令和7年度 厚生労働省 臨床研究総合促進事業
臨床研究・治験従事者研修

2025年10月18日

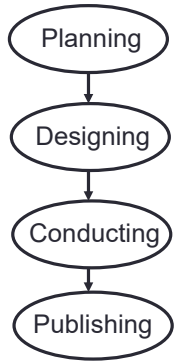
# 生物統計学
# 試験デザイン・サンプルサイズ設計

東京大学大学院医学系研究科
公共健康医学専攻　生物統計学分野
松山　裕

1

## 臨床研究に必要なもの

- 研究仮説：Unmet Medical Needs
- 研究計画書（プロトコル）
  - 診断・評価の基準
  - 疾患の定義、有効性・安全性の評価基準
- 実施システム
  - スポンサー、施設（IRBや事務局）
    およびサポート体制（CRO、SMO）
- 統計家のインプットと解析計画書（SAP）
- CRF（調査票）とその標準化
- データマネージメントのシステム
- 品質保証のシステム
  - モニタリングと監査




2

2

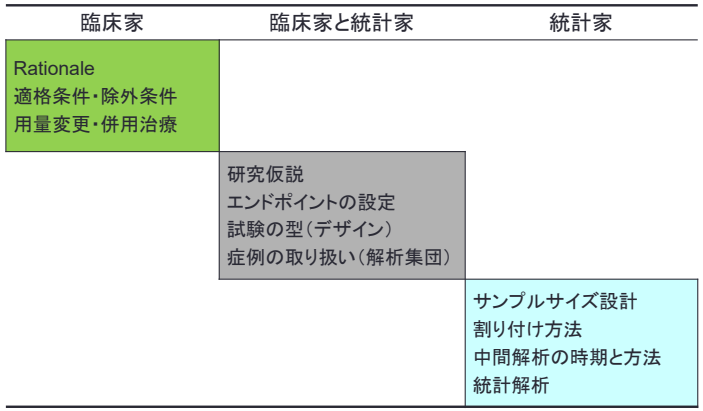
## プロトコルの章構成と内容

| 章 | 内容 |
|---|---|
| 試験の意義 | 背景、目的 |
| 患者選択と登録 | 適格条件、登録手順 |
| 治療と有害事象 | 治療計画、予想される有害事象、治療変更規準 |
| 評価 | 臨床評価項目、臨床検査、効果判定の方法 |
| 解析とデータ管理 | 調査票、エンドポイントの定義、統計解析の計画 |
| 品質管理と品質保証 | モニタリング、中央判定（施設外判定）、監査 |
| 倫理や規制要件 | インフォームドコンセント、有害事象の報告 |
| 管理責任体制 | 研究組織、結果公表の方法 |

3

3

## 臨床家と統計家の協力



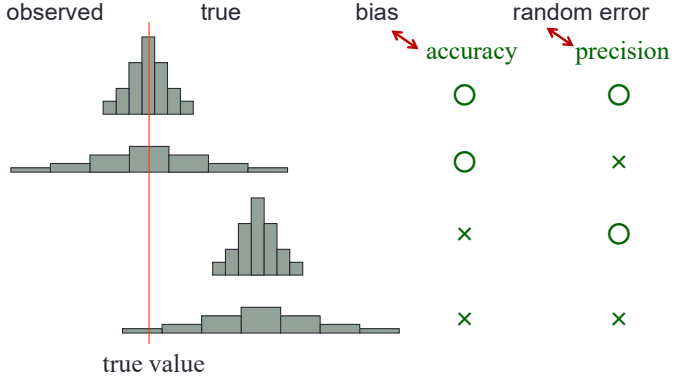

4

4

## 精密度と正確度 Precision and Accuracy

観測値 ＝ 真の値 ＋ バイアス ＋ 誤差的バラツキ

observed　true　bias　random error




5

5

## なぜバラツキ?

進行・再発大腸癌患者に対する5FUの成績

| Authors | # of patients | response (%) 奏効割合（CR+PR） |
|---|---|---|
| Sharp and Benefeil | 13 | 85 |
| Hall and Good | 19 | 63 |
| Rochlin et al. | 47 | 55 |
| Allaire et al. | 17 | 47 |
| Cornell et al. | 13 | 46 |
| Every | 12 | 41 |
| Field | 37 | 41 |
| Bell | 22 | 36 |
| Weiss and Jackson | 37 | 35 |
| Ferguson and Humphrey | 12 | 33 |
| Hurley | 150 | 31 |
| ECOG | 48 | 26 |
| Talley | 271 | 21 |
| Hyman et al. | 30 | 20 |
| Moore et al. | 80 | 19 |
| Ansfield | 141 | 17 |
| Mayo | 358 | 17 |
| Ellison | 87 | 12 |
| Kennedy | 22 | 9 |
| Knoepp et al. | 11 | 9 |
| Olson and Green | 12 | 8 |

8%から85%まで結果がばらついている

6

6

## バイアスと誤差によるバラツキ

- さまざまな理由による「バイアス」
  - 例えば、進行癌に対するある化学療法の奏効率
    - 患者全身状態（PS）とADMEの違い
    - 腫瘍の性質（病理・分子生物学的）
    - dose-intensityとコース、評価部位と評価方法、出版バイアス、…
    - 施設差
- モデルとしての「誤差的バラツキ」
  - バラツキの理由は同定できない、あるいはあえてしない
  - 確率変数としてのモデル化：確率論の応用が可能となる
- 誤差的バラツキとバイアスの相対性
  - 知識が深まれば / 情報が得られれば、誤差的バラツキは制御可能なバイアス要因に転化

7

7

## 医学研究デザインで必要なこと

- 誤差的バラツキを小さくすること（精度を高くすること）
  - サンプルサイズを増やす
  - 測定の精度を上げる
- 偏り(バイアス)を小さくすること
  - デザイン上の工夫が必要（サンプルサイズを増やしても減少しない）
  - その中で最強の方法が、ランダム化（randomization）
  - 統計解析でも制御可能
- そして、得られた結論の一般化可能性を高めること

8

8

## エンドポイントの型による統計手法の分類

| 解析目的 | エンドポイントの統計的型 | | |
|---|---|---|---|
| | 二値 | 連続 | 生存時間 |
| 分布の記述 | 頻度集計<br>分割表 | ヒストグラム<br>平均·SD、相関 | 生存曲線<br>Kaplan-Meier法 |
| 単純な<br>群比較 | $\chi^2$検定、<br>リスク推定 | $t$ 検定<br>平均値の推定 | Log-rank検定<br>ハザードの推定 |
| 層別解析 | MH法<br>標準化 | 分散分析 | 層別 log-rank |
| 回帰モデル | Logistic回帰 | 分散分析<br>重回帰分析 | Cox回帰 |

層別解析、回帰モデル：バイアス補正のための方法





## サンプルサイズ設計

- 臨床試験は実験
  - 研究者が考えている仮説
    - 正しい　：肯定的な結論
    - 誤り　：否定的な結論
- 灰色の結論
  - 肯定的な結論も否定的な結論も得られない
  - 参加者の協力をむだに…（倫理的に問題）
  - サンプルサイズの問題





## 統計的有意差と臨床的有意差

- 対象者数が非常に多い場合
  - わずかな治療効果（差）であっても統計的に有意となる
  - わずかな差が臨床的に重要？
- 対象者数が少ない場合
  - （非常に）大きな差であっても統計的有意とはならない可能性あり
  - 臨床的に重要な差であれば、無視すべきでなく、更なる検討の対象?
- 臨床的に意味のある差を検出し得るようなデザイン
  - そのためのサンプルサイズ設計



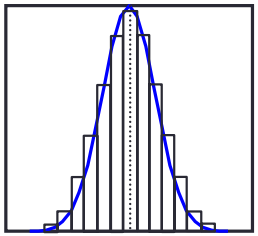


## サンプルサイズの設計法

- 仮説検定ベースで考えるのが普通
- 統計的仮説検定では、
  - データから否定したいための「帰無仮説」を考える
    - 「新治療と標準治療の結果は等しい」
    - 例えば、2群でのリスクの差がゼロ（$P_1 - P_0 = 0$）
  - この仮説が正しいと仮定すると
    - 群間差（リスク差）はゼロを中心に分布する（ランダムにバラツク）
  - 観察されたリスク差が
    - ゼロから大きくはずれていたら、帰無仮説が間違っていると考える

## 「帰無仮説」が正しいとき

- 観察された「群間差」はゼロを中心に左右対称にバラツクはず
  - 研究を繰り返さなくても結果の分布は予測できる：中心極限定理



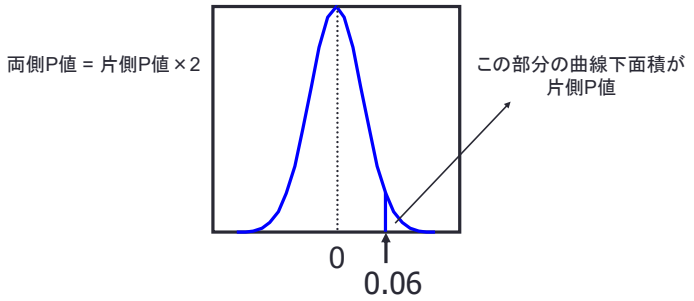

仮想的に同じ規模の研究を繰り返したときの群間差の分布

13

13

## 「帰無仮説」が正しいとき

- このとき、観察された群間差が、ゼロから十分離れていたら、帰無仮説が間違っているのでは?
  - ある人数で研究を行った結果の観察された群間差を 0.06（6%）とする




14

14

## 統計的仮説検定における2つの誤り

- 第1種の過誤（αエラー）
  - 帰無仮説が正しいのに、誤って棄却する（偽陽性）
- 第2種の過誤（βエラー）
  - 帰無仮説が間違っているのに、誤って棄却できない（偽陰性）

- 統計的仮説検定では、αエラーを5%以下に抑えて、できる限り検出力（Power）を高くする
  - 検出力：1−β
    - 帰無仮説が間違っているときに、有意差がつく確率
    - 通常は、80%以上

15

15

## 統計的仮説検定では、

- サンプルサイズ
  - 以下の3つが決まれば事前に算出可能
- αエラー
  - 両側5%が慣習
- βエラー（あるいは、検出力）
  - 検出力80%以上が慣習
- 群間差
  - 医学的に意味のある「最小値」
  - 期待したい値 ??

16

16

## 2グループの割合の差

- 1グループに必要なサンプルサイズ N

$$N = \left[ \frac{Z_\alpha \sqrt{2P(1-P)} + Z_\beta \sqrt{P_T(1-P_T) + P_C(1-P_C)}}{P_T - P_C} \right]^2$$

- $P_T$：試験治療でのイベント発生割合
- $P_C$：コントロール治療での発生割合
- $P=(P_T+P_C)/2$

17

17

## αエラーと$Z_\alpha$

| α | $Z_\alpha$ 片側検定 | $Z_\alpha$ 両側検定 |
|---|---|---|
| 0.10 | 1.282 | 1.645 |
| 0.05 | 1.645 | 1.960 |
| 0.025 | 1.960 | 2.240 |
| 0.01 | 2.326 | 2.576 |

18

18

## 検出力と$Z_\beta$

| 1-β | $Z_\beta$ |
|---|---|
| 0.50 | 0.00 |
| 0.60 | 0.25 |
| 0.70 | 0.53 |
| 0.80 | 0.84 |
| 0.85 | 1.036 |
| 0.90 | 1.282 |
| 0.95 | 1.645 |
| 0.975 | 1.960 |
| 0.99 | 2.326 |

19

19

## 2群の平均値の差の場合

- 1グループに必要なサンプルサイズ N

$$N = 2(Z_\alpha + Z_\beta)^2 \times \left( \frac{\sigma}{\Delta} \right)^2$$

- Δ：2群の平均値の差
- σ：各群の（共通）標準偏差

20

20

## 2群のハザード比の場合

- 必要イベント数：d
  - HR（Hazard Ratio：ハザード比）
- Freedman公式

$$d = \frac{\{z_\alpha + z_\beta\}^2 (HR+1)^2}{2(HR-1)^2}$$

- Shoenfeld公式

$$d = \frac{\{z_\alpha + z_\beta\}^2 \times 2}{(\log(HR))^2}$$

21

21

## サンプルサイズへの影響度

- 群間差の2乗に反比例
  - 群間差が2倍になれば必要なサンプルサイズは1/4倍
- 標準偏差の2乗に比例
  - バラツキが2倍になれば必要なサンプルサイズは4倍

22

22

## サンプルサイズ設計?

- 型通りに計算できる?
  - 平均値の比較
  - 割合の比較
  - 生存時間の比較
- 群間差の指定?
  - 検証試験でも曖昧なことがある
- 探索的研究?
  - 検証試験を行う前に、少数例で傾向をみる?
  - 5例、10例、15例??

23

23

## 例えば、

- ある疾患に対する新しい診断法を開発した。スクリーニング目的なので、感度（sensitivity）に関心がある。
  - 感度：真に疾患のあるものを疾患ありと判断する確率
- これまでの研究成果（ヒストリカル・コントロール）から、感度が80%あれば十分
  - 最低でも、60%、あるいは70%程度は必要。
  - 何例で行う?

24

24

## 信頼区間ベースでも計算可能

- 例えば、期待する感度を80%、その95%信頼区間の下限を60%とするのであれば、
  - 信頼区間：データと矛盾しない結果の範囲

$$0.8 - 1.96\sqrt{\frac{0.8\times(1-0.8)}{n}} = 0.6$$

  - 上式を n について解けば、n = 15.36.. = 16

25

25

## 観察研究では?

- サンプルサイズ設計
  - ランダム化試験では必須
  - 交絡バイアスの調整（リスク調整）が必要な観察研究では?
    - 交絡バイアスの大きさも事前に見積もる必要があるが、現実的にはその大きさがわからない
- 実際には
  - できる限り多くのサンプル数を集めることになるが…
  - バイアス調整済みの期待したい群間差の大きさを想定して、サンプルサイズ設計

26

26

## サンプルサイズの決め方

- 最大で何名集められるかを確認する
  - 臨床試験なら、参加施設における過去1年間の適格患者数を調査
- コントロールグループと試験治療グループで期待される結果を見積もる
- 見積もりを間違えた場合のシナリオを準備する
  - 感度解析（いくつかの群間差を想定して計算）
- 募集期間、募集人数を決定する

27

27

## 研究仮説は明確?

- 証明すべき仮説は明確に、データ解析によって検証可能な形で述べられているか?
  - 標準治療確立のための大規模イベント試験においては、比較的容易
    - 新規イベント、再発、あるいは死亡までの時間延長
- 優越性試験
  - 新治療が標準治療よりも優っていることを生存時間解析により証明
- 非劣性試験
  - 新治療が標準治療よりも劣っていないことを生存時間解析で証明

28

28

## 優越性試験（Superiority trial）

- 帰無仮説：「試験群Aと対照群Bとの差はない」
  - $H_0: P_A - P_B = 0$
  - 帰無仮説を否定することで、一方の群の他方に対する優越性を証明
- 統計的仮説検定
  - 「有意差あり」＝帰無仮説は間違い（2群で何らかの差がある）
  - 「有意差なし」≠帰無仮説は正しい（2群は同等）

29

29

## 検定結果が有意

- 統計学的に「差がある」といえる
  - Statistically Significant
- 問題は、
  統計的有意な群間差が、臨床的に意味があるかどうか?

30

30

## 有意でなかった場合は…

- 2群は同等と判断してはいけない
- 帰無仮説「2群間で結果は同じ」は否定するためのもの
  - 否定できなければ、（積極的には）何も統計学的にいえない
- さもなければ、
  - いい加減な（人数の少ない）研究を行えば、常に同等の（帰無仮説が正しい）結果が得られることになる!!

31

31

## 非劣性試験（Non-inferiority trial）

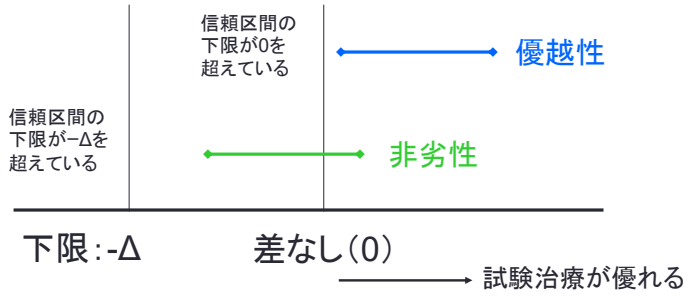
- 適用場面
  - 主たるエンドポイントにおいて、標準治療に対して同等で、かつ薬物有害反応の程度の少なさ・軽さやQOL、投与方法の利便性などの点で優れることが予想される場合
- 臨床的に許容できる差（Δ）の設定
- 帰無仮説：「試験治療 A は標準治療 B より Δ だけ劣る」
  - $H_0: P_A - P_B = -\Delta$
  - この帰無仮説を否定できれば（片側検定）、
    試験治療は著しく（Δ以上）劣ることはない（$P_A > P_B - \Delta$）を証明

32

32

# 優越性試験と非劣性試験

- 治療効果の差（test - control）と95%信頼区間




信頼区間：データと矛盾しない結果の範囲

33

33

# エンドポイント

- 有効性、安全性を評価するための臨床結果
  - 評価項目
- 「真のエンドポイント」と「代替エンドポイント」
  - True　：イベント発症
  - Surrogate　：血液マーカー、検査値など
- 「主要評価項目」と「副次評価項目」
  - Primary　：当該試験において、検出力が確保された関心のある項目（基本的には1つだが、必ずしもそうではない）
  - Secondary：検出力は確保されていないかもしれないが、関心のある項目

34

34

# ARISTOTLE trial (国際共同第III相試験)

## 心房細動症例に対するアピキサバンの有効性と安全性

脳卒中危険因子
- 年齢75歳以上
- 脳卒中、TIA、全身性塞栓症既往
- 心不全または左室駆出率≦40%
- 糖尿病
- 高血圧

二重盲検
ランダム化
ダブルダミー
N = 18,201

主な除外基準
- 人工心臓弁
- 重症腎不全
- アスピリンとチエノピリジンを必要とする症例

アピキサバン 5mg 経口1日2回
（特定患者には2.5mg 1日2回）

ワルファリン
（INR 2～3）

研究仮説
（1）脳卒中または全身性塞栓症における非劣性
（2）脳卒中または全身性塞栓症における優越性
（3）大出血における優越性
（4）全死亡率における優越性

35

Lopes RD, et al.: Am Heart J 2010; 159: 331-9

35

# いくつかのエンドポイント

- 有効性
  - Primary　：脳卒中（虚血性、出血性または特定不能）、または全身性塞栓症の初発までの期間
  - Secondary　：全死亡、心筋梗塞の発症までの期間など
- 安全性
  - Primary　：大出血（ISTH基準）の初発までの期間
  - Secondary　：大出血、または非大出血イベントの複合事象、およびすべての出血イベント
- 症例数設計
  - 2つのPrimary endpointに関する検出力を確保

36

36

## エンドポイントの設定・評価

- 客観性と信頼性
  - 定義・基準が明確か?
  - 評価に施設間差、医師間差はないか?
- PROBE法（Prospective Randomized Open-labeled Blinded Endpoints）
  - エンドポイントの評価を盲検化、中央判定
    - 二重盲検割り付けが困難な場合には標準的評価方法
- 複合エンドポイント
  - イベント数が少ないと予想される場合にはやむを得ない
  - 主観的要素が強いエンドポイントはできる限り含めない

37

37

## ExplanatoryかPragmaticか

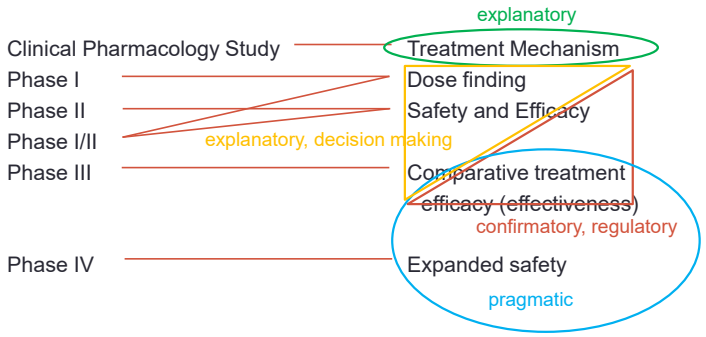
- Explanatoryな（説明的）臨床試験
  - 介入法の作用機序などを解明する目的で、実験条件をある程度厳しく設定して実施する試験
  - 薬剤の有効性（efficacy）を実験室的環境で検証
  - 併用療法・用量変更などは許さない
- Pragmaticな（実践的）臨床試験
  - 実施条件を緩く設定し、日常診療に近い状況で介入法を評価するために実施する試験
  - 実践医療の中での薬剤の有用性（effectiveness）を検証
  - 併用療法・用量変更などは許される（プロトコルで規定）

38

38

## Clinical Trial: Design Types

Piantadosi A (1997), *Clinical Trials*, Wiley




39

39

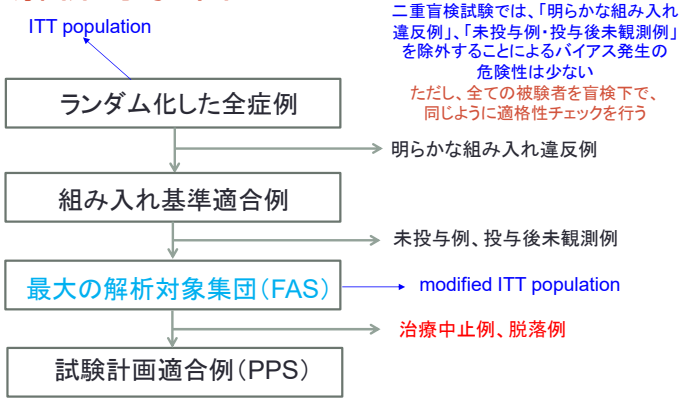
## Pragmatic trialにおける解析方針

- ITT（Intention-to-Treat）原則
  - 割付けた被験者は、たとえ早期中止やコンプライアンス不良でも、その割付け群として解析する方針
    - エンドポイントの測定は全登録例で必須
  - 検証的第3相試験では標準的な解析方針（有効性に関して）
    - ランダム化によって達成された「比較可能性」を被験者除外によって崩さないため
- ITT集団にもいくつかの考え方がある
  - プロトコル、論文ごとに異なる場合もある
- ICH-E9 統計ガイドラインでは、
  - 最大の解析対象集団（FAS：Full analysis set）の定義を要求

40

40

# 解析対象集団




FAS：Full analysis set、PPS：Per-protocol set

41

41

# まとめ

- 臨床研究を行う際には、
  - 生物統計学による「支援」が必須
  - 生物統計家との対話を通じた研究計画立案
- 生物統計家に期待される役割
  - 単なるデータ解析者だけでなく、methodologistとしての参画
- 生物統計学の本質的な目的
  - 研究目的の明確化と研究の効率化を通した実際の研究支援
  - 生物統計＝データ解析では「ない」

42

42