学　位　論　文

臨床試験において既存対照データを利用するための

検定併合法の研究

(Studies on a test-then-pool method

for incorporating historical control data in clinical trials)

２０２４年３月

北　海　道　大　学

岡　田　和　史

Kazufumi Okada

学　位　論　文

臨床試験において既存対照データを利用するための

検定併合法の研究

(Studies on a test-then-pool method

for incorporating historical control data in clinical trials)

２０２４年３月

北　海　道　大　学

岡　田　和　史

Kazufumi Okada

目　次

発表論文目録および学会発表目録

本研究の一部は以下の論文に発表した.

1. Kazufumi Okada, Shiro Tanaka, Jun Matsubayashi, Keita Takahashi, Isao Yokota
   De-coupling power and type I error rate considerations when incorporating historical control data using a test-then-pool approach
   Biometrical Journal, 2024, 66(1), 2200312.

本研究の一部は以下の学会に発表した.

1. Kazufumi Okada, Isao Yokota
   Two one-sided test-then-pool method for clinical trials
   43rd Annual Conference of the International Society for Clinical Biostatistics, 21th -25th August 2022. Newcastle, UK.
2. 岡田和史，横田勲
   Test-then-pool 法と傾向スコア重み付け法を組み合わせた既存試験データを併合するための二段階アプローチ
   2023 年度日本計量生物学会年会，2023 年 4 月 20-21 日．札幌.

要旨

【背景と目的】

近年，無作為化臨床試験の対照群の一部に，既に得られているデータを用いる臨床試験デザインが注目されている．このような試験デザインは，ハイブリッド対照デザインと呼ばれる．用いられる対照データは，過去に行われた臨床試験や観察研究のデータ，リアルワールドデータなどである．本研究では，これらのデータを既存対照データと呼ぶ．既存対照データを新規試験の対照群の一部に用いることで，新規試験の参加者数を削減し，臨床試験の実施可能性を高めることができる．

既存対照データを新規試験の解析に利用する際の重要な仮定は，新規試験の対照データと既存対照データの従う母集団の同一性である．母集団が同一でない場合，治療効果の推定にバイアスが生じ，その方向に応じて第一種の過誤確率が上昇する，あるいは検出力が低下する恐れがある．

母集団の同一性を評価する定性的な項目として，データの収集時期や施設，地域，及び各データの患者背景分布等があげられる．これらの項目はできるだけ同じであることが望ましい．しかしながら，これらが完全に一致した既存対照が存在することは稀であり，また一致していたとしても，未知の要因によって母集団にずれが生じる可能性は否定できない．

したがって，ハイブリッド対照デザインを利用する際は，動的な既存対照併合法の利用が望ましい．動的な既存対照併合法は，既存対照と新規対照の母集団の同一性を，得られたデータから評価し，解析に与える既存対照の影響を制御する統計的な方法である．動的な既存対照併合法は，多数の方法が提案されている．どの方法も，既存対照データと新規対照データの差異が小さい場合に，既存対照データの影響を強め，差異が大きい場合に，既存対照データの影響を弱める方法である．

本研究では，動的な既存対照併合法の 1 つである，検定併合法に注目する．検定併合法は，既存対照と新規対照の差異を，統計的仮説検定を用いて評価する．帰無仮説が棄却された場合，併合が受け入れ不可能であると判断し，既存対照を併合しない解析を行う．一方で，帰無仮説が棄却されない場合，併合が許容できると判断し，既存対照を単純に併合した解析を行う．検定併合法の利点は，その他の既存対照併合法と比較して，既存対照の利用法に透明性があり，臨床家にも解析結果の解釈が容易である点にある．しかし，検定併合法にはいくつか課題も存在しており，ハイブリッド対照デザインの標準的な方法として位置づけられているわけではない．本研究の目的は，検定併合法の課題を解決する方法の提案を通して，検定併合法の利便性を高めることである．

【方法】

検定併合法の課題の 1 つに，第一種の過誤確率と検出力の制御に関して柔軟性が乏しい点があげられる．第二章では，この課題を解決するために，新しい検定併合法を提案する．従来の検定併合法は，既存対照の併合可否を判断するために，両側検定を用いており，併合のされやすさを制御するチューニングパラメータである検定の有意水準が 1 つのみである．そのため，この値を決めると，第一種の過誤確率と検出力が同時に定まる．そこで，既存対照の併合可否を判断するための両側検定を，2 つの片側検定に分割し，それぞれに異なる片側有意水準を与える方法を提案する．片方の有意水準が第一種の過誤確率の制御を担当し，もう片方の有意水準が検出力の制御を担当する．

検定併合法の批判の 1 つは，既存対照の併合可否を判断するための検定の有意水準の決め方にコンセンサスが得られた方法はないという点である．先行研究では，0.1 や 0.2 が用いられているが，それらの値を用いる根拠はない．提案検定併合法では，有意水準が 2 つになるため，ますます有意水準を決めることが難しい．そこで，第二章では，最高第一種の過誤確率と最低検出力に基づいて検定併合法の有意水準を決める方法を併せて提案する．この方法は，最高第一種の過誤確率と最低検出力が許容水準に収まる有意水準の中で，新規対照と既存対照の差異がない場合の検出力が最も大きくなるものを選択する．

第三章では，検定併合法が，新規対照と既存対照のアウトカム分布の違いのみを考慮しており，患者背景の分布を考慮していない点を改良する．患者背景の違いを考慮するために，傾向スコアを用いた調整法と既存対照借用法を組み合わせる 2 段階法に注目する．2 段階法では，1 段階目に傾向スコア法を用いて，既存対照の患者背景を新規試験参加者に類似させた疑似集団を作成し，2 段階目に既存対照借用法を用いて解析を行う．1 段階目と 2 段階目に用いる方法の組合せは自由である．先行研究では，2 段階目にベイズ流の既存対照借用法を利用した 2 段階法がいくつか提案されている．本研究では，1 段階目に傾向スコア重み付け法，2 段階目に検定併合法を利用する 2 段階法を提案する．

【結果と考察】

第二章では，仮想的な臨床試験を想定した数値実験を行った．数値実験の結果，提案有意水準選択法を用いると，提案検定併合法の 2 つの片側有意水準の内，1 つは最高第一種の過誤確率を制御するように決まることが分かった．一方で，もう 1 つの片側有意水準は，最低検出力ではなく，新規対照と既存対照の差異がない場合の検出力を最大化するように決まることが分かった．提案有意水準選択法を用いると，提案検

定併合法は，従来の検定併合法と比較して，新規対照と既存対照の差異がある場合の検出力の低下を防ぐ，あるいは新規対照と既存対照の差異がない場合の検出力を高めることが分かった．臨床試験の設定に依っては，第一種の過誤確率と検出力が，従来の検定併合法とほとんど変わらないにも関わらず，提案検定併合法の既存対照併合確率が，従来の検定併合法に劣る場合があった．これは，提案有意水準選択法が第一種の過誤確率と検出力のみを考慮しており，併合確率を考慮していないからである．提案有意水準選択法には，改良の余地が残されていると考えられる．

第三章では，膵臓癌の患者を対象とした非劣性試験を参考にした数値実験を行った．数値実験の結果，極端な重みの影響を緩和する安定化重みを用いた傾向スコア重み付け法と検定併合法を組み合わせた 2 段階法が，検定併合法単体や傾向スコア重み付け法単体と比較して，バイアスの小さい推定を可能とし，第一種の過誤確率の上昇と検出力の低下を軽減することが分かった．本研究では，傾向スコア重み付け法と検定併合法の組合せのみを対象としている．傾向スコアマッチング法と検定併合法を組み合わせた 2 段階法や，ベイズ流の既存対照借用法を用いた 2 段階法との比較が望まれる．

【結論】

本研究では，検定併合法のいくつかの課題を解決するための方法を提案した．これらの提案は，互いに対立するものではなく，組み合わせて利用可能である．本研究の提案は，検定併合法の利便性を向上させるものであり，研究対象者数の制約による臨床試験の実施可能性が乏しい状況において，検定併合法を用いたハイブリッド対照デザインという選択肢を提供する．本研究は，臨床試験の実施可能性を向上させることを通して，医学研究の発展に寄与すると考えられる．

略語表

本文中および図中で使用した略語は以下のとおりである.

AWCS	adequate and well-controlled study
ATT	average treatment effect of the treated
CC	current control
CDD	covariate distribution difference
CTTP	conventional one-sided test-then-pool method
ETTP	equivalence-based test-then-pool method
FDA	Food and Drug Administration
HC	historical control
HAMA	Hamilton rating scale for anxiety
HAMD	Hamilton depression rating scale
IPW	inverse probability weighting
MSE	mean squared error
PHC	pseudo historical control
PSM	propensity score matching
PSW	propensity score weighting
PSWTTP	propensity score weighting + test-then-pool method
SPSW	stabilized propensity score weighting
SPSWTTP	stabilized propensity score weighting + test-then-pool method
T	treatment
TTP	test-then-pool method
TTTP	two one-sided test-then-pool method

第一章

序論

1.1 臨床試験における既存対照の利活用

無作為化比較試験は，新規治療の有効性を示すためのゴールドスタンダードとされている（Meldrum, 2000; Moher et al., 2012; Folefac and Desmond, 2022）．被験者を無作為に新規治療群と対照群に割り付けることで，患者背景の類似した 2 つの集団を作成できる．観察された治療効果の違いは，治療の違いによってもたらされたと考えられるため，因果的な解釈が可能となる（Rubin, 2008）．医薬品の承認においても，無作為化比較試験が求められることが多い．U.S. Food and Drug Administration（FDA）は承認の根拠を，（1）adequate and well-controlled study（AWCS）と（2）AWCS の再現性を評価するための検証的な substantial evidence としているが（U.S. Food and Drug Administration, 2019a），AWCS の要件から，無作為化比較試験が最も適切なデザインであると言われている（野村他，2023）．

近年，対照群に関して，既存対照データを新規試験の解析にも利用するという試みが注目されている（Schmidli et al., 2020; Wu et al., 2020; Mishra-Kalyani et al., 2022; Beckman et al., 2022; Campbell et al., 2022）．本研究では，既存対照データとは，過去に行われた臨床試験や観察研究のデータや，リアルワールドデータを指す．リアルワールドデータは“患者の健康状態及び医療の提供に関連する，様々なデータソースから日常的に収集されるデータ”と定義される（U.S. Food and Drug Administration, 2022）．無作為化比較試験は優れた研究デザインではあるものの，その実施可能性がしばしば問題になる．例えば，患者数の限られている希少疾患においては，1 つの研究内で無作為化して治療間を比較することが難しい（Carlin and Nollevaux, 2022）．金銭的なコストも問題となる．近年，臨床試験の品質を担保するためのコストかさみ，臨床試験のための費用は，製薬企業の生産性を超えて増大している（Paul et al., 2010）．本邦では，2018 年の臨床研究法施行後，臨床試験の金銭的負担が増加しており，臨床試験の実施数の減少が予想されている（Nakamura and Shibata, 2020）．臨床試験の実施可能性を向上させる手段として，既存対照の利用は魅力的な選択肢の 1 つとなり得る．通常，標準治療に関する情報は，医学研究や日常診療から十分に蓄積されている．これらの情報は，標準的な臨床研究において，症例数設計など研究計画の面で活用されてきたものの，解析に直接的に用いられることは少ない（Roychoudhury and Neuenschwander, 2020）．臨床研究の解析においても，既存対照データを積極的に活用することで，治療効果の推定精度の向上につながり，新規試験参加者の削減に寄与する可能性がある

（Berry, 2006; Viele et al., 2014）．欧米では，既に既存対照を利用した承認申請も散見されており（Gross, 2021; Wang et al., 2023），本邦の承認申請においても，既存対照の利活用が進んでいくと予想される．

既存対照データの解析での利用は，新規試験の必要参加者数の削減につながる．臨床試験において, 倫理的な観点から参加者数は少なすぎても多すぎても好ましくなく，必要十分な参加者数で試験を実施する必要がある（Feißt at al., 2020）．参加者数の少なすぎる試験は，実質的な検出力が低くなりやすく，そのような研究に患者の協力を仰ぐことは倫理的ではない（Halpern et al., 2002）．一方で，参加者数の多すぎる試験は，不必要な人数の患者に負担を強いることになる（Altman, 1980）．もし，既存対照データを解析で利用できれば，必要十分な新規試験の参加者数を減らすことができる．また，新規試験参加者数の縮小は，試験期間の短縮につながる．試験の早期終了は，有効な治療を受ける患者を増やす，あるいは無効な治療を受ける患者を減らすことになる．さらに，参加者数の減少や試験期間の短縮は，開発コストの低下につながり，患者数の少ない疾病であっても, 研究者や製薬企業が治療法を開発する動機となり得る．このように，既存対照データの利用は，必要参加者数の削減を通して，臨床試験の倫理性と実施可能性に好ましい影響を与える．

既存対照データの解析における利用方法の 1 つに，単群試験の比較対照として用いる方法があげられる．単群試験自体は，無作為化比較試験とともに広く行われている研究デザインであり，主に医薬品開発の早期の段階や希少疾患が対象の場合に用いられる（Li et al., 2023）. 近年, FDA の Accelerated Approval（U.S. Food and Drug Administration, 2014）を利用した分子標的薬の承認申請では，単群試験に基づく承認が増加している（Ribeiro et al., 2023）．しかしながら，単群試験は同時対照を設定しないため，標準治療との比較のためには，既存対照を用いる必要がある．しかし，無作為化されていないため，未測定の交絡因子によるバイアスを取り除くことは不可能であり，この点が単群試験の限界である（Burger et al., 2021）.

近年，既存対照データを使用した研究デザインとして，ハイブリッド対照デザインが注目を浴びている（Pocock, 1976; Viele et al., 2014; Burger et al., 2021）．このデザインは，対照群にも無作為割付を行う無作為化試験の一種である．この時，割付比を対照群の人数が少なくなるように設定し，不足した対照群の情報を補うために，既存対照データを利用する．既存対照と新規対照を併合して解析を行う点がハイブリッド対照デザインと呼ばれる所以である．最も単純なハイブリッド対照デザインは，新規試験の対照群のデータと既存対照データを単純に併合するものである．しかしながら，無作為化した集団と既存対照データの集団が大きく異なっている恐れがあるため，単純併合は治療効果の推定に大きなバイアスを招く恐れがある（Li et al., 2023）．ハイブリッド対照デザインの利点の 1 つは，新規試験においても対照群の情報を集めることか

ら，新規試験と既存対照の集団の差異をデータに基づいてある程度評価可能である点にある．ハイブリッド対照デザインでは，2 つの集団の異質性の大きさをデータから評価し，その大きさに基づいて既存対照データが解析に寄与する程度を制御できる．このような既存対照利用法は数多く提案されており，その一部を 1.3 節で紹介する．

1.2 併合受容性の評価

既存対照データの利用を計画する際，併合受容性を注意深く検討する必要がある．ここで，併合受容性とは，既存対照を新規対照に併合したとしても，試験結果の妥当な解釈可能であるといえる性質と定義する．Pocock（1976）は，併合受容性が満たされる条件として，6 つの条件を挙げている．ここでは武田ら（2015）による日本語訳を紹介する．

1. 詳細に定義された，新規試験の対照治療群と同じ治療を受けている．
2. 新規試験と同じ適格基準を用いた最近実施された臨床試験である．
3. 治療の評価方法が新規試験と同じである．
4. 重要な患者背景の分布が新規試験と同様である．
5. 新規試験とほぼ同じ組織で実施されている．
6. 新規試験との結果の違いを生むと予想される他の要因が存在しない．例えば，新規試験において症例登録スピードが予想よりも上がらないとき，適格基準をみたすぎりぎりの症例が登録されることも起こりうる．そういった場合，新規試験と既存試験で共通の適格基準を用いたとしても登録患者の背景が二つの試験間で異なってくる可能性がある．

条件 1 と条件 3 は，併合解析の臨床的な解釈を可能とするために必要な基本的な要件であると考えられる．条件 1 について，薬剤の種類や治療方針が異なるデータを併合したとしても臨床的な解釈はできないであろう．条件 3 について，例えば異なる QOL 指標を比較に用いることはできない．条件 2 と条件 4 から 6 は，母集団の同一性を確保するために規定される条件であると解釈できる．条件 2 について，既存対照データが収集された時期が新規試験を実施する時期と離れている場合，医療技術の進歩のような時代効果が生じているかもしれない．条件 4 について，アウトカムに関連する重要な因子の分布が既存対照と新規対照で異なっていた場合，適切に調整しなければ交絡によるバイアスが生じてしまう．条件 5 について，研究の実施組織が異なることで，研究対象集団に違いが生じてしまうかもしれない．条件 6 について他に考えられる例を挙げるとすれば，既存対照と新規試験の研究目的の違いが患者背景の違いにつながる恐れがある．例えば，既存試験が承認を目的とした第 3 相試験，新規試験が

承認後の薬剤の投与方法に関する追加的な研究である場合，既存試験に比べて新規試験では，予後の良い，重篤ではない参加者の割合が増えてしまい，患者背景の違いを生じさせる恐れがある．

母集団の同一性は，併合受容性の一部であり，統計的に極めて重要な仮定である．文献によっては，新規対照と既存対照の一貫性や類似性，ベイズ統計学の文脈では交換可能性と呼ばれることもあるが，本研究では母集団の同一性と表現する．新規試験と既存対照の母集団が同一であるならば，既存対照の併合によって，バイアスをもたらすことなく推定精度の向上および検出力の上昇を実現することができる．しかしながら，母集団が同一ではない場合，治療効果の推定にバイアスが生じ，その方向に応じて第一種の過誤確率の上昇や検出力の低下が起こる（Cuffe, 2011; Viele et al., 2014）．母集団の同一性は，既存対照の従う分布と新規対照の従う分布のパラメータの違いとして解釈される．ここで，既存対照をHC，新規対照をCCと表すこととし，各対照が従う分布を規定するパラメータをθ_g $(g \in \{HC, CC\})$とする．θ_gは，具体的に平均値や奏効割合が考えられる．$\theta_{HC} = \theta_{CC}$であれば，母集団が同一であることを意味し，$\theta_{HC} \neq \theta_{CC}$であれば，母集団が異なることを意味する．

ここで，母集団の同一性に関連する概念として”drift”を紹介する．Drift は既存対照と新規対照の差異を表す（Viele et al., 2018）．Viele らは drift を観察されている既存対照と新規対照の分布を規定する真のパラメータと定義した（Viele et al., 2018）．Lim らは，観察された既存対照と新規試験のパラメータが，新規試験の動作特性（第一種の過誤確率や検出力など）を規定することから， Viele らと同様に drift を定義した（Lim et al., 2020）．しかし，単に真のパラメータの差を drift と呼ぶ研究者もいれば（例えば Wang et al., 2022），観察されたデータの違いという意味で drift を用いる研究者もいる（例えば Jemielita et al., 2020）．Drift は”既存対照と新規対照の違い”を表す概念として認識されていると思われる．真に重要なことは，$\theta_{HC} = \theta_{CC}$が満たされているかどうかである．これらのパラメータは未知であるため，本当に成り立っているかどうかを証明する方法はない．”Drift”という言葉が使われたときに，必ずしもθ_{HC}とθ_{CC}の差異が意図されていないことに注意されたい．

Viele らの drift の定義に基づいて第一種の過誤確率の上昇と第二種の過誤確率の上昇（すなわち検出力の低下）を説明した概念図を図 1.1 に示す．ここでは，治療効果に関する仮説として，

$$H_0: \mu_T - \mu_{CC} > 0 \text{ versus } H_1: \mu_T - \mu_{CC} \leq 0. \tag{1.1}$$

を考えている．ここで，μ_Tは新規治療の母平均，μ_{CC}は新規対照の母平均を表す．すなわち，平均値が下がると治療効果があると考える．図 1.1 の 4 つの曲線は，既存対照を新規対照に単純併合した場合の治療効果に対する検定統計量の分布を表す．このうち点線で示した 2 つの分布は，それぞれ帰無仮説H_0と対立仮説H_1の下での，drift が 0

である場合の検定統計量の分布である．また，実線で示した 2 つの分布は，drift が 0 ではない場合の検定統計量の分布である. 帰無仮説と対立仮説の位置関係は, 仮説(1.1)に基づいている．Drift が正，すなわち$\mu_{CC} < \bar{x}_{HC}$の時は，治療効果を過大評価する方向にバイアスが生じるため，H_0の下での検定統計量は左にシフトし，第一種の過誤確率が上昇する．逆に，drift が負，すなわち$\mu_{CC} > \bar{x}_{HC}$の時は，治療効果を過小評価する方向にバイアスが生じるため，H_1の下での検定統計量は右にシフトし，第二種の過誤確率が上昇する，すなわち検出力が低下する．

このように，母集団の同一性が満たされない場合，第一種の過誤確率の上昇や検出力の低下が引き起こされる．Pocock は，6 つの条件をすべて満たせば既存対照を安全に利用できるが，1 つでも満たされない場合はバイアスが生じる可能性を否定できないと指摘した（Pocock, 1976）．武田らは，心筋梗塞患者を対象とした無作為化比較試験である AMIHOT I 試験と AMIHOT II 試験（O'Neill et al., 2007; Stone et al., 2009）を，Pocock の条件を完全に満たす例として紹介した（武田他, 2015）．しかし，Pocock の条件は非常に厳しい条件であり，現実的にこれらの基準をすべて達成することは困難である（Li et al., 2023）．また，Pocock の条件がすべて満たされていたとしても，併合受容性が真に満たされている保証はない．既存対照を新規試験の解析に利用する際は，Pocock の条件等を用いた併合受容性の検討に加えて，併合受容性を担保できる統計的な手法を適用する必要がある．

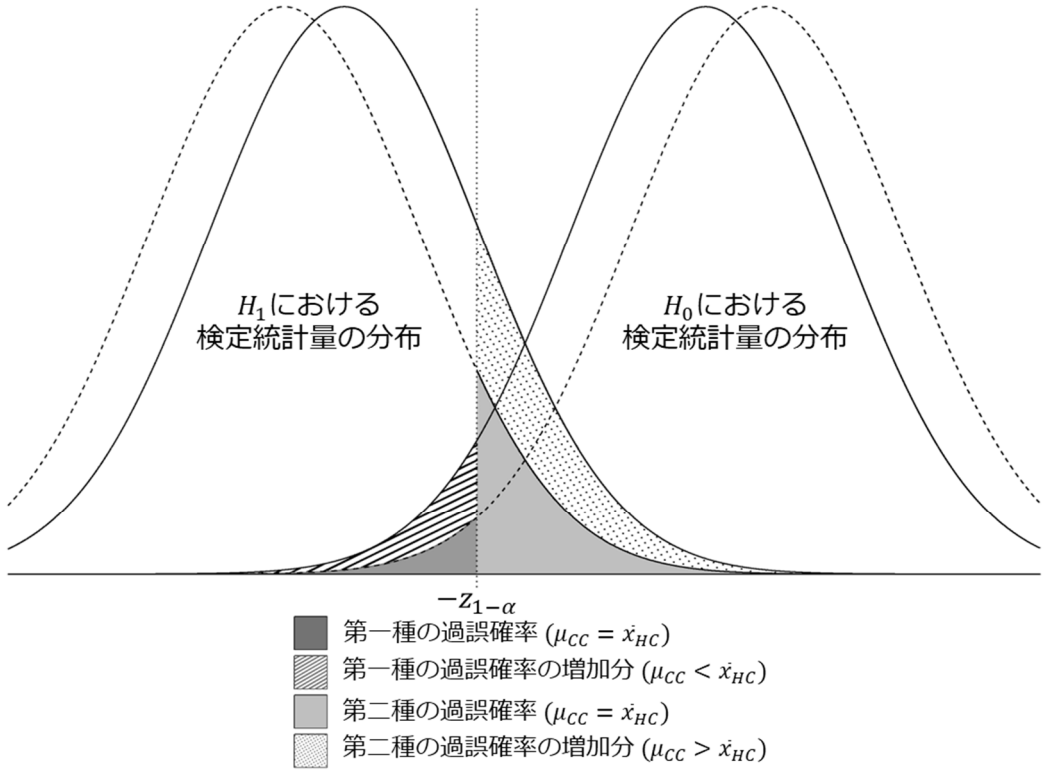

図 1.1　Drift が 0 ではない場合の第一種の過誤確率と第二種の過誤確率の増加を表した概念図．μ_{CC}は新規対照の母平均，$\bar{x}_{HC}$は既存対照の標本平均，H_0は帰無仮説，H_1は対立仮説を表す．

1.3　既存法

本節では既に提案されている既存対照借用法の一部を紹介する．

1.3.1　静的な借用法

静的な借用法（static borrowing）は，既存対照と新規対照の差異にかかわらず，既存対照を適当に重み付けしたうえで利用する方法である（Berger et al., 2021）．単純併合法は，既存対照の重みを 1 にした最も素朴な併合方法であるといえる．

既存対照データに重みを与える方法として，ベイズ流の方法である conditional power prior が有名である（Ibrahim and Chen, 2000; Chen et al., 2000）．この方法は，0 から 1 の値をとる power parameter を，既存対照の尤度にべき乗することで，既存対照の解析に与える影響を制御する．Power parameter が 0 であれば既存対照の尤度が 1 となるため，既存対照が無視される．Power parameter が 1 であれば，既存対照の尤度がそのまま残るため，ベイズ流に単純併合法を行う解析と同様になる．Power parameter と既存対照

のサンプルサイズの積は，解析に影響を与える実質的なサンプルサイズとして解釈できる（Morita et al., 2008）．Conditional power prior に基づく推論は，初期事前分布の影響を無視すれば, 頻度論の枠組みにおける重み付き尤度による推論と同等である（Psioda et al., 2018）．Power parameter を専門家の意見等を参考に決定するなど，定数として power parameter を設定する方法が考えられる．このように，power parameter を事前規定する場合，conditional power prior は静的な借用法に分類できる．

ハイブリッド対照デザインにおいて，静的な借用法は推奨されない．なぜならば，新規対照と既存対照の母集団の同一性が成り立たない場合であっても既存対照の持つ情報の大部分を解析に反映させてしまうため，大きなバイアスが生じてしまう危険性があるからである．つまり，静的な借用法は，バイアスに関して，単群試験における既存対照との比較と同様のリスクがある（Berger et al., 2021）．ハイブリッド対照デザインでは，新規試験でも対照群のデータを収集する点が特徴であるため，新規対照と既存対照の drift をデータから評価できる．母集団の同一性が疑わしい場合には，適切に既存対照の情報を割り引くことが可能な動的な借用法（dynamic borrowing）を利用すべきである．

1.3.2 動的な借用法

動的な借用法（dynamic borrowing）は，新規対照と既存対照の母集団の同一性を観察されたデータから評価し，借用の程度を決定する方法であり，静的な方法の限界を克服した方法である（Kotalik et al., 2021）．動的な借用法では，既存対照と新規対照の差が大きい時（すなわち drift が大きい時），既存対照の併合がバイアスを生じさせる可能性があるため, 借用の程度を小さくする. 動的な借用法は盛んに研究されており，様々な方法が提案されている．

頻度論の枠組みで適用可能な方法として最も有名な動的借用法は検定併合法（test-then-pool method）である（Viele et al., 2014）．この方法は，既存対照と新規対照を特徴づけるパラメータが等しいという帰無仮説を設定し，適当な有意水準γを用いて検定を行う．帰無仮説が棄却されなかった場合，既存対照と新規対照の差異は大きくないと判断し，単純併合による解析を行う．帰無仮説が棄却された場合，併合不可能な差異があると判断し, 新規試験データのみを用いて解析を行う．ここで有意水準γは併合のされやすさを制御するチューニングパラメータとして働く．γが小さければ併合されやすく，大きければ併合されにくくなる．Drift の検定について，優越性検定ではなく同等性検定を行うことによって，サンプルサイズが小さい場合でも不適切な併合を防ぐことができる同等性検定ベースの検定併合法も提案されている（Li et al., 2020）．そのほかの頻度論に基づく方法として，adaptive LASSO（least absolute shrinkage and selection operator）（Zou, 2006; Zhang and Lu, 2007）を用いた方法も提案されている（Li

et al., 2023)．

Conditional power prior を利用して動的に既存対照を借用する方法として，power parameter を既存対照と新規対照の類似度に応じてデータから推定する方法がいくつか提案されている．Gravestock らは，power parameter の周辺尤度を最大化する power parameter を利用する empirical Bayes power prior を提案した（Gravestock et al., 2017; Gravestock and Held, 2019）．Nikolapoulos らは，事前予測分布を用いて第一種の過誤確率が一定の値に抑えられるように power parameter を決定する prior-data conflict calibrated power prior を提案した（Nikolakopoulos et al., 2018）．どちらの方法も drift が大きいと，power parameter は小さく推定される．

Power parameter を定数ではなく，変数として治療効果に関するパラメータと同時に推定することで動的な借用を行う方法もある．もともと joint power prior として Ibrahim らによって conditional power prior と同時に提案されていた（Ibrahim and Chen, 2000）．この方法は，power parameter を確率変数として考え，治療効果に関するパラメータと同時に推定する．しかし，この方法は，尤度に正の定数を乗ずると power parameter と治療効果に関するパラメータの同時事前分布が変わってしまい尤度原理に反する，既存対照と新規対照の drift が小さくても既存対照の情報をほとんど利用できない，といった欠点があった（Duan et al., 2006）．そこで Duan らはこの点を修正した modified power prior を提案した（Duan et al., 2006）．この方法は normalized power prior と呼ばれることもある（Neuenschwander et al., 2009; Ibrahim et al., 2015）．さらに，modified power prior を既存対照データが複数存在する場合に拡張し，データ間の関連を考慮する dependent modified power prior も提案されている（Banbeta et al., 2019; Banbeta et al., 2022）．

ベイズ流階層モデルを応用した方法も多数提案されている．Hobbs は power prior が既存対照と新規対照の類似性を直接的に評価していない点を批判し，”共通性（commensurability）”を直接モデル化する commensurate power を提案した（Hobbs et al., 2011; Hobbs et al., 2012）．メタアナリシスの方法を応用した情報借用法として，meta-analytic predictive prior と meta-analytic combined prior がある（Spiegelhalter et al., 2004; Neuenschwander et al., 2010）．両者の違いは，解析手順にある．前者は既存対照データを用いて事前分布を構成したうえで，新規試験データを集め，事後分布を導出するのに対し，後者は新規試験データを入手した段階で既存対照データと併せて事後分布を導出する（野村他, 2022）．Schmidli らは meta-analytic predictive prior と弱情報事前分布を混合する robust meta-analytic predictive prior を提案した（Schmidli et al., 2014）．Kaizer らは，複数の既存対照データを想定し，ベイズ流モデル平均化（Hoeting et al., 1999）の重みに，新規対照データと母集団が同一である確率（交換可能性が成り立つ確率）を用いる multisource exchangeability model を提案した（Kaizer et al., 2018）．Ohigashi らは，複数の既存対照データのうち一部のデータの異質性が大きい場合に，そのデータ

の影響を軽減できる方法として horseshoe prior を用いる情報借用法を提案した（Ohigashi et al., 2022）.

1.3.3 患者背景の違いを調整する方法

1.3.2 節で紹介した方法は，観察されたアウトカムの違いに応じて借用の強度を調節する方法である．アウトカム分布だけではなく，共変量分布も考慮することで，併合受容性が高まると考えられている（Psioda et al., 2018）．特に Pocock の条件の 4 つ目が満たされていない場合でも，妥当な併合を達成できる可能性がある．共変量調整の方法は主に，回帰分析による方法と傾向スコア（Rosenbaum and Rubin, 1983）を用いた方法の 2 つがあるが（Fu et al., 2023），近年は傾向スコアを用いた方法が特に注目されている. 一般に, 傾向スコアを用いた解析には, マッチング, 重み付け（Sato and Matsuyama, 2003），層別，回帰モデルの説明変数に加える方法の 4 つが考えらえれる（Austin, 2011; Lin and Lin, 2022）．この中でも前者 3 つは，傾向スコア調整と解析を分離できるという点でハイブリッド対照デザインと相性が良い（Wang et al., 2022）．この性質を利用して，傾向スコアによる調整と 1.3.2 節で紹介した方法を組み合わせる 2 段階法がいくつか提案されている．例えば，Lin らは傾向スコアマッチングと power prior を組み合わせた方法（Lin et al., 2018），Wang らは傾向スコアを用いた層別と power prior を組み合わせた方法（Wang et al., 2019），Zhao らは傾向スコアマッチングと commensurate prior を組み合わせた方法を提案した（Zhao et al., 2016）．Wang らは複数の情報借用法と傾向スコア調整法の組合せを比較したシミュレーション研究を行った（Wang et al., 2022）.

1.3.4 頻度流とベイズ流

ハイブリッド対照デザインの文脈では，ベイズ流の方法が数多く提案されているものの，頻度流の方法は少ない（Zhou and Ji, 2021; Li et al., 2023）．ベイズ流の方法では，ベイズの定理に基づいて，事前情報を解析に明示的に反映できる．既存対照データは事前情報の 1 種であるため，既存対照データの情報を事前分布に集約するベイズ流の方法はハイブリッド対照デザインに自然に適していると考える研究者もいる（Zhang et al., 2023）．ベイズ流の方法は，既存対照と新規対照の差異に基づいて，借用の程度を柔軟に調整可能である（Ghadessi et al., 2020）．事前分布の構成方法に工夫の余地があり，また事後分布に基づいて推論を行うことから柔軟な試験デザインが実現可能となるため，ベイズ流の方法が多数提案されていると思われる.

しかしながら，ベイズ流の方法には課題も多い．ベイズ流の方法では，事前分布の構成方法が直感的ではなく，既存対照データがどの程度解析に寄与したかも分かりにくい（Bennett et al., 2021）．有効サンプルサイズを推定する方法がいくつか提案されているものの（Morita et al., 2008; Hobbs et al., 2013; Neuenschwander et al., 2020; Wiesenfarth

and Calderazzo, 2020)，計算方法によって異なる値になる（Bennett et al., 2022）．またベイズ流の方法は，計算コストが高い点に加えて，高度な解析スキルが要求される（Lin et al., 2023）．

一方で，頻度流の方法は柔軟性に欠けるとの批判があるものの（Berry, 2006），裏を返せば試験デザインがシンプルになりやすく，統計家ではない人達にとっても結果の解釈は容易であると考えられる．例えば，検定併合法であれば，既存対照と新規対照が似ていればすべての既存対照データを使う，似ていなければすべて使わないという all or nothing アプローチであり，併合される理由/されない理由が明確である．したがって，臨床家にとっても理解がしやすいと言われている（Li et al., 2020）．近年，規制当局は，ベイズ流の方法に基づく臨床試験を受け入れる姿勢を見せているものの，伝統的な方法である頻度論に基づく方法の方が，依然として受け入れられやすいと思われる．本研究では，数ある併合法の中でも，検定併合法に焦点を当てて研究を行った．

1.4 検定併合法を用いた事例

検定併合法を用いた事例として，PAN-01 試験（Yamaue et al., 2017）を紹介する．この試験では先行研究である GEST 試験（Ueno et al., 2013）のデータを既存対照として新規対照との併合が計画されていた．

まず，既存試験に当たる GEST 試験を紹介する．GEST 試験は，局所進行性および転移性膵臓癌の患者を対象に，点滴静注であるゲムシタビン単独療法に対する経口薬である S-1 単独療法の非劣性，ゲムシタビン＋S-1 併用療法の優越性を検証した多施設共同無作為化化第 3 相試験である．登録期間は 2007 年 7 月から 2009 年 10 月であった．832 人の患者が試験に参加し，ゲムシタビン群に 277 人，S-1 群に 280 人，ゲムシタビン+S-1 群に 275 人割り付けられた．主要評価項目は全生存であった．結果として，ゲムシタビン+S-1 の優越性は示されなかったものの，S-1 の非劣性が示され，ゲムシタビンに代わる第一選択薬としての S-1 の有用性が示された．

次に新規試験に当たる PAN-01 試験の説明をする．PAN-01 試験は，切除不能進行膵臓癌の患者を対象に，標準投与法である S-1 毎日服用に対する S-1 隔日服用の非劣性を検証する多施設共同，無作為化，非盲検の第 2 相試験である．登録期間は 2012 年 8 月から 2013 年 8 月であった．S-1 毎日服用群は，GEST 試験における S-1 群に相当する．研究グループ，参加施設，対象患者がほぼ同じであるため，GEST 試験の S-1 群と S-1 毎日服用群の併合が計画された．併合方法は検定併合法が用いられ，全生存期間と無増悪生存期間のハザード比がともに 0.87 より大きく 1.15 より小さければ（かつ比例ハザード性が成り立つという帰無仮説が両側有意水準 5%で棄却されなければ）2 つのデータを併合して解析し，そうでなければ GEST 試験のデータは用いないとされていた．検定併合法は，統計学的検定の結果に基づいて併合可否を判断する

方法であり，推定値に基づいて併合可否を判断する方法ではない．しかし，信頼区間と検定が表裏一体の関係にあることを踏まえると，PAN-01 試験の併合判定基準も検定併合法であるといってよいであろう．サンプルサイズは，毎日服用群で 60 人，隔日服用群で 120 人であった．この値は，GEST 試験のデータが併合された場合に検出力が 80%以上となる値として設定された．

試験の結果として，S-1 毎日服用群と GEST 試験の S-1 群のハザード比は，全生存期間で 0.86，無増悪生存期間で 0.66 であり，基準を満たさなかったため，GEST 試験のデータを併合した解析は行われなかった．また，S-1 毎日服用に対する S-1 隔日服用の非劣性も示されなかった．PAN-01 試験の Kaplan-Meier 曲線（Kaplan and Meier, 1958）を図 1.2 に示す．

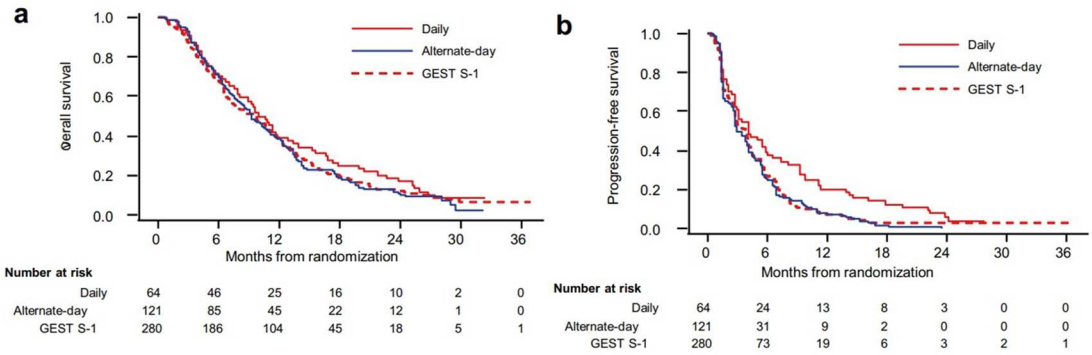

図 1.2　PAN-01 試験と GEST 試験の S-1 群の Kaplan-Meier 曲線 (a 全生存期間，b 無増悪生存期間)

図 1.2 から明らかなように，アウトカム分布は PAN-01 試験の S-1 毎日服用群と GEST 試験の S-1 群に違いがあり，前者の方が成績が良い．そのため，検定併合法による併合基準を満たすことができなかった．

このようなアウトカム分布の違いがどうして生じてしまったのだろうか．考えられる理由の 1 つに，PAN-01 試験と GEST 試験で組み入れられた患者集団に違いがあった可能性がある．GEST 試験は S-1 の第三相試験である一方，PAN-01 試験は S-1 承認後の投与法に関する後続研究であった．したがって，PAN-01 試験の方がより予後の良いと見込まれる，重症度の低い患者も研究に参加しやすかったかもしれない．しかし，Yamaue らが報告した患者背景の要約では，両試験に大きな違いは見られていない．そのほかに考えられる理由として，PAN-01 試験と GEST 試験の参加施設に違いがあった点も注目すべきであろう．この点については Yamaue らも報告の限界で触れている．GEST 試験では，日本と台湾の 75 施設が研究に参加した一方で，PAN-01 試験では日本の 37 施設であった．PAN-01 試験は後続研究であり，参加施設が医

療水準の高い施設に限定された可能性がある．また，PAN-01 試験では台湾の施設が参加していなかった点も，アウトカム分布に差が生じてしまった原因かもしれない．

PAN-01 試験は，完全ではないものの，Pocock の条件の大部分を満たしていた．治療内容とその評価方法が同じであるため，条件 1 と条件 3 は満たされている．両試験で適格基準は同じであり，実施時期も，登録時期としておよそ 5 年間の期間があるものの，この期間で医療・看護技術が劇的に進歩するとは考えづらく，条件 2 も満たされてるとしてよいであろう．測定された因子の要約値では，両試験に大きな違いはなかったことから，条件 4 も満たされている．研究参加施設は，全く同じではなかったが，GEST 試験の参加施設の一部が PAN-01 試験の参加施設となっている点で，条件 5 は部分的に満たされている．しかしながら，アウトカム分布には大きな違いがみられた．もし，動的な借用法である検定併合法を用いずに，単純に併合していた場合，第一種の過誤確率が生じていたかもしれない．PAN-01 試験の事例は，動的な借用法の必要性を示す良い例であったといえる．

1.5 検定併合法の課題

検定併合法にはいくつか課題が残されている．実際の運用において特に問題となる点は，検定併合法の有意水準であるγの設定方法である．γは併合のされやすさを制御するチューニングパラメータとして働く．γの決め方について定まった方法はなく，何人かの研究者が選択肢を提示しているに過ぎない．Viele らは，γの候補として 0.2，0.1，0.05，0.01 を提示している（Viele et al., 2014）．Liu は十分な併合確率を確保するために 0.05 か 0.1 を推奨し（Liu, 2018），Li らはシミュレーション実験において，第一種の過誤確率の上昇を抑制するために，0.2 を用いた（Li et al., 2020）．しかし，これらの選択には，個々の試験の動作特性を考慮していないという問題がある．臨床試験の動作特性は， チューニングパラメータだけではなく，サンプルサイズ，標準偏差，治療効果の大きさにも依存する．個々の臨床試験ごとに，第一種の過誤確率や検出力等の動作特性を考慮して，チューニングパラメータを決定することが望ましいと考える．

検定併合法は動作特性の制御に関して柔軟性が乏しいという点を指摘したい．検定併合法はチューニングパラメータを 1 つしか持たず，この値が，第一種の過誤確率と検出力を同時に規定する．1.2 節で説明した通り，drift の正負に応じて上昇する過誤確率の種類が異なる．検定併合法のγは，$\mathrm{drift} > 0$の場合の第一種の過誤確率の上昇の制御と，$\mathrm{drift} < 0$の場合の第二種の過誤確率の上昇の制御をどちらも担当している．Drift の正負に応じて借用のされやすさを別々に設定することで，同じ第一種の過誤確率の下で，検出力の向上が可能となるかもしれない．なお，この指摘は既存の借用法すべてに共通するものである．

検定併合法は，アウトカムの分布のみに着目して，データ間の分布の乖離を評価す

る．これは，既存対照と新規対照の患者背景の分布は同じと仮定し，アウトカム分布の乖離は，時代効果によって引き起こされるという立場をとっていると解釈できる．しかし，臨床試験において，既存対照の患者背景分布が，新規対照の患者背景分布と同一になるとは限らない．アウトカムに影響を与える因子の分布がデータ間で異なるために，見かけ上アウトカム分布に違いが生じている場合も考えられる．患者背景の違いを適切に調整することで，バイアスのない治療効果の推定が期待できる．

そのほかの検定併合法の批判として，Li らは，検定併合法は既存対照と新規対照の差異を判定するために優越性検定を採用しているため，サンプルサイズが小さい場合には検定が有意になりづらく，誤って既存対照を併合してしまう恐れがある点と，既存対照と新規対照の従う分布が同一であったとしても，第一種の過誤確率が名目水準に一致しない点を指摘した（Li et al., 2020）．そして，この点を解消するため，優越性検定ではなく同等性検定に基づいた，同等性検定ベースの検定併合法を提案した．また，第一種の過誤確率を名目水準に一致させるため，治療効果に関する仮説に対する有意水準を調整する方法を提案した.

1.6 本研究の目的

本研究では，1.5 節で述べた課題を解決するために，検定併合法を 2 つの方向に拡張する．第二章では，第一種の過誤確率と検出力の制御に柔軟性を与えるために，検定併合法の両側帰無仮説を 2 つの片側帰無仮説に分割し，それぞれの仮説に対して別々の有意水準を設定する検定併合法を提案する．また，チューニングパラメータを，最高第一種の過誤確率と最低検出力に基づいて選択する方法を提案する．そして，提案するチューニングパラメータ選択法を適用した場合の提案法の動作特性を，数値実験を通して評価する．第四章では，患者背景の調整のために，検定併合法に傾向スコア法を組み合わせた 2 段階法を提案し，数値実験によって動作特性を評価する．第五章で研究全体の考察を述べ，第六章に結論を記す.

本研究では，目的の異なる 2 種類の検定とそれらの有意水準が登場することに注意されたい．1 つは，治療効果の有無を判断するための検定であり，もう 1 つは，既存対照の併合可否を判断するための, 検定併合法の検定である. 前者を意図する場合は，必ず「治療効果に関する検定」や「治療効果に関する検定の有意水準」のように明示するが，後者を意図する場合は単に「有意水準」と表記する場合がある．また，それぞれの検定に対して，第一種の過誤確率と検出力を考えることができるが，本研究では，治療効果に関する検定に対してのみ第一種の過誤確率と検出力を考え，検定併合法の検定に対しては一切考えない.

第二章

2 つの片側検定を組み合わせた検定併合法

2.1 緒言

ハイブリッド対照デザインは，drift に応じて第一種の過誤確率と検出力が変化するという性質がある．そのため，第一種の過誤確率と検出力は，必ずしも名目有意水準・名目検出力通りとはならない．すなわち，名目有意水準として片側 5%を設定した場合であっても，ある drift では，実質的な第一種の過誤確率が 5%を超えてしまう．考えられるすべての drift について第一種の過誤確率を名目有意水準に制御すること（すなわち“厳密な”第一種の過誤確率の制御）は，ハイブリッド対照デザインにおいては不可能であることが理論的に示されている（Kopp-Schneider et al., 2020）．そのため，ハイブリッド対照デザインの方法論の目的は，第一種の過誤確率の上昇を抑えつつ，検出力を高くすることである．

1.5 節では，検定併合法が抱える課題の 1 つとして，第一種の過誤確率と検出力の制御について柔軟性が乏しい点を指摘した．これは，検定併合の動作特性に関するチューニングパラメータが検定併合法の有意水準γのみであることに由来する．治療効果の判断に用いる検定の第一種の過誤確率に関して，ある水準となるようγを決めることは可能である．ただし，同時に検出力が自動的に決まってしまう．一方，検出力がある水準を満たすようにγを決めると， 第一種の過誤確率も自動的に決まってしまう．もし，第一種の過誤確率と検出力の制御を分離することができれば，臨床試験の計画における自由度が向上し，検定併合法がより使いやすい方法になると考えた．本章では，従来の検定併合法の両側仮説を，2 つの片側仮説に分割し，それぞれに異なる片側有意水準を与えることで，第一種の過誤確率と検出力の制御に柔軟性を持たせる方法を提案する．

検定併合法が抱えるもう 1 つの課題として，γの選択方法があげられる．既存の研究では 0.01 から 0.2 が検討されていたが，これらの値を用いる根拠はない．臨床試験において検定併合法を用いる場合， γの選択根拠を理論的に説明できる必要があると考える．そこで，本章では，drift に対する第一種の過誤確率の最高値と検出力の最低値に基づいて検定併合法の有意水準を選択する方法を提案する．

本章の構成は以下の通りである．2.2 節で従来の検定併合法を紹介する．2.3 節で提案法である 2 つの片側検定を組み合わせた検定併合法を提案し，提案法が第一種の過誤確率と検出力の制御について柔軟性を持つことを示す．2.4 節で検定併合法の有意水準を, 目標としたい第一種の過誤確率と検出力に基づいて選択する方法を提案する．

2.5 節で仮想的な臨床試験を想定した数値実験を通して，提案法の有用性を示す．2.6 節で提案法の特性の理解を深めるために，うつ病の臨床試験のデータに提案法を適用する．2.7 節で考察を記す．

2.2 従来の検定併合法

本章では，連続量アウトカムを想定し，新規治療の優越性の証明を目的とする，次の形で表される仮説を考える．

$$H_0: \mu_T - \mu_{CC} > 0 \text{ versus } H_1: \mu_T - \mu_{CC} \leq 0. \tag{2.1}$$

ここで，Tは試験治療（treatment），CCは新規対照（current control）であり，μ_Tは新規治療の母平均，μ_{CC}は新規対照の母平均を表す．仮説(2.1)の片側有意水準をαと表す．

従来の検定併合法（conventional test-then-pool method; CTTP）は，既存対照と新規対照の分布の同一性を，検定を用いて判定する．具体的に，既存対照データと新規対照データの乖離に関する次の仮説を，有意水準γで検定する．

$$H_{0,CTTP}: \mu_{HC} = \mu_{CC} \text{ versus } H_{1,CTTP}: \mu_{HC} \neq \mu_{CC}. \tag{2.2}$$

ここで，HCは既存対照（historical control），μ_{HC}は既存対照の母平均を表す．帰無仮説$H_{0,CTTP}$が棄却されなかった場合，既存対照データを新規対照データに単純併合する方法（単純併合法，pooling）で解析する．$H_{0,CTTP}$が棄却された場合，新規試験データのみ用いる方法（分離法，no polling）で解析する．仮説(2.2)の有意水準γは，既存対照の併合確率のチューニングパラメータとして解釈可能である．γが高いと，$H_{0,C\,TTP}$が棄却されやすくなり，分離法が選択されやすくなる．逆に，γが低いと，$H_{0,CTTP}$が棄却されにくくなり，単純併合法が選択されやすくなる．

アウトカムが分散既知の連続分布に従う場合について，数式を用いて CTTP の具体的な手順を説明する．$\bar{X}_g$ $(g \in (T, CC, HC))$を群gの標本平均を表す確率変数とする．確率変数の実現値は小文字を用いて表す．例えば，$\bar{X}_g$の実現値は$\bar{x}_g$と表す．十分大きなサンプルサイズの下で，中心極限定理より，$\bar{X}_g$は平均μ_g，分散σ_g^2/n_gの正規分布に従う．ここで，σ_g^2は，それぞれ群gのデータが従う分布の母分散を表し，n_gはサンプルサイズを表す．仮説(2.2)に対する検定統計量Uは，

$$U = \frac{\bar{X}_{HC} - \bar{X}_{CC}}{\sigma} \tag{2.3}$$

と表される．ここで，σは$\bar{X}_{HC} - \bar{X}_{CC}$の標準偏差を表し，

$$\sigma = \sqrt{\frac{\sigma_{HC}^2}{n_{HC}} + \frac{\sigma_{CC}^2}{n_{CC}}}$$

である．Uの実現値uの値に応じて，次のように意思決定を行う．

$$\begin{cases} |u| > z_{1-\gamma/2} \rightarrow \text{分離法で解析} \\ |u| \leq z_{1-\gamma/2} \rightarrow \text{単純併合法で解析} \end{cases} \tag{2.4}$$

ここで，z_{1-x}は標準正規分布の上側 $100x$%点を表す．仮説(1.1)に対する検定統計量Tは，分離法の検定統計量をT_s，単純併合法の検定統計量T_pとして，

$$T = \begin{cases} T_s = \dfrac{\bar{X}_T - \bar{X}_{CC}}{\sigma_s}, \text{if not pooling} \\ T_p = \dfrac{\bar{X}_T - \dfrac{n_{CC}\bar{X}_{CC} + n_{HC}\bar{X}_{HC}}{n_{CC} + n_{HC}}}{\sigma_p}, \text{if pooling} \end{cases} \tag{2.5}$$

と表される．ここで,

$$\sigma_s = \sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_{CC}^2}{n_{CC}}}, \sigma_p = \sqrt{\frac{\sigma_T^2}{n_T} + \frac{n_{CC}\sigma_{CC}^2 + n_{HC}\sigma_{HC}^2}{(n_{CC} + n_{HC})^2}}$$

である．仮説(2.1)に対する検定を行うため，Tの実現値tと$z_{1-\alpha}$の大小を比較する．すなわち,

$$\begin{cases} t < -z_{1-\alpha} \rightarrow \text{帰無仮説} H_0 \text{を棄却する} \\ t \geq -z_{1-\alpha} \rightarrow \text{帰無仮説} H_0 \text{を棄却しない} \end{cases} \tag{2.6}$$

という意思決定を行う．

実践的には，分散は未知であるため，σ_T^2，σ_{CC}^2，σ_{HC}^2には推定値を当てはめる．この時，σ，σ_s，σ_pについて，等分散性を仮定し，プールした分散を用いてもよい．また，棄却限界値は標準正規分布ではなく適切な自由度の t 分布から与えられる．具体的に，単純併合法が選択された場合の自由度は$n_T + n_{CC} + n_{HC} - 2$，分離法が選択された場合の自由度は$n_T + n_{CC} - 2$である．

2.2.1 従来の検定併合法の動作特性

臨床試験を行う際，試験開始前に試験の動作特性を評価しておく必要がある．特に臨床試験では第一種の過誤確率と検出力が重要視される．また，CTTP 特有の指標として，既存試験データの併合確率も事前に評価しておきたい．ここでは，分散既知の連続アウトカムに対して CTTP を適用した場合について, 検出力, 第一種の過誤確率, 併合確率の数式表現を示す.

まず, 併合確率を示す. 意思決定基準(2.6)より, 試験計画段階での併合確率P_{CTTP}は,

$$P_{CTTP} = P(|U| \leq z_{1-\gamma/2} | \bar{X}_{HC} = \bar{x}_{HC})$$

と表される．$\bar{X}_{HC} = \bar{x}_{HC}$で条件付けている理由は後述する．$\bar{X}_{HC} = \bar{x}_{HC}$のもとで,

$$U|\bar{X}_{HC} = \bar{x}_{HC} \sim N\left(\frac{\bar{x}_{HC} - \mu_{CC}}{\sigma}, \frac{1}{\sigma^2} \cdot \frac{\sigma_{CC}^2}{n_{CC}}\right)$$

となる．したがって，

$$P_{CTTP} = \Phi\left((\sigma z_{1-\gamma/2} - \bar{x}_{HC} + \mu_{CC})/\sqrt{\sigma_{CC}^2/n_{CC}}\right) - \Phi\left((-\sigma z_{1-\gamma/2} - \bar{x}_{HC} + \mu_{CC})/\sqrt{\sigma_{CC}^2/n_{CC}}\right)$$

と表せる．ここで，Φは標準正規分布の累積分布関数を表す．

次に，第一種の過誤確率と検出力を示す．意思決定基準(2.6)より，帰無仮説H_0が棄却される確率をr_{CTTP}とすると，

$$\begin{aligned} r_{CTTP} &= P(T_s \le -z_{1-\alpha}, \text{no pooling} | \bar{X}_{HC} = \bar{x}_{HC}) + P(T_p \le -z_{1-\alpha}, \text{pooling} | \bar{X}_{HC} = \bar{x}_{HC}) \\ &= P\left(T_s < -z_{1-\alpha}, (U < -z_{1-\gamma/2} \cup z_{1-\gamma/2} < U) | \bar{X}_{HC} = \bar{x}_{HC}\right) \\ &\quad + P(T_p < -z_{1-\alpha}, -z_{1-\gamma/2} < U < z_{1-\gamma/2} | \bar{X}_{HC} = \bar{x}_{HC}) \end{aligned}$$

と表される．$\bar{X}_{HC} = \bar{x}_{HC}$で条件付けたもとで，$(T_s, U)$，$(T_p, U)$の同時分布はそれぞれ次に示すに二変量正規分布になる．

$$\begin{pmatrix} T_S \\ U \end{pmatrix} | \bar{X}_{HC} = \bar{x}_{HC} \sim N\left(\begin{pmatrix} \frac{\mu_T - \mu_{CC}}{\sigma_s} \\ \frac{\bar{x}_{HC} - \mu_{CC}}{\sigma} \end{pmatrix}, \begin{pmatrix} 1 & \frac{1}{\sigma\sigma_s} \cdot \frac{\sigma_{CC}^2}{n_{CC}} \\ \frac{1}{\sigma\sigma_s} \cdot \frac{\sigma_{CC}^2}{n_{CC}} & \frac{1}{\sigma^2} \cdot \frac{\sigma_{CC}^2}{n_{CC}} \end{pmatrix}\right),$$

$$\begin{pmatrix} T_P \\ U \end{pmatrix} | \bar{X}_{HC} = \bar{x}_{HC} \sim$$

$$N\left(\begin{pmatrix} \frac{\mu_T - \frac{n_{CC}\mu_{CC} + n_{HC}\bar{x}_{HC}}{n_{CC} + n_{HC}}}{\sigma_p} \\ \frac{\bar{x}_{HC} - \mu_{CC}}{\sigma} \end{pmatrix}, \begin{pmatrix} \frac{1}{\sigma_p^2} \cdot \left(\frac{\sigma_T^2}{n_T} + \frac{n_{CC}\sigma_{CC}^2}{(n_{CC} + n_{HC})^2}\right) & \frac{1}{\sigma\sigma_p} \cdot \frac{n_{CC}\sigma_{CC}^2}{n_{CC} + n_{HC}} \\ \frac{1}{\sigma\sigma_p} \cdot \frac{n_{CC}\sigma_{CC}^2}{n_{CC} + n_{HC}} & \frac{1}{\sigma^2} \cdot \frac{\sigma_{CC}^2}{n_{CC}} \end{pmatrix}\right). \quad (2.7)$$

以上より，r_{CTTP}は，(T_S, U)，(T_P, U)の同時分布を適当な領域で積分したものの和として計算できる．具体的に，(T_s, U)の密度関数を$f_s(t_s, u)$，(T_p, U)の密度関数を$f_p(t_p, u)$として，

$$\begin{aligned} r_{CTTP} &= \int_{-\infty}^{-z_{1-\gamma/2}} \int_{-\infty}^{-z_{1-\alpha}} f_s(t_s, u)\, dt_s du \\ &\quad + \int_{-z_{1-\gamma/2}}^{z_{1-\gamma/2}} \int_{-\infty}^{-z_{1-\alpha}} f_p(t_p, u)\, dt_p du \\ &\quad + \int_{z_{1-\gamma/2}}^{\infty} \int_{-\infty}^{-z_{1-\alpha}} f_s(t_s, u)\, dt_s du \end{aligned} \quad (2.8)$$

と計算される．上式の右辺第 1 項と第 3 項が分離法の棄却確率，第 2 項が単純併合法

の棄却確率である．治療効果がない，すなわち$\mu_T - \mu_{CC} = 0$のとき，r_{CTTP}は第一種の過誤確率となる．また，$\mu_T - \mu_{CC} > 0$のとき，r_{CTTP}は検出力である．

新規試験の計画段階において，通常，既存対照データを考慮して動作特性を評価する（Lim et al., 2020）．このことは，数式上，$\bar{X}_{HC} = \bar{x}_{HC}$で条件付けることに対応する．$\bar{X}_{HC} = \bar{x}_{HC}$で条件付けなければ，既存対照データを考慮した動作特性を正確に評価できない．例えば，CTTP では，両側有意水準γで検定を行うため，帰無仮説が成り立つ場合（すなわち$\mu_{CC} = \mu_{HC}$）であれば，既存対照が併合される確率が$1 - \gamma$となると思われる．しかし，これは$\bar{X}_{HC} = \bar{x}_{HC}$で条件付けていない値であり，実際には$1 - \gamma$と一致しない．なぜならば，$\bar{x}_H$は既に観察されていて，変動することはないためである．同様の問題は第一種の過誤確率や検出力でも生じる．動作特性を数式で正確に評価するためには，$\bar{X}_{HC} = \bar{x}_{HC}$で条件付ける必要がある．

2.3　2 つの片側検定を組み合わせた検定併合法

1.2 節では，図 1.1 を用いて，(i) $\mu_{CC} < \bar{x}_{HC}$のとき，第一種の過誤確率が上昇し，(ii) $\mu_{CC} > \bar{x}_{HC}$のとき，検出力が減少すると説明した．式(2.7)から次のようにも説明できる．(i)のとき，治療効果が 0（$\mu_T - \mu_{CC} = 0$）であっても，式(2.7)におけるT_Pの平均値が負になる．その結果，式(2.8)の第 2 項が上昇するため，第一種の過誤確率が上昇する．このことは，図 1.1 において，右側の点線の分布が左方向にシフトすることに対応する．一方で，(ii)のとき，式(2.7)におけるT_Pの平均値が正になる．その結果，式(2.8)の第 2 項が減少するため，検出力が減少する．このことは，図 1.1 において，左側の点線の分布が右方向にシフトすることに対応する．

実践的には，μ_{CC}は未知であるため，その推定値である$\bar{x}_{CC}$を評価する．CTTP の併合基準は，仮説(2.2)で表される両側検定である．したがって，併合されるかどうかの判定は，$\bar{x}_{CC}$と$\bar{x}_{HC}$の正負には依存せず，差の絶対値の大きさに依存する．ゆえに，γを決めると，第一種の過誤確率と検出力が同時に決定される．このことは，CTTP が，第一種の過誤確率と検出力の制御の柔軟性が乏しいことを意味する．詳細は 2.3.2 項で説明する．$\bar{x}_{CC}$と$\bar{x}_{HC}$の正負に応じて，併合基準を変えることで，第一種の過誤確率の制御と検出力の制御を別々に行うことができると考えた．

本節では，仮説(2.2)で表される両側仮説を次の 2 つの片側仮説に置き換える，2 つの片側検定を組み合わせた検定併合法（two one-sided test-then-pool method; TTTP）を提案する．

$$H_{0,TTTP,1}: \mu_{HC} \leq \mu_{CC} \text{ vs. } H_{1,TTTP,1}: \mu_{HC} > \mu_{CC}, \tag{2.9}$$

$$H_{0,TTTP,2}: \mu_{HC} \geq \mu_{CC} \text{ vs. } H_{1,TTTP,2}: \mu_{HC} < \mu_{CC}. \tag{2.10}$$

$H_{0,TTTP.L}$と$H_{0,TTTP.R}$の有意水準をそれぞれγ_1とγ_2とする．TTTP では，2 つの帰無仮説がどちらも棄却されなかった場合に，既存対照を併合する．γ_1とγ_2の範囲はともに 0

以上 0.5 以下とする．0.5 以上の値を許容すると，$\bar{x}_{CC} = \bar{x}_{HC}$であっても併合されない場合が生じてしまい望ましくない. γ_1とγ_2がともに 0 である場合, $H_{0,TTTP.L}$と$H_{0,TTTP.R}$は必ず棄却されないため, 単純併合法と同一になる. γ_1とγ_2がともに 0.5 である場合, $H_{0,TTTP.L}$と$H_{0,TTTP.R}$は必ず棄却されるため，分離法と同一になる.

2.3.1 2 つの片側検定を組み合わせた検定併合法の動作特性

TTTP は，$\gamma_1 = \gamma_2$とすると，CTTP と一致する．したがって，TTTP の動作特性は, CTTP と同様の枠組みで表すことができる. 仮説(2.9)と仮説(2.10)の検定統計量はCTTP と同様に式(2.3)で表されるUである．また, 治療効果に関する仮説である仮説(2.1)の検定統計量は, CTTP と同様に式(2.5)で表されるTである. 式(2.4)の CTTP の併合基準は, TTTP では次のように変更される.

$$\begin{cases} \left(u < -z_{1-\gamma_2}\right) \cup \left(z_{1-\gamma_1} < u\right) \rightarrow \text{分離法で解析} \\ -z_{1-\gamma_2} < u < z_{1-\gamma_1} \rightarrow \text{単純併合法で解析} \end{cases} \tag{2.11}$$

式(2.11)と式(2.4)を比べることで，TTTP は，$\gamma/2 = \gamma_1 = \gamma_2$とした CTTP と一致することが分かる．式(2.11)より，既存対照の併合確率P_{TTTP}は，P_{CTTP}と同様に考えると，

$$P(-z_{1-\gamma_2} < u < z_{1-\gamma_1} | \bar{X}_{HC} = \bar{x}_{HC})$$
$$= \Phi\left((\sigma z_{1-\gamma_1} - \bar{x}_{HC} + \mu_{CC}) / \sqrt{\sigma_{CC}^2 / n_{CC}}\right) - \Phi\left((-\sigma z_{1-\gamma_2} - \bar{x}_{HC} + \mu_{CC}) / \sqrt{\sigma_{CC}^2 / n_{CC}}\right)$$

となる．また，帰無仮説を棄却する確率r_{TTTP}は，r_{CTTP}と同様に考えると，

$$\begin{aligned} &P(T_s \le -z_{1-\alpha}, \text{no pooling} | \bar{X}_{HC} = \bar{x}_{HC}) + P(T_p \le -z_{1-\alpha}, \text{pooling} | \bar{X}_{HC} = \bar{x}_{HC}) \\ &= P\left(T_s < -z_{1-\alpha}, (U < -z_{1-\gamma_2} \cup z_{1-\gamma_1} < U) | \bar{X}_{HC} = \bar{x}_{HC}\right) \\ &+ P(T_p < -z_{1-\alpha}, -z_{1-\gamma_2} < U < z_{1-\gamma_1} | \bar{X}_{HC} = \bar{x}_{HC}) \\ &= \int_{-\infty}^{-z_{1-\gamma_2}} \int_{-\infty}^{-z_{1-\alpha}} f_s(t_s, u)\, dt_s du \\ &+ \int_{-z_{1-\gamma_2}}^{z_{1-\gamma_1}} \int_{-\infty}^{-z_{1-\alpha}} f_p(t_p, u)\, dt_p du \\ &+ \int_{z_{1-\gamma_1}}^{\infty} \int_{-\infty}^{-z_{1-\alpha}} f_s(t_s, u)\, dt_s du \end{aligned} \tag{2.12}$$

となる. 式(2.8)と同様に, r_{TTTP}は, $\mu_T - \mu_{CC} = 0$のとき第一種の過誤確率, $\mu_T - \mu_{CC} > 0$のとき検出力となる.

2.3.2 第一種の過誤確率と検出力の制御の柔軟性

CTTP の併合基準である式(2.6)は，信頼水準をγとした$\mu_{HC} - \mu_{CC}$の両側信頼区間が 0 を含むかどうかと言い換えられる. $\mu_{HC} - \mu_{CC}$の両側信頼区間は次のように表される.

$$\left[\bar{x}_{HC} - \bar{x}_{CC} - \sigma z_{1-\gamma/2}, \bar{x}_{HC} - \bar{x}_{CC} + \sigma z_{1-\gamma/2}\right]$$

この区間が，0 を含んでいた場合は既存対照を併合し，0 を含んでいない場合は既存対照を併合しない．一方で，TTTP の併合基準は，2 つの片側検定を組み合わせていることから，信頼水準をγ_1とした$\mu_{HC} - \mu_{CC}$の片側信頼区間が 0 を含み，かつ，信頼水準をγ_2とした$\mu_{HC} - \mu_{CC}$の片側信頼区間が 0 を含む，と言い換えられる．すなわち，次のように表される区間が 0 を含んでいるかどうかである．

$$[\bar{x}_{HC} - \bar{x}_{CC} - \sigma z_{1-\gamma_1}, \bar{x}_{HC} - \bar{x}_{CC} + \sigma z_{1-\gamma_2}]$$

本研究では，これらの区間のように，0 を含む場合に既存対照を併合する区間を，併合区間と呼ぶ．図 2.1 に TTTP と CTTP の併合区間の例を示す．併合区間を計算する際の数値は 2.6 節のうつ病データを利用した．

併合区間の式から，CTTP と TTTP について，最高第一種の過誤確率と最低検出力の制御に関する柔軟性の違いを説明できる．$\bar{x}_{CC} < \bar{x}_{HC}$のとき，第一種の過誤確率の上昇が懸念されるのであった．図 2.1 では Case 1 に相当する．このとき，併合有無は，併合区間の下限に依存する．CTTP ではγが，TTTP ではγ_1が併合区間の下限を規定する．したがって，TTTP では，主にγ_1が $\mu_{CC} > \bar{x}_{HC}$における併合有無を制御している．一方で，$\bar{x}_{CC} > \bar{x}_{HC}$のとき，検出力の減少が懸念される．図 2.1 では Case 2 に相当する．このとき，併合有無は，併合区間の上限に依存する．CTTP ではγが，TTTP ではγ_2が併合区間の上限を規定する．したがって，TTTP では，主にγ_2が $\mu_{CC} > \bar{x}_{HC}$における併合有無を制御している．CTTP では，下限と上限の両方を 1 つのパラメータγが規定する一方で，TTTP では，下限と上限をそれぞれγ_1とγ_2が
別々に担当するため，最高第一種の過誤確率と最低検出力の制御をより柔軟に行うことができる．

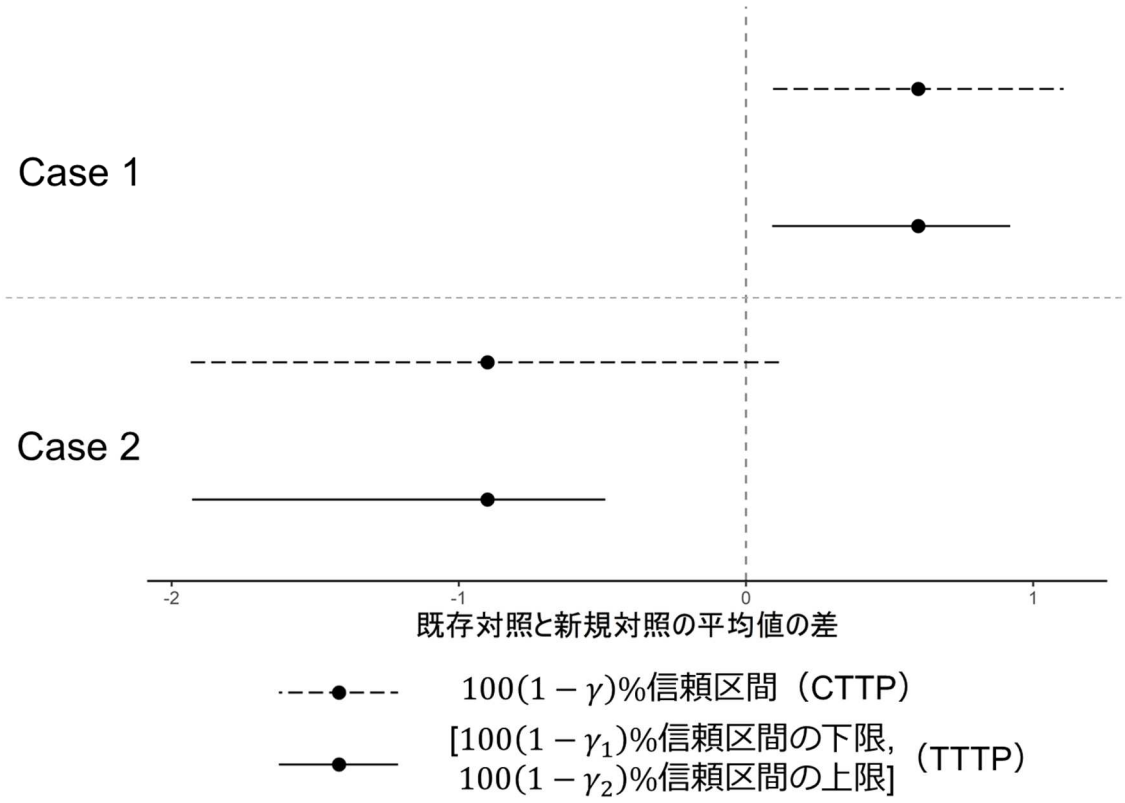

図 2.1　TTTP と CTTP の併合区間.
TTTP：2 つの片側検定を組み合わせた検定併合法
CTTP：従来の検定併合法
γ_1，γ_2：TTTP の有意水準
γ：CTTP の有意水準

2.4　第一種の過誤確率と検出力に基づく有意水準選択法

従来の検定併合法（CTTP）の批判の 1 つは，仮説(2.2)の有意水準γに合意の得られた値はないという点であった．0.1 や 0.2 が用いられることが多いが，その根拠は乏しい．提案する検定併合法（TTTP）は，有意水準がγ_1とγ_2の 2 つとなるため，これらの定め方はより難しくなる．何らかの基準を設けて，これらの有意水準（γ，γ_1とγ_2）の決める必要がある．

本節では，最高第一種の過誤確率と最低検出力を適当な値に制御するように有意水準を求める方法を提案する．既存対照データを新規試験の解析に利用する際の統計的な問題点は，1.2 節で説明したように，drift が 0 ではない場合の第一種の過誤確率の上昇と検出力の低下であった．そこで，drift を$\bar{x}_{HC} - \mu_{CC}$と定義し，第一種の過誤確率と検出力を drift の関数として考え，第一種の過誤確率の最高値がある許容限界以下に，検出力の最低値がある許容限界以上になるように有意水準を決める．TTTP の場合で

説明するが，$\gamma_1 = \gamma_2 = \gamma/2$という制約を加えることで，CTTP のγを決定することもできる．

既存対照の併合を考慮しない名目の検出力を$1-\beta$，治療効果に関する仮説である仮説(2.1)の有意水準をαと表す．第一種の過誤確率の許容上限を$\delta_e\ (\geq \alpha)$，検出力の許容下限を$\delta_p\ (\leq 1-\beta)$とする．式(2.12)で表される第一種の過誤確率と検出力の関数を，改めて$TIE(\mu_{CC}, \gamma_1, \gamma_2)$，$Pw(\mu_{CC}, \gamma_1, \gamma_2)$と表す．第一種の過誤確率と検出力は，これら 3 つの値以外に，サンプルサイズ，標準偏差，α，検出する治療効果の大きさにも依存するが，表記の単純化のためここで関係のある 3 つのみ表記する．既存対照で条件付けた第一種の過誤確率の最高値と検出力の最低値はそれぞれ次のように表現できる．

$$T1E_{\max}(\gamma_1, \gamma_2) = \max_{\mu_{CC}}\ T1E(\mu_{CC}, \gamma_1, \gamma_2)$$

$$Pw_{\min}(\gamma_1, \gamma_2) = \min_{\mu_{CC}}\ Pw(\mu_{CC}, \gamma_1, \gamma_2)$$

また，drift が 0 である場合の検出力を

$$Pw_0(\gamma_1, \gamma_2) = Pw(\mu_{CC} = \bar{x}_{HC}, \gamma_1, \gamma_2)$$

と表記する．

次の不等式を満たす(γ_1, γ_2)を求めたい．

$$T1E_{\max}(\gamma_1, \gamma_2) \leq \delta_e\ \cap\ Pw_{\min}(\gamma_1, \gamma_2) \geq \delta_p$$

この不等式を満たすγ_1とγ_2の組は無数に存在する．例えば，γ_1とγ_2をともに 0.5 とすると，既存対照データは必ず併合されないため，$T1E_{\max}(\gamma_1, \gamma_2) = \alpha < \delta_e$，$Pw_{\min}(\gamma_1, \gamma_2) = 1-\beta > \delta_p$となり，不等式は満たされる．不等式を満たす$\gamma_1$と$\gamma_2$を 1 組に定めるためには，もう 1 つ条件が必要となる．

既存対照データを利用する際は，新規試験と既存対照の母集団が同一であると仮定する．この仮定が正しいときの検出力である$Pw_0(\gamma_1, \gamma_2)$を最高にするようにγ_1とγ_2を定めると良いと考えた．これは，次の制限付き最大値問題を解くことに対応する．

$$\underset{\gamma_1, \gamma_2}{\operatorname{argmax}}\ Pw_0(\gamma_1, \gamma_2)$$

$$\text{s. t.}$$

$$T1E_{\max}(\gamma_1, \gamma_2) \leq \delta_e\ \cap\ Pw_{\min}(\gamma_1, \gamma_2) \geq \delta_p$$

この問題は，統計ソフトウェア R の Rsolnp パッケージ（Ghalanos and Theussl, 2015; Ye, 1987）を用いて解くことができる．γ_1とγ_2を求める際，初期値を指定する必要がある．臨床試験の設定によっては，初期値の設定に応じて異なるγ_1とγ_2が得られる場合がある．複数の初期値を用いて求めたγ_1とγ_2の組のうち，最も$Pw_0(\gamma_1, \gamma_2)$を最高にする組を選択するとよい．

上記の最適化を用いたγ_1とγ_2の選択は，検定統計量が（近似的に）正規分布に従い，

既存対照で条件付けた第一種の過誤確率と検出力を，数式を用いて表現できる場合に適用できる．例えば，連続量アウトカムや，サンプルサイズが十分に大きい場合の 2 値アウトカムが該当する．生存時間アウトカムの場合，既存対照で条件付けた第一種の過誤確率と検出力の計算が難しい．したがって，最適化ではなく，Monte Carlo シミュレーションとグリッドサーチを用いて最適なγ_1とγ_2の組を探索する必要がある．

2.5 数値実験

2.4 節で紹介した第一種の過誤確率と検出力に基づく有意水準選択法を用いた場合の CTTP と TTTP の動作特性（第一種の過誤確率，検出力，既存対照の併合確率）を評価するために数値実験を行った．

2.5.1 設定

治療効果に関する仮説が仮説(2.1)で表される無作為化試験の仮想的な計画を通して，既存対照が併合される確率，第一種の過誤確率，検出力を評価した．新規試験の試験治療群のサンプルサイズ（n_T）は，50，100，200，400 を検討した．各n_Tに対して，新規試験の対照群のサンプルサイズ（n_{CC}）は，n_Tの 0.25 倍，0.5 倍，0.75 倍，1 倍（小数点切り上げ）を検討した．すなわち，4:1 割付，2:1 割付，4:3 割付，1:1 割付を検討した．標準偏差は 1 とし，既知として扱った．治療効果の指標は平均値の差$\theta = \mu_T - \mu_{CC}$を考える．対立仮説が真の場合の$\theta$の大きさは，各サンプルサイズの組 (n_T, n_{CC})に対して，既存対照を考慮しない検出力が 75%になるように設定した．帰無仮説が真の場合のθは 0 とした．既存対照データとして，標本平均$\bar{x}_{HC}$が 0，標準偏差が 1 の既存対照データが利用可能であるとした．既存対照のサンプルサイズn_{HC}は，10 から 500 まで 10 刻みで検討した．CTTP のγと TTTP のγ_1とγ_2は，2.4 節で示した 第一種の過誤確率と検出力に基づく有意水準選択法を用いて探索した．δ_eは 0.05 と設定し，δ_pは 0 また 0.75 とした．$\delta_p = 0$は，検出力について制約を設けないことを意味し，$\delta_p = 0.75$は検出力の低下を一切許容しないことを意味する．

上記の検討に加えて，第一種の過誤確率と検出力に基づく有意水準選択法を用いずに，慣例的な値である$\gamma = 0.2$を与える CTTP の動作特性も検討した．

2.5.2 結果

2.3.2.1 $\gamma = 0.2$とした CTTP の動作特性

まずは慣例に従いγに 0.2 を与えた場合の CTTP の動作特性を検討する．図 2.2 は，$\gamma = 0.2$とした，$n_T = 200$，$n_{CC} = 200, 100$，$n_{HC} = 50, 100, 200, 400$ の場合の CTTP の動作特性を示している．各動作特性は，新規対照の母平均μ_{CC}の関数として示されている．既存対照の標本平均$\bar{x}_{HC}$が 0 であることから，μ_{CC}が 0 のとき drift が 0 である．

Drift が 0 であるとき，第一種の過誤確率は 概ね名目水準である 2.5%であり，検出力は 75%を大きく超えているため，望ましい動作特性が得られている．しかし，μ_{CC}が負のとき, 第一種の過誤確率が名目水準から大きく上昇している箇所がある. 特に, 新規対照のサンプルサイズが小さい$n_C = 100$の場合に顕著である．図 2.2 に示したすべての設定で, 最高第一種の過誤確率が 5%を超えている. 一方で, μ_{CC}が正のときは, 第一種の過誤確率は上昇しないものの，検出力が名目水準の 75%から減少している. この結果は，CTTP において，動作特性を考慮せず，γに慣例的に適当な値を与えることの危険性を示唆している.

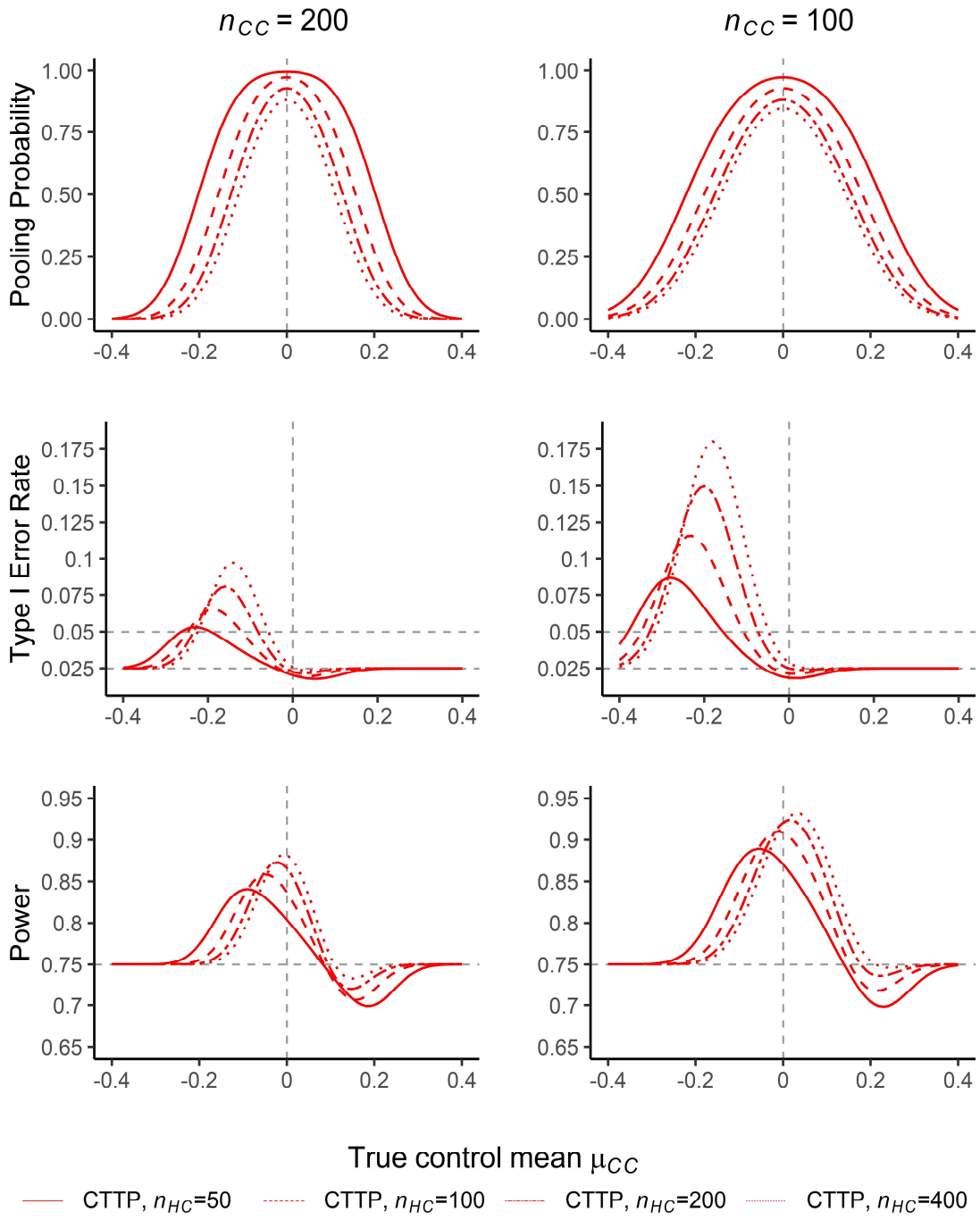

図 2.2 $\gamma = 0.2$とした CTTP の動作特性

CTTP：従来の検定併合法

δ_p：最低検出力の許容下限

n_T，n_{CC}，n_{HC}：試験治療，新規対照，既存対照のサンプルサイズ

μ_{CC}：新規対照の母平均

2.5.2.2　動作特性を考慮して有意水準を定めた検定併合法の動作特性

第一種の過誤確率と検出力を考慮して有意水準を定めた場合の結果を示す．表 2.1 と表 2.2 は，2.4 節の方法を用いて選択されたγ_1とγ_2，γである．

まずは, CTTP のγに注目する．表 2.1 と表 2.2 の違いは, $\delta_p = 0$であるか, $\delta_p = 0.75$であるかの違い，すなわち，検出力の最小値に関して条件を設定するかしないかの違いである．$\delta_p = 0$のとき, γは最高第一種の過誤確率のみに依存して決まる．一方$\delta_p = 0.75$のとき, γは最高第一種の過誤確率と最低検出力の両方を満たすように決まる．したがって, $\delta_p = 0.75$とすると, 選択されるγは$\delta_p = 0$と比べて大きな値が選択される．表 2.1 と表 2.2 を比較すると，γの値が異なっているのは$n_{CC} = 200$，$n_{HC} = 50, 100, 200$のときである．新規対照のサンプルサイズに比べて既存対照のサンプルサイズが小さいときに検出力が減少しやすいと予想できる．

次に，TTTP のγ_1とγ_2に注目する．表 2.1 のγ_1は，多くの設定で$\gamma/2$と同じ値となっている．$\delta_p = 0$のとき，γは最高第一種の過誤確率を制御するように選択されていたことを踏まえると，TTTP においてはγ_1が最高第一種の過誤確率の制御を担当することが分かる．表 2.2 の$n_{CC} = 200$，$n_{HC} = 50, 100, 200$における，γ_2と$\gamma/2$は一致していない．$\delta_p = 0.75$のとき，γは最低検出力が名目水準以上になるように決まる．γ_2と$\gamma/2$とが一致していないということは，γ_2が最低検出力を制御しているわけではないことを意味している．表 2.1 と表 2.2 から，$\delta_p = 0$とした場合と$\delta_p = 0.75$とした場合で選択されたγ_1とγ_2に違いはない．γ_2は drift が 0 の場合の検出力を最大化するように選ばれるのであり，そのようなγ_2は drift が 0 ではないときの検出力を減少させないと予想できる．この予想は多くの場合正しいが，2.5.2.3 節で述べるように，一部の設定では正しくないことが分かっている．

図 2.3 と図 2.4 に drift の関数としてみた CTTP と TTTP の動作特性を示す．図 2.3 は$\delta_p = 0$，図 2.4 は$\delta_p = 0.75$の結果である．CTTP について，図 2.3 の$n_{CC} = 200$では，検出力が 75%を下回る箇所が存在しており，本節 2 段落目の予想が正しいことが分かる．$\delta_p = 0$で検出力が減少する設定では，図 2.4 に示す$\delta_p = 0.75$とした場合，検出力の減少は防ぐことができるが，$\mu_C = 0$の場合の検出力も減少してしまう．また，最高第一種の過誤確率は5%に達していない．CTTP はチューニングパラメータがγのみであるため，最高第一種の過誤確率と最低検出力の一方のみを任意の値に制御することしかできないことが分かる．一方で, TTTP では, すべての設定で最高第一種の過誤確率が 5%となりつつ，$\delta_p = 0$の場合であっても検出力が減少しない．TTTP では，最高第一種の過誤確率をδ_e以下に制御するようにγ_2が決まり， drift が 0 のときの検出力を最大化するようにγ_1が決まる．そのようにして決まるγ_1とγ_2の組は検出力を減少させない.

TTTP と CTTP は，n_{HC}が小さいときに動作特性に違いが現れるが，n_{HC}が大きくな

ると検出力と第一種の過誤確率に違いはなくなる．このとき，γ_2は 0.5 が選択されている． これは，$\bar{x}_{CC}$が$\bar{x}_{HC}$よりわずかでも小さい場合，既存対照を併合しないことを意味する．検出力と第一種の過誤確率がほとんど変わらないことから，n_{HC}が大きいときは，併合確率が大きくなる CTTP で十分かもしれない．TTTP は， n_{HC}が小さいときに，CTTP と比較して，検出力を減少させない（$\delta_p = 0$のとき），あるいは drift がない場合の検出力が高い（$\delta_p = 0.75$のとき）方法であるといえる．

表 2.1 選択された有意水準一覧（$\delta_p = 0$）

			TTTP		CTTP
n_{CC}	θ	n_{HC}	γ_1	γ_2	$\gamma/2$
200	-0.26	50	0.118	0.375	0.118
		100	0.176	0.338	0.176
		200	0.231	0.300	0.230
		400	0.278	0.267	0.278
100	-0.32	50	0.258	0.312	0.258
		100	0.294	0.500	0.327
		200	0.329	0.500	0.374
		400	0.353	0.500	0.399

TTTP：2 つの片側検定を組み合わせた検定併合法

CTTP：従来の検定併合法

n_{CC}：新規対照のサンプルサイズ

θ：治療効果の差

γ_1，γ_2：TTTP の有意水準

γ：CTTP の有意水準

表 2.2 選択された有意水準一覧（$\delta_p = 0.75$）

			TTTP		CTTP
n_{CC}	θ	n_{HC}	γ_1	γ_2	$\gamma/2$
200	-0.26	50	0.118	0.375	0.329
		100	0.176	0.338	0.282
		200	0.231	0.300	0.236
		400	0.278	0.267	0.278
100	-0.32	50	0.258	0.312	0.258
		100	0.294	0.500	0.327
		200	0.329	0.500	0.374
		400	0.353	0.500	0.399

TTTP：2 つの片側検定を組み合わせた検定併合法

CTTP：従来の検定併合法

n_{CC}：新規対照のサンプルサイズ

θ：治療効果の差

γ_1，γ_2：TTTP の有意水準

γ：CTTP の有意水準

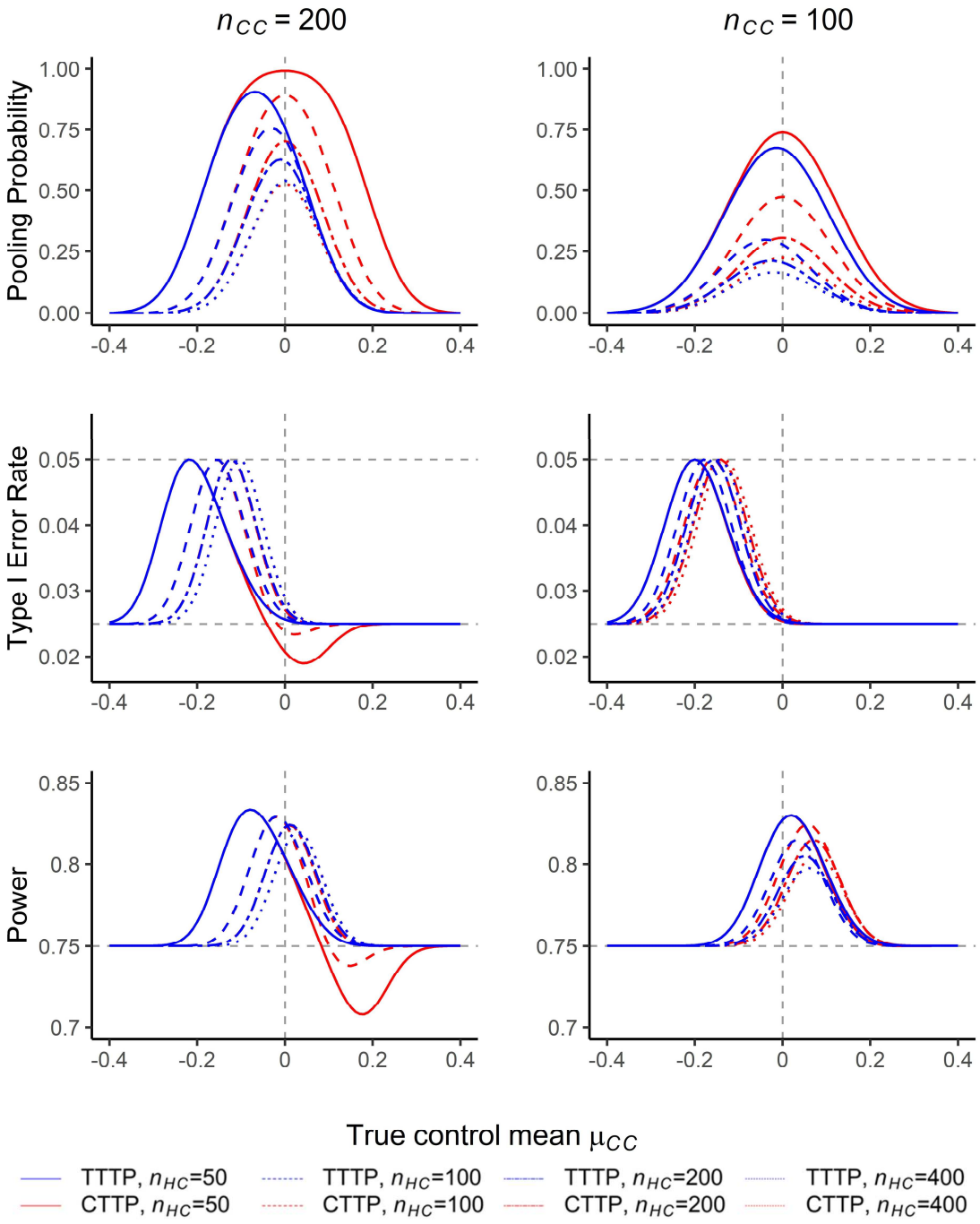

図 2.3 CTTP と TTTP の動作特性（$n_T = 200$，$\delta_p = 0$）.

TTTP：2 つの片側検定を組み合わせた検定併合法

CTTP：従来の検定併合法

δ_p：最低検出力の許容下限

n_T，n_{CC}，n_{HC}：試験治療，新規対照，既存対照のサンプルサイズ

μ_{CC}：新規対照の母平均

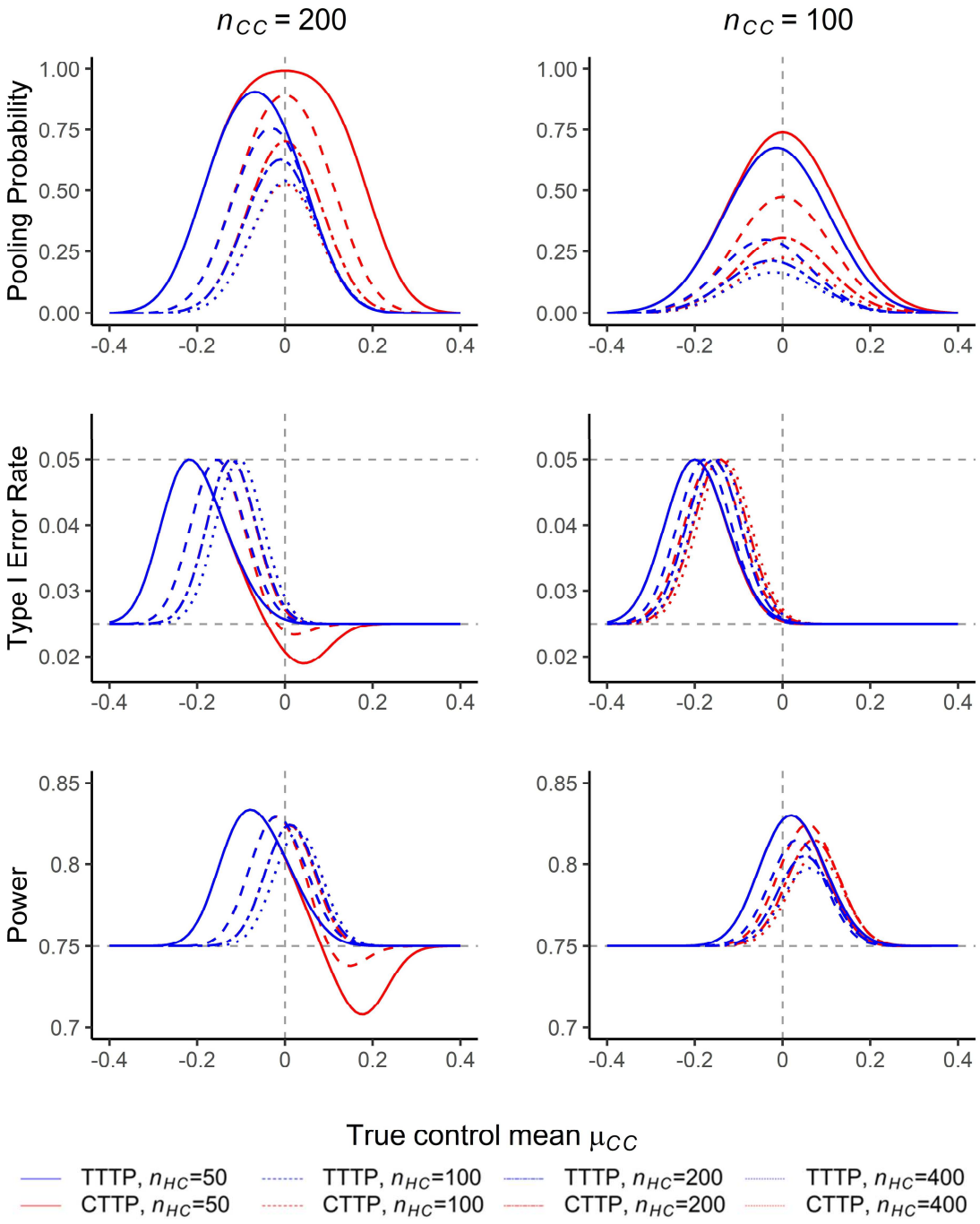

図 2.4 CTTP と TTTP の動作特性（$n_T = 200$，$\delta_p = 0.75$）.

TTTP：2 つの片側検定を組み合わせた検定併合法

CTTP：従来の検定併合法

δ_p：最低検出力の許容下限

n_T，n_{CC}，n_{HC}：試験治療，新規対照，既存対照のサンプルサイズ

μ_{CC}：新規対照の母平均

2.5.2.3 既存対照のサンプルサイズの大きさと動作特性の関係

図 2.5 から図 2.8 及び付録に示す図 A1.1 から図 A1.28 は，既存対照のサンプルサイズと選択された有意水準, 各動作特性の関係を表す. 図 2.5 と図 2.7 は$n_T = 200$, $n_C = 200$，図 2.6 と図 2.8 は$n_T = 200$，$n_C = 100$の結果である．図 2.5 と図 2.6 は$\delta_p = 0$, 図2.7 と図2.8 は$\delta_p = 0.75$の結果である. そのほかの設定の結果を図A1.1 から図A1.28 に示す．各動作特性について，drift が 0 である場合の値，最大値，最小値をプロットした．最高第一種の過誤確率と最低検出力に基づいて有意水準を選択している点と 2.5.2.2 節の結果から，検定併合法の動作特性の理解において，drift が 0 の場合の値と最大値，最小値に注目すれば十分だと考えた． CTTP の最高第一種の過誤確率について，TTTP のものと完全に重なっている箇所がある．

選択された有意水準について，δ_pが 0 であるとき，CTTP の$\gamma/2$は，n_{HC}について連続的に上昇する．n_{HC}が大きいと，第一種の過誤確率の上昇も大きくなるため，γを厳しく設定する必要があると考えられる．TTTP のγ_1も同様に増加するが，n_{CC}の大きさによってその挙動は異なる．n_Tとの比が 1:1 の場合， n_{HC}の増加に伴って$\gamma/2$と同様に上昇する．このとき，γ_2はn_{HC}の増加に伴って減少している．一方で，n_Tに比べてn_{CC}が小さいとき，n_{HC}が小さいと$\gamma/2$と同じ値をとるが，あるn_{HC}において乖離し始め，$\gamma/2$よりも小さくなる．このとき，γ_2は 0.5 となる．γ_2は，主に検出力の制御にかかわるが，$\gamma_2 = 0.5$は既存対照の併合に関して厳しい設定であり，第一種の過誤確率にまで影響を与えるため，γ_1は$\gamma/2$と比べて小さな値になると思われる．

次に各動作特性を説明する．$\delta_p = 0$のとき，CTTP では，最高第一種の過誤確率と drift が 0 の場合の第一種の過誤確率が，n_{HC}が小さいと名目水準の 2.5%を下回る．このとき，drift が 0 の場合の検出力は TTTP のものと変わらないが，最低検出力は名目水準の 75%を下回っている．$\delta_p = 0.75$とすると，CTTP の第一種の過誤確率・検出力低下は生じないが，drift が 0 の場合の検出力は，TTTP のものより低くなる．

図 2.5 から図 2.8 及び図 A1.1 から図 A1.28 より，既存対照のサンプルサイズn_{HC}の増大に伴って，検出力が上昇するわけではないことが分かる．あるn_{HC}の値までは検出力が上昇するが，以降は低下する．この現象は 2 つの要因が関係していると考えられる．第一に，n_{HC}が大きいと，第一種の過誤確率の上昇を抑えるために，より大きなチューニングパラメータを設定する必要があり, 併合確率が低下してしまう. 第二に，対照群のデータサイズのみが大きくなることによる追加的な検出力の上昇分は低下していく．n_{HC}の増大による併合確率低下が引き起こす検出力の低下が，n_{HC}の増大による追加的な検出力の上昇を上回るため，このような現象が観察されると考えられる．

Li et al. (2020)が指摘したように，drift が 0 であっても第一種の過誤確率の上昇が確認された．図 2.5 から図 2.8 と図 A1.1 から図 A1.28 より，新規対照のサンプルサイズが大きくなるほど，あるいは，既存対照のサンプルサイズが大きくなるほど，第一種

の過誤確率が上昇することが分かる.

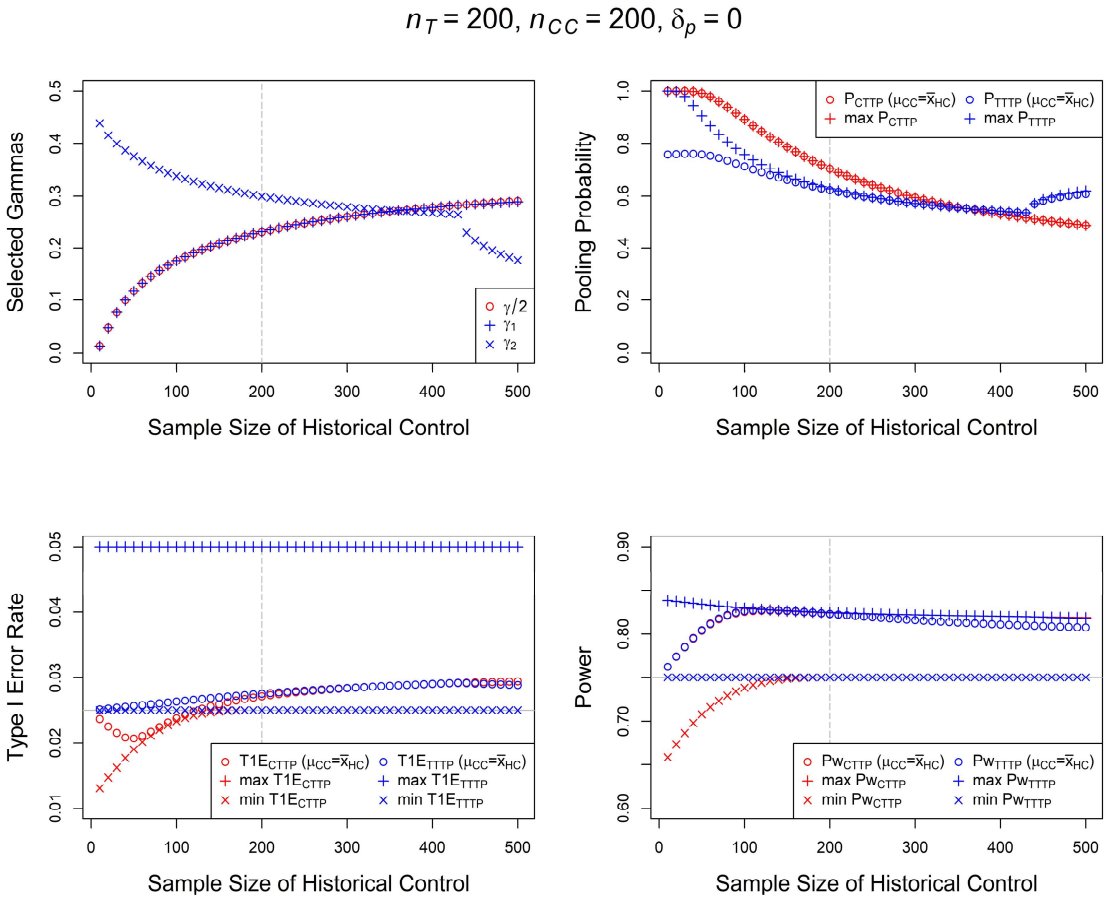

図 2.5 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係（$n_T = 200$，$n_{CC} = 200$，$\delta_p = 0$）

TTTP：2 つの片側検定を組み合わせた検定併合法

CTTP：従来の検定併合法

γ_1，γ_2：TTTP の有意水準

γ：CTTP の有意水準

δ_p：最低検出力の許容下限

μ_{CC}：新規対照の母平均

$\bar{x}_{HC}$：既存対照の標本平均

n_T，n_{CC}：試験治療，新規対照のサンプルサイズ

P_{TTTP}，P_{CTTP}：併合確率

$T1E_{TTTP}$，$T1E_{CTTP}$：第一種の過誤確率

Pw_{TTTP}，Pw_{CTTP}：検出力

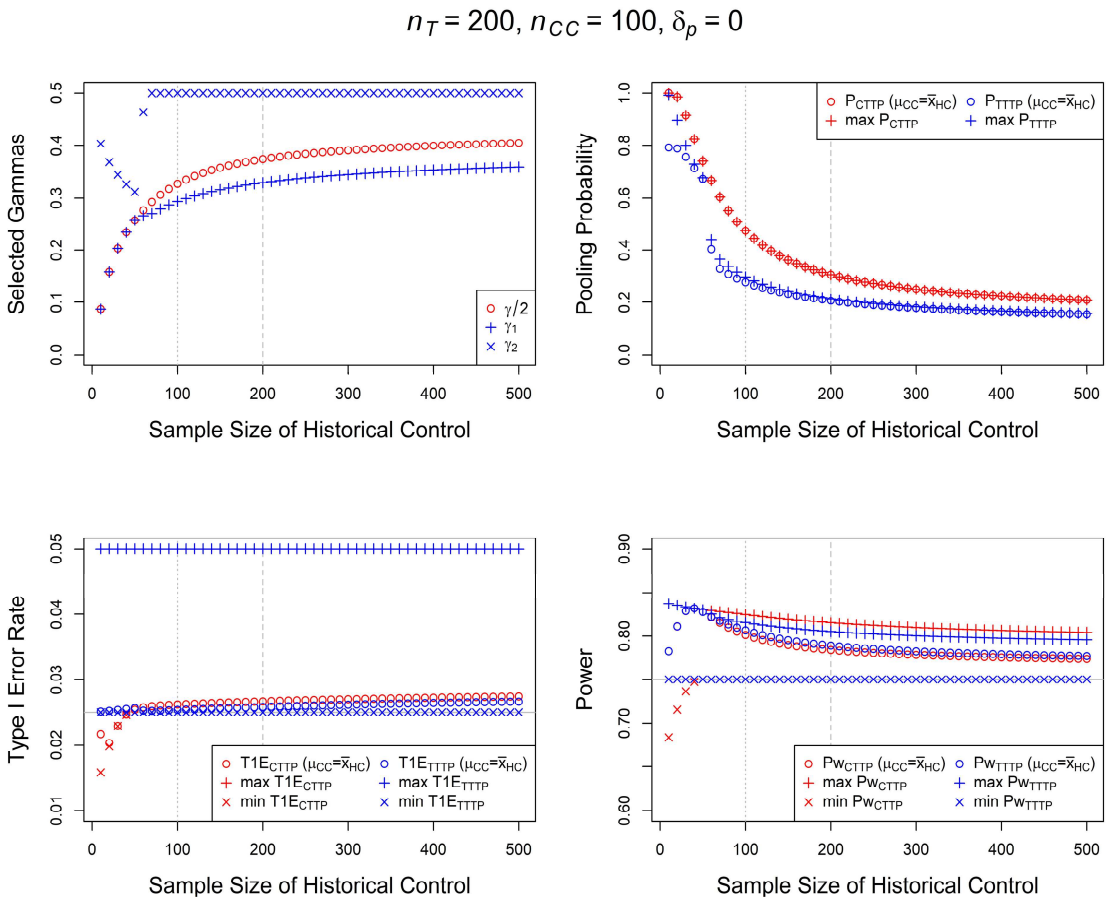

図 2.6 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係（$n_T = 200$，$n_{CC} = 100$，$\delta_p = 0$）

TTTP：2 つの片側検定を組み合わせた検定併合法

CTTP：従来の検定併合法

γ_1，γ_2：TTTP の有意水準

γ：CTTP の有意水準

δ_p：最低検出力の許容下限

μ_{CC}：新規対照の母平均

$\bar{x}_{HC}$：既存対照の標本平均

n_T，n_{CC}：試験治療，新規対照のサンプルサイズ

P_{TTTP}，P_{CTTP}：併合確率

$T1E_{TTTP}$，$T1E_{CTTP}$：第一種の過誤確率

Pw_{TTTP}，Pw_{CTTP}：検出力

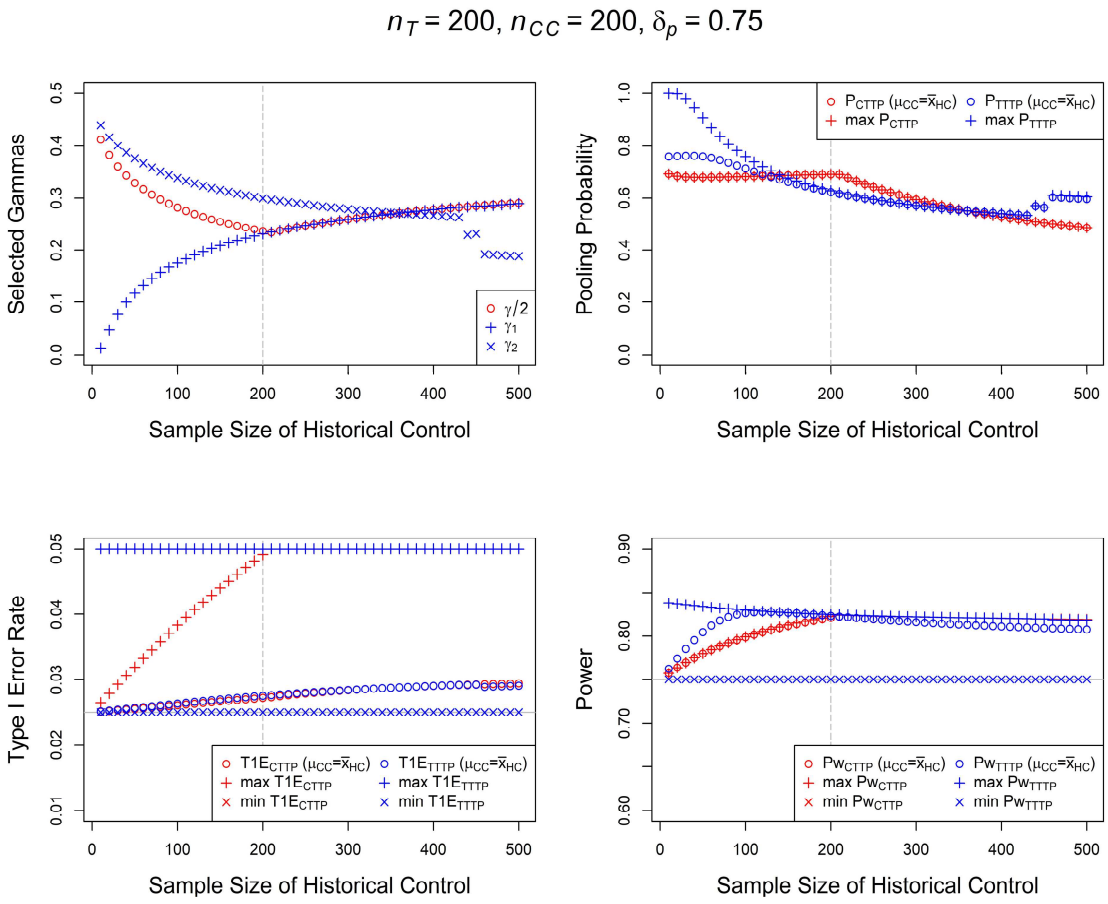

図 2.7 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係（$n_T = 200$，$n_{CC} = 200$，$\delta_p = 0.75$）

TTTP：2 つの片側検定を組み合わせた検定併合法

CTTP：従来の検定併合法

γ_1，γ_2：TTTP の有意水準

γ：CTTP の有意水準

δ_p：最低検出力の許容下限

μ_{CC}：新規対照の母平均

$\bar{x}_{HC}$：既存対照の標本平均

n_T，n_{CC}：試験治療，新規対照のサンプルサイズ

P_{TTTP}，P_{CTTP}：併合確率

$T1E_{TTTP}$，$T1E_{CTTP}$：第一種の過誤確率

Pw_{TTTP}，Pw_{CTTP}：検出力

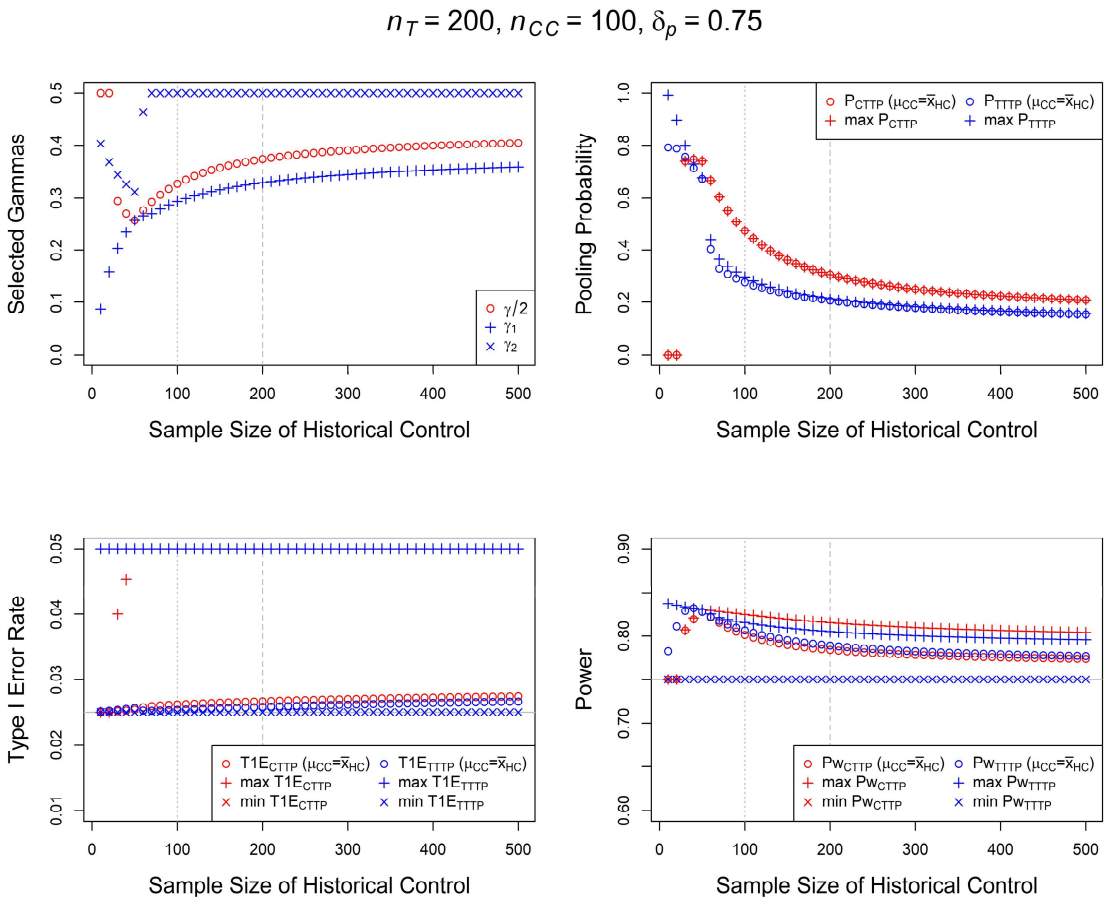

図 2.8 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係（$n_T = 200$，$n_{CC} = 100$，$\delta_p = 0.75$）

TTTP：2 つの片側検定を組み合わせた検定併合法

CTTP：従来の検定併合法

γ_1，γ_2：TTTP の有意水準

γ：CTTP の有意水準

δ_p：最低検出力の許容下限

μ_{CC}：新規対照の母平均

$\bar{x}_{HC}$：既存対照の標本平均

n_T，n_{CC}：試験治療，新規対照のサンプルサイズ

P_{TTTP}，P_{CTTP}：併合確率

$T1E_{TTTP}$，$T1E_{CTTP}$：第一種の過誤確率

Pw_{TTTP}，Pw_{CTTP}：検出力

2.5.3 γ_2の選択に関する追加検討

2.5.2.2 節，2.5.2.3 節で説明したように，$\delta_p = 0$と設定した場合に選択されるγ_2には解釈の難しい現象が確認された. γ_1は, 最高第一種の過誤確率の制御に関与しており，既存対照データのサンプルサイズに上昇に伴って，γ_1も高い水準が選択されるという単純な挙動を示した．一方で，γ_2は，既存対照のサンプルサイズの増大に伴ってγ_2も低下する場合と，あるn_{HX}を境に上昇に転じる場合があった.

また, 選択されるγ_2は必ずしも検出力の低下を防ぐわけではないようである. 図 2.5 から図 2.8 に示した$(n_T, n_{CC}) = (200, 200), (200, 100)$の設定では, δ_pの値にかかわらず，選択された TTTP のγ_1とγ_2の値は変わらず，検出力を低下させなかった．したがって, TTTP を用いる場合, 最低検出力に関する設定は不要であるように思われる．検出力の条件を設定しないことは，計算時間の観点からも好ましい．しかしながら，図 A1.3，図 A1.7，図 A1.10，図 A1.13 で確認できるように，一部の設定では，$\delta_p = 0$と設定すると，最低検出力が名目水準を下回る場合がある．Drift が 0 の場合の検出力を最高にするようなγ_2が，必ずしも drift が 0 ではない場合の検出力を低下させないわけではない.

$\delta_p = 0$におけるγ_2の選択について追加検討を行った．以下に示す 3 つのサンプルサイズの組合せを検討した.

$$(n_T, n_{CC}, n_{HC}) = (200, 200, 200), (200, 150, 150), (200, 100, 200)$$

1 つ目は，図 2.5 に示す，n_Tとn_{CC}が等しく，n_{HC}の増大に伴ってγ_2が低く選択される設定の 1 つである．2 つ目は，図 A1.10 に示す，検出力が名目水準を下回る設定の 1 つである．3 つ目は，図 2.6 に示す，n_Tがn_{CC}より大きく，γ_2として 0.5 が選択される設定である. それぞれの設定について, γ_1とγ_2を 0.01 刻みで 0 から 0.5 まで変化させ，それぞれの組で drift が 0 の場合の検出力（Pw_0）を計算した.

図 2.9 に，Pw_0，γ_1，γ_2の 3 次元プロットとPw_0とγ_2の 2 次元プロットを示す．γ_1とγ_2の組は，(i)最高第一種の過誤確率が 5%以下になるγ_1とγ_2の組のうち，(ii) Pw_0が最高になるものが選択される．図 2.9 では，(i)を満たす(γ_1, γ_2)を水色，満たさないものをオレンジ色で示した．(i)を満たすもののうち，(ii)を満たす (γ_1, γ_2)の組を赤色で示した．(i)を満たす組と満たさない組の境界を詳細に調べるため，特定のγ_1の値に注目して，Pw_0とγ_2の 2 次元プロットを作成した．γ_1は，3 次元プロットにおいて既存対照を借りる場合と借りない場合の境界が含まれるように選択した. 例えば, 図 2.9(a) では, 3 次元プロットから$\gamma_1 = 0.22, 0.23, 0.24$の点を抜き出し, 横軸を$\gamma_2$，縦軸を$Pw_0$とした 2 次元プロットを作成した.

図 2.9 から，Pw_0の曲面は，(I) γ_1について減少関数であり，(II) γ_2について上に凸な関数であることが分かる．(I)は，γ_1を高くすると検出力が上昇することを意味しており，第一種の過誤確率の制約を満たす中で最も高いγ_1が選ばれる．(II) は，γ_2は高

すぎても低すぎても検出力が低下することを意味する．ただし，γ_2の変化に対しPw_0はあまり変わらない.

図 2.9(a)では，Pw_0を最高にするγ_2の候補として，0.36 と 0.21 があり，わずかな検出力の差で 0.36 が採用されている．一方で，図 2.9(b)では，γ_2の候補は 0.37 と 0.13 であり， 0.13 が採用されている．2 つの候補のPw_0の差はわずかであるため，n_{HX}の増大に伴って，切り替わりが生じることがあると考えられる．このことが，γ_2の変化がn_Hに関して離散的に変化する場合が生じる原因である．例えば，図 2.5 の設定では，$n_{HX} \leq 420$において，γ_2の候補のうち高い値が採用されるため，連続的に減少しており，$n_{HX} \geq 430$では，低い値が採用されるため，γ_2の傾向に変化が生じている．$\delta_p = 0$のとき，最低検出力の制約はないため，図 2.9(b)のように，γ_2に低い値が選択された結果，最低検出力が 75%を下回ることもあるようである．また，n_{CC}がn_Tよりも小さくn_{HC}が十分に大きいと，図 2.9(c)のように，Pw_0を最高にするγ_2が 0.5 になる．

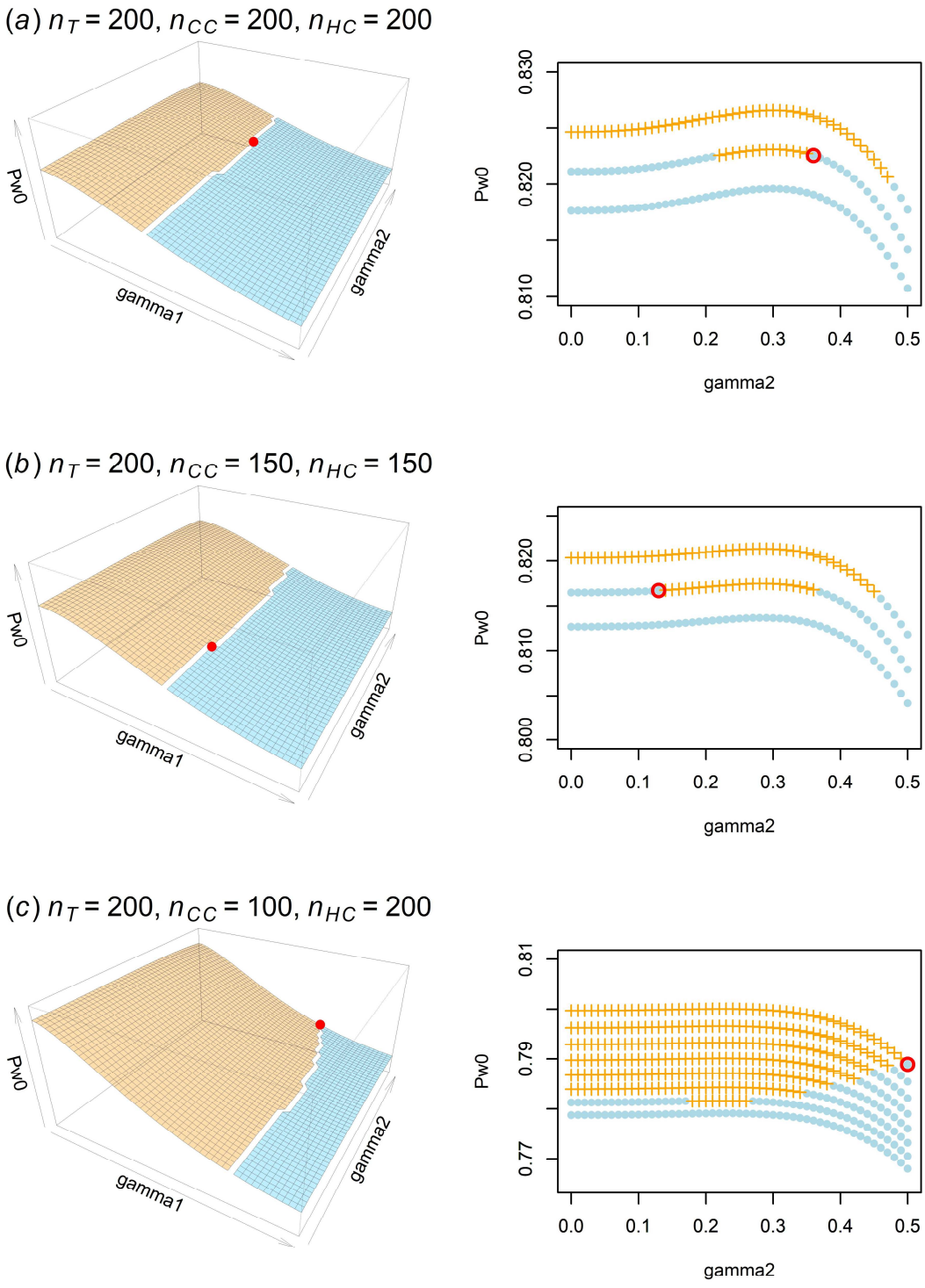

図 2.9　drift が 0 の場合の TTTP の検出力，γ_1，γ_2のプロット（$\delta_p = 0$）

δ_p：最低検出力の許容下限，Pw_0：drift が 0 の場合の検出力

n_T，n_{CC}，n_{HC}：試験治療，新規対照，既存対照のサンプルサイズ

gamma1，gamma2：γ_1，γ_2，TTTP の有意水準

2.6 うつ病データへの適用

2 つの片側帰無仮説を組み合わせた検定併合法（TTTP）と第一種の過誤確率と検出力に基づく有意水準選択法の特性を明らかにするため，従来の検定併合法（CTTP）と併せて数値例に適用し比較する．数値例は，Li et al. (2020) が case study で用いた P061 試験と P059 試験（Keller et al., 2006）のデータ及びこれらのデータを一部修正した仮想データを利用する．

P061 試験と P059 試験は，うつ病の外来患者を対象にニューロキノン 1 受容体拮抗薬であるアプレピタントが独自の抗うつ作用を持つかどうかを確認するための一連の 5 つの第 3 相試験の一部である．P061 試験と P059 試験は，アプレピタント 80mg 群とプラセボ群にパロキセチン 20mg 群を加えた 3 群試験であった．主要評価項目は Hamilton Depression Rating Scale (HAMD; Hamilton, 1960)の最初の 17 項目（HAMD-17）であり，副次評価項目の 1 つは Hamilton Rating Scale for Anxiety (HAMA; Hamilton, 1959) であった．HAMD-17 と HAMA はどちらもベースラインからの変化量を評価した．どちらの試験においても，片側有意水準 5%の下で HAMD-17 及び HAMA についてアプレピタント群とプラセボ群に統計的な有意差はなかった．しかし，パロキセチン群とプラセボ群には HAMD-17 について有意差が認められ，HAMA では有意差はないもののパロキセチンの効果を示唆する結果であった．パロキセチン群の HAMA における有意差なしの結果はサンプルサイズの不足が考えられる．プラセボ群のデータを統合することで，有意差が示されるかもしれない．

本節では，P059 試験のプラセボ群のデータを既存対照として扱い，P061 試験が検定併合法を用いたハイブリッド対照デザインであったと仮定した検討を行う．表 2.3 に適用したデータを一覧に示す．本研究では 2 つのケースを想定した．Case 1 は，P059 試験と P061 試験の実際に得られたデータである．Case 2 では，Case 1 のデータを一部修正している．2.5 節の検討から，TTTP は新規対照の母平均μ_{CC}が既存対照の標本平均$\bar{x}_{HC}$よりも小さく，既存対照のサンプルサイズが小さい場合に，検出力の低下を防ぐことが示された．そこで Case 2 では，Case 1 の既存対照のサンプルサイズを 149 から 50 に，既存対照の標本平均を-8.1 から-9.6 に，パロキセチン群の変化量を-9.9 から -10.3 に変更した．表 2.3 には，併合された場合とされなかった場合の治療効果に関する仮説に対する片側 P 値も掲載した．Case 1 では，併合しない場合の片側 P 値は 0.095 であるが，併合した場合は 0.032 となる．一方 Case 2 では，併合しない場合は 0.040 であるが，併合した場合は 0.058 となる．

パロキセチンがプラセボと比較して HAMA を低下させることを示すための 1:1 割付の無作為化試験の計画を想定する．どちらのケースにおいても，変化量の差を-2.3，標準偏差を 8.3，有意水準を片側 5%として，既存対照を考慮しない検出力が 75%となるように，新規試験のサンプルサイズを 280 人と設定したとする（表 2.3 に示すよう

に，実際にはプラセボ群で 140 人，パロキセチン群で 137 人が解析対象であったとする）．第一種の過誤確率と検出力に基づく有意水準選択法を用いて選択されたγ_1とγ_2を表 2.4 に示す．表 2.4 には，仮説(2.9)と仮説(2.10)の片側 P 値（p_1とp_2）と併合有無（○：併合する，×：併合しない），治療効果に関する仮説に対する両側 P 値も示す．既存対照を併合すると判定される場合は表 2.3 に示す併合する場合の両側 P 値が，併合しないと判定される場合は表 2.3 に示す併合しない場合の両側 P 値が表示される．CTTP の有意水準γと仮説(2.2)の両側 P 値（p）の表記は，TTTP の表記に合わせてγ_1とγ_2，p_1とp_2に分けて表示した．γはγ_1とγ_2の和であり，pはp_1とp_2の和である．

まずは Case 1 について検討する．Case 1 は，新規対照の標本平均$\bar{x}_{CC}$が$\bar{x}_{HC}$よりも大きいため，2.5 節の結果より，第一種の過誤確率上昇が懸念される状況である．TTTP では，$(\gamma_1, \gamma_2) = (0.290, 0.266)$，$(p_1, p_2) = (0.258, 0.742)$である．$p_2 > \gamma_2$ではあるが$p_1 < \gamma_1$であるため，TTTP では既存対照は併合されない．第一種の過誤確率を主に制御するγ_1の基準が満たされていないことからも，潜在的な第一種の過誤確率の上昇を懸念して既存対照が併合されないことが分かる．また，CTTP では，$\gamma = \gamma_1 + \gamma_2 = 0.580$，$p = p_1 + p_2 = 0.516$である．$p < \gamma$であるため，CTTP においても既存対照は併合されない．もし CTTP において慣例的な値である$\gamma = 0.2$（すなわち$\gamma_1 = \gamma_2 = 0.1$）と設定していた場合，$p_1 > \gamma_1$であるため既存対照は併合され，治療効果に関する仮説の P 値は 0.032 となり有意差が得られる．しかし，$\gamma = 0.2$という設定は第一種の過誤確率の上昇を一切考慮していない点に問題がある．

Case 2 は既存対照のサンプルサイズが小さく，$\bar{x}_{CC} < x_{HC}$である．2.5 節の結果より，この状況では，CTTP の検出力は低下する恐れがある．Case 2 では，既存対照が併合されない場合の治療効果に関する仮説の P 値は 0.040 であり，併合される場合は 0.058 である．すなわち，既存対照を併合すると有意差が消えてしまう．CTTP では，$\gamma = \gamma_1 + \gamma_2 = 0.438$，$p = p_1 + p_2 = 0.472$であり，$p > \gamma$であるため，既存対照が併合されてしまう．一方，TTTP では，$(\gamma_1, \gamma_2) = (0.219, 0.379)$，$(p_1, p_2) = (0.764, 0.236)$であり，$p_1 < \gamma_1$であるため既存対照は併合されない．最低検出力に関与するγ_1 の基準が満たされていないことから，TTTP は検出力低下を懸念して併合しなかったと解釈できる．

表 2.3 検討したデータ一覧

	新規試験				
	プラセボ群		パロキセチン群		P 値
	n	mean (SD)	n	mean (SD)	
Case 1: **実際のデータ**					
P059 のプラセボ群	149	-8.1 (8.3)			
併合しない場合	140	-8.7 (7.3)	137	-9.9 (7.9)	0.095
併合した場合	289	-8.4 (7.8)			0.032
Case 2: **仮想的なデータ**					
既存対照	50	-9.6 (8.3)			
併合しない場合	140	-8.7 (7.3)	137	-10.3 (7.9)	0.040
併合した場合	190	-8.9 (7.6)			0.058

n：サンプルサイズ

mean：HAMA のベースラインからの変化量 の平均

SD：標準偏差

P 値：Student の t 検定の片側 P 値

表 2.4 選択された有意水準と併合有無

	方法	γ_1	γ_2	p_1	p_2	併合	P 値
Case 1	TTTP	0.290	0.266	0.258	0.742	×	0.095
	CTTP	0.290	0.290	0.258	0.258	×	0.095
Case 2	TTTP	0.219	0.379	0.764	0.236	×	0.040
	CTTP	0.219	0.219	0.236	0.236	○	0.058

γ_1，γ_2：TTTP の有意水準

p_1：仮説(2.9)の片側 P 値

p_2：仮説(2.10)の片側 P 値

○，×：併合有り，併合無し

P 値：Student の t 検定の片側 P 値

2.7 考察

本章では，既存の検定併合法（CTTP）について，(i)第一種の過誤確率と検出力の制御に柔軟性がない，(ii)併合可否を判断するための検定の有意水準が場当たり的に決められている，という 2 つの課題を指摘した．そして，(i)を改善するために，CTTP の両側仮説を 2 つの片側仮説に分割し，それぞれの仮説に異なる有意水準を与える方法（TTTP）を提案した．また，(ii)を改善するために，最高第一種の過誤確率と最低検出力に基づいて有意水準を選択する方法を提案した．

TTTP の本質的で画期的なアイデアは，既存対照の併合可否を評価する際，新規対照の標本平均（$\bar{x}_{CC}$）と既存対照の標本平均（$\bar{x}_{HC}$）の差の絶対値だけではなく，符号まで考慮する差を評価する点にある．CTTP は$|\bar{x}_{CC} - \bar{x}_{HC}|$を評価している．しかし，この方法だと 2.3.2 項で説明したように, drift の関数として考えた場合の第一種の過誤確率と検出力の制御が柔軟に行えない．第一種の過誤確率と検出力のどちらか一方を制御しようとすると，もう片方は自動的に定まってしまうからである．この特性は，CTTP に限らず既存の情報借用法に共通する特性である．$\bar{x}_{CC} - \bar{x}_{HC}$の正負に応じて借用の基準を変えることで第一種の過誤確率と検出力を別々に制御できると考えた．本研究では，検定併合法において$\bar{x}_{CC} - \bar{x}_{HC}$の正負を考慮するために，2 つの片側帰無仮説を組み合わせる TTTP を提案した. $\bar{x}_{CC} - \bar{x}_{HC}$の正負に応じて借用の基準を変えるというアイデアは，ベイズ流の情報借用法にも応用できるかもしれない．ベイズ流の方法への拡張は今後の研究課題の 1 つである．

$\bar{x}_{CC} - \bar{x}_{HC}$の正負に応じて借用の基準を変える既存対照併合法は，筆者の知る限り他に提案されているものはない．既存の既存対照併合法はどれも，$|\bar{x}_{CC} - \bar{x}_{HC}|$が小さいときに既存対照の情報を多く利用し，大きいときには既存対照の情報の利用を控えるという考え方に基づいている．新規対照と既存対照の異質性が大きいと，第一種の過誤確率や検出力が悪化する恐れがあるため，類似性が高いときに既存対照の情報を多く利用するという発想は自然である．この考え方に基づくと，$\bar{x}_{CC} - \bar{x}_{HC}$の正負を考慮するという発想には至らないであろう．本章では，第一種の過誤確率と検出力をより良く制御するためにはどうしたらよいか, という観点から研究を進めた. その結果, $\bar{x}_{CC} - \bar{x}_{HC}$の正負を考慮することで第一種の過誤確率と検出力の制御に柔軟性を持たせることが可能となることを発見したのである．

$\bar{x}_{CC} - \bar{x}_{HC}$の正負を考慮することの欠点は，$\mu_{CC} = \bar{x}_{HC}$であったとしても，治療効果の推定量にバイアスが生じる恐れがある点にある．極端な例として，$\gamma_1 = 0.5$，$\gamma_2 = 0$と設定した場合を考える．この時，併合区間は,

$$[-\infty\ , \bar{x}_{HC} - \bar{x}_{CC}]$$

となる．すなわち，$\bar{x}_{CC} \leq \bar{x}_{HC}$のときは常に既存対照を併合するが，$\bar{x}_{CC} > \bar{x}_{HC}$のときは既存対照を一切併合しないという設定である．このようにγ_1とγ_2を設定すると，

$\mu_{CC} = \bar{x}_{HC}$のとき，2 分の 1 の確率で，$\bar{x}_{CC}$よりも大きな$\bar{x}_{HC}$ばかりが併合される．そのため平均すると，対照群の標本平均は，$\bar{x}_{CC}$のみから計算される標本平均よりも大きくなる．したがって，治療効果の推定量にもバイアスが生じる．この性質は，$\bar{x}_{HC} - \bar{x}_{CC}$に対する併合区間の非対称性に由来する．$\gamma_1$と$\gamma_2$の差が大きいと併合区間の非対称性も大きくなる. この欠点を回避するためには,γ_1とγ_2をできるだけ近い値にする，検定のみ既存対照を利用し，推定では既存対照を利用しない等の対策が考えられる．

γ_1とγ_2の設定方法として，最高第一種の過誤確率と最低検出力に基づく有意水準選択法を提案したが，必ずしもこの方法で選択されたγ_1とγ_2を用いる必要はない．提案する有意水準選択法は, 最高第一種の過誤確率と最低検出力を任意の値に制御しつつ, drift が 0 の場合の検出力（Pw_0）を最高にするようにγ_1とγ_2を選択する方法であった．この方法は，既存対照の併合確率やバイアスを一切考慮していない．2.5 節の結果から，Pw_0を最高にするγ_2の値として 0.5 が選択される場合があった．この設定は，$\bar{x}_{CC}$が$\bar{x}_{HC}$より少しでも小さい場合は既存対照を併合しないという設定である．このようにγ_2を設定してしまうと，既存対照の併合確率が低くなってしまうだけでなく，前述のバイアスの問題も懸念される．幸い，2.5.3 項の検討から，γ_2がPw_0に与える影響は大きくない．γ_2の値に変更することで，Pw_0の低下という欠点を上回る，併合確率やバイアスの利点があるのであれば，γ_2の値を変更してもよい．また，このような事情から，提案した有意水準選択法は，さらなる改善の余地が残されていることが示唆される．

各データ（新規試験治療，新規対照，既存対照）のサンプルサイズの大きさの最適な組み合わせには議論の余地がある．2.5 節では，新規試験における割付比として，4:1 割付，2:1 割付，4:3 割付，1:1 割付を検討した．割付比に応じて動作特性は変化することが分かったが，最適な割付比についての議論は本章ではできていない．また，2.5.2.3 節の検討より，必ずしも大きなサンプルサイズの既存対照が，高い検出力をもたらすわけではないことが示された．新規治療，新規対照，既存対照のサンプルサイズの大きさの最適な組み合わせについて，さらなる研究が期待される．

前述の議論に関連して，本章では考慮していないが，既存対照データに 0 から 1 の値で重み付けすることで，既存対照データの解析に寄与する情報量としての実質的なサンプルサイズを縮小させることができる（ベイズ流の文脈では conditional power prior に相当する）．例えば，既存対照に 0.5 の重みをつけて解析に用いることは，サンプルサイズ換算で既存対照の情報量を半分に割り引くことに相当する．TTTP のチューニングパラメータであるγ_1とγ_2に加えて， 既存対照の割引率も考慮した最適なパラメータ選択は，今後の研究課題の 1 つである．

近年の検定併合法の発展の 1 つに，同等性検定に基づいた検定併合法（equivalence-based test-then-pool method, ETTP; Li et al., 2020）がある．ETTP は，CTTP が併合可否を

判定するために用いていた優越性検定を，適当な同等性マージンを設定した同等性検定に置き換えた．ETTP は，同等性検定において有意差がある場合に，既存対照を併合するという方法である．優越性検定において，有意差がないということは，必ずしも新規対照と新規対照が類似していることを意味するわけではない．Li らは，この点を CTTP の欠点として指摘した．同等性検定における有意差は，新規対照と既存対照の同等性を積極的に主張する．したがって，確かに ETTP の方が論理的整合性に優れていると考えられる．同等性検定の仮説は，2 つの片側仮説を組み合わせた仮説に変形できる（Schuirmann, 1987）．これらの仮説に別々の有意水準を与えることで， TTTP と同様のアイデアを適用可能である．

しかしながら，本研究では ETTP ではなく CTTP を改良する形で TTTP を提案した．理由は次の通りである．Li らが報告しているように，優越性検定と同等性検定は理論的に等価である．すなわち，CTTP の有意水準を適切に設定することで，ある同等性マージンと同等性検定の有意水準を設定した ETTP と同一な性能をもつ CTTP を実現可能である．本研究の提案の 1 つは，第一種の過誤確率と検出力に基づいて各手法のチューニングパラメータを設定するというものであった．この提案法に従う限り，CTTP と ETTP，及び TTTP と 2 つの同等性片側仮説に別々の有意水準を与える方法は，同等の性能を持つようにチューニングパラメータが選択される． すなわち，併合区間は同一となる．また，2 つの同等性片側仮説に別々の有意水準を与えると，同等性マージンの解釈が困難になる．以上の理由から，CTTP に基づいて TTTP を提案したほうが良いと判断した．

第三章

傾向スコア重み付け法と検定併合法を組み合わせた 2 段階法の提案

3.1 緒言

第二章で紹介した研究は，アウトカムの要約値さえ分かっていれば利用できる．したがって，文献のみの情報を既存対照データとして利用できる点が利点である．しかしながら，もし既存対照の 1 人 1 人のデータ（個人患者データ）が利用可能であれば，より併合受容性を高められる可能性がある．

患者個人データが利用可能な場合，新規試験と既存対照の患者背景の違いを解析で考慮できる．観察研究では，比較群間の患者背景の違いは，交絡によるバイアスを生じさせるため，適切な方法を用いて患者背景の違いを調整する必要がある．ハイブリッド対照デザインにおいても，新規試験と既存対照の患者背景の違いは，交絡によるバイアスを生じさせる原因となる．そのため，Pocock (1976)は，併合受容性を高める条件の 1 つに，新規試験と既存対照の患者背景の類似を挙げていた．この条件が満たされていなくとも，患者背景の調整を行うことで，併合受容性を高めると考えられる（Han et al., 2017; Psioda et al., 2018）．患者個人データが利用可能であるならば，統計的な手法を用いて疑似的に新規試験と既存対照の患者背景を類似させた状態で試験治療と対照治療の比較が可能となる．

傾向スコア（Rosenbaum and Rubin, 1983）を用いた患者背景の調整は，観察研究における交絡の調整法としてしばしば利用される．傾向スコアは，観察研究の文脈においては，ある治療を受ける確率として定義される．傾向スコアを用いた交絡の調整法には，主にマッチング（PSM, propensity score matching），重み付け（PSW, propensity score weighting; Sato and Matsuyama, 2003），層別，共変量調整の 4 つがある（Austin, 2011）．傾向スコアは，ハイブリッド対照デザインにおいても，既存対照と新規試験の患者背景の違いを調整する方法として利用できる．

ハイブリッド対照デザインの文脈において，傾向スコア法で患者背景を調整と，適当な情報借用法を組み合わせる方法を，2 段階法と呼ぶ（Wang et al., 2022）．傾向スコア調整法と情報借用法の組合せは自由であり，様々な 2 段階法を考えることができる．Zhao et al. (2016)は傾向スコアマッチングと commensurate prior を組み合わせた方法，Wang et al. (2019)は傾向スコアを用いた層別と power prior を組み合わせた方法，Lin et al. (2018)は傾向スコアマッチングと power prior を組み合わせた方法をそれぞれ提案し

た．Wang et al. (2022)は，複数の傾向スコア調整法（マッチング，重み付け，層別）と複数の情報借用法（power prior と commensurate prior）を組み合わせた方法のシミュレーション研究を行い，マッチング又は重み付けと commensurate prior を組み合わせた方法が，リスク（第一種の過誤確率の上昇や検出力の低下）とリターン（検出力の上昇）のバランスが良い方法であると結論付けた．しかしながら，既存の提案手法はすべてベイズ流の情報借用法を利用したものであり，頻度論に基づく 2 段階法は提案されていない．

第三章では，生存時間アウトカムを想定し，傾向スコア重み付け法と検定併合法を組み合わせた 2 段階法を提案する．本章の構成は以下の通りである．3.2 節で傾向スコアを紹介する．3.3 節で傾向スコア重み付け法と検定併合法を組み合わせた 2 段階法を提案する．3.4 節で数値実験を行い，提案法の有用性を示す．3.5 節で考察を記す．

3.2 傾向スコア

交絡によるバイアスは，グループ間の患者背景の偏りによって引き起こされる．交絡を引き起こす因子の組を交絡因子と呼び，Xで表すことにする．交絡によるバイアスを取り除く最も本質的な方法は層別解析に基づく標準化である（佐藤・松山, 2011）．層に分けることで，各層では交絡がないと考えられるため，層ごとの解析結果を適切に統合することで，交絡によるバイアスを排除した推定値が得られる．層別解析に基づく標準化は，交絡因子の数が多い場合に層ごとの人数が少なくなり推定が不安定になる（特に交絡因子が連続変数の場合）という欠点がある．層別解析の代わりに良く用いられる方法の 1 つが，回帰モデルに基づく標準化である．しかし，回帰モデルには多くの仮定を必要とし，それらが正しい保証はない．

実は，交絡を調整する際，層別解析において各層内で交絡因子が均一である必要はなく，ある種のバランスとれていれば十分である（Rosenbaum and Rubin, 1983; 丹後・松井, 2018）．Rosenbaum and Rubin (1983)は，次の式を満たすスコア$b(X)$をバランススコアとして定義した．

$$X \coprod Z \mid b(X) \tag{3.1}$$

ここでZは治療変数を表す．この式は，交絡要因Xを与えた下で，Xの条件付き分布が治療に依らず同じであることを意味する．すなわち，式(3.1)を満たす$b(X)$が見つかりさえすれば，$b(X)$を用いて条件付けることで交絡を排除できる．式(3.1)を満たすバランススコアは複数考えられる．最も精緻なバランススコアがXそのものであり，最も粗いバランススコアが，傾向スコアと呼ばれる，治療を受ける条件付き確率

$$e(X) = P(Z = 1|X) \tag{3.2}$$

である．

傾向スコアは多くの場合において未知であるが，無作為化試験においてのみ真値が

分かり，その値は割付確率である．無作為化試験では，個々の患者のXにかかわらず全員に同じ既知の割付確率を与えるため，未知の交絡因子を含めて調整が可能である．一方で観察研究では傾向スコアは未知であるため，データから推定する必要がある．ロジスティック回帰を用いて推定されることが多いが，より柔軟に機械学習の手法を用いて推定することもできる（Lee et al., 2010）．

Rousenbaum and Roubin (1983)は，傾向スコアを用いて因果効果を推定できるための条件として「強く無視できる割り当て（strongly ignorable treatment assignment）」を定義した．この条件は次の 2 つの条件が成り立つときに成立する．

$$(Y(1), Y(0)) \coprod Z \mid X \tag{3.3}$$

$$0 < P(Z = 1|X) < 1 \tag{3.4}$$

ここで，$Y(1)$，$Y(0)$は潜在結果変数を表す．式(3.3)は，交絡因子で条件付けると，治療割付と観察された潜在結果変数が独立となることを意味しており，式(3.4)は，どちらかの治療に必ず割り付けられる被験者はいないことを意味する．式(3.3)は，未測定の交絡因子がないと言い換えられる．これらが成り立つならば，傾向スコアで条件付けた推定量は不偏性を持つ.

ハイブリッド対照デザインの文脈においては，式(3.2)に基づく定義のほかに，既存対照データと新規試験のうち，患者が新規試験に属する確率として定義することもできる．すなわち，新規試験参加を表す指示変数をQと表すと，

$$p(X) = P(Q = 1|X) \tag{3.5}$$

と定義することもできる．ハイブリッド対照デザインではなく，新規試験が単群試験であれば, 標本は新規試験治療である$(Z = 1, Q = 1)$と既存対照である$(Z = 0, Q = 0)$の 2 つであるため，$e(X)$と$p(X)$に違いはない．しかし，ハイブリッド対照デザインでは，さらに新規対照である$(Z = 0, Q = 1)$が追加される．$(Z = 1)$と$(Z = 0)$の患者背景の差異を調整するか，$(Q = 1)$と$(Q = 0)$の患者背景の差異を調整するかで，式(3.2)と式(3.5)の 2 通りが考えられるのである．$(Z = 1, Q = 1)$と$(Z = 0, Q = 1)$は，無作為化されているため，患者背景の分布は期待的に同一である．したがって，$e(X)$と$p(X)$のどちらの定義を用いたとしても交絡の調整は可能となる.

ハイブリッド対照デザインの先行研究では，どちらの定義も用いられているが，元々の定義に従った式(3.2)による研究が多い．しかし，本研究では，式(3.5)による定義を採用したい．ハイブリッド対照デザインにおいて興味のある集団（ターゲット集団）は新規試験の集団であると考えられる．式(3.5)の定義において， ATT (average treatment effect of the treated) weight を推定することで，新規試験の集団をターゲット集団とした解析が可能となる．また，$e(X)$は，新規試験参加者について，治療割付確率が既知であるにも関わらず，新たに割付確率を推定している点が懸念事項である.

$p(X)$を推定する際，$(Z = 1, Q = 1)$を含めてすべてのデータを用いればよい（Li and

Jemielita, 2023)．$(Z = 0, Q = 0)$と$(Z = 1, Q = 0)$のみから推定する方法も考えられるが，$(Z = 1, Q = 1)$を含めたとしても$p(X)$の期待値は変わらない．$p(X)$の推定精度の観点から$(Z = 1, Q = 1)$を用いたほうが良い．

傾向スコアを用いた交絡の調整法には，主にマッチング（PSM, propensity score matching），重み付け（PSW, propensity score weighting），層別，共変量調整の 4 つがある（Austin, 2011）．PSM は，傾向スコアが同程度の被験者をマッチングする．PSW は，傾向スコアから構成した重みをつけて解析を行う．層別は，傾向スコアの値によって適当な数の層に分け，層別解析を行う．共変量調整は，回帰モデルの説明変数に傾向スコアを加えた解析を行う．

PSM，PSW，層別に共通する利点は，研究デザインと解析を分離できる点である（Austin, 2011）．これら 3 つの方法は，傾向スコアを正しく推定できてさえいれば，マッチング，重み付け，層別された標本に対して任意の解析を，傾向スコアと目的変数の関係を指定することなく適用できる．この利点は，ハイブリッド対照デザインにおいても魅力的である（Wang et al., 2022）．

多くのシミュレーション研究から，層別と共変量調整よりも，PSM と PSM の方が，患者背景の調整に関して性能が良いことが分かっている（Austin, 2011）．PSM と PSW はどちらにも利点と欠点があり，一長一短である．この 2 つの性能は同程度であることが多いが，一部の状況では重み付けよりもマッチングの性能が良いとの指摘がある（Austin, 2009）．臨床家にとっては，マッチングの方が直感的に分かりやすいためか，PSW よりも PSM の方が多くの使用事例がある．しかし，PSM は，マッチされなかったデータは解析に使用されない，マッチングのアルゴリズムとその設定に恣意性が入りやすい，ターゲット集団が不明瞭という懸念事項もある．また，ハイブリッド対照デザインにおいては，既存対照のサンプルサイズが新規対照よりも小さいと適用できない．一方で PSW は，すべてのデータを利用でき，重みの構成方法を工夫することでターゲット集団を容易に変更できる．既存対照のサンプルサイズが小さくとも，適用は可能である．しかし，PSW は，傾向スコアの推定精度に敏感であるという指摘もある（Rubin, 2004）．本研究では，PSW を用いたハイブリッド対照デザインに注目する．

3.2.1 傾向スコア重み付け法

傾向スコアから計算された適当な重みを用いてデータを重み付けすることで，患者背景の類似した疑似集団を作成可能である．重みの計算方法はターゲット集団に応じていくつか提案されている．もっとも一般的な重みは，傾向スコアの逆確率を用いた次の形で表される重みである．

$$\frac{1}{p(X_i)}Q_i + \frac{1}{1-p(X_i)}(1-Q_i) \tag{3.6}$$

ここで添え字iは患者を表す．この重みを用いた重み付け法はIPW（inverse probability weighting）と呼ばる．IPWのターゲット集団は，集団全体である．すなわち，観察研究の文脈では，集団全体が曝露を受けた場合と受けなかった場合の比較を行う．ハイブリッド対照デザインにおいて式(3.6)の重みを用いると，患者背景を調整する集団と比較する治療集団が同一ではないため，ターゲット集団の解釈が難しい．

IPWの重みの亜種として，次の重みを用いることもできる．

$$Q_i + \frac{p(X_i)}{1-p(X_i)}(1-Q_i) \tag{3.7}$$

この重みを用いると，観察研究の文脈では，曝露を受けた患者集団に対する平均因果効果（ATT, average treatment effect for treated; $E[Y(1)-Y(0)|Z=1]$）が推定対象となる．すなわち，曝露を受けなかった集団が曝露を受けた場合の平均因果効果が推定対象となる．ハイブリッド対照デザインでは，既存対照の患者が新規試験に参加していた場合の治療効果を推定することに対応する．ハイブリッド対照デザインにおいて，興味のあるターゲット集団は新規試験の集団である．したがって，式(3.7)に基づく重みが適しており，本研究ではこちらを採用する．

傾向スコア重み付け法の問題点の1つに，重みが極端に大きな値をとる患者が存在すると，その患者に推定が強く依存してしまい，推定が不安定になる恐れがある点があげられる．この問題を解決する方法として，極端な重みをもつ患者を除外する方法（trimmed weight または truncated weight; Lee et al., 2011）と安定化重みを用いる方法（stabilized weight; Cole and Hernán, 2008）が有名である．Trimmed weightは，適当な閾値（例えば1%点と99%点）を設定し，閾値を超えた重みをもつ患者を解析から除外する．安定化重みでは，次のように各グループの人数の割合を各項に乗じることで，極端な重みの影響を緩和する．

$$Q_i \times P(Q=1) + \frac{p(X_i)}{1-p(X_i)}(1-Q_i) \times P(Q=0) \tag{3.8}$$

trimmed weight は閾値の設定に恣意性が混入するが，安定化重みはそのようなことはない．また，サンプルサイズが小さくならない点も魅力的である．本研究では，数値実験において，安定化重みを使用しない方法に加えて，式(3.8)の安定化重みを利用した方法も検討する．

傾向スコアで重み付けた解析を行う際の推定量の分散には注意が必要である．傾向スコアが真の値ではなく推定値であることから，通常の重み付き回帰による信頼区間は誤ったものとなってしまう（佐藤・松山, 2011）．分散の推定方法として，ロバストサンドイッチ分散（Lin and Wei, 1989）を用いる方法と，ブートストラップサンプルを用いる方法が考えられる．ブートストラップサンプル法が最も良い方法であるという報告（Austin, 2016）があるが，本研究では実装の簡便さからロバストサンドイッチ分

散を用いる．

3.3 傾向スコア重み付け法と検定併合法を組み合わせた 2 段階法

傾向スコア法の利点の 1 つは，患者背景の調整と解析を分離できる点にあった．この特性を利用し，傾向スコア法と 1.3 節で紹介した既存対照借用法を組み合わせる 2 段階法がいくつか提案されている．2 段階法では，1 段階目に傾向スコア法を用いて既存対照の患者背景を新規対照に類似させた疑似集団を作成し，2 段階目に既存対照借用法を適用する．先行研究では，2 段階目の既存対照借用法として，commensurate prior を用いた方法と power prior を用いた方法が提案されている（Zhao et al., 2016; Lin et al., 2018; Wang et al., 2019; Wang et al., 2022）が，その他の既存対照借用法を適用することも理論的に可能である．

本研究では，PSW と検定併合法を組み合わせた方法（propensity score weighting + test-then-pool method; PSWTTP）を提案する．1 段階目として，PSW により，過去試験の共変量分布を新規試験に調整した疑似既存対照 PHC（pseudo historical control）を得る．この時，重みは式(3.7)に基づく重みを用いる．PHC に対する新規試験の対照群の対数ハザード比をθ^*とすると，過去試験と新規試験の対照群の観察されたアウトカム分布の違いが，観察された共変量で説明できるのであれば，$\theta^* = 0$となるはずである．したがって，2 段階目として，次の仮説を有意水準γで検定する．

$$H_{0,PSWTTP}: \theta^* = 0 \; vs. H_{1,PSWTTP}: \theta^* \neq 0.$$

$H_{0,PSWTTP}$が棄却されなかった場合，新規対照と PHC を併合して新規治療と比較する．すなわち傾向スコア重み付け解析を行う．$H_{0,PSWTTP}$が棄却された場合，単に新規試験治療と新規対照を比較する．

1 段階目において，式(3.7)に基づく重みの代わりに，式(3.8)に基づく安定化重みを用いた方法（stabilized propensity score weighting + test-then-pool method; SPSWTTP）も合わせて提案する．SPSWTTP は，傾向スコアの推定が不安定となるサンプルサイズが小さい場合に，良い性能を示すと予想される．

3.4 数値実験

Wang ら (2022)の数値実験の設定を参考に，PSWTTP，SPSWTTP の性能を評価するための数値実験を行った．

3.4.1 設定

PAN-01 試験と GEST 試験を参考に，非劣性試験を想定した数値実験を行った．生存時間アウトカムを考え Cox の比例ハザードモデル（Cox, 1972）でハザード比を推定する．非劣性マージンを対数ハザード比としてlog (1.33)，組み入れ期間を 3 年，フォ

ローアップ期間を 2 年，片側有意水準を 5%とした．新規試験のサンプルサイズは，対照群の中央生存期間を 9.7 か月と仮定した下で，割付比率 2:1 の新規試験の検出力が，既存対照を考慮せずに 70%となるように，285人とした．この点は PAN-01 試験の実際のサンプルサイズとは異なる．また，既存対照のサンプルサイズn_Hは GEST 試験の 280 人に加えて 70，140，560 人を検討した．

アウトカムY_iの生成モデルは，患者iについて次の平均構造を持つパラメータλ_iの指数分布とした．

$$\log(E(Y_i)) = \log(\lambda_i) = \beta_0 + \xi Q_i + \delta Z_i + \beta_1 X_{1i} + \beta_2 X_{2i}.$$

ここで，Z_iは割付を表す変数である．既存対照のZ_iは 0 を与え，新規試験参加者のZ_iは生起確率 2/3 の Bernoulli 分布から生成した．δは治療効果である対数ハザード比を表すパラメータであり，帰無仮説が真の時は$\log(1.33)$，対立仮説が真の時は$\log(1)$とした．ξは time trend（時期効果や観測できない共変量の影響）を表すパラメータである．X_{1i}, X_{2i}はそれぞれ次の分布に従うとした．

$$X_{1i} \sim \text{Bernoulli}(\pi_C \cdot Q_i + 0.5 \cdot (1 - Q_i)),$$

$$X_{2i} \sim N\left(m_C \cdot Q_i + 6.5 \cdot (1 - Q_i), 0.5^2\right).$$

ここでπ_Cとm_Cの添え字Cは新規試験（current trial）を表す．X_{1i}とX_{2i}が生存時間に与える影響を表すβ_1とβ_2はそれぞれ，$\log(0.5)$と$\log(1.2)$に設定した．β_0は上記の設定の下で既存対照の中央生存期間が 9.7 か月となるように log(0.0303)とした．この値は Monte Carlo シミュレーションを用いて探索した．

π_Cとm_C，ξを調整することで，drift の発生メカニズムが異なる，次の 3 つのシナリオを検討した．ここでは drift を，既存対照と比較した新規試験の対照群のナイーブな対数ハザード比，すなわち患者背景の違いを考慮せずに計算される見かけの対数ハザード比と定義する．

(1) 未測定の共変量の分布のみの違い（time trend）

$$\pi_C = 0.5 \text{ and } m_C = 6.5,\ \xi \neq 0$$

(2) 観察される共変量の分布のみの違い（covariate distribution difference, CDD）

$$\pi_C \neq 0.5 \text{ and/or } m_C \neq 6.5,\ \xi = 0$$

(3) (1)と(2)の両方（CDD + time trend）

$$\pi_C \neq 0.5 \text{ and/or } m_C \neq 6.5,\ \xi \neq 0$$

Time trend は，共変量分布は共通であるが，未測定の因子や時代効果によって drift が生じる設定である．CDD では，アウトカム分布の違いがすべて共変量分布の違いに起因しており，傾向スコア重み付け法を用いた共変量調整でバイアスなく治療効果を推定できると期待される設定である．

Drift は，ハザード比のスケールで 1/1.4 から 1.4 の範囲を検討した．exp(drift)が 1 よ

りも小さい場合，新規対照の治療成績が既存対照よりも悪く，第一種の過誤確率の上昇が懸念される．逆に exp(drift)が 1 よりも大きい場合，新規対照の治療成績が既存対照よりも悪く，検出力の低下が懸念される．今回の数値実験で用いたパラメータの値を表 3.1 に示す．各 drift を満たすパラメータπ_C，m_C，ξについて，次の手順で適当な値を探索した．CDD では，$\xi = 0$とし，π_Cを 0.05 刻みで変化させたもとで，exp(drift)にもっとも近くなるm_Cを探索した．Time trend では，$\pi_C = 0.5$，$m_C = 6.5$としたもとで，exp(drift)になるようにξを指定した．CDD + time trend では，ξを time trend のものの半分とし，π_Cを 0.025 刻みで変化させたもとで，exp(drift)にもっとも近くなるm_Cを探索した．なお，この時 CDD と CDD＋time trend では Monte Carlo シミュレーションを利用した．例えば，CDD の exp(drift)が 1.10 となるシナリオの各パラメータは，まず$\xi = 0$，$\pi_C = 0.4$を与えたもとで，多数の乱数を発生させ，新規対照と既存対照の見かけのハザード比（exp(drift)）が 1.1 になるm_Cの値として，6.70 を見つけた．

比較した方法は，検定併合法（TTP），傾向スコア重み付け法（PSW），安定化傾向スコア重み付け法（SPSW），傾向スコア重み付け法 + 検定併合法（PSWTTP），安定化傾向スコア重み付け法 + 検定併合法（SPSWTTP），既存対照を用いない方法（no borrowing）の 6 つとした．各方法について，バイアス，検出力，第一種の過誤確率，MSE（mean squared error）の比（no borrowing の MSE を分母とした比），併合確率の 5 つの動作特性を評価した．検定併合法の有意水準γは 0.1，0.2，0.3，0.4 を検討した．傾向スコアはロジスティック回帰を用いて推定した．動作特性推定のための繰り返し回数は 10000 回とした．

表 3.1　各シナリオのパラメータ設定一覧

	CDD			Time Trend			CDD + Time Trend		
exp (drift)	π_C	m_C	ξ	π_C	m_C	ξ	π_C	m_C	ξ
0.71	0.90	5.86	0	0.50	6.50	0.336	0.700	6.12	0.168
0.74	0.85	5.90	0	0.50	6.50	0.300	0.675	6.14	0.150
0.77	0.80	5.96	0	0.50	6.50	0.262	0.650	6.18	0.131
0.80	0.75	6.04	0	0.50	6.50	0.223	0.625	6.22	0.112
0.83	0.70	6.10	0	0.50	6.50	0.182	0.600	6.28	0.091
0.87	0.65	6.18	0	0.50	6.50	0.140	0.575	6.30	0.070
0.91	0.60	6.28	0	0.50	6.50	0.095	0.550	6.36	0.048
0.95	0.55	6.38	0	0.50	6.50	0.049	0.525	6.42	0.024
1.00	0.50	6.50	0	0.50	6.50	0.000	0.500	6.50	0.000
1.05	0.45	6.62	0	0.50	6.50	-0.049	0.475	6.56	-0.024
1.10	0.40	6.70	0	0.50	6.50	-0.095	0.450	6.64	-0.048
1.15	0.35	6.78	0	0.50	6.50	-0.140	0.425	6.68	-0.070
1.20	0.30	6.82	0	0.50	6.50	-0.182	0.400	6.72	-0.091
1.25	0.25	6.86	0	0.50	6.50	-0.223	0.375	6.76	-0.112
1.30	0.20	6.86	0	0.50	6.50	-0.262	0.350	6.80	-0.131
1.35	0.15	6.86	0	0.50	6.50	-0.300	0.325	6.82	-0.150
1.40	0.10	6.84	0	0.50	6.50	-0.336	0.300	6.84	-0.168

π_C：二値変数X_1の生起確率

m_C：連続変数X_2の母平均

ξ：時代効果を表すパラメータ

3.4.2 結果

本項には，$\gamma = 0.2$の結果を示す（図 3.1 から図 3.4）．これらの図は，見やすさを考慮して，結果の一部見切れてしまうものの，縦軸の縮尺を拡大調整した．縮尺未調整の図は，付録の図 A2.1 から図 A2.4 に示す．また，$\gamma = 0.1, 0.3, 0.4$の結果は，付録の図 A2.5 から図 A2.16 に示す.

3.4.2.1 Time trend の結果

まず time trend のシナリオ（1 列目）に注目する．このシナリオでは，drift が未測定の因子のみによって生じるため，共変量調整はバイアスの抑制に意味を持たない．したがって，傾向スコア重み付け法のみを行う PSW と SPSW では，drift が 0 ではないときに，その大きさに比例してバイアスが大きくなっている．それに伴って，drift が正の時は検出力が低下，負の時は第一種の過誤確率が上昇している．一方で検定併合法を用いた方法（TTP，PSWTTP，SPSWTTP）では，drift が 0 ではない場合のバイアスが完全に 0 になるわけではないものの，PSW，SPSW と比較するとバイアスは小さく，drift が一定程度大きくなると低下に転じている．併合確率の結果から，drift が大きくなるにつれて既存対照データを併合する確率が低下しており，この特性がバイアスを低下させている. そのため, 第一種の過誤確率はどこまでも上昇することはなく，十分に既存対照データのサンプルサイズn_Hが大きければ，検出力は no borrowing と比較して低下することはない．Time trend のシナリオでは，検定併合法単体（TTP）と傾向スコア重み付け法を組み合わせた方法（PSWTTP と SPSWTTP）で動作特性に大きな違いはない.

3.4.2.2 CDD の結果

次に CDD のシナリオ（2 列目）を説明する．このシナリオでは，drift が観察されている因子ですべて説明できる．そのため， PSW と SPSW では，n_Hが十分に大きければ，第一種の過誤確率の上昇は起こらず，検出力も全体的に検定併合法を用いた場合よりも高くなる．PSWTTP と SPSWTTP では，drift が 0 の場合の検出力は TTP と同じであるものの，drift が 0 ではない場合の検出力は TTP よりも高い．併合確率は，どのような drift の大きさでも，理論的な検定併合法の併合確率に近い値である$1 - \gamma$にほぼ一致している.

PSW を用いる方法について，n_Hが小さいときには注意が必要である．図 3.1 では，drift が負の時に，傾向スコア重み付け法を用いた方法は検出力が低下している．特に安定化重みを用いない PSW と PSWTTP で顕著である．これは，表 3.1 に示した設定では，drift が負の方向に大きい場合に患者背景の分布の違いが大きく，n_Hが小さいと傾向スコアの推定が不安定になるからと考えられる. 傾向スコアの推定が不安定だと，

重みが極端な値に推定されてしまい，特定の患者に治療効果の推定が強く影響を受けてしまうため，治療効果の推定も不安定になってしまう．極端な大きさの重みに対処法の 1 つである安定化重みを用いた SPSW と SPSWTTP では，検出力低下の程度は幾分緩和されている．n_Hが大きい場合では安定化重みを用いない方法と用いた方法の動作特性に違いは表れないことから，安定化重みの利用に動作特性上の欠点はない．したがって，安定化重みの利用が推奨される．

3.4.2.3　CDD + time trend の結果

3 つ目のシナリオとして，CDD + time trend（3 列目）を説明する．このシナリオは，drift の半分が未測定の因子によって引き起こされ，もう半分が観察されている因子で説明できるシナリオである．このシナリオでは，傾向スコアを用いた方法の動作特性は, time trend のみの動作特性と CDD のみの動作特性の中間的なものになる. 例えば, PSW のバイアスの drift に対する傾きは time trend のみの場合と比較して小さくなっている．また，PSW TTP の第一種の過誤確率は，CDD のみの時のように上昇しないわけではないものの, その最高値は, time trend の場合と比較して低い. 加えて, SPSWTTP の検出力は，drift が 0 の時に TTP と同一となり，drift に対して全体的に TTP よりも高くなるという性質は CDD の場合と変わらない．さらに，PSWTTP の併合確率は，drift に対して$1-\gamma$で一定ではないが，全体的に TTP よりは高い．CDD と time trend が混ざったシナリオであっても，患者背景の違いで説明できる drift を傾向スコア重み付け法によって調整できるため，このような結果が得られたと考えられる．

3.4.2.3　各シナリオに共通して確認された結果

最後に，シナリオ共通で確認された結果に言及する．図 3.1 から図 3.4 より，n_Hが大きくなるにつれて検出力が高くなる一方で，検定併合法を用いた方法の最高第一種の過誤確率も高くなる．既存対照データのサンプルサイズが大きいと，drift が存在する場合の負の影響が増大されてしまう. また, 図 3.1 から図 3.4 と図 A2.5 から図 A2.16 より，既存対照の併合されやすさを調整するチューニングパラメータであるγが高くなるほど，検出力と第一種の過誤確率の低下の程度が小さくなる．γの設定は，既存対照のリスクとベネフィットのトレードオフである．γの決定方法の 1 つに，第二章で提案した動作特性に基づいて選択する方法が考えられる．

検定併合法を用いた 4 つの方法は，drift が 0 であったとしても no borrowing と比較して第一種の過誤確率が上昇している．この結果は，Li ら(2020)の指摘通りである．TTP では, drift が 0 の時に第一種の過誤確率がおよそ 10%になっている. PSW と SPSW では，CDD のシナリオにおいて，drift の大きさにかかわらず 10%であった．今回の数値実験から，n_Hが大きくなるほど，第一種の過誤確率の上昇の程度が大きくなること

が分かった.

検定併合法の第一種の過誤確率の上昇とは別に，no borrowing においても第一種の過誤確率の上昇が確認された．治療効果に関する仮説の有意水準は片側 5%であるものの，推定された no borrowing の第一種の過誤確率はおよそ 7~8%であった．これは，ハザード比が collapsibility のない指標であるためと考えられる．Collapsibility とは，共変量で調整する前と後で効果の指標の値が変化しない性質である（Greenland et al., 1999）.

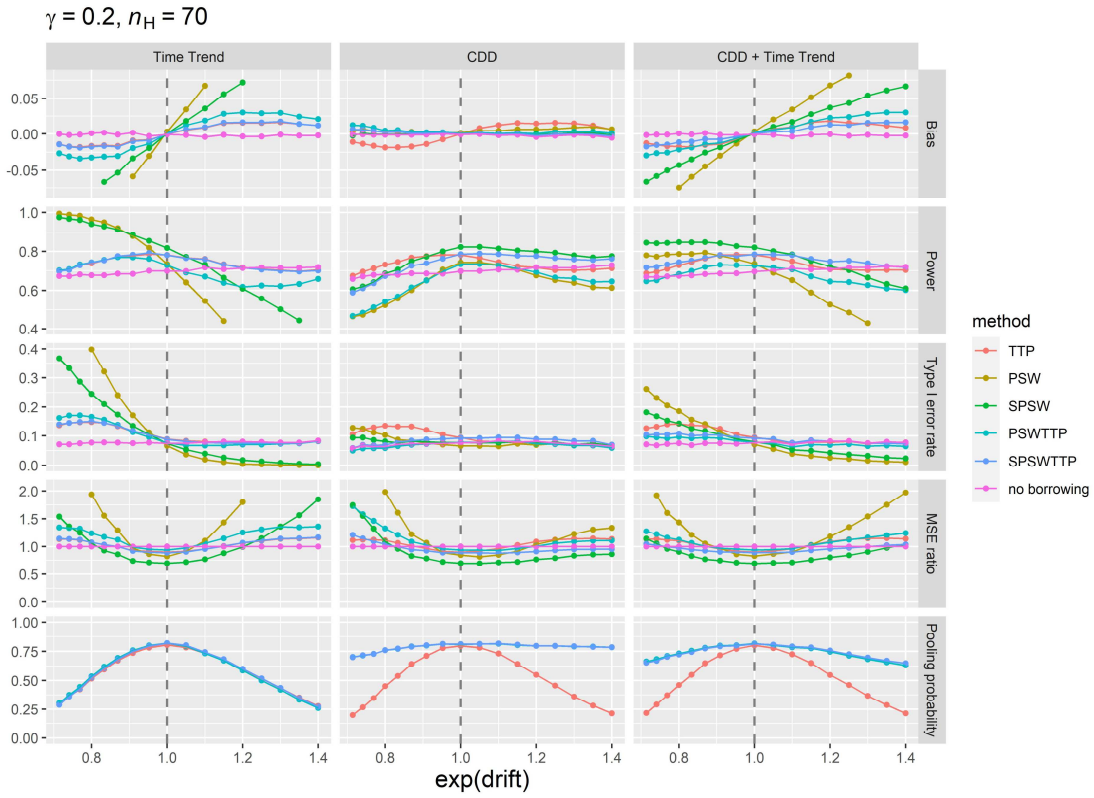

図 3.1 推定された動作特性（$\gamma = 0.2$，$n_H = 70$，縦軸縮尺調整）

γ：検定併合法の有意水準

n_H：既存対照のサンプルサイズ

TTP：検定併合法

PSW：傾向スコア重み付け法

SPSW：安定化重みを用いた傾向スコア重み付け法

PSWTTP：傾向スコア重み付け法＋併合検定法

SPSWTTP：安定化重みを用いた傾向スコア重み付け法＋併合検定法

no borrowing：既存対照を用いない方法

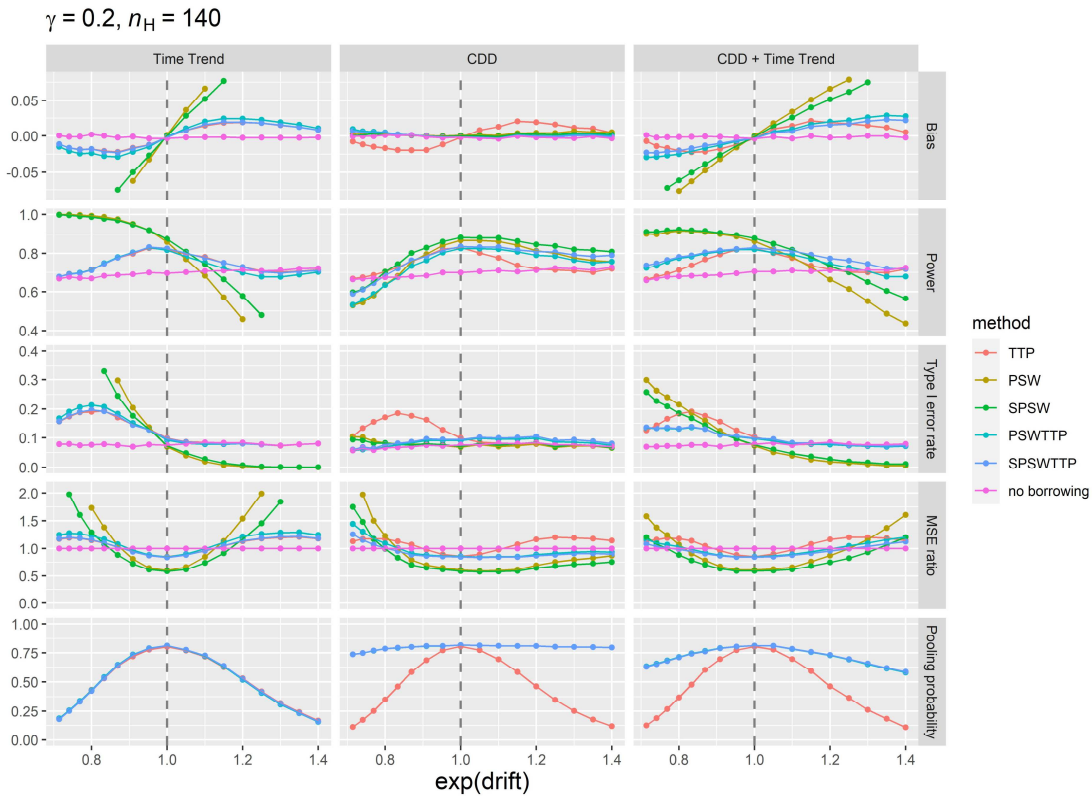

図 3.2 推定された動作特性（$\gamma = 0.2$，$n_H = 140$，縦軸縮尺調整）

γ：検定併合法の有意水準

n_H：既存対照のサンプルサイズ

TTP：検定併合法

PSW：傾向スコア重み付け法

SPSW：安定化重みを用いた傾向スコア重み付け法

PSWTTP：傾向スコア重み付け法＋併合検定法

SPSWTTP：安定化重みを用いた傾向スコア重み付け法＋併合検定法

no borrowing：既存対照を用いない方法

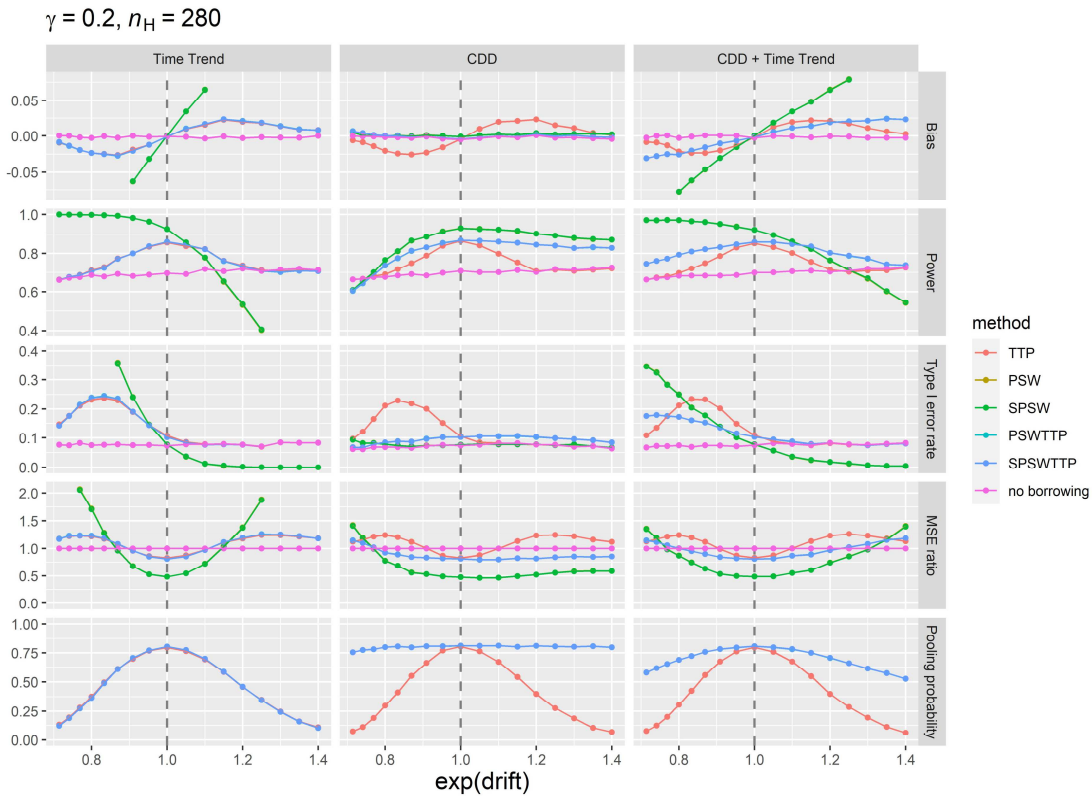

図 3.3. 推定された動作特性（$\gamma = 0.2$，$n_H = 280$，縦軸縮尺調整）

γ：検定併合法の有意水準

n_H：既存対照のサンプルサイズ

TTP：検定併合法

PSW：傾向スコア重み付け法

SPSW：安定化重みを用いた傾向スコア重み付け法

PSWTTP：傾向スコア重み付け法＋併合検定法

SPSWTTP：安定化重みを用いた傾向スコア重み付け法＋併合検定法

no borrowing：既存対照を用いない方法

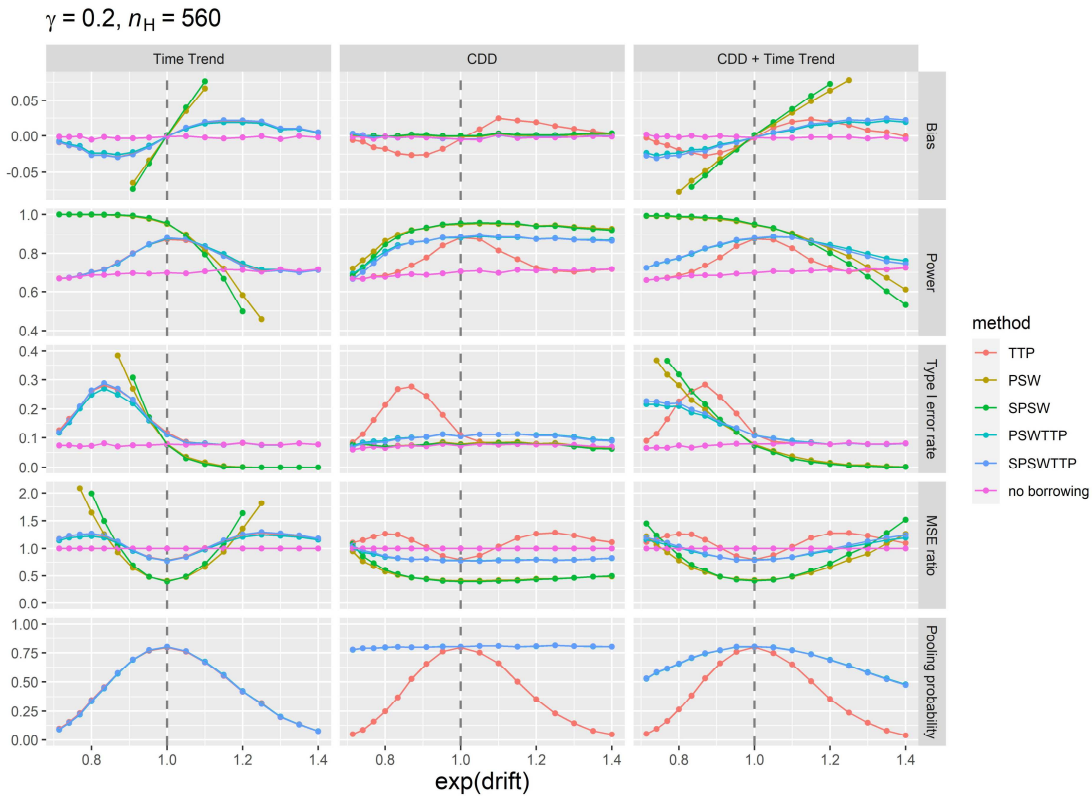

図 3.4 推定された動作特性（$\gamma = 0.2$，$n_H = 560$，縦軸縮尺調整）

γ：検定併合法の有意水準

n_H：既存対照のサンプルサイズ

TTP：検定併合法

PSW：傾向スコア重み付け法

SPSW：安定化重みを用いた傾向スコア重み付け法

PSWTTP：傾向スコア重み付け法＋併合検定法

SPSWTTP：安定化重みを用いた傾向スコア重み付け法＋併合検定法

no borrowing：既存対照を用いない方法

3.5 考察

本章では，観察研究において交絡を調整するための方法の 1 つである傾向スコア法を紹介した．また，傾向スコア重み付け法をハイブリッド対照デザインに応用し，検定併合法を組み合わせた 2 段階法を提案した．さらに，数値実験により提案法の動作特性を評価した．数値実験の結果，傾向スコア法は，患者背景の分布が異なっていたとしても，バイアスの小さな推定が可能であることが分かった．また，未測定の因子によって drift が生じる場合，検定併合法を組み合わせることで，バイアスを軽減できることが分かった．

3.4 節で検討した設定においては，傾向スコア重み付け法と検定併合法を組み合わせた 2 段階法（PSWTTP と SPSWTTP）は，検定併合法のみと比較して大きな欠点はない．サンプルサイズが小さく，drift が大きいときに検出力低下する恐れがあるが，安定化重みを利用する 2 段階法（SPSWTTP）である程度改善可能である．Drift が負の方向に大きいと，検定併合法と比較して第一種の過誤確率が上昇することがあるが，最大第一種の過誤確率の観点からは，検定併合法単体よりも良い．3.4 節で検討した設定の中では，多くの状況において，2 段階法の性能は，検定併合法と同じか上回っている．ハイブリッド対照デザインにおいて，傾向スコア重み付け法＋検定併合検定併合法の 2 段階法を利用する場合，安定化重みを用いた傾向スコア重み付け法を組み合わせることが推奨される．ただし，3.4 節の検討では，傾向スコアを推定するための因子の数は 2 つしか考慮していない．より多くの因子が存在する場合，提案法の性能に影響を与える恐れがある．より詳細な検討は今後の課題の 1 つである．

傾向スコア重み付け法と検定併合法を組み合わせた 2 段階法は，実装が簡便な点も魅力的である．ハイブリッド対照デザインにおけるマッチングはその手順が複雑となるため，ある程度の解析スキルが要求される．ベイズ流の方法は，事後分布の推定が解析的に行えず，適当な方法（マルコフ連鎖モンテカルロ法が良く用いられる）を用いて事後分布を近似する必要があるため，統計の専門家でなければ解析は難しいだろう．一方，PSWTTP と SPSWTTP は，一般的な統計ソフトウェアに実装されている機能の組み合わせで容易に適用できる．

本研究では，傾向スコア調整法として重み付けを利用したが，その他の調整方法も考えられる．特にマッチングは重み付けとともに，シミュレーション研究において良好な結果が得られており，ハイブリッド対照デザインにおいてもよい性能をもたらす可能性がある．重み付けを用いた方法とその他の調整法を用いた場合の比較は今後の課題の 1 つである．

2 段階目の既存対照借用法について，ベイズ流の方法との比較も検討課題の 1 つである．検定併合法とベイズ流の方法は，併合の考え方が異なる．前者は既存対照データをすべて使うことを念頭に置いているが，既存対照と新規対照の類似度が低い場合

は利用をあきらめる．ベイズ流の方法は，どの手法も基本的に類似度の程度に応じて利用する既存対照の情報量を制御する．したがって，手法の選択は，個々の手法の動作特性だけではなく，既存対照利用の目的や，結果の解釈性を考慮して選択されるべきである．しかしながら，対等な条件下で検定併合法とベイズ流の方法の動作特性を比較しておくことは，手法選択の一助になると考えらえれる．

本章の数値実験では，Li らの提案した有意水準の調整を行わなかった．結果として，drift が 0 の時であっても，検定併合法の第一種の過誤確率が名目水準よりも上昇していた．Li らの有意水準調整法を生存時間アウトカムに適用する場合，シミュレーションによる推定が必要となる．実際の臨床試験で検定併合法を利用する際は，drift を 0 と仮定した際の有意水準の調整を行うことが望ましい．

また，本章の数値実験では，no borrowing において，ハザード比の non-collapsibility に由来する第一種の過誤確率の上昇が確認された．数値実験の設定として，治療効果の指標は共変量で調整した対数ハザード比であるδを考えている．一方，no borrowing の解析では，推定対象が共変量で調整しない対数ハザード比である．ハザード比は collapsibility がないため，δと共変量で調整しない対数ハザード比に違いが生じてしまい，数値実験において第一種の過誤確率が名目水準に一致しなかった．治療効果が 0 であれば collapsibility の問題は生じないため，優越性試験を想定した数値実験では第一種の過誤確率の上昇は起こらない．本研究の数値実験の目的は，no borrowing を基準とした検定併合法や傾向スコア法の性能評価であったため，参考とした事例を尊重し非劣性試験で数値実験行い，non-collapsibility による第一種の過誤確率の上昇は許容した．

第四章

全体の考察

本研究では，第一章でハイブリッド対照デザインの利点と欠点を説明したうえで，既存の手法を紹介した．その中でも特に検定併合法に注目し，併合検定法の課題をまとめた．第二章では，検定併合法が第一種の過誤確率と検出力の制御に関して柔軟性が乏しいという点を改善するため，検定併合法の併合基準である両側検定を，2 つの片側検定に分割する検定併合法を提案した．それぞれの仮説に異なる片側有意水準を与えることで第一種の過誤確率と検出力の制御に関する柔軟性を獲得できると考えた．また，併合基準の検定に関する有意水準を選択する方法の 1 つとして，既存対照の新規対照のあらゆる差異を考慮した下で，最高第一種の過誤確率と最低検出力を任意の値に制御することを目的とした有意水準の選択法を提案した．提案検定併合法と提案有意水準選択法を組み合わせることで，従来の検定併合法と比較して，既存対照の新規対照に差異がある場合の検出力の低下が起こらない，あるいは既存対照の新規対照に差異がない場合の検出力が向上することを，数値実験を通して示した．第三章では，従来の検定併合法が患者背景の差異を考慮していない点を解決するために，傾向スコア重み付け法を組み合わせた 2 段階法を提案した．数値実験を通して，安定化重みを利用した傾向スコア法と検定併合法を組み合わせる方法が，最も第一種の過誤確率の上昇と検出力の低下を防ぐことが示唆された．本章では，研究全体の考察を記す．

検定併合法には，本研究で提案したものを含めて，以下の工夫が提案されている．

- 同等性検定に基づいた検定併合法
- 既存対照と新規対照に差異がない場合における第一種の過誤確率を，名目水準に一致させるための，治療効果に関する仮説に対する有意水準の調整
- 併合判断のための検定を 2 つの片側検定に分割し，別々の有意水準を与える方法
- 併合判断のための検定の有意水準を動作特性に基づいて選択する方法
- 傾向スコア重み付け法を組み合わせた 2 段階法

前者 2 つは Li ら（2020）による提案であり，残りの 3 つは本研究の提案である．これらの工夫は，互いに対立するものではなく，自由に組み合わせることが可能である．3.5 節に記した通り，同等性検定に基づく検定併合法において，併合判断のための同等性検定を 2 つの片側仮説に分割できるため，1 つ目の工夫に 3 つ目の工夫を自然と適用できる．2 つ目の工夫について，本研究では提案手法の純粋な性能を評価するために適用しなかった．しかし，実際の臨床試験では，既存対照と新規対照の差異はない場合の第一種の過誤確率は，名目水準に一致していることが望ましく，適用したほう

が良いだろう．この工夫は，単に治療効果に関する仮説に対する有意水準を調整する方法であるため，その他の工夫に本質的な影響を与えるわけではない．4 つ目の工夫について，第二章では，既存対照の標本平均と新規対照の母平均の差を drift と考え，すべての drift に対して望ましい動作特性が得られるような有意水準を選択する方法を提案した．傾向スコアを用いて患者背景を調整する場合，測定された因子で調整しきれなかった既存対照と新規対照の差異が drift となる．4 つ目の工夫と傾向スコアを用いた患者背景の調整を組み合わせる場合，有意水準の選択に関して特に患者背景因子を考慮する必要はなく，患者背景因子を考慮しない場合と同様に，単に既存対照の標本平均と新規対照の母平均の差を drift と考えて有意水準を選択すればよい．

Li ら（2020）の提案した，治療効果に関する仮説の検定に対する有意水準の調整法には，改良の余地があると考える．Li ら（2020）の提案は，治療効果に関する仮説に対する検定に対して，既存対照の併合有無にかかわらず同一の調整有意水準α^*を使用するというものであった．しかし，既存対照が併合される場合と併合されない場合で，同一のα^*を使用する必要はない．検定併合法は，既存対照と新規対照の差異が大きいとき，既存対照を併合しない場合が多くなる．この時，Li ら（2020）の方法は，治療効果に関する仮説に対して不必要に厳しい有意水準を適用することになる．別の調整方法として，既存対照が併合される場合のみ，治療効果に関する仮説に対して調整した有意水準を使用し，既存対照が併合されない場合は，治療効果に関する仮説に対して調整しない有意水準を使用する方法も考えられる．こちらの方法は，既存対照と新規対照の差異が大きい場合における検出力の低下が小さいと考えられる．併合される場合のみ治療効果に関する仮説に対する有意水準を調整する方が，研究者の理解も得られやすいであろう．詳細な検討は今後の課題である．

本研究では，既存対照データが 1 つのみの場合を取り扱った．実際には，既存対照データが複数存在する場合も考えられる．筆者の知る限り，複数の既存対照データが存在する場合の検定併合法の発展はない．1 つ 1 つの既存対照データに対して検定併合法を適用する方法や，メタアナリシスの手法を用いて複数の既存対照データを 1 つのデータにまとめたうえで検定併合法を適用する方法が考えられる．検定併合法ではないが，all-or-nothing アプローチの 1 つとして，大東らは，2023 年度日本計量生物学会年会において，クラスタリングの手法を用いて複数の既存対照データから新規対照データと同じ分布に従うデータを選択する方法を提案している（大東他, 2023）．複数の既存対照データが存在する場合における，検定併合法的手法の発展が期待される．

本研究の提案は，アウトカムの型に依存せず適用可能である．本研究において，第二章では連続量アウトカム，第三章では生存時間アウトカムを対象に研究を行った．理論的には，第二章の提案法を生存時間アウトカムや二値アウトカムに，第三章の提案を連続量アウトカムや二値アウトカムに適用できる．しかしながら，異なるアウト

カムに提案法を適用した場合の利点の大きさは，本研究では検討できていない．本研究の提案を実際の臨床試験で利用する場合，臨床試験の計画段階において状況に合わせた数値実験を行い，試験デザインの動作特性を十分に評価すべきである．

検定併合法では，既存対照と新規対照の従う分布が同一であっても，第一種の過誤確率を名目水準に制御できない．その理由は，検定併合法が予備検定方式だからである．予備検定方式とは，予備的な検定を行い，その結果に応じて，興味のある仮説に対する検定方法を変える検定手順である．予備検定方式では，第一種の過誤確率を名目水準に制御することができないことが知られている（Bancroft, 1964）．ここでは，予備検定方式に伴う，検定併合法における第一種の過誤確率の上昇を，数式を用いて説明する. 以下では, 既存対照と新規対照の従う分布が同一である場合を考える. また, 治療効果に関する仮説に対する検定の有意水準をαと表す．検定併合法は，検定の結果に応じて，単純併合法と分離法（既存対照を併合せず新規試験データのみを用いて解析する方法）を使い分ける方法である．検定併合法における第一種の過誤確率の上昇は，一見すると奇妙な現象である．というのは，単純併合法の第一種の過誤確率（$T1E_p$とする）と分離法の第一種の過誤確率（$T1E_s$とする）は共にα以下であるにも関わらず, 既存対照と新規対照の類似度に応じて単純併合法と分離法を使い分けると，第一種の過誤確率（$T1E_{TTP}$とする）がαを超えてしまう場合があるからである. 式(2.8)を利用すると，$T1E_p$，$T1E_s$，$T1E_{TTP}$は，それぞれ,

$$T1E_p = \int_{-\infty}^{\infty}\int_{-\infty}^{-z_{1-\alpha}} f_p(t_p, u)\, dt_p du,$$

$$T1E_s = \int_{-\infty}^{\infty}\int_{-\infty}^{-z_{1-\alpha}} f_s(t_s, u)\, dt_s du,$$

$$\begin{aligned} T1E_{TTP} = &\int_{-\infty}^{-z_{1-\gamma/2}}\int_{-\infty}^{-z_{1-\alpha}} f_s(t_s, u)\, dt_s du \\ &+ \int_{-z_{1-\gamma/2}}^{z_{1-\gamma/2}}\int_{-\infty}^{-z_{1-\alpha}} f_p(t_p, u)\, dt_p du \\ &+ \int_{z_{1-\gamma/2}}^{\infty}\int_{-\infty}^{-z_{1-\alpha}} f_s(t_s, u)\, dt_s du \end{aligned}$$

と表すことができる．単純併合法と分離法は，uの値にかかわらず併合する，あるいは併合しないという方法であるから，uを$-\infty$から∞ の範囲で積分消去している．$T1E_p$と$T1E_s$は，図 1.1 に示したような帰無仮説の下での検定統計量の分布において，治療効果に関する仮説に対する検定の閾値（$-z_{1-\alpha}$）以下となる面積を表しているに過ぎない．$T1E_p$と$T1E_s$は，どちらもα以下であることが保証される．一方で$T1E_{TTP}$では，治療効果に関する仮説に対する検定の統計量と，既存対照の併合可否を判断する検定の統計量の二変量分布において，治療効果に関する仮説の検定統計量の実現値が閾値

以下となる部分の体積を考えている．単純併合法が選択された場合と，分離法が選択された場合の 2 つの二変量分布を考え，uの値に応じてそれぞれの二変量分布の体積を切り取り，足し合わせたものが$T1E_{TTP}$である．この$T1E_{TTP}$が，α以下であることは保証されていないのである．第三章の数値実験から，既存対照のサンプルサイズが大きくなると，第一種の過誤確率の上昇が大きくなることが確認された．既存対照が大きくなるほど，$T1E_{TTP}$の第 2 項の値が大きくなるため，全体としての第一種の過誤確率の上昇が，αを超えて大きくなると考えられる．逆に既存対照のサンプルサイズが小さいと，$T1E_{TTP}$の第 2 項の値が小さくなるため，第二章の数値実験で確認されるように，全体としての第一種の過誤確率がαよりも低くなることがある．

第五章

結論

本研究では，既存対照データを臨床試験に用いるための方法である検定併合法を拡張した．本研究から得られた新知見は以下の通りである．

第二章

・2 つの片側仮説を組み合わせた検定併合法を提案した．
・検定併合法の有意水準を，動作特性に基づいて選択する方法を提案した．
・提案検定併合法と提案有意水準選択法を組み合わせることで，検出力について，従来の検定併合法よりも望ましい特性が得られることを示した．

第三章

・傾向スコア重み付け法と検定併合法を組み合わせた 2 段階法を提案した．
・提案する 2 段階法は，患者背景の違いに基づくアウトカム分布の差異を調整することで，第一種の過誤確率の上昇と検出力の低下の程度を軽減することを確認した．

本研究の最も大きな意義は，既存対照と新規対照の純粋な類似度ではなく，新規試験の動作特性を考慮して，既存対照の利用を制御するアイデアを示した点にあると考える．既存対照と新規対照の類似度を評価する際，これらの差とその正負を考慮することで，差の絶対値のみを考慮する方法と比較して，検出力に関してより望ましい性質が得られることを示した．本研究では，検定併合法にこのアイデアを適用したが，その他の既存対照併合法に対しても適用可能であると考えられ，今後の研究課題の 1 つである．

本研究で提案した有意水準選択法は，検定併合法の有意水準選択に明確な根拠を与えるものであり，既存の検定併合法が有意水準を場当たり的に決めているという批判に対する 1 つの回答となる．しかし，提案した有意水準選択法は，第一種の過誤確率の最高値と検出力の最低値のみに注目しており，バイアスや既存対照の併合確率を考慮していない．よりよい有意水準選択法を開発する余地が残されており，今後の研究課題の 1 つである．

本研究で提案した傾向スコア重み付け法と検定併合法を組み合わせた 2 段階法は，患者背景の調整を通して既存対照データの併合受容性を向上させる．本研究では，傾

向スコア重み付け法と検定併合法を組み合わせた 2 段階法のみを検討した．今後は，傾向スコアマッチング法と検定併合法を組み合わせた 2 段階法の検討や，ベイズ流の既存対照併合法を組み合わせた 2 段階法との性能比較などを通して，提案した 2 段階法の相対的な立ち位置を評価する必要がある．

本研究の提案は，総じて検定併合法の利便性を向上させるものである．本研究の提案は，研究対象者数の制約による臨床試験の実施可能性が乏しい状況において，検定併合法を用いたハイブリッド対照デザインという選択肢を提供する．本研究は，臨床試験の実施可能性を向上させることを通して，医学研究の発展に寄与すると考えられる．

謝辞

本論文は，筆者が北海道大学大学院医学院博士課程に在学中に行った研究をまとめたものです．指導教員である北海道大学大学院医学研究院医学統計学教室の横田勲准教授には，本研究を通して，研究が正しい方向に進むように導いてくださっただけではなく，研究発表技法や論文作成技法についてもご指導，ご鞭撻を賜りました．

筆者は，博士課程在籍中に，北海道大学病院にて生物統計担当者として勤務しました．医療・ヘルスサイエンス研究開発機構の機構長である佐藤典宏教授には，業務をこなしながらの博士課程在籍をご許可いただいただけではなく，スムーズに学位取得ができるように最大限のご配慮をいただきました．

京都大学大学院医学研究科社会健康医学系専攻医療統計学の佐藤俊哉教授，土居正明准教授，臨床統計学の田中司朗特定教授，大宮將義特定助教，及び国立循環器病研究センターデータサイエンス部臨床統計室の大前勝弘室長には，筆者が専門職学位課程在学中にお世話になりました．先生方のご指導が，博士課程の研究を支える基礎となりました．

田中司朗特定教授には，専門職学位課程修了後も，web 会議での研究会を毎週開いていただきました．田中司朗特定教授，滋賀医科大学医学部附属病院臨床研究開発センターの松林潤特任助教，及び研究会にご参加の皆さまから，私の拙い研究アイデアに，たくさんのご意見，ご指摘，ご助言を頂戴しました．

北海道大学大学院医学院医学統計学教室の髙橋圭太氏は，専門職学位課程の後輩であり，最も身近な研究仲間として，何度も研究の相談に乗っていただきました．

本研究は，多くの方々に支えられてまとめることができました．この場を借りて心より感謝申し上げます．誠にありがとうございました．

利益相反

利益相反はない.

引用文献

Altman DG (1980) Statistics and ethics in medical research: III How large a sample?. British Medical Journal, 281(6251), 1336.

Austin PC (2009) Type I error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. The international journal of biostatistics, 5(1), 13.

Austin PC (2011) An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behavioral Research, 46(3), 399–424.

Austin PC (2016) Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. Statistics in Medicine, 35(30), 5642-5655.

Banbeta A, Van Rosmalen J, Dejardin D and Lesaffre E (2019) Modified power prior with multiple historical trials for binary endpoints. Statistics in Medicine, 38(7), 1147-1169.

Banbeta A, Lesaffre E and Van Rosmalen J (2022) The power prior with multiple historical controls for the linear regression model. Pharmaceutical Statistics, 21(2), 418-438.

Bancroft TA (1964) Analysis and inference for incompletely specified models involving the use of preliminary test (s) of significance. Biometrics, 20(3), 427-442.

Beckman RA, Natanegara F, Singh P, Cooner F, Antonijevic Z, Liu Y, Mayer C, Price K, Tang R, Xia A, et al (2022) Advancing innovative clinical trials to efficiently deliver medicines to patients. Nature Reviews Drug Discovery, 21(8), 543–544.

Bennett M, White S, Best N and Mander A (2021) A novel equivalence probability weighted power prior for using historical control data in an adaptive clinical trial design: A comparison to standard methods. Pharmaceutical Statistics, 20(3), 462–484.

Berry DA (2006) Bayesian clinical trials. Nature Reviews Drug Discovery, 5(1), 27-36.

Burger HU, Gerlinger C, Harbron C, Koch A, Posch M, Rochon J and Schiel A (2021) The use of external controls: To what extent can it currently be recommended? Pharmaceutical Statistics, 20(6), 1002–1016.

Campbell G, Irony T, Pennello G and Thompson L (2023) Bayesian Statistics for Medical Devices: Progress Since 2010. Therapeutic Innovation & Regulatory Science, 57(3), 453-463.

Carlin BP and Nollevaux F (2022) Bayesian complex innovative trial designs (CIDs) and their use in drug development for rare disease. The Journal of Clinical Pharmacology, 62, S56-S71.

Chen MH, Ibrahim JG and Shao QM (2000) Power prior distributions for generalized linear models. Journal of Statistical Planning and Inference, 84(1-2), 121-137.

Cole SR and Hernán MA (2008) Constructing inverse probability weights for marginal structural models. American Journal of Epidemiology, 168(6), 656-664.

Cox DR (1972) Regression models and life-tables. Journal of the Royal Statistical Society: Series B, 34(2), 187-202.

Cuffe RL (2011) The inclusion of historical control data may reduce the power of a confirmatory study. Statistics in Medicine, 30(12), 1329-1338.

Duan Y, Ye K and Smith EP (2006) Evaluating water quality using power priors to incorporate historical information. Environmetrics, 17(1), 95-106.

Feißt M, Krisam J and Kieser M (2020) Incorporating historical two-arm data in clinical trials with binary outcome: a practical approach. Pharmaceutical Statistics, 19(5), 662-678.

Folefac CA and Desmond H (2022) Clinical equipoise: Why still the gold standard for randomized clinical trials?. Clinical Ethics, 14777509221121107.

Fu C, Pang H, Zhou S and Zhu J (2023) Covariate handling approaches in combination with dynamic borrowing for hybrid control studies. Pharmaceutical Statistics, 22(4), 619-632.

Ghadessi M, Tang R, Zhou J, Liu R, Wang C, Toyoizumi K, Mei C, Zhang L, Deng C Q and Beckman RA (2020) A roadmap to using historical controls in clinical trials - by Drug Information Association Adaptive Design Scientific Working Group (DIA-ADSWG). Orphanet Journal of Rare Diseases, 15(1), 69.

Ghalanos A and Theussl S (2015) Rsolnp: general non-linear optimization using augmented Lagrange multiplier method. R package version, 1.16.

Gravestock I, Held L and COMBACTE-Net consortium (2017) Adaptive power priors with empirical Bayes for clinical trials. Pharmaceutical Statistics, 16(5), 349-360.

Gravestock I and Held L (2019) Power priors based on multiple historical studies for binary outcomes. Biometrical Journal, 61(5), 1201–1218.

Greenland S, Pearl J and Robins JM (1999) Confounding and collapsibility in causal inference. Statistical Science, 14(1), 29-46.

Gross AM (2021) Using real world data to support regulatory approval of drugs in rare diseases: A review of opportunities, limitations & a case example. Current Problems in Cancer, 45(4), 100769.

Halpern SD, Karlawish JH and Berlin JA (2002) The continuing unethical conduct of underpowered clinical trials. JAMA, 288(3), 358–362.

Hamilton M (1959) The assessment of anxiety states by rating. The British Journal of Medical Psychology, 32(1), 50–55.

Hamilton M (1960) A rating scale for depression. Journal of Neurology, Neurosurgery, and Psychiatry, 23(1), 56.

Han B, Zhan J, John Zhong Z, Liu D and Lindborg S (2017) Covariate-adjusted borrowing of historical control data in randomized clinical trials. Pharmaceutical Statistics, 16(4), 296–308.

Hoeting JA, Madigan D, Raftery AE and Volinsky CT (1999) Bayesian model averaging: a tutorial. Statistical Science, 14(4), 382-417.

Hobbs BP, Carlin BP, Mandrekar SJ and Sargent DJ (2011) Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. Biometrics, 67(3), 1047-1056.

Hobbs BP, Sargent DJ and Carlin BP (2012) Commensurate Priors for Incorporating Historical Information in Clinical Trials Using General and Generalized Linear Models. Bayesian Analysis, 7(3), 639–674.

Hobbs BP, Carlin BP and Sargent DJ (2013) Adaptive adjustment of the randomization ratio using historical control data. Clinical Trials, 10(3), 430–440.

Ibrahim JG and Chen MH (2000) Power prior distributions for regression models. Statistical Science, 46-60.

Ibrahim JG, Chen MH, Gwon Y and Chen F (2015) The power prior: theory and applications. Statistics in Medicine, 34(28), 3724-3749.

Jemielita T, Tse A and Chen C (2020) Oncology phase II proof-of-concept studies with multiple targets: Randomized controlled trial or single arm ?. Pharmaceutical Statistics, 19(2), 117-125.

Kaizer AM, Koopmeiners JS and Hobbs BP (2018) Bayesian hierarchical modeling based on multisource exchangeability. Biostatistics, 19(2), 169-184.

Kaplan EL and Meier P (1958) Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53(282), 457-481.

Keller M, Montgomery S, Ball W, Morrison M, Snavely D, Liu G, Hargreaves R, Hietala J, Lines C, Beebe K, et al (2006) Lack of efficacy of the substance p (neurokinin1 receptor) antagonist aprepitant in the treatment of major depressive disorder. Biological Psychiatry, 59(3), 216–223.

Kopp-Schneider A, Calderazzo S and Wiesenfarth M (2020) Power gains by using external information in clinical trials are typically not possible when requiring strict type I error control. Biometrical Journal. 62(2), 361–374.

Kotalik A, Vock DM, Donny EC, Hatsukami DK and Koopmeiners JS (2021) Dynamic borrowing in the presence of treatment effect heterogeneity. Biostatistics, 22(4), 789-804.

Lee BK, Lessler J and Stuart EA (2010) Improving propensity score weighting using machine learning. Statistics in Medicine, 29(3), 337-346.

Lee BK, Lessler J and Stuart EA (2011) Weight trimming and propensity score weighting. PloS one, 6(3), e18174.

Li W, Liu F and Snavely D (2020) Revisit of test-then-pool methods and some practical considerations. Pharmaceutical Statistics, 19(5), 498–517.

Li L and Jemielita T (2023) Confounding adjustment in the analysis of augmented randomized

controlled trial with hybrid control arm. Statistics in Medicine, 42(16), 2855–2872.

Li R, Lin R, Huang J, Tian L and Zhu J (2023) A frequentist approach to dynamic borrowing. Biometrical Journal. 65(7), e2100406.

Lim J, Wang L, Best N, Liu J, Yuan J, Yong F, Zhang L, Walley R, Gosselin A, Roebling R, et al (2020) Reducing Patient Burden in Clinical Trials Through the Use of Historical Controls: Appropriate Selection of Historical Data to Minimize Risk of Bias. Therapeutic Innovation & Regulatory Science, 54(4), 850–860.

Lin DY and Wei LJ (1989) The robust inference for the Cox proportional hazards model. Journal of the American Statistical Association, 84(408), 1074-1078.

Lin J and Lin J (2022) Incorporating propensity scores for evidence synthesis under bayesian framework: review and recommendations for clinical studies. Journal of biopharmaceutical statistics, 32(1), 53–74.

Lin J, Gamalo-Siebers M and Tiwari R (2018) Propensity score matched augmented controls in randomized clinical trials: A case study. Pharmaceutical Statistics, 17(5), 629–647.

Liu GF (2018) A dynamic power prior for borrowing historical data in noninferiority trials with binary endpoint. Pharmaceutical Statistics, 17(1), 61-73.

Meldrum ML (2000) A brief history of the randomized controlled trial: From oranges and lemons to the gold standard. Hematology/Oncology Clinics of North America, 14(4), 745-760.

Mishra-Kalyani PS, Amiri Kordestani L, Rivera DR, Singh H, Ibrahim A, DeClaro RA, Shen Y, Tang S, Sridhara R, Kluetz PG, et al (2022) External control arms in oncology: current use and future directions. Annals of Oncology, 33(4), 376–383.

Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG and CONSORT (2012) CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. International Journal of Surgery, 10(1), 28–55.

Morita S, Thall PF and Müller P (2008) Determining the effective sample size of a parametric prior. Biometrics, 64(2), 595-602.

Nakamura K and Shibata T (2020) Regulatory changes after the enforcement of the new Clinical Trials Act in Japan. Japanese Journal of Clinical Oncology, 50(4), 399-404.

Nikolakopoulos S, van der Tweel I and Roes KC (2018) Dynamic borrowing through empirical power priors that control type I error. Biometrics, 74(3), 874-880.

Neuenschwander B, Branson M and Spiegelhalter DJ (2009) A note on the power prior. Statistics in Medicine, 28(28), 3562-3566.

Neuenschwander B, Capkun-Niggli G, Branson M and Spiegelhalter DJ (2010) Summarizing historical information on controls in clinical trials. Clinical Trials, 7(1), 5-18.

Neuenschwander B, Weber S, Schmidli H and O'Hagan A (2020) Predictively consistent prior effective sample sizes. Biometrics, 76(2), 578-587.

Ohigashi T, Maruo K, Sozu T and Gosho M (2022) Using horseshoe prior for incorporating multiple historical control data in randomized controlled trials. Statistical Methods in Medical Research, 31(7), 1392-1404.

O'Neill WW, Martin JL, Dixon SR, Bartorelli AL, Trabattoni D, Oemrawsingh PV, Atsma DE, Chang M, Marquardt W, Oh JK, et al (2007) Acute Myocardial Infarction with Hyperoxemic Therapy (AMIHOT): a prospective, randomized trial of intracoronary hyperoxemic reperfusion after percutaneous coronary intervention. Journal of the American College of Cardiology, 50(5), 397–405.

Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR and Schacht AL (2010) How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nature reviews. Drug Discovery, 9(3), 203–214.

Pocock SJ (1976) The combination of randomized and historical controls in clinical trials. Journal of Chronic Diseases, 29(3), 175-188.

Popat S, Liu SV, Scheuer N, Hsu GG, Lockhart A, Ramagopalan SV, Griesinger F and Subbiah V (2022) Addressing challenges with real-world synthetic control arms to demonstrate the comparative effectiveness of Pralsetinib in non-small cell lung cancer. Nature Communications, 13(1), 3500.Psioda MA, Soukup M and Ibrahim JG (2018) A practical Bayesian adaptive design incorporating data from historical controls. Statistics in Medicine, 37(27), 4054-4070.

Ribeiro TB, Bennett CL, Colunga-Lozano LE, Araujo APV, Hozo I and Djulbegovic B (2023) Increasing FDA-accelerated approval of single-arm trials in oncology (1992 to 2020). Journal of Clinical Epidemiology, 159, 151–158.

Rosenbaum PR and Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55.

Roychoudhury S and Neuenschwander B (2020) Bayesian leveraging of historical control data for a clinical trial with time-to-event endpoint. Statistics in Medicine, 39(7), 984-995.

Rubin DB (2004) On principles for modeling propensity scores in medical research. Pharmacoepidemiology and Drug Safety, 13(12), 855-857.

Rubin DB (2008) For objective causal inference, design trumps analysis. Annals of Applied Statistics, 2(3) 808-840.

Sato T and Matsuyama Y (2003) Marginal structural models as a tool for standardization. Epidemiology, 14(6), 680–686.

Schmidli H, Gsteiger S, Roychoudhury S, O'Hagan A, Spiegelhalter D and Neuenschwander B (2014) Robust meta-analytic-predictive priors in clinical trials with historical control

information. Biometrics, 70(4), 1023-1032.

Schmidli H, Häring DA, Thomas M, Cassidy A, Weber S and Bretz F (2020) Beyond randomized clinical trials: use of external controls. Clinical Pharmacology & Therapeutics, 107(4), 806-816.

Schuirmann DJ (1987) A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. Journal of Pharmacokinetics and Biopharmaceutics, 15(6), 657–680.

Spiegelhalter DJ, Abrams KR and Myles JP (2004) Bayesian approaches to clinical trials and health-care evaluation. John Wiley & Sons.

Stone GW, Martin JL, de Boer MJ, Margheri M, Bramucci E, Blankenship JC, Metzger DC, Gibbons RJ, Lindsay BS, Weiner BH, et al (2009) Effect of supersaturated oxygen delivery on infarct size after percutaneous coronary intervention in acute myocardial infarction. Circulation. Cardiovascular Interventions, 2(5), 366–375.

Ueno H, Ioka T, Ikeda M, Ohkawa S, Yanagimoto H, Boku N, Fukutomi A, Sugimori K, Baba H, Yamao K, et al (2013) Randomized phase III study of gemcitabine plus S-1, S-1 alone, or gemcitabine alone in patients with locally advanced and metastatic pancreatic cancer in Japan and Taiwan: GEST study. Journal of Clinical Oncology, 31(13), 1640–1648.

U.S. Food and Drug Administration (2014) Expedited programs for serious conditions – drugs and biologics. draft guide to industry. https://www.fda.gov/media/86377/download. Accessed 20 September 2023.

U.S. Food and Drug Administration (2019) Draft guidance for industry: demonstrating substantial evidence of effectiveness for human drug and biological products. https://www.fda.gov/media/133660/download. Accessed 2 November 2023.

U.S. Food and Drug Administration (2022) Real-World Evidence. https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence. Accessed 19 September 2023.

Viele K, Berry S, Neuenschwander B, Amzal B, Chen F, Enas N, Hobbs B, Ibrahim JG, Kinnersley N, Lindborg S, et al (2014) Use of historical control data for assessing treatment effects in clinical trials. Pharmaceutical Statistics, 13(1), 41–54.

Viele K, Mundy LM, Noble RB, Li G, Broglio K and Wetherington JD (2018) Phase 3 adaptive trial design options in treatment of complicated urinary tract infection. Pharmaceutical Statistics, 17(6), 811-822.

Wang C, Li H, Chen WC, Lu N, Tiwari R, Xu Y and Yue LQ (2019) Propensity score-integrated power prior approach for incorporating real-world evidence in single-arm clinical studies. Journal of Biopharmaceutical Statistics, 29(5), 731–748.

Wang X, Suttner L, Jemielita T and Li X (2022) Propensity score-integrated Bayesian prior approaches for augmented control designs: a simulation study. Journal of Biopharmaceutical Statistics, 32(1), 170–190.

Wang X, Dormont F, Lorenzato C, Latouche A, Hernandez R and Rouzier R (2023) Current perspectives for external control arms in oncology clinical trials: Analysis of EMA approvals 2016-2021. Journal of Cancer Policy, 35, 100403.

Wiesenfarth M and Calderazzo S (2020) Quantification of prior impact in terms of effective current sample size. Biometrics, 76(1), 326-336.

Wu J, Wang C, Toh S, Pisa F E and Bauer L (2020) Use of real-world evidence in regulatory decisions for rare diseases in the United States-Current status and future directions. Pharmacoepidemiology and Drug Safety, 29(10), 1213–1218.

Yamaue H, Shimizu A, Hagiwara Y, Sho M, Yanagimoto H, Nakamori S, Ueno H, Ishii H, Kitano M, Sugimori K, et al (2017) Multicenter, randomized, open-label Phase II study comparing S-1 alternate-day oral therapy with the standard daily regimen as a first-line treatment in patients with unresectable advanced pancreatic cancer. Cancer Chemotherapy and Pharmacology, 79(4), 813–823.

Ye Y (1987) Interior algorithms for linear, quadratic, and linearly constrained non-linear programming (Doctoral dissertation, Ph. D. thesis, Department of ESS, Stanford University).

Zhang HH and Lu W (2007) Adaptive lasso for Cox's proportional hazards model. Biometrika, 94(3), 691–703.

Zhang H, Shen Y, Li J, Ye H and Chiang AY (2023) Adaptively leveraging external data with robust meta-analytical-predictive prior using empirical Bayes. Pharmaceutical Statistics, 22(5), 846–860.

Zhao H, Hobbs BP, Ma H, Jiang Q and Carlin BP (2016) Combining Non-randomized and Randomized Data in Clinical Trials Using Commensurate Priors. Health Services and Outcomes Research Methodology, 16(3), 154–171.

Zou H (2006) The adaptive lasso and its oracle properties. Journal of the American statistical association, 101(476), 1418–1429.

Zhou T and Ji, Y (2021) Incorporating external data into the analysis of clinical trials via Bayesian additive regression trees. Statistics in Medicine, 40(28), 6421-6442.

大東智洋, 丸尾和司, 寒水孝司, 五所正彦 (2023) Dirichlet 過程混合モデルを用いたクラスタリングによる既存データ利用法の提案. 2023 年度日本計量生物学会年会, 2023 年 4 月 20-21 日. 札幌.

佐藤俊哉, 松山裕 (2011) 交絡という不思議な現象と交絡を取りのぞく解析—標準化と周辺構造モデル—. 計量生物学, 32 (Special Issue), S35-S49.

武田健太朗, 大庭真梨, 柿爪智行, 坂巻顕太郎, 田栗正隆, 森田智視 (2015) 臨床試験におけるヒストリカルコントロールデータの利用. 計量生物学, 36(1), 25-50.
丹後俊郎, 松井茂之編 (2018) 医学統計学ハンドブック. 朝倉書房.
野村尚吾, 大東智洋, 澤本涼 (2022) Hybrid control アプローチを用いるランダム化比較試験の計画と解析: 外部データが要約統計量の場合. 計量生物学, 43(1), 63-96.

付録

付録 1　既存対照のサンプルサイズの大きさと動作特性の関係

2.5.2.3 目の数値実験の結果のうち，本文の掲載できなかった結果を，図 A1.1 から図 A1.28 にすべて示す．TTTP は 2 つの帰無仮説を組み合わせた検定併合法，CTTP は従来の検定併合法，δ_pは最低検出力の許容下限，μ_{CC}は新規対照の母平均，$\bar{x}_{HC}$は既存対照の標本平均，n_Tは試験治療のサンプルサイズ，n_{CC}は新規対照のサンプルサイズ，P_{TTTP}は TTTP の併合確率，P_{CTTP}は CTTP の併合確率，$T1E_{TTTP}$は TTTP の第一種の過誤確率，$T1E_{CTTP}$は CTTP の第一種の過誤確率，Pw_{TTTP}は TTTP の検出力，Pw_{CTTP}は CTTP の検出力を表す．

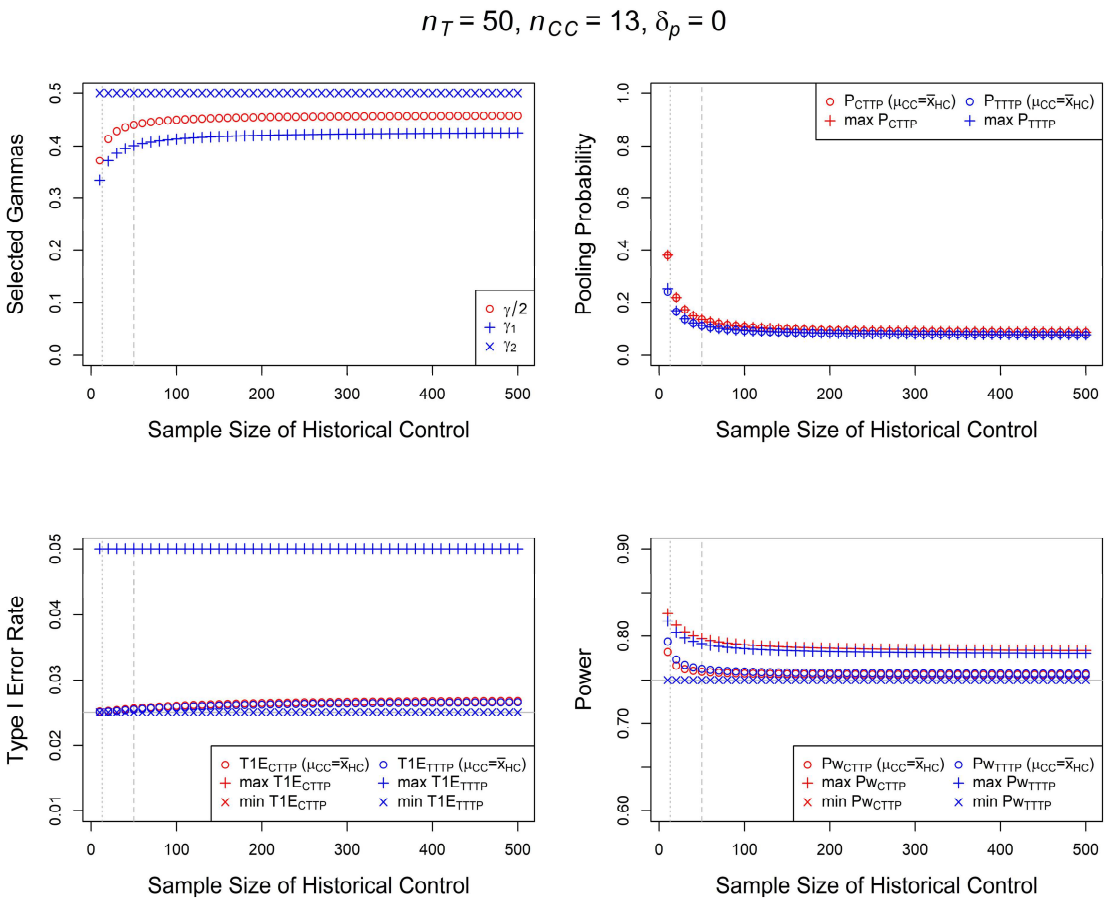

図 A1.1 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係
（$n_T = 50$，$n_{CC} = 13$，$\delta_p = 0$）

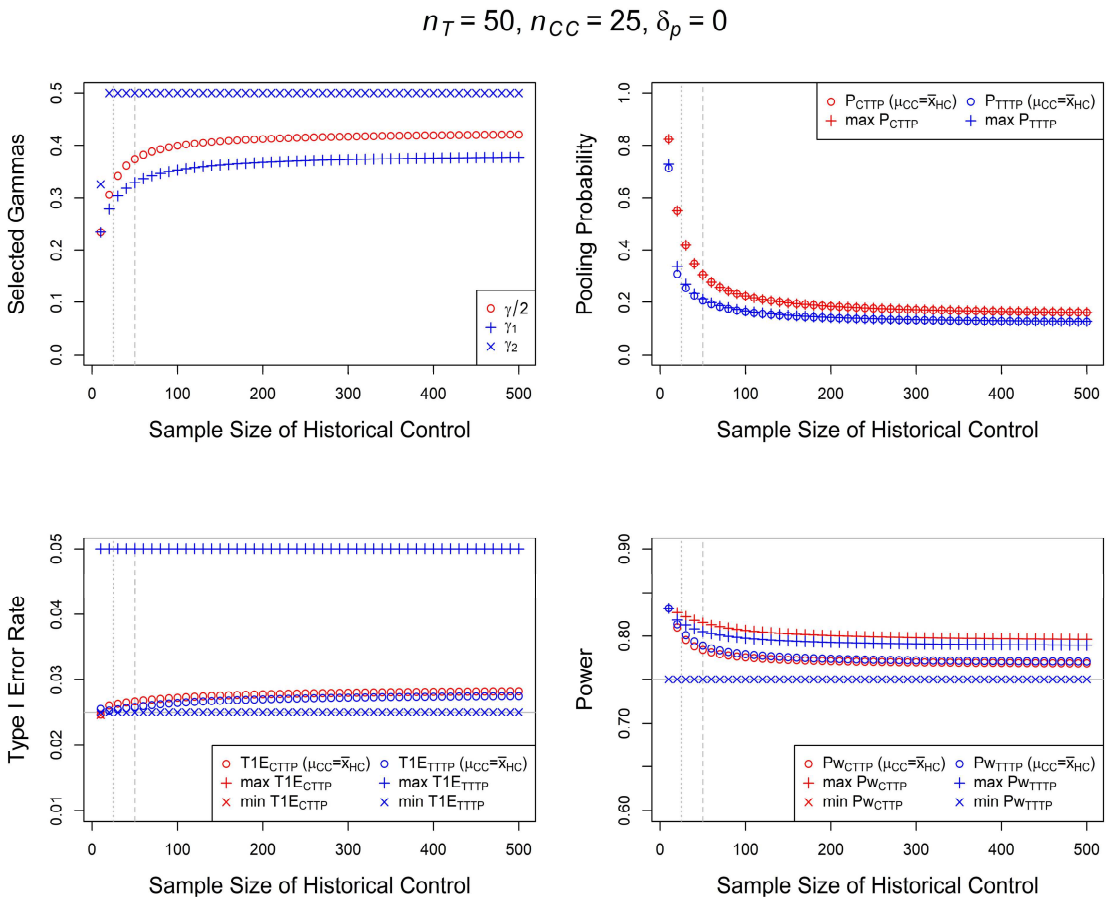

図 A1.2 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係
（$n_T = 50$，$n_{CC} = 25$，$\delta_p = 0$）

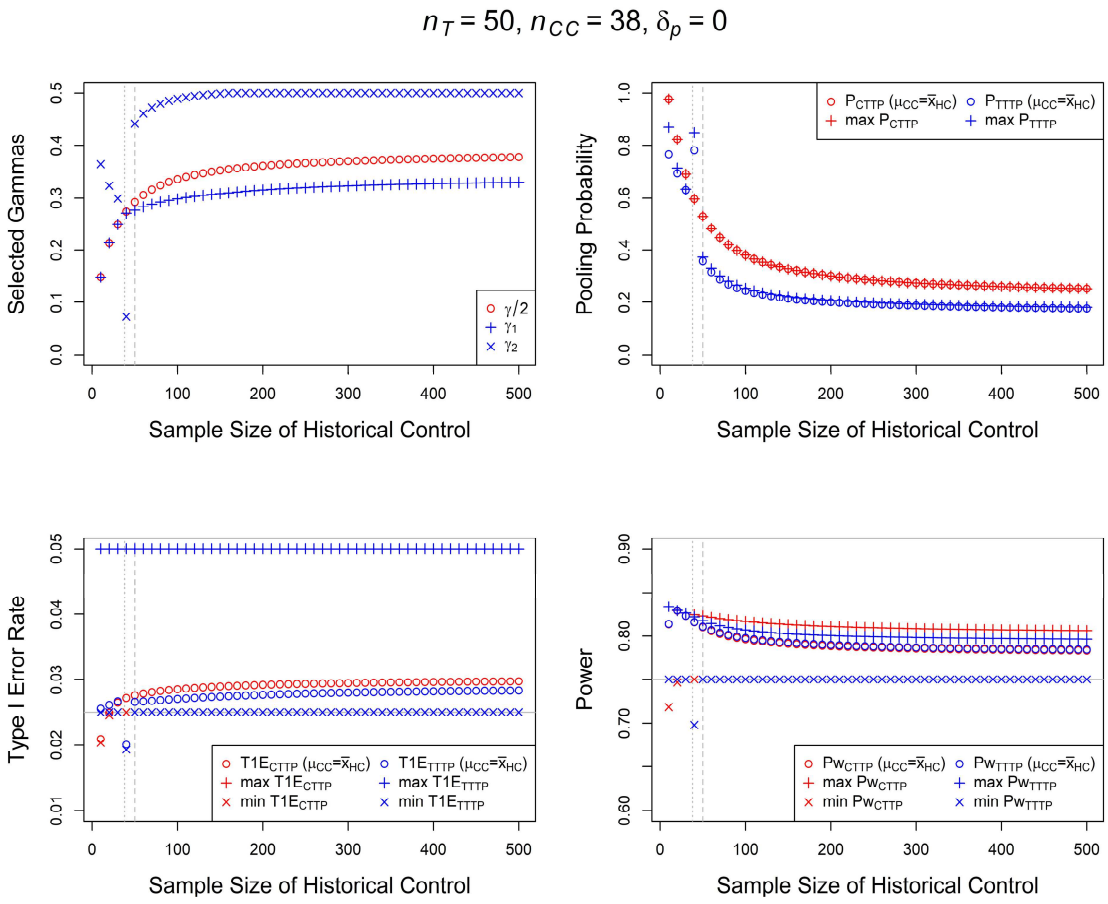

図 A1.3 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係
（$n_T = 50$，$n_{CC} = 38$，$\delta_p = 0$）

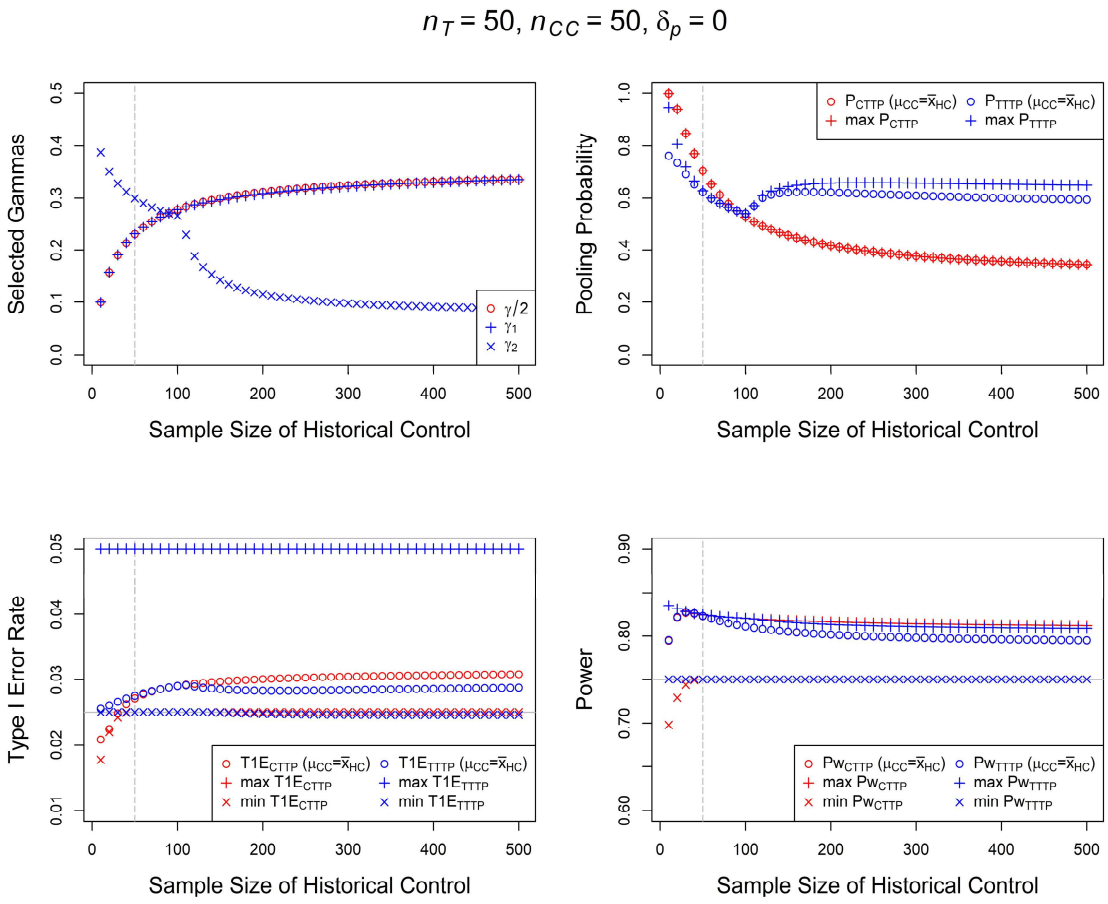

図 A1.4 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係
（$n_T = 50$，$n_{CC} = 50$，$\delta_p = 0$）

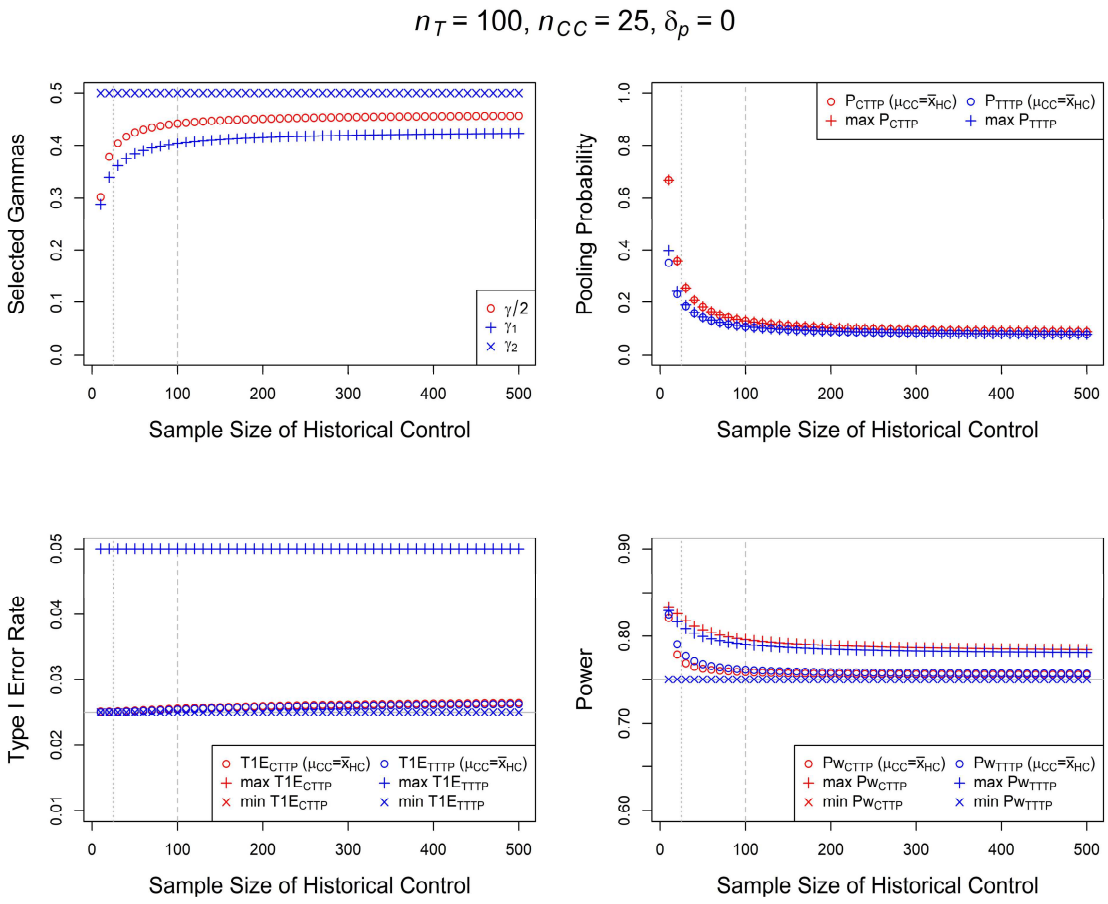

図 A1.5 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係
（$n_T = 100$，$n_{CC} = 25$，$\delta_p = 0$）

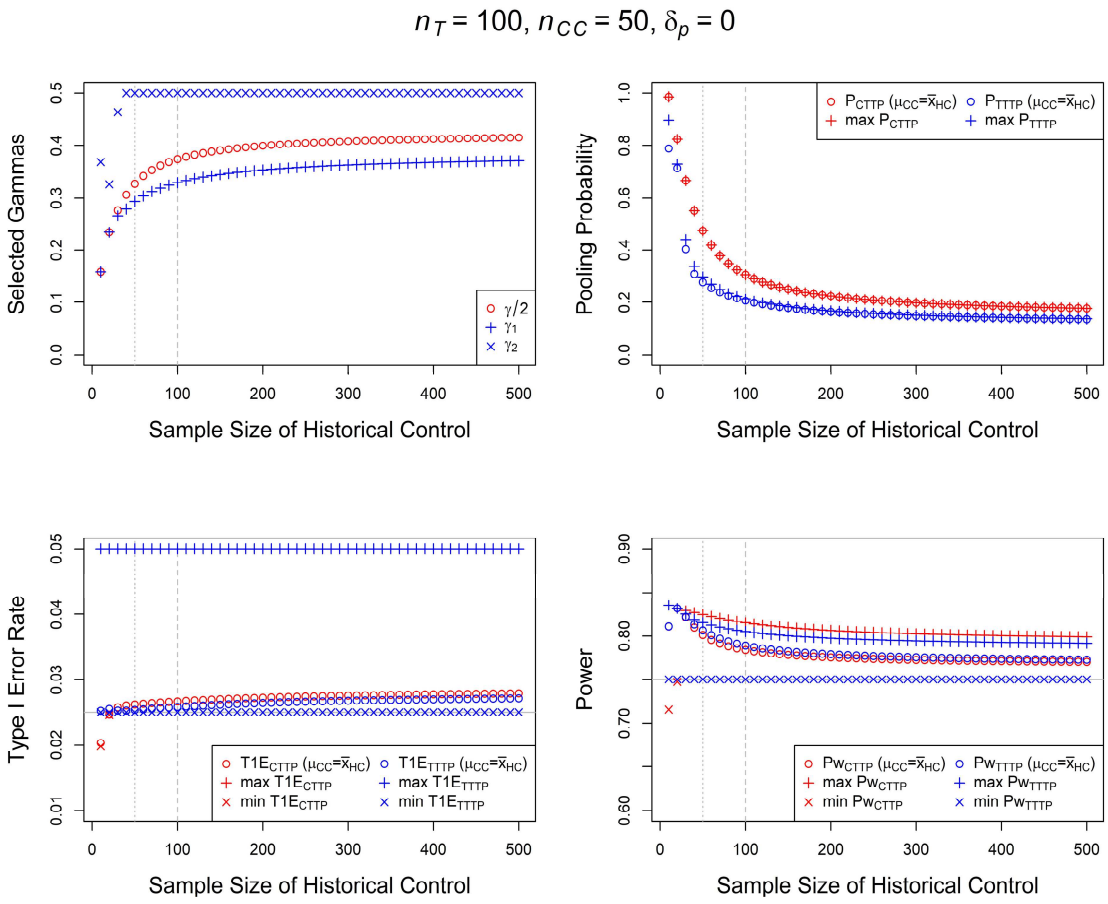

図 A1.6 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係
（$n_T = 100$，$n_{CC} = 50$，$\delta_p = 0$）

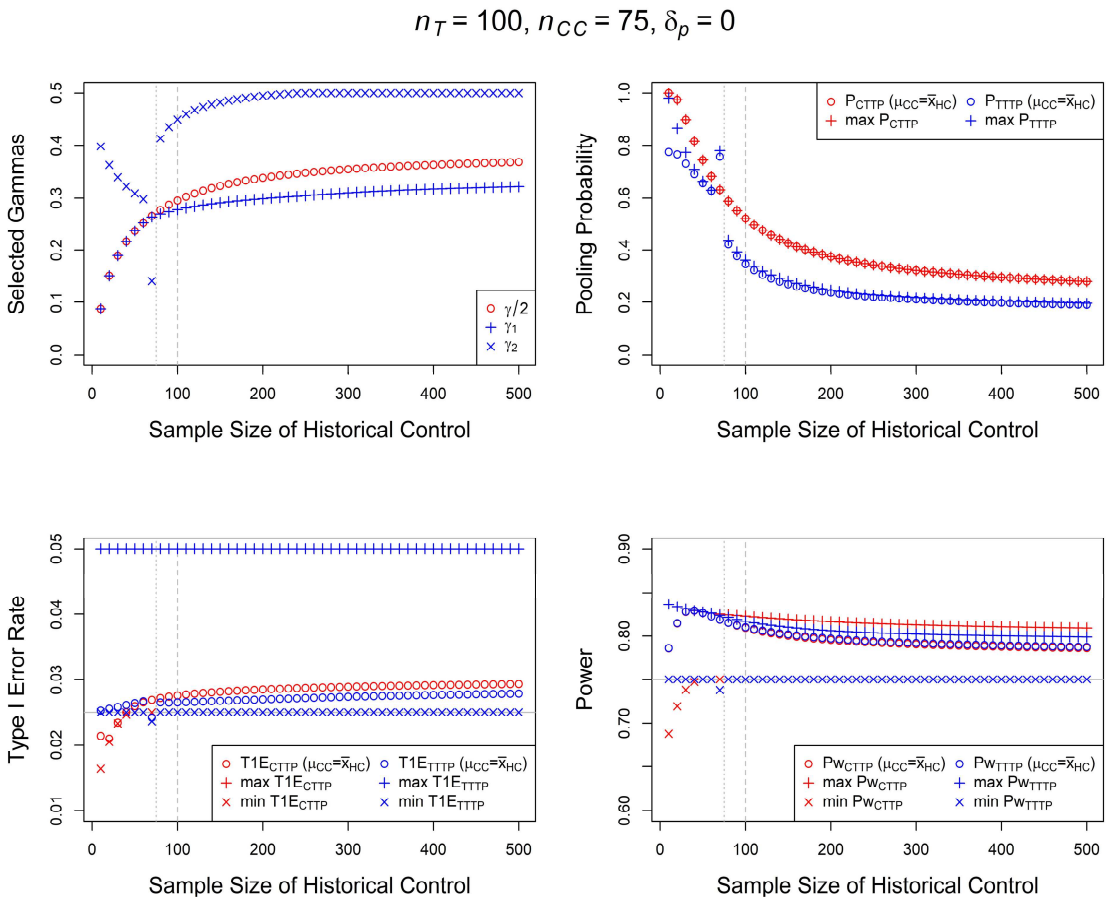

図 A1.7 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係
（$n_T = 100$，$n_{CC} = 75$，$\delta_p = 0$）

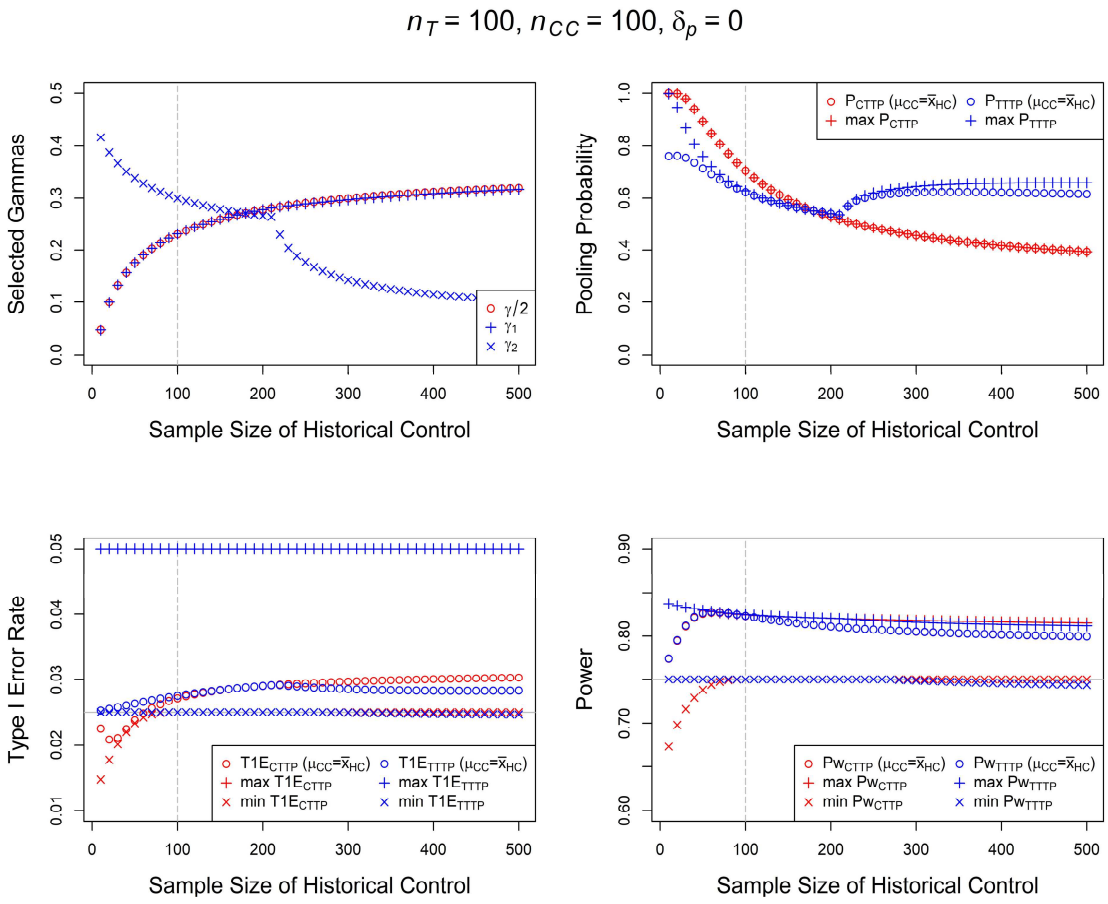

図 A1.8 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係
（$n_T = 100$，$n_{CC} = 100$，$\delta_p = 0$）

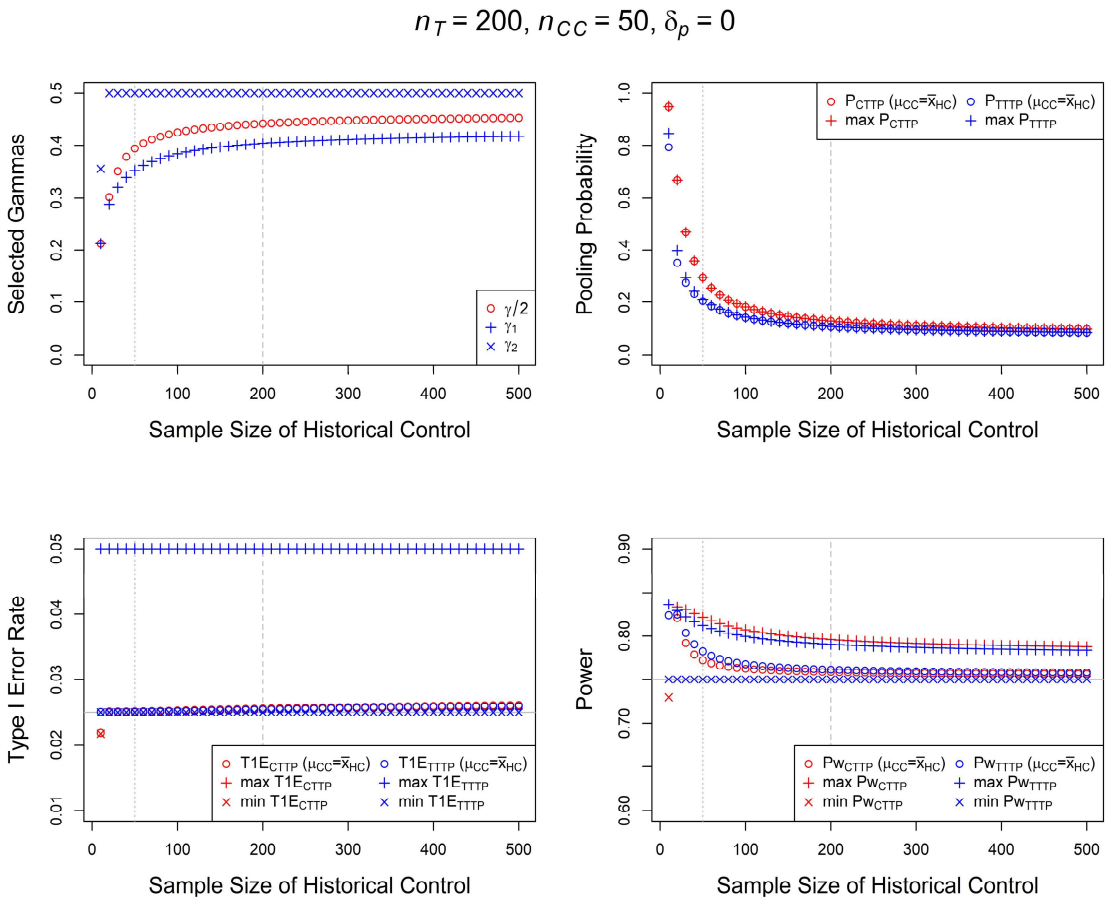

図 A1.9 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係
（$n_T = 200$，$n_{CC} = 50$，$\delta_p = 0$）

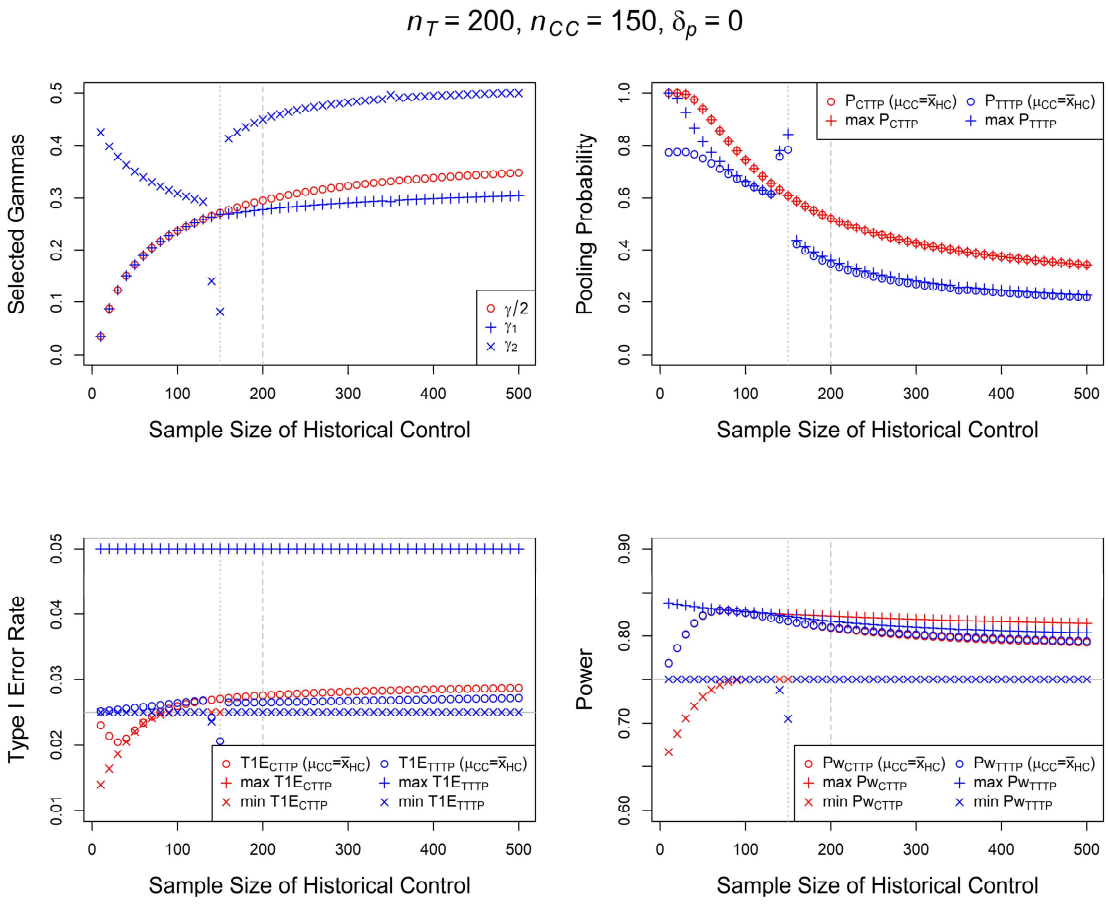

図 A1.10 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係
（$n_T = 200$，$n_{CC} = 150$，$\delta_p = 0$）

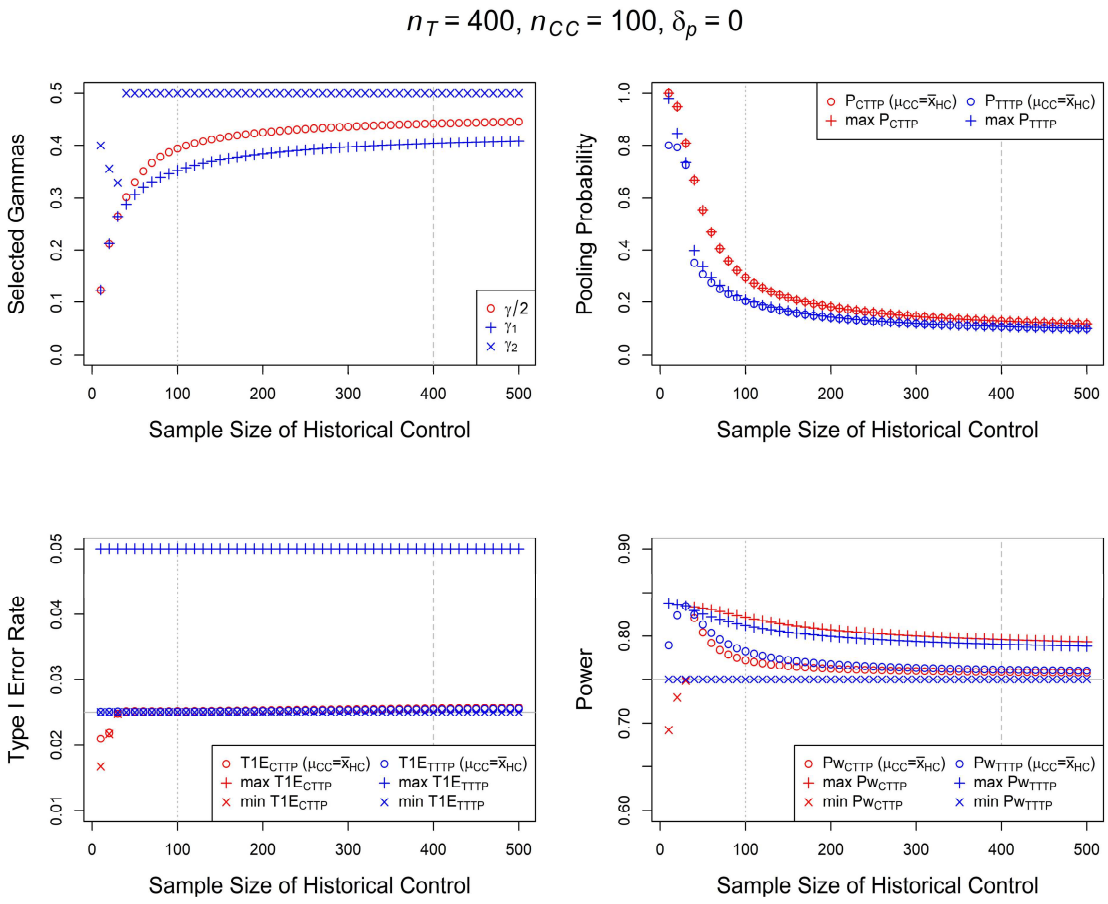

図 A1.11 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係
（$n_T = 400$，$n_{CC} = 100$，$\delta_p = 0$）

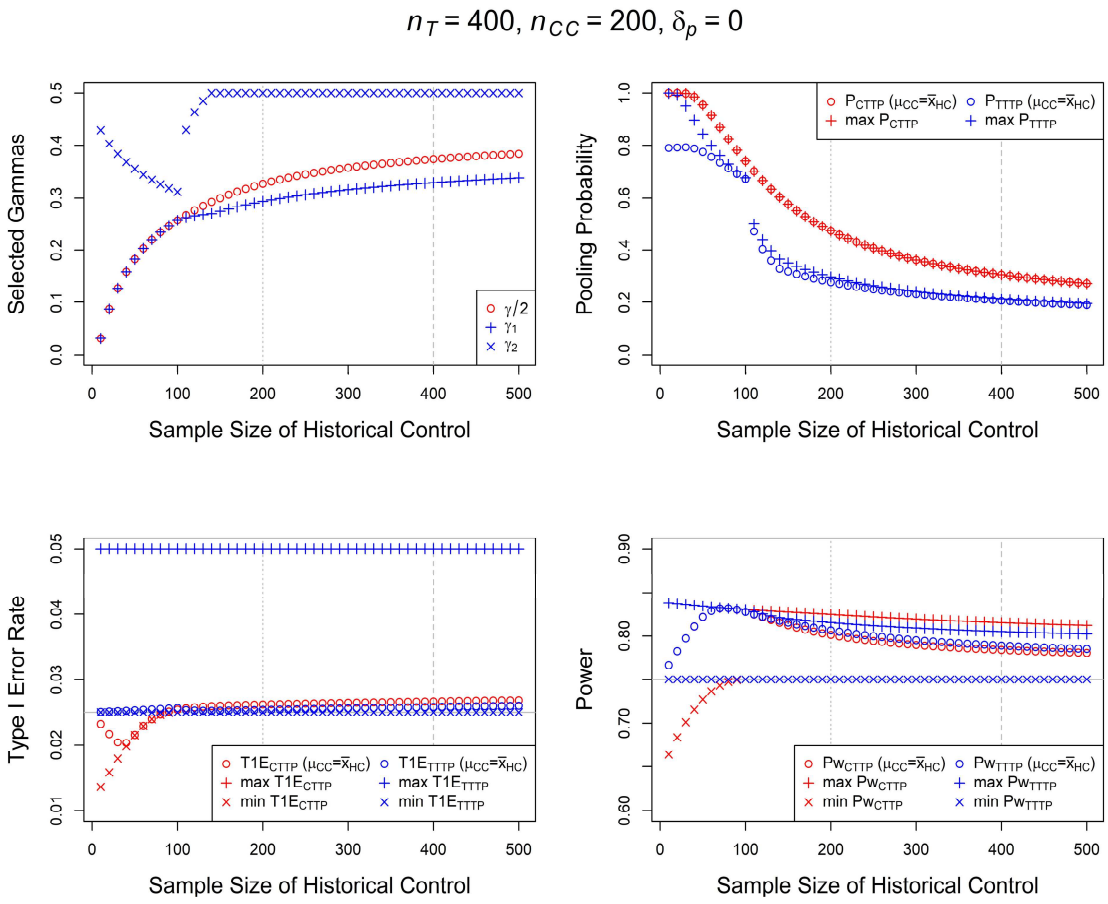

図 A1.12 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係
（$n_T = 400$，$n_{CC} = 200$，$\delta_p = 0$）

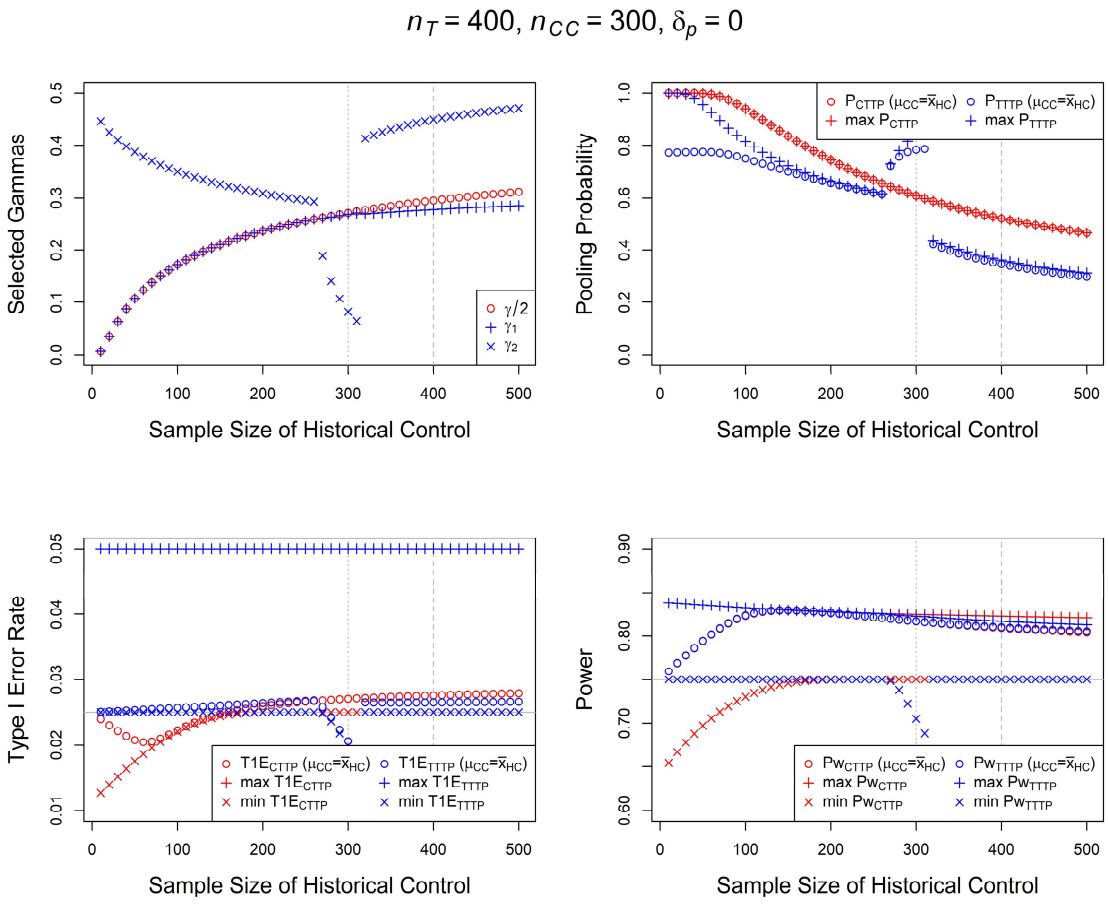

図 A1.13 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係
（$n_T = 400$，$n_{CC} = 300$，$\delta_p = 0$）

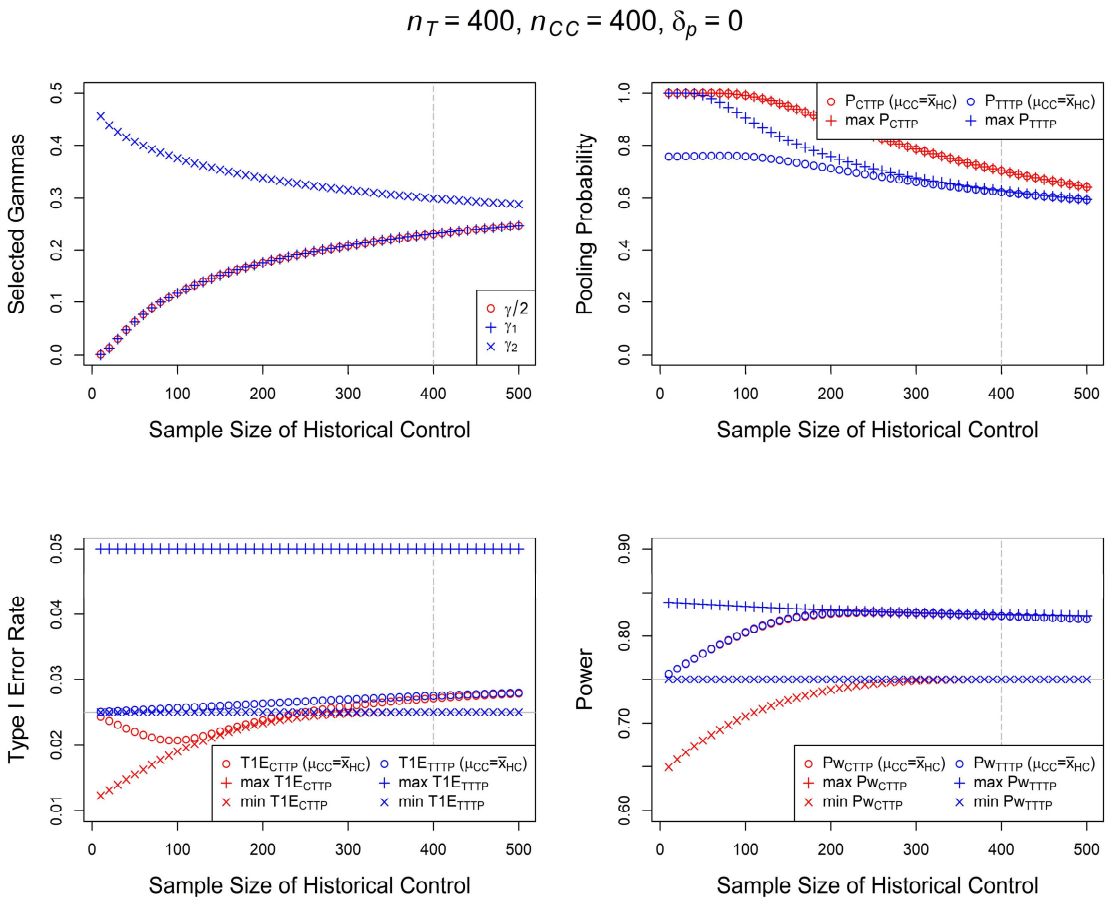

図 A1.14 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係
（$n_T = 400$，$n_{CC} = 400$，$\delta_p = 0$）

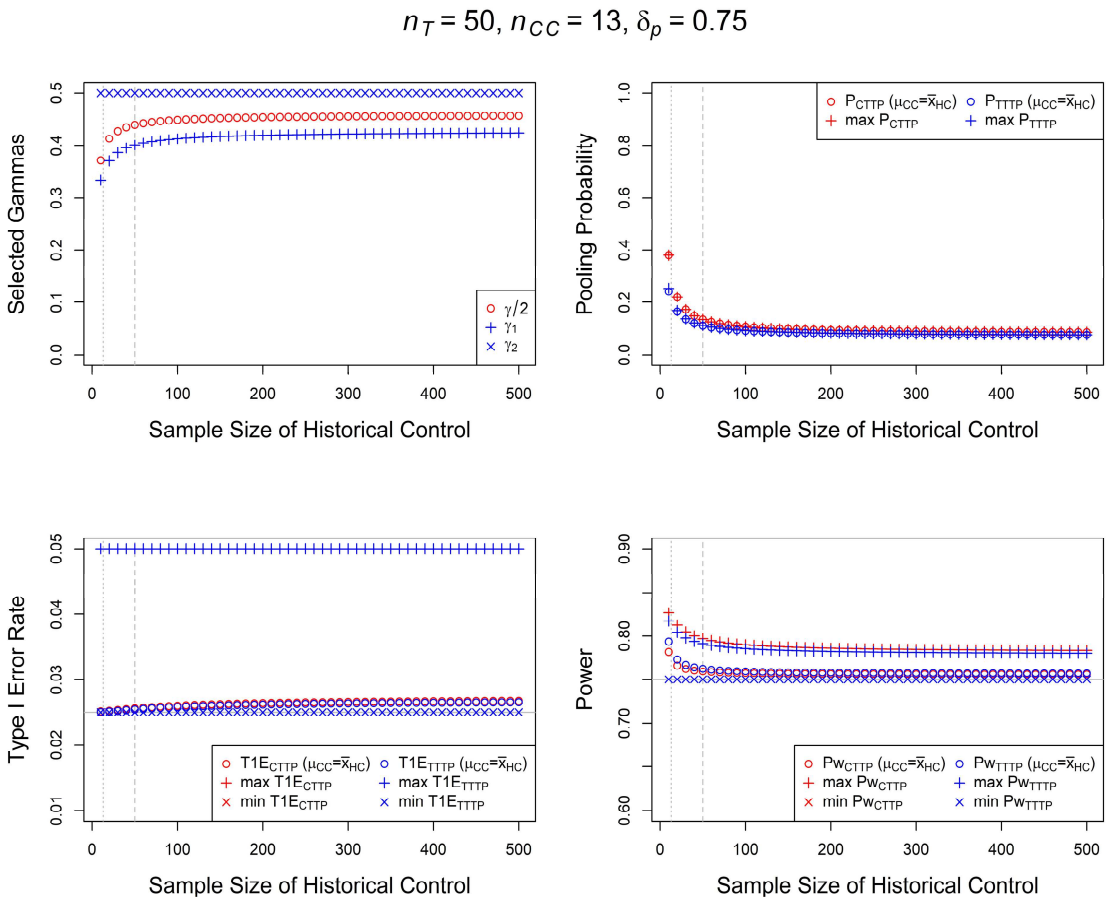

図 A1.15 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係
（$n_T = 50$，$n_{CC} = 13$，$\delta_p = 0.75$）

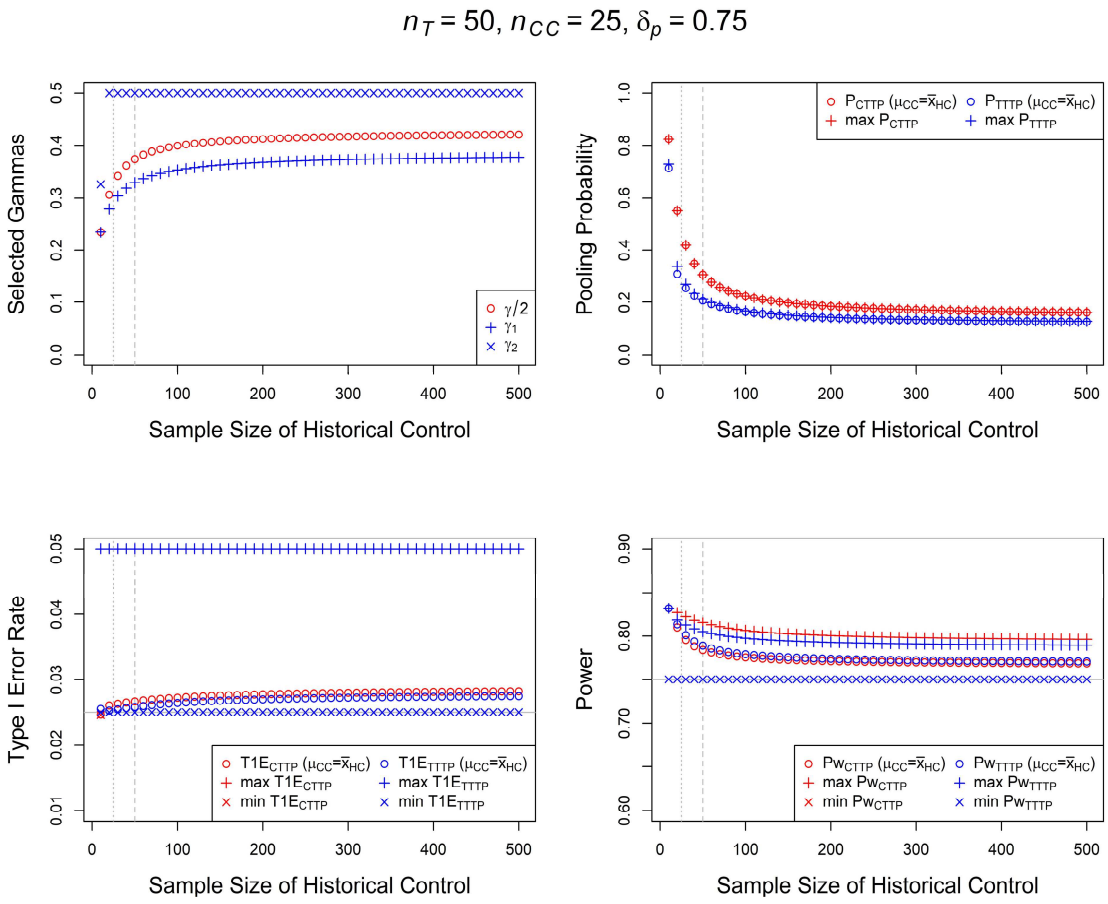

図 A1.16 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係
（$n_T = 50$，$n_{CC} = 25$，$\delta_p = 0.75$）

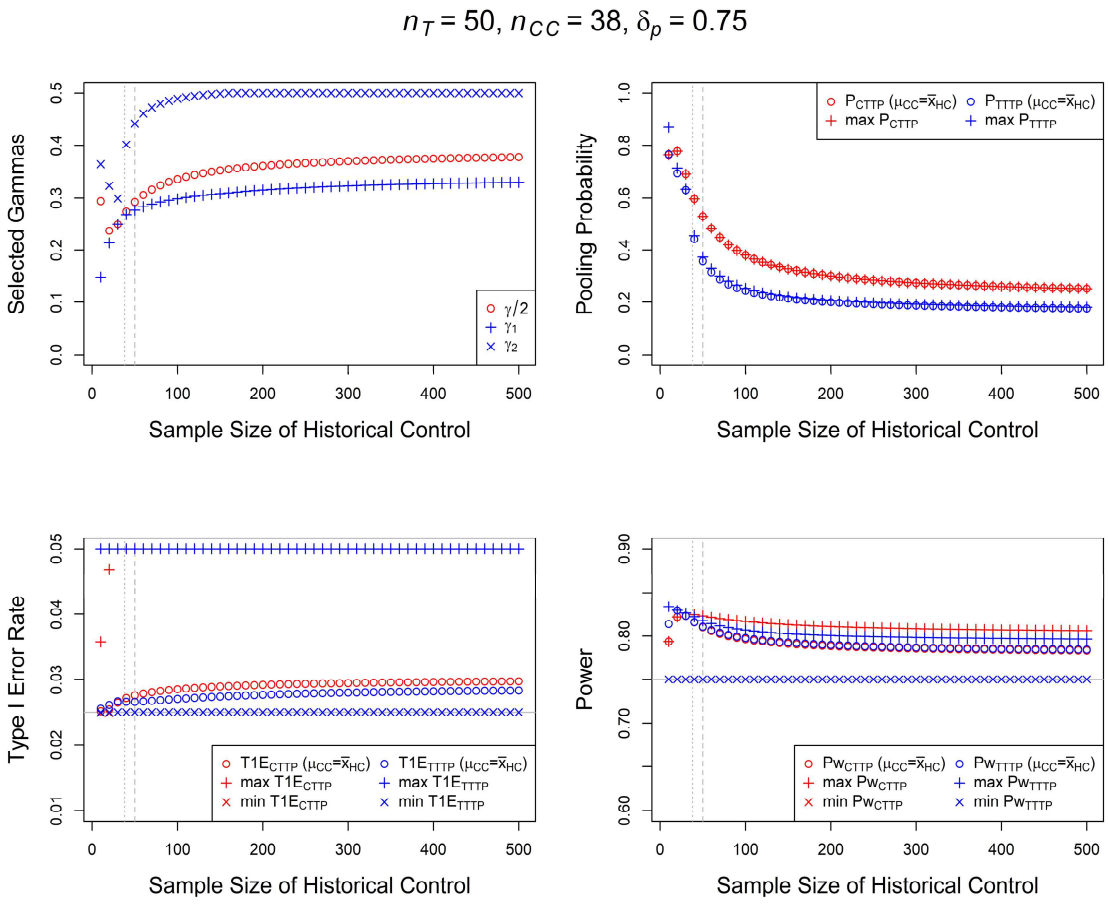

図 A1.17 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係
（$n_T = 50$，$n_{CC} = 38$，$\delta_p = 0.75$）

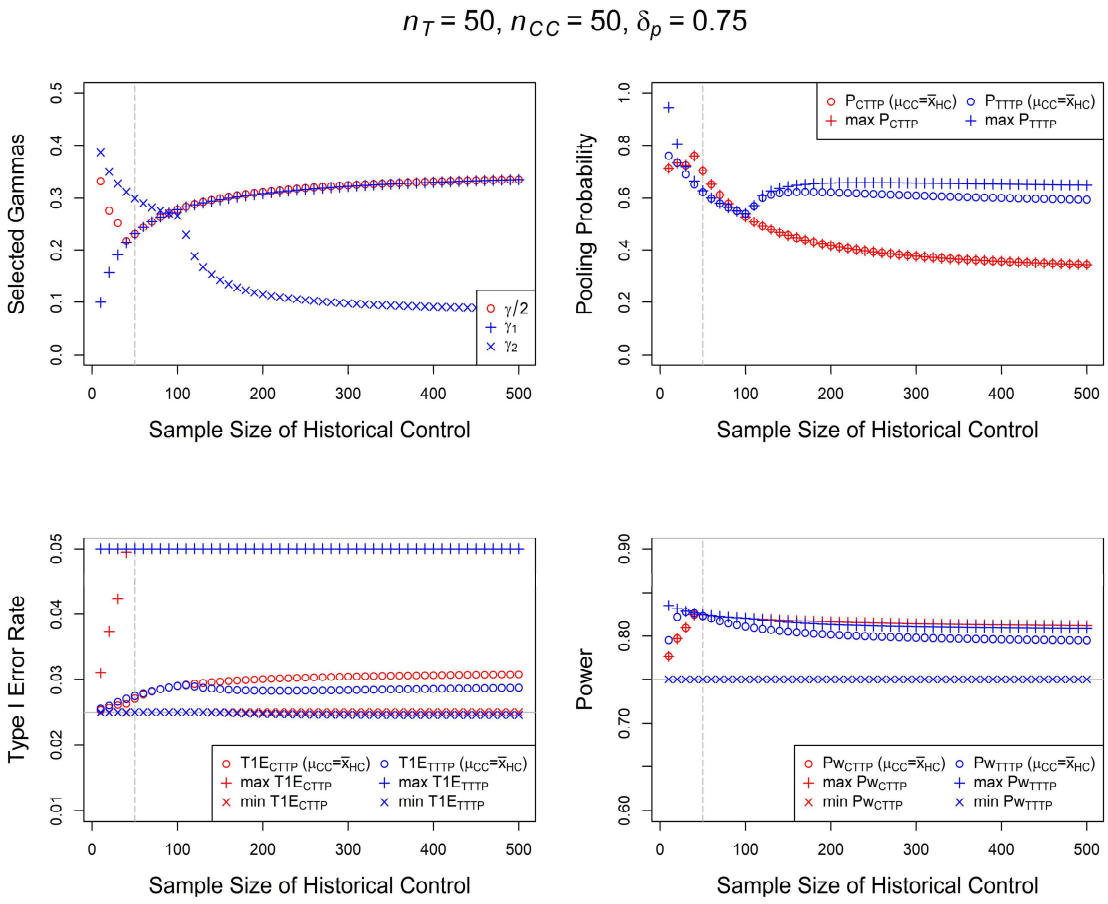

図 A1.18 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係
（$n_T = 50$，$n_{CC} = 50$，$\delta_p = 0.75$）

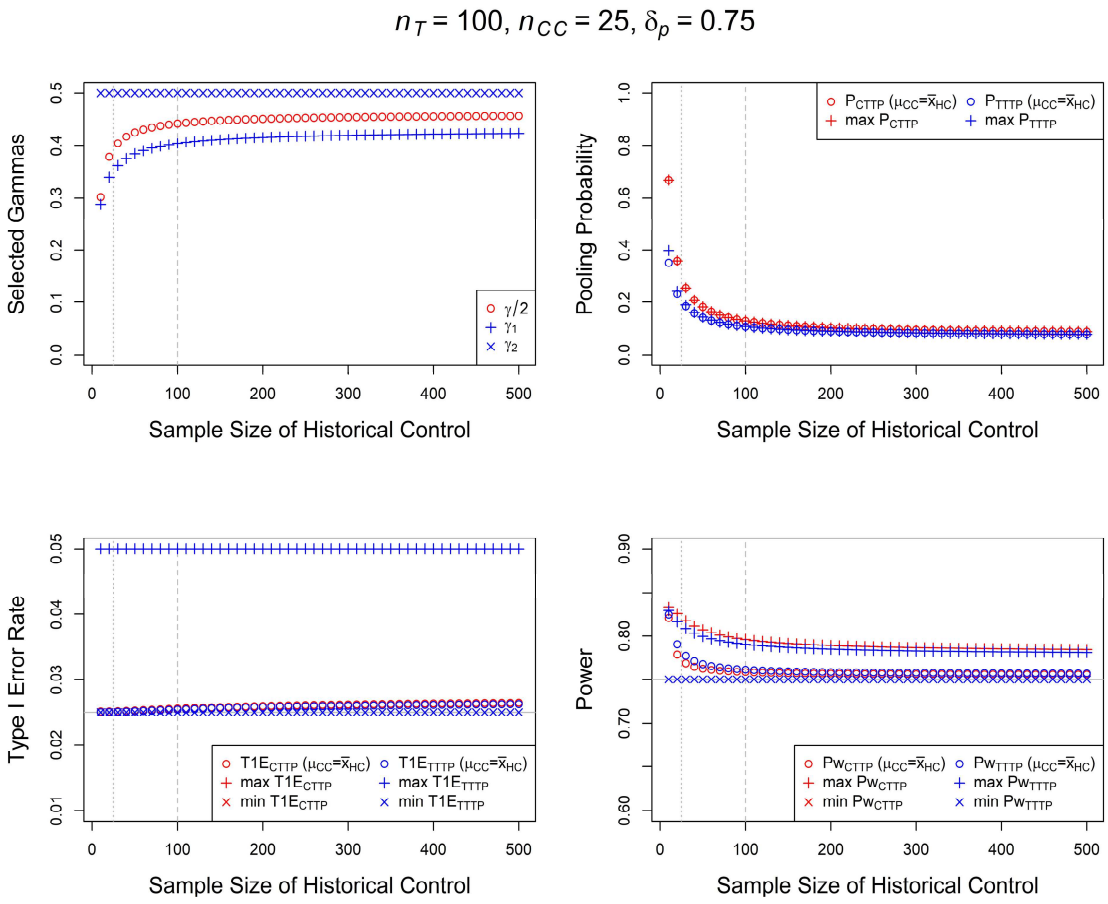

図 A1.19 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係
（$n_T = 100$，$n_{CC} = 25$，$\delta_p = 0.75$）

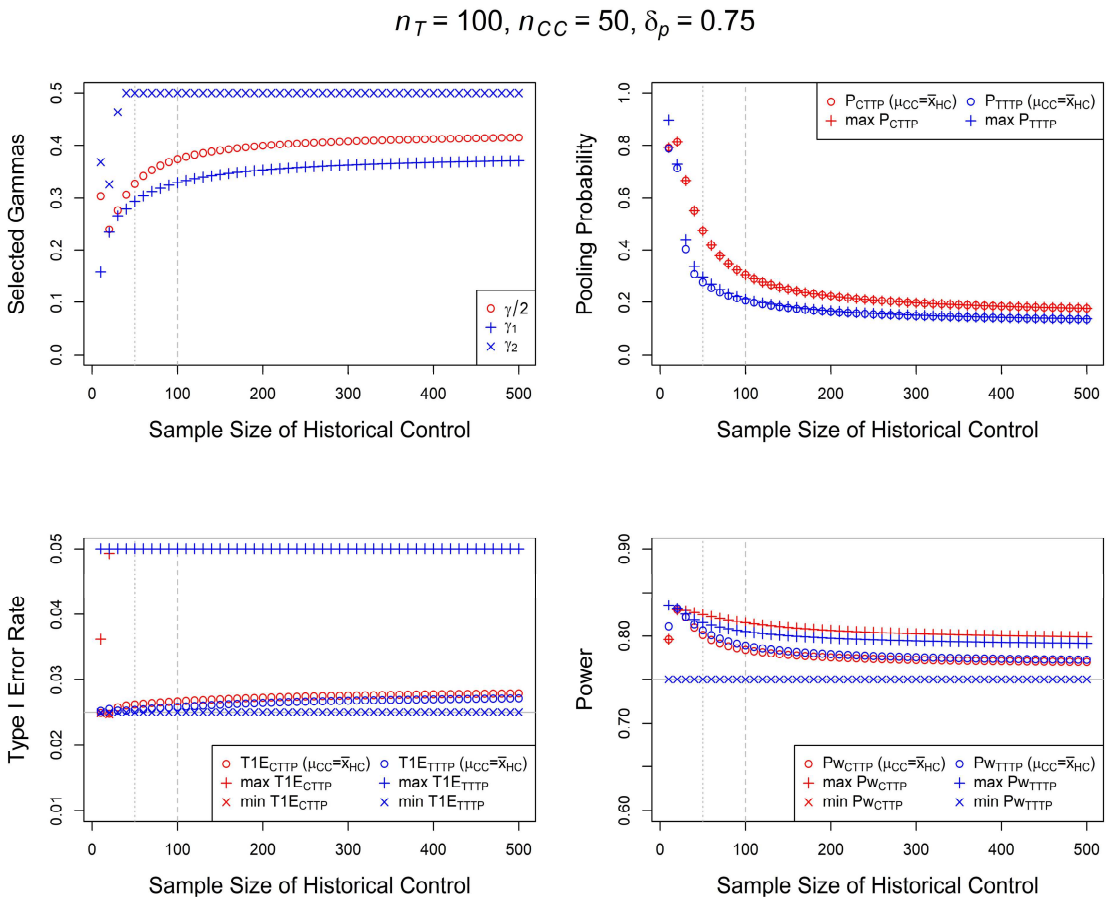

図 A1.20 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係
（$n_T = 100$，$n_{CC} = 50$，$\delta_p = 0.75$）

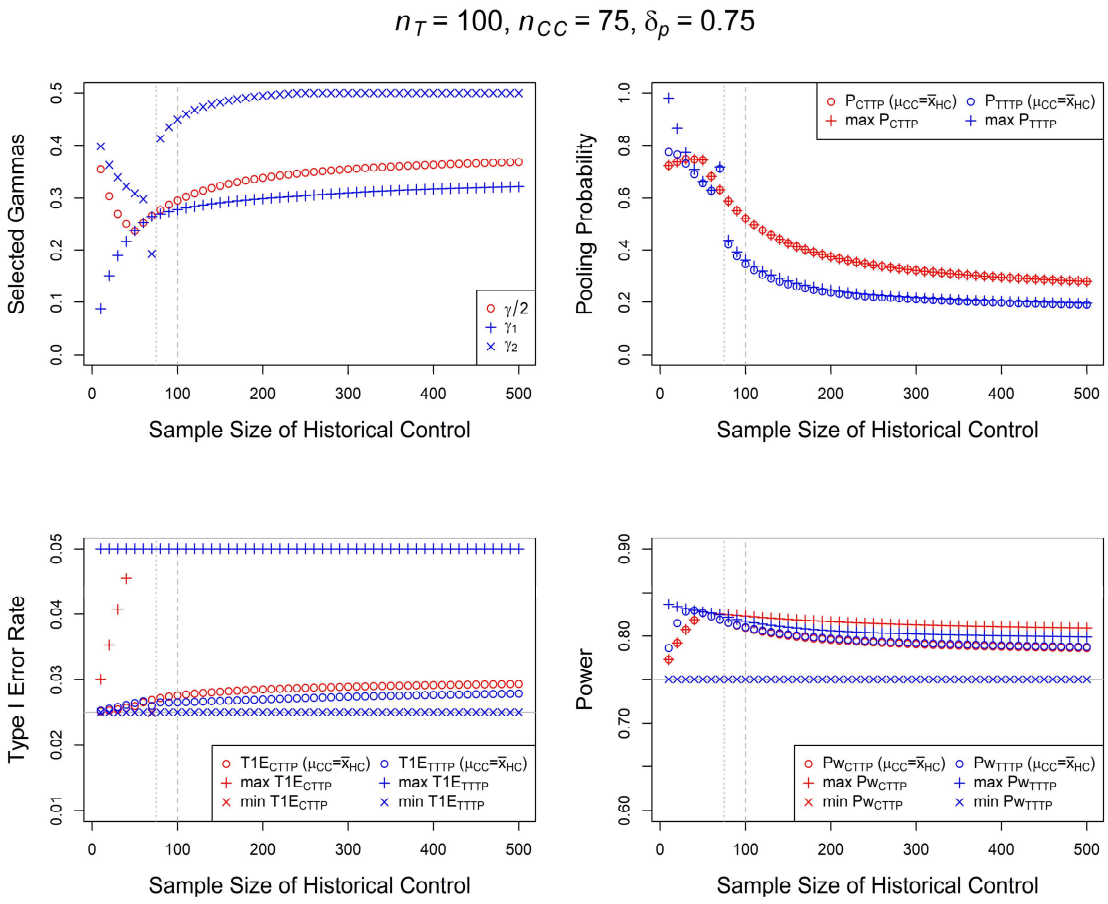

図 A1.21 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係
（$n_T = 100$，$n_{CC} = 75$，$\delta_p = 0.75$）

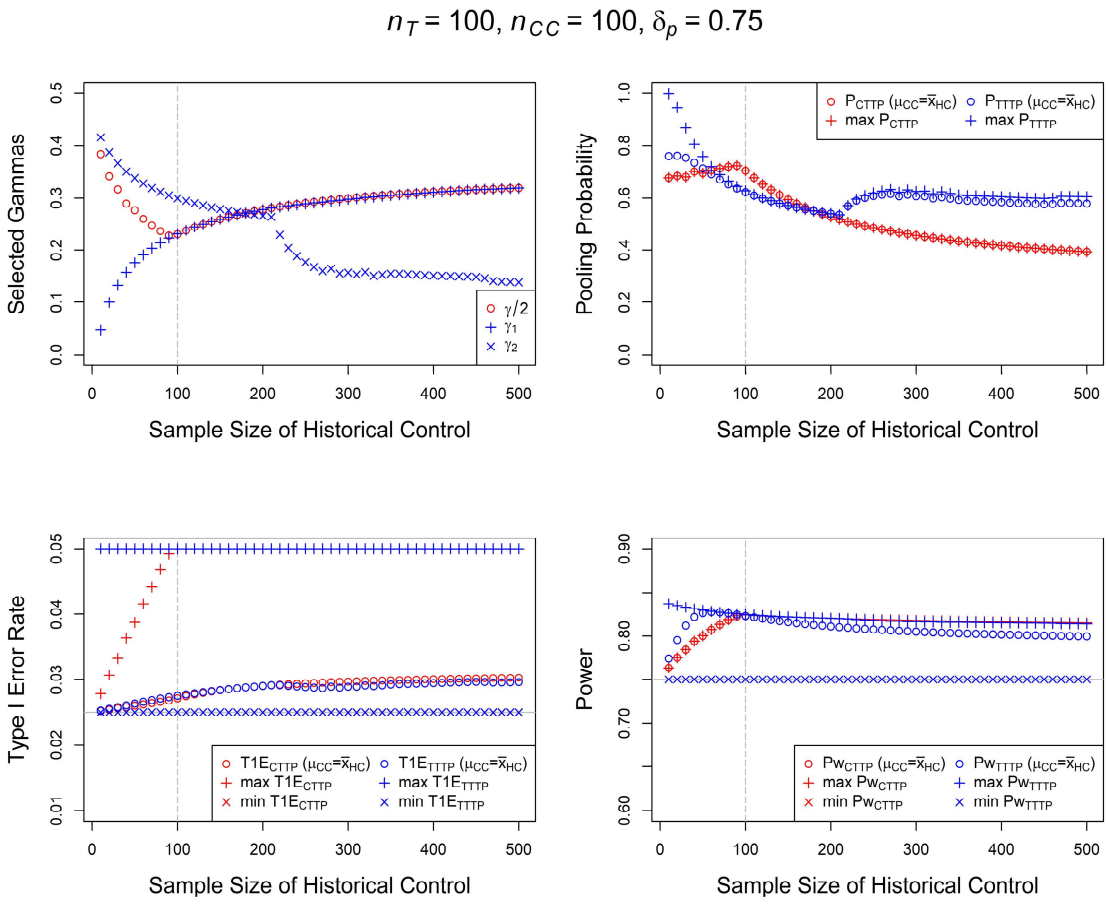

図 A1.22 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係
（$n_T = 100$，$n_{CC} = 100$，$\delta_p = 0.75$）

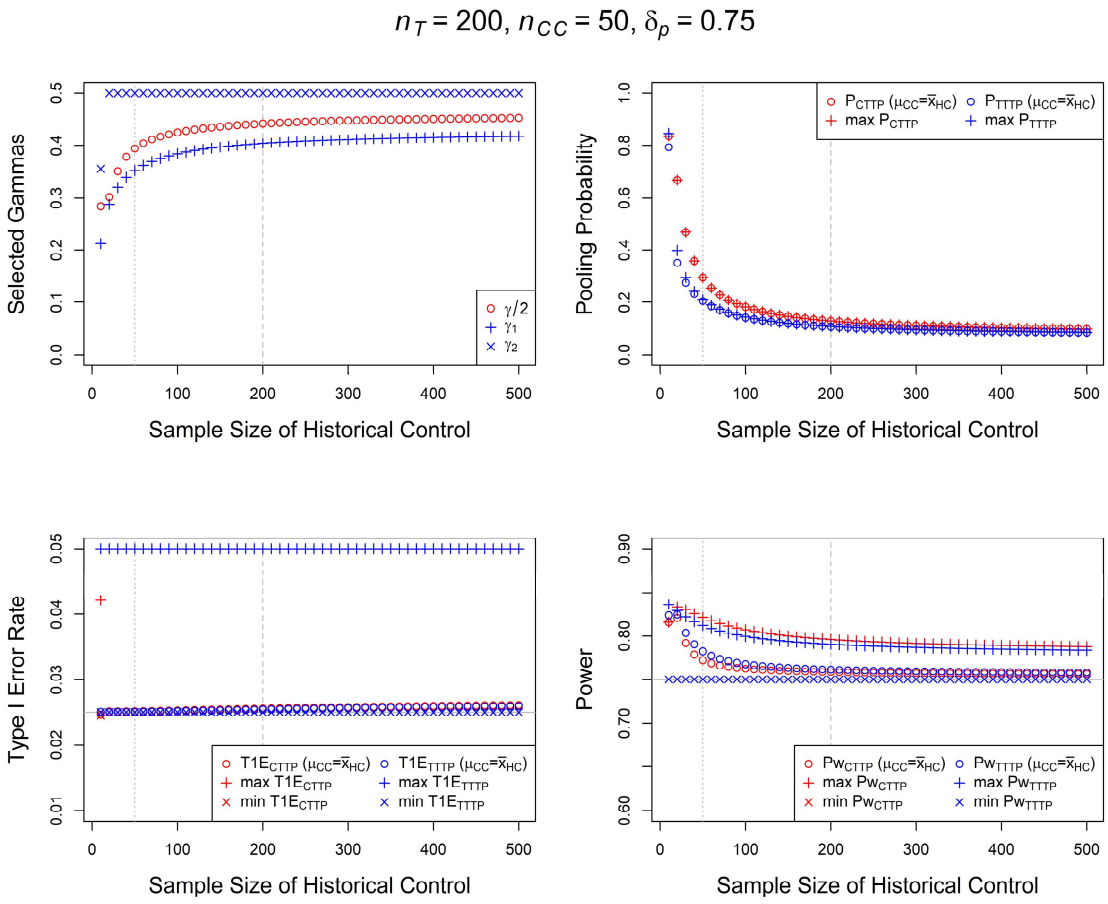

図 A1.23 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係
（$n_T = 200$，$n_{CC} = 50$，$\delta_p = 0.75$）

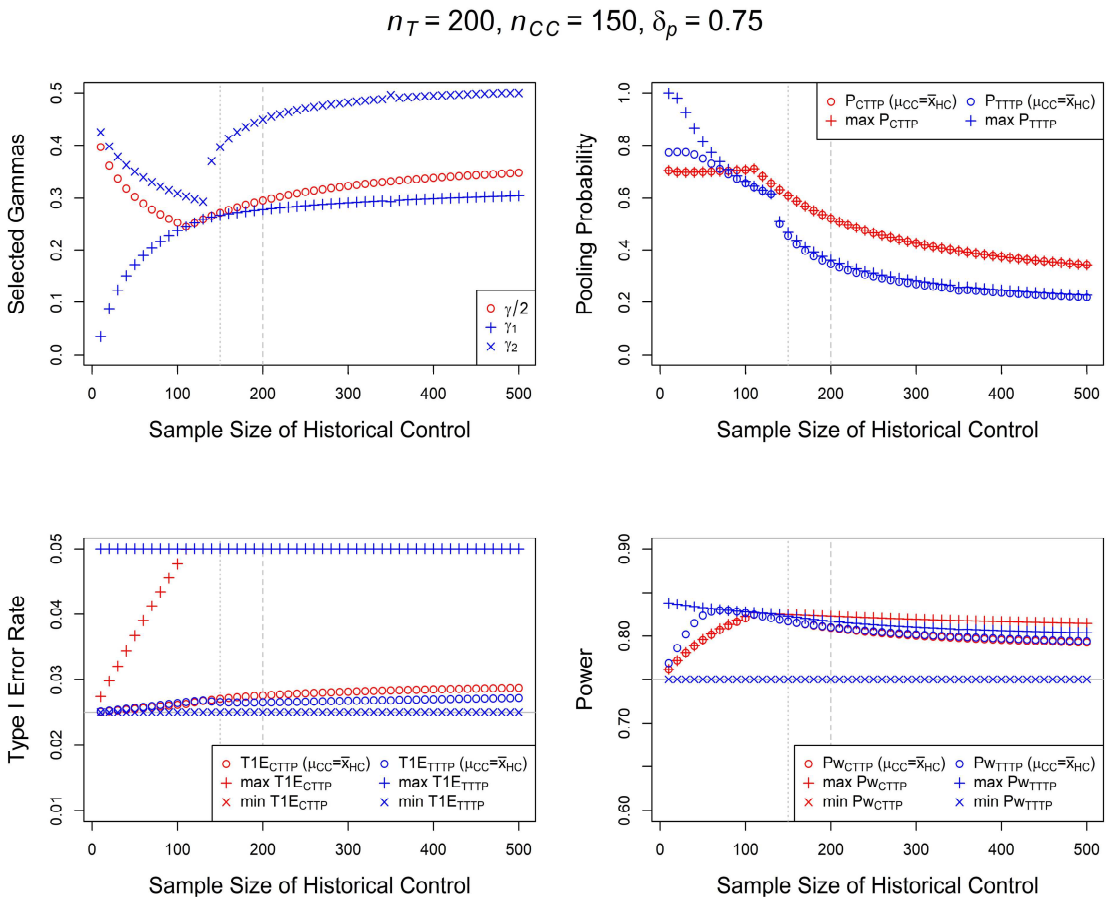

図 A1.24 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係
（$n_T = 200$，$n_{CC} = 150$，$\delta_p = 0.75$）

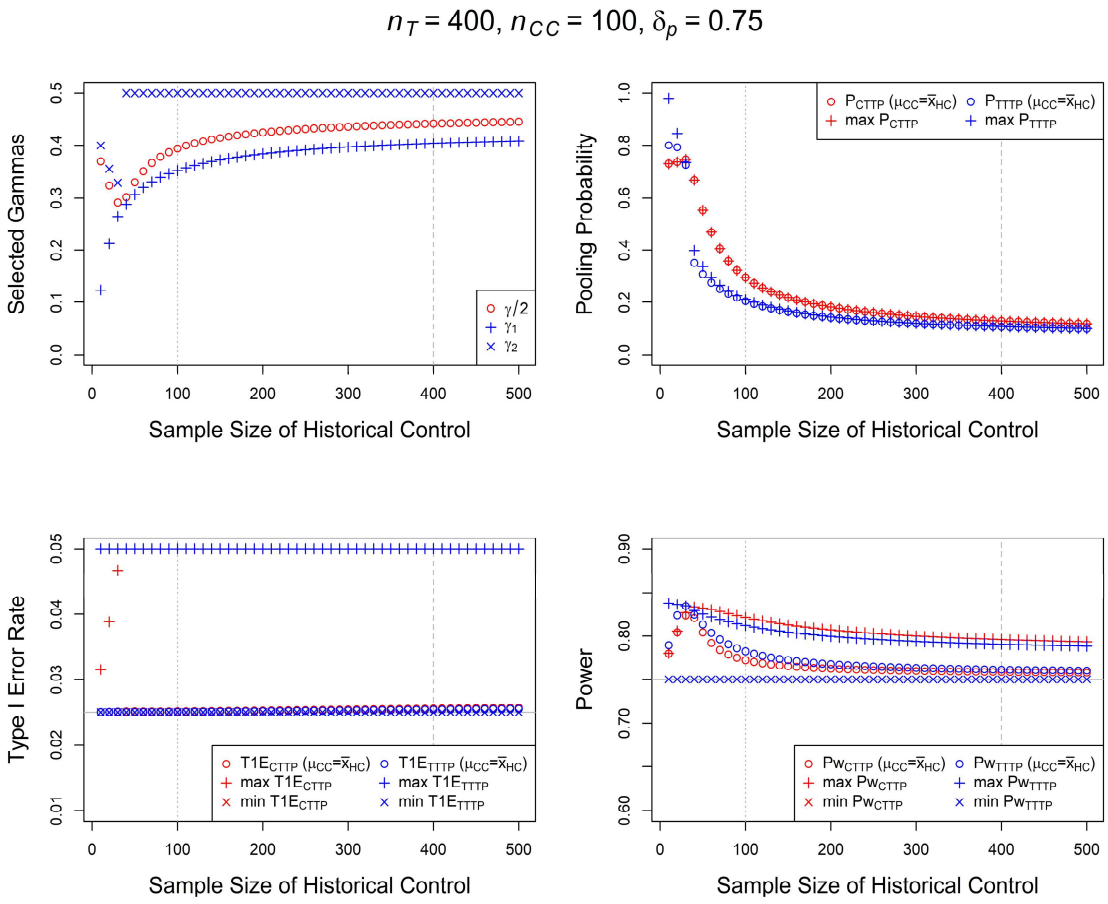

図 A1.25 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係
（$n_T = 400$，$n_{CC} = 100$，$\delta_p = 0.75$）

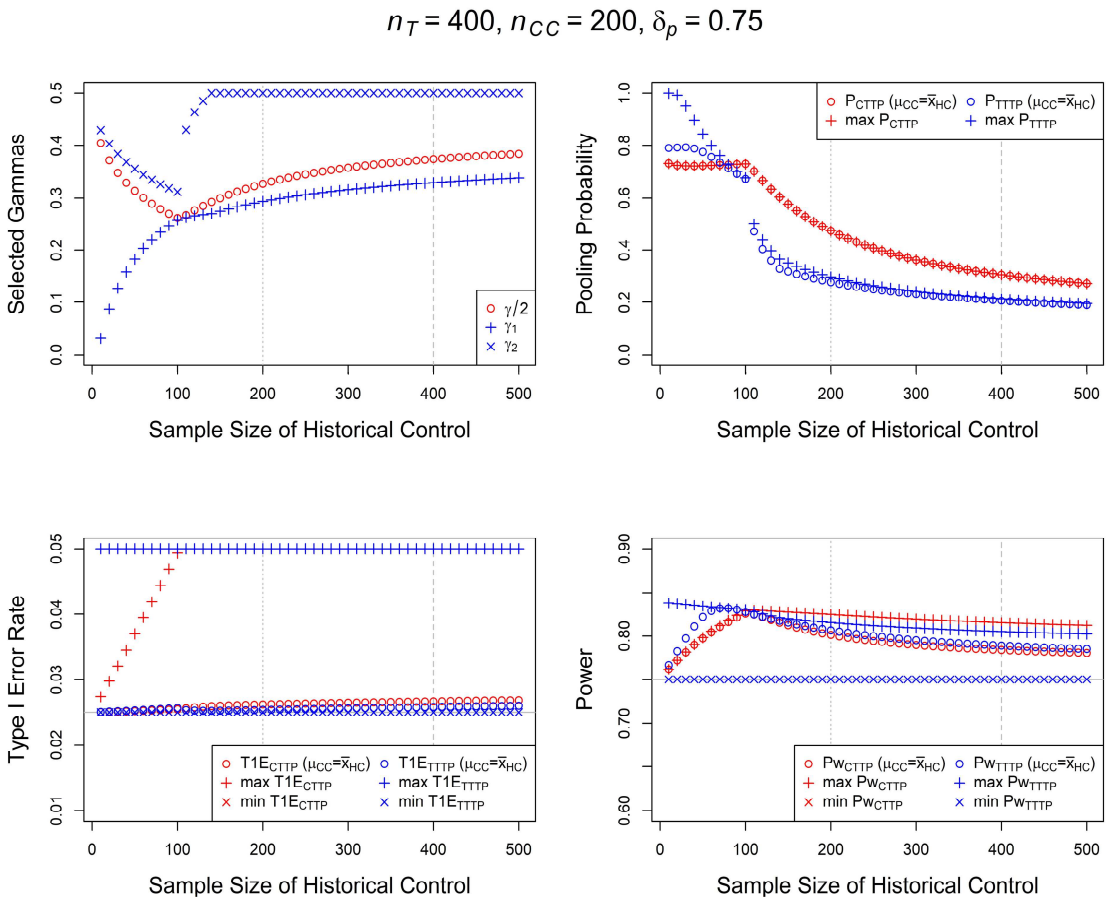

図 A1.26 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係
（$n_T = 400$，$n_{CC} = 200$，$\delta_p = 0.75$）

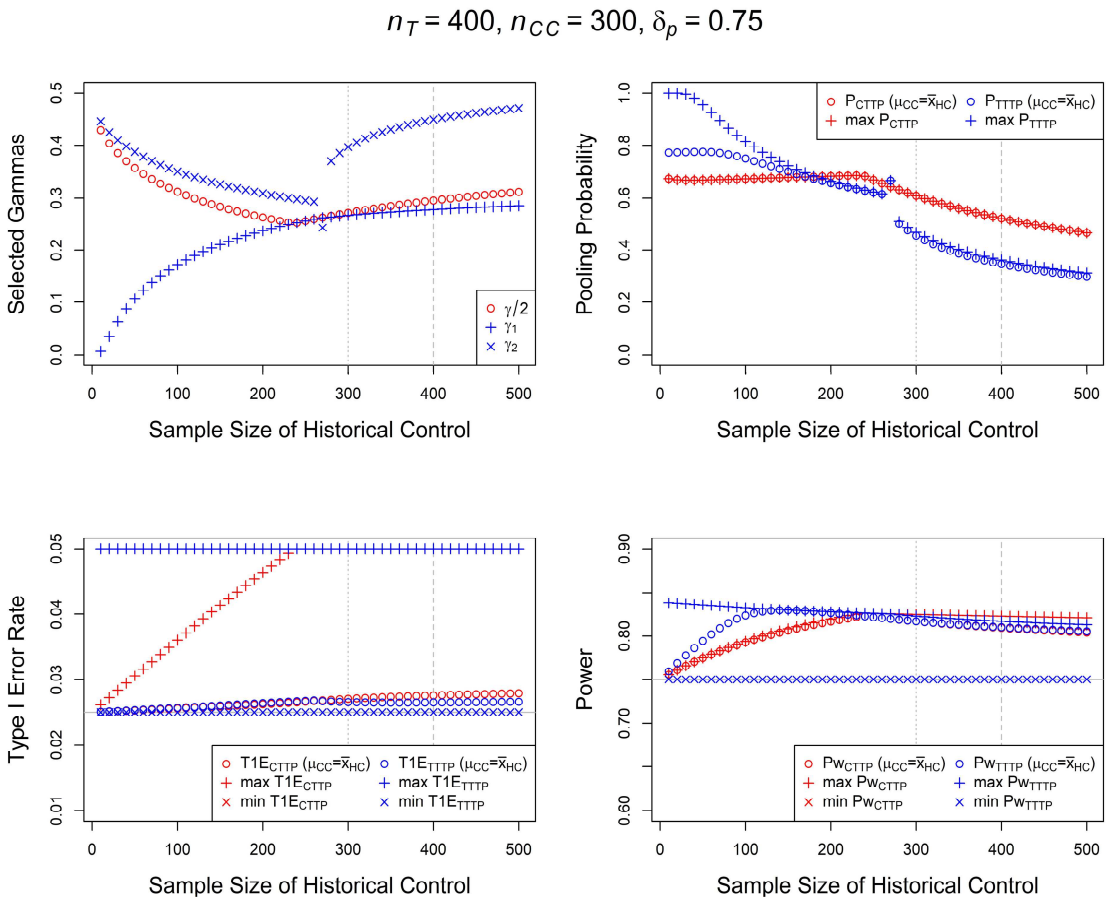

図 A1.27 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係
（$n_T = 400$，$n_{CC} = 300$，$\delta_p = 0.75$）

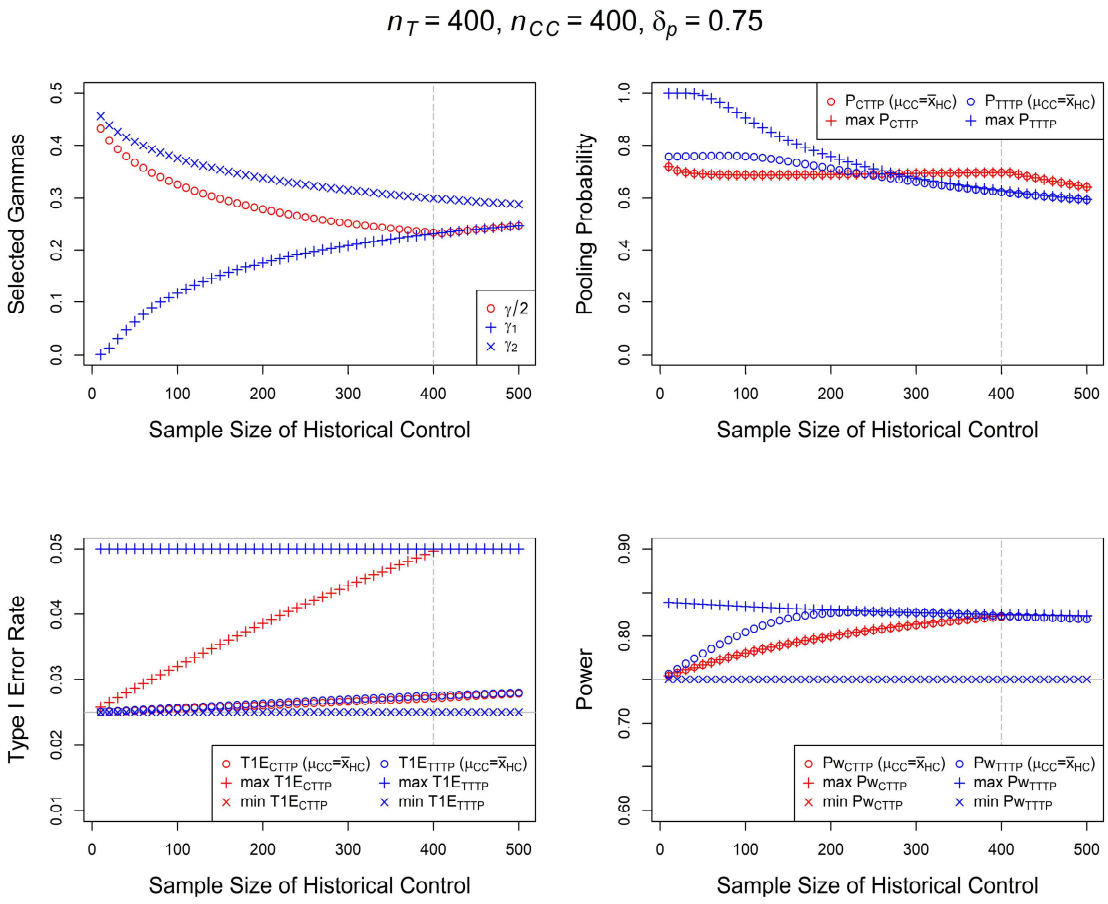

図 A1.28 既存対照のサンプルサイズと選択された有意水準及び動作特性の関係
（$n_T = 400$，$n_{CC} = 400$，$\delta_p = 0.75$）

付録 2　3.4 節の数値実験の結果

3.4 節の数値実験の結果をすべて示す．図 A2.1 から図 A2.4 は，本文に示した図 3.1 から図 3.4 の縮尺を調整しない，すべての点がプロットされた結果である．図 A2.4 から図 A2.16 は，本文に示さなかった$\gamma = 0.1, 0.3, 0.4$の結果であり，縦軸を拡大調整して示している．γは検定併合法の有意水準，n_Hは既存対照のサンプルサイズ，TTP は検定併合法，PSW は傾向スコア重み付け法，SPSW は安定化重みを用いた傾向スコア重み付け法，PSWTTP は傾向スコア重み付け法＋併合検定法，SPSWTTP は安定化重みを用いた傾向スコア重み付け法＋併合検定法，no borrowing は既存対照を用いない方法を表す.

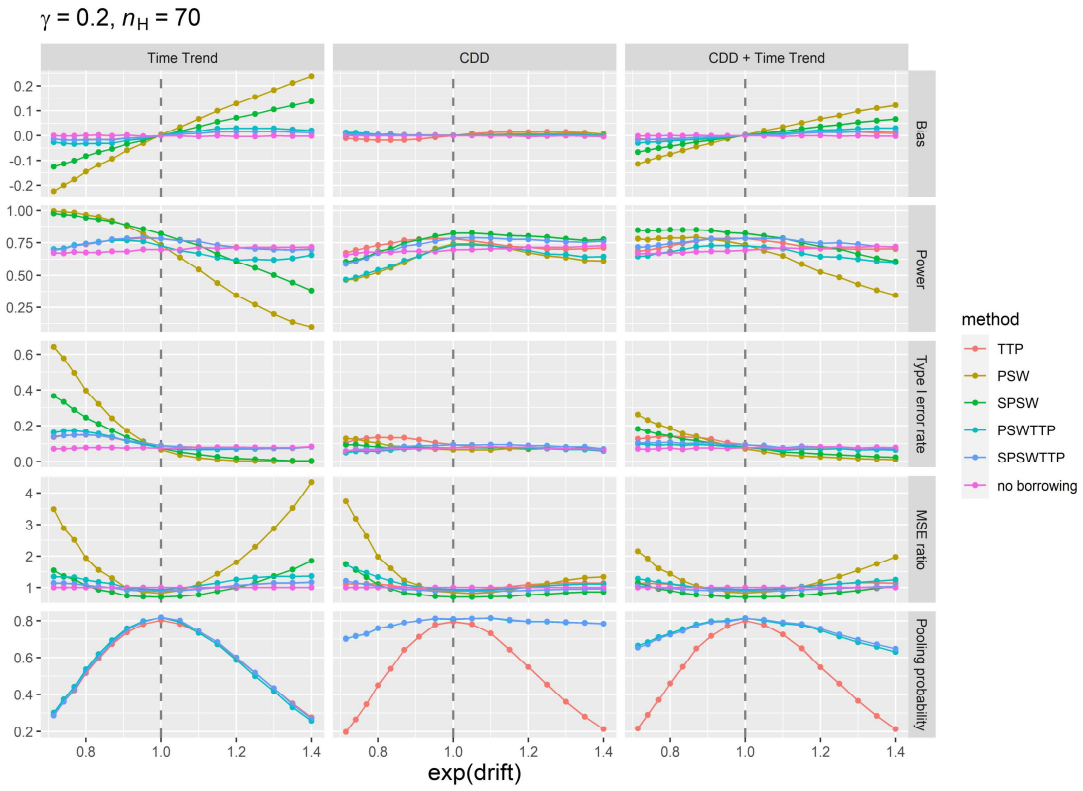

図 A2.1 推定された動作特性（$\gamma = 0.2$，$n_H = 70$）

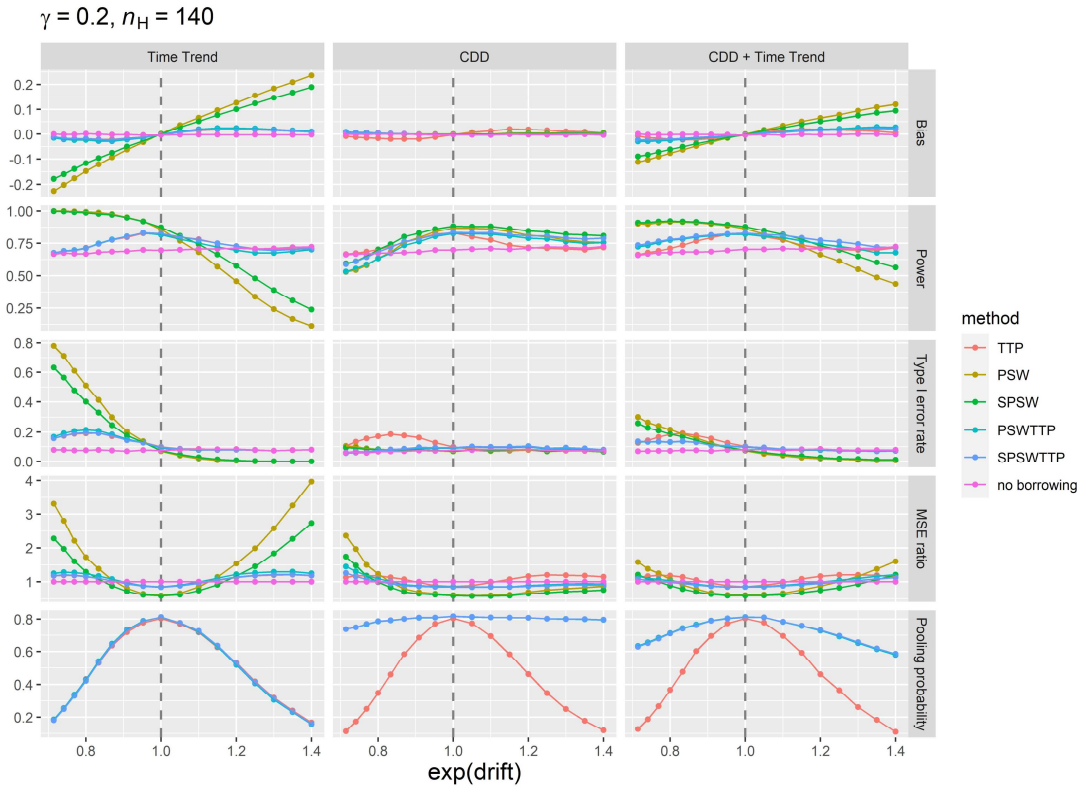

図 A2.2 推定された動作特性（$\gamma = 0.2$，$n_H = 140$）

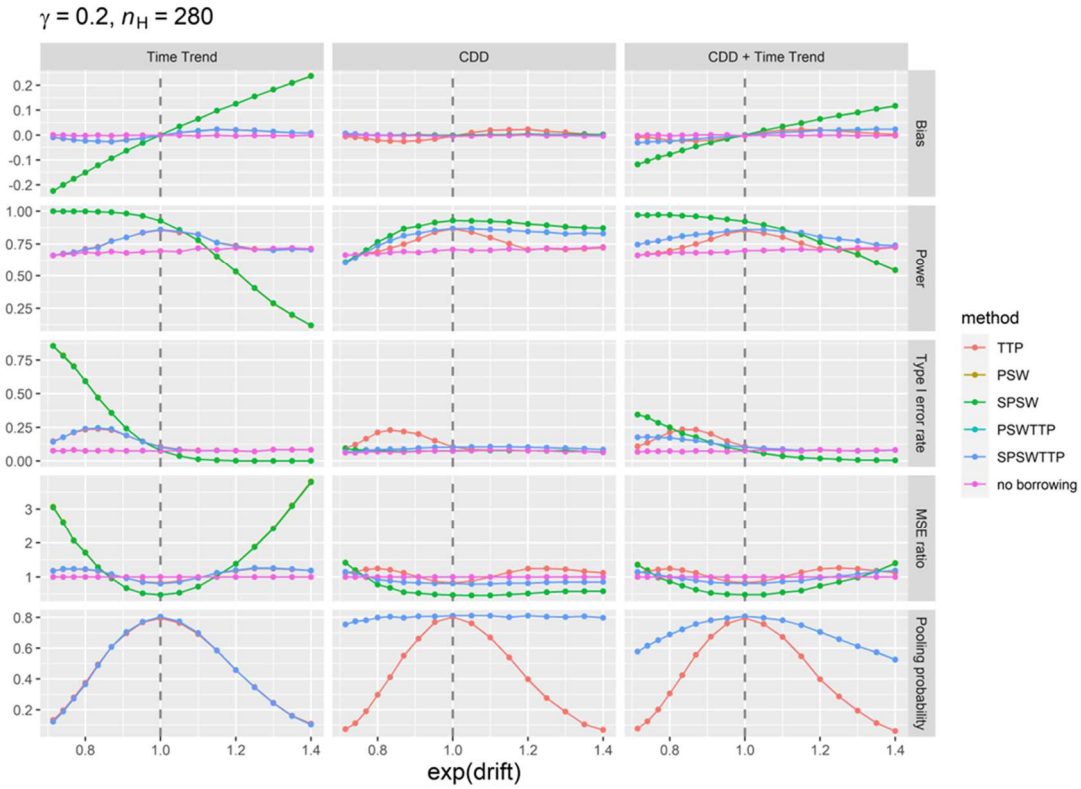

図 A2.3 推定された動作特性（$\gamma = 0.2$，$n_H = 280$）

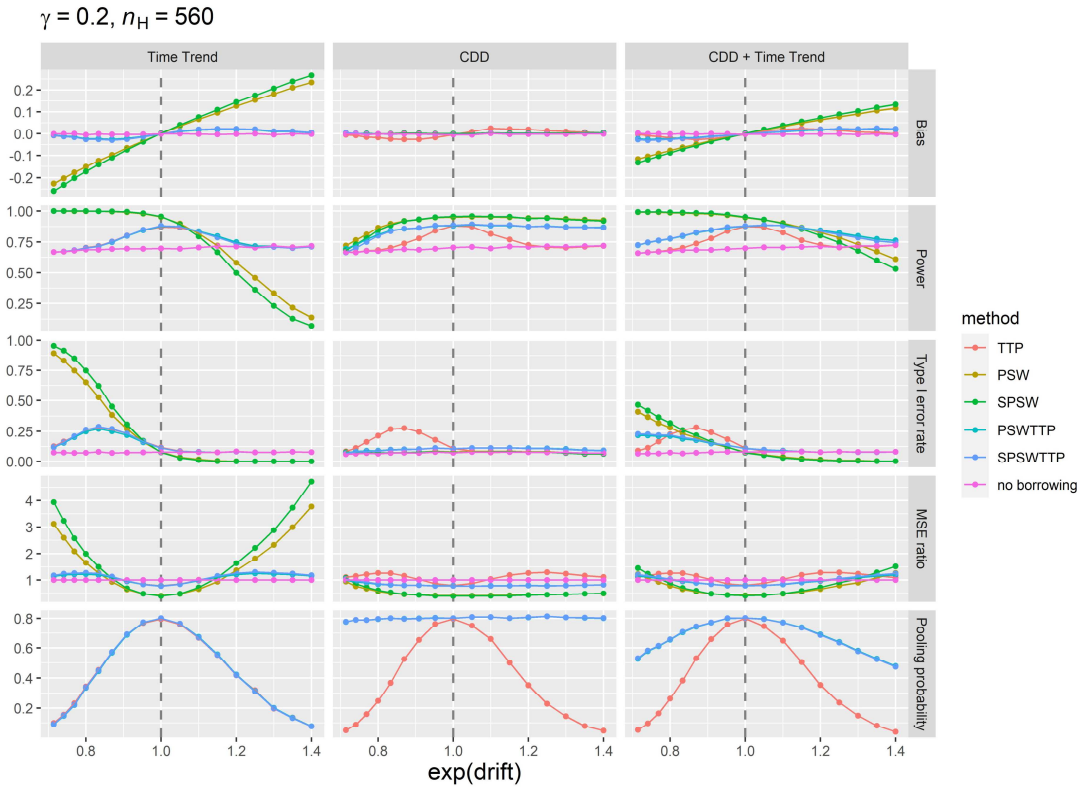

図 A2.4 推定された動作特性（$\gamma = 0.2$，$n_H = 560$）

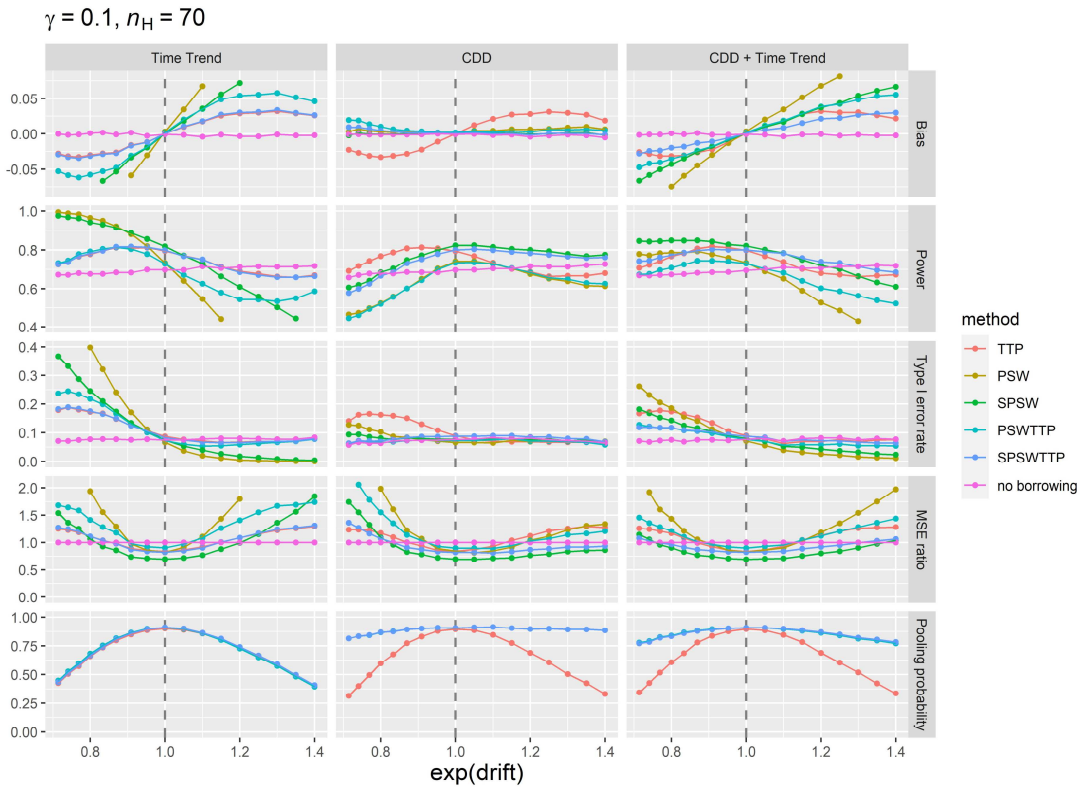

図 A2.5 推定された動作特性（$\gamma = 0.1$，$n_H = 70$，縦軸縮尺調整）

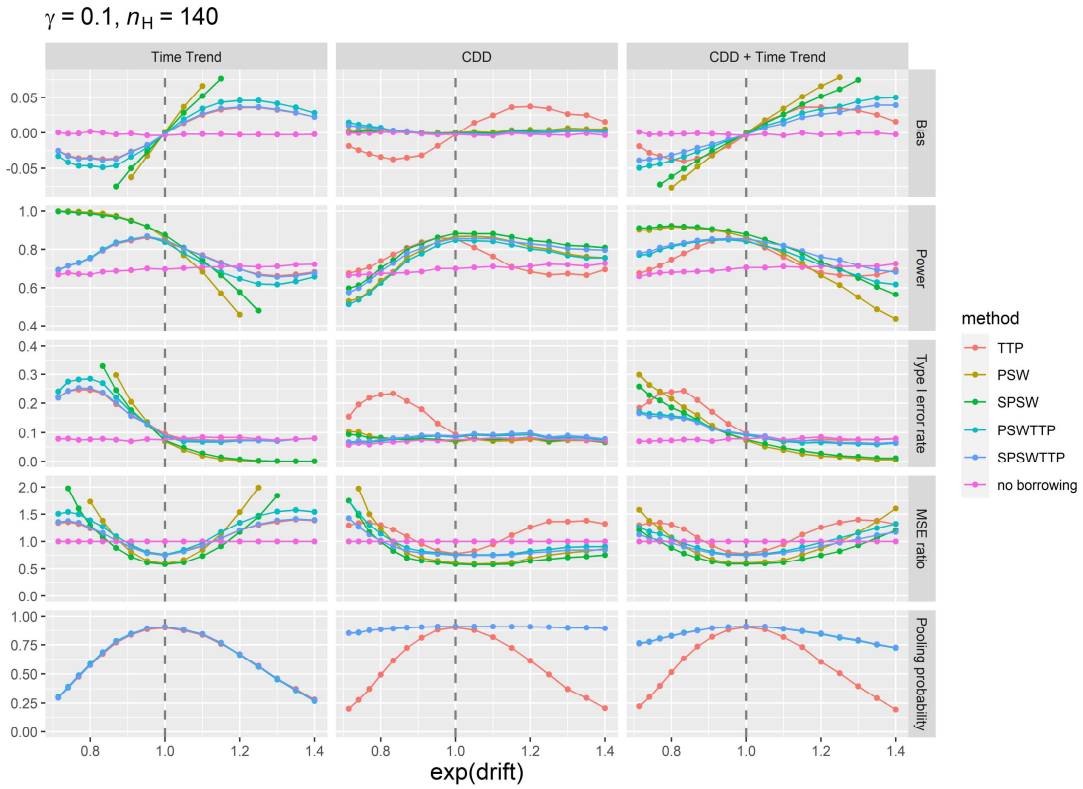

図 A2.6 推定された動作特性（$\gamma = 0.1$，$n_H = 140$，縦軸縮尺調整）

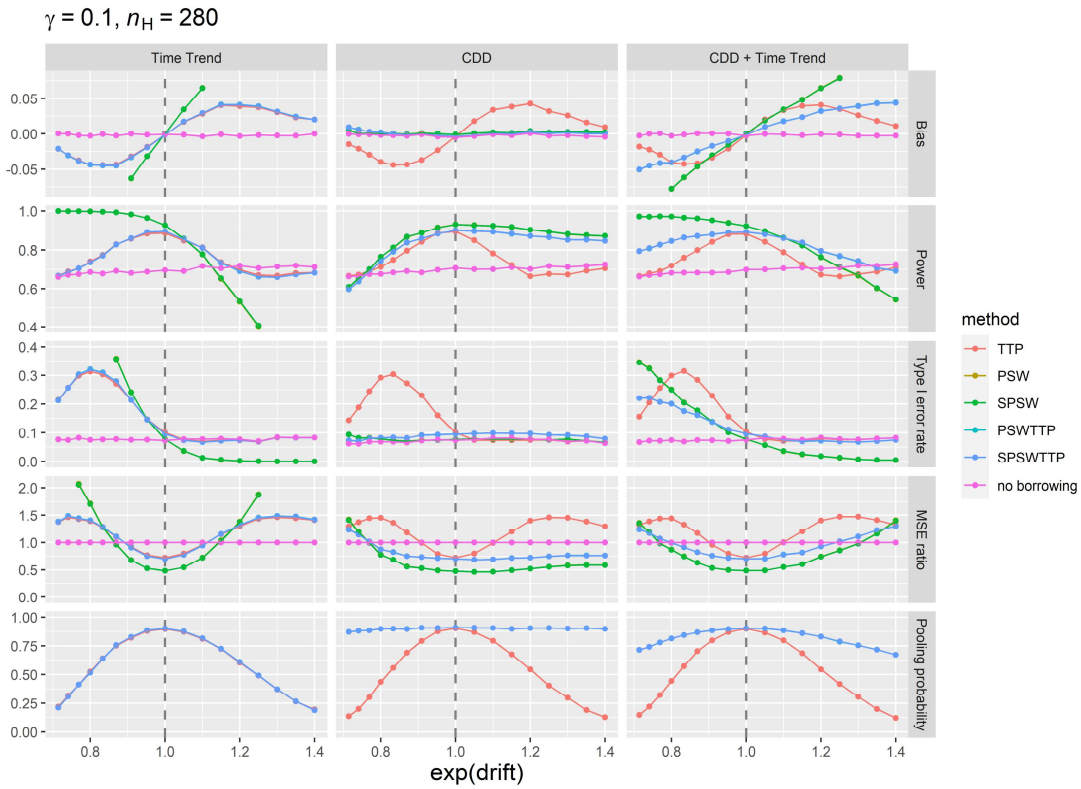

図 A2.7 推定された動作特性（$\gamma = 0.1$，$n_H = 280$，縦軸縮尺調整）

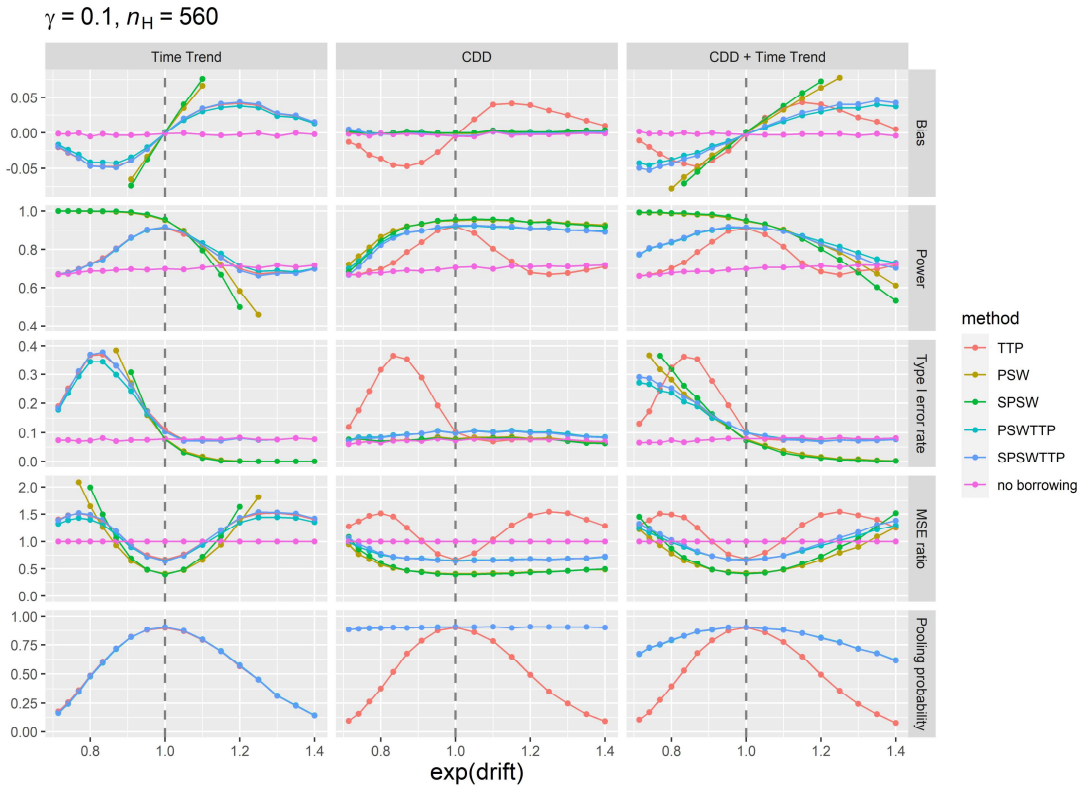

図 A2.8 推定された動作特性（$\gamma = 0.1$，$n_H = 560$，縦軸縮尺調整）

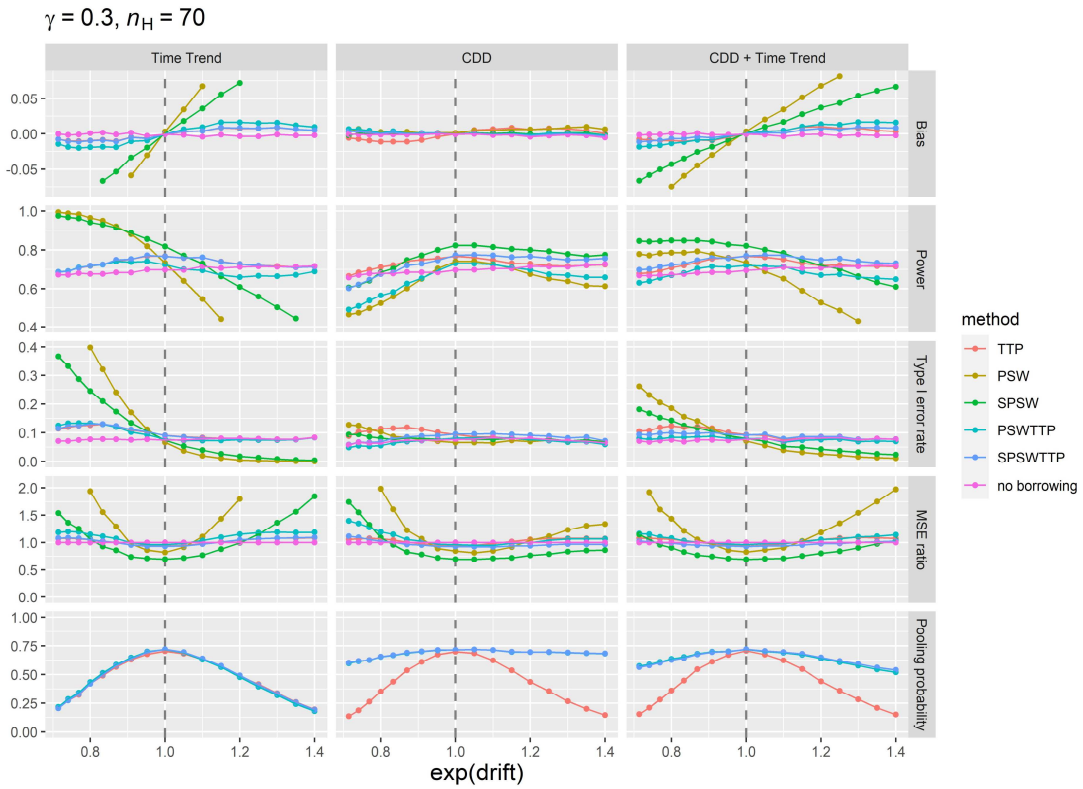

図 A2.9 推定された動作特性（$\gamma = 0.3$，$n_H = 70$，縦軸縮尺調整）

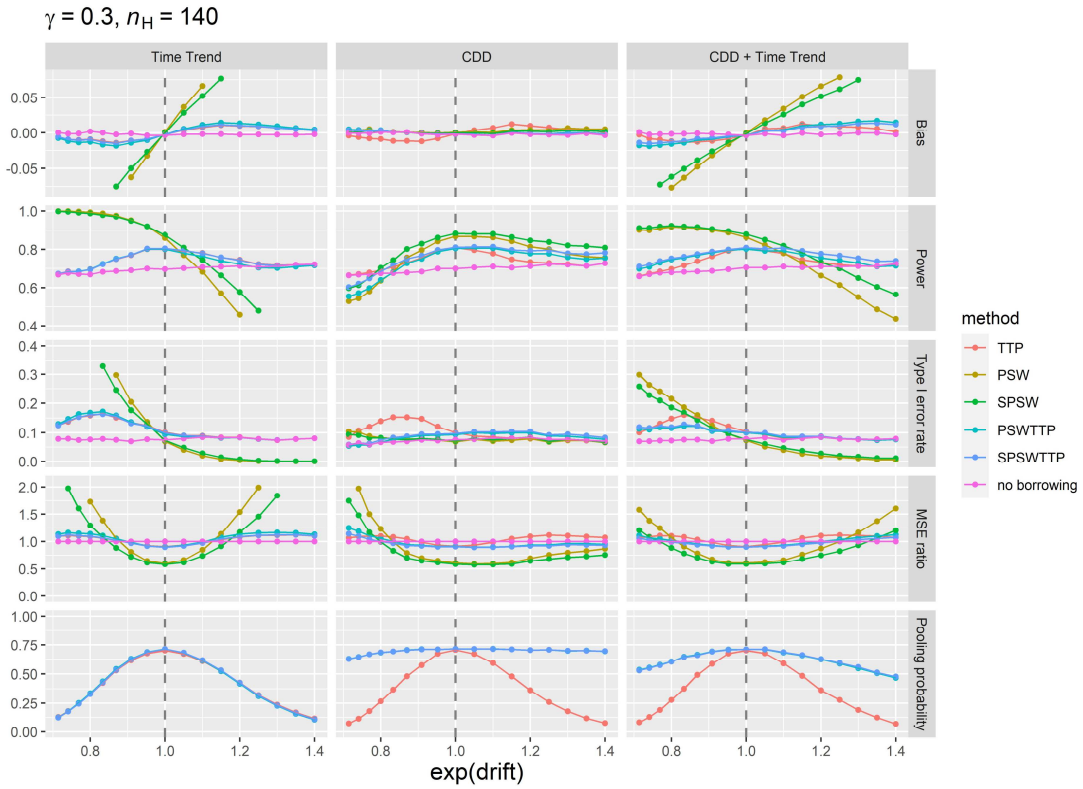

図 A2.10 推定された動作特性（$\gamma = 0.3$，$n_H = 140$，縦軸縮尺調整）

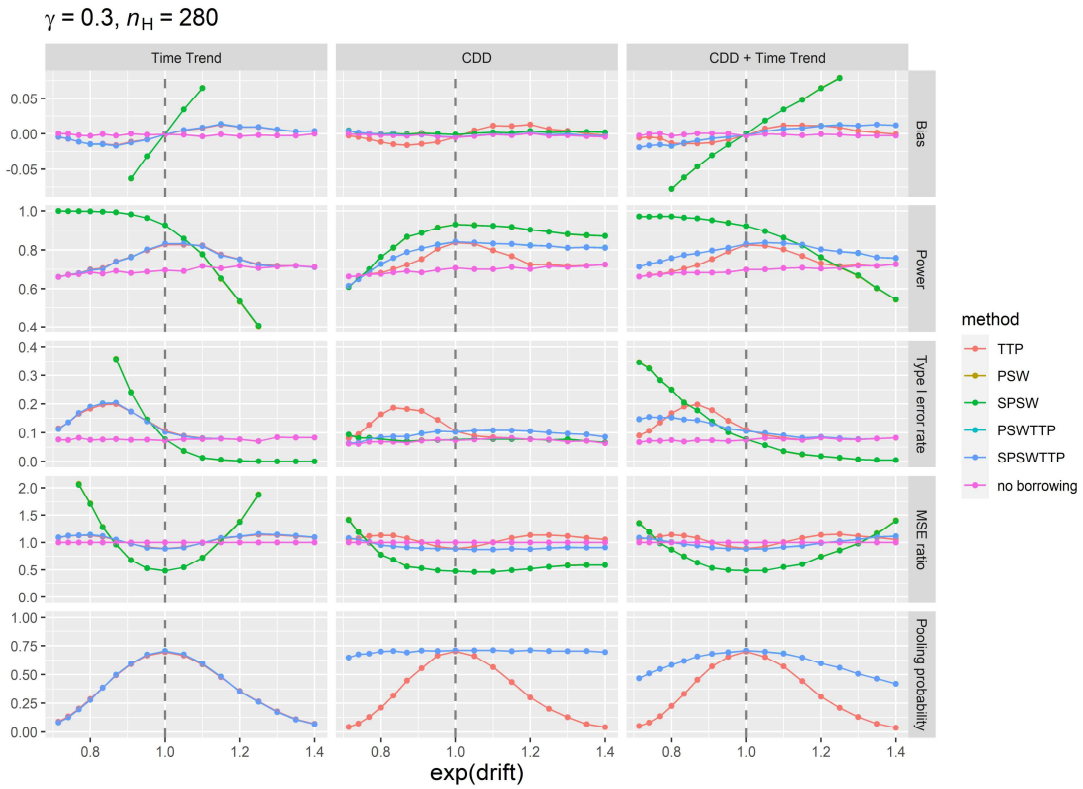

図 A2.11 推定された動作特性（$\gamma = 0.3$，$n_H = 280$，縦軸縮尺調整）

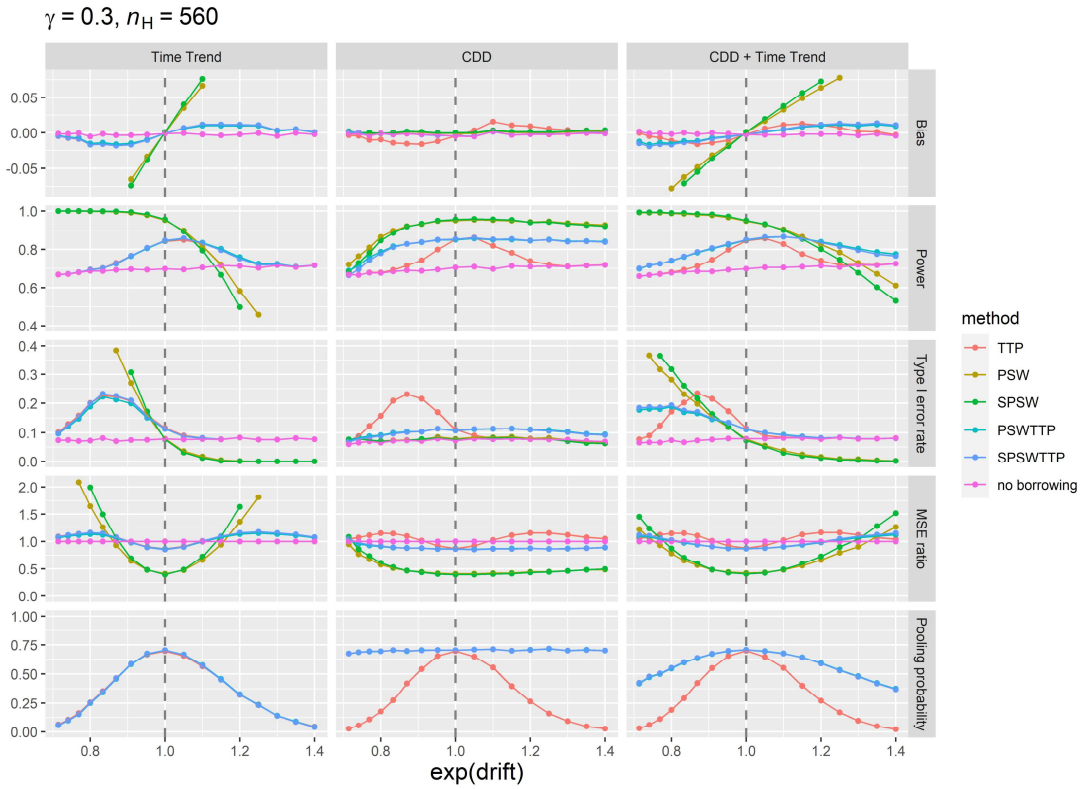

図 A2.12 推定された動作特性（$\gamma = 0.3$，$n_H = 560$，縦軸縮尺調整）

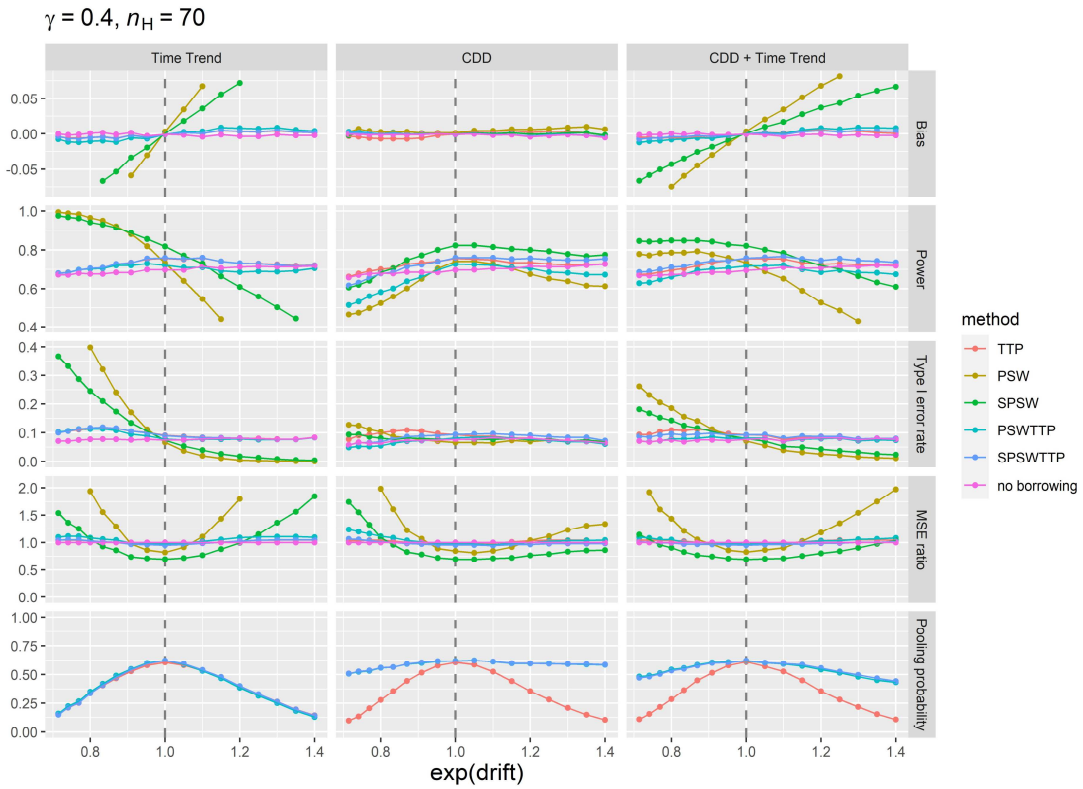

図 A2.13 推定された動作特性（$\gamma = 0.4$，$n_H = 70$，縦軸縮尺調整）

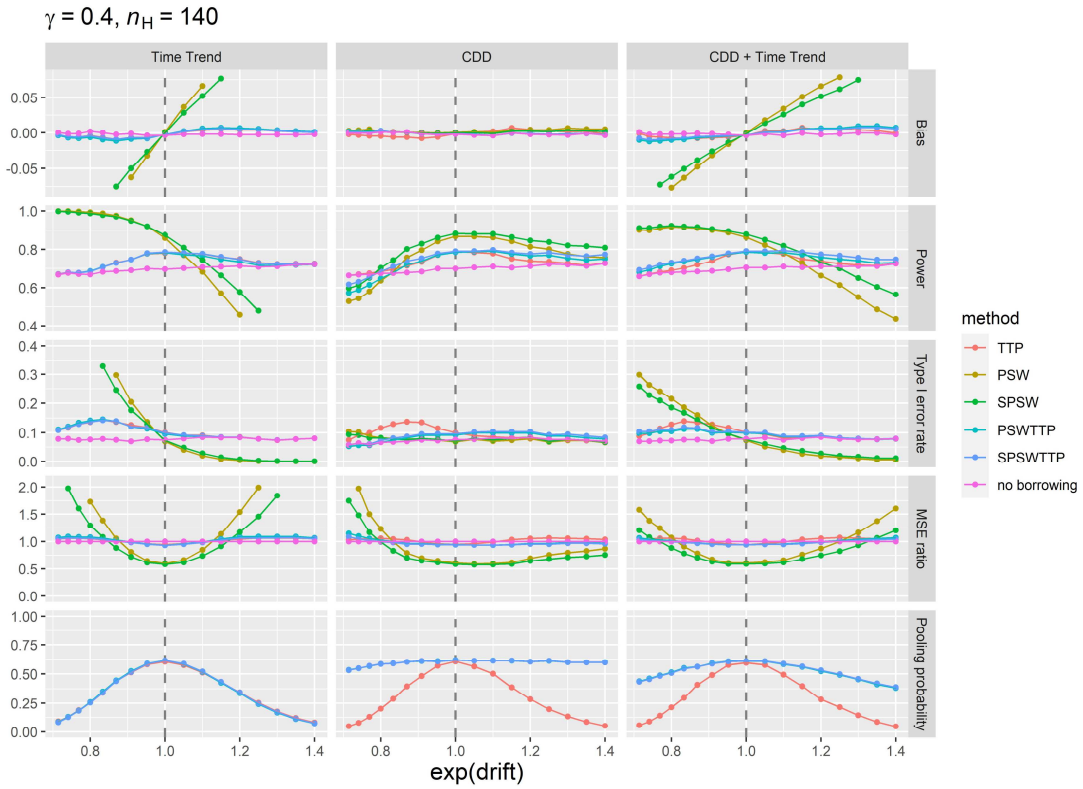

図 A2.14 推定された動作特性（$\gamma = 0.4$，$n_H = 140$，縦軸縮尺調整）

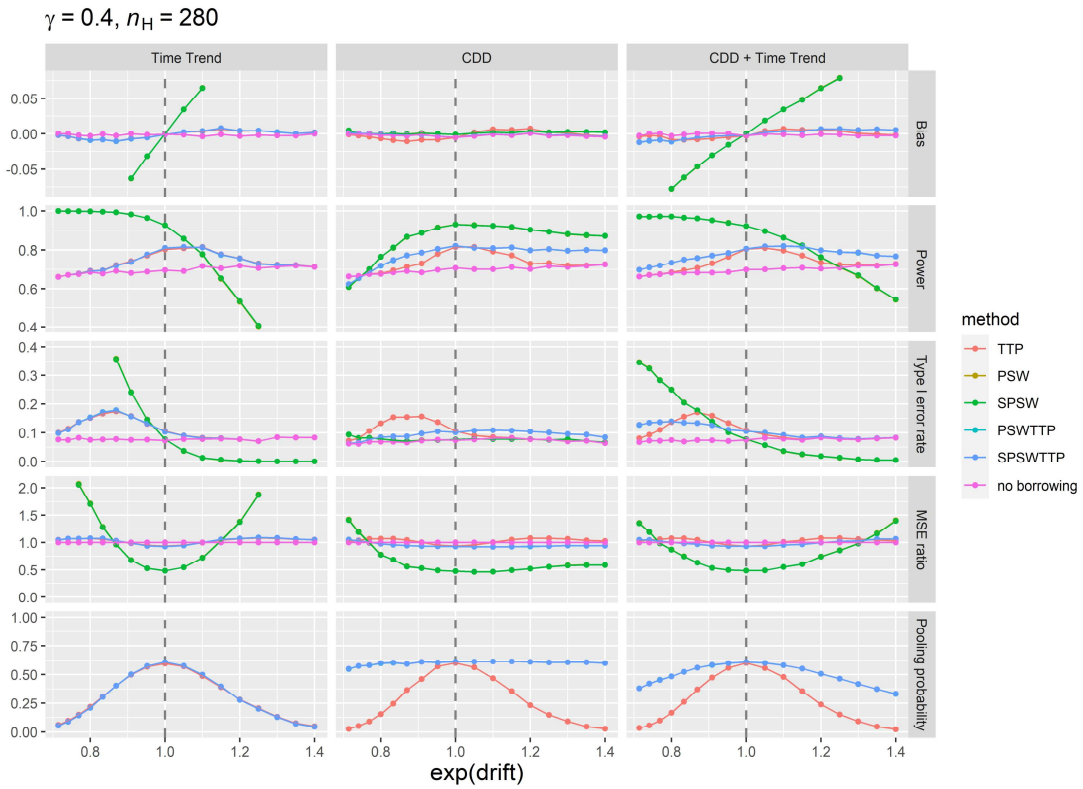

図 A2.15 推定された動作特性（$\gamma = 0.4$，$n_H = 280$，縦軸縮尺調整）

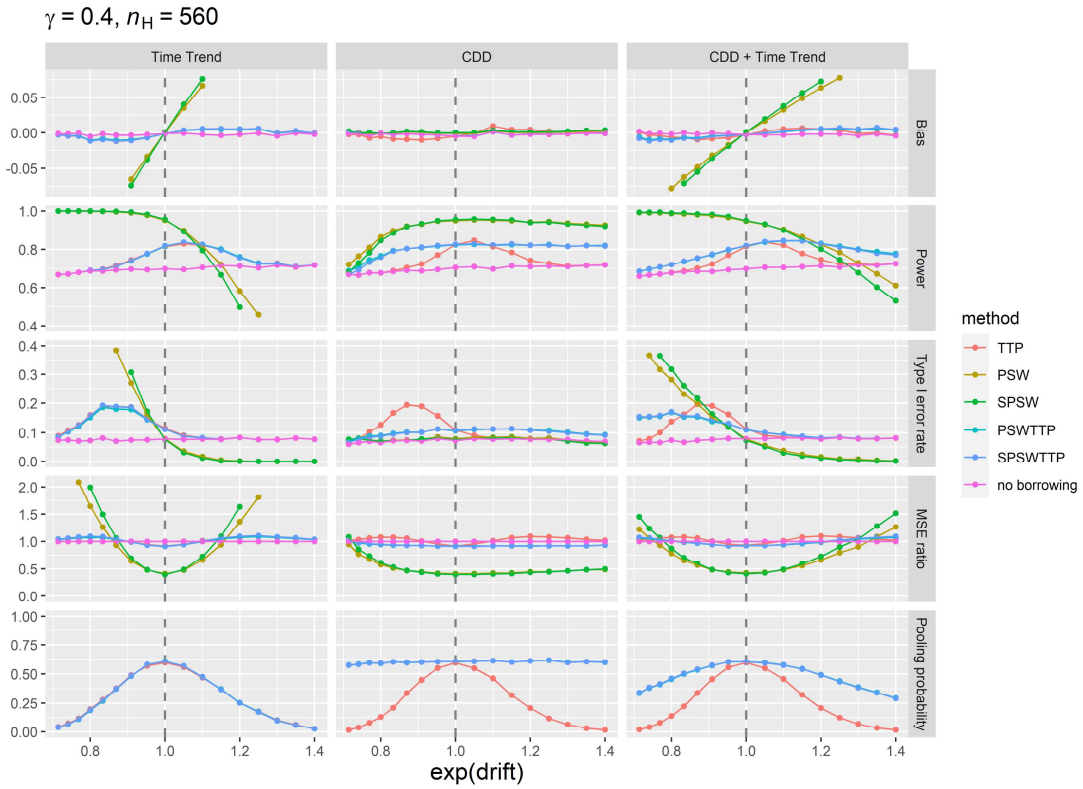

図 A2.16 推定された動作特性（$\gamma = 0.4$，$n_H = 560$，縦軸縮尺調整）