

氏名（本籍） いし はら たく ま 石 原 拓 磨（岐阜県）
学位の種類 博士（理学）
学位記番号 乙第1300号
学位授与の日付 2023年3月19日
学位授与の要件 学位規則第4条第2項該当
学位論文題目 **Various Methods for Analysis in Clinical
Trials with Categorical Endpoints**
(カテゴリカル変数を評価項目とした臨床試験における種々の解析法)

論文審査委員 (主査) 教授 田畑 耕治
教授 明石 重男 教授 宮本 暢子
教授 鈴木 知道 教授 寒水 孝司

論文内容の要旨

ヒトを対象として新規の治療、薬、医療機器等の有効性および安全性を確認するために行われる試験を臨床試験という。特に薬の承認申請を目的とした臨床試験の事を治験とよぶ。代表的な治験に関する統計的ガイドラインとして「ICH E9 臨床試験のための統計的原則」が挙げられる。このガイドラインでは、主たる評価に用いられる変数(主要評価変数)は、試験の主要な目的に直結した臨床的に最も適切で説得力のある証拠を与えうる変数で通常ただ1つに定めるべきであるとしている。一方で、複数の主要変数を用いることが望ましい場合もあるという内容も注意点と共に述べられている。実際にアレルギー薬の開発や関節リウマチの治療に関する分野では、臨床試験の検証的な段階において新規治療の効果を評価するために主要評価変数として複数の評価変数を用いる場合がある。したがって、臨床試験の計画を行う際にはどのような評価変数を用いるか、複数の評価変数を用いる必要があるかをまず議論するべきである。

特に検証的な臨床試験の前段階では、評価変数の特徴を明らかにすることは重要である。関節リウマチを例にすると、疾患活動性を示す評価変数として Disease Activity Score28 erythrocyte sedimentation rate (DAS28-ESR)と Simplified Disease Activity Index (SDAI)という指標が用いられることがある。いずれも連続変数であるが一般的には「重症度」を示す指標として4つのカテゴリで疾患の程度が評価される。これらの指標の類似性

を統計学的に明らかにしておくことは臨床試験において重要である。変数間の類似性を評価する1つの方法として分割表解析が考えられる。2つの変数が共に同じ分類でR水準から構成されているとき、 $R \times R$ 型の表をえることができる。特にこの表は $R \times R$ 正方分割表とよばれる。たとえば、カテゴリ化されたDAS28-ESRとSDAIは、寛解、軽度、中等度、重度の4つに分類され、このときえられる分割表は 4×4 正方分割表とよぶ。多くの場合、同じ分類からなる正方分割表の解析では分類間の関連性が高く、統計的独立性は成り立たない。したがって、本研究では、正方分割表を使った2種類の変数の対称性に着目し、変数の特徴を明らかにすることを考える。

治療効果の判定においては、多くの場合「治療効果がない」という帰無仮説に対して統計的仮説検定が適用され、第一種の過誤確率をあらかじめ宣言した(名目の)有意水準以下に抑える必要がある。また、複数の評価変数を用いる場合、正しい帰無仮説のうち少なくとも1つが誤って棄却されてしまう確率をFamilywise error rate (FWER)とよび、これを名目の有意水準以下に抑えることが要求される。複数の検定を繰り返すことによりFWERが名目の有意水準を超える(インフレーションする)ことを多重性の問題とよび、この問題への対処法を多重性の調整とよぶ。我々が複数の評価変数から治療効果に関する1つの結論をえる際には、検定における多重性の問題に対処することが不可欠である。

複数の主要評価変数を用いて新たな治療の有効性を検証する方法として、(1)すべての主要評価変数に対する効果が認められたときのみ新たな治療の効果を認める、(2)少なくとも1つの主要評価変数に対する効果を認めたときに新たな治療の効果を認める、という枠組みが考えられる。(1)の枠組みにおける検定の帰無仮説は主要評価変数それぞれに対する検定の帰無仮説の和集合で表現される。一方で、対立仮説は主要評価変数それぞれに対する検定の対立仮説の積集合で表現される。この和集合の帰無仮説と積集合の対立仮説に対してはIntersection-Union (IU)検定を適用することができ、個々の検定において多重性の調整は必要としない(Berger, 1982)。(2)の枠組みでは、個々の主要評価変数に対する検定の帰無仮説の積集合とそれぞれの対立仮説の和集合に対して検定することになるが、この場合には多重性の調整が必要とされる。

一般的に、検証的な臨床試験において新たな治療の効果は標準治療と比較することで評価される。特に本研究においては2つの治療の比較を考え、治療効果の差を「新たな治療による効果-標準治療による効果」と定義し、効果は正である程高いことを示すとす。代表的な比較の形式としては優越性試験と非劣性試験がある。優越性試験は標準治療に対して新たな治療が優れていることを示すことが目的であり、帰無仮説は「治療効果の差が0以下である」と設定される。一方で非劣性試験は標準治療に対して新たな治療は劣っていないことを示すことが目的であり、帰無仮説は「治療効果の差が非劣性マージン以下である」と設定される。ここで非劣性マージンは臨床的に許容できると判断しうる最大の差として定義される(ICH E9 臨床試験のための統計的原則より)。

本論文は、5つの章から構成されている。

第1章では、本研究に関する動機、歴史的な背景、各章の概要を述べた。

第2章では、正方分割表における対称性からの隔たりを測る新たな尺度を提案した。

Bowker (1948) は正方分割表のセル確率の対称性を示すモデルとして対称モデルを導入し、対称性に関する検定を与えた。また、2つの評価変数の対称性が成り立たない場合には、対称性からの隔たりの程度に関心がある。対称性からの隔たりを測る尺度は Tomizawa (1994) が提案した。分割表の主対角セルを中心に対称的に位置するセルの対称性に関しては部分尺度を用いて測ることができ、Tomizawa (1994)は部分尺度を Kullback-Leibler 情報量の形式で表現している。対称性からの隔たりを測る尺度は、主対角セルに対して対称的に位置する各セルのペアについての部分尺度の加重平均としても表現できる。本研究では、正方分割表における対称性からの隔たりを測る尺度として、部分尺度の加重平均のクラスにおいて分散が最小となる尺度を提案した。分割表の観測度数が多項分布に従ってえられていると仮定したもとの部分尺度ベクトルの推定量が漸近的に多変量正規分布に従うことを示し、部分尺度ベクトルの推定量の共分散行列を導出した。その結果から、部分尺度の推定量が漸近的に互いに独立であることがわかった。また、変数間に対称性が成り立たない場合は、どの水準で対称性が成り立っていないのかを特定する必要がある。したがって、事例への適用では部分尺度の信頼区間が 0 を含むか否かで部分対称性を判断した。さらに、既存の尺度と提案尺度の部分尺度に対する重みを比べることで提案尺度の性質を明らかにした。

第 3 章では、複数の二値評価変数のうち、少なくとも 1 つの変数についての優越性と、すべての変数の非劣性が同時に成り立つ場合のみ治療効果を認める試験の枠組みに適用できる新たな検定を提案した。複数の主要評価変数を用いて有効性を検証する臨床試験では、新たな治療の優越性と非劣性を同時に評価する場合がある。例えば、緊急の外科的処置を必要とする患者には血栓塞栓症の治療のために抗凝固薬という薬剤が投与される。この薬剤の服用者は血管内で血液が固まりにくくなるため、凝固薬という種類の薬剤を服用することで出血傾向を是正する必要がある。Goldstein et al. (2015)の臨床試験では 2 つの凝固薬の効果を、止血の有無と International Normalized Ratio (INR)が 1.3 以下であること、という 2 つの主要な二値評価変数を用いて検証した。分析では最初に両方の主要評価変数に対して非劣性が評価された後に優越性が評価された。本研究では上述の例のように複数の主要評価変数を用いて優越性と非劣性を同時に評価することを考える。上記の枠組みにおける検定としては Bloch et al. (2001)が Hotelling の T^2 検定を優越性の検定として用いた手法を提案している。Perlman and Wu (2014)は Tang et al. (1989) の片側近似尤度比検定を UI-IU 検定に用いた手法を提案している。さらに、Nakazuru et al. (2014)は Perlman and Wu (2004)の手法よりも検出力が高い UI-IU 検定を提案している。ただし、これらは評価変数が連続型の場合における手法であり、二値変数に対する手法は十分に提案されていない。したがって、二値評価変数にベルヌーイ分布を仮定し、優越性の検定に用いる統計量としてスコア型とワルド型の 2 種類提案した。非劣性の検定には比率の検定に一般的に用いられている Farrington and Manning (1990)の検定を用いた。検定手法の性能は、二値評価変数間の相関をいくつか設定し、モンテカルロシミュレーションを用いて第一種の過誤確率と経験検出力で評価した。また、実際の事例を基に、提案手法の中で用いた優越性の UI 検定のみを検出力と提案した UI-IU 検定の検出力を比較し、検出力の低下の程度を評価した。優越性検定の統計量にワルド型を用いた場合、スコア型と比べて検出力が高いことが示

された。しかしながら、ワルド型では第一種の過誤確率がインフレーションするケースがあることが分かった。

第4章では、第3章と同様の試験の枠組みにおいて、複数の二値評価項目の背後に潜在的な連続分布が仮定できる場合の検定手法を提案した。臨床の現場で使用される二値評価変数には、連続型の分布をもつ評価変数を二値化してえられるものが考えられる。例えば、関節リウマチの分野では疾患の改善の度合いを示す評価変数として American College of Rheumatology criteria (ACR)が代表的であり、現場では圧痛の数と腫れた関節の数がある割合（例えば20%）以上減ったか否かで症状の改善が判断される。このように潜在的に連続的な分布をもつ二値変数に関しての手法の提案が必要である。特に本研究では、潜在的な連続変数を二値化するためのカットポイントの推定方法の違いにより、優越性検定に用いる統計量を2種類提案した。検定統計量の推定の手順をステップに分けて説明し、検定手法の性能は第3章と同様にシミュレーションにより示した。第一種の過誤確率と検出力は第3章で紹介する手法の特徴と類似していた。複数の帰無仮説の包含関係を基に効率的に第一種の過誤確率のインフレーションを抑えることができる手法として閉検定手順が Marcus et al. (1976)によって提案されおり、提案手法はこの閉検定手順と比べて検出力が高い結果となった。

第5章では本研究の総括を行った。

論文審査の結果の要旨

ヒトを対象として新規の治療、薬、医療機器等の有効性及び安全性を確認するために行われる試験を臨床試験という。臨床試験では、主たる評価に用いられる変数（主要評価変数）は、試験の主要な目的に直結した臨床的に最も適切で説得力のある証拠を与える変数で通常ただ1つに定めるべきであるとされている（ICH-E9 ガイドライン）。一方で、複数の主要評価変数を用いることが望ましい場合があることも知られている。実際にアレルギー薬の開発や関節リウマチの治療に関する分野では、臨床試験の検証的な段階において新規治療の効果を評価するために主要評価変数として複数の評価変数を用いる場合がある。したがって、どのような評価変数を用い、1つで評価すればよいか、複数で評価すべきかをまず議論する必要がある。さらに、評価変数を複数用いる場合には、統計的仮説検定の多重性の問題が起こることがあるため、検定全体における有意水準を、あらかじめ決定した有意水準以下に抑えることが重要になる。

臨床試験において治療効果を評価するにあたっては優越性試験と非劣性試験の形式がよく用いられる。優越性試験とは、被験薬への反応が比較薬剤よりも臨床的に優れることを示す試験であり、非劣性試験とは、被験薬への反応が比較薬剤よりも臨床的に劣らないことを示す試験である。実際の臨床試験では、評価変数や試験形式をどのようにするかをよく検討した上で実施する。その際、当該試験の目的達成に必要なサンプルサイ

ズを試験実施前に決定する必要があるが、より少ないサンプルサイズでの試験が望ましい。これを実現するための一つの方法が検出力の高い解析手法（検定統計量）の開発であり、この分野において非常に重要な課題である。

本論文では、評価変数の選択に関する1つの方法を提案し、さらに、複数の主要評価変数、とくに二値の評価変数に対して、優越性試験と非劣性試験を組み合わせた形式による治療効果の評価を行う状況において、新たな検定統計量を提案している。本論文は、以下の5つの章から構成されている。

第1章では、本研究に関する動機、歴史的な背景、各章の概要について述べている。

第2章では、治療効果を認めるために複数の評価項目が必要か否か、統計的・臨床的に議論することをサポートするための指標が提案されている。具体的には、二つのカテゴリカル変数に対する対称性からの逸脱度を測る尺度を提案している。行と列が同じ分類からなる正方分割表においては統計的独立性よりも対称性に関心があり、Bowker (1948) が対称性のカイ二乗検定を導出した。また、得られたデータに対して対称性が当てはまらない場合には、対称性からの逸脱度を測るための尺度が Tomizawa, Seo and Yamamoto (1998) などにより提案された。本章では、Tomizawa et al. (1998) の部分尺度の性質を明らかにし、部分尺度の加重平均で定義されるクラスで分散が最小になる重みを与えた。さらに、その重みを用いた対称性からの逸脱度を測る尺度を与えている。

第3章では、複数の二値評価変数のうち、少なくとも1つの変数についての優越性と、すべての変数の非劣性が同時に成り立つ場合にのみ治療効果を認める試験の枠組みに適用できる新たな検定統計量を2種類提案している。また、モンテカルロシミュレーションにより提案手法の第一種の過誤確率と経験検出力を評価し、どちらの手法がより優れているかについても検討しており、実用上有益な情報が提示されている。

第4章では、第3章と同様の試験の枠組みにおいて、複数の二値評価項目の背後に潜在的な連続分布が仮定できる場合の検定統計量を提案している。これは臨床の現場で使われる二値評価変数には、潜在的に連続型の分布をもつ評価変数を二値化して得られるものが少なくないからである。本章で提案された手法についても、シミュレーションを通してその性能が評価されている。さらに、想定している試験の枠組みにおいて用いることのできる既存の多重比較法との性能比較も行われており、提案手法はより高い検出力を示すことが述べられている。これは本手法を用いることで既存手法を用いるよりもサンプルサイズを削減できることを示しており、応用上極めて意義のある結果である。

第5章では、本論文の総括が示されている。また、第3章と第4章で提案された方法及び既存手法との比較が、シミュレーション実験を用いて行われ、各手法の性能評価・考察がなされている。さらに、複数の二値評価変数または連続型評価変数を用いた関連研究を紹介している。

以上、本論文は理論面と応用面の両面において大変高く評価できるものであり、臨床試験における統計解析及び分割表統計解析の分野に、独創的な新しい解析法を与えてお

り，この分野に大きな貢献をしている．よって理学的に価値ある知見と成果を得たもので博士（理学）の学位論文として十分価値あるものと認める．