総 説

臨床試験における被験者数再設定
—方法論の概説と統計学的留意点—

Sample Size Re-estimation in Clinical Trials:
Overview of Methods and the Statistical
View Points to Consider

上村 鋼平*
Kohei Uemura

独立行政法人 医薬品医療機器総合機構 審査センター生物統計領域
Biostatistics Group, Center for Product Evaluation,
Pharmaceuticals and Medical Devices Agency
e-mail: uemura-kohei@pmda.go.jp

Recently, it is becoming more challenging to cope with low success probabilities, much cost and severe competitions in a new drug development. An adaptive design is considered to be a promising tool for efficient development for its various types of adaptations and applications in multiple development stages. However, its flexibility does not necessarily make the development more efficient and it may always allow a risk of operational bias to use such a design. Thus, to consider the specific benefit and risk of each adaptive design using quantitative measures under a concrete setting, it is important to make intensive discussions and to share experiences between each side persons in charge. In this review, focusing on sample size re-estimation which is a relatively simple and basic adaptive design, I will give some outlines of methods and review statistical points to consider.

Key words: adaptive design; sample size re-estimation; type I error; operational bias; trial simulation.

1. はじめに

近年、医薬品開発の場面において、開発成功確率の低下、開発コストの増加及び開発競争の激化といった現状を背景に、効率的な開発戦略・方法への関心が高まっている (FDA, 2004). 効率化を目指した臨床開発の方法の 1 つとして、アダプティブデザインの利用が挙げられる (O'Neill, 2006). アダプティブデザインをどの程度広く又は狭く捉えるべきかについては議論があるが、本稿では以下のように定義する。アダプティブデザインとは、臨床試験実施中のある時点までに集積されたデータの間解分析結果や盲検下レビューの情報に基づき、試験継続の可否及び/又は試験

* 本稿は、著者の個人的見解に基づくものであり、PMDA の公式見解を示すものではない。

Received May 2012. Revised July 2012. Accepted July 2012.

のデザインの側面に係わる、何らかの決定又は修正の判断(以下、「アダプテーション」と表現)を行う機会を有する臨床試験デザインである。ここで、アダプティブデザインでは、アダプテーションの手順や方法の詳細が、アダプテーションのための情報を確認する以前の前向きな段階において(以下、「事前に」と表現)計画されていることが必要である。

アダプティブデザインに係わる提案は、1990年代初頭より学術誌を中心に数多くなされ、2000年代に入ってから、*Biometrical Journal* において2006年8月の特集号が刊行される等、世界的に産官学の間で非常に活発な議論がなされてきた。このような流れの中で、米国研究製薬工業協会(Pharmaceutical Research and Manufacturers of America, PhRMA)が、アダプティブデザイン作業部会(adaptive design working group)を2005年3月に立ち上げ、その提言内容や文献レビューを白書として整理し(Gallo et al., 2006)、欧州医薬品庁(European Medicines Agency, EMA)のヒト用医薬品委員会(Committee for Human Medicinal Products, CHMP)が2007年10月にアダプティブデザインを用いた検証的臨床試験における方法論的な問題点に関する Reflection Paper(CHMP, 2007)を示している。さらに、米国食品医薬品局(Food and Drug Administration, FDA)が2010年2月にアダプティブデザイン臨床試験に関する企業向けガイダンス案(FDA, 2010)を発出した。ここで、本稿の目的及び本稿における被験者数再設定に係わる議論の基軸となる背景を示す目的で、まずFDAの行った整理の全般的な特徴から言及していくと、これまでに提案されてきた各タイプのアダプティブデザインに対し、acceptable か non-acceptable かという軸ではなく、well-experienced か less-well-experienced かを軸に、well-understood か less-well-understood かという概念に基づくグループ分けを行ったということがある。従来の議論の多くでは、アダプティブデザインを利用すれば確実に試験が効率的となり、その一方で試験のオペレーションに伴うバイアスを確実にもたらすということが、ある程度前提となっており、アダプティブデザインの利用の是非のみが特段に問われてきた傾向があった。これに対し、このガイダンス案では、統計学的な問題(妥当性のみならず効率性の問題も含む)及びオペレーション上の問題が存在することを認識しておくだけでなく、それらの起き易さ、起きた場合の深刻さと対処可能性等も含め、個々のアダプティブデザインの利点と欠点が定量的あるいは具体的に十分見極められ、経験に基づく理解と考察が今後より深まっていくことの重要性が強調されていると捉えることもできる。このような解釈を前提にすれば、真に開発の効率化を目指していくためには、単純に less-well-understood なアダプティブデザインの利用の是非を問うというよりは、個々の試験デザインや開発戦略について計画段階で十分な議論を行い、個々の臨床試験や開発品目における経験についても十分な分析と考察の共有を可能な限り積み重ねていくことが、今後重要になってくるものと考えられる。よって、次に触れる本邦での話にも係わるが、産官学の間の良いコミュニケーションと具体的で開かれた議論の場が、基本的に国内外を問わずさらに発展していくことに期待したい。

本邦の医薬品医療機器総合機構(Pharmaceuticals and Medical Devices Agency, PMDA)は、アダプティブデザインに関するガイダンスを未だ発表していないものの、これまで治験相談や審査を通し、個別の議論や経験をある程度蓄積しつつある。アダプティブデザインに係わる相談内容としては、過去にはアダプティブデザインの導入の可能性について試験計画の詳細が十分に議論

されていない段階で相談事項に挙げられることもあったが、現在ではアダプテーションを伴う臨床試験計画の詳細が提出され、具体的な議論が可能な事例も多くなっている(安藤, 2012)。しかしながら、治験相談における議論には、時間的又は進行上の制約もあり、詳細に統計学的な議論を行っておくべきと思われる側面等については、必ずしも十分な議論や意見交換がなされているとは言い切れない。臨床試験に係わるプラクティスを産官学全体にわたり発展させていくためには、実務担当者間での議論をより充実させ、より具体的に検討内容を深めていくプロセスを互いに重要視していく必要があると考える。

そこで、本稿では、アダプティブデザインの理論の基礎となっており、アダプテーションの内容が限定的で比較的議論が行い易い被験者数再設定の方法論について、実務担当者間での議論の発展に期待し、方法論の解説及び統計学的な論点の整理を目指した総説を示す。2 節は、well-understood なグループに入る盲検下の被験者数再設定を、3 節は、less-well-understood なグループに入る非盲検下の被験者数再設定を対象とする。4 節で被験者数再設定の実用上の課題について考察する。なお、本稿は、2011 年 11 月に開催された計量生物セミナーにおける著者の講演内容に基づくものである。本稿は、著者の個人的見解に基づくものであり、PMDA の公式見解を示すものではない。

2. 盲検下の被験者数再設定

盲検下の被験者数再設定とは、臨床試験の主要評価変数が連続量の場合の分散や 2 値の場合の群を併合した反応割合又は対照群の反応割合といった局外パラメータについて、その事前の見積もり値が、被験者数の設定根拠として妥当なものであったか否か、盲検下で得られた情報に基づく確認を行い、局外パラメータの設定に対し修正が必要な場合には、その修正に従い被験者数を再設定するための機会が設けられた試験計画のことである。一方、2.2 節の議論における対照デザインであり、3 節で方法論の解説を行う非盲検下の被験者数再設定とは、中間解析時点で非盲検下のデータへアクセスし推定された群間差又は局外パラメータと群間差の両方に基づき被験者数再設定を実施する機会を有する試験計画のことである。盲検下の被験者数再設定が提案される以前に、非盲検下で局外パラメータに基づき被験者数を再設定する内部パイロット試験の提案がなされており、2.1 節では、内部パイロット試験及び盲検下の被験者数再設定の手法について概説し、2.2 節で非盲検下の被験者数再設定との考え方の比較について議論を行う。

2.1 内部パイロット試験及び盲検下の被験者数再設定の概説

内部パイロット試験(Wittes and Brittain, 1990)による被験者数再設定は、非盲検下の中間解析データに基づく局外パラメータの推定値を利用するため、第 1 種の過誤確率が上昇し得ることがわかっている。Wittes et al. (1999) は、最終解析で t 検定を用いる状況にて、内部パイロット試験に基づく被験者数再設定を行うと、以下の式に基づき、帰無仮説の下で群内分散の推定量に過小評価方向のバイアスが発生することを示した。

$$E[s_{final}^2] = \sigma^2 + (n_1 - 1) \text{cov} \left[s_1^2, \frac{1}{\hat{N}_{final} - 1} \right], \quad (1)$$

s_{final}^2 , $\hat{N}_{final}$ は最終解析の t 検定に用いる併合分散及び 1 群あたりの被験者数, σ^2 は分散の真値, n_1 , s_1^2 は内部パイロット試験データの 1 群あたりの被験者数及び併合分散とする. s_1^2 が想定より大きいほど $\hat{N}_{final}$ を増加させるため, (1) 式の共分散は負の値となり, $E[s_{final}^2] < \sigma^2$ となる. 分散推定量のバイアスが第 1 種の過誤確率へ与える影響は, 症例数の増加のみを検討する場合 (restricted design) は小さく, 症例数の減少も検討する場合 (unrestricted design) には大きくなる. unrestricted design では第 1 種の過誤確率が上昇し, 特に内部パイロット試験のサイズが小さい場合には無視できない程度になる (Wittes et al., 1999; Birkett and Day, 1994). Stein (1945), Zucker et al. (1999), Proschan and Wittes (2000) では, 最終解析で用いる分散推定量を調整することにより第 1 種の過誤確率が制御可能になる方法が提案されている.

内部パイロット試験による被験者数再設定に対し, Gould (1992), Gould and Shih (1992), Gould (1995) は, 非盲検下のデータにアクセスすることの危険性及びアクセスを可能にするためのオペレーション上の負担等を重要視し, 盲検下データに基づき局外パラメータを見直すための, いわゆる盲検下の被験者数再設定方法を提案した. 結果変数が連続量の場合について説明する. 提案法の 1 つは, 群を併合した盲検下データより単純に推定した分散 s_{lumped}^2 と事前に仮定した群間差 δ_{pre} を用いて最終解析の t 検定に用いる併合分散を以下のように推定することにより, 被験者数を調整するものである.

$$\hat{s}_{final}^2 = \frac{(2n_1 - 1)s_{lumped}^2 - n_1\delta_{pre}^2/2}{2(n_1 - 1)}, \quad (2)$$

$s_{lumped}^2 = (2n_1 - 1) \sum_{i=1}^{2n_1} (Y_i - \bar{Y}_1)^2$, $\bar{Y}_1 = (2n_1)^{-1} \sum_{i=1}^{2n_1} Y_i$, Y_i は対象者 i の結果変数とする. この方法は, 事前に仮定した群間差が過大評価されていると, 併合分散を過小評価してしまうことがあるため, 盲検下データに基づく分散 s_{lumped}^2 を単純に用いて被験者数を調整する, より保守的な方法も提案されている. 群間差の真値によって s_{lumped}^2 と s_{final}^2 がどの程度食い違うかは, 分散分析の平方和の分解より, 以下のようになり, effect size δ/σ に依存することがわかる.

$$E[s_{lumped}^2] = \frac{(n_1/2)\{\text{var}(\hat{\delta}_1) + \delta^2\} + 2(n_1 - 1)\sigma^2}{2(n_1 - 1)} \approx \sigma^2 \left\{ 1 + \frac{(\delta/\sigma)^2}{4} \right\}, \quad (3)$$

$\hat{\delta}_1$ は盲検下データの盲検を解除した場合に得られる群間差の推定値を指す. 例えば, $\delta/\sigma = 1/3$ の場合, s_{lumped}^2 と s_{final}^2 との間で生じる違いは, (3) 式より平均的には 2.8% の分散の上昇, 1.4% の標準偏差の上昇に相当する程度になる. 従って, effect size があまり大きくなければ, s_{lumped}^2 を単純に利用して被験者数再設定を行っても実務上は大きな問題は無いと考えられる. 2 値データの場合を含み, より詳細には Proschan (2005) を参照されたい. また, Gould and Shih (1992), Gould (1995) は, EM アルゴリズムを用いたより複雑な分散推定量も提案しており, さらにこの提案に係わる議論 (Friede and Kieser, 2002; Gould and Shih, 2005) や他の提案 (Kieser and Friede, 2003; Govindarajulu, 2003) もなされている.

2.2 盲検下 vs. 非盲検下

被験者数再設定の実施を検討する際に検討すべき点として, 盲検下で実施すべきか, それとも非盲検下で実施すべきかという点がある. オペレーショナルバイアスの危険を踏まえると, 盲検下で実施する方が望ましいことは明らかである. 一方, 最近では, 被験者数再設定といえば非盲

検下で実施する場合を指すことが多いと思われるものの、被験者数再設定自体の利用の必要性や有用性を議論する際には、被験者数再設定を非盲検下で実施することを前提とするのではなく、盲検下と非盲検下のどちらで実施すべきかという点についても、議論を詰めておくことに意義があると考えられる(Shih, 2001)。その理由は、その議論の本質には、オペレーション上の観点だけでなく、局外パラメータを重視すべきか、それとも局外パラメータ及び群間差の両方又は後者を重要視すべきかという観点も含まれており、そこが必ずしも明らかではないと考えられるからである。

そこで、①盲検下の被験者数再設定又はその背景にある内部パイロット試験の考え方と、②非盲検下の被験者数再設定の考え方の比較について議論を行ってみる。まず、Proschan (2005) や Stein et al. (2006) によれば、①の想定している状況としては以下のような捉え方があるようである。被験者数の設定に用いる群間差は、臨床的意義及び類似する試験の結果の両方に基づくものである。例えば、慢性閉塞性肺疾患(chronic obstructive pulmonary disease, COPD)の患者における呼吸機能の改善を目指した開発を行う場合に、FEV₁ (努力性呼気量の一秒量)のベースラインからの変化量において 100 mL が臨床的に意義のある最小の群間差とされていたとする。一方で、他試験で 150 mL 以上の群間差でもって医薬品の有効性が証明された試験があったとする。これらの情報に基づく、実施予定の本試験の計画の際には、100~150 mL の群間差が仮説として設定されるかもしれない。しかし、仮に他試験で臨床的意義のある FEV₁ 変化量を下回る効果しか認められなかったとした場合には、臨床的意義のある最小の群間差に基づき 100 mL 付近の群間差が本試験の仮説として設定されるかもしれない。これらを言い換えるとすると、被験者数設定において設定される群間差の仮説は、必ずしも群間差の真値の良い見積もりとなっているとは限らず、臨床的価値を見出せると思われる値が反映されるかもしれないということである。

それに対し局外パラメータについては、正確な値を推定した上で被験者数を設定せざるを得ない。そこで、本試験の対象者の局外パラメータを推定する上では本試験のデータ以上に最適なデータは無いという考え方(Wittes and Brittain, 1990)に基づき、盲検下の被験者数再設定又は内部パイロット試験による被験者数再設定の方法論が展開されてきた。特に、内部パイロット試験では、群間差を途中で確認できるにもかかわらず敢えてそこは調整しようとは考えられていない点を踏まえると、あくまで群間差は検証的試験の仮説そのものなので、事前に定義された値より変更すべきではなく、その仮説を検証するために必要な試験精度を維持するために局外パラメータを調整することで被験者数を再設定するものと考えることができる。

次に、②非盲検下の被験者数再設定では、想定している状況が異なるのかもしれないが、被験者数設定に係わる考え方が、①とは異なっている(Stein et al., 2006)。ここでは、事前に設定された群間差は、局外パラメータと同様に被験者数設定(検出力の維持)に必要な仮定の 1 つとして位置づけられ、それが試験の仮説として固定されるというよりは、事前情報に基づき仮定した値が真値の良い見積もりとなっているかどうか問題にされる。また、臨床的意義の観点からは、真の差が仮定した群間差よりも若干小さい差であっても、依然として臨床的意義を有する差であるならば、検出力を維持すべく仮定を下方修正し被験者数を再設定するという状況を想定した考え方になる。この考え方は、臨床的意義のある最小の群間差が設定可能な状況においては成り立たないことに留意が必要(Hung et al., 2006)であるものの、事前情報による推定値(例えば、被験

薬の探索的試験における有効率)や期待値(例えば、類薬の試験成績を踏まえた被験薬の期待有効率)に基づき仮定した群間差が、臨床的に意味のある最小の差よりも大きい差になると考えられる状況においては、実際に仮説を下方修正する余地が多少残されている。実務上の観点からは、最初から保守的な値や臨床的に意義のある最小の群間差を仮定すれば問題無いとの比較的一般的な考えがある一方で、臨床的意義のある最小の群間差に基づく被験者数は非現実的な規模や試験期間が長期間となるために最初に設定不可能である場合や、最初から保守的に設定可能な被験者数という意味でも一定の限界が生じる場合等において、やはり念のため設定した群間差の仮定の適切性を確認しておく機会が必要である、といった考え方を②のデザインでは想定しているのかもしれない(Schäfer, 2006)。

被験者数の設定根拠は非常にシンプルに示されるものではあるが、実際には、様々な要素が関係してくるため、非常に複雑な過程を経た結果、最終的に決定されているものと捉えることは1つ有用なことである。①と②における考え方の比較を踏まえても、被験者数再設定に対する捉え方には、個別の開発状況、開発戦略、疾患領域の特徴等に依存する部分が存在するだけでなく、学術的な立場、実務上の立場、規制上の立場による違いや各立場内での違いが存在することは、容易に想像できる。規制上の立場からは、探索的試験で十分な検討を経た上で設定された仮説に対し検証的試験で最終的に証明する、というプロセスを踏んでいることが重要視されており、検証的試験の仮説を試験実施途中に変更してしまった場合、検証的試験としては成立しなくなるという考え方もある。このような、仮説を変更すべきではないという考え方は、どちらかということと内部パイロット試験や盲検下の被験者数再設定における群間差に対する考え方を支持するものと考えられる。ただし、誤解が無いようにすべき点は、非盲検下の被験者数再設定においても、証明すべき仮説が変更される訳ではなく、あくまで群間差が少なくとも0又はある一定以上の値を下回ることは無いという仮説を証明することに変更は生じないため、必ずしも前述の検証的試験の考え方と齟齬が生じる訳では無いということである。つまり、検証すべき仮説と被験者数の設定根拠に用いる仮説とを区別して捉える必要性は少なくともある。一方、被験者数の設定根拠に掲げた仮説の変更は、いわば結果の期待値が下方修正されることになる。当初より臨床的意義の観点から結果の期待値があまり有望なものとも言えない状況等においては、仮説が有意水準付近で検証されるような(マージナルな)結果が得られた際の結果解釈の問題がより深刻になることが懸念される。マージナルに有意差を示せない場合を救うために被験者数再設定を実施したいと考えること自体には、一定の理解が成立する場合もあると考えられるものの、検定の有意性と臨床的有意性の乖離問題の起き易さとそれを踏まえた被験者数再設定の有用性については、事前にシミュレーション等によって十分確認し、慎重に検討しておくことが必要になると考えられる(Hung, Wang and O'Neill, 2011)。

3. 非盲検下の被験者数再設定の統計学的課題

この節では、被験者数再設定における基本的な統計学的課題である第1種の過誤確率の制御及び群間差の推定問題に関する方法論について解説すると共に、実務上の統計学的留意点について検討する。

3.1 第1種の過誤確率の制御

中間解析時点で非盲検下のデータへアクセスし推定された群間差に基づき被験者数再設定を実施する機会を有する試験計画では、固定デザイン(最終解析時の被験者数が固定された中間解析を伴わない計画)や群逐次デザイン(最大情報量が固定された中間解析を伴う計画)で用いられる通常の最終解析方法(以下、被験者数再設定の最終解析においては「ナイーブな解析」と表現)を用いると、第1種の過誤確率が上昇し得ることが知られている(Proschan and Hunsberger, 1995; Cui, Hung and Wang, 1999; Shun et al., 2001). これまでに、第1種の過誤確率を名義水準以下に制御することが可能な方法がいくつか提案されている(Bauer and Köhne, 1994; Proschan and Hunsberger, 1995; Fisher, 1998; Cui, Hung and Wang, 1999; Lehmacher and Wassmer, 1999; Chen, DeMets and Lan, 2004). これらの方法は、どのようなフレームワークで提案されているか(2ステージデザイン又は群逐次デザイン、用いる検定統計量のタイプ等)に起因する差異はあるものの、それらの基づく原理の観点から次の4つのタイプに分けて理解することが可能である。ナイーブな解析による第1種の過誤確率の上昇の原理を含め、順に説明する。

- ① Combination Test の原理
- ② Conditional Error Function の原理
- ③ 制約付き被験者数再設定
- ④ シミュレーションに基づく第1種の過誤確率の制御

プラセボ対照2群比較試験の実施を想定し、原理の理解が容易になるよう、最も単純化した2ステージデザインのフレームワークを用いる。すなわち、中間解析を1回のみ実施する状況を考え、中間解析以前のデータを第1ステージ、中間解析以降最終解析までのデータを第2ステージと呼ぶこととする。また、簡単のため、最初は中間解析における試験中止は行わない状況を考え、その後試験中止を伴う場合への拡張についての説明を補う。群 j ($j=1$: 被験薬群, 2 : プラセボ群)に割り付けられた i 番目の被験者の主要評価変数 Y_{ij} は、独立に平均がそれぞれ $\mu_1 = \delta$, $\mu_2 = 0$, 共通の分散 $\sigma^2 = 1$ の正規分布に従うとし、添え字無しの δ は真値を意味する記号とし、 $\delta \geq 0$ とする。第1種の過誤確率の名義水準を片側 α , 検出力を $1 - \beta$ とする。試験開始以前に仮定した群間差を δ_{pre} , ナイーブな最終解析方法を2群の平均値の差の Z 検定とすると、1群あたりの初期被験者数 $N_{initial}$ は、以下のとおり設定することとする(以下、被験者数はすべて1群あたりで示す)。

$$N_{initial} = 2 \left(\frac{z_\alpha + z_\beta}{\delta_{pre}} \right)^2, \quad (4)$$

z_u は標準正規分布の $(1 - u)$ 分位数とする。ナイーブな最終解析の検定統計量 Z_{naive} は、

$$Z_{naive} = \frac{1}{\sqrt{2\hat{N}_{final}}} \sum_{i=1}^{\hat{N}_{final}} (Y_{i1} - Y_{i2}), \quad (5)$$

$\hat{N}_{final}$ は、被験者数再設定によって最終的に決定された、両ステージを併せた総被験者数とし、初期被験者数に対し最終被験者数と呼ぶこととする。ここで、被験者数再設定の第1種の過誤確率の検討を行っていく上では、対照となるデザインとして固定デザインを考えることが有用になる。対照として考えるべき固定デザインの被験者数は、被験者数再設定の初期被験者数に一致す

るものとする、被験者数再設定において中間解析結果に依らず値を $\hat{N}_{final} = N_{initial}$ とセットすれば、ナイーブな解析は対照とする固定デザインの最終解析に一致する。対照となる固定デザインの場合、(5) 式内の $\hat{N}_{final}$ に相当する部分は $\text{set } \hat{N}_{final} = N_{initial}$ より定数項となっており、帰無仮説 ($\delta = 0$) の下で Y_{ij} が独立に同一の標準正規分布に従うことより Z_{naive} の分布も標準正規分布に従うことが導かれる。しかしながら、被験者数再設定の場合には、この最終被験者数 $\hat{N}_{final}$ は、第 1 ステージより求めた群間差 $\hat{\delta}_1$ 及び事前に計画されたアダプテーションルールに依存する確率変数であり、アダプテーションルールが完全に定式化され、かつそのルールがどんな中間解析結果が得られた場合でも完全に遵守されるという前提が成立しない限り、その確率分布は特定できない。従って、 Z_{naive} の分布はもはや標準正規分布に従わず、次式で定義される実際の(ある試験デザインの下で消費される)第 1 種の過誤確率 α_{actual} が計算できない以上、第 1 種の過誤確率が名義水準 α 以下であることは、固定デザインのように保証されない。

$$\alpha_{actual}(Z_{naive}) = E_{\hat{\delta}_1, Z_1, \hat{N}_{final}} [\Pr(Z_{naive} > z_\alpha \mid Z_1, \hat{N}_{final}, \delta = 0)], \quad (6)$$

$\hat{\delta}_1, Z_1$ はそれぞれ第 1 ステージの群間差の点推定値及び Z 検定統計量とする。 Z_{naive} を用いた最終解析の $\alpha_{actual}(Z_{naive})$ は、実際には正確に計算できないだけでなく、場合によっては名義水準を超えて上昇し得ることがわかっている。Cui, Hung and Wang (1999) は、あるアダプテーションルールの下、名義水準に対し 30% から 40% の第 1 種の過誤確率の増大が生じ得ることを示している。また、第 1 種の過誤確率の増大は、無制限に被験者数再設定を行うと最大で名義水準の 2 倍以上にまで及ぶことが示されている (Proschan and Hunsberger, 1995; Shun et al., 2001)。誤解が無いようにしなければならない点は、被験者数再設定に起因する第 1 種の過誤確率の上昇は、中間解析と最終解析の間の多重性に因るものとは別物であり、中間解析で有効中止を想定せず、中間解析による第 1 種の過誤確率の消費が無い場合でも生じるという点である。これは、極端な場合を考えてみると理解しやすい。例えば、中間解析時点で $\hat{\delta}_1 \gg \delta$ のような結果^{*1}が得られており、中間解析時点の検定統計量 Z_1 が O'Brien and Fleming の境界付近であったとすると、有効中止基準の設定の有無にかかわらず、被験者のエントリーを即中止し、中間解析以降に固定された少数のデータを追加されただけの最終解析を実施すれば^{*2}、ほぼ確実に有意な結果を得ることができるだろう。従って、直観的には、偶然ポジティブな方向へ偏った中間解析データに対して、最終解析では選択的に大きな重みが ad hoc に与えられることになり、その結果として第 1 種の過誤確率の増大が起きていると理解できる。実際、 $(\hat{\delta}_1, Z_1)$ で条件づけた下での第 1 種の過誤確率における、固定デザインとの差の挙動をシミュレーション又は数値積分によって調べれば、

- i. 中間解析結果が事前に想定していたよりもポジティブな傾向の場合に被験者数を減少させるというアダプテーションが計画されていると、深刻な第 1 種の過誤確率の上昇が生じること
- ii. 中間解析結果がネガティブな傾向の場合に被験者数を増加させるアダプテーションでは第 1 種の過誤確率が一定の上昇を示すこと
- iii. 一方、中間解析結果が被験薬の有効性を示せる見込みのあるものであるが、期待していたほどはポジティブな傾向ではない場合に被験者数を増加させるというアダプテーションの場合

^{*1} 帰無仮説の下では $\delta = 0$.

^{*2} 最終解析の境界は両側 2α 付近。

合には、第1種の過誤確率は逆に減少を示すこと

といった、一定の特徴を有することが確認できる。このような被験者数再設定に起因する第1種の過誤確率の性質は、より明確には、③の制約付き被験者数再設定の項で示す。

① Combination Test の原理

Combination Test の原理とは、各ステージの結果を事前に定義された重みで統合した統計量で最終解析を行えば、ナイーブな解析と同一の棄却限界値の下で第1種の過誤確率を名義水準以下に制御可能であるという原理である。統合の仕方によりいくつかの方法が提案されている。

1) 各ステージの p 値をその積で統合する、Fisher の p 値統合検定 (Bauer and Köhne, 1994), 2) $\hat{N}_{final} = N_{initial}$ とセットしたときのナイーブな最終解析 (対照として考える固定デザイン) におけるステージ間の重みと同じ重みを利用して各ステージの Z 検定統計量を統合する、重み付き Z 検定 (Cui, Hung and Wang, 1999), 3) 重み付き Z 検定の Z 検定以外の検定への一般化が可能な、各ステージの p 値を逆正規法を利用し統合した重み付き Z 検定 (Lehmacher and Wassmer, 1999) 等が提案されている。各ステージの重みが第1ステージの非盲検データへアクセスする以前に、対照として考えるデザインに応じて決定される 1)~3) の方法に対し、4) 各ステージの重みを各々当該ステージの非盲検データへアクセスする直前までに逐次割り付けてゆき、事前に定義しておいた重みの総和 (二乗和) を消費し切る時点が最終解析となる、分散消費アプローチによる統合検定 (Fisher, 1998; Shen and Fisher, 1999) も提案されている。このアプローチの特徴は、ステージの数、各解析時点、次ステージの被験者数、総被験者数といったデザインパラメータを逐次決定することが可能である点にあり、デザインアダプテーションの自由度の高い方法論になっている。2), 3) 及び 4) は群逐次試験、1) は 2 ステージデザインのフレームワークで提案されており、1) を再帰的に繰り返すことで 3 ステージ以上の群逐次試験へ拡張した方法 (Brannath, Posch and Bauer, 2002) も提案されている。

Combination Test の原理について、簡単のため 2) 重み付き Z 検定の 2 ステージバージョンにて説明する。まず、ナイーブな解析が各ステージにどのような重みを与えているかについて検討する。(5) 式は以下のように各ステージの Z 検定統計量の重み付和として表すことができる。

$$Z_{naive} = Z_1 \sqrt{\frac{n_1}{\hat{N}_{final}}} + Z_2 \sqrt{\frac{\hat{n}_2}{\hat{N}_{final}}}, \quad (7)$$

$n_1, \hat{n}_2, Z_2$ はそれぞれ第1ステージの被験者数、被験者数再設定によって決定された第2ステージの被験者数及び第2ステージの Z 検定統計量とする。中間解析結果に依らず値を $\hat{N}_{final} = N_{initial}$ とセット^{*3}したときの Z_{naive}, Z_2 を、それぞれ $Z_{naive}(\text{set } \hat{N}_{final} = N_{initial})$, $Z_2(\text{set } \hat{N}_{final} = N_{initial})$ と表すとする、

$$Z_{naive}(\text{set } \hat{N}_{final} = N_{initial}) = Z_1 \sqrt{\frac{n_1}{N_{initial}}} + Z_2(\text{set } \hat{N}_{final} = N_{initial}) \sqrt{\frac{n_2}{N_{initial}}}, \quad (8)$$

$n_2 = N_{initial} - n_1$ とする。(7), (8) 式を見比べると、帰無仮説 ($\delta = 0$) の下では、(8) 式において Z_1 及び $Z_2(\text{set } \hat{N}_{final} = N_{initial})$ は独立に同一の標準正規分布に従い、 $Z_{naive}(\text{set } \hat{N}_{final} = N_{initial})$ が標準正規分布に従うことから、(7) 式においても Z_1 及び Z_2 はいずれも独立に同一の標準正規分布

^{*3} 「 $\hat{N}_{final} = N_{initial}$ で条件付ける」とは異なる。

に従うという前提が成り立つのであれば, (7) 式の重み $\sqrt{\frac{n_1}{\hat{N}_{final}}}, \sqrt{\frac{\hat{n}_2}{\hat{N}_{final}}}$ の代わりに (8) 式の重み $\sqrt{\frac{n_1}{N_{initial}}}, \sqrt{\frac{n_2}{N_{initial}}}$ を利用した以下の重み付き Z 検定統計量 $Z_{weighted}$ の帰無分布は, 標準正規分布に従うことは明らかである.

$$Z_{weighted} = Z_1 \sqrt{\frac{n_1}{N_{initial}}} + Z_2 \sqrt{\frac{n_2}{N_{initial}}} = Z_1 \sqrt{t} + Z_2 \sqrt{1-t}, \quad (9)$$

t は中間解析の時期を表す情報量時間(又は分数)であり, 第 1 ステージのデータへアクセスする以前に決定されるデザインパラメータである. 従って, 重み付き Z 検定に基づく最終解析は, 各ステージの結果を事前に定義された重みで統合した統計量で最終解析を行うという Combination Test の原理に基づくものであり, ナイーブな解析と同一の棄却限界値の下で第 1 種の過誤確率を名義水準以下に制御可能である. この場合, 理論上 $Z_{weighted}$ と $Z_{naive}(\text{set } \hat{N}_{final} = N_{initial})$ の帰無分布は完全に一致することから, 重み付き Z 検定に基づく被験者数再設定デザインの下で消費される第 1 種の過誤確率 $\alpha_{actual}(Z_{weighted})$ は, 正確に名義水準 α に一致する.

Combination Test の原理に基づく方法のメリットは, 対照として考えるデザインを固定デザインではなく, 群逐次デザインに設定することで, 群逐次法に基づくあらゆる第 1 種の過誤確率の調整法(Pocock の境界, O'Brien and Fleming の境界, α 消費関数法, Wang and Tsiatis の境界, etc.)をそのまま適用可能であるという点である. 例えば, ある境界の列を c_1, c_2 , 最大情報量を N_{max} とすると, 群逐次法により下式を満たす棄却限界値 c_1, c_2 が, 各々解析に適用されるデータへアクセスする以前に定義されることが担保されている.

$$\begin{aligned} \alpha_{actual}(Z_{naive}(\text{set } \hat{N}_{final} = N_{max})) \\ = \Pr(Z_1 \geq c_1 \mid \delta = 0) + E_{\delta_1, Z_1}[\Pr\{Z_{naive}(\text{set } \hat{N}_{final} = N_{max}) \geq c_2 \mid Z_1 \hat{N}_{final}, \delta = 0\} I(Z_1 < c_1)] \\ = \alpha, \end{aligned} \quad (10)$$

$I(\cdot)$ は括弧内が真のときのみ 1 をとり, それ以外は 0 をとる指示変数とする. N_{max} をベースとする情報量時間 t に基づく $Z_{weighted}$ を (9) 式に準じて構成すれば, 被験者数再設定を伴う群逐次検定統計量の列 $(Z_1, Z_{weighted})$ の帰無分布は, 被験者数再設定を伴わない群逐次検定統計量の列 $[Z_1, Z_{naive}(\text{set } \hat{N}_{final} = N_{max})]$ の帰無分布に理論上完全に一致することから, 以下のとおり群逐次境界を直接適用することで, 被験者数再設定を伴う群逐次検定の第 1 種の過誤確率は, 正確に名義水準に一致する.

$$\begin{aligned} \alpha_{actual}(Z_{weighted}) \\ = \Pr(Z_1 \geq c_1 \mid \delta = 0) + E_{\delta_1, Z_1, \hat{N}_{final}}[\Pr(Z_{weighted} \geq c_2 \mid Z_1, \hat{N}_{final}, \delta = 0) I(Z_1 < c_1)] = \alpha. \end{aligned} \quad (11)$$

なお, K ($K > 2$) ステージデザインの場合については, Cui, Hung and Wang (1999) を参照されたい.

ただし, ここで留意しなければならない点は, 被験者数再設定を伴わない群逐次法自体を含み, Combination Test の原理並びに後述の②以降の第 1 種の過誤確率の制御の理論的妥当性は, すべて各ステージ間のデータが独立で同一の分布に従っているかどうかにかかっているということ

である。情報漏洩やアダプテーションの結果によって生じ得る、各施設内の被験者の組み入れ及び各種オペレーションの変化や変更等により、第1ステージの結果が作用して第2ステージの被験者集団が変容したりエンドポイントの評価にバイアスが生じる等、いわゆるオペレーショナルバイアスが発生すれば、第1ステージのデータと第2ステージのデータが上記の条件を満たさなくなると捉えるべきである。そうすると、群逐次法の理論的大前提である独立増分過程、並びに検討中の被験者数再設定と、対照として考える固定デザインや群逐次デザインとの間の各ステージの帰無分布の一致性が保証できなくなってしまう。結果的に、第1ステージと第2ステージのデータを比較し異なる傾向が確認された場合、第1ステージの結果が作用してそのようなデータが得られてしまったのか、それとも、仮に被験者数再設定を伴わないデザインであっても生じるような、組み入れ時期の違いに因る被験者集団の推移と考えるべきなのかどうかは、データからは識別困難であることが考えられる。このように、評価の際には厄介な問題へと発展する可能性がある点には、特段の留意が必要である。

② Conditional Error Function の原理

Conditional Error Function の原理とは、第1ステージのデータに関する期待値を取ると名義水準に一致するよう、事前に定義しておいた Conditional Error Function に基づき、第1ステージの結果に応じてナイーブな検定統計量に対する棄却限界値を調整し、条件付第1種の過誤確率を制御することにより、第1種の過誤確率を名義水準以下に制御可能であるという原理である。ここでも、2ステージデザインにてこの原理について説明する。Conditional Error Function $A(Z_1)$ とは、以下を満たし $[0, 1]$ の値を取る、第1ステージのデータ Z_1 に関する何らかの増加関数として事前に定義されるものである。

$$\int_{-\infty}^{\infty} A(Z_1) \phi(Z_1) dZ_1 = \alpha, \quad (12)$$

$\phi(Z_1)$ は Z_1 の帰無分布を表す確率密度関数とし、ここでは標準正規分布となる。最終解析の棄却限界値 c は、 $A(Z_1)$ により、以下のように第1ステージの結果に応じて調整される。

$$c = \frac{Z_1 \sqrt{n_1} + z_{A(Z_1)} \sqrt{\hat{n}_2}}{\sqrt{\hat{N}_{final}}}, \quad (13)$$

$z_{A(Z_1)}$ は、標準正規分布の $(1 - A(Z_1))$ 分位数を表す。ここで、 $Z_{naive} \geq c \Leftrightarrow Z_2 \geq z_{A(Z_1)}$ より、

$$\Pr(Z_{naive} \geq c \mid Z_1, \delta = 0) = \Pr(Z_2 \geq z_{A(Z_1)} \mid Z_1, \delta = 0) = A(Z_1). \quad (14)$$

(12), (14) 式より、(13) 式に従って最終解析でナイーブな検定統計量に対する棄却限界値を調整すれば、第1種の過誤確率が制御可能であることがわかる。Conditional Error Function は、第1ステージの結果に応じて、第2ステージの有意水準が自動的に決定される方法と解釈することができ、第1ステージの結果に応じて第2ステージの被験者数 $\hat{n}_2$ を調整しても、帰無仮説の下では偏りは生じないと言える。

Conditional Error Function は、群逐次デザインへの適用も可能である。中間解析で有効中止境界に達した場合には $A(Z_1) = 1$ 、無益性中止境界に達した場合には $A(Z_1) = 0$ と定義した上で、(12) 式を満たすように $A(Z_1)$ を定義しておけばよい。

また, Conditional Error Function の原理は, ①と同様に対照となる固定デザイン又は群逐次デザインを想定し, $Z_{naive}(\text{set } \hat{N}_{final} = N_{initial})$ 又は $Z_{naive}(\text{set } \hat{N}_{final} = N_{max})$ が消費する条件付第 1 種過誤確率を $A(Z_1)$ として定義 (Proschan and Hunsberger, 1995; Denne, 2001; Müller and Schäfer, 2001, 2004) しておけば, 対応する Combination Test に帰着する. 例えば, 固定デザインの場合作で考え,

$$A(Z) \equiv \Pr(Z_{naive}(\text{set } \hat{N}_{final} = N_{initial}) \geq z_\alpha \mid Z_1, \delta = 0) \quad (15)$$

と定義すると,

$$\begin{aligned} A(Z_1) &= \Pr(Z_1\sqrt{t} + Z_2(\text{set } \hat{N}_{final} = N_{initial})\sqrt{1-t} \geq z_\alpha \mid Z_1, \delta = 0) \\ &= \Pr\left(Z_2(\text{set } \hat{N}_{final} = N_{initial}) \geq \frac{z_\alpha - Z_1\sqrt{t}}{\sqrt{1-t}} \mid Z_1, \delta = 0\right) = 1 - \Phi\left(\frac{z_\alpha - Z_1\sqrt{t}}{\sqrt{1-t}}\right) \quad (16) \\ &\Leftrightarrow z_{A(Z_1)} = \frac{z_\alpha - Z_1\sqrt{t}}{\sqrt{1-t}} \end{aligned}$$

となる. よって, $Z_{naive} \geq c \Leftrightarrow Z_2 \geq z_{A(Z_1)}$ より,

$$Z_{naive} \geq c \Leftrightarrow Z_2 > \frac{z_\alpha - Z_1\sqrt{t}}{\sqrt{1-t}} \Leftrightarrow Z_1\sqrt{t} + Z_2\sqrt{1-t} \geq z_\alpha \Leftrightarrow Z_{weighted} \geq z_\alpha \quad (17)$$

となることから, Conditional Error Function に基づく検定と重み付き Z 検定の棄却域が一致することがわかる. このように, 同一のデザインをベースとして考えている場合には, Combination Test の原理と Conditional Error Function の原理との間には一貫性を持たせることが可能である. 一方で, 固定デザイン又は群逐次デザインをベースとしない Conditional Error Function も提案されており (Proschan and Hunsberger, 1995), メタアナリシスにおける p 値の統合検定について, Conditional Error Function の原理を利用することにより, 中間解析を伴う被験者数再設定へ応用した例が, Bauer and Köhne (1994) の方法になる. また, 効率の観点から最適な Conditional Error Function の検討もなされている (Brannath and Bauer, 2004).

③ 制約付き被験者数再設定

制約付き被験者数再設定は, 第 1 種の過誤確率を制御するために解析上の調整が必要であった, ①や②の原理に基づくアプローチとは異なり, 最終解析の方法は, 対照として考える固定デザインや群逐次デザインと同じ, ナイブな解析を維持したまま, 最終被験者数 $\hat{N}_{final}$ のために事前に計画しておくアダプテーションルールに対し一定の制約を設けることによって, 第 1 種の過誤確率の制御を可能にする方法である. Shun et al. (2001) は, いくつかのシナリオに対する第 1 種の過誤確率を計算し, これらのうちいくつかのアダプテーションルールの下では, 第 1 種の過誤確率が名義水準以下になることを確認した. Shun et al. (2001) によって第 1 種の過誤確率の特徴がある程度明らかになったと言えるが, 実際の応用上は, 個々に異なるアダプテーションルールに対し第 1 種の過誤確率を計算し直す必要があるだけでなく, 実際の最終被験者数とそのアダプテーションルールを完全に遵守可能かどうかが問題となる. この点については, ④のシミュレーションに基づく第 1 種の過誤確率の制御の項でも触れる.

Shun et al. (2001) に対し, Chen, DeMets and Lan (2004) は, 中間解析時の条件付き検出力が 50%以上であれば, 被験者数を何例増加させても第 1 種の過誤確率は名義水準以下に保たれる,

制約付き被験者数再設定を提案した。これを 50 %-CP (Conditional Power) アプローチと呼ぶこととする。50 %-CP アプローチでは、最終解析の方法は、対照として考える固定デザインや群逐次デザインと同じな解析(検定統計量及び棄却限界値)を維持したままでよく、条件付き検出力が 50 %以上であり、かつ被験者数を減少させてはならないという制約以外は、任意のアダプテーションルールに対応可能な制約付き被験者数再設定である。しかし、50 %-CP アプローチは、その制約下で被験者数を増加させる程第 1 種の過誤確率はむしろ減少する、という特徴に基づいているため、このアプローチの第 1 種の過誤確率は、名義水準に対し過度に保守的となる可能性がある。Uemura, Matsuyama and Ohashi (2008) では、50 %-CP アプローチの保守的な特徴を活かし、条件付き検出力が 50 %以上という制約を、50 %よりも低い閾値へと制約を緩めるための拡張がなされている。Mehta and Pocock (2011) は、この制約によって限定された、アダプテーションが可能となるような中間解析結果の範囲のことを “promising zone” と呼んでおり、Gao, Ware and Mehta (2008) に基づき、50 %-CP アプローチの “promising zone” を拡張する提案を行っている。

このように、制約付き被験者数再設定では、統計学的な調整が必要となるような領域ではアダプテーションを控え、“promising zone” に入るような中間解析結果が得られた場合へアダプテーションの機会を限定してしまて良いという考え方に立っている。このことは、初期被験者数ある程度保守的に設定しておけば大幅な検出力不足を想定しなくても良いという見通しや、臨床的に意味の無い差まで検出してしまうという over-power の観点からは、一見都合良く思われる。ただし、“promising zone” は、観察された群間差を真と仮定した際の条件付き検出力等の指標に応じ、あくまで第 1 種の過誤確率の制御の観点から特異的に閾値が決められた領域に過ぎない。実務上の観点から “promising zone” を「有望な領域」と解釈すべきかどうかは別問題であり、個々のデザインや状況に依存している点に留意すべきである。また、制約付き被験者数再設定では、“promising zone” に入る程結果は思わしくなくアダプテーションの実施は許容されないものの、無益性中止基準には該当しない場合に、検出力が低いと予測された初期被験者数のまま試験継続することを許容せざるを得ない」といった、判断のグレーゾーンのような領域を有する場合が考えられる。試験実施上の観点から、このような領域がどのように位置づけられるべきかが明確ではなく、留意が必要である。

④ シミュレーションに基づく第 1 種の過誤確率の制御

実施が計画された複数のアダプテーションの 1 つとして被験者数再設定を含むような、より複雑なアダプティブデザインにおいては、①～③の第 1 種の過誤確率の制御方法を直接適用することが困難な場合が考えられる。そのような場合にはシミュレーションに基づく制御方法の適用が想定されるが、それには第 1 種の過誤確率の解析的な評価が困難になるような複雑なアダプテーションは、基本的には避けるべきであるという点以外に、以下のような問題点が考えられる。

- i. シミュレーションのみに基づき第 1 種の過誤確率が制御されていることを示すことについての広いコンセンサスは、未だ得られていないこと
- ii. 実際に消費される第 1 種の過誤確率は、アダプテーションルールが完全に遵守されるかどうかには依存すること

- iii. シミュレートされた第1種の過誤確率が、複数のアダプテーションの過程(例えば、用量選択ルールとある用量の選択確率等)に対するモデリングの仕方に依存してしまうこと
- iv. 複数の仮説を伴うアダプティブデザインの場合、複数の仮説に対し強く FWER (Family Wise Error Rate) を制御するためには、仮説の配置毎に FWER をシミュレートすることになるが、帰無仮説以外の対立仮説の値(例えば、他の用量群におけるプラセボ群との差に関する期待値等)にアダプテーションの確率が依存するため、正確な FWER は計算できないことが考えられること
- v. あるアダプテーションによる第1種の過誤確率の上昇と、あるアダプテーションによる第1種の過誤確率の減少とが相殺されるようなシミュレーションの場合には、いずれか一方における想定の間違いが発生してしまった際に第1種の過誤確率が上昇するというリスクがあり、頑健性の問題を有すると考えられること

これらの問題点を踏まえると、より複雑なアダプティブデザインにおいては、より保守的で単純な第1種の過誤確率の調整法の適用や、アダプテーション過程の不確実性に関しては少なくとも第1種の過誤確率の最大値をシミュレートしておくこと等、より頑健性の高いと思われる対処方法を検討することが少なくとも必要になってくると考えられる。

3.2 群間差の推定問題

被験者数再設定においてナイーブな解析を行うと、第1種の過誤確率の上昇と同様の理由により、点推定値にバイアスが入ったり、信頼区間の被覆確率が名義水準より小さくなったりする。ここでは、被験者数再設定における点推定値及び信頼区間の推定方法について概説するが、群逐次デザインにおける被験者数再設定では、中間解析における早期中止基準がもたらすバイアスも併せて考慮しなければならないため、早期中止基準の有無別に考える。なお、ここで扱うバイアスはすべて統計学的なバイアスであり、3.1.①の終わりで述べたように、検定同様、群間差の推定問題において厄介となる決定的なバイアスは、オペレーショナルバイアスに因るものであり、統計学的な対処法が存在しないことに留意すべきである。

① 中間解析における早期中止を伴わない場合

ナイーブな最終解析の点推定値 $\hat{\delta}_{naive}$ は、

$$\hat{\delta}_{naive} = \frac{1}{\hat{N}_{final}} \sum_{i=1}^{\hat{N}_{final}} (Y_{i1} - Y_{i2}) = \frac{n_1}{\hat{N}_{final}} \hat{\delta}_1 + \frac{\hat{n}_2}{\hat{N}_{final}} \frac{\sum_{i=n_1+1}^{\hat{N}_{final}} (Y_{i1} - Y_{i2})}{\hat{n}_2} \quad (18)$$

となる。中間解析の結果 $\hat{\delta}_1$ 及びアダプテーションの結果 $\hat{n}_2$, $\hat{N}_{final}$ で条件付けた下では、

$$E[\hat{\delta}_{naive} | \hat{\delta}_1, \hat{n}_2, \hat{N}_{final}] = \frac{n_1}{\hat{N}_{final}} \hat{\delta}_1 + \frac{\hat{n}_2}{\hat{N}_{final}} \delta \quad (19)$$

となり、ナイーブな最終解析の点推定値 $\hat{\delta}_{naive}$ の期待値は、 $\hat{\delta}_1$ と独立ではない $\hat{n}_2$, $\hat{N}_{final}$ の分布に依存することがわかる。仮に $\hat{\delta}_1$ と $\hat{n}_2$, $\hat{N}_{final}$ が独立であるとする、 $\hat{\delta}_{naive}$ のバイアスは、

$$E[\hat{\delta}_{naive} - \delta] = E\left[\left(\frac{n_1}{\hat{N}_{final}}\right)(\hat{\delta}_1 - \delta)\right] = E\left[\frac{n_1}{\hat{N}_{final}}\right] E[\hat{\delta}_1 - \delta] + Cov\left[\frac{n_1}{\hat{N}_{final}}, \hat{\delta}_1\right] = 0 \quad (20)$$

となる (Brannath, König and Bauer, 2006). よって, アダプテーションを伴わない固定デザインのようになり, $\hat{\delta}_1$ と $\hat{n}_2$, $\hat{N}_{final}$ が独立でない限り, ナイーブな点推定値にはバイアスが存在することがわかる. このバイアスの詳細については, Brannath, König and Bauer (2006) の Appendix A.1 を参照されたい.

被験者数再設定に伴う点推定値のバイアス補正には, 平均値不偏推定量と中央値不偏推定量の 2通りの考え方がある. 前者については, 3.1. ①の Combination Test の原理と同様の考え方が適用できる. すなわち, 対照として考える固定デザインの推定量 $\hat{\delta}_{naive}$ (set $\hat{N}_{final} = N_{initial}$) における重みを利用し, 以下の重み付き推定量 $\hat{\delta}_{weighted}$ を考える.

$$\hat{\delta}_{weighted} = \frac{n_1}{N_{initial}} \hat{\delta}_1 + \frac{n_2}{N_{initial}} \frac{\sum_{i=n_1+1}^{\hat{N}_{final}} (Y_{i1} - Y_{i2})}{\hat{n}_2}. \quad (21)$$

$$E[\hat{\delta}_{weighted} | \hat{\delta}_1, \hat{n}_2, \hat{N}_{final}] = \frac{n_1}{N_{initial}} \hat{\delta}_1 + \frac{n_2}{N_{initial}} \delta \text{ より,}$$

$$E[\hat{\delta}_{weighted}] - \delta = \frac{n_1 + n_2}{N_{initial}} \delta - \delta = 0. \quad (22)$$

(21) 式における重みは, 事前に定義された他の重みでも良い (Liu, Proschan and Pledger, 2002; Proschan, Liu and Hunsberger, 2003). なお, $\hat{\delta}_{weighted}$ の分散推定量は, アダプテーションの結果 $\hat{n}_2$, $\hat{N}_{final}$ の分布に依存してしまうため, 正確な推定量は求められない点に留意が必要である. 詳細については, Brannath, König and Bauer (2006) の Appendix A.1 を参照されたい.

後者の中央値不偏推定量は, Brannath, Posch and Bauer (2002), Lawrence and Hung (2003), Proschan, Liu and Hunsberger (2003), Cheng and Shen (2004) で提案されている. (9) 式の重み付き Z 検定統計量はある仮説 $\delta = \delta_H$ に対しても同様に妥当な検定であり, 以下のように表すことができる.

$$Z_{weighted} = (\hat{\delta}_1 - \delta_H) \sqrt{\frac{n_1 t}{2}} + (\hat{\delta}_2 - \delta_H) \sqrt{\frac{\hat{n}_2 (1-t)}{2}}. \quad (23)$$

中央値不偏推定量 $\hat{\delta}_{median}$ は, $Z_{weighted} = 0$ となる仮説 δ_H の値を解くことによって, 以下のようになれる.

$$\hat{\delta}_{median} = \tilde{w} \hat{\delta}_1 + (1 - \tilde{w}) \hat{\delta}_2, \quad \tilde{w} = \frac{\sqrt{tn_1}}{\sqrt{tn_1} + \sqrt{(1-t)\hat{n}_2}}. \quad (24)$$

Cheng and Shen (2004) はモーメント法により $\hat{\delta}_{median}$ を導いている.

信頼区間については, 中央値不偏推定量と同様, 検定との対応で考えれば良い (Lehmacher and Wassmer, 1999; Brannath et al., 2003). (23) 式において, $Z_{weighted} \leq z_\alpha$ 及び $Z_{weighted} \geq -z_\alpha$ を解くと, 重み付き Z 検定をベースとした両側信頼区間 $CI_{weighted}$ は,

$$CI_{weighted} = \left[\hat{\delta}_{median} - \frac{z_\alpha}{\sqrt{2n_1 t} + \sqrt{2\hat{n}_2 (1-t)}}, \hat{\delta}_{median} + \frac{z_\alpha}{\sqrt{2n_1 t} + \sqrt{2\hat{n}_2 (1-t)}} \right] \quad (25)$$

となる. 重み付き Z 検定の第 1 種の過誤確率が正確に名義水準片側 α に一致するように, $CI_{weighted}$ の被覆確率も正確に名義水準 $1 - 2\alpha$ に一致する.

② 中間解析における早期中止を伴う場合

中間解析における早期中止がもたらすバイアスは、被験者数再設定によるバイアスとは別物である。従って、群逐次デザインにおける被験者数再設定の推定問題では、①で説明した重み付き推定量による不偏推定又は重み付き Z 検定をベースとした中央値不偏推定や信頼区間によって被験者数再設定によるバイアスへ対処した上で、群逐次デザインで提案されてきたバイアス修正法をいかに適用するかが基本的な考え方となる。

群逐次デザインでは、試験終了時点のステージまでのデータから (21) 式のような重み付き推定量や (24) 式のような中央値に基づく推定量を求めることが考えられるが、これらはいずれも平均値不偏推定量及び中央値不偏推定量にはならない。平均値不偏推定については、そのステージで有効中止となったということ又はそのステージまでは有効中止とならなかったということによるバイアスが存在し、中央値不偏推定については、試験終了時のデータの順位を考える際に、仮に以前のステージで中止していた場合の試験終了時のデータの順位を無視しているため、真の中央値との間のバイアスが存在する。前者のバイアスについては、実際にはあまり汎用されていないものの、中間解析結果によらず第 1 ステージの群間差の推定量を拡張した平均値不偏推定量 (Emerson and Kittelson, 1997) や試験終了時点のステージまでのデータに基づく推定量のバイアス補正項の提案 (Whitehead, 1986) が、被験者数再設定を伴わない群逐次デザインのフレームワークでなされており、被験者数再設定を伴う場合への拡張は可能と考えられる (Brannath, König and Bauer, 2006)。後者のバイアスについては、ステージ間の結果の順位付け (ordering) を予め定義しておくことにより、中央値不偏推定量、 p 値の補正方法、信頼区間の構成方法が、群逐次デザインのフレームワークにおいて提案されており (Fairbanks and Madsen, 1982; Tsiatis, Rosner GL and Mehta, 1984; Rosner and Tsiatis, 1988; Chang, 1989; Emerson and Fleming, 1990)、被験者数再設定を伴う場合の中央値不偏推定量等への拡張は、Brannath, Posch and Bauer (2002) において提案されている。さらに、Lehmacher and Wassmer (1999) は、Jennison and Turnbull (1984, 1989) の繰り返し信頼区間について、被験者数再設定を伴う場合への拡張を行っており、Mehta, et al. (2007) は、保守的な点推定値や繰り返し p 値についての同様の拡張を行っている。他の信頼区間の構成方法については Brannath, König and Bauer (2006) を参照されたい。

4. 考察とまとめ

本稿では、アダプティブデザインの理論の基礎となっており、アダプテーションの内容が限定的で比較的議論が行い易い被験者数再設定の方法論について、実務担当者間でのより良い議論を念頭に、方法論の解説を行うと共に、実務上の問題を意識し、統計学的な留意点についての検討を行った。被験者数が不足するかもしれないという懸念のみに基づき、安易に被験者数再設定を計画することは、規制上の観点及び統計学的な観点のいずれにおいても推奨されない。一方、開発の置かれている状況や疾患領域によっては、不十分な被験者数の設定根拠に基づき試験を開始せざるを得ないこと、同じプロトコルでも患者集団とそれに伴う有効性の評価可能性 (プラセボや対照治療における疾患の進行の速度やイベントの発生状況, etc.) が試験によってばらつく領域もあること、臨床的意義のある有効性の大きさが定義し難いこと又は試験の失敗や追試験が許

容できないこと等、設定した被験者数の適切性を確認しておく機会があった方が望ましい場合も考えられる。2.2 節では、盲検下と非盲検下の被験者数再設定の考え方の比較・検討を行ったが、被験者数再設定を検討する際には被験者数の設定自体についてより深く検討しておくことが前提として必要になり (Hung, Wang and O'Neill, 2011), 事前の十分な検討を実務担当者がリードして行くことがより重要になる (Gallo, and Maurer, 2006; Kock, 2006; Wittes, and Lachenbruch, 2006)。被験者数再設定を含むアダプティブデザインについて、今後実務担当者間での議論をより深めていくためには、実用上の具体的な議論を丁寧に進めていくことが不可欠となるだろう。本節では、被験者数再設定の実用上の課題を少し考えてみたい。

被験者数再設定の実用上の課題には、オペレーション上の問題や試験実施上の意思決定問題等に係わる個別の状況も複雑に関係してくるため、一般的な議論を単純に行うことは困難であるが、検討すべき課題をいくつか挙げることは可能である。1 つ目は、中間解析結果に基づく被験者数再設定が必ずしも正確な被験者数を導いてくれるとは限らないという点についての課題である。例えば、中間解析結果の良し悪しを考慮した予測検出力の指標として、中間解析結果を与えた下での条件付き検出力に基づき、必要被験者数を算出することが提案されているが (Proschan and Hunsberger, 1995), 条件付き検出力は中間解析で認められた群間差が真である ($\delta = \delta_1$) という仮説の下での予測値に過ぎないため、 $\hat{\delta}_1$ のばらつきの影響を強く受けることがわかっている (Bauer and Köenig, 2006)。従って、中間解析結果には一定の不確実性が存在するという点を前提に、被験者数再設定の機会が被験者数を決定する上でもたらすベネフィットの有無について、よく議論しておかなければならない。また、一定の不確実性があるという前提の下、中間解析結果を用いてより頑健な意思決定を行うためには、どのような予測指標を用いてどのようなアダプテーションルールを計画しておく方が良いか、といったことを検討することも重要になるだろう。

2 つ目は、群逐次デザインの利用によって被験者数設定の不確実性へ対処する方が、統計的な効率性がより良いという指摘 (Tsiatis and Mehta, 2003; Jennison and Turnbull, 2003) もあることを踏まえ、効率性の観点からの検討も行っておくべきであるという点である。両者のデザインの優劣については、比較する条件や想定するシナリオ、何を指標として最適化するかといった点にも依存するため (Mehta and Patel, 2006; Eimermacher and Hommel, 2007; Bartroff and Lai, 2008; Liu, Zhu and Cui, 2008), 一概に結論付けることは困難であるものの、群逐次デザインや 2 節で議論した盲検下の被験者数再設定で対応可能かどうかは、計画時に必ず考慮されるべき事項である。オペレーション上は、中間解析結果を確認した上で被験者数を実施限度内で増加させるよりも、実施限度が明確になっているのであれば、最初から最大被験者数を設定しておいて、想定どおりかそれ以上の有効性が存在するのであれば、中間解析で有効中止となって被験者数が予定より少なくて済むことを期待するデザインのほうが、より実施し易いということも十分に考えられる。そういう意味においても、被験者数再設定の機会を設けることの意義は十分に吟味しておくべきと言える (Hung, Wang and O'Neill, 2011; Mehta and Pocock, 2011a, b; Emerson, Levin and Emerson, 2011)。

3 つ目は、Posch, Bauer and Brannath (2003) 及び Burman and Sonesson (2006) が指摘しているように、3.1.①や②の原理に基づく検定では有意差が得られるが、ナイーブな解析では逆に有

意差が得られない, いわゆる “inconsistency regions” の存在が, 試験結果の解釈を困難にする可能性があるという点である. ①の原理については, 以下で示すように $Z_{weighted}$ と Z_{naive} との間で比較すると, 第1ステージと第2ステージとの間で, 1被験者あたりに与えられる重みに不均衡が発生することがわかっている. ②の原理についても実質的には同じことで, 最終解析の棄却限界値の越え易さに対する1被験者あたりの寄与の程度がステージ間で異なってくる. このようなステージ間の重みの不均衡が “inconsistency regions” が発生する要因ではあるが, いつどの程度生じ易いかは, 個々の試験のデザインやアダプテーションルールに依存するため, 事前に試験シミュレーションによって確認しておくことが推奨されるだろう. また, このような “inconsistency regions” の問題は結果次第でもあるため, ①や②の検定とナイーブな解析との両方で有意差が示されたときのみ有意差ありと判断する, Dual Test (Chen, DeMets and Lan, 2004; Burman and Sonesson, 2006) というアプローチも提案されており, Hung, Wang and O'Neill (2011) でも言及がある.

$$Z_{weighted} = \sqrt{\frac{1}{2N_{initial}}} \sum_{i=1}^{n_1} (Y_{i1} - Y_{i2}) + \sqrt{\frac{1}{2N_{initial}}} \sqrt{\frac{n_2}{\hat{n}_2}} \sum_{i=n_1+1}^{\hat{N}_{final}} (Y_{i1} - Y_{i2}). \quad (26)$$

$$Z_{naive} = \sqrt{\frac{1}{2\hat{N}_{final}}} \sum_{i=1}^{n_1} (Y_{i1} - Y_{i2}) + \sqrt{\frac{1}{2\hat{N}_{final}}} \sum_{i=n_1+1}^{\hat{N}_{final}} (Y_{i1} - Y_{i2}). \quad (27)$$

さらに, (26) 式と (27) 式を見比べると, 状況によっては第1種の過誤確率の制御のためのペナルティーが大きくなることが想定される. 例えば, 初期被験者数 $N_{initial}$ をかなり少なめに設定しておき, 試験の終盤の中間解析結果に基づき被験者数を大幅に増加させた場合, (19) 式における重みと比較し, 第1ステージにかかる重みがより大きく, 第2ステージにかかる重みはより極端に小さくなることになり, 被験者数を増加させた割には検定統計量の値は増加し難くなることがわかる. つまり, 統計学的効率の損失が生じることになる. 被験者数を減少させる場合にも, アダプテーションの仕方によっては同様の非効率性が生じ得る. このような, ステージ間の重みがステージ間の被験者数に比例していないことにより生じ得る, 統計学的な効率の損失の程度についても, 個々の試験のデザインやアダプテーションルールに依存するため, 事前に試験シミュレーションによって確認しておくことが推奨される.

最後に, 本節では, 被験者数再設定の実用上の課題に係わる留意点を挙げたが, 実際にどういった点がどのような状況で問題になるので特に重要視すべきか, あるいはもっと別の観点をより重要視すべきかについては, well-understood ではない. また, 2節や3節で触れた論点についても, 本邦のプラクティスにおいては, 現時点で見られる限りにおいては十分に well-understood な状況とは言えない. 被験者数再設定に限らず, 他のアダプティブデザインや複数のアダプテーションを組み合わせた複雑なデザインについても, まずは実務担当者間の理解とコミュニケーションを深めていき, 実績を積むことが本邦のプラクティスの発展にとって必要なステップであると考ええる.

謝 辞

本稿について、貴重なご意見を頂きました査読者の先生方へ深く御礼申し上げます。また、本稿の内容は、2011年11月に開催された計量生物セミナーでの著者の発表内容に基づいており、このような発表の機会だけでなく、貴誌への投稿機会も併せて頂いたこと、日本計量生物学会の上記セミナー開催に係わった先生方と編集委員会の先生方へ深く御礼申し上げます。

参考文献

- 安藤友紀. (2012). 審査の立場からみたアダプティブ試験. 腫瘍内科 **9**, 128–131.
- Bartroff, J. and Lai, T. L. (2008). Efficient adaptive designs with mid-course sample size adjustment in clinical trials. *Statistics in Medicine* **27**, 1593–1611.
- Bauer, P. and Köhne, K. (1994). Evaluations of experiments with adaptive interim analyses. *Biometrics* **50**, 1029–1041.
- Bauer, P. and König, F. (2006). The reassessment of trial perspectives from interim data—a critical view. *Statistics in Medicine* **25**, 23–36.
- Birkett, M. A. and Day, S. J. (1994). Internal pilot studies for estimating sample size. *Statistics in Medicine* **13**, 2455–2463.
- Brannath, W., Posch, M. and Bauer, P. (2002). Recursive combination tests. *Journal of the American Statistical Association* **97**, 236–244.
- Brannath, W., Maurer, W., Posch, M. and Bauer, P. (2003). Sequential tests for non-inferiority and superiority. *Biometrics* **59**, 106–114.
- Brannath, W. and Bauer, P. (2004). Optimal conditional error functions for the control of conditional power. *Biometrics* **60**, 715–723.
- Brannath, W., König, F. and Bauer, P. (2006). Estimation in flexible two stage designs. *Statistics in Medicine* **25**, 3366–3381.
- Brannath, W., König, F. and Bauer, P. (2007). Estimation in flexible two stage designs. *Statistics in Medicine* **25**, 3366–3381.
- Burman, C. F. and Sonesson, C. (2006). A flexible design is sound? *Biometrics* **62**, 664–683.
- Chang, M. N. (1989). Confidence intervals for a normal mean following a group sequential test. *Biometrics* **45**, 247–254.
- Chen, Y. H. J., DeMets, D. L. and Lan, K. K. G. (2004). Increasing sample size when the unblinded interim result is promising. *Statistics in Medicine* **23**, 1023–1038.
- Cheng, Y. and Shen, Y. (2004). Estimation of a parameter and its exact confidence interval following sequential sample size reestimation trials. *Biometrics* **60**, 910–918.

- CHMP. (2007). Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. <http://home.att.ne.jp/red/akihiro/emea/245902enadopted.pdf>
- Cui, L., Hung, H. M. J. and Wang, S. J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics* **55**, 853–857.
- Denne, J. S. (2001). Sample size recalculation using conditional power. *Statistics in Medicine* **20**, 2645–2660.
- Eimermacher, A. J. and Hommel, G. (2007). Performance of adaptive sample size adjustment with respect to stopping criteria and time of interim analysis. *Statistics in Medicine* **26**, 1450–1461.
- Emerson, S. S. and Fleming, T. R. (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika* **77**, 875–892.
- Emerson, S. S., Levin, G. P. and Emerson, S. C. (2011). Comments on ‘Adaptive increase in sample size when interim results are promising: A practical guide with examples.’ *Statistics in Medicine* **30**, 3285–3301.
- Fairbanks, K. and Madsen, R. (1982). *P* values for tests using the repeated significance test design. *Biometrika* **69**, 69–74.
- FDA. (2004). Innovation or stagnation? Challenge and opportunity on the critical path to new medical products.
<http://www.fda.gov/downloads/ScienceResearch/SpecialTopics/CriticalPathInitiative/CriticalPathOpportunitiesReports/ucm113411.pdf>
- FDA. (2010). Adaptive design clinical trials for drugs and biologics. *Guidance for Industry*, Draft Guidance.
<http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm201790.pdf>
- Fisher, L. (1998). Self-designing clinical trials. *Statistics in Medicine* **17**, 1551–1562.
- Friede, T. and Kieser, T. (2002). On the inappropriateness of an EM-algorithm based procedure blinded variance estimation. *Statistics in Medicine* **21**, 165–176.
- Gallo, P. and Maurer, W. (2006). Challenges in implementing adaptive designs: comments on the viewpoints expressed by regulatory statisticians. *Biometrical Journal* **48**, 591–597.
- Gallo, P., Stein, C. C., Dragalin, V., Gaydos, B., Krams, M. and Pinheiro J. (2006). Executive summary of the PhRMA working group on adaptive designs in clinical drug development. *Journal of Biopharmaceutical Statistics* **16**, 275–283.
- Gao, P., Ware, J. H. and Mehta, C. (2008). Sample size re-estimation for adaptive sequential design in clinical trials. *Journal of Biopharmaceutical Statistics* **18**, 1184–1196.

- Gould, A. L. (1992). Interim analyses for monitoring clinical trials that do not materially affect the type I error rate. *Statistics in Medicine* **11**, 55–66.
- Gould, A. L. and Shih, W. J. (1992). Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Communications in Statistics (A)* **21**, 2833–2853.
- Gould, A. L. (1995). Planning and revising the sample size of a trial. *Statistics in Medicine* **14**, 1039–1051.
- Gould, A. L. (2001). Sample size re-estimation: recent developments and practical considerations. *Statistics in Medicine* **20**, 2625–2643.
- Gould, A. L. and Shih, W. J. (2005). Letter to the editor. *Statistics in Medicine* **24**, 147–156.
- Govindarajulu, S. (2003). Robustness of sample size re-estimation procedure in clinical trials (arbitrary populations). *Statistics in Medicine* **22**, 1819–1828.
- Hung, H. M. J., O'Neill, R. T., Wang, S. J. and Lawrence, J. (2006). A regulatory view on adaptive/flexible clinical trial design. *Biometrical Journal* **48**, 565–573.
- Hung, H. M. J., Wang, S. J. and O'Neill, R. T. (2011). Flexible design clinical trial methodology in regulatory applications. *Statistics in Medicine* **30**, 1519–1527.
- Jennison, C. and Turnbull, B.W. (1984). Repeated confidence intervals for group sequential clinical trials. *Controlled Clinical Trials* **5**, 33–45.
- Jennison, C. and Turnbull, B.W. (1989). Interim analyses: the repeated confidence interval approach (with discussion). *Journal of the Royal Statistical Society, Series B* **51**, 305–361.
- Jennison, C. and Turnbull, B. W. (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine* **22**, 971–993.
- Kieser, M. and Friede, T. (2000). Re-calculating the sample size in internal pilot study designs with control of the type I error rate. *Statistics in Medicine* **19**, 901–911.
- Kieser, M. and Friede, T. (2003). Simple procedures for blinded sample size adjustment that do not affect the type I error rate. *Statistics in Medicine* **22**, 3571–3581.
- Koch, A. (2006). Confirmatory clinical trials with an adaptive design. *Biometrical Journal* **48**, 574–585.
- Lawrence, J. and Hung, H. M. J. (2003). Estimation and confidence intervals after adjusting the maximum information. *Biometrical Journal* **45**, 143–152.
- Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculation in group sequential trials. *Biometrics* **55**, 1286–1290.
- Liu, Q., Proschan, A. A. and Pledger, G. W. (2002). A unified theory of two-stage adaptive designs. *Journal of the American Association* **97**, 1034–1041.
- Jpn J Biomet Vol.33, No.1, 2012

- Liu, G. F., Zhu, G. R. and Cui, L. (2008). Evaluating the adaptive performance of flexible sample size designs with treatment difference in an interval. *Statistics in Medicine* **27**, 584–596.
- Mehta, C. R. and Patel, N. R. (2006). Adaptive, group sequential and decision theoretic approaches to sample size determination. *Statistics in Medicine* **25**, 3250–3269.
- Mehta, C. R., Bauer, P., Posch, M. and Brannath, W. (2007). Repeated confidence intervals for adaptive group sequential trials. *Statistics in Medicine* **26**, 5422–5433.
- Mehta, C. R. and Pocock, S. J. (2011a). Adaptive increase in sample size when interim results are promising: A practical guide with examples. *Statistics in Medicine* **30**, 3257–3284.
- Mehta, C. R. and Pocock, S. J. (2011b). Authors’ response to “Comment on adaptive increase in sample size when interim results are promising.” *Statistics in Medicine* **30**, 3302–3303.
- Müller, H. H. and Schäfer, H. (2001). Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* **57**, 886–891.
- Müller, H. H. and Schäfer, H. (2004). A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine* **23**, 2497–2508.
- O’Neill, R. T. (2006). FDA’s critical path initiative: a perspective on contributions of biostatistics. *Biometrical Journal* **48**, 559–564.
- Posch, M., Bauer, P. and Brannath, W. (2003). Issues in designing flexible trials. *Statistics in Medicine* **22**, 953–969.
- Proschan, M. and Hunsberger, S. (1995). Designed extension of studies based on conditional power. *Biometrics* **51**, 1315–1324.
- Proschan, M. A. and Wittes, J. (2000). An improved double sampling procedure based on the variance. *Biometrics* **56**, 1183–1187.
- Proschan, M. A., Liu, Q. L. and Hunsberger, S. (2003). Practical midcourse sample size modification in clinical trials. *Controlled Clinical Trials* **24**, 4–15.
- Proschan, M. A. (2005). Two-stage sample size re-estimation based on a nuisance parameter: A review. *Journal of Biopharmaceutical Statistics* **15**, 559–574.
- Rosner, G. L. and Tsiatis, A. A. (1988). Exact confidence limits following group sequential tests. *Biometrika* **75**, 723–729.
- Schäfer, H. (2006). Adaptive designs from the viewpoint of an academic biostatistician. *Biometrical Journal* **48**, 586–590.
- Emerson, S. S. and Kittelson, J. M. (1997). A Computationally Simpler Algorithm for the UMVUE of a Normal Mean Following a Group Sequential Trial. *Biometrics* **53**, 365–369.

- Shen, Y. and Fisher, L. (1999). Statistical inference for self-designing clinical trials with a one-sided hypothesis. *Biometrics* **55**, 190–197.
- Shih, W. J. (2001). Sample size re-estimation — journey for a decade. *Statistics in Medicine* **20**, 515–518.
- Shun, Z., Yuan, W., Brady, W. E. and Hsu, H. (2001). Type I error in sample size re-estimations based on observed treatment difference. *Statistics in Medicine* **20**, 497–513.
- Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics* **16**, 243–258.
- Stein, C. C., Anderson, K., Gallo, P. and Collins, S. (2006). Sample size reestimation: A review and recommendations. *Drug Information Journal* **40**, 475–484.
- Tsiatis, A. A., Rosner, G. L. and Mehta, C. R. (1984). Exact confidence intervals following a group sequential test. *Biometrics* **40**, 797–803.
- Tsiatis, A. A. and Mehta, C. (2003). On the efficiency of the adaptive design for monitoring clinical trials. *Biometrika* **90**, 367–378.
- Uemura K., Matsuyama, Y. and Ohashi, Y. (2008). A modification of the 50 %-conditional power approach for increasing the sample size based on an interim estimate of treatment difference. *Japanese Journal of Biometrics* **29**, 19–34.
- Whitehead, J. (1986). On the bias of maximum likelihood estimation following a sequential test. *Biometrika* **73**, 573–581.
- Wittes, J. and Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine* **9**, 65–72.
- Wittes, J. and Lachenbruch, P. A. (2006). Opening the adaptive toolbox. *Biometrical Journal* **48**, 598–603.
- Wittes, J., Schabenberger, O., Zucker, D., Britain, E. and Proschan, M. (1999). Internal pilot studies I: type I error rate of the naive t -test. *Statistics in Medicine* **18**, 3481–3491.
- Zucker, D., Wittes, J., Schabenberger, O. and Brittain E. (1999). Internal pilot studies II: comparison of various procedures. *Statistics in Medicine* **18**, 3493–3509.