博士論文

第Ⅲ相臨床試験における治療効果とその予測マーカーを評価するための統計的仮説検定とその基準に関する研究

野中　孝浩

目次

要旨

治療効果の予測因子（ predictive marker 、以下、当該因子の候補も含め、単にマーカーと呼ぶ）[1] が存在するものの、マーカーの予測性能に関する証拠が十分に信頼できる状況にない場合には、検証的試験として位置づけられる第Ⅲ相臨床試験が実施される際に、マーカーで層別するデザインで試験が実施されることが多い。このようなデザインの試験は一般にマーカー層別第Ⅲ相臨床試験などと呼ばれる。第Ⅲ相臨床試験がマーカー層別第Ⅲ相臨床試験として計画される状況下では、一般に、マーカーで定義される部分集団間で治療効果が異なり、不均一性が認められることがより自然な状況であると考えられるため、本研究では、このような自然な状況を仮定した上で、マーカー層別第Ⅲ相臨床試験における統計的仮説検定（ 多重検定 ）の枠組みと方法論を開発する。

本研究で開発するマーカー層別第Ⅲ相臨床試験の統計的仮説検定の方法では、複数の部分集団に対する仮説検定における第一種の過誤確率の制御に、強い制御ではなく弱い制御を許容する多重検定の枠組みことを提案する。

また、マーカー層別第Ⅲ相臨床試験では一般に、マーカーで定義される特定の部分集団で治療効果が示されることが期待され、実際に特定の部分集団のみで新薬などの新規治療法の治療効果が検証された場合には、当該新規治療法の適応を特定の部分集団に制限するためにマーカーが用いることになる。したがって、マーカーの治療効果

予測能、すなわち、マーカーの臨床的妥当性（ clinical validity ）の評価もマーカー層別第Ⅲ相臨床試の中で併せて行われることが期待される。そこで、本研究では、マーカーの臨床的妥当性の評価を、明確な基準に基づいて行う統計的仮説検定の方法として、マーカー陰性集団に対する無効中止基準に基づく方法を提案する（提案法）。

さらに、本研究では、全集団またはマーカー陽性集団のいずれかに対して治療効果を主張する確率として三種類の確率（ P_{overall}、P_{subgroup} および P_{success} ）を導入し、これらに基づいて、第Ⅲ相臨床試験が満たすべき統計的仮説検定の作用特性に関する評価基準も提案する。その上で、提案する作用特性に関する評価基準に基づいて提案法、すなわち、マーカー陰性集団に対する無効中止基準に基づく方法と既存の様々な統計的仮説検定の方法（ 固定順序法、フォールバック法、ハイブリッド法、交互作用検定に基づく方法 ）との比較を数値評価で行った。その結果、第Ⅲ相臨床試験を計画する時点において、マーカーの治療効果予測能の信頼性が比較的高い場合には固定順序法およびハイブリッド法が推奨された一方、低い場合には保守的な無効中止基準を用いた提案法を選択することが妥当と考えられた。

第 1 章　序論

がん患者に対する抗がん剤による薬物治療は、従来、がん細胞のみならず正常細胞を含め、活発に増殖や分裂する細胞内で起こるデオキシリボ核酸（ DNA ）やタンパク質の合成などを阻害し､非選択的に細胞毒性を示す殺細胞性薬剤が主流であったが、近年、発がんやがん細胞の増殖などに関連する特定の分子を標的とし、がん細胞の増殖などを選択的に阻害する分子標的薬剤の使用に移行している。分子標的薬剤によって治療効果が得られる患者は、特定の部分集団、すなわち、薬剤の標的となる分子に変異などが認められる患者集団に限定される可能性が高い。したがって、特にがんの分子標的薬剤の開発に際しては、治療の恩恵を受ける患者集団を同定することができる治療効果の予測因子（ predictive marker 、以下、当該因子の候補も含め、単にマーカーと呼ぶ ）を開発することが重要であると考えられている。

新薬などの新規治療法の開発においては、当該新規治療法の臨床的有用性を確立するために、臨床導入される前の最終段階で検証的試験として位置づけられる第Ⅲ相臨床試験の実施が求められることが一般的である。マーカーの予測性能に関する証拠が十分に信頼できる状況にない場合には、第Ⅲ相臨床試験が実施される際に、マーカーで層別するデザインで試験が実施されることが多く、このようなデザインの試験は一般にマーカー層別第Ⅲ相臨床試験などと呼ばれる。

マーカー層別第Ⅲ相臨床試験のデザインは、マーカーを用いずに実施される伝統的な第Ⅲ相臨床試験のデザインと比較して複雑な試験となる。したがって、マーカーとなる分子を確実に測定可能である、すなわち、マーカーの分析的妥当性（ analytical validity ）が示され、かつ、新規治療法によりマーカーで定義される特定の部分集団、例えばマーカー陽性集団（ または、マーカー陰性集団 ）で治療効果が得られることを支持する信頼性の高い証拠が十分に得られている場合には、ランダム化する患者の適格基準をマーカー陽性集団に限定するエンリッチメントデザイン [2] [3] [4] で第Ⅲ相臨床試験を実施することが最善である [5] [6] [7] [8] [9] [10] 。エンリッチメントデザインの臨床試験は、特にマーカー陽性割合が低い場合や、マーカー陰性集団に対して新規治療法の治療効果が小さい場合には、ランダム化に必要な患者数を少なくできるという観点から効率的なデザインである [5] 。

しかしながら、マーカーによっては、マーカーとなる分子を測定する際に生じる測定誤差や、マーカー陽性集団をより適切に定義できる他の閾値の存在、マーカー以外の経路等による治療効果の出現（ オフターゲット効果 ）など、第Ⅲ相臨床試験を計画する時点ではマーカーの治療効果予測能、すなわち、マーカーの臨床的妥当性（ clinical validity ）の信頼性が十分に高くなく、マーカー陰性集団が新規治療法により治療効果を得るかもしれないという可能性を排除できない場合がある。そのような場合には、ランダム化する患者の適格基準を全患者とし（ all-comers ）、かつ、マーカー陽性集団と陰性集団の両者を含めてマーカーで層別する試験（ marker-stratified all-

comers trials ）として第Ⅲ相臨床試験が実施される。このようなマーカー層別第Ⅲ相臨床試験において、全患者集団やマーカーに基づく部分集団に対して統計的仮説検定を行う方法がこれまでに多く提案されている。これらは、大きく、固定順序（ fixed-sequence ）法、アルファ（α）分割法に分類される [6] [7] [8] [9] [10] （ 第 3 章を参照 ）。

マーカーで層別する第Ⅲ相臨床試験では、新規治療法の適応をマーカー陽性集団など、マーカーで定義される特定の部分集団に限定される可能性があるため、マーカーに基づく部分集団での新規治療法の治療効果を検証することに加えて、マーカーが確かに治療効果を予測するものであるという意味でのマーカーの臨床的妥当性（ clinical validity ）を全生存期間のような真のエンドポイントに基づいて示すことも重要である。近年、全集団とマーカー陽性集団のどちらで治療効果の統計的仮説検定を行うかを決定するために、予備的に治療効果の交互作用検定を行う方法が提案された（ treatment-by-marker interaction 法 ） [10] 。この交互作用検定は、マーカーの臨床的妥当性を評価する基準に相当する。本研究では新たに、新規治療法の治療効果を検証することを目的とした第Ⅲ相臨床試験において、併せてマーカー陰性集団における治療効果に基づいてマーカーの臨床的妥当性を評価する基準を提案する [1] 。

マーカーで層別する第Ⅲ相臨床試験における統計的仮説検定法のすべてに共通す

る重要な特徴は、全集団またはマーカー陽性集団のいずれかで治療効果があることを主張できる可能性があることである。しかしながら、従来の統計的仮説検定の枠組みは、検定全体に対しての検出力といった包括的な基準が主であり、マーカー集団別の基準は十分には検討されていない枠組みである。そこで、本研究では、マーカーに基づく部分集団間での治療効果を主張する確率に基づく基準を提案する [1] 。

本論文は以下のように構成されている。第 2 章では、マーカー層別第Ⅲ相臨床試験における主要な統計解析（ 主要解析 ）のための枠組みを述べる。その中で、帰無仮説および第一種の過誤確率の制御（ 弱い制御または強い制御 ）に関する詳細な理論的根拠も述べる。 第 3 章では、既存の統計的仮説検定の方法（ 固定順序法、フォールバック法、ハイブリッド法 ）に関する概要を述べる。 第 4 章では、マーカーの臨床的妥当性の評価の基準を含んだ新しい統計的仮説検定の方法に関する概要を述べる。第 5 章では、マーカーに基づく部分集団別、すなわち、全集団またはマーカー陽性集団のいずれかに対して治療効果を主張する三種類の確率（ P_{overall} 、P_{subgroup} および P_{success} ）を導入し、当該確率に基づいて、第Ⅲ相臨床試験における統計的仮説検定の方法が満たすべき作用特性に関する評価基準を提案する [1] 。 第 6 章では、提案する基準をサンプルサイズ設計に適用し、様々な統計的仮説検定の方法を数値評価によって比較する。最後に、第 7 章では本研究の考察および結論を述べる。

第 2 章　マーカー層別第Ⅲ相臨床試験における治療効果の統計的仮説検定の枠組み

効果予測因子であることが期待されるマーカーで層別する第Ⅲ相臨床試験を想定し、生存期間に関する評価項目に基づいて新規治療と既存治療の治療効果を比較する場合を考える。マーカーの値は二値、すなわち、「陽性」と「陰性」に分けられるとする。連続的なマーカーの場合は、カットオフ値によって「陽性」か「陰性」のいずれかに分けられるものとする。また、本研究では、新規治療による治療効果がマーカー陽性集団でより期待できるものと仮定する。マーカーによる層別は、マーカーが治療効果の予測能に加えて、予後効果をもつ際にも有効である。なお、マーカーを層別因子として第Ⅲ相臨床試験を実施することにより、ランダム化されたすべての患者でマーカーが測定されていることになるため、マーカー陽性集団、陰性集団での治療効果を評価することが可能となる。

2.1 節　帰無仮説

マーカーで層別する第Ⅲ相臨床試験において、「マーカー陽性集団における治療効果の大きさが、マーカー陰性における治療効果の大きさと比較して小さくない」、と仮定することは、一般的に理に適っている。そのため、この論文（本研究）では、これを「マーカー仮説」と呼び、すべての議論においてこの仮説が成立することを前提とする。

マーカー陽性集団および陰性集団に対する「治療効果なし」の帰無仮説をそれぞれ

$H_0^{(+)}$ および $H_0^{(-)}$ で表すとする。マーカーの部分集団別の治療効果に関するマーカー仮説が成立するもとでは、部分集団別の帰無仮説（ $H_0^{(+)}$ および $H_0^{(-)}$ ）に基づいて想定される帰無仮説シナリオを考えることが自然である。具体的には、以下の 2 種類の帰無仮説シナリオが想定される [11] [12] 。なお、帰無仮説シナリオとは、想定される帰無仮説（ null ）のシナリオであり、個々の帰無仮説を意図していない。

- 帰無仮説シナリオ 1（ グローバル帰無仮説 ）： $H_0^{(+)}$ が真、かつ、 $H_0^{(-)}$ が真
- 帰無仮説シナリオ 2 ： $H_0^{(+)}$ が偽、かつ、 $H_0^{(-)}$ が真

一方、既存の多くのマーカー層別第Ⅲ相臨床試験のための統計的仮説検定の方法（ 図 1 を参照 ）では全集団に対する帰無仮説にも関心が持たれ、これを $H_0^{(O)}$ と表すこととする。しかしながら、マーカーに基づく部分集団間で治療効果が異なる可能性があるため、 $H_0^{(O)}$ は単独では使用されるべきではないと考えられる。 実際、 $H_0^{(O)}$ は、予測マーカーに基づく部分集団別の帰無仮説（ $H_0^{(+)}$ および $H_0^{(-)}$ ）と必ずしも一致しない。例えば、対照治療（ 既存治療 ）に対する新規治療のハザード比を考える。ここで、① マーカー陽性集団、② マーカー陰性集団、および、③ 全集団における治療効果のハザード比をそれぞれ① $HR^{(+)}$ 、② $HR^{(-)}$ 、および、③ $HR^{(O)}$ と表す。 両側検定において、 $H_0^{(O)}$ ： $HR^{(O)} = 1$ は、帰無仮説シナリオ 1（ $HR^{(+)} = 1$ かつ $HR^{(-)} = 1$ ）のもとでは成立するが、 帰無仮説シナリオ 2 のもとで常に成立するというわけではない（ $H^{(+)} \neq 1$ かつ $HR^{(-)} = 1$ ）。 同様に、

片側検定において、 $H_0^{(O)}$ ： $HR^{(O)} > 1$ が帰無仮説シナリオ 2 のもとで常に成立するというわけではない（ $HR^{(+)} < 1$ かつ $HR^{(-)} > 1$ ）。

以上が示唆するように、部分集団に対する帰無仮説と全集団に対する帰無仮説との間でもともと一貫性がないために、全集団での統計的仮説検定によってマーカー集団別での治療効果の判断を誤る可能性がある。特に、 $H_0^{(O)}$ の棄却が必ずしも帰無仮説シナリオ 2 の棄却とならないという事実は、全集団での統計的有意差（ $H_0^{(O)}$ の棄却 ）は、マーカー陰性集団で実際には治療効果がないにもかかわらず、治療効果があると誤って評価してしまう可能性があるという、よく知られた問題を引き起こす [10] [13] [14] [15] 。上記の議論は、統計的仮説検定の方法を評価する際にも当てはまる（ 第 6 章 ）。すなわち、統計的仮説検定の作用特性は、全集団での治療効果というよりはむしろ部分集団での治療効果に基づいて評価されるべきである。

上記の議論は、予測マーカーで層別する第Ⅲ相臨床試験における全集団での統計的仮説検定（ $H_0^{(O)}$ の仮説検定 ）の役割を見直すことにつながる。すなわち、全集団での統計的仮説検定（ $H_0^{(O)}$ の仮説検定 ）は、部分集団での帰無仮説（ $H_0^{(+)}$ および $H_0^{(-)}$ ）を直接的に検定するものではないが、（ マーカー陽性集団に加えて ）マーカー陰性集団にも適応を拡げる、すなわち、全集団で適応を取得するための“作業基準”または“操作基準”として機能していると考えることができる。これは、 $H_0^{(O)}$ の仮説検定が“主要解析”として位置づけられるマーカーを用いない伝統的な第Ⅲ相臨床

試験の枠組みの場合とは対照的である。この点は次節で改めて議論する。

2.2 節　弱い制御または強い制御

一般に、多重検定における第一種の過誤確率の制御には弱い制御と強い制御がある[16]。マーカー仮説（治療効果はマーカー陰性集団と比較してマーカー陽性集団で高い）が成立するもとで、弱い制御およびと強い制御はそれぞれ以下のことを意味する。

- 弱い制御：

  $H_0^{(+)}$ と $H_0^{(-)}$ の両方が成立しているもとで、$H_0^{(+)}$ または $H_0^{(-)}$ のいずれかを誤って棄却してしまう確率（第一種の過誤確率）を制御する。すなわち、帰無仮説シナリオ1（グローバル帰無仮説）のもとでのみ第一種の過誤確率を制御する。

- 強い制御：

  $H_0^{(+)}$ または $H_0^{(-)}$ を含む帰無仮説（$H_0^{(+)}$、$H_0^{(-)}$ および $H_0^{(+)} \cap H_0^{(-)}$）が成立しているという仮定のもとでの仮説検定において、$H_0^{(+)}$、$H_0^{(-)}$ または $H_0^{(+)} \cap H_0^{(-)}$ のいずれかを誤って棄却してしまう確率（第一種の過誤確率）。マーカー仮説のもとでは、帰無仮説シナリオ1（グローバル帰無仮説）と帰無仮説シナリオ2の両方に対して第一種の過誤確率を制御する。

なお、弱い制御であっても、マーカー仮説が成立するもとで、少なくともマーカー陽性集団での治療効果の存在に関しては第一種の過誤確率を厳格に制御しているとみなすことができる点は注目に値する。

一方、強い制御の場合に追加される帰無仮説シナリオ2 のもとでの制御は、マーカー陰性集団に対して治療効果がない場合に、新規治療の適応を誤ってマーカー陰性集団に拡げてしまう、すなわち、全集団で適応を取得してしまう第一種の過誤確率の制御と捉えることができる点も注目に値する。

一般的には、弱い制御のもとでは、帰無仮説シナリオ2 のもとでの第一種の過誤確率は制御されないので、強い制御、すなわち、帰無仮説シナリオ1 と帰無仮説シナリオ 2 の両方のもとでの第一種の過誤確率の厳格な制御が受け入れられている [11] [12] 。しかしながら、帰無仮説シナリオ1 が成立しているもとで第一種の過誤確率を厳格に制御（ 弱い制御 ）している状況のもとで、帰無仮説シナリオ2 が成立しているもとでも第一種の過誤確率を厳格に制御する（ つまり、強い制御が ）必要であるかについては議論がある。

具体的には、マーカー陽性集団に対して極めて大きな治療効果が示されている場合に、他に有効な治療法がないマーカー陰性の進行癌に対する $H_0^{(-)}$ の統計的仮説検定において第一種の過誤確率の制御を緩めることは考慮に値するだろう。実際には、帰無仮説シナリオ2 が成立しているもとでの第一種の確率の制御の程度は、多くの外的

な要因（ ① マーカーの分析性能、② マーカーの陽性割合、③ 新規治療によって発現する可能性のある有害事象、④ 他の治療法の有無、⑤ 治療を受けるために必要な費用など ）に依存して個別に検討すべきものと考えられる [10] 。これは、治療効果の検証と、新規治療が使用されるべきである患者の同定を分離するアプローチ [13] に近いものである。そこでは、治療効果の検証は、帰無仮説シナリオ 1（ グローバル帰無仮説 ）が成立しているもとでのみ厳格に第一種の過誤確率を制御した上で（ 弱い制御 ）、統計的に有意な結果が示された場合にのみ、新規治療を使用すべき患者集団の同定が別途行われる [13] 。

2.3 節　提案する多重検定の枠組み

以上の議論のもと、本研究で考える多重検定の枠組みでは、弱い制御を許容する。

ただし、本研究では、想定しうるマーカー部分集団別の治療効果を仮定した上で、帰無仮説シナリオ 2 が成立しているもとでの第一種の過誤確率（ さらには検出力 ）を評価または一定水準以下になるように評価・モニタリングすることも提案する [1] 。これより、第一種の過誤確率の弱い制御を許容した枠組みであっても、マーカー部分集団別のもっともらしい治療効果のもとで、第一種の過誤や検出力に関して望ましい作用特性を達成することができる第Ⅲ相臨床試験をデザインすることが可能になる（ 第 5 章を参照 ）。

第 3 章　マーカー層別第Ⅲ相臨床試験のための統計的仮説検定の方法

マーカー層別第Ⅲ相臨床試験における既存の主な統計的仮説検定の方法を 図 1 に示す。これらは、治療効果を予測する性能の観点でマーカーの信頼性の程度が異なる [6] [7] [8] [9] [10] 。また、これらの方法は、全集団またはマーカー陽性集団のいずれかに対して治療効果を主張することを可能とする。

以降、統計的仮説検定の方法として、大きく、固定順序法（ 3.1 節を参照 ）、アルファ（α）分割法（ 3.2 節を参照 ）およびハイブリット法（ 3.3 節を参照 ）に分けて紹介する。なお、広義には、ハイブリッド法（ 3.3 節を参照 ）、交互作用検定に基づく方法（ 4.1 節を参照 ）、本研究における提案法（ 4.2 節を参照 ）もアルファ（α）分割法に含まれるが、本研究では、アルファ（α）分割法としては co-primary 解析とフォールバック法を意図する。

すべての公称有意水準は片側で表記する。また、以降の検討に用いる統計的仮説検定の方法は、（ 生存期間に係るイベント数における ）マーカー陽性割合（ p ）の情報が必要であるが、解析の開始前に統計解析計画を更新し、公称有意水準を再計算する際には、マーカーの陽性割合は既知であると考えることが可能である。

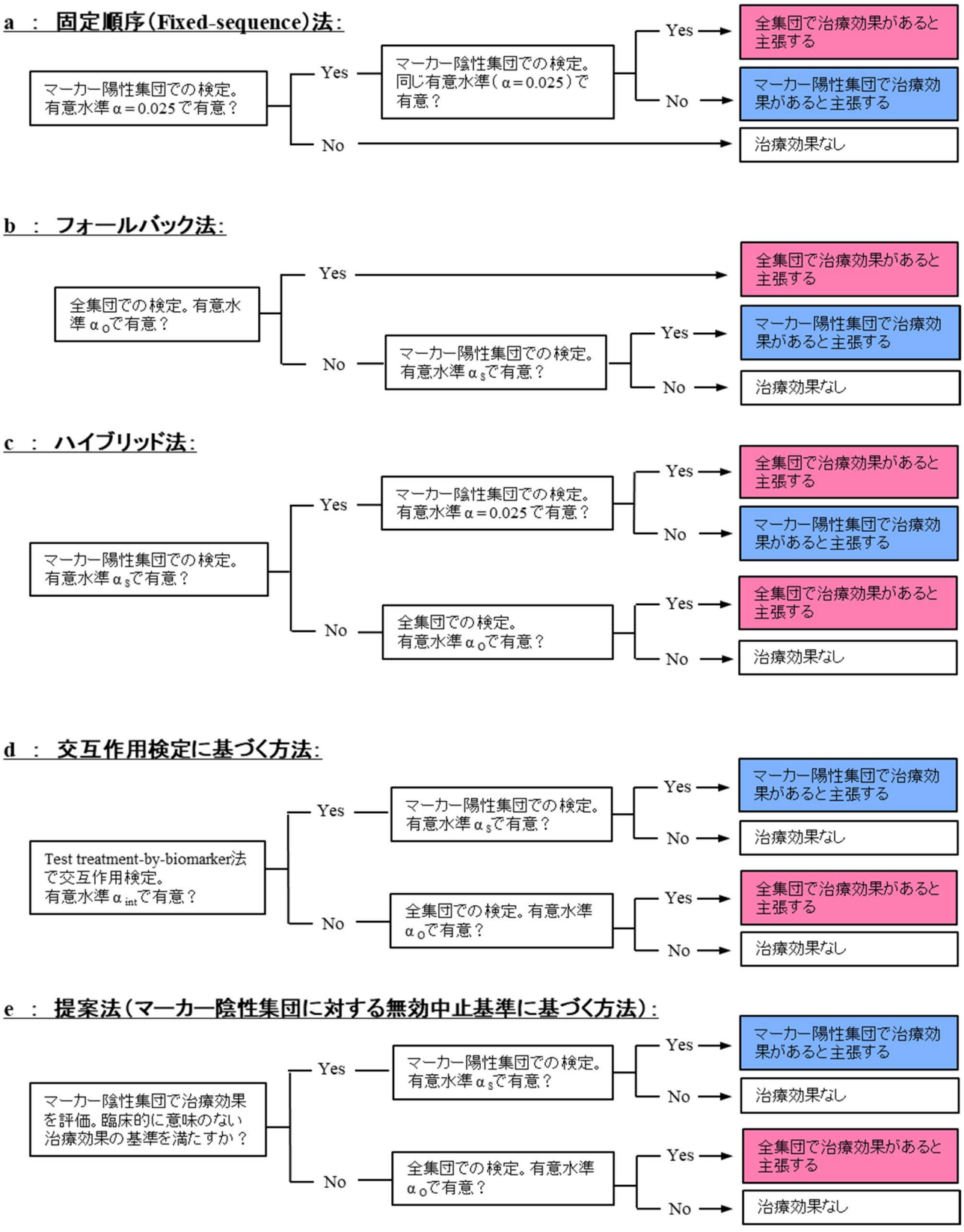


図 1　様々なマーカー層別第Ⅲ相臨床試験のための仮説検定の方法

3.1 節　固定順序（ fixed-sequence ）法

「マーカー陰性集団での治療効果と比較してマーカー陽性集団での治療効果の方が大きい」というマーカー仮説（ 第 2 章を参照 ）が成立しているもとで、さらに、マーカー陽性集団に対して新規治療の治療効果があることを支持する比較的強い情報が得られている場合、すなわち、マーカーの治療効果予測能に関して十分なエビデンスがある場合には、最初にマーカー陽性集団での治療効果を仮説検定する固定順序法を採用するのが合理的である [9] [17] 。すなわち、マーカー陽性集団での仮説検定で有意水準 $\alpha = 0.025$ で統計的に有意である場合、マーカー陰性集団に対して同じ有意水準（ $\alpha = 0.025$ ）を用いて仮説検定を行う（ 17ページ、図 1 a を参照 ）。

例えば、この方法は、化学療法未治療の転移性結腸・直腸癌に対する FOLFOX（ フルオロウラシル、ロイコボリンおよびオキサリプラチンの併用投与 ）単独と FOLFOX へのパニツムマブ（ 遺伝子組換え ）の上乗せを比較したランダム化第Ⅲ相臨床試験に用いられた [18] 。より具体的には、最初に KRAS 野生型腫瘍（ マーカー陽性集団 ）の患者における無増悪生存期間（ PFS ）に関して仮説検定が行われ、KRAS 野生型腫瘍の患者で統計的に有意であった場合には、KRAS 変異型腫瘍（ マーカー陰性集団 ）の患者で仮説検定を行うこととされた。このように、固定順序法は、帰無仮説シナリオ1 および帰無仮説シナリオ2 の両方の第一種の過誤確率を厳格に制御する方法である（ 強い制御 ） [12] 。 検出力は、エンリッチメント試験のように、最初の仮説検定、すなわち、マーカー陽性集団に対する仮説検定で決定される。したがっ

て、固定順序法は、マーカー陽性集団に対してかなり治療効果が大きいと想定される場合に良好に機能することが期待される [10] 。

3.2 節　アルファ（α）分割法

マーカーが新規治療の治療効果をうまく予測できるか否かについて強い証拠がなく、より広範囲の患者において新規治療の治療効果が得られる可能性が十分あるという、より一般的な状況を考える。そのような場合、アルファを分割し、一部を全集団、残りを部分集団に配分するアルファ分割法を採用するのが合理的である。

例えば、進行性の非小細胞肺癌に対する白金系抗悪性腫瘍剤による治療後に病勢進行を認めなかった患者に対する維持療法でのエルロチニブの治療効果を評価した SATURN 試験では、ランダム化後の無増悪生存期間（ PFS ）に関して全集団に対する有意水準は $\alpha_{\text{。}} = 0.015$ 、EGFR タンパクが過剰発現した腫瘍の患者（ マーカー陽性集団 ）に対する有意水準は $\alpha_s = 0.01$ で仮説検定が行われた [19] 。なお、この仮説検定の効率は、全集団と部分集団との間での検定統計量の相関関係を考慮した、より緩めの有意水準を用いることで改善することが可能である [20] [21] 。

上記のように二つの仮説検定（ 全集団での仮説検定とマーカー陽性集団での仮説検定 ）にアルファを分割する方法は、それらを共に主要解析と位置づける、すなわち、co-primary 解析として位置づけることを意図しているが、一般に、部分集団での仮説検定（ の結果 ）よりも全集団での仮説検定（ の結果 ）により重きが置かれる傾向

がある。おそらくこの傾向は、全集団での仮説検定は、マーカーを用いない伝統的な第Ⅲ相臨床試験での主要解析であることを反映しているものと考えられる。さらに、全集団で治療効果が示されることの方がより望ましいとみなされることも考えられる。このような状況は、最初に全集団で仮説検定を行い、統計的に有意でない場合に残りの有意水準でマーカー陽性集団での仮説検定を行う 2 段階の手順である、フォールバック法に近いものとなる [9] [17] （ 17ページ、図 1 b を参照 ）。 実際に、 co-primary 解析で全集団での仮説検定の方が優先される状況では、フォールバック法とほぼ同じ作用特性になると考えられる。

しかしながら、マーカー陽性集団での治療効果が大きく、マーカー陰性集団で治療効果がない場合でも全集団に関する仮説検定で統計的に有意になる可能性が十分あるため、co-primary 解析やフォールバック法はマーカー陰性集団に対しても誤って治療効果を主張してしまうことが懸念される（ 2.2 節を参照 ）。したがって、マーカー陰性患者を過剰治療から保護するために（ 主要解析には含まれていなかった ）マーカー陰性集団における治療効果の評価が別途必要となる [10] [13] [14] [15] 。

3.3 節　ハイブリット法

Co-primary 解析やフォールバック法に関しての前述の問題は、帰無仮説シナリオ2（ $H_0^{(+)}$ が偽、 かつ $H_0^{(-)}$ が真 ）が成立しているもとでの試験あたりの第一種の過誤確率（ study-wise alpha rate ）の増大として見ることができる。この問題を解決する

ために、 Freidlin らは、マーカー逐次検定（ marker sequential test; MaST ）法と呼ばれる逐次手順法を提案している [12] 。

MaST 法は、固定順序法と同様に、帰無仮説シナリオ1 が成立しているもとのみではなく、帰無仮説シナリオ2 が成立しているもとでの試験あたりの第一種の過誤確率も厳格に制御することを意図している。より具体的には、最初に、マーカー陽性集団での治療効果に関する仮説検定を有意水準 $\alpha_s = 0.022$ を用いて行う。治療効果が統計的に有意である場合には、マーカー陰性集団での治療効果の仮説検定を有意水準 0.025 で行う。一方、最初のマーカー陽性集団での治療効果の仮説検定が有意でない場合には、全集団での治療効果の仮説検定を有意水準 $\alpha_o = 0.003$ （ $= 0.025 - 0.022$ ）で行う（ 17ページ、図 1 c を参照 ）。

この MaST 法は、固定順序法とアルファ分割法との混合（ ハイブリッド ）法とみなすことができる。Co-primary 解析やフォールバック法と同様に MaST 法の効率は、全集団とバイオマーカー陽性集団の間での検定統計量の相関関係を考慮することで多少改善することが可能である。

なお、2.2 節で議論したように、本研究で提案する枠組みでは、帰無仮説シナリオ2 が成立しているもとでは厳格な第一種の過誤確率の制御を求めず、第一種の過誤確率の弱い制御を許容する。したがって、以降の本研究では、有意水準をより緩く変更（ 特に、部分集団に対しては 0.022 より小さく、全集団に対しては 0.003 より大きく ）

した MaST 法（ すなわち、固定順序法とフォールバック法のハイブリット法 ）を用いて検討する。

第 4 章　マーカーの臨床的妥当性に関する基準を含む統計的仮説検定の方法

第 4 章では、マーカーの臨床的妥当性に関する基準を含む統計的仮説検定の方法について論じる。

4.1 節　治療とマーカーの交互作用検定に基づく方法

マーカーの臨床的妥当性に関する基準を含む統計的仮説検定の方法について検討するにあたって、別の方法を検討することが可能である。この方法では、新規治療の治療効果をマーカー陽性集団または全集団のどちらで仮説検定を行うかを決定するために最初にマーカーに関する交互作用検定を予備的に行うのが特徴的である [10] 。マーカー陽性集団の治療効果は陰性集団のそれよりも大きいという片側の交互作用検定において有意差が認められた場合には、マーカー陽性集団で治療効果の仮説検定を行い、 一方、交互作用検定で有意差が認められない場合には、全集団で治療効果の仮説検定を行う（ 17ページ、図 1 d を参照 ）。帰無仮説シナリオ1（ グローバル帰無仮説 ）が成立しているもとで、試験あたりの第一種の過誤確率を厳格に制御するために三種類の仮説検定（ 交互作用検定、マーカー陽性集団での治療効果の仮説検定、全集団での治療効果の仮説検定 ）における有意水準が決定される。第 3 章に纏めた統計的仮説検定の方法と比較した場合、この方法の魅力的な特徴は、治療とマーカーの交互作用検定が、治療効果の仮説検定を行うべき適切な集団を選択する指針を提供するという役割を果たすだけでなく、マーカーの効果予測能、つまり、臨床的妥当性を示

す基準の一つとして見なせることである。つまり、マーカー陽性集団で治療効果が有意であった場合には、その集団を規定するマーカーの臨床的妥当性も同時に示されていることになる。

4.2 節　マーカー陰性集団に関する無効中止基準に基づく方法（ 提案法 ）

上記 4.1 節で述べた治療とマーカーの交互作用検定に基づく方法の欠点は、交互作用検定が必ずしも適切な適応を得ることにつながらないということである。特に、マーカーに関する交互作用検定は、臨床的に意味のある交互作用、すなわち、マーカー集団間での臨床的に意味のある治療効果の差を検出するための基準に必ずしも対応していない。

本研究では、統計学的な交互作用と区別して、臨床的に意味のある交互作用に関する用語として以下の二つを導入する。

- 臨床的な質的交互作用（ clinically qualitative interaction; cQL ）:

  マーカー陽性集団では臨床的に意味のある治療効果が存在する一方、マーカー陰性集団では臨床的に意味のある治療効果が存在しないこと。つまり、治療効果の観点からは、新規治療はマーカー陽性集団だけに対して推奨される。

- 臨床的な量的交互作用（ clinically quaantitative interaction; cQT ）:

  マーカー陽性集団、陰性集団ともに臨床的に意味のある治療効果が存在するが、両集団間で治療効果の大きさが異なること。ただし、マーカー仮説が成立してい

るもとでは、マーカー陰性集団での治療効果がより小さい。つまり、治療効果の観点からは、新規治療は全集団に対して推奨される。

治療とマーカーの交互作用検定は、単に、マーカー集団間での治療効果の差を検出するものであり、その差が臨床的な量的交互作用（ cQT ）なのか、あるいは、臨床的に質的交互作用（ cQL ）なのかを判定するものではない。特に問題となるのは、臨床的な量的交互作用（ cQT ）を検出してしまうことでマーカー陽性集団での治療効果を仮説検定し、それが有意となることとでマーカー陽性集団のみで治療効果を主張することになってしまうことである。

この問題に対処する一つの方法は、 cQL に関する明示的な基準を使用することである。具体的には、マーカー陽性集団における治療効果は、臨床的に意味のある治療効果の水準（ 臨床的に意味のある一定の大きさの治療効果を表す閾値 ）よりも大きく、一方で、マーカー陰性集団における治療効果は臨床的に意味のある最小の治療効果の水準（ 別の小さな閾値 ）よりも小さい、といった基準である。仮にこのような基準が満たされた場合には、マーカー陽性集団での治療効果に関する仮説検定を行い、一方、満たされない場合には、全集団での治療効果に関する仮説検定を行うという手順となる。しかしながら、マーカー陽性と陰性集団のそれぞれで治療効果の閾値を設定する必要がある。マーカー陰性集団で用いる臨床的に意味のある最小の治療効果の

設定についてはサンプルサイズ設計等でよく議論されるが、マーカー陽性集団で用いる最小の治療効果よりも大きな効果を表す閾値を指定するのはより難しい。具体的には、期待される効果サイズを指定することが考えられるが、臨床試験の開始に先立ってこれを指定するのは難しい。

その上、この問題と部分的に関連して、マーカー陽性集団に対して新規治療の治療効果が臨床的に意味のある大きさで存在するものの、それが事前に指定した治療効果の閾値（ 例えば、期待される治療効果サイズ ）よりも大きくなく、マーカー陰性集団では臨床的に意味のある治療効果が存在しない場合には、交互作用検定の基準を満たすことに失敗する可能性が高くなる。これは、マーカー陰性集団で臨床的に意味のある治療効果が存在しないにもかかわらず、全集団で治療効果の仮説検定が行われる可能性が高くなることを意味する。

以上の方法の問題を回避するために、本研究では、マーカー陰性集団のみに対して治療効果を評価するというより簡単な基準を提案する [1] 。すなわち、マーカー陰性集団での治療効果が臨床的に意味のある最小の治療効果よりも小さい場合には、マーカー陽性集団に対して治療効果の仮説検定を行う。一方、臨床的に意味のある最小の治療効果よりも小さくない場合には､全集団で治療効果の仮説検定を行う（ 17ページ、図 1e を参照 ）。この基準を、マーカーの臨床的妥当性（ 治療効果予測能 ）に関する証拠として解釈できるものとするために、本研究では、マーカー陰性集団での治療

効果が少なくとも臨床的に意味のある最小の水準以上であるベイズ流事後確率に基づく基準を提案する [1] 。この基準は中間解析の無効中止基準としてよく用いられるものである [22] が、他の基準を用いてもよい。ただし、臨床試験の開始に先立って事前に規定する必要がある。

例えば、生存期間を評価する第Ⅲ相臨床試験を考える。その際、一般的に、臨床的に意味のある最小の治療効果の大きさを $\mathrm{HR}^{(-)}=0.8$ 、または、対数ハザード比 $\delta=\log(0.8)$ と設定することが妥当である。より厳しい値として、$\delta=\log(0.7)$ などとしてもよい。臨床的に意味のある最小の治療効果の大きさは、疾患の特性、標準治療を含めた既存治療、想定される新規治療の安全性プロファイルや費用等を踏まえて臨床試験の計画時に設定する。非劣性試験として実施する場合の非劣性マージンを参考にしてもよいだろう。マーカー陰性集団での治療効果の判定基準として、臨床的に意味にある最小の治療効果 δ よりも大きな治療効果が存在する事後確率がある小さな値 γ （ $=$ 0.05 ～ 0.2 ）以下となるとき、マーカー陰性集団で臨床的に意味のある治療効果は存在しないと判定する [22] 。治療効果の大きさに関して無情報の事前分布を指定する場合、この条件は以下のように表される。

$$\hat{\theta}^{(-)} > c^{(-)} \quad \text{および} \quad c^{(-)} = \delta + z_{\gamma}\sqrt{V^{(-)}} \tag{1}$$

γ　　：　事後確率に対する閾値（0.05 ～ 0.2）

$\hat{\theta}^{(-)}$　：　推定対数ハザード比

$V^{(-)}$　：　マーカー陰性集団における $\hat{\theta}^{(-)}$ の推定分散

Z_{γ}　　：　標準正規分布における上側 γ 点

ここで、事後確率に対する閾値 γ に小さい値（ 0.05 ～ 0.2 ）を使用することで、（ マーカー陰性集団での無効中止基準が満たされた後に ）マーカー陽性集団に対する治療効果が統計的に有意になれば、マーカー陽性集団に適用を絞る際のマーカーの臨床的妥当性に関する証拠として役に立つ基準となる。

表 1 は、検定統計量の漸近分布に基づいて、帰無仮説シナリオ1（ グローバル帰無仮説 ）が成立しているもとで、試験あたりの第一種の過誤確率（ study-wise alpha rate ）を $\alpha = 0.025$ に制御することが可能となる全集団での治療効果の仮説検定の有意水準 α_o、部分集団での治療効果の仮説検定の有意水準 α_s を纏めたものである（ 第 8 章を参照 ）。 提案した基準では、 α_s の値が $\alpha = 0.025$ より大きいことは興味深い。このように、 $\alpha = 0.025$ より大きい第一種の過誤確率は、 $H_0^{(+)}$ および $H_0^{(O)}$ に対する強い制御を伴う多重検定では許容されない。以上の方法で強い制御を満たす、より厳しい有意水準を用いることは可能であるが（ 例えば、閉手順の方法 [21] [23] を採用する ）、その代償として検出力の低下を招く。 同様に、第 2 章で論じたよう

に、そもそも $H_0^{(O)}$ の仮説検定には、マーカーによる部分集団別の帰無仮説 $H_0^{(+)}$ および $H_0^{(-)}$ と整合しないという問題があり、そのような性質をもつ $H_0^{(O)}$ を含む多重検定において厳格な制御を行うこと自体、あまり大きな価値を見いだせないと考えられる。

表 1　帰無仮説シナリオ 1（グローバル帰無仮説）のもとでの試験あたりの第一種の過誤確率 α を 0.025 とする α_o および α_s

δ	γ	$E^{(-)}$	$c^{(-)}$	α_o	α_s
log (0.7)	0.05	100	−0.028	0.005	0.0360
				0.010	0.0271
				0.020	0.0092
		300	−0.167	0.005	0.0222
				0.010	0.0181
				0.020	0.0111
	0.15	100	−0.149	0.005	0.0260
				0.010	0.0197
				0.020	0.0077
		300	−0.237	0.005	0.0224
				0.010	0.0204
				0.020	0.0175
log (0.8)	0.05	100	0.106	0.005	0.0670
				0.010	0.0503
				0.020	0.0168
		300	−0.033	0.005	0.0326
				0.010	0.0245
				0.020	0.0084
	0.15	100	−0.016	0.005	0.0376
				0.010	0.0282
				0.020	0.0095
		300	−0.103	0.005	0.0247
				0.010	0.0189
				0.020	0.0079

検定統計量の漸近分布と $V^{(-)} = 4 / E^{(-)}$ を仮定する。ここで $E^{(-)}$ はマーカー陰性集団におけるイベント数（ 第 8 章を参照 ）。マーカー陽性割合を $p = 0.4$ とした場合で、$c^{(-)}$ は 4.2 節における式（1）で定義される。

臨床的な意味に関する同様の基準をマーカー陽性集団に組み込むことも可能である。そのような基準は、マーカー仮説が成立しているもとで、マーカー陽性集団と陰性集団の両方に対する無効中止の基準として取り扱われることになるだろう（ [9] を参照 ）。すなわち、これは試験全体に対する無効中止基準という位置づけとなり、提案法で用いているマーカー陰性集団に対する効果判定基準（ 無効中止基準 ）が「患者選択」の基準として位置づけられていることとは異なる。

一方、多くの臨床試験（ 第Ⅲ相臨床試験 ）の中間解析において、新規治療の臨床的な意味に基づく無益中止基準は、典型的には非公式なものとして取り扱われ、保守的な解析として主要解析に取り込むことは明確にはなされていない。この種の非公式の無益性の中止基準は、第 3 章 および 第 4 章で述べた統計的仮説検定の方法のすべてに共通して適用できる。

4.3 節　統計的仮説検定の方法の例

第 3 章および第 4 章で述べた様々な統計的仮説検定の方法（ 17ページ、図 1 を参照 ）を、転移性結腸・直腸癌患者を対象に、 FOLFIRI（ フルオロウラシル、ロイコボリンおよびイリノテカンの併用投与 ）単独（ 対照群 ）と FOLFIRI へのパニツムマブ（遺伝子組換え）の上乗せ（ 新規治療群 ）を、無増悪生存期間（ PFS ）を主要評価項目として比較するランダム化第Ⅲ相臨床試験に適用した [24] 。

まず、臨床試験論文 [24] で報告されているマーカー陽性（ KRAS 野生型腫瘍 ）、

陰性集団（ KRAS 変異型腫瘍 ）別の治療効果（ ハザード比 ）の推定値とその分散の推定量（ あるいは 95% 信頼区間 ）、PFSのイベント数（%）、中央値［95% 信頼区間］（カ月）およびハザード比［95% 信頼区間］の情報を基に 表 2 の臨床試験成績が得られた。

表 2　臨床試験論文 [24] で報告されている情報に基づく臨床試験成績

	マーカー陽性集団 （KRAS 野生型腫瘍） 597例 マーカー陽性割合　0.55		マーカー陰性集団 （KRAS 変異型腫瘍） 486例 マーカー陰性割合　0.45		全集団 1,083例	
	新規治療群 303例	対照群 294例	新規治療群 238例	対照群 248例	新規治療群 541例	対照群 542例
PFSのイベント数（%） 中央値（カ月） ［95% 信頼区間］	178（59%） 5.9 ［5.5, 6.7］	203（69%） 3.9 ［3.7, 5.3］	162（68%） 5.0 ［3.8, 5.6］	161（65%） 4.9 ［3.6, 5.5］	－ 	－
PFSのハザード比 ［95% 信頼区間］	0.73 ［0.59, 0.9］		0.85 ［0.68, 1.06］		0.78 ［0.67, 0.91］	
PFSの対数ハザード比 ［95% 信頼区間］	－0.31 ［－0.53, －0.11］		－0.16 ［－0.39, －0.058］		－0.25 ［－0.40, －0.094］	
標準誤差	0.11		0.11		0.079	
分散	0.011		0.013		0.0062	
片側 P 値	0.0016		0.075		0.00079	

また、各仮説検定の有意水準設定（ 第 8 章を参照 ）のためにマーカーの陽性割合が既知であれば、第 3 章 および 第 4 章で述べたすべての統計的仮説検定の方法を適用した際の結果を得ることができる（ 表 3 を参照 ）。

表 3　様々な統計的仮説検定の方法を適用した際の結果

	固定順序法	フォールバック法	ハイブリッド法	交互作用検定に基づく方法	提案法 $\delta = \log(0.8)$ $\gamma = 0.15$
集団	マーカー陽性集団	全集団	マーカー陽性集団	マーカー 陽性 対 陰性	マーカー陰性集団
有意水準	$\alpha = 0.025$	$\alpha_{\mathrm{o}} = 0.015$	$\alpha_{\mathrm{s}} = 0.022$	$\alpha_{\mathrm{int}} = 0.05$	$\theta^{(-)} = -\log \mathrm{HR}^{(-)} = -0.16$ ∧ $c^{(-)} = -0.1064$ ↓ 基準を満たさない
検定統計量	$Z^{(+)} = -2.95$	$Z^{(\mathrm{O})} = -3.16$	$Z^{(+)} = -2.95$	$Z^{(\mathrm{int})} = -0.98$	
片側 P 値	0.016	0.00079	0.0016	0.16	
帰無仮説	棄却 （有意である）	棄却 （有意である）	棄却 （有意である）	採択 （有意でない）	
⇓	⇓	⇓	⇓	⇓	⇓
集団	マーカー陰性集団	⇓	マーカー陰性集団	全集団	全集団
有意水準	$\alpha = 0.025$	⇓	$\alpha = 0.025$	$\alpha_{\mathrm{o}} = 0.015$	$\alpha_{\mathrm{o}} = 0.015$
検定統計量	$Z^{(-)} = -1.44$	⇓	$Z^{(-)} = -1.44$	$Z^{(\mathrm{O})} = -3.16$	$Z^{(\mathrm{O})} = -3.16$
片側 P 値	0.075	⇓	0.075	0.0008	0.0008
帰無仮説	採択 （有意でない）	⇓	採択 （有意でない）	棄却 （有意である）	棄却 （有意である）
⇓	⇓	⇓	⇓	⇓	⇓
結論	マーカー陽性集団	全集団	マーカー陽性集団	全集団	全集団

主要評価項目と設定された PFS に対して、KRAS 野生型腫瘍の患者（ マーカー陽性集団 ）およびKRAS 変異型腫瘍の患者（ マーカー陰性集団 ）でのハザード比［95% 信頼区間］の推定値は、それぞれ 0.73［0.59, 0.90］ および 0.85［0.68, 1.06］であった。また、全体集団でのハザード比［95% 信頼区間］の推定値は、部分集団での推定値に基づいて 0.78［0.67, 0.91］と算出した（ 第 8 章を参照 ）。観察された PFS のイベントは、マーカー陽性集団および陰性集団でそれぞれ 381 件および 323 件であり、マーカー陽性の陽性割合は 0.55 であった。

フォールバック法、ハイブリット法、交互作用検定に基づく方法、および、本研究での提案法における全集団に対する仮説検定の有意水準（ α_o ）を $\alpha_o = 0.015$（ $\alpha = 0.025$ の 60% ）、交互作用検定に基づく方法における交互作用検定の有意水準（ α_{int} ）を $\alpha_{int} = 0.05$ 、提案法におけるマーカー陰性集団での無効中止の基準を（ δ, γ ） = （ log（ 0.8 ）, 0.15 ）と設定した。

固定順序法およびハイブリット法（ 17ページ、それぞれ 図 1a 、図 1c を参照 ）では、マーカー陽性集団についてのみ治療効果が示されたと主張できるが、フォールバック法、交互作用検定に基づく方法、および本研究での提案法（ 17ページ、それぞれ 図 1b 、図 1d 、図 1e を参照 ）では、全集団について治療効果が示されたと主張できる結果となった（ 表 3を参照 ）。

なお、マーカー陰性集団で臨床的に意味のある最小の治療効果を δ = log（ 0.7 ）

とすると、本研究での提案法はマーカー陽性集団についてのみ治療効果が示されたと主張できる。一方、マーカー陰性集団ではより小さい治療効果の大きさ（ ハザード比 0.8 または $\delta = \log(0.8)$ ）でも臨床的に意味があると考えられる場合には、全集団について治療効果が示されたと主張するという結果となった。

第 5 章　統計的仮説検定の方法の作用特性に関する評価基準

第 5 章では、統計的仮説検定の方法の作用特性を評価するための基準を提案する [1] 。

改めて、第 3 章 および第 4 章 で議論した多重検定法では、全集団またはマーカー陽性集団でいずれかに対して新規治療の治療効果を示すことが可能という特徴を持っている。図 1 に関連して、「治療効果を主張する確率」として三種類の確率（ P_{overall} 、P_{subgroup} および P_{success} ）を導入する [10] 。 P_{overall} 、P_{subgroup} および P_{success} の定義はそれぞれ以下のとおりである。

- P_{overall}　：　全集団に対して治療効果を主張する確率。
- P_{subgroup}　：　マーカー陽性集団のみに対して治療効果を主張する確率。
- P_{success}　：　全集団、マーカー陽性集団のいずれかに対して治療効果を主張する確率（試験が何らかの意味で positive な結果となる確率という意味では試験の成功確率といえる）

なお、 P_{overall} と P_{subgroup} は、それぞれの統計的仮説検定、すなわち、 $H_0^{(O)}$ および $H_0^{(+)}$ の統計的仮説検定が有意となる確率ではないことに注意が必要である。例えば、固定順序法では、 $H_0^{(O)}$ の仮説検定は実施しないものの、 P_{overall} はマーカー陽性および陰性の両方の集団で有意となる確率に対応する（ 図 1 を参照 ）。 一般に、治療効

果を主張する確率では、 $P_{success} = P_{overall} + P_{subgroup}$ が成立する。 $P_{success}$ は、新規治療の成功確率と解釈でき、少なくともマーカー陽性集団に対して治療効果が示されたと主張できる確率を表している。なお、$P_{success}$ が与えられたもとで、$P_{overall}$ と $P_{subgroup}$ との間にはトレードオフの関係がある。

本研究では、2.2 節で述べたように、帰無仮説シナリオ2 が成立しているもとで、試験あたりの第一種の過誤確率（ study-wise alpha rate ）を評価・モニタリングすることを提案する [1] 。すなわち、帰無仮説シナリオ2が成立しているもとで（ マーカー陽性集団でのみ治療効果が存在するもとで ）、$P_{overall}$ が十分に小さいかを評価することに対応する。

一方、検出力を高くする際には、マーカー陽性集団に対して大きな治療効果が存在し、マーカー陰性集団に対しては治療効果が存在しない、または、臨床的に意味のある治療効果が存在しないシナリオのもとでは、マーカー陰性集団に対して過剰な治療を行ってしまう機会を減らすために、$P_{subgroup}$ を大きくし、$P_{overall}$ を小さくすように試みるべきである。他方、マーカー陽性集団及び陰性集団の両方に対して治療効果が存在するシナリオ（ あるいは、 cQT ）のもとでは、マーカー陰性集団が有効な治療を受ける機会を逃してしまうことを避けるために、$P_{overall}$ を大きくし、$P_{subgroup}$ を小さくすることが望まれる。同時に、マーカー陰性集団での治療効果に対して臨床的に意味があるか否かについて判断することが困難な状況も存在する。

より具体的に、第 2 章で取り上げたマーカー仮説が成立しているもとで、例えば、$HR^{(+)} \in (\ 0.5, 0.7\ )$ および $HR^{(-)} \in (\ 0.7, 1.0\ )$ の治療効果が想定できるものとして以下の議論を行う。

ここで、マーカー陰性集団に対する治療効果の大きさとして、$\log(\ HR^{(-)}\ )$ を考え、また、$\log(\ HR^{(-)}\ )$ における二種類の閾値として ϕ_1 および ϕ_2 を導入する（ ただし、$\phi_1 < \phi_2$ とする） 。 具体的には、$\log(\ HR^{(-)}\ ) \leqq \phi_1$ の場合には、マーカー陰性集団に対して新規治療の臨床的に意味のある治療効果が存在するとする。また、$\log(\ HR^{(-)}\ ) > \phi_2$ の場合には、マーカー陰性集団に対して臨床的に意味のある治療効果が存在しないとする。 上記のいずれにも該当しない場合、 すなわち、$\log(\ HR^{(-)}\ ) \in (\phi_1, \phi_2\ ]$ の場合は、マーカー陰性集団に対して臨床的に意味のある治療効果が存在するか否かについての判断が難しい領域（グレーゾーン）に対応すると考えられる。

以下では、議論を分かりやすくするために二種類の閾値 ϕ_1 および ϕ_2 をより具体的に、例えば、$\phi_1 = \log(\ 0.8\ )$ および $\phi_2 = \log(\ 0.95\ )$ として述べる。マーカー仮説が成立しているもとで､統計的仮説検定の方法が満たすべき作用特性として、以下の評価基準（Ⅰ）～（Ⅲ）を同時に満たすことを提案する [1] 。

（Ⅰ）　（ $HR^{(+)}, HR^{(-)}$ ）＝（0.5, 0.8）、（0.7, 0.7）、（0.7, 0.8）など、log（ $HR^{(+)}$ ）≦ ϕ_1 かつ log（$HR^{(-)}$）≦ ϕ_1 を満たす治療効果プロファイルの場合（ 全集団に対して新規治療の効果を主張すべきケース ）では、$P_{overall}$ が十分大きく（ 例えば、0.8 以上 ）、$P_{subgroup}$ が十分小さい（ 例えば、0.2 以下 ）。

（Ⅱ）　（ $HR^{(+)}, HR^{(-)}$ ）＝（0.5, 0.95）、（0.7, 0.95）など、log（ $HR^{(+)}$ ）≦ ϕ_1、log（ $HR^{(-)}$ ）＞ ϕ_2 を満たす治療効果プロファイルの場合（ マーカー陽性集団のみに対して新規治療の効果を主張すべきケース ）では、 $P_{subgroup}$ が十分大きく（ 例えば、0.8 以上 ）、$P_{overall}$ が十分小さい（ 例えば、0.2 以下 ）。

ここで、［0.5, 0.9］や［0.7, 0.9］のようなグレーゾーンの治療効果プロファイルの場合に対応する $P_{overall}$ や $P_{subgroup}$ に関する基準については特に設けないが、試験全体の成功確率を意味し、全集団の平均治療効果のみを仮説検定する伝統的な第Ⅲ相臨床試験において用いられてきた基準に相当する $P_{success}$ に関する以下の評価基準（Ⅲ）を置く。

（Ⅲ）　可能性のあるすべての治療効果のプロファイルのもとで、$P_{success}$ が十分大きい（ 例えば、0.9 以上 ）。

この評価基準（Ⅲ）の場合には、 $P_{\text{success}} = P_{\text{overall}} \geqq 0.9$ であり、また、上記で使用されたハザード比の閾値はあくまで例示であり、実際には疾患や利用可能な対照群の治療によってケースバイケースで設定する必要がある。

次の第 6 章では、上記で提案した、統計的仮説検定の方法が満たすべき作用特性に関する評価基準に基づいて、 図 1 で述べた様々な統計的仮説検定の方法について比較する。

第 6 章　数値評価

生存期間を評価する第Ⅲ相臨床試験を想定する。生存期間を評価する際には打ち切りも含めて考える必要がある。打ち切りの割合が治療群間で異なる場合には、試験の総患者数におけるマーカー陽性の割合と試験で観察される総イベント数におけるマーカー陽性の割合（ p ）が異なる可能性がある。第 6 章では、便宜上、両者がほぼ一致する状況を想定する。例えば、進行癌を対象とした臨床試験で十分な経過観察がなされ、打ち切りが少なく、イベント数と患者数がほとんど一致する場合や、イベントの発生率と打ち切り時間の分布がマーカー陽性集団と陰性集団との間でほぼ同じ場合などがこれに該当する。なお、マーカーの陽性割合の値としては $p=$ 0.2 、0.4 、または 0.6 を指定する。

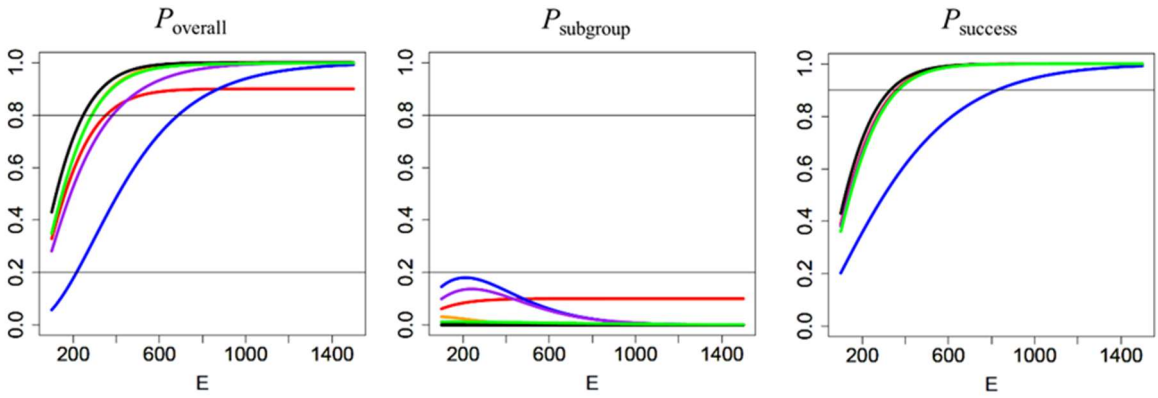
検討に用いた統計的仮説検定の方法は、第 3 章で述べた固定順序法、フォールバック法（ $\alpha_o = 0.015$ ）、ハイブリット法（ $\alpha_o = 0.015$ ）（ 緩やかな有意水準を用いた MaST 法 ）、第 4 章で述べた交互作用検定に基づく方法（ $\alpha_o = 0.015$ 、$\alpha_{int} = 0.1$ ）、提案法（ $\alpha_o = 0.015$ 、$(\delta, \gamma) =$（ log（ 0.8 ）， 0.15 ） ）である。全集団における治療効果の仮説検定にはマーカー層別ログランク検定を、マーカー陽性集団における治療効果の仮説検定には通常の非層別ログランク検定を想定した。交互作用検定に基づく方法では、マーカー部分集団間での対数ハザード比（ 治療効果 ）の差異に基づいて行った。フォールバック法、ハイブリット法、交互作用検定に基づく方法、お

よび、提案法では、検定統計量間での相関関係を考慮して有意水準を設定した。また、参照のため、伝統的な第Ⅲ相臨床試験での治療効果の仮説検定（ 有意水準 α を用いた全集団での治療効果の仮説検定を一回だけ行う ）も併せて実施した（ この場合は、$P_{\text{success}} = P_{\text{overall}}$ ）。マーカー陽性割合、治療効果プロファイルが与えられたもとでの三種類の確率 P_{overall} 、P_{subgroup} 、および、P_{success} の計算は、ログランク統計量の漸近分布に基づいて行う（ 第 8 章を参照 ）。帰無仮説シナリオ1（ グローバル帰無仮説 ）が成立しているもとで、試験あたりの第一種の過誤確率が $\alpha = 0.025$ となるように各検定手順の有意水準を設定した。

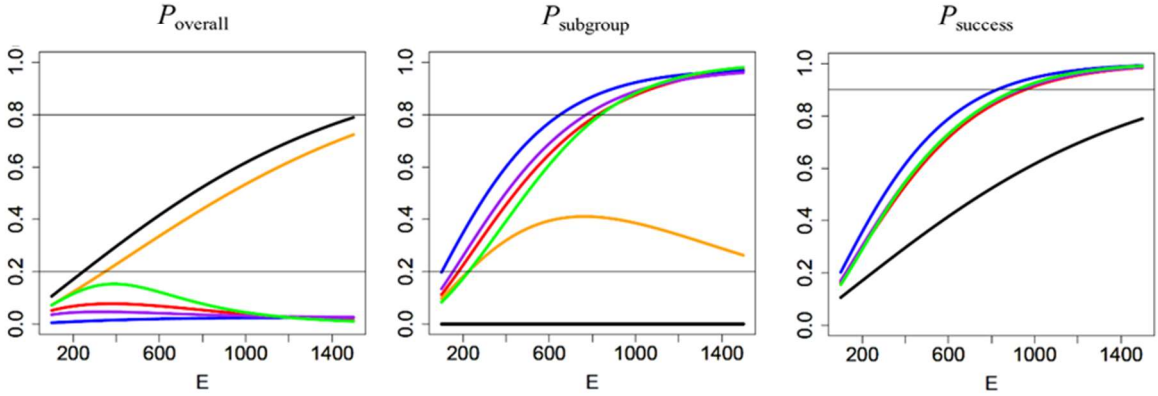
最初に、二種類の極端なシナリオのもとでの作用特性を評価することで、それぞれの仮説検定の方法の基本的な特徴について確認した。マーカー陽性集団での治療効果は、マーカー陰性集団での治療効果よりも小さくなることはないというマーカー仮説のもとで、一つのシナリオは、治療効果がマーカー陽性集団と陰性集団との間で一定の場合である（ $\text{HR}^{(+)}$, $\text{HR}^{(-)}$ ）=（ 0.7, 0.7 ）、もう一つのシナリオは、マーカー陰性集団に対して治療効果がなく、 cQL が成立する場合である（ $\text{HR}^{(+)}$, $\text{HR}^{(-)}$ ）=（ 0.7, 1.0 ）。

図 2 の a） および b） は、マーカー陽性割合 $p = 0.4$ の場合の様々なイベント数（ E ）に対する三種類の治療効果を主張する確率である P_{overall} 、P_{subgroup} および P_{success} の曲線を示している（ イベント数に対するこれらの確率の値については表 4 を参照 ）。

a） 治療効果が一定: $(HR^{(+)}, HR^{(-)}) = (0.7, 0.7)$



b） cQL交互作用　マーカー陰性で治療効果がない: $(HR^{(+)}, HR^{(-)}) = (0.7, 1.0)$



c） cQT 交互作用　$(HR^{(+)}, HR^{(-)}) = (0.5, 0.7)$

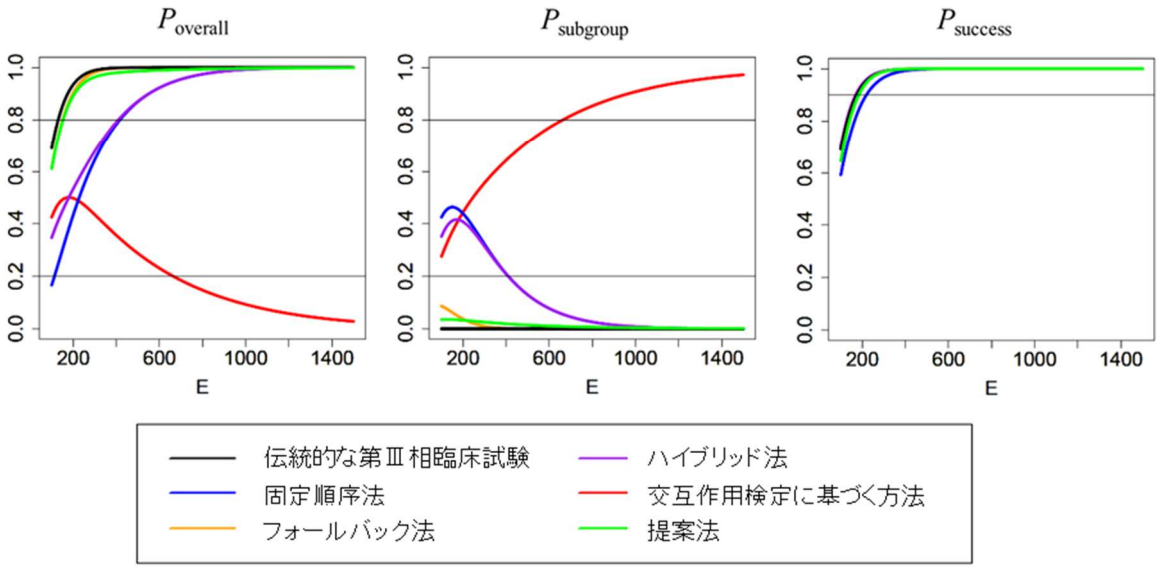


図 2. $(HR^{(+)}, HR^{(-)}) = (0.7, 0.7)$、$(0.7, 1.0)$ および $(0.5, 0.7)$のシナリオのもとでの治療効果を主張する三種類の確率の曲線 （縦：治療効果を主張する確率、横：イベント数）

表 4　様々な全イベント数の場合の 図 2 における $P_{overall}$ 、$P_{subgroup}$ および $P_{success}$ の値

$HR^{(+)}$	$HR^{(-)}$	E	確率	伝統的	FS	FB	HB	INT	提案法
0.7	0.7	300	$P_{overall}$	0.871	0.332	0.821	0.703	0.749	0.818
			$P_{subgroup}$	0.000	0.165	0.013	0.131	0.093	0.011
			$P_{success}$	0.871	0.497	0.834	0.834	0.842	0.829
		500	$P_{overall}$	0.979	0.621	0.965	0.888	0.872	0.960
			$P_{subgroup}$	0.000	0.092	0.003	0.080	0.099	0.009
			$P_{success}$	0.979	0.713	0.968	0.968	0.971	0.968
		700	$P_{overall}$	0.997	0.809	0.995	0.960	0.896	0.989
			$P_{subgroup}$	0.000	0.038	0.001	0.035	0.100	0.006
			$P_{success}$	0.997	0.847	0.995	0.995	0.996	0.995
0.7	1.0	300	$P_{overall}$	0.234	0.012	0.175	0.047	0.076	0.145
			$P_{subgroup}$	0.000	0.485	0.257	0.385	0.342	0.281
			$P_{success}$	0.234	0.497	0.432	0.432	0.418	0.425
		500	$P_{overall}$	0.358	0.018	0.283	0.044	0.074	0.143
			$P_{subgroup}$	0.000	0.695	0.365	0.603	0.559	0.505
			$P_{success}$	0.358	0.713	0.647	0.647	0.633	0.648
		700	$P_{overall}$	0.471	0.021	0.389	0.038	0.061	0.099
			$P_{subgroup}$	0.000	0.826	0.408	0.758	0.723	0.698
			$P_{success}$	0.471	0.847	0.797	0.797	0.785	0.797
0.5	0.7	300	$P_{overall}$	0.989	0.645	0.981	0.678	0.435	0.964
			$P_{subgroup}$	0.000	0.322	0.010	0.314	0.557	0.025
			$P_{success}$	0.989	0.967	0.991	0.991	0.991	0.989
		500	$P_{overall}$	1.000	0.869	1.000	0.871	0.287	0.986
			$P_{subgroup}$	0.000	0.129	0.000	0.129	0.713	0.014
			$P_{success}$	1.000	0.998	1.000	1.000	1.000	1.000
		700	$P_{overall}$	1.000	0.955	1.000	0.955	0.184	0.992
			$P_{subgroup}$	0.000	0.045	0.000	0.045	0.816	0.008
			$P_{success}$	1.000	1.000	1.000	1.000	1.000	1.000

有意水準 α_o、α_s、α_{int} および パラメータ（δ, γ）が 図 2 で使用された値と同様に設定された。

伝統的：伝統的な第Ⅲ相臨床試験、FS：固定順序法、FB：フォールバック法、HB：ハイブリット法、INT：交互作用検定に基づく方法

Matsui らの先行研究によって指摘されているように [9]、固定順序法では、一般的に、治療効果が一定のシナリオのもとで P_{overall} （ および P_{success} ）が低くなるという問題が生じた。

また、フォールバック法（ および 伝統的な第Ⅲ相臨床試験の仮説検定 ）では、第 2 章で述べたように、 cQL が認められるシナリオもとで P_{overall} を大きくすることが困難であった。ハイブリット法、交互作用検定に基づく方法、および、提案法では、治療効果が一定のシナリオもとで P_{overall} が大きくなる一方、 P_{subgroup} は小さくなり、また、cQL が認められるシナリオもとで P_{subgroup} が大きくなる一方、P_{overall} は小さくなり、比較的良好な性能であることを確認できた。

図 2 の c） は、マーカー陽性集団および陰性集団の両方で臨床的に意味のある治療効果があるものの、cQT が認められる（ $\mathrm{HR}^{(+)}$, $\mathrm{HR}^{(-)}$ ）=（ 0.5, 0.7 ）のシナリオもとでの三種類の治療効果を主張する確率の曲線を示している。交互作用検定に基づく方法は、他の方法と比較して、まったく異なる曲線を示した。すなわち、イベント数（ E ）の増加とともに、P_{overall} は小さくなる一方、P_{subgroup} は大きくなった（第 3 章で述べたように、部分集団の選択における交互作用検定に基づく方法の問題点が反映されていると考えられる）。

次に、第 5 章で提案した統計的仮説検定の方法が満たすべき作用特性に関する評価基準を適用し、各統計的仮説検定の方法を比較した。

表 5　様々な治療効果プロファイルのもとでの、既存の様々な統計的仮説検定の方法を用いた場合の $TE_{overall}$、$TE_{subgroup}$、および、$TE_{success}$（マーカー陽性割合 $p = 0.4$ ）

	最大のイベント数	全集団で治療効果あり			マーカー陽性集団のみで治療効果あり		マーカー陽性集団で治療効果あり、マーカー陰性集団で治療効果がある可能性あり		TE_{max}
		[0.5, 0.8]	[0.7, 0.7]	[0.7, 0.8]	[0.5, 0.95]	[0.7, 0.95]	[0.5, 0.9]	[0.7, 0.9]	
固定順序法									
	$TE_{overall}$	1,051	688	1,131	<100	<100	–	–	
	$TE_{subgroup}$	1,051	<100	<100	181	744	–	–	
	$TE_{success}$	219	826	826	219	826	219	826	
	最大値	1,051	826	1,131	219	826	219	826	1,131
フォールバック法									
$\alpha_o = 0.01$	$TE_{overall}$	238	316	525	<100	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	>2,000	>2,000	–	–	
	$TE_{success}$	205	392	589	230	849	225	788	
	最大値	238	392	589	>2,000	>2,000	225	788	>2,000
$\alpha_o = 0.02$	$TE_{overall}$	199	264	439	<100	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	>2,000	>2,000	–	–	
	$TE_{success}$	213	346	549	262	946	249	825	
	最大値	213	346	549	>2,000	>2,000	249	825	>2,000
ハイブリッド法									
$\alpha_o = 0.003$	$TE_{overall}$	1,051	499	1,075	<100	<100	–	–	
	$TE_{subgroup}$	1,051	<100	<100	183	751	–	–	
	$TE_{success}$	213	511	696	220	825	219	807	
	最大値	1,051	511	1,075	220	825	219	807	1,075
$\alpha_o = 0.02$	$TE_{overall}$	1,051	346	999	<100	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	245	1,024	–	–	
	$TE_{success}$	213	345	549	262	946	249	825	
	最大値	1,051	346	999	262	1,024	249	825	1,051
交互作用検定に基づく方法									
$\alpha_{int} = 0.05,\ \alpha_o = 0.01$	$TE_{overall}$	>2,000	349	>2,000	<100	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	257	1,116	–	–	
	$TE_{success}$	231	403	632	252	947	251	888	
	最大値	>2,000	403	>2,000	257	1,116	251	888	>2,000
$\alpha_{int} = 0.05,\ \alpha_o = 0.025$	$TE_{overall}$	>2,000	277	>2,000	<100	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	330	1,313	–	–	
	$TE_{success}$	250	335	568	339	1,199	315	981	
	最大値	>2,000	335	>2,000	339	1,313	315	981	>2,000
$\alpha_{int} = 0.15,\ \alpha_o = 0.01$	$TE_{overall}$	>2,000	477	>2,000	<100	<100	–	–	
	$TE_{subgroup}$	>2,000	<100	<100	199	793	–	–	
	$TE_{success}$	207	394	594	231	854	226	794	
	最大値	>2,000	477	>2,000	231	854	226	794	>2,000
$\alpha_{int} = 0.15,\ \alpha_o = 0.025$	$TE_{overall}$	>2,000	391	>2,000	<100	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	263	1,013	–	–	
	$TE_{success}$	248	349	601	305	1,106	292	966	
	最大値	>2,000	391	>2,000	305	1,106	292	966	>2,000
伝統的な第Ⅲ相臨床試験									
		249	331	550	443	1,398	363	992	1,398

表 5 は、マーカー陽性割合 $p=0.4$ 、様々な治療効果プロファイルのもとで、各仮説検定の方法に対して、第 5 章の作用特性に関する評価基準を満たすに必要な最小のイベント数（ E ）を纏めた表である。

予測されたように、固定順序法に対する TE_{overall} 、TE_{subgroup} 、および、TE_{success} は、マーカー陽性集団のみで臨床的に意味のある治療効果が存在する場合に最小値となり、全集団で臨床的に意味のある治療効果が存在する場合に非常に大きくなった。言い換えれば、固定順序法は、評価基準（Ⅱ）（ および（Ⅲ） ）を達成することは（ 比較的少ないイベント数で ）容易であるが、評価基準（Ⅰ）を達成することは困難であった（ 図 2 の a）を参照 ）。

対照的に、フォールバック法では、TE_{overall} 、TE_{subgroup} および TE_{success} が、全集団で臨床的に意味のある治療効果が存在する場合に小さくなり、マーカー陽性集団で臨床的に意味のある治療効果が存在する場合にはとても大きな値となった。すなわち、評価基準（Ⅰ）を達成することは容易であったが、マーカー陽性集団のみで治療効果が認められる場合、特に $P_{\text{subgroup}} < 0.2$ （ 結果として $TE_{\text{subgroup}} > 2{,}000$ となる ）となり、評価基準（Ⅱ）を達成できなかった（ 図 2 の b）を参照 ）。

ハイブリット法の性能は、固定順序法とフォールバック法のほぼ中間であった。オリジナルの MaST 法を近似した $\alpha_o = 0.003$（ 注： 全集団とマーカー陽性集団との相関関係を考慮して有意水準を定めたことからオリジナルとは若干異なる ）での仮説検定では、固定順序法とほぼ同様の結果であった。これに対して、より緩めた有意

水準 $\alpha_o = 0.02$ を用いると、治療効果が一定（ $HR^{(+)}$, $HR^{(-)}$ ）=（ 0.7, 0.7 ）のシナリオのもとで性能が改善したが（同様に（ $HR^{(+)}$, $HR^{(-)}$ ）=（ 0.7, 0.8 ）のシナリオのもとでも改善した ）、 マーカー陰性集団に対して治療効果が存在せず、cQL が認められる場合（ $HR^{(+)}, HR^{(-)}$ ）=（ 0.5, 0.95 ）や（ 0.7, 0.95 ）のシナリオのもとでは性能が悪化した。

交互作用検定に基づく方法の性能は、二種類の極端なシナリオ、すなわち、治療効果が一定（ $HR^{(+)}$, $HR^{(-)}$ ）=（ 0.7, 0.7 ）のシナリオのもと、および、 cQL が存在する（ $HR^{(+)}$, $HR^{(-)}$ ）=（ 0.5, 0.95 ）のシナリオのもとで比較的良好であった（ 図 2 の a）および b）を参照 ）。しかし、治療効果プロファイルがこれら二種類のシナリオから離れた場合にはその性能は大きく低下するケースがあった。

提案法の性能は、 表 6 に示されているように、臨床的に意味のある最小の治療効果を特定する閾値（ δ ）に依存して大きく変化する傾向が認められた。

マーカー陰性集団に対する無効判定基準で用いられる、臨床的に意味のある効果の閾値 δ に対して、比較的緩やかな（ 無効判定がなされやすい ） 値、 $\delta = \log$（ 0.7 ）を用いると、提案法の性能は、マーカー陽性集団のみで治療効果が存在する場合に良好であったが、全集団で治療効果が存在する場合には良好ではなかった。これは、固定順序法で認められた結果と類似しているが、全集団で治療効果が存在する場合に、提案法では P_{overall} を大きくすることはより多くのイベント数を必要とする。

表 6　様々な治療効果プロファイルのもとでの、提案法を用いた場合の $TE_{overall}$、$TE_{subgroup}$、および、$TE_{success}$（マーカー陽性割合 $p = 0.4$ ）

	最大のイベント数	全集団で治療効果あり			マーカー陽性集団のみで治療効果あり		マーカー陽性集団で治療効果あり、マーカー陰性集団で治療効果がある可能性あり		TE_{max}
		[0.5, 0.8]	[0.7, 0.7]	[0.7, 0.8]	[0.5, 0.95]	[0.7, 0.95]	[0.5, 0.9]	[0.7, 0.9]	
提案法									
$\delta = \log(0.7)$									
$\gamma = 0.05$, $\alpha_o = 0.02$	$TE_{overall}$	>2,000	321	>2,000	442	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	448	742	–	–	
	$TE_{success}$	234	393	611	250	844	249	818	
	最大値	>2,000	393	>2,000	448	844	249	818	>2,000
$\gamma = 0.05$, $\alpha_o = 0.04$	$TE_{overall}$	>2,000	269	>2,000	442	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	453	757	–	–	
	$TE_{success}$	215	347	621	262	856	250	833	
	最大値	>2,000	347	>2,000	453	856	250	833	>2,000
$\gamma = 0.15$, $\alpha_o = 0.02$	$TE_{overall}$	>2,000	375	>2,000	<100	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	295	666	–	–	
	$TE_{success}$	209	402	707	231	832	226	823	
	最大値	>2,000	402	>2,000	295	832	226	823	>2,000
$\gamma = 0.15$, $\alpha_o = 0.04$	$TE_{overall}$	>2,000	318	>2,000	249	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	309	672	–	–	
	$TE_{success}$	215	377	712	260	834	250	826	
	最大値	>2,000	377	>2,000	309	834	250	826	>2,000
$\delta = \log(0.8)$									
$\gamma = 0.05$, $\alpha_o = 0.02$	$TE_{overall}$	242	316	534	1,396	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	1,396	1,443	–	–	
	$TE_{success}$	278	407	630	328	866	321	820	
	最大値	278	407	630	1,396	1,443	321	820	1,443
$\gamma = 0.05$, $\alpha_o = 0.04$	$TE_{overall}$	202	264	447	1,396	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	1,396	1,459	–	–	
	$TE_{success}$	239	348	553	295	948	283	828	
	最大値	239	348	553	1,396	1,459	283	828	1,459
$\gamma = 0.15$, $\alpha_o = 0.02$	$TE_{overall}$	282	317	624	797	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	797	1,004	–	–	
	$TE_{success}$	244	400	594	265	851	264	789	
	最大値	282	400	624	797	1,004	264	789	1,004
$\gamma = 0.15$, $\alpha_o = 0.04$	$TE_{overall}$	240	266	529	797	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	797	1,036	–	–	
	$TE_{success}$	219	346	556	265	923	253	847	
	最大値	240	346	556	797	1,036	253	847	1,036
伝統的な第Ⅲ相臨床試験									
		249	331	550	443	1,398	363	992	1,398

対照的に、より保守的に δ = log（ 0.8 ）とすると、提案法の性能は、全集団で治療効果が存在する場合には良好であったが、マーカー陽性集団のみで治療効果が存在する場合には良好でなかった。これは、一見、フォールバック法で認められた結果と類似している。ただし、マーカー陽性集団のみで治療効果が存在する場合に、フォールバック法では P_{subgroup} を制御できなかったが（ 第 2 章 および 図 2 の b）を参照 ）、提案法では一定のイベント数を確保することで P_{subgroup} を制御することができるという結果であった。

すべての治療効果シナリオが同等に起こり得ると仮定すると、すべての統計的仮説検定の方法の中で、 $\delta = \log$（ 0.8 ）、 $\gamma = 0.15$ および $\alpha_o = 0.06$ と設定した提案法において、必要なイベント数の最大値（ $TE_{\max}$ ）が最小の値となった（ $TE_{\max} = 1{,}009$ ）。つまり、提案法は、想定される治療効果シナリオの範囲に対して、最も小さなイベント数で評価基準（Ⅰ）～（Ⅲ）を達成できることを意味する。

以上の結果は、マーカー陽性割合（ p ）が他の値の場合にもほぼ同様であった（ $p = 0.2$ の場合は 表 7 および 表 8、$p = 0.6$ の場合は 表 9 および 表 10 を参照)。

表 7　様々な治療効果プロファイルのもとでの、既存の様々な統計的仮説検定の既存の方法を用いた場合の $TE_{overall}$、$TE_{subgroup}$、および、$TE_{success}$（マーカー陽性割合 $p = 0.2$ ）

	最大のイベント数	全集団で治療効果あり			マーカー陽性集団のみで治療効果あり		マーカー陽性集団で治療効果あり、マーカー陰性集団で治療効果がある可能性あり		TE_{max}
		[0.5, 0.8]	[0.7, 0.7]	[0.7, 0.8]	[0.5, 0.95]	[0.7, 0.95]	[0.5, 0.9]	[0.7, 0.95]	
固定順序法									
	$TE_{overall}$	801	1,235	1,358	<100	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	381	>2,000	–	–	
	$TE_{success}$	438	1,652	1,652	438	1,652	438	1,652	
	最大値	801	1,652	1,652	438	>2,000	438	1,652	>2,000
フォールバック法									
$\alpha_o = 0.01$	$TE_{overall}$	400	316	644	<100	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	>2,000	>2,000	–	–	
	$TE_{success}$	358	400	762	468	1690	225	788	
	最大値	400	400	762	>2,000	>2,000	225	788	>2,000
$\alpha_o = 0.02$	$TE_{overall}$	334	264	538	<100	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	>2,000	>2,000	–	–	
	$TE_{success}$	362	347	688	549	1,912	249	825	
	最大値	362	347	688	>2,000	>2,000	249	825	>2,000
ハイブリッド法									
$\alpha_o = 0.003$	$TE_{overall}$	793	477	975	<100	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	384	>2,000	–	–	
	$TE_{success}$	399	570	1,015	439	1,641	219	807	
	最大値	793	570	1,015	439	>2,000	219	807	>2,000
$\alpha_o = 0.02$	$TE_{overall}$	781	288	682	<100	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	519	>2,000	–	–	
	$TE_{success}$	359	352	694	531	1,862	249	825	
	最大値	781	352	694	531	>2,000	249	825	>2,000
交互作用検定に基づく方法									
$\alpha_{int} = 0.05,\ \alpha_o = 0.01$	$TE_{overall}$	>2,000	349	>2,000	<100	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	415	1,729	–	–	
	$TE_{success}$	378	403	782	481	1,768	251	888	
	最大値	>2,000	403	>2,000	481	1,768	251	888	>2,000
$\alpha_{int} = 0.05,\ \alpha_o = 0.025$	$TE_{overall}$	>2,000	277	>2,000	<100	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	625	>2,000	–	–	
	$TE_{success}$	477	344	749	732	>2,000	–	–	
	最大値	>2,000	344	>2,000	732	>2,000	466	1,491	>2,000
$\alpha_{int} = 0.15,\ \alpha_o = 0.01$	$TE_{overall}$	>2,000	477	>2,000	<100	<100	466	1,491	
	$TE_{subgroup}$	<100	<100	<100	372	1,426	–	–	
	$TE_{success}$	362	401	769	468	1,698	–	–	
	最大値	>2,000	477	>2,000	468	1,698	678	1,932	>2,000
$\alpha_{int} = 0.15,\ \alpha_o = 0.025$	$TE_{overall}$	>2,000	391	>2,000	<100	<100	678	1,932	
	$TE_{subgroup}$	<100	<100	<100	515	1,948	–	–	
	$TE_{success}$	499	389	909	637	>2,000	–	–	
	最大値	>2,000	391	>2,000	637	>2,000	447	1,425	>2,000
伝統的な第Ⅲ相臨床試験									
		249	331	550	443	1,398	363	992	1,398

表 8　様々な治療効果プロファイルのもとでの、提案法を用いた場合の $TE_{overall}$、$TE_{subgroup}$、および、$TE_{success}$（マーカー陽性割合 $p = 0.2$ ）

	最大のイベント数	全集団で治療効果あり			マーカー陽性集団のみで治療効果あり		マーカー陽性集団で治療効果あり、マーカー陰性集団で治療効果がある可能性あり		TE_{max}
		[0.5, 0.8]	[0.7, 0.7]	[0.7, 0.8]	[0.5, 0.95]	[0.7, 0.95]	[0.5, 0.9]	[0.7, 0.95]	
提案法									
$\delta = \log(0.7)$									
$\gamma = 0.05$, $\alpha_o = 0.02$	$TE_{overall}$	>2,000	317	>2,000	<100	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	463	1,237	–	–	
	$TE_{success}$	361	400	1,477	468	1,652	447	1,651	
	最大値	>2,000	400	>2,000	468	1,652	447	1,651	>2,000
$\gamma = 0.05$, $\alpha_o = 0.04$	$TE_{overall}$	>2,000	265	>2,000	<100	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	488	1,237	–	–	
	$TE_{success}$	374	349	1,477	506	1,652	489	1,651	
	最大値	>2,000	349	>2,000	506	1,652	489	1,651	>2,000
$\gamma = 0.15$, $\alpha_o = 0.02$	$TE_{overall}$	>2,000	346	>2,000	<100	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	395	1,235	–	–	
	$TE_{success}$	379	466	1598	457	1,652	449	1,652	
	最大値	>2,000	466	>2,000	457	1,652	449	1,652	>2,000
$\gamma = 0.15$, $\alpha_o = 0.04$	$TE_{overall}$	>2,000	291	>2,000	<100	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	407	1,235	–	–	
	$TE_{success}$	397	452	1,598	466	1,652	458	1,652	
	最大値	>2,000	452	>2,000	466	1,652	458	1,652	>2,000
$\delta = \log(0.8)$									
$\gamma = 0.05$, $\alpha_o = 0.02$	$TE_{overall}$	401	316	646	<100	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	1,052	1,585	–	–	
	$TE_{success}$	460	409	781	571	1,691	567	1,511	
	最大値	460	409	781	1,052	1,691	567	1,511	1,691
$\gamma = 0.05$, $\alpha_o = 0.04$	$TE_{overall}$	335	264	540	<100	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	1,057	1,607	–	–	
	$TE_{success}$	399	350	688	563	1,714	528	1,559	
	最大値	399	350	688	1,057	1,714	528	1,559	1,714
$\gamma = 0.15$, $\alpha_o = 0.02$	$TE_{overall}$	437	316	704	<100	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	662	1,377	–	–	
	$TE_{success}$	400	405	770	481	1,661	469	1,586	
	最大値	437	405	770	662	1,661	469	1,586	1,661
$\gamma = 0.15$, $\alpha_o = 0.04$	$TE_{overall}$	368	264	593	<100	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	695	1,384	–	–	
	$TE_{success}$	365	348	741	549	1,662	503	1,588	
	最大値	368	348	741	695	1,662	503	1,588	1,662
伝統的な第Ⅲ相臨床試験									
		249	331	550	443	1,398	363	992	1,398

表 9　様々な治療効果プロファイルのもとでの、既存の様々な統計的仮説検定の方法を用いた場合の $TE_{overall}$、$TE_{subgroup}$、および、$TE_{success}$（マーカー陽性割合 $p = 0.6$）

	最大のイベント数	全集団で治療効果あり			マーカー陽性集団のみで治療効果あり		マーカー陽性集団で治療効果あり、マーカー陰性集団で治療効果がある可能性あり		TE_{max}
		[0.5, 0.8]	[0.7, 0.7]	[0.7, 0.8]	[0.5, 0.95]	[0.7, 0.95]	[0.5, 0.9]	[0.7, 0.95]	
固定順序法									
	$TE_{overall}$	1,577	688	1,577	<100	<100	–	–	
	$TE_{subgroup}$	1,577	<100	<100	119	465	–	–	
	$TE_{success}$	146	551	551	146	551	146	551	
	最大値	1,577	688	1,577	146	551	146	551	1,577
フォールバック法									
$\alpha_o = 0.01$	$TE_{overall}$	158	316	437	>2,000	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	>2,000	>2,000	–	–	
	$TE_{success}$	143	383	475	150	560	149	541	
	最大値	158	383	475	>2,000	>2,000	149	541	>2,000
$\alpha_o = 0.02$	$TE_{overall}$	132	264	365	>2,000	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	>2,000	>2,000	–	–	
	$TE_{success}$	147	343	453	165	606	160	562	
	最大値	147	343	453	>2,000	>2,000	160	562	>2,000
ハイブリッド法									
$\alpha_o = 0.003$	$TE_{overall}$	1,577	638	1,577	<100	<100	–	–	
	$TE_{subgroup}$	1,577	<100	<100	120	469	–	–	
	$TE_{success}$	144	448	515	147	551	146	545	
	最大値	1,577	638	1,577	147	551	146	545	1,577
$\alpha_o = 0.02$	$TE_{overall}$	1,577	570	1,576	<100	<100	–	–	
	$TE_{subgroup}$	1,577	<100	<100	160	630	–	–	
	$TE_{success}$	149	339	452	171	623	165	572	
	最大値	1,051	570	1,576	171	630	165	572	1,577
交互作用検定に基づく方法									
$\alpha_{int} = 0.05$, $\alpha_o = 0.01$	$TE_{overall}$	>2,000	349	>2,000	<100	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	251	1,106	–	–	
	$TE_{success}$	167	403	530	175	658	174	635	
	最大値	>2,000	403	>2,000	251	1,106	174	635	>2,000
$\alpha_{int} = 0.05$, $\alpha_o = 0.025$	$TE_{overall}$	>2,000	277	>2,000	239	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	269	1,141	–	–	
	$TE_{success}$	164	333	462	201	721	190	633	
	最大値	>2,000	333	>2,000	269	1,141	190	633	>2,000
$\alpha_{int} = 0.15$, $\alpha_o = 0.01$	$TE_{overall}$	>2,000	477	>2,000	<100	<100	–	–	
	$TE_{subgroup}$	>2,000	<100	<100	162	676	–	–	
	$TE_{success}$	149	392	494	154	578	153	563	
	最大値	>2,000	477	>2,000	162	676	153	563	>2,000
$\alpha_{int} = 0.15$, $\alpha_o = 0.025$	$TE_{overall}$	>2,000	391	>2,000	<100	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	194	778	–	–	
	$TE_{success}$	162	338	470	189	687	182	623	
	最大値	>2,000	391	>2,000	194	778	182	623	>2,000
伝統的な第Ⅲ相臨床試験									
		249	331	550	443	1,398	363	992	1,398

表 10　様々な治療効果プロファイルのもとでの、提案法を用いた場合の $TE_{overall}$、$TE_{subgroup}$、および、$TE_{success}$（マーカー陽性割合 $p = 0.6$）

		全集団で治療効果あり			マーカー陽性集団のみで治療効果あり		マーカー陽性集団で治療効果あり、マーカー陰性集団で治療効果がある可能性あり		TE_{max}
	最大のイベント数	[0.5, 0.8]	[0.7, 0.7]	[0.7, 0.8]	[0.5, 0.95]	[0.7, 0.95]	[0.5, 0.9]	[0.7, 0.95]	
提案法									
$\delta = \log(0.7)$									
$\gamma = 0.05$, $\alpha_o = 0.02$	$TE_{overall}$	199	327	>2,000	663	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	663	740	–	–	
	$TE_{success}$	172	391	481	180	561	179	543	
	最大値	199	391	>2,000	663	740	179	543	>2,000
$\gamma = 0.05$, $\alpha_o = 0.04$	$TE_{overall}$	162	274	>2,000	663	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	663	755	–	–	
	$TE_{success}$	152	343	457	169	602	165	568	
	最大値	162	343	>2,000	663	755	165	568	>2,000
$\gamma = 0.15$, $\alpha_o = 0.02$	$TE_{overall}$	>2,000	403	>2,000	379	<100	–	–	
	$TE_{subgroup}$	>2,000	<100	<100	380	564	–	–	
	$TE_{success}$	153	383	478	158	561	157	545	
	最大値	>2,000	403	>2,000	380	564	157	545	>2,000
$\gamma = 0.15$, $\alpha_o = 0.04$	$TE_{overall}$	>2,000	344	>2,000	379	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	380	580	–	–	
	$TE_{success}$	147	353	480	165	585	160	566	
	最大値	>2,000	353	>2,000	380	585	160	566	>2,000
$\delta = \log(0.8)$									
$\gamma = 0.05$, $\alpha_o = 0.02$	$TE_{overall}$	163	317	453	>2,000	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	>2,000	>2,000	–	–	
	$TE_{success}$	188	405	528	211	627	206	613	
	最大値	188	405	528	>2,000	>2,000	206	613	>2,000
$\gamma = 0.05$, $\alpha_o = 0.04$	$TE_{overall}$	137	265	379	>2,000	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	>2,000	>2,000	–	–	
	$TE_{success}$	162	347	460	185	609	179	568	
	最大値	162	347	460	>2,000	>2,000	179	568	>2,000
$\gamma = 0.15$, $\alpha_o = 0.02$	$TE_{overall}$	201	323	557	1,195	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	1,195	1,206	–	–	
	$TE_{success}$	170	396	494	178	571	177	557	
	最大値	201	396	557	1,195	1,206	177	557	1,206
$\gamma = 0.15$, $\alpha_o = 0.04$	$TE_{overall}$	172	272	476	1,195	<100	–	–	
	$TE_{subgroup}$	<100	<100	<100	1,195	1,210	–	–	
	$TE_{success}$	151	344	453	168	606	164	564	
	最大値	172	344	476	1,195	1,210	164	564	1,210
伝統的な第Ⅲ相臨床試験									
		249	331	550	443	1,398	363	992	1,398

以上を要約すると、第Ⅲ相臨床試験をデザインする段階において、試験治療の効果がマーカーによって十分捉えられることを示唆する生物学的根拠や先行研究のデータ（試験治療に関する早期試験のデータ等）が存在し、マーカーの治療効果予測能が十分高いと想定できる場合には、固定順序法、および、ハイブリット法が推奨されると考えられた。一方、そうでない場合には、マーカー陰性集団に対して保守的な無効中止基準を用いた提案法を選択することが妥当であると考えられた。

第 7 章　考察および結論

本研究では、マーカーが規定する部分集団間での治療効果の不均一性に関するマーカー仮説のもと、検証的試験として位置づけられるマーカー層別第Ⅲ相臨床試験の計画と解析に関する検討を行い、大きく三つの提案を行った [1] 。

一つ目は、マーカー層別第Ⅲ相臨床試験における統計的仮説検定（ 部分集団間の多重検定 ）に対する新しい枠組みの提案である。本研究で検討するマーカー層別第Ⅲ相臨床試験では、第一種の過誤確率の弱い制御を許容する。しかし、たとえ弱い制御であっても、少なくともマーカー陽性集団に対して治療効果を主張することに対する厳格な第一種の過誤の制御は保証されている。その一方で、マーカー陰性集団に対して治療効果があると誤って主張してしまう（ すなわち、誤って新規治療の適応を全集団に広げてしまう ）第一種の過誤については厳格に制御すること（ 強い制御 ）は求めない。その代わりに、試験の計画段階において、適応拡大に関する第一種の過誤を一定の水準に制御するか、または、解析の段階において、想定される治療効果の範囲のもとでそれらを評価・モニタリングすることを提案した [1] 。

マーカーにより規定される部分集団間の多重検定法については、これまで多くの提案があるが、そのほとんどは、 $H_0^{(O)}$ と $H_0^{(+)}$ の多重検定を強い制御のもとで行うものである。しかし、その一方で、マーカー陰性集団に対する事後的な解析が現場では推奨されていることから、治療効果の検証的な位置づけの解析の枠組みとして不

完全、あるいは、過度に強いものと言わざるをえない。一方、本研究で提案する多重検定の枠組みは、マーカー仮説のもと、マーカーで規定された集団での治療効果の評価により成り立っており、事後的な評価を必要とする $H_0^{(O)}$ の仮説検定に依存しない枠組みである。さらに、検証の基準（ 第一種の過誤の制御 ）に関しては、従来の枠組みでは、$H_0^{(O)}$ と $H_0^{(+)}$ から構成されるすべての帰無仮説シナリオに対して第一種の過誤の制御を考えることになり、しかも、それらはマーカー仮説から導かれる帰無仮説シナリオと合致していない（ 第 2 章を参照 ）。これに対して、提案する枠組みは、マーカー陽性集団での治療効果の推測に限定して第一種の過誤を制御するので解釈がより明確である。併せて、全集団への適応拡大に対しては厳密な基準を設けず、その判断はケースバイケースでなされることを最初から許容している。提案する枠組みは、結果として、従来の枠組みより効率的なデザインが可能となる。なお、以上のメリットをもつ提案する枠組みはマーカー仮説の成立を拠り所としている。マーカーが存在する多くの新しい分子標的薬剤にとってこの仮説は自然で、合理的であると考えられる。これより、多くの分子標的薬剤の臨床試験で提案する枠組みの採用を推奨したい。

本研究の二つ目の提案は、第 4 章で与えたマーカーの治療効果予測能（ 臨床的妥当性 ）の評価を組み込んだ仮説検定法の提案である [1] 。これより、マーカー陽性集団で治療効果が有意となった際、マーカーを使って集団を限定することの臨床的妥当

性についても明示的な根拠が得られていることになる。第 6 章の数値評価において、マーカー陽性集団での保守的な無効中止基準（ 例えば、 $\delta = \log(0.8)$ ）を用いるときの効率は、既存の仮説検定の方法と比較して高いという結果が得られた。ここで、マーカー陽性集団で治療効果が示された場合には、当該無効中止の基準は、マーカー陽性集団に適応を制限するための証拠となり得る。

なお、第 4 章の（1）式において、マーカー陰性集団での治療効果に対して無情報の事前分布が使用されているが、信頼できる事前情報がある場合には、情報のある事前情報も使用できる。また、本研究の提案法では、マーカー陰性集団に対しては、治療効果の優越性ではなく無効中止の基準を用いるため、一般的な無効中止に関する中間解析での評価・モニタリングで見られるように、厳格な無効中止の基準（ または、無効中止に関する統計的に有意な結果 ）は要求されないと考えることは妥当であろう。つまり、フォーマルな統計的な評価を行う基準でなく、臨床的な評価も反映した様々な無効中止基準を考えてよい。ただし、検証的試験としての重要な要素は、無効中止基準（ 例えば、臨床的に意味のある治療効果の基準 ）を事前に規定することである。なお、無効中止基準を事前に規定する試験の計画段階では、マーカー陰性集団も含めたマーカー規定集団に対して第 5 章で提案した作用特性に関する評価基準を満たすようなサンプルサイズの設計も同時になされることに注意する。

本研究の三つ目の提案は、全集団、マーカー陽性集団それぞれ、あるいは、いずれ

かに対して治療効果を主張する確率（ P_{overall} 、P_{subgroup} 、および、P_{success} ）を導入し、それに基づいて、臨床的に意味のある治療効果の観点から、試験の作用特性を規定することである [1] 。先行研究では、事後的な評価の枠組みでこれに近い基準が提案されている [10] [13] [14] [15] 。例えば、新規治療の適応をマーカー陽性集団に限定するかどうかについての意思決定において、 Millen らは、本研究で提案した基準に近い基準の使用を提案している [14] 。具体的には、影響条件（ influence condition ）と呼ばれるものが一例であり、そこでは、全集団とマーカー陽性集団の両方で治療効果が有意であった場合には、マーカー陰性集団での治療効果の評価を行う。もし影響条件を反映した基準をフォールバック法や co-primary 解析（ つまり、主要解析の中 ）に導入されたなら、4.2 節で提案した方法と同様の性能を示すことが期待される。

繰り返しになるが、臨床試験を実施する者（ 研究者または製薬会社 ）によって規定された特定の作用特性を持つ第Ⅲ相臨床試験における検証的な位置づけの解析では、部分集団に対する評価を事前に規定すること（ 例えば、マーカー陰性集団に対する無効中止基準の事前規程 ）は必須である。 その一方で、マーカー部分集団での事後的な評価基準など（ 例えば、影響条件などを用いた事後評価基準 ）は、むしろ、臨床試験の評価者（ 規制当局者など ）が臨床試験の結果を評価または解釈する際の道具として役に立つものであろう。

最後に今後の研究の発展についてであるが、マーカー陰性集団に対して無効中止基

準を使用することは有効であることから、これを中間解析の評価・モニタリングで使用することが考えられる。この方向性では、アダプティブな適応患者の選択など、いくつかの提案がある [23] [25] 。さらに、最終解析で、アダプティブな適応患者の選択と様々な統計的仮説検定（ 本研究での提案法を含む ）の組み合わせを検討することも可能である。このような複合的な仮説検定の方法の評価については今後の課題にしたい。

第 8 章　付録 A：統計的仮説検定の漸近分布

生存期間を評価するマーカーで層別した第Ⅲ相臨床試験を想定し、治療群間で比例ハザード性を仮定する。治療群間の比較のためにログランク検定を行う。群間での同一の割り付け率、および、同一の観察期間のもとで、ログランク検定における検定統計量 $S \sim N(\theta, 4/E)$ の漸近分布を用いる [26] 。ここで、θ は新規治療でのハザード関数と既存治療（ 対照 ）でのハザード関数の比の対数であり（ 負の値は新規治療が有効であることを示す ）、E は観察されたイベント数の総数である。与えられたイベント数について、マーカー陽性集団における治療効果の統計的仮説検定のために標準化した検定統計量を $Z^{(+)} = \hat{\theta}^{(+)} / \sqrt{V^{(+)}}$ と表現する。 ここで、$\hat{\theta}^{(+)}$ とは $\theta^{(+)}$ の推定値、$V^{(+)} = 4/E^{(+)}$ である。また、マーカー陰性集団における治療効果の統計的仮説検定のために、同様に標準化した検定統計量 $Z^{(-)}$ を考える。さらに、全集団おける治療効果の統計的仮説検定のためには、標準化した検定統計量 $Z^{(o)} = \hat{\theta}^{(o)} / \sqrt{V^{(o)}}$ を用いる。ここで、$\hat{\theta}^{(o)}$ は $\theta^{(O)}$ の推定値、$V^{(O)} = 4/E^{(O)} = 4/(E^{(+)} + E^{(-)})$ であり、以下の近似を用いる。

$$\hat{\theta}^{(o)} \approx \{(1/V^{(+)})\hat{\theta}^{(+)} + (1/V^{(-)})\hat{\theta}^{(-)}\} / (1/V^{(+)} + 1/V^{(-)}) = (E^{(+)}\hat{\theta}^{(+)} + E^{(-)}\hat{\theta}^{(-)}) / (E^{(+)} + E^{(-)})$$

これより、全集団に対する層別ログランク検定における検定統計量は以下のように表される。

$$Z^{(o)} = \frac{\sqrt{V^{(o)}}}{V^{(+)}}\hat{\theta}^{(+)} + \frac{\sqrt{V^{(o)}}}{V^{(-)}}\hat{\theta}^{(-)}$$

交互作用検定に基づく方法で用いる新規治療とマーカーの交互作用検定では、以下の検定統計量を用いる。

$$Z^{(int)} = \frac{\hat{\theta}^{(+)} - \hat{\theta}^{(-)}}{\sqrt{V^{(+)} + V^{(-)}}}$$

前述した分散 1 に標準化した検定統計量に関して正規性を仮定する。 ① $Z^{(+)}$、② $Z^{(-)}$、③ $Z^{(O)}$ 、および、④ $Z^{(int)}$ の期待値は、それぞれ ① $\theta^{(+)}/\sqrt{V^{(+)}}$ 、② $\theta^{(-)}/\sqrt{V^{(-)}}$ 、③ $\sqrt{V^{(o)}}(\theta^{(+)}/V^{(+)} + \theta^{(-)}/V^{(-)})$ 、および、 ④ $(\theta^{(+)} - \theta^{(-)})/\sqrt{V^{(+)} + V^{(-)}}$ となる。

検定統計量の共分散（ または相関 ）に関して、 $Z^{(+)}$ と $Z^{(-)}$ は、互いに独立であると仮定でき、共分散は 0 となる。 $Z^{(+)}$ と $Z^{(o)}$との間の共分散は $\sqrt{p}$ とみることができる [14] [15] 。 交互作用検定と全集団/部分集団での検定統計量との間の共分散は、以下のように表すことができる。

$$\mathrm{cov}(Z^{(\mathrm{int})}, Z^{(\mathrm{o})}) = 0$$

$$\mathrm{cov}(Z^{(\mathrm{int})}, Z^{(+)}) = \sqrt{V^{(+)}/(V^{(+)}+V^{(-)})} = \sqrt{E^{(-)}/(E^{(+)}+E^{(-)})} = \sqrt{R/(1+R)},$$

ここで、 $R = E^{(-)}/E^{(+)}$ である。また、第 6 章で仮定したように（ イベント数の観点からの陽性割合として ） $E^{(+)} = pE$ または $R = (1-p)/p$ としたとき、$\mathrm{cov}(Z^{(\mathrm{int})}, Z^{(+)}) = \sqrt{1-p}$ となる。

一般的に、近似である $R=(1-p)/p$ と $\mathrm{cov}(Z^{(\mathrm{int})}, Z^{(+)}) = \sqrt{1-p}$ を用いるよりはむしろ、帰無仮説シナリオ 1（ グローバル帰無仮説 ）が成立しているもとで期待されるイベント率 R を用いて、共分散 $\mathrm{cov}(Z^{(\mathrm{int})}, Z^{(+)}) = \sqrt{R/(1+R)}$ に基づき、有意水準の値を求めることができる。期待されるイベント率 R は、一般に、それぞれの（ ベースラインの ）イベント率（ 若干の予後効果の可能性がある ）、および、マーカー規定集団内での打ち切りの分布に依存する。

マーカー陽性集団および陰性集団のいずれに対しても治療効果がない（ その結果、全集団に対して治療効果がない ）というに帰無仮説シナリオ 1（ グローバル帰無仮説 ）が成立しているもとで、フォールバック法、ハイブリット法（ 緩やかな有意水準を用いた MaST 法 ）、交互作用検定に基づく方法、および、提案法での試験あたりの第一種の過誤確率（ study-wise alpha rate ）を制御するために有意水準 α_0（ 検定統

計量： $Z^{(o)}$ ）、 α_s （ 検定統計量： $Z^{(+)}$ ）、および、α_{int} （ 検定統計量： $Z^{(int)}$ ）を定めることができる。

提案法では、マーカー陰性集団での無効中止基準（1）は以下により与えられる。

$$Z^{(-)} = \hat{\theta}^{(-)} / \sqrt{V^{(-)}} > \delta / \sqrt{V^{(-)}} + z_\gamma = c^{(-)} / \sqrt{V^{(-)}} = c_Z^{(-)}$$

帰無仮説シナリオ 1（ グローバル帰無仮説 ）$H_{G,0}$ が成立しているもとで、試験あたりの第一種の過誤確率（ study-wise alpha rate ）は、以下のように表される。

$$\Pr(Z^{(-)} > c_Z^{(-)} \,\&\, Z^{(+)} < c_Z^{(+)} \mid H_{G,0}) + \Pr(Z^{(-)} \le c_Z^{(-)} \,\&\, Z^{(o)} \le c_Z^{(o)} \mid H_{G,0})$$

前半部分と後半部分はそれぞれ $P_{subgroup}$ と $P_{overall}$ に対応する。 $P_{subgroup}$ の計算は、$Z^{(+)}$ と $Z^{(-)}$ との間の独立性に基づいているのに対して、$P_{overall}$ の計算では、$\mathrm{cov}(Z^{(o)}, Z^{(-)}) = \sqrt{1-p}$ をもつ多変量正規分布に基づいている。同様に、帰無仮説が成立しない場合の治療効果に対してもこれらの確率を計算することが可能である。

第 9 章　謝辞

本研究を進めていく上で研究の方向性に関して指導ならびに議論をしてくださった国立大学法人名古屋大学大学院医学系研究科臨床医薬学講座生物統計分野の松井茂之教授、に心より感謝申し上げます。

本研究のために有用な助言をくださった共同研究者の兵庫医科大学医学部 井桁正尭 助教に深く御礼申し上げます。

また、研究を継続するにあたって、研究が一向に進まないときでも見捨てず、様々な点で支えてくださった統計数理研究所のみなさま、家族に感謝を表します。

参考文献

1. Nonaka T, Igeta M and Matsui S (2019) Statistical testing strategies for assessing treatment efficacy and marker accuracy in phase III trials. Pharm Stat 18:459-475. doi: 10.1002/pst.1937

2. Slamon DJ, Leyland-Jones B, Shak S, Fuchs H, Paton V, Bajamonde A, Fleming T, Eiermann W, Wolter J, Pegram M, Baselga J and Norton L (2001) Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. N Engl J Med 344:783-92. doi: 10.1056/NEJM200103153441101

3. Chapman PB, Hauschild A, Robert C, Haanen JB, Ascierto P, Larkin J, Dummer R, Garbe C, Testori A, Maio M, Hogg D, Lorigan P, Lebbe C, Jouary T, Schadendorf D, Ribas A, O'Day SJ, Sosman JA, Kirkwood JM, Eggermont AM, Dreno B, Nolop K, Li J, Nelson B, Hou J, Lee RJ, Flaherty KT, McArthur GA and Group B-S (2011) Improved survival with vemurafenib in melanoma with BRAF V600E mutation. N Engl J Med 364:2507-16. doi: 10.1056/NEJMoa1103782

4. Shaw AT, Kim DW, Nakagawa K, Seto T, Crino L, Ahn MJ, De Pas T, Besse B, Solomon BJ, Blackhall F, Wu YL, Thomas M, O'Byrne KJ, Moro-Sibilot D, Camidge DR, Mok T, Hirsh V, Riely GJ, Iyer S, Tassell V, Polli A, Wilner KD and Janne PA

(2013) Crizotinib versus chemotherapy in advanced ALK-positive lung cancer. N Engl J Med 368:2385-94. doi: 10.1056/NEJMoa1214886

5. Simon R and Maitournam A (2004) Evaluating the efficiency of targeted designs for randomized clinical trials. Clin Cancer Res 10:6759-63. doi: 10.1158/1078-0432.CCR-04-0496

6. Mandrekar SJ and Sargent DJ (2009) Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. J Clin Oncol 27:4027-34. doi: 10.1200/JCO.2009.22.3701

7. Hoering A, Leblanc M and Crowley JJ (2008) Randomized phase III clinical trial designs for targeted agents. Clin Cancer Res 14:4358-67. doi: 10.1158/1078-0432.CCR-08-0288

8. Buyse M, Michiels S, Sargent DJ, Grothey A, Matheson A and de Gramont A (2011) Integrating biomarkers in clinical trials. Expert Rev Mol Diagn 11:171-82. doi: 10.1586/erm.10.120

9. Freidlin B and Korn EL (2014) Biomarker enrichment strategies: matching trial design to biomarker credentials. Nat Rev Clin Oncol 11:81-90. doi: 10.1038/nrclinonc.2013.218

10. Matsui S, Choai Y and Nonaka T (2014) Comparison of statistical analysis plans in randomize-all phase III trials with a predictive biomarker. Clin Cancer Res 20:2820-

30. doi: 10.1158/1078-0432.CCR-13-2698

11. Rothmann MD, Zhang JJ, Lu L and Fleming TR (2012) Testing in a Prespecified Subgroup and the Intent-to-Treat Population. Drug Inf J 46:175-179. doi: 10.1177/0092861512436579

12. Freidlin B, Korn EL and Gray R (2014) Marker Sequential Test (MaST) design. Clin Trials 11:19-27. doi: 10.1177/1740774513503739

13. Simon RM (2013) Genomic clinical trials and predictive medicine. Cambridge University Press, Cambridge.

14. Millen BA, Dmitrienko A, Ruberg S and Shen L (2012) A Statistical Framework for Decision Making in Confirmatory Multipopulation Tailoring Clinical Trials. Therapeutic Innovation & Regulatory Science 46:647-656.

15. Millen BA, Dmitrienko A and Song G (2014) Bayesian assessment of the influence and interaction conditions in multipopulation tailoring clinical trials. J Biopharm Stat 24:94-109. doi: 10.1080/10543406.2013.856025

16. Dmitrienko A, Tamhane, A. C., Bretz, F. (2010) Multiple Testing Problems in Pharmaceutical Statistics. Chapman & Hall/CRC.

17. Simon R (2010) Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. Per Med 7:33-47. doi: 10.2217/pme.09.49

18. Douillard JY, Siena S, Cassidy J, Tabernero J, Burkes R, Barugel M, Humblet Y,

Bodoky G, Cunningham D, Jassem J, Rivera F, Kocakova I, Ruff P, Blasinska-Morawiec M, Smakal M, Canon JL, Rother M, Oliner KS, Wolf M and Gansert J (2010) Randomized, phase III trial of panitumumab with infusional fluorouracil, leucovorin, and oxaliplatin (FOLFOX4) versus FOLFOX4 alone as first-line treatment in patients with previously untreated metastatic colorectal cancer: the PRIME study. J Clin Oncol 28:4697-705. doi: 10.1200/JCO.2009.27.4860

19. Cappuzzo F, Ciuleanu T, Stelmakh L, Cicenas S, Szczesna A, Juhasz E, Esteban E, Molinier O, Brugger W, Melezinek I, Klingelschmitt G, Klughammer B, Giaccone G and investigators S (2010) Erlotinib as maintenance treatment in advanced non-small-cell lung cancer: a multicentre, randomised, placebo-controlled phase 3 study. Lancet Oncol 11:521-9. doi: 10.1016/S1470-2045(10)70112-1

20. Song Y and Chi GY (2007) A method for testing a prespecified subgroup in clinical trials. Stat Med 26:3535-49. doi: 10.1002/sim.2825

21. Spiessens B and Debois M (2010) Adjusted significance levels for subgroup analyses in clinical trials. Contemp Clin Trials 31:647-56. doi: 10.1016/j.cct.2010.08.011

22. Berry SM, Carlin, B.P., Lee, J.J., Muller P. (2010) Bayesian Adaptive Methods for Clinical Trials. Chapman & Hall/CRC.

23. Brannath W, Zuber E, Branson M, Bretz F, Gallo P, Posch M and Racine-Poon A (2009) Confirmatory adaptive designs with Bayesian decision tools for a targeted

therapy in oncology. Stat Med 28:1445-63. doi: 10.1002/sim.3559

24. Peeters M, Price TJ, Cervantes A, Sobrero AF, Ducreux M, Hotko Y, Andre T, Chan E, Lordick F, Punt CJ, Strickland AH, Wilson G, Ciuleanu TE, Roman L, Van Cutsem E, Tzekova V, Collins S, Oliner KS, Rong A and Gansert J (2010) Randomized phase III study of panitumumab with fluorouracil, leucovorin, and irinotecan (FOLFIRI) compared with FOLFIRI alone as second-line treatment in patients with metastatic colorectal cancer. J Clin Oncol 28:4706-13. doi: 10.1200/JCO.2009.27.6055

25. Magnusson BP and Turnbull BW (2013) Group sequential enrichment design incorporating subgroup selection. Stat Med 32:2695-714. doi: 10.1002/sim.5738

26. Tsiatis AA (1981) The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time. Biometrika 68:311–315.