Application of Resampling Method in Futility Analysis Based on Multiple Studies

Jun Zhao[1], Jingshan Zhang[2] and Annpey Pong[1]

[1]Merck & Co., Inc., 126 E. Lincoln Avenue, Rahway, NJ 07065
[2]Celgene Corporation, 86 Morris Avenue, Summit, NJ 07901

Abstract

In late stage drug development, it is common to conduct multiple studies simultaneously in a development program in order to fulfill regulatory filing requirements as early as possible. However, the design elements such as sample sizes and corresponding powers are determined based on information from past studies at earlier phases. In addition, the assumptions may vary in terms of timing or conditions of conducting these new studies. Considering limited resource, the futility analysis is usually to be considered at interim to stop a study or a program early if the expected probability of success is relatively low. In this research work, a resampling method is proposed to assess the expected probability of success when a futility evaluation of multiple studies is performed in an interim analysis in a development program.

Key Words: futility, interim analysis, probability of success, Bayesian methods, resampling

1. Introduction

The development of a new drug is a lengthy, expensive and risky process. In order to fulfill regulatory requirement to provide substantial evidence of the effectiveness and safety for marketing approval of a new drug, sponsors may need to design an extensive program that includes two or more well-controlled large clinical studies (FDA, 1988). To design the new studies, the key elements, such as sample sizes and statistical powers, are primarily determined based on the information obtained from early stage clinical studies. However, in certain therapeutic areas such as psychiatric area, these design elements may change over time or with conditions at time of conducting new studies. Additionally, the high variability of treatment response may be expected in the population being studied. The phenomenon can be seen in a recent literature that pointed out the placebo response increased over time in the treatment of schizophrenia (Khin, Chen, et al, 2012). Therefore, due to large variation and uncertainty, it is quite unreliable to conduct a large program based on assumptions from the previously completed studies.

In the past few decades, interim analysis has been broadly utilized to evaluate the performance of a trial prior to completion of the study (Jennison and Turnbull, 1990). By definition in the ICH E6 guideline (1996), an interim clinical trial/study report is defined as a report of intermediate results and their evaluation based on analyses performed during the course of a trial. Interim analysis can be used to evaluate safety, efficacy, or

both safety and efficacy of a study drug. Furthermore, interim analysis can be used for assessing the study futility in order to prevent patients from being excessively exposed to an ineffective and/or unsafe treatment. Meanwhile, the early termination of an undesirable trial would also be beneficial for the sponsor in saving time and resource. Statistically speaking, an undesirable trial/program is the one with a low expected probability of success.

In general, the possible decisions at an interim analysis are: i) continue the study as it is (GO); ii) early termination (NOGO); (iii) continue the study but with modifications (Pong and Chow, 2010). These decision rules should be well planned and properly documented before the trial starts in order to ensure the integrity and validity of the clinical trial (Gallo, 2006). Note that the interim analysis methods and corresponding decision rules are usually applied to a single clinical trial. In this paper, we extend them to multiple clinical trials within a developing clinical program. Futility decision rules and various considerations on multiple studies by using probability of study success and probability of program-wise success have been discussed by Zhao and Pong (2011). In this paper, we use the resampling methods to calculate the study and/or program-wise probability of success based on multiple studies. For simplicity, the mid-course design modification, e.g. sample size modification (Jennison, 2003), is not considered in this research.

In Section 2, two examples from therapeutic area of psychiatry are presented to illustrate the background of the development program, and the rationale of conducting the program futility. The program-wise decision rules are discussed in Section 3. The algorithms of calculating the probabilities of success at study and/or program level are proposed in Section 4. Examples from clinical programs are also demonstrated in this section. The conclusion remarks of this paper are shown in Section 5.

2. Examples

Two examples in the treatment of schizophrenia and bipolar disorder in the therapeutic area of psychosis are given. Schizophrenia is a mental disorder characterized by a breakdown of thought process and poor emotional responsiveness. It most commonly manifests itself as auditory hallucinations, paranoid or bizarre delusions, or disorganized speech and thinking. In addition, it is accompanied by significant social or occupational dysfunction. Bipolar disorder is a chronic, typically cyclic mood disorder and is associated with marked social and occupational dysfunction, high rates of disability, frequent psychiatric comorbidity, and an increased risk of suicide.

[Example 1] Schizophrenia Program and Bipolar I Disorder Program

A drug candidate in the course of drug discovery and development had shown the potential benefit to the treatment of schizophrenia based on the analysis results from a well powered phase IIb efficacy and safety study for patients with acute schizophrenia. In addition, the drug candidate was expected to have similar potential for the treatment of manic or mixed episodes of bipolar I disorder as for the treatment of Schizophrenia, based on the clinical efficacy of antipsychotics with similar pharmacological properties (D2-dopamine antagonists). Therefore, two phase III large programs were initiated to investigate the efficacy and safety of this drug candidate in patients with schizophrenia, as well as in patients with manic or mixed episodes associated with bipolar I disorder, simultaneously.

Based on preclinical and clinical data from schizophrenia patients, the decision on phase III programs for fulfilling the regulatory requirement were: schizophrenia program to conduct three studies with different comparators; bipolar program to conduct two identical studies. It was expected that both programs would be at high risk due to uncertainty of placebo effect and high variation from the expected patient populations. For the bipolar program, one more concern was that there was no information available from early phases to support the program. In Section 4, we will demonstrate an ad-hoc interim futility analysis for the two identical studies in bipolar program.

[Example 2] Bipolar Depression Program

In bipolar disorder, depressive symptoms contribute significantly longer periods of time and contribute more suicide risk and functional disability than manic symptoms. A drug has been approved for treatment of manic or mixed episodes associated with bipolar I disorder. In addition to pre-clinical and early clinical data that show the drug may have clinical benefit for treatment of bipolar depression (BP-D), post hoc analyses on the patients with baseline depressive symptoms show strong signor of clinical benefit.

According to the regulatory agency, at least 2 positive studies are required to confirm the effectiveness of a new indication for a drug which has been approved in the market. Due to uncertainty and relatively high study failure rate in this area, the proposed clinical development program for this new indication was to conduct 4 studies simultaneously. Three identical studies included one targeted dose and placebo. The 4^{th} study contained the multiple doses of the tested drug that include the target dose and the matched placebo. Except the different number of treatment arms which result with different total sample sizes in study, the study design in the 4^{th} study are identical to the other 3 studies. The primary objective of these 4 studies is to show superiority of the target dose to placebo in reducing symptoms of bipolar depression.

It was expected that at least 2 out of 4 studies could reach the significance of the primary endpoint on the targeted dose. However, there is high risk of uncertainty on patient population which may cause a failed program. In order to minimize the time and resource for the possibility of not getting 2 positive studies, an interim analysis was proposed to assess the program-wise futility analysis to make the GO/NOGO decision for the entire program at interim.

3. Decision Rules of a Futility Analysis

Many methods have been proposed on futility assessment of a single study on different types of endpoints (Jennison and Turnbull, 2000). Our focus in this section is to develop a quantitative evaluation of program-wise futility at the middle of a program. It is to be noted that any interim analysis requires upfront plans and an independent data monitoring committee (DMC) to keep the integrity and validity of the development program. In general, the DMC conducts the interim analysis and communicates with the sponsor on the recommendation based on assessment of both efficacy and safety data (Ellenberg et al, 2002). The GO/NOGO decision rule at interim should be pre-defined clearly, and may include both efficacy and safety assessments.

It is well known that the probability of study success (PoSS), a conditional probability (Lachin, 2005) based on interim data can be calculated as follows:

$$CP(\delta) = \Pr(\text{test will reject null} \mid \text{interim observed data}).$$

The parameter δ is an assumed trend value, e.g. a z-score. Most commonly we use the estimated value at the interim analysis. Since this estimated value sometimes is inaccurate (e.g. Jennison and Turnbull, 2003), the null and alternative hypothesis values are also often considered.

Bayesian methods and simulations can also be used to estimate the PoSS. The approach is a weighted probability by the posterior distribution of δ, that is:

$$PP = \int CP(\delta)\, \pi\{\delta \mid \text{interim data}\}\, d\delta,$$

where the weight function $\pi\{\delta \mid \text{interim data}\}$ is the posterior distribution of the parameter of interest given data accumulated up to time of interim analysis. When multiple studies are conducted in a program, the posterior distribution can be estimated through either a single study or all the studies.

As mentioned previously, it is uncommon to have program-wise interim analyses because it contains a great benefit on the resources saving for a time-consuming program in clinical development. Our interest here is to develop the probability of program success (PoPS) that can be used to make a interim decision for the entire program, such as the probability of having two or more positive studies from the program eventually, based on the observed data at interim analysis. Then the PoPS can be written as, PoPS = Prob (two or more positive studies at the end | interim observed data).

Let p_j be the conditional probability of success for the j^{th} study in the program, where (j=1,…, J). Since the goal is to achieve at least 2 positive studies, the PoPS can be computed as below:

Case I: 2 studies

$$PoPS = p_1 p_2 \qquad (1)$$

Example: If $p_1=p_2=0.5$ then PoPS = 0.25. If $p_1=p_2=0.3$ then PoPS = 0.09.

Case II: 3 studies

$$PoPS = 1 - (1-p_1)(1-p_2)(1-p_3) - p_1(1-p_2)(1-p_3) - (1-p_1)p_2(1-p_3) - (1-p_1)(1-p_2)p_3 \qquad (2)$$

Case III: 4 studies (Example 2 in Section 2)

$$PoPS = 1 - (1-p_1)(1-p_2)(1-p_3)(1-p_4) - p_1(1-p_2)(1-p_3)(1-p_4) - (1-p_1)p_2(1-p_3)(1-p_4) - (1-p_1)(1-p_2)p_3(1-p_4) - (1-p_1)(1-p_2)(1-p_3)p_4 \qquad (3)$$

Assuming the conditional probabilities of each study are the same, e.g. $p_j=p$ (j = 1, …, J), the relationships between PoSS and PoPS can be shown as in Figure 1. For example, in Case III (4 studies), there is a 35% chance in program-wise success for a p=0.3; if p=0.4, then there is a 52% chance for a program-wise success. While in Case I (two studies), there is a 25% chance in program-wise success for a common p=0.5.

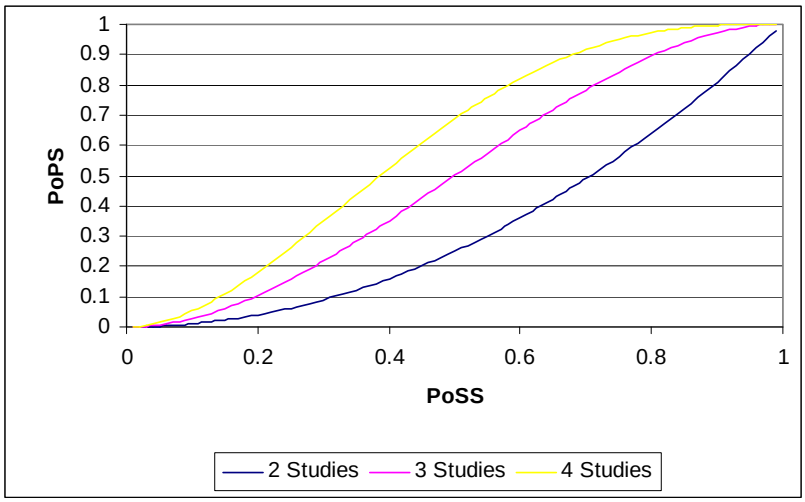

Figure 1 The relationship between PoPS and PoSS assuming an equal PoSS from each study.

Based on Zhao and Pong (2011), when a program includes 4 parallel studies, two strategies can be used to perform interim analysis on the futility assessment, namely single program-wise interim analysis and two-stage program-wise interim analysis.

Strategy #1

Single program-wise interim analysis can be applied when all studies have similar percentage of recruitment at interim stage. The interim analysis will be done when pre-defined percentage, e.g. 40% (Gould, 2005) of patients are included in the analysis. Case III may be applied to calculate the PoPS.

Strategy #2

Two-stage program-wise interim analysis is to be applied when studies have large difference in recruitment. In the approach, the 4 studies in the development program are divided by two sets: they are (a) two studies in the quick recruitment set (quick set), and (b) two studies in the slow recruitment set (slow set).

[Stage One]

The interim analysis is conducted when pre-defined percentage (e.g. 40%) of subjects are accrued from the first two studies. Let p_1 and p_2 be the conditional probability of success for the first 2 studies in the program. Case I above may be applied to calculate the stage one PoPS. At this stage, the high probability of study success (PoSS) from both studies is required in order to make a GO decision and skip stage two. Assuming the two studies are similar (e.g. $p_1=p_2=p$), the stage one PoPS is $PoPS_1 = p^2$. That means there is a 25% chance in program-wise success at stage one when p=0.5; similarly, there is a 49% chance in program-wise success at stage one when p=0.7. Based on the clinical experiences, we would recommend a probability greater than 49% to continue the entire program. Otherwise we need to get more data from stage two to further assessment on program futility (GO/NOGO).

[Stage Two]

In order to get comfortable level of making GO/NOGO decision, the interim analysis on stage two collects the data from the other two studies (slow set) in addition to the stage one data (quick set) which has been collected. For more precise estimations at interim

analysis, the sufficient sample sizes from the slow set are required. The recommended interim timing is when the study with the slowest recruitment rate has at least 25% of patients being randomized and having the endpoint assessment. Using the interim data from all four studies, the PoPS can be calculated based on the equation for Case III. The decision of stopping the program (NOGO) will be adopted when two/more out of the four studies have low probability of program success based on interim data.

The two-stage design can save the resource by more than half potentially. In addition, the two-stage design takes the differences in recruitment rates into consideration. This approach is particularly useful when recruitment rates are at two significantly different levels among studies. It is to be noted that there is only one interim analysis at most for each study. On stage two, the interim data contains the data from studies on stage one and studies on stage two. However, there is no new data to be added for interim analysis from studies on stage one although the studies are still on-going.

4. Algorithms for Calculating Conditional Probability

In previous section, we proposed a conditional probability based on interim data which can be calculated as follows:

$$CP(\delta) = \Pr(\text{study being positive at the end} \mid \text{interim observed data}).$$

There are many ways to calculate this conditional probability. If we know the underlying distribution of the test statistics of the primary hypothesis, e.g. normality assumption holds, the CP(δ) may be easy to calculate. However in reality, there are several concerns that include the skewness of the data, missing data due to early discontinuation, or the validity of modeling assumptions. In most cases, simulations using Bayesian methods or non-parametric methods may be used to calculate the conditional probability. Here we propose Bayesian type of resampling method for conditional probability based on interim data.

Efron (1979) discussed the use of bootstrap to generate sampling distribution of statistics and thereby to draw inferences about parameters. The Bayesian bootstrap (Rubin, 1981) is a natural Bayesian analogue of the bootstrap. Application to this technique has been found in Gu et al (2008) that applied it to ROC curve, and Merlo-Pich, et al (2009) that applied posterior probability, predictive power and on risk analysis based on non-parametric bootstrap simulation to make decision of whether stopping longitudinal trails on antidepressant drugs.

The bootstrapping methods and their application may be found in Davison and Hinkley (2003). The method to calculate the conditional probability for the PoPS is proposed as below:

There are two scenarios at the patient/subject level from multiple studies:

1) Responses from each study vary - therefore, the resampling of the prior data need to be from each individual study;
2) Responses from each study are similar - therefore the resampling of the prior data can be from the pooling of multiple studies.

Note that the regional differences in terms of baseline characteristics and treatment response are expected for the multi-regional trials (Hung et al, 2010). The resampling method would require more considerations on multi-regional clinical trials.

The simulated data of a whole study include two parts:

Part 1: data at interim analysis; and
Part 2: simulated data (from interim analysis to the study end)

The part 2 data are derived based on the resampling technique. The final analysis for treatment comparison and statistical significance are based on the combination of the above two parts.

The algorithms for simulations are shown below. An independent DMC should be established in order to unblind the treatment and perform analysis. Furthermore, the trial integrity should be maintained during the interim analysis.

Prior data from the same study

1. Pre-define percentage (%) of subjects recruited or randomized for the interim analysis in the current study.
2. Unblind the treatments. Without loss of generality,
    a. Assuming n_1 is the numbers of subjects (each treatment group, balanced design). n_1 represents the above percentage of total subjects to be used for interim analysis allocated to the treatment group;
    b. Assuming n_2 is the numbers of subjects (each treatment group, balanced design). n_2 is represents the number of simulated subjects such that the total number of subjects $n = n_1+n_2$ planned for the treatment group.
3. Get Data:
    a. Part I data: n_1 subjects from the current study.
    b. Part II data: n_2 subjects resampled from the interim n_1 subjects (within the same treatment group).
    c. If needed, considering resampling within "REGION", if regional difference is expected.
4. Simulate *r* times, and calculate number of times that the study/treatment claims positive (*m*). The statistical method used to calculate the significance should be the same method that stated in the protocol.
5. Estimated PoSS = *m/r*.

Prior data from the multiple studies

1. Pre-define percentage (%) of subjects recruited or randomized for the interim analysis in the whole program including the current study.
2. Unblind the treatments for all studies. Without loss of generality
    a. Assuming n_1 is the numbers of subjects (each treatment group, balanced design) of the current study. n_1 represents the above percentage of total subjects to be used for interim analysis allocated to the treatment group of the current study;
    b. Assuming n_2 is the numbers of subjects (each treatment group, balanced design). n_2 is represents the number of simulated subjects such that the total

number of subjects $n = n_1+n_2$ planned for the treatment group of the current study.

3. Get Data:
    a. Part I data: n_1 subjects from the current study. Note that n_1 may not exactly the same percentage of total subjects defined above in the current study.
    b. Part II data: n_2 subjects resampled from the interim data from the whole program.
    c. If needed, considering resampling within "REGION", if regional difference is expected.
4. Simulate *r* times, and calculate number of times that the study/treatment claims positive (*m*). The statistical method used to calculate the significance should be the same method that stated in the protocol.
5. Estimated PoSS = *m/r*.

Example

To illustrate the algorithms and corresponding results based on different settings, the studies in the bipolar program in the Example 1 of the Section 2 are demonstrated below. We retrospectively studied the bipolar program assuming an interim futility analysis was performed. The program contains two parallel studies; each study has more than 90% power to detect treatment difference. The endpoint to be analyzed is the change from baseline of the Y-MRS total score. Assuming there is no missing data and an ANCOVA model below was applied:

response ~ treatment + baseline value.

Note that these assumptions can be expanded for longitudinal data analysis methods, time to event survival analysis, or others. Since this is a retrospective study, the adjustment of type I or type II error due to interim analysis are not considered (Chang and Chuang-Stein, 2004; DeMets and Lan, 1994). Assuming the pre-defined percentage of patients/subjects to be included in the interim analysis, the inclusion of those subjects is based on the order of recruitment dates.

To well understand the results output from the simulation, the following are some of the background information on the two studies:

- Both studies were eventually positive ($p<0.01$ and $p<0.0001$). Therefore, the sponsor fulfilled the regulatory requirement in order to get an indication.
- Post hoc analysis showed that region A was less effective than region B in both studies, the difference was partly due to high placebo response in region A; however, both regions showed efficacy in the same direction.
- Operationally, region A started recruitment earlier than region B; and the recruitment of region A was faster than region B. Therefore, there were more patients than planned in the first half of each study.

The simulation results based on the observed data from the bipolar program are presented below.

Table 1 illustrated the results of PoSS/PoPS using the resampling method from the same study. The cut off % (40%, 50%, 55%) represented the percentage of subjects required in the interim analysis (resampling) for each study. Two methods were applied: 1) resampling from all interim subjects by treatment; 2) resampling interim subjects within region by treatment. Two test treatment groups were considered in study. Each test treatment was compared with placebo.

Table 1 Calculation of PoSS/PoPS – Prior Data from the Same Study

	Simple Resampling			Resampling within Region		
Cut Off %	Study #1	Study #2	PoPS	Study #1	Study #2	PoPS
Treatment 1 vs Placebo						
40%	<1.0	65.0	<1.0	<1.0	79.0	<1.0
55%	2.4	98.0	2.3	5.6	98.0	5.5
Treatment 2 vs Placebo						
40%	6.0	96.5	5.8	18.0	98.7	17.8
50%	47.8	>99.9	47.8	61.4	>99.9	41.4

Table 2 illustrated the analysis result of PoSS/PoPS for the multiple studies. The cut off % (40%, 50%, 55%) represented the percentage of subjects required in the interim analysis (resampling) from all studies. Therefore, the percentages of each study may be different due to recruitment rate difference for each study. The same as in Table 1, two methods were applied: 1) resampling from all interim subjects by treatment; 2) resampling interim subjects within region by treatment. Each test treatment group was compared with placebo.

Table 2 Calculation of PoSS/PoPS – Prior Data from Multiple Studies

	Simple Resampling			Resampling within Region		
Cut Off %	Study #1	Study #2	PoPS	Study #1	Study #2	PoPS
Treatment 1 vs Placebo						
40%	<1.0	53.4	<1.0	1.9	74.3	1.4
55%	5.4	79.4	4.3	11.2	90.4	10.1
Treatment 2 vs Placebo						
40%	44.2	93.0	41.1	64.6	96.9	63.0
50%	66.9	99.8	66.8	77.6	99.9	77.5

The analysis of PoSS for different % of cut-off data per study were illustrated in Figure 2 (Treatment 1 vs Placebo) and Figure 3 (Treatment 2 vs Placebo) by different resampling methods.

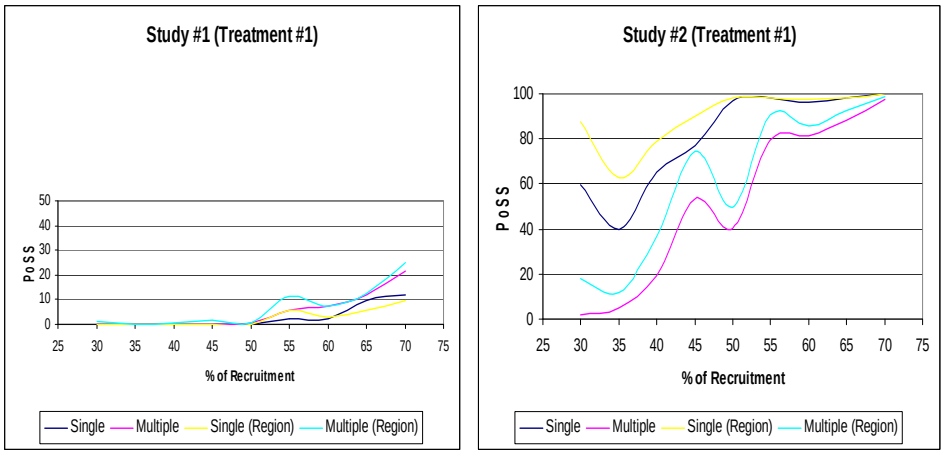

Figure 2 PoSS results per study by different cut-off and different resampling method (Treatment 1 vs Placebo).

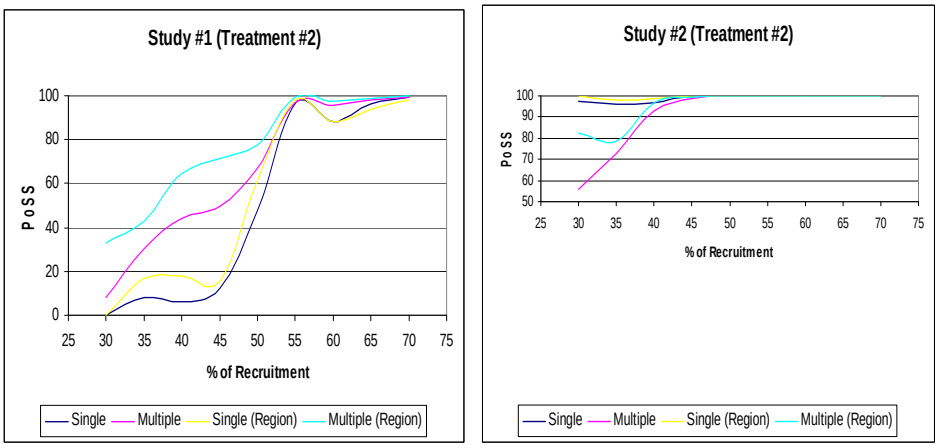

Figure 3 PoSS results per study by different cut-off and different resampling method (Treatment 2 vs Placebo).

5. Conclusion Remarks

The proposed methods/algorithms provide a quantitative solution to futility decision on multiple studies in a development clinical program. While the program-wise interim analysis in clinical development has clear benefit, proper plan can enable trials to achieve the objective in an efficient manner. PoSS/PoPS is useful to quantify the expected probability of study/program success at interim, even though no clear direction on how large the probability is supposed to be. In addition, the resampling method provides a 'non-parametric' way in calculating PoSS/PoPS for one or multiple studies. The proposed methods are practical for the decision rules for program futility which do not restrict to the analysis method from each individual study.

To legitimate the analysis result, the interim data should be representive to the whole population. If the same performance from multiple studies is expected, an interim

analysis at early time maybe considered. To develop a successful clinical program, some intrinsic/extrinsic factors such as regional difference in multi-regional global trials need to be carefully evaluated, at interim analysis.

References

Chang, W, Chuang-Stein, C. (2004). Type I error and power in trials with one interim futility analysis. Pharmaceutical Statistics, 3: 51–59

Davison, A.C. and Hinkley, D. V. (2003). Bootstrap methods and their application. Cambridge University Press

DeMets, D., and Lan, K.K.G. (1994). Interim analysis: The alpha spending function approach. Statistics in Medicine, 13, 1341-1352

Efron, B (1979). Bootstrap methods: another look at the jackknife. Ann. Statistics, 7, 1-26

Ellenberg, S.S., Fleming, T.R., DeMets, D.L. (2002). Data monitoring committees in clinical trials: a practical perspective, Wiley: Chichester, 2002

FDA (1988). Guideline for the format and content of the clinical and statistical sections of new drug applications, U.S. Food and Drug Administration, Rockville, MD

Gallo, P. (2006). Confidentiality and trial integrity issues for adaptive designs. Drug Information Journal, 40, 445-449

Gould AL. (2005). Timing of futility analyses for `proof of concept' trials. Statistics in Medicine, **24**: 1815-1835.

Gu, J, Ghosal, S, Roy, A (2008). Bayesian bootstrap estimation of ROC curve. Statistics in Medicine. Vol 27 (26), 5407-5420

Hung, J, Wang, SJ, and O'Neill, R. (2010). Consideration of regional difference in design and analysis of multi-regional trials. Pharmaceutical Statistics, 9:173–178

Huson, L (2009). The Bayesian bootstrap in a predictive power analysis. CS-BIGS 3(1): 18-22

ICH E6 (1996). Guideline for good clinical practice. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH)

Jennison C., and Turnbull, B.W. (1990). Statistical approaches to interim monitoring of medical trials: A review and commentary. Stat. Sci., 5, 209-317

Jennison, C., Turnbull, BW. (2000). Group Sequential Methods with Applications to Clinical Trials. Chapman & Hall/CRC.

Jennison, C and Turnbull, BW. (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect, Statistics in Medicine, **22**: 971-993.

Khin, Ni A., Chen, Yeh-Fong, Yang Yang, Yang, Peiling, and Laughren, Thomas P. (2012). Exploratory Analyses of Efficacy Data From Schizophrenia Trials in Support of New Drug Applications Submitted to the US Food and Drug Administration. J Clin Psychiatry 2012;73(6):856–864

Lachin, J. (2005). A review of methods for futility stopping based on conditional power. Statistics in Medicine, 24: 2747-2764

Merlo-Pich E, Bettica P, Gomeni R. (2009). Bayesian Monitoring and Bootstrap Trial Simulation: A New Paradigm to Implement Adaptive Trial Design for Testing Antidepressant Drugs. The Open Psychiatry Journal, 3, 20-32

Pong, A. and Chow, S-C (2010). Handbook of adaptive designs in pharmaceutical and clinical development. CRC Press: Taylor & Francis

Rubin, D (1981). The Bayesian bootstrap. The Annals of Statistics. Vol 9 (1), 130-134

Zhao, J and Pong, A. (2011). Considerations on Decision Rules for Futility Based on Multiple Studies. In JSM 2011 Proceedings, Biometrics Section. Alexandria, VA: American Statistical Association, pp. 3147-3156.