アダプティブデザインの

統計的推測に関する検討

日本製薬工業協会

データサイエンス部会

2022 年度 継続タスクフォース 1

Ver 1.0

2023 年 2 月

目次

1　はじめに

本報告書は、医薬品の承認申請における検証的試験に適用可能なアダプティブデザインの基本的な統計的推測法に関してまとめたものである。

規制当局によるアダプティブデザインのガイダンスは、European Medicine Agency にて 2007 年に最終化[1]、Food and Drug Administration にて 2010 年にドラフトが発出され、10 年以上前よりアダプティブデザインは注目されていた。Food and Drug Administration によるガイダンスは 2019 年に最終化[2][3]され、さらに、現在 ICH-E20 にてアダプティブデザインが検討されるなど、再び議論が広がっている。FDA ガイダンス[2][3]は、アダプティブデザインを「臨床試験に参加した被験者の蓄積されたデータに基づいて、試験デザインの 1 つ以上の側面について、予め計画された変更を行うことができる臨床試験デザイン」と定義している。被験者の蓄積されたデータに基づいて修正する計画の内容は、試験がもつ様々な不確実性に応じて様々なアダプテーションが考えられる。FDA ガイダンスの V 章は、アダプテーションのタイプを以下の項目に分類して解説している。

- 群逐次デザイン
- 症例数に対するアダプテーション
- 患者集団に対するアダプテーション
- 治療群の選択に対するアダプテーション
- 患者割付に対するアダプテーション
- 評価項目の選択に対するアダプテーション

本報告書は、症例数の変更を伴うアダプティブデザイン（以下、症例数再推定と呼ぶ）と、複数の試験治療を中間解析の結果から絞り込んで検証する治療群の選択（いわゆる，シームレス第 2/3 相試験）のアダプティブデザインに注目した。これら 2 つのアダプティブデザインは治験への適用を検討する機会が比較的多く、かつ、これら 2 つのアダプティブデザインを理解しておくことで、他のアダプテーションを理解する基本となると考えた。

アダプティブデザインにより与えられる柔軟性により、試験の参加者がより良い治療を受けられる機会が増え、より効率的な医薬品開発、さらには利用可能なリソースの活用といった利点が期待される[4]。その一方、統計手法を適切に用いなければ、第一種の過誤確率の増大、点推定値へのバイアスの発生、信頼区間の被覆確率が名義上の信頼係数よりも下回るなど、統計的妥当性の観点から望ましくない現象が起こる可能性がある。ICH-E9「臨床試験のための統計的原則」[5]は、検証的試験にて第一種の過誤確率を制御すること、偏りを最小とすることをガイドラインの原則として述べている。アダプティブデザインを適用した臨床試験（以後、アダプティブ臨床試験とする）に固定標本試験で用いるナイーブな解析（2 標本 t 検定、標本平均等）を適用すると、第一種の過誤確率の増大や点推定値にバイアスが生じ、信頼区間が名義上の被覆確率を有さないことがあるため、統計

学的推測上の目標を満たすための手法が提案されてきた。中間解析の方法としてはICH-E9「臨床試験のための統計的原則」[5]にも述べられているように、群逐次デザイン[6]が広く用いられている。アダプティブデザインの統計的推測法は、1990年頃から数多く提案されている。例えば、第一種の過誤確率の増大に対応する統計手法としては、統合検定に基づく方法[6][8][9]や条件付き過誤関数に基づく方法[10]、各ステージのp値を直接統合する方法[11]などが提案されている。また、アダプティブデザインに適用できる点推定値や信頼区間の構成方法についても研究がなされている。点推定値としては中央値不偏推定量、信頼区間の構成方法としては、繰り返し信頼区間やステージに基づく順序を利用した信頼区間などがある[12][13][14][15][16][17]。治療群の選択のアダプティブデザインとしてはDunnett検定の拡張[18][19]、一様最小分散条件付き不偏推定のアプローチによる不偏推定量[20][21][22][23][24][25]と信頼区間の提案[21][26][27][28]などがある。近年においてもアダプティブデザインに関する統計的推測法の論文が発表されており、今もなお、より適切な方法が模索されている状況である。現状のすべての統計的推測法を網羅し、最適な方法を述べることは難しいため、本報告書では現時点で適用可能と考えられる統計的推測法を解説する。アダプティブ臨床試験を計画する際には、適用するアダプテーションルールの複雑さによって統計的妥当性の性質が変化するため、本報告書の内容を基本として、最新の議論からより適切な方法を調べることが望ましい。

本報告書は、以下のように構成されている。2章は、結果変数が連続変数であり、1回の中間解析において症例数再推定を実施する場合の症例数再推定の方法及び統計的推測法（検定、点推定、区間推定）に関するいくつかの方法を解説した。症例数再推定にて参照するパラメータ情報を、非盲検下と盲検下に分けて統計手法を解説した。3章は、2章と同じく連続変数と1回の中間解析を仮定し、治療群の選択のアダプティブデザインを対象としている。アダプテーションの要点である治療群を選択するルールを解説し、第一種の過誤確率の増大やバイアスの発生などが起こる状況を説明した上で、いくつかの統計的推測法を解説した。2章と3章それぞれには、アダプティブデザインの計画や解析に適用することができるRパッケージrpactを紹介した。4章にまとめを記載した。参考文献は章ごとにまとめ、章末に列記した。

本報告書の読者として、計画や実施に携わる統計解析担当者を想定しているが、一部の内容は臨床試験に携わる統計解析担当者以外のメンバーにも参考になる。非盲検下での症例数再推定の事例は2.2.2.3項、盲検下での分散に基づく症例数再推定の事例は2.3.5項、治療群の選択の事例は3.4項に示した。また、アダプティブ臨床試験の計画が第一種の過誤確率やバイアスに影響すること、それぞれに対処する方法が提案されていることは、統計解析担当者以外のメンバーにも是非理解を頂きたい。これらの解説は、非盲検下での症例数再推定について2.2.1項、2.2.3項、2.2.4項に、盲検下での分散に基づく症例数再推定については2.3.3項と2.3.4項に、治療群の選択については3.3項と3.5項に示した。

本報告書が、アダプティブデザインの理解と適切な試験実施の一助となれば幸いである。

参考文献

[1] European Medicines Agency. Reflection Paper on Methodological Issues in Confirmatory Clinical Trials Planned with an Adaptive Design (CHMP/EWP/2459/02) 2007. https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-methodological-issues-confirmatory-clinical-trials-planned-adaptive-design_en.pdf［2022 年 6 月 22 日アクセス］

[2] Food and Drug Administration. Adaptive Design Clinical Trials for Drugs and Biologics Guidance for Industry. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry［2021 年 7 月 8 日アクセス］

[3] 日本製薬工業協会，アダプティブデザインに関する FDA ガイダンスの邦訳，2021 年 8 月 . https://www.jpma.or.jp/information/evaluation/results/allotment/lofurc000000n5a8-att/adaptive_design.pdf ［2022 年 6 月 22 日アクセス］

[4] 小宮山 靖，越水 孝，菅波 秀規，酒井 弘憲，渡橋 靖，東宮 秀夫．（2009）医薬品の臨床開発におけるアダプティブ・デザイン —米国研究製薬工業協会ワーキング・グループのエグゼクティブ・サマリー邦訳—．臨床薬理．40 巻 6 号 p. 303-310.

[5] 厚生省．臨床試験のための統計的原則．医薬審第 1047 号，平成 10 年 11 月 30 日．

[6] Jennison C and Turnbull BW. Group Sequential Methods with Applications to Clinical Trials. Chapman and Hall/CRC, Boca Raton, FL, 1999. 森川 敏彦，山中 竹春訳．臨床試験における群逐次法 理論と応用．EP クロア．東京．2012.

[7] Bauer P and Kohne K. Evaluation of experiments with adaptive interim analyses. Biometrics 1994;50:1029-1041.

[8] Cui L, Hung, HMJ, Wang SJ. Modification of sample-size in group sequential trials. Biometrics 1999;55:853-857.

[9] Lehmacher W, Wassmer G. Adaptive sample-size calculations in group sequential trials. Biometrics 1999;55:1286-1290.

[10] Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. Biometrics 1995;51:1315-1324.

[11] Chang M. Adaptive design theory and implementation using SAS and R. 2nd ed, Boca Raton., FL: CRC Press, 2014.

[12] Brannath W, König F, Bauer P. Improved repeated confidence bounds in trials with a maximal goal. Biometrical Journal 2003;45:311-324.

[13] Brannath W, König F, Bauer P. Estimation in flexible two stage designs. Statistics in Medicine

2006;25:3366-3381.

[14] Brannath W, Mehta CR, Posch M. Exact confidence bounds following adaptive group sequential tests. Biometrics 2009;65(2):539-546.

[15] Gao P, Liu L, Mehta C. Exact inference for adaptive group sequential designs. Statistical in Medicine 2013;32(23):3991-4005.

[16] Liu Q, Proschan MA, Pledger GW. A unified theory of two-stage adaptive designs. Journal of the American Statistical Association 2002;97:1034-1041.

[17] Mehta CR, Bauer P, Posch M, Brannath W. Repeated confidence intervals for adaptive group sequential trials. Statistics in Medicine 2007;26(30):5422-5433.

[18] Koenig F, Brannath W, Bretz F, Posch M. Adaptive Dunnett tests for treatment selection. Statistics in Medicine 2008;10;27(10):1612-25.

[19] Magirr D, Jaki T, Whitehead J. A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection. Biometrika 2012;99(2):494-501.

[20] Cohen A, Sackrowitz HB. Two stage conditionally unbiased estimators of the selected mean. Statistics & Probability Letters 1989;8(3):273-278.

[21] Bowden J, Glimm E. Unbiased estimation of selected treatment means in two-stage trials. Biometrical Journal 2008;50(4):515-527.

[22] Kimani PK, Todd S, and Stallard N. Conditionally unbiased estimation in phase II/III clinical trials with early stopping for futility. Statistics in Medicine 2013;32(17):2893-910.

[23] Robertson DS, Prevost AT, Bowden J. Unbiased estimation in seamless phase II/III trials with unequal treatment effect variances and hypothesis-driven selection rules. Statistical in Medicine 2016;35:3907-22.

[24] Stallard N, Kimani PK. Uniformly minimum variance conditionally unbiased estimation in multi-arm multi-stage clinical trials. Biometrika 2018;105(2):495-501.

[25] Robertson DS and Glimm E. Conditionally unbiased estimation in the normal setting with unknown variances. Communications in Statistics: Theory and Methods 2019;48:616-627

[26] Stallard N, Todd S. Point estimates and confidence regions for sequential trials involving selection. Journal of Statistical Planning and Inference 2005;135(2):402-419.

[27] Sampson AR and Sill MW. Drop-the-losers design: Normal case. Biometrical Journal 2005;47(3):257-68.

[28] Posch M, Koenig F, Branson M, Brannath W, Dunger-Baldauf C, Bauer P. Testing and estimation in flexible group sequential designs with adaptive treatment selection. Statistics in Medicine 2005;24:3697-3714.

2　症例数再推定

症例数の変更を伴うアダプティブデザイン（以下、症例数再推定と呼ぶ）は、アダプティブデザインの方法論に関する研究が進み始めた当初から検討がなされてきた。

本章では、結果変数が連続変数であり、1 回の中間解析において症例数再推定を実施する場合の統計学的推測法（検定、点推定、区間推定）に関するいくつかの方法を紹介する。2.2 項では、非盲検下での症例数再推定を実施し、通常の検定統計量を用いて検定を実施した場合の第一種の過誤確率への影響について紹介する。症例数再推定には、被験者の治療群の割付情報を用いる方法と用いない方法が存在する。慣習的な表現に基づき、本報告書では、前者を非盲検下の症例数再推定、後者を盲検下の症例数再推定と呼ぶこととし、非盲検下の症例数再推定に関する統計学的推測法は 2.2 項、盲検下の症例数再推定に関する統計学的推測法を 2.3 項にて解説する。

2.1　第 2 章で用いる記号の整理

本章で用いる記号の説明を表 1 に示す。

表 1　本章で用いる記号の説明

記号	説明
n_1, n_2	試験計画時の第 1、第 2 ステージにおける 1 群あたりの症例数
$\hat{n}_2$	症例数再推定後の第 2 ステージにおける 1 群あたりの症例数
$N_{initial}$	試験計画時の試験全体の 1 群あたりの症例数（$= n_1 + n_2$）
$\widehat{N}_{final}$	症例数再推定後の試験全体の 1 群あたりの症例数（$= n_1 + \hat{n}_2$）
N_{min}	試験全体の 1 群あたりの最小症例数
N_{max}	試験全体の 1 群あたりの最大症例数
w_1, w_2	各ステージの重み。条件$w_1^2 + w_2^2 = 1$を満たすものとする。
t	第 1 ステージにおける症例数の比 $t = n_1/N_{initial}$
μ	1 標本平均値の真値
δ	平均値の群間差の真値
δ_{pre}	試験計画時に想定した平均値の群間差
$\hat{\delta}_1$	第 1 ステージのデータに基づく平均値の群間差の推定値
$\hat{\delta}_2$	第 2 ステージのデータに基づく平均値の群間差の推定値
$\hat{\delta}_{naive}$	試験全体のデータに基づく平均値の群間差の推定値
$\bar{x}_1$	第 1 ステージのデータに基づく各群の平均値の推定値
$\bar{x}_2$	第 2 ステージのデータに基づく各群の平均値の推定値
$\bar{x}$	試験全体のデータに基づく各群の平均値の推定値
σ^2	群内分散の真値

記号	説明
s_{final}^2	最終解析の併合分散
s_{lumped}^2	群併合の盲検下データに基づく単純に推定した分散
s_1^2	第 1 ステージのデータに基づく併合分散
Z_1	第 1 ステージのデータに基づく標準化検定統計量
Z_2	第 2 ステージのデータに基づく標準化検定統計量
Z_{naive}	ナイーブな最終解析の標準化検定統計量
$Z_{weighted}$	重み付き標準化検定統計量；$Z_{weighted} = w_1 Z_1 + w_2 Z_2$
p_1	第 1 ステージのデータに基づく p 値
p_0	第 2 ステージのデータに基づく p 値
α_1	第 1 ステージのデータに基づく、有効と判断するための閾値（$p_1 < \alpha_1$ならば有効と判断）
α_0	第 1 ステージのデータに基づく、無効と判断するための閾値（$p_1 > \alpha_0$ならば無効と判断）
c_Z	標準化検定統計量で表したときの棄却限界値（c_Zより大きい場合に棄却）
c_p	p 値で表したときの棄却限界値（c_pより小さい場合に棄却）
$Q(p_1, p_2)$	Overall p 値

2.2 非盲検下での症例数再推定

2.2.1 第一種の過誤確率への影響

非盲検下での症例数再推定を伴うアダプティブデザインを適用し、通常の検定統計量を用いて検定を実施する場合、たとえ有効中止を伴う中間解析を実施しない場合であっても、第一種の過誤確率が増大する可能性がある。例えば、(帰無仮説の下で）中間解析の結果が非常に良好なときに、速やかに最終解析を実施する場合（症例数を減少させる場合)、有意な結果が得られやすくなり、その結果、第一種の過誤確率が増大することなる。また、場合によっては、第一種の過誤確率が減少することもある。第一種の過誤確率が増大するかどうか、増大の程度は、症例数再推定の条件（例：症例数の上限、症例数の変更を許容する範囲等）によって異なる。

本項では、主要評価項目が連続変数の二群比較試験で、中間解析時の治療効果の推定値を用いて症例数を算出し、最終解析時にナイーブな検定統計量を用いて検定を実施する場合を想定する。また、中間解析は、症例数再推定を目的として実施することとし、有効中止は考慮しないこととする。このような条件下における、さまざまな症例数の決定規則における第一種の過誤確率への影響について、Shun (2001)[1]の論文に沿って解説する。

なお、本項で提示する計算結果やグラフ作成に用いたRプログラムは本報告書の補遺に添付しているので、適宜参照されたい。

2.2.1.1 2.2項で用いる記号の整理

- 帰無仮説$H_0: \mu_1 = \mu_0$、対立仮説$H_1: \mu_1 \neq \mu_0$の両側検定について考える。
- 試験開始前に設定された症例数を$N_{initial}$例とし、中間解析においてn_1例のデータに基づき、中間解析時に算出された試験全体の1群あたりの症例数をN例とし、$N_{initial} = n_1 + n_2$、$N = n_1 + K$と表すこととする。Kは非負整数の確率変数である。
- 試験開始前に設定された症例数$N_{initial}$は、名義有意水準（両側）をα、群間差の想定値をδ_{pre}とし、検出力$1-\beta$を確保できるように設定された症例数とする。このとき$N_{initial}$は、次式から算出することができる。ここで、z_pは標準正規分布の上側p%点とし、σは既知とする。

$$N_{initial} = \left(z_{\alpha/2} + z_{\beta}\right)^2 \frac{2\sigma^2}{\delta_{pre}^2}$$

- 中間解析時に算出された症例数Nは、名義有意水準をα、群間差の想定値を$\hat{\delta}_1$とし、検出力$1-\beta$を確保できるように設定された症例数とする。このときNは、次式から算出することができる。

$$N = \left(z_{\alpha/2} + z_{\beta}\right)^2 \frac{2\sigma^2}{\hat{\delta}_1^2}$$

- n_1例のデータに基づく、中間解析時の検定統計量をZ_{n_1}とする。$Z_{n_1} = \sqrt{\frac{n_1}{2\sigma^2}}\hat{\delta}_1$と表すことができる。
- $C_\beta = z_{\alpha/2} + z_{\beta}$と定義する。
- 中間解析時に算出された症例数Nに関する等式について、両辺をn_1で割って整理すると、中間解析時の検定統計量Z_{n_1}は以下のように表すことができる。

$$Z_{n_1} = \left(z_{\alpha/2} + z_{\beta}\right)\sqrt{\frac{n_1}{N}} = C_\beta\sqrt{\frac{n_1}{N}}$$

2.2.1.2 第一種の過誤確率の計算

2.2.1.2.1 症例数の変更が伴わない場合

帰無仮説下における、症例数の変更が伴わない場合（$K = n_2$）の第一種の過誤確率は次のように表される。ここで、$\phi(\cdot)$は標準正規分布の確率密度関数とする。

$$\alpha = 2Pr\left(Z_N > z_{\alpha/2}\right) = 2\int_{-\infty}^{\infty} Pr\left(Z_N > z_{\alpha/2} \middle| Z_{n_1}\right)\phi\left(Z_{n_1}\right)dZ_{n_1}$$

$Z_N = \sqrt{\frac{n_1}{N}} Z_{n_1} + \sqrt{\frac{K}{N}} Z_K$[1] を代入して整理すると

$$\alpha = 2\int_{-\infty}^{\infty} \Phi\left(\sqrt{\frac{n_1}{n_2}} Z_{n_1} - \sqrt{1+\frac{n_1}{n_2}} z_{\alpha/2}\right) \phi\left(Z_{n_1}\right) dZ_{n_1}$$

となる。ここで、$\Phi(\cdot)$は標準正規分布の累積分布関数とする。ここに、試験計画時の試験全体の症例数に対する中間解析時の症例数の比を$t = n_1/N_{initial}$とする。

また、$t < 1$において、

$$f_0^{\pm}(z,t) = \pm\sqrt{\frac{t}{1-t}} z - \sqrt{\frac{1}{1-t}} z_{\alpha/2}、\ G_0^{\pm}(z,t) = \Phi\left(f_0^{\pm}(z,t)\right)\phi(z)$$

と定義する。このとき、

$$\lim_{t\to 1} f_0^{+}(z,t) = \begin{cases} +\infty, & z > z_{\alpha/2} \\ 0, & z = z_{\alpha/2} \\ -\infty, & z < -z_{\alpha/2} \end{cases}$$

$$\lim_{t\to 1} f_0^{-}(z,t) = \begin{cases} +\infty, & z < -z_{\alpha/2} \\ 0, & z = -z_{\alpha/2} \\ -\infty, & z > -z_{\alpha/2} \end{cases}$$

が成り立つ。

さらに、$t = 1$とき、$G_0^{+}(z,t)$、$G_0^{-}(z,t)$をそれぞれ以下のように定義する。

$$G_0^{+}(z,1) = \begin{cases} \phi(z), & z > z_{\alpha/2} \\ 0, & z \le z_{\alpha/2} \end{cases} \qquad G_0^{-}(z,1) = \begin{cases} 0, & z < -z_{\alpha/2} \\ \phi(z), & z \ge -z_{\alpha/2} \end{cases}$$

症例数の変更が伴わない場合の第一種の過誤確率は、次式で表すことができ、中間解析の症例数の比tに依存する、$G_0^{+}(z,t)$の曲線下面積となる。なお、$G_0^{+}(z,t)$の形状はtに依存して変化するものの、曲線下面積はtに依らず、αとなる。

$$\alpha = 2\int_{-\infty}^{\infty} \Phi\left(f_0^{+}(z,t)\right)\phi(z)dz = 2\int_{-\infty}^{\infty} G_0^{+}(z,t)dz$$

2.2.1.2.2 症例数の変更が伴う場合

次に症例数の変更が伴う場合の第一種の過誤確率について考える。中間解析時の治療効果の推定値を用いて症例数を算出した場合の第一種の過誤確率は、次式から算出することができる。

---

[1] $Z_N = \sqrt{\frac{N}{2\sigma^2}}\hat{\delta}_N = \sqrt{\frac{n_1}{N}} \cdot \sqrt{\frac{n_1}{2\sigma^2}}\hat{\delta}_1 + \sqrt{\frac{K}{N}} \cdot \sqrt{\frac{K}{2\sigma^2}}\hat{\delta}_K = \sqrt{\frac{n_1}{N}} Z_{n_1} + \sqrt{\frac{K}{N}} Z_K$. ここで、$Z_{n_1} = \sqrt{\frac{n_1}{2\sigma^2}}\hat{\delta}_1$、 $Z_K = \sqrt{\frac{K}{2\sigma^2}}\hat{\delta}_K$ である。

なお、α^*の導出に関しては 2.2.1.5 項を参照されたい。

$$\alpha^* = Pr\left(|Z_N| > z_{\alpha/2}\right) = \int_{-\infty}^{\infty} Pr\left(|Z_N| > z_{\alpha/2}|Z_{n_1}\right)\phi\left(Z_{n_1}\right)dZ_{n_1}$$

$$= \int_{-\infty}^{\infty} Pr\left(Z_N > z_{\alpha/2}|Z_{n_1}\right)\phi\left(Z_{n_1}\right)dZ_{n_1} + \int_{-\infty}^{\infty} Pr\left(Z_N < -z_{\alpha/2}|Z_{n_1}\right)\phi\left(Z_{n_1}\right)dZ_{n_1}$$

また、$|z| < C_\beta$において、

$$f^{\pm}(z) = \left(\pm z - \sqrt{1+\frac{K}{n_1}}z_{\alpha/2}\right)\Big/\sqrt{\frac{K}{n_1}}, \qquad G_\infty^{\pm}(z) = \Phi\left(f^{\pm}(z)\right)\phi(z)$$

と定義する。

このとき、症例数の変更が伴う場合の第一種の過誤確率は、次式で表すことができる。

$$\alpha^* = 2\Phi\left(-C_\beta\right) + 2\int_{-C_\beta}^{C_\beta} G_\infty^{+}(z)dz$$

表 2 に$\alpha = 0.05$、検出力を 80%、85%、90%、95%とした場合のα^*を示した。いずれの場合もα^*は 0.05 を超えており、第一種の過誤確率の増大が確認できる。また、第一種の過誤確率の増大は検出力にのみ依存することがわかる。また、検討した検出力の範囲では、検出力が小さいほど、第一種の過誤確率の増大が大きくなっている。

表 2　症例数の変更が伴う場合の第一種の過誤確率の増大の程度（制約条件なし）

検出力（%）	C_β	第一種の過誤確率（α^*）	第一種の過誤確率の増大の割合（%）
80	2.8016	0.07996	60
85	2.9963	0.07624	52
90	3.2415	0.07244	45
95	3.6048	0.06824	36

中間解析時の検定統計量zと、$G_\infty^{+}(z)$と$G_0^{+}(z,t)$の関係を図 1 に示した。症例数の変更が伴わない場合については、中間解析時点$t = 1/2$、$1/3$、$2/3$について示した。症例数の変更が伴わない場合、中間解析の時期が遅くなるにつれて（t が大きいほど）、右にシフトしている。

図からわかるように、$G_0^{+}(z,t)$と$G_\infty^{+}(z)$の形状は大きく異なっており、第一種の過誤確率が増大する可能性があることがわかる。$G_0^{+}(z,t)$とG_∞^{+}の交点は、各 t についてそれぞれ 2 点存在し、2 つの交点の間の区間では、$G_0^{+}(z,t)$の方が大きく、それ以外の区間では、$G_\infty^{+}(z)$の方が大きくなっている。

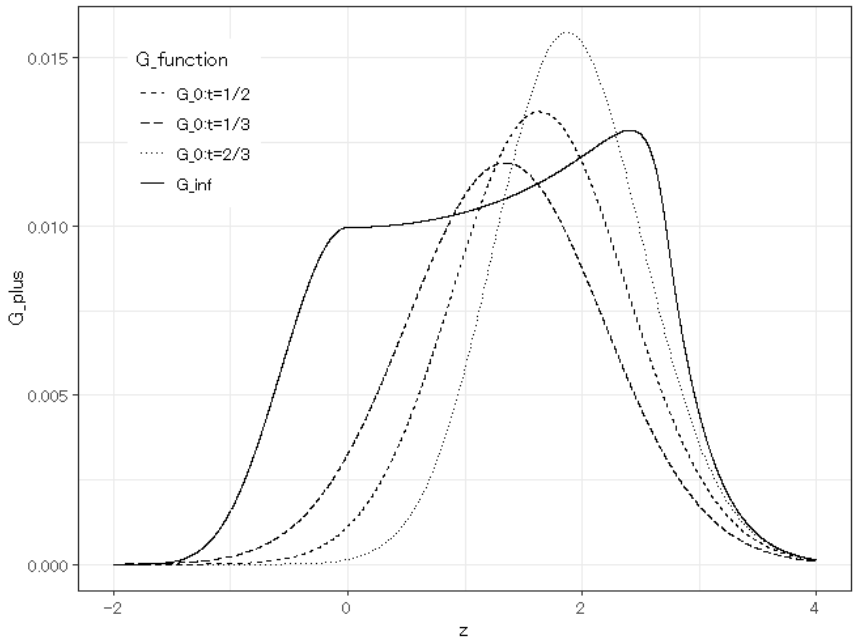


図 1　zと、$G_{\infty}^{+}(z)$と$G_{0}^{+}(z,t)$との関係

2.2.1.3　　症例数の上限と下限が設定されている場合の第一種の過誤確率

前項では、再推定する症例数に制約がない（上限や下限が設定されていない）場合の第一種の過誤確率への影響について説明した。しかしながら、実際上は、臨床試験に費用や期間の制約が課されることになり、症例数の上限が設定されることが考えられる。そこで、本項では、症例数に制約が課されている場合における第一種の過誤確率について考える。なお、Shun (2001)[1]の論文に従って、検出力はいずれの場合も 85%に設定した。

まず、症例数の上限と下限がそれぞれN_{max}とN_{min}に設定されている場合における、第一種の過誤確率について考える。ここでは、中間解析時の治療効果の推定値$\hat{\delta}_1$に基づいて、症例数は以下の決定規則に従って設定されるものとする。

$$\widehat{N}_{final} = \begin{cases} n_1 + K = C_{\beta}^{2}\dfrac{2\sigma^2}{\hat{\delta}_1^2}, & N_{min} \le N \le N_{max} \\ N_{max}, & N > N_{max} \\ N_{min}, & N < N_{min} \end{cases}$$

$N = N_{min}$、N_{max}であるときの、中間解析時の症例数n_1との例数比をそれぞれ$t_{min} = n_1/N_{min}$、$t_{max} = n_1/N_{max}$とし、Z_{n_1}を$z(N_{min}) = C_{\beta}\sqrt{t_{min}}$、$z(N_{max}) = C_{\beta}\sqrt{t_{max}}$とする。ここで、前項で示した$G_{\infty}^{\pm}(z)$、$G_{0}^{\pm}(z,t)$を利用して、次のような関数$G^{\pm}(z,t)$を定義する。

$$G^{\pm}(z,t)=\begin{cases} G_{\infty}^{\pm}(z), & z(N_{max})\le|z|\le z(N_{min}) \\ G_{0}^{\pm}(z,t_{max}), & |z|<z(N_{max}) \\ G_{0}^{\pm}(z,t_{min}), & |z|>z(N_{min}) \end{cases}$$

このとき、第一種の過誤確率の増大の程度α_{diff}は次式で表される。

$$\alpha_{diff}=2\int_{-\infty}^{\infty}G^{+}(z,t)dz-\alpha=2\int_{-\infty}^{\infty}[G^{+}(z,t)-G_{0}^{+}(z,t)]dz$$

表 3 は、N_{min}を試験計画時の症例数$N_{initial}$、N_{max}を$rN_{initial}$とした場合の第一種の過誤確率の増大の程度である。中間解析時点が遅いほど（tが大きいほど)、症例数の上限が大きいほど、第一種の過誤確率の増大の程度が大きくなっていることがわかる。

表 3　症例数の上限N_{max}と下限N_{min}が設定されている場合の
第一種の過誤確率の増大の程度
（$t=n_1/N_{initial}$、$N_{min}=N_{initial}$、$N_{max}=rN_{initial}$）

t	r	第一種の過誤確率の増大の程度
1/3	1.5	0.00466
	2	0.00737
	3	0.01028
	4	0.01175
	5	0.01262
1/2	1.5	0.00538
	2	0.00879
	3	0.01275
	4	0.01492
	5	0.01625
2/3	1.5	0.00531
	2	0.00909
	3	0.01375
	4	0.01645
	5	0.01818

2.2.1.3.1　　症例数の上限を超えた場合は症例数を変更しない場合

中間解析時に算出された症例数Nが症例数の上限N_{max}を超えた場合は、試験計画時の症例数$N_{initial}$とする、次のような症例数の決定規則の場合について考える。

$$\widehat{N}_{final} = \begin{cases} n_1 + K, & N_{initial} \le N \le rN_{initial} \\ N_{initial}, & N < N_{initial} \ \ or \ \ N > rN_{initial} \end{cases}$$

このとき、$z(N_{min}) = z(N_{initial}) = C_\beta \sqrt{t}$、$z(N_{max}) = z(rN_{initial}) = C_\beta \sqrt{t/r}$である。また、$G^{\pm}(z,t)$を以下のように定義したとき、

$$G^{\pm}(z,t) = \begin{cases} G_{\infty}^{\pm}(z), & z(N_{max}) \le |z| \le z(N_{min}) \\ G_0^{\pm}(z,t), & |z| < z(N_{max}) \ \ or \ \ |z| > z(N_{min}) \end{cases}$$

第一種の過誤確率の増大の程度は、次のように表される。

$$\alpha_{diff} = 2\left\{\int_{z(N_{max})}^{z(N_{min})} [G_{\infty}^{+}(z) - G_0^{+}(z,t)]dz + \int_{-z(N_{min})}^{-z(N_{max})} [G_{\infty}^{+}(z) - G_0^{+}(z,t)]dz\right\}$$

表 4 に、中間解析時点$t = 1/3$、$1/2$、$2/3$、$r = 1.5$~12（最大）の場合の第一種の過誤確率の増大の程度を示した。中間解析時点$t = 1/3$の場合は症例数の上限が試験計画時の症例数の 5 倍の場合に第一種の過誤確率の増大し、中間解析時点$t = 1/2$、$2/3$の場合は 9 倍、12 倍の場合に第一種の過誤確率が増大した。その他の場合は、第一種の過誤確率が減少した。

表 4　症例数の上限を超えた場合は症例数を変更しない場合の
第一種の過誤確率の増大の程度

t	*r*	第一種の過誤確率の増大の程度
1/3	1.5	-0.00048
	2	-0.00091
	3	-0.00046
	4	-0.00050
	5	0.00009
1/2	1.5	-0.00118
	2	-0.00242
	3	-0.00335
	4	-0.00315
	8	-0.00032
	9	0.00036
2/3	1.5	-0.00229
	2	-0.00489
	3	-0.00693
	4	-0.00682

t	r	第一種の過誤確率の増大の程度
	5	-0.00600
	11	-0.00035
	12	0.00039

2.2.1.3.2 症例数を変更しない、または減少させる場合

中間解析時に算出された症例数Nが、中間解析時の症例数未満であればn_1、計画時の症例数を上回るならば$N_{initial}$、それ以外は再計算後の症例数とする、次のような症例数の決定規則の場合について考える。

$$\widehat{N}_{final} = \begin{cases} n_1 + K, & n_1 \le N \le N_{initial} \\ n_1, & N < n_1 \\ N_{initial}, & N > N_{initial} \end{cases}$$

このとき、$z(N_{min}) = z(n_1) = C_\beta$ 、$z(N_{max}) = z(N_{initial}) = C_\beta \sqrt{t}$である。

また、$G^{\pm}(z,t)$を以下のように定義する。

$$G^{\pm}(z,t) = \begin{cases} G_\infty^{\pm}(z), & z(N_{max}) \le |z| \\ G_0^{\pm}(z,t), & |z| < z(N_{max}) \end{cases}$$

2.2.1.3 項のα_{diff}の式を用いて、中間解析時点を変えた場合の、第一種の過誤確率の増大の程度を表 5 に示した。中間解析時点が遅いほど、第一種の過誤確率の増大の程度が小さいことがわかる。

$z = z(N_{max}) = C_\beta \sqrt{t}$の場合、$G_0^{+}(z,t) = G_\infty^{+}(z)$が満たされる。また、$z > z(N_{max})$における$G^{\pm}(z,t)$は$G_\infty^{\pm}(z)$であり、図 1 からこの区間では$G_\infty^{+}(z) > G_0^{+}(z,t)$であることから、第一種の過誤確率が増大するのは明らかである。中間解析の時点が早いほど（tが小さいほど)、G_0^{+}は左方向へシフトし、G_∞^{+}とG_0^{+}の間の面積が大きくなっており、計算結果と傾向が一致していることがわかる。

表 5　症例数は変更しない、または減少させる場合の
第一種の過誤確率の増大の程度

t	第一種の過誤確率の増大の程度
1/4	0.01443
1/3	0.01102
1/2	0.00564

t	第一種の過誤確率の増大の程度
2/3	0.00194
3/4	0.00078

2.2.1.3.3 試験計画時の症例数以下で継続するか、中止する場合

前項で検討した症例数の決定規則のうち、中間解析時に算出された症例数Nが計画時の症例数$N_{initial}$を上回るならばn_1とする、次のような症例数の決定規則の場合について考える。

$$\widehat{N}_{final} = \begin{cases} n_1 + K, & n_1 \le N \le N_{initial} \\ n_1, & N < n_1 \ \ or \ \ N > N_{initial} \end{cases}$$

また、$G^{\pm}(z,t)$を以下のように定義する。

$$G^{\pm}(z,t) = \begin{cases} G_{\infty}^{\pm}(z), & z(N_{initial}) \le |z| \\ G_0^{\pm}(z,t), & |z| < z(N_{initial}) \end{cases}$$

2.2.1.3 項のα_{diff}の式を用いて、表 6 に中間解析時点$t = 1/4、1/3、1/2、2/3、3/4$に対する第一種の過誤確率の増大の程度α_{diff}を示した。いずれの場合も第一種の過誤確率が減少し、中間解析が早い方がより第一種の過誤確率が減少した。

表 6　試験計画時の症例数で継続するか、中止する場合の
第一種の過誤確率の増大の程度

t	第一種の過誤確率の増大の程度
1/4	-0.01782
1/3	-0.02275
1/2	-0.01526
2/3	-0.00263
3/4	-0.0007

2.2.1.3.4 症例数の上限を設定し、ある一定の検出力が確保される場合に、症例数を増加させる場合

2.2.1.3.1 項の症例数の決定規則は、中間解析時に算出された症例数Nが症例数の上限を超えた場合は、試験計画時の症例数$N_{initial}$に設定する規則であった。本項では、Nが症例数の上限を超えた場合であっても、症例数の上限（$N_{max} = rN_{initial}$）における検出力が

$1-\beta_{min}$以上確保される場合は$\widehat{N}_{final}=rN_{initial}$とする、次のような症例数の決定規則の場合について考える。

$$\widehat{N}_{final}=\begin{cases} n_1+K, & n\le N\le rN_{initial} \\ rN_{initial}, & rN_{initial}<N<n(\beta_{min}) \\ N_{initial}, & N<n_1 \ \ or \ \ N>n(\beta_{min}) \end{cases}$$

ここで、$n(\beta_{min})$は、名義有意水準をα、群間差の想定値を$\hat{\delta}_{n_1}$とし、検出力$1-\beta_{min}$を確保できるように設定された症例数である。また、$G^{\pm}(z,t)$を以下のように定義する。

$$G^{\pm}(z,t)=\begin{cases} G_{\infty}^{\pm}(z), & z(rN_{initial})\le|z|\le\ z(N_{initial}) \\ G_0^{\pm}(z,t/r), & z\big(n(\beta_{min})\big)<|z|<z(rN_{initial}) \\ G_0^{\pm}(z,t), & |z|>z(n)\ \ or\ \ |z|<z\big(n(\beta_{min})\big) \end{cases}$$

表 7 に、$\alpha=0.05$、$1-\beta_{min}$を 75%、65%、50%、$t=1/3$、$1/2$、$2/3$とし、rを1.5以上の様々な値に設定した場合の第一種の過誤確率の増大の程度を示した。$1-\beta_{min}$ =75%の場合、中間解析時点$t=1/3$の場合は症例数の上限が試験計画時の症例数の 4 倍の場合に第一種の過誤確率の増大が確認され、中間解析時点$t=1/2$、$2/3$の場合は 7 倍、9 倍の場合に第一種の過誤確率の増大が確認された。その他の場合は、第一種の過誤確率の減少が確認された。また、$1-\beta_{min}$を小さくするほど、第一種の過誤確率の増大の程度が大きくなることが確認された。

表 7　　症例数の上限を設定し、ある一定の検出力が確保される場合に
症例数を増加させる場合の、第一種の過誤確率の増大の程度

t	*r*	$1-\beta_{min}$		
		75%	65%	50%
1/3	1.5	-0.00079	-0.00086	-0.00066
	2	-0.00101	-0.00086	-0.00031
	3	-0.00059	-0.00006	0.00096
	4	0.00018	0.00091	0.00215
1/2	1.5	-0.00207	-0.00242	-0.00231
	2	-0.00305	-0.00309	-0.00247
	3	-0.00318	-0.00261	-0.00125
	4	-0.00243	-0.00148	0.00027
	5	-0.00147	-0.00031	0.00165
	6	-0.00052	0.00077	0.00283
	7	0.00037	0.00174	0.00383
2/3	1.5	-0.00421	-0.00503	-0.0048

t	r	$1-\beta_{min}$		
		75%	65%	50%
	2	-0.00625	-0.0064	-0.00528
	3	-0.0068	-0.00592	-0.00383
	4	-0.0058	-0.00442	-0.0019
	5	-0.00449	-0.00284	-0.00014
	6	-0.00319	-0.0014	0.00138
	7	-0.00198	-0.00013	0.00268
	8	-0.00088	0.001	0.0038
	9	0.00011	0.002	0.00477

2.2.1.4 まとめ

本項では、Shun (2001) [1]の論文に沿って、非盲検下での症例数再推定による第一種の過誤確率への影響について説明した。

症例数に制限が課されていない場合、検討した検出力が 80%～95%の範囲では、第一種の過誤確率が約 1.8%～3%の増大がみられ、検出力が小さいほど、増大の程度は大きかった（2.2.1.2 項参照）。有効中止を伴う中間解析を実施しない場合であっても、第一種の過誤確率が増大することが示された。

症例数の上限や下限が設定されている場合、第一種の過誤確率の増大の程度は、症例数の決定規則に従って定義される区分関数で表すことができ、中間解析の時期、症例数の上限、症例数の決定規則により異なっている。症例数の上限と下限が設定されている場合の第一種の過誤確率の挙動を、2.2.1.3 項で説明した症例数の決定規則別に表 8 にまとめた。表 8 の①と②は、中間解析時に算出された症例数Nが症例数の上限を超えた場合の$\hat{N}_{final}$の定め方に違いがあり、それにより第一種の過誤確率の挙動が異なっていると考えられる。中間解析時に算出された症例数Nが上限を超えた場合に、症例数の上限（N_{max}=$rN_{initial}$）を$\hat{N}_{final}$と設定する場合（表 8 の①の場合）、第一種の過誤確率の増大がみられた。また、中間解析の時期が遅いほど、症例数の上限が大きいほど、増大の程度は大きかった（2.2.1.3 項参照）。一方、Nが上限を超えた場合に症例数を変更しない場合（表 8 の②の場合）、症例数の上限が大きい場合（計画被験者数の 5 倍以上）を除き、第一種の過誤確率の増大がみられなかった（2.2.1.3.1 項参照）。症例数の上限を設定し、ある一定の検出力が確保される場合のみ、症例数を増加させる場合、症例数の増加を許容する検出力の大きさを小さくするほど、第一種の過誤確率の増大の程度が大きくなることが確認された（2.2.1.3.4 項参照）。

表 8　　症例数の上限と下限が設定している場合の第一種の過誤確率の挙動

	症例数の決定規則（$\widehat{N}_{final}$）	第一種の過誤確率の挙動
①	$\widehat{N}_{final} = \begin{cases} n_1 + K, & N_{initial} \le N \le rN_{initial} \\ rN_{initial}, & N > rN_{initial} \\ N_{initial}, & N < N_{initial} \end{cases}$	中間解析時点が遅いほど、症例数の上限が大きいほど、第一種の過誤確率が大きい。
②	$\widehat{N}_{final} = \begin{cases} n_1 + K, & N_{initial} \le N \le rN_{initial} \\ N_{initial}, & otherwise \end{cases}$	中間解析時点が早いほど、症例数の上限が大きいほど、第一種の過誤確率が大きい。 検討した範囲では、多くの場合は、第一種の過誤確率は名義水準以下に制御される。
③	$\widehat{N}_{final} = \begin{cases} n_1 + K, & n_1 \le N \le N_{initial} \\ n_1, & N < n_1 \\ N_{initial}, & N > N_{initial} \end{cases}$	中間解析時点が早いほど、第一種の過誤確率が大きい。
④	$\widehat{N}_{final} = \begin{cases} n_1 + K, & n_1 \le N \le N_{initial} \\ n_1, & otherwise \end{cases}$	中間解析時点が遅いほど、第一種の過誤確率が大きい。 検討した範囲では、いずれの場合も第一種の過誤確率が名義水準以下に制御される。
⑤	$\widehat{N}_{final} = \begin{cases} n_1 + K, & n \le N \le rN_{initial} \\ rN_{initial}, & rN_{initial} < N < n(\beta_{min}) \\ N_{initial}, & N < n_1 \ \ or \ \ N > n(\beta_{min}) \end{cases}$	中間解析時点が早いほど、症例数の上限が大きいほど、第一種の過誤確率が大きい。 検討した範囲では、多くの場合は、第一種の過誤確率は名義水準以下に制御される。

2.2.1.5　　式の導出

$\alpha^* = 2\Phi(-C_\beta) + 2\int_{-C_\beta}^{C_\beta} \Phi\left(\left(z - \sqrt{1+\frac{K}{n_1}}z_{\alpha/2}\right)\Big/\sqrt{\frac{K}{n_1}}\right)\phi(z)dz$の導出

$K > 0$の場合、$Z_N = \sqrt{\frac{n_1}{N}}Z_{n_1} + \sqrt{\frac{K}{N}}Z_K$ を代入して整理すると

$$Pr(Z_N > z_{\alpha/2}|Z_{n_1}) = Pr\left(\sqrt{\frac{n_1}{N}}Z_{n_1} + \sqrt{\frac{K}{N}}Z_K > z_{\alpha/2}|Z_{n_1}\right)$$

$$= Pr\left(Z_K > \left(z_{\alpha/2} - \sqrt{\frac{n_1}{N}}Z_{n_1}\right)\Big/\sqrt{\frac{K}{N}}|Z_{n_1}\right) = Pr\left(Z_K < \left(Z_{n_1} - \sqrt{1+\frac{K}{n_1}}z_{\alpha/2}\right)\Big/\sqrt{\frac{K}{n_1}}|Z_{n_1}\right)$$

$$= \Phi\left(\left(Z_{n_1} - \sqrt{1+\frac{K}{n_1}}z_{\alpha/2}\right)\Big/\sqrt{\frac{K}{n_1}}\right)$$

となる。ここに、$\Phi(.)$は標準正規分布の累積分布関数とする。
同様にして、

$$Pr(Z_N < -z_{\alpha/2}|Z_{n_1}) = \Phi\left(\left(-Z_{n_1} - \sqrt{1+\frac{K}{n_1}}z_{\alpha/2}\right)\Big/\sqrt{\frac{K}{n_1}}\right)$$

が導かれる。
$K = 0$の場合も考慮し、次のような条件付き第一種の過誤確率の関数を定義する。

$$H^+(z) = \begin{cases} 1, & z \geq C_\beta \\ \Phi\left(f^+(z)\right), & |z| < C_\beta \\ 0, & z \leq -C_\beta \end{cases}$$

$$H^-(z) = \begin{cases} 0, & z \geq C_\beta \\ \Phi\left(f^-(z)\right), & |z| < C_\beta \\ 1, & z \leq -C_\beta \end{cases}$$

ここに、$f^+(z) = \left(z - \sqrt{1+\frac{K}{n_1}}z_{\alpha/2}\right)\Big/\sqrt{\frac{K}{n_1}}$、$f^-(z) = \left(-z - \sqrt{1+\frac{K}{n_1}}z_{\alpha/2}\right)\Big/\sqrt{\frac{K}{n_1}}$とする。
このとき、$f^-(z) = f^+(-z)$、$H^-(z) = H^+(-z)$の関係が成り立つ。
さらに、$G_\infty^\pm(z)$を次のように定義する。

$$G_\infty^\pm(z) = H_\infty^\pm(z)\ \phi(z)$$

このとき、第一種の過誤確率は、$G_\infty^\pm(z)$を用いて次式で表すことができる。

$$\alpha^* = \int_{-\infty}^{\infty} G_{\infty}^{+}(z)dz + \int_{-\infty}^{\infty} G_{\infty}^{-}(z)dz = 1Pr\big(|z| > C_\beta\big) + \int_{-C_\beta}^{C_\beta} G_{\infty}^{+}(z)dz + \int_{-C_\beta}^{C_\beta} G_{\infty}^{-}(z)dz$$

$$= \Pr\big(|z| > C_\beta\big) + \int_{-C_\beta}^{C_\beta} H^{+}(z)\phi(z)dz + \int_{-C_\beta}^{C_\beta} H^{-}(z)\phi(z)dz$$

$$= 2\Phi\big(-C_\beta\big) + 2\int_{-C_\beta}^{C_\beta} H^{+}(z)\phi(z)dz \qquad \text{（終）}$$

2.2.2 検定手法と症例数再推定の方法

本項では、特に断りがないかぎり、2 ステージデザインで、有効・無益性中止を判断する中間解析を伴う非盲検下での症例数再推定を想定することとする。中止基準を以下のように設定し、最終解析で$p_2 < c_p$の場合に帰無仮説は棄却されるとする。

$$\begin{cases} \text{有効中止:} if\ \ p_1 < \alpha_1 \\ \text{無益性中止:} if\ \ p_1 > \alpha_0 \\ \text{アダプテーション後試験継続:} if\ \ \alpha_1 \le p_1 \le \alpha_0 \end{cases}$$

症例数を減少させるアダプテーションは考慮しないこととする。

3 ステージ以上での適用に関しては、以下の各項にて参照している論文又はテキストに紹介があるため適宜参照されたい。

2.2.2.1 第一種の過誤確率を制御するための方法

本項では、非盲検下での症例数再推定における第一種の過誤確率の制御方法について紹介する。第一種の過誤確率の制御方法は複数提案されており、大きく分けて以下の 3 種のアプローチに分類される。第 1 のアプローチは、**統合検定 (Combination test) による方法**であり、各ステージの部分標本について、それぞれ独立して得られた検定統計量（又はp値）を事前に規定した関数形式によって統合する検定手法である。Bauer and Kohne (1994)[2]、Cui, Hung and Wang (1999)[3]、Lehmacher and Wassmer (1999)[4]、Chang (2006)[5]等があげられる。第 2 のアプローチは、**条件付き過誤関数 (Conditional error function ; CEF) による方法**であり、事前に規定した条件付き過誤関数に基づき、棄却域を調整する方法である。Proschan and Hunsberger (1995)[6] 等があげられる。第 3 のアプローチは、アダプテーションのルールを変更し、固定標本デザインと同様のナイーブな検定手法を用いても、**第一種の過誤確率が増大しない範囲で症例数の変更を許容することで、第一種の過誤確率を制御する方法**である。Chen (2004)[7]や Mehta and Pocock (2011)[8]等があげられる。

2.2.2.1.1 統合検定 (Combination test)による方法

本項では、統合検定による代表的な方法を紹介する。Bauer and Kohne (1994)[2]は、Fisher の基準に基づき、各ステージのp値の積を用いてp値を統合し、棄却域を導出した。Cui, Hung and Wang (1999)[3]は、各ステージの Z 検定統計量を事前に規定した重みで統合する方法を提案した。また、同年に提案された Lehmacher and Wassmer (1999)[4]は、逆正規法を用いてこの重み付け法を一般化した方法となっている。Chang (2006)[5]は各ステージのp

値の和を用いて検定統計量を構築し、有効・無益性中止の境界、調整済みp値を計算するための閉形式を導出した。また、Chang (2014)[9]のテキストで、p値に基づく方法を一般化し紹介されている。

(1) Bauer and Kohne (1994)[2]の方法

帰無仮説下において、p値は$[0,1]$の一様分布に従うため、$-2\log p$は、自由度 2 のカイ二乗分布に従う。よって、

$$-2\log p_1 p_2 = (-2\log p_1) + (-2\log p_2) \sim \chi_4^2$$

となり、棄却域c_pを以下のように設定できる。

$$p_1 p_2 < c_p = e^{-\frac{1}{2}\chi_{4,1-\alpha}^2}$$

すなわち、$p_2 < c_p/p_1$であるときに帰無仮説を棄却する。

このとき、帰無仮説下で第 2 ステージの帰無仮説が棄却される確率は、

$$\Pr\left(\alpha_1 \le p_1 \le \alpha_0, p_2 < c_p/p_1\right) = \int_{\alpha_1}^{\alpha_0}\int_0^{c_p/p_1} dp_2 dp_1 = c_p \log(\alpha_0/\alpha_1)$$

となり、試験全体の第一種の過誤確率 αを用いて、以下の式を満たす範囲で、α_0、α_1を設定する。

$$\alpha = \alpha_1 + c_p \log(\alpha_0/\alpha_1)$$

(2) Cui, Hung and Wang (1999)[3]、Lehmacher and Wassmer (1999) [4]の方法

Cui, Hung and Wang (1999) [3]は、以下の各ステージの Z 検定統計量を情報時間で重み付けし統合した方法を提案し、症例数再推定の決定ルールにかかわらず、帰無仮説下で第一種の過誤確率が制御されることを示した。

$$Z_{CHW} = \sqrt{\frac{n_1}{N_{initial}}} Z_1 + \sqrt{\frac{n_2}{N_{initial}}} Z_2$$

よって、各ステージのp値は以下のようになる。

$$\begin{cases} 1-\Phi(Z_1) & : \text{第 1 ステージ} \\ 1-\Phi(Z_{CHW}) & : \text{第 2 ステージ} \end{cases}$$

ここで、再推定後の症例数$\hat{n}_2$、N_{final}ではなく元の症例数n_2、$N_{initial}$を用いている点に注意が必要である。また、この方法は中間解析で症例数の増加が発生しない場合、固定標本デザインと同じナイーブな検定統計量Z_{naive}と一致する点が魅力の一つである。

同年に提案された、Lehmacher and Wassmer (1999) [4]は、各ステージのp値を逆正規法に基づき統合することにより、正規分布に従うエンドポイントのみでなく二値や生存時間などあらゆるエンドポイントに利用可能な一般化された方法となっている。

$$Z_{LW} = w_1\Phi^{-1}(1-p_1) + \mathrm{w}_2\Phi^{-1}(1-p_2) = w_1 Z_1 + w_2 Z_2$$

ここで、重みw_1、w_2は事前に規定され、$w_1^2 + w_2^2 = 1$である。Lehmacher and Wassmer(1999)[4]では、各ステージの重みを同等とした$w_1 = w_2 = \frac{1}{\sqrt{2}}$を提案した。

よって、各ステージのp値は以下のようになる。

$$\begin{cases} 1 - \Phi(Z_1) & : \text{第 1 ステージ} \\ 1 - \Phi(Z_{LW}) & : \text{第 2 ステージ} \end{cases}$$

これらの方法は、有効・無益性中止を伴う中間解析を実施した場合でも、上記のp値を利用して、Pocock や O'Brien and Fleming 等の群逐次デザインで導出された中止境界α_0、α_1と最終解析の棄却域c_pをそのまま利用可能である。

(3) p値に基づく方法の一般化

本項では、Chang (2006)[5]、Chang (2014)[9]のテキストによって概説されたp値に基づく方法の一般化について紹介する。各ステージの部分標本によって得られたp値の組み合わせを利用して、検定統計量、有効・無益性中止の境界、調整済みp値を導出する。

有効・無益性中止を伴う中間解析を実施することとし、第 1、第 2 ステージでp値に基づく検定統計量T_1、T_2が得られたときの中止基準を以下のように設定し、最終解析で$T_2 < c_p$の場合に帰無仮説が棄却されるとする。

$$\begin{cases} \text{有効中止:} if \;\; T_1 < \alpha_1 \\ \text{無益性中止:} if \;\; T_1 > \alpha_0 \\ \text{アダプテーション後試験継続:} if \;\; \alpha_1 \le T_1 \le \alpha_0 \end{cases}$$

帰無仮説下で第 1 ステージの帰無仮説が棄却される確率を$\psi_1(\alpha_1)$、第 2 ステージの帰無仮説が棄却される確率を$\psi_2(c_p)$とする。$\psi_1(\alpha_1)$はT_1の確率密度関数f_{T_1}を用いて、以下のように与えられる。

$$\psi_1(\alpha_1) = \Pr(T_1 < \alpha_1) = \int_0^{\alpha_1} f_{T_1} dt_1 = \alpha_1$$

また、第 2 ステージに到達するためには、第 1 ステージで試験継続が判断される必要があり、$\psi_2(c_p)$は、T_1、T_2の同時密度関数f_{T_1,T_2}を用いて、以下のように与えられる。

$$\begin{aligned} \psi_2(c_p) &= \Pr(\alpha_1 \le T_1 \le \alpha_0, \; T_2 < c_p) \\ &= \int_{\alpha_1}^{\alpha_0} \int_0^{c_p} f_{T_1,T_2} dt_2 dt_1 \end{aligned}$$

よって、試験全体の第一種の過誤確率αは以下のようになる。

$$\alpha = \alpha_1 + \psi_2(c_p)$$

また、第 1、2 ステージの帰無仮説であるH_{01}とH_{02}の積仮説$H_0: H_{01} \cap H_{02}$に対するp値である調整済みp値は、各ステージの検定統計量$T_1 = t_1$、$T_2 = t_2$を用いて以下のように算出される。

$$\begin{cases} t_1 & : \text{第 1 ステージ} \\ \alpha_1 + \psi_2(t_2) & : \text{第 2 ステージ} \end{cases}$$

第 2 ステージまで到達した場合、$t_2 < c_p$すなわち$\alpha_1 + \psi_2(t_2) < \alpha$の場合に帰無仮説は棄却される。

本項では、各ステージのp値を統合する以下の 3 種の方法について紹介する。

① 個々のp値に基づく方法(method based on individual p-values; MIP)

② p値の和に基づく方法(method based on the sum of p-values; MSP)

③ p値の積に基づく方法(method based on the product of p-values; MPP)

① 個々のp値に基づく方法 (MIP)

各ステージの個々のp値をそのまま検定統計量として利用する。

$$T_1 = p_1$$

$$T_2 = p_2$$

T_1、T_2は独立なので、帰無仮説のもとでは、$f_{T_1,T_2} = 1$となり、各ステージにおけるαの消費はそれぞれ、

$$\psi_1(\alpha_1) = \Pr(T_1 < \alpha_1) = \int_0^{\alpha_1} dt_1 = \alpha_1$$

$$\psi_2(c_p) = \Pr(\alpha_1 \le T_1 \le \alpha_0, T_2 < c_p) = \int_{\alpha_1}^{\alpha_0} \int_0^{c_p} dt_2 dt_1 = c_p(\alpha_0 - \alpha_1)$$

よって、試験全体の第一種の過誤確率αは以下のようになり、下式を満たす範囲でα_0、α_1、c_pを設定する。

$$\alpha = \alpha_1 + c_p(\alpha_0 - \alpha_1)$$

また、各ステージの調整済みp値は以下のようになる。

$$\begin{cases} T_1 & : \text{第 1 ステージ} \\ \alpha_1 + (\alpha_0 - \alpha_1)T_2 & : \text{第 2 ステージ} \end{cases}$$

MIP は、非常にシンプルであり、後述する MSP、MPP などの方法を比較するための「基準」として有用である。この方法は、異なるステージからの統合されたデータを使用しない点で特徴的である。

② p値の和に基づく方法(MSP)

Chang (2006)[5]は、各ステージのp値の和を検定統計量として利用する方法を提案した。

$$T_1 = p_1$$

$$T_2 = p_1 + p_2$$

$T_1 = p_1 > c_p$のとき、$T_2 = p_1 + p_2 > T_1 = p_1 > c_p$となるので、$T_1 = p_1 > c_p$が無益性中止の境界となる。

よって、各ステージにおけるαの消費はそれぞれ、

$$\psi_1(\alpha_1) = \Pr(T_1 < \alpha_1) = \int_0^{\alpha_1} dt_1 = \alpha_1$$

$$\psi_2(c_p) = \Pr(\alpha_1 \le T_1 \le \alpha_0, T_2 < c_p) = \begin{cases} \int_{\alpha_1}^{\alpha_0} \int_{t_1}^{c_p} dt_2 dt_1 : \alpha_0 < c_p \\ \int_{\alpha_1}^{c_p} \int_{t_1}^{c_p} dt_2 dt_1 : \alpha_0 \ge c_p \end{cases}$$

積分計算を行うと、全体の第一種の過誤確率αは以下のようになる。下式を満たす範囲で、α_0、α_1、c_pを設定する。

$$\alpha = \begin{cases} \alpha_1 + c_p(\alpha_0 - \alpha_1) - \frac{1}{2}(\alpha_0^2 - \alpha_1^2) : \alpha_0 < c_p \\ \alpha_1 + \frac{1}{2}(c_p - \alpha_1)^2 : \alpha_0 \ge c_p \end{cases}$$

また、各ステージの調整済みp値は以下のようになる。

$$\begin{cases} T_1 : \text{第 1 ステージ} \\ \alpha_1 + T_2(\alpha_0 - \alpha_1) - \frac{1}{2}(\alpha_0^2 - \alpha_1^2) : \text{第 2 ステージ} \quad and \quad \alpha_0 < c_p \\ \alpha_1 + \frac{1}{2}(T_2 - \alpha_1)^2 : \text{第 2 ステージ} \quad and \quad \alpha_0 \ge c_p \end{cases}$$

③ p値の積に基づく方法(MPP)

検定統計量を各ステージのp値の積として利用する。

$$T_1 = p_1$$
$$T_2 = p_1 p_2$$

$p_1 < c_p$のとき、$p_1 p_2 < p_1 < c_p$となり、最終解析で必ず帰無仮説は棄却されるため、試験を継続しても意味がない、そのため、$c_p > \alpha_1$と設定することは望ましくなく、$\alpha_0 > \alpha_1 > c_p$と設定する。

各ステージにおけるα消費はそれぞれ、

$$\psi_1(\alpha_1) = \Pr(T_1 < \alpha_1) = \int_0^{\alpha_1} dt_1 = \alpha_1$$

$$\psi_2(c_p) = \Pr(\alpha_1 \le T_1 \le \alpha_0, T_2 < c_p) = \int_{\alpha_1}^{\alpha_0} \int_0^{c_p} \frac{1}{T_1} dt_2 dt_1$$

積分計算を行うと、全体の第一種の過誤確率αは以下のようになる。下式を満たす範囲で、α_0、α_1、c_pを設定する。

$$\alpha = \alpha_1 + c_p \log\left(\frac{\alpha_0}{\alpha_1}\right)$$

Bauer and Kohne (1994) [2]の方法は、上式の$c_p = e^{-\frac{1}{2}\chi^2_{4,1-\alpha}}$とした特別な場合である。

また、各ステージの調整済みp値は以下のようになる。

$$\begin{cases} T_1 & : \text{第 1 ステージ} \\ \alpha_1 + T_2 \log\left(\frac{\alpha_0}{\alpha_1}\right) & : \text{第 2 ステージ} \end{cases}$$

計算例

ここでは、Chang (2014)[9]で紹介された仮想的事例に基づいて、p値の和に基づく方法(MSP)における計算例を紹介する。

喘息患者を対象とした、実薬群及びプラセボ群の 2 群を設定した第 III 相試験を想定する。主要評価項目は、FEV1 のベースラインからの変化率である。ベースラインからの変化率はプラセボ群、実薬群でそれぞれ 5%、12%と想定され、標準偏差は 22%と想定した。従って、1 群あたり 208 例であれば、片側$\alpha = 0.025$で 90%の検出力が期待される。50%の被験者が集積された時点で MSP を使用した中間解析を計画している。中止基準を$\alpha_1 = 0.01$、 $\alpha_0 = 0.15$と設定すると、下記表より$c_p = 0.1871$である。試験の結果、第 1 ステージで$p_1 = 0.012$が得られ（$\alpha_1 \le p_1 \le \alpha_0$のため試験を継続した）、第 2 ステージで$p_2 = 0.18$が得られたとする。従って、MSP に基づく検定統計量は、$0.012 + 0.18 = 0.192 > c_p = 0.1871$であり、帰無仮説は棄却されなかった。また、試験全体の積仮説に対する調整済み p 値は、$\alpha_1 + t_2(\alpha_0 - \alpha_1) - \frac{1}{2}(\alpha_0^2 - \alpha_1^2) = 0.01 + 0.192 \times (0.15 - 0.01) - \frac{1}{2}(0.15^2 - 0.01^2) = 0.02568 > \alpha = 0.025$となり、同様の結論が導かれる。

表 9　MSP におけるα_0、α_1、c_p

α_0	α_1	0.0025	0.005	0.010	0.015	0.020
0.05		0.4999	0.4719	0.4050	0.3182	0.2017
0.10		0.2820	0.2630	0.2217	0.1751	0.1225
0.15	c_p	0.2288	0.2154	0.1871	0.1566	0.1200
0.20		0.2152	0.2051	0.1832	0.1564	0.1200
0.25		0.2146	0.2050	0.1832	0.1564	0.1200

Note: 片側 $\alpha = 0.025$

2.2.2.1.2　条件付き過誤関数 (Conditional Error Function；CEF) による方法

Proschan and Hunsberger (1995)[6]は、条件付き過誤関数を利用し、第一種の過誤確率を制御する方法を提案した。第 1 ステージで得られた検定統計量（又はp値）に基づき、第一種の過誤確率の条件付き確率を計算し、最終解析の検定の棄却域c_{CEF}を調整する方法である。

$A(p_1)$をp_1のもとで、第 1 ステージで試験継続が判断された場合の第 2 ステージの第一種の過誤確率の条件付き確率とすると、p_1は帰無仮説下で[0, 1]の一様分布に従うため、試験全体の第一種の過誤確率αは以下のようになる。

$$\alpha = \alpha_1 + \int_{\alpha_1}^{\alpha_0} A(p_1)dp_1$$

$A(p_1)$は、pスケールの条件付き過誤関数と呼ばれ、$0 \le A(p_1) \le 1$の任意の減少関数（Zスケールの場合は増加関数）である。Proschan and Hunsberger (1995)[6]は、以下の関数を提案した。

$$A(p_1) = 1 - \Phi\left(\sqrt{[\Phi^{-1}(1-\alpha_1)]^2 - [\Phi^{-1}(1-p_1)]^2}\right), \quad \alpha_1 \le p_1 \le \alpha_0$$

最終解析でナイーブな検定統計量Z_{naive}を用いる場合、症例数再推定後の第一種の過誤確率の条件付き確率は以下のようになる。

$$\Pr(Z_{naive} > c_{CEF} | H_0) = 1 - \Phi\left(\frac{c_{CEF}\sqrt{n_1 + \hat{n}_2} - Z_1\sqrt{n_1}}{\sqrt{\hat{n}_2}}\right)$$

ここで、$\Pr(Z_{naive} > c_{CEF} | H_0) = A(p_1)$なので、最終解析の棄却域$c_{CEF}$は以下のようになる。

$$c_{CEF} = \frac{\sqrt{n_1}Z_1 + \sqrt{\hat{n}_2}\Phi^{-1}\left(1 - A(p_1)\right)}{\sqrt{n_1 + \hat{n}_2}}$$

異なる条件付き過誤関数を選択することは、各ステージからのデータに異なる重みを用いることを意味しており、統合検定による方法は、それぞれ対応する条件付き過誤関数による方法に変換することができる。例えば、Bauer and Kohne (1994)[2]の方法であれば、

$c_p = e^{-\frac{1}{2}\chi^2_{4,1-\alpha}}$としたとき、$A(p_1) = \frac{c_p}{p_1}$となる。すなわち、

$$Z_{naive} = \frac{\sqrt{n_1}}{\sqrt{n_1 + \hat{n}_2}}Z_1 + \frac{\sqrt{\hat{n}_2}}{\sqrt{n_1 + \hat{n}_2}}Z_2 > \frac{\sqrt{n_1}Z_1 + \sqrt{\hat{n}_2}\Phi^{-1}\left(1 - \left(\frac{c_p}{p_1}\right)\right)}{\sqrt{n_1 + \hat{n}_2}} \Leftrightarrow p_1 p_2 < c_p$$

となることから、Bauer and Kohne (1994)[2]の方法と一致することがわかる。

Cui, Hung and Wang (1999)[3]の方法であれば、対応する$A(p_1)$は以下のようになる。

$$A(p_1) = 1 - \Phi\left(\frac{c_Z - Z_1\sqrt{n_1/N_{initial}}}{\sqrt{n_2/N_{initial}}}\right)$$

すなわち、

$$Z_{naive} = \frac{\sqrt{n_1}}{\sqrt{n_1 + \hat{n}_2}}Z_1 + \frac{\sqrt{\hat{n}_2}}{\sqrt{n_1 + \hat{n}_2}}Z_2 > \frac{\sqrt{n_1}Z_1 + \sqrt{\hat{n}_2}\left(\frac{c_Z - Z_1\sqrt{n_1/N_{initial}}}{\sqrt{n_2/N_{initial}}}\right)}{\sqrt{n_1 + \hat{n}_2}}$$

$$\Leftrightarrow \sqrt{\frac{n_2}{N_{initial}}}Z_1 + \sqrt{\frac{n_1}{N_{initial}}}Z_2 > c_Z \Leftrightarrow Z_{CHW} > c_Z$$

となることから、Cui, Hung and Wang (1999) [3]の方法と一致することがわかる。

2.2.2.1.3 第一種の過誤確率が増大しない範囲で症例数の変更を許容する方法

これまで紹介してきた、統合検定による方法や条件付き過誤関数による方法は、各ステージの部分標本に基づき解析され、症例数再推定前に登録された被験者と症例数再推定後に登録された被験者の結果を別々に取り扱い、特殊な統計手法を必要とした。また、各ステージの被験者が異なる重み付けがされ解析されることが、すべての被験者は平等であるという“one patient one vote”の原則に反するため、望ましくないという主張もある（Hung et al. (2014)[13]）。従って、アダプテーションの有無にかかわらず、固定標本デザインと同様のナイーブな検定手法($Z_{naive} > c_Z$)を用いることは、依然として魅力的である。そのため、アダプテーションのルールに制約を設定し、第一種の過誤確率が増大しない範囲で症例数の変更を許容することで、従来の検定統計量と棄却域は変更せずに第一種の過誤確率を制御することを目的とした方法が提案された。Chen (2004)[7] は、条件付き検出力CPが 0.5 以上の場合にのみ症例数を増加させると、ナイーブな検定手法を用いても、第一種の過誤確率は増大しないことを示した。Gao (2008) [14]は、CPが 0.5 を下回る場合にも拡張した。Mehta and Pocock (2011)[8]は、Gao (2008) [14]の方法を 2 ステージデザインの枠組みで整理し、ナイーブな検定手法を用いても第一種の過誤確率が増大しない条件付き検出力の領域内で症例数を増加させる方法を提案し、これらの明示的なカットオフ値を示すことで、実務担当者がより利用しやすくした。本項では、Mehta and Pocock (2011) [8]ついて紹介する。

Liu (2021) [10] によれば、非盲検下の症例数再推定は、”Five-zone” アダプティブデザインと呼ばれることがある。”Five-zone”は 5 種類の中間解析の結果のシナリオを表しており、一般的に、有効・無益性中止の基準や条件付き検出力によって判断される。

Efficacy zone（有効領域）

有効中止の基準を満たした場合、試験は早期に有効中止される。

Futility zone（無益領域）

無益性中止の基準を満たした場合、無益であるとして試験は早期に無益性中止される。

Favorable zone（良好領域）

目標の条件付き検出力を超えている場合、試験は計画通りの症例数で継続される。

Promising zone（有望領域）

目標の条件付き検出力には到達していない、しかし、結果が有望な場合、症例数の上限を決め、試験の成功確率を向上させるために症例数を増やす。

Unfavorable zone（不利領域）

目標の条件付き検出力には到達してない、かつ、有望ではないが無益でもない場合、試験は当初の計画された症例数を維持し継続する。

通常、症例数再推定は、上限N_{max}又は$n_2^{max}(= N_{max} - n_1)$を設定する。ここで、症例数を増加させる状況は、目標の条件付き検出力に対して十分ではない状態であり、目標検出力を達成するために、条件付き検出力が目標とする検出力を満たす$\widehat{N}'_{final}$又は$\hat{n}_2'$に基づいて、$\widehat{N}_{final} = \min(\widehat{N}'_{final}, N_{max})$又は$\hat{n}_2 = \min(\hat{n}_2', n_2^{max})$の制限下で決定される。

本項では、以降、簡単のため、有効・無益性中止を伴う中間解析を実施せず、最終解析で$Z_{naive} > z_\alpha$を用いる場合を想定して議論する（すなわち，Efficacy zone 及び Futility zone を設定しない)。有効・無益性中止を伴う中間解析を設定した場合でも、Pocock や O'Brien and Fleming 等の群逐次デザインで導出された中止境界をそのまま利用可能であり、以降紹介する式のz_αをc_zに置き換えればよい。

Mehta and Pocock (2011) [8]では、条件付き検出力の大きさに基づき、Favorable zone、Promising zone、Unfavorable zone を以下のように設定した。

Favorable

$CP(Z_1, n_2) \geq 1 - \beta$で規定される。$1 - \beta$は、目標とする条件付き検出力であり、試験計画時の症例数設計で用いた検出力が利用されることが多い。条件付き検出力を満たしているため、症例数は変更しない。

Promising

$CP_{min} \leq CP(Z_1, n_2) < 1 - \beta$で規定される。目標とする条件付き検出力に到達していないため、症例数を増やす。ここで、CP_{min}はナイーブな検定手法を用いた際に第一種の過誤確率を制御するために、小さ過ぎてはいけない。その要件を満たす最小のCP_{min}は、$n_{max}/N_{initial}$、$n_1/N_{initial}$及び$1 - \beta$に依存する。症例数の増加をこの範囲のサブセットに制限することは自由であるが、その場合、試験全体の検出力が多少低下する。

Unfavorable

$CP(Z_1, n_2) < CP_{min}$で規定される。目標とする条件付き検出力に到達していないが、有望ではないため、症例数は変更せず、試験を継続する。

なお、本項では、条件付き検出力は以下の中間解析時の治療効果の推定値を用いる方法を利用する。条件付き検出力については，2.2.2.2.2 項で詳しく紹介する

$$CP(Z_1, n_2) = \Phi\left(\frac{Z_1\sqrt{n_2}}{\sqrt{n_1}} - \frac{z_\alpha\sqrt{n_1 + n_2} - Z_1\sqrt{n_1}}{\sqrt{n_2}}\right)$$

Promising zone($CP_{min} \leq CP(Z_1, n_2) < 1 - \beta$)の範囲でナイーブ検定手法を用いても第一種の過誤確率が制御される妥当性

Gao (2008)[14]は、以下に示す$b(Z_1,\hat{n}_2)$に基づき、$Pr(Z_{naive} > b(Z_1,\hat{n}_2)|H_0) = \alpha$と棄却域を設定すれば、第一種の過誤確率が制御されることを示した。

$$b(Z_1,\hat{n}_2) = (n_1+\hat{n}_2)^{-0.5}\left[\sqrt{\frac{\hat{n}_2}{n_2}}\ \left(z_\alpha\sqrt{N_{initial}} - Z_1\sqrt{n_1}\right) + Z_1\sqrt{n_1}\right]$$

これは、棄却域を$b(Z_1,\hat{n}_2)$に設定すれば、ナイーブな検定統計量Z_{naive}がそのまま利用可能であることを意味している。しかしながら、シンプルでわかりやすい最終解析という理念に基づき、最終解析では、棄却域はz_αを用いたい。これを可能にするために、$b(Z_1,\hat{n}_2) \le z_\alpha$となる範囲として、

$$\mathcal{D} = \{CP(Z_1,\hat{n}_2): b(Z_1,\hat{n}_2) \le z_\alpha\}$$

を設定する。$\mathcal{D}$の範囲では以下の式が成立するため、ナイーブな検定手法（$Z_{naive} > z_\alpha$）が利用可能である。

$$\alpha = Pr(Z_{naive} > b(Z_1,\hat{n}_2)|H_0) \ge Pr(Z_{naive} > z_\alpha|H_0)$$

CP_{min}の探索方法

本項では、仮想的事例に基づき、CP_{min}の算出方法を紹介する。

仮想的事例として、試験薬群とプラセボ群のランダム化第 III 相試験を想定する。有効性の主要評価項目は連続変数である。1 群 221 例で、各投与群の評価項目の標準偏差を 7.5 と仮定し、群間差 2.0 を 80%の検出力で検出するために計画されたとする。臨床的に群間差 1.6 という低い値でも意味があると考えられるが、その場合 1 群 345 例が必要である。そのため、1 群 104 例が観測された段階で、中間解析を実施し、症例数再推定を実施することとした。なお、最大症例数は 1 群 442 例と設定した。CP_{min}は、以下のように探索される。

① 任意の$CP(Z_1) = \Phi\left(\frac{Z_1\sqrt{117}}{\sqrt{104}} - \frac{z_\alpha\sqrt{221}-Z_1\sqrt{104}}{\sqrt{117}}\right) \in (0,1)$に対応する$Z_1$を導出する。

② ①で導出したに各Z_1対応する$\widehat{N}_{final}$を$\hat{n}_2'(Z_1) = \left[\frac{104}{Z_1^2}\right]\left[\frac{z_\alpha\sqrt{221}-Z_1\sqrt{108}}{\sqrt{117}} + z_\beta\right]^2$と$\widehat{N}_{final} = \min(n_2'(Z_1)+104, N_{max}=442)$から導出する。

③ ①及び②を用いて、各$CP(Z_1)$に対応する$b(Z_1,\hat{n}_2)$を

$b(Z_1,\hat{n}_2) = (104+\hat{n}_2)^{-0.5}\left[\sqrt{\frac{\hat{n}_2}{117}}\ \left(z_\alpha\sqrt{221} - Z_1\sqrt{104}\right) + Z_1\sqrt{104}\right]$から導出する。

図 2 は、①～③の計算結果であり、$CP(Z_1) = 0.3603$を境界に、$b(Z_1,\hat{n}_2) \le 1.96$となっており、$0.3603 \le CP(Z_1) < 0.80$の範囲で、症例数の増加が行われていることがわかる。

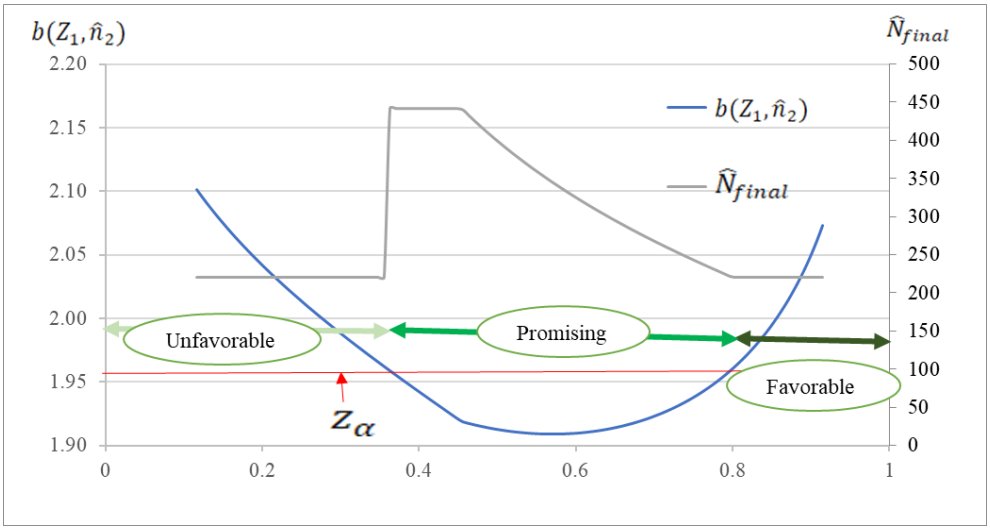


図 2　$CP(Z_1)$、$b(Z_1, \hat{n}_2)$、　$\hat{N}_{final}$の関係性

2.2.2.2　　症例数再推定の方法

2.2.2.2.1　　症例数の公式に基づく方法

Chen (2004) [7]や Cui, Hung and Wang (1999) [3]は、以下の症例数の公式に基づく方法で議論した。

これは、真の治療効果が、中間解析で観察されたものと全く同じであると仮定している。また、第 1 ステージの検定統計量の最終解析に対する検出力への影響を考慮していない。

$$\hat{N}'_{final} = 2\sigma_{pre}^2 \left[\frac{z_\alpha + z_\beta}{\hat{\delta}_1}\right]^2 = N_{initial}\left[\frac{\delta_{pre}}{\hat{\delta}_1}\right]^2$$

2.2.2.2.2　　条件付き検出力に基づく方法

条件付き検出力は、中間解析で検定統計量Z_1が与えられた場合の最終解析で帰無仮説が棄却される確率として定義される。

最終解析で用いる任意の検定統計量を以下のように設定したとする。

$$Z_{final} > c_Z \Leftrightarrow Z_2 > B(c_Z, Z_1)$$

このとき、条件付き検出力CPは以下のようになる。

$$CP(\delta, Z_1, n_2) = \Pr(Z_2 > B(c_Z, Z_1)) = \Phi\left(\frac{\delta\sqrt{n_2}}{\sqrt{2}\sigma} - B(c_Z, Z_1)\right)$$

例えば、Bauer and Kohne (1994)[2]の方法では、$p_1 p_2 < c_p$、$Z_1 = \Phi^{-1}(1 - p_1)$から、以下のようになる。

$$Z_2 > \Phi^{-1}\left(1 - \frac{c_p}{p_1}\right) = B(c_p, Z_1)$$

$Z_{naive} > c_Z$を用いた場合の条件付き検出力は、

$$\frac{\sqrt{n_1}}{\sqrt{n_1+n_2}}Z_1+\frac{\sqrt{n_2}}{\sqrt{n_1+n_2}}Z_2>c_Z \Leftrightarrow Z_2>\frac{c_Z\sqrt{n_1+n_2}-Z_1\sqrt{n_1}}{\sqrt{n_2}}=B(c_Z,z_1)$$

であるため、条件付き検出力は以下のようになる。

$$CP(\delta,Z_1,n_2)=\Pr(Z_{naive}>c_Z)=\Phi\left(\frac{\delta\sqrt{n_2}}{\sqrt{2}\sigma}-\frac{c_z\sqrt{n_1+n_2}-Z_1\sqrt{n_1}}{\sqrt{n_2}}\right)$$

上記に示される通り、CPは真の治療効果δの関数であるが、未知数であるため、ここでは 2 つの選択肢がある。1 つめは当初の症例数設計の仮定に用いたδ_{pre}を用いること。2 つめは、中間解析の時の治療効果の推定値$\hat{\delta}_1$を用いることである。σが既知のもと、すなわち、$Z_1=\frac{\sqrt{n_1}\hat{\delta}_1}{2\sigma}$のもとで、それぞれ以下のようになる。

$$CP\left(\delta_{pre},Z_1,n_2\right)=\Pr(Z_{naive}>c_Z)=\Phi\left(\frac{\delta_{pre}\sqrt{n_2}}{\sqrt{2}\sigma}-\frac{c_Z\sqrt{n_1+n_2}-Z_1\sqrt{n_1}}{\sqrt{n_2}}\right)$$

$$CP(Z_1,n_2)=\Pr(Z_{naive}>c_Z)=\Phi\left(\frac{Z_1\sqrt{n_2}}{\sqrt{n_1}}-\frac{c_Z\sqrt{n_1+n_2}-Z_1\sqrt{n_1}}{\sqrt{n_2}}\right)$$

Mehta and Pocock (2011)[8]は、$\delta=\hat{\delta}_1$を用い、$c_Z=z_\alpha$及び条件付き検出力の目標値を$1-\beta$としたときに、$CP(Z_1,\hat{n}_2')=Pr(Z_{naive}>z_\alpha)=1-\beta$を満たす、$\hat{n}_2'$を導出するために以下の式を提案した。

$$\hat{n}_2'(Z_1)=\left[\frac{n_1}{Z_1^2}\right]\left[\frac{z_\alpha\sqrt{n_1+n_2}-Z_1\sqrt{n_1}}{\sqrt{n_2}}+z_\beta\right]^2$$

なお、Liu (2021) [10]が指摘するとおり、上式は、実際には、$Pr(Z_{naive}>z_\alpha)=1-\beta$ではなく、$Pr(\sqrt{n_1/N_{initial}}Z_1+\sqrt{n_2/N_{initial}}Z_2>z_\alpha)=1-\beta$を満たす式であるが、その影響は一般的に非常に小さい。

治療効果の推定値$\hat{\delta}_1$を利用する欠点は、条件付き検出力の算出に際して中間データである$\hat{\delta}_1$及びZ_1を用いるため，中間データが二度使用される点にある。従って、条件付き検出力が不安定になりやすく、特にn_1/n_2が小さい場合に影響が大きい。Bauer and Koenig (2006)[11]は、$\hat{\delta}_1$及びδ_{pre}使用した場合の条件付き検出力の動作特性について詳細に議論した。また、Glimm (2012) [12]は$\hat{\delta}_1$が確率変数であるため、二度使用することは、望ましくない試験変更の可能性があることを懸念し、$\hat{\delta}_1$を用いた条件付き検出力を利用する際は、デザインの動作特性の慎重に調査することを推奨した。

2.2.2.2.3 最適化問題として取り扱い症例数再推定を行う方法

2.2.2.2 項では、症例数再推定の方法について紹介した。これらは、条件付き検出力のみに基づいて再推定されている。しかし、スポンサーの立場からすると、統計的検出力が必ずしも規制上あるいは商業上の成功につながるとは限らない。現実的には、症例数の増加による投資と条件付き検出力の増加はトレードオフの関係にある。Jennison and Turnbull (2015)[15] は、検出力を高めるために最も効果的なときに症例数の増加を行うために、症例

数の増加と条件付き検出力の増加のトレードオフの程度を表すパラメータγを含む以下の目的関数を最大化する症例数再推定のデザインを提案した。

$$\Phi\left(\frac{\delta_{pre}\sqrt{\hat{n}_2}}{\sqrt{2}\sigma}-\frac{z_\alpha\sqrt{n_1+\hat{n}_2}-Z_1\sqrt{n_1}}{\sqrt{\hat{n}_2}}\right)-\gamma(2\hat{n}_2-2n_2)$$

Jennison and Turnbull (2015) [15]は、Chen (2004)[7]や Mehta and Pocock (2011)[8]らの第一種の過誤確率が増大しない範囲で症例数の変更を許容する方法は、症例数の増加させることで最大の利益が得られるかもしれない状況で症例数の増加がなされない特徴があることから、これらの方法を、分が悪いとし、Lehmacher and Wassmer (1999) [4]の統合検定と上記の目的関数を最大化する方法を組み合わせた方法が効率的であると述べた。Hsiao, Liu and Mehta (2019) [16]は、症例数の増加は、最小の許容可能な条件付き検出力に依存する、というもっともらしい概念に基づいて、制約条件を設定し、これを Constrained promising zone design (CPZ)と呼んだ。本項では、Hsiao, Liu and Mehta (2019) [16]について紹介する。

Hsiao, Liu and Mehta (2019) [16]は、スポンサーが、症例数の増加することを承認する前に、中間解析時にある特定のマイルストーンを達成していることを望んでいるとして、以下のような目的関数と制約条件からなる制約付き最適化問題として整理した。

目的関数：Maximize$\{CP_{\delta_{min}}(Z_1,\widehat{N}_{final})\}$

制約条件 1：$N_{initial}\le\widehat{N}_{final}\le N_{max}$

制約条件 2：$CP_{\delta_{min}}(Z_1,\hat{n})\ge CP_{min}$

制約条件 3：$CP_{\delta_{min}}(Z_1,\hat{n})\le CP_{max}$

N_{max}：最大症例数

δ_{min}：臨床的に意味のある最小の治療効果

CP_{min}：Promising zone 内の条件付き検出力が満たすべき最低要件。もし、$\delta=\delta_{min}$、$\widehat{N}_{final}=N_{max}$としたときの条件付き検出力が$CP_{min}$に達していない場合、その中間解析結果は“Unfavorable zone”にあるとみなし、症例数は変更しない。

CP_{max}：Promising zone 内の条件付き検出力の最大値。$\delta=\delta_{min}$、$\widehat{N}_{final}\le N_{max}$としたときの条件付き検出力が$CP_{max}$に達した場合、症例数を$\widehat{N}_{final}$に設定する。

CPZ design では、Mehta and Pocock (2011) [8] と比較し、以下のような特徴がある。

- 条件付き検出力を算出する際に、中間解析時の治療効果の推定値ではなく、臨床的に意味のある最小の治療効果を用いて算出する。
- スポンサーが任意で予め規定した条件付き検出力の範囲に到達する場合に、症例数を増加させる。
- 統合検定や条件付き過誤関数等を用いて第一種の過誤確率を制御する必要がある。

なお、ここでは、第一種の過誤確率を制御するために、Cui, Hung and Wang (1999) [3]の方法よる調整を行うこととし、調整を考慮した以下の条件付き検出力を用いる。

$$CP(\delta_{min}, Z_1, \hat{n}_2) = \Phi\left(\frac{\delta_{min}\sqrt{\hat{n}_2}}{\sqrt{2}\sigma} - \frac{z_\alpha\sqrt{n_1+n_2} - Z_1\sqrt{n_1}}{\sqrt{n_2}}\right)$$

CPZ デザインの特徴は、条件付検出力の計算にδ_{min}を用いることにより、臨床的に意味のある最小の治療効果を示す最小のしきい値を超えた場合にのみ、症例数が追加されるため、スポンサーにとって魅力的である。また、CP_{min}、CP_{max}は、スポンサーによって、柔軟に設定される。

計算例

2.2.2.1.3 項で設定した仮想的事例を元に、$CP_{max} = 0.8$、$CP_{min} = 0.7$、$\delta_{min} = 1.6$と設定し、Mehta and Pocock (2011)[8]：MP 法、Jennison and Turnbull (2015)[15]：JT 法及び Hsiao, Liu and Mehta (2019)[16]：CPZ 法、それぞれについて、$\hat{N}_{final}$及び症例数再推定後の条件付き検出力CPを算出した。なお、JT 法では $\gamma = 0.25/4\sigma^2$と設定したが、これは MP 法における真の治療効果$\delta = 1.6$のときの検出力 0.658 を JT 法でも満たすためのパラメータである。

MP 法は、$\delta = \hat{\delta}_1$に基づき条件付き検出力を算出しており、JT 法及び CPZ 法は、$\delta = \delta_{min}$に基づき条件付き検出力を算出しているため、それぞれ、収束先の条件付き検出力が異なることに注意したい。Jennison and Turnbull (2015)[15]の述べるとおり、JT 法は MP 法で症例数が増加されない範囲（MP 法での Unfavorable zone）でも大きく症例数を増加させ、逆に MP 法での Promising zone においては、症例数の増加の程度は小さくなっている。CPZ 法は MP 法よりZ_1が小さいときでも、症例数の増加がされていた。これは、$\delta = \delta_{min}$を用いているため、Z_1が小さいときでも、中間解析の治療効果の推定値を用いた場合に比べ、条件付き検出力が大きくなることや、MP 法での Promising zone における制約がないことなどが原因だと考えられる。

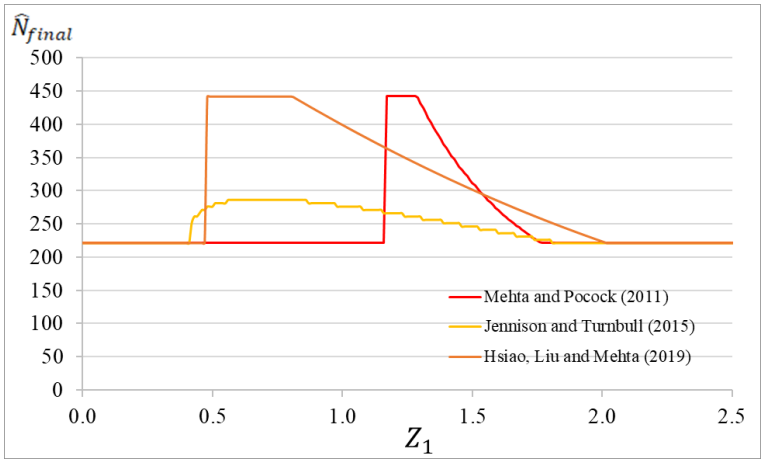


図 3　Z_1と$\hat{N}_{final}$の関係

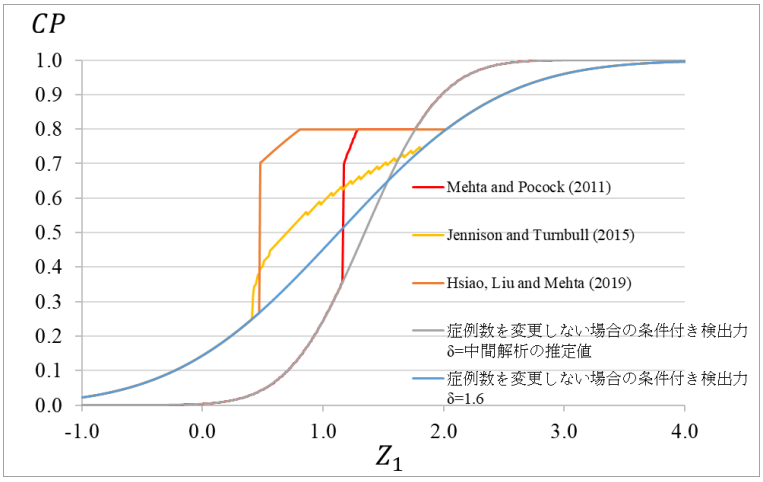


図 4　Z_1と症例数再推定後の条件付き検出力の関係

2.2.2.3　　事例紹介

Bauer and Kohne (1994)[2]の方法の適用例 Diener (2014)[17]

開腹手術後の術後感染症の発生を抑えるために開発されたトリクロサンコーティングの縫合糸に関する臨床試験（PROUD 試験）の事例を紹介する。被験者は、トリクロサンコーディングされた縫合糸（PDS II 群）又はコーディングされていない縫合糸（PDS Plus 群）に 1:1 の比でランダム化された。主要評価項目は、術後 30 日以内のアメリカ疾病予防管理センターの定義に基づく表在性または深在性の手術部位感染症の発生である。計画された 1 群の症例数 375 例（合計 750 例）は、PDS II 群、PDS Plus 群の発生割合をそれぞれ 0.12、0.06 と見積り、片側有意水準 0.025 の片側カイ二乗検定で 80%の検出力が期待される症例数として設定された。

非盲検下での中間解析は、合計 375 例（予定症例数の 50%）を対象として実施され、Bauer and Kohne (1994) [2]に基づき、無益性中止の基準$a_0 = 0.5$、有効中止の基準$a_1 = 0.0102$、最終解析の棄却域$c_p = 0.0038$と設定された。中間解析後に試験が継続した場合、条件付き検出力が 80%になるように症例数が最大 1 群 600 例（合計 1200 例）を上限に再推定される計画であった。

なお、本試験は中間解析の結果に基づいて、データ安全監視委員会は合計 1200 人の被験者が登録された後にさらに中間解析を行うことを推奨した。追加の中間解析実施による第一の過誤確率の制御は、Brannath (2002)[18]を用いて調整された。最終的に、2 回目の中間解析後に試験は中止され、論文ではプールデータの結果のみが報告されている。

Lehmacher and Wassmer (1999)[4]の方法の適用例 Wessels(2019)[19]

早期アルツハイマー病患者を対象とした、Lanabecestat に関する第 II/III 相臨床試験（AMARANTH 試験）の事例を紹介する。被験者は、実薬 20 mg 群、実薬 50 mg 群、プラセボ群に 1:1:1 の比でランダム化された。主要評価項目は Alzheimer's Disease Assessment

Scale-Cognitive Subscale 13(ADAS-Cog13)スコアのベースラインから 104 週目までの変化量である。計画された 1 群の症例数 734 例（合計 2202 例）は、プラセボに対する効果量として、実薬 20 mg 群、実薬 50 mg 群をそれぞれ 0.21、0.27 と見積り、片側有意水準 0.025 の Mixed effect Models for Repeated Measures (MMRM) による解析で、90%の検出力が期待される症例数として設定された。

本試験では 4 回の中間解析が予定されており、1 回目の中間解析は安全性の確認のみを目的として計画され、有効性に関する判断は伴わない。2 回目の中間解析は安全性の確認及び症例数再推定を目的として計画された。計画では合計 1000 例（予定症例数の約 45%）を対象として、ベイズ流予測確率が 55%未満なら症例数を増加させず試験続行、55 – 80%なら最大合計 3000 例を上限に症例数を増加、80%以上なら症例数を増加させず試験続行させる計画であった。もし、症例数を増加させた場合の第一種の過誤確率は、Lehmacher and Wassmer (1999)[4]で調整され、重みは、症例数の比の関数としてそれぞれ、$w_1 = \sqrt{1000/2202} = 0.67, w_2 = \sqrt{1202/2202} = 0.74$を用いる予定であった（なお、実際には登録は不確実性を伴うため、第 1 ステージ（2 回目の中間解析）に含まれる被験者数を 2202 で割った平方根をw_1とし、1 から第 1 ステージの情報量を引いた値の平方根が第 2 ステージの重みw_2とされる）。

なお、本試験は中間解析の結果に基づいて、症例数の変更は行われず、その後の中間解析で無益性中止されており、上記の調整は実際には利用されず、論文ではその時点のプールデータの結果のみが報告されている。

Gao(2008)[14]及び Mehta and Pocock (2011)[8]の方法の適用例 Bhatt (2013)[20]

緊急・待機的な経皮的冠動脈インターベンションを受ける患者を対象とした、抗血小板薬 cangrelor 静脈内投与及び点滴に関する第 III 相臨床試験（CHAMPION PHOENIX 試験）の事例を紹介する。被験者は、cangrelor 群、clopidogrel 群に 1:1 の比でランダム化された。主要評価項目は、ランダム化 48 時間後の死亡、心筋梗塞、虚血による再血行再建術、ステント血栓症の複合エンドポイントの発生率である。計画された 1 群の症例数 5450 例(合計 10900 例)は、cangrelor 群、Clopidogrel 群の複合エンドポイントの発生率をそれぞれ 3.1%、5.1%と見積り、検出力 85%が期待される症例数として、有効中止に関する中間解析の実施も考慮し設定された。70%の症例が集積された際に有効中止及び症例数再推定の中間解析の実施を計画し、Gamma family の α 消費関数を用い、有意水準を中間解析と最終解析、それぞれ、0.0109、0.0473 と設定した。症例数再推定は、条件付き検出力が 80%未満の場合に、目標検出力が 80%となるように、最大 1 群 22500 例（合計 45000 例）を上限に再推定される計画であった。この中間解析では、Gao(2008)[14]及び Mehta and Pocock (2011)[8]に基づき、相対リスクの低下率を用い、Uunfavorable zone (<13.6%)、Promising zone (≥13.6% to ≤21.2%)、Favorable zone (>21.2%)に分割された。なお、本試験は、中間解析で結果が Favorable zone に入ったため、症例数の増加はなかった。最終解析では、cangrelor 群

の統計的有意性が示された。アダプティブデザインのより詳細に関しては、Bhatt (2016)[21]で議論されているため適宜参照されたい。

2.2.3 区間推定

中間解析における早期中止や症例数再推定を伴う場合、その性質を考慮しない通常の信頼区間は目標とする名目上の被覆確率を有しない場合がある。

本項では、早期中止や症例数再推定の性質を考慮した片側信頼区間を構築するための2つのアプローチについて Wassmer and Brannath (2016)[22]の 8.2 節を参考に解説する。これらのアプローチは、群逐次デザインのために用いられる方法と同様のものである。

最初の方法は、事前に指定された中止基準に従った場合、臨床試験の終わりに報告できる正確な被覆確率を持つ信頼区間を生成する。

2 つ目の方法は、いつ、どのような理由で試験が中止されたかに関わらず、試験の各ステージで報告可能な繰り返し信頼区間を生成する。

また、2.2.2 項で説明した手法のうち、よく利用される統合検定に焦点を当てた。2 ステージ統合検定の関数を$C(p_1, p_2)$と表記する。この統合検定の関数はいくつかの提案法があり、その詳細は 2.2.2.1.1 項を参照のこと。

信頼区間の構築について議論するために、有効性パラメータδを設定し、帰無仮説H_0：$\delta \leq 0$を主要な興味とする。信頼区間と仮説検定の双対性を利用し、全ての帰無仮説$H_0^{\tilde{\delta}}$：$\delta \leq \tilde{\delta}$ （$\tilde{\delta}$は任意の実数）とする検定問題を考えて信頼区間を構成する。この章のほとんどの部分では、各$H_0^{\tilde{\delta}}$（H_0含む）は、ステージごとの片側 p 値 $p_{k,\tilde{\delta}} = 1 - \Phi((\hat{\delta}_k - \tilde{\delta})/\mathrm{se}_k)$ で検定されるとする（k=1、2 はステージ、$\hat{\delta}_k$は各ステージの推定値、se_kは各ステージの$\hat{\delta}_k$の標準誤差)。推定値$\hat{\delta}_k$およびse_kも、独立した患者のコホートから計算されたものと仮定する。なお、p 値$p_{k,\tilde{\delta}}$は、すべての標本点において$\tilde{\delta}$に対して非減少、すなわち、全ての$\tilde{\delta} \leq \tilde{\delta}'$に対して$p_{k,\tilde{\delta}} \leq p_{k,\tilde{\delta}'}$となる。

2.2.3.1 統合検定の正確な信頼限界

統合検定のためのステージに基づく順序を少し修正し、一貫性のある正確な下側信頼限界 $l^e_{k,\delta}$を定義し、前述した$p_{k,\tilde{\delta}}$に基づき構築する。ここでステージに基づく順序は、実際に観察される結果に対し、より「極端である」とする結果p'_1, p'_2の方向を次のように定義するものである。

第 1 ステージで中止する場合（$p_1 \leq \alpha_1$ または $p_1 > \alpha_0$）：$p'_1 \leq p_1$

第 2 ステージに進む場合（$\alpha_1 < p_1 \leq \alpha_0$）：$p'_1 \leq \alpha_1$,または,

$$\alpha_1 < p'_1 \leq \alpha_0 \text{かつ} C(p_1, p_2) \leq C(p'_1, p'_2)$$

信頼区間$\left(l^e_{k,\delta}; \infty\right)$は、$\delta = \tilde{\delta}$の状況下で$p_{1,\tilde{\delta}}, p_{2,\tilde{\delta}}$が独立に一様分布に従う場合に、被覆確率が正確に$1 - \alpha$に等しくなる。

ステージ k で H_0 を棄却する場合に限り$l^e_{k,\delta} > 0$となり、統合検定の結果と一致する。

$\delta = \tilde{\delta}$に対する p 値の関数$Q_{\tilde{\delta}}(p_{1,\tilde{\delta}}, p_{2,\tilde{\delta}})$は以下のように定義する。

$$Q_{\delta}(p_{1,\tilde{\delta}}, p_{2,\tilde{\delta}}) = \begin{cases} p_{1,\tilde{\delta}} & \text{if } p_1 \le \alpha_1 \ \ or \ \ p_1 > \alpha_0 \\ \alpha_{1,\tilde{\delta}} + \int_{\alpha_{1,\tilde{\delta}}}^{\alpha_{0,\tilde{\delta}}} \int_0^1 \mathbf{1}_{\{C(x,y) \le C(p_{1,\tilde{\delta}}, p_{2,\tilde{\delta}})\}} dxdy & \text{if } \alpha_1 < p_1 \le \alpha_0 \end{cases}$$

ここで $\alpha_{1,\tilde{\delta}} = 1 - \Phi\left(\Phi^{-1}(1-\alpha_1) - \frac{\tilde{\delta}}{\mathrm{se}_1}\right)$, $\alpha_{0,\tilde{\delta}} = 1 - \Phi\left(\Phi^{-1}(1-\alpha_0) - \frac{\tilde{\delta}}{\mathrm{se}_1}\right)$

正確な下側信頼限界 $l^e_{k,\delta}$は、以下のように$Q_{\tilde{\delta}}(p_{1,\tilde{\delta}}, p_{2,\tilde{\delta}}) = \alpha$となる$\tilde{\delta}$の値を解くことにより算出される。第 1 ステージで中止する場合は、通常用いられる信頼区間である。

<u>第 1 ステージで中止する場合（$p_1 \le \alpha_1 \ \ or \ \ p_1 > \alpha_0$）</u>

$$l^e_{1,\delta} = \hat{\delta}_1 - \Phi^{-1}(1-\alpha)\mathrm{se}_1$$

<u>第 2 ステージに進む場合（ $\alpha_1 < p_1 \le \alpha_0$）</u>

以下の方程式を$\tilde{\delta}$に関して数値積分を数値解析で解く。

$$\alpha_{1,\tilde{\delta}} + \int_{\alpha_{1,\tilde{\delta}}}^{\alpha_{0,\tilde{\delta}}} \int_0^1 \mathbf{1}_{\{C(x,y) \le C(p_{1,\tilde{\delta}}, p_{2,\tilde{\delta}})\}} dxdy = \alpha$$

ここで重要なのは、事前に規定された中止基準を順守する場合にのみ、下側信頼区間$l^e_{k,\delta}$が計算できることである。これは、中止基準に違反した結果に対しては、事前に規定したステージに基づく順序から外れるため、結果の順序関係が定義されないままだからである。

2.2.3.2 統合検定の繰り返し信頼限界

Lehmacher and Wassmer (1999)[4]と Brannath et al. (2002)[18]は、統合検定に繰り返し信頼区間アプローチを適用することを提案した。群逐次デザインにおける繰り返し信頼区間と同様に、$H_0^{\tilde{\delta}}$に対応する第 k ステージの p 値を$p_{k,\tilde{\delta}} = 1 - \Phi\left(\left(\hat{\delta}_k - \tilde{\delta}\right)/\mathrm{se}_k\right)$とし、統合検定にも適用することができる。

第 1 ステージの片側信頼区間の下側信頼限界$l^r_{1,\delta}$は、信頼係数$1-\alpha_1$として以下のように与えられる。

$$l^r_{1,\delta} = \hat{\delta}_1 - \Phi^{-1}(1-\alpha_1)\mathrm{se}_1$$

第 2 ステージの片側信頼区間の下側信頼限界$l^r_{2,\theta}$は、無効中止の有無（$\alpha_0 = 1$又は$\alpha_0 < 1$）により以下のように与えられる。

$$l^r_{2,\delta} = \begin{cases} \tilde{l}_{2,\delta} & \text{if } \alpha_0 = 1 \\ \min\{\tilde{l}_{0,\delta}, \tilde{l}_{2,\delta}\} & \text{if } \alpha_0 < 1 \end{cases}$$

ここで、$\tilde{l}_{0,\delta} = \hat{\delta}_1 - \Phi^{-1}(1-\alpha_0)\mathrm{se}_1$、$\tilde{l}_{2,\delta}$は$C(p_{1,\tilde{\delta}}, p_{2,\tilde{\delta}}) = c_p$の$\tilde{\delta}$についての解である。$\tilde{l}_{0,\delta}$の式は$p_{1,\tilde{\delta}} = \alpha_0$を解くことにより導出される。$\tilde{l}_{2,\delta}$は、$p_{k,\tilde{\delta}}$が$\tilde{\delta}$の増加関数であり、$C(p_{1,\tilde{\delta}}, p_{2,\tilde{\delta}})$も$p_{1,\tilde{\delta}}, p_{2,\tilde{\delta}}$の増加関数であることから上記の式が一意な解をもつ。解析的な解が得られない場合には、数値解析により導出する。

第 1 ステージでの中止基準がない（α_1=0, α_0=1）場合、正確な信頼区間l^e_{δ}と一致する。

統合関数$C(p_{1,\tilde{\delta}}, p_{2,\tilde{\delta}})$が逆正規法によるものである場合、$C(p_{1,\tilde{\delta}}, p_{2,\tilde{\delta}}) = c_p$を$\tilde{\delta}$について解くことで$\tilde{l}_{2,\delta}$は明示的に得られる。

$$C(p_1, p_2) = 1 - \Phi(w_1 z_1 + w_2 z_2) = 1 - \Phi\left(w_1 \Phi^{-1}(1-p_1) + w_2 \Phi^{-1}(1-p_2)\right)$$

$$p_{k,\tilde{\delta}} = 1 - \Phi\left(\frac{\hat{\delta}_k - \tilde{\delta}}{\mathrm{se_k}}\right)$$

$$\tilde{l}_{2,\delta} = \hat{\delta}_{weighteda} - \Phi^{-1}(1-c_p) \Big/ \left(\frac{w_1}{\mathrm{se}_1} + \frac{w_2}{\mathrm{se}_2}\right)$$

$$\text{with} \quad \hat{\delta}_{weighteda} = \left(\hat{\delta}_1 \frac{w_1}{\mathrm{se}_1} + \hat{\delta}_2 \frac{w_2}{\mathrm{se}_2}\right) \Big/ \left(\frac{w_1}{\mathrm{se}_1} + \frac{w_2}{\mathrm{se}_2}\right)$$

ここで$\hat{\delta}_{weighteda}$は、後述する 2.2.4.3 項のアダプティブな重み付き最尤推定量に相当する。

2.2.3.3 両側信頼区間（Two-Sided Confidence Intervals）

両側信頼区間を算出するには、$H_0^{(-)}$: $\delta \le 0$と$H_0^{(+)}$: $\delta \ge 0$の 2 つの片側検定を考える。ここまでの項では下側限界の導出について説明したが、上側限界については、$H_0^{(+)}$に対する p 値を算出し同様の原理を適用することで算出可能である。

2.2.4 点推定

本項では、点推定に関して、ナイーブな推定値及び重要な 3 つの提案法をレビューし、バイアスと平均二乗誤差（MSE）に関して各方法の比較結果について、Wassmer and Brannath (2016)[22]の 8.3 節の内容を参考に解説する。

ほとんどの文献（規制ガイドラインを含む）では、特に逐次デザインやアダプティブデザインを扱う場合、点推定値のバイアスに大きな重点が置かれている。一方で、分散は推定精度に影響を与えるもう一つの重要なものである。そのため、「良い」点推定値は、バイアスだけでなく、分散も適度に小さいことが求められる。

バイアスを除去するほとんどの方法は、（ナイーブな推定値と比較して）分散の増加につながる。したがって、バイアス除去法は、分散と平均バイアスの二乗の合計である MSE などで定量化されるように、全体的な精度に関して慎重に評価する必要がある。これらの理由から、各推定値の性能比較とその結論では MSE（またはその平方根）に焦点を当てるが、バイアスについても考慮する。

2.2.4.1 最尤推定量（ナイーブな推定値）

2 標本の群間差δの最尤推定量 MLE は試験全体のデータに基づく平均値の群間差$\hat{\delta}_{naive}$である。

$$\hat{\delta}_{naive} = (n_1 \hat{\delta}_1 + \hat{n}_2 \hat{\delta}_2) / N_{final}$$

第 1 ステージのデータに基づく中止基準又は症例数再推定が計画されている場合、$\hat{\delta}_{naive}$はバイアスのある推定値となる。このバイアスは、$\hat{\delta}_1, \hat{n}_2$が与えられたもとで$\hat{\delta}_2$の条件付き平均が$\delta$となることを用い、以下のように得られる（Liu et al. (2002)[23]）。

$$E_\delta(\hat{\delta}_{naive}) - \delta = Cov_\delta(n_1/N_{final}, \hat{\delta}_1)$$

$\hat{\delta}_1$の増加に伴い第 2 ステージの症例数が減る場合（症例数再推定や、有効中止など）、この共分散は正となり、バイアスも正となる。一方、$\hat{\delta}_1$が小さいことにより無効中止する場合においては負となる。

前述したように、MSE も推定量の重要な性質である。ナイーブな推定値の MSE は以下の式で与えられる（Brannath et al. (2006)[24]）。

$$E_\delta\left(\hat{\delta}_1 - \delta\right)^2 = E_\delta\left(\frac{\left(\hat{\delta}_1 - \delta\right)^2}{\left(N_{final}/n_1\right)^2}\right) + \frac{2\sigma^2}{n_1} E_\delta\left(\frac{n_1 \cdot \hat{n}_2}{N_{final}^2}\right)$$

この式より、MSE も$\hat{n}_2$のルールに依存することがわかり、したがって同様に一般的に未知となる。

2.2.4.2 固定重み付き最尤推定量

試験が中間解析後も必ず継続される（つまり、症例数再推定のみを目的として中間解析が行われる）場合を考える。この場合、δについてバイアスのない推定値として、第 1 ステージと第 2 ステージの平均値の群間差$\hat{\delta}_1, \hat{\delta}_2$、事前に規定した重み$u$（$0 < u < 1$）を用いた固定重み付き最尤定量として以下が提案されている（Liu et al. (2002)[23]）。

$$\hat{\delta}_{weighted} = u\hat{\delta}_1 + (1-u)\hat{\delta}_2$$

固定重み付き最尤推定量$\hat{\delta}_{weighted}$の分散は以下の式で与えられ、$\hat{n}_2$のアダプテーションルールに依存することがわかる（Brannath et al. (2006)[24]）。

$$Var_\delta(\hat{\delta}_{weighted}) = \frac{2\sigma^2}{n_1}\left(u^2 + (1-u)^2 E_\delta\left(\frac{n_1}{\hat{n}_2}\right)\right)$$

2.2.4.3 アダプティブ重み付き最尤推定量（Adaptively Weighted ML-Estimate）

Brannath et al. (2006)[24]、Cheng and Shen (2004)[25]、Lawrence and Hung (2003)[26]らは、アダプティブ重み付き最尤推定量として、以下のような各ステージの最尤推定量$\hat{\delta}_1, \hat{\delta}_2$の重み付き平均を提案している[2]。

$$\hat{\delta}_{weighteda} = \tilde{w}\hat{\delta}_1 + (1-\tilde{w})\hat{\delta}_2, \tilde{w} = \frac{w_1/\mathrm{se}_1}{w_1/\mathrm{se}_1 + w_2/\mathrm{se}_2}$$

ここで、se_kは$\hat{\delta}_k$の誤差、w_kは$w_1^2 + w_2^2 = 1$を満たす事前に決めた重みであり、$w_k = \sqrt{n_k/N_{initial}}$が自然な重みである。

この推定量$\hat{\delta}_{weighteda}$は中間解析での中止基準がない場合に中央値不偏推定量となる。これは、2.2.3.2 項より逆正規法の繰り返し信頼区間が、中間解析での中止基準がない場合は$l_{2,\delta}^r = \tilde{l}_{2,\delta} = \hat{\delta}_{weighteda} - \Phi^{-1}(1-c_p)/\left(\frac{w_1}{\mathrm{se}_1} + \frac{w_2}{\mathrm{se}_2}\right)$となり、この片側 50%信頼区間の下側信頼限界に一致する（α=0.5 の場合に$\Phi^{-1}(1-c_p) = 0$となる）ためである。

[2] 本推定量は Brannath et al. (2006)[24]で示されている「中央値不偏推定量」に該当する。

2.2.4.4 中央値不偏点推定量

アダプテーションルールとは無関係に中央値がδに等しい中央値不偏点推定量$\hat{\delta}_{median}$を用いること提案されている(Brannath et al. (2003)[27]; Lawrence and Hung (2003)[26]; Proschan (2003)[28])。

中央値不偏点推定量$\hat{\delta}_{median}$は、正確な片側信頼区間$\left(l_{k,\delta}^{e};\infty\right)$の被覆確率が 50%となる$l_{k,\delta}^{e}$を求めること、つまり、2.2.3.1 項の p 値の関数$Q_{\tilde{\delta}}\left(p_{1,\tilde{\delta}},p_{2,\tilde{\delta}}\right)$ を用いて以下の方程式をδについて解くことにより得られる。

$$Q_{\tilde{\delta}}\left(p_{1,\tilde{\delta}},p_{2,\tilde{\delta}}\right)=0.5$$

この推定値は、試験の終了時（試験が中止した時点）に一度だけ計算することができる。

中央値不偏点推定量として、厳密に保守的な信頼区間（例えば、50%繰り返し信頼区間から算出される$l_{0.5}$）に基づくものも適用可能である。繰り返し信頼区間の実際の被覆確率は 0.5 より大きくなるため、この推定量は有効性を過少評価する方向にバイアスが生じる。そのバイアスの大きさは、パラメータの真値とアダプテーションルールに依存し、その検討には一般に数値シミュレーションが必要である。

2.2.4.5 点推定値の性能評価

Wassmer and Brannath (2016)[22]はナイーブな推定量、固定重み付き最尤推定量、アダプティブ重み付き最尤推定量、中央値不偏推定量の 4 つについて、症例数再設計を伴う 2 段階デザインで、中間解析での有効中止基準（α_1=0.005）の設定有無×無効中止基準（α_0=0.5）の設定有無による 4 つのシナリオでのシミュレーションによる評価を行っている。このとき、症例数再推定は、z 検定の条件付き検出力が 80%となるよう設定し、また、各ステージの症例数の比（$\hat{n}_2/n_1$）の最小値を 0.1、最大値を 5 と設定した。アダプティブな重みづけ ML-推定量、中央値不偏推定量算出の逆正規統合検定に用いる各ステージの重みは$w_1=w_2=1/\sqrt{2}$とした。本項の冒頭に述べたように、バイアスと平均二乗誤差（MSE）による評価・考察を行い、以下のように結論づけている。

いずれの推定量も平均バイアスは$\sqrt{\mathrm{MSE}}$よりかなり小さいことから、ばらつきも重要な関心事である。

中央値不偏推定量が最も良い性能（バイアスと MSE が共に小さい）を示し、少なくとも条件付き検出力に基づいて症例数再推定される試験に最も適していると思われる。

ナイーブな推定量とアダプティブ重み付き最尤推定量は、MSE に関して性能がよく、計算が単純であることから、同様に合理的な推定値と考えられる。

固定重み付き最尤推定量は、バイアスについては有効・無効中止基準の設定がない場合のみバイアスが 0 となるがそれ以外では中央値不偏推定量よりバイアスが大きく、MSE については全てのシナリオで他の推定量より劣っており、特にバイアスが 0 となる中止基準

のないシナリオで顕著であった。従って、この推定値はバラつきが大きすぎるため、中間解析での中止基準がない場合でも使用することを勧められない。

これらの推定値には、ある種の平均値と中央値のバイアスがあることが想定される。推定値の分布はアダプテーションルールに依存するため、試験の計画段階で MSE（または残差の四分位数）と同様に平均値と中央値のバイアスを調査し、最も合理的なアダプテーションルールを見つけることを推奨する。

2.2.5 その他

2.2.3 項では、2 つのステージ（中間解析：1 回）から構成される症例数再推定を伴う 2 群比較群逐次デザインを用いる試験を想定して、p 値や信頼区間を算出する方法について解説した。2.2.2.1.1 項で解説した統合検定に対する原則を利用することで、3 つ以上のステージ（中間解析：2 回以上）から構成される試験においても、p 値（正確な p 値および繰り返し p 値）の算出は比較的容易である。一方、3 つ以上のステージから構成される試験において、正確な信頼区間の算出はそれほど容易ではない。

Brannath et al. (2009)[29]は、条件付き棄却確率（conditional rejection probability：CRP）の原則を利用することで正確な信頼区間を算出する方法を提案した。Tsiatis et al. (1984)[30]が提案した古典的な群逐次デザインにおけるステージに基づく順序を利用して算出する調整信頼区間を拡張した方法である。特別な状況（最終解析よりも一つ前の中間解析でアダプテーションを実施し、その後、最終解析を実施する場合）では、正確な被覆確率を有することが保証される。それ以外の状況では、正確な被覆確率を有するかどうかは単調性の仮定の成立に依存し、この仮定の成立は数理的に示すことができない。数理的に示すことができるのは、被覆確率が一般に保守的になることを保証することだけである。しかしながら、広範に渡るシミュレーション実験を通して、実用的な状況下では、正確な被覆確率を有する信頼区間と中央値不偏推定量を得られることが示されている。なお、単調性の仮定とは、$h \in (-\infty, \infty)$に対して $H_h : \delta \leq h$ versus $\delta > h$ とした場合、ステージに基づく順序を利用して算出した p 値が h に対して単調であることを意味し、この仮定により正確な信頼区間が一意的に定まる。本手法により算出可能な信頼区間は片側であり、事前規定した中止規則に従って試験が早期中止したステージでのみ算出可能で妥当である。その後、Gao et al. (2013)[31]が、アダプテーション実施後のデザイン下で得られる最終解析の検定統計量を、当初予定していたデザイン下に逆にマッピングすることで正確な信頼区間を算出する方法を提案した。本手法により算出可能な信頼区間は両側である。

一方、Mehta et al. (2007)[32]も、CRP の原則を利用することで信頼区間を算出する方法を提案した。Jennison and Turnbull (1989)[33]が提案した古典的な群逐次デザインに対する繰り返し信頼区間を拡張した方法である。しかしながら、古典的な繰り返し信頼区間は試験全体の第一種の過誤確率をすべて消費しない（例えば、中間解析で早期中止となった場合、実際には実施されないそれ以降の中間解析も考慮に入れる）ため、一般に保守的になる。本手法を用いた場合でも信頼区間は保守的になり、点推定ではバイアスを伴う。なお、本

手法では早期中止したステージのみではなく、すべてのステージで信頼区間の算出が可能である。

また、本項で述べた方法で考慮可能なアダプテーションには、症例数再推定のほか、得られたデータに基づく中間解析の回数や間隔、消費関数の変更を含む。

2.2.6 rpact を用いた症例数再推定の実装

本項では、R パッケージの 1 つである rpact[34]を用いて症例数再推定を実装する R コードについて解説する。rpact を利用すれば、評価項目の型（連続変数、2 値変数またはイベント発生までの期間）にかかわらず、2.2.3 項で解説した症例数再推定の実施を考慮した p 値や信頼区間の算出が可能となる。本項では、そのうち、評価項目が連続変数の場合で、試験途中に症例数再推定を計画していることから、群間比較に逆正規法を利用した統合検定を用いる試験に焦点を当て、rpact を用いて p 値や信頼区間を算出する方法について解説する。本項で解説する内容は、rpact のサイトでも事例として紹介されているので、合わせて参照いただきたい[35]。なお、rpact により 3 つ以上のステージから構成される試験でも実装可能であるものの、症例数再推定を行った場合、正確な信頼区間の出力値は妥当ではないため、ご注意いただきたい。

試験デザインの設定

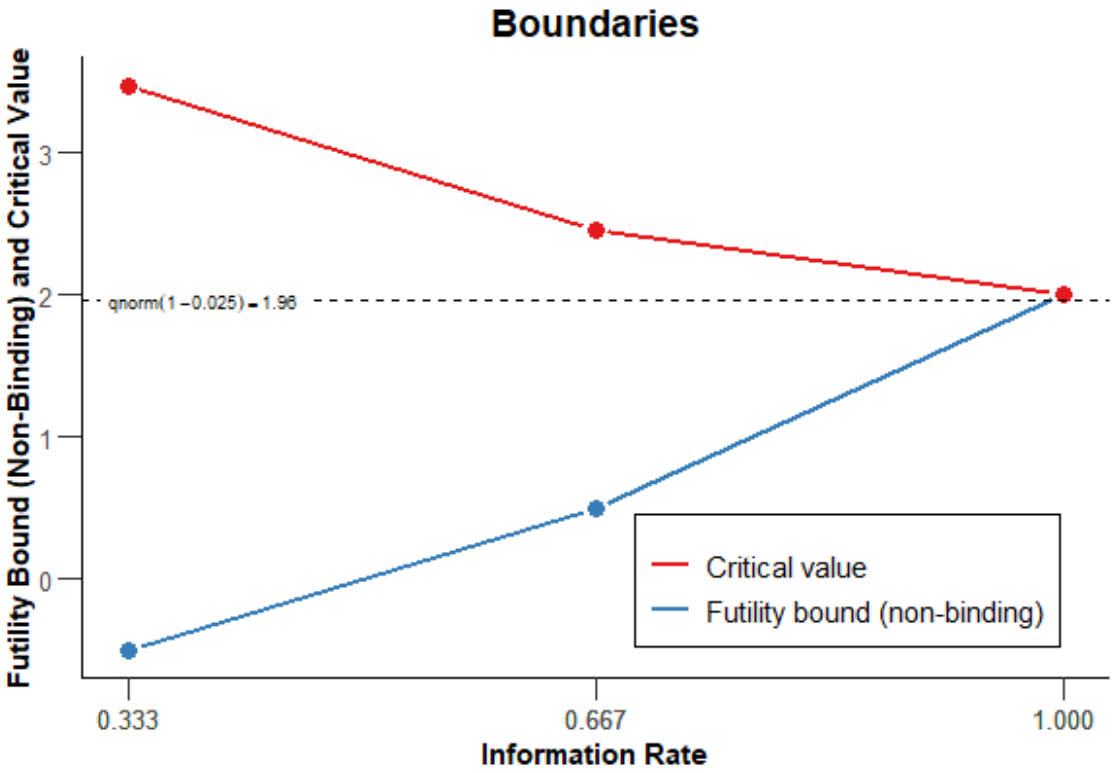
3 つのステージから構成される試験（中間解析を 2 回、最終解析を 1 回実施する試験）を想定する。中間解析時に症例数再推定を検討するため、群間比較には逆正規法を利用して各ステージの p 値を統合する統合検定を用いる。中間解析では、有効中止および無益性中止の両方を考慮し、有効中止の境界には O‘Brien & Fleming 型の境界を用い、無益性中止の境界（non-binding の境界）には 1 回目、2 回目の中間解析でそれぞれ-0.5 および 0.5 の Z 値に基づく棄却限界値を設定する。rpact では getDesignInverseNormal()関数を用いて、以下のとおりに設定する。

なお、デフォルトでは、ステージ間の情報分数は等間隔であり、試験全体の有意水準は片側 0.025 である。また、有効中止の境界は O‘Brien and Fleming 型の境界であるため、本事例では、明示的にコードを指定する必要はない。本事例では、デフォルトである等間隔の情報分数からなる統合検定を用いてデータ解析を実施する。

```
# Example of an inverse normal combination test:
designIN <- getDesignInverseNormal(futilityBounds = c(-0.5, 0.5))
```

各解析時点における棄却限界値を、plot()関数を用いることで図示することができる。

```
plot(designIN, type = 1)
```



無益性中止の境界は non-binding の境界であるため、有効中止の境界に影響を及ぼさないものの、条件付き検出力には影響を及ぼす。また、getDesignInverseNormal()関数を用いることにより、群間比較に逆正規法を利用して各ステージの p 値を統合する統合検定を用いることが自動的に設定される。

データの入力

rpact において、症例数再推定を伴う中間解析時のデータを利用するには、各ステージの要約統計量を算出する必要がある。一般に getDataset()関数が用いられ、入力した要約統計量のタイプによって、自動的に評価項目の型と治療群の数が判別される。2 群の並行群間比較試験で平均を比較する場合、means1、means2、stDevs1、stDevs2、n1、n2 をベクトル形式（ベクトルの長さは中間のステージ数）で指定する。例えば、第 1 ステージおよび第 2 ステージにおいて、治療群および対照群で以下の結果が得られたとする。

第 1 ステージ

群	症例数（n）	平均（means）	標準偏差（stDevs）
治療群	34	112.3	44.4

群	症例数（n）	平均（means）	標準偏差（stDevs）
対照群	37	98.1	46.7

第 2 ステージ

群	症例数（n）	平均（means）	標準偏差（stDevs）
治療群	31	113.1	42.9
対照群	33	99.3	41.1

rpact では、以下のコードを用いてデータセットを作成する。

```
datasetExample <- getDataset(
    means1 = c(112.3, 113.1),
    means2 = c(98.1, 99.3),
    stDevs1 = c(44.4, 42.9),
    stDevs2 = c(46.7, 41.1),
    n1 = c(34, 31),
    n2 = c(37, 33))
```

また、別の方法として、各ステージまでに集積された累積データの要約統計量を指定することによっても、データセットを作成することができる。

```
getDataset(
    overallMeans1 = c(112.3, 112.68),
    overallMeans2 = c(98.1, 98.67),
    overallStDevs1 = c(44.4, 43.35),
    overallStDevs2 = c(46.7, 43.84),
    overallN1 = c(34, 65),
    overallN2 = c(37, 70))
```

データ解析

getAnalysisResults()関数を用いてデータ解析を実施する。その際、上述の［試験デザインの設定］で定義した designIN と［データの入力］で定義した datasetExample を指定す

る。なお、デフォルトでは、データ解析による第 1 ステージと第 2 ステージの結果が出力される。第 1 ステージの結果のみに出力を制限する場合、getAnalysisResults()関数において、stage = 1 と明記する必要がある。

```
getAnalysisResults(design = designIN, dataInput = datasetExample)
```

その結果、以下の結果が出力される。第 2 ステージにおいて、逆正規法を用いて統合した検討統計量は 1.837 となり、棄却限界値である 2.454 より小さくなった。従って、帰無仮説は棄却できない。また、繰り返し p 値は 0.0785 となり、繰り返し信頼区間の下限値は -4.803 となった（0 よりも小さく、信頼区間は 0 を含んでいる）。

```
## Analysis results (means of 2 groups, inverse normal combination test design):
##
## Design parameters:
##     Fixed weights                          : 0.577, 0.577, 0.577
##     Critical values                        : 3.471, 2.454, 2.004
##     Futility bounds (non-binding)          : -0.500, 0.500
##     Cumulative alpha spending              : 0.0002592, 0.0071601, 0.0250000
##     Local one-sided significance levels    : 0.0002592, 0.0070554, 0.0225331
##     Significance level                     : 0.0250
##     Test                                   : one-sided
##
## User defined parameters: not available
##
## Default parameters:
##     Normal approximation                   : FALSE
##     Direction upper                        : TRUE
##     Theta H0                               : 0
##     Equal variances                        : TRUE
##
## Stage results:
##     Cumulative effect sizes                : 14.20, 14.02, NA
##     Cumulative (pooled) standard deviations : 45.61, 43.60, NA
##     Stage-wise test statistics             : 1.310, 1.314, NA
##     Stage-wise p-values                    : 0.09721, 0.09680, NA
##     Combination test statistics            : 1.298, 1.837, NA
```

```
##
## Analysis results:
##   Assumed standard deviation                : 43.6
##   Actions                                   : continue, continue, NA
##   Conditional rejection probability         : 0.06767, 0.19121, NA
##   Conditional power                         : NA, NA, NA
##   Repeated confidence intervals (lower)     : -25.271, -4.803, NA
##   Repeated confidence intervals (upper)     : 53.67, 32.80, NA
##   Repeated p-values                         : 0.29776, 0.07854, NA
##   Final stage                               : NA
##   Final p-value                             : NA, NA, NA
##   Final CIs (lower)                         : NA, NA, NA
##   Final CIs (upper)                         : NA, NA, NA
##   Median unbiased estimate                  : NA, NA, NA
```

<u>症例数再推定の検討</u>

rpact において、nPlanned を指定することにより中間解析時に条件付き検出力の算出が可能である。nPlanned では、残りの各ステージにおける両群を合わせた症例数をベクトル形式で指定する。例えば、nPlanned = 60 とした場合、最終ステージにおいて両群合わせて 60 例の患者が登録された場合の条件付き検出力が算出される。デフォルトの割付け比は 1 対 1 であるものの、allocationRatioPlanned により変更可能である。

```
results <- getAnalysisResults(design = designIN, dataInput = datasetExample, stage = 2, nPlanned = 60)
```

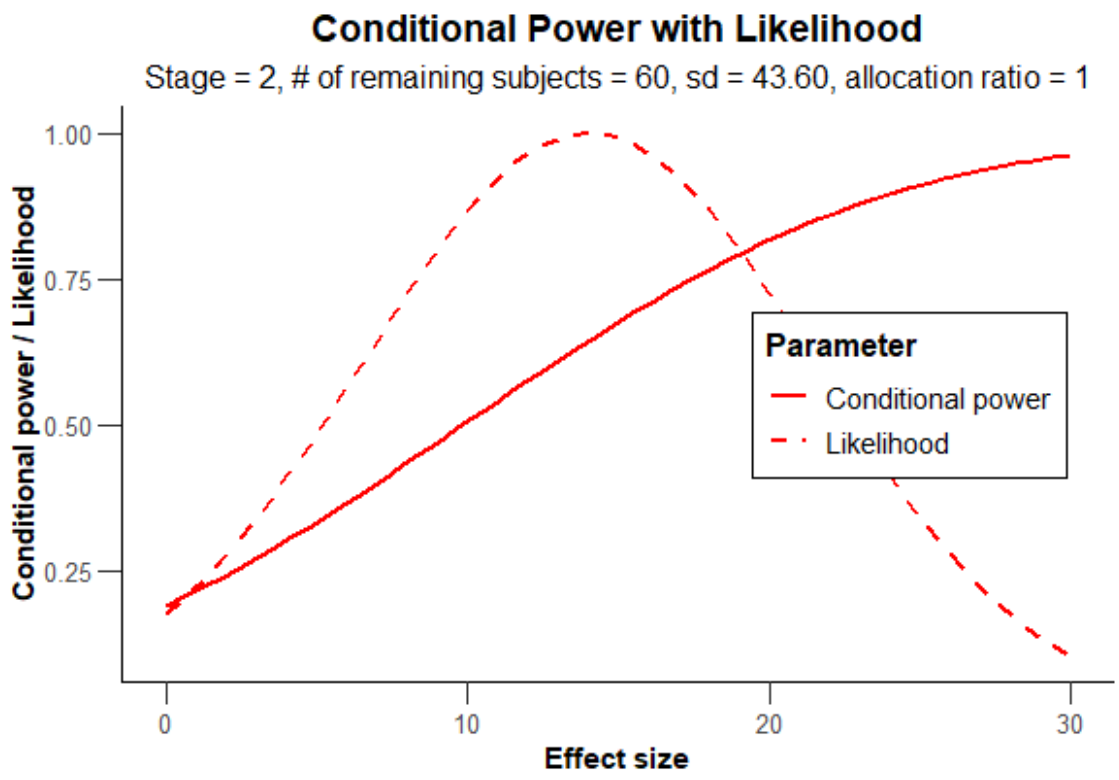
その結果、以下の結果が出力される。条件付検出力の値は 0.645 となり、十分な大きさを有していないため、症例数の変更を検討することが適切な状況と言える。しかしながら、条件付き検出力の算出は、中間解析で観察された群間差と標準偏差に基づいて行われるため、まずはエフェクトサイズやそのばらつきの大きさに伴う条件付き検出力や尤度関数の変化を確認することが適切と考える。

```
## Analysis results (means of 2 groups, inverse normal combination test design):
##
## Design parameters:
##   Fixed weights                             : 0.577, 0.577, 0.577
##   Critical values                           : 3.471, 2.454, 2.004
##   Futility bounds (non-binding)             : -0.500, 0.500
```

```
##    Cumulative alpha spending              : 0.0002592, 0.0071601, 0.0250000
##    Local one-sided significance levels    : 0.0002592, 0.0070554, 0.0225331
##    Significance level                     : 0.0250
##    Test                                   : one-sided
##
## User defined parameters:
##    Planned sample size                    : NA, NA, 60
##
## Default parameters:
##    Normal approximation                   : FALSE
##    Direction upper                        : TRUE
##    Theta H0                               : 0
##    Planned allocation ratio               : 1
##    Equal variances                        : TRUE
##
## Stage results:
##    Cumulative effect sizes                : 14.20, 14.02, NA
##    Cumulative (pooled) standard deviations : 45.61, 43.60, NA
##    Stage-wise test statistics             : 1.310, 1.314, NA
##    Stage-wise p-values                    : 0.09721, 0.09680, NA
##    Combination test statistics            : 1.298, 1.837, NA
##
## Analysis results:
##    Assumed standard deviation             : 43.6
##    Actions                                : continue, continue, NA
##    Conditional rejection probability      : 0.06767, 0.19121, NA
##    Conditional power                      : NA, NA, 0.6449
##    Repeated confidence intervals (lower)  : -25.271, -4.803, NA
##    Repeated confidence intervals (upper)  : 53.67, 32.80, NA
##    Repeated p-values                      : 0.29776, 0.07854, NA
##    Final stage                            : NA
##    Final p-value                          : NA, NA, NA
##    Final CIs (lower)                      : NA, NA, NA
##    Final CIs (upper)                      : NA, NA, NA
##    Median unbiased estimate               : NA, NA, NA
```

rpact では、例えば、thetaRange = c(0,30)と指定することで、エフェクトサイズを変化させたときの条件付き検出力を以下のとおりに図示することができる。

```
plot(results, thetaRange = c(0, 30))
```



また、エフェクトサイズのみでなく、標準偏差も変化させた場合の条件付き検出力の算出も可能である。例えば、以下のコードにより、エフェクトサイズが 15、標準偏差が 35 の場合の条件付き検出力の算出が可能である。なお、getAnalysisResults()関数で、thetaH1 = 15、assumedStDev = 35 と指定することで、同様の結果を得ることができる。

```
stageResults <- getStageResults(design = designIN, dataInput = datasetExample)
getConditionalPower(stageResults, nPlanned = 60, thetaH1 = 15, assumedStDev = 35)
```

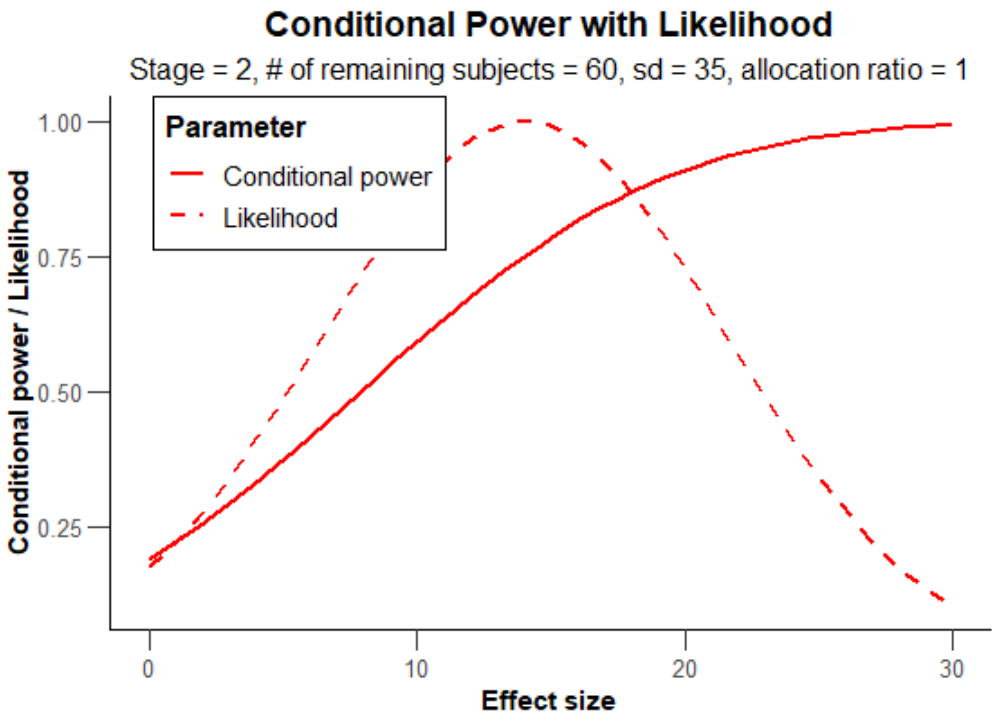
その結果、以下の結果が出力される。条件付き検出力は 0.784 となり、標準偏差が中間解析で観察された値よりも小さい 35 程度であれば、80%に近い条件付き検出力となる。

```
## Conditional power results means:
##
```

```
## User defined parameters:
##     Planned sample size                 : NA, NA, 60
##     Assumed effect under alternative    : 15
##     Assumed standard deviation          : 35
##
## Default parameters:
##     Planned allocation ratio            : 1
##
## Output:
##     Conditional power                   : NA, NA, 0.7842
```

また、先ほどと同様、thetaRange = c(0,30)と指定することで、標準偏差は 35 に固定して、エフェクトサイズを変化させた場合の条件付き検出力を以下のとおりに図示することができる。

```
plot(stageResults, nPlanned = 60, thetaRange = c(0,30), assumedStDev = 35)
```



最終解析

上述の症例数再推定について検討した結果から、症例数を変更せずに、当初の予定どおりの症例数（最終ステージの両群合わせた症例数は 60 例）で最終解析を実施することにしたとする。その結果、最終ステージで以下の結果が得られたとする。なお、本事例では、症例数を変更せずに最終解析を実施する状況を想定するが、症例数を変更する場合も、getDataset()関数を用いて最終ステージで得られたデータを入力することで、最終解析の実施が可能である。

最終ステージ

群	症例数（n）	平均（means）	標準偏差（stDevs）
治療群	32	111.3	41.4
対照群	31	100.1	39.5

先ほどと同様に、getAnalysisResults()関数を用いてデータ解析を実施する。

```
datasetExample <- getDataset(
    means1 = c(112.3, 113.1, 111.3),
    means2 = c(98.1, 99.3, 100.1),
    stDevs1 = c(44.4, 42.9, 41.4),
    stDevs2 = c(46.7, 41.1, 39.5),
    n1 = c(34, 31, 32),
    n2 = c(37, 33, 31))
getAnalysisResults(design = designIN, dataInput = datasetExample)
```

その結果、以下の結果が出力される。最終解析において、逆正規法を用いて統合した検討統計量は 2.128 となり、棄却限界値である 2.004 より大きくなった。従って、帰無仮説は棄却できる。また、繰り返し p 値は 0.0183 となり、繰り返し信頼区間の下限値は 0.768 となった（0 よりも大きく、信頼区間は 0 を含んでいない）。

なお、3 ステージ以上から構成される試験において症例数の変更を行った場合、正確な信頼区間は妥当でないことから、以下のような警告文が表示される。今回の場合、症例数の変更を実施していないことから、最終解析で算出された調整信頼区間（0.62, 24.52）は妥当である。

```
## Calculation of final confidence interval performed for kMax = 3 (for kMax > 2, it is
theoretically shown that it is valid only if no sample size change was performed)
```

```
## Analysis results (means of 2 groups, inverse normal combination test design):
##
## Design parameters:
##      Fixed weights                        : 0.577, 0.577, 0.577
##      Critical values                      : 3.471, 2.454, 2.004
##      Futility bounds (non-binding)        : -0.500, 0.500
##      Cumulative alpha spending            : 0.0002592, 0.0071601, 0.0250000
##      Local one-sided significance levels  : 0.0002592, 0.0070554, 0.0225331
##      Significance level                   : 0.0250
##      Test                                 : one-sided
##
## User defined parameters: not available
##
## Default parameters:
##      Normal approximation                 : FALSE
##      Direction upper                      : TRUE
##      Theta H0                             : 0
##      Equal variances                      : TRUE
##
## Stage results:
##      Cumulative effect sizes              : 14.20, 14.02, 13.12
##      Cumulative (pooled) standard deviations : 45.61, 43.60, 42.43
##      Stage-wise test statistics           : 1.310, 1.314, 1.098
##      Stage-wise p-values                  : 0.09721, 0.09680, 0.13826
##      Combination test statistics          : 1.298, 1.837, 2.128
##
## Analysis results:
##      Assumed standard deviation           : 42.43
##      Actions                              : continue, continue, reject
##      Conditional rejection probability    : 0.06767, 0.19121, NA
##      Conditional power                    : NA, NA, NA
##      Repeated confidence intervals (lower) : -25.2714, -4.8030, 0.7676
##      Repeated confidence intervals (upper) : 53.67, 32.80, 25.31
##      Repeated p-values                    : 0.29776, 0.07854, 0.01828
##      Final stage                          : 3
```

##	Final p-value	: NA, NA, 0.01968
##	Final CIs (lower)	: NA, NA, 0.6209
##	Final CIs (upper)	: NA, NA, 24.52
##	Median unbiased estimate	: NA, NA, 12.62

2.3　盲検下での分散に基づく症例数再推定

本章では盲検下での症例数再推定について紹介する。2.3.1 項から 2.3.5 項までは連続データにおける症例数再推定について紹介し、2.3.6 項では二値データにおける症例数再推定について紹介する。

連続データでの一般的な盲検下での症例数再推定の方法として、中間解析において治療群を併合したデータによって推定した分散（一標本分散推定値）を用いた症例数を調整する方法が提案されている（Gould and Shih (1992)[36]）。一標本分散推定値は真の治療効果の差が大きい場合に過大推定される事が知られているため、他にも盲検下の分散の推定法が提案されており、不偏な推定値として、登録順とランダム化ブロックサイズを用いてブロック間の分散を推定する事により全体の分散を推定し、症例数を調整する方法（Xing and Ganju (2005)[37]）が提案されている。ただし、この方法では第一種の過誤確率の増大が指摘されている（Friede and Kieser (2013)[38]）。また、EM アルゴリズムを用いた分散推定の方法（Gould and Shih (1992)[36]）も提案されており、この方法では E ステップにおいて暫定的な治験薬が割り付けられる条件付き確率を求め、M ステップにおいて中間データに基づき分散の最尤推定値を推定する事ができる。この方法においては推定値が初期値に依存する事、収束条件に到達する前に手順が止まってしまう場合がある事、真の治療効果の大きさに応じてバイアスが生じる事、ブロックランダム化などにはそのまま適用する事が出来ない事などの問題が指摘されている（Friede and Kieser (2002)[39]）。そのため、ここでは第一種の過誤確率の増大が小さく、推定が単純で説明が容易である一標本分散推定値を用いた症例数再推定の方法のみを紹介する。

2.3.1　優越性試験における方法

平均値の差に関する優越性の検定を行う場合、仮説を以下のように設定する。

$$H_0: \delta_T - \delta_C = 0 \ \ vs. H_1: \delta_T - \delta_C > 0$$

なお、治療群の真の治療効果をδ_Tとし、対照群の真の治療効果をδ_Cとする。この時、一群あたりの症例数は以下の通り計算できる。

$$N_{initial} = \frac{2\sigma^2\left(z_\alpha + z_\beta\right)^2}{\delta_{pre}^2}$$

これに対して、n_1例の評価が完了した時点で、治療群を併合した一標本分散推定値を以下の通り推定する。

$$s_{lumped}^2 = \frac{1}{2n_1 - 1}\sum_{i=1}^{2n_1}(X_i - \bar{X}_1)^2$$

なお、$\bar{X}_1$はn_1例の治療群を併合した時の平均値とする。この一標本分散推定値を用いて、以下の通り、最終解析に必要な症例数を再推定する。

$$\widehat{N}_{final} = \frac{2s^2_{lumped}\left(z_{\alpha/2} + z_{\beta}\right)^2}{\delta^2_{pre}}$$

この時、治療効果に群間差がなければバイアスがないが、群間差がある場合には分散の過大推定を生じる事が報告されている（Kieser and Friede (2003)[40]）。バイアスの程度は以下の式で表される。

$$E\left(s^2_{lumped}\right) - \sigma^2 = \frac{n_1}{2(2n_1 - 1)}\delta^2$$

なお、δは真の治療効果の差を表す。そこで、以下に表す通り、一標本分散推定値から想定される治療効果に基づく群内分散を引く事で一標本分散推定値固有のバイアスを減少させる事ができる（Zucker et al. (1999)[41]）。

$$s^2_{adj} = s^2_{lumped} - \frac{n_1}{2(2n_1 - 1)}\delta^2_{pre}$$

なお、$\delta = \delta_{pre}$の場合には不偏となる。この調整された一標本分散推定値を用いて、最終解析に必要な症例数を再推定する事もできる。ただし、調整された一標本分散推定値は最初に想定した治療効果δ_{pre}が真の治療効果δより小さい場合には分散を過小推定してしまう（Proschan (2005)[42]）。そのため、未調整の一標本分散推定値よりバイアスは小さいが、検出力の低下が指摘されている（Kieser and Friede (2003)[40]）。なお、最終の検定統計量は最終の分散推定値を用いて推定される。

2.3.2 同等性や非劣性試験における方法

平均値の差に対する同等性の検定を行う場合、同等性の範囲を（m_1, m_2）とし、同等性の仮説を以下のように定義する。

$$H_0: \delta_T - \delta_C \le m_1 \ \ or \ \ \delta_T - \delta_C \ge m_2 \ \ vs. H_1: m_1 < \delta_T - \delta_C < m_2$$

この仮説は以下の二つの片側仮説の同時検定となる

$$H_{01}: \delta_T - \delta_C \le m_1 \ \ vs. H_{11}: \delta_T - \delta_C > m_1$$
$$H_{02}: \delta_T - \delta_C \ge m_2 \ \ vs. H_{12}: \delta_T - \delta_C < m_2$$

この仮説に対して、以下の検定統計量を考える。

$$T_i = \frac{\left(\hat{\delta}_T - \hat{\delta}_C - m_i\right)}{\sqrt{2s^2/n}}, i = 1,2$$

片側有意水準α/2 に対して、$T_1 \ge t_{n-2,1-\alpha/2}$、又は$T_2 \le -t_{n-2,1-\alpha/2}$なら、帰無仮説$H_{0i}$は棄却される。そして、二つの片側検定において、両方の仮説が片側有意水準$\alpha/2$で棄却されれば、同等性の仮説H_0も片側有意水準$\alpha/2$で棄却される。そのため、$\delta = \delta_{pre}$を想定し、検出力$1-\beta$でH_{0i}を棄却するための症例数は、

$$N_{initial} = \max\left\{\frac{2\sigma^2\left(z_{\alpha/2} + z_{\beta}\right)^2}{\left(\delta_{pre} - m_1\right)^2}, \frac{2\sigma^2\left(z_{\alpha/2} + z_{\beta}\right)^2}{\left(\delta_{pre} - m_2\right)^2}\right\}$$

で推定される。これに対して、一標本分散推定値を用いて、以下の通り、最終解析に必要な症例数を推定する事ができる。

$$\widehat{N}_{final} = \max\left\{\frac{2s_{lumped}^2\left(z_{\alpha/2}+z_{\beta}\right)^2}{\left(\delta_{pre}-m_1\right)^2}, \frac{2s_{lumped}^2\left(z_{\alpha/2}+z_{\beta}\right)^2}{\left(\delta_{pre}-m_2\right)^2}\right\}$$

なお、一標本分散推定値s_{lumped}^2の代わりに調整済み一標本分散推定値s_{adj}^2を用いる事もできる。また、優越性試験の場合と同様に最終の検定統計量は最終の分散推定値を用いて推定される。

平均値の差に対する非劣性の検定を行う場合は非劣性マージンをm_1とし、非劣性の仮説を以下のように定義する事で同等性の検定と同様に扱う事ができる。

$$H_0: \delta_T - \delta_C \leq m_1 \ \ vs. H_1: \delta_T - \delta_C > m_1$$

2.3.3 第一種の過誤確率への影響

2.3.1 項で示した一標本分散推定量に基づく症例数再推定では正規分布に従うデータに対する優越性の仮説検定において第一種の過誤確率の増大は無視できるほど小さい（Kieser and Friede (2003)[40]）。また、症例数が十分に大きければ任意の分布に対しても第一種の過誤確率の増大はない。一方、非劣性試験や同等性試験においては、第一種の過誤確率の増大が生じる可能性が指摘されている（Kieser and Friede (2003)[40]）。2019 年に発行された FDA ガイダンスにおいても非劣性又は同等性の仮説検定を含んだ試験では、第一種の過誤確率の限定的な増大が生じる可能性があると指摘されており、第一種の過誤確率の増大の程度を評価すべきであると記載されている（FDA (2019)[44]）。特に、中間解析までの症例数が小さい時に第一種の過誤の増大は大きくなり、最大症例数が大きくなる程、第一種の過誤確率の増大は大きくなる事が報告されている（Glimm et al. (2020)[45], Friede and Stammer (2010)[46]）。また、最終解析時の治療群間差にバイアス（同等性においては正と負のバイアス、非劣性の場合は正のバイアス）がある場合に第一種の過誤確率が増大する（Kieser and Friede (2003)[40]）。そこで、シミュレーションによって最大の第一種の過誤確率の増大を確認し、第一種の過誤確率の増大を抑えるため、有意水準を引き下げるといった方法が提案されており、FDA や EMA に受け入れられたという報告もある（Glimm et al. (2020)[45]）。この事例については 2.3.5 項で紹介する。さらに、優越性試験において、最終の症例数に対して最初に計画した症例数から増加のみを許容する場合、第一種の過誤確率の増加は小さいが、減少も許容する場合には第一種の過誤確率が増大すると報告されている（Wittes et al. (1999)[47]）。また、非劣性試験や同等性試験において、最終の症例数に対して下限及び上限を設定する事により、第一種の過誤確率の増大は減少する事が報告されている（Glimm et al. (2020)[45]）。

2.3.4 推定値の性質について

盲検下の症例数再推定では最終解析時の平均値や分散の推定値にバイアスが生じる事が知られている（Posch et al. (2018)[48]）。

最終解析時の平均値の推定値に対するバイアスは、以下の式で与えられる。

$$\int_0^{\infty}\int_0^{\infty}\frac{t(2n_1-2)\big(r(q,\delta-1)-r(q,\delta+t)\big)}{\sigma^2[1+r(q,\delta+t)][1+r(q,\delta-t)]}\phi_{0\sqrt{2\sigma^2/n_1}}(t)\chi^2_{2n_1-2}\left(\frac{q(2n_1-2)}{\sigma^2}\right)dtdq$$

なお、$r(x,y)$は以下の式で与えられる。

$$r(x,y)=n_2\left(\frac{2(n_1-1)}{2n_1-1}x+\frac{n_1}{2(2n_1-1)}y^2\right)/n_1$$

この時、真の治療効果δが 0 の場合には最終解析時の非盲検データにおける平均値の差の推定値は不偏であるが、真の治療効果δが正の値を取る場合に真値より負の方向へバイアスを生じ、真の治療効果δが負の値を取る場合に真値より正の方向へバイアスを生じる。なお、第 1 ステージの症例数n_1が大きくなるにつれて、バイアスは小さくなる。また、真の治療効果δが大きくなるにつれて、バイアスの程度は小さくなる。

最終解析時の分散の推定値に対するバイアスは、以下の式で与えられる。

$$\int_0^{\infty}\int_{-\infty}^{\infty}\frac{1}{\big(n_1-1+n_1r(q,t+\delta)\big)}\left((n_1-1)(q-\sigma^2)+\frac{t^2n_1/2-\sigma^2}{2+2r(q,t+\delta)}\right)$$
$$\times\phi_{0,\sqrt{2\sigma^2/n_1}}(t)\chi^2_{2n_1-2}(q\gamma)\gamma dtdq$$

ここで、$\gamma=(2n_1-n)/\sigma^2$で表される。この時、真の分散σ^2が大きいほど最終解析時の非盲検データにおける分散の推定値S^2は真値より負の方向へバイアスを生じ、第 1 ステージの症例数n_1が大きくなるにつれて、バイアスは小さくなる。また、真の治療効果δが大きくなるにつれて、バイアスの程度は小さくなる。

最終解析時の信頼区間の被覆確率においては、第 1 ステージの症例数n_1が少ない場合に平均値及び分散のバイアスが大きくなり、名目上の被覆確率を有さない。ただし、真の治療効果δが 0 の場合には名目上の被覆確率を有する。また、正の治療効果を持つ場合、信頼区間の下限は広がる一方、信頼区間の上限は狭まり、治療効果の大きさによって両側信頼区間が広がるか狭まるかが決まる。

なお、調整された一標本分散推定値S^2_{adj}は未調整の一標本分散推定値S^2_{lumped}よりもバイアスの程度は大きくなる事が知られている。

盲検下の症例数再推定によって生じる最終解析時の平均値や分散の推定値にバイアスや信頼区間の被覆確率をシミュレーションによって計算する blindConfidence という R package が公開されている。

2.3.5 事例

ここでは Glimm et al. (2020)[45]で紹介されている二重盲検ランダム化三群同等性試験の盲検下での症例数再推定の事例を紹介する。

一般に、PK/PD 試験では PK の血中濃度の AUC が主要評価項目として用いられ、AUC は対数正規分布に従う。また、薬理学では、データは幾何平均や変動係数（CV）で要約される。そして、幾何平均に対して、以下のような仮説が設定される。

$$H_{01}:\frac{\tau_B}{\tau_j}\le 0.80\ \text{ vs.}\ H_{\mathrm{A1}}:\frac{\tau_B}{\tau_j}>0.80,$$

$$H_{02}:\frac{\tau_B}{\tau_j} \geq 1.25 \text{ vs.} H_{\mathrm{A2}}:\frac{\tau_B}{\tau_j} < 1.25$$

ここで、τ_Bとτ_jは治療群と対照群j $(j = 1,2)$の PK の血中濃度の AUC の幾何平均を表し、多くの規制当局では 0.80 と 1.25 が同等性マージンとして要求される。

ここで紹介する事例では、ある対照薬に対して、いくつかの異なる集団や目的を持つ PK 試験が実施されていたが、常に異なる結果が与えられていたため、盲検下の中間症例数再推定による臨床試験が計画された。この試験では 1:1:1 にランダム化され、各群 20 例（全 60 例）が予定していた評価を終えた時に盲検下の中間症例数再推定を実施する事とした。第 2 ステージの症例数mは表 10 に示した観測された PK の血中濃度の AUC の CV に基づいて選択された。

表 10　第 2 ステージの症例数

中間時点の併合 CV	第 2 ステージの各群の症例数
$< CV_1$	m_1
CV_1以上,CV_2未満	m_2
CV_2以上,CV_3未満	m_3
$> CV_3$	m_4

なお、具体的な$CV_1 < CV_2 < CV_3$と$m_1 < m_2 < m_3 < m_4$は機密保持のため、ここでは与えられていない。これに対して、第一種の過誤確率の増大を調査するため、片側有意水準を 5%とし、シミュレーションが実施された。真の CV が 45%より小さい値である場合に一万回のシミュレーションにおいて、帰無仮説下（$H_{01}:\tau_B/\tau_j \leq 0.80$,及び$H_{02}:\tau_B/\tau_j \geq 1.25$）における棄却する確率を計算した結果を図 5 に示した。実線と破線は二つの対照薬に対して帰無仮説を棄却する確率を表す。

真の CV が 0.175 と 0.375 の間にある時、間違って棄却してしまう割合は真の割合が期待されていた値（5%であった場合に、二つの赤い破線）を明らかに超える。実際、真の CV が 0.21 の時、間違って棄却してしまう割合は 5.32%となる。これに対して、第一種の過誤確率の増大を抑えるため、片側有意水準を 5%から 4.5%に引き下げ、この条件でのシミュレーションが実施された（図 6）。これにより、最大の第一種の過誤確率は 5%以下に抑える事ができ、この手法は FDA と EMA にも受け入れられたという事が報告されている。

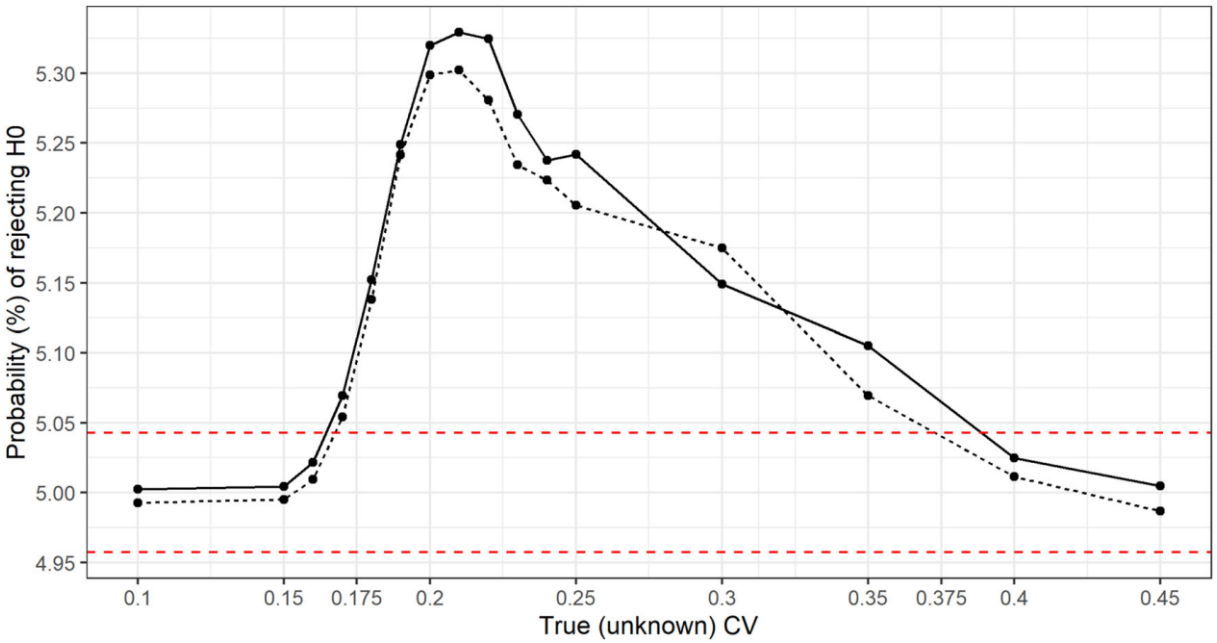


図 5　片側有意水準 5%に対する対照薬 1（実線）と対照薬 2（破線）との同等性の誤判断の割合

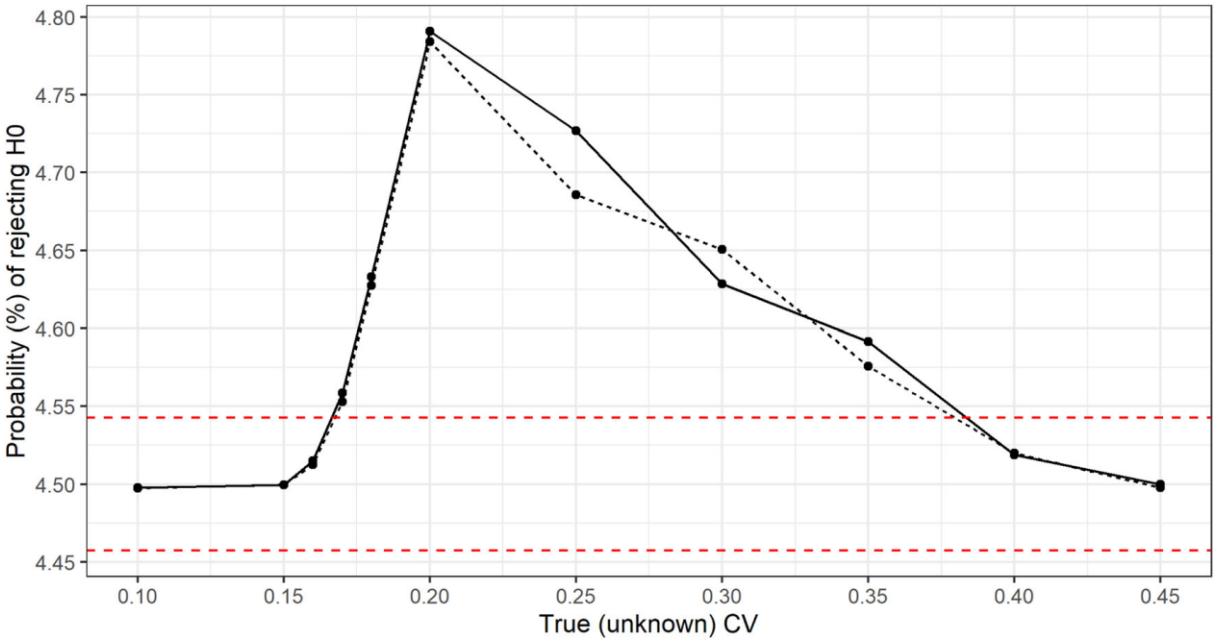


図 6　片側有意水準 4.5%に対する対照薬 1（実線）と対照薬 2（破線）との同等性の誤判断の割合

2.3.6　二値データの評価項目における方法

本節では評価項目を二値データとした場合の優越性試験の症例数再推定の方法を紹介する。各患者で観測される二値データの評価項目において、群$i(i = 1,2)$に対してイベント数X_iが発生割合π_iの二項分布に従う事とする。この時、帰無仮説は以下の通りに表される。

$$H_0: \pi_1 = \pi_2 \text{ vs. } H_1: \pi_1 \neq \pi_2$$

この仮説に対して、以下のχ^2検定統計量を考える。

$$\chi^2 = \frac{2n\left(X_1(n - X_2) - X_2(n - X_1)\right)^2}{n^2 X(2n - X)}$$

ここで、$X = X_1 + X_2$とする。この時、片側有意水準を$\alpha/2$、検出力を$1 - \beta$、全体のイベント発生割合を$\pi = (\pi_1 + \pi_2)/2$とし、対立仮説の下でのイベント発生割合の差をθ_{pre}とすると、一群の必要症例数は以下の通りに表される[49]。

$$N_{initial} = \frac{2\left(z_{\alpha/2} + z_\beta\right)^2}{\theta_{pre}^2} \cdot \pi(1 - \pi)$$

これに対して、n_1例の評価が完了した時点で全体のイベント発生割合を以下の通り推定する。

$$p = \frac{X_{11} + X_{12}}{n_1}$$

ここで、X_{11}とX_{12}はn_1例に対する各群のイベント数とする。これを用いて、以下の通り、最終解析に必要な症例数を推定する。

$$\widehat{N}_{final} = \frac{2\left(z_{\alpha/2} + z_\beta\right)^2}{\theta_{pre}^2} \cdot p(1 - p)$$

これは、最初に設定した症例数$N_{initial}$を用いて表すと、以下の通りとなる。

$$N_{final} = N_{initial} \cdot \frac{p(1 - p)}{\pi(1 - \pi)}$$

この方法においても第一種の過誤確率の固定標本デザインと比較して、過度な増大はない事が知られている（Friede and Kieser (2004)[50]）。しかしながら、厳密に第一種の過誤確率を制御したい場合には未知の分散において、全ての取り得る値に対して最大の第一種の過誤確率を計算し、最大の第一種の過誤確率を名義有意水準以下に抑えられるよう、有意水準を調整する方法が提案されている（Kieser and Friede (2000)[51]）。また、症例数を再推定するための中間時点の症例数n_1が最終解析に必要な症例数N_{final}に近い、又は真のイベント発現割合が 0 に近い場合は検出力が高くなり、群間で症例数の割付比に不均衡がある場合には検出力が低下する事が知られている（Friede and Kieser (2004) [50]）。

参考文献

[1] Shun Z, Yuan W, Brady WE, Hsu H. Type I error in sample size re-estimations based on observed treatment difference. Statistics in Medicine 2001;20(4):497-513.

[2] Bauer P and Kohne K. Evaluation of experiments with adaptive interim analyses. Biometrics 1994;50:1029-1041.

[3] Cui L, Hung, HMJ, Wang SJ. Modification of sample-size in group sequential trials. Biometrics 1999;55:853-857.

[4] Lehmacher W, Wassmer G. Adaptive sample-size calculations in group sequential trials. Biometrics 1999;55:1286-1290.

[5] Chang M. Adaptive design method based on sum of p-values. Statistics in Medicine 2006;26(14): 2772-2784.

[6] Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. Biometrics 1995;51:1315-1324.

[7] Chen YHJ, DeMets DL, Lan KKG. Increasing the sample size when the unblinded interim result is promising. Statistics in Medicine 2004;23(7):1023-1038.

[8] Mehta CR, Pocock SJ. Adaptive increase in sample size when interim results are promising: a practical guide with examples. Statistics in Medicine 2011;30(28):3267-3284.

[9] Chang M. Adaptive design theory and implementation using SAS and R. 2nd ed, Boca Raton., FL: CRC Press, 2014.

[10] Liu Y, Xu H. Sample size re-estimation for pivotal clinical trials, Contemporary Clinical Trials 2021;102:106215.

[11] Bauer P, Koenig F. The reassessment of trial perspectives from interim data—a critical view. Statistics in Medicine 2006;25(1):23-36.

[12] Glimm E. Comments on ‘adaptive increase in sample size when interim results are promising: a practical guide with examples’ by C. R. Mehta and S. J. Pocock, Statistics in Medicine 2012; 31(1):3198-3199.

[13] Hung HMJ, Wang SJ, Yang P. Some challenges with statistical inference in adaptive designs. Journal of Biopharmaceutical Statistics 2014;24:1059-1072.

[14] Gao P, Ware JH, Mehta CR. Sample size re-estimation for adaptive sequential design in clinical trials. Journal of Biopharmaceutical Statistics 2008;18:1184-1196.

[15] Jennison C, Turnbull BW. Adaptive sample size modification in clinical trials: Start small then ask for more? Statistics in Medicine 2015;34:3793-3810.

[16] Hsiao ST, Liu L, Mehta CR. Optimal promising zone designs. Biometrical Journal 2018;61(3): 1175–1186.

[17] Diener MK, Knebel P, Kieser M, et al. Effectiveness of triclosan-coated PDS Plus versus uncoated PDS II sutures for prevention of surgical site infection after abdominal wall closure: The randomised controlled PROUD trial. Lancet 2014;384:142-152.

[18] Brannath W, Posch M, Bauer P. Recursive combination tests. Journal of the American Statistical Association 2002;97:236-244.

[19] Wessels AM, Tariot PN, Zimmer JA, et al. Efficacy and safety of lanabecestat for treatment of early and mild Alzheimer disease: the AMARANTH and DAYBREAK-ALZ randomized clinical trials. JAMA Neurology. 2020;77(2):199-209.

[20] Bhatt DL, Stone GW, Mahaffey KW, et al. Effect of platelet inhibition with cangrelor during PCI on ischemic events. New England Journal of Medicine 2013;368:1303-1313

[21] Bhatt DL, Mehta CR. Adaptive designs for clinical trials. New England Journal of Medicine 2016;375:65-74

[22] Wassmer G, Brannath W. Group Sequential and Confirmatory Adaptive Designs in Clinical Trials. Springer, 2016.

[23] Liu Q, Proschan MA, Pledger GW. A unified theory of two-stage adaptive designs. Journal of the American Statistical Association 2002;97:1034-1041.

[24] Brannath W, König F, Bauer P. Estimation in flexible two stage designs. Statistics in Medicine 2006;25:3366-3381.

[25] Cheng Y, Shen Y. Estimation of a parameter and its exact confidence interval following sequential sample size reestimation trials. Biometrics 2004;60:910-918.

[26] Lawrence J, Hung H. Estimation and confidence intervals after adjusting the maximum information. Biometrical Journal 2003;45:143-152.

[27] Brannath W, König F, Bauer P. Improved repeated confidence bounds in trials with a maximal goal. Biometrical Journal 2003;45:311-324.

[28] Proschan MA. The geometry of two-stage tests. Statistica Sinica 2003;13:163-177.

[29] Brannath W, Mehta CR, Posch M. Exact confidence bounds following adaptive group sequential tests. Biometrics 2009;65(2):539-546.

[30] Tsiatis AA, Rosner GL, Mehta CR. Exact confidence intervals following a group sequential test. Biometrics 1984;40(3):797-803.

[31] Gao P, Liu L, Mehta C. Exact inference for adaptive group sequential designs. Statistical in Medicine 2013;32(23):3991-4005.

[32] Mehta CR, Bauer P, Posch M, Brannath W. Repeated confidence intervals for adaptive group sequential trials. Statistics in Medicine 2007;26(30):5422-5433.

[33] Jennison C, Turnbull BW. Interim Analyses: The Repeated Confidence Interval Approach. Journal of the Royal Statistical Society: Series B (Methodological) 1989;51(3):305-334.

[34] rpact https://www.rpact.com [accessed 14 September 2022].

[35] rpact vignettes https://www.rpact.com/vignettes/rpact_continuous_analysis_example [accessed 14 September 2022].

[36] Gould AL, Shih WJ. Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. Communications in Statistics - Theory and Methods 1992;21(10):2833-2853.

[37] Xing B, Ganju J. A method to estimate the variance of an endpoint from an on-going blinded trial. Statistics in Medicine 2005;24(12):1807-14.

[38] Friede T, Kieser M. Blinded sample size re-estimation in superiority and noninferiority trials: bias versus variance in variance estimation. Pharmaceutical Statistics. 2013;12(3):141-6.

[39] Friede T, Kieser M. On the inappropriateness of an EM algorithm based procedure for blinded sample size re-estimation. Statistics in Medicine 2002;21(2):165-76.

[40] Kieser M, Friede T. Simple procedures for blinded sample size adjustment that do not affect the type I error rate. Statistics in Medicine 2003;22(23):3571-81.

[41] Zucker DM, Wittes JT. Oliver Schabenberger, and Erica Brittain. Internal pilot studies II: comparison of various procedures. Statistics in Medicine 1999;18(24):3493-3509.

[42] Proschan MA. Two-Stage Sample Size Re-Estimation Based on a Nuisance Parameter: A Review. Journal of Biopharmaceutical Statistics 2005;15:559-574.

[43] Friede T, Kieser M. Blinded sample size reassessment in non-inferiority and equivalence trials. Statistics in Medicine 2003;22(6):995-1007.

[44] Food and Drug Administration. Adaptive Designs for Clinical Trials of Drugs and Biologics Guidance for Industry. 2019.

[45] Glimm E, Yau L, Woehling H. Type I Error Inflation of Blinded Sample Size Re-Estimation in Equivalence Testing. Statistics in Biopharmaceutical Research 2020;13(2):156-169.

[46] Friede T, Stammer H. Blinded Sample Size Recalculation in Noninferiority Trials: A Case Study in Dermatology. Drug Information Journal 2010;44(5):599-607.

[47] Wittes J, Schabenberger O, Zucker D, Brittain E, Proschan M. Internal pilot studies I: Type I error rate of the naive t-test. Statistics in Medicine 1999;18:3481.3491.

[48] Posch M, Klinglmueller F, König F, Miller F. Estimation after blinded sample size reassessment. Statistical Methods in Medical Research 2018;27(6):1830-1846.

[49] Sahai H, Khurshid A. Formulae and tables for the determination of sample sizes and power in clinical trials for testing differences in proportions for the two-sample design: a review. Statistics in Medicine 1996;15(1):1-21.

[50] Friede T, Kieser M. Sample size recalculation for binary data in internal pilot study designs. Pharaceutical statistics. 2004;3:269-279.

[51] Kieser M, Friede T. Re-calculating the sample size in internal pilot study designs with control of the type I error rate. Statistics in Medicine 2000;19:901-911.

3　治療群の選択

この章では、複数の治療群と 1 つの対照群があり、中間解析の結果に基づいて治療群を 1 つ以上選択し、選択された治療群と対照群を最終解析にて比較する、2 ステージを基本とするアダプティブデザインを解説する。これは、試験目的が 2 つある単一の検証的試験である。通常、proof of concept や用量反応性の確認を通して検証する治療群を絞る探索的目的の第二相試験と、治療群と対照群の比較が目的である検証目的の第三相試験は分けて実施される。このアダプティブデザインでは、中間解析をはさんで探索的試験と検証的試験を統合し、最終解析にて試験全体の被験者データを用いて治療群と対照群を比較する。ここでは、試験開始から中間解析までを第 1 ステージ、中間解析から最終解析までを第 2 ステージと呼ぶことにする。図 7　治療群の選択のアダプティブデザインの例にこのアダプティブデザインの例を図示した。この例では、3 つの治療群と 1 つの対照群を設定して開始され、第 1 ステージの集積データによる中間解析の結果から治療群 B を選択（言い換えれば、治療群 A と治療群 C を中止）している。第 2 ステージでは治療群 B と対照群のデータをさらに集積し、最終解析にて第 1 ステージと第 2 ステージのデータを統合して治療群 B と対照群を比較している。この例の最終解析は、第 1 ステージと第 2 ステージのデータを統合して検証しているが、第 2 ステージのみのデータを用いて検証することもあり得る。

この目的のアダプティブデザインは、シームレス第 2／3 相試験、drop the loser デザイン、pick the winner デザイン、multi-arm multi-stage デザインなど、様々な名称で呼ばれている。この報告書では、検証する試験治療を選択する目的から、治療群の選択のアダプティブデザインと呼ぶことにする。

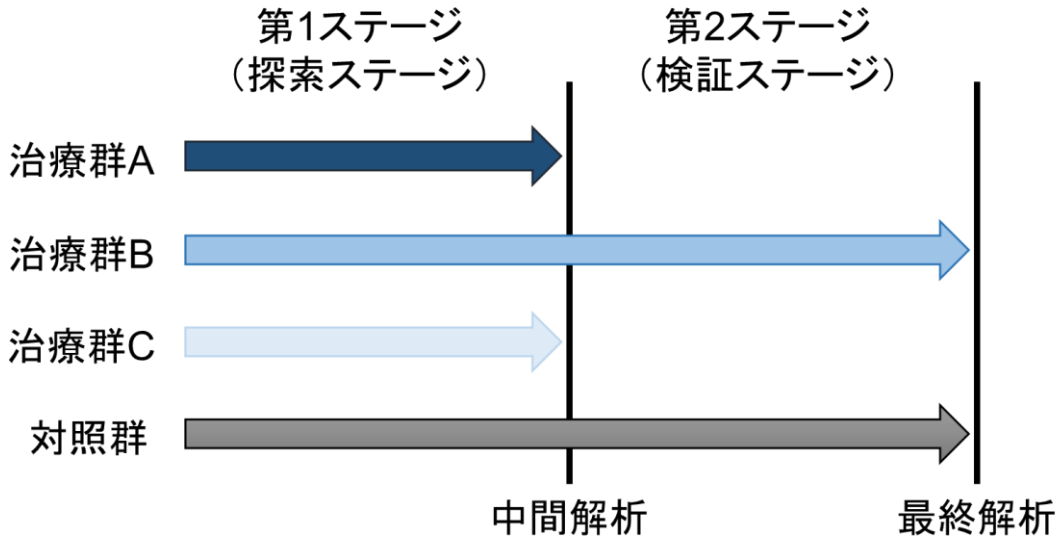


図 7　治療群の選択のアダプティブデザインの例

この章の構成は以下の通りである。3.1 項では、この章を通した表記法をまとめた。3.2 項では、中間解析にて治療群を選択するルールについて述べる。治療群の選択のアダプティブデザインでは、他のタイプのアダプティブデザインと同様に、一般的な解析方法を用いるとバイアスの発生や第一種の過誤確率の増大が懸念される。3.3 項にて、これらの課題が生じる原因を解説する。3.4 項は、治療群の選択のアダプティブデザインが適用された二つの臨床試験の事例を紹介する。3.5 項では、治療選択のアダプティブデザインに適用可能な仮説検定、点推定、区間推定の統計的方法を解説する。3.6 項には一部の統計手法について R の rpact パッケージと本タスクフォースが作成した SAS マクロを含めた。

3.1　第 3 章で用いる記号の整理

この章では、以下の記号法を用いている。

表 11　本章の表記法

記号	説明
j	治療群の番号（$j = 1, \ldots, k$）。対照群は$j = 0$
n_1、n_2	試験計画時の各ステージ、各治療群の症例数。症例数は各群で同じ。
μ_j	群 j の治療効果の真値
$\mu_{(j)}$	中間解析で治療効果が j 番目に優れた群の、治療効果の真値
$\bar{X}_j$	第 1 ステージでの群 j の治療効果の推定値
$\bar{X}_{(j)}$	第 1 ステージで治療効果が j 番目に優れた群の、治療効果の推定値
$\delta_{(j)}$	第 1 ステージで治療効果が j 番目に優れた治療群と対照群との治療効果の差の真値（つまり、$\mu_{(j)} - \mu_0$）
$\tilde{\mu}_{(j)}$	第 1 ステージで治療効果が j 番目に優れた治療群の、治療効果の一様最小分散条件付き不偏推定量（UMVCUE）
$\sigma^2_{1,(j)}$	第 1 ステージで治療効果が j 番目に優れた群の、第 1 ステージでの治療効果の分散
$\sigma^2_{2,j}$	第 1 ステージで治療効果が j 番目に優れた群の、第 2 ステージでの治療効果の分散
$\bar{Y}_j$	第 2 ステージでの治療群 j の治療効果の推定値
Z_j	治療群 j と対照群の群間差の検定統計量
$z_j^{(1)}, z_j^{(2)}$	第 1 ステージ、第 2 ステージそれぞれの被験者データを用いたZ_jの実現値
z_j	両ステージの被験者データを用いたZ_jの実現値
ϕ, Φ	標準正規分布の確率密度関数、累積分布関数
$\Phi^{-1}(1-\alpha)$	標準正規分布の累積分布関数の $100(1-\alpha)$%点

記号	説明
$\mathcal{S}$	治療群の集合
$\mathcal{T}_1$	全ての治療群の集合 $\{1, \ldots, \mathrm{k}\}$
$\mathcal{T}_2$	第 2 ステージに進んだ治療群の集合
$A_{\mathcal{S}}$	条件付き過誤確率
$H_{\mathcal{S}}$	積帰無仮説
H_j	治療群 j と対照群の群間差についての基本帰無仮説
$\varphi_{\mathcal{S}}$	積仮説の検定（$\varphi_{\mathcal{S}} = 1$ は$H_{\mathcal{S}}$を棄却、$\varphi_{\mathcal{S}} = 0$ は$H_{\mathcal{S}}$を採択）
$q_{\mathcal{S}}$	積仮説の検定の p 値
$d_{\mathcal{S}}$	Dunnett 検定の棄却限界値

3.2　治療群を選択するルール

治療群の選択のアダプティブデザインでは、中間解析の結果に基づいて 1 つ以上の治療群を選択し、最終解析にて対照群と比較する。治療群を選択するルールは、対象患者、試験の目的、開発状況などに応じて、様々な考え方がある。有効性の観点だけでも、特に有効性の高い治療だけを選ぶ、もしくは特に有効性の低い群を中止して多くの治療群を選択するなど、様々考えられる。第 1 ステージの目的を用量反応性の確認とする場合、観察された有効性がプラトーに達した用量群のうち低い方を選択するなど、用量反応性の結果によって選択する用量群は異なることが考えられる。ここでは、以下の二つのルールに分ける。

ルール 1：最も効果の高い 1 つの治療群を選択する
ルール 2：2 番目以降に効果の高い治療群を含む，1 つ以上の治療群を選択する

ルール 1 は、点推定値が最も大きい群を選択するルールと言い換えることができる、単純なルールである。しかし現実には、より柔軟なルールを適用したい場合がある。例えば、安全性など有効性以外の結果を考慮して 2 番目に効果の高い治療群を選択する、用量反応性を考慮して効果がプラトーに達する用量群のうち用量の低い 2 つの群を選択する、対照群と比較して一定以上の効果がある 2 つの治療群を選択する，などが考えられる。ルール 1 よりも柔軟なルール 2 を適用したい場面がある。

治療群の選択のアダプティブデザインを伴う臨床試験に未調整の解析方法を適用すると、治療群の選択ルールなど試験計画の様々な側面に依存して、第一種の過誤確率やバイアスの性質が変化する。次項にて、これらの性質がアダプティブデザインの計画とどのように関係しているかを説明する。

3.3　特別な統計的方法が必要な理由

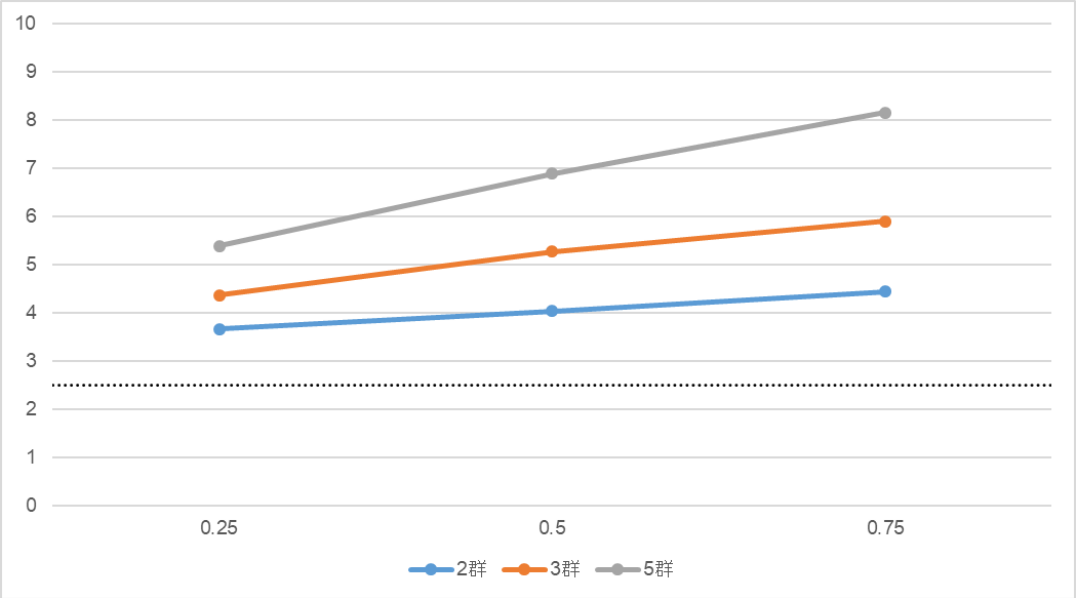
Bauer et al. (2010)[1]は、治療群の選択のアダプティブデザインに発生するバイアスに、選択バイアスと報告バイアスがあることを述べている。選択バイアスとは、中間解析にて選択された治療群について最終解析の平均に生じるバイアスである。報告バイアスとは、中間解析にて選択された治療群であるかどうかに関わらず、それぞれの治療群の平均に生じるバイアスである。有効性が低かったことによって中間解析にて中止した治療群の有効性は最終解析で観察されず、中間解析の結果のみが報告されることで、報告バイアスが発生する。一方、中間解析にて選択された治療群は、中間解析で有効性が高かったことに基づいて最終解析まで進むことにより、選択バイアスが発生する。Bauer et al. (2010) [1]は、選択バイアスと報告バイアスそれぞれのバイアスと平均二乗誤差（MSE）を定式化し、中間解析の時期と治療群の数によってバイアスと平均二乗誤差が変化することを図示している。

Wang et al. (2010)[2]は、第一種の過誤確率が上昇することを、シミュレーションを用いた分かりやすい例で説明している。J 個の治療群があり、反応はすべて同じ標準正規分布に従うとする（つまり、すべての治療群の真の効果は同じ）。検証に進む試験治療は、1 回の中間解析にて、最大の反応が得られた試験治療を選ぶという、対照群との比較ではないルールに基づくとする。検証に進んだ試験治療は、最終解析にて片側 2.5%の有意水準により仮説検定がされる。試験治療の数 J を 2、5、7、中間解析の実施時期を 1/6、2/6、3/6、4/6、5/6 とした、10 万回繰り返しのシミュレーション結果を示している。例えば、J=2、中間解析の時点を 3/6（つまり、被験者数として 50%の時点）としたとき、有意水準片側 2.5%に対する第一種の過誤確率は 4%弱（Wang et al. (2010)[2]の図 1 から目視で読み取った）であり、適切に制御できていない。

本タスクフォースにて、Wang et al. (2010) [2]の図 1 と似た設定によるシミュレーションを実行した。対照群を 1 つ、治療群の数を 2、3、5、各群の被験者数は 100、中間解析の時点は情報分数（中間解析時の被験者数／100）として 0.25、0.5、0.75 とする。各群の治療の反応は同じ標準正規分布に従うとし、つまり、すべての群の真値は等しいことを仮定する。治療群の選択ルールに前項のルール 1、つまり、中間解析にて、対照群との平均の差が最も大きい治療群が検証へ進み、それ以外の治療群は中間解析にて中止する。最終解析では、検証に進んだ治療群の第 1 ステージと第 2 ステージの被験者データを統合し、群間差の真値を 0 とする帰無仮説に基づく有意水準片側 2.5%の Z 検定、平均と 95%信頼区間を算出する。信頼区間には、通常の算術平均と分散既知に基づく正規分布による方法を用いる。シミュレーションの繰り返し数は 10 万回である。中間解析にて選択された治療群の、最終解析での有意水準片側 2.5%に対する第一種の過誤確率、最終解析での 95%信頼区間の被覆確率、中間解析時と最終解析の点推定値のバイアスのシミュレーションの結果を、それぞれ図 8～10 に示した。この例では、すべての群で真の平均は同じであるため、中間解析時点に観測されたランダムなエラーにより治療群が選択される。すべての治療群

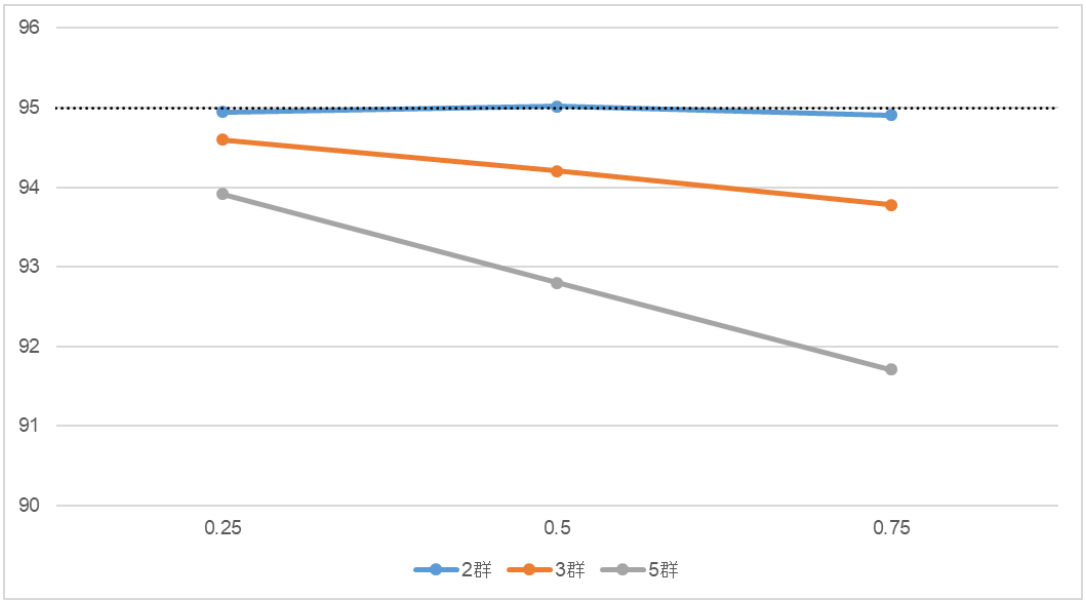
で真の効果は同じであっても、ランダムなエラーが効果として確定して治療群が中間解析時にて選択されるため正のバイアスが生じ、最終解析時の平均にも影響する。このバイアスの発生に伴い、第一種の過誤確率の上昇、被覆確率の低下が起こる。第一種の過誤確率を適切に制御し、被覆確率を名義の水準近くに維持し、バイアスを抑えるためには、特別な解析方法が必要である。

バイアス、被覆確率、第一種の過誤確率の程度は、治療群の選択ルール、中間解析の回数と時期、各群の反応値の真の平均や分散など、様々な側面の影響を受ける。これらの計画と妥当な解析方法を試験開始前に決めておいても、規定した計画と異なるルールによって治療群を選択すると、統計的な妥当性を常に保証できるとは限らない。そのため、治療群の選択ルール、中間解析の回数と時期などのアダプティブ臨床試験の計画を試験開始前に決めておくだけでなく、計画を守って試験を実施することが重要である。



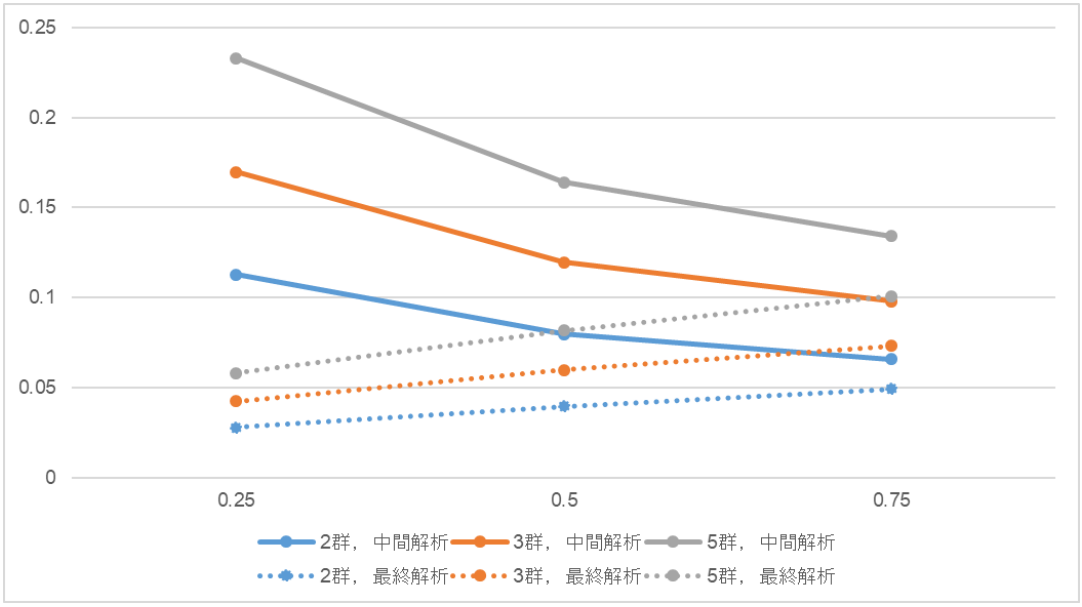
横軸：情報分数，縦軸：第一種の過誤確率（%）

図 8　第一種の過誤確率のシミュレーション結果



横軸：情報分数，縦軸：最終解析の被覆確率（%）

図 9　95%信頼区間の被覆確率のシミュレーション結果



横軸：情報分数，縦軸：バイアス

図 10　点推定値のバイアスのシミュレーション結果

3.4　試験例

3.4.1　ESCAMI：急性心筋梗塞患者の比較試験

ESCAMI 試験は、血栓溶解療法または ST 上昇型急性心筋梗塞に対する初回血管形成術を受けている患者を対象に、eniporide を試験治療とした二重盲検プラセボ対照ランダム化比較第二相試験であり、2 ステージのアダプティブデザインを採用している。

試験計画は Zeymer et al. (2001)[3]、試験結果は Zeymer et al. (2001)[4]から報告されている。Bauer et al. (2016)[5]は、ESCAMI 試験が治療群の選択のアダプティブデザインとして初めて注意深く計画した試験であると紹介している。これらの文献を参照し、ESCAMI 試験の計画と結果を以下にまとめる。

有効性の主要評価項目は α-HDBH（alpha-hydroxybutyrate dehydrogenase）の AUC（投与後 72 時間の α-HDBH の累積放出）によって測定される梗塞サイズ、副次評価項目は CK（クレアチンキナーゼ）の AUC、CK-MB の AUC などであった。中間解析の目的を 3 つ設定しており、1 つめは有効性の評価項目の初めてのエビデンスを得ること、2 つめは第 2 ステージに進む用量を選択すること、3 つめは第 2 ステージにて集積する患者数を決定することであった。治療群の選択のアダプティブデザインとしては、2 つめの目的が相当する。なお、3 つめの目的として症例数再推定も計画している。

中間解析の結果から 1 つまたは 2 つの用量群を選択し、最終解析にて、選択された用量群と対照群とを比較することを計画した。文献から治療群の選択ルールの詳細を読み取ることはできなかった。中間解析では線形傾向の用量反応について片側検定を実施し、p 値が 0.0080 未満の場合、第 1 ステージのみで明らかな有効性のエビデンスを示したとして試験を中止し、p 値が 0.7（Bauer et al. (2016) [5]では 0.5 と記載されている）より大きい場合、特定の用量にて有効性が示される可能性が低いとして試験を中止することとした。第 2 ステージに進む場合、中間解析にて選択された用量群とプラセボ群との差を参照し、90%の検出力を満たすように症例数再推定を行うこととしていた。さらに第 2 ステージにて、臨床的に適切な用量（詳細は不明）を追加することも可能とする計画をしていた。最終解析の主要な検定は、すべての群の積仮説に対して第 1 ステージと第 2 ステージを統合したデータについて Fisher の統合検定[6]を用いて有意であった場合、最も有効である 1 つの治療群とプラセボ群の比較を Fisher の統合検定にて比較する計画であった。

第 1 ステージと第 2 ステージそれぞれの平均と標準偏差の結果を、それぞれ表 12 と表 13 に示した。第 2 ステージの結果は、第 2 ステージにて集積された被験者を対象にしており、第 1 ステージの被験者を含めていない。いずれの評価項目も、数値が低いほど治療効果が高いことを意味する。中間解析の結果、傾向性検定は片側 $p = 0.12$ であった。点推定値は 150mg 群が最も低く、50mg 群と 200mg 群はプラセボ群との差が小さかった。有効性の副次解析と安全性の観点から、試験実施計画書にて事前に決めたルールに基づき 100mg 群と 150mg 群が選択され、第 2 ステージに進んだ。中間解析での傾向性検定を条件付き過誤関数に適用し、第 2 ステージの必要例数を各群 316 名と再推定した。最終解析では、全

体の積仮説に対する Fisher の統合検定が有意でなかったため、個々の治療群とプラセボ群を比較する検定は実施されなかった。

表 12　ESCAMI 試験の第 1 ステージの結果（平均±SD）

	プラセボ (N=88)	Eniporide 50mg (N=86)	Eniporide 100mg (N=91)	Eniporide 150mg (N=74)	Eniporide 200mg (N=91)
α-HDBH AUC (U/ml×h)	44.2 ± 26.0	45.3 ± 31.8	40.2 ± 22.5	33.9 ± 20.5	43.9 ± 27.0
CK AUC (U/ml×h)	70.9 ± 56.5	69.5 ± 54.9	58.1 ± 39.2	48.1 ± 36.0	69.5 ± 49.2
CK-MB AUC (U/ml×h)	6.8 ± 7.0	6.0 ± 5.3	5.9 ± 5.3	4.5 ± 4.1	6.5 ± 5.2

表 13　ESCAMI 試験の第 2 ステージの結果（平均±SD）

	プラセボ (N=322)	Eniporide 100mg (N=321)	Eniporide 150mg (N=316)
α-HDBH AUC (U/ml×h)	41.2 ± 28.5	40.2 ± 22.5	33.9 ± 20.5
CK AUC (U/ml×h)	70.9 ± 56.5	58.1 ± 39.2	48.1 ± 36.0
CK-MB AUC (U/ml×h)	6.8 ± 7.0	5.9 ± 5.3	4.5 ± 4.1

3.4.2　INHANCE：慢性閉塞性肺疾患患者の比較試験

INHANCE 試験は、慢性閉塞性肺疾患（Chronic Obstructive Pulmonary Disease、COPD）の患者を対象に、indacaterol を試験治療、プラセボと formoterol と tiotropium を対照治療（formoterol と tiotropium は陽性対照）として実施された並行群間比較試験である。INHANCE 試験は 2 ステージのアダプティブデザインを採用している。第 1 ステージの目的は、第 2 ステージにて検証する indacaterol の用量を選択することである。試験全体の主目的は、第 1 ステージにて選択された 1 つ以上の indacaterol の用量とプラセボを、第 1 ステージと第 2 ステージを通して比較し、優越性を検証することである。

アダプティブデザインに関連する複数の文献にて INHANCE 試験が取り上げられている。Lawrence et al. (2014)[6]は、INHANCE 試験の計画、用量選択のガイドライン、中間解析と最終解析の結果、データモニタリング委員会の運用など、試験に関する情報を詳細に述べており、最も参考になる。New England Journal of Medicine の The Changing Face of Clinical Trials のシリーズの 1 つである Bhatt et al. (2016)[8]は、治療群の選択のアダプティブ

デザインの試験例として INHANCE 試験を挙げ、Lawrence et al. (2014)[6]を引用して試験計画の概要をまとめている。Bauer et al. (2016)[5]は、INHANCE 試験でのデータモニタリング委員会の運用を中心に報告している。Lawrence and Bretz (2014)[9]は至適用量を選択するアダプティブデザインの第三相試験の例として INHANCE 試験を挙げ、Robertson and Glimm (2019)[10]は提案する一様最小分散条件付き不偏推定量（Uniformly Minimum Variance Conditionally Unbiased Estimator；以下 UMVCUE）の方法を INHANCE 試験のデータに適用した例を示している。INHANCE 試験の中間解析の結果は Barnes et al. (2010)に、最終解析の結果は Donohue et al. (2010)[11]にて報告されており、FDA の審査レポートも参考になる[12][13]。本項は、Lawrence et al. (2014)[6]と FDA 審査レポート[12][13]を中心に、INHANCE 試験の計画及び結果をまとめる。

INHANCE 試験の第 1 ステージと第 2 ステージを通した試験全体の主目的は、慢性閉塞性肺疾患患者を対象に、12 週時点のトラフ FEV_1 を評価項目として、プラセボに対する 1 つ以上の用量の indacaterol の優越性を検証することである。試験開始時、indacaterol に 4 つの用量群を設定していた。中間解析にて indacaterol の用量を選択する 2 つの基準を設定していた。

基準 1：2 週時点のトラフ FEV_1（L）を評価項目として、プラセボとの群間差の点推定値が 0.12 L よりも大きく、かつ点推定値が formoterol 群と tiotropium 群よりも大きい。
基準 2：2 週時点の FEV_1 $AUC_{1–4\,h}$ を評価項目として、点推定値が formoterol 群と tiotropium 群よりも大きい。

2 つの評価項目トラフ FEV_1 と FEV_1 $AUC_{1–4\,h}$ は、数値が大きいほど効果が高いことを意味している。第 2 ステージに進む用量を選択するルール、両方の基準を満たす用量が 2 つ以上ある場合、両基準を満たす最低の用量と次に高い用量が次に進むなど、2 つの基準を満たす状況に応じて事前に決めていた。治療群の選択ルールの詳細は、Lawrence et al. (2014)[6]を参照されたい。なお、中間解析の目的は治療群選択のみであり、症例数再推定は計画していなかった。中間解析は 770 名（各群 110 名）が 2 週間の治療を終えたときとした.

中間解析の結果を表 14 に示す。Indacaterol 75μg 群の点推定値が formoterol 群よりも小さく基準 2 を満たさなかったが、それ以外の用量群は両基準を満たした。選択ルールに基づき、両基準を満たす最低用量である 150μg 群と次に多い用量群である 300μg 群が第 2 ステージに進むこととした。

表 14　INHANCE 試験の中間解析の結果

群	症例数	2 週時点のトラフ FEV_1（L）の最小二乗平均のプラセボ群との群間差と 95%CI	2 週時点の FEV_1 $AUC_{1-4\,h}$（L）の最小二乗平均のプラセボ群との群間差と 95%CI
Indacaterol 75μg	104	0.15 (0.90, 0.20)	0.20 (0.14, 0.27)
Indacaterol 150μg	105	0.18 (0.12, 0.24)	0.23 (0.16, 0.29)
Indacaterol 300μg	110	0.21 (0.15, 0.27)	0.28 (0.22, 0.34)
Indacaterol 600μg	108	0.20 (0.14, 0.25)	0.23 (0.17, 0.29)
Formoterol	105	0.11 (0.06, 0.17)	0.22 (0.16, 0.28)
Tiotropium	112	0.14 (0.08, 0.19)	0.19 (0.13, 0.25)
プラセボ	104	-	-

最終解析の結果を表 15 に示す。最終解析では、Bonferroni の方法を用いて有意水準を α/4（α=0.05）とした indacaterol の用量群とプラセボ群との優越性の検定を計画していた。Bonferroni の調整を用いて有意水準を α/4 としたのは、試験開始時に indacaterol が 4 用量あるためであった。より検出力の高い方法を適用することは可能であったが、試験計画が複雑になるため採用しなかったと Lawrence et al. (2014)[6]は述べている。プラセボとの群間差は、indacaterol と tiotropium のいずれも臨床的に意義のある最小の差 0.12L を上回り、indacaterol はプラセボに対して統計的に有意な優越性を示した。

表 15　INHANCE 試験の最終解析の結果

群	被験者数	12 週時点のトラフ FEV_1（L）の最小二乗平均のプラセボ群との群間差と 98.75% CI	プラセボに対する優越性を比較する両側 p 値
Indacaterol 150μg	389	0.18 (0.14, 0.22)	< 0.001
Indacaterol 300μg	389	0.18 (0.14, 0.22)	< 0.001
Tiotropium	393	0.14 (0.10, 0.18)	< 0.001
プラセボ	376	-	-

3.5　統計的推測の方法

本項は、治療群の選択のアダプティブデザインに対して提案されている仮説検定、点推定、区間推定の方法を紹介する。測定する反応に連続データを、反応の値が大きいほど効果が高いことを意味している。

3.5.1　仮説検定

治療群の選択のアダプティブデザインに適用できる検定方法として、統合検定アプローチ（2.2.2.1.1 項）と、Dunnett 検定を拡張したアプローチがある。Dunnett 検定を拡張したアプローチとして、Koenig et al. (2008)[14]と Magirr et al. (2012)[15]による提案方法がある。Friede and Stallard (2008)[16]は、Step-down Dunnett 検定、統合検定、Koenig et al. (2008) [14]による Adaptive Dunnett 検定を対象に、検出力を比較している。Adaptive Dunnett 検定は R の rpact パッケージで実行可能であり、rpact パッケージの使用例を 3.6.1 項で解説する。

3.5.1.1　Koenig et al. (2008)による Adaptive Dunnett 検定

本項では、Koenig et al. (2008)[14]により提案された Adaptive Dunnett 検定を詳細に解説する。本手法は、一段階デザインにおける閉検定手順の Step-down Dunnett 検定に、条件つき過誤確率の考え方を導入したものである

治療群の選択を行うアダプティブデザインにおける単純なアプローチは、治療群と対照群の比較に Dunnett 検定[15]を用いることである。治療群の選択が行われた場合、最終解析において選択されなかった治療群の統計量を$-\infty$として Dunnett 検定を適用することで、全体の第一種の過誤確率を α（名義水準）に保つことが可能である[18]。Dunnett 検定を閉検定手順によって改良した Step-down Dunnett 検定を用いる場合でも第一種の過誤確率は制御されるが、治療群の選択以外のアダプテーションを行うことができない。Adaptive Dunnett 検定はこの Step-down Dunnett 検定の性能を改善し、かつ治療群の選択以外のアダプテーションも適用可能な形に改善したものである。

Adaptive Dunnett 検定の手順は次の通りである。中間解析にて、1 つまたは 2 つ以上の治療群を脱落させる。ここで、治療群を選択する方法には仮定を置いておらず、中間データだけでなく外部情報を用いる意思決定も許容している点に留意する。治療群の脱落が起きた場合、積仮説の一部は脱落した群の欠測となった第 2 ステージのデータに依存するため、閉検定手順を直接適用することはできない。第 2 ステージに進んだ治療群（$\mathcal{S} \subseteq \mathcal{T}_2$）を含む積仮説は、元々計画されていた通りに検定を行う。第 1 ステージで脱落した治療群（$\mathcal{S} \nsubseteq \mathcal{T}_2$）を含む検定に関しては、元々計画されていた検定を修正する必要がある。

まず、第 1 ステージのデータに基づく条件付き過誤確率$A_{\mathcal{S}} = P_{H_{\mathcal{S}}}\left(\varphi_{\mathcal{S}} = 1 | X(\mathcal{S})\right)$を計算する。ここで、$X(\mathcal{S})$は積帰無仮説に対する第 1 ステージのデータである。条件付き過誤確率は第 1 ステージの情報のみに基づくため、脱落した群を含む積仮説においても計算可能である。次に、選択された群のみのデータに基づく p 値$q_{\mathcal{S}}$を求めて、第 2 ステージにおける積仮説$H_{\mathcal{S}}$に対する検定を定義する。最終解析において、p 値が条件付き過誤確率より小さい（$q_{\mathcal{S}} \leq A_{\mathcal{S}}$）場合、その積仮説を棄却する。この p 値$q_{\mathcal{S}}$は p-clud（Brannath, Posch, and Bauer (2002)[19]）の性質を満たしている。p-clud とは、第 1 ステージのデータを与えたもとでの第 2 ステージの p 値の条件つき分布が、標準一様分布[0, 1]と等しいかそれよりも確率

的に大きいという性質のことを指し、2 ステージデザインにおいて第一種の過誤確率が α（名義水準）以下となる。

Adaptive Dunnett 検定における条件付き過誤確率の計算方法を説明する。k 個の治療群と 1 つの対照群における、正規分布（分散既知）の観測値に対する片側の Step-down Dunnett 検定を考える。$\mathcal{S} \subseteq \mathcal{T}_i$の帰無仮説$H_{\mathcal{S}}$の検定に対する条件つき過誤関数は以下のとおり与えられる。

$$A_{\mathcal{S}} = P_{H_{\mathcal{S}}}\left(\phi_{\mathcal{S}} = 1 | X_1(\mathcal{S})\right) = P_{H_{\mathcal{S}}}\left(\max_{i \in \mathcal{S}} Z_i \geq d_{\mathcal{S}} | z_i^{(1)},\ i \in \mathcal{S}\right)$$

$$= 1 - \int_{-\infty}^{\infty} \left[\prod_{i \in \mathcal{S}} \Phi\left(d_{\mathcal{S}} \sqrt{\frac{2(n_1 + n_2)}{n_2}} - z_i^{(1)} \sqrt{\frac{2n_1}{n_2}} + x \right) \right] \phi(x) dx$$

次に、第 2 ステージに進んだ治療群のデータに基づく積仮説の p 値を求める。Koenig et al. (2008) [13]は、p 値の算出方法を 2 通り提案している。一方は、選択群の第 2 ステージのデータのみを用いる方法で、もう一方は、選択群の第 1 ステージおよび第 2 ステージの両方のデータを用いる方法である。R の rpact パッケージでは、オプションでいずれの p 値も算出可能である。最終解析において、p 値が条件付き過誤確率より小さい場合、その積仮説を棄却する。

選択群の第 2 ステージのデータのみを用いる場合、積仮説$H_{\mathcal{S}}$に対する Dunnett 調整された p 値は以下のとおり与えられる。

$$q_{\mathcal{S}}^{*} = 1 - \int_{-\infty}^{\infty} \left[\prod_{i \in \mathcal{S} \cap \mathcal{T}_2} \Phi\left(z_{\mathcal{S} \cap \mathcal{T}_2}^{(2),\ max} \sqrt{2} + x \right) \right] \phi(x) dx$$

$$q_{\mathcal{S}}^{*} = 1 \text{ if } \mathcal{S} \cap \mathcal{T}_2 = \emptyset$$

選択群の第 1 ステージおよび第 2 ステージの両方のデータを用いる場合、積仮説$H_{\mathcal{S}}$の p 値$q_{\mathcal{S}}$は下記の通り与えられる。$q_{\mathcal{S}} \leq A_{\mathcal{S}}$ならば、積仮説$H_{\mathcal{S}}$を棄却する。

$$q_{\mathcal{S}} = P_{H_{\mathcal{S}}}\left(\max_{i \in \mathcal{S} \cap \mathcal{T}_2} Z_i \geq z_{\mathcal{S} \cap \mathcal{T}_2}^{max} | z_i^{(1)},\ i \in \mathcal{S} \cap \mathcal{T}_2\right)$$

$$= 1 - \int_{-\infty}^{\infty} \left[\prod_{i \in \mathcal{S} \cap \mathcal{T}_2} \Phi\left(z_{\mathcal{S} \cap \mathcal{T}_2}^{max} \sqrt{\frac{2(n_1 + n_2)}{n_2}} - z_i^{(1)} \sqrt{\frac{2n_1}{n_2}} + x \right) \right] \phi(x) dx$$

$$q_{\mathcal{S}} = 1 \text{ if } \mathcal{S} \cap \mathcal{T}_2 = \emptyset$$

上記では中間解析の時期を事前に固定した各群n_1の観察終了後としたが、より自由に適用が可能である。また、治療群の選択以外にも、症例数再推定も組み込める。

中間解析で治療群の脱落が起きた場合、Adaptive Dunnett 検定は Step-down Dunnett 検定より性能が改善されている。Adaptive Dunnett 検定における改善点は、閉検定手順に従って基本仮説の検証の前に積仮説の検定を行う部分である。中間解析で投与群が脱落した場

合、Adaptive Dunnett 検定は積仮説の検定に脱落した群の第 1 ステージのデータも用いるため、Step-down Dunnett 検定の条件付き過誤確率は Adaptive Dunnett 検定のそれより非常に小さくなる。他の利点として、治療群の選択ルールを柔軟に設定することができ、ベイズ流など異なる決定ツールも使用可能である。また、中間解析の時期を固定する必要がなく、データ依存の方法で決めてよい。唯一の条件は、事前規定した Dunnett 検定の条件付き過誤確率が計算可能であることである。しかし、Koenig et al. (2008) [13]では検証試験の integrity を確保するため、起こりうるアダプテーションや治療群の選択のシナリオを試験開始時にプロトコルに明記しておくことを推奨している。

3.5.1.2 Friede and Stallard (2008)による検定方法の比較

Friede and Stallard (2008)[16]は、統合検定、群逐次デザイン、Step-down Dunnett 検定、Adaptive Dunnett 検定の 4 つの手法について、中間解析で選択する治療群の数を表すパラメータε（$z_i^{(1)} \geq z^{(1),max} - \varepsilon;\ \varepsilon = 0$ならば有効性が最大の 1 群のみ選択、$\varepsilon = \infty$ならば全ての治療群を選択）を設定し、いくつかのシナリオの下での第一種の過誤確率および検出力を評価した。第一種の過誤確率は、Adaptive Dunnett 検定はその定義からεの値によらずほぼ名義水準を維持していた一方、他 3 つの手法はεが一部の値のとき過度に保守的となった。検出力は、治療効果の真値の設定に応じて異なる結果となった。試験開始時点ですべての治療群が有効ではないことが想定される場合や、治療群の選択ルールが無効な治療を第 2 ステージに進めてしまう可能性が低いと思われる場合には、統合検定、群逐次デザイン、Adaptive Dunnett 検定はほとんど同一で群逐次デザインが他よりもわずかに高い検出力であり、Step-down Dunnett 検定は保守的で低い検出力であった。一方、試験開始時点で多くの治療群が有効であることが期待され、複数の治療群を第 2 ステージに進める可能性が高い場合、もしくは、安全性に対する懸念などの理由から最も有効な群以外の治療群を第 2 ステージに進める可能性が高い場合、統合検定、Adaptive Dunnett 検定、Step-down Dunnett 検定は互いに近い検出力となり、群逐次デザインより 3%程度高くなった。Friede and Stallard(2008) [16]は、1 つの手法の検出力が他より常に高いわけではないため、どの手法を選択するかについては、真の治療効果と治療群の選択ルールに関する事前信念に依存するとともに、検出力以外にも実装のしやすさやソフトの利用可能性、試験関係者の知識等にも関わってくると述べている。

3.5.2 点推定

2 ステージデザインでは、中間解析後に複数の治療群から最良の治療効果を示す治療群を選択し、次の段階で選択された治療群と対照群を比較する戦略が広く用いられる。この戦略を用いる場合、正しい補正が行われないと、選択された治療効果の点推定値はバイアスを含むことが知られている。UMVCUE は、このバイアスを正確に補正した推定値である。2 ステージデザインにおける UMVCUE を最初に提案したのは Cohen and Sackrowitz(1989)[19]であり、ステージ間で症例数が均等であるという仮定に基づく推定量で

ある。その後、ステージ間で不均等の症例数に拡張した手法（Bowden and Glimm (2008)[21]）、2 段階から多段階に拡張した手法（Bowden and Glimm (2014)[22]）が提案された。無益性による早期中止を伴うデザインへ拡張した手法（Kimani et al. (2013)[23]）、等分散性を仮定せず仮説検定による検証ステージへ移行できる試験デザインへ拡張した手法（Robertson et al. (2016)[23]）が提案され、さらに Kimani et al. (2013)[23]や Robertson et al. (2016)[24]の手法で最小分散にならない場合の状況を改善し、多段階に拡張することのできる手法（Stallard and Kimani (2018)[25]）が提案された。さらに、分散が既知ではなく未知の状況を想定した場合の手法（Robertson and Glimm (2019)[10]）も提案された。Cohen and Sackrowitz (1989)[19]が提案した UMVCUE は 3.2 項に記載する治療群の選択ルール 1 の基づく、最良の治療群を選択する方法である。本項では、Bowden and Glimm (2008) [21]は Cohen and Sackrowitz (1989) [19]の提案する UMVCUE を拡張することで、より柔軟な選択ルールに適用可能な Bowden and Glimm (2008)[21]を紹介する。

3.5.2.1 Bowden and Glimm (2008)による一様最小分散条件付き不偏推定量

Cohen and Sackrowitz (1989)[19]により最初に提案された 2 ステージデザインにおける UMVCUE は以下の式で定義される。

$$\frac{Z}{2}-\frac{1}{\sqrt{2}}\,\frac{\phi\,(W)}{\Phi(W)}$$

Z= $\bar{X}_{(1)}+\bar{Y}$、W= $\sqrt{2}\ (\frac{Z}{2}-\bar{X}_{(2)})$)、$\bar{Y}$：第 2 ステージの推定値

この UMVCUE は、Rao-Blackwell の定理を応用し、不偏推定量である第 2 ステージの推定値を各治療群の推定値を$\bar{X}_{(1)}>\cdots>\ \bar{X}_{(k)}$として条件づけ、いくつかの完備十分統計量を与えることで導出されている。Cohen and Sackrowitz (1989) [20]が提案した上記の UMVCUE は第 1 ステージと第 2 ステージの平均値の分散が等しい場合に、k 個の統計量の中で最も優れた治療に対する推定量である。

Bowden and Glimm (2008) [21]は Cohen and Sackrowitz (1989) [20]の UMVCUE を第 1 ステージと第 2 ステージの平均値の分散が不等である場合に拡張することで、任意の中間解析時点における UMVCUE の推定を可能とした。また、最も優れた治療群ではなく k 個の治療の内 j 番目に良い治療群を選択することが可能であり、有効性以外の基準に基づく治療選択も考慮している。上記条件を踏まえ、Bowden and Glimm (2008) [21]は UMVCUE の拡張法を以下の式で提案した。

$$\tilde{\mu}_{(j)}=\frac{\sigma_{2,j}^2\bar{X}_{(j)}+\sigma_{1,(j)}^2\bar{Y}_j}{\sigma_{1,(j)}^2+\sigma_{2,j}^2}-\frac{\sigma_{2,j}^2}{\sqrt{\sigma_{1,(j)}^2+\sigma_{2,j}^2}}\frac{\{\phi\left(W_{j,j+1}\right)-\phi\left(W_{j,j-1}\right)\}}{\{\Phi\left(W_{j,j+1}\right)-\Phi(W_{j,j-1})\}}$$

$\bar{X}_i(i=1,..,k)$：N(μ_i, $\sigma_{1,i}^2$)の互いに独立な正規分布に従う群 i の第 1 ステージの治療効果の推定値

$\sigma_{1,i}^2$：第 1 ステージでの群 i の治療効果の推定値の分散

$\bar{Y}_i(i = 1, .., k)$：N($\mu_{(i)}$, $\sigma^2_{2,i}$)の正規分布に従う i 番目に優れた治療効果の第 2 ステージの推定値$\sigma^2_{2,i}$：第 2 ステージでの群 i の治療効果の推定値の分散

また、$W_{s,t}$は以下の式で与えられる。

$$W_{s,t} = \frac{1}{\sigma^2_{1,(s)}}\left(\frac{\sigma^2_{2,s}\bar{X}_{(s)} + \sigma^2_{1,(s)}\bar{Y}_s}{\sqrt{\sigma^2_{1,(s)} + \sigma^2_{2,s}}} - \bar{X}_{(t)}\sqrt{\sigma^2_{1,(s)} + \sigma^2_{2,s}}\right)$$

$s = 1, .., k$、$t = 0, .., k+1$、

$\bar{X}_{(0)} := \infty > \bar{X}_{(1)} \geq \cdots \geq \bar{X}_{(k)} > \bar{X}_{(k+1)} := -\infty$、

$\sigma^2_{1,(s)}$：s 番目に優れた治療効果の分散

治療群の選択基準を推定量ではなく p 値の大きさとする場合、上記$W_{s,t}$中の$\bar{X}_{(t)}$を $\frac{\sigma_{1,(s)}\bar{X}_{(t)}}{\sigma_{1,(t)}}$と置き換えることで UMVCUE を推定することが可能である。

Bowden and Glimm (2008)[21]は、提案する UMVCUE の性能をシミュレーションにより検証した。中間解析時点を変えた場合の影響として、中間解析時点が遅いほど UMVCUE の MSE が大きくなることを示した。これは、中間解析時点が遅いほど Rao-Blackwell 化するための不偏な情報量が少ないためである。一方、MLE（第 1 ステージ及び第 2 ステージの治療効果の平均値の重み付き平均）の MSE は中間解析時点の影響を受けない。ただし、中間解析時点が早いと正しい治療を選択する確率は小さくなるため、MSE を小さくするために中間解析時点を早めることが必ずしも良いとは言えない。

Bowden and Glimm (2008)[21]は、治療群を 5 群とした場合の、k 個の治療の内 j 番目に良い治療を選択とする場合の$\mu_{(j)}$の MSE 及び分布の形状を検討している。真の治療効果の大きさが等間隔の場合及び、等間隔でない場合を検討したところ、UMVCUE の MSE は$\mu_1 = \mu_2 = \cdots = \mu_5$の場合に最大となり、一部の場合を除き、$j$ の中央値（$\mu_{(3)}$）でより大きくなる傾向が見られた。MSE が$\mu_1 = \mu_2 = \cdots = \mu_5$の場合に最大となることは Sill and Sampson (2007)[26]の結果とも整合するが、その証明はされていない。また、$\mu_1 = \mu_2 = \cdots = \mu_5 = 0$と仮定した場合の UMVCUE の 20000 回分の推定値をプロットしたところ、$\mu_{(1)}$と$\mu_{(5)}$の時 UMVCUE の分布のピークが他と比較して高くなり、なおかつ少し非対称で裾が長くなったが、それぞれ正規分布に近似された。

3.5.2.2 一様最小分散条件付き不偏推定量以外のアプローチ

これまで 2 ステージデザインにおける効果の推定量のバイアスに対処する方法として、UMVCUE を用いるアプローチを解説した。推定量のバイアスを減らすことはその分散を増加させる、すなわち推定の精度を下げることにつながることがある。不偏推定量を求めることにより標準誤差が増大し、信頼区間幅が広がることは実用的ではないかもしれない。ここでは、完全にバイアスを除去するのではなくバイアスを低減するアプローチをいくつか紹介する。なお、バイアスを低減する、またはバイアスを調整するような推定量は

反対方向にバイアスを導く、つまり過修正を引き起こす可能性があることに留意する必要がある。

Carreras and Brannath (2013)[27]は、Lindley の縮小推定量を利用し、4 つ以上の治療群を含む 2 ステージデザインに適用可能な縮小推定量を提案している。著者らはシミュレーションを用いて、提案法と Cohen and Sackrowitz (1989)[19]による UMVCUE 及び Stallard and Todd (2005)[28]によるバイアスを低減する推定量を比較した。シミュレーションの結果、Cohen and Sackrowitz (1989)[19]による UMVCUE は選択バイアスを完全に除去可能であるが、MLE よりも MSE が大きくなった。Stallard and Todd (2003) [29]による推定量はバイアスを過補正する傾向があり、MSE に関しても MLE 及び縮小推定量に劣る結果であった。著者ら提案の縮小推定量は、選択バイアスを減らすことを可能とし、MSE に関しても MLE と同等かもしくはより改善すると述べている。

Pickard and Chang (2014) [30]は、正規分布及び二項分布に従う評価項目における 2 ステージデザインに対してパラメトリックブートストラップ法によりバイアスを低減することを提案している。著者らはシミュレーションを用いて、提案法、MLE 及び正規分布に従う評価項目に対しては Carreras and Brannath (2013)[27]による縮小推定量及び Bowden and Glimm (2008) [21]による UMVCUE、二項分布に従う評価項目に対しては Tappin (1992)[31]による UMVCUE との比較をバイアス及び MSE の観点から評価した。真の平均が等しい正規分布の場合、提案法は MLE と比較してバイアス及び RMSE を低減し、Carreras and Brannath (2013)[27]による縮小推定量と同程度のバイアスを示した。真の平均が異なる場合、提案法は Carreras and Brannath (2013)[27]による縮小推定量よりもバイアスを低減し、RMSE は同程度であった。二項分布の場合にも同様の結果が得られた。

3.5.3 区間推定

本項では、治療群の選択のアダプティブデザインに対して提案されている、区間推定の方法を解説する。

Stallard and Todd (2005) [28]は、標本空間順序と p 値の構成に基づく同時信頼区間を提案している。Sampson and Sill (2005)[32]は、一様最強力条件付き不偏検定に基づく信頼区間を提案している。Posch et al. (2005)[33]は、統合検定に対応する同時信頼区間を提案している。Bowden and Glimm (2008)[21]は UMVCUE アプローチによる点推定を提案するとともに、最大の分散にて信頼区間を構成する方法を述べている。Wu et al. (2010)[34]は、確率順序アプローチに基づく方法を提案した。 Neal et al. (2011)[35]は、治療群の選択ルール 1 に基づく Wu et al. (2010) [34]の方法を、より柔軟な選択ルール 2 へ一般化している。Magirr et al. (2013)[36]は、Posch et al. (2005)[33]の方法が閉検定手順に対応していない点を改善している。信頼区間を計算する群が選択した群のみの方法では、選択されずに中間解析で中止した群には適用できないことに注意が必要である。また、ステージ数が 3 以上の場合に拡張した方法として、Bowden and Glimm (2014)[22]、Lu et al. (2018)[37]がある。治療群の選択ルー

ルと、区間推定の対象とする群（ルールに基づいて選択した群、又は選択されなかった群を含めたすべての群）について、区間推定の方法を表 3.4 に提案法をまとめた。

表 16　区間推定の方法のまとめ

治療群の選択ルール	区間推定をする群	
	すべての群	選択した群のみ
ルール 1	Stallard and Todd (2005) [28]	Sampson and Sill (2005)[32]、 Wu et al. (2010)[34]
ルール 2	Posch et al. (2005)[33]、 Bowden and Glimm (2008)[21]、 Magirr et al. (2013)[36]*	Neal et al. (2011)[35]*

* 区間下限のみを計算。

3.5.3.1　Kimani et al (2014)による区間推定方法の比較

本項では、信頼区間の構成方法の特性を比較した論文 Kimani et al. (2014)[38]の概要を解説する。Kinami et al. (2014) [38]では治療群の選択ルール 1 を前提として、Sampson and Sill (2005)[32]、Wu et al. (2010)[34]、Stallard and Todd (2005) [28]、Posch et al. (2005)[33]の Dunnett 補正（以下 Posch et al. with Dunnett）、Posch et al. (2005)[33]の Sidak 補正（以下 Posch et al. with Sidak）及び未調整（以下ナイーブ）の計 6 つの信頼区間の特性を、被覆確率及び真の群間差を上回る確率と下回る確率の観点からシミュレーションによって評価している。なお、治療群の選択ルール 1 の下では、Neal et al. (2011)[35]は Wu et al. (2010)[34]、Magirr et al. (2013)[36]は Posch et al. (2005)[33]と同様な信頼区間の構成方法であるため、Kinami et al. (2014) [38]はこれらを比較する方法に含めていない。シミュレーションの設定として、治療群の数を 2 から 4、共通分散を 1、被験者数を対照群 400 人、最終的に選択された治療群 400 人、情報時間を 0.2 から 0.8 及びシミュレーション回数を 10000 回としている。群間差は 0 から 0.25 付近を想定した結果を主に載せている。

Kinami et al. (2014) [38]によるシミュレーション結果の考察、ナイーブの信頼区間は第 1 ステージでの治療群の数が多いほど、そして情報時間が大きいほど上記に挙げた特性評価の観点から不適切であることを示している。6 つの信頼区間の中では Sampson and Sill (2005)[32]の信頼区間が最も良い性能であったが、他の方法もほとんどのシナリオで良く機能することを述べている。下限が 0 を上回る確率（以下、検出力）の観点では、Wu et al. (2010)[34]と Posch et al. (2005)[33] with Dunnett がよい構成方法であることが示された。また、下限を検出力の高い Wu et al. (2010)[34]または Posch et al. (2005)[33]の構成方法とし、上限は Sampson and Sill (2005)[32]の方法とするハイブリッドな方法もより良い信頼区間の構成方法と考えられる、というアイデアを述べている。

Kinami et al. (2014) [38]は、第 1 ステージで終わるということなく必ず第 2 ステージに進むこと、第 1 ステージと第 2 ステージの症例数が固定であることを仮定している。

Sampson and Sill (2005)[32]の方法は、これらの仮定でのみ信頼区間が得られることに注意されたい。また、Stallard and Todd (2005) [28]の方法は、第 1 ステージの中間解析の結果に基づく無益性又は有効性による早期中止を組み込んだ場合でも（ただし、症例数再推定は許容していない）信頼区間を構成できる。Posch et al. (2005)[33]と Wu et al. (2010)[34]の方法は、第 1 ステージの中間解析の結果に基づく無益性又は有効性による早期中止と症例数再推定の両方を組み込むことができることが述べられている。

3.5.3.2 Sampson and Sill (2005)による信頼区間

本項は、Kinami et al. (2014) [38]にて治療群の選択ルール 1 の状況で推奨された方法である Sampson and Sill (2005)[32]を説明する。

仮説 H_0: $\mu_{(1)} - \mu_0 \leq \Delta_{(1)0}$ vs H_1: $\mu_{(1)} - \mu_0 > \Delta_{(1)0}$ を考える。$\Delta_{(1)0}$ は中間解析の結果から最も効果が大きい群と対照群との群間差の優越性マージンを示している。ナイーブなアプローチとして、検定は下記の検定統計量$\bar{Z} - \bar{Y_0}$に基づく。$\bar{Z}$はナイーブな統計量 $\bar{Z} = \frac{Z}{n_1+n_2} = \frac{n_1\bar{X}_{(1)}+n_2\bar{Y}}{n_1+n_2}$、$n_0$は対照群の第 1 ステージと第 2 ステージを併せた症例数、$\bar{Y_0}$は第 1 ステージと第 2 ステージを併合した対照群の推定値である。

$$\bar{Z} - \bar{Y_0} \sim N\left(\Delta_{(1)0}, \sigma^2\left(\frac{1}{n_1+n_2} + \frac{1}{n_0}\right)\right)$$

この帰無仮説は以下を満たすと棄却される。

$$\bar{Z} - \bar{Y_0} > \Delta_{(1)0} + \Phi^{-1}(1-\alpha)\sigma\sqrt{\frac{1}{n_1+n_2} + \frac{1}{n_0}}$$

このナイーブな分布は第一種の過誤確率を増大させる可能性があり、対応する 1 つの方法として Bonferroni の方法を使うことが考えられ、つまり$\Phi^{-1}(1-\alpha/k)$とすることが一つの方法である。

第 2 ステージのデータだけを用いる別の検定も考えられる。

$$\bar{Y} - \bar{Y_0} \sim N\left(\Delta_{(1)0}, \sigma^2\left(\frac{1}{n_2} + \frac{1}{n_0}\right)\right)$$

これは、治療群の第 1 ステージのデータを含めず、第 2 ステージのデータに基づく検定である。ただし、対照群は第 1 ステージと第 2 ステージの両方のデータを用いている。これは第一種の過誤確率の問題はないが、検出力に欠けてしまう問題がある。

Sampson and Sill (2005)[32]は$\delta_{(1)}$や$\mu_{(1)}$のパラメータについての一様最強力条件付き不偏検定（uniformly most powerful conditionally unbiased test, UMPCU）を開発し、検定から信頼区間を構成する方法を提案した。UMPCU に基づくアプローチでは$Q = \{X: \bar{X}_1 > \bar{X}_2 > \cdots > \bar{X}_k\}$を仮定した場合、条件付き確率密度関数 W は次のようになる。

$$f_Q\left(W|\delta_{(1)},X^*,T\right) = \ C_N\exp\left\{-\frac{\sigma^2}{n_0}\left(W-\frac{n_0\delta_{(1)}}{2\sigma^2}\right)\right\}\Phi\left\{\frac{\sqrt{n_1}\left(\sigma^2W+n_0\left(T-\bar{X}_{(2)}\right)\right)}{\sigma\sqrt{n_0n_2}}\right\}$$

ここでX^*は$\{X^*:\bar{X}_2>\cdots>\bar{X}_k\}$、T は$(\bar{Z}-\overline{Y_0})/2$、$C_N$は正規化定数、W は$n(\bar{Z}-\bar{Y}_0)/2\sigma^2$である。この式は$X^*,T$の条件付き分布に基づいた$\Delta_{(1)}$の UMPCU であると著者は示している。選択された治療群と対照群との群間差の$100(1-\alpha)$%信頼区間は、$f_Q(\cdot)$の累積分布関数の上限と下限の$100\times\alpha/2$%点から構成される。

3.5.3.3 Bowden and Glimm (2008)による信頼区間

より柔軟な治療群の選択ルールに適用できる信頼区間として、3.5.2.1 項で解説した、Bowden and Glimm (2008)[21]の UMVCUE アプローチを用いた信頼区間を紹介する。記法は 3.5.2.1 項を参照されたい。本タスクフォースが作成した、この方法を実行する SAS コードを 3.6.2 項に示した。

$\mu_{(j)}$の信頼区間は$\tilde{\mu}_{(j)}$の漸近正規性、及びその分散の上限が$\mu_1=\mu_2=\cdots=\mu_k$の場合に得られるという仮定の下で、提案法の UMVCUE に対する保守的な$100(1-\alpha)$%信頼区間は以下の式で与えられる。

$$\tilde{\mu}_{(j)}\pm\Phi^{-1}(1-\ \alpha/2)V$$

V^2は$\mu_1=\mu_2=\cdots=\mu_k$の場合に得られる最大の点推定値の分散である。治療群が 5 群の場合のシミュレーション結果では、$\mu_{(1)}$と$\mu_{(5)}$の信頼区間の被覆確率は名目水準をわずかに上回るが、$\mu_{(2)}$, $\mu_{(3)}$, $\mu_{(4)}$の信頼区間ではわずかに名目水準を下回る結果であった。治療群間にて真の治療効果に若干の差がある場合、すべての推定量の被覆確率は名目上に保たれる、もしくは上回っていた。

3.6 解析プログラム

3.6.1 R による rpact パッケージ

本項では、R パッケージの 1 つである rpact を用いて、治療群の選択のアダプティブデザインを実装する R コードの解説について要約する。また、具体的な実装例として二値のエンドポイントを持つ試験デザインについて紹介する。本項で解説する内容は、2.2.6 項でも既に例示されている通り、rpact のサイト[39]で事例として掲載されているので、合わせて参照いただきたい。方法や理論的背景については、Wassmer and Brannath (2016)[40]を参照いただきたい。

実装の前に、rpact のパッケージを読み込む。

```
library(rpact)
```

［試験デザインの設定］

3 つのステージから構成される試験（中間解析を 2 回、最終解析を 1 回実施する試験）で対照群に対する 2 つの治療群の治療効果の比較を想定する。$\alpha = 0.025$、治療群の比較の検出力は 80%に設定する。さらに、O'Brien and Fleming 型の α 消費関数を使用して棄却限界値を計算し、各ステージの情報量は均等に分散されていると仮定する。さらに、拘束力のない（non-binding）無益性の限界は、二値の治療効果の相対的な割合の減少が第 1 ステージで 5%しかない場合（つまり、比率が 0.95）と、第 2 ステージで 10%しかない場合（つまり、比率が 0.9）と設定する。簡単にするために、無益の範囲は独立していると想定する。rpact は主に z スケールの無益性の境界を使用するため、上記の無益性の限界値に対応する z スケール値を与える必要がある。z スケールでの（おおよその）無益性の限界値として、第 1 ステージは 0.14915、第 2 ステージは 0.41381 と計算され、これらの無益境界を加味した試験デザインを次のように設定する。kMax= 3 は、3 ステージの試験デザインを示す。なお、getDesignConditionalDunnett()関数を用いることで Adaptive Dunnett による多重性の調整を行うデザインを検討することができる。

```
#GSD with futility bounds according to above calculations
d <- getDesignGroupSequential(
      kMax=3,
      alpha=0.025,
      beta = 0.2,
      sided=1,
      typeOfDesign = "asOF",
      informationRates=c(1/3, 2/3, 1),
      futilityBounds=c(0.149145,0.41381),
      bindingFutility=FALSE)
summary(d)
```

実行結果を以下に示す.

```
## Sequential analysis with a maximum of 3 looks (group sequential design)
##
## O'Brien & Fleming type alpha spending design, non-binding futility,
## one-sided overall significance level 2.5%, power 80%, undefined endpoint.
##
## Stage                                1      2      3
## Information rate                     33.3%  66.7%  100%
## Efficacy boundary (z-value scale)    3.710  2.511  1.993
## Futility boundary (z-value scale)    0.149  0.414
```

## Cumulative alpha spent	0.0001	0.0060	0.0250
## Overall power	0.0213	0.4471	0.8000

定義されたオブジェクトを出力することで、関連するすべての試験デザインパラメータの概要がわかる。最終ステージの調整済み α（= 0.0231、下記の症例数設計の項の「One-sided local significance level」を参照）は、事前定義された α よりもわずかに低く、1.993 の棄却限界値が 1.96 よりわずかに大きいこと。これは、O'Brien and Fleming 型の α 消費関数の調整によるものである。さらに、基本的な入力パラメータの概要が出力される。

［症例数設計］

二値のエンドポイントを持つ試験デザインの症例数計算に関しては、rpact はコマンド getSampleSizeRates()を提供している。rpact での症例数の計算は、常に 1 つの投与群の比較のみを参照することに注意していただきたい。このセクションでは、二値のエンドポイントを使用した治療群の選択のアダプティブデザインでの症例数の計算についてさらに詳しく説明する。

H_0 での治療効果が 0 であるか、治療群での発現率が増加していると仮定する。π_1 を治療群の想定有効割合を表し、π_2 は対照群の想定有効割合とすると、$\pi_2 - \pi_1 \geq 0$ を意味する。ここでは、$\pi_2 = 0.1$ が与えられた場合に、イベント発生の予想される相対減少が 50%であると想定する（つまり $\pi_1 = 0.05$）。リスクを直接比較する場合は、riskRatio を TRUE に設定する。これにより、特に $H_1 : \pi_1 / \pi_2 < 1$ に対して $H_0 : \pi_1 / \pi_2 \geq 1$ が検定される。この場合、以下のコマンドを使用して、1 つの投与群を比較するためのステージごとの症例数を計算できる。

```
c_rate <- 0.1 #assumed rate in control
effect <- 0.5 #relative reduction that is to be detected with probability of 0.8

#rates indicate binary endpoint
d_sample <- getSampleSizeRates(
      design=d,
      riskRatio = TRUE,
      pi1 = c_rate*(1-effect),
      pi2 = c_rate)
summary(d_sample)
```

実行結果を以下に示す.

```
## Sample size calculation for a binary endpoint
```

```
##
## Sequential analysis with a maximum of 3 looks (group sequential design), overall
## significance level 2.5% (one-sided).
## The sample size was calculated for a two-sample test for rates
## (normal approximation),
## H0: pi(1) / pi(2) = 1, H1; treatment rate pi(1) = 0.05, control rate pi(2) = 0.1,
## power 80%.
##
## Stage                                  1       2       3
## Information rate                       33.3%   66.7%   100%
## Efficacy boundary (z-value scale)      3.710   2.511   1.993
## Futility boundary (z-value scale)      0.149   0.414
## Overall power                          0.0213  0.4471  0.8000
## Expected number of subjects            751.7
## Number of subjects                     313.8   627.5   941.3
## Exit probability for futility          0.0625  0.0108
## Cumulative alpha spent                 0.0001  0.0060  0.0250
## One-sided local significance level     0.0001  0.0060  0.0231
## Efficacy boundary (t)                  0.061   0.476   0.643
## Futility boundary (t)                  0.950   0.903
## Overall exit probability (under H0)    0.5594  0.1828
## Overall exit probability (under H1)    0.0838  0.4366
## Exit probability for efficacy (under H0) 0.0001 0.0059
## Exit probability for efficacy (under H1) 0.0213 0.4258
## Exit probability for futility (under H0) 0.5593 0.1769
## Exit probability for futility (under H1) 0.0625 0.0108
##
## Legend:
##      (t): treatment effect scale
```

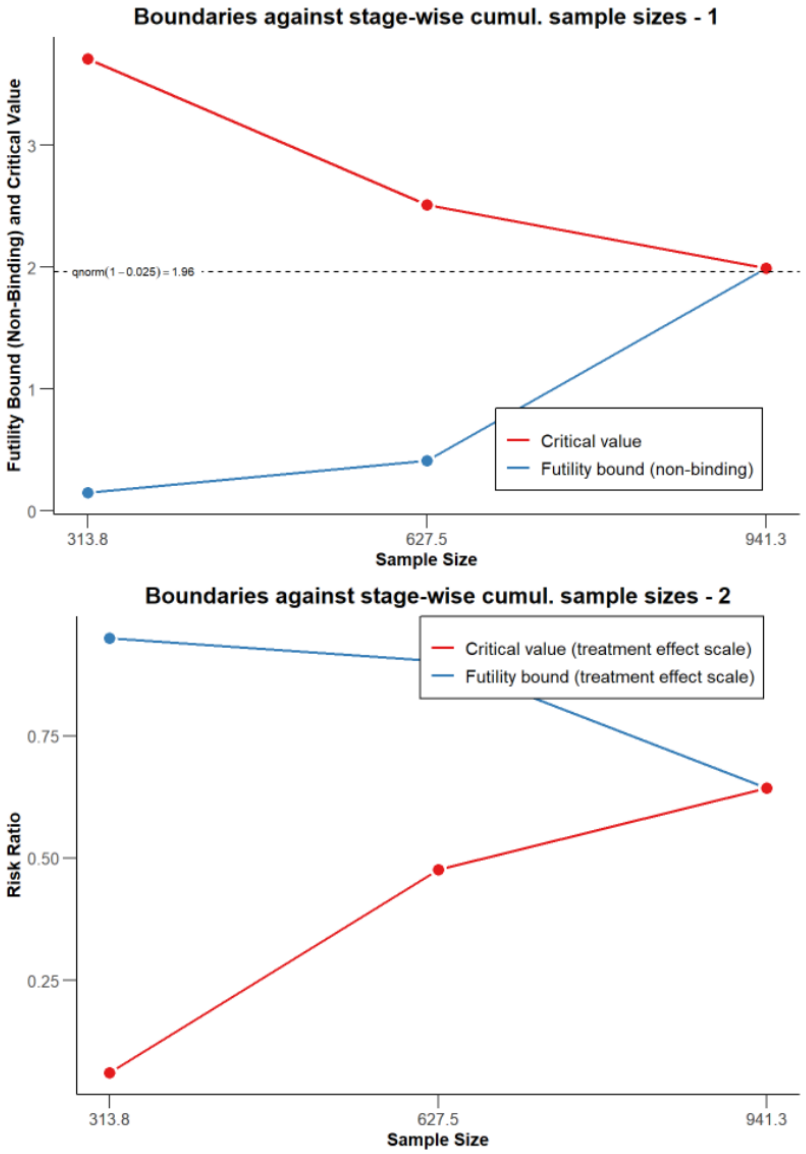
「Futility boundary (t) 」は、治療効果スケールに変換された無益性の境界を表す。したがって、上記で計算し事前定義された境界は、第 1 ステージで 0.950 の比率（つまり、5%の相対減少）と第 2 ステージで 0.903 の比率（つまり、約 10%の相対減少）に対応することがわかる。実際に意図したものにかなり近い値である。上記の結果とのわずかな違いは、無益の範囲に仮定された独立性によるものである。ただし、その単純化をしても、少なくとも 3 桁の精度の値を得ることができる。

出力によって得られるもう 1 つの重要な情報は、αエラーを 0.025 に制御しながら、$1-\beta = 0.8$ の検出力を確保するなど、試験の目的を達成することである。「Number of subjects」に出力されている値は、2 群比較した場合の 2 群の必要合計症例数が出力されているため、群ごとには 314 / 2 = 157 の症例が必要となる。したがって、3 つの投与群（2 つは治療群、1 つは対照群）を使用した試験デザインを想定した場合、1 つの対照群と比較して、3 つのステージと 2 つの治療群があるため、最大の被験者は 3 群×3 ステージ×157 = 1413 となる。ここで、これは $1-\beta = 0.8$ の検出力を達成するために必要な症例数の近似値にすぎないことに注意していただきたい。ここでの検出力は、想定される効果を正常に検出する確率として定義され、複数の治療群を使用した試験のシミュレーションを検討する場合は通常、検出力は少なくとも 1 つの治療群で有意となる確率である。

無益性による早期中止の確率「Exit probability for futility (under H1)」はかなり低いことが示されている（第 1 ステージ 0.0625、第 2 ステージ 0.108）が、これは想定される治療効果と定義された境界に依存する。「Efficacy boundary (t) 」は、第 1 ステージで H_0 を棄却するために治療群でイベントの大幅な減少を検出する必要があることを示している（発現率比 0.061、つまり相対減少 1 − 0.061 = 0.939）。これもまた、第 1 ステージでの O'Brien and Fleming 型の α 消費関数による保守性を反映しているが、ステージの特徴に応じてかなり自由に単調減少な値を設定できる。

```
#boundary plots
par(mar = c(4, 4, .1, .1))
plot(d_sample, main="Boundaries against stage-wise cumul. sample sizes - 1", type=1)
plot(d_sample, main="Boundaries against stage-wise cumul. sample sizes - 2", type=2)
```

実行結果を以下に示す.



適用可能な棄却限界値に対して累積症例数をプロットすることは、重要な試験特性をうまく視覚化する方法である。上の図では、y 軸を z スケールの境界を示したプロットを見ることができる。破線は、片側検定と $\alpha = 0.025$ を使用した固定標本試験の棄却限界値を表す。下の図には基本的に上の図と同じ情報が含まれているが、y 軸は治療効果のスケールで示されている。赤い線は、該当するステージで統計的有意性を得るために交差する必要がある有効性の限界を示し、青い線は無益性の限界を示す。

［試験デザインのシミュレーション］

試験を実施する前にシミュレーションを実行することは、さまざまなシナリオまたは治療効果を想定して、検出力（上記の定義を参照）などの特性を評価できるため、多くの場合合理的である。したがって、シミュレーションにより、さまざまなシナリオまたは治療

効果の想定が仮に正しいとした場合に、計画された試験がどこにどのように進むことができるかについて、より全体的な見方を得ることができる。

試験計画のシミュレーションは、関数 getSimulationMultiArmRates()を使用して実行される。

```
#design as above, just as inverse normal
d_IN <- getDesignInverseNormal(
      kMax=3,
      alpha=0.025,
      beta = 0.2,
      sided=1,
      typeOfDesign = "asOF",
      informationRates=c(1/3, 2/3, 1),
      futilityBounds=c(0.149145,0.41380800),
      bindingFutility=FALSE)
```

例えば、検出力や成功確率の評価のために試験シミュレーションを実行するには、2 つの治療群で（異なる）有効割合を想定する必要がある。これは、行列オブジェクトで定義する。二値データの場合、行列は各群の実際の有効割合を参照するが、対照群と治療群の有効割合の差は参照しない。さらに、反復回数を事前に定義する必要がある。例えば、シミュレーション回数を 10000 回、2 つの試験治療における有効割合を 3 のシナリオ（シナリオ 1：[0.1, 0.1], シナリオ 2：[0.05,0.05], シナリオ 3：[0.055, 0.045]）について検討する場合、以下の様に指定する。

```
#set number of iterations to be used in simulation
maxNumberOfIterations <- 10000

#specify the scenarios, nrow: number of scenarios, ncol: number of treatment arms
effectMatrix <- matrix(c(0.100, 0.100, 0.05, 0.05, 0.055, 0.045),
      byrow = TRUE, nrow = 3, ncol = 2)

#first column: first treatment arm, second column: second treatment arm
show(effectMatrix)
```

```
##          [,1]  [,2]
## [1,] 0.100 0.100
```

```
## [2,] 0.050 0.050
## [3,] 0.055 0.045
```

対照群と比較される 2 つの治療群を備えた試験デザインを考慮すると、行列には、それぞれ最初の列の最初の治療群の有効割合と、2 番目の列の 2 番目の治療群の有効割合を含める。

次のステップで実際にシミュレーションを実行する。必要に応じてシミュレーションを設定するには、いくつかの変数を設定する必要がある。いくつかのオプションについて設定方法を紹介するが、さらに細かな各オプションの設定については R package の説明を参照していただきたい。

引数	説明
directionUpper	仮説の向きを指定する。デフォルトは TRUE で、検定統計量の値が大きくなるほど p 値が小さくなることを意味する。本項では率を下げることが有効であるため、FALSE に設定する。
intersectionTest	交互作用の検定に使用する方法。”Bonferroni”, “Dunnett”, “Simes”が選択できる。
typeOfSelection	治療群の選択の方法を設定する。"best", "rBest", "epsilon", "all", "userDefined"が選択できる。userDefined を選択することで "selectArmsFunction" あるいは "selectPopulationsFunction"の設定が必要になるが、ここでは治療群で想定される最大効果のベクトル (piMaxVector)が参照される。一般的に userDefined を使用することが推奨される。
successCriterion	有効性の早期中止の条件。"all", “atLeastOne”が選択できる。all の場合は中間解析時点のすべての投与群が有効性を示す必要がある。
plannedSubjects	計画段階での各ステージの 1 群辺りの症例数。

```
#first simulation
simulation <- getSimulationMultiArmRates(
      design = d_IN,
      activeArms = 2,
      effectMatrix = effectMatrix,
      typeOfShape = "userDefined",
      piControl = 0.1,
      intersectionTest = "Simes",
```

```
        directionUpper = FALSE,
        typeOfSelection = "rBest",
        rValue=2,
        effectMeasure = "testStatistic",
        successCriterion = "all",
        plannedSubjects = c(157, 314, 471),
        allocationRatioPlanned = 1,
        maxNumberOfIterations = maxNumberOfIterations,
        seed = 145873,
        showStatistics = TRUE)
summary(simulation)
```

```
## Simulation of a binary endpoint (multi-arm design)
##
## Sequential analysis with a maximum of 3 looks
## (inverse normal combination test design), overall significance level 2.5%
## (one-sided).
## The results were simulated for a multi-arm comparisons for rates
## (2 treatments vs。 control),
## H0: pi(i) - pi(control) = 0, power directed towards smaller values,
## H1: treatment rate pi_max as specified, control rate pi(control) = 0.1,
## planned cumulative sample size = c(157, 314, 471), intersection test = Simes,
## effect shape = user defined, selection = r best, r = 2,
## effect measure based on test statistic, success criterion: all,
## simulation runs = 10000, seed = 145873.
##
## Stage                                1      2      3
## Fixed weight                         0.577  0.577  0.577
## Efficacy boundary (z-value scale)    3.710  2.511  1.993
## Futility boundary (z-value scale)    0.149  0.414
## Reject at least one [1]              0.0181
## Reject at least one [2]              0.8550
## Reject at least one [3]              0.8695
## Rejected arms per stage [1]
##    Treatment arm                     0.0001 0.0031 0.0087
##    Control arm                       0 0.0024 0.0079
```

## Rejected arms per stage [2]			
## Treatment arm	0.0086	0.3839	0.3954
## Control arm	0.0080	0.3796	0.3996
## Rejected arms per stage [3]			
## Treatment arm	0.0043	0.3090	0.3968
## Control arm	0.0145	0.4650	0.3686
## Success per stage [1]	0	0.0007	0.0034
## Success per stage [2]	0.0029	0.2735	0.4437
## Success per stage [3]	0.0014	0.2645	0.4228
## Exit probability for futility [1]	0.5797	0.1920	
## Exit probability for futility [2]	0.0482	0.0068	
## Exit probability for futility [3]	0.0429	0.0064	
## Expected number of subjects [1]	776.2		
## Expected number of subjects [2]	1232.8		
## Expected number of subjects [3]	1243.7		
## Overall exit probability [1]	0.5797	0.1927	
## Overall exit probability [2]	0.0511	0.2803	
## Overall exit probability [3]	0.0443	0.2709	
## Stagewise number of subjects [1]			
## Treatment arm 1	157.0	157.0	157.0
## Treatment arm 2	157.0	157.0	157.0
## Control arm	157.0	157.0	157.0
## Stagewise number of subjects [2]			
## Treatment arm 1	157.0	157.0	157.0
## Treatment arm 2	157.0	157.0	157.0
## Control arm	157.0	157.0	157.0
## Stagewise number of subjects [3]			
## Treatment arm 1	157.0	157.0	157.0
## Treatment arm 2	157.0	157.0	157.0
## Control arm	157.0	157.0	157.0
## Selected arms [1]			
## Treatment arm	1.0000	0.4203	0.2276
## Control arm	1.0000	0.4203	0.2276
## Selected arms [2]			
## Treatment arm	1.0000	0.9489	0.6686
## Control arm	1.0000	0.9489	0.6686

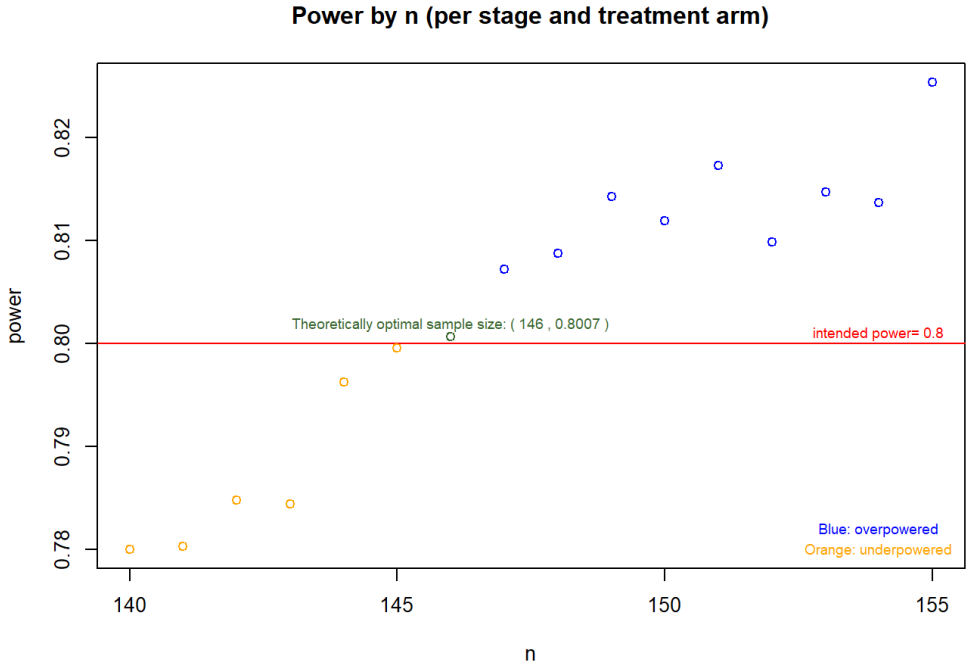
## Selected arms [3]			
## Treatment arm	1.0000	0.9557	0.6848
## Control arm	1.0000	0.9557	0.6848
## Number of active arms [1]	2.000	2.000	2.000
## Number of active arms [2]	2.000	2.000	2.000
## Number of active arms [3]	2.000	2.000	2.000
## Conditional power (achieved) [1]	0.0602	0.1379	
## Conditional power (achieved) [2]	0.4254	0.6771	
## Conditional power (achieved) [3]	0.4290	0.6911	
##			
## Legend:			
## (i): treatment arm i			
## [j]: effect matrix row j (situation to consider)			

この出力では、最初に設定した 3 つの治療効果の結果がそれぞれ[j], j = 1,2,3 で示される。

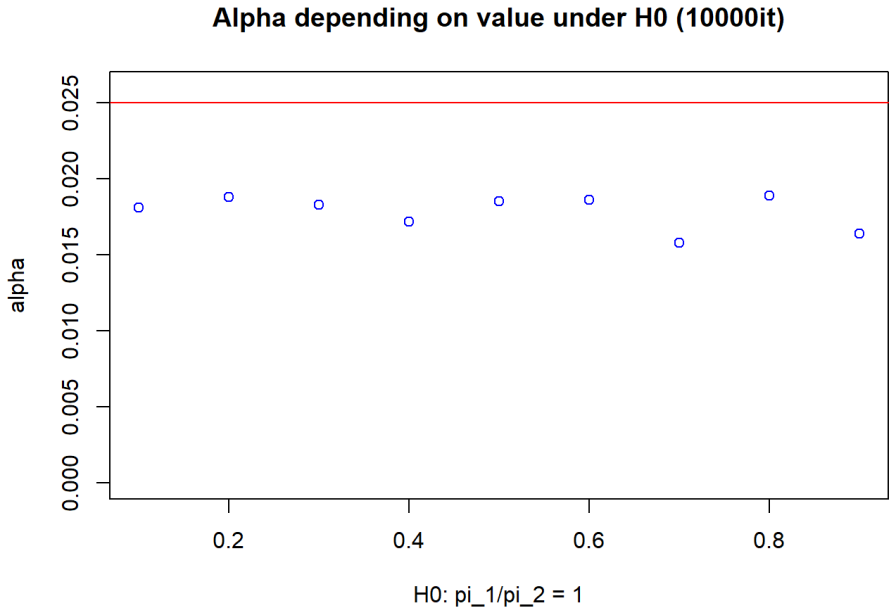
想定される治療効果が対照群（シナリオ 1）の想定される効果と等しい H_0 の下では、少なくとも 1 つを棄却する確率（Reject at least one [1]）は低くなる（つまり、第一種の過誤確率が 0.0181）。特に両方の治療群で高い治療効果を仮定すると高くなる（対立仮説の下では、シナリオ 2 で少なくとも 1 つの仮説を棄却する確率 Reject at least one [2] は 0.8550）。また、いずれの場合も、rValue = 2 として両方の投与群が選択されることは、一方または両方の投与群が該当する無益性の範囲を満たしているかどうかに関係なく、2 つの最良の投与群（この場合はすべての群）が選択されることを意味する。さらに、シミュレーション出力にはいくつかの情報が含まれている。つまり、たとえばステージごとの症例数が計算され、各ステージで選択される投与群の確率が、想定される有効割合ごとに提供される。さらに、これまでに観察されたデータから、統計的に有意な結果が得られる確率として定義される条件付き検出力（Conditional power (achieved) [1]～[3]）と、無益確率や条件付き検出力を利用して試験を停止する試験デザインの最適性基準として一般的に使用される期待症例数（Expected number of subjects [1]～[3]）などが一覧表示される。シナリオ 1（第 1 ステージ：0.5797、第 2 ステージ：0.1920）では無益停止の確率（Exit probability for futility [1]）がかなり高いため、対応する期待症例数は約 776 であり、他のシナリオの期待症例数をはるかに下回っている。

ベースとなる投与群の選択方法が特殊で、特に rpact で利用可能で事前定義されたオプション（つまり、best、rBest、epsilon、すべて）で簡単にカバーされていない場合、typeOfSelection を userDefined に設定し、selectArmsFunction-argument の入力として使用される関数を定義する必要がある。

次の図は、検出力の要件を達成するために選択できる最小症例数を示している。これは、約 146 と読み取れるが、この数値は、偏差のために最適な解からわずかにずれている可能性があることに注意が必要である。



さらに、帰無仮説の下で、第一種の過誤確率が $\pi_2 = 0.1$ を想定するだけでなく、$\pi_2 \neq 0.1$ の場合の他のシナリオも考慮して制御されているかどうかを適切に確認する必要がある。シミュレーションによって得られた次のプロットは、第一種の過誤確率がさまざまな有効割合の仮定の下で制御されることを示している。



［実データの解析］

ここでは、getAnalysisResults()コマンドを使用して、さまざまな架空の二値のエンドポイントデータセットを生成し分析する。ここでの解析は、実際の有効割合の値に基づいて実施される。なお、getClosedConditionalDunnettTestResults ()関数を用いることで Adaptive Dunnett による多重性の調整を実施することができる。

第 1 ステージの解析では、最初に、試験で観察されたデータのデータセットを手動で入力する。ここでは二値のデータを仮想的に生成して解析する。

```
genData_1 <- getDataset(
    events1      = 4,
    events2      = 8,
    events3      = 16,
    sampleSizes1 = 153,
    sampleSizes2 = 157,
    sampleSizes3 = 156)
summary(genData_1)
```

```
## Dataset of multi-arm rates
##
## The dataset contains the sample sizes and events of
## two treatment groups and one control group.
##
## Stage                1    1    1
## Group                1    2    3
## Sample size          153  157  156
## Number of events     4    8    16
```

このデータセットは、$\pi_2 = 0.1$ が与えられた場合の治療群で 50%の割合の減少があるという仮定の下で、二値のエンドポイントを持つ試験の第 1 ステージのデータを仮想的に生成する。 最も高い発現率である Index（3）は、対照群で発生したイベントとベースとなる症例数を意図している。他の Index（1, 2）は、治療群を表しており、治療群でより低い有効割合が見られるように想定されている。治療の中止またはエントリーの問題が原因で発生する可能性のある症例数のわずかな不均衡に留意すべきである。

交互作用の検定として Simes を使用した第 1 ステージの実際の分析は次のようになる。なお、

```
results_1 <- getAnalysisResults(
        design = d_IN,
        dataInput = genData_1,
        directionUpper = FALSE,
        intersectionTest="Simes")
```

```
## [PROGRESS] Stage results calculated [0.019 secs]
## [PROGRESS] Closed test calculated [0.01 secs]
## [PROGRESS] Conditional power calculated [0.019 secs]
## [PROGRESS] Conditional rejection probabilities (CRP) calculated [0.002 secs]
## [PROGRESS] Repeated confidence intervals for stage 1 calculated [0.8796 secs]
## [PROGRESS] Repeated p-values for stage 1 calculated [0.6753 secs]
```

```
summary(results_1)
```

```
## Multi-arm analysis results for a binary endpoint (2 active arms vs. control)
##
## Sequential analysis with 3 looks (inverse normal combination test design).
## The results were calculated using a multi-arm test for rates (one-sided),
## Simes intersection test, normal approximation test.
## H0: pi(i) - pi(control) = 0 against H1: pi(i) - pi(control) < 0.
##
## Stage                              1       2       3
## Fixed weight                       0.577   0.577   0.577
## Efficacy boundary (z-value scale)  3.710   2.511   1.993
## Futility boundary (z-value scale)  0.149   0.414
## Cumulative alpha spent             0.0001  0.0060  0.0250
## Stage level                        0.0001  0.0060  0.0231
## Cumulative effect size (1)         -0.076
## Cumulative effect size (2)         -0.052
## Cumulative treatment rate (1)      0.026
## Cumulative treatment rate (2)      0.051
## Cumulative control rate            0.103
## Stage-wise test statistic (1)      -2.730
## Stage-wise test statistic (2)      -1.716
## Stage-wise p-value (1)             0.0032
```

```
## Stage-wise p-value (2)                    0.0431
## Adjusted stage-wise p-value (1, 2)        0.0063
## Adjusted stage-wise p-value (1)           0.0032
## Adjusted stage-wise p-value (2)           0.0431
## Overall adjusted test statistic (1, 2)    2.493
## Overall adjusted test statistic (1)       2.730
## Overall adjusted test statistic (2)       1.716
## Test action: reject (1)                   FALSE
## Test action: reject (2)                   FALSE
## Conditional rejection probability (1)     0.2907
## Conditional rejection probability (2)     0.1204
## 95% repeated confidence interval (1)      [-0.212; 0.043]
## 95% repeated confidence interval (2)      [-0.191; 0.079]
## Repeated p-value (1)                      0.1150
## Repeated p-value (2)                      0.2429
##
## Legend:
##     (i): results of treatment arm i vs. control arm
##     (i, j, ...): comparison of treatment arms 'i, j, ...' vs. control arm
```

すでに第 1 ステージで明らかに低い有効割合であるが、2 つの仮説のいずれも棄却することはできない。これは、たとえば、有効性の限界まで治療群のいずれにも効果がないという全体の積仮説の全体的な調整済み検定統計量（Overall adjusted test statistic (1, 2)）が 2.493 < 3.710、つまり全体の積仮説を棄却できないことを意味する。第 1 ステージでは無益性の境界を越えていないため、無益による停止はされない。これは、両方の治療群が第 2 ステージの解析に繰り越されることを意味する。

第 2 ステージに進むと、データセットを次のように生成し、第 2 ベクトルの入力がその第 2 ステージのデータを表す。

```
#assuming there was no futility or efficacy stop study proceeds to randomize subjects
genData_2 <- getDataset(
      events1       = c(4,7),
      events2       = c(8,7),
      events3       = c(16,15),
      sampleSizes1 = c(153,155),
      sampleSizes2 = c(157,155),
```

```
    sampleSizes3 = c(156,155))
summary(genData_2)
```

```
## Dataset of multi-arm rates
##
## The dataset contains the sample sizes and events of
## two treatment groups and one control group.
## The total number of looks is two; stage-wise and cumulative data are included.
##
## Stage                          1   1   1   2   2   2
## Group                          1   2   3   1   2   3
## Stage-wise sample size         153 157 156 155 155 155
## Cumulative sample size         153 157 156 308 312 311
## Stage-wise number of events    4   8   16  7   7   15
## Cumulative number of events    4   8   16  11  15  31
```

ここでも、有効割合は治療群で低くなると想定している。

```
results _2 <- getAnalysisResults(
    design=d_IN,
    dataInput = genData_2,
    directionUpper = FALSE,
    intersectionTest="Simes")
```

```
## [PROGRESS] Stage results calculated [0.016 secs]
## [PROGRESS] Closed test calculated [0.008 secs]
## [PROGRESS] Conditional power calculated [0.016 secs]
## [PROGRESS] Conditional rejection probabilities (CRP) calculated [0.002 secs]
## [PROGRESS] Repeated confidence intervals for stage 1 calculated [0.9176 secs]
## [PROGRESS] Repeated confidence intervals for stage 2 calculated [0.8726 secs]
## [PROGRESS] Repeated p-values for stage 1 calculated [0.7006 secs]
## [PROGRESS] Repeated p-values for stage 2 calculated [0.7001 secs]
```

```
summary (results_2)
```

```
## Multi-arm analysis results for a binary endpoint (2 active arms vs. control)
##
```

Sequential analysis with 3 looks (inverse normal combination test design).
The results were calculated using a multi-arm test for rates (one-sided),
Simes intersection test, normal approximation test.
H0: pi(i) - pi(control) = 0 against H1: pi(i) - pi(control) < 0.
##

## Stage	1	2	3
## Fixed weight	0.577	0.577	0.577
## Efficacy boundary (z-value scale)	3.710	2.511	1.993
## Futility boundary (z-value scale)	0.149	0.414	
## Cumulative alpha spent	0.0001	0.0060	0.0250
## Stage level	0.0001	0.0060	0.0231
## Cumulative effect size (1)	-0.076	-0.064	
## Cumulative effect size (2)	-0.052	-0.052	
## Cumulative treatment rate (1)	0.026	0.036	
## Cumulative treatment rate (2)	0.051	0.048	
## Cumulative control rate	0.103	0.100	
## Stage-wise test statistic (1)	-2.730	-1.770	
## Stage-wise test statistic (2)	-1.716	-1.770	
## Stage-wise p-value (1)	0.0032	0.0384	
## Stage-wise p-value (2)	0.0431	0.0384	
## Adjusted stage-wise p-value (1, 2)	0.0063	0.0384	
## Adjusted stage-wise p-value (1)	0.0032	0.0384	
## Adjusted stage-wise p-value (2)	0.0431	0.0384	
## Overall adjusted test statistic (1, 2)	2.493	3.014	
## Overall adjusted test statistic (1)	2.730	3.182	
## Overall adjusted test statistic (2)	1.716	2.464	
## Test action: reject (1)	FALSE	TRUE	
## Test action: reject (2)	FALSE	FALSE	
## Conditional rejection probability (1)	0.2907	0.7911	
## Conditional rejection probability (2)	0.1204	0.5133	
## 95% repeated confidence interval (1)	[-0.212; 0.043]	[-0.130; -0.005]	
## 95% repeated confidence interval (2)	[-0.191; 0.079]	[-0.119; 0.011]	
## Repeated p-value (1)	0.1150	0.0086	
## Repeated p-value (2)	0.2429	0.0274	

##
Legend:

```
##     (i): results of treatment arm i vs. control arm
##     (i, j, ...): comparison of treatment arms 'i, j, ...' vs. control arm
```

第 1 ステージと同じ比較を実行すると、次のことが分かる。全体の積仮説（Overall adjusted test statistic (1, 2)）は棄却され（3.014> 2.511）、その後（Overall adjusted test statistic (1)）、3.182> 2.511 であるため、最初の治療群の仮説は棄却される。0.0086 の繰り返し p 値（Repeated p-value (1)）が 0.025 を下回ることによっても、同じ結果を得ることができる。その結果、この治療群は有効性のために早期に中止することができる。ただし、2 番目の治療群の検定（Overall adjusted test statistic (2)）は有意にならず（2.464 < 2.511）、試験は 2 番目の治療群のみで最終ステージまで継続される。ここで、調整されたステージごとの p 値（Adjusted stage-wise p-value）を直接検定に使用することはできず、その値は第 2 ステージのデータにのみ基づいていることに留意していただきたい。ここでは 13 行目で与えられる重みを使用した逆正規の統合検定が実行されるため、第 1 ステージと第 2 ステージの p 値を使用して、全体的な調整済み検定統計量（Overall adjusted test statistic (1, 2)）が計算され、その値は 3.014 となる。

さらに、第 3 ステージのデータを入力する。

```
genData_3 <- getDataset(
      events1       = c(4,7,NA),
      events2       = c(8,7,6),
      events3       = c(16,15,16),
      sampleSizes1 = c(153,155,NA),
      sampleSizes2 = c(157,155,156),
      sampleSizes3 = c(156,155,160))
summary(genData_3)
```

```
## Dataset of multi-arm rates
##
## The dataset contains the sample sizes and events of
## two treatment groups and one control group.
## The total number of looks is three; stage-wise and cumulative data are included.
##
## Stage                    1   1   1   2   2   2   3   3   3
## Group                    1   2   3   1   2   3   1   2   3
## Stage-wise sample size   153 157 156 155 155 155 156 160
## Cumulative sample size   153 157 156 308 312 311 468 471
```

```
## Stage-wise number of events      4   8  16   7   7  15   6  16
## Cumulative number of events      4   8  16  11  15  31  21  47
```

最終解析は次の通りとなる。

```
results_3 <- getAnalysisResults(
    design=d_IN,
    dataInput = genData_3,
    directionUpper = FALSE,
    intersectionTest="Simes")
```

```
## [PROGRESS] Stage results calculated [0.018 secs]
## [PROGRESS] Closed test calculated [0.008 secs]
## [PROGRESS] Conditional power calculated [0.016 secs]
## [PROGRESS] Conditional rejection probabilities (CRP) calculated [0.002 secs]
## [PROGRESS] Repeated confidence intervals for stage 1 calculated [0.8906 secs]
## [PROGRESS] Repeated confidence intervals for stage 2 calculated [0.8746 secs]
## [PROGRESS] Repeated confidence intervals for stage 3 calculated [0.467 secs]
## [PROGRESS] Repeated p-values for stage 1 calculated [0.7046 secs]
## [PROGRESS] Repeated p-values for stage 2 calculated [0.7456 secs]
## [PROGRESS] Repeated p-values for stage 3 calculated [0.352 secs]
```

```
summary (results_3)
```

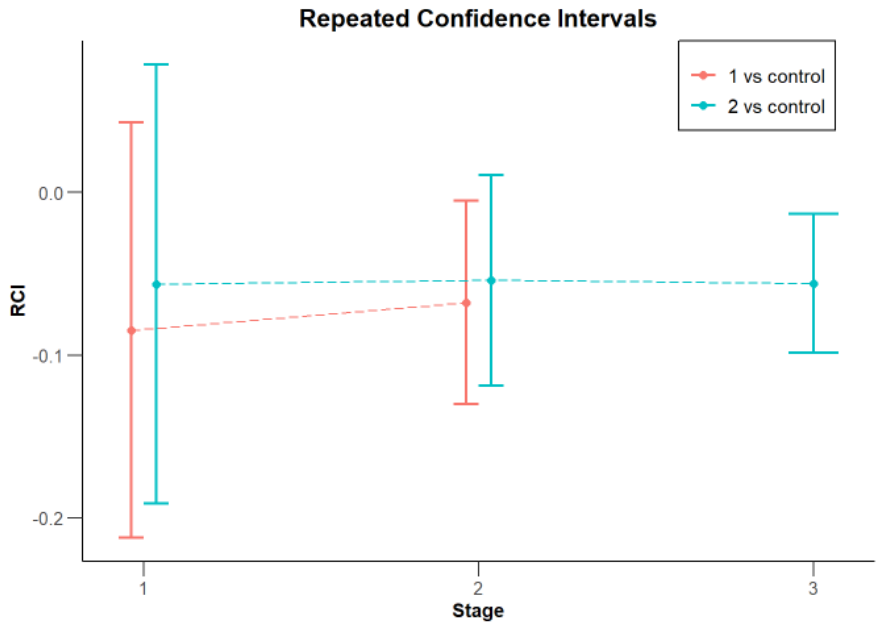
```
## Multi-arm analysis results for a binary endpoint (2 active arms vs. control)
##
## Sequential analysis with 3 looks (inverse normal combination test design).
## The results were calculated using a multi-arm test for rates (one-sided),
## Simes intersection test, normal approximation test.
## H0: pi(i) - pi(control) = 0 against H1: pi(i) - pi(control) < 0.
##
## Stage                            1       2       3
## Fixed weight                     0.577   0.577   0.577
## Efficacy boundary (z-value scale) 3.710  2.511   1.993
## Futility boundary (z-value scale) 0.149  0.414
## Cumulative alpha spent           0.0001  0.0060  0.0250
## Stage level                      0.0001  0.0060  0.0231
```

## Cumulative effect size (1)	-0.076	-0.064	
## Cumulative effect size (2)	-0.052	-0.052	-0.055
## Cumulative treatment rate (1)	0.026	0.036	
## Cumulative treatment rate (2)	0.051	0.048	0.045
## Cumulative control rate	0.103	0.100	0.100
## Stage-wise test statistic (1)	-2.730	-1.770	
## Stage-wise test statistic (2)	-1.716	-1.770	-2.149
## Stage-wise p-value (1)	0.0032	0.0384	
## Stage-wise p-value (2)	0.0431	0.0384	0.0158
## Adjusted stage-wise p-value (1, 2)	0.0063	0.0384	0.0158
## Adjusted stage-wise p-value (1)	0.0032	0.0384	
## Adjusted stage-wise p-value (2)	0.0431	0.0384	0.0158
## Overall adjusted test statistic (1, 2)	2.493	3.014	3.702
## Overall adjusted test statistic (1)	2.730	3.182	
## Overall adjusted test statistic (2)	1.716	2.464	3.253
## Test action: reject (1)	FALSE	TRUE	TRUE
## Test action: reject (2)	FALSE	FALSE	TRUE
## Conditional rejection probability (1)	0.2907	0.7911	
## Conditional rejection probability (2)	0.1204	0.5133	
## 95% repeated confidence interval (1)	[-0.212; 0.043]	[-0.130; -0.005]	
## 95% repeated confidence interval (2)	[-0.191; 0.079]	[-0.119; 0.011]	[-0.099; -0.013]
## Repeated p-value (1)	0.1150	0.0086	
## Repeated p-value (2)	0.2429	0.0274	0.0006
##			
## Legend:			
## (i): results of treatment arm i vs. control arm			
## (i, j, ...): comparison of treatment arms 'i, j, ...' vs. control arm			

ここでは、最初の治療群がすでに第 2 ステージで有意となっているため、2 番目の治療群の交差および単一の仮説または繰り返し p 値の全体的な調整済み検定統計量（Overall adjusted test statistic (1, 2)）、または Test action: reject (2)を直接参照することから、2 番目の治療群も最終ステージで対照群よりも優れていると考えられ、試験は成功したと結論付けることができる。

繰り返し信頼区間（repeated confidence interval）を示すことによって解析の視覚化することができる。累積症例数が増加し、ステージ全体で一貫した傾向、またステージに沿って間隔がどのように狭くなるかを確認できる。

```
plot(results_3, type=2)
```



3.6.2 Bowden and Glimm (2008)による UMVCUE と信頼区間

Bowden and Glimm (2008)[21]が提案した 2 ステージデザインにおける UMVCUE に関して、推定値のバイアス、MSE 及び UMVCUE の信頼区間の被覆確率を計算する JPMA タスクフォースのメンバーが作成した SAS マクロを紹介する。このマクロは SAS 9.4, SAS/STAT 14.3 の下で作成された。表 17 に SAS マクロの入力引数を示す。このマクロは、補遺 C に示している。

表 17　UMVCUE の性能を評価する SAS マクロの入力引数

引数	説明
k	治療群の数。この引数に入力する数は、mu1, m2, …で指定する真値の数の最大値と一致させる必要がある（例：5 群のシミュレーションを行う場合、k=5 とし、mu1, …, mu5 を定める）。
mu1, m2, …	各治療群における治療効果の真値。10 群まで設定可能。

引数	説明
comsig	シミュレーションデータの標準偏差。ステージ 1 及びステージ 2 を通して全治療群で共通とする。
nsub1	ステージ 1 の各治療群のサンプルサイズ。各治療群のサンプルサイズは共通とする。
nsub2	ステージ 2 の治療群のサンプルサイズ。
scenario	各シミュレーションで用いるシナリオに対する特定の番号。
sim	シミュレーションでの試験数。デフォルトは 100000。
sim2	全ての mu を 0 とした場合のシミュレーションでの試験数。UMVCUE の信頼区間を推定するために利用する、UMVCUE の分散の最大値を求めるために設定する。デフォルトは 50000。

想定するエンドポイントは正規分布に従うことを仮定した。各治療群の標準偏差は comsig で設定し、ステージ 1 及びステージ 2 を通して全治療群で共通とした。また、第 1 ステージの各治療群の症例数は全ての群で共通とする。二段階デザインは、ステージ 1 における k 個の治療群の内、l 番目に大きい平均値を持つ治療群を第 2 ステージに進めることを想定する。

本マクロでは、ステージ 1 において l 番目に大きい平均値を持つ治療群の UMVCUE 及び MLE の bias（$b(\mu_{(l)})$）及び MSE（$MSE(\mu_{(l)})$）を以下の式に基づき算出する。

$$b(\mu_{(l)}) = \sum_{i=1}^{k} E\left[\mu_{(l)} - \mu_i \mid \bar{X}_{(i)} = \bar{X}_i\right] P(\bar{X}_{(l)} = \bar{X}_i)$$

$$MSE(\mu_{(l)}) = \sum_{i=1}^{k} E\left[(\mu_{(l)} - \mu_i)^2 \mid \bar{X}_{(i)} = \bar{X}_i\right] P(\bar{X}_{(l)} = \bar{X}_i)$$

シミュレーション回数を除く入力引数に加え、表 18 の結果をデータセット Bias_mse_sX 及び Coverage_sX（X は引数で指定する scenario の番号）に出力する。バイアス、MLE 及び信頼区間の被覆確率は、l 番目に大きい治療群に対して、1, 2, …, k 番目の順に出力される。

表 18　UMVCUE の性能を評価する SAS マクロの出力

データセット	変数名	説明
共通	l	k 個の治療群の内 l 番目の UMVCUE。合計 k 行のオブザベーションが発生する。
Bias_mse_sX	b_umvcue	UMVCUE のバイアス
Bias_mse_sX	b_mle	MLE のバイアス
Bias_mse_sX	mse_umvcue	UMVCUE の MLE
Bias_mse_sX	mse_mle	MLE の MLE
Coverage_sX	PERCENT	UMVCUE の信頼区間の被覆確率

図 11　Bowden and Glimm (2008)の図 3（右）を複製した図（実線は UMVCUE、点線は MLE）と図 12 は、Bowden and Glimm (2008)[21]の図 3（右）と図 5 を複製した図である。図 11 は、5 つの治療群があり、3 種類の平均の仮定に基づく$\mu_{(i)}$の平均二乗誤差を、図 12 は信頼区間に被覆確率をプロットしている。

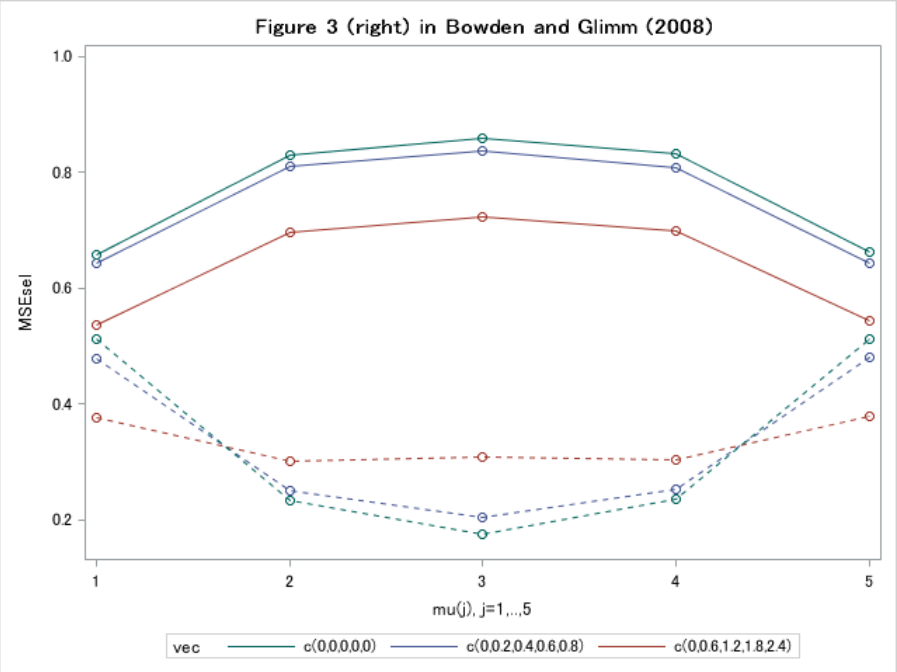


図 11　Bowden and Glimm (2008)の図 3（右）を複製した図（実線は UMVCUE、点線は MLE）

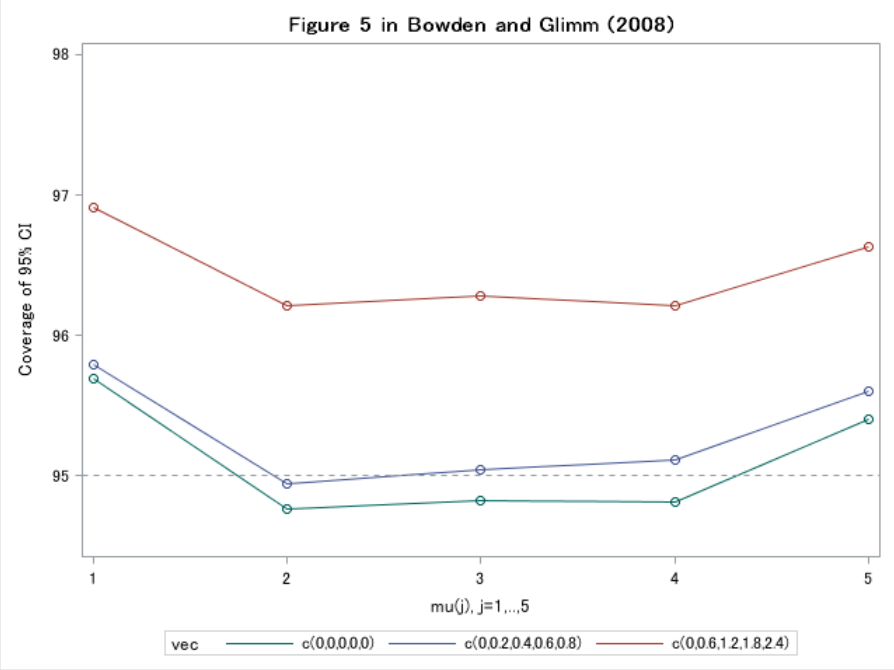


図 12　Bowden and Glimm (2008)の図 5 を複製した図

参考文献

[1] Bauer P, Koenig F, Brannath W, Posch M. Selection and bias—two hostile brothers. Statistics in Medicine 2010;29(1):1-13.

[2] Wang SJ, James Hung HM, O'Neill RT. (2010) Impacts on type I error rate with inappropriate use of learn and confirm in confirmatory adaptive design trials. Biometrical Journal 52(6):798-810. doi: 10.1002/bimj.200900207.

[3] Zeymer U, Suryapranata H, Monassier JP, Opolski G, Davies J, Rasmanis Get al. Evaluation of the safety and cardioprotective effects of eniporide, a specific Sodium/Hydrogen exchange inhibitor, given as adjunctive therapy to reperfusion in patients with acute myocardial infarction. Heart Drug 2001;1(2):71-76.

[4] Zeymer U, Suryapranata H, Monassier JP, Opolski G, Davies J, Rasmanis G et al. The Na+/H+ exchange inhibitor eniporide as an adjunct to early reperfusion therapy for acute myocardial infarction: Results of the evaluation of the safety and cardioprotective effects of eniporide in acute myocardial infarction (ESCAMI) trial. Journal of the American College of Cardiology 2001;38(6):E1644-E1650.

[5] Bauer P, Bretz F, Dragalin V, König F, Wassmer G. Twenty‐five years of confirmatory adaptive designs: opportunities and pitfalls. Statistics in Medicine 2016;35(3):325-347.

[6] Bauer P and Kohne K. Evaluation of experiments with adaptive interim analyses. Biometrics 1994;50:1029-1041.

[7] Lawrence D, Bretz F, Pocock S. INHANCE: An adaptive confirmatory study with dose selection at interim. In Indacaterol - The First Once-Daily Long-Acting Beta2 Agonist for COPD, Trifilieff A (ed.) Springer: Basel, 2014; 77–93.

[8] Bhatt DL, Mehta C. Adaptive designs for clinical trials. New England Journal of Medicine 2016; 375(1):65-74.

[9] Lawrence D, Bretz F. Approaches for optimal dose selection for adaptive design trials. In Practical Considerations for Adaptive Trial Design and Implementation, He W, Pinheiro J, Kuznetsova OM (eds). Springer: New York, Heidelberg, Dordrecht, London, 2014; 77–93.

[10] Robertson DS and Glimm E. Conditionally unbiased estimation in the normal setting with unknown variances. Communications in Statistics: Theory and Methods 2019;48:616-627

[11] Donohue JF, Fogarty C, Lötvall J, Mahler DA, Worth H, Yorgancioglu A et al. Once-daily bronchodilators for chronic obstructive pulmonary disease: indacaterol versus tiotropium. American journal of respiratory and critical care medicine 2010;182(2):155-162.

[12] Food and Drug Administration. Medical Review, APPLICATION NUMBER:022383Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2011/022383Orig1s000MedR.pdf [accessed 14 September 2022]

[13] Food and Drug Administration. Statistical Review, APPLICATION

NUMBER:022383Orig1s000.
https://www.accessdata.fda.gov/drugsatfda_docs/nda/2011/022383Orig1s000StatR.pdf [accessed 14 September 2022]

[14] Koenig F, Brannath W, Bretz F, Posch M. Adaptive Dunnett tests for treatment selection. Statistics in Medicine 2008;10;27(10):1612-25.

[15] Magirr D, Jaki T, Whitehead J. A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection. Biometrika 2012;99(2):494-501.

[16] Friede T, Stallard N. A comparison of methods for adaptive treatment selection. Biometrical Journal 2008;50(5):767-81.

[17] Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association 1955;50:1096–1121.

[18] Marcus R, Eric P, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. Biometrika 1976;63(3):655-660.

[19] Brannath W, Posch M, Bauer P. Recursive combination tests. Journal of American Statistical Association 2002;97(457):236-244

[20] Cohen A, Sackrowitz HB. Two stage conditionally unbiased estimators of the selected mean. Statistics & Probability Letters 1989;8(3):273-278.

[21] Bowden J, Glimm E. Unbiased estimation of selected treatment means in two-stage trials. Biometrical Journal 2008;50(4):515-527.

[22] Bowden J, Glimm E. Conditionally unbiased and near unbiased estimation of the selected treatment mean for multistage drop-the-losers trials. Biometrical Journal 2014;56(2):332-349.

[23] Kimani PK, Todd S, and Stallard N. Conditionally unbiased estimation in phase II/III clinical trials with early stopping for futility. Statistics in Medicine 2013;32(17):2893-910.

[24] Robertson DS, Prevost AT, Bowden J. Unbiased estimation in seamless phase II/III trials with unequal treatment effect variances and hypothesis-driven selection rules. Statistical in Medicine 2016;35:3907-22.

[25] Stallard N, Kimani PK. Uniformly minimum variance conditionally unbiased estimation in multi-arm multi-stage clinical trials. Biometrika 2018;105(2):495-501.

[26] Sill M, Sampson A. Extension of a two-stage conditionally unbiased estimator of the selected population to the bivariate normal case. Communications in Statistics-Theory and Methods 2007;36:801–813.

[27] Carreras M, Brannath W. Shrinkage estimation in two‐stage adaptive designs with midtrial treatment selection. Statistics in Medicine 2013;32(10):1677-1690.

[28] Stallard N, Todd S. Point estimates and confidence regions for sequential trials involving selection. Journal of Statistical Planning and Inference 2005;135(2):402-419.

[29] Stallard N, Todd S. Sequential designs for phase III clinical trials incorporating treatment

selection. Statistics in Medicine 2003;22(5):689-703.

[30] Pickard MD, Chang M. A flexible method using a parametric bootstrap for reducing bias in adaptive designs with treatment selection. Statistics in Biopharmaceutical Research 2014;6(2):163-174.

[31] Tappin L. Unbiased Estimation of the Parameter of a Selected Binomial Population. Communications in Statistics: Theory and Methods 1992;21:1067–1083

[32] Sampson AR and Sill MW. Drop-the-losers design: Normal case. Biometrical Journal 2005;47(3):257-68.

[33] Posch M, Koenig F, Branson M, Brannath W, Dunger-Baldauf C, Bauer P. Testing and estimation in flexible group sequential designs with adaptive treatment selection. Statistics in Medicine 2005;24:3697-3714.

[34] Wu SS, Wang W, Yang MCK. Interval estimation for drop-the-losers design. Biometrika 2010;97:405-418.

[35] Neal D, Casella G, Yang, MCK, Wu SS. Interval estimation in two-stage drop-the-losers clinical trials with flexible treatment selection. Statistics in Medicine 2011;30:2804-2814.

[36] Magirr D, Jaki T, Posch M, Klinglmueller F. Simultaneous confidence intervals that are compatible with closed testing in adaptive designs. Biometrika 2013;100(4):985-996.

[37] Lu X, He Y, Wu SS. Interval estimation in multi-stage drop-the-losers designs. Statistical Methods in Medical Research 2018;27(1):221-233.

[38] Kimani PK, Todd S, Stallard N. A comparison of methods for constructing confidence intervals after phase II/III clinical trials. Biometrical Journal 2014;56(1):107-128.

[39] Rpact package. https://www.rpact.com/vignettes/rpact_mams_design_and_analysis [access 2022/09/09]

[40] Wassmer G, Brannath W. Group Sequential and Confirmatory Adaptive Designs in Clinical Trials. Springer, 2016.

4　おわりに

単純な中間解析の場合にも推定量が影響を受けることは、日本の通知でも触れられている[1]。本報告書では、症例数再推定と治療群の選択のそれぞれのアダプティブデザインでの推定量が受ける影響のみならず、適用できる統計的推測法を解説した。第一種の過誤確率を制御する検定方法、バイアスを低減させる点推定の方法、被覆確率を名目水準に維持する区間推定の方法を紹介し、一部ではいくつかの方法を比較した内容も述べた。

アダプティブデザインの基本的な統計的推測法を理解するために、いくつかの成書を参考にすることができる。Chang (2014)[1]は、症例数再推定や治療群の選択などのアダプテーションルール別に適用できる方法を検定中心に述べており、SAS と R によるソースコードを豊富に載せている。Wassmer and Brannath (2016)[3]は、アダプティブデザインの理解の基本である群逐次デザインと、アダプティブデザインを適用する検証的試験のデザインをまとめており、特に 2 ステージのアダプティブデザインについて分かりやすく記載している良書である。また、和書としては、Chow and Chang (2006)[4]の邦訳として平川・五所（2018）がある。総説論文もいくつか公表されている。検証的試験におけるアダプティブデザインに関する、これまでの発展や展望等については Bauer et al. (2016)[5]を参照されたい。治療群の選択のアダプティブデザインとしては、Statistics in Medicine 誌の Tutorial in Biostatitics のシリーズにある Bretz et al. (2009)[6]がまとまっている。

試験に適用するアダプテーションルール（例えば、中間解析の回数や時期、治療群の選択ルールなど）によって、統計的推測法の性質が変化することが考えられる。シミュレーションなどを用いて、アダプテーションルールに対する統計的推測法の性質を予め調べておくことが重要であると考えられる。また、本文書で紹介した統計的推測法によっては想定しているアダプテーションルールが異なる場合があるため、試験に適用する統計的推測法と試験デザインのアダプテーションルールが整合しているかを確認しておくことも重要である。本文書の冒頭で述べたように、アダプティブデザインの統計的推測法が今もなお模索されている状況であり、最新の情報を参照しつつ適切な方法を調べることが望ましい。

参考文献

[1] 厚生労働省，薬食審査発 0404 第 1 号　平成 25 年 4 月 4 日 データモニタリング委員会のガイドラインについて

[2] Chang M. Adaptive design theory and implementation using SAS and R. 2nd ed, Boca Raton., FL: CRC Press, 2014.

[3] Wassmer G, Brannath W. Group Sequential and Confirmatory Adaptive Designs in Clinical Trials. Springer, 2016.

[4] Chow S-C and Chang M. Adaptive design methods in clinical trials. Chapman and Hall/CRC,

2006. 平川 晃弘，五所 正彦訳．臨床試験のためのアダプティブデザイン．朝倉書店．東京．2018.

[5] Bauer P, Koenig F, Brannath W, Posch M. Selection and bias—two hostile brothers. Statistics in Medicine 2010;29(1):1-13.

[6] Bretz F, Koenig F, Brannath W, Glimm E, Posch M. Adaptive designs for confirmatory clinical trials. Stat Med 2009; 28(8):1181-217.

補遺 A　2.2.1 項の図表を作成する R コード

```
rm(list = ls())
options(warn=-1)
windowsFonts(Meiryo = windowsFont("Meiryo"))
library(tidyverse)

# Define G functions ----------------------------------------
G_plus <- function(z, Cb, za) {
   f_z_plus <- (z * abs(z) - Cb * za) / sqrt (Cb **2 - z **2)
   if (abs(z) < Cb) {
      H_plus <- pnorm(f_z_plus)
   } else if (z >= Cb) {
      H_plus <- 1
   } else {
      H_plus <- 0
   }
   H_plus * dnorm(z)
}

G0_plus <- function(z, t, za) {
   if (t != 1) {
      f0_plus <- sqrt(t / (1-t)) * z - sqrt(1 / (1-t)) * za
      pnorm(f0_plus) * dnorm(z)
   } else {
      dnorm(z)
   }
}

# Table 2. overall inflation (power, Cb, alpha*, Inflation)
table2 <- function() {
   alpha <- 0.05
   za <- qnorm(1 - alpha/2)
   for (beta in seq(0.2,0.05,-0.05)){
      power <- (1 - beta) * 100
      zb <- qnorm(1-beta)
```

```
    Cb <- za + zb
    alpha_star <- 2 * pnorm(-Cb) + 2 * integrate(G_plus,-Cb,Cb,Cb,za)$value
    inflation <- round( (alpha_star - alpha)/ alpha * 100 )
    cat(power, Cb, alpha_star, inflation, "¥n")
  }
}
table2()

# Figure 1 ------------------------------------------------
figure1 <- function() {
  alpha <- 0.05
  beta <- 0.2
  za <- qnorm(1 - alpha/2)
  zb <- qnorm(1-beta)
  Cb <- za + zb
  i <- 1
  z_range <- seq(-2,4,0.001)
  df_fig1 <- data.frame(matrix(nrow = length(range), ncol = 5))
  colnames(df_fig1) <- c("z", "G_inf","G_0:t=1/3", "G_0:t=1/2", "G_0:t=2/3")
  for (z in z_range){
    G_inf <- G_plus(z, Cb, za)
    G_0_13 <- G0_plus(z,1/3, za)
    G_0_12 <- G0_plus(z,1/2, za)
    G_0_23 <- G0_plus(z,2/3, za)
    df_fig1[i, 1] <- z
    df_fig1[i, 2] <- G_inf
    df_fig1[i, 3] <- G_0_13
    df_fig1[i, 4] <- G_0_12
    df_fig1[i, 5] <- G_0_23
    i <- i + 1
  }
  # Convert to long-type data
  df_fig1_L <- df_fig1 %>%
    pivot_longer(c("G_inf","G_0:t=1/3", "G_0:t=1/2", "G_0:t=2/3"),
                 names_to  = "G_function",
                 values_to = "G_plus")
```

```
  ggplot(df_fig1_L) +
    geom_line(aes(x = z, y = G_plus, linetype = G_function)) +
    scale_linetype_manual(values = c("dashed", "longdash", "dotted", "solid")) +
    theme_bw() +
    theme(legend.position = c(0.15, 0.8)) +
    theme(text=element_text(family = "Meiryo", size=12)) +
    theme(title=element_text(family = "Meiryo", size=12))
}
figure1()

# 2.2.1.3 ----------------------------------------------------------
alpha_diff_31 <- function() {
  alpha <- 0.05
  beta <- 0.15
  za <- qnorm(1 - alpha/2)
  zb <- qnorm(1-beta)
  Cb <- qnorm(1 - alpha/2) + qnorm(1 - beta)
  i <- 1
  result <- data.frame(matrix(nrow = 15, ncol = 3))
  colnames(result) <- c("t", "r","alpha_diff")
  for (t in c(1/3, 1/2, 2/3)){
    for (r in c(1.5, 2.0, 3.0, 4.0, 5.0)){
      t_L <- t
      t_U <- t/r;
      z_L <- Cb * sqrt(t_L)
      z_U <- Cb * sqrt(t_U);

      alpha_star_31 <- 2 * (integrate(G_plus, -z_L, -z_U, Cb, za)$value +
          integrate(G_plus, z_U, z_L, Cb, za)$value + integrate(G0_plus, -z_U, z_U, t/r, za)$value +
          integrate(G0_plus, -Inf, -z_L, t, za)$value + integrate(G0_plus, z_L, Inf, t, za)$value )
      alpha0_31 <- 2 * integrate(G0_plus, -Inf, Inf, t, za)$value
      alpha_diff_31 <- alpha_star_31 - alpha0_31
      result[i, 1] <- round(t, digits = 3)
      result[i, 2] <- r
      result[i, 3] <- round(alpha_diff_31, digits = 5)
      i <- i + 1
```

```
    }
  }
  result
}
alpha_diff_31()

# 2.2.1.3.1 ----------------------------------------------------
alpha_diff_32 <- function(t,x) {
  alpha <- 0.05
  beta <- 0.15
  za <- qnorm(1 - alpha/2)
  zb <- qnorm(1-beta)
  Cb <- za + zb
  i <- 1
  result <- data.frame(matrix(nrow = length(x), ncol = 3))
  colnames(result) <- c("t", "r","alpha_diff")
  for (r in as.list(x)){
    #t_L <- n1 / n_L; t_U <- n1 / n_U;
    t_L <- t; t_U <- t/r;
    z_L <- Cb * sqrt(t_L); z_U <- Cb * sqrt(t_U);

    alpha_diff <- 2 * ( integrate(G_plus, z_U, z_L, Cb, za)$value - integrate(G0_plus, z_U, z_L, t, za)$value +
                        integrate(G_plus, -z_L, -z_U, Cb, za)$value - integrate(G0_plus, -z_L, -z_U, t, za)$value)
    result[i, 1] <- round(t, digits = 3)
    result[i, 2] <- r
    result[i, 3] <- round(alpha_diff, digits = 5)
    i <- i + 1
  }
  result
}
alpha_diff_32(1/3, c(1.5, 2.0, 3.0, 4.0, 5.0))
alpha_diff_32(1/2, c(1.5, 2.0, 3.0, 4.0, 8.0, 9.0))
alpha_diff_32(2/3, c(1.5, 2.0, 3.0, 4.0, 5.0, 11, 12))

# 2.2.1.3.2 ----------------------------------------------------
alpha_diff_33 <- function(tseq) {
```

```
  alpha <- 0.05
  beta <- 0.15
  za <- qnorm(1 - alpha/2)
  zb <- qnorm(1-beta)
  Cb <- za + zb
  i <- 1
  result <- data.frame(matrix(nrow = 5, ncol = 2))
  colnames(result) <- c("t", "alpha_diff")
  for (t in as.list(tseq)){
    t_L <- 1
    t_U <- t
    z_L <- Cb * sqrt(t_L)
    z_U <- Cb * sqrt(t_U)
    alpha_star <- 2 * (integrate(G_plus, -z_L, -z_U, Cb, za)$value + integrate(G0_plus, -z_U, z_U, t, za)$value +
                         integrate(G_plus, z_U, z_L, Cb, za)$value + integrate(dnorm, z_L, Inf)$value)
    alpha0 <- 2 * integrate(G0_plus, -Inf, Inf, t, za)$value
    alpha_diff <- alpha_star - alpha0
    result[i, 1] <- round(t, digits = 3)
    result[i, 2] <- round(alpha_diff, digits = 5)
    i <- i + 1
  }
  result
}
alpha_diff_33(c(1/4, 1/3, 1/2, 2/3, 3/4))

# 2.2.1.3.3 ----------------------------------------------------
alpha_diff_34 <- function(tseq) {
  alpha <- 0.05
  beta <- 0.15
  za <- qnorm(1 - alpha/2)
  zb <- qnorm(1-beta)
  Cb <- za + zb
  i <- 1
  result <- data.frame(matrix(nrow = 5, ncol = 2))
  colnames(result) <- c("t", "alpha_diff")
  for (t in as.list(tseq)){
```

```
    t_U <- t
    z_L <- Cb
    z_U <- Cb * sqrt(t_U)
    if (za < z_U) {
      alpha_star <- 2 * (integrate(G_plus, -z_L, -z_U, Cb, za)$value + integrate(dnorm, za, z_U)$value +
                         integrate(G_plus, z_U, z_L, Cb, za)$value + integrate(dnorm, z_L, Inf)$value)
    }
    else {
      alpha_star <- 2 * (integrate(G_plus, -z_L, -z_U, Cb, za)$value +
                         integrate(G_plus, z_U, z_L, Cb, za)$value + integrate(dnorm, z_L, Inf)$value)
    }
    alpha0 <- 2 * integrate(G0_plus, -Inf, Inf, t, za)$value
    alpha_diff <- alpha_star - alpha0
    result[i, 1] <- round(t, digits = 3)
    result[i, 2] <- round(alpha_diff, digits = 5)
    i <- i + 1
  }
  result
}
alpha_diff_34(c(1/4, 1/3, 1/2, 2/3, 3/4))

# 2.2.1.3.4 ----------------------------------------------------
alpha_diff_36 <- function(beta_min,t,x) {
  alpha <- 0.05
  beta <- 0.15
  za <- qnorm(1 - alpha/2)
  zb <- qnorm(1-beta)
  Cb <- za + zb
  i <- 1
  result <- data.frame(matrix(nrow = length(x), ncol = 3))
  colnames(result) <- c("t", "r",paste0("1-Bmin=",(1-beta_min)*100,"%"))
  for (r in as.list(x)){
    zb_min <- qnorm(1-beta_min)
    Cb_min <- za + zb_min
    z_L <- Cb * sqrt(t)
    z_U <- Cb * sqrt(t/r)
```

```
      z_min <- Cb_min * sqrt(t/r)
      alpha_star <- 2 * (integrate(G_plus,-z_L,-z_U, Cb, za)$value + integrate(G_plus, z_U, z_L, Cb, za)$value +
         integrate(G0_plus, -z_U, -z_min, t/r, za)$value + integrate(G0_plus, z_min, z_U, t/r, za)$value +
         integrate(G0_plus, -Inf, -z_L, t, za)$value + integrate(G0_plus, -z_min, z_min, t, za)$value +
         integrate(G0_plus, z_L, Inf, t, za)$value )
      alpha0 <- 2 * integrate(G0_plus, -Inf, Inf, t, za)$value    # (18)
      alpha_diff <- alpha_star - alpha0
      result[i, 1] <- round(t, digits = 3)
      result[i, 2] <- r
      result[i, 3] <- round(alpha_diff, digits = 5)
      i <- i + 1
   }
   result
}
alpha_diff_36(0.25, 1/3, c(1.5, 2.0, 3.0, 4.0))
alpha_diff_36(0.25, 1/2, c(1.5, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0))
alpha_diff_36(0.25, 2/3, c(1.5, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0))
alpha_diff_36(0.35, 1/3, c(1.5, 2.0, 3.0, 4.0))
alpha_diff_36(0.35, 1/2, c(1.5, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0))
alpha_diff_36(0.35, 2/3, c(1.5, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0))
alpha_diff_36(0.5, 1/3, c(1.5, 2.0, 3.0, 4.0))
alpha_diff_36(0.5, 1/2, c(1.5, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0))
alpha_diff_36(0.5, 2/3, c(1.5, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0))
```

補遺 B　3.3 項のシミュレーションを実行する SAS マクロ

```
libname out "C:¥xxx" ; *  実行結果を保存するフォルダ;


%macro sim(ngroup, n, ratio, seed, iteration) ;
data ds1 ;
   call streaminit(&seed) ;
   do rep = 1 to &iteration ;
      do i = 1 to &n ;
%if &ngroup = 2 %then %do ;
         x1 = rand('normal', 0, 1) ;
         x2 = rand('normal', 0, 1) ;
%end ;
%if &ngroup = 3 %then %do ;
         x1 = rand('normal', 0, 1) ;
         x2 = rand('normal', 0, 1) ;
         x3 = rand('normal', 0, 1) ;
%end ;
%if &ngroup = 5 %then %do ;
         x1 = rand('normal', 0, 1) ;
         x2 = rand('normal', 0, 1) ;
         x3 = rand('normal', 0, 1) ;
         x4 = rand('normal', 0, 1) ;
         x5 = rand('normal', 0, 1) ;
%end ;
         y = rand('normal', 0, 1) ;
         output ;
end ;
   end ;
run ;


proc summary data = ds1 noprint ;
   where i <= &n * &ratio ;
   by rep ;
%if &ngroup = 2 %then %do ;
   var x1 x2 ;
```

```
%end ;
%if &ngroup = 3 %then %do ;
   var x1 x2 x3 ;
%end ;
%if &ngroup = 5 %then %do ;
   var x1 x2 x3 x4 x5 ;
%end ;
   output out = interim ;
run ;

data interim ;
   set interim ;
   where _STAT_ = "MEAN" ;
%if &ngroup = 2 %then %do ;
   if max(x1, x2) = x1 then select = "x1" ;
   else if max(x1, x2) = x2 then select = "x2" ;
%end ;
%if &ngroup = 3 %then %do ;
   if max(x1, x2 , x3) = x1 then select = "x1" ;
   else if max(x1, x2, x3) = x2 then select = "x2" ;
   else if max(x1, x2, x3) = x3 then select = "x3" ;
%end ;
%if &ngroup = 5 %then %do ;
   if max(x1, x2 , x3, x4, x5) = x1 then select = "x1" ;
   else if max(x1, x2, x3, x4, x5) = x2 then select = "x2" ;
   else if max(x1, x2, x3, x4, x5) = x3 then select = "x3" ;
   else if max(x1, x2, x3, x4, x5) = x4 then select = "x4" ;
   else if max(x1, x2, x3, x4, x5) = x5 then select = "x5" ;
%end ;
run ;

data ds2 ;
   merge ds1 interim ;
   by rep ;
%if &ngroup = 2 %then %do ;
   if select = "x1" then x = x1 ;
```

```
   else if select = "x2" then x = x2 ;
%end ;
%if &ngroup = 3 %then %do ;
   if select = "x1" then x = x1 ;
   else if select = "x2" then x = x2 ;
   else if select = "x3" then x = x3 ;
%end ;
%if &ngroup = 5 %then %do ;
   if select = "x1" then x = x1 ;
   else if select = "x2" then x = x2 ;
else if select = "x3" then x = x3 ;
   else if select = "x4" then x = x4 ;
   else if select = "x5" then x = x5 ;
%end ;
run ;

proc summary data = ds2 noprint ;
   by rep ;
   var x y ;
   output out = final ;
run ;

data final ;
   set final ;
   where _STAT_ = "MEAN" ;
   lower = (x - y) - probit(1 - 0.025) * sqrt(2/&n) ;
   upper = (x - y) + probit(1 - 0.025) * sqrt(2/&n) ;
   if lower > 0 then sig = "Y" ;
   else sig = "N" ;
   if lower < 0 and 0 < upper then coverage_flg = "Y" ;
   else coverage_flg = "N" ;
run ;

* Type I error ;
proc freq data = final ;
   table sig ;
```

```
    ods output OneWayFreqs = error ;
run ;

data error ;
    set error ;
    where sig = "Y" ;
    error = Percent ;
    keep error ;
run ;

* Coverage probability    ;
proc freq data = final ;
    table coverage_flg ;
    ods output OneWayFreqs = coverage ;
run ;

data coverage ;
    set coverage ;
    where coverage_flg = "Y" ;
    coverage = Percent ;
    keep coverage ;
run ;

* Bias at interim analysis ;
proc summary data = ds2 noprint ;
    where i <= &n * &ratio ;
    by rep ;
    var x ;
    output out = interim_mean mean = interim_mean ;
run ;

proc summary data = interim_mean noprint ;
    var interim_mean ;
    output out = bias_interim mean = bias_interim ;
run ;
```

```
data bias_interim ;
   set bias_interim ;
   keep bias_interim ;
run ;

* Bias at final analysis ;
proc summary data = ds2 noprint ;
   by rep ;
   var x ;
   output out = final_mean mean = final_mean ;
run ;

proc summary data = final_mean noprint ;
   var final_mean ;
   output out = bias_final mean = bias_final ;
run ;

data bias_final ;
   set bias_final ;
   keep bias_final ;
run ;

data result ;
   merge error coverage bias_interim bias_final ;
   ngroup = &ngroup ;
   n = &n ;
   ratio = &ratio ;
   seed = &seed ;
   iteration = &iteration ;
run ;

data out.result ;
   set out.result result ;
run ;
%mend ;
```

```
data out.result ;
   stop ;
run ;
%sim(2, 100, 0.25, 1, 100000) ;
%sim(2, 100, 0.5, 2, 100000) ;
%sim(2, 100, 0.75, 3, 100000) ;
%sim(3, 100, 0.25, 4, 100000) ;
%sim(3, 100, 0.5, 5, 100000) ;
%sim(3, 100, 0.75, 6, 100000) ;
%sim(5, 100, 0.25, 7, 100000) ;
%sim(5, 100, 0.5, 8, 100000) ;
%sim(5, 100, 0.75, 9, 100000) ;
```

補遺 C　3.6.2 項の Bowden and Glimm (2008)による UMVCUE と信頼区間のシミュレーションを実行する SAS コード

```
/*
Conditionally unbiased estimation in the normal setting with known variances, Bowden and Glimm (2008)

Assumption for this macro:
- Treatment selection is based on mean of treatment effect.
- Variance for data is common among treatments and both stage1 and stage 2.
- Sample size across treatment group is common.

Included macros:
%run_all: Macro to run all major macro (%makedata, %biasmse, %coverci)
  Need to define following variables.
  k: number of treatments. k SHOULD be equal to max of number of mu
  mu1,...muk: true treatment effect "mu" for each treatment up to k=10. If you want more than 10 treatments,
please add macro viables (e.g, mu11=, mu12=, ...) in %run_all
  comsig: common SD of data among treatments and both stage1 and stage 2
  nsub1: N /arm for    stage 1, assuming same sample size across arms. If not, then set maximum sample size and
delete unnecessary sample size by modifing %makedata;
  nsub2: N /arm for    stage 2
  scenario: to use unique number of scenario in each simulation setting

Major macro:
%makedata: to make simulation data based on both given true treatment effect (mu1,...,muk) and null treatment
effect (all mu = 0) in order to estimate maximal variance for confidence interval
%biasmse: to calculate bias and mse for all ranked treatment (1st, 2nd,...kth)
%coverci: to calculate coverage proability of confidence interval for all ranked treatment (1st, 2nd,...kth)

Nested macro:
%estimate_l: to estimate UMVCUE given lth rank treament of mean
%_biasmse: to calculate bias and mse for lth rank treament
%umvci:    to calculate coverage proability of confidence interval for lth rank treament
*/
```

```
/* x for stage 1 data, y for stage 2 data */
%macro makedata;
/*data for UMVCUE*/
data data01;
   call streaminit(20200120);
   do i=1 to &sim.;
      %do trt = 1 %to &k.;
            trt=&trt;
            mu_true = &&mu&trt;
      do sub1 = 1 to &nsub1.;
         x=RAND('NORMAL') * &comsig. + &&mu&trt;
            output;
            end;
            %end;
   end;
run;

data data02;
   call streaminit(9987633);
   do i=1 to &sim.;
      %do trt = 1 %to &k.;
            trt=&trt;
            mu_true = &&mu&trt;
      do sub1 = 1 to &nsub2.;
         y=RAND('NORMAL') * &comsig. + &&mu&trt;
         output;
            end;
            %end;
   end;
run;

/* data for confidence interval of UMVCUE */
data data03;
   call streaminit(8457940);
   do i=1 to &sim2.;
```

```
    %do trt = 1 %to &k.;
        trt=&trt;
        mu_true = 0;
    do sub1 = 1 to &nsub1.;
      x=RAND('NORMAL') * &comsig. + 0;
         output;
        end;
        %end;
  end;
run;

data data04;
  call streaminit(117083);
  do i=1 to &sim2.;
    %do trt = 1 %to &k.;
        trt=&trt;
        mu_true = 0;
    do sub1 = 1 to &nsub2.;
      y=RAND('NORMAL') * &comsig. + 0;
      output;
        end;
        %end;
  end;
run;
%mend;

%macro estimate_l(l=); * lth highest rank is your interest ;
/*calculate mean of X, mean of Y*/
proc univariate data=tmp01 noprint;
  var x;
  by i trt;
  output out=stat01 n=n_x mean=mean_x;
run;

proc univariate data=tmp02 noprint;
  var y;
```

```
   by i trt mu_true;
   output out=stat03 n=n_y mean=mean_y;
run;

proc rank data=stat01 out=tmp11 DESCENDING;
   var mean_x;
   ranks mean_xs;
   by i;
run;

data tmp12;
   merge tmp11 stat03;
   by i trt;
run;

data tmp13;
   set tmp12(where=(mean_xs = &l.)) ;
   Zl = n_x* mean_x + n_y * mean_y;
run;

/*need to search Wl,l+1. Wl,l-1, which requires the divided cases*/
%macro rqserch();
%if &l. = &k. %then %do;
data tmp14;
   set tmp12;
   if mean_xs = %eval(&l. - 1);
   keep i mean_x;
   rename mean_x=mean_x_m1;
run;

data wl;
   merge tmp13 tmp14;
   by i;
   wlp1 = 99999999; * infinity ;
   wlm1 = sqrt(n_x * (n_x + n_y)/n_y) / &comsig. *(Zl / (n_x + n_y) - mean_x_m1);
   keep i wlp1 wlm1;
```

```
run;
%end;

%else %if &l. = 1 %then %do;
data tmp14;
   set tmp12;
   if mean_xs = %eval(&l. + 1);
   keep i mean_x;
   rename mean_x=mean_x_p1;
run;

data wl;
   merge tmp13 tmp14;
   by i;
   wlp1 = sqrt(n_x * (n_x + n_y)/n_y) / &comsig. *(Zl / (n_x + n_y) - mean_x_p1);
   wlm1 = -99999999; * -infinity ;
   keep i wlp1 wlm1;
run;
%end;

%else %do;
data tmp14;
   set tmp12;
   if mean_xs = %eval(&l. - 1);
   keep i mean_x;
   rename mean_x=mean_x_m1;
run;

data tmp14_;
   set tmp12;
   if mean_xs = %eval(&l. + 1);
   keep i mean_x;
   rename mean_x=mean_x_p1;
run;

data wl;
```

```
   merge tmp13 tmp14 tmp14_;
   by i;
   wlp1 = sqrt(n_x * (n_x + n_y)/n_y) / &comsig. *(Zl / (n_x + n_y) - mean_x_p1);
   wlm1 = sqrt(n_x * (n_x + n_y)/n_y) / &comsig. *(Zl / (n_x + n_y) - mean_x_m1);
   keep i wlp1 wlm1;
run;
%end;
%mend;

%rqserch();

data tmp21;
   merge tmp13 wl;
   by i;
   l=&l.;
   mle =Zl/(n_x + n_y) ;
   umvcue = Zl/(n_x + n_y) - sqrt(n_x/(n_y * (n_x +n_y))) * &comsig. * (PDF('NORMAL', wlp1) -
PDF('NORMAL', wlm1)) / (CDF('NORMAL', wlp1) - CDF('NORMAL', wlm1));
   b_mle = mle - mu_true;
   b_umvcue = umvcue - mu_true;
   mse_mle = (mle - mu_true)**2;
   mse_umvcue = (umvcue - mu_true)**2;
run;

proc sort data=tmp21 out=tmp22;
   by trt;
run;
%mend;

%macro _biasmse(lt=);
data tmp01; set data01; run;
data tmp02; set data02; run;

%estimate_l(l=&lt.);

%macro _calc(param=);
```

```
/*Calculate expectation(mu_l - mu)*/
proc univariate data=tmp22 noprint;
   var &param.;
   by trt;
   output out=stat05 n=n mean=mean;
run;

data tmp23;
   set stat05;
   e_p = mean * n / &sim.;
run;

proc univariate data=tmp23 noprint;
   var e_p;
   output out=stat_&param. sum=&param. ;
run;
%mend;

%_calc(param=b_mle);
%_calc(param=b_umvcue);
%_calc(param=mse_mle);
%_calc(param=mse_umvcue);

data tmp31_&lt.;
   attrib _ALL_ label=" ";
   merge stat_:;
   l=&lt.;
run;
%mend;

%macro _umvci(lt=);
data tmp01; set data03; run;
data tmp02; set data04; run;

%estimate_l(l=&lt.);
```

```
proc univariate data=tmp22 noprint;
   var umvcue;
   output out=_mean_umv n=n mean=mean;
run;
data mean_umv_&lt.;
   attrib _ALL_ label=" " ;
   set _mean_umv;
   l=&lt.;
run;

proc univariate data=tmp22 noprint;
   var umvcue;
   output out=_sd_umv0 std=sd_umv;
run;

data sd_umv0_&lt.;
   attrib _ALL_ label=" " ;
   set _sd_umv0;
   l=&lt.;
   v=sd_umv**2;
run;

data tmp01; set data01; run;
data tmp02; set data02; run;

%estimate_l(l=&lt.);
data tmp41;
   merge tmp21 sd_umv0_&lt.;
   by l;
   ci_l =umvcue - sd_umv * quantile('NORMAL', .975);
   ci_u =umvcue + sd_umv * quantile('NORMAL', .975);
   if ci_l <= mu_true <= ci_u then cover=1;
   else cover=2;
run;

proc freq data=tmp41 noprint;
```

```
   tables cover/ out =freq01_&lt.;
run;

data _coverage_&lt.;
   set freq01_&lt.;
   if cover=1;
   l=&lt.;
run;
%mend;

%macro biasmse;
%do trt=1 %to &k.;
%_biasmse(lt=&trt.);
%end;

data _bias_mse;
   set tmp31_:;
   %do i = 1 %to &k.;
   mu&i = &&mu&i;
   %end;
   comsig=&comsig.;
   nsub1=&nsub1.;
   nsub2=&nsub2.;
run;
%mend;

%macro coverci;
%do trt=1 %to &k.;
%_umvci(lt=&trt.);
%end;

data _coverage;
   set _coverage_:;
   %do i = 1 %to &k.;
   mu&i = &&mu&i;
   %end;
```

```
   comsig=&comsig.;
   nsub1=&nsub1.;
   nsub2=&nsub2.;
   drop cover count;
   label percent =;
run;
%mend;

%macro run_all(k=, mu1=, mu2=, mu3=, mu4=, mu5=, mu6=, mu7=, mu8=, mu9=, mu10=, comsig=, nsub1=,
nsub2=, scenario=, sim=100000, sim2=50000);

%makedata;

%biasmse;

data bias_mse_s&scenario.;
   set _bias_mse;
   scenario=&scenario.;
run;

%coverci;

data coverage_s&scenario.;
  set _coverage;
  scenario=&scenario.;
run;

%mend;
```

本項の図 11，図 12 を作成する実行文

```
/*
Conditionally unbiased estimation in the normal setting with known variances
Bowden and Glimm 2008
*/

/*Make sure to run "UMVCUE_Bowden2008_macro.sas" before running this program*/
```

```
/*For running macro, you can use below */
/*%run_all(k=,mu1=,mu2=,mu3=,mu4=,mu5=,mu6=,mu7=,mu8=,mu9=,mu10=, comsig=,
nsub1=,nsub2=,scenario=);*/

/*
Illustration of runnig simulation program for Bowden and Glimm (2008)
*/

/*Figure 1(left)
Situateion:
two-stage trial involving three candidate treatments
(mu1, mu2, mu3) = (0, 1/2, 0)
variane of estimate are sigma1^2 = 1/n1 for 1st stage, sigma2^2 = 1/n2 for 2nd stage => common variance =1
n1 + n2 =10
Information at interim, I = (1/sigma1^2)/(1/sigma1^2 + 1/sigma2^2) = n1/(n1+n2)
Figure 1 (left) shows Monte-Carlo estimates for the MSE as I is varied between 0 and 1
*/

proc datasets lib=work memtype=data kill nolist;
quit;
%run_all(k=3,mu1=0,mu2=0.5,mu3=0, comsig=1, nsub1=1,nsub2=9,scenario=1);
%run_all(k=3,mu1=0,mu2=0.5,mu3=0, comsig=1, nsub1=2,nsub2=8,scenario=2);
%run_all(k=3,mu1=0,mu2=0.5,mu3=0, comsig=1, nsub1=3,nsub2=7,scenario=3);
%run_all(k=3,mu1=0,mu2=0.5,mu3=0, comsig=1, nsub1=4,nsub2=6,scenario=4);
%run_all(k=3,mu1=0,mu2=0.5,mu3=0, comsig=1, nsub1=5,nsub2=5,scenario=5);
%run_all(k=3,mu1=0,mu2=0.5,mu3=0, comsig=1, nsub1=6,nsub2=4,scenario=6);
%run_all(k=3,mu1=0,mu2=0.5,mu3=0, comsig=1, nsub1=7,nsub2=3,scenario=7);
%run_all(k=3,mu1=0,mu2=0.5,mu3=0, comsig=1, nsub1=8,nsub2=2,scenario=8);
%run_all(k=3,mu1=0,mu2=0.5,mu3=0, comsig=1, nsub1=9,nsub2=1,scenario=9);

data tmp53;
   set bias_mse_s:;
   vec = "c("||trim(left(mu_1))||","||trim(left(mu_2))||","||trim(left(mu_3))||")";
   if l=1;
   info=nsub1/(nsub1 + nsub2);
run;
```

```
title 'Figure 1 (left) in Bowden    and Glimm (2008)';
proc sgplot data=tmp53 ;
   series x=info y=mse_umvcue /;
   scatter x=info y=mse_umvcue / legendlabel= "UMVCUE";
   series x=info y=mse_mle /;
   scatter x=info y=mse_mle / legendlabel ="MLE"    markerattrs=(symbol=square);
   keylegend /EXCLUDE=("mse_umvcue" "mse_mle");
run;

/*Figure 3, 5
Situation:
A trial with k = 5 treatments
Consider several sets of (mu1, mu2, mu3, mu4, mu5)
Each stage 1 treatment mean is given an underlying variance of 1/2, stage 2 statistic with a variance of 1
Information at interim, I = (1/sigma1^2)/(1/sigma1^2 + 1/sigma2^2) = 2/3
Based on above situation, this simulation is done with common SD =5, n1= 50, n2=25
*/

proc datasets lib=work memtype=data kill nolist;
quit;
%run_all(k=5, mu1=0, mu2=0.2, mu3=0.4, mu4=0.6, mu5=0.8, comsig=5, nsub1=50, nsub2=25, scenario=1);
%run_all(k=5, mu1=0, mu2=0.6, mu3=1.2, mu4=1.8, mu5=2.4, comsig=5, nsub1=50, nsub2=25, scenario=2);
%run_all(k=5, mu1=0, mu2=0, mu3=0, mu4=0, mu5=0, comsig=5, nsub1=50, nsub2=25, scenario=3);

data tmp51;
   set bias_mse_s:;
   vec = "c("||trim(left(mu1))||","||trim(left(mu2))||","||trim(left(mu3))||","||trim(left(mu4))||","||trim(left(mu5))||")";
run;

title 'Figure 3 (right) in Bowden and Glimm (2008)';
proc sgplot data=tmp51;
   series x=l y=mse_umvcue / group=vec    GROUPORDER=ASCENDING    ;
   scatter x=l y=mse_umvcue / group=vec    GROUPORDER=ASCENDING;
   series x=l y=mse_mle / group=vec    GROUPORDER=ASCENDING LINEATTRS=(pattern=2) ;
   scatter x=l y=mse_mle / group=vec    GROUPORDER=ASCENDING;
```

```
   xaxis label='mu(j), j=1,..,5' ;
   yaxis min=0.15 max=1.0 label="MSEsel";
run;

data tmp52;
   set coverage_s:;
   vec = "c("||trim(left(mu1))||","||trim(left(mu2))||","||trim(left(mu3))||","||trim(left(mu4))||","||trim(left(mu5))||")";
run;

title 'Figure 5 in Bowden and Glimm (2008)';
proc sgplot data=tmp52;
   series x=l y=percent / group=vec    GROUPORDER=ASCENDING    ;
   scatter x=l y=percent / group=vec    GROUPORDER=ASCENDING;
   xaxis label='mu(j), j=1,..,5' ;
   refline 95/ axis=Y LINEATTRS=(pattern=2) ;
   yaxis min=94.5 max=98 label="Coverage of 95% CI";
run;
```

執筆者

日本製薬工業協会　医薬品評価委員会　データサイエンス部会　2022 年度継続タスクフォース 1

タスクフォースメンバー

氏名	所属	担当
田中　勇輔	アステラス製薬株式会社	2.2.2 項
大澤　志乃	中外製薬株式会社	2.2.3, 2.2.4 項
吉田　瑞樹	ファイザーR&D 合同会社	2.2.5, 2.2.6 項
青木　誠	ノバルティス ファーマ株式会社	2.3 項
高津　正寛	持田製薬株式会社	3.5.1 項
中村　将俊	ファイザーR&D 合同会社	3.5.1, 3.6.1 項
沖野　邦明	日本ベーリンガーインゲルハイム株式会社	3.5.2 項
中山　高志	グラクソ・スミスクライン株式会社	3.5.2, 3.6.2 項
飯塚　政人	田辺三菱製薬株式会社	3.5.3 項

タスクフォースリーダー兼推進委員

角野　修司	武田薬品工業株式会社	1, 2.1, 2.2.1 項
棚瀬　貴紀	大鵬薬品工業株式会社	1, 3.1-3.4, 3.5.3, 4 項

担当副部会長

菅波　秀規	興和株式会社