群間差に関する事前情報を考慮した
ベイズ流被験者数再設定方法と新しい評価系の提案

上村 鋼平

目次

１章　緒言　1

1.1 検証的臨床試験における被験者数設定の現状……2

1.2 被験者数設定の不確実性……3

1.3 被験者数設定の不確実性に対する試験デザインアプローチ……5

1.3.1 群逐次デザイン……7

1.3.1.1 データモニタリングの必要性……7

1.3.1.2 群逐次デザインの方法論……9

1.3.2 被験者数再設定……11

1.3.2.1 被験者数再設定における第一種の過誤確率の制御方法……12

1.3.2.2 被験者数再設定の問題点……14

1.4 本研究の目的……16

２章　群逐次デザイン　17

2.1 想定する検証的臨床試験……17

2.2 中間解析及び最終解析……17

３章　被験者数再設定　19

3.1 被験者数再設定を伴う検証的臨床試験……19

3.2 第一種の過誤確率の制御……20

3.3 既存の被験者数再設定基準……20

3.3.1 delta-replacement 基準……21

3.3.2 条件付検出力基準……21

3.3.3 条件付検出力の不確実性……23

3.3.4 ベイズ流予測検出力……24

3.3.5 無情報事前分布に基づく予測検出力基準……25

3.4 提案する予測検出力基準……26

４章　シミュレーション研究　30

4.1 想定する試験設定……30

4.1.1 全般的な試験環境設定……30
4.1.2 群逐次デザインに関する設定……31
4.1.3 被験者数再設定に関する設定……32
4.2 検証的試験デザインの評価方法……33
4.2.1 包括的評価指標……33
4.2.2 直接評価指標……35
4.3 結果……37
4.3.1 検出力……37
4.3.2 期待被験者数……38
4.3.3 期待被験者数100例あたりの検出力……39
4.3.4 期待後悔度……39
4.3.5 平均不足検出力……40
4.3.6 平均過剰被験者数……40
4.3.7 平均後悔度……41
4.3.8 PR 内及び全治療効果範囲内の要約結果……42
５章　考察 **43**
６章　結論 **50**
謝辞 **51**
参考・引用文献 **52**
図表 **63**
表1. シミュレーション研究の全般的試験環境設定……63
表2. 治療効果のPR（$\delta \in [0.2,0.3]$）における各評価指標の要約値：平均値（ 最小値 – 最大値 ）……64
表3. 治療効果の全範囲（$\delta \in [0.15,0.35]$）における各評価指標の要約値：平均値（ 最小値 – 最大値 ）……65
図1. 条件付検出力の分布……66
図2. 検出力　$Power$……67

図3. 期待被験者数　ASN……..68

図4. 期待被験者数 100 例あたりの検出力　$Efficiency_{100}$…….....69

図5. 期待後悔度　$ER(\%)$…….....70

図6. 平均不足検出力　MUP…….....71

図7. 平均過剰被験者数　MOS…….....72

図8. 平均後悔度　$MR(\%)$…….. ……....….….....73

１章　緒言

今日の生物医科学の革命的な進歩により、生活習慣病を含む難治性疾患の予防、治療、治癒への期待が高まっているにもかかわらず、そのことがより有効で、より安く、より安全な医薬品を早く患者さんの手に渡ることに必ずしも繋がっていない。この背景に、医薬品の研究開発が世界中でますます成功しにくく、非効率的で、費用、時間のかかるものになりつつあるという現状がある。

日本における、医薬品の候補化合物が上市に至る開発成功確率は、およそ一万分の一と言われている。日本製薬工業協会による調査では、1995~2000 年の間の 18 社合計で、422,653 個の化合物のうち承認されたものはわずか 63 個で、自社製品に限ると 35 個（1/12,076）まで減少する [1)]。また、開発費は年々増加傾向にあり、大手製薬企業 10 社の平均研究開発費は 1995 年に 334 億円であったものが、2003 年には 622 億円にまで増加しており [1)]、開発全体のなかで臨床開発に最も多くの費用と期間が投じられている。

米国も同様の現状にあり、医薬品の研究開発には、前臨床試験開始から承認取得まで 15~17 年もの長い期間を要し、推定 8~17 億 US ドル [2)]もの巨額の研究開発費が費やされている。さらに、1995~2000 年に承認申請に至った薬剤にかかった平均 11 億ドルの開発費用に対し、2000~2002 年でのそれは平均 17 億ドルと約 1.6 倍 [2)]に増加しており、検証を目的とする第三相試験及び第二相試験が、その費用増大のほとんどを占めている。研究開発費の高騰は、薬価 [3)]へ反映されて医療費が増大 [4)]するだけでなく、革新的でリスクの高い薬剤や希少疾患の治療法開発への投資の主な障壁となる [5)]ため、国民の不利益にも繋がる。

このような医薬品開発の厳しい現状を受け、2004 年 3 月に、米国 FDA (Food and Drug Administration)は、‘Innovation or Stagnation? – Challenge and Opportunity on the Critical Path to New Medical Products[6)]’という白書を公表した。この‘Critical Path’とは、候補化合物選択から市販・製造に至るまでの医薬品開発の道程を指している。FDA の‘Critical Path’イニシアティブは、上述の深刻な状況に対し、候補物質の有効性及び安全性に関する新しい評価方法（方法論）の開発・利用と得られる情報の共有・公開の必要性を強く訴えている。FDA Biostatistics 部 Director の O‘Neill 氏は、臨床試験の成功確率を高める新しい方法論の候補の一つとして、本研究のテーマである被験者数再設定を始めとする適応的デザイン

（Adaptive Design）[7-14]に対し、これらの適用が今後増加していくだろうと述べている[15]。

1.1 検証的臨床試験における被験者数設定の現状

新薬の有効性の判断は、十分慎重に計画されたランダム化比較試験（Randomized Controlled Trials RCTs）において示された標準薬（あるいは、プラセボ）に対する優越性を根拠になされることが一般的である。しかし、より早期の第二相試験で有望な結果が観察されたにもかかわらず、引き続くより規模の大きな RCTs において統計的な有意差が得られないことがしばしばある[16]。

研究の検出力とは、真に群間差が存在する場合に、被験薬群と対照薬群との間の臨床的に意味のある差を検出できる確率のことである。この研究の検出力が十分確保された試験を行うべきであるが、被験薬の有効性の根拠となる統計的有意差を示すことに失敗した場合、たとえその試験で臨床的有意差を示すことができるのに十分な被験者数が確保されていなかったとしても、研究者は被験薬のベネフィットはないという誤った結論をくだすかもしれない。実際、一般医学雑誌及び専門誌に報告されたネガティブな結果に終わった RCTs 研究の多くは、臨床的に意味のある群間差を検出するための十分な被験者数が確保されていないことが示されている[17-26]。

Bedard et al. (2007)[26]は、1995~2003 年までに ASCO (American Society of Clinical Oncology) 年次総会で発表されたすべての二群比較第三相 RCTs のうち、ネガティブな結果（$P > .05\ or\ 95\%CI\ including\ 1.0$）が得られたすべての試験の検出力を評価した。主要評価項目の変数型は、イベント発生割合またはイベント発生までの時間となっているもので95%以上を占めており、効果サイズ（本論文では、効果サイズ、治療効果、試験治療群間差、治療群間差、群間差は同義とする）の指標はそれぞれオッズ比（OR）、ハザード比（HR）であった。そこで、小さな、中程度の、大きな効果サイズをそれぞれ$\mathrm{OR/HR} \geq 1.3, \mathrm{OR/HR} \geq 1.5, \mathrm{OR/HR} \geq 2.0.$と定義し、事後的に被験者数の再設定を行った。

その結果、評価可能な 423 試験のうち、45 試験（10.6%）が小さな効果サイズを検出するのに十分な被験者数であり、138 試験（32.6%）、233 試験（55.1%）がそれぞれ中程度、大きな効果サイズならば検出力が確保できただろう試験であることが判明した。つまり、

ネガティブ試験の大半が、臨床上意味のある群間差が真に存在しなかったというより、小さめ又は中程度の群間差を検出する被験者数設定がなされていなかった可能性が示唆された研究結果といえる。それにもかかわらず、35 のネガティブ RCTs (7.1%)しか、不十分な被験者数を原因として報告していなかったことは注目に値する。また、多変量モデルによる要因分析の結果、口頭セッションで発表された研究（$P = .0038$）、共同研究グループにより支持された多施設研究（$P < .0001$）、主要評価項目がイベント発生までの時間の研究（$P < .0001$）の方がより十分な被験者数が確保される傾向、すなわち保守的な被験者数設定が実施される傾向が示唆された。

研究の設定被験者数の妥当性に関するこのような系統的評価は、被験者数設定の不確実性の現状を捉える上で重要である。個別の試験結果の評価では、統計的有意差が観察されたかどうかの二値結果しか得られないため、どの程度の検出力を保持していたかを事後的に評価することはできない[27]。故に設定被験者数の妥当性を二値的結果から判断すべきでないと考えられる。また、これらの系統的な試験デザインの事後評価[17-26]により、質の高い研究と思われる多くの試験の現状として、大きめの効果サイズを期待した不十分な被験者数が設定され検出力不足に陥っている可能性が示唆される。

1.2 被験者数設定の不確実性

この節では、前節の検証的臨床試験の現状を受け、被験者数設定の不確実性について論じる。

検証的臨床試験における被験者数設定は、試験計画の要素のなかで最も重要なものの一つである。被験薬の有効性を検証するのに十分な精度を確保するためには、被験者数を多く設定しなければならない（International Conference on Harmonisation ICH E9 ガイドライン「臨床試験のための統計的原則」[28]）。一方、倫理的側面を考えると、被験薬もしくは対照薬によって引き起こされるかもしれない有害事象のリスクに曝される人数を意味する被験者数は、可能な限り少なく設定すべきである。被験者数設定の別の側面として、安全性に関する情報や開発費用も勘案しなければならない。被験薬の承認を行う規制当局や将来の患者の立場に立てば、被験薬の安全性情報を十分に確保するため、多くの被験者数が要求

されることがあるかもしれない。一方、被験薬の申請者または試験実施者の立場に立てば、被験薬の有効性を検証できる範囲内でできる限り被験者数は少なく済む方が経済的と考えられる。このように、複数の互いに矛盾した要求の対立軸が存在するもとでは、試験の検証的目的を達成可能な必要最小限の被験者数設定が、その原理・原則とならざるを得ない[29, 30]。

必要最小限の被験者数を設定するためには、被験薬及び対照薬による試験治療群間での結果の差とばらつきに関する情報が必要となる。それらを用い必要被験者数の推定値Nは、基本的に以下のように群間差δをそのばらつきσで標準化したものの二乗に反比例する形で設定される。

$$N \propto (\delta/\sigma)^{-2}.$$

δ/σを算出するために必要となる試験パラメータは、結果変数の型によって異なる。結果変数が連続量の場合[31]には、平均値の群間差と群内分散を、二値の場合[32]には、反応割合の群間差と対照群での反応割合を、イベント発症までの時間の場合[33, 34]には、群間のハザード比と対照群でのイベント発生率を用いてδ/σを計算する。しかし、これらの試験パラメータに関する情報・知識は、試験開始前の時点では不確実である場合がほとんどである。

それには、いくつかの理由が考えられる。第一の明らかな理由は、新薬の場合、特にその試験でターゲットとしているものと同じ併用薬剤を用いかつ同質の患者集団への適用という意味では、これまでの適用例がほとんどないことがあり得るということである。この状況は、例えば多剤併用療法による HIV（ヒト免疫不全ウイルス）感染に対する治療においては珍しいことではない[35]。それは、新薬が多剤併用療法のうちの一剤として検証試験の対象となると、併用薬剤のなかに新しく承認された別の薬剤が含まれていたりすることがあるためである。

第二の理由は、急激な医療環境の変化[36]が、対象疾患のイベント発生率などの試験パラメータを大きく変えてしまうかもしれないということである。ベースラインリスクが変化すれば、被験薬の有効性としての群間差のサイズも当然変わり得る。よって、このような不確実な事前情報に基づき設定した被験者数は、適切でないかもしれない。また、試験開始後に急速な環境変化が生じる場合もあり、長期にわたる試験において試験開始前に設定

した初期被験者数はもはや適切ではなくなってしまう可能性も考えられる。

第三の理由は、検証的試験の前に通常実施される第二相試験の規模がそれほど大きくなく、そこから試験パラメータの情報を得る場合[37]には、小規模データの不確実性を伴うことが考えられる[38]。特に、多重エンドポイント[39]や、複合エンドポイント[40]等の特殊なエンドポイントを用いる試験では、複数ある結果変数間の相関の強さも検出力へ影響してくるため、小規模データに基づく検出力評価の精度がさらに悪くなることが考えられる。また、イベントを発症するまでの時間をエンドポイントとする場合には、小規模データより見積もることがしばしば困難な被験者登録率や打ち切り割合等が検出力へ影響するため、被験者数設定の不確実性は他のエンドポイントを用いる場合よりもさらに大きくなってしまう[41]。

このように、検証的臨床試験における被験者数は、その設定に必要な情報が試験開始前に得られていなければならないこともあり、基本的に不確実な情報に基づいているといえる。1.1 節の検証的臨床試験の被験者数設定の現状及び不確実な事前情報を考慮すると、その不確実性に対しより頑健な試験デザイン方法論の開発が望まれる。

1.3 被験者数設定の不確実性に対する試験デザインアプローチ

検証的臨床試験の被験者数設定の不確実性に対し、臨床上意味のある最小の群間差に基づき被験者数を設定する方法が考えられている[42-46]。最小の群間差に基づく被験者数は、原理的に検出力不足に陥ることはないが、要求される被験者数が膨大な数となり、被験者の集積しやすさや開発費といった実施可能性を勘案すると、設定が困難であることが少なくない。検証的臨床試験の被験者数設定の現状は、このような実施可能性が反映された結果であるとも考えられる。一方、臨床上有意な群間差は、安全性とのバランスにも依存するため、被験薬の安全性に関する情報が十分でない場合には明確に定義できないこともしばしばあるだろう[44, 45]。以上より、実際には、最小の差に基づくというより、実施上可能な範囲においてできる限り保守的に多めの被験者数を設定する場合が多くなる[46]。このような通常行われる被験者数設定は、基本的に試験開始前に定めた初期被験者数を試験終了まで変更することは想定しておらず、固定デザインと呼ばれている。

固定デザインで設定可能な被験者数は、多く設定しすぎると試験実施者側の経済的リスクが大きくなってしまうため、一般にその設定数には限度がある。また固定デザインは、途中で試験中止すること（有効性/安全性）は想定していないため、初期被験者数を多く設定しすぎることに倫理的問題が発生する場合もあるかもしれない。そこで、そのような固定デザインに対し、初期被験者数が必ずしも最終被験者数（試験終了時の被験者数）に一致するとは限らない、柔軟性を備えた試験デザインアプローチが二種類提案されている。

一つは、試験途中に盲検を一度解除し群間比較を行う中間解析を実施し、その結果被験薬の優越性が疑いなく立証された場合、又は適切な治療群間差を示す見込みのないことが判明した場合、あるいは許容できない有害事象が明らかになった場合に試験を早期に中止することが可能な群逐次デザインアプローチである（臨床試験のための統計的原則[28] 4.5節）。群逐次デザインを用いた臨床試験では、被験薬群が対照薬群と比較して有効性が優れているか又は劣っているかといった判断をくだすために必要な情報を逐次集積していき、その結論に至った時点で試験を中止することが可能になる。このような早期中止の可能性は、固定デザインと同等の検出力を維持しつつ被験者数を平均的に減少させられることがある。そのため、試験開始前に小さめの群間差を想定し被験者数を多めに設定しておいた場合には、（実際には観測できないが）真の群間差がそれよりも大きいならば、予定した初期被験者数よりも少ない中間解析時の被験者数でもって検証的試験の目的を達成できる可能性がより大きくなると考えられる。つまり、固定デザインよりも多くの初期被験者数を設定した群逐次デザインは、被験者数設定の不確実性にある程度対応可能な試験デザイン上のアプローチの一つになる。

もう一つは、試験途中に得られる情報に基づき必要な被験者数の調整を行う被験者数再設定アプローチである。被験者数再設定は、盲検を維持したまま実施するアプローチと盲検を解除して行われた中間解析結果に基づくアプローチの二つに大別される。前者のアプローチは、主に両群併せた全体のイベント発生率が予想よりも低い場合、又は群間差のばらつきが予想よりも大きい場合に、割付結果を明らかにすること又は試験治療間の比較を行うことなく被験者数を見直すことができ、盲検下被験者数再設定（Blinded Sample Size Re-estimation[47]）や内的パイロット研究（Internal Pilot Study[37, 48, 49]）と呼ばれている。臨床

試験のための統計的原則[28]（ICH E9）4.4 節の必要な被験者数の調整では、「盲検性を維持するために行う手続きと共に、可能であれば、第一種の過誤と信頼区間の幅に対する被験者数の変更による影響を説明すべきである」とあるが、盲検下被験者数再設定では、群間差の情報を用いないため、基本的に第一種の過誤確率への影響はほとんどないことが示されている[50]。

本研究で焦点をあてるのは、後者の非盲検下での被験者数再設定（Unblinded Sample Size Re-estimation）であり、これは第一種の過誤確率への影響を考慮した最終解析を行う必要がある。以後、単に被験者数再設定と書いた場合はこの非盲検下のアプローチを指すこととする。

1.3.1 群逐次デザイン

一般に、長い期間をかけて徐々にデータが蓄積されるような試験では、何らかの試験側面に対し途中のデータをモニタリングすることは必須である。データモニタリングの必要性に関する様々な側面には、1.3.2 節の被験者数再設定に関する側面も存在するため、1.3.1.1 節にデータモニタリング自体の必要性を述べる。

有効性による試験中止の判断を伴うデータモニタリングを行う場合には、最終解析時の有効性に関する判断を含め、複数回検定を行うため、検定の多重性が生じる。このような検定の多重性を調整するために提案された統計的モニタリング手法が、逐次デザインと一般に呼ばれている。逐次デザインの理論的背景として、被験者 1 例ごとの連続モニタリングを想定した手法[51, 52]が最初に考えられたが、多くの場合ある一定の被験者数（逐次群）のデータが集積するごとに中間モニタリングを実施する形式の方が、より実際的である。そこで、逐次群ごとにデータが集積される状況に特化した一連の手法として、後に群逐次デザイン[53, 54]が提案された。

1.3.1.1 データモニタリングの必要性

一般に、臨床試験において有効性及び安全性データをモニタリングする場合、その必要性にはいくつかの側面がある[55-57]。一つは、被験者が有害な治療に曝されていないことを

保証しなければならないという倫理的要求により、試験治療が重大な副作用を引き起こすことが判明したら早急に臨床試験を中止しなければならないという側面である。一般に、死亡のような不可逆的なエンドポイントを用いる臨床試験では、中間安全性評価を行うことが要求される。

二つ目は、研究開発費や資源の無駄をなくし最適な試験計画を実施するという経済的考慮の側面である。試験規模がより大きな後期第二相、第三相試験だけでなく、より早期に実施される、研究の最終結果の予測のために行う Proof of concept (POC)試験においても、有効性エンドポイントの中間解析がしばしば実施される。もし、研究が予定した終了時点までにその目的を達成できる可能性が極めて低いことが判明した場合、例えば被験薬の臨床的有効性が存在するという証拠がほとんど得られない場合等、これ以上試験を継続することが資源の無駄であるという理由で早期に試験を中止することがある。このような中止を無効中止（Futility Stop[58-62)]）という。

三つ目は、試験の実施管理上の側面である。長期大規模臨床試験等では、試験が計画通りに進行しているか、治療レジメンに関するプロトコル規定が遵守されているかといったことを試験途中で確認する必要がしばしばある。中間集積データを早期に調べることによって、多くの代償を払う以前の修正が可能な問題点が明らかになることがしばしばある。例えば、ある癌に対する栄養補強食品の発症予防効果を検証するランダム化臨床試験において、一部の被験者でプロトコル不遵守の問題が疑われたとする。このような場合、仮にその予防効果が存在していたとしても、その検出力は失われているかもしれない。そこで、試験のもつ検証という目的を達成するために、潜在的不遵守者の同定を目的とする試走期間を新たに設定するという実施管理上のデザイン変更を行い、そのような対象者を割り付け前に研究から除外するといった対策を講じることが可能になる。

モニタリングの別の重要な実施管理的側面は、試験開始前になされた試験計画上の前提（仮定）を確認することにある。初期被験者数や中間解析の時期や回数を設定する際に必要となる仮定がその主たるものとなる。実施管理上モニタリングが必要な長期大規模臨床試験等では、試験の持つ情報量が関心のあるイベント数に比例することがほとんどである。必要なイベント数が観察されるために必要な試験期間、被験者登録期間、中間解析時期を

予測するために用いる仮定である、全体のイベント発生割合、被験者集積率、打ち切り割合を試験途中で確認し、必要ならば各期間や中間解析時期の変更を行う。同様に、被験者数再設定も、モニタリングの実施管理的側面より動機づけられた試験デザインの仮定の確認及び変更という見方ができる。特に、群間差に基づく被験者数再設定は、有効性モニタリングのための中間解析と密接に絡むため、群逐次デザインの枠組みにおける被験者数再設定方法論が多く提案されている[63-65]。

1.3.1.2 群逐次デザインの方法論

群逐次デザインに適用可能な方法論は、大きく境界アプローチ（Boundary Approach）と繰り返し検定アプローチ（Repeated Significance Testing Approach）に大別される。境界アプローチは、Wald の SPRT (Sequential Probability Ratio Test)[51]を基礎として発展した。Wald の SPRT は、Neyman and Pearson の尤度比検定を逐次モニタリングへ拡張したものである。尤度比検定統計量を縦軸に横軸に情報量をとり、検定統計量が上側境界を越えれば有効性中止、下側境界を越えれば無効中止と判断する片側仮説に関し構成された境界である。この方法論では、第一種の過誤確率の名義水準を片側 5%、検出力が 95%と設定されており、期待症例数又は期待停止情報量はすべての逐次デザインのなかで最小となるという理論上の最適性を有する[52]。

しかし、Wald の SPRT は、結果が一例得られるごとに解析する連続モニタリングを想定して構成された境界であるため、一定の被験者数ごとに中間解析を実施する群逐次デザインへ適用した場合、その理論上の最適性は崩れてしまう。また、設定した検出力 95%は、上下の境界のいずれかを越えるまでデータが集積し続ける、つまり最大被験者数を無限大とすることを前提としたものであるため、実際の有限被験者数の下での検出力は 95%よりも低いものになる。

その後、Wald の SPRT の効率的性能をある程度維持したまま、実際の適用場面に合うように改良された境界アプローチ[52, 66-70]が、Whitehead や Stratton らを中心として数多く提案された。有限サンプルへ拡張した境界アプローチには、Truncated SPRT や Triangular Test があり、それらを両側仮説へ拡張した Double Truncated SPRT や Double Triangular Test があ

る。これらは、いずれも連続モニタリングを想定した手法であり、群間差が大きいほど群逐次モニタリングを想定した手法よりも期待症例数が少なくて済むという良い性質を持つ。そのため、これらの境界アプローチは、致死性又は重篤な副作用を引き起こす可能性のある薬剤を使用せざるを得ない疾患領域、治療域（Therapeutic Range）の狭い抗がん剤の試験において、倫理的観点より群間での安全性の偏りが発覚した時点で即中止の判断を行わなければならないような場面、そもそも少ない被験者数しか確保できない小児疾患や希少疾患領域の試験などに適したアプローチといえる。これらの連続モニタリングを想定した境界を、群逐次デザインへ適用した際に生じる検出力の低下を防ぐために、中間解析ごとに境界の形状をクリスマスツリー型に修正する方法[52)]も提案されている。

一方、群逐次デザインの主要なアプローチは、試験全体の第一種の過誤確率の計算するための再帰的数値積分方法を提案した Armitage の成果[71)]を基に発展した、繰り返し検定アプローチである。代表的な繰り返し検定アプローチには、Pocock[72)]及び O'Brien and Fleming[73)]の境界があり、群逐次デザインとして最初に提案された正式なモニタリング手法といえる。これらの境界は離散的で、各中間解析における検定の棄却限界値の列をなしたものである。この点において、連続モニタリングへの適用を想定した、境界アプローチの連続的な境界線とは異なっている。

繰り返し検定アプローチの境界は、一般に、中間解析と最終解析において検定を繰り返すことによる、検定の多重性を調整するための数値計算に基づいて構成される。多重性を調整する分、最終解析における棄却限界値は、固定デザインの場合よりも大きい（厳しい：有意になりにくい）ものとなっている。Pocock の境界は水平な形状をしており、固定デザインと比較した場合、解析時点によらず一様に厳しい棄却限界値の列で構成されている。また、Pocock の境界の良い性質として、真の治療群間差がより大きい場合にはより早期の中間解析で中止しやすくなるという特徴がある。しかし、治療効果が想定した大きさしかない場合には、最終解析で固定デザインよりも厳しい棄却限界値で検定をすることとなり、検出力が落ちるという欠点がある（Jennison and Turnbull 2000[54)], p. 31-39.）。

そこで、中間解析で有意差が観察されずに最終解析へ至った場合に、固定デザインとほぼ同じ棄却限界値となるよう、中間での第一種の過誤確率の消費を抑えた境界として、

O'Brien and Fleming の境界が提案された。これは、Pocock よりも早期中止となりにくく、最終解析の棄却限界値も固定デザインに近いという点で、より保守的なデザインであり、実際にこの方法がよく用いられている。本論文のシミュレーション研究では、O'Brien and Fleming の境界を用いた群逐次デザインを評価する。

O'Brien and Fleming の保守性を緩和した Haybittle (1971) [74] と Peto et al. (1976) [75] の境界や、Pocok と O'Brien and Fleming の境界を両極端に持ち、その間を一つのパラメータ値を用いて保守性の程度を自由に変えられる、Wang and Tsiatis (1987) [76] の one parameter boundary family も提案されている。さらに、Lan and DeMets (1983) [77] は、それまでに提案されたどの群逐次境界も解析回数や解析時点を予め（試験開始前に）定めておかなければならなかったのに対し、予定した解析回数や時点と異なった場合でも、事後的に棄却限界値を調整することができる α 消費関数を提案した。これによって、群逐次デザインの柔軟性が大幅に上昇した上、O'Brien and Fleming 型及び Pocock 型の α 消費関数も提案されている。

統計的モニタリングの他のアプローチとして、条件付検出力を用いた確率打ち切り法（Stochastic Curtailment[58-62]）や、ベイズ流のモニタリング手法 [78, 79] が提案されている。これらは、境界アプローチや正式な群逐次デザイン手法である繰り返し検定アプローチよりも、直感的にわかりやすいという利点をもつ。確率打ち切り法は、主に無効中止に用いられるモニタリング手法であり、無効中止基準は第一種の過誤確率を低下（deflate）させることはあっても、上昇（inflate）させることはないため、正式な第一種の過誤確率の計算は行わないのが通常である。また、ベイズ流のモニタリング手法は、提案する方法とも関連があるため、3 章のなかでも触れる。

1.3.2 被験者数再設定

被験者数再設定をはじめとする Adaptive Designs の方法論が、1990 年代中頃から急速に発展してきた。これは、近年の薬剤開発の開発成功確率の低下及び開発費増大という現状を受け、Adaptive Design の早期臨床試験及び検証的臨床試験への適用にかかる期待と可能性の大きさを反映している [80]。しかし、Adaptive Designs の種類によっては、より後期の試験のもつべき検証的性格自体を脅かすものもあり、検証的臨床試験への適用の是非を巡

る議論は絶えない[81-89]。そのため、欧州医薬品審査庁 European Medicines Agency (EMEA)、及び米国 FDA は、Adaptive Design に関するドラフトガイダンス[90, 91]を公布しており、Adaptive Design の適用に関する規制側のコメントを述べている。生産者側の動向としては、米国研究製薬工業協会 Pharmaceutical Research and Manufacturers of America (PhRMA) が Adaptive Design Working Group[92-100]を 2004 年に結成し、Adaptive Design の定義と開発上の適正な位置付けに関し、産官学でのコンセンサスを広げるべくその活発な活動を展開している。PhRMA によると、Adaptive Design は、「試験の validity（妥当性）及び Integrity（完全性）を損なうことなく、試験デザインのいくつかの側面を集積中のデータに基づき修正・決定する試験デザイン」と定義されている。EMEA、FDA の定義もほとんど同じである。ここで、妥当性の意味するところとは、第一種の過誤確率の保持等正確な統計的推測を行うこと、試験の異なるステージ間の一貫性を保証すること、Operational Bias（実施上のバイアス）を最小にすることである。完全性の意味するところは、より広い科学の場へ信頼できる結果を提供すること、計画された修正内容でもって可能な限り事前に計画すること、データの情報漏洩を防ぐことである。

被験者数再設定に関して、各規制当局とも審査経験の蓄積が足りないこともあり[81]、検証的な場面への適用に対しては依然として慎重な姿勢を取っているのが現状であるが、適用事例は徐々に増えつつある[100, 101]。一方、第一種の過誤確率の制御の代償としての統計的効率の損失への指摘や、推測の原理の妥当性侵犯に関する議論もなされている。さらに、被験者数設定の不確実性に対する代替的手法として、既に検証試験への適用に関するコンセンサスが十分に得られた群逐次デザインと比較し、被験者数再設定の適用上の利点は存在するか、もしくはどのような状況で被験者数再設定の実施が有用といえるか、または必要になるかということは、未だに決着のついていない議論である。

1.3.2.1 被験者数再設定における第一種の過誤確率の制御方法

中間解析で観察された試験治療群間差に基づき、試験開始前に設定した初期被験者数を調整し最終被験者数を再設定した場合、固定デザインと同様の最終解析を行うと、試験全体の第一種の過誤確率が名義水準以下に必ずしも保たれないことが示されている[63, 102, 103]。

そのため、第一種の過誤確率を制御する方法論が、これまで数多く提案されてきている。

これらは､統合検定の原理に基づく制御方法､条件付過誤関数の原理に基づく制御方法、デザインに基づく制御方法の三つのアプローチに大別できる。統合検定の原理とは、異なるステージ間のデータを別々に解析し､予め定めた統合ルールに基づく統合解析を行えば、第一種の過誤確率が保たれるという原理である。統合の仕方が異なるいくつかの方法が提案されている。最初に提案された統合検定は、Bauer and Köhne (1994) [104)]の方法で、これはメタ･アナリシスに用いる方法として提案されたFisherの統合p値を拡張した方法である。さらに、二段階デザインの枠組みで提案された Bauer and Köhne (1994) [104)]の方法を、再帰的に行うことで多段階デザインへ拡張した方法も提案されている（Brannath, Posch and Bauer, 2002[105)]）。一方、二段階デザインの Bauer and Köhne (1994) [104)]の方法に対し、群逐次デザインの枠組みで提案された以下に挙げる手法の方が、適用上の利便性が高いと考えられる。これは、中間解析の実施を予定した検証的臨床試験では、群逐次デザインを適用する場合が多いからである [28)]。Cui, Hung and Wang (1999) [63)]の重み付 Z 検定は、群逐次デザインを採用した試験において予定された各ステージの被験者数を重みとした統合検定であり、Lemacher and Wassmer (1999) [64)]は、中間解析が等間隔に予定された群逐次試験を想定した逆正規法による統合検定である。他に連続モニタリングの場合に適用可能な Fisher (1998) [106)]の分散消費アプローチによる重み付検定も提案されている。

二つ目のアプローチである条件付過誤関数の原理は、Proschan and Hunsberger (1995) [102)]によって提案された概念で、条件付過誤の不変性を保つことで無条件付の第一種の過誤確率を制御する一般的原理である。条件付過誤関数は、中間解析結果を表した検定統計量又はp値の関数となっている。その関数の満たすべき条件は、中間データに対する関数の期待値が第一種の過誤確率の名義水準に一致することである。Proschan and Hunsberger (1995) [102)]では、事前に定義した条件付過誤関数を用いて最終解析の棄却限界値を調整する方法が提案されている。他に被験者数を変更したときとしなかったときとで条件付過誤を不変に保つアプローチ（Denne, 2001[107)]）やそれを群逐次デザインへ拡張したアプローチ（Müller and Schäfer, 2001[108)], 2004[109)]）も提案されている。一般に、一つ目の統合検定の原理に基づく方法は、条件付過誤関数を用いて統一的に表現できる [110)]ことが示されており、両者の原

理は、本質的には同等なアプローチといえる[111]。

三つ目のアプローチは、第一種の過誤確率が名義水準以下となるように、被験者数再設定に関するデザイン上の制約を設ける方法である。Shun et al. (2001)[103]の方法は、次ステージの被験者数に対する事前に定めたサンプリングルールが完全に遵守される場合にのみ第一種の過誤確率が名義水準以下に保たれる方法である。実際には、再設定後の被験者数の決め方を事前に完全に特定することは困難なため、この方法はあまり実際的ではない。それに対し、Chen, DeMets and Lan (2004)[112]の方法は、中間解析時の条件付検出力が 50%以上であれば、被験者数を何例増加させても第一種の過誤確率は名義水準以下に保たれるという方法である。Uemura, Matsuyama and Ohashi (2008)[113]によって、条件付検出力が 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。これらのデザインに基づく方法は、被験者数再設定を行える状況に制約が設けられてはいるが、最終解析時に固定デザインと同じ通常の解析方法を実施できるという利点をもつ。よって、一つ目及び二つ目のアプローチに見られる第一種の過誤確率の代償としての統計的効率の損失[114, 115]がない方法である。

1.3.2.2 被験者数再設定の問題点

被験者数再設定の統計的問題のうち、第一種の過誤確率の上昇に関する問題は、前節で述べたとおり、多くの方法が提案されたことにより、ほとんど解決したといえる。近年議論されているのは、被験者数再設定の効率に関する問題である。その議論の発端となったのは、Jennison and Turnbull (2003)[114]及び Tsiatis and Mehta (2003)[115]である。

Tsiatis and Mehta (2003)[115]は、ある被験者数再設定デザインに対し、より効率の良い尤度比検定を用いた群逐次デザインが常に存在し得ることを理論的に示した。Jennison and Turnbull (2003)[114]は、従来の群逐次デザインと被験者数再設定を比較したシミュレーション研究を行っており、被験者数再設定の方が比較的検出力が低く期待被験者数も多くなるといった、効率の低下が著しいような状況があることを示した。被験者数再設定における効率低下は、最終解析の検定統計量が十分統計量に基づかない、より端的にいえば被験者一人あたりにかかる重みが中間解析前後のステージ間で等しくないことに起因する[116]。こ

れは、第一種の過誤確率を制御するための代償といえる。

しかしながら、これまでに群逐次デザインと被験者数再設定に関する統計的な最適性（効率）の議論や比較が他にも多くなされており、未だ十分な結論に至っているとはいえない[117-122]。また、理論的な最適性を比較した研究のなかには、中間解析の回数や最大症例数の上限が適切に考慮されておらず、あまり実際的ではないことも多い[123, 124]。さらに、群逐次デザインにおける最適な試験パラメータ（中間解析時期や回数）及び最適な境界の選択、あるいは被験者数再設定における重み付検定の最適な重みの選択及び最適な条件付過誤関数の選択は、試験治療群間差の真値に必ず依存してしまうため、実際には各デザインにおける最適なパラメータ設定が可能であるとは限らない[125]。

Jennison and Turnbull (2003)[114]のシミュレーション研究における環境設定は、最大被験者数が初期被験者数の 16 倍で中間解析が 1 回のみの被験者数再設定と、ある程度多めの初期被験者数が設定された上で中間解析が複数回予定された群逐次デザインとを比較するというものである。実際の試験効用としては、中間解析回数を増やせば試験費用は膨らむと考えられるが、彼らの研究ではそのことを考慮した評価指標が用いられていない。また、初期被験者数を小さく、最大被験者数を大きく設定すると、被験者数再設定で用いる重み付検定統計量の非効率性が実際に用いる場合よりも誇張されてしまう[123]。よって、実際的な状況設定のもとで比較の妥当性をできる限り保った上、群逐次デザインと被験者数再設定との比較を行うべきである。

第一種の過誤確率の制御をするための代償以外にも、被験者数再設定の効率を低下させる別の要因が指摘されている。それは、中間解析時の群間差の推定値に基づく被験者数再設定基準を用いると、一般に期待被験者数が多くなってしまうという指摘[126]である。被験者数再設定基準とは、中間解析後のステージ又は最終被験者数を何例に設定すべきかの指針及び根拠のことであり、具体的には目標検出力を達成するのに必要な被験者数の推定方法のことをさす。既に提案されている被験者数再設定基準は、Cui, Hung, Wang (1999)[63]の delta-replacement 基準（3.3.1 節）と Proschan and Hunsberger (1995)[102]の条件付検出力基準（3.3.2 節）、Wang (2006)[127]の無情報事前分布に基づく予測検出力基準（3.3.5 節）がある。Jennison and Turnbull (2003)[114]のシミュレーション研究では、Cui, Hung, Wang (1999)[63]の

delta-replacement 基準を用いており、その期待被験者数が初期被験者数の 12 倍まで増大してしまっている。また、Bauer (2008) [128]では、中間解析時点で計算される条件付検出力は、群間差の推定誤差の影響を受けるため、不安定になることが示されている。Wang (2006) [127]の予測検出力基準は、無情報事前分布を仮定しているため、中間解析時に観察された試験治療群間差の点推定値を真とみなした方法であり、この点において前者二つの基準と同様にその推定誤差を考慮していない。また、これまでに被験者数再設定基準に関する十分な検討を行った研究は未だない。

1.4 本研究の目的

固定デザインは、試験開始前に仮定した群間差に基づく被験者数を試験終了まで固定したままのデザインであり、事前情報の不確実性を考慮できない。既存の被験者数再設定方法は、事前情報を中間解析結果に基づき修正できるという点で事前情報の不確実性を考慮できるが、中間解析の結果として得られる群間差の推定値を真値とみなした被験者数再設定基準を用いるため、その推定誤差がもたらす中間データの不確実は考慮できていない。そこで、本研究では、事前情報を中間データで置き換えるのではなく、ベイズの定理に基づき事前情報を中間データで更新できる、いわば事前情報と中間データの両方を用いる新しいベイズ流被験者数再設定方法を提案し、その性能をシミュレーション研究により評価することを目的とする。また、そのシミュレーション研究において、近年盛んに議論がなされている群逐次デザインと被験者数再設定という対立軸、及び新しい被験者数再設定と既存の被験者数再設定という対立軸のそれぞれについて性能の比較検討を行う。その際用いる評価指標として、検出力や期待被験者数といった既存の指標に限らず、これまでに議論されてきた統計的効率及び実際の試験効用に与える影響を考慮できるような一連の評価指標からなる、群逐次デザイン又は被験者数再設定といった Adaptive Sample Size Design に特化した新しい評価系の提案も行う。

２章　群逐次デザイン

2.1 想定する検証的臨床試験

プラセボ又は標準薬との比較に基づき被験薬の優越性検証を目的とする標準的な第三相臨床試験の状況を考える。試験治療群$j(j = 1$: 被験薬群, 2: 対照薬群)に割り付けられた被験者iの結果変数Y_{ij}は、一般性を損なうことなく独立に平均がそれぞれ$\mu_1 = \delta, \mu_2 = 0$、共通の群内分散$\sigma^2 = 1$の正規分布に従うとする。被験薬に有効性が存在しないという帰無仮説$H_0: \delta = 0$に対し、対立仮説$H_1: \delta > 0$とする。第一種の過誤確率の名義水準を片側α、目標検出力を$1 - \beta$とする。

試験開始前に仮定した治療効果サイズをδ_{pre}とすると、一般に試験開始前に設定する片群あたりの初期被験者数$N_{initial}$は、以下の式に基づき設定される [31)]。

$$N_{initial} = 2\left(\frac{z_\alpha + z_\beta}{\delta_{pre}}\right)^2, \tag{1}$$

z_uは標準正規分布の第$(1 - u)th$分位点、以下特に断らない限り被験者数はすべて片群あたりの人数で表わすとする。ここで、初期被験者数設定の不確実性を考慮するために、群逐次デザイン（GSD: Group Sequential Design）の被験者数N_{GSD}は、以下のように多めに設定する。

$$N_{GSD} \geq N_{initial}.$$

どの程度多めに設定するかは、4 章で具体的に示す。

簡単のため、中間解析回数は 1 回とし、中間解析までに集積したデータを第一ステージ、中間解析以降最終解析時点までに集積されたデータを第二ステージと呼ぶこととする。第一ステージの被験者数は、$n_1 = tN_{GSD}(0 < t < 1)$とすると、第二ステージの被験者数は、$n_2 = (1 - t)N_{GSD}$となる。ここで、$t = n_1/N_{GSD}$は、中間解析時期を意味する情報量時間または情報量分数と呼ばれる。

2.2 中間解析及び最終解析

中間解析時の二群の平均値の差の検定統計量Z_1は、両群併せて$2n_1$人の被験者の結果変数$Y_{11}, \dots, Y_{n_1 1}, Y_{12}, \dots, Y_{n_1 2}$より、

$$Z_1 = \frac{1}{\sqrt{2n_1}}\sum_{i=1}^{n_1}(Y_{i1} - Y_{i2}).$$

となる。最終解析時の検定統計量Z_{final}も同様に、

$$Z_{final} = \frac{1}{\sqrt{2N_{GSD}}}\sum_{i=1}^{N_{GSD}}(Y_{i1} - Y_{i2}).$$

となる。

中間解析時の有効性中止基準に使用する棄却限界値をc_1、最終解析時をc_2とおくと、代表的な Pocock (1977) [72]、O'Brien and Felming (1979) [73]、Wang and Tsiatis (1987) [76]の有効性中止境界は、それぞれ以下のように表わされる。

$$c_1 = \begin{cases} c_{\mathrm{P}}(\alpha, k_2) \\ c_{\mathrm{OF}}(\alpha, k_2) \\ c_{\mathrm{WT}}(\alpha, k_2, \rho) \end{cases}, c_2 = \begin{cases} c_{\mathrm{P}}(\alpha, k_2) \\ c_{\mathrm{OF}}(\alpha, k_2){k_2}^{-1/2} \\ c_{\mathrm{WT}}(\alpha, k_2, \rho){k_2}^{\rho-1/2} \end{cases},$$

$k_2 = 1/t$とし、$c_{\mathrm{P}}(\alpha, k_2), c_{\mathrm{OF}}(\alpha, k_2), c_{\mathrm{WT}}(\alpha, k_2, \rho)$は各境界の形状に応じて定められる定数であり、試験全体の第一種の過誤確率が名義水準片側αに一致するように求められる。試験全体の第一種の過誤確率Aは、

$$A = \int_{c_1}^{\infty} \varphi(Z_1)\, dZ_1 + \int_{c_2}^{\infty}\int_{0}^{c_1} \varphi(Z_1)\, f\left(Z_{final}\middle|Z_1\right) \partial Z_1 \partial Z_{final},$$

$\varphi(\cdot)$は標準正規分布の確率密度関数、$f\left(Z_{final}\middle|Z_1\right)$は$Z_1$で条件付けた$Z_{final}$の確率密度関数を表す。Wang and Tsiatis の境界とは、ρが$0 \le \rho \le 0.5$の値をとる 1 パラメータ境界族をさし、その特殊な場合に$\rho = 0.5$の O'Brien and Felming の境界、$\rho = 0$の Pocock の境界を含む。4 章のシミュレーション研究では、実際の中間解析を伴う検証的臨床試験でよく使われる O'Brien and Fleming の境界を用いることとする。

中間解析結果に基づく試験中止・継続の判断は、

$$\begin{cases} \text{無効中止}, \ if\ Z_1 \le 0 \\ \text{有効中止}, \ if\ Z_1 \ge c_1 \\ \text{試験継続}, \ if\ 0 < Z_1 < c_1 \end{cases}. \quad (2)$$

よって、試験終了時の被験者数をN_{final}とおくと、

$$N_{final} = \begin{cases} n_1, \ if\ Z_1 \le 0 \text{ or } Z_1 \ge c_1 \\ N_{GSD}, \ if\ 0 < Z_1 < c_1 \end{cases}. \quad (3)$$

となる。また、中間解析又は最終解析において、$Z_1 \ge c_1$ or $Z_{final} \ge c_2$ならば、被験薬の有効性ありと判断される。

３章　被験者数再設定

3.1 被験者数再設定を伴う検証的臨床試験

想定する検証的臨床試験は、基本的には 2.1 節に示したとおりで、1 回の中間解析及び最終解析を実施することとする。ただし、被験者数再設定を想定しているため、初期被験者数は多めに設定することなく、(1)式の$N_{initial}$とする。また、それに基づき第一ステージの被験者数は、$n_1 = tN_{initial}(0 < t < 1)$とすると、予定する第二ステージの被験者数は、$n_2 = (1-t)N_{initial}$となる。

第一ステージn_1人のデータに基づき推定された試験治療群間差を$\hat{\delta}_1$とし、中間解析を実施した結果(2)式に従い試験継続と判断されたとする。このとき例えば、群間差$\hat{\delta}_1$が試験開始前に想定したδ_{pre}よりもいくらか小さく、このままの状況で試験継続し予定した試験終了時点の解析を行っても統計的に有意になりにくいような状況が生じたとする。そのような状況下では、試験開始前に仮定したδ_{pre}が楽観的すぎたと考えられるが、それでもなお被験薬の効果サイズは依然として臨床的に意味があると判断されたとする。このような場合に、目標検出力を達成できるよう被験者数を調整した結果、予定していた被験者数$N_{initial}$よりも多くの最終被験者数$N_{re-estimated}(> N_{initial})$まで総被験者数を増加させたとする。ただし、このときの被験者数再設定基準に関する詳細は次章で説明する。

$N_{re-estimated}$人の被験者データに基づく、第一種の過誤確率の制御を行わない固定デザインで用いられる検定統計量Z_{naive}は以下のようになる。

$$Z_{naive} = \frac{1}{\sqrt{2N_{re-estimated}}}\sum_{i=1}^{N_{re-estimated}}(Y_{i1} - Y_{i2}). \quad (4)$$

被験者数は、通常の固定デザインにおける最終解析では定数として扱われるが、この最終被験者数$N_{re-estimated}$は、第一ステージより求めた治療群間差$\hat{\delta}_1$と相関のある確率変数として扱われる。従って、Z_{naive}の分布はもはや標準正規分布に従わず、第一種の過誤確率が名義水準以下になることは保証されない。Cui, Hung and Wang (1999)[63]は、3.3 節に示す delta-replacement 基準による被験者数再設定を行った場合、名義水準に対し 30%から 40%の第一種の過誤確率の増大が生じることを示している。また、第一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二倍以上にまで及ぶことが示されている（Proschan and Hunsberger, 1995[102]; Shun et al., 2001[103]）。

3.2 第一種の過誤確率の制御

本研究では、1.3.2.1 節で提案された手法のうち、群逐次デザインの枠組みで用いることのできる Cui, Hung and Wang (1999)[63]の重み付 Z 検定を用いた第一種の過誤確率の制御を行う。被験者数再設定により設定された被験者数のうち、第二ステージの被験者数$n_2^* = N_{re-estimated} - n_1$に基づく検定統計量を$Z_2^*$とおくと、

$$Z_2^* = \frac{1}{\sqrt{2n_2^*}} \sum_{i=n_1+1}^{N_{re-estimated}} (Y_{i1} - Y_{i2}).$$

となる。(4)式のZ_{naive}は、以下のようにZ_1とZ_2^*を用いた重み付き和で表すことができる。

$$Z_{naive} = \sqrt{n_1/N_{re-estimated}}\, Z_1 + \sqrt{n_2^*/N_{re-estimated}}\, Z_2^*. \quad (5)$$

Z_{naive}は、第一ステージと第二ステージの統合方法を意味する重み$\sqrt{n_1/N_{re-estimated}}$及び$\sqrt{n_2^*/N_{re-estimated}}$が、第一ステージに得られた結果であるに依存してしまうため、第一種の過誤確率が保たれない。そこで、統合検定の原理に従い、第一ステージの結果と独立な、試験開始前に設定した被験者数に基づく重み$\sqrt{t}, \sqrt{1-t}$を用いることを考える。

$$Z_{weighted} = \sqrt{t} Z_1 + \sqrt{1-t} Z_2^*. \quad (6)$$

この$Z_{weighted}$に基づく検定が Cui, Hung and Wang (1999)の重み付 Z 検定である。帰無仮説$H_0: \delta = 0$のもとでは、Z_1及びZ_2^*は互いに独立に標準正規分布に従うことより、$Z_{weighted}$の従う分布も標準正規分布となる。よって、第一種の過誤確率は、正確に名義水準に一致する。なお、試験終了時の被験者数N_{final}は、

$$N_{final} = \begin{cases} n_1, & if\ Z_1 \le 0 \text{ or } Z_1 \ge c_1 \\ N_{re-estimated}, & if\ 0 < Z_1 < c_1 \end{cases},$$

となる。また、中間解析又は最終解析において、$Z_1 \ge c_1$ or $Z_{weighted} \ge c_2$ならば、被験薬の有効性ありと判断される。

3.3 既存の被験者数再設定基準

被験者数再設定基準とは、中間解析結果に基づき第二ステージの被験者数を何例にすれば良いかの指針及び根拠となるものである。既に提案されている被験者数再設定基準は、基本的に中間解析時に観察された試験治療群間差の点推定値を真値とみなしたもとで構成されたものである。それらには、Cui, Hung and Wang (1999)[63]の delta-replacement 基準（3.3.1 節）と Proschan and Hunsberger (1995)[102]の条件付検出力基準（3.3.2 節）があり、3.3.3 節で

その問題点に触れる。3.3.4 節で、統計的モニタリングの場面で提案された概念であるベイズ流予測検出力について述べ、3.3.5 節で Wang (2006) [127]の無情報事前分布に基づくベイズ流予測検出力基準を説明し、提案する予測検出力基準は 3.4 節に示す。

3.3.1 delta-replacement 基準

Cui, Hung and Wang (1999) [63]の提案した delta-replacement 基準は、予定していた第二ステージ被験者数$n_2 = (1-t)N_0$を以下のように再設定するというものである。

$$n_2^* = \left(\frac{\delta_{pre}}{\hat{\delta}_1}\right)^2 N_{initial} - n_1, \tag{7}$$

n_2^* は、被験者数再設定基準により算出された第二ステージ必要被験者数を表すとする。(7)式右辺第一項は、総被験者数の再推定値を意味し、(1)式で表わされる初期被験者数設定基準に対し、試験開始前における治療群間差の初期推定値δ_{pre}を、中間解析時点の不偏推定値$\hat{\delta}_1$で置換することによって導出される。delta-replacement 基準は、簡便であるが、第一ステージの結果を初期被験者数設定における仮定を修正するためにしか用いていない。つまりそれは、既に試験開始前の時点で評価された検出力が目標水準$1-\beta$に一致するような総被験者数を再推定する基準といえる。従って、最終解析の検定統計量の一部として使用されるだろう第一ステージの結果が、検出力へ与える影響は考慮されていない。第一ステージの結果の影響を考慮した検出力のことを条件付検出力と呼び、次節のものはこれに基づく被験者数再設定基準である。

3.3.2 条件付検出力基準

条件付検出力とは、第一ステージの結果を与えたもとでの検出力のことである。(5), (6)式のように、最終解析の検定統計量は、第一ステージの検定統計量Z_1と第二ステージの検定統計量Z_2との重み付和で表わされる。条件付検出力の“条件付”は、Z_1は実際に観察された値とし、Z_2のみを確率変数として扱うことを意味し、そのもとでZ_1, Z_2の重み付和で表わされる最終解析時の検定統計量が棄却限界値c_2を超える確率を計算したものが条件付検出力である。被験者数を再設定した場合に用いられる、(6)式に示す最終解析の重み付検

定で有意差が観察される確率を規定する要因には、第一ステージの結果であるZ_1、その重みを決める中間解析時点の情報量時間t、第二ステージの情報量を表すn_2^*、そして第二ステージの真の治療群間差δ_2がある。通常、$\delta_2 = \hat{\delta}_1$を仮定する。条件付検出力を$CP\left(Z_1, t, n_2^*;\ \delta_2 = \hat{\delta}_1\right)$とすると、

$$CP\left(Z_1, t, n_2^*;\ \delta_2 = \hat{\delta}_1\right) = Pr\left(Z_w > c_2 \middle| Z_1, t, n_2^*;\ \delta_2 = \hat{\delta}_1\right)$$

$$= 1 - \Phi\left(\frac{c_2 - \sqrt{t}Z_1}{\sqrt{1-t}} - \hat{\delta}_1\sqrt{\frac{n_2^*}{2}}\right). \quad (8)$$

条件付検出力が目標水準$1-\beta$に一致するような推定第二ステージ被験者数n_2^*は、$CP\left(Z_1, t, n_2^*;\ \delta_2 = \hat{\delta}_1\right) = 1-\beta$を解き、

$$n_2^* = \frac{2}{\hat{\delta}_1{}^2}\left(\frac{c_2 - \sqrt{t}Z_1}{\sqrt{1-t}} + z_\beta\right)^2. \quad (9)$$

となる。(9)式に表わされる必要第二被験者数の推定方法のことを条件付検出力基準と定義する。

ここで、Proschan and Hunsberger (1995)[102)]にて、第一種の過誤確率を制御するために提唱された概念である条件付過誤関数$A(Z_1)$との関連で、条件付検出力基準を考えてみる。条件付過誤関数$A(Z_1)$の定義は、第一ステージの結果Z_1を与えたもとで、帰無仮説のもとで誤って有意差が観察される確率である。Cui, Hung, Wang (1999)[63)]の重み付 Z 統計量Z_wを用いた第一種の過誤確率の制御は、以下で表わされる条件付過誤関数による第一種の過誤確率の制御と同等である[102, 110)]。

$$A(Z_1) = 1 - \Phi\left(\frac{c_2 - \sqrt{t}Z_1}{\sqrt{1-t}}\right).$$

となる。$z_{A(Z_1)}$はこれまでの表記法にならい、標準正規分布の$\{1 - A(Z_1)\}$th分位点を表すとすると、

$$\frac{c_2 - \sqrt{t}Z_1}{\sqrt{1-t}} = \Phi\left(1 - A(Z_1)\right) = z_{A(Z_1)}.$$

と表記できる。これを(9)式に代入すると、

$$n_2^* = 2\left(\frac{z_{A(Z_1)} + z_\beta}{\hat{\delta}_1}\right)^2. \quad (10)$$

となる。これは、(1)式の初期被験者数設定基準とよく似た形をしており、第二ステージの治療群間差に関する仮定$\delta_2 = \hat{\delta}_1$のもとで、第一種の過誤確率の名義水準$\alpha$の代わりに、条

件付過誤関数$A(Z_1)$で置換したものとなっていることがわかる。つまり、第一種の過誤確率に関し検出力と同様に条件付推測を行ったもとで、自然に導かれる被験者数推定方法といえる。

3.3.3 条件付検出力の不確実性

delta-replacement 基準及び条件付検出力基準は、中間解析で得られた推定値$\hat{\delta}_1$が治療群間差の真値に一致すると仮定したもとで必要被験者数を推定する方法である。しかし、中間解析データの誤差的変動の影響により、$\hat{\delta}_1$が真の群間差より乖離していた場合には、被験者数推定を誤る危険性が存在する。この節では条件付検出力の不確実性について説明する。

Bauer, Koenig (2006) [128]は、第一ステージに集積するデータを確率変数Z_1で表わしたとき、Z_1の関数という意味において条件付検出力を確率変数として扱い、その不確実性について論じている。仮に試験開始前に初期被験者数を正確に推定できた場合、その時点で評価される検出力は、目標水準$1-\beta$である。しかし、第一ステージデータの誤差的変動により、中間解析時点で評価される条件付検出力は、必ずしも$1-\beta$ に一致しない。このことを詳細に検討するため、以下で確率変数として扱った条件付検出力の分布を示す。

簡単のため、第二ステージの被験者数の再設定は行わない状況$n_2^* = n_2$で、(8)式の条件付検出力$CP_{\delta_2=\hat{\delta}_1}$の確率密度関数$f_{\delta_1}\left(CP_{\delta_2=\hat{\delta}_1}\right)$は以下のようになる。

$$f_{\delta_1}\left(CP_{\delta_2=\hat{\delta}_1}\right) = \frac{\sqrt{1-t}}{\sqrt{t}}\varphi\left(\frac{c_2}{\sqrt{t}} - z_{CP_{\delta_2=\hat{\delta}_1}}\frac{\sqrt{1-t}}{\sqrt{t}} - \hat{\delta}_1\sqrt{\frac{n_2(1-t)}{2t}} - \delta_1\sqrt{\frac{n_1}{2}}\right)\frac{1}{\varphi\left(z_{CP_{\delta_2=\hat{\delta}_1}}\right)},$$

$\varphi(\cdot)$は、標準正規分布の確率密度関数を表し、$z_{CP_{\delta_2=\hat{\delta}_1}}$は標準正規分布の$\left\{1-CP_{\delta_2=\hat{\delta}_1}\right\}$th分位点を表す。同様に、第二ステージの治療群間差の真値δ_2を既知とした場合、条件付検出力の確率密度関数$f_{\delta_1}\left(CP_{\delta_2}\right)$は以下のようになる。

$$f_{\delta_1}\left(CP_{\delta_2}\right) = \frac{\sqrt{1-t}}{\sqrt{t}}\varphi\left(\frac{c_2}{\sqrt{t}} - z_{CP_{\delta_2}}\frac{\sqrt{1-t}}{\sqrt{t}} - \delta_2\sqrt{\frac{n_2(1-t)}{2t}} - \delta_1\sqrt{\frac{n_1}{2}}\right)\frac{1}{\varphi\left(z_{CP_{\delta_2}}\right)},$$

$z_{CP_{\delta_2}}$は標準正規分布の第$\left(1-CP_{\delta_2}\right)th$分位点を表す。

図 1 は、$f_{\delta_1}\left(CP_{\delta_2=\hat{\delta}_1}\right)$及び$f_{\delta_1}\left(CP_{\delta_2}\right)$を、異なる中間解析時期ごとに示したものである（$t = 0.1, 0.2, 0.5, 0.8$）。ただし、治療群間差は、$\delta_1 = \delta_2 = \sqrt{2}$となる。また、試験開始前に

初期被験者数$N_{initial} = n_1 + n_2$を正確に推定できた状況を想定しており、その時点で評価される包括的な検出力は、目標水準$1 - \beta = 0.8$に一致する。よって、条件付検出力CP_{δ_2}（実線）の分布の期待値は、$E_{Z_1}\left(CP_{\delta_2}\right) = 0.8$となる。しかし、試験の初期（$t = 0.1$）以外では、$CP_{\delta_2} = 0.6$くらいの値は十分取り得る。さらに、破線で表わされている (8)式の条件付検出力$CP_{\delta_2=\hat{\delta}_1}$は、常にU字型に広く分布している上、試験の中期くらいまでは$(t = 0.1, 0.2, 0.5)$、実線で表わされる真値CP_{δ_2}より乖離してしまう確率は小さくない。このことより、$CP_{\delta_2=\hat{\delta}_1}$を用いる条件付検出力基準は、被験者数の再推定値が不安定となり、必要被験者数の再設定値に誤りを生じる危険性をもつことが示唆される。

3.3.4 ベイズ流予測検出力

ベイズ流予測検出力とは、条件付検出力の第二ステージ治療群間差δ_2に関する不確実性を考慮したものである。ここでは、元々統計的モニタリングツールとして提案された両指標の間の関係性に着目した上、ベイズ流予測検出力を説明する。

長期の臨床試験の途中で中間解析結果が得られた際に、共同研究グループは試験の早期中止が適切かどうかを判断するのに多くの要因を考慮する。Halperin et al. (1982) [59] は、この難しい判断の助けとなる統計的なツールを、“Stochastically Curtailed Tests”という枠組みに基づき提案した。特に、彼らは試験が予定通り継続された場合に帰無仮説が棄却される2種類の条件付確率を推定することを提唱した。1つ目の確率は、治療効果が存在しないという帰無仮説のもとで、第一ステージデータの条件付確率で、もう一つは試験開始前に期待された治療群間差δ_{pre}が真であるという対立仮説の下での条件付確率である。これら二種類の確率は、第二ステージ群間差δ_2の関数である条件付検出力関数CP_{δ_2}のいろいろな値の一部とみることができる。Andersen (1987) [60]で生存時間への拡張もなされている。

Halperin et al. (1982) [59] の条件付検出力を用いたモニタリングアプローチに対し、より直感的に解釈のしやすいベイズ流予測検出力アプローチが、Spiegelhalter, Freedman and Blackburn (1986) [79]によって提案された。予測検出力とは、条件付検出力の“条件付き”解析を、試験が継続した場合の結果に対する“無条件付き”予測へ拡張したものだということを強調している。つまり、二点の仮説$\delta_2 = 0, \delta_{pre}$のもとでしか評価されない条件付検出

力よりも、あり得る仮説δ_2の範囲に対し条件付検出力関数CP_{δ_2}をプロットしたほうが、試験継続の是非を判断するのに役立つだろう。その要約指標としてCP_{δ_2}を平均したものが予測検出力であると解釈することができる。予測検出力$PP\ (Predicitive\ Power)$は、ベイズの定理より得られる中間解析時点でのδ_2の事後分布$p(\delta_2|Z_1)$を重みとする以下の重み付き平均により計算される。

$$PP = \int CP_{\delta_2}\, p(\delta_2|Z_1) d\delta_2.$$

PPの別の見方として、第二ステージの予測分布に基づく条件付検出力という見方ができる。以下、第二ステージの予測分布及びそれに基づく推測の結果得られる統計量を添え字のpを用いて区別する。第二ステージの治療群間差$\hat{\delta}_2$の予測確率密度関数$f_p(\hat{\delta}_2|Z_1)$は以下のように計算される。

$$f_p(\hat{\delta}_2|Z_1) = \int f(\hat{\delta}_2;\, \delta_2) p(\delta_2|Z_1) d\delta_2. \tag{12}$$

$f_p(\hat{\delta}_2|Z_1)$の期待値δ_pと分散σ_p^2を、第二ステージの群間差$\hat{\delta}_2$の予測平均、予測分散とすると、これらに基づく条件付検出力CP_pは、以下のようになり、

$$CP_p = 1 - \Phi\left(\frac{c_2 - \sqrt{t}Z_1}{\sqrt{1-t}} - \frac{\delta_p}{\sigma_p}\right), \tag{13}$$

$CP_p = PP$が成り立つことが示されている。

3.3.5 無情報事前分布に基づく予測検出力基準

Wang (2006) [127]は、被験者数再設定基準に関してはベイズ流プローチをとり、最終解析の検定手法つまり第一種の過誤確率の制御に関しては既存の頻度論的アプローチをとることを提案している。頻度論と Bayesian の相反する統計学上の理論的立場に対し、彼は、一つの臨床試験においてそれぞれの立場をデザインと解析方法という異なる場面で利用するという折衷案的アプローチを“semi”-Bayesian と呼んでいる。モニタリングの場面では、ベイズ流予測検出力を試験の中止・継続の判断に用いることを想定しているため、本試験のデータのみに基づく予測となるよう、δの事前分布に無情報事前分布を考えることが一般に支持される（Lindley, 1970[129]）。これに従い、Wang (2006) [127]は、無情報事前分布に基づく予測検出力基準を提案している。

試験治療群間差δの事前分布$p(\delta)$は、以下のように事前平均をδ_0、事前分散をσ_0^2とした

正規分布に従うものとする。

$$p(\delta) \sim N(\delta_0, \sigma_0^2).$$

第一ステージの結果Z_1を与えたもとでの治療群間差δの事後分布$p(\delta_2|Z_1)$は、

$$p(\delta_2|Z_1) = \frac{L(\delta_2|Z_1)p(\delta_2)d\delta_2}{\int L(\delta_2|Z_1)p(\delta_2)d\delta_2}$$

$$\sim N\left(\frac{\frac{\delta_0}{\sigma_0^2}+\frac{n_1}{2}\hat{\delta}_1}{\frac{1}{\sigma_0^2}+\frac{n_1}{2}}, \frac{1}{\frac{1}{\sigma_0^2}+\frac{n_1}{2}}\right). \quad (14)$$

δの無情報事前分布$\sigma_0^2 \sim \infty$より、

$$p(\delta_2|Z_1) \sim N\left(\hat{\delta}_1, \frac{2}{n_1}\right).$$

また、$f(\hat{\delta}_2;\ \delta_2) \sim N\left(\delta_2, \frac{2}{n_2}\right)$より、(12)式と同様に積分を行うと、第二ステージの予測分布$f_p(\hat{\delta}_2|Z_1)$は、

$$f_p(\hat{\delta}_2|Z_1) \sim N\left(\hat{\delta}_1, \frac{2}{n_1}+\frac{2}{n_2}\right). \quad (15)$$

となる。よって、(13), (15)式より、第二ステージの必要被験者数推定値は、

$$n_2^* = \left[\frac{1}{2}\left(\frac{\hat{\delta}_1}{z_{A(Z_1)}+z_\beta}\right)^2 - \frac{1}{n_1}\right]^{-1}. \quad (16)$$

となる。これを無情報事前分布に基づく予測検出力基準と定義する。

3.4 提案する予測検出力基準

3.3.5 節に示した無情報事前分布に基づくベイズ流予測検出力基準は、第二ステージの治療群間差δ_2の不確実性は考慮されているが、第二ステージの治療群間差の予測値は、$\delta_p = \hat{\delta}_1$であり、第一ステージデータの誤差的変動による予測の不安定さは改善されていない。そこで、第二ステージの予測の安定化のために、治療群間差に関する事前情報を用いるアプローチを提案する。

一般に、試験開始前に被験者数設定を行う際、試験治療群間差δに対し、確信をもって$\delta = \delta_{pre}$の一点に特定できるとは限らない。むしろ、δに関する取り得る複数のシナリオのなかから、開発費用、開発期間、患者や医師の試験に対するインセンティブ等を考慮した上で、現実的に妥当な、あるいは妥当な範囲内で最も保守的な、すなわち群間差を小さく

見積もった結果、$\delta = \delta_{pre}$を設定する状況が多いだろう。その際、設定する被験者数の不確実性を最小限にする努力として、治療レジメン、併用薬剤、剤型、試験規模、疾患領域が異なっているかもしれないが、同被験薬または類似薬の国内外における過去の臨床試験成績が収集される。それらの参照可能な情報量や被験者背景及び試験環境、試験目的自体の類似性に応じ、群間差δのとり得そうな範囲 *Prior Range of* δ (以下、*PR*) を以下のように特定することは可能であると考えられる。

$$PR = [\delta_{lower}, \delta_{upper}],\ \delta_{lower} < \delta_{upper}.$$

例えば、通常の初期被験者数設定で試験開始前に検討したδに関する取り得る複数のシナリオのうち、最も悲観的なシナリオをδ_{lower}、最も楽観的なシナリオをδ_{upper}と設定する場合が考えられる。このような場合には、$\delta_{lower} < \delta_{pre} < \delta_{upper}$となるだろう。他の例としては、臨床的に意味のある最小の治療群間差をδ_{lower}、初期被験者数を設定する上で実施可能性を失わない範囲で最も保守的な（小さい）治療群間差をδ_{upper}として設定する場合、$\delta_{lower} < \delta_{pre} \approx \delta_{upper}$となるだろう。また、これらの組み合わせもあるだろう。いずれにしても、以下のようにその範囲の間に初期被験者数設定に用いる群間差δ_{pre}は含まれることは確かである。

$$\delta_{pre} \in [\delta_{lower}, \delta_{upper}].$$

実際には、δ_{pre}に基づき初期被験者数を設定して試験を開始せざるを得ないが、設定した被験者数には、前述のPRに相当する不確実性を依然として持つものと考えることは妥当である。中間解析結果に基づき被験者数再設定を実施することを考慮した試験デザインでは、試験開始前に設定した初期被験者数が、どの程度の不確実性を含んだものであったかを考慮することには意味がある。本研究では、治療群間差のPRに基づく事前分布の設定方法及び、それに基づく被験者数再設定基準を提案し、それに基づき治療群間差の不確実性へ対処することを提唱したい。

δ_{pre}及び$PR = [\delta_{lower}, \delta_{upper}]$を用いた事前分布の設定方法には、さまざまなアプローチが考えられるが、妥当と思われる設定方法を二通り提案する。前節までと同様、事前分布を正規分布$p(\delta) \sim N(\delta_0, \sigma_0^2)$で表わし、ベイズ流予測検出力を用いた事前情報の考慮の仕方を考える。一つ目の事前分布の設定方法は、

$$\delta_0 = \delta_{pre},$$

$$\sigma_0 = \frac{\left|\delta_{upper} - \delta_{lower}\right|}{2z_{u/2}}.$$

σ_0の設定方法は、事前分布の$100u\%$被覆区間幅がPR幅である$\left|\delta_{upper} - \delta_{lower}\right|$に一致するように設定するものである。本論文では、$u = 2\alpha$とする。事前分散$\sigma_0^2$の大きさは、その逆数が事前情報量に相当または比例すると考えられるので、PR幅の広さに比例してσ_0を設定することは自然である。

次に二つ目の事前分布の設定方法を説明する。(14)式をみると、第一ステージの結果Z_1を与えたもとでの治療群間差δの事後分布$p(\delta_2|Z_1)$の事後平均は、事前平均δ_0と中間解析時点の点推定値$\hat{\delta}_1$に対し、各分散の逆数を重みとした重み付平均となっていることがわかる。よって、事前平均δ_0の重みをデータに基づき、以下のように決定する方法が考えられる。

$$\delta_0 = \delta_{pre},$$

$$\sigma_0 = \left|\hat{\delta}_1 - \delta_{pre}\right|.$$

この事前分散の設定方法に従えば、事前平均δ_0にかかる重みを次のように考慮することが可能である。例えば、中間解析を実施した結果、治療群間差のPRを大きく外れた$\hat{\delta}_1$が得られた場合、データを尤もらしいとすれば、事前平均δ_0にかかる重みを小さくしたいと考えるのは自然である。この場合、$\left|\hat{\delta}_1 - \delta_{pre}\right|$が大きい状況なので、それに応じ重み$1/\sigma_0^2$は減少する。反対に、$\delta_{pre}$の近傍に$\hat{\delta}_1$が得られた場合を考えると、事前に設定した$\delta_{pre}$は尤もらしかったという意味より、重みを大きく設定することが考えられるが、$\left|\hat{\delta}_1 - \delta_{pre}\right|$が小さい状況なので、それに応じ重み$1/\sigma_0^2$は増加する。近傍な値間の重み付平均値は、そもそも重みの影響をあまり受けないという性質を考慮すると、いずれの状況においてもデータ$\hat{\delta}_1$と事後平均がかけ離れることは比較的生じにくい。このような事前情報の設定方法は、基本的にはデータを尤もらしいとする立場をとっていると言える。

一つ目の事前分布の設定方法に基づく被験者数再設定のための予測検出力は、前節と同様に計算していくと以下のようになる。

$$n_2^* = \left[\frac{1}{2} \left\{ \frac{\delta_{pre} / \left(\frac{|\delta_{upper} - \delta_{lower}|}{2z_\alpha} \right)^2 + \frac{n_1}{2} \hat{\delta}_1}{1 / \left(\frac{|\delta_{upper} - \delta_{lower}|}{2z_\alpha} \right)^2 + n_1/2} \Big/ \left(z_{A(Z_1)} + z_\beta \right) \right\}^2 - 1 \Big/ \left\{ n_1 + 1/2 \left(\frac{|\delta_{upper} - \delta_{lower}|}{2z_\alpha} \right)^2 \right\} \right]^{-1}. \quad (17)$$

二つ目の事前分布の設定方法に基づく被験者数再設定のための予測検出力は、

$$n_2^* = \left[\frac{1}{2} \left\{ \frac{\delta_{pre} / |\hat{\delta}_1 - \delta_{pre}|^2 + \frac{n_1}{2} \hat{\delta}_1}{1 / |\hat{\delta}_1 - \delta_{pre}|^2 + n_1/2} \Big/ \left(z_{A(Z_1)} + z_\beta \right) \right\}^2 - 1 \Big/ \left\{ n_1 + 1/2 |\hat{\delta}_1 - \delta_{pre}|^2 \right\} \right]^{-1}. \quad (18)$$

となる。

４章　シミュレーション研究

本章では、被験者数設定の不確実性へ対処可能なデザインとして 2 章で述べた群逐次デザイン及び 3 章で述べた提案法を含むいくつかの被験者数再設定基準を用いた被験者数再設定の性能及び性質を比較・検討するシミュレーション研究を行う。4.1 節でシミュレーションデータを発生させるために想定する試験設定を示し、4.2 節では、発生させた仮想試験データに対し各種試験デザインを適用した結果に対し、被験者数設定の不確実性へ対処可能なデザインとしての性能をどのような指標を用いて評価すべきかを論じる。4.3 節で結果を述べる。

4.1 想定する試験設定

4.1.1 全般的な試験環境設定

各試験デザインタイプの性能及び性質を議論する目的においては、想定する試験は、より一般性の高い示唆が得られるよう単純な状況を想定すべきである。一方、実際に個々の臨床試験デザインを決定する目的で試験シミュレーションを実施する際には、可能な限り実際に即した複雑なシナリオ設定をすべきであろう。

被験薬の対照薬に対する優越性を検証することを目的としたランダム化二群比較試験を想定する。試験治療群$j(j = 1$: 被験薬群, 2: 対照薬群)に割り付けられた被験者iの結果変数Y_{ij}は、平均がそれぞれ$\mu_1 = \delta, \mu_2 = 0$、共通の群内分散$\sigma^2 = 1$に従う正規乱数より発生させた。臨床的に意味のある最小の治療群間差を$\delta_{min} = 0.15$と設定し、群間差パラメータδは$\delta = 0.15, 0.17, \ldots, 0.33, 0.35$と 0.02 刻みで設定した。試験開始前に得られる事前情報より、本仮想試験の真の治療群間差の取りうる範囲は、$PR = [\delta_{lower} = 0.2, \delta_{upper} = 0.3]$と絞り込むことができると想定した。以上に設定した試験パラメータは、設定の仕方を変更しても特に一般性を失うことはなく、各設定の絶対値に意味はない。一方、以下で設定する試験パラメータは、上記パラメータ設定値に対する相対値に意味をもち、試験デザイン間の性能比較に一定の影響を与える試験設定の違いを表す。従って、現実にあり得そうな一般的な試験環境のシナリオをうまく代表するよう設定を行う必要がある。

初期被験者数は、従来のデザインでは当然のこと、被験者数を再設定する場合でも、それを基準に何倍に再設定するかということになるので、最終的な検出力を決定づける大き

な要因となるパラメータである。初期被験者数設定に関する一般的試験環境としては、多めに設定する場面と少なめに設定する場面と二つのパターンが一般に考えられる。丁度良いという状況は、シミュレーション研究の目的上考えないこととする。上記二つを代表するように、初期被験者数設定に用いる群間差の事前推定値δ_{pre}は、$PR = [\delta_{lower} = 0.2, \delta_{upper} = 0.3]$に対し、$\delta_{pre} = 0.225, 0.275$の二通りを設定した。これに基づき(1)式より、初期被験者数は$N_{initial} = 310, 208$と設定した。最大被験者数N_{max}は、被験者数の上限を決める試験パラメータであり、実際の試験環境において潜在的には常に存在する。N_{max}を決定する要素には、当試験の出資者が提供可能な予算の上限、被験者数登録の進捗具合、一被験者の結果変数が観察されるまでに要する平均観察期間、全試験期間の上限、被験者の潜在的母集団等が考えられる。これらの要素からN_{max}を設定する実際上のやり方を一般的に定式化することは不可能であるため、本研究では便宜上、臨床的に意味のある最小の治療群間差$\delta_{min} = 0.15$に基づき(1)式に従って$N_{max} = 698$と設定した。同様に、被験者数の再設定値の下限を決める最小被験者数N_{min}は、$\delta_{upper} = 0.3$に基づき$N_{min} = 174$と設定した。ただし、設定によっては$n_1 > 174$となり、その場合は$N_{min} = n_1$とした。中間解析の実施時期は、$N_{initial}$に対する第一ステージ被験者数の分数を意味する情報量時間を用い、$t = 0.25, 0.5, 0.75$と設定した。

4.1.2 群逐次デザインに関する設定

群逐次デザインは、治療群間差が予想よりも小さな場合に備えて初期被験者数を多めに設定しておき、治療群間差が予想よりも大きな場合に備えて中間解析を実施するというデザインである。ある一点の治療群間差のもとでは、設定すべき初期被験者数の最適値を理論的に求めることは簡単である。しかし、本シミュレーション研究の目的は、試験治療群間差の真値に関する一定の不確実性$\delta \in PR$のもとで生じ得る従来の被験者数設定の脆弱性に対し、試験デザイン上の頑健性をどの程度向上させることができるかを評価することにあるため、ある一点の治療群間差のもとでの群逐次デザインの初期被験者数の最適値を設定することはしない。また、群逐次デザインでは、(3)式より初期被験者数と最大被験者数が等しいため、実際の臨床試験では、どの程度初期被験者数を多めに設定するかは、前述

のN_{max}設定と同様、その一般的定式化を不可能にする不確定要素に左右される。そこで、被験者数再設定との比較可能性を可能な限り担保すべく、$N_{GSD} = N_{initial}, (N_{initial} + N_{max})/2, N_{max}$と三つの代表値を設定し、対応する三パターンの群逐次試験デザインを、順にGSD_S, GSD_M, GSD_Lと表記する。$N_{initial} = 310,208;\ N_{max} = 698$より、$GSD_S, GSD_M, GSD_L$の初期被験者数$N_{GSD}$は、順に$N_{GSD} = 310,504,698;\ 208,453,698$となる。なお、中間解析時及び最終解析の有効性中止境界c_1, c_2は、実際の臨床試験の中間解析で汎用される O'Brien & Fleming 型の境界とし、$t = 0.25$ならば$c_1 = 3.92, c_2 = 1.96$、$t = 0.5$ならば$c_1 = 2.78, c_2 = 1.98$、$t = 0.75$ならば$c_1 = 2.33, c_2 = 2.02$と設定した。早期中止基準、最終被験者数はそれぞれ(2), (3)式に従うとする。

4.1.3 被験者数再設定に関する設定

有効性及び無効性による早期中止基準は、群逐次デザインに準ずる。(7)式の delta-replacement 基準による被験者数再設定デザインを$D_{replace}$、(9)式の条件付検出力基準による被験者数再設定デザインをCP、(16)式の無情報事前分布に基づく予測検出力基準による被験者数再設定デザインを$Noninfo$、(17)式の一つ目に提案する事前分布に基づく予測検出力基準による被験者数再設定デザインを$PP_{info}A$、(18)式の二つ目に提案する事前分布に基づく予測検出力基準による被験者数再設定デザインを$PP_{info}B$と表記する。(9)式の条件付検出力基準で用いられる試験治療群間差の推定値$\hat{\delta}_1$の代わりに、真値δを用いた仮想的被験者数再設定を理想デザインとして、性能参照の基準として設置した。このデザインを$True$と表記する。各種の被験者数再設定基準により推定された第二ステージ被験者数n_2^*に基づき、最終的な総被験者数N_{final}は以下のように設定した。

$$N_{final} = \begin{cases} n_1, & if\ Z_1 \le 0 \text{ or } Z_1 \ge c_1 \\ N_{min}, & if\ 0 < Z_1 < c_1 \text{ and } n_2^* < N_{min} - n_1 \\ n_1 + n_2^*, & if\ 0 < Z_1 < c_1 \text{ and } N_{min} - n_1 \le n_2^* \le N_{max} - n_1 \\ N_{max}, & if\ 0 < Z_1 < c_1 \text{ and } n_2^* > N_{max} - n_1 \end{cases}'$$

最終解析は、(6)式の重み付 Z 検定を行った。

なお、4.1.1 節及び 4.1.2 節で示した試験パラメータ$\delta_{min}, \delta, PR, \delta_{pre}, N_{initial}, N_{GSD}, N_{max}, N_{min}, t$の設定値を表 1 に示す。

4.2 検証的試験デザインの評価基準

被験薬の効果の大きさにある程度の不確実性が存在する場合、検証的臨床試験の計画段階において、従来通り固定デザインを採用して問題がないか、それとも群逐次デザインや被験者数再設定デザインといった Adaptive Sample Size Design を適用すべきか、適用するだけの効用が得られるかを試験実施者及びその統計責任者は判断しなければならない場面がある。そのような場合に、試験デザインのさまざまな側面のうち、統計的評価が可能な側面について詳細に検討することは、大いに判断の助けとなると考えられる。従来デザインの統計的評価は、最終的な総被験者数が固定されているため、(1)式を検出力について解析的に解くだけで評価可能である。しかし、最終的な総被験者数が、中間解析結果次第で決定する Adaptive Sample Size Design を評価する際、検出力が明示的に表現できないため、デザイン選択及びデザインの善し悪しを決定していくための評価指標として、検出力だけでは不十分である。本節では、各デザインのもつ一般的性能・性質の比較・検討という本シミュレーション研究のためだけでなく、実際の検証的臨床試験の計画の助けとなるような試験デザインの評価系として、一連の統計的評価指標を提案する。提案する指標は、既存の指標を含む包括的評価指標（4.2.1 節）と、中間解析結果の不確実性を考慮するための直接評価指標（4.2.2 節）に分類できる。

4.2.1 包括的評価指標

包括的評価指標とは、ある試験パラメータ設定のもとで複数回繰り返し発生させたシミュレーションデータに、毎回各試験デザインを適用した場合に得られる最終的な試験結果の経験分布に基づく統計量である。s回目$(s = 1,2, \dots, S$: 本シミュレーション研究では $\mathrm{S} = 100{,}000)$に発生させたシミュレーションデータに対し、ある試験デザインを適用したもとで、中間解析と最終解析のいずれかにおいて統計的に有意な試験治療群間差が観察された場合は 1 を、観察されなかった場合は 0 をとる指示変数をR_sとすると、検出力$Power$を以下のように定義する。

$$Power = \frac{\sum_{s=1}^{S} R_s}{S}. \tag{19}$$

同様に、期待被験者数ASNは、

$$ASN = \frac{\sum_{s=1}^{S} N_{final\_s}}{S}. \quad (20)$$

N_{final_s}は、第s番目のシミュレーションデータにあるデザインを適用した際の最終総被験者数を表す。1.3.2.2 節で示した通り、Jennison and Turnbull (2003)[114]及び Tsiatis and Mehta (2003) [115]が(6)式の重み付 Z 検定を用いた被験者数再設定は、十分統計量に基づかないことより、群逐次デザインと比較して統計的効率が劣るという指摘をしている。Jennison and Turnbull (2003)[114]のシミュレーション研究では、(19), (20)式と同様に定義された検出力、期待被験者数を対比させることにより、各デザインの統計的効率を議論していた。そこで、本研究では、先行研究での議論と比較できるよう、(19), (20)式の$Power, ASN$を用い、期待被験者数 100 例あたりの検出力$Efficiency_{100}$という指標を以下のように定義する。

$$Efficiency_{100} = \frac{Power}{ASN} \times 100. \quad (21)$$

一般に、検証的臨床試験の主目的である被験薬の対照薬に対する優越性を検証することに関する成功確率を左右する因子の一つである$Power$はできるだけ高く、一方で経済的効用を左右する因子の一つであるASNはできる限り少ない試験デザインが良いデザインと言える。これら二つの指標はトレード・オフの関係にあるので、それを考慮できる統合指標による評価をしたいと考えることは自然だろう。$Efficiency_{100}$も一つの統合指標ではあるが、統計的効率が高いからといって必ずしも目標とする検出力$1-\beta$を達成できているとは限らない。そこで、理想のデザインからどの程度乖離しているかという視点から、$Power, ASN$の二指標を統合した Liu, Shu, Cui (2008)[130]と同様な包括的評価指標を提案する。

検出力と被験者数に関する目標又は理想的水準からの‘乖離’を各々考える。まず、期待不足検出力$EUP(Expected\ UnderPower)$を以下のとおり定義する。

$$EUP = -[Power - (1-\beta)]_{-},$$

$[\cdot]_{-}$は、$[\cdot] < 0$の時のみ値を返し、それ以外は0を返す関数である。次に期待過剰被験者数$EOS(Expected\ OverSize)$を以下のとおり定義する。

$$EOS = \left[ASN - N_{initial\_ideal}(\delta)\right]_{+},$$

$[\cdot]_{+}$は、$[\cdot] > 0$の時のみ値を返し、それ以外は0を返す関数である。(1)式より、

$$N_{initial\_ideal}(\delta) = 2\left(\frac{z_\alpha + z_\beta}{\delta}\right)^2.$$

EUP及びEOSという試験効用に関する二種の損失は、前者は確率、後者は被験者数という

異なる次元で表わされたものなので、統合するために次元を揃える必要がある。そこで、EUPを被験者数の次元に以下のように変換する。

$$N_{EUP} = -\left[N_{Power} - N_{initial\_ideal}(\delta)\right]_{-},$$

N_{Power}は、各デザインで得られる$Power$と同じ検出力を得る固定デザインの被験者数を意味し、以下のように計算する。

$$N_{Power} = 2\left(\frac{z_{\alpha}+z_{1-Power}}{\delta}\right)^{2}.$$

検出力不足を被験者数の次元に変換したN_{EUP}とEOSの統合の仕方は、いろいろ考えられるが、本シミュレーション研究では以下の統合方法を用いる。

検証的試験を実施する上で許容できない不足検出力の基準を決め、前述の通りそれを被験者数の次元に変換したものを基準N_{EUP}とする。検出力 50%を基準と考え、

$$\text{基準}N_{EUP} = \left|N_{0.5} - N_{initial\_ideal}(\delta)\right|,$$

$N_{0.5} = 2\left(\frac{z_{\alpha}+z_{0.5}}{\delta}\right)^{2}$とする。検証的試験なので、検出力の損失を基準とし、それと等価な期待過剰被験者数の基準EOSを決める。真に必要な被験者数の 2 倍を検出力 50%と同等とみなし、

$$\text{基準}EOS = N_{initial\_ideal}(\delta).$$

本シミュレーション研究で提案する包括的統合評価指標を期待後悔度$ER(Expected\ Regret)$と呼び、以下の通り定義する。

$$\text{期待後悔度}ER = \left(\frac{N_{EUP}}{\text{基準}N_{EUP}} + \frac{EOS}{\text{基準}EOS}\right) \times 100(\%), \tag{22}$$

4.2.2 直接評価指標

前節の包括的評価指標は、$S = 100{,}000$個のシミュレーションデータセットに対し、各試験デザインごとに100,000回臨床試験を仮想的に繰り返した際の結果を、(19), (20)式のように平均して得られる指標及びそれらを(21), (22)式のように統合した指標である。ここで、(19)式を以下のように書き換える。

$$Power = \frac{\sum_{s=1}^{S} R_s}{S} = \frac{\sum_{s=1}^{S}(R_s|Z_{1s})}{S} = \frac{\sum_{s=1}^{S}\left(\frac{\sum_{s=1}^{S} R_s}{S}\middle|Z_{1s}\right)}{S} \simeq \frac{\sum_{s=1}^{S} CP\left(Z_{1s},t,n_{2final\_s};\,\delta_2=\delta\right)}{S},$$

第s回目のシミュレーションデータに対する、各デザインの真の群間差のもとで評価した条

件付検出力$CP\left(Z_{1s}, t, n_{2final_s};\ \delta_2 = \delta\right)$は、(8)式と同様にして、

$$CP\left(Z_{1s}, t, n_{2final\_s};\ \delta_2 = \delta\right) = Pr\left(R_s = 1 \middle| Z_{1s}, t, n_{2final\_s};\ \delta_2 = \delta\right)$$

$$= 1 - \Phi\left(\frac{c_2 - \sqrt{t}Z_{1s}}{\sqrt{1-t}} - \delta\sqrt{\frac{n_{2final\_s}}{2}}\right),$$

となり、Z_{1s}は第s回目のシミュレーションデータに対する中間解析時の検定統計量を、n_{2final_s}は、$N_{final_s} - n_1$を表すとする。また、$CP\left(Z_{1s}, t, n_{2final_s};\ \delta_2 = \delta\right)$を$CP_s(\delta)$と以後略記する。よって、各試験デザインにおける中間解析時点の条件付検出力$CP_s(\delta)$は、第s回目のシミュレーションデータに対し、どの被験者数再設定方法が適切か、それとも群逐次デザインの方が適切かを、発生させたデータごとに直接評価した指標である。

試験デザインの平均的性能である 4.2.1 節の包括的評価指標では､その性能のばらつきが評価されない。それに対し直接評価指標$CP_s(\delta)$を評価することの利点は、100,000回臨床試験を仮想的に繰り返した際、発生させたデータごとに各試験デザインの性能がどの程度ばらつくかを考慮できる点にある。ばらつきを考慮できる統計量には、分散、平均二乗誤差、偏差の絶対値がある。$CP_s(\delta)$の目標検出力$1-\beta$との偏差の絶対値を以下のように分解する。

$$|CP_s(\delta) - (1-\beta)| = [CP_s(\delta) - (1-\beta)]_+ - [CP_s(\delta) - (1-\beta)]_-.$$

今、検出力の不足に興味があるので、平均不足検出力$MUP(Mean\ UnderPower)$という直接評価指標を以下のように定義する。

$$MUP = -\frac{\sum_{s=1}^{S}[CP_s(\delta) - (1-\beta)]_-}{S}. \quad (23)$$

同様に被験者数に関する直接指標、平均過剰被験者数$MOS(Mean\ OverSize)$を以下のように定義する。

$$MOS = \frac{\sum_{s=1}^{S}[n_{2final\_s} - n_{2ideal\_s}]_+}{S}, \quad (24)$$

$n_{2ideal_s} = \frac{2}{\delta^2}\left(\frac{c_2 - \sqrt{t}Z_{1s}}{\sqrt{1-t}} + z_\beta\right)^2$とする。ここで、期待後悔度と同様に平均不足検出力$MUP$と平均過剰被験者数$MOS$を統合した指標を考える。第$s$回目のシミュレーションデータにおいて中間解析後の検出力不足を被験者数の次元に変換したN_{UP_s}は、

$$N_{UP\_s} = -\left[\frac{2}{\delta^2}\left(\frac{c_2 - \sqrt{t}Z_{1s}}{\sqrt{1-t}} + z_{1-CP_s(\delta)}\right)^2 - n_{2ideal\_s}\right]_- = -\left[n_{2final\_s} - n_{2ideal\_s}\right]_-.$$

第s回目のシミュレーションデータにおける基準$N_{UP_s}h$は、

$$基準N_{UP\_s} = -\left[\frac{2}{\delta^2}\left(\frac{c_2-\sqrt{t}Z_{1s}}{\sqrt{1-t}} + z_{0.5}\right)^2 - n_{2ideal\_s}\right]_-.$$

第s回目のシミュレーションデータにおける過剰被験者数OS_s及び基準OS_sは、

$$OS_s = \left[n_{2final\_s} - n_{2ideal\_s}\right]_+, 基準OS_s = n_{2ideal\_s}.$$

よって、直接指標である第s回目のシミュレーションデータに対し直接評価した後悔度を平均した、平均後悔度$MR(Mean\ Regret)$を以下のように定義する。

$$MR = \frac{\sum_{s=1}^{S}\left(\frac{N_{UP\_s}}{基準N_{UP\_s}} + \frac{OS_s}{基準OS_s}\right)}{S} \times 100(\%). \quad (25)$$

4.3 結果

4.3.1 検出力

検出力$Power$を比較した結果を図 2 に示す。6 つの試験環境設定（$t = 0.25,0.5,0.75;\ \delta_{pre} = 0.225,0.275$）に対する図が三行二列に配置されており、左列が$\delta_{pre} = 0.225; N_{inititial} = 310$、右列が$\delta_{pre} = 0.275; N_{initial} = 208$、上行が$t = 0.25$、中行が$t = 0.5$、下行が$t = 0.75$となっている。横軸に対する 3 本の参照線のうち、左側がPRの下側境界（$\delta_{lower} = 0.2$）、右側がPRの上側境界（$\delta_{upper} = 0.3$）、間の参照線は$\delta_{pre} = 0.225,0.275$を表す。縦軸に対する参照線は、目標検出力$1 - \beta = 0.8$を表す。

まず、群逐次デザインと被験者数再設定の対比軸で検出力を比較する。被験者数再設定よりも多くの初期被験者数を設定した群逐次デザインGSD_M, GSD_Lは、被験者数再設定よりも一貫して検出力が高く、すべての設定においてPR内のどの群間差に対しても目標検出力を達成していた。一方、被験者数再設定と同じ初期被験者数を設定したGSD_Sは、被験者数再設定よりも一貫して検出力が低く、どの設定においてもPR内で常に目標検出力を達成するということはなく、特に少なめの初期被験者数を設定した右列では、PR内の多くの群間差に対して目標検出力を大きく下回っていた。それに対し、被験者数設定では、すべての設定においてPR内のどの群間差に対しても目標検出力を達成していたのは、無情報事前分布に基づく予測検出力基準$Noninfo$のみではあったものの、初期被験者数を多めに設定した左列であるか、あるいは中間解析の時期がより後期であるほど、PR内のほとんどの群間

差に対し目標検出力を達成できるということがより多くなった。

次に被験者数再設定の各方法間の比較を行う。最も検出力が高かった$Noninfo$の次に検出力が高かった方法は、順に delta-replacement 基準$D_{replace}$、条件付検出力基準CPであったが、両者の検出力はほぼ同等であった。$D_{replace}$及びCPは、ほとんどの設定においてPR内のどの群間差に対しても検出力が目標水準に達していたが、特に早期の中間解析で少なめの初期被験者数という設定（$t = 0.25, \delta_{pre} = 0.275$）では$PR$内のより小さな群間差に対する検出力が目標水準を下回っていた。

提案する予測検出力基準の$PP_{info}A$及び$PP_{info}B$は、被験者数再設定の中で最も検出力が低く、$CP, D_{replace}$と同様、少なめの初期被験者数を設定した右列であるか、あるいは中間解析の時期がより早期であるほど、PR内のより小さな群間差に対する検出力が目標水準を下回ることが多くなった。一方$PP_{info}A$と$PP_{info}B$を比較すると、事前分布の重みをデータに依存させる方法$PP_{info}B$は、PRの幅に基づく重みを常にもつ$PP_{info}A$よりも検出力が高くなる傾向があり、中間解析時期が早期でなければ、初期被験者数を少なめに設定した場合でもPR内のほとんどの群間差に対し検出力が目標水準よりも 5%以上下回ることはなかった。仮想的被験者数再設定デザイン$True$は、群間差の真値を既知とした理想の被験者数再設定である。$True$は、PR内の異なる群間差に対し、検出力の変動がすべてのデザインのなかで最も少なく、目標検出力を大きく下回ることも大きく上回ることもないという意味で、被験者数設定の不確実性に対する頑健性をもつデザインであった。提案する予測検出力基準の$PP_{info}A$及び$PP_{info}B$は、他のデザインより比較的検出力が低い傾向にあったが、全体を通して$True$からの乖離が小さい傾向にあった。

4.3.2 期待被験者数

期待被験者数ASNを比較した結果を図 3 に示す。縦軸に対する二本の参照線のうち、上側は群間差がPRの下側境界（$\delta_{lower} = 0.2$）に等しい場合に固定デザインで目標検出力を丁度達成する初期被験者数（$N_{initial} = 392$）を表し、下側は群間差がPRの上側境界（$\delta_{upper} = 0.3$）に等しい場合に固定デザインで目標検出力を丁度達成する初期被験者数（$N_{initial} = 174$）を表す。図 2 と比較すると、期待被験者数は検出力が高いデザインほど

多く、低いデザインほど少なかった。PR内で各デザインの期待被験者数と固定デザインの被験者数とを比較すると、群逐次デザインGSD_M, GSD_L及び無情報事前分布に基づく予測検出力基準$Noninfo$は、固定デザインの被験者数の上側境界と同等かそれを超える期待被験者数となっていることが示された。提案する予測検出力基準の$PP_{info}A$及び$PP_{info}B$は、検出力の結果と同様、全体を通して$True$からの乖離が小さい傾向にあった。

4.3.3 期待被験者数 100 例あたりの検出力

統計的効率を評価するための指標、期待被験者数 100 例あたりの検出力$Efficiency_{100}$を比較した結果を図 4 に示す。この指標により、前節に述べた検出力と期待被験者数のトレード・オフが、各方法間でどの程度異なるかを正確に検討することが可能である。縦軸に対する二本の参照線は、PR内の群間差の各値に対し固定デザインで目標検出力を丁度達成する初期被験者数を設定したときの$Efficiency_{100}$の最大値と最小値を示したものである。最も統計的効率が高いデザインは$True$であったが、提案する$PP_{info}A$はPR内のほとんどの群間差に対し$True$と同等か若干劣る程度の高い効率で、他のデザインと比較し最も高い効率を示した。無情報事前分布に基づく予測検出力基準$Noninfo$は、被験者数再設定のなかで最も低い効率を示し、また検出力が$Noninfo$よりも高かったGSD_Lよりもその効率は低かった。被験者数再設定よりも多くの初期被験者数を設定した群逐次デザインGSD_M, GSD_Lは、検出力は高かったものの、$Noninfo$以外の被験者数再設定デザインと比較すると、PR内のすべてにおいて効率がより低いことが示された。

4.3.4 期待後悔度

検出力と期待被験者数を試験効用の損失として統合した包括的指標である期待後悔度$ER(\%)$を比較した結果を図 5 に示す。縦軸の期待後悔度（%）は、検出力が 50%まで低下したときの試験効用または期待被験者数が必要最小限より 2 倍まで増大したときの試験効用と、検出力が目標の 80%でかつ期待被験者数が必要被験者数に一致した理想状態の試験効用との差を 100%の期待後悔度とし、それに対する相対値で各デザインの期待後悔度を表した評価指標である。群逐次デザインと被験者数再設定を比較すると、特にGSD_M, GSD_L

は、PR内のほとんどにおいて$Noninfo$以外の被験者数再設定よりも期待後悔度が大きい傾向にあった。提案する$PP_{info}A$及び$PP_{info}B$は被験者数再設定のなかでPR内の期待後悔度が最も低い傾向にあり、理想のデザイン$True$に最も近い期待後悔度を示した。また、$PP_{info}A$は試験開始前に仮定したδ_{pre}近傍の群間差に対し期待後悔度が最小となり、真の群間差がδ_{pre}と乖離しても期待後悔度の上昇が緩やかであった。それに対し、群逐次デザインGSD_M, GSD_LはPR外の群間差に対し期待後悔度が最小となり、δ_{pre}近傍の期待後悔度は大変大きくなった。また、群逐次デザインの期待後悔度は、GSD_Sも含め群間差の変化に対し期待後悔度が上昇する傾きが急であった。

4.3.5 平均不足検出力

直接評価指標である平均不足検出力MUPを比較した結果を図 6 に示す。MUPは、4.3.1 節で結果を示した包括的評価指標の$Power$の意味する検出力とは異なった指標である。包括的評価指標の検出力は、仮想的に 100,000 回同じ臨床試験を繰り返した場合の統計的有意差の有無という結果を平均したものであった。それに対し、MUPは 1 回の臨床試験の中間解析結果（群逐次デザイン）又はその結果に基づき試験デザインを変更した結果（被験者数再設定）として得られる検出力を直接評価し、さらにその不足分の絶対値を平均することで臨床試験の繰り返し間のばらつきを考慮した不足検出力指標となっている。

包括的評価指標の$Power$が低い傾向にあった提案する$PP_{info}A$及び$PP_{info}B$は、PR内のほとんどの群間差に対し 10%以上検出力が不足することはなかった。さらに、試験開始前に多めの被験者数を設定した左列では、5%未満の検出力不足に抑えられている上、期待被験者数がかなり多くなった群逐次デザインGSD_M, GSD_L及び被験者数再設定$Noninfo$と比較しても同等かそれ以下であった。

4.3.6 平均過剰被験者数

直接評価指標である平均過剰被験者数MOSを比較した結果を図 7 に示す。MOSは、前節のMUP同様包括的評価指標ASNの意味するところの被験者数とは異なった指標である。MOSは、中間解析結果で条件付けたもとでその後の検出力が目標水準を丁度達成するため

に必要な第二ステージ被験者数を臨床試験 1 回ごとに直接評価し、各試験デザインの第二ステージ被験者数（被験者数再設定では再設定された第二被験者数）で過剰に集積された被験者数を平均することにより、臨床試験の繰り返し間のばらつきを考慮した過剰被験者数の指標となっている。縦軸の参照線は、PR内で生じうる固定デザインの過剰被験者数の最大値を表す。固定デザインでは、PR下側の群間差 0.2 に基づく初期被験者数（$N_{initial} = 392$）を設定したもとで、真の群間差がPR下側の 0.3 であった場合に過剰被験者数は最大となる。よって、その必要最小限の被験者数（$N_{initial} = 174$）と 392 例との差をとった、218 例が固定デザインの過剰被験者数の最大値となる。

群逐次デザインGSD_L及び被験者数再設定$Noninfo$は、固定デザインの最大過剰被験者数よりも、平均過剰被験者数がおよそ 100 以上多かった。$Noninfo$以外の被験者数再設定の平均過剰被験者数は、PR内のどの群間差に対しても参照線を下回った。特に提案するデザイン$PP_{info}A$は、比較するデザイン間でもっとも過剰被験者数が少なく、初期被験者数を少なめに設定した右列では、$True$と同等であり、PR内のどの群間差に対しても過剰被験者数が 50 例を下回っていた。

4.3.7 平均後悔度

直接評価指標の平均後悔度$MR(\%)$を比較した結果を図 8 に示す。MRは、4.3.5 節に示した直接評価指標の平均不足検出力と 4.3.6 節に示した直接評価指標の平均過剰被験者数とを、その試験効用に与えるトレード・オフを考慮して統合した指標である。無情報事前分布に基づく$Noninfo$以外の被験者数再設定は、PR内のほとんどの群間差に対し群逐次デザインよりも平均後悔度が低いことが示された。また、提案する$PP_{info}A$及び$PP_{info}B$は他の被験者数再設定デザイン$CP, D_{replace}, Noninfo$と比較し、特に多めの初期被験者数を設定した左列において平均後悔度がより低い傾向にあった。$PP_{info}A$より少し後悔度が高めの$PP_{info}B$は、中間解析データのみに基づく被験者数再設定基準を用いた$D_{replace}$及びCPと比較すると、被験者数を少なめに設定した右列においてはPR内の平均的後悔度が同等であったが、δ_{pre}のより近傍の群間差に対し後悔度が最小になる傾向を示した。

4.3.8 PR 内及び全治療効果範囲内の要約結果

表 2 及び表 3 に、PR内及び治療効果の全範囲における各評価指標の要約値として、平均値、最小値、最大値を比較した結果を示す。表 2 及び表 3 は、本研究で設定した 4 つの試験環境設定のうち、実際の臨床試験の設定としてより一般的と思われる、初期被験者数を多めに設定し試験中期に中間解析を実施する状況設定（$t = 0.5, \delta_{pre} = 0.225$）となっている。他の設定における要約結果は、同様の傾向を示したので割愛する。

包括的評価指標に関し、提案するデザイン以外では表 2 と表 3 とでどの指標もあまり変化しなかった。それに対し、提案するデザイン$PP_{info}A$及び$PP_{info}B$は、表 2 に示すPR内の期待後悔度の要約値（$PP_{info}A: 11[0-27], PP_{info}B: 15[0-33]$）は、表 3 に示す治療効果の全範囲における要約値（$PP_{info}A: 29[0-63], PP_{info}B: 30[0-57]$）と比較し、最小値は変化しないものの、平均値、最大値は半減した。直接評価指標に関し、提案するデザイン以外では包括的評価指標と同様に表 2 と表 3 とでどの指標もあまり変化しなかった。それに対し、提案するデザイン$PP_{info}A$及び$PP_{info}B$は、表 2 に示すPR内の平均後悔度の要約値（$PP_{info}A: 18[5-36], PP_{info}B: 22[9-40]$）は、表 3 に示す治療効果の全範囲における要約値（$PP_{info}A: 41[5-83], PP_{info}B: 41[9-77]$）と比較し、最小値は変化しないものの、平均値、最大値は半減した。

５章　考察

本研究では、被験者数設定の不確実性に対処可能な二つの試験デザインアプローチとして挙げられる群逐次デザインと、提案する方法を含むいくつかの被験者数再設定方法の性能の比較・検討を、シミュレーション研究により行った。既存のシミュレーション研究は、臨床的に意味のない値を含めた広い範囲の群間差に対する検出力曲線及び期待被験者数曲線を比較するか、あるいはある群間差の代表値に対する検出力及び期待被験者数を比較したものであった。これに対し本シミュレーション研究では、ある程度限定された群間差の範囲PR内における各試験デザインの性能評価・比較を行い、被験者数設定の現状において問題となっていた群間差に関するある一定の不確実性に対し、各試験デザインがどの程度の適応力を発揮できるかを評価することに焦点を絞った。それは、検証的臨床試験の計画段階において、群間差に関する事前情報は可能な限り収集され、真の群間差の不確実性の程度に関し、あり得るいくつかのシナリオや、あるいは取り得る値の範囲がある程度絞り込こまれることを想定したためである。このような背景より設定されたPR内の群間差に焦点を絞って性能評価を実施したもう一つの理由は、広い群間差の範囲に対しより性能の良い試験デザインを追及したとしても、そもそもそのような理想的な性能を備えた試験デザインは存在しない可能性もあり、限定された範囲内での性能を向上させるほうが先決でかつ実際的な場面で役立つ機会も多いと考えたからである。本研究で提案した群間差に関する事前情報を考慮したベイズ流被験者数再設定方法は、まさにPR内の群間差に対する性能を向上させることを念頭に提案した方法である。表 2 に示したPR内の性能を要約した結果と表 3 に示したPR外を含めたより広い範囲で性能を要約した結果を比較すると、提案した二つの方法のみその性能が向上しており、その目的に叶った方法が提案できたといえる。

群逐次デザインは、表 1 に示すとおり全部で 5 通り（結果の 6 枚の図のうちGSD_Lは左右の列で設定が一致する）の初期被験者数が設定されたが、その検出力は異なる中間解析時期によってほとんど変化せず、初期被験者数の設定によってその性能がほとんど決定づけられるデザインであることが示された。設定された 5 通りのうちのいくつかは、PRのどの群間差に対しても検出力が目標水準を達成できるという意味で被験者数設定の不確実性へ対処可能なデザインといえる。結果の図の右列に示した$GSD_M(N_{GSD} = 453)$は、それらのな

かで最も被験者数が少なくなる設定であった。しかし、そのGSD_Mは無情報事前分布を仮定した$Noninfo$以外のいずれの被験者数再設定よりも期待被験者数が多くなった。さらに、検出力と期待被験者数を併せた統合指標を評価した場合でも、また被験者数再設定よりも性能が良いとされていた統計的効率を評価した場合でさえ、その性能は$Noninfo$以外のいずれの被験者数再設定よりも劣ることが示された。一方、5 通りのうち被験者数再設定と同じ初期被験者数を設定した$GSD_S(N_{GSD} = 208{,}310)$は、期待被験者数、統計的効率、期待後悔度のいずれも、最も性能の良かった被験者数再設定と同等の性能を示したが、反対に肝心な性能指標である検出力がその他の性能の良さに替えられないほどに低下していた。従って、群逐次デザインによって被験者数設定の不確実性へ対処する場合には、被験者数再設定を行う場合よりもある程度初期被験者数を多く設定しておく必要がある。また、そのように設定された群逐次デザインは不確実性へ対処できる代わりに、事前に仮定した群間差が妥当であった場合（実際には 1 回の臨床試験データによって確認することはできないが、事前に仮定した群間差が真の群間差と近い場合）には、目標検出力を達成するために真に必要な被験者数よりも多くの被験者数を過剰に設定してしまったという後悔を伴う可能性を受け入れなければならない。本論文のシミュレーション研究では、実際に最もよくある状況を想定し、中間解析回数を 1 回と設定した群逐次デザインの性能評価を実施したが、中間解析を複数回行うことによって最終解析まで到達してしまう状況（被験者数が過剰となってしまう状況）をある程度回避することも可能である。しかしその一方で、中間解析回数を増加させたことによる費用の増加は避けられない。よって、群逐次デザインを用いて被験者数設定の不確実性へ対処する場合には、資源に多少の余裕がある試験環境である方が望ましいと考えられる。このような群逐次デザインの特徴を考慮すると、中間解析回数が少なく、被験者数の無駄は許されないが検出力不足も可能な限り避けたい場合には、被験者数再設定を適用する方が望ましいと考えられる。

真のパラメータは未知であり、それ故試験開始前のデザインパラメータの設定にある一定の不確実性が生じるのは必然的でもあり、そのことは検証的臨床試験では避けられないデザイン上の問題といえる。このような問題に対処すべく提案された Adaptive Design の性能を正当に評価するためには、その性能を表す各評価指標の絶対値の高低に強く影響する、

事前パラメータの設定の違いを適切に差し引く必要がある。つまり、試験デザインのもつ適応的性能（Adaptive Performance[130]）によって得られる試験効用の改善なのか、それとも事前の計画の慎重さ、あるいは当該試験のもつ資源の量や実施可能性に依存した試験効用であるのかを区別する必要がある。Adaptive Sample Size Design に関しその適応的性能を正当に評価するためには、結果的に得られる検出力の高低を直接比較するのではなく、群間差の事前の見積もりを中間解析結果に基づき正しく修正できたことによって検出力不足を回避できたのか、それとも妥当な群間差の見積もりに基づき十分な初期被験者数が確保されていたために目標検出力が達成されたのかを区別して評価すべきである。本研究では、実際には観測できない真の群間差に基づく被験者数再設定を各デザインと同じ事前設定のもとで適用したものを、理想的な適応的性能を備えた Adaptive Sample Size Design（$True$）として評価の参照デザインとしたが、この理想デザインと各被験者数再設定との乖離を評価することで各デザインのもつ適応的性能を適切に評価できると考えられる。提案する被験者数再設定方法の検出力の絶対値は、本研究で設定した初期被験者数設定のもとでは比較的低い傾向を示したものの、既存の評価指標である検出力と期待被験者数のいずれにおいても理想のデザインとの乖離が最も小さく、他のデザインと比較し最も適応的性能に優れたデザインであるといえる。また、本研究で新たに提案した評価指標においても、同様に最も適応的性能に優れたデザインであることが確認された。このような性能が得られた理由として、提案する被験者数再設定基準は、ベイズの定理に基づき事前情報を中間解析データで更新することが唯一可能なデザインであることが考えられる。つまり、事前情報を中間データで置き換えるという形の既存の被験者数再設定と比較すると、事前情報と中間データのいずれかが用いられるのではなく、両者が適切に統合されたより多くの情報を活かした上で必要被験者数の再見積もりがなされた結果、提案法の適応的性能が得られたものと考えられる。実際の臨床試験で提案する方法を適用する際の注意点としては、試験の初期に中間解析が実施される場合、シミュレーション結果より検出力不足になる傾向が認められたので、そのような試験計画の際には初期被験者数をより多めに設定する方が安全であると考えられる。

提案する二つの方法のうち$PP_{info}A$は、事前情報の不確実性の度合いを表すPR幅の逆数

に比例した重みを事前分布に与える方法である。このような重みの設定方法により、$PP_{info}A$は、事前情報がより不確実な場合には既存の被験者数再設定に近い結果が得られ、一方より確実な事前情報が得られる場合には固定デザインに近い結果が得られる。このように、提案した$PP_{info}A$は、事前情報の不確実性の程度に応じたデザイン特性を発揮できるという柔軟性を有する方法であるといえる。これに対し提案するもう一つの方法である$PP_{info}B$は、事前情報の重みをデータに依存させて決める方法であることから、試験開始前に仮定した群間差に対する信念があまり強くない場合に有用な方法であると考えられる。同じPR幅に相当する事前情報の不確実性が存在すると考えられる場合でも、そのPR内で事前に仮定する群間差（δ_{pre}）に対する事前の信念の強さは、状況によって異なってくると考えられる。仮定した群間差の値に対する信念が強いわけではなく、実施可能性等の他の試験環境要素を考慮して決められた作業値としての役割が大きい群間差を仮定（設定）した場合には、事前分布の重みをデータに依存させた$PP_{info}B$を用いる方が望ましいと考えられる。$PP_{info}A$は事前に仮定した群間差δ_{pre}の近傍で最も性能が良いという結果であったが、そこから離れたときの性能の悪化が$PP_{info}B$よりも著しい傾向にあった。一方、$PP_{info}B$は、必ずしもδ_{pre}の近傍における最適性は有しないが、PR内の群間差全般に渡って安定した性能の良さを発揮するという傾向がみられた。このようなシミュレーション結果からも、事前に仮定した群間差に対する信念の度合いに応じて事前分布の設定方法を選択すればよいと考えられる。事前の信念の強さを事前分布に反映させるという考え方自体はベイズ流アプローチ全般でよく用いられており、初期被験者数の選択基準、中間解析における試験中止基準、中間解析数の決定基準といった被験者数再設定以外の試験デザイン上のアプローチに対し、専門家の意見といった事前の信念の強さを反映させたベイズ流の臨床試験デザインアプローチがこれまでにもいくつか提案されている[131-137]。

本研究では、シミュレーション研究に基づく、Adaptive Sample Size Design に特化した新しい評価系を提案した。実際の臨床試験計画において詳細な試験デザインの決定を行う際に試験シミュレーションを実施することの重要性が指摘されており[6, 138]、提案した評価系を実際の試験計画へ役立てることが可能である。提案した評価系では、後悔度という概念を用いて検出力と被験者数が試験効用へ与えるトレード・オフを考慮することが可能にな

った。従って、実際の試験計画においても、既存の評価指標である検出力と期待被験者数のみを別々に評価する場合と比較し、提案した後悔度に基づく評価を追加することで最適な試験デザインの選択・決定を明示的に行うことが可能になると考えられる。さらに、複数回発生させたシミュレーションデータに各試験デザインを適用した際の適応的性能のばらつきを考慮した平均後悔度の提案により、実際には 1 回限りでしかない臨床試験において中間解析結果のみに基づき誤った被験者数を再設定してしまう危険性を直接評価することが可能となったことも提案した評価系の利点といえる。既存の被験者数再設定方法では、適応的性能のばらつきを考慮しない期待後悔度とそのばらつきがペナルティーとして反映される平均後悔度を比較すると、どの方法も平均後悔度が大きく上昇してしまうという結果が示された。また、群逐次デザインでは、さらに大きく平均後悔度が上昇することが示された。これは、中間データに含まれる誤差的ばらつきに対し、それに対応する選択肢として試験の中止・継続しか無いという、デザイン的柔軟性に乏しい群逐次デザインの特徴が反映されたものと考えられる。一方、提案するベイズ流被験者数再設定方法では、期待後悔度と比較し平均後悔度がほとんど上昇せず、中間データのばらつきに対する頑健性が確認された。これは、群間差に関する事前情報を考慮したことによって中間データの持つ不確実性が補われた結果得られたものと考えられる。

提案する後悔度において検出力と被験者数のトレード・オフを決める重みは、基準N_{EUP}（検証的試験を実施する上で許容できない不足検出力の基準を被験者数の次元に変換したもの）と基準EOS（許容できない過剰被験者数の基準）の逆数に比例して定めることを提案した。基準N_{EUP}を小さく設定するほど検出力をより重視した後悔度となり、基準EOSを小さく設定するほど被験者数をより重視した後悔度となる。実際の臨床試験計画では、より厳密に試験効用を反映させた費用換算に基づく重みの設定も可能である。例えば、検証的臨床試験の実施上の費用を詳細に分析すれば、「第二種の過誤確率」×「統計的有意差が観察されなかった場合の損失を費用換算した値」と「過剰被験者数」×「被験者一人あたりの追跡やデータマネジメント等に要する費用」との対比を考えた重みの設定も可能と考えられる。後悔度という概念は、実際の効用や損失が理想の効用または最小の損失からどの程度乖離しているかを表す概念で、意思決定理論の分野で用いられている [139]。また、後

悔度や効用を用いた意思決定理論に基づき最適な初期被験者数選択を行うアプローチも提案されており [140-142]、本研究ではそのような後悔度の概念を、被験者数再設定及び群逐次デザインといった Adaptive Sample Size Design の適応的性能に対する評価指標へ応用した点で新規性があった。

今後の検討課題として考えられるものを 3 点挙げる。まず 1 点目は、本研究で提案した被験者数再設定方法の実データへの適用を検討することである。本研究で行ったシミュレーション研究の設定は、一般的な試験環境を模式的に想定したものであるため、実際の適用を想定した結果の解釈には限度がある。そこで、既に実施された臨床試験で仮に提案するベイズ流被験者数再設定方法を適用し被験者数を増加又は減少させていた場合にどのような結果が得られていたかを検討する目的で、観測データの re-sampling に基づき疑似データを発生させ、その試験と同じ試験環境設定のもとで提案法を適用した場合の結果をシミュレーションすることを今後の検討課題としたい。2 点目は提案法の適用条件に関する課題である。提案法の適用条件として一般に考えられるものがいくつか挙げられる。提案法は群間差の事前情報を必要とするため、過去の臨床試験成績や海外試験成績が入手可能な状況が必要となる。標準的な第三相試験のような検証的臨床試験の場合、開発のより早期の臨床試験成績または海外試験成績等が得られていることが通常であるため、提案法の適用可能性が高いと考えられる。また、特に癌化学療法の領域では、新規の薬剤でも異なる癌腫、異なる対象集団、併用薬や投与方法に関し異なる治療レジメンのもとでの使用成績は得られている場合が多く、その薬剤の効果の Prior Range をある程度特定することは可能である。一方、希少疾患のように事前情報に乏しいことが多い疾患領域では、提案法の適用は困難であり、さらに試験へ登録可能な被験者数にも限度があるため、被験者数を再設定すること自体が実際的ではない可能性が考えられる。また、登録可能な被験者数に余裕が存在する場合でも、例えば心筋梗塞や脳卒中の 1 次予防試験、乳癌患者の生存をエンドポイントとした試験のように、患者登録期間と比較してイベントが観測されるまでの期間が極端に長いような疾患においては、被験者数再設定の実施は困難であると考えられる。そのような試験の場合には、症例追跡期間を変更することによって、被験者数の代わりに観測イベント数を再設定するという試験デザインへ提案法を拡張する必要性が生じる。実

際に提案法を適用する際には、以上で述べた一般的な適用可能性以外にもより詳細な適用条件を検討すべきと考えられる。よって、2 点目の課題は、現在のところ被験者数再設定の実施例は極めて少ないが今後増加すると予想されるため、それらをレビューすることによって提案法の適用が必要とされる状況及び提案法が適用可能と考えられる状況を検討することとしたい。3 点目は、結果変数の型の違いが今回得られた結果に与える影響を評価することである。本研究で扱った結果変数の型は連続量であったが、検証的臨床試験のエンドポイントとしてはイベント発生割合やイベント発生までの時間の場合が多いことを考えると、今回提案した方法がどのような挙動を示すかを今後検討することに意味があると考えられる。

６章　結論

本研究では、群間差に関する事前情報を考慮した新しいベイズ流被験者数再設定方法を提案し、その性能をシミュレーション研究により評価した。その結果、提案法は既存の被験者数再設定方法及び群逐次デザインと比較し、検出力及び期待被験者数がいずれも真の群間差に基づく理想の被験者数再設定デザインに最も近く、検証的臨床試験における被験者数設定の不確実性へ対処可能な方法として、より優れた適応的性能を有することが示された。また、群逐次デザイン及び被験者数再設定という Adaptive Sample Size Design に特化した評価系を新たに提案したことにより、不足検出力と過剰被験者数が試験効用に与えるトレード・オフ及び試験デザインの性能のばらつきを考慮した評価が可能になった。これによって、提案するベイズ流被験者数再設定方法が、中間データの不確実性に対する頑健性を持つことが確認された。

謝辞

本研究を進めていく上で大変有用なご助言をくださった東京大学大学院医学系研究科疫学・予防保健学分野の大橋靖雄教授、研究の方向性についてのご指導、ならびに研究内容について議論をしてくださった指導教員の松山裕准教授、及び研究室のスタッフの皆様に深く感謝申し上げます。

参考・引用文献

1) 日本製薬工業協会. *Data Book*, 2004.

2) Windhover Information Inc. Windhover's *in vivo*: The business and medicine report. Bain drug economics model, 2003.

3) 立石佳代. *日本大学大学院総合社会情報研究科紀要* 2007; **8**:367-378.

4) 厚生労働省. *新医薬品ビジョン～イノベーションを担う国際競争力のある産業をめざして*, 2007. http://www.mhlw.go.jp/houdou/2007/08/dl/h0830-1b.pdf.

5) 米国製薬工業協会. *ニュースリリース*, 2008. http://www.phrma-jp.org/images/uploads/080402_J_2007R&D.pdf.

6) FDA. *Innovation or stagnation? Challenge and opportunity on the critical path to new medical products*, 2004. http://www.fda.gov/oc/initiatives/criticalpath/whitepaper.html.

7) Gallo P, Krams M. PhRMA working group on adaptive designs: introduction to the full white paper. *Drug Information Journal* 2006; **40**:421-423.

8) Dragalin V. Adaptive designs: terminology and classification. *Drug Information Journal* 2006; **40**:425-435.

9) Quinlan J, Krams M. Implementing adaptive designs: logistical and operational considerations. *Drug Information Journal* 2006; **40**:437-444.

10) Gallo P. Confidentiality and trial integrity issues for adaptive designs. *Drug Information Journal* 2006; **40**:445-450.

11) Gaydos B, Krams M, Perevozskaya I, Bretz F, Liu Q, Gallo P, Berry D, Stein C-C, Pinheiro J, Bedding A. Adaptive dose-response studies. *Drug Information Journal* 2006; **40**:451-461.

12) Maca J, Bhattacharya S, Dragalin V, Gallo P, Krams M. Adaptive seamless phase II/II designs-background, operational aspects, and examples. *Drug Information Journal* 2006; **40**: 463-473.

13) Stein C-C, Anderson K, Gallo P, Collins S. Sample size re-estimation: a review and recommendations. *Drug Information Journal* 2006; **40**:475-484.

14) Gallo P, Stein C-C, Dragalin V, Gaydos B, Krams M, Pinheiro J. Executive summary of the

PhRMA working group on adaptive designs in clinical drug development. *Journal of Biopharmaceutical Statistics* 2006; **16**:275-283.

15) O'Neill RT. FDA's critical path initiative: a perspective on contributions of biostatistics. *Biometrical Journal* 2006; **48**:559-564.

16) Zia MI, Siu LL, Pond GR, Chen EX. Comparison of outcomes of phase II studies and subsequent randomized control studies using identical chemotherapeutic regimens. *Journal of Clinical Oncology* 2005; **23**:6982-6991.

17) Freiman JA, Chalmers TC, Smith H, Kuebler RR. The importance of beta, the type II error, and sample size in the design and interpretation of the randomized controlled trial: a survey of 71 'negative' trials. *New England Journal of Medicine* 1978; **299**:690-694.

18) Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. *Journal of the American Medical Association* 1994; **272**:122-124.

19) Hebert RS, Wright SM, Dittus RS, Elasy TA. Prominent medical journals often provide insufficient information to assess the validity of studies with negative results. *Journal of Negative Results in Biomedicine* 2002; **1**:1.

20) Brown CG, Kelen GD, Ashton JJ, Werman HA. The beta error and sample size determination in clinical trials in emergency medicine. *Annals of Emergency Medicine* 1987; **16**:183-187.

21) Edmund MJ, Overall JE, Rhoades HM. Beta, or type II error in psychiatric controlled clinical trials. *Journal of Psychiatric Research* 1985; **19**:563-567.

22) Mengel MB, Davis AB. The statistical power of family practice research. *Family Practice Research Journal* 1993; **13**:105-111.

23) Dimick JB, Diener-West M, Lipsett PA. Negative results of randomized clinical trials published in the surgical literature: equivalency or error? *Archives of Surgery* 2001; **136**:796-800.

24) Williams HC, Seed P. Inadequate size of 'negative' clinical trials in dermatology. *British Journal of Dermatology* 1993; **128**:317-326.

25) Keen HI, Pile K, Hill CL. The prevalence of under-powered randomized clinical trials in rheumatology. *Journal of Rheumatology* 2005; **32**:2083-2088.

26) Bedard PL, Krzyzanowska MK, Pintilie M, Tannock IF. Statistical power of negative randomized controlled trials presented at American society for clinical oncology annual meetings. *Journal of Clinical Oncology* 2007; **25**:3482-3487.

27) Piantadosi S. *Clinical Trials: A Methodologic Perspective* (2nd edn). Wiley: New York, 2005.

28) 医薬審 第 1047 号. *臨床試験の統計的原則*, 1998.

29) Altman DG. Statistics and ethics in medical research III. How large a sample? *British Medical Journal* 1980; **281**:1336-1338.

30) Brown BR Jr. Statistical controversies in the design of clinical trials－some personal views. *Controlled Clinical Trials* 1980; **1**:13-27.

31) Lachin JM. Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials* 1981; **2**:93-113.

32) Fleiss JL, Tytun A, Ury HK. A simple approximation for calculating sample sizes for comparing independent proportions. *Biometrics* 1980; **36**:343-346.

33) Freedman LS. Tables of the number of patients required in clinical trials using the logrank test. *Statistics in Medicine* 1982; **1**:121-129.

34) Schoenfeld DA. Sample size formula for the proportional-hazards regression model. *Biometrics* 1983; **39**:499-503.

35) Leth FV, Phanuphak P, Ruxrungtham K, Baraldi E, Miller S, Gazzard B, et al. Comparison of first-line antiretroviral therapy with regiments including nevirapine, efavirenz, or both drugs, plus stavudine and lamivudine: a randomized open-label trial, the 2NN Study. *Lancet* 2004; **363**:1243-1263.

36) Gebhart MJP, Procter M, Jones BL, Goldhirsch A, Untch M, Smith I, et al.. Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. *New England Journal of Medicine* 2005; **353**:1659-1672.

37) Wittes J, Brittain E. The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine* 1990; **9**:65-72.

38) Wang SJ, Hung J, O'Neil R. Uncertainty in planning phase III trial based on phase II data:

sample size. *Proceedings of the Biopharmaceutical Section*, ASA, Toronto, 2004; 895–901.

39) Senn S, Brets F. Power and sample size when multiple endpoints are considered. P*hermaceutical Statistics* 2007; **6**:161-170.

40) Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. Composite Outcomes in Randomized Trials. Greater Precision but with Greater Uncertainty? *Journal of the American Medical Association,* 2003; **289**:2554-2559.

41) Shih JH. Sample size calculation for complex clinical trials with survival endpoints. *Controlled Clinical Trials* 1995; **16**:395-407.

42) Woodward M. Formulae for sample size, power and minimum detectable relative risk in medical studies. *Statistician* 1992; **41**:185-196.

43) Walraven CV, Mahon JL, Moher D, Bohm C, Laupacis A. Surveying physicians to determine the minimal important difference: implications for sample-size calculation. *Journal of Clinical Epidemiology* 1999; **52**:717-723.

44) Detsky AS, Sackett DL. Establishing therapeutic equivalency. What is a clinically significant difference? *Archieves of Internal Medicine* 1986; **5**:861-862.

45) Naylor CD, Thomas HAL. Can there be a more patient-centered approach to determining clinically important effect sizes for randomized treatment trials?, *Journal of Clinical Epidemiology* 1994; **47**:787-795.

46) Detsky AS. Using economic analysis to determine the resource consequences of choices made in planning clinical trials. *Journal of Chronic Disease* 1985; **38**:753-765.

47) Shih WJ, Long J. Blinded sample size re-estimation with unequal variances and center effects in clinical trials. *Communications in Statistics (A) －Theory and Methods* 1998; **27**:395-408.

48) Wittes J, Schabenberger O, Zucker D, Britain E, Proschan M. Internal pilot studies I: type I error rate of the naive t-test. *Statistics in Medicine* 1999; **18**:3481-3491.

49) Zucker D, Wittes J, Schabenberger O, Brittain E. Internal pilot studies II: comparison of various procedures. *Statistics in Medicine* 1999; **18**:3493-3509.

50) Gould AL. Interim analyses for monitoring clinical trials that do not materially affect the type I

error rate. *Statistics in Medicine* 1992; **11**:55-66.

51) Wald A. *Sequential Analysis*. Wiley: New York, 1947.

52) Whitehead J. *The Design and Analysis of Sequential Clinical Trials* (revised 2nd edn). Wiley: Chichester, 1987.

53) Armitage P. *Sequential Medical Trials* (2nd edn). Blackwell: Oxford, 1975.

54) Jennison C, Turnbull BW. *Group sequential methods with applications to clinical trials*. Chapman& Hall: Boca Raton, 2000.

55) Enas GG, Dornseif BE, Sampson CB, Rockhold FW, Wuu J. Monitoring versus interim analysis of clinical trials: Perspective from the pharmaceutical industry. *Controlled Clinical Trials* 1989; **10**:57-70.

56) Jennison C, Turnbull BW. Statistical approaches to interim monitoring of medical trials: A review and commentary. *Statistical Science* 1990; **5**:299-317.

57) Ellenberg SS, Fleming TR, DeMets DL. *Data Monitoring Committees in Clinical Trials: A Practical Persective.* Wiley: New York, 2002.

58) Lan KKG, Simon R, Halperin, M. Stochastically curtailed tests in long-term clinical trials. *Sequential Analysis* 1982; **1**:207-219.

59) Halperin M, Lan KKG, Ware JH, Johnson NJ, DeMets DL. An aid to data monitoring in long-term clinical trials. *Controlled Clinical Trials* 1982; **3**:311-323.

60) Anderson PK. Conditional power calculations as an aid in the decision whether to continue a clinical trial. *Controlled Clinical Trials* 1987; **8**:67-74.

61) Halperin M, Lan KKG, Wright EC, Foulkes MA. Stochastic curtailing for the comparison of slopes in longitudinal studies. *Controlled Clinical Trials* 1987; **8**:315-326.

62) Hunsberger S, Sorlie P, Geller NL. Stochastic curtailing and conditional power in matched case-control studies. *Statistics in Medicine* 1994; **13**:663-670.

63) Cui L, Hung HMJ, Wang S. Modification of sample size in group sequential clinical trials. *Biometrics* 1999; **55**:853-857.

64) Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials.

Biometrics 1999; **55**:1286-1290.

65) Müller HH, Schäfer H. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* 2001; **57**:886-891.

66) Anderson, TW. A modification of the sequential probability ratio test to reduce sample size. *Annals of Mathematical Statistics* 1960; **31**:165-197.

67) Lai, TL. Optimal stopping and sequential tests which minimize the maximum expected sample size. *Annals of Statistics* 1973; **1**:659-663.

68) Whitehead J, Stratton I. Group sequential clinical trials with triangular continuation regions. *Biometrics* 1983; **39**:227-236.

69) Whitehead J. Use of the triangular test in sequential clinical trials. In *Handbook of Statistics in Oncology*, Crowley J (ed.). Dekker: New York, 2001.

70) Whitehead, J. and Stratton, I. (1983), "Group Sequential Clinical Trials with Triangular Continuation Regions," Biometrics, 39, 227-236.

71) Armitage P, McPherson CK, Rowe BC. Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society* 1969; **A132**:235-244.

72) Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977; **64**:191-199.

73) O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979; **35**:549-556.

74) Haybittle JL. Repeated Assessment of Results in Clinical Trials of Cancer Treatment. *British Journal of Radiology* 1971; **44**:793-797.

75) Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. Design and Analysis of Randomized Clinical Trials Requiring Prolonged Observation of Each Patient: I. Introduction and Design. *British Journal of Cancer* 1976; **34**:585-612.

76) Wang SK, Tsiatis AA. Approximately optimal one-parameter boundaries for group sequential

trials. *Biometrics* 1987; **43**:193-199.

77) Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**:659-663.

78) Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation.*Wiley: Chichester, 2004.

79) Spiegelhalter DJ, Freedman LS, Blackburn PR. Monitoring clinical trials: conditional or predictive power? *Controlled Clinical Trials* 1986; **7**:8-17.

80) Röhmel J. Editorial－adaptive designs: expectations are high. *Biometrical Journal* 2006; **48**:491-492.

81) Hung HMJ, O'Neill RT, Wang SJ, Lawrence J. A Regulatory View on Adaptive/Flexible Clinical Trial Design. *Biometrical Journal* 2006; **48**: 565-573.

82) Koch A. Confirmatory Clinical Trials with an Adaptive Design. *Biometrical Journal* 2006; **48**: 574-585.

83) Gallo P, Maurer W. Challenges in Implementing Adaptive Designs: Comments on the Viewpoints Expressed by Regulatory Statisticians. *Biometrical Journal* 2006; **48**: 591-597.

84) Wittes J, Lachenbruch PA. Opening the Adaptive Toolbox *Biometrical Journal* 2006; **48**:598-603.

85) Cyrus R, Mehta CR, Jemiai Y. A Consultant's Perspective on the Regulatory Hurdles to Adaptive Trials. *Biometrical Journal* 2006; **48**:604-608.

86) Bauer P. Methodological Developments vs. Regulatory Requirements. *Biometrical Journal* 2006; **48**:609-612.

87) Hung HMJ, O'Neill RT, Wang SJ, Lawrence J. Rejoinder. *Biometrical Journal* 2006; **48**:613-615.

88) Koch A. Rejoinder. *Biometrical Journal* 2006; **48**:616-622.

89) Phillips AJ, Keene ON. Adaptive designs for pivotal trials: discussion points from the PSI Adaptive Design Expert Group. *Pharmaceutical Statistics* 2006; **5**:61-66.

90) EMEA. Reflection paper on methodological issues in confirmatory clinical trials with flexible

design and analysis plan. CHMP/EWP/2459/02. London, 2006.

91) FDA. Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials - Draft Guidance for Industry and FDA Staff, 2006. http://www.fda.gov/cdrh/osb/guidance/1601.html

92) Gallo P, Stein C-C, Dragalin V, Gaydos B, Krams M, Pinheiro J. Adaptive designs in clinical drug development－An executive summary on the PhRMA working group. *Journal of Biopharmaceutical Statistics* 2006; **16**:275-283.

93) Gallo P, Krams M. PhRMA working group on adaptive designs: introduction to the full white paper. *Drug Information Journal* 2006; **40**:421-423.

94) Dragalin V. Adaptive designs: terminology and classification. *Drug Information Journal* 2006; **40**:425-435.

95) Quinlan JA, Krams M. Implementing adaptive designs: logistical and operational considerations. *Drug Information Journal* 2006; **40**:437-444.

96) Gallo P. Confidentiality and trial integrity issues for adaptive designs. *Drug Information Journal* 2006; **40**:445-450.

97) Gaydos B, Krams M, Perevozskaya I, Bretz F, Liu Q, Gallo P, et al.. Adaptive dose-response studies. *Drug Information Journal* 2006; **40**:451-461.

98) Maca J, Bhattacharya S, Dragalin V, Gallo P, Krams M. Adaptive seamless II/III designs－background, operational aspects and examples. *Drug Information Journal* 2006; **40**:463-473.

99) Stein C-C, Anderson K, Gallo P, Collins S. Sample size reestiamtion: A review and recommendations. *Drug Information Journal* 2006; **40**:475-484.

100) Bauer P, Einfalt J. Application of adaptive designs－a review. *Biometrical Journal* 2006; **48**:493-506.

101) Taylor AL, Ziesche S, Yancy C, Carson P, D'Agostino R, Ferdinand K, et al.. Combination of isosorbide dinitrate and hydralazine in blacks with heart failure. *New England Journal of Medicine* 2005; **351**:2049-2057.

102) Proschan M, Hunsberger S. Designed extension studies based on conditional power. *Biometrics* 1995; **51**:1315-1324.

103) Shun Z, Yuan W, Brady W, Hsu H. Type I error in sample size re-estimation based on observed treatment difference. *Statistics in Medicine* 2001; **20**:497-513.

104) Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics* 1994; **50**:1029-1041.

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

106) Fisher LD. Self-designing clinical trials. *Statistics in Medicine* 1998; **17**:155-1562.

107) Denne JS. Sample size recalculation using conditional power. *Statistics in Medicine* 2001; **20**:2645-2660.

108) Müller HH, Schäfer H. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* 2001; **57**:886-891.

109) Müller HH, Schäfer H. A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine* 2004; **23**:2497-2508.

110) Posch M, Bauer P. Adaptive two stage designs and the conditional error function. *Biometrical Journal* 1999; **41**:689-696.

111) Wassmer, G. A comparison of two methods for adaptive interim analysis in clinical trials. *Biometrics* 1998; **54**:696-705.

112) Chen YH, DeMets DL, Lan KKG. Increasing the sample size when the unblended interim result is promising. *Statistics in Medicine* 2004; **23**:1023-1038.

113) Uemura K, Matsuyama Y, Ohashi Y. 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* 2008; **29**:19-34.

114) Jennison C, Turnbull BW. Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine* 2003; **22**:971-993.

115) Tsiatis AA, Mehta C. On the efficiency of the adaptive design for monitoring clinical trials. *Biometrika* 2003; **90**:367-378.

116) Burman CF, Sonesson C. Are flexible designs sound? (with Discussion). *Biometrics* 2006; **62**:664-683.

117) Schmits N. *Optimal Sequentially Planned Decision procedures.* Lecture Notes in Statistics, col. 79. Springer: New York, 1993.

118) Jennison C, Turnbull BW. Efficient group sequential designs when there are several effect sizes under consideration. *Statistics in Medicine* 2006; **25**:917-932.

119) Brannath W, Bauer P, Posch M. On the efficiency of adaptive design for flexible interim decisions in clinical trials. *Journal of Statistical Planning and Inference* 2006; **136**:1956-1961.

120) Bauer P, König F. The reassessment of trial perspectives from interim data－a critical view. *Statistics in Medicine* 2006; **25**:23-36.

121) Bartroff J, Lai TL. Efficient adaptive designs with mid-course sample size adjustment in clinical trials. *Statistics in Medicine* 2008; **27**:1593-1611.

122) Bauer P. Adaptive designs: Looking for a needle in the haystack－A new challenge in medical research. *Statistics in Medicine* 2008; **27**:1565-1580.

123) Hung HMJ, Cui L, Wang SJ, Lowrence J. Adaptive statistical analysis following sample size modification based on interim review of effect size. *Journal of Biopharmaceutical Statistics* 2005; **15**:693-706.

124) Shih WJ. Commentary on: Group sequential, sample size re-estimation and two-stage adaptive designs in clinical trials: A comparison. *Statistics in Medicine* 2006; **25**:933-941.

125) Brannath W, Bauer P. Optimal conditional error functions for the control of conditional power. *Biometrics* 2004; **60**:715-723.

126) Posch M, Bauer P, Brannath W. Issues in designing flexile trials. *Statistics in Medicine* 2003; **22**:953-969.

127) Wang MD. Sample size reestimation by Bayesian prediction. *Biometrical Journal* 2007; **49**:365-377.

128) Bauer P, Koenig F. The reassessment of trial perspectives from interim data－a critical view. *Statistics in Medicine* 2008; **25**:23-36.

129) Lindley, D. V. (1970). *Introduction to probability and statistics from a Bayesian viewpoint*. Part 2: Inference. Cambridge: Cambridge University Press.

130) Liu DF, Cui L. Evaluating the adaptive performance of flexible sample size designs with treatment difference in an interval. *Statistics in Medicine* 2008; **27**:584-596.

131) Day NE. Two-stage designs for clinical trials. *Biometrics* 1969; **25**:111-118.

132) Herson J. Predictive probability early termination plans for phase2 clinical trials. *Biometrics* 1979; **35**:775-783.

133) Thompson MS. Decision-analytic determination of study size. The case of electronic fetal monitoring. *Medical Decision Making* 1981; **1**:165-179.

134) Atkison EN, Brown BW, Herson J. KSTAGE: an interactive computer program for designing phase2 clinical trials. *Biometrics* 1982; **15**:220-227.

135) McPherson K. On choosing the number of interim analyses in clinical trials. *Statistics in Medicine* 1982; **1**:25-36.

136) Freedman LS, Spiegelhalter DJ. The assessment of subjective opinion and its use in relation to stopping rules for clinical trials. *The Statisticians* 1983; **33**:153-160.

137) Spiegelhalter DJ, Freedman LS. A predictive approach to selecting then size of a clinical trial, based on subjective clinical opinion. *Statistics in Medicine* 1986; **5**:1-13.

138) Chang M. *Adaptive Design Theory and Implementation Using SAS and R*. Chapman & Hall: Boca Raton, 2008.

139) Lindgren BW. *Elements of Decision Theory*. Macmillan: New York, 1971.

140) Lindley DV. The choice of sample size. *Statistician* 1997; **46**:129-138.

141) Pally A. A decision analytic approach to determining sample size in a phase III program. *Drug Information Journal* 2000; **34**:365-377.

142) Pezeshk H. Bayesian techniques for sample size determination in clinical trials: a short review. *Statistical Methods in Medical Research* 2003; **12**:489-504.

表1. シミュレーション研究の全般的試験環境設定

δ_{min}	δ	PR	δ_{pre}	$N_{initial}$	N_{GSD} / S,M,L	N_{max}	N_{min}†	t
0.15	0.15 to 0.35 by 0.02	$\delta_{lower} = 0.2$	0.225	310	310,504,698	698	174	0.25
								0.5
		$\delta_{upper} = 0.3$	0.275	208	208,453,698			0.75

† $n_1 > 174$の場合には$N_{min} = n_1$とする。

表2. 治療効果のPR（$\delta \in [0.2,0.3]$）における各評価指標の要約値：平均値（ 最小値 – 最大値 ）

種類	評価指標	CP	$D_{replace}$	GSD_L	GSD_M	GSD_S
包括	*Power*	0.89 (0.83 - 0.94)	0.91 (0.86 - 0.95)	0.99 (0.97 - 1.00)	0.96 (0.91 - 1.00)	0.86 (0.74 - 0.95)
	ASN	320 (266 - 375)	339 (281 - 397)	460 (401 - 525)	376 (333 - 418)	264 (246 - 278)
	$Efficiency_{100}$	0.28 (0.22 - 0.35)	0.28 (0.22 - 0.34)	0.22 (0.19 - 0.25)	0.26 (0.22 - 0.30)	0.33 (0.27 - 0.39)
	期待後悔度	25 (6 - 43)	33 (12 - 51)	81 (48 - 115)	49 (18 - 79)	15 (0 - 32)
直接	平均不足検出力	0.02 (0.00 - 0.05)	0.02 (0.01 - 0.03)	0.01 (0.00 - 0.03)	0.04 (0.00 - 0.09)	0.13 (0.05 - 0.22)
	平均過剰被験者数	158 (125 - 175)	176 (142 - 194)	272 (234 - 302)	168 (132 - 197)	71 (49 - 93)
	平均後悔度	36 (29 - 45)	46 (28 - 65)	291 (153 - 448)	168 (87 - 266)	81 (65 - 113)

種類	評価指標	$Noninfo$	$PP_{info}A$	$PP_{info}B$	$TRUE$
包括	*Power*	0.93 (0.89 - 0.96)	0.88 (0.79 - 0.94)	0.89 (0.81 - 0.95)	0.84 (0.81 - 0.88)
	ASN	398 (333 - 462)	273 (238 - 309)	290 (247 - 336)	252 (202 - 320)
	$Efficiency_{100}$	0.24 (0.19 - 0.29)	0.33 (0.26 - 0.40)	0.31 (0.24 - 0.38)	0.35 (0.25 - 0.43)
	期待後悔度	56 (30 - 78)	11 (0 - 27)	15 (0 - 33)	2 (0 - 8)
直接	平均不足検出力	0.01 (0.00 - 0.03)	0.00 (0.00 - 0.02)	0.00 (0.00 - 0.02)	0.00 (0.00 - 0.00)
	平均過剰被験者数	271 (236 - 291)	46 (1 - 79)	95 (61 - 116)	0 (0 - 0)
	平均後悔度	85 (54 - 117)	18 (5 - 36)	22 (9 - 40)	0 (0 - 0)

試験環境設定： $t = 0.5, \delta_{pre} = 0.225$

表3. 治療効果の全範囲（$\delta \in [0.15,0.35]$）における各評価指標の要約値：平均値（ 最小値 – 最大値 ）

種類	評価指標	CP	$D_{replace}$	GSD_L	GSD_M	GSD_S
包括	*Power*	0.86 (0.62 - 0.97)	0.88 (0.65 - 0.98)	0.96 (0.79 - 1.00)	0.92 (0.65 - 1.00)	0.81 (0.46 - 0.99)
	ASN	323 (208 - 448)	341 (218 - 468)	471 (361 - 617)	374 (284 - 459)	259 (215 - 285)
	$Efficiency_{100}$	0.29 (0.14 - 0.47)	0.28 (0.14 - 0.45)	0.21 (0.13 - 0.28)	0.26 (0.14 - 0.35)	0.32 (0.16 - 0.46)
	期待後悔度	34 (6 - 63)	38 (0 - 70)	84 (3 - 181)	56 (0 - 121)	38 (0 - 77)
直接	平均不足検出力	0.04 (0.00 - 0.17)	0.03 (0.00 - 0.16)	0.04 (0.00 - 0.18)	0.07 (0.00 - 0.28)	0.16 (0.01 - 0.42)
	平均過剰被験者数	130 (7 - 175)	148 (13 - 194)	256 (130 - 322)	158 (62 - 220)	69 (17 - 116)
	平均後悔度	44 (29 - 70)	53 (24 - 97)	333 (65 - 771)	205 (65 - 479)	110 (65 - 219)

種類	評価指標	$Noninfo$	$PP_{info}A$	$PP_{info}B$	$TRUE$
包括	*Power*	0.90 (0.69 - 0.98)	0.84 (0.53 - 0.98)	0.85 (0.58 - 0.98)	0.83 (0.68 - 0.93)
	ASN	395 (248 - 528)	276 (198 - 361)	294 (201 - 400)	282 (173 - 490)
	$Efficiency_{100}$	0.25 (0.13 - 0.40)	0.33 (0.15 - 0.50)	0.31 (0.14 - 0.49)	0.34 (0.14 - 0.54)
	期待後悔度	54 (6 - 93)	29 (0 - 63)	30 (0 - 57)	13 (0 - 34)
直接	平均不足検出力	0.03 (0.00 - 0.15)	0.05 (0.00 - 0.26)	0.04 (0.00 - 0.21)	0.02 (0.00 - 0.14)
	平均過剰被験者数	242 (103 - 291)	44 (0 - 86)	77 (0 - 117)	0 (0 - 2)
	平均後悔度	91 (37 - 169)	41 (5 - 83)	41 (9 - 77)	4 (0 - 24)

試験環境設定： $t = 0.5, \delta_{pre} = 0.225$

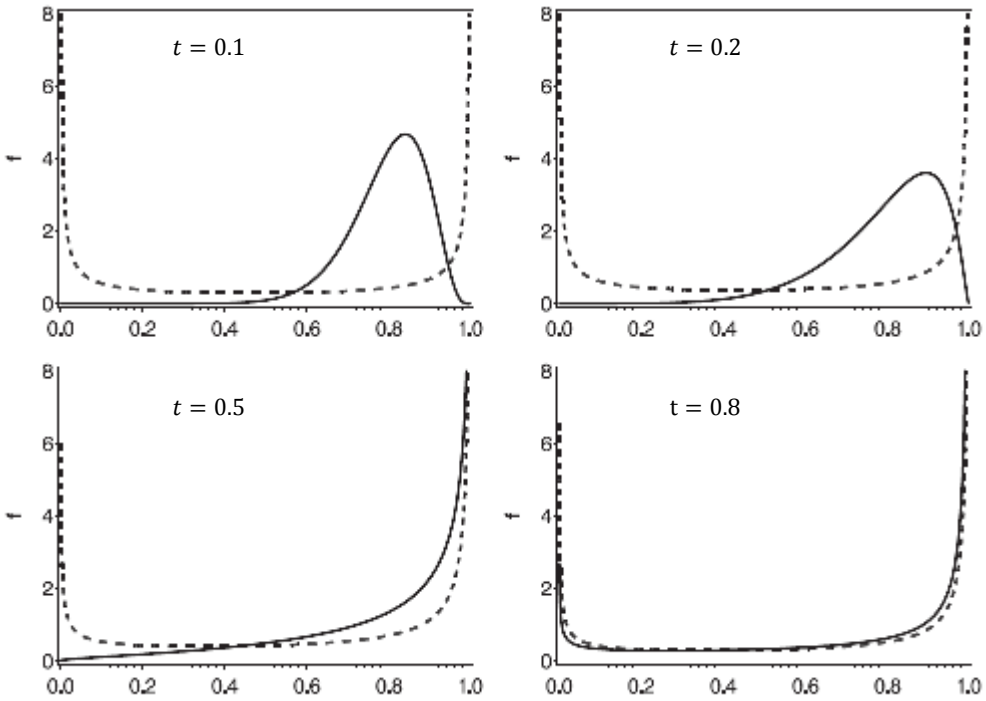

図1. 条件付検出力の分布

実線が群間差の真値に基づく条件付検出力CP_{δ_2}の確率密度関数$f_{\delta_1}(CP_{\delta_2})$を表し、破線が群間差の推定値$\hat{\delta}_1$に基づく条件付検出力$CP_{\delta_2=\hat{\delta}_1}$の確率密度関数$f_{\delta_1}(CP_{\delta_2=\hat{\delta}_1})$を表す（Bauer, Koenig (2008)の Figure 1 を改編）

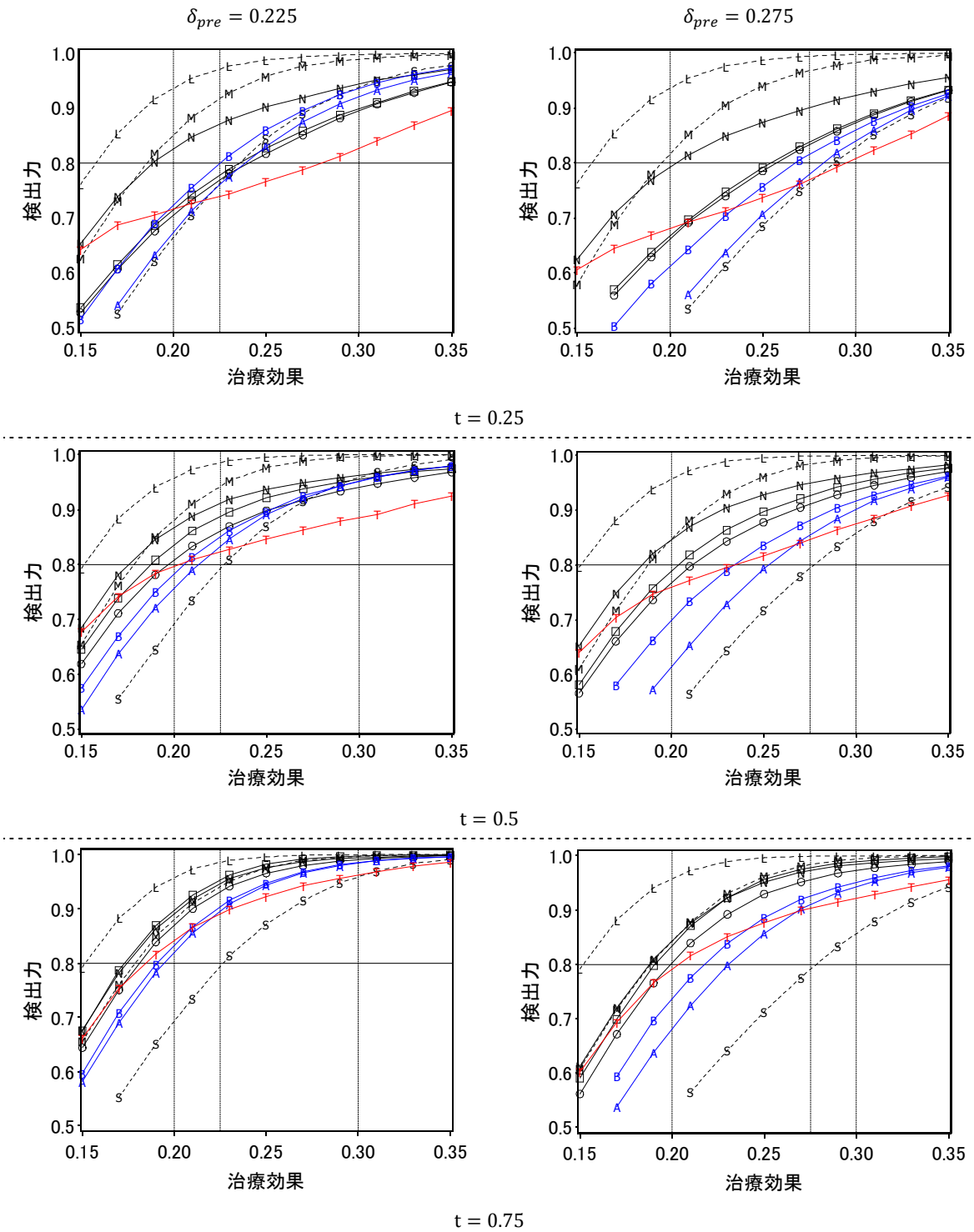
δ_{pre} = 0.225
δ_{pre} = 0.275
検出力
治療効果
t = 0.25
t = 0.5
t = 0.75
CP
D_Replace
GSD_L
GSD_M
GSD_S
Noninfo
PP_info A
PP_info B
True

図2. 検出力　*Power*

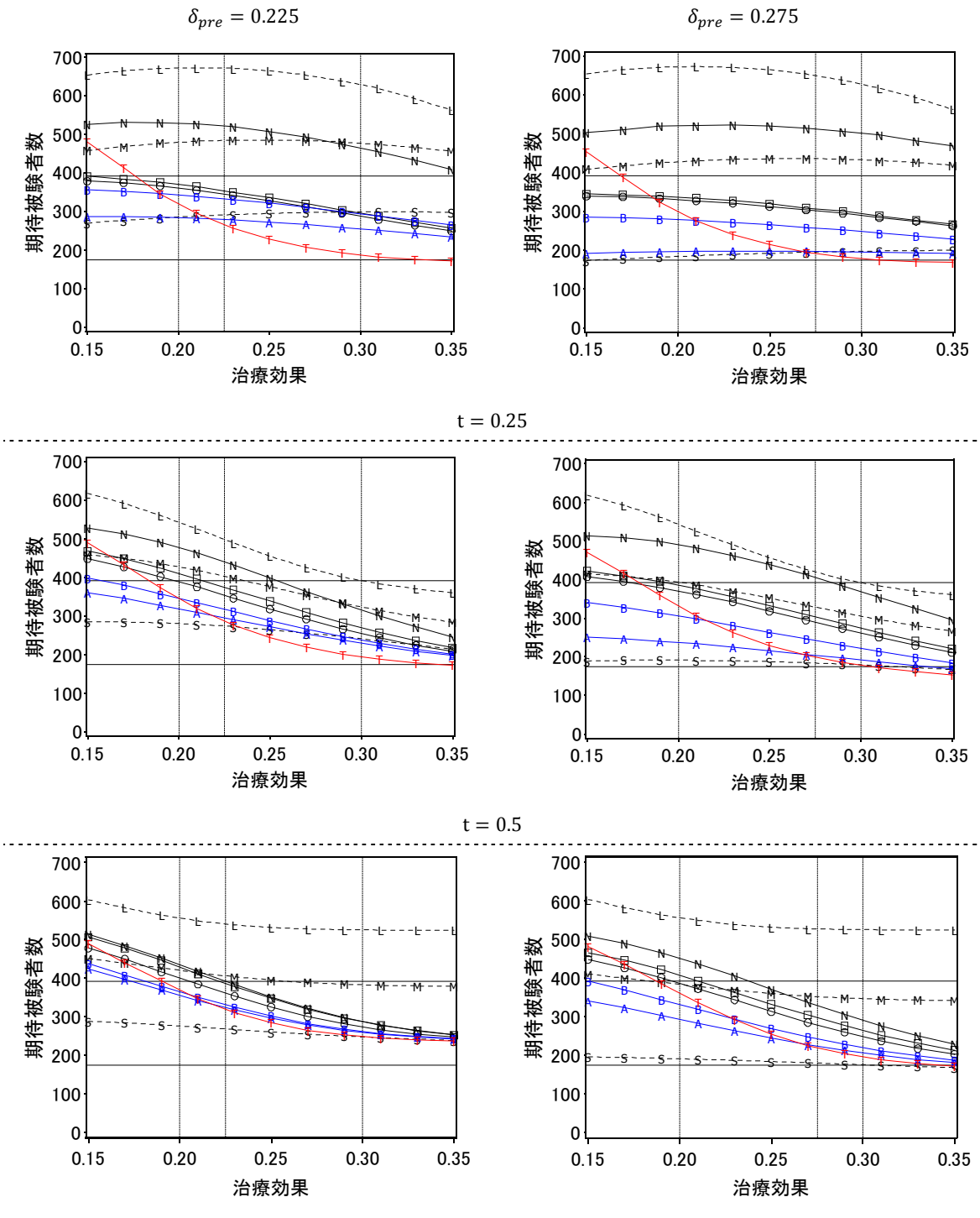

図3. 期待被験者数　ASN

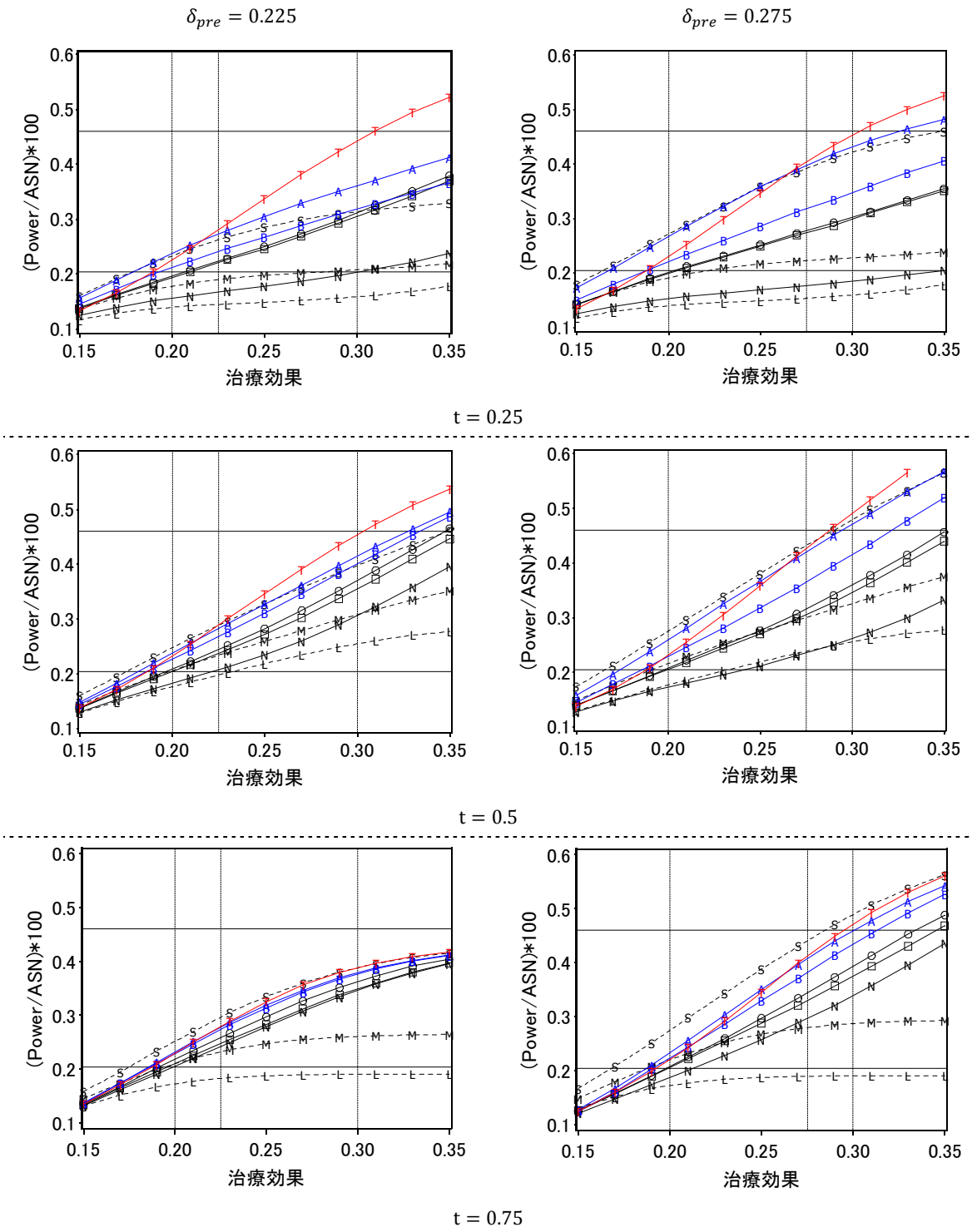

図4. 期待被験者数 100 例あたりの検出力　$Efficiency_{100}$

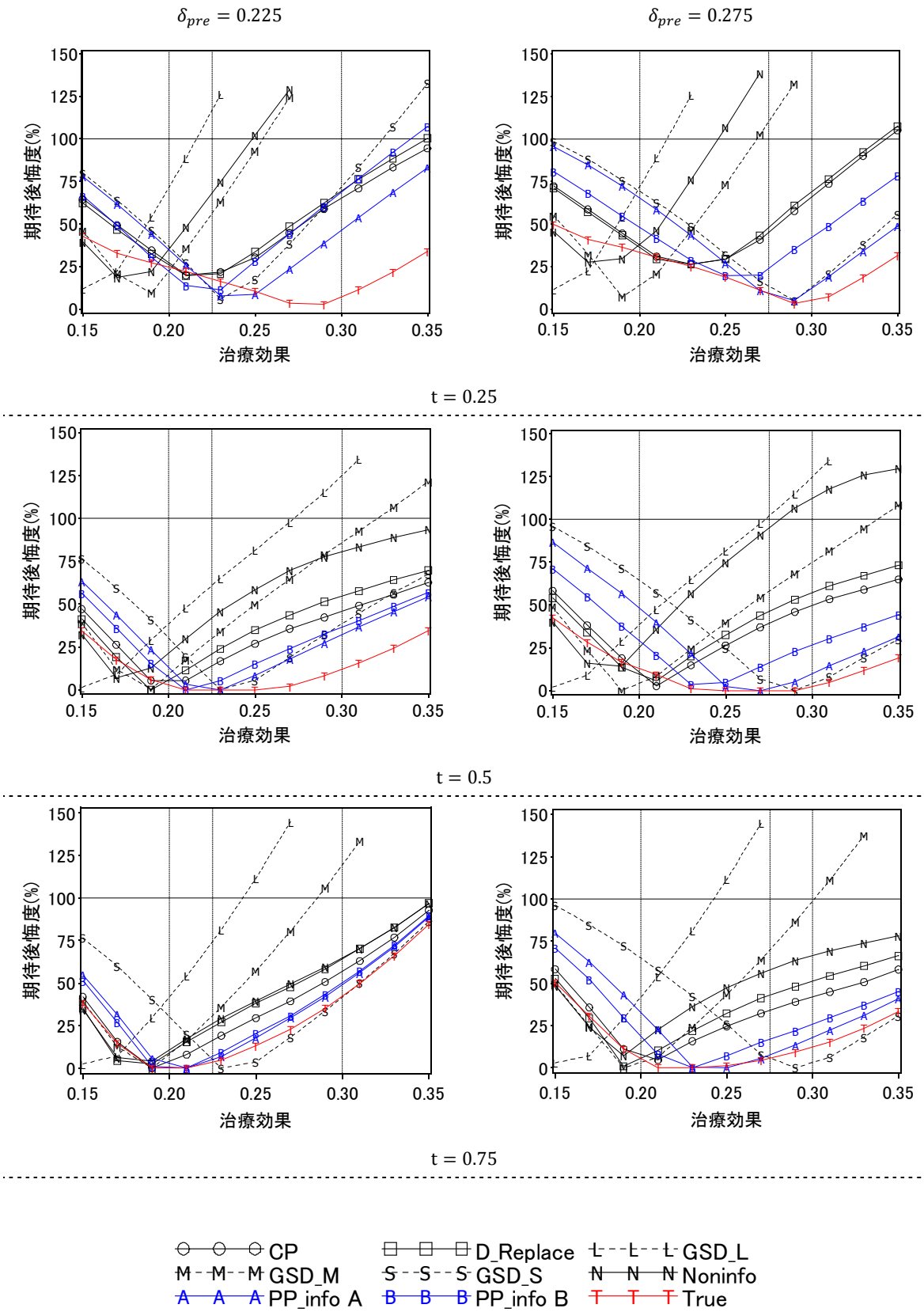
$\delta_{pre} = 0.225$
$\delta_{pre} = 0.275$
期待後悔度(%)
治療効果
t = 0.25
t = 0.5
t = 0.75
CP
D_Replace
GSD_L
GSD_M
GSD_S
Noninfo
PP_info A
PP_info B
True

図5. 期待後悔度　$ER(\%)$

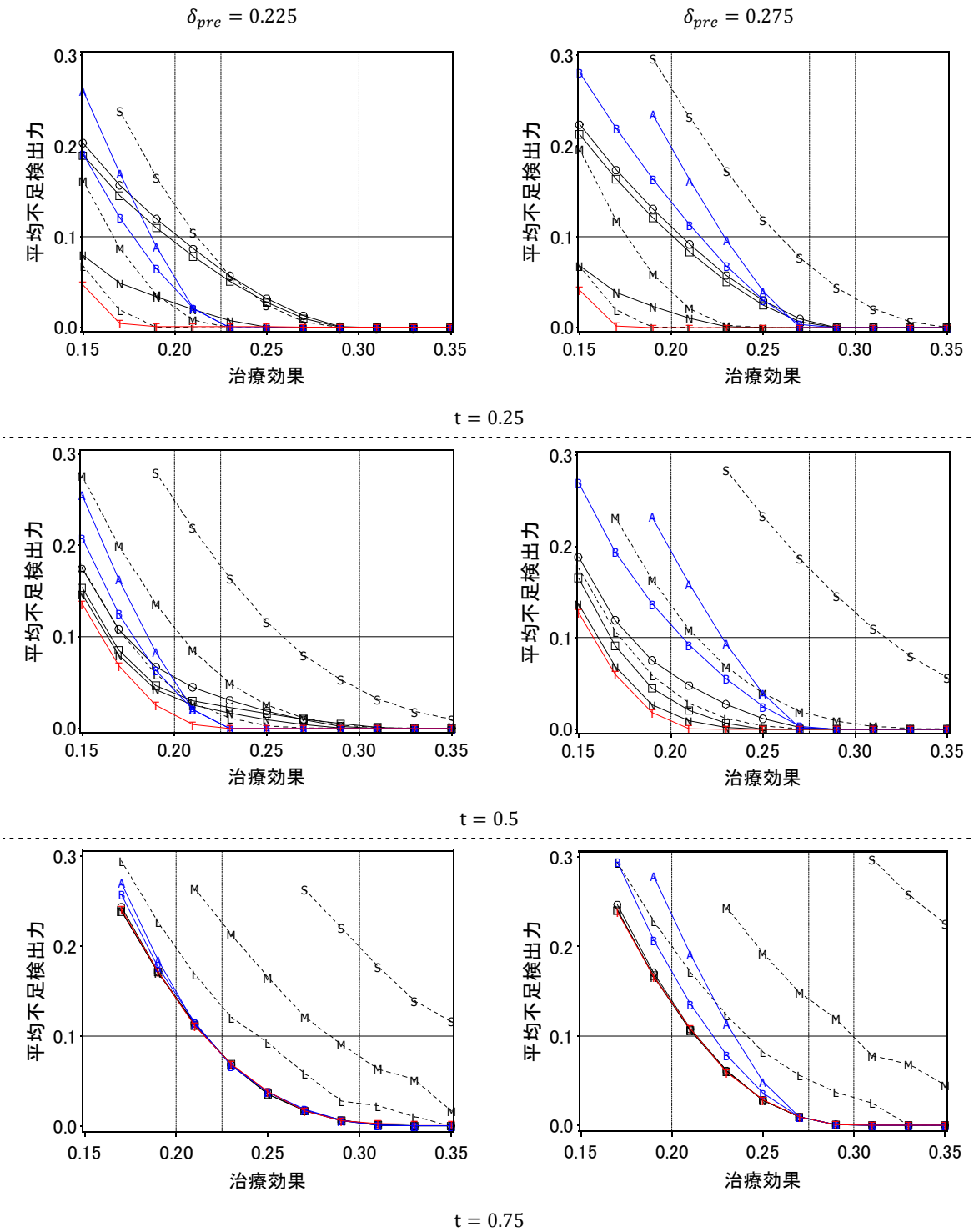

図6. 平均不足検出力　MUP

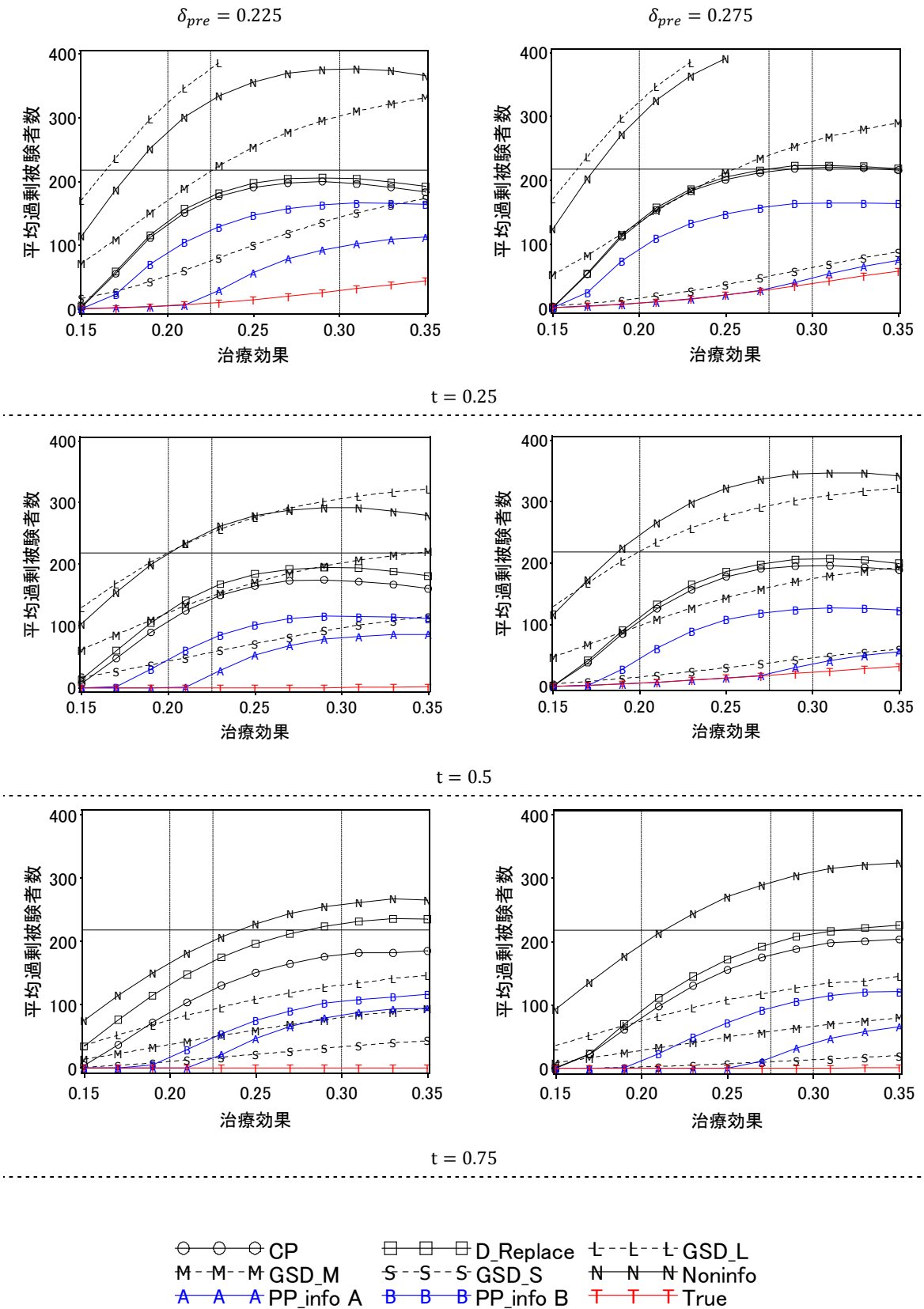

図7. 平均過剰被験者数　MOS

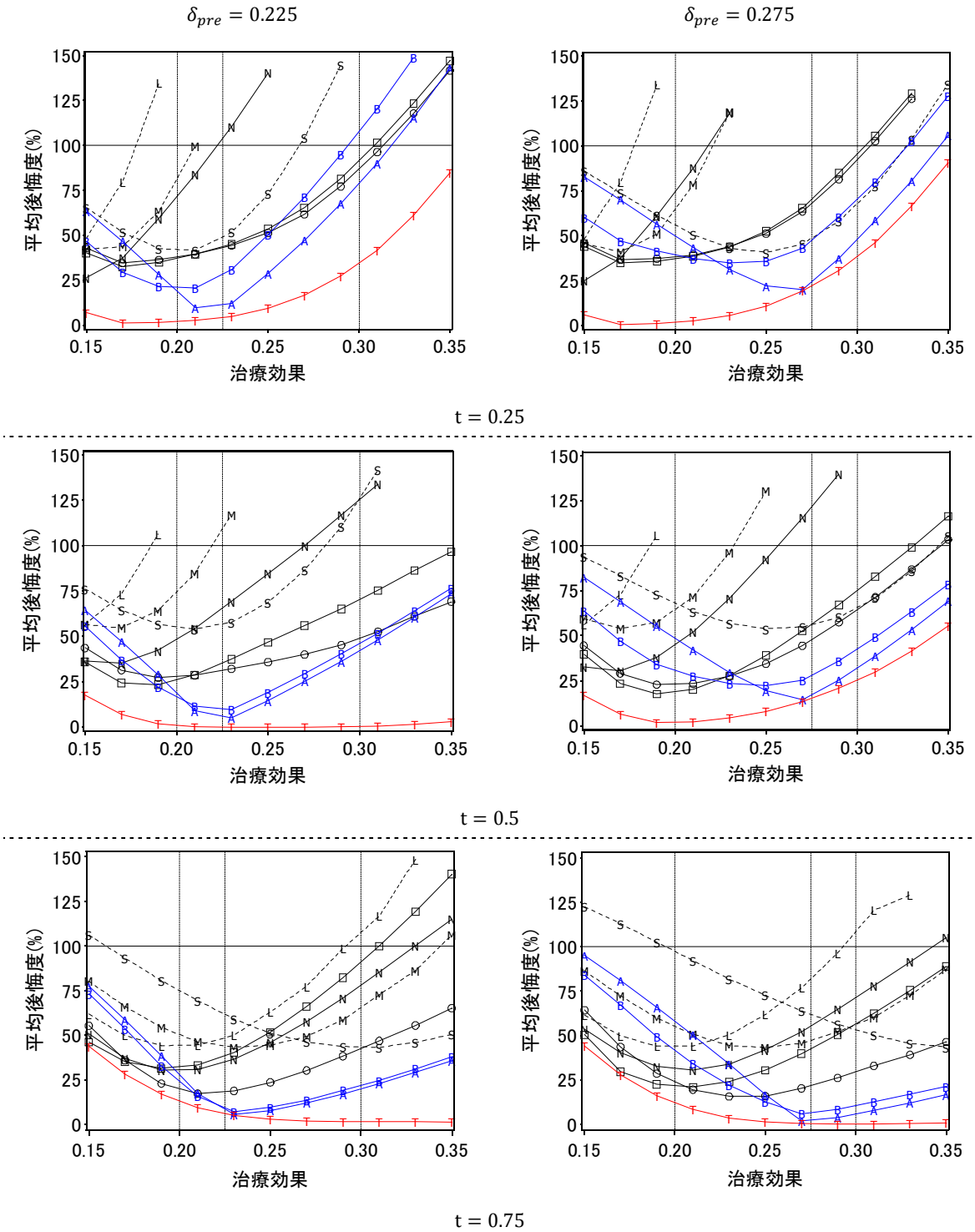
$\delta_{pre} = 0.225$
$\delta_{pre} = 0.275$
平均後悔度(%)
治療効果
t = 0.25
t = 0.5
t = 0.75
CP
D_Replace
GSD_L
GSD_M
GSD_S
Noninfo
PP_info A
PP_info B
True

図8. 平均後悔度　$MR(\%)$