CLINICAL STUDY DESIGN TO ASSESS BOTH SHORT- AND LONG-TERM EFFICACY IN ADDITION TO GROUP SEQUENTIAL TEST ON SAFETY

JIACHENG YUAN[1], PETER MESENBRINK[2], JIHAO ZHOU[1]
JEEN LIU[1], RAY ZHU[1], GARY KOCH[3]

[1] ALLERGAN INC.
[2] NOVARTIS PHARMACEUTICALS CORPORATION
[3] UNIVERSITY OF NORTH CAROLINA

INTRODUCTION

> In pharma industry, sometimes there is interest to perform a H2H comparison with an active comparator, comparing for short-term efficacy, long-term efficacy, and/or safety

> We focus on the situation with one endpoint for each of short-term efficacy (E_s) , long-term efficacy (E_l), and safety assessments (S)

> The sequentially rejective graphical procedures can be used where ordered hypotheses are tested repeatedly in time, under mild monotonicity conditions on error spending functions

> A delayed recycling method is further proposed that allocates the recycled significance level from rejected hypotheses to unrejected hypotheses from Stage r onward, where r is prespecified

> This paper proposes a graphical testing procedure that incorporates tests of short-term and long-term efficacy, as well as safety which is tested twice, once at E_s and once at E_l

INTRODUCTION (CTD)

> The study to be discussed has two periods

> The first period is from baseline to the time of the final analysis of short-term efficacy endpoint

- also being the time of the interim analysis for the safety endpoint

> The second period is from the time of the first period analysis until the time of the final analysis

- the second period analysis addresses the final analysis of the long-term efficacy endpoint and the final analysis of the safety endpoint if needed

INTRODUCTION (CTD)

> The proposed method is applicable to multiple indications.

> We will provide detailed discussion for the rheumatoid/psoriatic arthritis indication

- E_s is the American College of Rheumatology 20% improvement (ACR20) or ACR50 which is on signs and symptoms
- E_l is the modified total Sharp/van der Heijde score (SHS)
- S is major adverse cardiac events (MACE)

EVIDENCE OF EFFECTIVENESS FROM A SINGLE STUDY

> FDA had guidelines on adequacy of a single trial to support approval

> Potential suitable situations limited to a trial that has demonstrated

- clinically meaningful effect on mortality
- irreversible morbidity, or
- prevention of a disease with potentially serious outcome and
- confirmation of the result in a second trial would be practically or ethically impossible

EVIDENCE OF EFFECTIVENESS FROM A SINGLE STUDY (CTD)

> Such single adequate and well-controlled study may have one or more of following characteristics

- Large multicenter study
- Consistency across study subsets
- Multiple studies in a single study
- Multiple endpoints involving different events
- Statistically very persuasive finding

THE TESTING PROCEDURE

> Let H_s represent the null hypothesis for S

> Let H'_{es} and H_{es} be the null hypotheses of non-inferiority and superiority, respectively, for E_s

> Let H'_{el} and H_{el} be the null hypotheses of non-inferiority and superiority, respectively, for E_l

> Notation $H_e \mid H'_e$ means that for the same endpoint, represented by subscript, superiority is tested upon success of non-inferiority

> We assume E_s is tested at the same time as the interim safety analysis, E_l is tested at the final safety analysis, and S is tested twice

> Taking the same approach as Maurer and Bretz (2013), if H_s is rejected at the time of the interim safety, we consider it as fully rejected

THE TESTING PROCEDURE (CTD)

> For the interim safety assessment, we define the information fraction as $t = F / D$

- F is the number of events at the interim, and D is the number of events at the final assessment

> At the beginning of the study, a fraction κ $(0 < \kappa < 1)$ of the full alpha, $\kappa\alpha$, is allocated to the $H_{es} \mid H'_{es}$ test, and the rest $(1-\kappa)\alpha$ to the H_s test with an alpha spending function (ASF) applied

> No alpha is assigned for the $H_{el} \mid H'_{el}$ test at the beginning

> Let $A(t, \pi)$ be the alpha spent for the interim analysis of the safety endpoint

- t $(0 \le t \le 1)$ is the information fraction, and π is the total alpha that can be spent on the safety test at the interim and final analyses

> The decision boundaries can be expressed as nominal significance levels $A^*(t, \pi)$ such that H_s is rejected if the nominal p-value at time t is smaller than $A^*(t, \pi)$

> It is computed such that the overall probability of rejection up to time point t does not exceed $A(t, \pi)$

> In our setting, at the interim $A^*(t, \pi) = A^*(F / D, \pi) = A(F / D, \pi)$

THE TESTING PROCEDURE (CTD)

1) If we fail to reject H'_{es} at $\kappa\alpha$ level, the trial fails and is terminated

2) If we successfully reject H'_{es} but fail to reject H_{es} at $\kappa\alpha$ level, then the group sequential test (GST) of the safety endpoint will be performed at a significance level of $(1-\kappa)\alpha$

2a) If we fail the GST, i.e. failing to reject H_s both at the level $A(F/D,(1-\kappa)\alpha)$ at the interim and at the $A^*(1,(1-\kappa)\alpha)$ level at the final assessment, then the trial fails

2b) If the GST succeeds, $H_{el}|H'_{el}$ can be tested at the $(1-\kappa)\alpha$ level

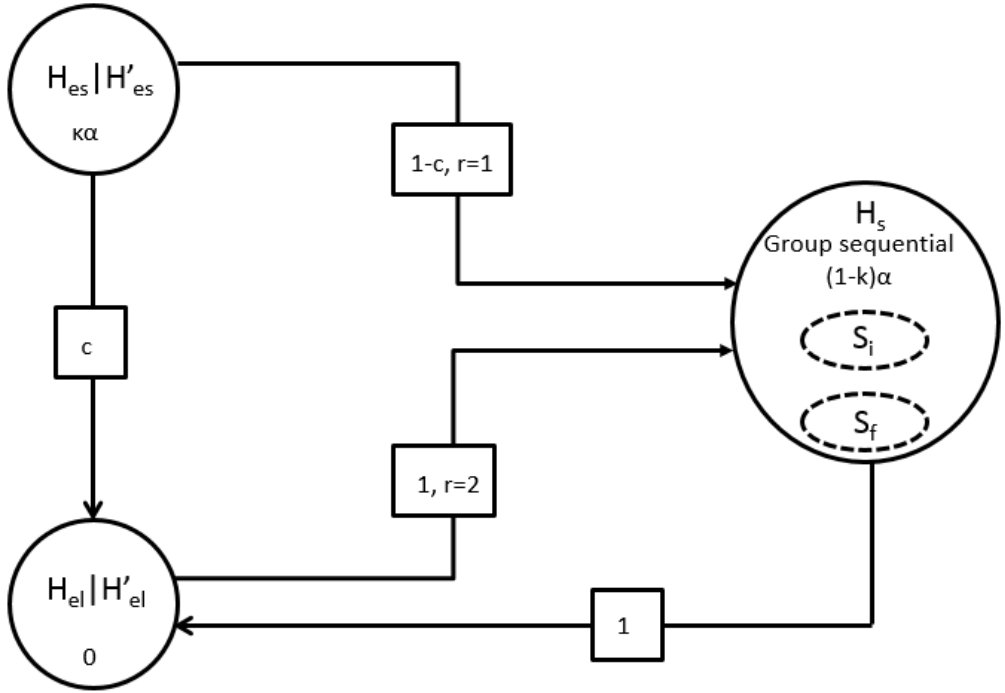
3) If both H'_{es} and H_{es} are rejected at $\kappa\alpha$ level, a portion $c\kappa\alpha$ $(0<c<1)$ will be passed to $H_{el}|H'_{el}$, and the rest $(1-c)\kappa\alpha$ will be passed to the safety tests, being spent from the beginning, i.e. $r=1$

3a) If the GST succeeds, i.e. H_s is rejected at the $A(F/D,(1-c\kappa)\alpha)$ level at the interim or at the $A^*(1,(1-c\kappa)\alpha)$ level at the final, $H_{el}|H'_{el}$ can be tested at the full α level

3b) Otherwise, $H_{el}|H'_{el}$ can only be tested at $c\kappa\alpha$ level

3c) If both H'_{el} and H_{el} are rejected at $c\kappa\alpha$ level, while H_s has yet been rejected, the alpha of $c\kappa\alpha$ can be passed to the test of H_s

FIGURE. GRAPHICAL TESTING PROCEDURE FOR SHORT- AND LONG-TERM EFFICACY AS WELL AS SAFETY WHICH IS TESTED TWICE.



CONTROL OF FAMILY-WISE TYPE I ERROR RATE

> To show that the family wise type I error rate is controlled in the suggested design, we treat the suggested graph as a special case of methods proposed elsewhere

> Maurer and Bretz (2013) extend the graphical approach by Bretz et al (2009) to group sequential designs with multiple endpoints

> They prespecify error spending functions for a range of possible levels $0 \le \gamma \le \alpha$ instead of a single significance level $\gamma = \alpha$, and then they consider $h > 1$ one-sided hypotheses, $H_i,\ i \in I = \{1, \ldots, h\}$, in a group sequential trial at k time points, $t = 1, \ldots, k$

> Each hypothesis $H_i,\ i \in I = \{1, \ldots, h\}$ is assigned a local significance level α_i such that $\sum \alpha_i = \alpha$, and define univariate testing strategies with appropriate ASFs $A(t, \alpha_i)$, separately for each of the α_i's

CONTROL FAMILY-WISE TYPE I ERROR RATE (CTD)

> Our testing strategy can be put in the Maurer and Bretz (2013) framework as follows

> The initial alpha allocation is $\kappa\alpha$, 0, and $(1-\kappa)\alpha$ for $H_{es}|H'_{es}$, $H_{el}|H'_{el}$ and H_s, respectively

> Formally, a GST is performed on each of them

- A regular ASF is applied for H_s
- A special ASF that spends all at the interim is applied for $H_{es}|H'_{es}$
- Another special ASF that spends none at the interim but all at the final is applied for $H_{el}|H'_{el}$

> After H_{es} is rejected, $c\kappa\alpha$ is propagated to $H_{el}|H'_{el}$, and $(1-c)\kappa\alpha$ is propagated to H_s

> After both H'_{el} and H_{el} are rejected, all their local alpha is propagated to H_s

> After H_s is rejected, all its local alpha is propagated to $H_{el}|H'_{el}$

> Please note, $H_{el}|H'_{el}$ does not have initial alpha

- However, after H'_{es} and H_{es} are rejected, a portion of their local alpha is propagated to $H_{el}|H'_{el}$
- and that alpha is not spent immediately but reserved to the final test, i.e. $r = 2$

EXAMPLE/APPLICATION

> In what follows, we proceed as a team that is developing a biologic similar to Novartis' secukinumab in the psoriatic arthritis (PsA) indication

- planning to conduct a H2H study versus the standard of care on the market, Abbott's adalimumab
- so that a submission can be made with a single confirmatory study

> Week 24 ACR50 is the short-term endpoint for signs and symptoms

- assessed with both non-inferiority test $H'_{es}: p_1 - p_2 \leq -0.05$, and superiority test $H_{es}: p_1 - p_2 \leq 0$
- p_1 and p_2 are Week 24 ACR50 response rate for the test and control medication, respectively

> Change of SHS from baseline to Week 52 is the long-term endpoint for structural damage

- assessed with both non-inferiority test $H'_{el}: \mu_1 - \mu_2 \geq 0.1$, and superiority test $H_{el}: \mu_1 - \mu_2 \geq 0$
- μ_1 and μ_2 are Week 52 change of SHS from baseline for the test and control medication, respectively

> The incidence of MACE is the safety endpoint

- assessed with test $H_s: \rho \geq 1.3$
- ρ is the cardiovascular risk ratio between the test and control medication

EXAMPLE/APPLICATION (CTD)

> For the test and control medication, respectively
  - we assume the expected Week 24 ACR50 response rate to be 40% and 34%
  - change from baseline of SHS to Week 52 to be 0.1 and 0.2
  - and annual MACE rate to be both 5%

> The family-wise type I error rate is controlled at one-sided 0.025 level

> At the beginning 0.005 is allocated for the ACR50 test

> the other 0.020 for the GST of MACE
  - where the Lan DeMets ASF with O'Brien Fleming boundaries is applied

> No alpha is assigned for the SHS test at the beginning

> The hypotheses testing follows the procedure as described before
  - with $\alpha = 0.025,\ \kappa = 0.2,$ and $c = 0.5$

EXAMPLE/APPLICATION (CTD)

> The sample size of this study is driven by the MACE assessment

> For the test of $H_s : \rho \geq 1.3$, the value of 1.3 is chosen according to the FDA Guidance for CV risk assessment for medical products for type 2 diabetes mellitus

> Assuming constant proportional hazards, a total of 5000 subjects $(2500$ per treatment arm$)$ enrolled uniformly over 3.3 years will result in an expected trial duration of approximately 4.5 years

> An interim analysis will be performed after 325 adjudicated events occur

- when approximately 4272 patients are enrolled

> and the final test will be performed after 649 adjudicated events occur

> Such a procedure provides at least 90% power to contradict $\rho \geq 1.3$ for the test over control medication

- true hazard ratio $\rho = 1$
- one-sided 0.020 significance level
- control MACE rate 5% annually
- lost to follow-up rate $\leq 1\%$ annually

EXAMPLE/APPLICATION (CTD)

> At interim analysis of MACE, 4272 patients will have been enrolled, which is at about 34 months in accrual period, hence patients enrolled in the first 28 months, i.e. about 3534 patients, will have been assessed for the Week 24 ACR50

> Assuming a reference rate of 34%, an actual test rate of 37%, and a non-inferiority margin of 5%, the sample size of 3534 patients (1767 per arm) yields 99% power with one-sided alpha of 0.005

> For SHS, the Mann-Whitney test is considered for the power calculation

> The study is planned to end at 54 months, and all enrollment finishes within 40 months, so all patients will have had the Week 52 (i.e. 12 months) SHS measured at the final analysis

> With sample size of 2500 per arm, there is 84% power to detect non-inferiority

- non-inferiority margin of 0.1
- true difference between means of -0.05 (test - control)
- common standard deviation of 1.4
- significance level of 0.0025

> However, there is only 39% power for the test of superiority at the alpha of 0.0025 level

> If safety endpoint passes and superiority is tested at 0.025 level instead, power will be 71%

EXAMPLE/APPLICATION (CTD)

> Let p'_{es}, p_{es}, p'_{el}, p_{el}, p_{s1}, and p_{s2} be the unadjusted p values for the tests of H'_{es}, H_{es}, H'_{el}, H_{el}, and H_s at the interim and final, respectively

> Possible sets of nominal significance levels at the interim and final safety tests are as follows

π	$A^*(0.5,\pi) = A(0.5,\pi)$	$A^*(1,\pi)$
0.0200	0.0010	0.0197
0.0225	0.0013	0.0221
0.0250	0.0015	0.0245

EXAMPLE/APPLICATION (CTD)

> Different sets of possible p values are shown below

Cases	p'_{es}	p_{es}	p'_{el}	p_{el}	p_{s1}	p_{s2}
C1	0.0004	0.0060	--	--	0.0020	0.0300
C2	0.0004	0.0060	0.0080	0.0400	0.0020	0.0150
C3	0.0004	0.0060	0.0020	0.0100	0.0020	0.0150
C4	0.0004	0.0060	0.0080	0.0400	0.0005	--
C5	0.0004	0.0060	0.0020	0.0100	0.0005	--
C6	0.0001	0.0030	0.0008	0.0040	0.0015	0.0300
C7	0.0001	0.0030	0.0008	0.0020	0.0015	0.0300
C8	0.0001	0.0030	0.0080	0.0400	0.0010	--
C9	0.0001	0.0030	0.0080	0.0200	0.0010	--
C10	0.0001	0.0030	0.0001	0.0020	0.0015	0.0230

EXAMPLE/APPLICATION (CTD)

> For C1, $p'_{es} = 0.0004 < 0.005$ and $p_{es} = 0.0060 > 0.005$, we succeed the H'_{es} test but fail the H_{es} test. The safety GST is conducted at 0.020 level. As $p_{s1} = 0.0020 > 0.0010$ and $p_{s2} = 0.0300 > 0.0197$, we also fail the GST. So the trial fails and $H_{el} \mid H'_{el}$ is not tested

> For C2, $p'_{es} = 0.0004 < 0.005$ and $p_{es} = 0.0060 > 0.005$, we succeed the H'_{es} test but fail the H_{es} test. The safety GST is conducted at 0.020 level. As $p_{s1} = 0.0020 > 0.0010$ but $p_{s2} = 0.0150 < 0.0197$, GST succeeds at the final test. So $H_{el} \mid H'_{el}$ is tested at the 0.020 level. Since $p'_{el} = 0.0080 < 0.020$ and $p_{el} = 0.0400 > 0.020$, the H'_{el} test succeeds but the H_{el} test fails

> C3 is the same as C2 until the test of H'_{el} and H_{el}. Since $p'_{el} = 0.0020 < 0.020$ and $p_{el} = 0.0100 < 0.020$, both H'_{el} and H_{el} tests succeed

> C4 is similar to C2, and C5 is similar to C3. The only difference is that the GST succeeds at the interim test, because $p_{s1} = 0.0005 < 0.0010$

EXAMPLE/APPLICATION (CTD)

> For C6, $p'_{es} = 0.0001 < 0.005$ and $p_{es} = 0.0030 < 0.005$, both H'_{es} and H_{es} tests succeed, and the $H_{el} \mid H'_{el}$ test hence receives 0.0025 alpha, and the GST can be tested at the 0.0225 level because the 0.0025 alpha from the successful H_{es} test is propagated to the GST from the start, i.e. $r = 1$. However, $p_{s1} = 0.0015 > 0.0013$ and $p_{s2} = 0.0300 > 0.0221$, the GST fails. $H_{el} \mid H'_{el}$ is tested at the 0.0025 level. Since $p'_{el} = 0.0008 < 0.0025$ and $p_{el} = 0.0040 > 0.0025$, the H'_{el} test succeeds but the H_{el} test fails

> C7 is the same with C6 until the test of H_{el}. It succeeds because $p_{el} = 0.0020 < 0.0025$

> For C8, $p'_{es} = 0.0001 < 0.005$ and $p_{es} = 0.0030 < 0.005$, both H'_{es} and H_{es} tests succeed, and the $H_{el} \mid H'_{el}$ test hence receives 0.0025 alpha, and the GST can be tested at the 0.0225 leveltest. As $p_{s1} = 0.0010 < 0.0013$, the GST succeeds. The $H_{el} \mid H'_{el}$ test receives another 0.0225 alpha and hence can be tested at the full 0.025 level. Since $p'_{el} = 0.0080 < 0.025$, H'_{el} succeeds, but the H_{el} still fails because $p_{el} = 0.0400 > 0.025$

EXAMPLE/APPLICATION

> C9 is the same with C8 until the test of H_{el}. It succeeds because $p_{el} = 0.0200 < 0.025$

> For C10, $p'_{es} = 0.0001 < 0.005$ and $p_{es} = 0.0030 < 0.005$, both H'_{es} and H_{es} tests succeed, and the $H_{el} \mid H'_{el}$ test hence receives 0.0025 alpha, and the GST can be tested at the 0.0225 leveltest. But $p_{s1} = 0.0015 > 0.0013$, the GST fails at the interim and would also fail the final if no additional alpha comes in because $p_{s2} = 0.0230 > 0.0221$. However, since $p'_{el} = 0.0001 < 0.0025$ and $p_{el} = 0.0020 < 0.0025$, both the H'_{el} and H_{el} tests succeed and their corresponding 0.0025 alpha is propagated to the second stage of the GST, i.e. $r = 2$, so the final safety test can be performed at the $0.0246 = 0.0221 + 0.0025$ level. So it is significant now

ACKNOWLEDGEMENT

- We thank Dr. Dong Xi for his review and comments on an earlier version of this paper, which have been incorporated into this version
- Discussions of a related problem in oncology studies with Mr. Jonathan Siegel and Mr. Daniel Haverstock helped forming the idea of this research

REFERENCES

- Yuan J, Mesenbrink P, Zhou J, Liu J, Zhu R, Koch G. (2018) Clinical Study Design to Assess Both Short- and Long-term Efficacy in Addition to Group Sequential Test on Safety. Therapeutic Innovation & Regulatory Science, 52: 690-695.
- Bretz F, Maurer W, Brannath W, and Posch M. (2009). A graphical approach to sequentially rejective multiple test procedures. Statistics in Medicine, 28, 586-604.
- Maurer W, Bretz F. (2013). Multiple testing in group sequential trials using graphical approaches. Statistics in Biopharmaceutical Research, 5, 311-320.
- Xi D, Tamhane AC. (2015). Allocating recycled significance levels in group sequential procedures for multiple endpoints. Biometrical Journal, 57, 90-107.
- FDA (1998). Guidance for industry: Providing Clinical Evidence and Effectiveness for Human Drug and Biologic Products.
- DeMets DL, Lan G. (1995). The alpha spending function approach to interim data analyses, in Recent Advances in Clinical Trial Design and Analysis, ed. Thrall PF, Boston, MA: Kluwer Academic Publishers, pp. 1-27.
- Ye Y, Li A, Liu L, Yao B. (2013). A group sequential Holm procedure with multiple primary endpoints. Statistics in Medicine 32, 1112–1124.

REFERENCES

- O'Brien PC, Fleming TR. (1979). A multiple testing procedure for clinical trials. Biometrics, 35, 549-556.
- FDA Guidance to Industry. (2008). Diabetes Mellitus—Evaluating Cardiovascular Risk in New Antidiabetic Therapies to Treat Type 2 Diabetes, available at http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm071627.pdf
- Munzel U, Hauschke D. (2003). A nonparametric test for proving noninferiority in clinical trials with ordered categorical data. Pharm Stat, 2: 31-37.
- Sun H, Kawaguchi A, Koch GG. (2016) Analyzing multiple endpoints in a confirmatory randomized clinical trial—an approach that addresses stratification, missing values, baseline imbalance and multiplicity for strictly ordinal outcomes, Pharmaceutical Statistics 2017;16:157-166.
- Mease PJ, McInnes IB, Kirkham B, et al. (2015) Secukinumab Inhibition of Interleukin-17A in Patients with Psoriatic Arthritis. New England Journal of Medicine, 373:1329-39.
- Gladman DD, Mease PJ, Ritchlin CT, et al. (2007) Adalimumab for long-term treatment of psoriatic arthritis: forty-eight week data from the adalimumab effectiveness in psoriatic arthritis trial. Arthritis Rheum 56, 476–88.
- van der Heijde D, Landewé RB, Mease PJ, et al. (2016) Brief report: secukinumab provides significant and sustained inhibition of joint structural damage in a phase III study of active psoriatic arthritis. Arthritis Rheumatol 68, 1914–1921.