German Medical Data Sciences: Bringing Data to Life
R. Röhrig et al. (Eds.)

© 2021 The authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI210043

Choosing Interim Sample Sizes in Group Sequential Designs

Sergey TARIMA[a] and Nancy FLOURNOY[b,1]
[a] *Division of Biostatistics, Medical College of Wisconsin*
starima@mcw.edu
[b] *Department of Statistics, University of Missouri-Columbia*
flournoyn@missouri.edu

Abstract. This manuscript investigates sample sizes for interim analyses in group sequential designs. Traditional group sequential designs (GSD) rely on "information fraction" arguments to define the interim sample sizes. Then, interim maximum likelihood estimators (MLEs) are used to decide whether to stop early or continue the data collection until the next interim analysis. The possibility of early stopping changes the distribution of interim and final MLEs: possible interim decisions on trial stopping excludes some sample space elements. At each interim analysis the distribution of an interim MLE is a mixture of truncated and untruncated distributions. The distributional form of an MLE becomes more and more complicated with each additional interim analysis. Test statistics that are asymptotically normal without a possibly of early stopping, become mixtures of truncated normal distributions under local alternatives. Stage-specific information ratios are equivalent to sample size ratios for independent and identically distributed data. This equivalence is used to justify interim sample sizes in GSDs. Because stage-specific information ratios derived from normally distributed data differ from those derived from non-normally distributed data, the former equivalence is invalid when there is a possibility of early stopping. Tarima and Flournoy [3] have proposed a new GSD where interim sample sizes are determined by a pre-defined sequence of ordered alternative hypotheses, and the calculation of information fractions is not needed. This innovation allows researchers to prescribe interim analyses based on desired power properties. This work compares interim power properties of a classical one-sided three stage Pocock design with a one-sided three stage design driven by three ordered alternatives.

Keywords. adaptive clinical trials, large sample properties, interim analyses, hypotheses testing, statistical distributions, likelihood functions

1. Introduction

1.1. Background

Influential books on group sequential methods, [1] and [2] rely on the asymptotic normality of test statistics to develop and justify group sequential designs (GSD). In Section 3.1 of [1], authors introduce the joint canonical distribution assumption, which if true, essentially implies that the central limit theorem (CLT) is applicable not just to

[1] Nancy Flournoy, 600 S State St., #408, Bellingham, WA 98225, United States of America; E-mail: flournoyn@missouri.edu.

test statistics calculated using independent data, but also to separate stage-specific data and data collected using non-ancillary interim stopping rules. This assumption allows authors to assume test statistics are approximately normal. Alternatively, the probability model of Brownian motion also implies the joint asymptotic normality of stage-specific test statistics [2]. The possibility of early stopping is informative which makes originally normal test statistics non-normal. This change of distributions changes the information. As shown on page 174-175 of [2] and in [3], the MLE in the presence of possible early stopping does not change, but the distributional form of the test statistic is not normal anymore. Moreover, the impact of early stopping is more profound in that non-normality holds asymptotically [3-5], as convergence to a stationary distribution continues to exist. Asymptotic non-normality has repeatedly been found before in other adaptive designs [6-10].

The equivalence between interim information ratios and interim sample size ratios for independent and identically distributed data is used to determine interim sample sizes in GSDs. But interim information ratios derived from from non-normally distributed data differ from those derived from normal data. The possibility of early stopping makes the normality assumption invalid. In this manuscript, GSDs relying on pre-determined fractional sample sizes are referred to as GSD-FSS.

Because, as previously suggested, the theoretical justification used to choose sample sizes for interim analyses is not valid, a new approach was suggested in [3] in which interim sample sizes are determined by a sequence of ordered alternative hypotheses (GSD-SOA). Section 2 introduces a three-stage Pocock GSD. Section 3 describes a GSD-SOA and develops two GSD-SOA designs based on different α-spending functions. Section 4 compares stopping probabilities of the designs via Monte-Carlo simulations. Finally, Section 5 concludes the manuscript with a short discussion.

2. A one-sided three-stage Pocock group sequential designs

Group sequential designs have been implemented in various statistical software including but not limited to the SEQDESIGN procedure in SAS, an R package "gsdesign" and Cytel's EAST software. All these programming products rely on the same theory and use information fractions to determine sample sizes of interim analyses. Then, to evaluate power properties their software relies Armitage's formula [11] which is a recursive sub-density formula that incorporates the possibility of early stopping at interim analyses. Armitage's algorithm correctly calculates overall statistical power under the alternative hypothesis. Thus, the software provides correct power calculations despite calculating "information fractions" from normal densities.

Consider a simple one arm study where a new treatment needs to be tested against a historically established level. Thus, the null hypothesis that the mean difference from historical controls, $\theta=0$, needs to be tested. The alternative hypothesis is defined on a standardized scale (mean divided by a standard deviation): $\theta=0.1$. The use a standardized scale (effect size) eliminates a need to estimate nuisance parameter (standard deviation) from the design problem. To design a three-stage clinical trial with possibility of early efficacy stopping one first chooses an α-spending function. Pocock's α-spending function is a poular choice determined by having the same critical values at all interim analyses. SAS SEQDESIGN syntax to design such as study is

```
proc seqdesign altref=0.1 pss stopprob errspend;
   OneSidedPocock: design nstages=3 alt=upper
     method=poc BETA=0.2 ALPHA=0.05 STOP=REJECT;
   samplesize model=onesamplemean(stddev=1);
```

The following R code using the “gsdesign” package leads to identical sample sizes and critical values:

```
gsDesign(k=3,test.type=1,sfu="Pocock",n.fix=NULL,
   alpha=0.05,beta=0.2,delta=0.1)
```

Output from this SAS procedure states that the first interim analysis should be performed after $n_{(1)}$=244 patients, the second is at $n_{(2)}$=488 if did not stop at stage one, and the third and final analysis is done at $n_{(3)}$=732 if the study did not stop before. These sample sizes are justified by information fractions 0.3333 at stage k=1, 0.6667 at k=2, and 1.0000 at k=3. At each interim analysis, the test statistic (sample mean multiplied by a square root of the sample size and divided by a sample standard deviation) is compared against the efficacy critical value 1.9922 (c_1=c_2=c_3): if above, the study is stopped for efficacy, if below, the study continues with additional data collection until next interim or final analysis. The α-spending function is defined by cumulative stopping probabilities 0.0232 at stage k=1, 0.0387 at k=2, and 0.0500 at k=3, under the null. More generally, the SAS output reports the following operational characteristics:

```
                ----Stopping Probabilities----

        CRef     Stage_1     Stage_2     Stage_3

      0.0000    0.02318     0.03866     0.05000
      0.5000    0.11289     0.22628     0.32918
      1.0000    0.33343     0.62094     0.80000
      1.5000    0.63698     0.91735     0.98455


                  ----Sample Size Summary----

   Test                                   One-Sample Mean
   Mean                                               0.1
   Standard Deviation                                   1
   Max Sample Size                               731.7011
   Expected Sample Size (Null Ref)               716.6188
   Expected Sample Size (Alt Ref)                498.9308
```

Note that this design is not driven by these stopping probabilities, but is determined by a chosen α-spending function and multiple fractional sample sizes. Monte-Carlo simulation results for this Pocock design are reported in Table 2.

In the next section, a pre-determined α-spending function and a sequence of ordered alternatives to be detected with the predetermined stopping probabilities are used to determine interim sample sizes and stage-specific critical values.

3. Group Sequential Design with Interim Sample Sizes Defined by Ordered Alternatives

In [4], the sample sizes of interim analyses were chosen to have desired power determined by several ordered alternatives. For a one-sided three-stage design considered in Section 2, they suggested choosing interim sample sizes to secure 80% statistical power at all three alternatives: θ=0.3, θ=0.2, and θ=0.1 regardless of when stopping occurs. They relied on equal stage-specific rejection probabilities (0.0172) under the null hypothesis. Their chosen sample sizes for interim analyses were $n_{(1)}$=98, $n_{(2)}$=196, and $n_{(3)}$=772; and stage specific critical values were c_1=2.12, c_2=2.02, and c_3=2.02. Rejection probabilities and expected sample sizes under various alternatives are reported in Table 1.

Note, Table 1 relied on an equal probability of rejecting the null hypothesis at each of three stages stage if θ=0. Let α_k denote the stage-specific rejection probability, that is, the probability of rejecting the null hypothesis at analysis k given the study did not stop at stage *k-1*. Then if α_1=.0172 and α_2= 0.0172, the probability to reject by or at stage 2 = 0.0172+(1-0.0172) 0.0172=0.0341. Similarly, if α_3= 0.0172, then the overall type I error is 0.0172+(1-0.0172)0.0172+(1-0.0172) (1-0.0172)*0.0172=0.0507. Due to rounding, type I error is not exactly 5%, but it is close enough for illustrative purposes. These results are consistent with Monte-Carlo simulations reported in Table 1 under θ=0. This, however, highlights the fact that Pocock's design does not have equal rejection probabilities at each stage: uniform critical values do not translate into equal rejection probabilities.

To make GSD-SOA comparable with the α-spending function used in Pocock's GSD-FSS (Table 2), we need to build a GSD-SOA design with the same α-spending function. It is easy to show that the α-spending function defined by cumulative rejection probabilities (0.0232, 0.0387, and 0.0500) is associated with stage-specific rejection probabilities α_1= 0.0232, α_2= 0.0158, and α_3 = 0.0118. This new α-spending function leads to a new SOA design with $n_{(1)}$=90, $n_{(2)}$=205, and $n_{(3)}$=863 with stage-specific critical values c_1= 1.9921, c_2=2.0216, and c_3=2.1812. Monte-Carlo simulations are reported in Table 3.

4. Monte-Carlo Simulation Experiments

Each Monte-Carlo simulation study in this section relied on 100,000 random sequences of standard normal random variables.

Table 1: GSD-SOA's cumulative rejection probabilities by stage and expected sample sizes using with equal probabilities of stage-specific stopping: α_1= α_2=α_3= 0.0172.

θ	Pr(Reject at k=1)	Pr(Reject at k≤2)	Pr(Reject at k≤3)	E(N)
0.0	0.0179	0.0337	**0.0509**	750.83
0.1	0.1320	0.3019	**0.7990**	585.18
0.2	0.4473	**0.7985**	0.9998	268.24
0.3	**0.8007**	0.9868	1.0000	125.17

Table 2: GSD-FSS's cumulative rejection probabilities by stage and expected sample sizes using with Pocock's α-spending function: α_1= 0.0232, α_2= 0.0158, and α_3 = 0.0118.

θ	**Pr(Reject at *k*=1)**	**Pr(Reject at *k*≤2)**	**Pr(Reject at *k*≤3)**	**E(*N*)**
0.0	0.0248	0.0407	**0.0515**	716.02
0.1	0.3351	0.6219	**0.8004**	498.49
0.2	0.8704	**0.9930**	0.9997	277.35
0.3	**0.9965**	1.0000	1.0000	244.87

Table 3: GSD-SOA's cumulative rejection probabilities by stage and expected sample sizes using with Pocock's α-spending function: α_1= 0.0232, α_2= 0.0158, and α_3 = 0.0118.

θ	**Pr(Reject at *k*=1)**	**Pr(Reject at *k*≤2)**	**Pr(Reject at *k*≤3)**	**E(*N*)**
0.0	0.0254	0.0409	**0.0516**	833.01
0.1	0.1513	0.3124	**0.8019**	638.40
0.2	0.4654	**0.8028**	0.9999	277.87
0.3	**0.8016**	0.9867	1.0000	119.74

5. Discussion

GSDs are predominantly defined by a triplet of (1) an α-spending function, (2) overall statistical power and (3) fractional sample sizes (FSS), whereas interim stopping probabilities are not directly controlled; but they are determined by the input triplet. The alternative illustrated in this paper is motivated by the recognition in [3] that possibility of early stopping alters finite-sample and asymptotic distribtions of test statistics; and this alteration invalidates the FSS assumption that sample size fractions are equal to information fractions calculated from normal densities. One option is to calculate information measures from the true asymptotic distributions, but this is a computationally intensive proposition and the relationship between the true information and the sample size may not be simple.

To avoid FSS as an input for GSDs, researchers can use a sequence of alternative hypotheses, each with a pre-determined stopping probability. In this paper, several Monte-Carlo simulation studies highlight these new GSD SOA designs. Examples demonstrate how to use stopping probabilities as design inputs at alternative hypotheses that are fixed for each interim test. As the clinical research community is familiar with the concept of statistical power, we anticipate that this new design will improve the clinical interpretation of design choices and facilitate the use of GSD in clinical research.

Conflict of Interest

The authors have no conflict of interests.

References

[1] C. Jennison, and B. W. Turnbull. Group Sequential Methods with Applications to Clinical Trials, *Chapman & Hall/CRC Interdisciplinary Statistics*, *CRC Pres*, New York, 1999.

[2] M. A. Proschan, K. K. G. Lan, and J. T. Wittes. Statistical Monitoring of Clinical Trials: A Unified Approach, *Springer*, 2006.

[3] S. S. Tarima and N. Flournoy. Effect of Interim Adaptations in Group Sequential Designs, https://arxiv.org/pdf/1908.01411.pdf, (2019), 1-30.

[4] S. S. Tarima and N. Flournoy. Asymptotic properties of maximum likelihood estimators with sample size recalculation. *Statistical Papers* **60** (2019), 373–394.

[5] S. S. Tarima and N. Flournoy. Distribution theory following blinded and unblinded sample size re-estimation under parametric models. *Communications in Statistics,* 2020, in press.

[6] A. Ivanova, W. F. Rosenberger, S. D. Durham and N. Flournoy. A birth and death urn for randomized clinical trials: asymptotic methods. *Sankhy, Series B* **62** 104-118, 2000.

[7] A. Ivanova and N. Flournoy A birth and death urn for ternary outcomes: stochastic processes applied to urn models. *Probability and Statistical Models with Applications* 583-600, 2001.

[8] C. May and N. Flournoy. Asymptotics in resonse-adaptive designs generated by a two-color, randomly reinforced urn. *The Annals of Statistics* **37** 1058-1078, 2009.

[9] A. Lane and N. Flournoy. Two-stage adaptive optimal design with fixed first-stage sample size. *Journal of Probability and Statistics* **2012**, 2012.

[10] N. Flournoy, C. May and C. Tommasi. The effects of adaptation on inference for non-linear regression models with normal errors. *arXiv preprint arXiv:1812.03970,* 2019.

[11] P. Armitage, C. McPherson, and B. Rowe. Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society: Series A* **132** 235–244, 1969.