総 説

主要評価変数が複数ある臨床試験の統計的諸問題 Statistical Issues in Clinical Trials with Multiple Primary Endpoints

寒水孝司^{*1}・杉本知之^{*2}・濱崎俊光^{*3}

Takashi Sozu^{*1}, Tomoyuki Sugimoto^{*2} and Toshimitsu Hamasaki^{*3}

^{*1}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

^{*2}弘前大学大学院 理工学研究科 数理科学

^{*3}大阪大学大学院 医学系研究科 内科系医学専攻 医学統計学

^{*1}Department of Biostatistics,
Kyoto University School of Public Health

^{*2}Department of Mathematical Sciences,
Hirosaki University Graduate School of Science and Technology

^{*3}Department of Biomedical Statistics,
Osaka University Graduate School of Medicine
e-mail:sozu.takashi.4s@kyoto-u.ac.jp

Clinical trials often employ two or more primary endpoints because a single endpoint may not provide a comprehensive picture of the intervention's effects. In such clinical trials, a decision is generally made as to whether it is desirable to evaluate the joint effects on *all* endpoints (i.e., co-primary endpoints) or *at least one* of the endpoints. This decision defines the alternative hypothesis to be tested and provides a framework for approaching trial design.

In this article, we discuss recent statistical issues in clinical trials with multiple primary endpoints. Especially, we introduce the methods for power and sample size determinations in clinical trials with co-primary endpoints, considering the correlations among the endpoints into the calculations. We also discuss the methods to alleviate conservativeness of testing co-primary endpoints.

Key words: clinical trials, drug development, multiplicity, power, primary endpoint, sample size.

1. はじめに

臨床試験では、試験治療の有効性や安全性を多面・包括的に評価するために、臨床症状、生理学的所見、血液検査、有害事象など、複数の評価指標が観察される(O'Brien 1984; Pocock, Geller, and Tsiatis, 1987; Cook and Farewell, 1996; Pocock, 1997). とくに、近年の診断技術の向上や薬物療法の多様・複雑化により、多くの評価指標が観察される傾向がある(Bloch et al., 2007). 通常の臨床試験では、そのような複数の評価指標から主要評価変数(以下、主要変数)を1つ選択し、

その変数に基づいて、試験の参加者数(以下、標準サイズ)や主解析などの試験デザインを設計するのが一般的である。とくに「臨床試験のための統計的原則(ICH, 1998; 以下「ICH 統計ガイドライン」と呼ぶ)」では、検証的臨床試験における主要変数をできるだけ1つにすることを原則としている。検証すべき仮説はできるだけ絞られていることや、統計的推測における多重性はできるだけ回避するのが望ましいからである(森川・平山, 2006)。

ところが、ICH 統計ガイドラインで述べられているように、複数の主要変数を用いるのが望ましいことがある。その主な理由は、つぎの3つである(Pocock, 1997; Sankoh, D'Agostino, and Huque, 2003; Chuang-Stein et al., 2007)。

- (1) 疾患の状態や治療効果の評価が多次的に特徴づけられる
- (2) 主要変数の重要性に関する医学的見解が一致していない
- (3) 疾患の原因が明らかでない

(1)の理由に関する典型的な疾患として、アルツハイマー病がある。実際に、いくつかの臨床試験では、認知機能に関する評価項目(例: Alzheimer's Disease Assessment Scale - cognitive subscale: ADAS-cog)と全般症状に関する評価項目(例: Clinician's Interview-Based Impression of Change plus caregiver input: CIBIC-plus)の2つを主要変数にしている(Rogers et al., 1998; Rösler et al., 1999; Tariot et al., 2000; Peskind et al., 2006)。米国研究製薬工業協会は、アルハイマー病を含む20の疾患を例示して、規制当局が複数の主要変数について「試験治療の対照治療に対する有効性(以下、単に「有効性」と呼ぶ)」を示すことを要求したことを報告している(Offen et al., 2007)。このような背景から、いくつかの疾患領域では、ガイドライン等で使用すべき主要変数が推奨されている(CHMP, 2008; FDA, 2007, 2012, 2013)。

複数の主要変数を用いる臨床試験では、試験の主要な目的(以下、単に「試験目的」と呼ぶ)に応じて、複数の主要変数の扱い方、すなわち結論の導き方が異なる。このとき、複数の主要変数の扱いは、合成変数に統合する場合とそれ以外に大別される。合成変数を使用することの利点は、統計的推測における多重性の問題を回避できることである。さらに、いくつかのイベントから構成される複合型の合成変数については、各イベントに対する治療効果や発生頻度に依存して、イベント数の増加により、事象時間データの解析のもとで試験の検出力が向上することがある。たとえば、心血管疾患の領域では、死亡、非致死的心筋梗塞、再血行再建術といったイベントから構成されるMACE (Major Adverse Cardiac Event)という合成変数が使用されることがある。合成変数を使用することの欠点は、解釈が難しくなることや誤解を招きやすいことである(Ferreira-Gonzalez et al., 2007a, b)。さらに、近年では、実際の臨床試験の報告のレビューにより、合成変数の定義が不明であったり、不適切であったりすることの問題が指摘されている(Lim et al., 2008; Cordoba et al., 2010; 佐藤, 2007, 2011)。

一方で、合成変数を使用せずに、複数の主要変数をそれぞれ解析する場合は、複数の仮説を扱うことによる多重性の問題への対応を検討する必要がある。このとき、多重性の問題の性質や程度は、試験目的、主要変数の数、主要変数間の治療効果や重要度の違い、主要変数間の関連の大きさなどに依って異なってくる。このような多重性の問題を扱う統計学的方法については、1980

年頃から、少なくとも1つの主要変数について有効性を示すことを目的として、第一種の過誤確率を調整する方法が多く提案・使用されている。2005年頃からは、すべての主要変数について有効性を示すことを目的として、主要変数間の関連を考慮した標本サイズ設計や検定の保守性を緩和させる方法が提案されている。

本稿では、すべての主要変数について有効性を示すことを目的とする臨床試験を中心に、最近の統計学的方法を整理する。とくに、標本サイズ設計に関して、著者らを中心とした最近の研究成果を紹介する。2節では、複数の主要変数の扱いを試験目的と統計的方法論によってそれぞれ分類し、多重性の問題を整理する。3節では、すべての主要変数について有効性を示すことを目的とする臨床試験の標本サイズ設計について、最近の研究成果を整理・解説する。4節では、検定の保守性を緩和させる方法の考え方を示し、提案されている2つの方法の数理的関係を示す。5節では、今後の研究の方向性や拡張性について述べる。

2. 複数の主要変数の扱い

2.1 試験目的による分類

合成変数を使用せずに、複数の主要変数をそれぞれ解析する場合は、試験目的をつぎの3に分けることができる(Sankoh et al., 2003)。

- (1) 少なくとも1つの主要変数の有効性を示す
- (2) 一部分の主要変数の有効性を示す
- (3) すべての主要変数の有効性を示す

実際の臨床試験では、(1)と(3)のどちらかを試験目的にすることが多い。この2つの試験目的は、表1に示すように、試験デザインや解析における多重性の問題の内容が異なる。

試験目的(1)のもとでは、第一種の過誤確率を適切に調整したもとの、どの主要変数であっても有効性が1つでも示されれば、試験治療と対照治療の有効性に違いがあると解釈される。第一種の過誤確率を調整する方法としては、 p 値を調整する方法や包括的な検定など多くの方法があり、SASやRなどのソフトウェアで利用できる(Westfall, Tobias, and Wolfinger, 2011; Bretz,

表 1. 試験目的(1)と(3)における多重性の問題の内容

	試験目的(1)	試験目的(3)
主要変数の名称	(Alternative) Primary endpoints ^[1,2]	Co-primary endpoints ^[1,2]
前提とする仮説検定	Union-intersection 検定	Intersection-union 検定
第1種過誤の調整	必要	不要
第2種過誤の調整	必要	必要
多重性の問題の名称	多重性問題 (Multiplicity problem)	逆多重性問題 (Reverse multiplicity problem ^[1])
検出力の名称	Minimal power ^[3] Disjunctive power ^[4,5]	Complete power ^[3] Conjunctive power ^[4,5]

[1] Offen et al. (2007), [2] Dmitrienko, Tamhane, and Bretz (2010), [3] Westfall, Tobias, and Wolfinger (2011), [4] Senn and Bretz (2007), [5] Bretz, Hothorn, and Westfall (2011)

図 1. 試験全体の第一種の過誤確率と相関係数の関係(左)と試験全体の検出力と相関係数の関係(右)

Hothorn, and Westfall, 2011). 最近では, 閉手順検定に基づく Gatekeeping 手順の研究・利用が顕著である (Dmitrienko et al., 2010). 試験全体の検出力は解析法に依存するため一意に定義できないが, 標準正規分布に従う検定統計量のもとで, 単純な Bonferroni 法を適用すると, つぎのように与えられる.

$$1 - \beta = \Pr \left[\bigcup_{k=1}^K \{Z_k > z_{\alpha/K}\} \right] = 1 - \Pr \left[\bigcap_{k=1}^K \{Z_k \leq z_{\alpha/K}\} \right] \quad (1)$$

Z_k は主要変数 $k (= 1, \dots, K)$ の検定統計量, $z_{\alpha/K}$ は棄却限界値 (標準正規分布の上側 $\alpha/K\%$ 点) である. 式 (1) は帰無仮説のもとでは試験全体の第一種の過誤確率を表わすので, 相関係数が大きくなると, 試験全体の第一種の過誤確率と検出力はともに小さくなる. つぎの条件のもとでの試験全体の第一種の過誤確率と相関係数の関係を図 1(左)に, 試験全体の検出力と相関係数の関係を図 1(右)に示す.

- 試験治療群と対照治療群の参加者数は等しい
- 試験全体の標本サイズは主要変数が 1 つの場合に検出力 80%以上を満たす最小の数
- 主要変数の数 $K = 2, 3, 4, 5, 10$
- 主要変数は連続量で分散が既知の正規分布に従う
- 効果サイズはすべて等しい (第一種の過誤確率の計算ではすべて 0 とする)
- 主要変数間の相関係数はすべて等しい
- 有意水準は片側 2.5%

試験全体の第一種の過誤確率は, Dmitrienko et al. (2010) が示すように

- 相関係数が高いほど小さい
- 主要変数の数が多いほど小さい

という特徴がある. 試験全体の検出力は, Senn and Bretz (2007) が示すように

- 相関係数が約 0.7 以下の場合

図 2. 試験全体の検出力と相関係数の関係

目標とする検出力 (80%) を上回り、主要変数の数が多いほど大きい

- 相関係数が約 0.7 を超える場合

目標とする検出力 (80%) を下回り、主要変数の数が多いほど小さい

という特徴がある。

標本サイズ設計では、実際の解析法のもとで、試験全体の検出力を計算すればよい。Bonferroni 法などの単純な方法以外では、試験全体の検出力の計算にシミュレーションが必要である。

試験目的 (3) のもとでは、すべての主要変数について有効性が示されれば、試験治療の対照治療に対する有効性が検証される。このような試験治療の効果の検証の手続きを “co-primary outcome measure approach” と呼ぶことがある (FDA, 2013)。標本サイズ設計では、試験目的 (1) と同様に、実際の解析法のもとで、試験全体の検出力を計算すればよい。標準正規分布に従う検定統計量のもとでは、試験全体の検出力は、各主要変数の個々の検出力の関数として次式で与えられる。

$$1 - \beta = \Pr \left[\bigcap_{k=1}^K \{Z_k > z_\alpha\} \right] \quad (2)$$

先ほどの条件のもとでの試験全体の検出力と相関係数の関係を図 2 に示す。

試験全体の検出力は、目標とする検出力 (80%) を下回る。相関係数が 1 に近いほど、または主要変数の数が少ないほど、試験全体の検出力は高い。

主要変数が互いに独立であるとき、試験の全体の検出力は、次式のように各主要変数の検出力の積で与えられる。

$$1 - \beta = \prod_{k=1}^K \Pr\{Z_k > z_\alpha\} \quad (3)$$

たとえば、2つの主要変数の検出力がともに 80% なら、試験全体の検出力は 64% ($= 0.8 \times 0.8 \times 100$) である。したがって、各主要変数の検出力を 80% より高くして、たとえば、90% にすると、試験全体の検出力は 81% ($= 0.9 \times 0.9 \times 100$) となる。ところが、実際には、主要変数間で (治療効果の大きさが異なり) 個々の検出力が異なるのが自然である。さらに、主要変数間にある程度の関連があるのが自然である。これより、表 2 に示すように、現実に近い状況で、主要変数間の関連を考

表 2. 複数の主要変数を用いる臨床試験の標本サイズ設計法

主要変数の 尺度	主要変数 の数	主要変数間の関連を 表現するモデル	検定法	論文
連続量	2	なし	検定統計量の 漸近正規性を利用	Xiong et al. (2005)
		二変量正規分布	T 検定	Sozu et al. (2006)
	K	K 変量正規分布	Z 検定* と T 検定	寒水 (2010) Sozu et al. (2011)
			Z 検定*	Julious and McIntyre (2012)
			Z 検定* 標本サイズの公式	Sugimoto et al. (2012)
二値変数	K	なし	割合の差の検定	Song (2009)
		(1) K 変量 Bernoulli 分布 (2) オッズ比 (3) 潜在分布 (K 変量正規分布)	5 つの検定法**	Sozu et al. (2010)
事象時間 データ	2	コピュラ・モデル (1) Clayton 族 (2) 正安定族 (3) Frank 族	対数ハザード比の 漸近正規性を利用	Hamasaki et al. (2013)
			ログランク検定	Sugimoto et al. (2013)
連続量と 二値変数の組	K	潜在分布 (K 変量正規分布)	連続量: Z 検定* 二値変数: 5 つの検定法**	Sozu et al. (2012)

* Z 検定: 各群の標本平均の差を標準誤差 (分散既知) で除した検定統計量による検定

** 5 つの検定法: Pearson のカイ二乗検定 (連続修正なし・あり), 逆正弦変換法 (連続修正なし・あり), Fisher の直接法

慮した標本サイズ設計法が提案されている。3 節では、主要変数の尺度ごとに、標本サイズ設計法の概要を紹介する。

2.2 統計学的方法による分類

Dmitrienko et al. (2010) は、複数の主要変数の扱いを、統計学的方法の観点から、(1) At-least-one, (2) Global, (3) All-or-none, (4) Superiority-noninferiority procedures の 4 つに分類している。(1) At-least-one procedure は試験目的 (1) と同じであり、(3) All-or-none procedure は試験目的 (3) と同じである。

主要変数が連続量の場合、試験治療と対照治療の応答変数の母平均の差を Δ_k とすると、(2) の Global procedure は、複数の主要変数による多変量解析の枠組みで「帰無仮説 $H_0: \Delta_k = 0$ がすべての k で成立する」を検定するものである。具体的な検定法として、Hottelling の T^2 が有名であるが、両側検定であるために、本来証明したい片側対立仮説

$$H_1: \Delta_k \geq 0 \text{ がすべての } k \text{ で成立し, 少なくとも 1 つの } k \text{ で } \Delta_k > 0 \text{ が成立する} \quad (4)$$

に対して検出力が低くなりやすい。そのため、Hottelling の T^2 に代わる方法がいくつか提案されている (Dmitrienko et al., 2010)。

(4) の Superiority-noninferiority procedure は、すべての主要変数の非劣性を示し、かつ少なくとも 1 つの主要変数の有効性を示せば、新しい治療法としての価値を認める、という考えに基づくものである。すべての主要変数の非劣性を示す (3) All-or-none の条件と、少なくとも 1 つの主要変数の有効性を示す (1) At-least-one の 2 つの条件を組み合わせるために、Intersection-

union 検定が適用される (Bloch et al., 2001; Perlman and Wu, 2004). (1) At-least-one の条件を評価するために, Bloch et al. (2001) は, Hottelling の T^2 の検定統計量の分散共分散行列を修正した方法とノンパラメトリック Bootstrap 法を用いることを提案している. Perlman and Wu (2004) は, Hottelling の T^2 が両側検定であることの不十分さを指摘し, Perlman (1969) による片側尤度比検定を用いることを提案している. ただし, Perlman (1969) の方法は, 主要変数間の相関構造によっては検定が保守的になりやすいことと, 検定統計量の計算が複雑であるという欠点がある. (3) All-or-none の条件を評価するためには, 比劣性マージンを考慮して, 通常の主要変数が 1 つの場合の解析法を適用すればよい.

3. 標本サイズ設計

3.1 連 続 量

Xiong et al. (2005) は, 分散共分散行列が既知である 2 つの主要変数に対する検定統計量が漸近的に正規分布に従うもとの, 標本サイズ設計法を導出し, アルツハイマー病の臨床試験への応用例を示している. これにより, 各主要変数の効果サイズの違いや主要変数間の相関が試験に必要な標本サイズに与える影響が定量的に明らかになった.

Sozu et al. (2006) は, 2 つの主要変数が 2 変量正規分布に従うもとの, 主要変数の分散を未知として, 標本サイズ設計法を導出している. 試験全体の検出力を正確に計算するには, 多変量正規分布と Wishart 分布に関する多重積分が必要であるが, それを単純なモンテカルロ法で計算する方法を示している. これにより, 主要変数や検定統計量の同時分布に従う乱数を発生させて検定統計量の実現値を計算し, 帰無仮説が棄却される割合を検出力の近似値とする単純な方法よりも, 短時間で精度の高い検出力が得られる. 試験に必要な標本サイズを計算する SAS のプログラムが <http://www.kbs.med.kyoto-u.ac.jp/sozu/software.html> に公開されている.

寒水 (2010) および Sozu et al. (2011) は, 前述した Sozu et al. (2006) の方法を主要変数の数が 3 つ以上の場合に一般化し, 主要変数の分散を既知または未知とした 2 つの方法による標本サイズを比較している. 主要変数の分散が既知の場合は, 検定統計量が多変量正規分布に従い, 検定統計量間と主要変数の相関が一致する. 試験全体の検出力は多変量正規分布の累積分布関数から比較的容易に計算できる. 有意水準を片側 2.5%, 目標とする検出力を 80% とした一般的な状況では, 分散を未知とすると, 通常の主要変数が 1 つの場合と同じように試験に, 必要な標本サイズは 1 群あたり 1 例多くなる.

Eaton and Muirhead (2007) は, 複数の主要変数が多変量正規分布に従うもとの, 主要変数の分散を未知として, 前提とする検定 (Intersection-union 検定) が極端に保守的になりやすく, 不偏検定でないことを考察している. さらに, 主要変数間の相関が 0 のとき (負の値を除く), 試験全体の検出力が最小になることを示している.

Julious and McIntyre (2012) は, 複数の検定に関連があり, すべての検定で有意であることが必要な臨床試験の枠組みで, 3 つの標本サイズ設計法をとりあげて, 標本サイズの挙動と変形性関節症の臨床試験への応用例を示している. その方法とは, (1) 主要変数ごとに試験に必要な標本サイズを計算してその最大値を採用する方法 (多重性を無視する方法), (2) 主要変数が互いに

図 3. 標本サイズの比と相関係数の関係

独立であると仮定する方法, (3) 主要変数が 1 つの場合の試験に必要な標本サイズに係数を乗じる方法 (主要変数間の関連を考慮する方法) である. (3) の方法は, 主要変数間の相関係数がすべて等しく, 効果サイズがすべて等しい (または, この条件を少し緩めた) 状況にしか適用できない.

これまでに紹介した標本サイズ設計法により, 試験に必要な標本サイズを求めるには, 基本的には, ソフトウェアによる逐次計算が必要である. 逐次計算のアルゴリズムとしては, 効果サイズが最小の主要変数に対する標本サイズを初期値として, 標本サイズを 1 例ずつ増加させて, 目標とする検出力を満たす標本サイズを求めるのが最も単純であるが, 主要変数の数が多く, 効果サイズが等しいときには, 計算に時間がかかる. このような逐次計算を必要としない方法として, Sugimoto et al. (2012) は, 複数の主要変数が多変量正規分布に従う場合について, 分散を既知として, 公式と数表を使って標本サイズを簡便に計算する方法を導出している. これにより, 通常の主要変数が 1 つの場合と同じように「 $n =$ 」と表現される公式に必要な値を代入すれば (逐次計算をしなくても) 試験に必要な標本サイズを計算できるようになった. さらに, 治療効果, 分散の大きさ, 有意水準, 目標とする検出力, といった試験デザインの要因が標本サイズに与える影響を評価しやすくなった.

つぎの条件のもとでの標本サイズの比 (= 各相関係数のもとでの標本サイズ/相関係数が 0 のもとでの標本サイズ) と相関係数の関係を図 3 に示す.

- 試験治療群と対照治療群の参加者数は等しい
- 主要変数は連続量で分散が既知の正規分布に従う
- 主要変数間の相関係数はすべて等しい
- 有意水準は片側 2.5%
- 目標とする検出力は 80%

図 3(左)は, 主要変数の効果サイズがすべて等しいもとで, 標本サイズの比を主要変数の数 $K = 2, 3, 4, 5, 10$ ごとに示している. たとえば, 主要変数の数が 2 つで ($K = 2$) 相関係数が 0.8 の場合, 標本サイズの比は約 0.89 である. これは, 主要変数間の相関を考慮すると, 試験に必要な

標本サイズが約 11% 少なくなることを意味する。標本サイズの比は、主要変数の数が多いほど小さい。

図 3(右)は、主要変数の数が 2 つのもとで、標本サイズの比を効果サイズの比 $\delta_1/\delta_2 = 1, 1.25, 1.5, 1.75, 2$ ごとに示している。試験に必要な標本サイズは、効果サイズの比が大きいほど相関係数に依存しない。したがって、効果サイズの比が大きいとき、すなわち、主要変数間で効果サイズが大きく異なるときは、効果サイズが小さい方の主要変数に対する標本サイズが試験全体の標本サイズとなる。

3.2 二値変数

Sozu et al. (2011, 2012) は、複数の主要変数が二値変数の場合について、主要変数間の関連指標として、多変量 Bernoulli 分布の相関、オッズ比、潜在分布の相関の関係を整理して、標本サイズ設計法を導出している。これら 3 つの関連の指標は、いずれも主要変数の成功確率と主要変数間の同時確率の関数として表現できる。どの指標を用いても、主要変数間の関連を規定できるが、主要変数の成功確率が与えられたもとで、とり得る値に範囲が生じる (Emrich and Piedmonte, 1991; Prentice, 1988)。解析法には、Pearson のカイ二乗検定(連続修正なし・あり)、逆正弦変換法(連続修正なし・あり)、Fisher の直接法の 5 つがとりあげられている。Fisher の直接法以外の 4 つの検定法の検定統計量の相関は主要変数間の相関の関数として与えられる。試験に必要な標本サイズと相関係数の関係は、解析法によらず、主要変数が連続量の場合と定性的に同じである。各解析法のもとでの試験に必要な標本サイズの関係は、通常的主要変数が 1 つの場合と同じである。たとえば、逆正弦変換法(連続修正あり)は Fisher の直接法に対する近似法として使用できる。Pearson のカイ二乗検定のもとで試験に必要な標本サイズを計算する SAS のプログラムが <http://www.kbs.med.kyoto-u.ac.jp/sozu/software.html> に公開されている。

Song (2009) は、複数の主要変数が二値変数の場合について、すべての主要変数の非劣性を示すことを目的としたもとでの標本サイズ設計法を導出している。ただし、主要変数間の関連を表現するモデルの定義が与えられてないので、そのままでは適用が難しい。

3.3 事象時間データ

主要変数が事象時間データの場合、対象となる確率変数は正の値をとり、イベント発生のみニズムの観点から、時間一定のハザード関数をもつ指数分布を前提とすることが多い。そのため、主要変数間の関連は、前述の連続量に対する多変量正規分布とは異なった形式で与えられる。その際、事象時間データでは、ログランク検定といった、いわゆるノンパラメトリック検定法への対応を考える必要がある。

事象時間データに対する 2 変量の確率変数を (T_1, T_2) とすると、2 変量間の関連指標として、次式の累積ハザード変量間の相関がしばしば使用される。

$$\rho_{\Lambda} = \text{corr}[\Lambda_1(T_1), \Lambda_2(T_2)] = \int_0^{\infty} \int_0^{\infty} S(t_1, t_2) d\Lambda_1(t_1) d\Lambda_2(t_2) - 1 \quad (5)$$

$\Lambda_1(t_1)$ と $\Lambda_2(t_2)$ は、 T_1 と T_2 の周辺累積ハザード関数であり、 $S(t_1, t_2)$ は (T_1, T_2) の同時生存関数である。事象時間データが指数分布に従うとき、 ρ_{Λ} は (T_1, T_2) の Pearson 相関係数となる。式 (5) より、 $S(t_1, t_2)$ のモデル化が (T_1, T_2) の関連指標 ρ_{Λ} を決定づけることがわかる。ただし、 ρ_{Λ} の値が同じであっても、同時生存関数 $S(t_1, t_2)$ の違いにより、2 変量間の関連性の特徴が異

なる. このとき, 既存のデータがあれば, 同時生存関数のモデルのあてはまりの良さを検討することで, 2 変量間の関連性の特徴を調べることができる. コピュラ・モデルは, このようなモデルのあてはまりを検討したり, 周辺生存関数から同時生存関数を生成するための便利なツールである. 代表的なコピュラとして, Clayton 族 (Clayton, 1978), 正安定族 (Hougaard, 1984, 1986), Frank 族 (Frank, 1979) があるが, 2 変量間の関連性についてそれぞれ異なった特徴を与える. 一般的に, Clayton 族では後期の時間の関連性が強く, 逆に, 正安定族では前期の時間の関連性が強い. Frank 族では, 前期と後期の関連性が対称的であり, 時間の中心から離れるほど関連性が弱い.

これらの背景のもと, Sugimoto et al. (2013) は, 2 つの主要変数が事象時間データの場合について, 2 変量ログランク検定統計量の漸近分散共分散の定式化とその計算法を与え, ログランク検定に基づく標本サイズ設計法を導出している. 2 変量間の関連指標の定式化には, 前述の 3 つのコピュラ (Clayton 族, 正安定族, Frank 族) のもとで同時生存関数 $S(t_1, t_2)$ を与えて, 各コピュラのもとでの標本サイズの数値例を示し, その特徴を考察している. ただし, 標本サイズ設計の公式は, 特定のコピュラ・モデルに依存した形式ではなく, 任意の同時生存関数 $S(t_1, t_2)$ に対応した形式で与えられている. 試験に必要な標本サイズを計算する R のプログラムが <http://www.st.hirosaki-u.ac.jp/~sugimoto/samplesize2.zip> に公開されている.

Hamasaki et al. (2013) は, 2 つの事象時間データがともに指数分布に従うもとで, 各事象に対する対数ハザード比が漸近的に 2 変量正規分布に従うことを利用して, 標本サイズ設計法を導出している. Sugimoto et al. (2013) と同様に, 同時生存関数 $S(t_1, t_2)$ のモデル化にコピュラを用いている. 通常の主要変数が 1 つの場合と同じように, 指数分布を仮定した場合とログランク検定に基づく場合の標本サイズの差は実質的にわずかである. したがって, Hamasaki et al. (2013) の方法は, 前述の Sugimoto et al. (2013) の方法を簡単かつ短時間で近似するものと捉えることができる.

これらの 2 つの研究では, 一方の事象を観測すれば, もう一方の事象が観察できなくなるといった状況は想定していない. そのような状況として, エイズ関連カポジ肉腫に関する臨床試験をとりあげて「カポジ肉腫の進行までの時間」と「ウイルス学的失敗 (virologic failure) までの時間」の 2 つの事象時間型の主要変数を紹介している. 片方もしくは両方とも致命的なイベントで打ち切りを伴う場合は, 中途打ち切りの独立性が崩れる. そのような状況に対する標本サイズ設計法の導出は今後の課題である.

3.4 連続量と二値変数の組

正規変数と二値変数の主要変数が混在する場合について, 両者の関連を表現するモデルがいくつか提案されている (Lev, 1949; Tate, 1954, 1955a, 1955b; Plackett, 1965; Dale, 1986; Pearson, 1909; Molenberghs et al., 2001). Sozu et al. (2012) は, 主要変数の潜在分布として多変量正規分布を仮定するモデル (Pearson, 1909; Tate, 1955a, 1955b) のもとで, 標本サイズ設計法を導出している. たとえば, 正規変数と二値変数の主要変数がそれぞれ 1 つずつある場合, 潜在分布は 2 変量正規分布となり, 正規変数は潜在分布に従う確率変数の実現値として観測され, 二値変数は潜在分布に従う確率変数の実現値をあるカットオフ値で二値化したものとして観察される.

主要変数の尺度が1つの場合は、主要変数間で効果サイズが等しければ、対応する個々の検出力が等しくなるので、3.1節の後半で述べたように、試験に必要な標本サイズが相関係数に依存するかどうかは、効果サイズの比に依存する。ところが、主要変数の尺度が異なる場合は、主要変数間で効果サイズが等しくても、対応する個々の検出力が異なるので、試験に必要な標本サイズが相関係数に依存するかどうかは、各主要変数の個々の検出力に注目しなければならない。

4. 検定の保守性を緩和させる方法

4.1 検定の保守性

すべての主要変数について有効性を示すことを目的とする臨床試験では、2節で述べたように、Intersection-union 検定を適用するため、検定が保守的になりやすい。たとえば、有意水準が片側2.5%のもとで、2つの主要変数が互いに独立で、第一種の過誤確率がともに2.5%の場合、試験全体の第一種の過誤確率は0.0625% ($= 0.025 \times 0.025 \times 100$)となる。このように、試験全体の第一種の過誤確率は有意水準を大きく下回り、検定は過度に保守的になる。さらに、主要変数の数が多くなると、検定はますます保守的になりやすい。これに対応して、図2に示したように、相関係数が0に近いほど、または主要変数の数が多いほど、試験全体の検出力は低くなる。これにより、通常の主要変数が1つの場合に比べて、試験に必要な標本サイズは過度に大きくなりやすく、標本サイズが実施可能数を超過してしまうことがある。

このような検定の保守性を緩和させるためには、主要変数の数を減らしたり、合成変数を使用したりするのが効果的である。ところが、そのような対応が常に可能であるとは限らない。そこで、ある基準のもとで、各主要変数に対する検定の有意水準を大きくして、検定の保守性を緩和させる方法が提案されている。

4.2 平均的な第一種の過誤確率を制御する方法：平均制御法

Chuang-Stein et al. (2007) と Offen et al. (2007) は「平均的な第一種の過誤確率」という概念を導入し、それを制御する方法(以下「平均制御法」と呼ぶ)を提案している。説明のために、主要変数の数を2つ、各効果サイズを δ_k ($k=1,2$)とする。ただし、 $\delta_k > 0$ であることが臨床的に好ましい状態であるとする。これより、帰無仮説 H_0 と対立仮説 H_1 をつぎのように表現できる。

$$\begin{aligned} H_0: \delta_k \leq 0 \text{ がすくなくとも1つの } k \text{ で成立する} \\ H_1: \delta_k > 0 \text{ がすべての } k \text{ で成立する} \end{aligned} \quad (6)$$

これを δ_1 と δ_2 に関する2次元の母数空間に図示すると図4(左)が得られる。

このとき、帰無仮説が正しいとしても、各主要変数の効果サイズが負の大きな値であることが現実的には起こりえないとし、図4(右)に示すように、帰無仮説に対する母数空間の領域を限定している。これは、効果サイズに関する事前情報(信念)を帰無仮説に反映させる考えに基づいている(Patel, 1991)。その際、1つの主要変数だけに効果があるときに、現実的に期待される最大の効果サイズ(M)を領域の閾値にしている。その上で、限定された空間上に (δ_1, δ_2) に対する同時確率分布 $f_0(\delta_1, \delta_2)$ を仮定し、第一種の過誤確率の期待値を「平均的な第一種の過誤確率(Average type I error)」として定義し、それを制御することを提案している。たとえば、主要変数間の相関を0、 $f_0(\delta_1, \delta_2)$ を一様分布にした場合、各主要変数の検定の有意水準を3.7%にすると、通常の

図 4. 帰無仮説と対立仮説に対応する母数空間(左)と平均制御法のもとでの母数空間(右)

試験全体の第一種の過誤確率は 3.7%以下に制御され, 平均的な第一種の過誤確率は 2.5%以下に制御される.

ただし, 著者らが挑戦的な試みであることを認識しているように, 実際の臨床試験で, 有意水準を高くすることに抵抗を感じるのが自然である. とくに, 平均的な第一種の過誤確率の概念が数学的に妥当であっても, それを実際の臨床試験(とくに医薬品開発における検証的臨床試験)で適用するには, 通常的第一種の過誤確率の枠組みとの定性的かつ定量的な違いを詳細に検討する必要がある. その際, 主要変数間の相関の誤特定の影響や $f_0(\delta_1, \delta_2)$ の決め方など, いくつかの統計学的な課題を検討する必要がある.

4.3 主要変数の有意性を均衡させる方法: 均衡法

Kordzakhia, Siddiquia, and Huque (2009) は, 平均制御法のように, すべての主要変数に対して検定の有意水準を一律に高くするのではなく, 主要変数の有意性を均衡させること(以下「均衡法」と呼ぶ)を提案している. 著者らは, 一方の主要変数に対する検定の p 値が小さければ, その分, 他の主要変数に対する検定の有意水準を大きくしてもよいという考えのもと, それを実現するための p 値に対する棄却域を示している. たとえば, 主要変数が 2 つの場合, 各主要変数の検定の p 値を p_1, p_2 , 試験全体の有意水準を α としたもとで, 各 p 値に対する検定の棄却域は次式で与えられる.

$$R(\alpha_L) = \{(p_1, p_2) \in (0, \alpha_L)^2 : p_1 + p_2 + cp_1p_2 < \alpha_L\}, \quad c = \frac{(\alpha_L - 2\alpha)}{\alpha^2}$$

α_L は $\alpha < \alpha_L < 2\alpha$ を満たす正の実数であり, その設定の仕方については, 平均制御法における各検定の有意水準をひとつの基準にしている.

著者らは, 実際のアルツハイマー病の臨床試験の結果に対して, 均衡法を仮想的に適用した結果を示している. そこでは, ADAS-cog と CIBIC-plus の 2 つの主要変数の検定の p 値がそれぞれ <0.0001 と 0.0305 という結果である. 通常の方法では, CIBIC-plus の p 値が有意水準である 2.5%を上回るため, すべての主要変数の有効性を示すことができない. ところが, 均衡法では, ADAS-cog の p 値がほぼ 0 に近いので, CIBIC-plus に対する有意水準が(平均制御法の基準のもとで)3.7%となり, p 値がこれを下回るため, すべての主要変数の有効性を示すことができる.

4.4 平均制御法と均衡法の数理的関係

検定の保守性を緩和させる方法の関係を明らかにするため, 保守性の緩和を意図しない方法(通常の方法), 平均制御法, 均衡法の 3 つの方法のもとでの検定の棄却域を図 5 に示す. 図 5 の横

軸の Z_1 と縦軸の Z_2 は各主要変数に対する検定統計量, z_α と $z_{\tilde{\alpha}}$ はそれぞれ標準正規分布の上側 $\alpha\%$ 点と上側 $\tilde{\alpha}\%$ 点を表わす. $\tilde{\alpha} (> \alpha)$ は, 平均制御法のもとで調整された有意水準を表わす. 通常の方法の棄却域は(1)の領域, 平均制御法の棄却域は(1)~(6)の領域, 均衡法の棄却域は(1)~(3)の領域である. 検定における棄却域は, 通常の方法, 均衡法, 平均制御法の順に広くなり, 結果として, 検出力が高くなり検定の保守性が緩和される. ただし, Sozu and Hamasaki (2012) は, 典型的な臨床試験の状況下では, 平均制御法と均衡法の標本サイズに実質的な差が生じないことを示している.

図 5. 3つの方法(通常の方法・平均制御法・均衡法)における検定の棄却域

5. おわりに

本稿では, すべての主要変数について有効性を示すことを目的とする臨床試験を中心に, 最近の統計学的方法を整理した. とくに, 標本サイズ設計について, 著者らを中心とした最近の研究成果を概説した. 本節では, 今後の研究の方向性や拡張性について述べる.

これまでの研究は, 表 2 に示すように, 主要変数が連続量の場合を中心に, 二値変数, 事象時間データの場合を対象としている. 臨床試験では, これらの尺度以外に, カテゴリカルデータが使用されることがある. たとえば, 1 節や 4.3 節で示したアルツハイマー病の全般症状に関する評価項目の CIBIC-plus は, 治療開始前の状態と比べた全般症状を 7 段階(著明改善・改善・軽度改善・不変・軽度悪化・悪化・著明悪化)で評価するものである. 実際の解析では, 7 段階の評価を連続量として扱ったり, 「著明改善」と「不変」または「不変」と「軽度悪化」で区切って二値化したりすることもあるが(Rogers et al., 1998; Rösler et al., 1999; Tariot et al., 2000; Peskind et al., 2006), 基本的には, 順序カテゴリカルデータとして解析される. これより, 主要変数がカテゴリカルデータの場合についての検討が必要である.

つぎに, これまでの研究は, 並行 2 群間比較試験の枠組みで検討されている. したがって, 他

の試験デザイン(例: 用量反応試験, プラセボ・実薬対照試験, 群逐次試験, 適応的試験)のもとでの検討が必要である。さらに, 副次変数, 合成変数の各項目, 部分集団といった階層性に伴う多重性に関する検討が必要である。このような状況では, 複数の仮説を扱うことにより, 多重性の問題が重なり, これまでの研究成果を統合的に発展・拡張させる必要がある。

最後に, 臨床試験の効率に関する課題がある。「新たな治験活性化五カ年計画の中間見直しに関する検討会 (2010)」では, 加速・強化すべき課題として「治験・臨床研究の効率化」をとりあげている。これからの臨床試験では, 試験の効率化についての検討が必要であり, 本稿が対象とする主要変数が複数ある臨床試験も例外ではない。

謝 辞

本稿について, 丁寧な査読を通して貴重なご意見を頂きました査読者の先生に心から謝意を表します。本研究は, JSPS(Japan Society for the Promotion of Science) 科研費 25870332 の助成を受けた。

参考文献

- Bloch, D. A., Lai, T. L., Su, Z., and Tubert-Bitter, P. (2007). A combined superiority and non-inferiority approach to multiple endpoints in clinical trials. *Statistics in Medicine* **26**, 1193–1207.
- Bretz, F., Hothorn, T., and Westfall, P. (2011) Multiple Comparisons Using R. Chapman & Hall/CRC, Boca Raton, FL.
- Chuang-Stein, C., Stryszak, P., Dmitrienko, A., and Offen, W. (2007). Challenge of multiple co-primary endpoints: A new approach. *Statistics in Medicine* **26**, 1181–1192.
- Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease. *Biometrika* **65**, 141–151.
- Committee for Medicinal Products for Human Use (CHMP) (2008). *Guideline on Medicinal Products for the Treatment Alzheimer's Disease and Other Dementias* (CPMP/EWP/553/95 Rev. 1). EMEA: London.
- Cook, R. J. and Farewell, V. T. (1996). Multiplicity considerations in the design and analysis of clinical trials. *Journal of the Royal Statistical Society, Series A* **159**, 93–110.
- Cordoba, G., Schwartz, L., Woloshin, S., Bae, H., and Gotzsche, P. C. (2010). Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. *BMJ* **341**, c3920.
- Dale, J. R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics* **42**, 909–917.
- Dmitrienko, A. Tamhane, A. C., and Bretz, F. (2010). Multiple Testing Problems in Pharmaceutical Statistics. Chapman & Hall/CRC, Boca Raton, FL.

- Eaton, M. L. and Muirhead, R. J. (2007). On a multiple endpoints testing problem. *Journal of Statistical Planning and Inference* **137**, 3416–3429.
- Emrich, L. J. and Piedmonte, M. R. (1991). A method for generating high-dimensional multivariate binary variates. *The American Statistician* **45**, 302–304.
- Ferreira-González, I., Permanyer-Miralda, G., Busse, J. W., Bryant, D. M., Montori, V. M., Alonso-Coello, P., Walter, S. D., and Guyatt, G. H. (2007a). Methodologic discussions for using and interpreting composite endpoints are limited, but still identify major concerns. *Journal of Clinical Epidemiology* **60**, 651–657.
- Ferreira-González, I., Busse, J. W., Heels-Ansdell, D., Montori, V. M., Akl, E. A., Bryant, D. M., Alonso-Coello, P., Alonso, J., Worster, A., Upadhye, S., Jaeschke, R., Schünemann, H. J., Permanyer-Miralda, G., Pacheco-Huergo, V., Domingo-Salvany, A., Wu, P., Mills, E. J., and Guyatt, G. H. (2007b). Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. *BMJ* **334**, 786.
- Food and Drug Administration (2007). *Guidance for Industry. Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics*. Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER), Food and Drug Administration, Rockville, MD.
- Food and Drug Administration (2012). *Guidance for Industry. Irritable Bowel Syndrome – Clinical Evaluation of Drugs for Treatment*. Center for Drug Evaluation and Research (CDER), Food and Drug Administration, Rockville, MD.
- Food and Drug Administration (2013). *Guidance for Industry. Alzheimer’s Disease: Developing Drugs for the Treatment of Early Stage Disease. Draft Guidance*. Center for Drug Evaluation and Research (CDER), Food and Drug Administration, Rockville, MD.
- Frank, M. J. (1979). On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$. *Aequationes Mathematicae* **19**, 194–226.
- Hamasaki, T., Sugimoto, T., Evans, S. R., and Sozu, T. (2013). Sample size determination for clinical trials with co-primary outcomes: Exponential event-times. *Pharmaceutical Statistics* **12**, 28–34.
- Hougaard, P. (1984). Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika* **71**, 75–83.
- Hougaard, P. (1986). A class of multivariate failure time distribution. *Biometrika* **73**, 671–678.
- Hung, H. M. and Wang, S. J. (2009). Some controversial multiple testing problems in regulatory applications. *Journal of Biopharmaceutical Statistics* **19**, 1–11.

- ICH (1998). ICH Harmonised Tripartite Guideline: Statistical principles for clinical trials (E9). Available at: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf (Accessed January 17, 2013).
- Julious, S. A. and McIntyre, N. E. (2012). Sample sizes for trials involving multiple correlated must-win comparisons. *Pharmaceutical Statistics* **11**, 177–185.
- Kordzakhia, G., Siddiqui, O., and Huque, M. F. (2010). Method of balanced adjustment in testing co-primary endpoints. *Statistics in Medicine* **29**, 2055–2066.
- 厚生労働省医政局研究開発振興課 (2010). 「新たな治験活性化五カ年計画の中間見直しに関する検討会」報告. http://www.mhlw.go.jp/shingi/2010/01/dl/s0119-10a_0001.pdf (Accessed June 25, 2013).
- Lev, J. (1949). The point biserial coefficient of correlation. *The Annals of Mathematical Statistics* **20**, 125–126.
- Lim, E., Brown, A., Helmy, A., Mussa, S., and Altman, D. G. (2008). Composite outcomes in cardiovascular research: a survey of randomized trials. *Annals of Internal Medicine* **149**, 612–617.
- Molenberghs, G., Geys, H., and Buyse, M. (2001). Evaluation of surrogate endpoints in randomized experiments with mixed discrete and continuous outcomes. *Statistics in Medicine* **20**, 3023–3038.
- 森川敏彦, 平山正史 (2006). 臨床試験ハンドブック, 丹後俊郎, 上坂浩之 編, .55–59, 朝倉書店.
- O'Brien, P. C. (1984). Procedure for comparing samples with multiple endpoints. *Biometrics* **40**, 1079–1087.
- Offen, W., Chuang-Stein, C., Dmitrienko, A., Littman, G., Maca, J., Meyerson, L., Muirhead, R., Stryszak, P., Boddy, A., Chen, K., Copley-Merriman, K., Dere, W., Givens, S., Hall, D., Henry, D., Jackson, J. D., Krishen, A., Liu, T., Ryder, S., Sankoh, A. J., Wang, J., and Yeh, C. H. (2007). Multiple co-primary endpoints: Medical and statistical solutions. *Drug Information Journal* **41**, 31–46.
- Patel, H. I. (1991). Comparison of treatments in a combination therapy trial. *Journal of Biopharmaceutical Statistics* **1**, 171–183.
- Pearson, K. (1909). On a new method for determining the correlation between a measured character A, and a character B. *Biometrika* **7**, 96–105.
- Perlman, M. D. (1969). One-sided testing problems in multivariate analysis. *Annals of Mathematical Statistics* **40**, 549–567.
- Perlman, M. D. and Wu, L. (2004). A note on one-sided tests with multiple endpoints. *Biometrics* **60**, 276–280.

- Peskind, E. R., Potkin, S. G., Pomara, N., Ott, B. R., Graham, S. M., Olin, J. T., and McDonald, S. (2006). Memantine treatment in mild to moderate Alzheimer disease: A 24-week randomized, controlled trial. *American Journal of Geriatric Psychiatry* **14**, 704–715.
 - Plackett, R. L. (1965). A class of bivariate distributions. *Journal of the American Statistical Association* **60**, 516–522.
 - Pocock, S. J. (1997). Clinical trials with multiple outcomes: A statistical perspective on their design, analysis and interpretation. *Controlled Clinical trials* **18**, 530–545.
 - Pocock, S. J., Geller, N. L., and Tsiatis, A. A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics* **43**, 487–498.
 - Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033–1048.
 - Rogers, S. L., Farlow, M. R., Doody, R. S., Mohs, R., Friedhoff, L. T., and The Donepezil Study Group. (1998). A 24-week, double-blind, placebo-controlled trial of donepezil in patients with Alzheimer’s disease. *Neurology* **50**, 136–145.
 - Rösler, M., Anand, R., Cicin-Sain, A., Gauthier, S., Agid, Y., Dal-Bianco, P., Stähelin, H. B., Hartman, R., and Gharabawi, M. (1999). Efficacy and safety of rivastigmine in patients with Alzheimer’s disease: International randomised controlled trial. *BMJ* **318**, 633–640.
 - Sankoh, A. J., D’Agostino, R. B., and Huque, M. F. (2003). Efficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple endpoint issues. *Statistics in Medicine* **22**, 3133–3150.
 - 佐藤俊哉 (2007). 複合エンドポイントは解釈が難しく誤解を招くことも. 解説「複合エンドポイントの問題を改善する工夫も必要」. *Mainichi Medical Journal* **3**, 894–895.
 - 佐藤俊哉 (2011). 妥当性や報告に問題のある複合評価項目が多い. 解説「複合エンドポイントの問題か、報告の問題か」. *Mainichi Medical Journal* **7**, 48–49.
 - Senn, S. and Bretz, F. (2007). Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics* **6**, 161–170.
 - Song, J. X. (2009). Sample size for simultaneous testing of rate differences in noninferiority trials with multiple endpoints. *Computational Statistics & Data Analysis* **53**, 1201–1207.
 - Sozu, T., Kanou, T., Hamada, C., and Yoshimura, I. (2006). Power and sample size calculations in clinical trials with multiple primary variables. *Japanese Journal of Biometrics* **27**, 83–96.
 - Sozu, T. and Hamasaki, T. (2012). Adjustment methods for testing co-primary endpoints in clinical trials. The 3rd East Asia Regional Biometric Conference 2012. Seoul, Korea. February 2–3, 2012.
 - Sozu, T., Sugimoto, T., and Hamasaki, T. (2010). Sample size determination in clinical trials with multiple co-primary binary endpoints. *Statistics in Medicine* **29**, 2169–2179.
- Jpn J Biomet Vol.34, No.1, 2013

- Sozu, T., Sugimoto, T., and Hamasaki, T. (2011). Sample size determination in superiority clinical trials with multiple co-primary correlated endpoints. *Journal of Biopharmaceutical Statistics* **21**, 650–668.
- Sozu, T., Sugimoto, T., and Hamasaki, T. (2012). Sample size determination in clinical trials with multiple co-primary endpoints including mixed continuous and binary variables. *Biometrical Journal* **54**, 716–729.
- 寒水孝司 (2010). 複数の主要評価変数をもつ検証的臨床試験. 臨床試験の統計的エッセンス, 濱崎俊光 編, 67–76, 大阪大学出版会.
- Sugimoto, T., Sozu, T., and Hamasaki, T. (2012). A convenient formula for sample size calculations in clinical trials with multiple co-primary continuous endpoints. *Pharmaceutical Statistics* **11**, 118–128.
- Sugimoto, T., Sozu, T., Hamasaki, T., and Evans, S. R. (2013). A logrank test-based method for sizing clinical trials with two co-primary time-to-events endpoints. *Biostatistics* **14**, 409–421.
- Tariot, P. N., Solomon, P. R., Morris, J. C., Kershaw, P., Lilienfeld, S., Ding, C., and The Galantamine USA Study Group (2000). A 5-month, randomized, placebo-controlled trial of galantamine in AD. *Neurology* **54**, 2269–2276.
- Tate, R. F. (1954). Correlation between a discrete and a continuous variable. Point-biserial correlation. *The Annals of Mathematical Statistics* **25**, 603–607.
- Tate, R. F. (1955a). The theory of correlation between two continuous variable when one is dichotomized. *Biometrika* **42**, 205–216.
- Tate, R. F. (1955b). Applications of correlation models for biserial data. *Journal of the American Statistical Association* **50**, 1078–1095.
- Westfall, P. H., Tobias, R. D., and Wolfinger, R. D. (2011). Multiple Comparisons and Multiple Tests Using SAS (2nd ed). SAS Institute Inc., Cary, NC, USA.
- Xiong, C., Yu, K., Gao, F., Yan, Y., and Zhang, Z. (2005). Power and sample size for clinical trials when efficacy is required in multiple endpoints: Application to an Alzheimer’s treatment trial. *Clinical Trials* **2**, 387–393.