental treatment with respect to reference" whereas Event 2 is defined as "successfully establishing superiority of reference over placebo and establishing non-inferiority of experimental treatment with respect to reference". Suppose a sample size in the placebo arm is taken to be 17, which reflects a total sample size of $N = 170$, where patients are randomly allocated across the experimental, reference and placebo groups in a ratio of 5:4:1. Based on Figure 1, that particular sample size will yield an assurance of 58 percent, when considering Event 2. On the other hand, the same 58 percent assurance can already be achieved at a smaller sample size, if one considers Event 1, which is optimistically misleading.

Note that when the values v_E, v_R, v_P and b are set to be very small, the assurance curve (see Event 1 in Figure 2) will match the power curve, due to the adoption of strong priors. Based on Figure 2, power of 80 percent is achieved when the sample size in placebo arm is set as 11, which brings a total sample size of 110, in the ratio of 5:4:1 across experimental, reference and placebo groups. The sample size which is derived based on the main objective

however, is too small to power a superiority trial of reference against placebo, conducted in the first stage. This is the reason why a two-stage testing in the three-arm non-inferiority trials suffer an inflation of type II error rate.

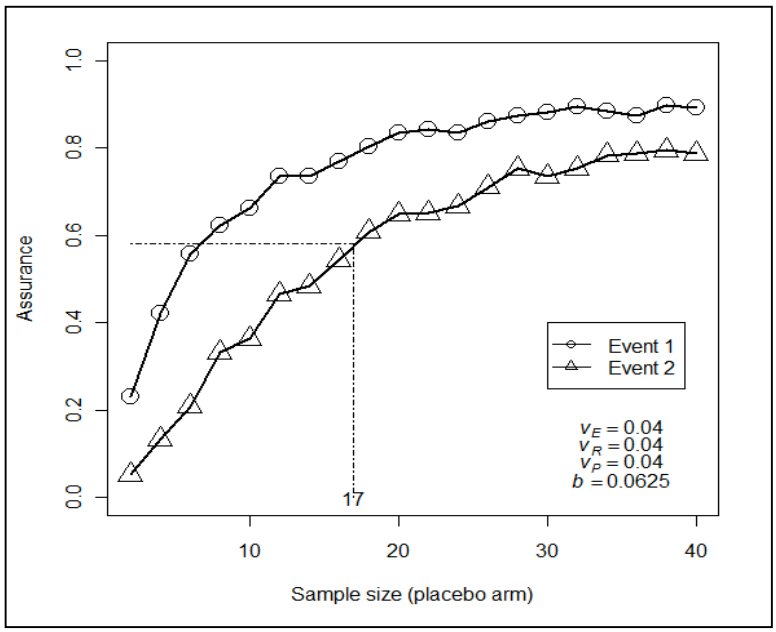

Figure 1 Comparison of assurance curves based on Event 1 and Event 2, using proper priors.

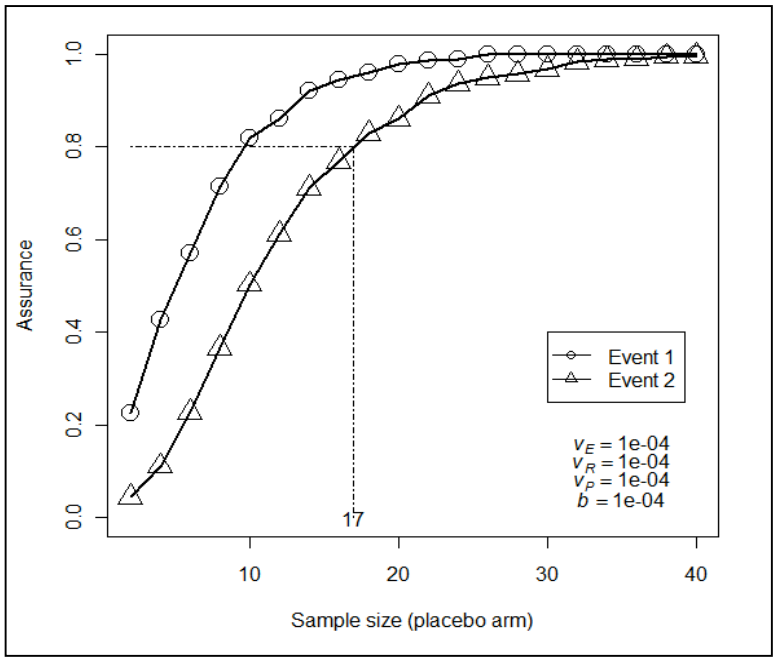

Figure 2 Comparison of assurance curves based on Event 1 and Event 2, using strong priors.

The inflation of a type II error rate can be tackled by considering assurance as probability of establishing superiority of reference against placebo and establishing non-inferiority of the experimental treatment (see Event 2, in Figure 2). For example, when the sample size in the placebo arm is 17, that

determined sample size is able to ensure that the desired 80 percent power is achieved. The result is in line with Pigeot et al. (2003), but the approach taken is different as it requires adjusting the type II error rate according to the ratio ρ, which works only for optimal allocation of sample size.

4. Conclusions

To conclude, the choice of a sample size based on the assurance concept is a subjective matter. Unlike power, it is not possible to fix an assurance of say γ for all situations, as different priors will lead to different values of assurance. This can be seen clearly in the simple example provided in Section 3. However, this approach is seen attractive as uncertainty is easier to be represented using a prior distribution rather than the point estimate. In particular, this paper has demonstrated that the common practice of determining the sample size based on the main objective (i.e. non-inferiority) is not sufficient to detect the desired effect of non-inferiority of the experimental treatment with respect to reference. Since the analysis of three-arm trial involves a two-stage testing, the sample size calculation must take into account both objectives; that is establishing superiority of reference against placebo and establishing non-inferiority of the experimental treatment.

References

Azmee, N.A., Mohamed, Z. and Ahmad, A. (2013) "Determination of the required sample size with assurance for three-arm non-inferiority trials", *Jurnal Teknologi*, 63, 89-93.

Brasher, P.M.A and Brant, R.F. (2007) "Sample size calculations in randomized trials: common pitfalls", *Canadian Journal of Anesthesia*, 54, 103-106.

Julious, S.A. (2004) "Sample sizes for clinical trials with normal data", *Statistics in Medicine*, 23, 1921-1986.

O'Hagan, A. and Stevens, J.W. (2001) "Bayesian assessment of sample size for clinical trials of cost-effectiveness", *Medical Decision Making*, 21, 219-230.

O'Hagan, A., Stevens, J.W. and Campbell, M.J. (2005) "Assurance in clinical trial design", *Biometrical Journal*, 48, 559-564.

Pigeot, I., Schafer, J., Rohmel, J. and Hauschke, D. (2003) "Assessing non-inferiority of a new treatment in a three-arm clinical trial including a placebo", *Statistics in Medicine*, 22, 883-899.

Temple, R. and Ellenberg, S.S. (2000) "Placebo-controlled trials and active-control trials in the evaluation of new treatments", *Annals of Internal Medicine,* 133, 455-463.