研究デザインと統計解析

科学的合理性の確保と評価

6NOV2023 順天堂大学研修

順天堂大学　健康データサイエンス学部
順天堂大学附属順天堂医院　臨床研究・治験センター

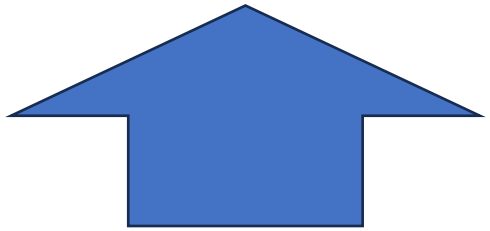
大津　洋

生物統計学とは

- 生物統計学：
  - ✓医学や生物学といった分野において、データの収取や解析、解釈を行うための統計学の応用
  - ✓疫学や臨床試験など、様々な医療・医学・公衆衛生の研究課題に対して、科学的な方法論を提供し、研究の質や信頼性を高める役割を持つ
- 目的: 不確実性を管理し、科学的根拠に基づいた決定をサポート

生物統計学についての誤解

データが決まれば、統計学での結果は同じである。
（誰がやっても、同じ結果になる）



残念ながら、この考え方は間違い

統計解析の結果が異なる原因(一因)

統計解析ソフトウェア/プログラム

- アルゴリズムの違い
- 欠測、外れ値…

手法・オプション

- 例) t-test: 正規性によりStudent/Welch

データの加工

- 変数変換や補完

統計解析の結果が異なる原因(一因)

科学的合理性

評価項目のあいまいさ

科学的合理性の重要性

- 正確な結論の導出
- 効率的なデザインによる資源の節約
- 結果の再現性と一貫性の確保

臨床研究の科学的合理性を見る
[P.E.C.O. / F.I.N.E.R.]

P.E.C.O.

- P: Patient
- E(I) : Exposure(Intervention)
- C:Comparison
- O:Outcome

F.I.N.E.R

- F:Feasible
- I: Interesting
- N:Novel
- E:Ethical
- R:Relevant

“FINER” criteria

- F: 実施可能性（Feasible）
  - **対象症例数**、科学的な裏打ち（レビュー）、資金
- I: 科学的な興味　（Interesting）
  - **サイエンスとしての重要性**
- N: 新規性　（Novel）
  - イノベーティブであるかどうか？
- E: 倫理性　（Ethical）
  - 研究倫理などに反する研究でないか？IRBなどが承認できるレベルか？
- R:必要性　（Relevant）
  - **今、実施する必要性**があるか

[Original Article]

人を対象とする医学研究における重篤な有害事象報告に関する研究

―国際基準からみた統合指針の特徴と多施設共同試験の運用―

■

Adverse Event Reporting in Clinical Investigations ―Comparing Roles and Responsibilities for Multi-center Clinical Trials in the ICH-GCP and the New Japanese Guidance―

■

信濃　裕美[*1]　大津　洋[*2,3]　松岡　浄[*4]　富野康日己[*1]　佐瀬　一洋[*1]

	ICH-E6 Good Clinical Practice	米国 Common Rules	米国 FDA規制	EU Clinical Trial Regulation	臨床試験の実施の基準 (J-GCP)	倫理指針 2003	倫理指針 2008	統合指針 2014	臨床研究法 2017
対象範囲	薬事申請目的とそれ以外の一部	公的助成金による全ての人を対象とする研究	薬事申請目的とそれ以外の一部	臨床試験，低介入臨床試験	企業治験，医師主導治験（薬事申請目的）	臨床研究	臨床研究	臨床研究と疫学研究	特定臨床研究（臨床試験）
根拠法令等	E6(R1) 1997/3/27 Step5 ICH-E6(R2) 2016/11/9 Step4 2017/6/14 Step5 (EMA)	45CFR46 PHS 2017/1/19 全面改訂	21CFR50 PHS 2016/4/1 21CFR56 IRB 2016/4/1 21CFR312 IND 21CFR812 IDE	EU Regulation (No.536/2014) 2014/4/16 EU Directive 2001/20/EC は廃止	薬機法 2014/11/17改正 省令GCP 1997/3/27 2016/1/22改正	厚生労働大臣告示 2003/7/30 2004/12/28改正	厚生労働大臣告示 2008/7/31改正	厚生労働大臣告示 2014/12/22 課長通知（ガイダンス） 2015/3/31改訂	臨床研究法 2017/4/14公布 省令（施行規則） 2018/2/28 課長通知 2018/2/28
規制当局		○計画審査 (FDA, IND/IDE) ○有害事象報告 (serious unexpected)	○計画審査 (FDA, IND/IDE) ○有害事象報告 (serious unexpected)	○計画審査 (MHRA等, CTN) ○有害事象報告 (SUSAR)	○計画審査 (PMDA, 治験届) ○有害事象報告 (SAE)	なし	○有害事象報告 (MHLW, serious unexpected)	○有害事象報告 (MHLW, serious unexpected)	○計画届 (PMDA, 計画) ○疾病等報告 (SUADR)
IRB/EC	○IRB/IEC	○法定IRB 原則として中央IRB	○法定IRB 中央IRBも可	○公的EC (仏CPP 40カ所，英国REC 69カ所)	○法定IRB 中央IRBも可	△IRB (倫理審査委員会，1.3.(10))	△IRB 中央IRB可 倫理審査委員会 (3.(16))	△IRB 中央IRB可 倫理審査委員会 (2.(15))	○法定中央IRB (認定臨床研究審査委員会) 法23
Informed Consent	○	○	○	○	○	○	○	○	○
Initial IRB approval	○	○	○	○	○	○	○	○	○
Continuing Review	○年次，○SUADR	△年次，○USH	○年次，○USH	△年次，○SUSAR	○年次，○SAE	○SAE	○年次，○SAE	○年次，○SAE	○年次，○SUSAR
Sponsor	試験の法的責任主体（スポンサー＋資金提供者）(E6 1.53)		試験全体の責任者 (IND Holder) 21CFR56.102	試験の法的責任主体（スポンサー＋資金提供者）(2.2.(14))	治験依頼者（法80条の2）(治験調整医師（省令17）設置可)				
Sponsor－Investigator	アカデミア試験の責任者（研究者，研究機関，助成機関等）(E6 1.54)		IND Holder（研究者，研究機関，助成機関等）	Sponsor (2.2.(14))	自ら治験を実施する者（法80条の2）(治験調整医師（省令2.16）が代表して届出)			研究代表者 (ガ5-2-11) 統括責任者 (ガ17-2-6)	研究代表医師 (規1.4, 規12.1) 各研究責任医師の共同責任 (通知2.(7)) 研究を統括する者 (通知2.(11)①)
Investigator	Investigator (E6 1.34)		各施設の責任医師 21CFR56.102	Principal investigator (2.2.(16))	治験責任医師 (省令2.3)	研究責任者 (1.3.(5))	研究責任者 (3.(12))	研究責任者 (2.(13))	研究責任医師 (規則1.2)
Sub-Investigator	Subinvestigator (E6 1.56)			Investigator (2.2.(15))	治験分担医師 (省令2.11)	研究者等 (1.3.(4))	研究者等 (3.(11))	研究者等 (2.(12))	研究分担医師 (規則1.5)
Site Administrator	Trial Site (E6 1.59)	公的資金を受ける研究実施機関			実施医療機関の長 (省令2.2)	臨床研究機関の長 (1.3.(10))	組織の代表者等 臨床研究機関の長 (3.(13), 3.(14))	研究機関の長 (2.(14))	実施医療機関の管理者 (規則11.1)
その他	IDMC/DSMC (E6 1.25)	Data and Safety Monitoring Policy: NIH 1998/6/10 NCI 2014/9/30	21CFR314 NDA 42CFR11 CTR 全てのSAEまとめ DMC ucm127073 2006/3/20	CTR (2.2.(25))	効果安全性評価委員会，独立データモニタリング委員会（省令19）(省令26の5)		CTR (2.(5)) 効果安全性評価委員会 (3.(8)細則)	効果安全性評価委員会 (ガ11.2)	CTR (規則24) 効果安全性評価委員会 (Q&A問6-2)

SUADR: Serious Unexpected Adverse Drug Reaction, USH: Unexpected Serious Harm, SUSAR: Suspected Unexpected Serious Adverse Reaction, SAE: Serious Adverse Event

効果安全性評価委員会/
独立データモニタリング委員会

ICH-GCP

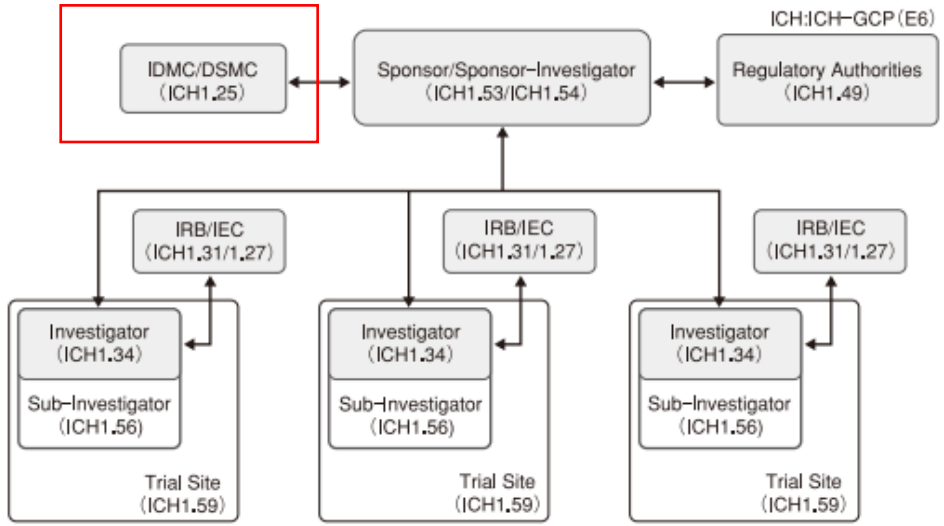
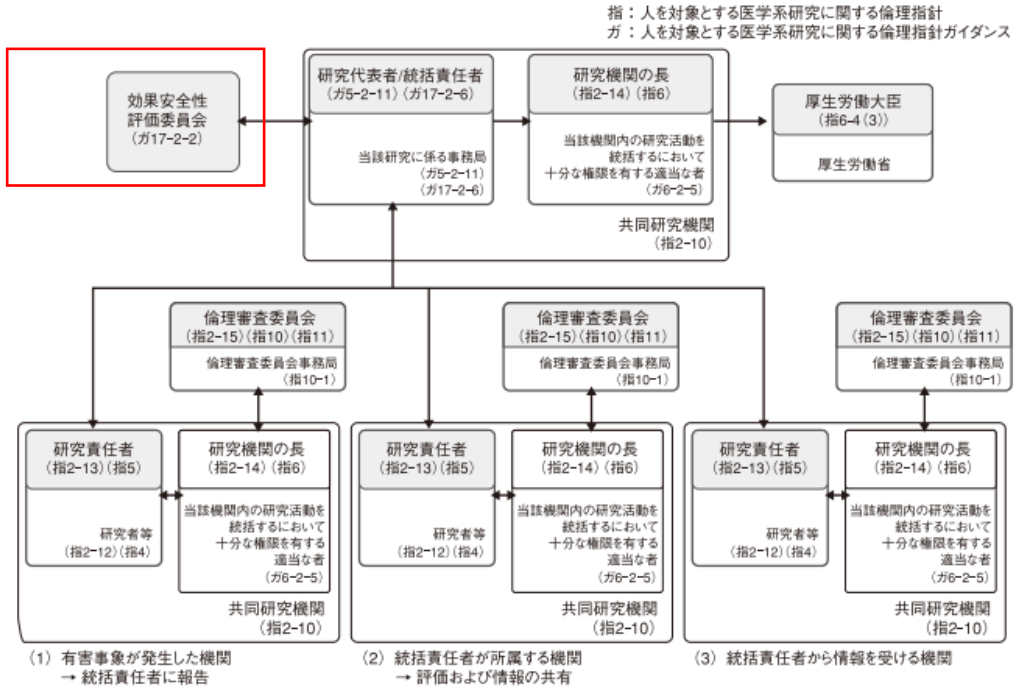


図 1　ICH-GCP における重篤な有害事象報告の流れ

共同研究における重篤な有害事象の報告（指針）



- 中間解析(Interim analysis)
  - 試験の正式な完了以前に，有効性または安全性に関して試験治療群間を比較することを意図して行われるあらゆる解析

SWOG

- South West Oncology Group (SWOG)
  - 1985年以前は、公式な試験中止規準も、モニタリグ委員会もなし
  - 当時は、毎年定期的にデータの開示があった
    - そうしたらどうなったか….
    - 14trial のうち
      - 5試験で途中の患者登録が鈍る
      - 2試験で目標参加者数を達成できなかった

運営側は公正な判断ができない
(Green,Benidetti and Crowley, 2004)

途中でデータを見るとどうなるか？

- 「当事者」が情報を見ると、その後の判断が狂う可能性がある。
  - 「効果がありそう」だったら、試験にエントリーではなく実臨床で使った方がよいだろう。
  - 「効果がなさそう」だったら、試験に対するモチベーションが下がる

当時者でなければ問題ないのか？

- FDA Advisory Committee　の事例
  - 2005年 Cox-2 阻害剤
    - 心血管リスクが問題、Cox-2阻害剤の安全性について議論
    - 撤退させるか否か？
      - Bextra,Celebrex(Phizer)
      - Vioxx(Merck)

投票結果はどうだったか？ (Steinbrook,2005)

- Bextra の取り下げ
  - 賛成　13 : 17 反対
  - コンサルティング・講演・研究資金を得ていた10名では、　賛成 1: 9 反対

FDAは取り下げにできず、自主回収とした

- VioXXは撤退支持
  - 賛成 15: 17 反対
  - コンサルティング・講演・研究資金を得ていた10名では、　賛成 1: 9 反対

僅差だったので、撤退することをFDAは選択

統計手法として考える中間解析

- まず、多くの研究で「有意水準を5%」とする、という前提で動いている

検定の多重性

- 事例
- A,B,C の 3 条件で得られた 3 つの平均値を比較
  - 比較パターン：A-B　B-C　C-A

この時、3つの比較がひとつでも間違っている確率はどの程度でしょうか？

答え：　$1-0.95^3 \fallingdotseq 0.142$

一般論として

- 「A と B との間に有意差が認められたが、B と C、および C と A との間に有意差は認められなかった」
  というような、複合的な結論を　有意水準5%で主張することは難しい。

  - そこで、様々な手法が提案されている。

中間解析が抱える問題

- 中間解析を行うことによる、質の変化の問題
- 利益相反の問題
- 統計手法としての、検定の多重性の問題

中間解析を行うに当たって

- 中間解析を行うことによる、質の変化の問題
- 利益相反の問題
  - 中間解析を行う体制の問題
    - ICH-GCP
    - JCOG Policy paper etc.
- 統計手法としての、検定の多重性の問題
  - 多くの手法が提案されている

統計家だって

- 中間解析を伴う場合は、試験統計家と独立した生物統計家を用意しておく

- ucm127073.pdf　(FDA ガイダンス)
  - その試験の決定や修正には加われない
  - スポンサーの統計家はダメ

生物統計学の限界

- データの品質に依存
- 統計的有意性 ≠ 臨床的有意性
- 複雑なモデルの過剰適合のリスク

検定手法に頼ることが必要なのか？

- そもそも統計手法のp値が伝えることは何か？を知る必要がある

- 検定で用いるp値は
  - 実験の結果から
  - 一般集団（母集団）での姿を
  - 比較すること
    - つまり、結果は一般化できるかどうか？

統計的

The American Statist

ISSN: 0003-1305 (Print) 1537-2731 (Onli

The ASA's Statement and Purpose

Ronald L. Wasserstein & Nicole

To cite this article: Ronald L. Wasserst
on *p*-Values: Context, Process, and Purp
10.1080/00031305.2016.1154108

To link to this article: https://doi.org/1

2023/10/26

ASA Statement on Statistical Significance and *P*-Values

1. Introduction

Increased quantification of scientific research and a proliferation of large, complex datasets in recent years have expanded the scope of applications of statistical methods. This has created new avenues for scientific progress, but it also brings concerns about conclusions drawn from research data. The validity of scientific conclusions, including their reproducibility, depends on more than the statistical methods themselves. Appropriately chosen techniques, properly conducted analyses and correct interpretation of statistical results also play a key role in ensuring that conclusions are sound and that uncertainty surrounding them is represented properly.

Underpinning many published scientific conclusions is the concept of "statistical significance," typically assessed with an index called the *p*-value. While the *p*-value can be a useful statistical measure, it is commonly misused and misinterpreted. This has led to some scientific journals discouraging the use of *p*-values, and some scientists and statisticians recommending their abandonment, with some arguments essentially unchanged since *p*-values were first introduced.

In this context, the American Statistical Association (ASA) believes that the scientific community could benefit from a formal statement clarifying several widely agreed upon principles underlying the proper use and interpretation of the *p*-value. The issues touched on here affect not only research, but research funding, journal practices, career advancement, scientific education, public policy, journalism, and law. This statement does not seek to resolve all the issues relating to sound statistical practice, nor to settle foundational controversies. Rather, the statement articulates in nontechnical terms a few select principles that could improve the conduct or interpretation of quantitative science, according to widespread consensus in the statistical community.

2. What is a *p*-Value?

Informally, a *p*-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value.

3. Principles

1. ***P*-values can indicate how incompatible the data are with a specified statistical model.**

   A *p*-value provides one approach to summarizing the incompatibility between a particular set of data and a proposed model for the data. The most common context is a model, constructed under a set of assumptions, together with a so-called "null hypothesis." Often the null hypothesis postulates the absence of an effect, such as no difference between two groups, or the absence of a relationship between a factor and an outcome. The smaller the *p*-value, the greater the statistical incompatibility of the data with the null hypothesis, if the underlying assumptions used to calculate the *p*-value hold. This incompatibility can be interpreted as casting doubt on or providing evidence against the null hypothesis or the underlying assumptions.

2. ***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.**

   Researchers often wish to turn a *p*-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The *p*-value is neither. It is a statement about data in relation to a specified hypothetical explanation, and is not a statement about the explanation itself.

3. **Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.**

   Practices that reduce data analysis or scientific inference to mechanical "bright-line" rules (such as "$p < 0.05$") for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision making. A conclusion does not immediately become "true" on one side of the divide and "false" on the other. Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis. Pragmatic considerations often require binary, "yes-no" decisions, but this does not mean that *p*-values alone can ensure that a decision is correct or incorrect. The widespread use of "statistical significance" (generally interpreted as "$p \leq 0.05$") as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.

4. **Proper inference requires full reporting and transparency**

   P-values and related analyses should not be reported selectively. Conducting multiple analyses of the data and reporting only those with certain *p*-values (typically those passing a significance threshold) renders the reported *p*-values essentially uninterpretable. Cherry-picking promising findings, also known by such terms as data dredging, significance chasing, significance questing, selective inference, and "*p*-hacking," leads to a spurious excess of statistically significant results in the published literature and should be vigorously avoided. One need not formally carry out multiple statistical tests for this problem to arise: Whenever a researcher chooses what to present based on statistical results, valid interpretation of those results is severely compromised if the reader is not informed of the choice and its basis. Researchers should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted, and all *p*-values computed. Valid scientific conclusions based on *p*-values and related statistics cannot be drawn without at least knowing how many and which analyses were conducted, and how those analyses (including *p*-values) were selected for reporting.

5. **A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.**

   Statistical significance is not equivalent to scientific, human, or economic significance. Smaller *p*-values do not necessarily imply the presence of larger or more important effects, and larger *p*-values do not imply a lack of importance or even lack of effect. Any effect, no matter how tiny, can produce a small *p*-value if the sample size or measurement precision is high enough, and large effects may produce unimpressive *p*-values if the sample size is small or measurements are imprecise. Similarly, identical estimated effects will have different *p*-values if the precision of the estimates differs.

6. **By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.**

   Researchers should recognize that a *p*-value without context or other evidence provides limited information. For example, a *p*-value near 0.05 taken by itself offers only weak evidence against the null hypothesis. Likewise, a relatively large *p*-value does not imply evidence in favor of the null hypothesis; many other hypotheses may be equally or more consistent with the observed data. For these reasons, data analysis should not end with the calculation of a *p*-value when other approaches are appropriate and feasible.

P値の適正な使用と解釈に関する６つの原則

佐藤俊哉2017 計量生物学Vol38.No.2 109-115

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ひとつ）が矛盾する程度をしめす指標のひとつである.
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.
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）を超えたかどうかにのみ基づくべきではない.
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.
5. P 値や統計的有意性は，効果の大きさや結果の重要性を意味しない.
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，よい指標とはならない.

生物統計学の限界

- データの品質に依存
- 統計的有意性 ≠ 臨床的有意性
- 複雑なモデルの過剰適合のリスク

これらの限界を分かったうえで、適切な統計手法の選択・報告が必要

Take Home message

- 臨床研究（試験）の研究計画を科学的合理性を見るための指標としてPE(I)CO,FINER　を紹介した。
  - 症例数の設定も、問題設定に応じて必要な症例数が必要。
  - 症例数ありきでなく、FINER criteria の一部として考えるとよい
  - 臨床研究の科学性と倫理性は関連している

- 解析方法が異なれば、同じデータであっても結果（有意or有意でない）の結果は異なることがある
  - きちんと評価できる生物統計家が必要
- IDMCの意義