保健医療科学　2011　Vol.60　No.1　p.27－32

特集：臨床試験・治験の最近の動向

＜総説＞

臨床試験の適応的デザイン

丹後俊郎

医学統計学研究センター

Adaptive design in clinical trials

Toshiro TANGO

Center for Medical Statistics

抄録

事前に effect size を見積もり，有意水準α，検出力 100（1−β）%を決めて最小限必要な症例数を計算し，その症例数を達成するまで試験を継続するというのが通常の臨床試験のデザインの基本である．しかし，そのデザインでは，治療効果があるものは早く市場に出し，効果のないものは早く中止すべき，という社会的要請に応えられない．また，臨床試験に必要不可欠な目標症例数を決定する際の因子である「臨床的に意味のある効果の大きさ」を正確に見積もることは必ずしも簡単ではない．特に，長い時間を要する試験では効果の大きさの誤った見積もりの影響は深刻である．したがって，試験途中で試験を終了できる，あるいは，試験デザインを変更できる適応的デザインが最近世界的な注目を浴びている．本稿ではその特徴を代表的なデザインで解説するとともに，その適用上の問題点を述べる．

キーワード：条件付 Type I エラー関数，グループ逐次デザイン，中間解析，サンプルサイズ再設定，無作為化比較試験

Abstract

In classical trial designs, it is necessary to determine the minimum number of eligible patients (sample size) required for detecting the pre-estimated effect size at a reasonable power and to continue the trial until the required number of patients are recruited. However, we often have a certain degree of uncertainty about the clinically relevant effect size and other important parameters when calculating the sample size at the design stage. Especially, for long-term trials (continuing for many years), an erroneously calculated sample size may have a devastating consequence. Therefore, adaptive designs are getting widespread international attention because they allow early termination because of inefficacy and/or futility, for the re-estimation of the sample size, and for other modifications of the trial design in the middle of a trial. In this paper, several important adaptive designs are briefly introduced in historical order. In addition, we discuss points to consider in their practical applications.

Keywords: conditional type I error function, group sequential design, interim analysis, sample size re-estimation, randomized controlled trial.

連絡先：丹後俊郎
〒 105-0021 東京都港区東新橋 2-9-6
汐留イタリア街 SAN ビル 4F
Shiodome Italia st. 4F 2-9-6 Higashi Shinbashi, Minato-ku, Tokyo, 105-0021, Japan
Email: tango@niph.go.jp
[平成 23 年 1 月 17 日受理]

Ⅰ．はじめに

新薬の有効性と安全性について，当局から承認を受けるための申請には，企業が行うヒトを対象とした臨床試験（治験）が必須である．それには，開発水準に応じたフェーズ（相）があり，健康なボランティアを基本的な対象として安全性の検討を行う第１相試験，用量反応性と至適用法用量を検索する第２相試験，最終的に選ばれた用法・用量でプラセボあるいは標準薬との比較を行う検証的な第３相試験があり，それぞれ，独立にデザイン・実施・解析が行われ，次の相に推移していく，あるいは，最終的な当局への申請にいく形態をとっている．特に，第２相，第３相の臨床試験では，患者の治療への無作為割り付け（random allocation）に基づく比較試験（RCT, randomized controlled trial）が必須である．臨床試験のデザインは相によって異なるが，第Ⅲ相の典型的な型は次のようなものである．

1. 当該試験に最も相応しい（複合）エンドポイントをひとつ決定する．
2. エンドポイントに期待される「臨床的に有効であると考えられる最小の効果の大きさ δ（effect size）」を慎重に見積もる．
3. エンドポイントの評価に適切な検定法を選択，有意水準 α，検出力 $100(1-\beta)\%$ を設定し，検定で「有意」となる最小の症例数（目標症例数）を計算する．
4. その症例数が試験に組み入れられ，必要なデータが観察されるまで試験を継続する．
5. 目標症例数に達した時点で，キーオープンして症例検討会を開催し，症例を固定した後，解析を実施する．

しかし，この伝統的なデザインでは，治療効果が事前の期待以上にあるものは早く市場に出し，効果のないものは早く試験を中止すべき，という社会的要請に応えられない．また，RCT に必要不可欠な目標症例数を決定する際の因子である「effect size」δ を過去の（類似薬も含めた）試験から正確に見積もることは必ずしも簡単ではない．特に，長い時間を要する RCT では effect size の誤った見積もりの影響は深刻である．したがって，試験途中で試験を良い意味でも，悪い意味でも終了できる，あるいは，試験デザインを試験途中で変更できる，などのように，上記の伝統的な試験デザインを，試験途中までの結果に応じて変更できる柔軟な様々な試験デザイン，総称としての適応的デザイン（adaptive design），が最近注目を浴びている．更に，上述したように，これまでは，第２相，第３相試験が，それぞれ，独立にデザインされ，実施・解析が行われてきたが，第２相と第３相を一つの試験デザインとして結合した適応的デザインも提案され始めた．これは，いわば，途切れのないデザインと言う意味で適応的シームレス・デザイン（adaptive seamless design）と呼ばれている．最近は適応的デザインに関するテキスト，ソフトウエアも次第に増加傾向がある（例：丹後，2003; Chang, 2008; ADDPLAN, 2009）．

本稿では，代表的な適応的デザインを歴史的な順序で紹介することを目的とするが，特に断らない限り，検定，p値，試験全体の有意水準 α（= 0:025）は片側とする．

Ⅱ．古典的なグループ逐次デザイン

適応的デザインの基礎を築いたのは Armitage et. al.（1969）の逐次検定（repeated significance tests）であり，それを集団に拡張した，Pocock（1977）のグループ逐次デザイン（group sequential design）である．Pocock は，世界的に有名な臨床試験のテキスト Clinical Trials; A Practical Approach（1983）の著書でもある．グループ逐次デザインの基本的な考え方は，試験期間に解析可能な症例が一定数集積される毎に，K 回の検定，つまり，中間解析（interim analysis）を行うデザインであった．彼は両側検定に基づく手順を考えたが，片側検定で表現すると，各 stage での片側有意水準を α_k と設定，しかし，試験期間全体での片側有意水準は α となるように調整する方法である．具体的手順は次のようになる：

1. 第 k（=1, ..., K）stage（一定の症例数が集積された時点を指す）での両側検定で新薬群が対照群に有意に優った場合，すなわち，片側 p 値が $p_k<\alpha_k$ を満たせば「有効」と判断し試験を終了する（有効早期中止），
2. 第 k stage での両側検定で新薬群が対照群に有意に劣った場合，すなわち，$p_k>1-\alpha_k$ であれば「無効」として試験を中止する（無効早期中止），
3. 第 k stage で $\alpha_k \le p_k \le 1-\alpha_k$ であれば，再び，一定の症例数が集積されるまで試験を継続する，
4. 事前に決められた最終 K stage でも $\alpha_K \le p_K \le 1-\alpha_K$ であれば試験は終了し，有意水準 α で帰無仮説を否定できる証拠は得られなかったと結論する．

この推測プロセスで重要な点は事前に宣言された有意水準 α がプロセス全体で保持するように $\alpha_1, ..., \alpha_k$ の値が設計されている点である．各 stage 毎に治療の安全性と有効性を評価する中間解析の実施と解釈は，試験とは独立に組織された独立データモニタリング委員会（IDMC, independent data monitoring commitee）によって行われ，有効性ばかりか安全性を検討し有害事象，副作用などが期待した以上の多ければ試験の中止を勧告できる．

以下では，今でもよく利用される Pocockの方法と O'Brien-Fleming の方法について，エンドポイントが正規分布にしたがう連続変数で，新薬群（A）と対照群（B）の間で等分散 $\sigma^2_A=\sigma^2_B=\sigma^2$ が仮定できる平均値の差の検定の例で紹介しよう．この場合，検定仮説は片側検定（片側有意水準 α）

$$H_0:\mu_A=\mu_B,\quad H_1:\mu_A>\mu_B$$

(1) Pocock の方法

Pocock（1977）は各群 n 例づつ計 $2n$ 例集積された時点で中間解析を最大 K 回繰り返すグループ逐次デザインを提唱した．その特徴を示すデータとして Pocock の論文の中の Table 1 を少々修正したものを表１に示す．彼の方法は両側検定で考えているが，

$$\alpha_1 = \alpha_2 = \quad \cdots \quad = \alpha_K = \alpha'/2 \qquad (1)$$

とすべての中間解析での有意水準を等しく$\alpha' = 2$と設定しているのが特徴である．その値はKが増加するにしたがって減少している．

表 1　Pocock (1977) によるグループ逐次デザインでの特徴．両側有意水準 5% で，effect size $\delta/\sigma=1$ の場合に検出力 95% を達成する症例数を計算している．

中間解析の回数K	各stageでの両側有意水準 α'	各stageでの棄却点 α	各 stageで必要な症例数* $2n$	最大症例数 $2nK$	対立仮設の下で試験終了までに期待される症例数*
1	.05	1.96	51.98	52.0	52.0
2	.0294	2.178	28.39	56.8	37.2
3	.0221	2.289	19.73	59.2	33.7
4	.0182	2.361	15.19	60.8	32.2
5	.0158	2.413	2.38	61.9	31.3
10	0.0106	2.555	6.50	65.0	29.8
20	0.0075	2.672	3.38	67.6	29.5

*：effect size ＝ $\delta/\sigma=(\mu_A-\mu_B)/\sigma$ に対する症例数は $(\delta/\sigma)^2$ 乗ずる．

[例 1] 例えば，effect size を$\delta/\sigma=(\mu A-\mu B)/\sigma=0.5$と見積もった場合，両側有意水準 5% の両側検定で検出力 95% で各群同数で割り付け，中間解析をしない場合に必要となる通常の症例数は表 1 の K=1 のところを参照して$2n=52\times(0.5)^{-2}=208$例となる．これに対して，中間解析の回数$K$を増やしていくと，この最大症例数$2nK$は若干増加するが，対立仮説が正しい場合に試験終了までに期待される症例数が減少している点に注目したい．例えば，最大で，3 回の中間解析を考えると，各 stage での有意水準は$\alpha'=0.0221$，各 stage で必要となる症例数は$2n=19.73\times 2^2=80$例，最大で合計 240 例と中間解析を考えないデザインに比べて 32 例ほど増えることになる．しかし，effect size の見積もりが正しければ，「有意差あり」と判断されるまでに要する期待症例数は$33.7\times 2^2=135$例と約 73 例の節約となる．また，表から$K=5$より増やしても対立仮説の下での期待症例数はあまり減少していないことがわかる．

(2) O'Brien-Fleming の方法

さて，Pocock の「すべての中間解析での有意水準が同じ」と仮定する方法は

1. グループ逐次デザインを採用する根拠の一つは，見積もりを超えた「驚くべき有意差」が検出された場合には試験を早期に終了すべきという方針であろう．しかし，常に同程度の驚くに足らない有意差で試験を早期に終了するのは妥当なデザインと言えるのだろうか？
2. 例えば，K=5，α'=0.0158（片側有意水準は 0.0158＝2＝0.0079）のデザインで，試験は最終 stage まで継続し，最後の片側p値は$p=0.015$であったとしよう．当然，グループ逐次デザインでは有意差はないと判定される．しかし，グループ逐次デザインを採用しなければ片側p値は$p<0.025$となり，有意差ありと判定されたのではないか？

などの点で受け入れがたい方法であるという非難も少なくない．この問題を解決するには早期のα_kはかなり小さくして，α_Kは全体の有意水準αにほとんど近い値にすることである．この一つの解が O'Brien-Fleming (1979) の提案である．彼らは各 stage の有意水準を等しく設定するのではなく，各 stage の検定統計量 S_k の棄却点（critical point）α_k ($|Sk|>\alpha_k$であれば有効と判定し試験終了）を等しく

$$\alpha_1 = \alpha_2 = \quad \ldots \quad = \alpha_K = \alpha \qquad (2)$$

と設定した．こうすると各 stage での有意水準 α_k は k が増加するにしたがって，α_k も増加する．例えば，K=5 の場合であれば，各 stage 毎の両側有意水準は α'_k= 0:00001, 0.0013, 0.0084, 0.0226, 0.0413 となり，K−stage では全体の両側有意水準 0.05 に近い事がわかる．この他にも，「有意差あり」と判定されるまでに期待される症例数を最小にする Wang-Tsiatis (1987) の方法もある．

Ⅲ．α 消費関数

古典的なグループ逐次デザインでは，1）各stage に同じ症例数 $2n$を仮定，2）事前に決められた中間解析の回数の最大値 K は変更できず，現実の臨床試験を考えると実用的とはいい難い．この二つの制約をはずした α 消費関数 (α-spending function) の概念が Lan-DeMets (1983) により提案されてからグループ逐次デザインの適用が広まったといっても過言ではない．その基本的なデザインは，次の通りである．

1. まず，適当な方法により試験に必要な目標症例数（あるいは期待イベント数）を設計する．
2. 次に，解析時点 t ($0\leq t\leq 1$) の関数で，値域 [0, α] を取る単調増加関数 $\alpha^*(t)$を設定する．$\alpha^*(0)=0$，$\alpha^*(1)=\alpha$を満たす増加関数であれば，原則，何でも良い．

このように試験デザインを設定しておくと，試験を開始してから，中間解析を行いたい時点を決めることができ，その中間解析 (検定) に必要な有意水準は，それまでにエンドポイントが観測されている症例数の目標症例数に対する割合を解析時点 t (情報時間，information time，と呼ぶ) とした消費関数 $\alpha^*(t)$ から簡単に計算できる．具体的には，第 k 回目の中間解析 (情報時間 t_k) の有意水準 α_k は

$$\alpha_1 = \alpha^*(t_1),$$
$$\vdots$$
$$\alpha_k = \alpha^*(t_k) - \alpha^*(t_k - 1),$$

と設定される．つまり，全体の有意水準 α を最終の解析を含めた中間解析に振り分ける，あるいは，それぞれの中間解析で消費する分量を事前に決めた消費関数で計算できる方法と考えることができる. 消費関数の考え方からすれば，古典的な Pocock の方法は各中間解析で同じ分量を消費する方法であり，O'Brein-Fleming の方法は最初は極めて少なく，徐々に増加させる方法の一つと考えることができる．

しかし，実際にどのような消費関数を用いるかは，任意であると言われても，迷うところであり，古典的な Pocock の方法，O'Brien-Fleming の方法とその消費パターンが類似している Pocock 型，O'Brien-Fleming 型と言われている消費関数が用意されている．

[例 2] Pocock のデザイン（K=5，$2n$=40）で試験を開始したとしよう．しかし，第 1 回目の中間解析に計画した症例数 $2n$=40 例を超えて A 群 n_A=25 例，B 群 n_B=30 例が解析対象となった状況を考えてみよう．この場合に従来の Pocock の方法は適用できない．ところが α 消費関数を利用すれば（事前に指定しておけば）問題は解決されるのである．つまり，Pocock 型の消費関数を使用するとプロトコールに定義しておくことにより，第 1 回の中間解析での情報時間は，それまでの症例数の割合は t_1=（25+30）=（100+100）=0.275 となるから有意水準を

$$\alpha_1 = \alpha^*{}_{\text{Pocock}}(0.275) = 0.0193$$

として検定を実施すればよい．

[例 3] CAST（Cardiac Arrythmia Suppression Trial）の DSMB（Data and Safety Monitoring Board）に採用された．α 消費関数は

$$\alpha^*(t) = \begin{cases} (\alpha/2)t, & t < 1 \\ \alpha, & t = 1 \end{cases}$$

が採択された．ここで α＝0.025 である．期待イベント数は当初 425 と推定された．第 1stage の中間解析で active 群 22 例，placebo 群 7 例，計 29 例のイベントが観測され $\alpha_1=\alpha^*(29/425)=0.0009$ と計算された．log-rank 検定の正規近似統計量は −2.82 であったが，イベントの発生数が少ない場合のこの統計量の正規近似が疑わしいので並べ替え検定（permutation test）で p 値を求めた（事前にプロトコールでそのように定義された）．その結果，$p_1>\alpha_1$= 0.0009 となり，試験は継続となった．第 2 回めの中間解析までに active 群 33 例，placebo 群 9 例，計 42 例のイベントが観測され，active 群の死亡が期待に反して増加した．第 2 stage の中間解析の有意水準は α_2 = 0.0011 と計算された．しかし，log-rank 検定の正規近似統計量は − 3.22 で，計算された p 値（p_2）が試験の中止勧告の棄却域に落ちた．DSMB は第 2 回の中間解析でこの検定の結果を参考に副作用の多発を理由に試験を中止したのである（CAST, 1989）．

Ⅳ．適応的グループ逐次デザイン

消費関数を含めたグループ逐次デザインでは，予想もしなかったような効果が観察された場合，あるいは逆に無効であったり，副作用が多発した場合には早期に試験を終了することができた．しかし，中間解析の結果から観察されたデータに基づいて，症例数を再設定するなどの試験デザインを変更をすることは原則できない．これに対して適応的グループ逐次デザイン（adaptive group sequential design）では早期の終了と中間解析の結果に基づいて途中での試験デザインの変更を可能にすることができる点で注目を浴びており，適応的デザインの最近の進展は目覚しいものがある．もちろん，全体の有意水準は一定値 α に保たれていることは言うまでもない．ここでは，「適応的デザイン」として解説しよう．

適応的デザインで用いる基本的な統計量は片側 p 値である．ここでは，なかでも最もよく利用される，中間解析を 1 回行う K=2 の 2-stage 適応的デザインを紹介しよう．第 1 stage で解析対象となった症例のデータに基づく片側 p 値を p_1, 第 2 stage で新たに解析対象となった症例のデータに基づく片側 p 値を p_2 とすると，次の手順で行われる．

1. Stage 1
   （a）$p_1>\alpha_0$ となれば帰無仮説 H_0 を採択する（無効早期中止）．
   （b）$p_1<\alpha_1$ であれば帰無仮説 H_0 を棄却する（有効早期中止）．
   （c）$\alpha_1 \leq p_1 \leq \alpha_0$ であれば試験を継続する．
2. Stage 2
   （a）$g(p_1, p_2) \geq c_\alpha$ であれば帰無仮説 H_0 を採択する（無効で試験終了）
   （b）$g(p_1, p_2) < c_\alpha$ 帰無仮説 H_0 を棄却する（有効で試験終了）

もちろん，試験全体での有意水準が α となるように，パラメータ（α_0, α_1, c_α）の調整が必要である．統計量 $g(p_1, p_2)$ については，次のような提案がされている：

$$g(p_1+p_2) = \begin{cases} p_1 p_2, & \text{(Bauer-Kohne, 1994)} \\ 1-\Phi(w_1 Z_{p1}+w_2 Z_{p2}), & \text{(Lehmacher-Wassmer, 1999)} \\ p_1+p_2, & \text{(Chang, 2007)} \end{cases}$$

ここで，Z_p は平均 0，分散 1 の標準正規分布の上側 100p パーセント点，$\Phi(.)$ は標準正規分部の分布関数，w_1, w_2 は重みで，$w_1^2+w_2^2$=1 を満たす．2-stage 適応的デザインを始めて提案した Bauer-Kohne（1994）は Fisher の p 値の統合検定を利用した．Lehmacher-Wassmer（1999）は逆正規分布に基づく p 値の統合検定を利用するものである．ただ，これらの方法は，2 − stage デザインのパラメータの値（$\alpha_0, \alpha_1, c_\alpha$）の計算が面倒である．一方，Chang（2007）の方法は，これらの値の計算が簡単にできる利点がある．しかし，いずれにしても，具体的な利用にあたっては，試験の目的に応じて，パラメータ（$\alpha_0, \alpha_1, c_\alpha$）の値の組み合わせを選ぶ必要がある．

さて，第 2 stage での帰無仮説の棄却条件は，いずれの方法でも

$$p_2 \leq \alpha(p_1)$$

と書きかえることができる．つまり，第 2 stage での有意水準を $\alpha(p_1)$ と設定して独立に試験を新しく始めることができることを意味する．この考え方を利用して Proschan-Hunsberger（1995）は

$$\int_0^1 \alpha(p_1)dp_1=\alpha$$

を満たす任意の関数 $\alpha(p_1)$ を導入し 2-stage デザインを一般化した．この $\alpha(p_1)$ を条件付 Type I エラー関数（conditional type I error function）と呼ぶ．上記の Bauer-Kohne の

2-stage デザインを条件付 Type Ⅰエラー関数で表現すると次のようになる：

$$\alpha(p_1) = \begin{cases} 0, & \text{if } p_1 > \alpha_0 \\ c_\alpha / p_1, & \text{if } \alpha_1 \le p_1 \le \alpha_0 \\ 1, & \text{if } p_1 < \alpha_1 \end{cases}$$

適応的 2-stage デザインで特徴的なのは検定統計量，症例数などはどこにも現れてないということである．言いかえれば，中間解析によって観察された全ての情報に基づいて第 2 stage の試験をデザインできることを意味する．例えば，中間解析の結果から，残りの症例数を再設定することができる．平均値の差の検定で言えば第2-stageの各群同数の症例数 n_2 は，

$$n_2 = 2\left(Z_{\alpha(p1)} + Z_\beta\right)^2 \left(\frac{\sigma}{\delta}\right)^2$$

と再設定できる．ここに 100 (1−β) % は第 2 stageで達成したい検出力である．最近は，様々な柔軟な適応的グループ逐次デザインが提案されている（例：Wassmer, 1999; Liu-Chi, 2001; Denne 2001; Muller-Schafer,2001, 2004; Brannath et al.,2002; Chang, 2008）．

Ⅴ．適応的シームレス II/III 相デザイン

シームレスデザイン（adaptive seamless phase II/III design）は，従来の用量反応と至敵用法用量の検索を行う第 II 相試験とプラセボあるいは標準治療との比較を行う第 III 相試験を一つの継ぎ目のない（seamless）臨床試験としてデザインするものである（例：Bauer-Kieser, 1999; Sampson-Sill, 2005; Shun-Lan-Soo, 2008）．従来の独立したデザインに比べ，試験を実施する際の様々な費用と時間（例：施設内審査委員会 IRB への提出など）の削減，必要な症例数の削減，結果として，有効性ある薬剤が早期に市場に登場できることにつながる，という点で製薬メーカーにとっては極めて魅力的なデザインであると言われている．また，統計学的な検定の有意水準，検出力という観点からは，企業にとっては魅力的かもしれない．なぜなら，ひとつに結合したシームレスデザインにおいても従来のように，同じ有意水準 α，同じ検出力 100 (1−β) % と設定して症例数を設定するとすれば，従来の第 II 相と第 III 相全体での有意水準と検出力は，二つの試験の独立性を仮定すれば（実際には相関があるが），それぞれ α^2，$100(1-\beta)^2$ となり検出力の点でもシームレスデザインの方が優れるからである．

これまでに主要なデザインの一つとして，次のようなデザインが検討されている：

1. 第 1 stage として，従来の第 II 相のように対照（プラセボ群）と k 個の用量群，合計 (k+1) 群で試験をスタートする
2. 中間解析（従来の第 II 相の解析時点）で最も反応の良かった用量群と対照群だけを残して試験を継続する（第 2 stage の試験を開始する）．他の用量群はここで試験から脱落させる．この意味で，drop-loser design とも呼ばれている．
3. 試験の最終解析時点では，残った群の比較を行い，有効性を評価する．

もちろん，上記のデザインは最も簡単な 2−stage デザインであるが，それぞれの stage において更なる中間解析を実施することも可能である．また，このデザインでは，第 1 stageで試験をストップする可能性を考慮していないので，前節で紹介した，2−stage デザインとは異なる．しかし，中間解析でたまたま最も反応の良かった用量群と対照群を比較するということは，用量反応関係が臨床で十分に確認されないまま認可される可能性につながるだけに，必ずしも適切とは言えない．中間解析での用量反応関係の証明と至敵用量の選び方には一層の工夫が必要となる．

Ⅵ．適応的デザインの問題点

これらの適応的デザインの議論は，type Ⅰエラーの確率である試験全体としての有意水準を名目の α に制御する（言いすぎを押さえる）方法が中心となって議論が展開されてきている．しかし，当初想定した試験デザインを試験途中で「変更する（せざるを得ない）」ということは，治療の効果を表す effect size，他のパラメータに関する十分な情報がない，不確実性が高い，ことを示唆している．したがって，適応的デザインはいわば試行錯誤デザインとも表現できるので，このようなデザインで行われた試験で（たまたま）有意な結果が得られたからと言って，その薬剤の効果が検証されたと言えるのだろうか？という大きな疑問が残る．Type Ⅰエラーを制御できていれば，なにをやっても良いということではない．つまり，従来の第 3 相試験に求められていた検証的試験（confirmatory trial）の性格が適応的デザインには乏しくなっている可能性が大であることが危惧される．臨床効果が十分に検証されていない薬剤が診療の現場で使用され始めると，危険にさらされるのは患者であることは言うまでもない．新薬の開発から認可までの時間を短縮することと薬剤の効果を十分に検証することとは，お互い重要な命題であるが，相反する事象かもしれない．企業の利益と患者の利益をどのようにバランスを取っていくか？これまで以上に今後の大きな課題であり，その視点からの統計的デザインの開発が必要不可欠な状況と考える．

参考文献

[1] ADDPLAN: adaptive designs- plans and analyses -, Release 4. ADDPLAN GmbH, 2009.

[2] Armitage P, McPherson CK, Rowe BC. Repeated significance tests on accumulating data. J Roy Stat Soc Series A 1969;132:235-44.

[3] Atkinson EN, Brown BW. Confidence limits for probability of response in multistage Phase II clinical trials. Biometrics 1985;41:741-4.

[4] Bauer P, Kohne K. Evaluation of experiments with adaptive interim analyses. Biometrics 1994;50:1029-41.
[5] Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. Statistics in Medicine 1999;18:1833-48.
[6] Brannath W, Posch M, Bauer P. Recursive combination tests. J Amer Stat Ass 2002;97:236-4.
[7] CAST Investigators. Effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. New England Journal of Medicine 1989; 312:406-12.
[8] Chang MN, O'Brein PC. Confidence intervals following group sequential tests. Contr Clin Trials 1986;7:18-26.
[9] Chang M. Adaptive design method based on sum of p-values. Statistics in Medicine 2007;26:2772-84.
[10] Chang M. Adaptive design theory and implementation using SAS and R. Boca Raton:Chapman & Hall/ CRC;2008.
[11] Denne JD. Sample size recalculation using conditional power. Statistics in Medicine 2001; 20:2645-60.
[12] Duffy DE, Santer TJ. Confidence intervals for a binomial parameter based on multistage tests. Biometrics 1987;43:81-93.
[13] Emerson SS, Kittelson JM. A computationally simpler algorithm for the UMVUE of a normal mean following a sequential trial. Biometrics 1997;53:365-9.
[14] Kim K, DeMets DL. Confidence intervals following group sequential tests in clinical trials. Biometrics 1987;43:857-4.
[15] Jennison C, Turnbull BW. Interim analysis: the repeated confidence interval approach. J R Statist Soc, Series B 1989;51:305-61.
[16] Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika 1983;70:659-63.
[17] Lan KKG, Zucker D. Sequential monitoring for clinical trials: the role of information and Brownian motion. Stat Med 1993;12:753-65.
[18] Lehmacher W, Wassmer G. Adapting sample size calculations in group sequential trials. Biometrics 1999;55:1286-90.
[19] Liu Q, Chi GYH. On sample size and inference for two-stage adaptive designs. Biometrics 2001;57:172-7.
[20] Liu A, Hall WJ. Unbiased estimation following a group sequential test. Biometrika 1999; 86:71-8.
[21] Mulier HH, Schafer H. Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches. Biometrics 2001;57: 886-91.
[22] Mulier HH, Schafer H. A general statistical principle of changing a design any time during the course of a trial. Stat Med 2004;23:2497-508.
[23] O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979;35:549-612.
[24] Pawitan Y, Hallstrom A. Statistical interim monitoring of the cardiac arrhythmia suppression trial. Stat Med 1990;9:1081-90.
[25] Pocock SJ. Clinical trials: a practical approach. Chicester:Wiley;1983.
[26] Pocock SJ. Group sequential method in the design and analysis of clinical trials. Biometrika 1977;64:191-9.
[27] Sampson A, Sill MW. Drop-the-losers Design: normal case. Biometrical Journal 2005;47:257-68.
[28] Shun Z, Lan KKG, Soo Y. Interim treatment selection using the normal approximation approach in clinical trials. Stat Med 2008;27:597-618.
[29] Tsiatis AA. The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time. Biometrika 1981; 68: 311-315.
[30] Tsiatis AA. Repeated significance testing for a general class of statistic used in censored survival analysis. Journal of the American Statistical Association 1982; 77:855-861.
[31] Tsiatis AA, Rosner,GL and Mehta,CR. Exact confidence intervals following a group sequential test. Biometrics 1984; 40:797-803.
[32] Wassmer G. Multistage adaptive test procedures based on Fisher's product criterion. Biometrical J 1999; 41:279-293.
[33] Wang SK and Tsiatis AA. Approximately optimal one parameter boundaries for group sequential trials. Biometrics 1987; 43:193-199.
[34] 丹後俊郎 . 無作為化比較試験，朝倉書店，2003.