—ヒト臨床試験（ヒト試験）において前後比較デザインを採用する場合の統計学的留意事項—

鈴木　直子 (*SUZUKI Naoko*)[1*]　田中　瑞穂 (*TANAKA Mizuho*)[1]　野田　和彦 (*NODA Kazuhiko*)[1]
波多野　絵梨 (*HATANO Eri*)[1]　金子　拓矢 (*KANEKO Takuya*)[1]　中村　駿一 (*NAKAMURA Shunichi*)[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：ヒト臨床試験，ヒト試験，特定保健用食品，機能性表示食品，前後比較試験

Current Status and Issues of Clinical Trials for Efficacy and Safety Evaluation of Health Foods —Statistical considerations when adopting a before-after comparative design in clinical trials—

Keywords: clinical trials, Foods for Specified Health Uses (FOSHU), Foods with Function Claims, Before-after trial

Authors:
Naoko Suzuki[1)]*, Mizuho Tanaka[1)], Kazuhiko Noda[1)], Eri Hatano[1)], Takuya Kaneko[1)], Shunichi Nakamura[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

第 2 回から第 5 回にかけて，ランダム化比較試験（RCT）について紹介してきた。RCT は，機能性表示食品の制度下において推奨されている試験デザインであるが，第 6 回でも紹介したように，RCT の実施が困難な状況は存在する。本稿では，RCT の代替法の一つである，単群での前後比較試験（Before-after trial）における統計学的留意事項について紹介する。

1．RCT の限界

RCT は，Fisher が提唱したランダム割付を伴う比較試験のことを指し，臨床試験において最も信頼性の高い試験デザインである。ランダム割付のメリットとして，比較群の背景を自動的に類似させることや，系統誤差を偶然誤差に転化させることにより，妥当な有意性検定を行えることなどが挙げられる。RCT 実施に向けて，盲検化の有無や，プラセボ対照かどうか，対象集団の大きさなどを試験計画時に検討するが，特定保健用食品，機能性表示食品制度で採用されることが多いのは，「二重盲検ランダム化プラセボ対照並行群間比較」である。しかしながら，実際にはこのデザインを設定できない状況がある。以下，いくつかの状況を説明する。

1-1．二重盲検の設定が困難な状況

盲検化は，試験結果に偏りを生じさせる危険性を減少又は最小化する重要な方法である [1)]。盲検化の内，試験に関わるすべての関係者が割り付けられた介入を知らずに実施される方法を二重盲検と呼ぶ。

臨床試験はヒトを対象とした治療を兼ねた試験で

[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

あるため，診療の一部として認識されている。それ故，特に医薬品の研究において，二重盲検の実施を，医師に受け入れてもらうのは容易ではない。さらに，心血管イベントを評価する臨床試験では数年にわたり研究を行うが，その間どのような介入を受けているかわからないまま試験を実施することは，患者の不安などの精神的負担が非常に大きい。そこで，1980 年代後半により実施可能性の高い PROBE（prospective randomized open blinded endpoint）と呼ばれるデザインが欧米において新しく開発された[2]。PROBE 試験はランダム化オープン比較試験であるが，エンドポイントの評価を，参加者がいずれの群に割り付けられたかを知らない臨床試験実施機関から独立した組織が行うことにより，オープン試験の弱点を補っている。二重盲検との違いを**表 1** にまとめた。

PROBE 試験はオープン試験であることから，試験期間中には，医師や患者のバイアスが入り込む余地がある。また，介入実施者が割付内容を知っているため，恣意的に参加者を脱落させる危険性があり，結果を解釈する際には特に脱落の評価に注意が必要である。アウトカムにおいても，疾患による入院などの評価者の主観が入り込むようなソフトエンドポイントが設定されている場合は，アウトカムの評価頻度の調整が可能であるため注意が必要である。従って，エンドポイントの評価は，主観的評価ではなく，血圧低下の程度のような客観性のあるハードエンドポイントであることが条件となる。

1-2. ランダム化ができない状況

ランダム化の目的は，対象の選択・割付けでのバイアスを避け，介入以外の全ての要因をそろえて比較可能な集団を作ることである。特に，プラセボを対照とした RCT は，研究上最も信頼度が高いといわれる方法であるが，プラセボの投与あるいはランダム化のそれぞれが研究倫理において争点となっている。何が問題であるかというと，プラセボは薬効成分を含まない偽薬であるため，プラセボを摂取する参加者の健康状態が改善されない可能性がある。従って，プラセボを摂取させることは，医療行為という点において倫理的問題がある。一方，ランダム化は，ランダムに 2 つ以上の群に割付け，それぞれの群に介入を行うが，タバコが肺癌の 1 つの原因であることを証明することを RCT で行うとすると，ランダムに割付けられたどちらかの群にタバコを与えなければならない。極端な例ではあるが，場合によってはランダム化に倫理的問題があると言える。実際には，様々な理由により RCT を実施することが不可能で，コホート研究などの観察研究の結果から，推測される関連について因果関係を評価しなければならない場面は多く存在する。

1-3. 対照集団を十分に確保できない状況

N- アセチルグルタミン生成酵素（NAGS）欠乏による高アンモニア血症の発症者は世界で約 50 例と報告されている[3]。このような希少疾患を対象とする臨床試験の場合，十分な参加者を収集することは困難である。通常の臨床試験に必要な症例数を，合理的な時間枠の中では集めることができない場合に実施される試験を Small Clinical Trial（SCT）と呼ぶ。

SCT は，少数例での試験になるため，適切なランダム化や盲検化がされていないゆえにバイアスを回避できず，大きい治療効果が得られる傾向があるというメタ・アナリシスの報告がある[3,4]。また，少数例の試験では一般化可能性が欠如してしまうなどの問題がある。しかしながら，日本製薬工業協会

表 1　PROBE 試験と二重盲検試験の比較

	PROBE	二重盲検
各群への割付け	ランダム割付	ランダム割付
エンドポイント評価の信頼性	高い	高い
費用	比較的安い	高い
研究者，対象者のバイアス	バイアスがかかる可能性あり	バイアスは排除できる
日常臨床との類似性	日常臨床に近い	日常臨床とは異なる
患者のアドヒアランス	高い	低下しやすい
エンドポイント	制約あり	制約は少ない

は，この状況において有効性を示すために，評価変数，比較対象，エビデンスの質と量，情報の表し方の 4 点を整理するよう提言している[3]。これら 4 点の試験計画時における留意事項を下記にまとめた。

SCT は，参加者が少ないため 1 例あたりのデータに重みがある。精度の低い測定を用いると，標準誤差が大きくなる要因となるので，精度の高い測定を積極的に採用することを検討すべきである。また，評価する変数において，通常の試験では，1 つの評価変数の最終時点だけを用いた推測が行われることが多いが，SCT では，参加者が少ない分，経時的な変化を測定するなど 1 例あたりの情報量を増やすことを検討すべきである。経時測定データの場合には，時間を含む統計モデルを利用することにより，個人差を誤差から分離した分析が可能となる。比較対象については，ヒストリカル・コントロールを用いた外部対照試験やベースライン対照試験を用いて，医薬品が開発された例があるが，少数例ではバイアスの影響を受けやすいため，可能な限り同時対照の設定を検討すべきある。

SCT では，十分な検出力（$1-\beta$）を確保できないため，有意水準（α）の設定も考慮すべき事項である。実際に米国で承認された事例でも，不十分な検出力の下で検定を行い，統計的有意差が得られていないにも関わらず承認されたものがある[3]。一般に，α は 5%，$1-\beta$ は 80% から 90% に設定されるが，これらは絶対的な数値ではない。試験計画時に，α と β の非線形の関係を示したうえで，調整された α に基づいて評価を行うことも重要な考え方である。

2. 前後比較試験とは？

RCT の限界について上述したが，食品の機能性評価における RCT の限界として挙げられるのが，生鮮食品等のプラセボの作製が難しい商品である。生鮮食品においては，機能性関与成分の作用機序が解明されており，その機能性関与成分におけるプラセボ対照 RCT の研究レビューを通して届出することができるが，最終製品を用いるとなると RCT の実施が難しい。その解決策として，RCT の代替法の 1 つである前後比較試験がある。本章では，単群での前後比較試験における，デザインの基本的な枠組みや，統計解析手法および解析における留意点を解説する。

2-1. デザインの基本的な枠組み

前後比較試験のデザインの概略図を**図 1** に示した。前後比較試験は，名称の通り，同一の試験参加者の摂取前と摂取後の結果値を比較するデザインである。個人内での比較のため精度は高く，症例数が少なくて済むメリットがある。一方，同時対照群がないため比較可能性の問題や，オープン試験になるため，評価にバイアスがかかってしまうなどのデメリットがある。

2-2. 対応のある *t* 検定

前後比較試験において，データが連続変数で標本が 2 つの場合に *t* 検定が用いられる。一般的に，対応のある 2 標本のデータは，差あるいは比を求めて 1 変量として扱う。例として，血圧低下を評価する試験の模擬データを**表 2** に示した。

まず，母集団における介入前の母平均（μ_b），介入後の母平均（μ_a）の帰無仮説（H_0）と対立仮説（H_1）は以下の通りである。

$$H_0: \mu_a = \mu_b$$
$$H_1: \mu_a \neq \mu_b \quad (1)$$

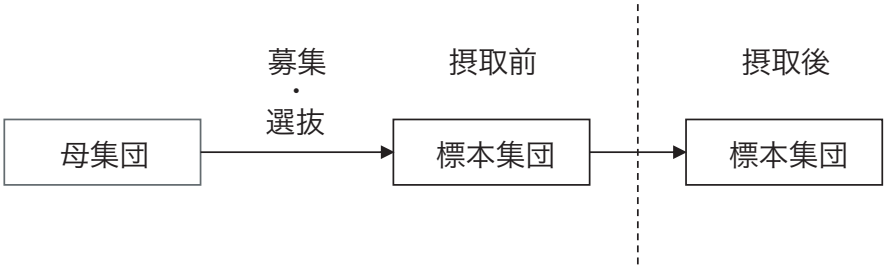

図 1　前後比較デザインの概略図

表 2　2 時点の前後比較デザインの模擬データ

試験参加者	摂取前	摂取後	平均	差（摂取後－摂取前）
No. 1	125	119	244	-6
No. 2	128	124	252	-4
No. 3	114	116	230	2
No. 4	120	114	234	-6
No. 5	130	110	240	-20
No. 6	119	123	242	4
No. 7	124	111	235	-13
No. 8	135	125	260	-10
No. 9	124	119	243	-5
No. 10	115	109	224	-6
平均	123	117	240	-6
標準偏差	6.26	5.62	9.94	6.56
標準誤差	1.98	1.78	3.14	2.07

表 3　3 時点の前後比較デザインの模擬データ

試験参加者	時点 1	時点 2	時点 3	行平均
No. 1	125	120	119	121
No. 2	128	127	124	126
No. 3	114	110	116	113
No. 4	120	119	114	118
No. 5	130	113	110	118
No. 6	119	117	123	120
No. 7	124	120	111	118
No. 8	135	132	125	131
No. 9	124	125	119	123
No. 10	115	110	109	111
平均	123	119	119	121

これらのデータは前後で対応があり，変化量（δ）を計算することにより 1 標本にすることができる。

$$H_0 : \delta = \mu_a - \mu_b = 0$$
$$H_1 : \delta = \mu_a - \mu_b \neq 0 \qquad (2)$$

この仮説を，模擬データを用いて両側有意水準 5% として検定する。対応のある *t* 検定は，変化量（δ），不変標準偏差（S），および自由度（n）を用いて以下の式で表される。

$$t = \frac{\sqrt{n}\delta}{S} \qquad (3)$$

この式（3）にデータを代入すると，「$|t| = 3.092 > t(9,0.05) = 2.262$」となり，有意確率 5% を下回っていることが分かる。ここで注意しなければならないことは，検定に用いるのは変化量の平均値とその標準偏差であって，介入前後の実測値の平均値とその標準偏差ではないということである。論文などで，介入前後の実測値のグラフを見かけることがあるが，そのようなグラフでは変化量の標準偏差についての情報は得られず，検定結果について見当をつけることはできない。従って，結果を表すには，変化量とその 95% 信頼区間で表現するのが合理的である。

2-3. 対応のある反復測定データにおける解析

上記で，2 標本に対する検定を紹介したが，実際には反復測定を実施する試験もしばしばある。そのような試験における解析手法についてまとめる。例として，**表 2** を拡張させ，測定時点を 3 つにした

模擬データを**表 3** に示した。このデータの場合，データを変動させる要因は，個人差と時点の 2 つと考えられる。例えば，試験参加者ごとの全ての時点の平均値がばらついているのは，個人差のためであり，時点ごとの全ての試験参加者の平均値がばらついているのは時点による変化と考えられる。このように，データを変動させる意味のある要因が 2 つある場合に，2 元配置分散分析が用いられる。この分析方法は，個人差を誤差から分離して効率の良い分析をすることができる。

表 3 のデータを一般化したものが**表 4** であり，それぞれのデータを y_{ab}（a：試験参加者，b：時点）とすると，2 元配置分散分析は以下の統計モデルで表すことができる。

$$y_{ij} = \mu + a_i + b_j + \varepsilon_{ij} \tag{4}$$

ここで，a，b，ε はそれぞれ試験参加者（個人差），時期の効果，誤差を表す。このモデルに従って，平方和，自由度，平均平方和などを算出すると，**表 5** のような分散分析表が作成される。以下，分散分析表作成のステップを紹介する。

まず，分散分析表を構成する要素である平方和を算出する。全体（S_T），個人差（S_A），時点（S_B），誤差（S_R）それぞれの平方和は以下の式を用いて算出する。

$$S_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\left(y_{ij} - m_T\right)^2 \tag{5}$$

$$S_A = \sum_{i=1}^{a}\sum_{j=1}^{b}\left(m_{i\cdot} - m_T\right)^2 \tag{6}$$

$$S_B = \sum_{i=1}^{a}\sum_{j=1}^{b}\left(m_{\cdot j} - m_T\right)^2 \tag{7}$$

$$S_R = S_T - S_A - S_B \tag{8}$$

全体，個人差，時点，誤差それぞれの自由度（ϕ）は，以下の通りである。

$$\phi_T = ab - 1 \tag{9}$$

$$\phi_A = a - 1 \tag{10}$$

$$\phi_B = b - 1 \tag{11}$$

$$\phi_R = \phi_T - \phi_A - \phi_B = \phi_A \times \phi_B \tag{12}$$

最後に，**表 5** に示した式で平均平方和および分散比を算出することにより，分散分析表は完成する。分散比まで求めることができれば，あとは F 確率分布にあてはめて有意確率を算出するだけである。**表 3** のデータを用いて分散分析表を作成したものが**表 6** である。ここで算出された結果を解釈するために，個人差，時点の効果の統計学的仮説を整理する。個人差の帰無仮説は「H_0：試験参加者ごとの平均値は全て等しい。」となり，時点の帰無仮説は「H_0：時点ごとの平均値は全て等しい。」となる。まず，個人差の検定では，有意確率 5% を下回ったことから，試験参加者ごとの平均値はばらついていることが分かる。しかし，この個人差は効果を分析するのが目的ではなく，誤差を減らすことが目的であるため，分散比の方が重要となる。今回の結果では，個人差の分散は，誤差の分散の約 6 倍であったことから，誤差から個人差を分離し，解析の効率が良くなっていることが理解できる。次に，この分析の主目的である時点の検定については，有意確率 5% を下回っていることから，帰無仮説は棄却され，対立仮説「H_1：時点ごとの平均値は異なる。」が採用さ

表 4　前後比較デザインの一般化データ

要因	B_1	B_j	B_b	計	平均
A_1	y_{11}	y_{1j}	y_{1b}	$T_{1\cdot}$	$m_{1\cdot}$
:	:	:	:	:	:
A_i	y_{i1}	y_{ij}	y_{ib}	$T_{i\cdot}$	$m_{i\cdot}$
:	:	:	:	:	:
A_a	y_{a1}	y_{aj}	y_{ab}	$T_{a\cdot}$	$m_{a\cdot}$
小計	$T_{\cdot 1}$	$T_{\cdot j}$	$T_{\cdot b}$	T_T	-
平均	$m_{\cdot 1}$	$m_{\cdot j}$	$m_{\cdot b}$	-	m_T

表 5　前後比較デザインの分散分析表

要因	平方和 (S)	自由度 (ϕ)	平均平方和（分散）(V)	分散比 (F)
A	S_A	ϕ_A	$V_A=S_A/\phi_A$	$F_A=V_A/V_R$
B	S_B	ϕ_B	$V_B=S_B/\phi_B$	$F_B=V_B/V_R$
残差	S_R	ϕ_R	$V_R=S_R/\phi_R$	-
全体	S_T	ϕ_T	-	-

表 6　前後比較デザインの模擬データにおける因子に時点と個人差を含む分散分析表

要因	平方和 (S)	自由度 (ϕ)	平均平方和（分散）(V)	分散比 (F)	有意確率
時点	210.200	2	105.100	6.468	0.000
個人差	888.033	9	98.670	6.073	0.008
誤差	292.467	18	16.248	-	0.001
全体	1390.700	29	-	-	-

表 7　前後比較デザインの模擬データにおける因子が時点のみの分散分析表

要因	平方和 (S)	自由度 (ϕ)	平均平方和（分散）(V)	分散比 (F)	有意確率
時点	210.200	2	105.100	2.404	0.109
誤差	1180.500	27	43.722	-	-
全体	1390.700	29	-	-	-

れる。

では，個人差を誤差から分離しなかったらどのようになるのかを検証する。**表 7** に個人差を誤差に含めた分散分析の結果を示した。**表 6** と **表 7** を比べると誤差の分散が大きくなり，その結果，時点の検定結果が有意確率 5% を上回るなど，検定や推定の精度が悪くなることが理解できる。ヒトを対象とした試験のデータは個人差が大きいものが多いため，個人差を誤差から分離して解析することは必要不可欠であると言える。

今回は，時点の水準数は 3 を想定したが，水準数を 2 にすると，両側検定の対応のある t 検定と一致する。対応のある t 検定よりも分散分析のほうが細かい分析が可能なため，対応のある 2 標本の平均値を両側検定で比較したい時は，分散分析を用いたほうが便利である。

3. 前後比較試験を採用する際の留意点

機能性表示食品に関する質疑応答集において，**表 8** に示した記載があった [5]。ここで, 前後比較は，機能性の科学的根拠として不十分であることについてまとめた。

ヒトを対象とした介入効果を検証する試験では，心理的なバイアスの制御が重要である。介入結果には，3 つの心理的要因が影響することが知られている。1 つ目は，効果を期待することにより効果が表れる「プラセボ効果」，2 つ目は，介入を受ける者が評価・観察者の期待に応えようとする「ホーソン効果」，介入実施者が期待をもって関わることによる「ピグマリオン効果」である [6]。二重盲検ランダム化プラセボ対照並行群間比較が推奨されている理由は，これらの心理的要因を排除できることが理由の 1 つであると考えられる。一方，単群の前後比較試験の場合，プラセボ効果やホーソン効果，ピグマリオン効果の影響を排除することができない。また，プラセボ効果は，実質的効果と分離して測定することは不可能で, 本来の結果とみなす必要がある。従って，前後比較試験で得られた結果は，このような心理的要因が含まれていることを念頭に置かなければならない。

まとめ

本稿は，機能性表示食品制度下における前後比較試験について統計学的留意事項をまとめた。前後比較試験は，対照を設定できない場合でも実施できることや，少人数での実施が可能のため，コストがか

表 8　機能性表示食品に関する質疑応答集 [5)]

問 45. ガイドラインにおいて、「本ガイドラインにおける「臨床試験（ヒト試験）」は、「特定保健用食品の表示許可等について」（平成 26 年 10 月 30 日付け消食表第 259 号消費者庁次長通知）の別添 2「特定保健用食品申請に係る申請書作成上の留意事項」で規定する「ヒトを対象とした試験」を指す。」とあるが、機能性については、試験食摂取群とプラセボ食摂取群との群間比較の差（有意差検定）で評価する必要はあるか。
最終製品を用いた臨床試験（ヒト試験）を科学的根拠とする場合は、特定保健用食品と同様に試験食摂取群とプラセボ食摂取群との群間比較により肯定的な結果が得られる必要がある。 研究レビューを科学的根拠とする場合は、レビューワーが適切に判断することが前提なので、研究レビューに前後比較の論文を含めることは差し支えないが、前後比較での有意差しかみられない論文のみでは、機能性の科学的根拠として不十分であるため注意する必要がある。

機能性表示食品に関する質疑応答集（2017 年 9 月 29 日付け消食表第 463 号）より引用

からないことなどのメリットがある。しかし，心理的要因などのバイアスを強く受けるため，エビデンスレベルは低くなってしまう。機能性表示食品制度への届出を目的とする最終製品を用いたヒト試験ならば，前後比較試験ではなく，RCT を実施するべきである。

参考文献

1. 厚生労働省．臨床試験の一般指針
（2021 年 6 月 28 日アクセス可能：https://www.pmda.go.jp/files/000156372.pdf）
2. Hansson L, Hedner T, Dahlöf B.: Prospective randomized open blinded end-point (PROBE) Study. A novel design for interventional trials. *Blood Pressure*, **1**: 113-9, 1992.（DOI: https://doi.org/10.3109/08037059209077502）
3. 日本製薬工業協会．Small Clinical Trials による薬効評価の考え方
（2021 年 6 月 28 日アクセス可能：http://www.jpma.or.jp/medicine/shinyaku/tiken/allotment/pdf/trials_01.pdf）
4. Nuesch E., Trelle S., *et al*.: Small study effects in meta-analyses of osteoarthritis trials: meta-epidemiological study. *BMJ*, 341: c3515, 2010.
5. 消費者庁．機能性表示食品に関する質疑応答集
（2021 年 6 月 28 日アクセス可能：https://www.caa.go.jp/policies/policy/food_labeling/foods_with_function_claims/assets/foods_with_function_claims_210322_0004.pdf）
6. 多賀谷昭．介護研究における介入効果の検証方法 . 長野県看護大学紀要．**16**: 13-23, 2014.