An Application of Adaptive Enrichment Design to CRT Trial Data

Yonghong Gao

Center for Devices and Radiological Health, Food and Drug Administration,
10903 New Hampshire Ave., Silver Spring, MD, 20993

Abstract

Rosenblum and Van der Laan (2011) proposed a new adaptive enrichment design for a two-stage randomized trial where the enrollment decision at the second stage and the hypothesis testing at the final are based on the first stage data. The proposed methodology provides a strong control of the family-wise type I error rate under a wide range of interim decision rules. In this paper we apply this new methodology to a CRT trial data under two interim decision rules. We also look at the effect of the decision rules to the overall power of this adaptive enrichment design through simulation studies. We conclude that decision rule at interim analysis is critical, and clinical input as well as statistical considerations should be taken into account when utilizing this new adaptive enrichment method for the study.

Key Words: adaptive, decision rule, enrichment, interim analysis, trial, data

1. Introduction

In many clinical trials experimental treatment is showed to have different treatment effect for different subpopulations, such as female subpopulation or diabetic subpopulation. For example, one study of a ventricular assist device (VAD) showed that female subjects treated with the investigational VAD were found to have much higher rate of stroke compared to male subjects. The heterogeneity among different stratum of subpopulation may be desirable in that the results of the trial can be generalizable to a wide class of patients. Sometimes, especially in regulatory setting, the main interest is to find any subgroup for which experimental treatment works. Under the traditional fixed sample size trial, the sample size is usually calculated to provide just enough power to reject null hypothesis of no treatment effect on the overall population, and therefore the trial is underpowered to detect possible treatment effect on some of the subpopulations. In regulatory setting, when trial data fail to provide significant evidence to reject the null hypothesis for the overall population, and the subgroup analysis indicates clinically, but not statistically, significant treatment effect on some of the subpopulation, the usual approach trial sponsor takes is to design a new trial that is specifically targeting at the promising subpopulation so that the new trial is powered to reject the null hypothesis for that subpopulation. Essentially the first trial serves as an exploratory study for the second trial that is tailed to the specific subpopulation. This two-trial approach is not cost-efficient and can be very time consuming for the industry to seek regulatory approval for their product. Adaptive design has been proposed in the literature to streamline trials to hopefully increase the chance of getting a success trial, in addition to shorten the trial time and decrease the cost. For example, Follmann (1997) presented a large class of

useful enrichment designs to adaptively change subgroup proportions in trials. See Rosenblum and Van der Laan (2011) for a comprehensive review of adaptive methodologies in the literature. In this paper, we focus only on the adaptive strategy proposed by Rosenblum and Van der Laan (2011), and apply this method to a clinical trial data set.

2. A Two-stage Adaptive Enrichment Design

Rosenblum and Van der Laan (2011) proposed a general method for constructing two-stage randomized enrichment designs that allow changes to the population enrolled based on interim data using a pre-specified decision rule. The aim of these designs is to improve overall power and better determine subpopulation-specific treatment effects, while the asymptotic, family-wise type I error rate is strongly controlled at a specified level α. Here is a simple description of the settings of the general clinical studies Rosenblum and Van der Laan (2011) focused on. A typical two-arm randomized clinical trial is considered where participants are enrolled and randomly assigned to one of the treatment arms: test arm treated with experimental treatment or therapy and control arm receiving some control therapy. The overall target population consists of two non-overlapping subpopulations, subpopulation 1 and subpopulation 2. For example, subpopulation 1 consists of all males and subpopulation 2 of all females. Under this setting, there are three questions the trial data can help answer: the experimental treatment works better than the control on overall population, experiment treatment only works on subpopulation 1 or only works on subpopulation 2.

The following is a brief description of Rosenblum and Van der Laan's (2011) general adaptive methodology tailed to the clinical trial setting mentioned above. Suppose the primary endpoint for the trial is some variable, denoted by X, with higher value of X being more desirable, and the treatment effect is assessed through looking at the mean difference of variable X between the two comparing treatments. Let H_{01} denote the null hypothesis for subpopulation 1, that the mean under the new treatment is less than or equal to the mean under the control. In an analogous manner, define the null hypothesis H_{02} corresponding to subpopulation 2, and the null hypothesis H_{0a} corresponding to the total population. The trial is divided into two stages, and the total number of subjects to be enrolled in stage 1 and stage 2 is pre-specified and can not be changed. At the end of stage 1, three test statistics, $T_a^{(1)}$, $T_1^{(1)}$, and $T_2^{(1)}$, corresponding to the standardized difference in mean values of X between treatment and control arms for subjects in the total population, in subpopulation 1 and in subpopulation 2, respectively; are calculated. According to a pre-specified decision rule, enrollment plan for stage 2 subjects can be decided from the two possible choices: (i) continue enrolling from both subpopulations in the same way as in stage 1 or (ii) enroll subjects only from the subpopulation $s \in \{1, 2\}$ corresponding to the larger of the stage 1 test statistics, $T_1^{(1)}$ and $T_2^{(1)}$. A test statistic $T_i^{(2)}$, based on stage 2 data only, can be calculated in a similar way when stage 2 data are available, where $i \in \{a, 1, 2\}$, depending on the actual enrollment plan for stage 2. At the end of the trial, a final test statistic T is computed leading to a possible rejection of one of the three null hypotheses $\{H_{0a}, H_{01}, H_{02}\}$. This final test statistic T is calculated in a conventional way, as it is a weighted combination of the test statistic $T_a^{(1)}$, which pools all stage 1 data from both subpopulations, and the test statistic $T_i^{(2)}$, which pools all stage 2 data. The corresponding weights are the proportions of enrolled subjects in each stage. If the final test statistic T exceeds a threshold c, which turns out to be the usual critical

value for fixed trial design, null hypothesis corresponding to the subpopulation, or the total population, selected for enrollment in stage 2, will be rejected.

There are some interesting features of this adaptive enrichment procedure that make it different from other designs in the literature. The first feature is that all stage 1 data are used in the final testing regardless the null hypothesis that end up being tested in the final. For example, enrollment decision after the interim data analysis is that only patients from subpopulation 1 would be enrolled for stage 2, but data from subpopulation 2 in stage 1 would still be used in the final test statistic, even when the final hypothesis tested concerns only subpopulation1. The second feature of this design is that the final decision is "random", here "random" means that we do not know, before we see the trial data, which null hypothesis, among the three possible null hypotheses, will be tested at the end of the trial. The third feature of this adaptive design is that the final test statistic is calculated in a conventional way and the testing of the hypothesis is conducted in a conventional way also, even when the trial is modified midway. Usually some type of penalty, either by combining data from two stages in an unconventional way or by raising the threshold in testing procedure or, is needed to control the type I error rate when some aspect of the trial design is adapted. It seems Rosenblum and Van der Laan's design requires no penalty to offset the gain of the power through adaptation. Rosenblum and Van der Laan (2011) provided some intuitive explanation for this. They argued that the aforementioned first feature of the design is actually the way this design pays the penalty when enrollment plan is changed, and their theoretical proof showed that the asymptotic, worst-case, family-wise Type I error rate for a wide range of enrichment designs is strongly controlled for their proposed trial design.

3. MADIT- CRT Trial

The MADIT-CRT trial is Boston Scientific's Multicenter Automatic Defibrillator Implantation Trial – Cardiac Resynchronization Therapy study. The goal of this randomized study is to determine whether Cardiac Resynchronization Therapy Defibrillators (CRT-D) in high-risk heart failure (HF) patients will reduce the combined endpoint of all cause mortality or HF intervention when compared to implantable cardioverter defibrillator (ICD) therapy. CRT-D provides two functions. As an implantable cardioverter defibrillator (ICD) it senses dangerous abnormal heart rhythms and then attempts to shock the heart back into a normal rhythm. As cardiac resynchronization therapy, it generates small electrical impulses to coordinate the beating of the left and right ventricles so that they work together more effectively to pump blood throughout the body. The MADIT-CRT trial enrolled a total of 1820 patients from 110 centers in 14 countries. Among them 1089 were randomized into CRT-D arm and 731 in ICD arm. The primary endpoint is all-cause mortality or heart failure intervention, whichever occurs first. The following table showed the data from the overall population.

Table 1: Data from Overall Population

	Test Arm	Control Arm	Hazard Ratio (HR), 95% CI for HR
Subject number	1089	731	HR=0.62
Event number	208	208	
Event Rate	19.1%	28.4%	95% CI of HR: (0.50, 0.75)

The above data demonstrated that early CRT intervention reduces the relative risk of all-cause mortality or first heart failure event when compared to ICD therapy. During the review of data for the pre-market approval (PMA) application, subgroup analysis for a wide range of different subgroups were conducted and a significant interaction between treatment and bundle branch block morphology was detected. Left Bundle Branch Block (LBBB) is a marker of an electrical conduction disorder in the heart and has been associated with a greater benefit in patients receiving CRT, further analyses revealed that LBBB is an objective discriminator of patient benefit from CRT-D (primary endpoint) regardless of other baseline characteristics. For MADIT-CRT trial, there were 1281 and 539 patients in LBBB subpopulation and no-LBBB subpopulation. The following two tables displayed the primary endpoint results for the two non-overlapping subpopulations.

Table 2: Data from LBBB subpopulation

	Test Arm	Control Arm	Hazard Ratio (HR), 95% CI for HR
Subject number	761	520	HR=0.43
Event number	120	162	
Event Rate	15.8%	31.1%	95% CI of HR: (0.33, 0.56)

Table 3: Data from no-LBBB subpopulation

	Test Arm	Control Arm	Hazard Ratio (HR), 95% CI for HR
Subject number	328	209	HR=1.32
Event number	81	41	
Event Rate	24.6%	19.6%	95% CI of HR: (0.85, 2.04)

The MADIT-CRT data indicated a quantitative interaction between treatment and the LBBB subgroup: the LBBB subpopulation benefits greatly from CRT-D, but not no-LBBB subpopulation. And it seemed that the observed statistically significant treatment effect on the overall population is largely driven by LBBB subpopulation which constituted 70% of the enrolled patients in the trial. Because of this finding, Boston Scientific's CRT-D indication is limited to sub-population of MADIT-CRT patients with left bundle branch block morphology. An interesting intellectual exercise is to apply the enrichment trial design proposed by Rosenblum and Van der Laan (2011) to the MADIT-CRT data and see whether the new enrichment design would lead to something different.

4. Application of the Enrichment Design to MADIT- CRT Trial Data

The application of the new enrichment trial design requires specification of some design parameters. Sample sizes for stage 1 and stage 2 are specified as 910, which is half of the total enrollment in MADIT-CRT trial. Proportion of LBBB subpopulation is specified as 70%, which is close to the observed proportion in MADIT-CRT trial. Interim decision rules need to be pre-specified before data available. In this paper, we look at two decision rules at the interim analysis:

- Decision rule (1): enroll from overall population if $T_a^{(1)} > 1.5$; else enroll only the subpopulation corresponding to larger $T_i^{(1)}$.

- Decision rule (2): If $T_1^{(1)}$ and $T_2^{(1)}$ have different signs, then two subpopulations are not poolable, enroll only the subpopulation corresponding to positive $T_i^{(1)}$; else enroll overall population.

The advantage of decision rule (1) is it provides more chance of testing null hypothesis for the overall population. Medical product industry that seeks regulatory approval for their product may prefer this decision rule.

Decision rule (2) emphasizes the poolability of the two subpopulations. If prior data indicated possible heterogeneity of the two subpopulations, trial designs utilizing decision rule (2) can provide higher power to reject the null hypothesis for one subpopulation. This can be an advantage and a disadvantage for the medical product company, since the product can have a higher chance to be approved for one subpopulation and at the same time the other subpopulation would be excluded from the indication.

In this paper, the MADIT-CRT data were randomly split into half and half, with the first half serving as stage 1 data. The following three tables showed the primary endpoint results for overall population, LBBB subpopulation and no-LBBB subpopulation based on the stage 1 data.

Table 4: Stage 1 Data from Overall Population

	Test Arm	Control Arm	Test Statistic
Subject number	542	368	$T_a^{(1)}$ =1.8638
Event number	94	83	
Event Rate	17.3%	22.6%	

Table 5: Stage 1 Data from LBBB Subpopulation

	Test Arm	Control Arm	Test Statistic
Subject number	378	265	$T_1^{(1)}$ =2.932
Event number	52	61	
Event Rate	17.3%	22.6%	

Table 6: Stage 1 Data from no-LBBB Subpopulation

	Test Arm	Control Arm	Test Statistic
Subject number	164	102	$T_2^{(1)}$ = - 0.6022
Event number	42	22	
Event Rate	25.6%	21.6%	

a) Under decision rule (1), the enrollment plan for stage 2 is to enroll 910 patients from the overall population as the test statistic for overall population, $T_a^{(1)}$ =1.8638, is larger than the cut-off value of 1.5. So there is no change in trial design and the trial is essentially a group sequential design. The following table presented the primary endpoint results for stage 2 data when enrolled from the overall population.

Table 7: Stage 2 Data from Overall Population

	Test Arm	Control Arm	Test Statistic
Subject number	547	363	$T_a^{(2)}$=4.1138
Event number	94	105	
Event Rate	17.1%	28.9%	

Data from two stages were combined using weighted average of the tests $T_a^{(1)}$ and $T_a^{(2)}$:

$$T = \sqrt{\frac{n_1}{n_1+n_2}}T_a^{(1)} + \sqrt{\frac{n_2}{n_1+n_2}}T_a^{(2)}$$
$$= \sqrt{.5}(1.8638 + 4.1138) = 4.2268$$

The final test statistic T is larger than the conventional critical value of 1.96, therefore we can reject H_{0a} at the 5% significance level and concluded that CRT-D provided more benefit than the control ICD for the overall population.

b) Under decision rule (2), we looked at the signs of the two statistics, $T_1^{(1)}$ and $T_2^{(1)}$ corresponding to two subpopulations. Table 5 and 6 indicated the two statistics, $T_1^{(1)}$=2.932 and $T_2^{(1)}$ = -0.6022, having different signs, so the enrollment plan for stage 2 is to enroll 910 patients from the LBBB subpopulation only as the test statistic for this subpopulation $T_1^{(1)}$ is positive. This is an enrichment design where the LBBB subpopulation, which showed greater treatment benefit of the CRT-D through the interim data, is enriched by stage 2 data. For the MADIT-CRT trial data, there were only 638 LBBB patients left in the remaining stage 2 data, but the enrollment plan requires enrolling a total of 910 LBBB patients. One possible solution is to bootstrap 910 subjects from the 638 LBBB patients. The following table gave the primary endpoint results for stage 2 data from LBBB subpopulation using bootstrap.

Table 8: Stage 2 Data from LBBB Subpopulation after Bootstrap

	Test Arm	Control Arm	Test Statistic
Subject number	542	368	$T_1^{(2)}$ =6.7364
Event number	79	117	
Event Rate	14.2%	33.2%	

To calculate the final test statistic T, we combine data from two stages using weighted average of the tests $T_a^{(1)}$ and $T_1^{(2)}$:

$$T = \sqrt{.5}(1.8638 + 6.7364) = 6.081$$

The final test statistic T is larger than the conventional critical value of 1.96, therefore we can reject H_{01} at the 5% level and concluded that CRT-D provided better benefit than ICD for LBBB subpopulation. Note that the final conclusion concerns LBBB subpopulation only, but data from no-LBBB subpopulation are used in reaching that conclusion through stage 1 test statistic $T_a^{(1)}$. From table 6 we can see that no-LBBB subpopulation in stage 1 contributed negatively to the overall stage 1 data $T_a^{(1)}$, therefore including this unfavorable data in the final testing is the built-in statistical correction (or penalty) to control the type I error rate after adaption.

5. Discussion

From section 4 we see that different conclusions are reached for the same MADIT-CRT trial data when the trial is designed under different decision paths. It shows the critical role of the interim decision rule for this adaptive enrichment trial design. As mentioned in section 4, medical product sponsor may prefer decision rule (1) so their product can get approval in a wider patient population. However, regulatory agency may prefer decision rule (2) as that decision rule provides more chance of protecting some subpopulation from exposing to the potentially risky treatment, if early data provided some evidence of no treatment benefit for this subpopulation. In the MADIT-CRT trial example of section 4, we see that if decision rule (2) is utilized in the trial, then no no-LBBB subpopulation patients would be enrolled in stage 2 and therefore the 272 (=910-638) no-LBBB patients could have avoided exposing themselves to the potentially risk treatment of CRT-D. At the trial design phase, if this enrichment adaptive methodology is being considered for the study, trial sponsor and regulatory agency should work closely together to decide and agree upon the right interim decision rule. The overall trial goal and the clinical relevance should be considered in working on the decision rule for the interim analysis.

Rosenblum and Van der Laan (2011) showed in their paper that the family-wise type I error rate is strongly controlled under the proposed class of decision rules. The two decision rules considered in section 4 are members of that class of decision rules, the critical value of 1.5 used in decision rule (1) is arbitrarily chosen, and the criteria of poolability of the two subpopulations used in decision rule (2) is quite arbitrary also. A natural question to ask is does the critical value in those decision rules have any impact on the operating characteristic of the trial design? We conducted two simulation studies to help answer this question.

Simulation study 1: the impact of critical value of c1 under decision rule (1) on the overall power of the enrichment adaptive design. In this simulation study, we focus on decision rule (1): enroll from overall population if $T_a^{(1)} > c1$; else enroll only the subpopulation corresponding to larger $T_i^{(1)}$. We let c1 take values in the range of (0.2, 1.8) in the simulation. A total of 920 subjects are simulated from some binomial distributions, among them, 70% are from subpopulation 1 and 30% are from subpopulation 2. The success rates for test arm of subpopulation 1, control arm of subpopulation 1, test arm of subpopulation 2 and control arm of subpopulation 2 are 0.86, .72, 0.75, and .80, respectively. Half of them are used as stage 1 subjects. The trial is simulated 100,000 times and the overall power (rejecting any null hypothesis) under

decision rule (1) is calculated. The following graph showed the plot of power vs. c1.

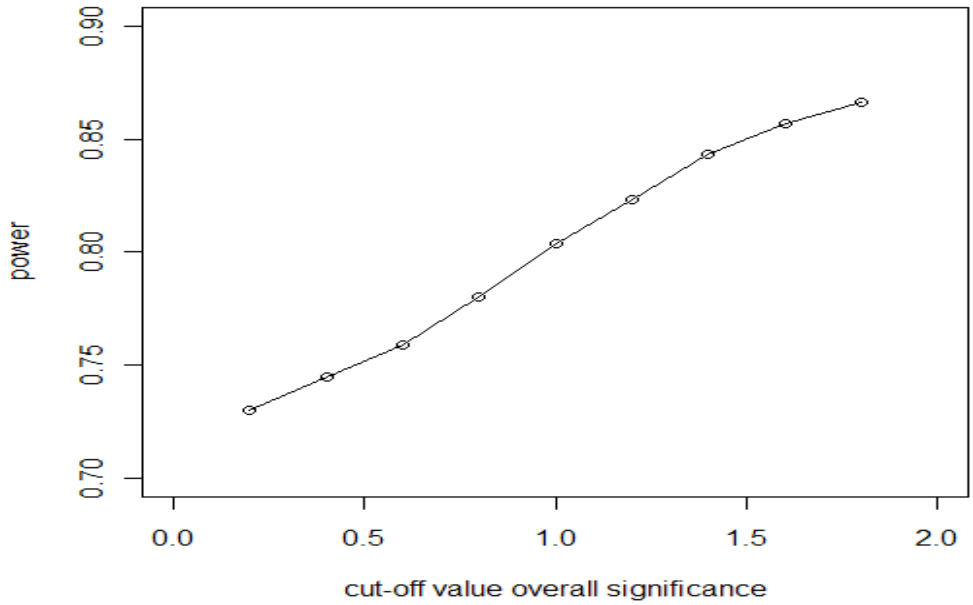

Figure 1: power as a function of cut-off value c1

It can be seen from Figure 1 that the overall power is an increasing function of cut-off value c1. Under decision rule (1), larger cut-off value c1 means higher standard of significance for overall population, which leads to smaller chance of enrolling overall population, and then more chance to enroll better subpopulation, and therefore leads to higher overall power.

Simulation study 2: the impact of poolablility criteria c2 under decision rule (2) on the overall power of the enrichment adaptive design. In this simulation study, we focus on revised decision rule (2): enroll from overall population if $|T_1^{(1)} - T_2^{(2)}| < c2$; else enroll only the subpopulation corresponding to larger $T_i^{(1)}$. We let c2 take values in the range of (0.2, 1.8) in the simulation. The simulation parameters used in simulation study 2 are quite similar to simulation study 1: a total of 920 subjects are simulated from some binomial distributions, among them, 70% are from subpopulation 1 and 30% are from subpopulation 2. Half of them are used as stage 1 subjects. The trial is simulated 100,000 times and the overall power (rejecting any null hypothesis) under decision rule (2) is calculated. We tried a range of success rates for the four strata: test arm of subpopulation 1, control arm of subpopulation 1, test arm of subpopulation 2 and control arm of subpopulation 2. Note that decision rule (2) provides higher power when there exists heterogeneity between the two subpopulations. Here we presented the simulation results when the two subpopulations are quite different. The following graph showed the plot of power vs. c2.

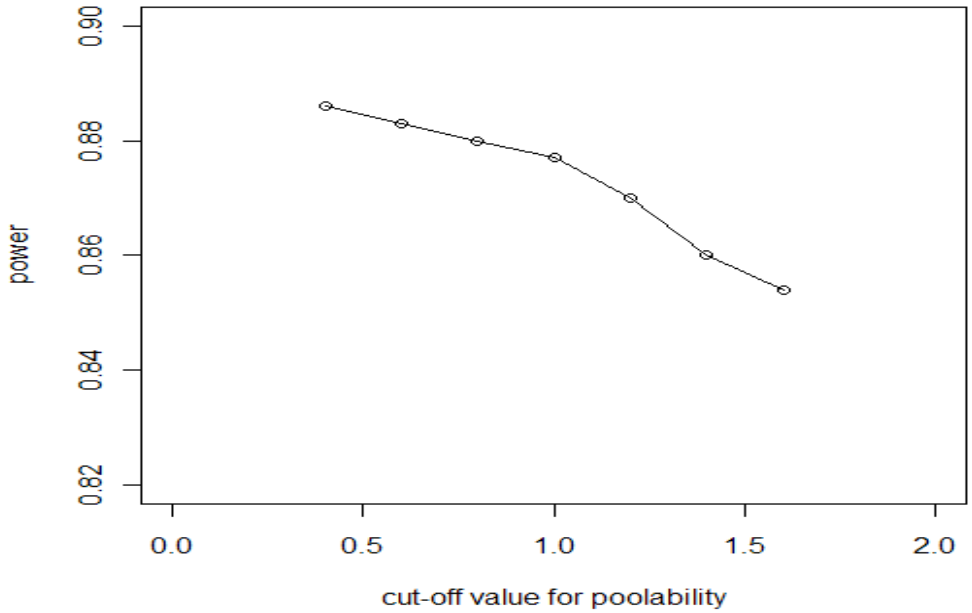

Figure2: power as a function of poolability criteria c2

It can be seen from Figure 2 that the overall power is a decreasing function of cut-off value c2. Under revised decision rule (2), smaller cut-off value c2 means higher standard for poolability, which leads to smaller chance of enrolling overall population, and then more chance to enroll better subpopulation, and therefore leads to higher overall power.

The two simulation studies further supported our position that the interim decision rule is really important for utilizing this adaptive enrichment design. Careful planning is essential in adopting this methodology to a clinical study.

References

1. M. Rosenblum and M. Van Der Lann, "Optimizing Randomized Trial Designs to Distinguish which subpopulations benefit from treatment", *Biometrika* (2011), **98**, 4, *pp.* 845–860
2. Follmann, "Adaptively changing subgroup proportions in clinical trials", statistica sinica, 1997
3. Thall, Simon and Ellenberg, "Two-stage selection and testing designs for comparative clinical trials", Biometrika, 1988
4. Kieser, Bauer, and Lehmacher "inference on multiple endpoints in clinical trials with adaptive interim analyses", Biometrical Journal, 1999
5. http://www.bostonscientific.com/cardiac-rhythm-resources/clinical/madit-crt-trial.htmlMADIT trial