—統計学的観点から見たランダム化比較試験—

鈴木　直子 (*SUZUKI Naoko*)[1]*　田中　瑞穂 (*TANAKA Mizuho*)[1]　佐野　友紀 (*SANO Yuki*)[1]
柿沼　俊光 (*KAKINUMA Toshihiro*)[1]　馬場　亜沙美 (*BABA Asami*)[1]　山本　和雄 (*YAMAMOTO Kazuo*)[1]

Key Words：ヒト試験，健康食品，特定保健用食品，機能性表示食品，ランダム化比較試験

Current Status and Issues of Clinical Trials for Efficacy and Safety Evaluation of Health Foods —Statistical analysis in randomized controlled trial—

Keywords: clinical trials, health food, Foods for Specified Health Uses (FOSHU), Foods with Function Claims, randomized controlled trial

Authors:
Naoko Suzuki[1)]*, Mizuho Tanaka[1)], Yuki Sano[1)], Toshihiro Kakinuma[1)], Asami Baba[1)], Kazuo Yamamoto[1)]
*Correspondence author: Naoko Suzuki

Affiliated institution
[1)] ORTHOMEDICO Inc.
2F Sumitomo Fudosan Korakuen Bldg., 1-4-1 Koishikawa, Bunkyo-ku, Tokyo, 112-0002, Japan.

はじめに

前回は，健康食品の有効性・安全性評価によく用いられる試験デザインの 1 つであるランダム化比較試験の概要と試験例を紹介した。第 3 回では，統計学的観点からランダム比較試験（RCT）について紹介する。

1．統計学の必要性

統計学とは

統計学は，方法論であり，作物の収穫率の向上方法，薬剤効果や治療効果を正確に評価するために発展してきた。統計学の柱は，推論と予測であり，科学的根拠を得る際に，矛盾を生じさせないように進めるための論理基盤を付与するものである。加えて，ばらつきを考慮する学問でもある。様々な事象に対して，全ての結果が一致する事象は少ない。例えば，ある睡眠薬の薬効を評価する際に，被験者の生理状態が同じであったとしても，たまたま疲れていて睡眠時間が長くなった者や実験時に神経が高ぶって眠れなかった者など偶然により誤差が生じてしまうことがある。また，ある漢方薬に不定愁訴を改善する効果があるという医師もいれば，全く効果がないという医師もいる。このように，生物学や医学における真実の追及には，ばらつきが混入するため，比較する 2 群間の差が偶然のばらつきよりも大きいかどうか，統計学を用いて評価する。

また統計学は，以下のようにデータの説明や記述に関する記述統計学と，標本から母集団の推測に関する推計統計学の 2 つの学問に大きく分けられる。

[1] 株式会社オルトメディコ　* 責任著者
〒 112-0002 東京都文京区小石川 1-4-1 住友不動産後楽園ビル 2 階
Tel: 03-3818-0610 / Fax: 03-3812-0670

< 記述統計学 >

記述統計学とは，あるデータに関して，そのデータが有する特性を記述，説明することを目的とした分野である。例えば，中学生の 50 人の身長データを取得したとする。この時，このデータを一見するだけでは，特性は見えてこない。ここで，このデータの中央値や平均値などを求めることで，この集団の身長の特性が見えてくる。このように，データの特性を記述することが記述統計学の大きな目的である。

< 推計統計学 >

推計統計学とは，標本から母集団を推定することを目的とした分野である。例えば，日本の中学生の身長特性を調査したいとする。この時，日本のすべての中学生の身長を取得し，記述統計学を用いることで解明することはできる。しかしながら，この方法は非常に困難である。そこで，推計統計学では，全国の中学生（母集団）からランダムに身長データ（標本）を取得する。そして，この標本の特性からこの母集団の特性を推測する。このようにして，標本から母集団を推測することが推計統計学の目的である。

臨床試験においては，後者の推計統計学が重要になってくるが，標本の記述統計から母集団の統計量を推測するため，この 2 つの分野は密接に絡み合っている。

研究デザインの設計における統計学の必要性

統計学者の臨床試験における主な役割は，提示されたデータに対して適切な統計手法を適用する知識を提供することである。そのため，正しい統計手法を適用することばかりに目が向きがちである。しかし，それ以前に収集されたデータに問題があるとの指摘が増えている。それは捏造のようなデータそのものにおける問題ではなく，研究自体は適切に実施されたにもかかわらず，研究デザインの組み方を誤ったために，収集されたデータにバイアスが生じてしまうということである。また，ずさんなデータ管理もバイアスやデータのノイズにつながるため，データの質の水準を維持するようなデータ管理も要求されており，特に大規模で長期な研究において重要となる。研究において，統計解析はバイアスが少なく，適切に取得された質の高いデータを用いなければ意味のある結果を得ることはできない。このような中で，統計学者が行うべき役割は，GSP（Good Statistical Practice）として位置づけられている。製薬企業ですでに用いられている GCP（Good Clinical Practice）や GLP（Good Laboratory Practice）と同様に，この GSP も近年重視されている [1]。欧州の製薬企業で働く統計学者らが，それを SOP（Standard Operating Procedures）としてまとめた。

ヒト試験を実施する際には，ほぼ必ず研究計画（プロトコル）を作成する。その内容は, 仮説, 対象, 評価指標, 統計学的考察（症例数や試験期間の根拠, 統計解析手法）から構成される。仮説は，できる限り具体的にし，対象を明確にする。また，評価指標の測定や観察を一定の方法で実施することも重要である。そのためには，判定する者を少なくする方法や客観的に測定できる指標を用いる方法などが考えられる。ほかに血圧のように 1 回測定しただけでは不安定な指標の場合は 2 回測定して平均をとるなどの操作も必要である。研究デザインや評価項目の特性を考慮し，適切なデータ取得方法や統計解析手法を選択するためには, 統計学の知識が不可欠である。したがって，プロトコルを作成する際，必要に応じて統計学者もメンバーに加えるとよい。特に，仮説の証明が困難な場合や大規模な研究の場合，関連因子が多く，複雑に絡み合っている場合にその必要性が増してくる。このように，統計学は解析段階だけでなくプロトコル作成から結果のデータを解釈する段階まで様々な段階でかかわってくる。これまで研究の計画段階や結果を解釈する段階において，統計学はあまりかかわってこなかった。また，**表 1** に示したように，医学雑誌である Lancet が実施した調査でも同様に，統計解析よりもデザインの記述で不備を指摘した例が多いことを報告している [2]。研

表 1　Lancet において統計学者がデザインの不備に関して指摘した個所の割合（一部改変）[2]

指摘箇所	不備の割合
適格条件	25%
検出力	14%
基本的なデザイン	14%
比較	10%
ランダム化	9%

究の失敗を防ぐために，どのような点を事前に注意しておくべきか。統計学は，そのための有益な考え方を提供することができる。

2. 研究デザインの設計と統計学

目的の選択

前述したように目的は，可能な限り具体的に設定することが求められる。仮説がある場合，まず先行研究があるかどうかを調査し，目的を検討する。次に，実際に細胞実験や動物実験，臨床試験を実施し仮説を検証するか決める。特に，臨床試験を実施する際は，事前にプロトコルを作成し，実施施設の倫理委員会で審査および被験者の同意を得ることが必須となる。これは，基本倫理要綱であるヘルシンキ宣言[3]でうたわれている。

仮説の証明には段階がある。ヒトを対象とした研究をする場合，不用意に多くのヒトを対象とした試験をはじめから計画し実施することには，倫理的問題も多い。そこで，パイロット試験を本試験の前に実施することもある。パイロット試験とは，本試験を実施する前の探索的な小規模試験であり，本試験を実施するか判断するために実施される。一方，本試験の目的は，仮説の検証である。パイロット試験から本試験の流れで臨床試験を実施した例として，ある血栓溶解剤の急性心筋梗塞患者への臨床効果を検証した研究グループは，まず 100 人の患者でパイロット試験を行い，薬剤起因性の脳内出血の発生頻度が高くない事を確認した後，本試験を実施した[4]。

母集団と標本とその選択

臨床試験において，母集団と標本は頻繁に耳にする用語だと思う。母集団は研究の対象となるすべての被験者を網羅する集団であり，標本は母集団からランダムに抽出された集団を指す。母集団を研究対象として用いることは理想であるが，すべてを網羅することは困難であるため標本を抽出する。この標本に対して，臨床試験をはじめとした調査を実施し，その結果を母集団へ回帰する。このように標本の結果から，母集団の結果を推測するため，母集団を代表できるような標本を選択することが非常に重要となる。

例えば，成人女性の痩身願望の強さと食習慣を調査したい場合，対象集団は成人女性全般とし，痩身願望が強い女性は，間食も少なく，少食で栄養バランスを考えた食事をしているのではないかという仮説を立てる。この時，栄養系大学の女性学生を標本とするとどうなるのであろうか。彼女らもほとんどが成人女性ではあろう。しかし，栄養に対する意識は高く，食生活は優れていると考えられるが，それは痩身願望が強いためではないかもしれない。このような可能性が考えられる場合，その標本は対象となる母集団を忠実に反映していない。つまり，偏った標本となる。適切な標本が選ばれると外部妥当性あるいは一般可能性の保証になる。

エンドポイントの選択

エンドポイントは目的に応じて，適切なものを選択する。例えば，感染症への感染を抑えることが目的なら，感染の有無がエンドポイントになる。他にも，目的が腸内環境の改善ならばビフィズス菌の占有率など，内臓脂肪の減少ならば，内臓脂肪面積などがエンドポイントになる。ほかにも主観的な評価として，MOS Short-Form 36-Item Health Survey（SF-36）[5,6]や MDQ（Menstrual Distress Questionnaire）日本語版[7,8]のような QOL を評価するアンケートを用いる場合もある。

実際に試験を計画する際には，エンドポイントを 1 つに絞ることが望ましい。様々なエンドポイントを評価，測定することはよいが，多重性の問題が生じてくる。例えば，A，B，C の食品があり，*t* 検定を用いて 3 つの食品間（A-B，A-C，B-C）の比較をしたとする。有意水準（差がないのに誤って差があると評価してしまう確率）を 5% とすると，3 つのうち 1 つでも有意になってしまう確率は 14.3% まで上がる。つまり，評価するエンドポイントや測定時点が増えるほど，誤って差があると評価してしまう危険性が高まる。また，機能性表示食品届出におけるこの問題に関して，「機能性表示食品に対する食品表示等関係法令に基づく事後的規制（事後チェック）の透明性の確保等に関する指針」では，「2. 科学的根拠として明らかに適切とは考えられない具体例」の「(1) 最終試験を用いた臨床試験（ヒト試験）及び研究レビューに共通する事項」に「ア . 届出資料において，表示する機能性に見合ったリサーチクエスチョン（PICO または PECO）は設定されているが，表示の内容が，科学的根拠の内容に比べ過大

である，又は当該根拠との関係性が認められない場合」の例として，プライマリーエンドポイントにおいて表示する有意な結果が得られていないものや表示する機能性について，プライマリーエンドポイントが複数設定されている場合であって，一部のエンドポイントで有意な結果が得られているが他のエンドポイントで有意な結果が得られていないときに，その関連性を踏まえた説明がされないものとある[9]。加えて，システマティックレビューを実施する際に生じる問題事項とその対処として，『「機能性表示食品」制度における機能性に関する科学的根拠の検証 - 届け出られた研究レビューの質に関する検証事業報告書」』の「第 2 項 . 不適正・研究論理に反すると考えられる注意事項」には，多重検定の問題に対する対処の事例として，「プライマリーエンドポイントとして，必要な項目のみに絞り込んでおくことや，複数設定されている場合には，Bonferroni 法などを活用して，厳しく有意差を設定している一次研究を抽出するなどの取り扱いが求められる」とある[10]。したがって，機能性表示食品届出のためのヒト試験を実施する際は，多重性について今後，より慎重に検討していくことが求められると考えられる。

前述したように，多重性の問題の対処法は，厳しく有意差を設定する方法とプライマリーエンドポイントの数を限定する方法の 2 つに大別される。前者の方法の 1 つとして，Bonferroni 法が挙げられる。この方法は，*t* 検定で得られた P 値に繰り返した検定数を掛け合わせ，補正する。その補正後の P 値が有意水準を下回れば有意差ありと評価する。他にも，Dunnett 法や Holm 法など，様々な方法が挙げられる。後者の方法は，複数のエンドポイントを測定する際に，研究において臨床的に最も重要なエンドポイントをプライマリーエンドポイントとし，薬理学的なエンドポイントや副次的に測定したいエンドポイントをセカンダリーエンドポイントとする方法である。例えば，癌治療では，プライマリーを死亡（延命効果）と QOL とし，セカンダリーを腫瘍縮小としたりする。加えて，エンドポイントは明確に定義しておく必要がある。内臓脂肪面積を例にすると，2 群間の比較するとき，最終検査の内臓脂肪面積の実測値または摂取前からの変化量で比較するのか，あるいは一定以上内臓脂肪が減少した者の人数を比較するのか決めておく必要がある。

ここで，機能性表示食品と医薬品の評価指標の設定の違いについて紹介する。機能性表示食品では，「学会等により健康の維持・増進に対する医学的及び栄養学的意義が十分に評価され，広く受け入れられているもの」とされている[11]。一方で，医薬品では，「プライマリーエンドポイントは，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エンドポイントであるべきである」とされている[11]。加えて，プライマリーエンドポイントは通常 1 つにし，先行研究または公表論文で使用された実績のある妥当性や信頼性を得られたエンドポイントを使用することが薦められる[11]。実際に，機能性表示食品と医薬品のコレステロールに関するプライマリーエンドポイントを比較すると，機能性表示食品では実測値や変化量などの記載がないのに対し，医薬品においては「12 週間後の変化率」と評価時期や評価方法が明確に定義されている（**表 2**，**表 3**）。現行制度上，機能性表示食品におけるアウトカムは「学会等により健康の維持・増進に対する医学的及び栄養学的意義が十分に評価され，広く受け入れられているもの」とされるが[11]，検証する対象物を評価する観点から考えれば，機能性表示食品も医薬品もその評価方法に違いはないはずである。言い換えれば，機能性表示食品においても，医薬品と同様に，実測値や変化量など使用された実績のある妥当性や信頼性が得られた評価方法を選択す

表 2　機能性表示食品の有効性に関する評価指標など (一部改変)[11]

ヘルスクレーム	プライマリーエンドポイント	試験期間
コレステロール	LDL- コレステロール	12 週
血圧	血圧	12 週
体脂肪	腹部脂肪面積，BMI，腹囲	12 週
整腸	排便回数，排便量，便性状，糞便菌叢	2 週以上

表 3　医薬品 (コレステロール) の有効性に関する評価指標など (一部改変)[11]

対象疾患名 (Clinical Trial ID)	プライマリーエンドポイント	試験期間
高コレステロール血症 (NCT02260648)	Percent Change from Baseline to Week 12 in Low-Density Lipoprotein Cholesterol (LDL-C) Measured by Beta Quantification (Time Flame: Baseline, Week 12)	12 週
高コレステロール血症 , 家族性高コレステロール血症 (NCT02550288)	治療期 12 週時における LDL-C のベースラインからの変化率 (12 週時での LDL-C 値を評価する)	12 週
高コレステロール血症 (NCT02741245)	LDL-C のベースラインからの変化率 (ベースラインと投与 12 週後の LDL-C の変化率を貯砂する)	2 週以上

れば，機能性表示食品制度における「学会等により健康の維持・増進に対する医学的及び栄養学的意義が十分に評価され，広く受け入れられているもの」[11] の記述に対する条件は満たすものと考えられるが，この点については今後の業界動向から慎重に解釈していく必要がある。

その他，エンドポイントに影響する可能性を示す変数となる因子を調査しておくことも必要がある。この因子は, エンドポイントにより大きく異なるが, 性別や年齢，運動習慣をはじめとした生活習慣などが挙げられる。しかしながら，このような因子になりうる項目は多く存在するが，エンドポイントに関連しない変数を増やすことは好ましくないので，プロトコル作成段階で検討し，調査する項目を明確にしておく必要がある。

3. データ解析と統計学

解析対象集団と欠損値の取扱い

すべての割付症例がプロトコル通りに試験を完了することは少ない。例えば不適格症例，介入の入れ替わり，途中脱落など様々な違反が生じる。このような症例の解析段階における取り扱いは慎重に行う必要がある。安易に除外すればよいわけではない。RCT では，ランダム化を行っているためである。ランダム化とは，各群に被験者をランダムに割り振ることによって，各群の被験者の背景因子を，未知のものを含めほぼ均一にし，各群を同一の集団として扱うための方法である。したがって，問題がある症例を安易に除外すると均一性がくずれ，同一の集団ではなくなってしまう危険性がある。そのため問題のある症例の取り扱いに応じて，解析対象集団は 3 種類に大別される。

<Intention-to-treat（ITT）>

意図された介入を受けた群を解析対象とするランダム化の保持を目的とした手法。この手法では，有害事象による脱落などの負の効果も含めて解析するため, 介入の実用的な価値を評価することができる。よって，ITT で効果ありと結論付けられた場合，その結果はより強固なものだといえる。ITT は理想的だが，達成は非常に困難である。

<Full analysis set（FAS）>

①介入を一度も受けていない，②割付後のデータが欠損という被験者を解析集団から除外する手法。ITT と同様に FAS も，ランダム化の保持を目的としている。ITT と異なる点は，ランダム化の保持をしつつも，最小限の被験者を除外することで介入の効果を検証できる点である。

<Per protocol set（PPS）>

①介入が遵守されている，②重大なプロトコル違反がない，③データの利用が可能というような被験者のみを解析対象とする手法。純粋に介入の効果を評価することができる。PPS では，途中脱落やデータ不足が生じた場合, 解析から除外されてしまう為, 研究開始時と終了時で患者背景が大きく変わる可能性があり，結果が現実世界に反映できるかは評価できない。非劣性試験や安全性の評価をする試験の場合は PPS が適切である。

実際的な立場に立てば，不用意なバイアスが生じることを防ぐために ITT 解析を行う[12]。つまり，割り付けられた症例はすべて解析し，割付後にいかなるプロトコル違反があったとしてもそのまま解

表 4　データ欠損のメカニズム

メカニズム	内容
MCAR	欠測が完全にランダムで生じた場合を指す。これは，欠測値の有無が他の分析に含まれる変数やその変数自体の値とは無関係であるということである。例えば，仕事の都合による転居で来院できなかったために欠測値が生じるような場合や測定装置がランダムな動作不具合を起こし欠測値が生じるような場合があり，これらはランダムに生じるため主要評価項目・有害事象の発現などとは全く無関係でありバイアスを生じない。
MAR	データが測定されている値に依存して欠測 (欠損データとは無関係) した場合を指す。つまり観測データに依存した欠測であり，欠測値の生じるメカニズムが，観測されている変数で全て完全に説明することができる状態である。原疾患の悪化や有害事象の発現などに関連しうる理由による欠測の場合などであり，特に試験を中止した時点で悪化しているデータが十分にあるか検査結果を見た上で中止を判断した場合である。
MNAR	データが欠損データに依存して欠損する場合である。これは，分析に含まれる他の変数を統制した後でも，欠測値の有無が欠測値を持つ変数自身と関係を持つケースを示している。例えば，喫煙を止めさせる研究において，来院しなくなったら喫煙を疑うことになる。このような場合，バイアスを生じさせることになる。主要評価項目・有害事象の発現などと関係し得る理由による欠測であり，来院間隔が広く欠測の原因となったデータが得られていないなど，欠測が発現した段階で欠測を説明しきれるだけの十分なデータが得られていない場合に当てはまる。

析する。長期的に死亡率や罹患率を評価するための試験では，多くの研究でこの ITT を主要な解析方針にしている。エンドポイントとして，有効性の指標が明示してあるにも関わらず，重大なプロトコル違反をしているからと言って，不用意に解析除外しない事が重要な点である。しかし，エンドポイントが欠損値であれば解析に問題が生じる場合がある。

欠測値が生じるメカニズムを**表 4** に示した。欠測が生じるメカニズムは 3 種類に分類され，完全にランダムに生じる場合（Missing Completely At Random; MCAR），測定されている値に依存している場合（Missing At Random; MAR），欠損データに依存している場合（Missing Not At Random; MNAR）である。特に，MAR と MNAR はバイアスが生じるので，試験の計画段階で可能な限りこのような欠損データが生じないように対策を取り，生じた場合の取り扱い方法も決めておくことが重要である。

前述した欠損データの取り扱い方法は，試験を完遂しなかった被験者のデータを削除して解析する方法（complete case 解析），完遂できなかった被験者の欠損データを補完して解析する方法 {Last Observation Carried Forward（LOCF）解析，Baseline Observation Carried Forward（BOCF）解析 }，完遂できなかった被験者の欠損データを補完せず観測された全データを用いて解析する方法（Mixed-effect Models for Repeated Measures; MMRM）などがある。ここで挙げた欠損データの取り扱いと欠損メカニズムの対応は**表 5** に示した。

表 5　欠損メカニズムと欠損に対する解析の対応表

	MCAR	MAR	MNAR
Complete case 解析	○	×	×
LOCF 解析	×	×	○
BOCF 解析	×	×	○
MMRM 解析	×	○	×

○：適用可，×：適用不可

< Complete case 解析 >

完遂していない被験者のデータを削除して，全評価時点の測定値が観測された被験者に対して解析計画に基づいた解析方法を適用する方法であり，欠測値の取り扱いとして最も単純なアプローチである。解析の実行が非常に容易である点がメリットである。Complete case 解析が根拠のある妥当な推定結果を与える状況は，欠測メカニズムが MCAR のもとでは幾つか考えられるが，一般的には不適切であることが多い。完了例ではない被験者のデータを用いないまま解析を行うため，バイアスを引き起こす可能性があること及び精度の低下の 2 点で問題があると考えられる。

<LOCF 解析 >

単一補完の方法であり単調な欠測と非単調な欠測の両方に対して適用可能である。また，プロトコルで規定された時点で評価を行う試験において被験者の応答変数が脱落後に変化しないという強い仮定が

成り立つ場合に導かれる応答変数の推定量にバイアスが入らないとされる補完法である。欠損値が存在しないデータを作成し，解析を実施するため比較的理解しやすい点がメリットである。LOCF を妥当とする欠測メカニズムの仮定は MNAR である。

<BOCF 解析 >

LOCF と同様の単一補完の手法である。完了例ではない被験者の欠測値をその被験者のベースライン測定値で補完する手法である。つまり，ベースラインからの変化量をエンドポイントとする場合には欠測値は全て 0 で置き換えられる。BOCF は，脱落後には介入効果が消え，介入前の状態に戻ると想定することが妥当と考えられる場合に選択が可能である。一方で LOCF に対する論説は BOCF に対しても同様にあてはまる。LOCF や BOCF といった単一補完の方法は，データの不確実性を単一の値で補完することで無視することに繋がり，真のデータの複雑さとは異なるものになってしまうと考えられる。

<MMRM 解析 >

MMRM は線形混合モデルを用いた解析手法である。これは経時測定データの場合に，欠損値を補完することなく，応答に対して特定の確率分布と回帰モデルを仮定することによって解析を行う手法である。そして，妥当な結果を得るためには，MAR の仮定が必要となる。また，MAR の場合においてバイアスのない解析を行うこと，および検定・信頼区間を通常の線形混合モデルと同様に取り扱うことが可能である。一方, モデルが非常に複雑であること，MNAR の場合にバイアスが入る可能性があるという欠点を有する。また，この解析手法は脱落割合，時点間の相関性の影響を受けるため，症例設計を適切に行わなければならない。そして，不十分な症例数であった場合，適切な推定が行えず，バイアスを含む結果となることが考えられる。

ここまで述べてきたように，臨床試験において，すべての被験者が試験をプロトコル通りに完遂することは少ない。そのため，適切な解析対象集団や欠損値に対する取り扱い方法を選択することが求められている。

経時的繰り返し測定の評価方法

臨床試験において，経時的にデータを測定することは少なくない。経時的な測定データの特徴は，同一対象に対し繰り返し測定することから必然的に測定値間に相関が生じることである。また，*t* 検定を用いた 2 群間の平均値の比較を各測定時点で繰り返し行った場合，時点数が多ければ多いほど有意差が生じやすくなり，時点の多重性の問題が生じる。このような問題の対処方法として，介入期間の最終時点での測定値の比較，事前に定めた改善の定義を達成した被験者の割合の比較などが考えられる。

4. 結果の解釈と統計学

統計的有意性と臨床的意義

有意確率（P 値）は，統計学的有意性を評価するための指標として通常用いられている。P 値が有意水準を下回った際に，有意差があると言える。一般的に有意水準は 0.05 に設定されることが多いが，その定義はあいまいである。例えば，Fisher は 0.02 だと帰無仮説を強く否定するが，0.05 であれば迷うことはないだろうとした [13]。また，Victor Cohn はフリップコインで 5 回続けてコインの表が出ると声を上げるとした。この確率は 0.03 であり，0.05 を下回ると人間は異常を感じるということから，有意水準が 0.05 となったのだろうとしている。有意水準を P 値が下回った際に有意差があるとして，多くの科学的な結論に用いられているが，誤用と誤解がしばしば見過ごされている。代表的な誤解として，有意差があれば効果があると判断してしまうこと，P 値が小さいほど効果が大きいと認識してしまうことの 2 つが挙げられる。例えば，体脂肪率を減少させる機能性が期待される緑茶を 8 週間摂取する臨床試験を実施したとする。その結果として，摂取 8 週間後の体脂肪率において，有効成分を含まない食品群（プラセボ群）と有効成分を含む食品群（被験食品群）に有意差が確認され，その群間差（被験食品－プラセボ群）が 0.1% であった際，被験食品は体脂肪率を減少させる効果があったと報告されることがある。しかし，実際に，8 週間で体脂肪率を 0.1% 減少させる効果があったとして，その差に意味があるのだろうか。その差に意味があるかどうかを評価するには，統計学的に差があるかどうかではなく，臨床的に意味のある差であるかどうかを評価する必要が生じる。つまり，P 値は 2 つの母集団が統計学的に異なった母集団かどうかを評価するのみで，そ

表 6　統計学的優位性と P 値に関する ASA 声明 [15, 16]

No.	内容
1	P 値はデータと特定の統計モデル（訳注 : 仮説も統計モデルの要素の一つ）が矛盾する程度を示す指標の一つである。 P-values can indicate how compatible the data are with a specified statistical model.
2	P 値は，調べている仮説が正しい確率や，データが偶然のみでえられた確率を測るものではない。 P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3	科学的な結論や，ビジネス，政策における決定は，P 値がある値（訳注 : 有意水準）を超えたかどうかのみに基づくべきではない。 Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4	適正な推測のためには，すべてを報告する透明性が必要である。 Proper inference requires full reporting and transparency.
5	P 値や統計学的有意性は，効果の大きさや結果の重要性を意味しない。 A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6	P 値は，それだけでは統計モデルや仮説に関するエビデンスの，良い指標とはならない。 By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

の効果の大きさや結果の重要性を示す値ではないのである。

このような誤解や誤用が見過ごされている中で，2016 年にアメリカ統計協会（American Statistical Association，ASA）が P 値の適切な使用と解釈の原則に関する声明を発表した [14, 15]。その声明には，**表 6** に示した 6 つの原則が提示されている。また 2019 年にはこの声明を踏まえ，自然科学雑誌である Nature において，統計学的な有意差のみで結果を判断するのは危険であるという考えに 800 人以上の科学者が賛同したとの記事が掲載された [16]。このように，P 値の誤用と誤解を理解し，統計学的有意性と臨床的意義を適切に取り扱っていくことで，より強い科学的根拠を有する結果を示すことができると考えられる。

おわりに

本稿では，統計学的観点からランダム化比較試験を紹介した。次回はクロスオーバー試験の概要および試験例を紹介する予定である。

参考文献

1. Wiles A, Atkinson G, Huson L, Morse P, Struthers L.: Good statistical practice in clinical research: Guideline standard operating procedures. *Ther. Innov. Regul. Sci.* **28** (2): 615-627, 1994.
2. Gore SM, Jones G, Thompson S.: The Lancet's statistical revview process: areas for improvement by authers. *Lancet* **340**: 100-102, 1992.
3. World Medical Association Declaration of Helsinki Ethical Principles for Medical Research Involving Human Subjects Adopted. *Bull. World Health Organ.* **79** (4): 373-374, 2001.
4. Topol T, Califf R, Van de Werf F, Armstrong P, Aylward P, *et al*.: An International Randomized Trial Comparing Four Thrombolytic Strategies for Acute Myogardial Infaction. *New English J. Med.* **329** (10): 673-682, 1993.
5. Fukuhara S, Bito S, Green J, Hsiao A, Kurokawa K: Translation, Adaptation, and Validation of the SF-36 Health Survey for Use in Japan. *J. Clin. Epidemiol.* **51** (11): 1037-1044, 1998.
6. Fukuhara S, Ware JE, Kosinski M, Wada S, Gandek B.: Psychometric and Clinical Tests of Validity of the Japanese SF-36 Health Survey. *J. Clin. Epidemiol.* **51** (11): 1045-1053, 1998.
7. Rudolf HM.: The development of a Menstrual Distress Questionnaire. *Psychosom. Med.* **30** (6): 853-867, 1968.
8. 秋山昭代，茅島江子：MDT (Mirror Drawing Test) からみた性周期の心身に及ぼす影響について．日看研会誌 **2** (2): 61-66, 1979.
9. 消費者庁：機能性表示食品に対する食品表示等関係法令に基づく事後的規制（事後チェック）の透明性の確保等に関する指針　2020.
10. 消費者庁：「機能性表示食品」制度における機能系に関する科学的根拠の検証 - 届け出られた研究レビューの質に関する検証事業 報告書　2016.
11. 種村菜奈枝，濱舘直史，漆原尚巳：レギュラトリーサイエンスの視点からみた医薬品と保健機能食品における有効性又は安全性の科学的根拠に必要な規制やその考え方の相違．日補完代替医療会誌 **14** (2): 47-60, 2017.
12. Gillings D, Koch G.: The Application of the Principle of Intention–to–Treat to the Analysis of Clinical Trials. *Drug Inf. J.* **25** (3): 411-424, 1991.
13. Fisher RA.: Statistical methods for research workers. Oliver and Boyd, London, p.80, 1950.
14. Wasserstein RL, Lazar NA. The ASA's Statement on p-Values: Context, Process, and Purpose. *Am. Stat.* **70** (2): 129-133, 2016.
15. 日本計量生物学会：統計的有意性と P 値に関する ASA 声明．2017.
16. Amrhein V, Greenland S, McShane B.: Scientists rise up against statistical significance. *Nature* **567** (7748): 305-307, 2019.