そこを知らなければ
プロトコルが理解できない
統計の話

（そこしら統計：CRA のための “やさしい統計のはなし”）

2025 年度 ICHE9 対応版

Ver3.0

2025 年 11 月

日本 CRO 協会
モニタリング WG
データサイエンス WG
政策委員会（No More Too Much タスクフォース QbD/CTQ 検討班）

はじめに
〜統計がちょっと苦手なあなたへ〜

CRA （Clinical Research Associate: 臨床開発モニター）として治験に関わっていると、プロトコルにはさまざまな“決まりごと”があることに気づくと思います。無作為化、盲検化、評価時期の設定、さらには「統計解析で有効性を判断する」など――。

これらの仕組みの背景には、すべて統計という考え方があります。統計は、思い込みや偶然に流されず、客観的なデータにもとづいて「この試験薬は効いたのか？」という問いに科学的に答えるためのルールです。つまり、治験とはサイエンスの営みであり、CRA はその一翼を担うサイエンスの仲間なのです。

この“統計のルール”を国際的に整理したもののひとつが ICHE9 です。統計家向けにまとめられたガイドラインですが、CRA にとってもプロトコルの背景を理解する手がかりになります。本書では、その考え方をかみ砕き、CRA がモニタリングで押さえておきたい部分だけを取り上げます。むずかしい数式や理論は扱いません。ここで学ぶのは、“そこを知らなければプロトコルが理解できない統計の話（そこしら統計）”です。

CRA に求められるのは、統計で何ができるのか、統計は何を示すのか、そして何を示さないのか――この“統計の本質”を理解して治験に向き合う姿勢です。「なぜこういう決まりがあるのか？」という視点を持つことで、プロトコルの構造や SOP の意味が見え、逸脱やミスの背景にも気づけるようになります。

「言われたとおりにやる」のではなく、「意味を理解して責任をもって実行する」
――それが、そこしら統計を学んだ CRA≒“そこしら CRA”への第一歩です。

各章では、学んだことを振り返れるように“理解度セルフチェック”を設けました。これは正解を出すことが目的ではなく、あえて突っ込みどころを残し、「考えること」を優先しています。これは、E6（R3）が掲げる「チェックリスト的な GCP から、考えさせる GCP へ」という理念に近づくための仕掛けです。

そして、もし本書で統計に少しでも興味が湧いたなら、ぜひ統計担当者に相談してみてください。彼らは数式という共通言語を使って治験を支えています。その世界をのぞいてみるのも、CRA としての視野を広げるきっかけになるでしょう。

さあ、今日からあなたも“そこしら CRA”への一歩を踏み出してみませんか？

もくじ

そこしら統計　コラムリスト

No.	タイトル	章・節
1	レギュラトリーサイエンスとは？	1.1.
2	そこしらのそこしら流定義	1.3.
3	FIH―なぜ海外ばかり？	2.3.
4	最近増えているフェーズがない臨床試験？	2.4.
5	企業治験はバイアスがかかりやすい？	2.5.
6	でもね第III相試験も比較試験だけではないのだよ	2.5.
7	中間解析とDMCの役割	3.3.
8	PPIと被験者募集 ― 例数設計を現実にするために	4.3.
9	Significant-itisってなに？	4.4.
10	変数って名前の秘密	5.4.
11	QbDとCTQってなんだろう？	5.4.
12	カテゴリ化した変数とCRAにとっての意味	5.4.
13	裁判と仮説検定	6.2.
14	Pivotal試験は1回でよいの？ ― FDAと日本の考え方	6.4.
15	へたな鉄砲も数撃ちゃあたる、はだめ	6.4.
16	E6（R3）で変わったデータマネジメントの役割	7.1.
17	感度分析とEstimand	8.2.

各章の統計的なねらい

第 1 章　治験はサイエンス
治験の枠組みはサイエンスであり、その科学性を支えるのが統計である。単一の感想や観察ではなく、集団比較による推測統計を用いることが「効いている」と言える根拠になることを示す。CRA にとっては「治験は科学的な営み」であることを理解する入口となる。

第 2 章　治験の物語と統計の役割
治験の各フェーズに応じた統計の役割を解説する。第 I 相では薬物動態・薬力学に基づく安全性評価、第IIa 相では信頼区間による効果の兆候把握、第IIb 相では用量－反応関係の検証、第III相では仮説の検証と承認申請への橋渡しを行う。CRA は開発全体のストーリーを意識することで、個々の試験の意味を理解する。

第 3 章　バイアスを避けて、ちゃんと比べる
無作為化による交絡因子の均衡化、盲検化による観察者バイアス排除、標準化や層別化による背景因子調整を解説する。さらに多施設共同治験では施設差を統計的に扱う必要があることを示す。CRA は SOP によるルール統一を確認する役割を担う。中間解析と DMC の役割、早期中止の仕組みについても理解し、「判断には関与しないが、その基盤を支えるのが CRA」という立場を押さえる。

第 4 章　例数設計とばらつきの考え方
必要症例数設計を中心に、第一種過誤（α）と第二種過誤（β）の制御、統計的検出力（Power）の意味を解説する。記録ミスや欠測が不要なばらつきを生むことを示し、CRA に「欠測を最小化することの重要性」を理解させる。

第 5 章　評価変数の種類と重み
個体差と集団差の関係、主要・副次・安全性変数の役割を説明する。CRA は「どのデータが CTQ か」を理解し、優先順位を持ってモニタリングすることが重要である。代替変数、カテゴリ化変数、多重性の問題についても理解し、データの扱い方や統計的背景を押さえる。さらに「すべてのデータは報告される」ことを通じて、主要も副次も安全性もすべて報告対象であるが、No More Too Much の観点で CTQ を優先し、現場を信頼する姿勢が求められることを強調する。

第 6 章　仮説検定の考え方
帰無仮説と対立仮説を背理法で検証する仕組みを解説する。有意差は「偶然では説明しにくい」証拠であるが真実の保証ではないことを強調する。優越性・同等性・非劣性の 3 種類の試験目的を比較し、治験の結論がどの前提に基づいているかを理解する。

第 7 章　SOP とデータの信頼性
ランダムエラーとシステマティックエラーの違い、欠測処理の真正性確保、監査証跡による完全性保証の重要性を示す。CRA には「なぜ逸脱や欠測が怖いのか」を理解させる。同時に、すべての欠測を均等に潰すのではなく、CTQ に関わるものを優先し、現場を信頼するリスクベースドアプローチとしての姿勢を持つことが大切であると説明する。

第 8 章　Estimand の世界へ
評価変数を土台に、対象集団・介入・比較・変数・要約方法を一本の問いとして定義する Estimand の枠組みを説明。Intercurrent Events への対応を事前に規定する意義、感度分析で仮定の頑健性を確認する考え方を解説する。CRA にとっては「なぜ現場の出来事を正しく記録するのか」を理解する章である。脱落例や逸脱を「すべて均等に追う」のではなく、CTQ に直結する部分を守り、現場を信頼することが正しいモニタリングであると理解する。

1. 実は治験ってサイエンスだったのです（今さらですが）

1.1. データを使って「効いた！」ってどうやって示すの？

「先生、この薬、すごく効きました！」

被験者さんがそう言ってくれたら、私たちも嬉しくなりますよね。

でも、治験では、それだけでは「この試験薬は効く」と証明したことにはなりません。実は、かつてはそういった「症例報告」――つまり、ある患者さんにこんな効果がありました、という１例ごとの体験報告だけで試験薬が承認される時代もありました。しかし現在は違います。新しい試験薬の承認は、サイエンスにもとづいた治験によって効果と安全性が科学的に証明されたうえで、行政（厚生労働省や FDA など）が認可するしくみになっています。

では、その「サイエンスにもとづく治験」とは、どのようなものなのでしょうか。たとえば、ある患者さんが試験薬を飲んで元気になったとします。でも、それが本当に試験薬のおかげだったのかどうかは分かりません。自然によくなったのかもしれませんし、別の治療が効いたのかもしれません。一人の患者さんだけの結果では、本当に効いたかどうかは判断できないのです。

そこで治験では、「試験薬を使ったグループ」と「使わなかったグループ（または別の薬を使ったグループ）」を用意して、同じ条件で比べるようにします。この「比べる」ことが、科学的に効果を確かめる第一歩なのです。そして、その違いがたまたま起こっただけなのか、それとも試験薬の効果と考えてよいのかを見きわめるために使うのが統計です。

コラム：レギュラトリーサイエンスとは？

最近、「レギュラトリーサイエンス（Regulatory Science）」という言葉を耳にすることが増えてきました。この考え方は、**私たちが日々取り組んでいる治験の根っこ**にあるものです。レギュラトリーサイエンスとは、「医薬品や医療機器などの安全性と有効性を、科学的に評価して、社会に適切に届けるための科学」のことです。つまり、「薬が効いた気がする」「患者さんが元気になった」――そうした印象や経験だけではなく、**ちゃんとしたデータを集めて、比べて、科学的に判断しよう**という姿勢のことです。

治験は承認申請のためのただの手続きではありません。**レギュラトリーサイエンスの実践の場であり、そこで得られた結果が、薬の承認という社会的な判断につながっていくのです。**CRA として治験に関わるということは、**サイエンスにもとづいて社会に新しい薬を届ける仕事を支えている**ということでもあります。サイエンスである限りは、CRA の役割は手続きを進めることだけではなく、統計が意図している“意味”をきちんと理解し、正しく治験を動かしていくことがより重要です。

1.2. 「ちゃんと比べる」ってどういうこと？

前の節では、「グループ」で「比べる」ことが、サイエンスの第一歩であるという話をしました。でも、ただ人数を並べて比べればよい、というわけではありません。**“ちゃんと”比べる**には、いくつか大切なポイントがあるのです。

たとえば――試験薬を使ったグループに元気な人が多くて、使わなかったグループに重症の人が多かったら、どうなるでしょう？試験薬を使ったほうが成績が良く見えても、それは試験薬の効果ではなく、「もともとの体調の違い」かもしれません。

このように、本来見たい“試験薬の効果”ではなく、**別の要因によって結果に差が出てしまうこと**を、統計の世界では「**バイアス（bias）**」と呼び、バイアスを引き起こす要因のことを交絡因子などと呼びます。**バイアスが入り込むと、公平な比較ができなくなってしまう**のです。ですから治験では、なるべく比べる条件が同じになるように、さまざまな工夫がされています。皆さんがよく知っている治験用語の中に、「**無作**

為化（ランダム化）」や「**盲検化（ブラインド化）**」がありますよね。こうした工夫があるからこそ、「この試験薬が効いた」と自信を持って言うことができるのです。つまり、「効いたかどうか」を見るには、ただ比べるのではなく、“バイアスを避けて、ちゃんと比べる”設計が必要なのです。

そしてこの設計こそが、プロトコルで細かく決められている内容なのです。

治験は、あらゆるバイアスとの闘いでもあります。統計が正しく働くためには、現場で集められるデータが、できるだけ“公平に”記録されていなければなりません。だからこそ、CRA の皆さんにもこの「バイアスを避ける」という視点を持っていただくことがとても大切です。

たとえば、割付手順の逸脱や、盲検破りにつながるような事象があったとき、それを**単なる手順違反として処理するのではなく、「これはバイアスにつながる可能性があるか？」という視点を持って対応すること**が、治験全体の信頼性を守ることにつながります。

そしてもう一つ大切なことがあります。**CRA 自身の行動や介入が、知らず知らずのうちに被験者や医療機関に影響を与え、結果的にバイアスの原因になってしまうこともある**のです。だからこそ、**客観性を保ち、プロトコルや SOP に基づいた中立的な対応を意識することが求められる**のです。

３章でもう少し詳しく説明しますね。

1.3. 「たまたま？それともホント？」見きわめるのが統計なんだよ

これまでの話で、「グループに分けて比べること」や「バイアスを避ける工夫」の大切さが見えてきました。でも、実はもう一つ、とても大事なポイントがあります。

それは「差があったように見えても、それがたまたま起こっただけかもしれない」ということです。たとえば、試験薬を使ったグループでは 30 人中 25 人が良くなった。

試験薬を使わなかったグループでは 30 人中 20 人が良くなった。

このとき、「お、効いているかも！」と感じるかもしれませんよね。でも、こういう差って、偶然でも起こることがあるのです。

そこで登場するのが統計という道具です。

統計はこう問いかけます。

「この差って、たまたま生まれた“偶然の差”なの？それとも、本当に試験薬のおかげなの？」

つまり統計は、「見えた差がホントの差かどうか」を見きわめるための道具なのです。ここでは「そういう道具がある」ということを知っておけば十分です。

詳しい仕組みやルール（どうやって判断しているのか）は、このあともう少し丁寧に説明していきましょう。

コラム：そこしらのそこしら流定義

そこしら統計

そこしら統計とは、「統計の数式を解き明かすこと」ではなく、「治験のプロトコルを理解するうえで最低限“そこを知らなければならない”統計の考え方」を指す。

言い換えれば、CRA が現場で判断するときに「なぜそう決まっているのか？」を理解するための統計の知識を、やさしく整理した学びの道具である。

そこしら CRA

そこしら CRA とは、「SOP をただ守る人」ではなく、「SOP やプロトコルの背後にある意味を理解し、責任をもって行動できる CRA」を指す。

統計の本質を少しでも理解し、治験の問い＝クリニカルクエスチョンをぶらさないように現場を支える姿勢を持つ人。数式を操る必要はないが、「統計の考え方」を現場に生かせる CRA のこと。

第１章理解度セルフチェック（そこしら式 記述テスト）

【問 1】
ある被験者さんが「この薬、すごく効きました！」と言っているのを CRC さんから聞きました。あなたは CRA として、その言葉を聞いてどんなことを考えるでしょうか？「そうか、よかった！」と思う気持ちはとても大切ですが、**そこしら CRA として、クリニカルデータサイエンティストとしての思いを示してください。**

【問 2】
「比較することが大事」と聞くと、多くの人は「数字を並べればいいのでしょ」と思いがちです。でも、そこしら CRA は“ちゃんと比べる”とはどういうことかを知っているはずです。あなたの言葉で説明してみましょう。

【問 3】
あなたが、ある被験者の割り付けが、手順から逸脱していることに気づいたとします。「まあ、ちょっとしたミスだから…」と済ませてしまいそうなこの場面、**そこしら CRA としては、どんな視点を持つべきか**を考えて書いてみてください。

【問 4】
P 値が 0.03 でした。「おお、試験薬が効いた！」と誰かが言いました。でも、そこしら CRA なら、心の中でそっとこうつぶやくはずです。**どんなひとことを心の中でつぶやくでしょうか？**

【問 5】
あなたが「治験における統計の役割って何ですか？」と聞かれたとします。
数字を使わず、日常の言葉で、**CRA の後輩に向けて説明するとしたらどう答えますか？**プロのクリニカルデータサイエンティストとしてちょっとカッコつけても OK です。あなたなりの表現で書いてみてください。

【問 6】
あなたは、治験データのモニタリングをしていて、副次評価項目に小さな欠測を見つけました。主要評価項目には問題はありません。このとき、そこしら CRA としてどんな考え方で対応するのが適切でしょうか？

2. すべてが比較試験というわけではないのだよ

2.1. 医薬品開発の壮大なストーリーを理解しよう

第 1 章では「グループに分けて、条件をそろえて、統計で見きわめる」という基本を確認しました。ここで皆さんにお伝えしておきたいのは、すべての臨床試験が同じ目的で設計されているわけではないということです。

最初は「安全なのか」を少人数で確認し（第 I 相)、次に「効きそうか」「どの用量がよいか」を探り（第 II 相)、そして「本当に効いているのか」を多数例で確かめます（第Ⅲ相)。このように臨床試験は、ひとつひとつの問いに答えながら、科学的に証拠を積み上げていく壮大なストーリーなのです。

この流れの中には、「探索的試験」と「検証的試験」という二つの顔があるのです。

2.2. 探索的試験と検証的試験 ～ヒント探しと答え合わせ～

ここでは、探索的試験（exploratory trial）と検証的試験（confirmatory trial）の違いを見ていきましょう。

2.2.1. 探索的試験 — ヒント探しのステージ

探索的試験は、新しい薬が「どのくらい効きそうか」「どの用量がよさそうか」といった仮説をつくるステージです。

第 II 相試験の多くは探索的な性格を持ち、設計や解析には柔軟性があります。

目的は「次に進めるかどうかの判断材料」を集めること。完璧な答えを出すのではなく、次のステップに進むための**ヒント探し**です。

2.2.2. 検証的試験 — 答え合わせのステージ

一方の検証的試験は、その仮説が正しいかどうかを科学的に確かめるステージです。第Ⅲ相試験の多くがこの性格を持ち、主要評価項目を事前に定め、プロトコルを厳格に守り、統計的に検証します。

ここでは「柔軟さ」よりも「一貫性と厳格さ」が求められます。もしプロトコルから外れる操作や評価が行われれば、積み上げた証拠そのものが揺らいでしまいます。

2.2.3. フェーズごとの位置づけ

この「探索」と「検証」の考え方は、フェーズの区切りとも重なります。

第 I 相：まず安全かどうかを少人数で確認（探索的)。

第 II 相：効果の兆しや用量を探る（探索的)。

第Ⅲ相：本当に効いているのかを多数例で厳密に検証（検証的)。

こうして試験薬は、臨床現場のクリニカルクエスチョンに答えながら、「探索」から「検証」へと歩みを進めます。

CRA にとっての意味

探索的試験ではある程度の柔軟性が許されますが、検証的試験では逸脱や欠測がそのまま「証拠の揺らぎ」につながります。だからこそ CRA は、「この試験は探索なのか、検証なのか」を意識してプロトコルを読み、モニタリングに臨む必要があります。

探索的試験では：

- データの一貫性よりも、幅広く正しく拾えているかを重視。
- 兆しを見逃さないために、背景情報やプロトコル修正の経緯も正確に把握する。
- 柔軟性のある設計を支える「正しい記録の積み上げ」が鍵になる。

検証的試験では：

- 主要評価項目に関するデータの正確性と完全性を最優先。
- プロトコルに定められた手順やタイミングを厳格に守らせることが使命。
- 逸脱や欠測を徹底的に防ぐことが、そのまま「承認を勝ち取る証拠」を守ることにつながる。

探索的試験なら「次の段階に進むための手がかりを正しく拾えているか」、検証的試験なら「承認に耐えうる証拠を守れているか」。この違いを理解してモニタリングに取り組むことが、そこしら CRA としての成長につながります。

2.3. 第Ⅰ相試験：はじめて人に使ってみる段階

新しい試験薬が、はじめて人に使われる。これを**「first in human（ファースト・イン・ヒューマン)」**と呼びます。まさにその名前のとおり、人にとっての“第一歩”となる試験。それが第Ⅰ相試験です。

でも、その前に、ちょっとだけ“裏舞台”の話をしましょう。

試験薬が人に届くまでには、**『非臨床試験（前臨床試験)』**という段階があります。ここでは、動物や細胞を使って「この試験薬、どれくらいの量で効きそう？」「毒性は？」「副作用は出そう？」など、人に使う前に最低限知っておきたいことを徹底的に調べておきます。つまり、**第Ⅰ相試験は「非臨床でのデータをふまえて、いよいよ人で安全性を確かめるステージ」**なのです。

この試験では、**ふつう健康な成人**（例外もあります）に、試験薬を少量から投与して、体の中でのふるまいや副作用を慎重に観察していきます。
このときの観察の中心になるのが**「臨床薬理」**という分野です。

たとえば、「この試験薬、飲んでからどのくらいで血液中に現れる？」「体から出ていくまでにどれくらい時間がかかる？」といった、体の中での動きを見るのが**「薬物動態（PK: Pharmacokinetics）」**。一方、「この試験薬、どのくらいの濃さになると体に変化が出る？」「どのくらいで効果や副作用が出る？」といった“効き目のあらわれ方”を見るのが**「薬力学（PD: Pharmacodynamics）」**です。

ここでの目的は、「この試験薬は“人にとって”どうふるまうのか？」を安全第一で明らかにすること。効果を比べるわけではありませんし、比較群もありません。1 つのグループだけで実施される単群試験が一般的です。

では、統計はここではどんなふうに活躍しているのでしょう？

第 I 相試験では、統計は“差を見つける”というより、“変化に気づく”ための道具として使われます。たとえば、「何ミリグラム投与したら血中濃度がどう変わった？」「副作用は何人中何人に出た？」そんなデータの中から、傾向や安全ラインを見きわめる目を持つのが統計の役割です。

また、少人数の試験だからこそ、小さなサインを見逃さないという感度も大切です。統計は、確定的な判断をするにはまだ早いけれど、「このデータ、ちょっと気になるかも」という“科学的な勘”を支える道具でもあります。

第 I 相試験は、新薬の物語における“第 1 章”。慎重に、でも確実に、安全性という土台を築く——その静かな始まりを支えているのが、統計なのです。

コラム：FIH—なぜ海外ばかり？

新しい薬の旅立ちには、必ず「最初の一歩」があります。それが First in Human（FIH：ファースト・イン・ヒューマン）試験。動物で安全性を確かめたあと、初めて人に投与して「大丈夫か？ 効きそうか？」を探る段階です。

ところがこの最初の舞台、じつは日本ではあまり行われていません。抗がん薬の例で言えば、2007 年から 2017 年までの 10 年間で、FIH 試験は日本で 22 件、アメリカで 261 件。数にして 10 分の 1 以下です。つまり、「主に海外で始まっている」というのが現実なのです。なぜでしょう？

理由はいくつもあります。実施施設や体制の整備が十分ではなかったこと、規制や環境の違いからスピード感で海外に後れを取ったこと…。でも一方で、日本には高い医療水準やまじめなデータ収集体制という強みがあります。

そこでいま、「日本に FIH を呼び戻そう」という動きが出てきています。厚生労働省も、FIH 試験を行える国際水準の施設整備を進め、2029 年の稼働を目標に掲げています。海外企業も日本を開発の拠点に選んでくれるように、GMP 製造や臨床研究インフラをまとめて整える計画が進んでいるのです。

つまり、FIH 試験は単なる「最初の試験」ではありません。「どこで最初の一歩を踏み出すのか」が、その国の医薬品開発力を映す鏡になっているのです。

そこしら CRA にとって大切なのは、「日本で行われる治験の背景には、世界の流れと競争がある」という視点。目の前のプロトコルも、そんな大きな文脈の一部だと気づくと、仕事の意味合いがまた違って見えてくるはずです。

2.4. 第 II 相試験：効果の兆しを探るステップ

第 I 相試験で「人に使っても大丈夫そうだ」ということがわかったら、次は「**効きそうかどうか**」を確かめにいきます。この段階で試験薬は、いよいよ**患者さんに投与される**ことになります。

その最初のステップが**第 II 相試験**です。

でもここでは、「効いた！確定！」とはまだ言えません。そう、第 II 相は“**まだ探索中**”のステージなのです。そしてこのステージは、開発のストーリーにおける**最初の山場**でもあります。

2.4.1. 第Ⅱ相にも「前半」と「後半」がある

第Ⅱ相試験は、試験薬の「効きそうかどうか」を探る大切なステージです。
この段階は、しばしば早期第Ⅱ相（第Ⅱa相）と後期第Ⅱ相（第Ⅱb相）に分けられることがあります。どちらも「この試験薬、効きそうかな？」を確認するという点では共通していますが、目的や設計、統計の使い方が少しずつ異なります。
ただし、この分け方は必ず採用されるわけではありません。疾患の特性や開発戦略によっては、第Ⅱa相と第Ⅱb相を区別せずに実施したり、第Ⅱb相を省略して第Ⅲ相へ進む場合もあります。

コラム：最近増えているフェーズがない臨床試験？

昔は臨床試験といえば、第Ⅰ相・第Ⅱ相・第Ⅲ相と階段を上がるように進むのが常識でした。ところが近年は「シームレス試験」や「プラットフォーム試験」のように、フェーズの区切りがはっきり見えないデザインが増えています。

シームレス試験では、第Ⅱ相と第Ⅲ相を一つにつなげ、途中の結果を見ながらスムーズに次のステージへ移行できます。プラットフォーム試験では、共通の枠組みの中で複数の薬を同時に試し、効かなければ外し、新しい薬を追加することもできます。がん領域やCOVID-19の開発でよく用いられました。

CRAにとって大切なのは「この試験は第何相ですか？」と数えることではなく、「この試験はどんな問いに答えようとしているのか？」を見極めること。フェーズレスの時代だからこそ、治験の本質＝クリニカルクエスチョンを理解する目がますます求められているのです。

2.4.2. 早期第Ⅱ相（第Ⅱa相）

この段階では、まず試験薬を患者さんに投与してみて、「効果の兆し」が見えるかどうかを探ります。多くの場合、1つの用量だけを使い、比較群を置かない単群試験で行われます。たとえば、「この試験薬を投与して30人中10人が改善した。これは偶然ではなさそうか？」といった問いに、統計が答えを出します。

この段階ではP値よりも、反応率や信頼区間といった“効果の大きさ”や“不確かさ”を見て判断することが多く、統計はまだ輪郭のあいまいな“効果の気配”を拾い上げる顕微鏡のような役割を果たします。ここでは、「比較」よりも「気づき」が大切。
効果の方向性を見て、「この試験薬、次のステップに進めるかも」という**感触をデータで裏づける**ことが目的です。

ここで得られた情報は、第Ⅱb相の設計（比較群や用量設定）に直結します。

2.4.3. 後期第Ⅱ相（第Ⅱb相）

ここからは、第Ⅲ相試験に向けた“足固め”のステージです。複数の用量群を設定し、比較対象（プラセボや既存治療）を含む無作為化試験が行われることもあります。

この段階で特に重要なのが「用量－反応関係」の確認です。試験薬の量が増えるにつれて効果がどのように変化するか――その流れを統計的に確かめることは、次の段階に進む判断に欠かせません。

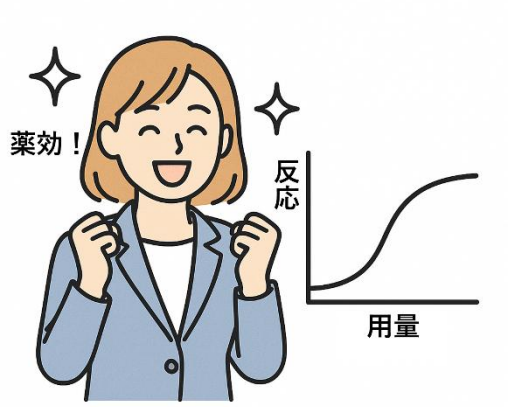
統計は、トレンド検定（trend test） などを用いて、「この効果の増え方は偶然ではなく、本当に用量に沿った変化なのか？」という問いに答えます。通常の検定は「ど

こかの群に差があるか」を見るのに対し、トレンド検定は「用量が増えるにつれて結果も並んでいるか」という一貫した流れを確かめる点が特徴です。

ここで得られた結果は、第Ⅲ相試験の用量選択や症例数設計に直結し、開発全体の方向性を決める根拠となります。

2.4.4. 効いていそうか？それを“かたち”にするのが用量－反応

試験薬には、「使う量（用量）」が増えれば効果も大きくなる――そんな関係がよく見られます。これが 用量－反応関係（dose–response relationship） です。



たとえば、5mg ではわずかな効果しか見られなかったけれど、10mg でははっきり効き、20mg ではさらに改善が大きくなる、といった具合です。この「だんだん効いてくる」という流れがきれいに確認できたとき、研究者は初めて「この物質はヒトにおいて薬効成分として働いている」と確信を持てます。

トレンド検定によって「偶然のばらつきではなく、一貫した上昇傾向である」と示されたとき、臨床開発における大きな“山場”を越えたことになります。ここで得られるのは、ヒトで薬効があることを裏づける科学的証拠です。そしてその成果が、第Ⅲ相比較試験に進むための強い根拠になるのです。

ただし、この段階で「薬効がある」と分かっても、それが患者さんにとって本当に役立つ薬になるかどうかは、まだ確定していません。その最終確認は、大規模な第Ⅲ相比較試験で行われます。

2.4.5. 統計の理解が、判断の分かれ道

このように、第Ⅱ相試験は、医薬品開発のストーリーにおいて、最初の大きな山場です。ここで「効いていそう！」という明確な兆しが見つかれば、次の第Ⅲ相試験へと進めますが、もし効果がはっきりしなければ、**開発中止**という厳しい判断もあり得ます。だからこそ、CRA としてもこのステージの意味と、統計の果たす役割をしっかり理解しておくことが大切なのです。

このステージでは、**“なんとなく効きそう”を“このくらい効きそう”に変換する力**が求められます。統計はその変換のツールであり、**データをストーリーに変える翻訳者**のような存在なのです。

第Ⅱ相試験は、試験薬が「効くかどうか」をはじめて問いかけられる、ちょっと緊張感のある場面です。でも、その問いかけに対して、「こういう反応が見えたから、次

に進んでいいと思うよ」とやさしく背中を押してくれる――
そんな存在が、統計なのです。

2.5. 物語のクライマックス（第Ⅲ相のはなし）

第Ⅰ相で「大丈夫そう」、第Ⅱ相で「効きそう！」という感触がつかめたら――いよいよ物語はクライマックス、第Ⅲ相試験です。ここでの目的はただひとつ。**「この薬は本当に効く」と科学的に胸を張って言える証拠をそろえること。**この段階では、患者さんの人数もぐっと増えます。

なぜなら、第Ⅱ相で見えた効果が“たまたま”ではないと確かめるためには、多くの患者さんを集めて、条件をそろえて、きちんと比べる必要があるからです。
第Ⅲ相は、開発の最終試験として承認申請の基盤になるステージ。言わば「判定試合」です。

第Ⅲ相試験は、試験薬とプラセボ、あるいは既存の治療薬を**同じ条件で比べる**ことが基本です。条件をそろえるのは、公平な比較をするため。そして、この比較は１つや２つの病院だけではなく、**複数の国や地域、数多くの施設が協力**して行われることが一般的です。なぜなら、より多くの患者さんに参加してもらうことで、結果の信頼性と一般化可能性が高まるからです。

さらに、この試験では**無作為化（ランダム化）と盲検化**が徹底されます。ここまで学んできた皆さんはこの意味がよくわかりますよね。これによって、治験担当者や患者さんの先入観が結果に影響を与える「バイアス」を極力減らすことができます。そして、試験の「勝敗」を決めるために、あらかじめ主要評価項目（Primary Endpoint）が１つ（または少数）に絞って設定されます。これがブレてしまうと、試験結果そのものの意味が揺らいでしまうため、最初にしっかりと決めておくことが重要なのです。まさに、バイアスとの戦いですね。この話も、３章でもう少し詳しくお話ししましょう。

第Ⅲ相試験は、このように大規模かつ厳密な設計で、「本当に効くのか？」という問いに対する**最終的な答え**を出すための舞台なのです。

コラム：企業治験はバイアスがかかりやすい？

これは完全な偏見や陰謀論ではなく、科学的・制度的な懸念として存在する実際の議論です。以下のような理由からです。

1. 選ばれた患者だけが対象になっている

治験では、組み入れ・除外基準が厳しく設定されており、高齢者、多疾患患者、社会的弱者などが除外される傾向があります。結果として、現実の臨床現場とは異なる「理想的な患者集団」での結果になるため、一般化可能性（external validity）が低くなる傾向があります。

2. 製薬企業が主導する「商業的インセンティブ」への懸念

開発中の薬剤を承認に導くための試験は、製薬企業にとって成功させたいプロジェクトです。一部では、「ネガティブな結果が発表されない（publication bias）」「比較対象が都合の良いものに設定される」などの設計上のバイアスがかかる可能性があると指摘されてきました。

こうした懸念を背景に、試験登録（ClinicalTrials.gov や jRCT、UMIN-CTR など）や結果報告の義務化が進められました。また、第三者評価（独立委員会）や事前プロトコル公開などの透明性確保も強化されています。

RWD（Real World Data）による検証の動き

まさにこの背景から、「治験だけでは不十分ではないか？」という問題意識が生まれ、RWD による検証や補完の重要性が世界的に認識されるようになっています。

FDA は 2018 年に Real World Evidence（RWE）プログラムを発表し、RWD の利用が承認審査や安全性評価に使える条件を明確化しようとしています。また、治験の限界を補う手段として、RWD を用いた外部対照群（external control arm）や長期追跡の補助的使用などが模索されています。

欧州医薬品庁（EMA）も同様に RWE の活用戦略を公開していますし、日本でも PMDA が RWD の品質評価や適用可能性に関する議論を進めています。

このように「企業主導の治験にバイアスがかかる」という指摘は、現代のレギュラトリーサイエンスが直視している現実の一部であり、その反省から「RWD を使って治験の結果を補完・検証しよう」という動きは、まさに国際的に進展している潮流です。治験と RWD は対立構造ではなく、互いに補い合うものとして制度設計が進んでいると理解するのが適切です。

2.5.1. CRA から見た第Ⅲ相の意味

第Ⅲ相は、開発のゴール目前――まさにクライマックスの試験です。
ここまで積み重ねてきた安全性と効果のデータを、最終的に「本当に効くのか？」という問いに対して証拠として示す舞台。その分、ひとつのミスや逸脱が試験全体の信頼性を揺るがすリスクも大きくなります。

CRA は、この大切な局面で「本当に公平な比較ができているか」を守る最後の砦です。比較試験であれば割付や盲検が崩れていないか、条件が揃っているかを現場で確認します。非比較試験であっても、「計画どおりの条件で、安全性や有効性の情報が集まっているか」を一貫して監視します。現場での小さな逸脱や記録の乱れが、最終的な結論を左右することもあるからです。

もっとも、集められたすべての安全性・有効性データが、そのまま試験や薬剤プロファイルの「結論」を左右するわけではありません。重要なのは、プロトコルで定められた主要評価項目を中心に、結論に直結するデータを確実に守ることです。その一方で、副次的な情報も薬剤の全体像を理解するためには欠かせない位置づけを持っています。

そして、集められたデータは最終的に統計解析されます。その結果に映し出されるのは、数字だけではありません。CRA が現場で守り抜いた“正しさ”そのものが、統計の枠組みの中で評価されるのです。

統計はこのステージで、まさに“判定員”として働きます。差が偶然でないことを確率（P 値や有意差）で示し、効果の大きさを数字（リスク比、差の大きさ、信頼区間）で示します。安全性についても、発現率や重症度を比較して評価します。比較試験なら効果の有無を明確にし、非比較試験なら安全性や有効性の信頼区間や発現率を使って「これだけの規模と期間で見ても問題がない」と判断します。

コラム：でもね第Ⅲ相試験も比較試験だけではないのだよ

第Ⅲ相試験と聞くと「大規模な比較試験」を思い浮かべますが、すべてが比較試験というわけではありません。確かに新薬とプラセボや既存治療を比べる試験が中心ですが、他にもいくつかの形があります。

比較試験以外の例

- 長期安全性試験：承認申請に必要な長期の安全性データを集める試験（比較群を置かず、単群で行うことが多い）
- 拡大治験（Expanded Access）／人道的使用：重篤な患者さんに早期に薬を提供しつつ、有効性・安全性の情報を集める
- 用量確認・集約試験：第Ⅱ相や途中解析の結果を受けて、最終的な用量を確定するための補足的試験

こうした非比較試験も、申請時の重要なデータとして扱われます。

だからこそ、CRA は統計の原則を理解しておく必要があります。統計の役割や数字の意味がわかれば、「なぜこの場面で厳密な条件を守らなければならないのか」「なぜこの記録が重要なのか」が自分の中で腑に落ちます。そしてその理解は、現場での判断力となり、試験の信頼性を守る力になります。

第Ⅲ相試験は、科学と努力と信頼の物語の最終章。統計がその物語の結末を書き上げるために、CRA は現場から正しいデータを届ける――それこそが、このステージでの最大の使命なのです。

2.5.2. 第Ⅲ相は物語の“答え合わせ”

第Ⅰ相と第Ⅱ相は「この試験薬、きっといける！」という期待を積み上げる過程。第Ⅲ相は、その期待が真実かどうかを全員で確かめる“答え合わせ”です。

比較試験でも非比較試験でも、ここで得られるデータが**試験薬の運命を左右する最終証拠**になります。ここで証拠がそろえば、いよいよ新薬として世に出て、次の舞台――市販後のリアルワールドへと物語が進みます。

第Ⅲ相試験は、新薬開発ストーリーのクライマックス。CRA にとっては、これまでの努力が実を結び、データの形で世に問われる瞬間でもあります。だからこそ、この試験の意義と、統計の役割を胸に刻んで、最後まで“科学の物語”を守り抜くことが大切なのです。

第 2 章理解度セルフチェック（そこしら式 記述テスト）

【問 1】
第 II 相試験は「探索的なステージ」と説明されました。もしあなたが CRA として第 II 相試験をモニタリングするときに心がけるべきことは何でしょうか？「効きそう！」という感触をどう支えるのか、自分の言葉で書いてください。

【問 2】
第 IIa 相試験と第 IIb 相試験では、目的や設計に違いがあると説明されました。それぞれの特徴を CRA の立場からどう理解しておくとよいか、自分の考えをまとめてみましょう。

【問 3】
用量－反応関係（dose–response relationship）は、第 II 相後半で重要な視点でした。もしプロトコルに「複数用量群での比較」と書かれていたら、CRA としてどんな点をモニタリングで特に確認すべきでしょうか？

【問 4】
第 2 章では「統計は“効きそう”を“こう効きそう”に変える翻訳者」とたとえられました。あなたなりの比喩を使って、「第 II 相試験における統計の役割」を表現してみてください。

【問 5】
もし後輩 CRA から「第 II 相試験って、まだ確定じゃないのにどうして重要なんですか？」と質問されたら、あなたはどう答えますか？短い説明で、しかし CRA としての責任感を込めて答えてみてください。

【問 6 】
探索は“ヒント探し”、検証は“答え合わせ”と表現されます。
あなた自身の言葉で、「モニタリングのアプローチがどう変わるか」を説明してみましょう。

3. プロトコルのウラにはワケ（理由）がある

第 1 章では、「治験はサイエンスだ」ということを確認しました。その中で大切なキーワードとして出てきたのが「ちゃんと調べる」でしたね。つまりただ「効いた！」という感想や体験談に頼るのではなく、試験薬を使ったグループと使わなかったグループを公平に比べ、その差が本物かどうかを統計で見きわめる。それが「ちゃんと調べる」ということでした。

そして第 2 章では、治験には比較試験以外の形もあることを少し寄り道しました。ここからは再び本題に戻り、この「公平に比べる」という仕組みを、プロトコルの視点からもう一歩掘り下げてみましょう。

CRA としてプロトコルを読むと、必ず出てくる言葉があります――「無作為化（ランダム化）」と「盲検化（ブラインド化）」。正直、「どうしてこんなにややこしいことをするのだろう？」と思ったことはありませんか？実はこの“面倒に見える工夫”こそが、治験をサイエンスにしている仕掛けなのです。

3.1. もう一度、「ちゃんと比べる」ってどういうこと？

第 1 章でお話ししたように、治験で大切なのは「ちゃんと調べる」こと。その中心になるのが「ちゃんと比べる」という考え方です。これは、治験に携わるすべてのひとにとって、もしかしたら一番重要なことかもしれませんね。でも、ここで注意が必要です。ただ人数を並べて比べればいい、というものではないのです。

え、比べるだけじゃダメなの？

たとえば、試験薬を使ったグループに軽症の患者さんばかりが集まって、対照群に重症の患者さんが多かったらどうなるでしょう？結果は試験薬のほうが良く見えるかもしれません。でも、それは薬の力ではなく軽症の患者さんばかりということで「もともとの体調の差」かもしれないのです。

こうした“公平でない差”を、統計の世界では「バイアス」と呼びます。治験はこのバイアスとの闘いでもあるのです。

3.1.1. バイアスを避けるための工夫

では、どうやって「公平に比べる」のでしょうか？実は、プロトコルに盛り込まれている仕組みはすべて、バイアスを防ぐための工夫なのです。その仕組みは、主に無作為化、盲検化、標準化、層別化の 4 つが挙げられます。

3.1.2.　無作為化（ランダム化）

治験をサイエンスとして成り立たせるうえで、最も重要な仕組み――それが**無作為化**です。

無作為化とは、患者さんを**くじ引きのように偶然で**グループに分ける方法です。もし「若い人はこちら」「重症の人はあちら」と人為的に分けてしまえば、結果に大きな偏りが入り込みます。そうなれば、試験薬が効いたのか、それとも「もともとの体調の差」なのか、区別がつかなくなってしまいます。

無作為化をすることで、年齢や病気の重さといった背景因子が両方のグループに散らばりやすくなり、**公平な比較に近づける**ことができます。　もちろん完全に均等になるわけではありません。**しかし、統計の考え方を使えば、ある程度の人数を集めれば、偶然の片寄りは小さくなり、統計的に“公平な比較”に近づけることができます。この“統計的に”という感覚が肝ですよね。**

だからこそ無作為化は、治験をただの経験談からサイエンスへと引き上げた仕組みなのです。統計は、この「公平に分けられたデータ」があるからこそ力を発揮できます。もし最初から偏ったデータしか集まっていなければ、統計はいくらがんばっても真実を示すことはできません。

無作為化は、治験を科学的に行うための“入口”であり、統計が正しく働くための“土台”なのです。

3.1.3.　盲検化（ブラインド化）

治験をサイエンスとして成り立たせるうえで、まず土台となるのは無作為化です。でも実は、それだけではまだ不十分なのです。せっかく患者さんをくじ引きのように分けても、「この人は新薬だ」「あの人はプラセボだ」と分かってしまったらどうでしょう？そこに人の思い込みや期待が入り込み、結果をゆがめてしまいます。

たとえば、患者さん自身は「新薬を飲んでいる」と思い込むと、症状が良くなった気がしてしまう（プラセボ効果）。医師も「プラセボ群だから副作用は出にくいだろう」と考えてしまえば、観察の厳しさに差が出てしまう。

これらはすべて立派なバイアスです。つまり、無作為化しただけでは、まだ“公平な比較”は守りきれないのです。そこで登場するのが盲検化。

誰が試験薬で、誰がプラセボなのか――それを患者さんにも医師にも分からないようにする仕組みです。これを二重盲検（ダブルブラインド）と呼びます。無作為化で「比較の土台」を作り、盲検化で「その土台を守る」。この二つがセットになって、はじめて治験はサイエンスとして成立します。

無作為化と盲検化は、治験を科学的に成り立たせる二本柱。
盲検化は、無作為化の力を「本物のサイエンス」に変えるための必須の仕組みなのです。

3.1.4. 標準化（プロトコルの統一ルール）

もう一つ大切な工夫が標準化です。
これは「症状をいつ、どうやって評価するか」をあらかじめ決めておく仕組みのことです。なぜこれが重要なのでしょうか？もし施設や医師ごとにやり方が違っていたら、試験薬の効果ではなく評価の仕方の違いが結果に出てしまうからです。
たとえば、ある病院では「痛みが軽くなったか」を 0〜10 のスケールで数値化して聞いている。別の病院では「先生、痛いですか？」と口頭で聞いて、その印象でカルテに記録している。同じ試験薬を使っても、この違いだけで「効いたように見える」「効いていないように見える」という差が生まれてしまいます。
また、観察のタイミングも大事です。ある施設では投与後 1 週間で評価し、別の施設では 4 週間後に評価していたら、症状の変化の出方がまったく違うタイミングで測られてしまうことになります。
こうした“評価のバラバラさ”を防ぐために、プロトコルで「いつ・どう評価するか」を厳密に定めるのです。だからこそ、プロトコルには「評価日は〇週目」「評価方法はこのスコア」「判定はこう記録」といった細かいルールが書き込まれています。なので、CRA にとっては、施設を訪問したときに「この評価がプロトコルどおりに行われているか」を確認することが、とても大切な役割になります。

そして最新の考え方を少し補足しておきましょう。
もちろん、プロトコルの規定を厳密に守ることが第一に重要です。ただし現実には、どうしても逸脱が起こることがあります。その逸脱には大きく二つのタイプがあります。たとえば、単なる手順上の問題（例：評価日を間違えた、記録漏れがあった)。もうひとつは、治験薬そのものが原因で起こるもの（例：副作用で投与を中止、効果が見られず別の薬に切り替え)。
最新の考え方では、これら両方を含めて、「逸脱があった場合に、それが最終的な評価にどう影響するのか」**をあらかじめ考えておくことが求められています。この発想を** Estimand（エスティマンド）と呼んでいます。ちょっと難しい話になりましたが、ここで覚えてほしいのは一つ。**“決められた通りにやること”と、“もし外れてしまったらどう評価するかを考えること”の両方が大切**——これが、いまの臨床試験の最新の姿勢なのです。Estimand に関しては第 8 章でもう少し詳しく説明しましょう。

3.1.5.　層別化（ストラティフィケーション）

層別化とは年齢や病期など、結果に大きな影響を与える要因ごとに分けて、その中で無作為化する仕組みです。これは「年齢」や「病期」など、結果に大きな影響を与えると分かっている要因ごとに分けて、その中で無作為化を行う仕組みです。

「え？ 無作為化していれば十分じゃないの？」と思うかもしれません。

たしかに無作為化は強力な方法で、基本的には公平さを守ってくれます。でも、少人数の試験や偶然の片寄りでは「大事な条件」が偏ってしまうことがあるのです。

たとえば、試験薬群に若い患者さんばかりが集まり、対照群に高齢の患者さんが多かったらどうでしょう？薬の効果ではなく「年齢の差」が結果に見えてしまいます。

そこで層別化。あらかじめ「年齢層ごと」「病期ごと」に分けて、その中でランダムに割り付けを行えば、重要な因子が片寄らないようにできます。これが層別割り付け（stratified randomization）です。

さらに解析の段階でも、層ごとに効果を見直す層別解析をあらかじめ計画に含めることで、「この試験薬は特にどんな患者さんに効きやすいのか？」といった追加の洞察も得られます。

ところで、「え？ 無作為化していれば十分じゃないの？」と思ったあなた。そういう疑問を持てるあなたは、もしかしたら統計家に向いているかもしれません。
統計家の中には「完全無作為化こそ最も科学的であり、層別化は余計な操作だ」と考える人も少なからずいるのです。「たとえ多少の片寄りが起こっても、統計解析で調整すればよい」という立場であり、それよりも無作為性が損なわれることの方が科学性からの逸脱であるという考え方です

一方で、「重要な因子が大きく偏ったら取り返しがつかないのだから、最初から層別化で均衡を保つべきだ」というより現実的な考え方もあります。つまり、層別化は“正義の必殺技”というより、場合によって使い分ける工夫なのです。こんな議論があることを知っていると、「なぜこの試験で層別化をしているのか？」とプロトコルを読むときの視点がひとつ増えるはずです。

CRA にとっての意味

こうした仕組みがあるからこそ、治験で得られた結果に「この試験薬が効いた」と胸を張って言えるのです。そして CRA は、この工夫がきちんと守られているかを現場で確かめる役割を担っています。モニタリングですよね。

だからこそ、「なぜこの手順があるのか？」を理解しておくことがとても大切です。手順を守ることは単なる作業ではなく、科学の信頼性を支える行為なのです。プロの CRA はこうじゃなくてはね。

プロトコルには、割付の手順、ラベルの扱い、ブラインド破りを防ぐ記録管理など、細かいルールが並んでいます。これらはすべて「バイアスを避けて公平に比べる」ための工夫です。

だからこそ CRA が「面倒だな」と感じるルールこそ、「あ、これはバイアスを防ぐためなんだ」と理解してほしいのです。

そしてもう一つ重要なのは、CRA 自身がバイアスの原因になり得るということです。不用意な発言や態度が、被験者や医療機関に影響を与え、結果をゆがめてしまうこともあります。だからこそ、CRA には中立を保ち、プロトコルに忠実である姿勢が求められます。

ただし、その「忠実さ」が形式的な確認や報告の繰り返しに偏りすぎると、現場本来の判断や医療行為の自然な流れを損ね、結果として新たなバイアスを生むことにもなりかねません。

「プロトコルのウラにはワケがある」とは、統計が科学として正しく働くための仕掛けが埋め込まれているということです。

ルールを守ることはもちろん重要ですが、CtQ（Critical to Quality）に関わる要素を見極め、守るべき点と柔軟に対応すべき点を区別することこそ、**No More Too Much** の精神に沿った GCP 実践といえるでしょう。

次の章では、「なんでこんなに被験者を集めなければならないの？」という疑問を解いていきましょう。

3.2. 多施設共同治験を“同じ楽譜”で演奏するということ

多施設共同治験のねらいは、より多様な患者さんで同じ治療を試し、結果を一般化できるようにすることです。

しかし施設が増えるほど運用に差が出やすくなり、同じ楽譜（プロトコルや SOP）を同じテンポで弾けるかどうかが勝負どころになります。

統計家は施設差を解析で扱うことができます。施設効果をモデルに入れたり、層別化や無作為化を取り入れたり、施設ごとの一貫性を確認したりすることも可能です。ただしこれには限界もあり、施設効果を解析に入れれば例数効率が下がったり、結果の解釈が複雑になったりします。そもそも施設ごとの手順や測定条件の不統一は、解析で完全に補正できるものではありません。

そして大前提として、施設差の扱い方はプロトコルや統計解析計画にあらかじめ定義されていなければなりません。結果を見てから「施設差が大きいので調整しよう」とするのは、バイアスの温床になります。主要解析の道筋は、最初に決めて最後まで守ることが原則です。

では、施設差を本当に防ぐにはどうすればよいのでしょうか。
それは個々の CRA がその場で調整することではなく、あらかじめ SOP やプロトコルに統一ルールを明文化し、全施設に同じやり方を求めることです。血圧測定なら体位や機器、測定時刻、再測定のルールまで、そして有害事象の定義や重症度の基準、データ入力の仕様や欠測の扱いまで、あらかじめ決められた設計図に従って進めることが大切です。

そのうえで、CRA に求められるのは、現場でそのルールがきちんと守られているかを確かめる目を持つことです。施設ごとに測定条件や逸脱の扱いにズレが出ていないかに気づく感覚も必要です。ただし、カテゴリ化の境界値のような統計的な後処理を意識しすぎて現場に影響を与えるのは逆効果です。境界を特別扱いするのではなく、すべての症例で同じルールを守っているかどうかを確認することが重要です。
そして、問題に気づいたときに自分だけでその場しのぎの修正をするのではなく、逸脱の根本原因を SOP や手順にフィードバックし、全体で改善する姿勢が求められます。

多施設共同治験のデータを守るのは、統計的な調整よりも、現場でのルールの統一と遵守です。統計家は施設差を解析で扱えますが、品質の穴は埋められません。施設の 24-25 を本気で防ぐのは、SOP やプロトコルという“同じ楽譜”です。CRA の役割は、全施設が同じ楽譜を同じテンポで演奏できているかを確かめ、ズレを早く見つけ、仕組みに戻して直すことです。

ここで忘れてはならないのは、「同じ楽譜を弾く」といっても、すべての施設で完全に同一の環境を再現することが目的ではないということです。
目指すのは、**同じメロディを奏でるための“柔らかな統一”**です。一定の統一性を保ちながらも、臨床現場の合理的な判断を尊重し、試験全体の信頼性を保つために必要十分な一貫性を確保することが理想です。
このような柔軟な統一のあり方は、国際的に求められる **Oversight の考え方（8.4 参照）**にも合致します。

3.3. 試験の目的はいろいろある

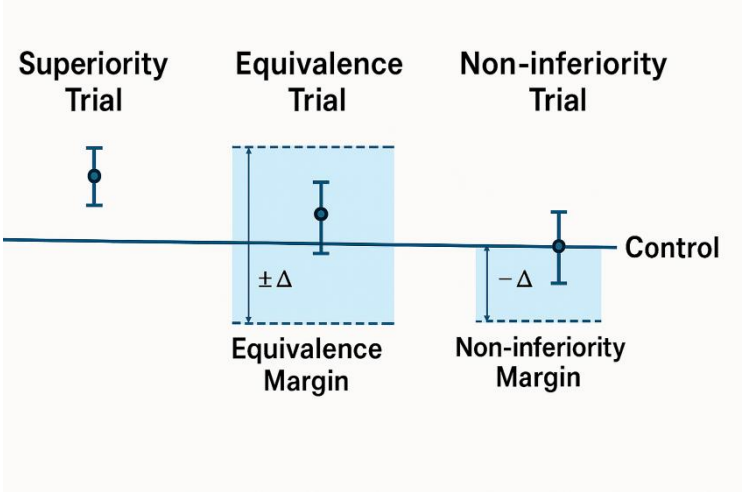
治験のプロトコルには「この試験で何を証明したいのか」という目的が必ず書かれています。ここでの目的によって、試験の設計や解析の考え方、そしてモニタリングで CRA が注目すべき点まで変わってきます。代表的なのが「優越性試験」「同等性試験」「非劣性試験」の三つです。

3.3.1. 優越性試験

問いは「新しい薬は、プラセボや既存薬よりはっきり良いか」です。統計家は、偶然では説明できない差があるかどうかを、信頼区間や 25 を使って判定します。CRA にとって大事なのは、この差を小さく見せてしまう要因を現場でできるだけ減らすことです。服薬を守れなかったり、救済治療を乱用したり、評価のタイミングを外したり、欠測が増えたりすれば、薬の効果が本当にあっても差が薄まってしまいます。主要評価のタイミングや盲検の維持、逸脱や欠測を最小限に抑えることが、優越性を支えるモニタリングの核心です。

3.3.2. 同等性試験

問いは「新しい薬と既存薬の効果の差は、臨床的に意味のない範囲に収まっているか」です。ここでは、差が大きすぎても小さすぎてもだめで、両側を同時にチェックします。CRA にとって重要なのは、データのばらつきを少しでも小さくすることです。測定の方法や条件が施設ごとに違えば、そのぶん差が大きく見えたり小さく見えたりします。測定条件を揃え、教育や再トレーニングで再現性を確保し、欠測を抑えることが同等性を支えるポイントです。もしデータにブレが多ければ、「同等である」と胸を張って言えなくなります。



3.3.3. 非劣性試験

問いは「新しい薬は既存薬より許容できないほど悪くはないか」です。非劣性試験では、「この差までなら許せる」という非劣性マージンをあらかじめ決め、その範囲内に収まっているかを示します。ここで怖いのは、差が小さく見える方向へのバイアスです。たとえば救済治療を多く使ったり、服薬が守られなかったり、評価が甘くなったりすると、効果の差が消えてしまいます。そうなると「負けていない」と誤って結論づけられる危険があるのです。だからこそ非劣性試験では、アドヒアランスや併用薬の管理、救済治療の扱いをきちんと記録し、手順を厳格に守ることが欠かせません。また、非劣性では FAS や ITT だけでなく、PPS でも同じ結論が得られるかを確認するのが定石です。さらに、比較対象の薬がきちんと効いていること（アッセイ・センシティビティ）を確保することも不可欠です。

参考（設定例）

優越性試験

主要評価項目について、試験薬の優越性を対照薬と比較して検証した。
検定仮説は以下のとおりである。
帰無仮説 H_0：$\mu_T - \mu_C = 0$（両群に差がない）
対立仮説 H_1：$\mu_T - \mu_C \neq 0$（両群に差がある）
有意水準は両側5%とし、標準偏差10、平均差5を検出する検出力80%で例数を算出した。

同等性試験

試験薬と対照薬の同等性を以下の仮説に基づいて検証した。
帰無仮説 H_0：$|\mu_T - \mu_C| \geq \Delta$（差が$\Delta$以上であり、同等でない）
対立仮説 H_1：$|\mu_T - \mu_C| < \Delta$（差がΔ未満であり、同等である）
同等性マージン Δ は 10%とし、両側5%に相当する片側2.5%の有意水準で両側検定を実施した。

非劣勢試験

試験薬が対照薬に対して非劣性であることを次の仮説に基づいて検証した。
帰無仮説 H_0：$\mu_T - \mu_C \leq -\Delta$（試験薬は対照薬より$\Delta$以上劣る）
対立仮説 H_1：$\mu_T - \mu_C > -\Delta$（試験薬は対照薬に対してΔ未満の劣性、すなわち非劣性である）
非劣性マージン Δ は10%とし、片側有意水準2.5%で検定を行った。検出力は90%とした。

コラム：中間解析と DMC の役割

治験の途中で集まったデータを確認し、安全性や有効性を評価することがあります。これを「中間解析」と呼びます。便利な仕組みですが、途中で結果をのぞき見すれば「偶然の差」を本当の効果だと誤解してしまう危険もあります。

そのため、中間解析を行うかどうか、いつ、どのように、誰が行うのかは、必ずプロトコルや統計解析計画に事前に定めておく必要があります。

このとき中心的な役割を担うのが「データモニタリング委員会（DMC)」です。DMC は独立した立場でデータを評価し、必要に応じて試験を続けるか、修正するか、中止するかをスポンサーに助言します。

つまり治験の進退を決めるのは DMC やスポンサーであって、個々の CRA ではありません。むしろ CRA や治験責任医師が中間結果を知ってしまうと、盲検が破れ、バイアスの原因となるため、結果には一切触れてはいけないのです。

治験が途中で中止される場合には、いくつかの典型的な理由があります。

ひとつは「有効性が圧倒的に明らかになった」ケースです。この場合、これ以上プラセボ群や既存治療群に患者を割り当てるのは倫理的に適切でないと判断されます。

もうひとつは「安全性に重大な懸念が生じた」場合です。深刻な有害事象が多発していれば、被験者を守るために試験を中止せざるを得ません。

そして三つ目は「有効性を示せる見込みが極めて低い（futility)」と判断されたときです。続けても結論に至らないなら、不要な負担を避けるために早期終了となります。

中止の 3 つの理由
有効性が圧倒的に明らかになった
安全性に重大な懸念が生じた
有効性を示せる見込みが極めて低い（futility）

統計的な観点から見ると、中間解析を繰り返すほど「のぞき見」の回数が増え、そのたびに偶然を有意と取り違える危険（タイプⅠエラー）が高まります。そこで、有意水準を少しずつ分割して使う方法（α-spending など）をあらかじめ決めておき、全体のエラー率を制御する工夫が必要になります。

こうした統計的な配慮も、プロトコルや統計解析計画に明記しておくことが望まれます。

では CRA には関係ない話かというと、そうではありません。中間解析に使われるデータは、日々のモニタリングを通じて現場から集められたものです。欠測や逸脱が多ければ、DMC の判断そのものが揺らいでしまいます。

また、中間解析は「ある症例数がそろった時点で実施」と定められていることが多いため、データ入力や症例登録の進捗を管理し、予定どおりデータがそろうように支えるのも CRA の重要な役割です。

さらに、現場では被験者の安全に直結する判断を迫られることもあります。定められた手順よりも臨機応変な対応を優先すべき場面では、その逸脱の背景を正確に記録し、科学的妥当性を維持する工夫（バイアス評価や補足説明など）を行う必要があります。

つまり、CRA は中間解析や DMC の「判断」には関与しませんが、その判断を支えるデータ品質の確保という基盤づくりに深く関わっています。

プロトコルを理解するうえでも、「なぜ自分が結果を知らされないのか」「誰が途中で試験を止める判断をするのか」を知っておくと、盲検維持やデータ品質確保の意味がより明確に見えてきます。

CRA にとっての共通の視点

三つの試験に共通して大事なのは、主要評価項目や解析集団の扱い、マージンや救済治療のルール、多重性の管理などをプロトコルや統計解析計画にあらかじめ定義しておくことです。結果を見てから条件を変えるのは恣意性の入り口になります。CRA の役割は、現場でそのルールが守られているかを確かめることです。主要評価を死守し、測定の条件を揃え、救済治療や併用薬の透明性を確保し、欠測をできるだけ減らす。そして施設ごとのズレをその場で抱え込むのではなく、仕組みに戻して改善する。こうした姿勢が結論の信頼性を支えます。

まとめると、優越性は差を薄めないこと、同等性はブレを増やさないこと、非劣性は差を消しやすい要因を封じることが大切です。そして三つすべてに共通して、事前定義を守り、同じルールを同じ精度でそろえてデータを集めることが、CRA に求められる視点です。ワンフレーズで言えば「優越性は勝ちを取りに行く、同等性はほぼ同じ、非劣性は負けていない——結論を守るのは、現場で差とブレを生まないあなたの運用」です。

3.4. 選択・除外基準を守る意味

治験に「誰を参加させるか」は、実はとても大切なポイントです。そのルールを決めているのが「選択基準」と「除外基準」です。選択基準は「この条件を満たしていれば参加できる」という入口の条件で、除外基準は「この条件に当てはまる人は参加してはいけない」という安全や科学性のためのストッパーです。つまり、選択・除外基準は治験という物語の“登場人物”を決める大前提なのです。

なぜここが重要かというと、治験の結論は「この基準を満たした患者さんに対して、この薬はどう効いたか」を示すものだからです。もし基準があいまいであり、現場でバラバラに解釈されてしまえば、どんな集団を対象にした結論なのかがわからなくなります。統計は「定義された集団から得られたデータ」を前提に推測を行います。入口がゆるければ、推測の対象がぶれてしまい、結果の一般化も危うくなるのです。

さらに注意が必要なのは、治験の途中で基準が変更される場合です。多くは新しい安全性情報が得られたときや、倫理的に修正が必要なときに起こりますが、これは対象集団の性質を変えてしまう可能性があります。前半は「重症例が多い集団」、後半は「軽症例が中心」となれば、結果をまとめて解釈するのは難しくなります。統計的には「集団の非一貫性」が生じ、解析では層別や感度分析が必要になるかもしれません。だから E9 でも「基準の変更はリスクがある」と明言されており、やむを得ない場合にはその理由を明確にし、プロトコルを改訂し、解析計画に反映させることが求められています。

CRA にとって大事なのは、この基準を「勝手に柔軟に解釈してはいけない」ということです。患者さんが「条件に近いからいいのでは」と思えても、決められた基準を守るのが鉄則です。そしてモニタリングでは、**組入れられた患者が基準を満たしてい**

たかどうかを必ず確認することが必要です。カルテや検査値を見て、組入れ判断の根拠がプロトコルに書かれた基準と一致しているかをチェックします。もし不一致があれば、それは重大な逸脱であり、後の解析に大きな影響を与えます。

また、基準が途中で変更された場合には、その改訂版プロトコルが IRB (Institutional Review Board、倫理審査委員会)や IEC (Independent Ethics Committee, 独立倫理委員会)で承認され、各施設に配布されているかを確認しなければなりません。医師や CRC が新しい基準を正しく理解しているか、過去の症例に誤って遡及的に適用していないかも点検します。つまり、CRA は「基準の変更を判断する立場」ではなく、「基準の変更が正しい手続きで行われ、現場で正しく運用されているか」をモニタリングする立場なのです。

統計的にみれば、選択・除外基準を守ることは **集団の同質性** を確保することです。入口が揃っていれば、得られた結果は「この集団ではこう効いた」と一貫して説明できます。逆に入口が揃っていなければ、試験結果は「結局どんな人に効いたのか」が曖昧になり、科学的な信頼性が崩れてしまいます。

選択・除外基準は、治験の入口を守るゲートです。このゲートがゆるければ、どんなに立派な統計解析をしても結果は意味を持ちません。逆に、このゲートをきちんと守れば、統計的にも臨床的にも試験結果の解釈に自信を持つことができます。CRA が日々モニタリングで確認している一つひとつの適格性チェックは、治験全体の科学的信頼性を支えているのです。

「選択・除外基準は治験の入口。入口が揃ってこそ統計は力を発揮する。守らせるのは CRA の大切な役割」。

第 3 章理解度セルフチェック（そこしら式 記述テスト）

問 1
ある試験で、割付の手順に小さな逸脱がありました。
そこしら CRA として、なぜ「小さいからいいや」とは考えずに、きちんと対応しなければならないのでしょうか？
統計の視点を交えて答えてください。

問 2
盲検化は「めんどうな仕組み」に見えることがあります。
しかし、そこしら CRA なら「盲検化が守っているものは何か？」と考えられるはずです。
あなたの言葉で説明してみてください。

問 3
プロトコルには「評価方法」や「評価の時期」が細かく書かれています。
そこしら CRA として、その理由を統計との関係でどう説明しますか？

問 4
層別化はいつでも必須というわけではありません。
そこしら CRA として、層別化を取り入れるときの意味やメリットを、統計の役割とあわせて説明してみてください。

問 5
CRA 自身の態度や言葉が、結果的にバイアスを生むことがあります。
そこしら CRA なら「自分の行動が統計にどう影響するか」をどう意識して現場に臨むでしょうか？

問 6
多施設共同治験では「施設ごとにやり方が少し違う」だけでも、統計の公平さが揺らいでしまいます。そこしら CRA なら、“この施設だけ特別ルール”を見逃さないためにどんな目を持つべきか？ あなたの言葉で書いてみてください。

問 7
中間解析や DMC は治験の進退を決める大きな役割を担いますが、CRA は結果をのぞいてはいけません。では、そこしら CRA としては、「知らないからこそできる支え方」 とは何でしょうか？ あなたの考えを述べてください。

4. なんでこんなに被験者を集めるの？

4.1. たくさん集めないと意味がないの？

「1 群 260 例、合計 520 例」――プロトコルを開いたとき、初心者 CRA ならきっとこう思うでしょう。「どうしてこんなに必要なの？ もっと少なくてもいいんじゃないの？」

ここで思い出してほしいのが、第 1 章から学んできたことです。治験はサイエンス。サイエンスで大切なのは、集めたデータが「試験薬の効果を本当に示しているのか」を確かめることでしたよね。

少人数だと何が起こるのでしょうか？

第 2 章や第 3 章で触れたように、グループに分けて比較するとき、少人数だと偶然のゆらぎに振り回されやすくなります。たとえば 10 人ずつで比べると、試験薬群にたまたま元気な人が多く集まって「効いたように見える」ことがあります。でも、それは薬の効果ではなく「偶然のかたより」かもしれません。統計は、その偶然をならして「本当に効いているのか」を見きわめる仕組みです。しかし、統計の力も、データが十分に集まらなければ発揮できません。

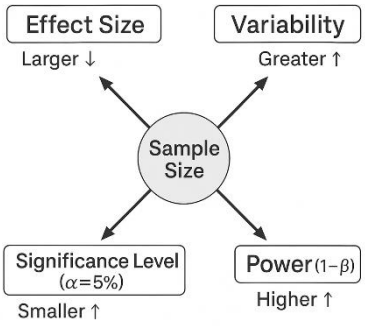
では、見極めるために必要な例数とはどのくらいなのでしょうか？

そこで統計家は、あらかじめ「このくらいの差が本当にあるときに、それを“偶然じゃない”と判断できるだけの人数」を計算します。これが必要例数です。たとえば「本当に効果があるなら、見逃さないで証明したい」し、「本当は効いていないのに効くと勘違いしたくない」。この 2 つのリスクをバランスよく小さくするように、統計家が人数を設計するのです。プロトコルに書かれている「1 群 260 例、合計 520 例」という数字は、ただの目安ではありません。それは「この試験薬が本当に効くなら、その証拠を“ちゃんとつかまえる”ための人数」なのです。統計家が“科学的な理由”で導いた数字なのです。

CRA がこの例数の数字を知っておくことは、ただの知識ではありません。「なぜ 520 例も必要なのか」を理解できていれば、モニタリングのときに「この症例ひとつも欠けてはいけない」という重みが伝わってきます。1 人ひとりのデータが、“本物と偶然を分ける証拠”になるのです。

4.2. 例数は“なんとなく”では決まらない

では、例数設計についてもう少し詳しく見てみますが、例数設計の理屈の話になりますので、興味のない人は飛ばしていただいても結構ですよ。例数設計は統計家が“科学的な理由”で導いた数字ということだけは理解しておいてくださいね。



さて、たとえば 2 つの薬剤の有効性の違いが「偶然の差」でないことを示すためにはたくさんのデータの数が必要ということをお話ししました。言い換えると、どんなに小さな差であっても、データがたく

さんあれば、差があると判断できます。これは直感的にわかりますよね。しかし、有効率を１％改善する薬が医学的に必要かどうかは疑問です。そこで、まず、事前に、医学的に意義のある差というものを設定します。例えば治験薬Ａが、これまでの既存薬よりも対象疾患を５％改善できるならば、医学的に意義がある、言い換えると、新薬として開発する意義があるならば、これを医学的に意義のある差として設定します。例数設計ではこの差を検出するための例数を算出しているのです。

やっと、例数設計の本題にはいれますね。繰り返しますが、治験で一番避けたいのは「まちがった結論」を出すことです。でも、その“まちがい”には２種類あります。**本当は効いていない薬を「効いた」と信じてしまうまちがい**、これは社会に役立たない薬を出してしまう危険があります。一方で**本当は効いている薬を「効かない」と見逃してしまうまちがい**というのもあり、患者さんを救えるチャンスを逃してしまう危険ということになります。

たくさんデータがあれば、小さな差でも検出することが出来るということをお話ししました。だったら、５％の差を確実に検出するために、500 例とか 1000 例の試験をやればいいじゃん、ということになります。でも、そうすると臨床的には意味のない小さな差まで「統計的に有意」と出てしまう危険があります。つまり、“数字の上では効いたことになるけれど、患者さんにとって本当に意味がある差なのか？”がわからなくなってしまうのです。そこで、例数設計では、効いていない薬を効くと判定する確率（第一種過誤）と効いている薬を効かないと判定する確率（第二種過誤）の両方をあらかじめ小さくコントロールしながら、臨床的に意味のある差を見逃さないための“ちょうど良い例数”、“必要以上に大きすぎない”例数を確率の理屈で算出します。したがって、この**２種類のまちがいをできるだけ小さくするための仕掛け**と考えても良いでしょう。だから例数は“なんとなく”ではなく、科学的な理由で導かれているのですよ。

4.3. なぜ5%なの？

治験の世界では「有意水準 5%」という数字がよく出てきます。

「5%未満なら有意」「5%を超えたら有意じゃない」と、まるでお役所が定めたルールのように扱われていますが、実はこれは行政が決めたものではありません。

では、なぜ「5%」なのでしょうか？

これは統計学の歴史の中で、「まちがいをどこまで許すか」の目安として自然に定着してきた数値なのです。統計学の父の一人とされるフィッシャーが、研究で「有意とみなす基準」として便宜的に 5%を提案したのが始まりといわれています。以来、多くの分野で使われるうちに、臨床試験の世界でも「国際的に共通する基準」として定着してきました。

5%というのはつまり、「100 回同じ試験をやったら、5 回くらいは偶然で“効いたように見える”結果が出ても仕方ないとみなそう」という線引きです。1%にすれば「まちがって効いたと判断するリスク」はもっと減らせますが、そのぶん「本当に効いている薬を見逃すリスク」が増えてしまいます。逆に 10%では「効いていない薬を効くと判定する危険」が大きすぎます。

そこで、5%というのは 「厳しすぎず、ゆるすぎない、社会的に納得できる基準」として、長い経験の中で実務上の標準になってきたわけです。

大事なのは、「5%だから正しい」ではなく、「5%という共通のものさしを使うことで、治験の判断を公平にし、国や時代を越えて比べられるようにしている」という点なのです。

コラム：PPI と被験者募集 ― 例数設計を現実にするために

症例数の計算は統計の世界で緻密に行われます。でも、どんなに科学的に「260 例必要」とはじき出しても、患者さんが集まらなければ試験は進みません。ここに現れるのが「統計」と「現実」のギャップです。

近年注目されているのが **PPI（Patient and Public Involvement）**、患者さんや市民が治験の計画段階から関わる仕組みです。たとえば「この条件なら参加しやすい」「通院回数が多すぎると続けられない」といった声は、統計の数式には出てこないけれど、実際のリクルートメントには決定的に影響します。

特に **希少疾患**では、対象となる患者数がもともと少なく、1 人の参加の有無が試験全体の成否を左右します。**小児領域**では、治療を受ける本人だけでなく、保護者の理解と同意が不可欠で、試験参加のハードルはさらに高くなります。こうした現場の事情を理解せずに「必要例数」を掲げても、計画倒れになってしまうのです。

CRA にとっても、例数設計の裏には「参加してくれる人がいて初めて成り立つ」という現実を意識することが大切です。被験者募集が滞れば、統計的検出力は机上の空論に終わります。だからこそ、統計をサイエンスとして尊重しながら、患者さんの声や生活の実情をどう拾いあげるか――そこに PPI の役割があるのです。

4.4. p 値の勘違い

さて、治験の結果として「p 値が 1%でした」と聞くと、「5%よりも小さいのだから、薬の効き目も 5%より強いのだ」と考えてしまう人がいます。

でも、これは大きな誤解なのです。

p 値が小さいというのは、「今回の結果が、もし本当に薬が効いていないとしたら、偶然にこんなに差が出てしまう確率がとても低い」ということを意味しています。つまり、「たまたま出た差ではなさそうだ」という根拠が強まった、ということです。

しかし、それはあくまで「偶然ではない」という話であって、「薬の効き目がどれくらい大きいのか」ということは教えてくれません。

コラム：Significant-itis ってなに？

臨床試験の論文を読んでいると「有意差あり！」という言葉が大きく取り上げられることがあります。すると「じゃあ、この薬は本当に効くんだ！」と短絡的に思ってしまう人も少なくありません。こうした「p 値が 0.05 を切ったかどうか」だけに過度にこだわる態度を、皮肉をこめて Significant-itis（シグニフィカントシス）と呼びます。

でも、本来の p 値は「差が偶然に出る確率」を示しているにすぎません。「有意差あり！」＝「効いている」と即断できるわけではないのです。効果の大きさ（臨床的意義）、再現性、ほかの試験との整合性などを見なければ、本当の答えは見えてきません。

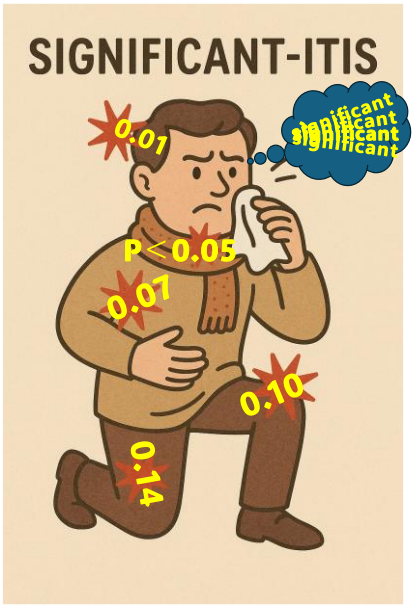
そこしら CRA に求められるのは、この Significant-itis に感染しないこと。「有意差あり」という結果を鵜呑みにせず、その裏にある統計の意味と限界を理解し、冷静に考えられる姿勢です。

たとえば、たくさんの患者さんに協力してもらえば、ごくごく小さな差でも統計的には“偶然ではない”と判断できてしまいます。逆に、参加人数が少なければ、差が大きくても“偶然かもしれない”と見なされることがあります。

つまり、p 値が小さいかどうかは、効き目の大きさそのものではなく、「偶然の可能性」をどの程度しぼり込めたかを示すだけなのです。

だから、p 値が 1%だからといって「5%よりも効いている」ということにはなりません。

大事なのは、「どれくらい差があったのか」という効果の大きさをしっかり見ることです。p 値はその差が偶然ではないらしいと示す数字にすぎないのです。



4.5. 真実は神のみぞ知る ― 仮説検定の枠組み

治験では、新しい試験薬が本当に効くのかどうかを確かめようとします。でも、ここで大事なことがあります。私たちは「絶対の真実」を直接つかむことはできない、ということです。治験を科学的に評価するための仮説検定という方法は、言ってみれば「裁判」のようなものです。「薬は効かない（帰無仮説）」を出発点にして、集めた証拠（データ）から「いや、効かないとするには無理がある、効いていると考えた方が自然だ」と判断できたときに、帰無仮説をしりぞけるのです。

でもこれは「絶対に効く」ことを証明するものではありません。あくまで「効いていないと考えるには無理がある」といった程度の論理的な判断にすぎません。だから、どんなに立派な第Ⅲ相試験であっても、「この試験薬は100%真実として効く」とは言い切れません。その証拠に、市販された後に「やっぱり思ったほど効かない」と分かったり、まれな副作用が見つかったりすることがあります。

つまり、治験のデータは“限られた視点での近似的な真実”にすぎないのです。「真実は神のみぞ知る」と言われるのはそのためです。だからこそ、日本では「市販後調査」や「再審査」「再評価」といった仕組みが制度として組み込まれています。

治験だけでは見えてこない長期の効果や副作用、日常診療の中での実際の効き目を、時間をかけて確かめ直していくのです。

承認は最終ゴールではなく、むしろ出発点なのです。治験で得られた「限られた真実」を社会に広げ、時間をかけて本当に妥当かどうかを確かめ続ける。そのために市販後調査や再審査の制度があるのです。

第４章理解度セルフチェック（そこしら式 記述テスト）

問 1.
プロトコルに「1 群 260 例」と書かれていました。
新人 CRA が「多すぎじゃない？」とつぶやきました。
その 260 例にはどんな“科学的な理由”が隠されているのか、新人 CRA に説明しましょう？

問 2.
「有効率が 1%改善する薬」――これって本当に新薬として開発する価値がある？
統計家はどうやって「医学的に意味のある差」を例数設計に組み込むのでしょうか？

問 3.
治験で避けたい“まちがい”は 2 種類ですよね。
どんなまちがいだったでしょう？

問 4.
「35 が 1 %で有意」という結果を聞いて、あなたの同僚が「5%より効いているってことね！」とドヤ顔で言いました。さあ、あなたならどうツッコミを入れますか？

問 5.
統計は「真実をズバリ言い当てる魔法の杖」ではありません。
じゃあ、仮説検定は何を示しているのでしょう？
市販後調査や再審査が必要な理由も併せて説明してください。

5. データって、そんなに素直じゃない

5.1. 同じ試験薬でも、効く人と効かない人がいる

治験の結果を見ていると、「この試験薬は効いた！」という人もいれば、「あまり変わらなかったな」という人もいます。そう、同じ試験薬を飲んでも、みんな同じように効くわけではないのです。

なぜでしょう？

それは、人間が一人ひとり違うからです。体質、年齢、性別、合併症、生活習慣…。いろいろな要素が試験薬の効き方に影響します。だから試験薬の効き目には自然と差が出てしまうのです。統計的に言えば、これは「ばらつき」として数字に表れます。効いた人と効かなかった人が混ざっているからこそ、平均値は揺れます。検定では、このばらつきを物差しにして「差が偶然なのか、本物の効果なのか」を見きわめようとするのです。CRA にとって、この「同じ試験薬でも効く人と効かない人がいる」という事実は単なる知識ではありません。「この試験は、誰に効くことを調べようとしているのか？」――そう問いかけながらプロトコルを読むことが大切です。

実はこの視点が、あとで出てくる Estimand（エスティマンド） に直結します。治験は「誰に、どんな状況で、どんな効き目を見たいのか」を最初に決めておく必要があるのです。個人差の理解は、その入り口に立つことなのです。

5.2. へんなデータ、抜けたデータ、どうするの？

治験の現場では、思ったとおりにデータがそろうことは、まずありません。来院を忘れる、体調不良で検査を受けられない、試験薬を途中でやめてしまう…。ときには「こんな値ありえない！」という入力ミスもあります。でも、これは不思議でも特別なことでもなく、人間が関わる以上、自然に起きることです。

5.2.1. 欠測や異常値を“きれいに埋める”と危険

「空いているところは埋めちゃえばいい」と思うかもしれません。でも、欠測をむりやり平均値などで埋めると、本来あるはずのばらつきが小さく見え、試験薬の効果が実際より大きく見積もられることがあ

ります。これがバイアス（偏り）です。つまり、「欠測をゼロにする」ことが正しいのではなく、「欠測がある前提でどう扱うか」を考えるのが科学的な姿勢なのです。

5.1.で人間が一人ひとり違うということを説明しましたが、それと密接に関係しています。

5.2.2. 国際ガイドラインはどう言っている？

E6(R3)による GCP では、欠測の埋め方よりも大事なのは、データを正直に残すこと（真正性）だと言っています。欠測があったら「なかったこと」にせず、理由とともに記録に残すことが求められています。大丈夫ですよね。

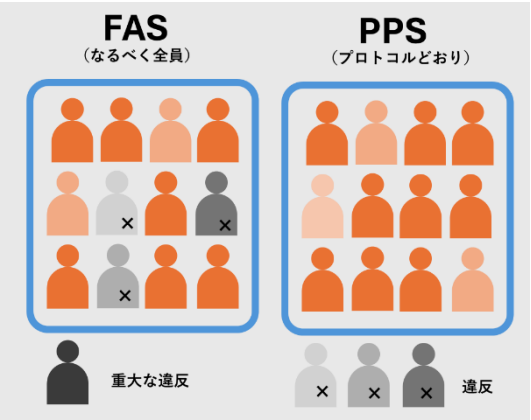
E8(R1) 一般原則では、欠測は「治験をゆがめる大きなリスク」と明記されています。だから、欠測が起きにくいように工夫すること、そして起きたときの扱いを最初から決めておくことが求められています。

E9 統計的原則はとても手厳しいですよ。欠測や外れ値は「統計解析を誤らせる犯人」だと名指しです。平均値で埋めるような安易なやり方はバイアスの温床になるので、取り扱い方法は事前に計画に書いておくべきとされています。

E9(R1) Estimand では、これからの新しいスタンダードが示されています。欠測を「どう埋めるか」よりも、**「この欠測は何を意味するのか」**をあらかじめ決めておくことが重要としています。たとえば「試験薬をやめたら“効果が切れた”と解釈するのか」「別の薬を使ったら“現実の経過”として数えるのか」といった具合です。欠測は単なる“穴”ではなく、出来事として意味づけするものだとされています。

5.2.3. 解析集団という考え方

では実際の解析ではどうするのでしょうか。統計家は、最初から「どの患者さんを解析に含めるか」をルール化しておきます。ええっ、欠測はイレギュラーなのだから、先にルール化するなんてできないよ。たしかにそうですよね、でもそうしなければならない理由があるのです。なぜなら、結果を見てから「この人は入れよう、外そう」と決めてしまえば、都合の良いデータだけを残すことになり、結論がゆがんでしまうからです。バイアスですよね。



そこで国際的なルール（E9）では、主に次の三つが定義されています。

FAS（Full Analysis Set）

無作為化された患者さんをできるだけ全員含める。途中でやめた人も可能な限り入れる。

→ 「現実の診療に近い姿」を反映させるための解析。

（背景にあるのが **ITT 原則** です。「無作為化したら原則全員含める」という理想を示した原則で、それを実務的に適用したものが FAS です。）
たとえば、100 人を無作為化した試験で、同意直後に中止して全くデータが得られなかった 2 人がいたとします。

- ITT 原則では「その 2 人も含めて 100 人全員」とするのが理想。
- しかし実務では解析が成立しないため、FAS では「98 人」で解析を行う。
  → **ITT が理想、FAS が実務** という関係です。

PPS（Per Protocol Set）
プロトコルどおりに治験薬を使い、必要なデータが揃った人だけを含める。
→ 「理想的に治験薬をきちんと使った場合の効果」を見る解析。

> ただし、「プロトコルどおりに使ったか」を形式的に捉えすぎると、実際の臨床状況を反映しない解析母集団となり、外的妥当性が損なわれるおそれもあります。
> 「プロトコルに定めた条件をおおむね遵守して治験薬を使用し、評価に必要なデータが揃った人を含める」と捉えたほうが適切でしょう。
> PPS の解釈にも、CtQ に基づくバランス感覚が必要になります。

安全性集団（Safety Set）
少なくとも 1 回は治験薬を投与された患者さんを含める。
→ 副作用や有害事象など、安全性を評価するための解析に用いられる。

どれが正しいということではなく、**FAS は“現実の姿”を、PPS は“理想の姿”を、安全性集団は“副作用リスクの姿”を示す**のです。治験ではこれらを見比べて判断します。
CRA にとって大事なのは、「途中でやめた人のデータも無駄ではない」ということ。どんな欠測も、適切に扱えば大切な情報になります。そのためにこそ、現場での出来事を正しく記録し報告する役割が欠かせないのです。
その欠測にはワケ（理由）がある、心してください。

5.3. もう一度“ばらつき”のはなし

第 5.1 でお話ししたように、薬の効き目には個人差があります。
「効いた！」という人もいれば、「あまり変わらなかった」という人もいる。
この個人差は数字の世界では「ばらつき」として表れます。集団の平均値はこのばらつきに揺さぶられ、検定はそのばらつきを使って「差が偶然か、本物か」を判断することになります。ここでは、その「ばらつき」と検定の関係をもう少し統計的な観点から見てみましょう。

5.3.1. 個人のばらつきと集団のばらつき

まず 個人のばらつき。これは「ある人には効いたが、別の人には効かなかった」という、一人ひとりの違いです。一方で、集団のばらつきは、こうした個人の差を全部まとめた「全体としての揺れ具合」を数字にしたものです。ここで言う“集団のばらつき”は、中学数学でも出てくる標準偏差や分散で表されます。

この集団のばらつきがあるために、同じ薬で同じ条件の試験を繰り返しても、平均値は毎回少しずつズレます。たとえば、1 回目は 60%、2 回目は 58%、3 回目は 62%…といった具合です。つまり、試験ごとの平均値の揺れは、集団にばらつきがあるから生じるのです。

5.3.2. 差を判断するには「集団のばらつき」と比べる

治験で私たちが見たいのは「集団として試験薬と既存薬に差があるかどうか」です。だから、個人の差そのものではなく、集団全体のばらつきを物差しにして差を判断する必要があります。集団のばらつきが小さければ 5%の差は「はっきりした差」に見える。集団のばらつきが大きければ5%の差は「偶然のブレかもしれない」と見えてしまう。こうして統計は、「差」そのものではなく「差と集団のばらつきの比率」を見て、本当に効いているのかどうかを判断するのです。

5.3.3. S/N 比という考え方

この考えをわかりやすく表すのが S/N 比（シグナル／ノイズ比） です。
シグナル（Signal）＝薬の効果の差（集団平均の差）
ノイズ（Noise）＝集団のばらつき

シグナルがノイズに比べて大きいとき、「薬の効果が偶然ではなく、はっきり見えている」と言えるのです。コンサート会場で大音量の音楽が流れている中、隣の友だちが何か話しかけてきました。聞き取れるときもあれば、雑音にかき消されてわからないときもありますよね。このとき大事なのは「声（シグナル）」と「雑音（ノイズ）」のバランスです。音楽が小さく、友だちの声が大きければはっきり聞き取れる。逆に音楽が大きすぎると、同じ声量でもほとんど聞こえません。

治験データも同じです。薬の効果というシグナルを、個人差や記録ミスといったノイズの中から拾い出すのが統計の役割です。ノイズが小さければ、ほんの少しの差でも「効いている」とはっきり見えてきます。ノイズが大きければ、大きな差があっても「偶然かもしれない」とぼやけてしまいます。

そこしら CRA にできるのは、このノイズを減らすこと。記録ミスを防ぎ、測定条件をそろえ、欠測を最小限にする。統計の物差しをクリアに保つことこそが、薬の効果を正しく見極めるための大切な現場の仕事なのです。

CRA として理解しておくこと

検定で使うのは「集団のばらつき」ですが、それは一人ひとりの「個人のばらつき」の積み重ねから計算されています。だから現場では、余計なばらつきを増やさない工夫、たとえば、記録ミスを防ぐ、測定条件をそろえる、単純な欠測が生じないような構造にするといったことがとても大切になります。

だからこそ、CRA が現場で“余計なばらつき”を減らすことは、単なる事務作業ではなく、科学的証拠を磨き上げる大切な役割なのです。

5.4. 評価変数の種類を見きわめよう ～何を一番大事に見るのか～

「評価変数」という言葉を聞いて戸惑う人も多いかもしれません。ここでいう「変数」とは、“患者さんから集めてくるデータの種類”のことです。そしてそのデータは一人ひとり違っていて、まさに「変動する数」だからこそ変数と呼ばれるのです。例えば「体重」「血圧」「生存の有無」といった項目は、人によって値が異なるからこそ統計で扱う対象になります。

このように変数は必ず変動するものなので、その集まりには「ばらつき」が生じます。前節で取り上げたばらつきの影響は、「どの変数で起きたか」によっても大きく変わるのです。つまり、臨床試験で集めるデータはすべてが同じ重みを持つわけではなく、そこには役割の違いがあるのです。

コラム：変数って名前のヒミツ

「変数」という言葉、なんだか数式のにおいがして身構えませんか？でもその語源は意外とシンプル。ラテン語の variabilis（変わりやすい）から来ていて、英語の variable も「変動するもの」という意味なんです。

たとえば治験で集める「体重」「血圧」「生存の有無」。患者さんごとに違う値をとるからこそ、それは“変わり得る数＝変数”と呼ばれています。

しかも「数」だけじゃありません。「男性／女性」といった区分や、「有／無」といった状態も立派な変数。要するに 観察のたびに変わり得る属性は全部「変数」なのです。

だから「変数」とは、治験で集める“変わるデータの種類”のこと。名前の由来を知れば、ちょっと親しみやすく感じられませんか？

5.4.1. 主要変数（Primary Endpoint）

試験の心臓部であり、「この薬は効いた」と言えるかどうかを決める最も重要なデータです。ここで欠測や逸脱が起きれば、試験全体の結論が揺らいでしまうほど大きな影響があります。まさに CTQ の中心であり、絶対に守らなければならない項目です。

5.4.2. 副次変数（Secondary Endpoints）

副次変数は薬の物語を豊かに彩る脇役です。主要変数が物語の主役だとすれば、副次変数は物語を補う大切な要素です。そして安全性変数は患者さんを守るためのデータであり、場合によっては主要変数と同じくらいの重みを持つこともあります。

ただし、No More Too Much（E6（R3）の考え方に従い、チェックリスト的な GCP から“考えさせる GCP”へ移行するために、治験において過剰な仕事を特定し排除する日本 CRO 協会の活動）といった考え方に立てば、すべてを主要項目と同じ厳しさでチェックするのは誤りです。主要は CTQ の中心として死守し、副次や安全性は役割に

応じて適切に確認する。現場で医療者が既に担っている部分も信頼しつつ、リスクに基づいて重点を置くことが CRA に求められます。

コラム：QbD と CTQ ってなんだろう？

最近の GCP Renovation では、**QbD（Quality by Design）** と **CTQ（Critical to Quality）** という言葉がよく出てきます。聞き慣れないかもしれませんが、考え方はシンプルです。

QbD とは「品質を設計する」という発想です。臨床試験では、後から不具合を修正するのではなく、最初から「何を大事にするのか」を決め、その品質が守られるように仕組みを組み込んでおくという考え方です。たとえば、主要評価項目をどうやって測定するか、安全性の情報をどうやって漏れなく集めるか、といった工夫がこれにあたります。

その QbD の考え方を具体的にしたのが CTQ です。CTQ は「品質にとって決定的に重要なもの」という意味で、試験の信頼性を左右する因子のことを指します。たとえば主要評価項目のデータ、誰を試験に入れるかという適格基準、治験薬の投与遵守、有害事象の記録などが代表的な例です。ここで欠測や逸脱があれば、試験全体の信頼性が揺らいでしまいます。

CRA にとって大切なのは、「全部を同じようにチェックすること」ではありません。限られた時間の中で、特に CTQ にあたる部分を見きわめ、そこを重点的に確認することが求められます。主要評価項目や安全性の核心部分は CTQ の中心であり、絶対に落としてはいけないデータです。一方で副次変数や補足的なデータは CTQ には含まれませんが、薬の物語を豊かにする要素として、丁寧に拾い上げる価値があります。

要するに、QbD は「品質を設計する考え方」、CTQ は「その中で特に守るべき決定的な因子」という位置づけです。CRA は「この試験で CTQ は何か」を意識し、モニタリングの力をそこに集中させることが、信頼性の高い臨床試験につながります。

5.4.3. 代替変数（Surrogate Endpoints）

本来は「生存期間が延びるか」「脳卒中の発症を減らすか」といった最終的な臨床アウトカムを見たいところですが、それには長い時間や多数の患者さんが必要です。そこで、がん治療では **腫瘍の縮小率** を生存期間の代わりに、高血圧では **血圧の下がり方** を脳卒中や心筋梗塞発症の代わりに使うことがあります。

代替変数は短期間で結果を得やすい利点がありますが、「本当に患者さんの利益につながっているのか」が科学的に検証されている必要があります。腫瘍が小さくなっても生存期間が延びるとは限らないし、血圧が下がっても脳卒中が減るとは限りません。主要評価項目として使う場合は、その妥当性が確認されていなければならないのです。

5.4.4. 安全性変数（Safety Endpoints）

有害事象や臨床検査値、バイタルサインなどがこれにあたり、薬が“安全に使えるかどうか”を判断するために不可欠なデータです。有効性と並ぶもう一つの柱であり、場合によっては主要変数と同じくらいの大切な重みを持つこともあります。患者さんの安心を守るためのデータとして、絶対におろそかにできません。

5.4.5. CRA にとっての意味

CRA にとって重要なのは、「この試験ではどのデータが心臓部＝CTQ なのか」を理解してモニタリングに臨むことです。

- 主要変数は勝負データとして死守する。
- 副次変数は薬の物語を補うデータとして丁寧に確認する。
- 代替変数は「真のアウトカムと結びついているか」を意識してチェックする。
- 安全性変数は患者さんを守るデータとして確実に押さえる。

同じ「データ」でも重みは違います。その違いを見きわめることこそ、CTQ を意識したモニタリングの第一歩であり、そこしら CRA に求められる視点なのです。

ここで言う、アウトカムとは、治験で「この薬は効いたか、安全か」を示すためにあらかじめ決めておく評価項目です。がん治療なら生存期間や腫瘍の大きさ、高血圧なら血圧値などがその例です。アウトカムは試験の結論を左右する“勝負データ”であり、プロトコルで厳密に定義されます。CRA にとっては、どのアウトカムが試験の中心なのかを理解することが、適切なモニタリングの第一歩となります。

一方で「副次変数」と呼ばれるデータもあります。これは主要変数を補う役割を持つもので、たとえば症状の改善度や生活の質（QOL）などが含まれます。承認の直接的な根拠にはなりにくいものの、薬の特徴やメリットを伝えるための大切な情報源となります。主要変数が物語の主役だとすれば、副次変数は物語を豊かに彩る脇役といえるでしょう。

そして忘れてはいけないのが「安全性変数」です。有害事象や臨床検査値、バイタルサインなどがこれにあたり、薬が“安全に使えるかどうか”を判断するために不可欠なデータです。安全性変数は有効性と並ぶもう一つの柱であり、場合によっては主要変数と同じくらいの大切な重みを持つこともあります。薬が“安全に使えるかどうか”を示すデータであり、患者さんの安心を守るために欠かせません。場合によっては主要変数と同じくらいの重みを持つこともあります。

CRA にとって重要なのは、**「この試験ではどのデータが心臓部＝Critical to Quality（CTQ）なのか」**を理解してモニタリングに臨むことです。

主要変数は CTQ の中心であり、欠測や逸脱は、特に主要変数に関しては CTQ に直

コラム：カテゴリ化した変数と CRA にとっての意味

血圧や年齢のように本来は連続的に測れるデータを、ある基準で区切って「高血圧／正常」「65 歳未満／65 歳以上」のようにラベルにすることがあります。これを カテゴリ化と呼びます。

なぜカテゴリ化するのか？

カテゴリ化の利点はシンプルさです。たとえば「140mmHg を超えたら高血圧」と示せば、直感的に理解できますし、臨床現場での判断に使いやすくなります。

しかし統計的には、カテゴリ化には弱点があります。情報が失われることです。たとえば、20 歳と 64 歳は大きく違いますが、65 歳未満でひとまとめ。逆に 139 と 141 は 2 しか違わないのに、境界を挟んで別グループになります。区切り方で結果が変わることもあります。たとえば、カットオフを 135mmHg にするか 140mmHg にするかで、治療効果の有無が変わってしまいますよね。

さらに、せっかく統計的に、論理的に処理をしようとしているのに、結果的にその効果が得られないことがあります。これを、統計家は統計的なパワーが落ちると言います。連続値の細かな違いを使えなくなるため、効果を見つけにくくなることがありますよね。

こうした理由から、E9 では「安易なカテゴリ化は避けるべき」とされています。どうしても臨床的に意味がある基準で使う場合には、その根拠を明確にすることが求められます。

さて、CRA にとってのカテゴリ化はどのような意味を持つのでしょうか？正直に言えば、**カテゴリ化そのものは統計家が後で処理する話であり、CRA が日常のモニタリングで意識する必要はありません。**むしろ境界値を特別視してしまうと、測定や記録の仕方に偏りが出て、かえってバイアスの温床になりかねません。

ただし「カテゴリ化にはもろさがある」ということを知っておくと、理解が深まります。それは、プロトコルに書かれた基準値がきちんと守られているかを確認する視点につながることになります。また、境界付近であっても「特別扱いするのではなく、すべての測定を同じルールで正確に行うことが大事」と理解しておくことが大切ですよね。

カテゴリ化は便利な道具ですが、同時に“情報を削ってしまう両刃の剣”です。CRA として直接扱う必要はありませんが、「結果の解釈にはこういう仕組みがあるのだ」という背景知識として知っておく価値はあります。

結するため、絶対に防ぎ、確実に守らなければなりません。副次変数は CTQ には含まれませんが、薬の物語を豊かにする要素として計画どおりに集められることが望まれます。ただし、No More Too Much といった考え方に立てば、副次まで主要と同じ厳しさで追いかける必要はありません。現場の医師や CRC が適切に記録している部分を信頼しつつ、リスクに応じて確認する姿勢が大切です。

安全性変数は患者さんを守るためのデータであり、誤りや漏れはリスク評価そのものを誤らせてしまうため、CTQ に匹敵する重要性を持ちます。ここは主要項目と並んで確実に押さえる必要があります。

つまり、同じ「データ」でも、その重みは違います。主要項目は勝負データとして死守、副次項目は薬の物語を補うデータとして必要に応じて確認し、安全性項目は患者さんを守るデータとして確実に押さえる。この優先順位を意識することで、モニタリングで何を最優先にすべきかが明確になります。

こうした役割の違いを見きわめ、**CTQ を中心に据えつつ現場を信頼する**ことこそ、リスクベースドアプローチの第一歩であり、そこしら CRA に求められる視点なのです。

- **主要変数（Primary Endpoint）**　：CTQ の中心
- **副次変数（Secondary Endpoints）**　：CTQ ではないが薬の物語を広げる要素
- **代替変数（Surrogate Endpoints）**　：真のアウトカムの代わりに用いる
妥当性の理解が必須
- **安全性変数（Safety Endpoints）**　：患者さんの安心を守る要素
- **アウトカム（Outcome）**　：薬の効果を示す勝負項目

5.5. すべてのデータは報告される

ここまで見てきたように、臨床試験で集められるデータにはそれぞれ重みや役割があります。主要変数は試験の結論を決める心臓部であり、副次変数は物語を補う要素、安全性変数は薬が安心して使えるかどうかを示すもう一つの柱です。代替変数やカテゴリ化した変数にも、それぞれの意味と限界がありました。

では、集められたデータのうち、最後に報告されるのはどれでしょうか。答えは「すべて」です。

ここでいう「すべて」とは、プロトコルで規定された収集対象データのことです。これらは最終的に承認申請の資料として規制当局に提出されることになります。E9 でも強調されているように、臨床試験の結果は、主要・副次・安全性を含め、事前に計画された解析方法に従ってすべて報告されなければなりません。都合のよいデータや“きれいな結果”だけを取り出すことは許されませんし、探索的に追加で行った解析も、区別して正しく示す必要があります。

ここで CRA にとって大切なのは、「プロトコルで規定されたどんなデータも最終的には承認申請に報告される」という事実を理解することです。しかし、だからといって **すべてを同じ重みで点検する必要はありません**。むしろ、No More Too Much といった考え方に従えば、CTQ 以外の項目まで過度に細かくチェックしすぎることは、本来守るべき CTQ の監視を手薄にするリスクがあります。

主要項目は CTQ の中心であり、欠測や逸脱を絶対に防ぐべき最優先の対象です。一方、副次や安全性のデータはもちろん大切ですが、これらは通常の診療の中ですでに医師や CRC によってチェックされているケースも多く、CRA が二重に過度の労力をかける必要はありません。大切なのは、**現場の医療者を信頼し、役割に応じた確認をすること**です。CRA は「すべてを等しく厳しく見る」役割ではなく、**リスクベースドアプローチとして、CTQ を優先しつつ、必要なところだけを適切に押さえる**役割を担っています。

つまり、CRA のモニタリングは「すべてをくまなく点検すること」ではありません。現場を信頼し、役割分担を踏まえて、CTQ を中心に確実に守る。それが治験を科学として成り立たせる、最も効果的なスタンスなのです。

プロトコルで規定されたすべてのデータは承認申請として報告される。
しかし CRA は、現場を信頼しつつ CTQ に集中する。
その優先づけこそが真のモニタリング。

第５章理解度セルフチェック（そこしら式 記述テスト）

問１
新人 CRA が「平均値だけを見て安心している」様子です。その項目に個人差があることを気にしていません。
あなたは先輩 CRA として、どのようにアドバイスしますか？

問２
「同じ試験薬で同じ条件なのに、平均値が毎回ちょっとずつ違うのは、データがいいかげんだよね」と同僚 CRA が言っています。
あなたなら、この“揺れ”の理由をどう説明し、CRA として現場で気をつけるべきことをどう伝えますか？

問３
「試験薬と既存薬で 5%の差があるなら、もう“効いている”って言えるんじゃないですか？」と質問されました。
あなたなら、どんな“物差し”と比べる必要があると説明しますか？

問４
新人 CRA が「記録ミスや欠測が少しくらいあっても、統計で処理できるんじゃないですか？」と気軽に言っています。
あなたは、シグナルとノイズの観点から、なぜ余計なノイズを減らすことが大切なのかをどう説明しますか？

問５
先輩 CRA が「結局、目標の差が付きさえすればいいんだよ」とつぶやきました。
あなたは「差そのもの」ではなく「差とばらつきの比べっこ」であることを思い出し、先輩 CRA に気づいてもらえるようにつぶやきました。なんとつぶやいたのでしょうか？

問６
後輩 CRA が、副次評価項目に欠測を見つけて「すぐに医療現場に調査して埋めるよう指示しなきゃ！」と張り切っています。
あなたなら、そこしら CRA としてどんなアドバイスをしますか？

6. 「この薬、ほんとに効くの？」のホントの意味

6.1. 「効いたっぽい」と「効いていると言える」の違い

前の章でお話ししたように、薬の効き方には人それぞれの違いがあります。ある人にはよく効いても、別の人にはそれほど効かない。だから集団の平均値はゆらぎます。このため、**「効いたように見える」ことと、「効いていると言える」**ことは、同じではないのです。

たとえば試験薬群の回復率が 60%、既存薬群が 55%だったとします。数字だけを見れば「5%の差があるじゃないか！」となりますよね。けれども、その差が**たまたまのゆらぎ**かもしれない。小さな試験ならなおさらです。逆に、同じような差が繰り返し観察され、しかも集団のばらつきに比べて十分大きければ、「偶然では説明できないぞ」と胸を張れるようになります。この「効いたっぽい」と「言える」の境目をつけてくれるのが、統計なのです。

何度も繰り返しますがここは重要なポイントなのです。

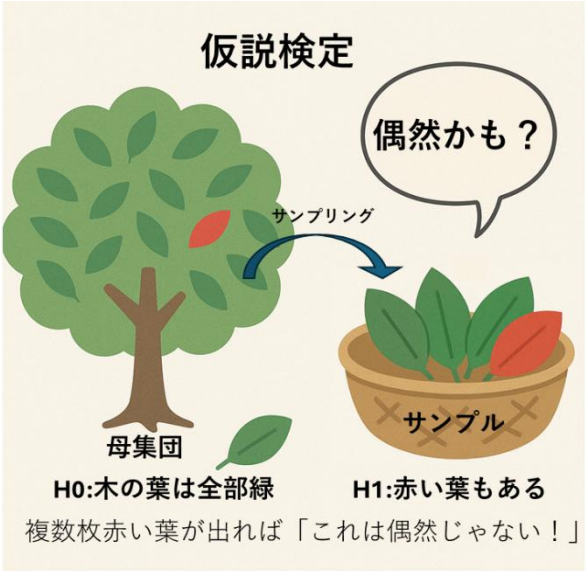
6.2. 仮説検定って意味わからないよね

では、統計はどうやって「効いていると言える」と判断するのでしょうか。その仕組みのカギが **仮説検定** です。仮説検定は、じつは **背理法** という考え方を使っています。つまり、いきなり「効いている」と証明しにいくのではなく、まずは逆を仮定してみるのです。

「よし、じゃあ“効いていない”ことにしてみよう」

――これを **帰無仮説** と呼びます。

そして考えます。「もし本当に効いていないのなら、今みたいに試験薬と既存薬で差が見えちゃう確率って、どのくらいあるんだろう？」。その確率がとても小さいとわかったら、「効いていない」という仮定の方がむしろおかしい、となりますよね。



ここで登場するのが **対立仮説** です。帰無仮説が「効いていない」だとしたら、その反対側にある「効いている（差がある）」という主張が対立仮説です。仮説検定の流れは、「帰無仮説を棄却できるかどうか」を調べ、その結果として対立仮説を支持する、という形になります。直接「効いている」と証明するのではなく、「効いていない」を否定することで「効いている」の可能性を強めていく、というのが統計の考え方です。

これが検定の基本の考え方です。

ここで素朴な疑問が出るかもしれません。

「効いているかどうかを知りたいのに、なぜ直接“効いている”と証明しないの？」と。

理由はシンプルです。**“効く”の姿はいろいろありすぎるから**です。

効果の大きさも人による違いも、本当は連続的で無限にありえます。だから「効く」という状態を１本針でピタッと定義して、直接証明するのは難しいのです。一方、「効いていない（差はゼロ）」は１本針で表しやすい。これを出発点にすれば、確率を計算して判定できます。つまり、**複雑な“効く”を直接追いかけるのではなく、シンプルな“効いていない”を否定していく**。その方が公平で、再現性のあるやり方なのです。だからこそ統計は、背理法を土台にした「仮説検定」という仕組みをとっているのです。

コラム：裁判と仮説検定

裁判ではまず「被告は無罪」と仮定するところから始まります。そして証拠を集めていき、「無罪だとすると、この状況は説明できない」というところまで来たら、有罪が認定されます。

仮説検定の仕組みも、これとよく似ています。治験では最初に「薬は効いていない（帰無仮説）」と置きます。そしてデータを集め、「効いていない」と仮定したのでは説明できないような結果が出たときに、はじめて「効いている」と結論づけるのです。

直接「効いている」と証明するのではなく、「効いていない」では説明がつかない、と背理法的に迫るのが統計のやり方。遠回りに見えますが、偶然に惑わされず科学的に判断するための、もっとも確かな仕組みなのです。

6.3. 有意差のホントの意味

仮説検定で出てくるのが「有意差」という言葉です。ちょっとカッコいい響きがありますが、実際の意味はとても慎ましいものです。有意差とは、「もし効いていないと仮定したら、今みたいな差が出ちゃう確率はすごく小さいですよ」という状態を指します。

言い換えると、**「偶然にしてはできすぎ」**。だから「効いているらしい」と胸を張れる、ということです。

でもここで大事な注意があります。

- ✓ 有意差がある＝**絶対に効いている保証**ではありません。
- ✓ 症例数が多ければ、ごく小さな差でも「有意」になります。
- ✓ 症例数が少なければ、大きな差でも「有意」にならないこともあります。
- ✓ １回の試験だけでは「真実」に近づいたとは言えず、再現性の確認が欠かせません。

		検定結果	
		差がある	差がない
真実	差がある	正しい判断 検出力（１－β） 検出力：８０％	間違った判定 第２種の過誤 （β）
	差がない	間違った判定 第１種の過誤 （α） 危険率：５％	正しい判断 （１－α）

有意水準
P＝0.05

つまり、有意差は「効いている可能性が高い」という**スタートライン**に過ぎないのです。だからこそ複数の試験が必要であり、第Ⅲ相試験の必然性もここから生まれます。そしてCRAにとってのポイントは、**この“偶然にしてはできすぎ”という証拠を濁らせないこと**です。記録のミス、欠測の多発、測定条件の不統一…これらはみな、統計の判定を曇らせるノイズになります。「余計なノイズを増やさない」というCRAの努力こそが、有意差の意味を支える大切な土台なのです。

6.4. 偶然とのたたかい

治験は「薬の効果」と「偶然のブレ」とのたたかいです。

小さな試験では、偶然の揺れに振り回されてしまいます。だからこそ、より大きな試験で、より多くの患者さんに参加してもらい、偶然のブレをならしていくことが必要になります。

ただし、偶然だけが敵ではありません。

本書の序盤でも触れたように、治験は「バイアスとのたたかい」でもあるのです。思い込みや先入観、不注意や不揃いな手続き――こうしたバイアスは、偶然以上にデータをゆがめてしまいます。つまり治験は、**偶然とバイアス、この二つの大敵に挑む科学の営み**なのです。

大規模な第Ⅲ相試験が必要とされるのは、偶然のブレを乗り越えるためです。そして厳密な手順やモニタリングが求められるのは、バイアスを封じるため。この二つがそろってはじめて、「ほんとに効く」と胸を張れる証拠が得られるのです。

コラム：Pivotal試験は1回でよいの？
― FDAと日本の考え方

昔は、FDAでは「独立した2つの主要試験（Pivotal試験）で同じ結果が得られること」を、有効性を証明する基本と考えていました。これは、「偶然の結果ではない」という再現性を重視していたためです。

しかし最近のFDAでは、“1つの信頼できる試験（A&WC：Adequate and Well-Controlled Trial）”に、十分な確認的な証拠を組み合わせることでも、有効性を示せるとしています。たとえば、強い用量反応関係や、他の試験・外部データとの一貫性などが確認的証拠になります。

一方、日本（PMDA）では、実施可能な患者数や試験環境を考慮して、単一試験が主要な根拠になることもあります。けれども、これは統計的な厳密さを緩めているわけではなく、試験全体のエビデンス（科学的根拠）を総合的に評価しているという考え方です。

つまり、国や制度が違っても目指すところは同じで、「科学的に納得できる一貫した結果を示すこと」が共通のゴールです。

大切なのは、結果の再現性をどのように確保するかを設計段階から考えることなのです。

CRAとして理解しておくこと

CRAが偶然をコントロールすることはできません。けれども、**人為的なゆがみやノイズを減らすことはできる**のです。記録ミスを防ぐ、測定条件をそろえる、欠測を最小限にする。ここで言う「欠測を最小限にする」とは、**欠測が出ないような仕組みを守ること**を指します。

たとえば来院スケジュールの工夫や入力漏れの防止などで、欠測を未然に防ぐことです。**すでに発生した欠測を“想像で埋める”ことは決して許されません。**どうして欠測になったのかを正しく記録し、必要があれば医師や患者本人に確認して正しい情報を補完する――その姿勢が大切なのです（この「欠測とSOPの関係」については、もう一度第7章で確認しましょう）。言い換えれば、CRAは「統計が偶然と公正に戦えるように、土俵を守る役割」を担っているのです。

コラム：へたな鉄砲も数撃ちゃあたる、はだめ（多重比較）

治験のプロトコルには「主要評価項目」「副次評価項目」「探索的評価項目」といった形で、複数の評価指標が並ぶことがあります。ここで気をつけなければならないのが「多重性の問題」です。

複数の検定を行うと、たまたま偶然の差が「有意」に見えてしまう危険が高まります。これを統計の言葉で「タイプⅠエラー（α エラー）が膨らむ」といいます。例えば、10 個の項目を独立に検定すれば、そのうち 1 つは偶然でも「有意差あり」と出てしまう可能性があるのです。

統計家にとっては、これは非常に重要な課題です。有意水準（α）を守るために、プロトコルや統計解析計画（SAP）の段階で「どの項目を主要とするか」を明確に定義し、多重性を制御する方法（Bonferroni 法や階層的検定など）を事前に決めておきます。

つまり、**多重性は設計と解析で対応すべき問題**であり、現場でのモニタリングが直接関与することはありません。

では、CRA にとっては無関係かというと、そうではありません。プロトコルに「主要評価項目はこれ」と明確に書かれているのは、多重性のリスクを避けるためでもあります。主要項目は試験の結論を決める“勝負データ”であり、まさに **CTQ（Critical to Quality）の中心** に位置づけられています。だからこそ欠測や逸脱は絶対に許されず、CRA が最も厳密に守るべき対象になります。一方、副次項目や探索項目は薬の物語を豊かにするデータではあっても、承認の直接的な根拠にはなりにくく、CTQ には含まれません。

この背景を知っておくと、「なぜこの項目だけは絶対に欠測できないのか」「なぜ評価項目に優先順位があるのか」がはっきりと理解できます。つまり、多重性の問題は統計家が解析上で取り組む課題ですが、その前提として**主要項目＝CTQ を死守すること**が、CRA のモニタリングで最優先の使命になるのです。

第６章理解度セルフチェック（そこしら式 記述テスト）

問 1
「試験薬の平均が既存薬より高い。これって効いていると言っていいのかな？」
そこしら CRA のあなたは自分なりに判断しました。どのように判断したのでしょう。

問 2
新人 CRA が、「検定ではまず“効いていない”と仮定するらしい。どうしてわざわざそんな回りくどいことをするんだろう？ “効いている”を直接証明すればいいんじゃない？」そこしら CRA のあなたは、どのように説明しますか？

問 3
「有意差あり！って出たとき、私はどこまで自信を持っていいんだろう？」
この悩みについて、自分なりの考えを示してください。

問 4
「予備的に行った試験で予想以上に大きな有意差が出た。これでもう十分な証拠だよ。第Ⅲ相試験をスキップして申請すれば大幅コストダウンだ」開発部長が満面の笑みを浮かべて言いました。あなたはどう思いますか？最低２つの問題点を指摘してみてください。

問 5
「“欠測を最小限にする”ってよく言われる。でも、欠測が出てしまったときに“埋める”ことと“正しく記録する”ことはどう違うんだろう？」
この悩みについて、自分なりの考えを示してください。

7. SOP はなぜそんなに厳しいの？

7.1. ここまでの振り返りと、SOP の本当の役割

ここまでの流れを振り返ってみましょう。

第 1 章では、治験はサイエンスであり、**みんなが納得できる公平な比べっこ**が必要だという話をしました。統計はそのための物差しでしたね。第 4 章では、無作為化や盲検化といった仕組みが、**思い込みや先入観（バイアス）**を避けるための工夫だと学びました。さらに第 5 章では、薬の効き方には個人差があって、集団の平均値もゆらぐ。だから**差はばらつきと比べる**ことで初めて意味を持つ、なので、CRA は現場で**余計なばらつき**を増やさない工夫をすることが大事だと確認しました。第 6 章では、仮説検定が**背理法**を使って「効いていない」を仮定し、それでも説明できない結果が出たときに「有意差あり」と判断することを学びました。そして、仮説検定で重要な要素となる有意差は、「偶然にしてはできすぎ」というサインにすぎず、真実の保証ではないことも確認しましたよね。そして最後に、治験は「偶然のブレ」と「バイアス」の両方と戦う営みであることを強調したと思います。

コラム：E6（R3）で変わったデータマネジメントの役割

昔の GCP（E6（R1））では、データマネジメント（DM）という職能ははっきり定義されていませんでした。スポンサーが「データの正確性・完全性を保証する」とだけ書かれていて、その裏で DM が地道にデータ処理を担っていたのです。

ところが最新の E6（R3）では状況が一変します。キーワードは QMS（Quality Management System） と データ完全性（Data Integrity）。治験データを「収集 → 処理 → 解析」する流れ全体を品質システムの中で管理し、リスクに応じたコントロールをかけることが求められるようになりました。

つまり DM は単なる「データ入力の番人」ではなく、治験全体の品質を守る仕組みの一部として公式に組み込まれたのです。欠測や矛盾を“後から直す”のではなく、最初から欠測やエラーが起きにくい構造を設計することが役割として強調されています。

そこしら CRA にとっても、DM は「統計家の後方支援」ではなく、「治験データの真正性を前線で支えるパートナー」。E6（R3）は、そのことをガイドラインとして明確に示したのです。

さて、ここで登場するのが SOP（標準業務手順書）です。

SOP が厳しく定められているのは、単なる形式や事務的な縛りのためではありません。データにウソをつかせないための仕組みだからです。

一方で、SOP の運用が硬直的になると、「逸脱のないこと」そのものが目的になってしまう危険があります。

本来の目的は、逸脱をゼロにすることではなく、**データの真正性と信頼性を守ること**です。

そのためには、SOP を守ることを前提としながらも、実際の現場で問題や非効率が見つかったときには、手順を見直してより良い形に更新していく姿勢も必要です。

この「目的を見失わない柔軟さ」こそ、No More Too Much の考え方に通じます。

もし手順がバラバラで記録があいまいなら、統計という判定員の目は簡単に曇ってしまいます。

小さな逸脱や不揃いが積み重なるだけで、試験薬が「効いたっぽい」のか「効いていると言える」のか――その境目はすぐに見えなくなってしまうのです。

だから SOP は厳しいのです。それは、治験をサイエンスとして成り立たせるための**優しさの裏返し**でもあります。

7.2. 「逸脱」の二つの顔：ランダムとシステマティック

逸脱と聞くと、どんなものでも一律に悪いと思いがちですが、実は性質が少し違います。ひとつは、**ランダムに起きる逸脱**です。たとえば、ある患者さんの来院時間が少しずれたとか、入力が一回だけ抜けてしまったとか。こうしたものは散発的で、群間に偏りがなければ「ばらつきを大きくする」方向に働きます。つまり統計的にはノイズが増えて差が見えにくくなるだけで、結論そのものを大きく曲げる危険は比較的少ないのです。

もうひとつは、システマティックに起きる逸脱です。特定の施設でいつも評価が甘いとか、ある群でだけ欠測が頻発するとか、同じ方向に偏って繰り返し起きるようなケースです。これは統計的にはとても怖い。なぜなら差があるように見せたり、逆に差を打ち消してしまったり、結論そのものをゆがめる「バイアス」になるからです。言い換えると、ランダムな逸脱は「ものが見えにくくなる」だけ。でもシステマティックな逸脱は「ものを間違って見せる」ことにつながります。だから CRA が注意すべきは、システマティックな逸脱の芽を早めに見つけて、きちんと正すことなのです。SOP が厳しく定められているのは、まさにこの“間違わせる敵”からデータを守るためでもあります。治験がバイアスとの闘いという意味がよくわかるのではないでしょうか。

7.3. 欠測と修正：正しい補完と、してはいけない穴埋め

欠測については 6 章でも少し触れましたが、SOP の文脈ではもう少し丁寧に整理しておきましょう。まず大事なのは、欠測を未然に防ぐ工夫です。来院スケジュールを無理のない形に整えたり、入力漏れがすぐ分かるシステムを導入したり。CRA の役割は、そうした仕組みが現場できちんと機能しているかを確認し、ほつれがあれば早めに直すことです。それでも欠測が出てしまうことはあります。そのときに大切なのは、「正しい補完」と「してはいけない穴埋め」をしっかり区別することです。

たとえば、評価は実際に行われていたのに記録が抜けていた場合、評価した医師本人に確認して正しく記録を残す。患者さんのアンケートが抜けていた場合も、プロトコルで認められた方法で患者さん本人に確認して追加入力する。こうしたやり直しや再確認は、真正なデータ収集の一部であり、正しい補完です。

一方で、想像で値を入れたり、過去のデータから「たぶんこれだろう」と埋めたりするのは、データを“つくりかえる”行為です。これは改ざんですので、統計の判定をゆがめてしまう大きなリスクです。こんな CRA はいませんよね。

さて、正しく補完できなければ、そのデータは欠測として記録し、その理由をはっきり残すことが大切です。統計家はその情報をもとに、解析で欠測をどう扱うかを考えます。つまり CRA の役割は、むやみに欠測をなくすことではなく、欠測の理由を透明にし、後から統計が正しく扱えるようにすることなのです。

7.4. “ちゃんとやった”を証明するということ

SOP が厳しいもうひとつの理由は、「ちゃんとやった」ということを**後から証明できるようにするため**です。治験は科学的な証拠を積み上げる営みですから、「そのとき

確かにそうやった」という証拠がなければ、いくら立派な結果でも信じてもらえません。だからこそ、記録が必要なのです。

誰が、いつ、どんな手順で、どんな判断をしたのか。それが紙なら署名や日付として、電子なら監査証跡として残ります。ときには二重チェックや承認フローが求められることもありますが、それは単なる事務作業ではなく、「あとから説明できるようにする」ための科学的な工夫なのです。

ここで大切なのは、記録がきちんと読めること、消えないこと、直したときには直した痕跡が残ること。こうした当たり前のような条件がそろって、初めて「このデータは信じられる」と胸を張れるのです。

CRA の役割は、この“証拠の鎖”に切れ目がないかを点検し、必要があれば現場と一緒に補っていくことです。欠測や逸脱の理由が明確に記録されているか、対応や是正が後から追いかけられるか。ひとつひとつ地道に確認することが、統計が公正に働ける土俵を守ることにつながるのです。

7.5. 統計とのつながり：判定員の目を曇らせない

統計は「偶然のブレ」と戦い、公平な比べっこで効果を見きわめます。SOP を守るということは、統計という判定員の視界をクリアに保つことと同じです。ランダムな逸脱は見えにくくする敵でしたよね。システマティックな逸脱や不適切な穴埋めは間違わせる敵でした。SOP の厳しさは、こうした敵から統計を守り、データにウソをつかせないための「優しさ」なのです。

次の章では、この土俵の上で「何を“効果”と呼ぶのか」をぶらさずに決めるための枠組み――Estimand について考えていきましょう。

第７章理解度セルフチェック（そこしら式 記述テスト）

問 1
「SOP ってどうしてこんなに細かくて厳しいんだろう？　現場ではちょっと窮屈に感じるんだけど……」
そこしら CRA のあなたは、この厳しさの本当の意味をどう理解し、どう説明しますか？

問 2
あるプロトコルで、重要な評価項目への評価方法の説明が誤解を招く表現であることがわかった。その結果、その項目は結果として、全体で同じ方向に偏った評価となってしまった。
あなたはこの問題をどう説明し、どう対応すべきと考えますか？

問 3
「欠測が出たときに、患者さんに後から聞いて記録するのは OK だよね？」
確かにプロトコルで認められた場合には正しい補完になりますが、後から聞くことに関してあなたは疑問を感じています
そこしら CRA のあなたは、あなたは、どのようなリスクや問題点を考えますか。

問 4
「記録って本当に面倒だなあ。いちいち署名や日付を残す必要ある？」
ある治験施設のスタッフがぼやいています。
そこしら CRA のあなたは、どのように説得しますか？

問 5
「SOP って、統計知らなくてもあんまり関係ないよね、統計に関しては別に統計家が作成した計画書があるのだから」新人 CRA にそう聞かれました。
あなたは、どのように説明しますか？

問 6
SOP を守ることは「全部を同じ厳しさで追いかける」ことではありません。そこしら CRA として、**リスクベースドアプローチとして**の視点から「SOP を守る」とはどういうことかを説明してください。

問 7
欠測や逸脱の扱いは Estimand の設計に影響します。そこしら CRA として、**「SOP を守ることが Estimand を守ることにつながる」** とはどういう意味だと考えますか？

8. 「誰に・どう効くか」をちゃんと聞く方法　Estimand の世界へ（導入）

ここまでの章で、統計が「偶然」や「バイアス」と戦いながら“効いていると言える”を見きわめる仕組みを学んできました。

5 章では、評価変数には“主役（主要）・脇役（副次）・身代わり（代替）・安全の柱（安全性）”という役割があることを確認しました。ここからは、その「何を測るのか（評価変数）」を出発点に、誰に・どの状況で・何と比べて・どう要約するのかをひとつの枠組み＝Estimand として描いていきます。

さて、実際の治験ではさらに厄介な問題に直面します。

途中で治験薬をやめてしまった患者さんはどう扱う？
他の薬を使い始めた人のデータは？
亡くなってしまった場合は？

治験のプロトコルに書かれている問いを突き詰めると、それは臨床現場から生まれるクリニカルクエスチョンに行きつきます。「この試験薬は誰に、どう効くのか？」という問いです。ところが、先ほどの「治療の途中で起こる出来事」があると、本当に知りたかったこと＝この試験薬は誰に・どう効くのかという問いが曖昧になってしまうのです。

この課題に対して、国際的に整理された**考え方の枠組み**を提示したのが ICHE9(R1)「Estimand と感度分析に関する補遺」 なのです。つまり Estimand とは、このクリニカルクエスチョンを“統計の言葉でぶれなく表現する”ための道具なのです。

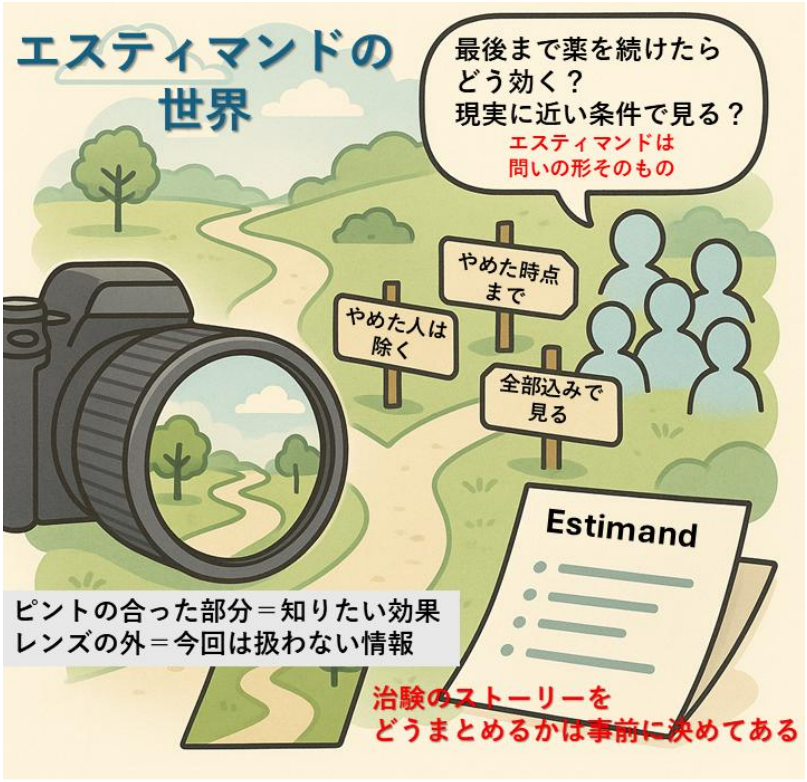
Estimand は、

What	：評価変数（主要／副次／安全性）
Who	：対象集団（誰に）
When/ICE*	：途中で起きる出来事の扱い（介入後事象の取り扱い）
How much	：要約の仕方（差・比・時間まで）

＊ICE：Intercurrent Events

を一本の“問い”として定義する作法です。5 章で説明した「主役／脇役／安全の柱」を土台に、問いの輪郭をイメージしましょう。

さて、「Estimand」という言葉は、皆さんも耳にしたことがあると思います。でも、「あれは統計の問題で、CRA にはあまり関係ないんじゃない？」と思っていませんでしたか？ここまで「そこしら統計」を学んできたあなたなら、もう気づいているはずです。Estimand は「誰に・どう効くか」をちゃんと聞くための方法であり、CRA も理解しておくべき大切な考え方なのです。



さあ、一緒に Estimand の世界をのぞいてみましょう。

8.1. 背景と他のガイドラインとの関係

ICHE9(R1)が出てきた背景と、他のガイドラインとの関係を少し整理してみましょう。

2019 年に正式に合意され、従来の ICHE9（統計的原則）に追加されました。

他の ICH 文書との関係を整理すると、

E6(R3)：GCP の最新の考え方を示し、品質マネジメントの枠組みを導入。

E8(R1)：治験デザインの基本原則を整理し、「Quality by Design」を強調。

E9(R1)：統計の立場から「治験で本当に知りたいことを、言葉でぶれなく定義しよう」と提案。

つまり E6 が「実施の基準」、E8 が「設計の原則」、E9(R1)が「統計的な問いの明確化」と位置づけられ、三つがそろって治験の信頼性を作っているのです。

8.2. CRA も Estimand を理解する意味

皆さんは Estimand は統計の問題なので CRA には直接関係ないと思っていたのではないですか？　たしかに E9(R1)は統計家向けの文書です。でも、CRA が Estimand を理解する意味を、もう一度整理しておきましょう。

コラム：感度分析と Estimand

治験の結論は、「この薬は効いた」と言えるかどうかを統計的に判断することです。でも、その結論は必ず何らかの「仮定」に乗っています。欠測をどう扱うか、途中で薬をやめた人をどう扱うか、あるいは共変量をどう調整するか。もし仮定を変えたら結論も変わってしまうとしたら、その試験の信頼性は揺らいでしまいます。そこで必要になるのが「感度分析」です。

感度分析とは、解析のやり方や前提条件を変えても結論が大きく揺れないかを確かめることです。FAS（ITT 原則）と PPS で同じ方向の結論が得られるか、欠測値の処理を変えても結果が安定しているか、モデルの仮定を変えても同じ結論が導けるか。複数の方法で確かめて、一貫性があるほど結論の信頼性は高まります。

Estimand の考え方では、この感度分析が特に重視されています。なぜなら、Estimand は「本当に知りたい問い」を正確に言葉にする枠組みだからです。例えば「途中で治療をやめた人を含めた効果を知りたい」のか、「きちんと最後まで治療を続けた人だけの効果を知りたい」のか。Estimand で定義した問いに対して、解析がどの仮定に依存しているかを示すのが感度分析なのです。もし感度分析で結論が揺らぐなら、「この試験の答えは仮定しだい」と正直に伝える必要があります。

CRA にとって解析そのものは専門外ですが、感度分析の意味を知っておくことには大きな価値があります。欠測や逸脱、途中の出来事（Intercurrent Events）が正しく記録されていなければ、感度分析は成り立ちません。つまり、CRA が現場で「なぜ欠測が起きたのか」「どういう経緯で治療をやめたのか」を丁寧に記録することが、感度分析を支える土台になるのです。

- **感度分析とは「結論はどこまで揺るがないか」を確かめる道具。**
- **そして Estimand では、その揺らぎの有無が問いの信頼性を決める。**
- **CRA は解析に関与しないが、出来事を正しく残すことが感度分析の生命線になる。**

CRA にとっても理解しておくことが大切なのは、Estimand が『何を知りたいのか』をぶらさずに表現する枠組みだからです。たとえば、Estimand では、プロトコルの規定によっては脱落した被験者のその後のデータを追いかける必要が出てきます。もし CRA がその意味を理解していなければ、「なぜこのデータを追いかけなければならないの？」となり、現場対応がちぐはぐになり、適切なモニタリングができなくなります。つまり Estimand を理解することで、CRA は自分の役割が「統計家のためにデータをそろえる」のではなく、治験全体の問いを守るためにデータの真正性を支えることだと実感できるのです。「CRA はルールの理由を理解し、統計が信じられる現場を守る“番人”」であるということをもう一度思い出してください。

8.3. そこしら CRA 向けの Estimand の 5 つの要点

ここでは、難解な ICHE9（R1）をそこしら CRA 向けに 5 つにまとめました。ちょっと難しいですが頑張って理解してください。

① Estimand とは「知りたいこと」を言葉で描いたもの
Estimand という言葉は「to estimate（推定する）」から派生した造語です。治験の目的を一言で「効いたかどうか」とせず、誰に・どの試験薬を比べて・どんな条件で・どの変数を・どう要約するのかを丁寧に言葉にする枠組みです。「この治験は、誰に対して何を明らかにしたいのか？」を描く道具だと考えてください。

② Intercurrent Events（途中の出来事：中間事象）の扱いを決める
実際の治験では、治験薬をやめる人や他の治療を始める人が必ず出てきます。これらをどう扱うかをあらかじめ決めておくのが Estimand の大きな役割です。「全部込みで効果を知りたいのか」「治験薬を続けられた人だけで見たいのか」、その選び方ひとつで結論は変わります。

③ 解析の計画は Estimand に沿って作られる

統計解析計画（SAP）は、Estimand で定義された「知りたいこと」に答えるための設計図です。Estimand があいまいだと解析もぶれてしまいます。つまりプロトコルに書かれた Estimand＝治験の問いそのもの。CRA はこの問いがぶれないよう、現場データを守るのです。

④ Sensitivity Analysis（感度分析）の必要性
欠測や途中の出来事を扱うときには、必ず『こうだったはずだ』という仮定が入ります。仮定の置き方が違っても結論が変わらないかを確かめるのが感度分析です。CRA の皆さんが手法を詳しく理解する必要はありませんが、「結論は一定の仮定のもとで成り立っている」ことを意識しておくのは重要です。

⑤ CRA の現場の役割は「Estimand を崩さないこと」
データの真正性が崩れれば、どんなに立派な解析計画も意味をなしません。欠測や逸脱が増えれば、Estimand で定義した「知りたいこと」に答えられなくなるのです。だから CRA の仕事は、統計家のための準備ではなく、治験全体の問い＝Estimand を守るための仕事なのです。ただし、ここでも大切なのは優先順位です。すべての欠測や逸脱を同じように扱う必要はありません。CTQ に関わる部分を最優先に確実に守り、それ以外は現場の医療者の適切な対応を信頼する。これが No More Too Much といった考え方に沿った CRA の役割です。

8.4. CRA のための「国際的に求められる Oversight の範囲」の理解

「Oversight」とは、治験全体の品質を確保するために、依頼者が実施状況を適切に監督することを指します。国際的に求められている Oversight は、「全てを確認する管理」から「重要な部分に焦点を当てる監督」への転換です。

ICH E6(R3)では、治験の品質を左右する要因を明確にし、それに基づいて監督・モニタリングの重点を定めるよう求めています。特に第 2 章（2.1〜2.4 Quality Management）では、品質管理を「プロトコルや SOP の遵守を機械的に追うことではなく、試験の目的（Estimand）を達成する上で重要な要素に焦点を当てること」と定義しています。

この考え方は、リスクベースド・クオリティマネジメント（RBQM）とも呼ばれ、すべてを細かく点検する従来型の監視（100%SDV）から脱却し、リスクに応じた Oversight が推奨されています。

CRA にとっての Oversight の範囲とは、

（1）被験者の安全とデータの信頼性に関わる部分（CtQ）を優先的に確認し、
（2）手順的な軽微な逸脱には柔軟に対応し、
（3）現場の医療判断を尊重しながら、全体の傾向やリスクを俯瞰的に把握する、

という三つの柱で構成されます。

つまり、「すべてを見る」ことよりも、「重要なことを正しい深さで見る」ことが国際標準です。

また、EMA の Reflection Paper（2013）や FDA の RBM ガイダンス（2019）も、いずれも「リスクに基づくモニタリングは、品質の確保と効率の両立を目的とする」と明言しており、逸脱の有無よりも、逸脱が治験の目的（Estimand）に影響するかどうかを判断基準としています。

このように、国際的に求められる Oversight の範囲は、「形式的な遵守管理」ではなく、リスクと品質のバランスをとりながら、考えて監督する文化（Thinking Oversight）へと進化しているのです。

CRA に求められるのは、チェックリストに沿う作業ではなく、リスクを見極め、治験の本質を守る判断者としての監督の目です。国際標準的な Oversight より過剰なチェックを実施すればバイアス要因となりうるということも十分に理解しておく必要があります。

8.5. Estimand は始まったばかり

さて、最後に大切なことを強調しておきましょう。

Estimand は「estimate（推定する）」から生まれた新しい造語であるということをお話ししましたが、国際的に使われ始めてまだ日が浅い考え方です。具体的にどう現場に実装し、どう運用していくかは、これから試行錯誤を重ねていく課題がたくさん残されています。そして Estimand は非常に難しい概念ですから、本書では、“そこを知らなければプロトコルを理解できない”という観点から、Estimand の概略だけを紹介しました。

CRA に求められるのは、統計的な専門用語を覚えることではなく、「治験は何を知りたいのか」という問いをぶらさずに守ることが自分の役割だと理解することです。

以上

第８章理解度セルフチェック（そこしら式 記述テスト）

問１
「Estimand って、新しく出てきた難しい言葉だなあ。結局、いったい何を表しているんだろう？」
そこしら CRA のあなたは、自分なりにどう説明しますか？

問２
「治験中に治験薬をやめたり、他の薬を使い始めたり、亡くなってしまったりする患者さんがいる。こういう途中の出来事はどう扱うべきなんだろう？」
あなたは、Estimand の考え方を踏まえて、どんな整理をしますか？

問３
「統計解析計画（SAP）は統計家が作るものだから、CRA には関係ないんじゃない？」
あなたは、Estimand と SAP の関係をどう理解し、CRA としてどんな役割を果たすべきだと考えますか？

問４
「欠測や脱落の扱いは、仮定の置き方次第で結論が変わるらしい。
もしそうなら、有意差ってどこまで信用できるんだろう？」
この悩みについて、感度分析の意味を踏まえて自分なりの考えを示してください。

問５
「CRA の役割はモニタリングでデータをチェックすること。Estimand なんて統計の専門概念は関係ないよね？」
そこしら CRA のあなたは、この言葉にどう反論しますか？
Estimand を崩さないという視点から、自分の役割を説明してみましょう。

問６
Estimand は「治験で本当に知りたいこと」を統計の言葉でぶれなく描く枠組みです。
そこしら CRA として、Estimand を理解することが CTQ を守ることにつながるとはどういう意味だと思いますか？

問７
ある症例で欠測や逸脱が発生しました。もし CRA が安易に「後から埋めればいい」と考えたら、Estimand は崩れてしまうかもしれません。そこしら CRA として、自分の対応が Estimand を守る／壊すとはどういうことか、説明してみてください。

おわりに　統計がちょっと味方に思えてきましたか？

ここまで長い道のりをご一緒いただき、ありがとうございました。
「治験はサイエンス」という出発点に立ち、そこから、ばらつきや偶然、バイアスと戦いながら、統計がどのように効き目を見きわめるかを学んできました。SOP の厳しさの意味を理解し、さらに Estimand の世界も少しのぞきました。

ふり返ってみれば、統計はいつも私たちのそばにありましたね。偶然やバイアスを見きわめる判定員であり、データに正直であるためのルールの支えでもありました。最初は冷たく見えた数字が、実は「人の声」を集めて物語に変える役割を果たしている、そう感じていただけたのではないでしょうか。

そこしら CRA へのメッセージです。

SOP を守るだけでなく、「なぜこのルールがあるのか」を理解して行動する――それが、そこしら CRA の第一歩です。これはまさに E6（R3）が目指す方向性であり、No More Too Much の精神とも通じます。すべてを同じ厳しさで点検するのではなく、CTQ を優先し、現場を信頼しながら効率的にモニタリングすることが求められています。

統計は冷たい数字に見えるかもしれませんが、実際にはひとりひとりの患者さんの声を集め、物語に変える翻訳者です。その物語を歪めないために、CRA が現場で守るのは「数字」ではなく「声」なのです。

Estimand の考え方に触れ、感度分析の意味を知った今、あなたはもう気づいているはずです。CRA の役割は統計家のためにデータを整えることではなく、治験全体の問いを守ること。その問いがぶれないように現場を支えることです。

今日からちょっとだけ、「そこしら CRA」になってみませんか。意味を理解して責任をもって行動する、その小さな一歩が治験の信頼性を守り、未来の患者さんへとつながるのです。

補足：そこしら統計における ICHE9 の扱い方

ICHE9 で CRA が理解しておくべき項目

そこしら統計で取り上げる ICHE9 の項目は、CRA がプロトコルを理解するうえで不可欠なものに限定している。具体的には、**無作為化・盲検化・評価変数・解析集団（FAS、PPS など）・欠測データ・推測の枠組み（有意差、信頼区間など）、そして Estimand の基本的な考え方**である。これらはモニタリング業務に直結し、CRA がプロトコルの意図を正しく読み取り、逸脱や欠測への対応を判断する際に欠かせない知識である。

CRA が理解すべき統計的姿勢（ICHE9 との関連）

治験の結果は、一つの p 値だけで判断できるものではない。この基本的な姿勢を理解しておくことは CRA にとっても重要である。その背景には多重性や感度分析といった統計家の専門領域が関わっており、詳細は統計家に委ねればよい。ただし、**「なぜ一つの p 値だけで結論できないのか」**という考え方を押さえておくことで、誤解や不適切な説明を避けることができる。そこしら統計では、多重性についてはコラムで触れるにとどめ、CRA 教育の範囲を超える詳細な解説は扱わない。

そこしら統計では扱わない ICHE9 の項目

一方で、ICHE9 で示されている統計解析の専門的な理論や設計・実装に関する詳細は、CRA の直接的な役割には含まれないため、そこしら統計では扱わない。たとえば、**検定手法の数理的背景、複雑な多重性調整法（Bonferroni 法や階層的検定手順の詳細）、サンプルサイズ算出の数式展開、ベイズ推論やモデリングの技術的側面**などである。これらは統計家の専門領域であり、CRA が深く習得する必要はない。ただし、その基本的な考え方や結果に及ぼす影響の方向性を理解しておくことは望ましいため、必要に応じてコラム等で補足する。

用語索引

氏名仮名順

そこしら統計制作チーム（制作・査読担当）				
No More Too Much　TF	QbD/CTQ検討班	井上和紀	エイツーヘルスケア株式会社	リーダー
		織田 勉	ICON クリニカルリサーチ合同会社	サブリーダー
		笹浪 和秀	シミック株式会社	
		内記 真木	富士通株式会社	
		小峰 知子	IQVIAサービシーズ ジャパン合同会社	サブリーダー
		渡辺 敏彦	日本CRO協会	リーダー
データサイエンスWG	データサイエンティスト育成検討チーム	乙黒 俊也	ClinChoice株式会社	DSWGリーダー
		隈本 滉一朗	シミック株式会社	リーダー
		設樂 俊也	IQVIAサービシーズ ジャパン合同会社	
		永井 玲奈	クリンクラウド株式会社	
		中易 知大	イーピーエス株式会社	
		三島 直	株式会社アスパークメディカル	
	Estimand検討チーム	大山 暁史	イーピーエス株式会社	
		古賀 優一	エイツーヘルスケア株式会社	
		野川 中	株式会社インテリム	
		藤川 桂	株式会社インテリム	
		室永 遼太郎	シミック株式会社	リーダー
モニタリングWG		小松 高幸	シミック株式会社	リーダー
		松原 一美	シミック株式会社	