

臨床研究の立案と 臨床評価のための統計的考え方について

下川敏雄

和歌山県立医科大学附属病院 臨床研究センター

Agenda

- 臨床研究の立案について
- 臨床研究のデザイン

臨床研究の立案について

臨床研究の基本理念 (臨床研究法 第9条)

臨床研究は、臨床研究の対象者の生命、健康及び人権を尊重し、次に掲げる事項を基本理念として実施しなければならない。

1. 社会的及び学術的意義を有する臨床研究を実施すること
2. 臨床研究の分野の特性に応じた科学的合理性を確保すること
3. 臨床研究により得られる利益及び臨床研究の対象者への負担その他の不利益を比較考量すること
4. 独立した公正な立場における審査意見業務を行う認定臨床研究審査委員会の審査を受けていること
5. 臨床研究の対象者への事前の十分な説明を行うとともに、自由な意思に基づく同意を得ること
6. 社会的に特別な配慮を必要とする者について、必要かつ適切な措置を講じること
7. 臨床研究に利用する個人情報 を適正に管理すること
8. 臨床研究の質及び透明性を確保すること

「Emanuelの7要件」^{*1}に則した内容になっている。Emanuelの7要件とは、

1) Social Value (社会的な価値),

2) Scientific Validity (科学的妥当性),

3) Fair Subject Selection (適正な被験者選択),

4) Favorable Risk/ Benefit Ratio (適正なリスクベネフィットバランス)

5) Independent Review (第三者による独立した審査),

6) Informed Consent (インフォームド・コンセント)

7) Respect for Potential and Enrolled Subject (被験者及び候補者の尊重)

である(中村, 2018^{*2})。

^{*1}: Emanuel et al.: What makes clinical research ethical? JAMA, 283, 2701- 2711, 2000.

^{*2}: 中村健一：研究倫理と臨床研究法, 58, 944-947, 2018

臨床研究の出発点：動機と研究目的 (Social valueを明確にする)

研究の背景

- 対象疾患の現況(疫学データを含む)
- 既存の治療法(標準治療)の内容および成績

- 対象疾患は「どのような疾患」であるのか？
- 症状等はどうなのか(予後など)？
- 現在は、どのように治療な治療法があるのか？

標準治療における課題及び不明点

- 現在の治療法にはどのような問題があるのか？

→ この部分が「本試験の動機」に繋がる

本試験治療の概要

- 今回、試験治療として何を選択し、それはどのようなものか？

研究の目的

本試験治療が、標準治療に対して優れていると期待すること(課題設定)が目的として記される。つまり、PICO(Patient, Intervention, Comparison, Outcome)が重要

- (1) 対象疾患のどのような患者に対して(Patient),
- (2) 本試験治療が (Intervention)
- (3) 標準治療に比べて (Comparison)
- (4) 何に優れているか (Outcome)**

について、試験デザインに繋がるように記載する。

研究のデザインの重要性 (Scientific Validityの担保)

偶然誤差
(random error)

偶然によって生じる誤差
ランダムサンプリングによる影響, 個体内変動



標本サイズが影響する
(標本サイズの増加は**定度(精度)**が増す)

系統誤差
(systematic error)

何らかのバイアスによって生じる誤差



研究のデザインが影響する
(適切なデザインは**真度(正確性)**が増す)

バイアスとは, 「偏って得られたデータに基づいて誤った結論を導く」おそれがある。つまり, 不適切な研究のデザインは「間違った知見」を発信する危険性を孕んでいる。

理想的な研究デザイン(Byrne, 2017)

■ バイアスのないサンプル

- サンプルサイズが十分に大きい(統計学的検出力がある).
- 研究テーマにふさわしい均一性がある(介入の多くは高リスクの患者を対象とする場合にのみ有効であるため, そうかもしくは予測モデルを用いて, リスクの高い患者を選別する必要がある).
- (母集団の)代表性が高い
- サンプルが複数の施設(多施設)から得られている.

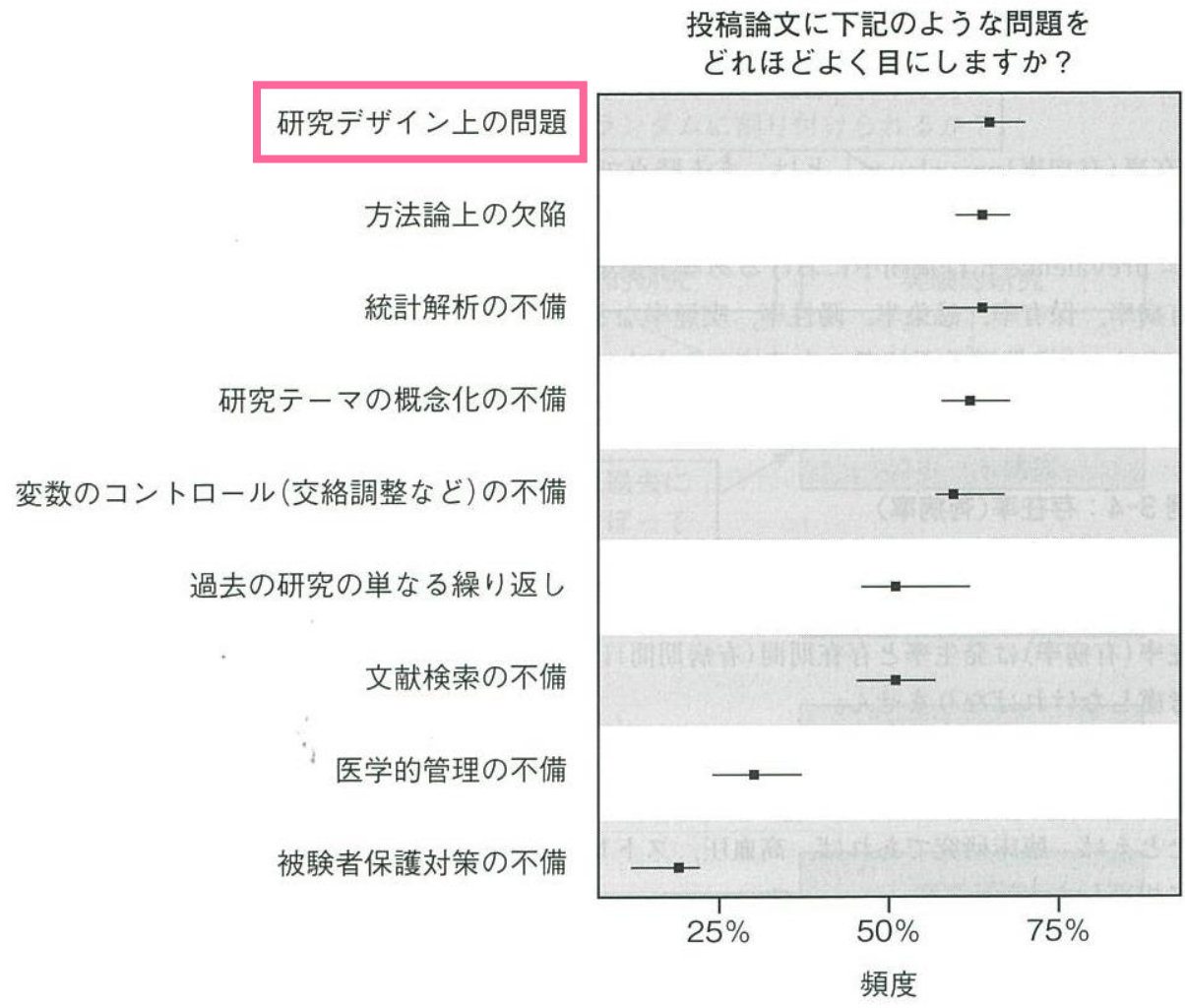
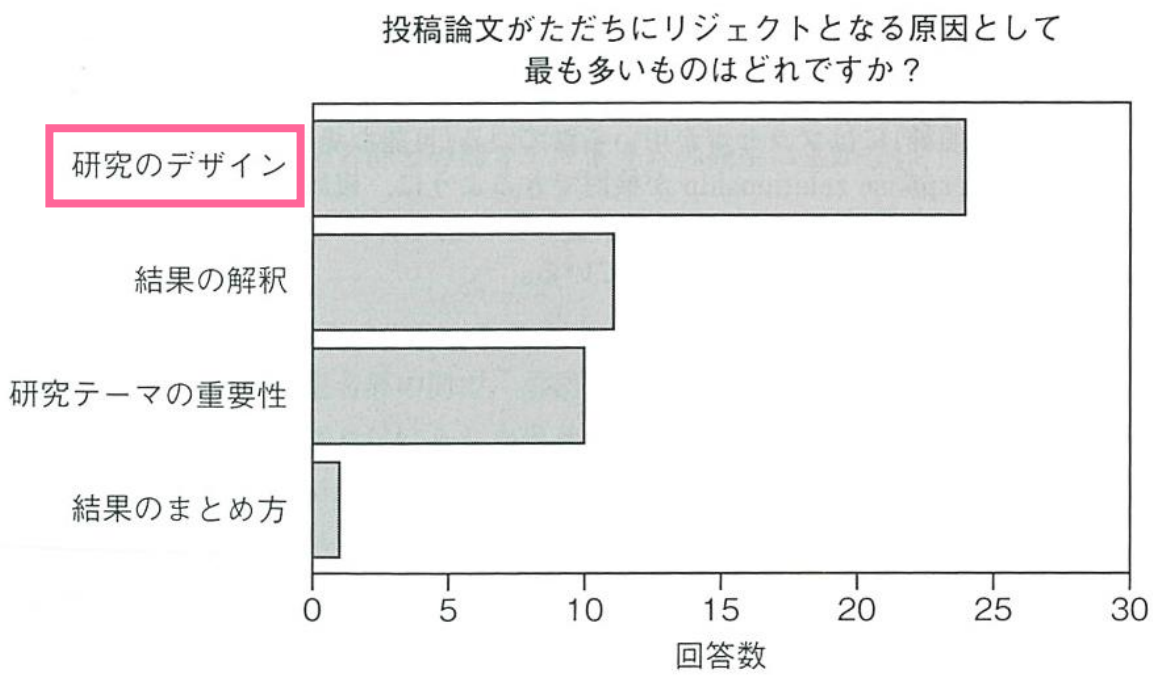
■ 介入

- ランダムに割り付けられている.
- コントロール群(対照群)にはプラセボが用いられている (可能な場合).
- 用量—反応関係(dose-response relationship)が検討できるように, 複数の要領が用いられている.
- 二重盲検化(double blind method)が用いられている.

■ バイアスのない厳密なアウトカムの測定

- アウトカムの定義が明確
- 測定の特異性が高い.
- 測定の客観性が高い.
- 測定方法の有効性が, 広く認められている.
- 中立的立場の観察者によって測定される.
- 長期のQOLを反映する指標である(自記式質問票によるものが望ましい).
- 前向きに測定される.
- あらゆる交絡可能因子を含めた包括的なデータセットの一部として測定され, かつ, 適切に定量化されている.

論文査読者に対するアンケート調査(Byrne, 2017)



すなわち、論文のAccept / Rejectは、執筆の段階ではなく、研究計画段階が最も重要である。

適格基準を適正に定める (Fair Subject Selectionの担保)

寄せ集めサンプル(適格基準が曖昧な被験者集団)の結果は、再現性が低くなるため、結果の信頼性を大きく損なう恐れがある。

● 選択基準

臨床研究の有効性が示された場合にその治療を適用することが妥当とみなされる集団を規定する基準であること。対象疾患、年齢、性別、症状、既往疾患、併存疾患に関する制限、臨床検査値等による閾値、同意能力等を明確に記述すること。例えば、特定の遺伝子変異を有する者を臨床研究の対象者として選択する場合にあっては、当該遺伝子変異の有無を記載すること。

● 除外基準

選択基準で示される集団に属するが、特定の状況下でリスクが高くなり臨床研究への参加が倫理的でない、また、臨床研究の有効性・安全性評価に影響を及ぼすと判断されることを規定する基準であること。

● 中止基準

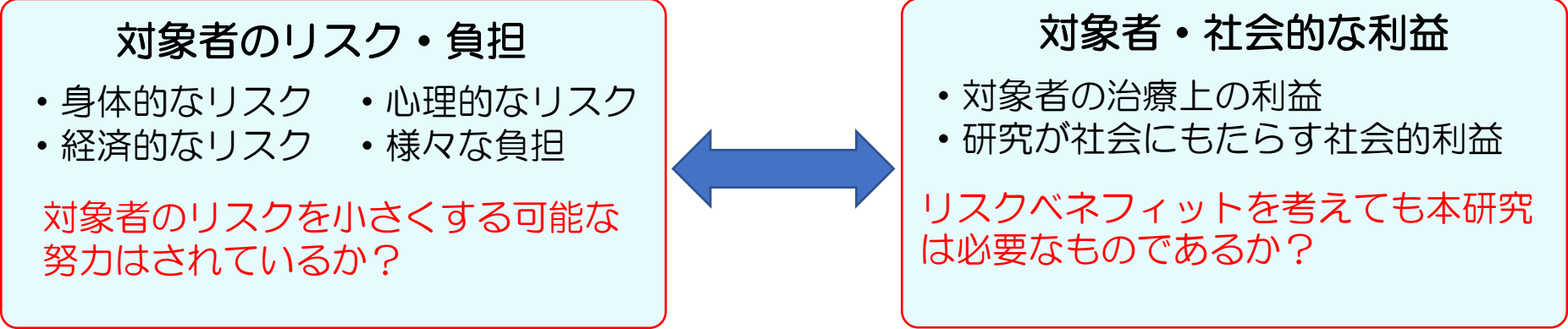
いつ、どのようにして臨床研究の対象者の参加を中止とするか、理由を含めて規定すること。また、中止後、どのようなデータをいつ集めるかも含めて記載すること。

適格基準の文言で気を付けなければならないところ

- ・「～と思われる」、「～と判断される」のような主観的判断を要する基準は避ける。
- ・「原則として」は基準が曖昧なので避ける。
- ・「かつ(and)」、「あるいは(or)」が明確であるかを確認する(一つの文書にandとorの両方が含まれないように気を付ける)。 [例：AかつBあるいはC (A and (B or C), (A and B) or Cなのかわからない]
- ・「～でないもの以外の〇〇」といった2重否定は避けること [Her2陽性ではない胃癌患者以外の標的病変を有する患者 → Her2陰性かつ標的病変を有さない胃癌患者]。

リスク・ベネフィット(Favorable Risk/ Benefit Ratio)

本試験におけるリスクベネフィットに基づいて「本研究参加に伴って予想される利益」，「本研究に伴って予想されるリスクと不利益」について明確に記載する。また，そのリスクベネフィットに基づいて適格基準を設定する。



- 対象者が被る可能性のある身体的・心理的リスクや負担（可能な範囲でその頻度、程度、持続期間などの詳細を含む。）を研究計画書に記載すること
- 対象者が被る可能性のある社会的・経済的リスクや負担（研究参加に伴い生じる社会的差別やプライバシー侵害、経済的損失など）が研究計画において適切に記載すること
- 対象者の被るリスクや負担が可能な限り小さくすること（実施体制の整備、医療スタッフの配置、適格基準や中止基準の適切な設定などの対策を含む。）
- 全体として、対象者の被るリスクや負担は、期待される利益（臨床研究の対象者への治療上の利益及び研究が社会にもたらす社会的利益）に照らして適切であること

インフォームドコンセントの取得 (Informed Consent)

■ 同意説明文書の要件

(1) 理解できるか

- ☐ 説明文書において、説明項目に過不足はなく、可能な限りわかりやすく記載されている
- ☐ 臨床研究の対象者の年齢等にも配慮されている

(2) 項目の妥当性

- ☐ 臨床研究と日常診療との違い（研究参加には社会貢献の要素が含まれること）
- ☐ 何を目指した研究なのか（研究の必要性や背景）
- ☐ どのような研究方法なのか（特にランダム化やプラセボの使用など）
- ☐ 研究に参加した場合、どのような利益や不利益があるのか
- ☐ 健康被害が生じた場合、どのように対処するのか（補償の内容も含む）
- ☐ 来院や検査のスケジュール
- ☐ 研究に参加しない場合の選択肢（具体的な治療法など）
- ☐ 研究者や研究組織について（特に企業が関与している場合には明確に）
- ☐ 研究についての問や相談ができる機会や場所、連絡先

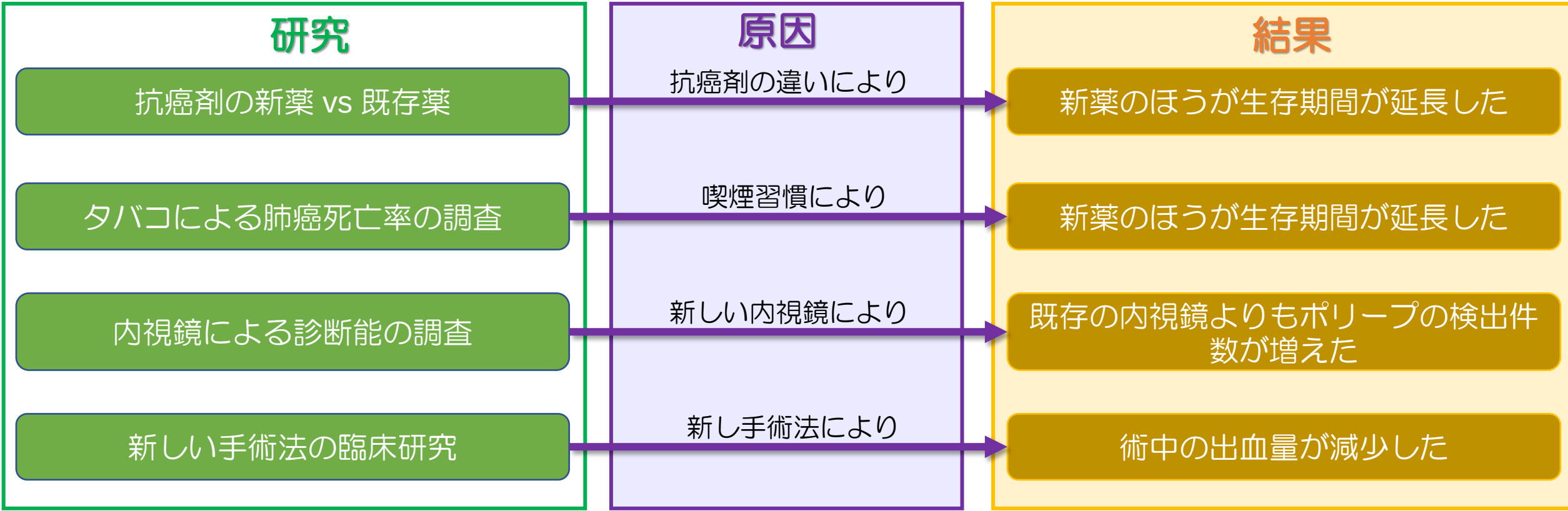
(3) 自発性が担保されているか

- ☐ 研究参加に関して、強制力が働いていない、又は強制力が働くことを取り除くような配慮がなされている
- ☐ 研究参加に対する不当な誘因がない（対象者の判断を狂わせるような過剰な医療サービスや物品・金銭等の提供は無い）

臨床研究のデザイン

臨床研究の多くは「因果」で成り立っている

因果とは、「原因」と「結果」が存在するような状況であり、臨床研究では、その因果関係を明らかにする研究が多い。とくに、特定臨床研究はすべて「因果」を取り扱う。



因果がない研究(調査)の一例



このような調査では、健康状態の因果関係を見ているのではなく、実態を把握するために行われるので、このような研究を記述的疫学研究という。

実際例：胃癌患者に対する栄養介入に関する研究

Ann Surg Oncol (2016) 23:2928–2935
DOI 10.1245/s10434-016-5221-4

Annals of
SURGICAL ONCOLOGY
OFFICIAL JOURNAL OF THE SOCIETY OF SURGICAL ONCOLOGY



ORIGINAL ARTICLE – GASTROINTESTINAL ONCOLOGY

Imamura et al., Ann. Surg. Oncol., 2016, 2928-2935.

Effects of an Oral Elemental Nutritional Supplement on Post-gastrectomy Body Weight Loss in Gastric Cancer Patients: A Randomized Controlled Clinical Trial

通常の治療(栄養介入なし)



手術



S-1投与により
体重が減少

体表面積が
減少

S-1の投与量
が減少

用量は体表面積に基づ
いて計算

仮定(本研究の枠外)

生存期間に
影響

当該試験の計画



手術



術後の体重減少を
抑制

結果

S-1の投与量の減
少を抑制

用量は体表面積に基づ
いて計算

生存期間に
影響

原因

成分栄養剤を
投与

体重減少を
抑制

研究の主体

評価項目(アウトカム)が適切であるか？

臨床研究(因果推論)では、適切にアウトカムを選択することが重要である。

切除可能病変を有する胃癌患者に対して、新たな術前補助化学療法(手術前に抗がん剤治療を行い、がんを縮小させたうえで手術を行う治療法)を計画している。本研究の目的は、新たな術前補助化学療法を実施することで、無再発生存期間(再発／死亡までの生存期間)を延長することにある。

■ 再考が必要かもしれない評価項目

手術時間、術中の出血量 etc...

→ 術前補助化学療法は、手術に対する影響を与えるものではないかもしれない(因果が成り立たない可能性が高い)。

■ 適切な評価項目

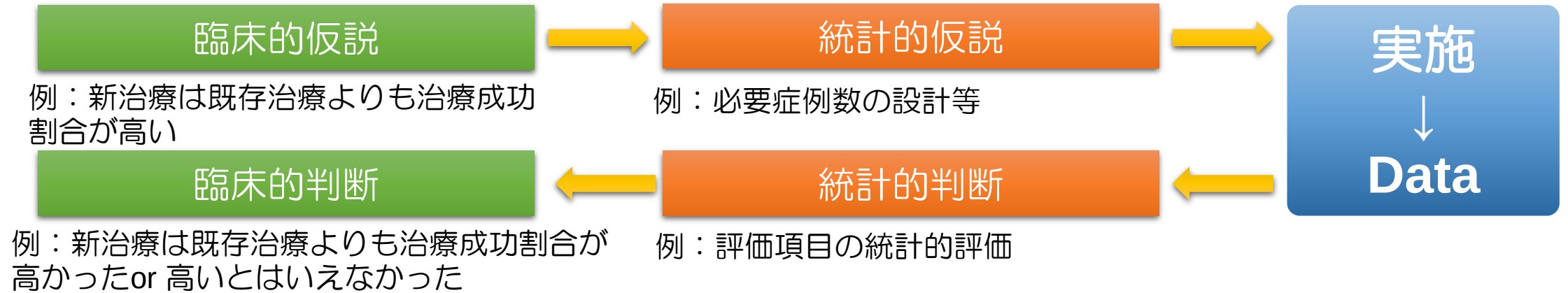
全生存期間(死亡までの期間)、無再発生存期間(再発／死亡までの生存期間) etc...

→ 術前補助化学療法を行うことで、手術後の再発を抑制することが期待され、その結果として死亡リスクを低減することができるため、因果が成り立つ。

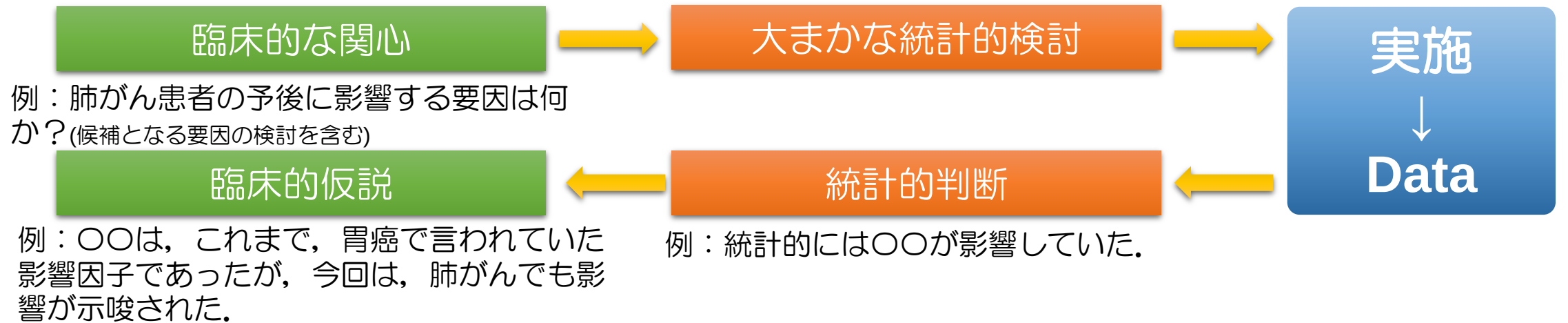
なお、がん臨床研究の真のエンドポイント(評価項目)の一つは、全生存期間(延命効果)である。ただし、全生存期間を得るには、(がん種や進行程度によるが)一般的に3年～5年程度の期間がかかるため、より早期に結果を得るために、奏効割合(がんが縮小した被験者の割合)、無増悪生存期間(増悪／死亡までの生存期間)、無再発生存期間などが用いられる。このような代替となる評価項目を代替エンドポイント(サロゲートエンドポイント)という。

臨床研究のスタイル：臨床試験といわゆる観察研究の違い

■ 臨床試験 (実験研究)



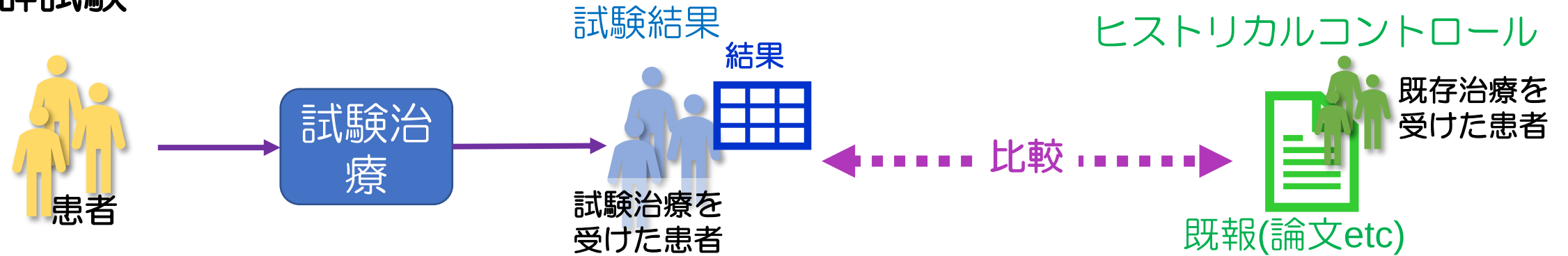
■ 観察研究



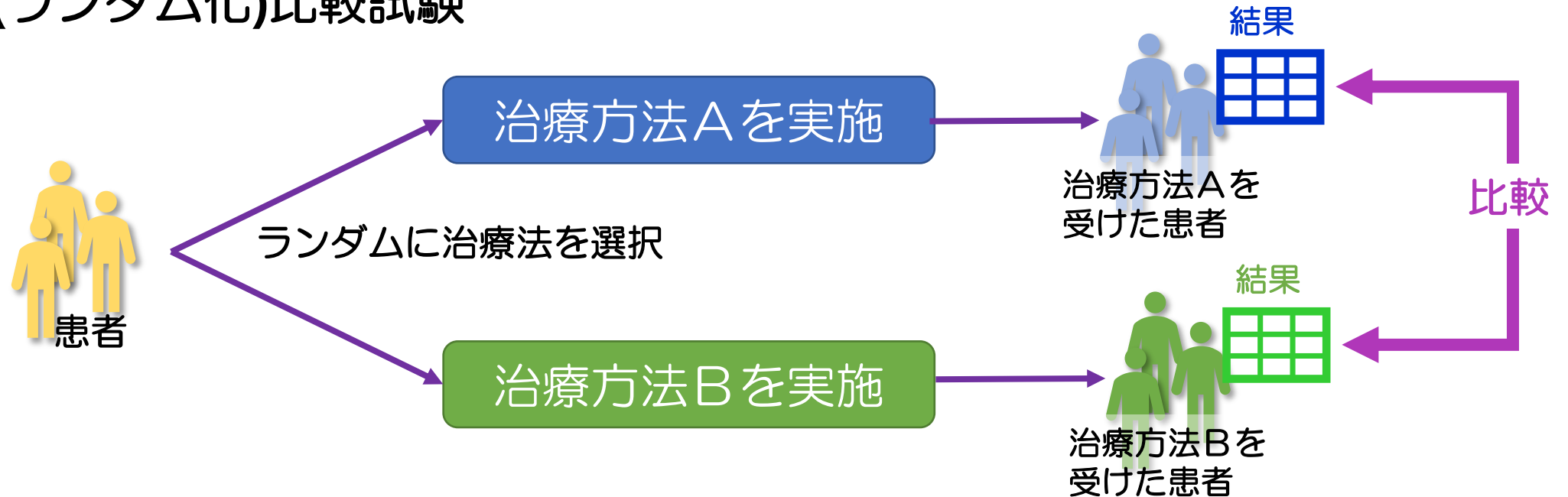
観察研究とは、臨床的仮説を新たに見出すために実施するものであり、臨床試験(特定臨床研究)とは、臨床的仮説を確認するものである。

臨床試験の主要なデザイン

■ 単群試験

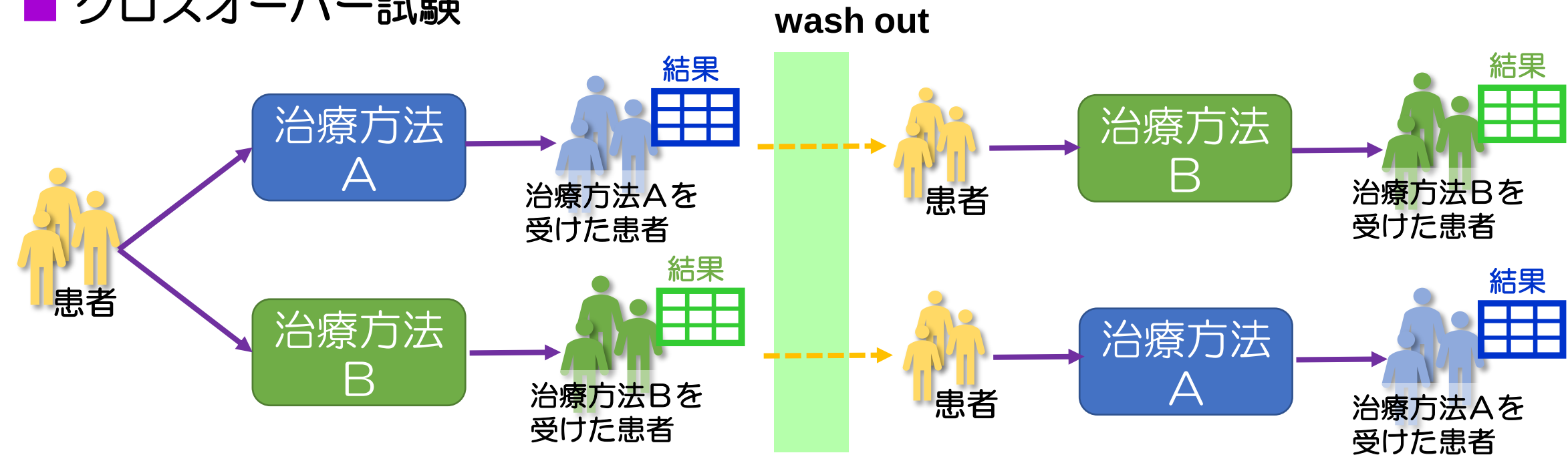


■ 無作為(ランダム化)比較試験

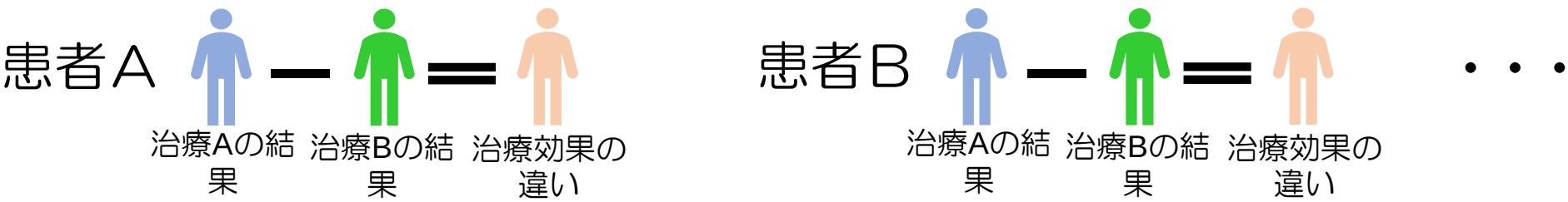


臨床試験の主要なデザイン

■ クロスオーバー試験



クロスオーバー試験では、すべての患者が二つの治療方法を実施している。そのため、治療効果の違いは、それぞれの治療での結果の違いを各患者において評価する。



単群試験の例示

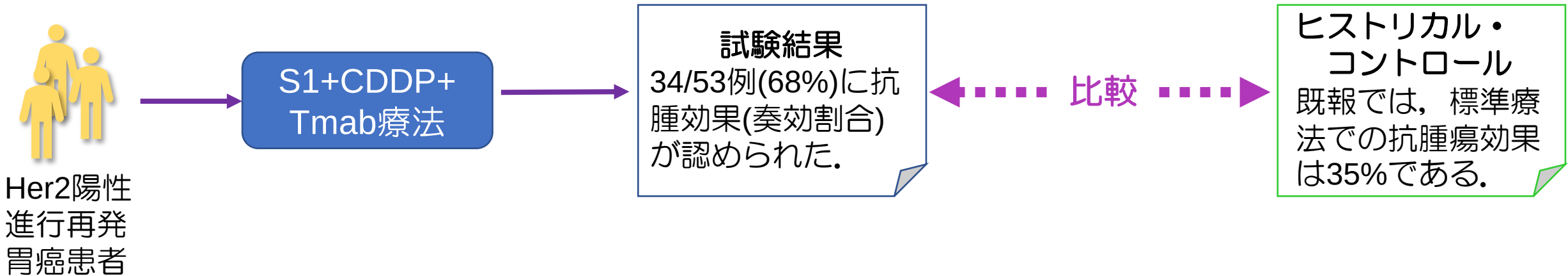
BJC

British Journal of Cancer (2014) 110, 1163–1168 | doi: 10.1038/bjc.2014.18

Keywords: trastuzumab; S-1; cisplatin; HERBIS

Phase II study of trastuzumab in combination with S-1 plus cisplatin in HER2-positive gastric cancer (HERBIS-1) Kurokawa, Y., et al., Br. J. Cancer, 110,1163-1168, 2014

HER2陽性の進行・再発胃癌患者に対するフッ化ピリミジン系抗がん剤とCDDPを併用したレジメンでの奏効率
が35%であると報告されている。新しい臨床試験では、S1+CDDP+Tmabの併用療法による新たなレジメンの有効
性を単群試験で実施した。



S1+CDDP+Tmab療法の抗腫瘍効果(奏効割合)が68%であった。これまでの標準療法の抗腫瘍効果が35%であることから、33%の上昇が認められた。

単群試験における留意点

単群試験では、① ヒストリカルコントロールでの主要評価項目(閾値)、② 期待される治療効果(期待値)の設定が必要である。

■ ヒストリカルコントロールには根拠が必要

単群試験では、閾値に対して臨床的な根拠が必要であり、統計解析は閾値(ヒストリカルコントロールにおける主要評価項目の値)よりもポジティブな結果か否かが評価される。したがって、ヒストリカルコントロールには、臨床的に許容される根拠が必要

- 既存の論文の結果
 - 当該論文の研究対象者は本試験と類似しているか？
 - プロトコル治療以外に当該論文との違いがないか(別の「原因」が存在しないか)？
- 院内でのこれまでの治療成績
 - これまでの治療成績がヒストリカルコントロールとして第3者的にも適切か？

HERBIS-1では、ToGA試験(Y-J Bang, et al. *The Lancet*, 28, 687 - 697, 2010.)の結果を参考にしている。ToGA試験では、化学療法群(5-FU /CDDP あるいはCapecitabine /CDDP)での**奏功率が35%**であった。

■ 期待される治療効果(期待値)は実現可能性及び臨床的有効性に基づいて設定

治療効果は、臨床的に合意されるような有効性(or 安全性)として設定される。

ToGA試験では、(1) Trastuzumab併用群の**奏功率が47%**であること、(2) HER2による選択基準(IHC 3+ or FISH positive + IHC 2+)、(3) 本邦の状況を鑑みて、閾値奏功率を50%と設定している。

無作為(ランダム化)比較試験 の例示

Ann Surg Oncol (2016) 23:2928–2935
DOI 10.1245/s10434-016-5221-4

Annals of
SURGICAL ONCOLOGY
OFFICIAL JOURNAL OF THE SOCIETY OF SURGICAL ONCOLOGY

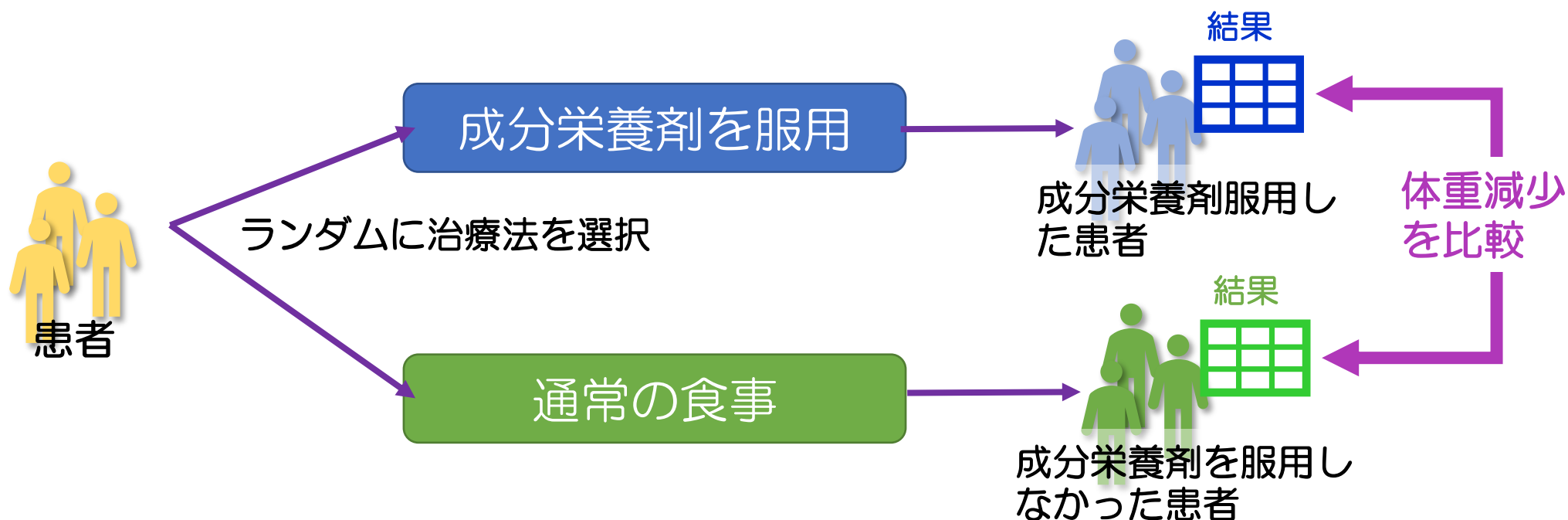


ORIGINAL ARTICLE – GASTROINTESTINAL ONCOLOGY

Effects of an Oral Elemental Nutritional Supplement on Post-gastrectomy Body Weight Loss in Gastric Cancer Patients: A Randomized Controlled Clinical Trial

Imamura et al., Ann. Surg. Oncol., 2016, 2928-2935.

胃癌患者において，外科治療後の補助化学療法(TS-1)を実施する予定の患者に対して，術後の食事に成分栄養剤を服用する群と，服用しなかった群での体重減少を比較する研究である。



無作為化比較試験における留意点

■ 無作為化比較試験では群間の均一化が重要

既存治療群と新規治療群の群間で研究対象者に偏りがないようにしなければならない。例えば、既存治療群に状態の悪い研究対象者が多く、新規治療群に状態の良い研究対象者が多い場合、治療による効果だったのか、研究対象者の状態による影響だったのかがわからなくなる。

そのため、無作為化比較試験では、群間で研究対象者が均一になるように、「**層別因子**」を設定する。そして、層別因子のもとで、均一化を行う。

■ CONSORT声明を遵守した研究になっているか？

CONSORT(Consolidated Standards of Reporting Trials：臨床試験報告に関する統合基準)とは、無作為化比較試験の報告の質の向上について記載された基準であり、論文とともにチェックリストを提出することが義務付けられるようになってきている。最新版は「CONSORT2010」である。

■ パイロット試験で無作為化比較試験を行うことについて

1群当たり10例程度の無作為化比較試験をパイロット試験と称して実施される場合が散見される。無作為化比較試験をざっくりと説明すると、「平均値の差をとると平均値に対するバラツキは2倍になる」。つまり、小規模な無作為化比較試験は、「誤差ばかり見ているだけ」になりかねない。そのため、小規模な無作為化比較試験では、

- ・ 無作為化比較試験でないといけない理由 (単群試験ではいけないのか？)
- ・ 小規模な症例数で有効性・安全性が評価できるのか？

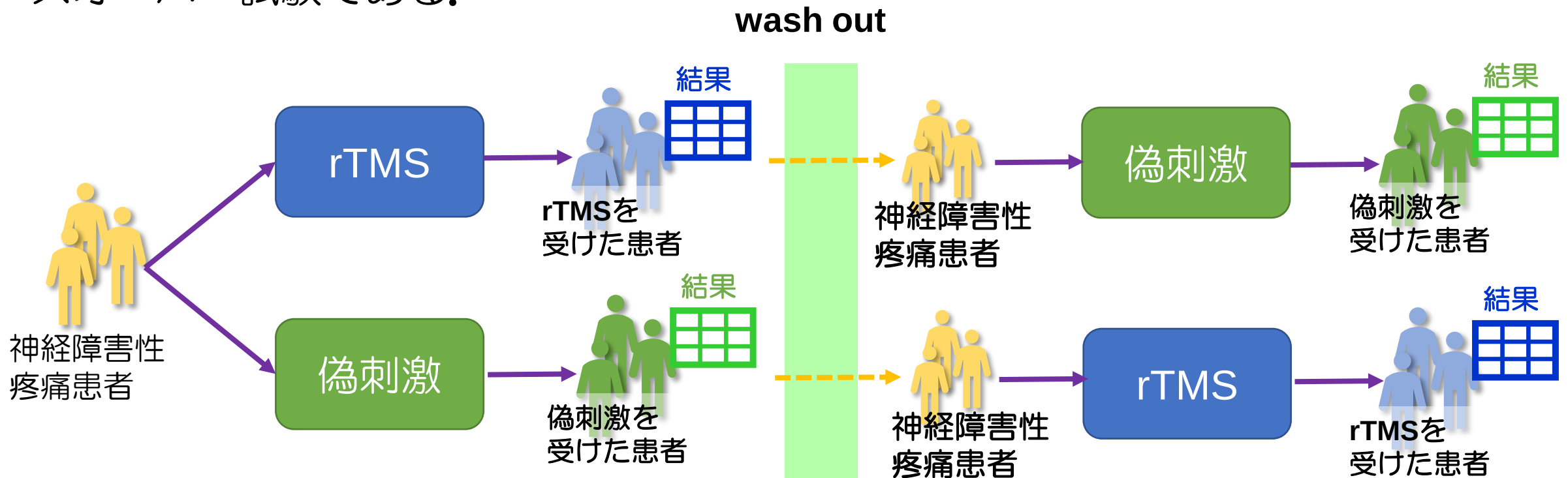
を吟味しないといけない(いまわかってないので、「とりあえずやってみます」では試験を行うべきでない。せめて、当該試験を受けて本試験(きちんと症例数を計算したもとで実施すること)が計画されていないといけない)。

クロスオーバー試験の例示

Daily repetitive transcranial magnetic stimulation of primary motor cortex for neuropathic pain: A randomized, multicenter, double-blind, crossover, sham-controlled trial

Hosomi et al., Journal of Stroke and Cerebrovascular Diseases., 2016, 1655-1664.

神経障害性疼痛患者に対するrTMS療法(反復的経頭蓋骨磁気刺激療法)と偽刺激のクロスオーバー試験である。



クロスオーバー試験における留意点

- 異なる治療を同一被験者に施行することで、群間の被験者の違いがなくなることから、標本サイズが大幅に削減できる。
- 持ち越し効果(carryover)が存在する場合には統計モデルが破綻するため注意が必要。
 - クロスオーバー試験ではウォッシュアウト期間において、最初の治療前と同様の状態に被験者が戻らないといけない。
 - したがって、がん臨床試験のように進行性の疾患にはクロスオーバー試験は実施できない。
- プラセボ対照試験の場合、心的持ち越し効果(盲検化の破綻)によるプラセボ効果の影響が分離できない可能性がある。
 - エンドポイントが主観的な評価の場合、研究対象者は前回に施行された治療と比較して相対的に評価を行う。そうすると、最初の治療と2番目の治療で評価基準が異なってしまうため、心的持ち越し効果が発生する。

統計的な考え方：サンプルサイズの設計と統計的評価について

仮説検定とは何か

無作為化比較試験で、200例の術後患者に抗生剤の新薬が投与され、別の200例の患者にプラセボが投与された。前者では10例が感染症に罹患し、後者では25例が罹患した。

カイ2乗検定の結果、p値は0.013であり、有意水準0.05のもとで有意だった。



統計学では、確率的な数値のもとで、決定論的な判断を行う

これは我々が自然に行っていること(例：降水確率が80%だったから傘を持っていこう)

帰無仮説と対立仮説

今回のリサーチクエスチョン：新薬とプラセボで罹患率に差があるか？

- 新薬とプラセボで罹患率に差がない

帰無仮説

罹患率に差がないとは、

$$(\text{新薬の罹患率}) - (\text{プラセボの罹患率}) = 0$$

である。「差がない」状況は、どのような研究でも同じである。

唯一の数値で設定できるこちら側を評価の基準とする(ただし研究の目的と逆)

- 新薬とプラセボで罹患率に差がある

対立仮説

罹患率に差があるとは、

$$(\text{新薬の罹患率}) - (\text{プラセボの罹患率}) \neq 0$$

である。「差がある」状況は、研究によって程度が違うかもしれない。

帰無仮説が誤っていることを示すことで、逆の仮説であるこちらが正しいと判断する。

どのように帰無仮説が誤っていると判断するのか (1/2)?

帰無仮説 H_0 ：新薬とプラセボの罹患率に違いがない

対立仮説 H_1 ：新薬とプラセボの罹患率に違いがある

帰無仮説 H_0 が正しいときに、データから得られた差が偶然に生じたものであるかどうかを表す確率のようなものが**P値**である。

p値が小さいとは、帰無仮説 H_0 が正しいとはいえないほどに差があることを意味する。つまり、対立仮説 H_1 が正しい状況にある。

先ほどの例では、無作為化比較試験の結果、新薬とプラセボの感染症に対する罹患率に**7.5%の差**があった。

新薬とプラセボの罹患率に違いがない(差がない)ことが真実ならば、臨床試験の結果で**7.5%の差**となる可能性は？

P値 = 0.013 (1.3%)

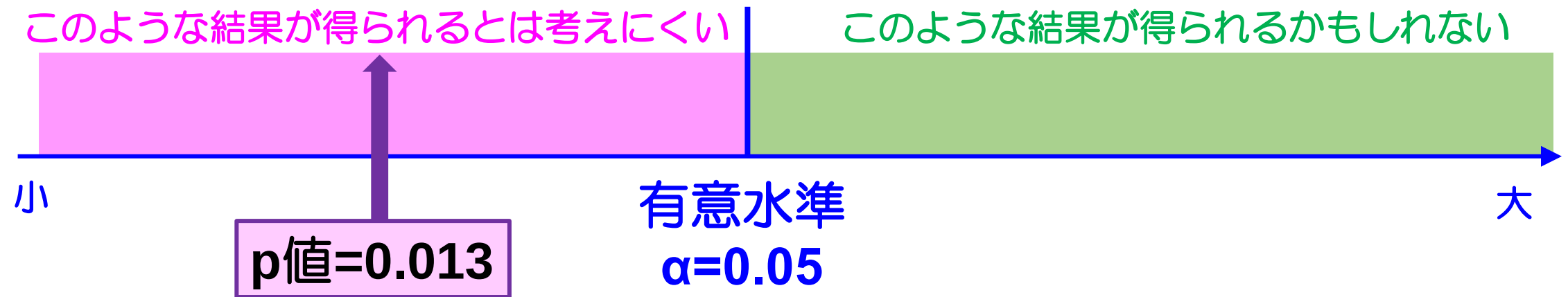
すなわち、新薬とプラセボの罹患率に違いがない(差がない)ことが真実であるとき、罹患率に7.5%の差が生まれるのは1.3%の可能性でしかない。

どのように帰無仮説が誤っていると判断するのか (2/2)?

P値 = 0.013 (1.3%)

すなわち、新薬とプラセボの罹患率に違いがない(差がない)ことが真実であるとき、罹患率に7.5%の差が生まれるのは1.3%の可能性でしかない。

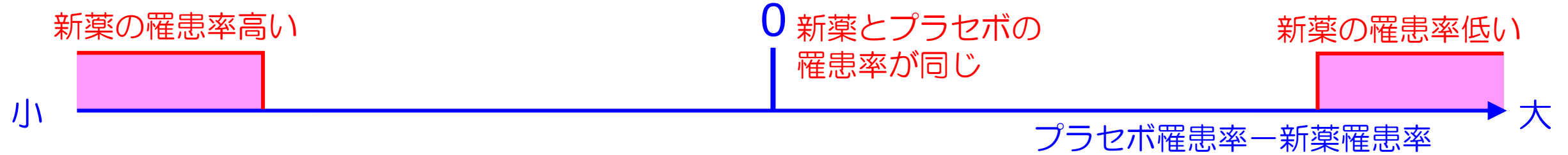
有意水準 α とは、この結果が得られる状況が考えられ得るものなのか、そうでないのかを判断するカットオフ値のようなもの。



新薬とプラセボの罹患率に違いがない(帰無仮説)とすると、本試験の結果が得られる確率*1は1.3%しかない。したがって、帰無仮説は誤っていると判断し(有意)、逆仮説である新薬とプラセボの罹患率に違いがある(対立仮説)と判断する。

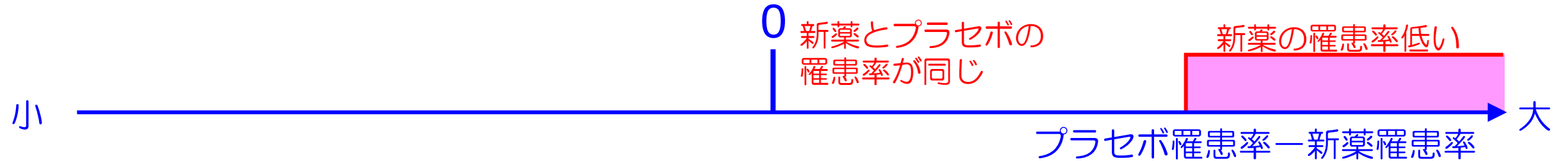
両側対立仮説(両側検定)と片側対立仮説(片側検定)

- 両側対立仮説：新薬とプラセボの罹患率に違いがある。



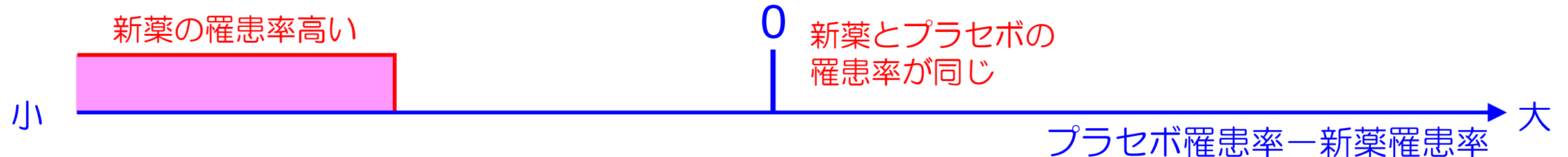
有意水準は「新薬がプラセボより罹患率が高い確率」「新薬がプラセボより罹患率が低い確率」の両方で設定されるので、片側では $\alpha/2$ になる。

- 片側対立仮説(1)：新薬のほうがプラセボの罹患率よりも低い。



有意水準は「新薬がプラセボより罹患率が低い確率」の両方で設定されるので、 α になる。

- 片側対立仮説(2)：新薬のほうがプラセボの罹患率よりも高い。



有意水準は「新薬がプラセボより罹患率が高い確率」の両方で設定されるので、 α になる。

両側検定だと有意なのに片側検定だと有意

例題：モルモットを用いて，L-アスコルビン酸を大量摂取させた場合の粗脂肪成分を調べた．9週にわたり毎日1000mgのアスコルビン酸を強制食餌させた群と対照群の体重減少のデータある．両群に体重減少に**差があるだろうか**．

(Brown & Hollander, 1976)

対照群	1000mg群
23.8	13.8
15.4	9.3
21.7	17.2
18.0	15.1

- 対照群での平均値 = 19.725
- 1000mg群の平均値 = 13.850

■2標本 t 検定での結果

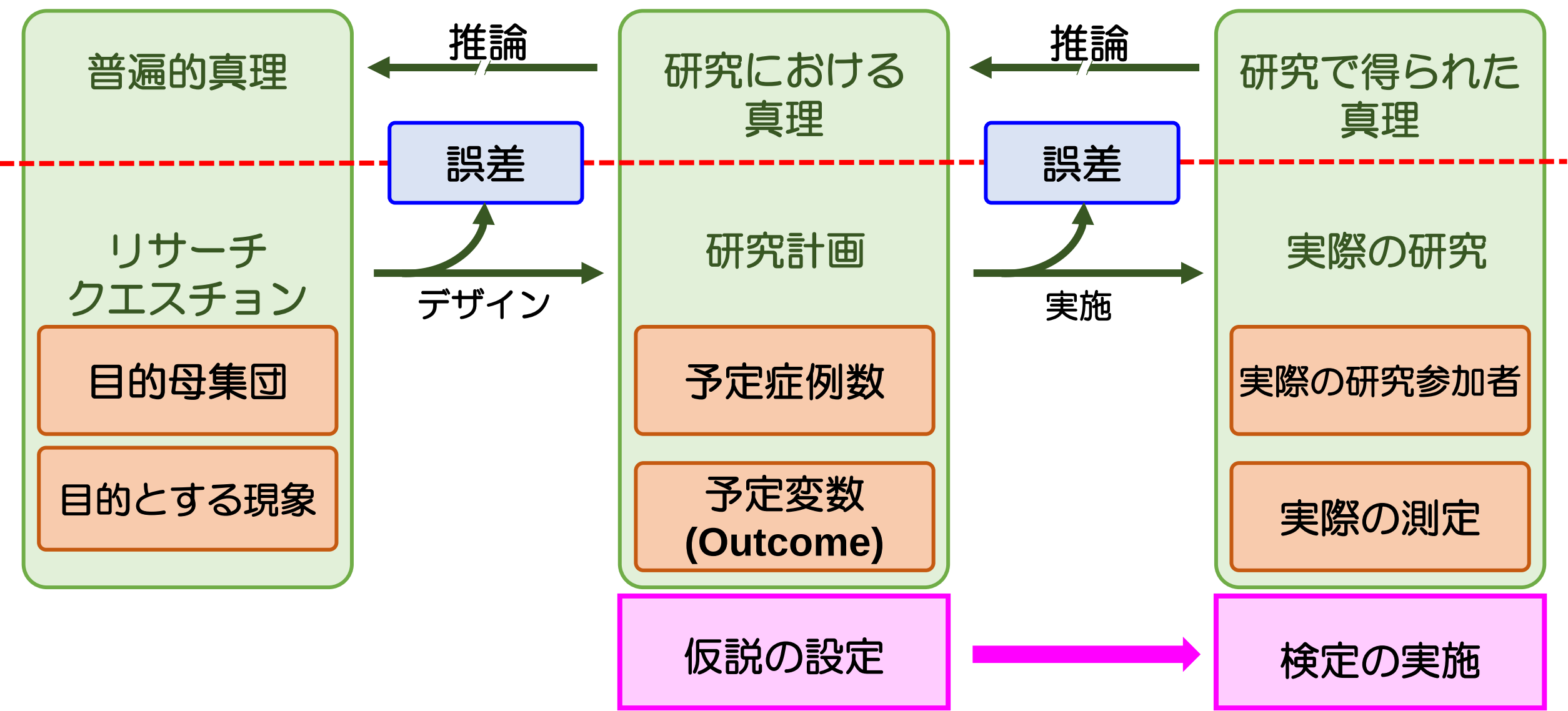
p値=0.057 (有意水準 $\alpha=0.05$ のもとで有意でない)

片側対立仮説（摂取によって体重が減少する）ならばp=0.0289



片側検定だったら有意になるからといって，片側検定に切り替えて解釈して良いか？

対立仮説は研究の計画段階で決めなければならない



両側・片側の選択は、研究計画で行われるため、実際の研究において、それを変更することは許されない。

サンプルサイズとp値

胃がんに対する二つの臨床研究があったとする。これらは、同一対象患者に対して、同一レジメンの術後補助化学療法が実施された。なお、既存治療での当該患者に対する奏効割合は40%だった。

帰無仮説 H_0 ：奏効割合は40%である(既存治療と同じ)

片側対立仮説 H_1 ：奏効割合は40%を上回る(既存治療を上回る)

Study A

症例数：40例, 奏功例：20例

奏効割合： $\hat{p} = 0.5$ (50%)

p 値(閾値0.4)でのp値 = 0.130 (n.s.)

Study B

症例数：80例, 奏功例：40例

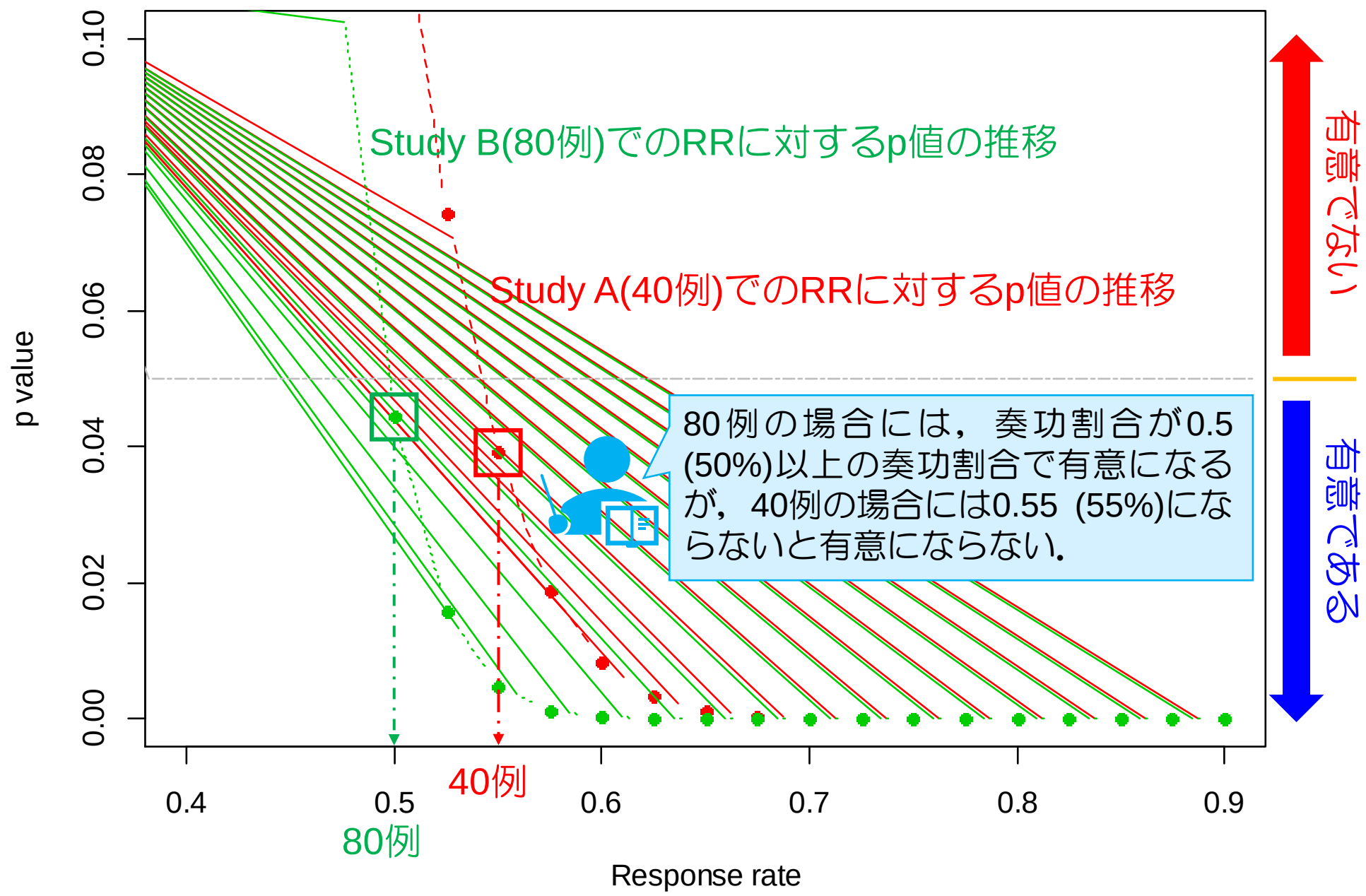
奏効割合： $\hat{p} = 0.5$ (50%)

p 値(閾値0.4)でのp値 = 0.045*

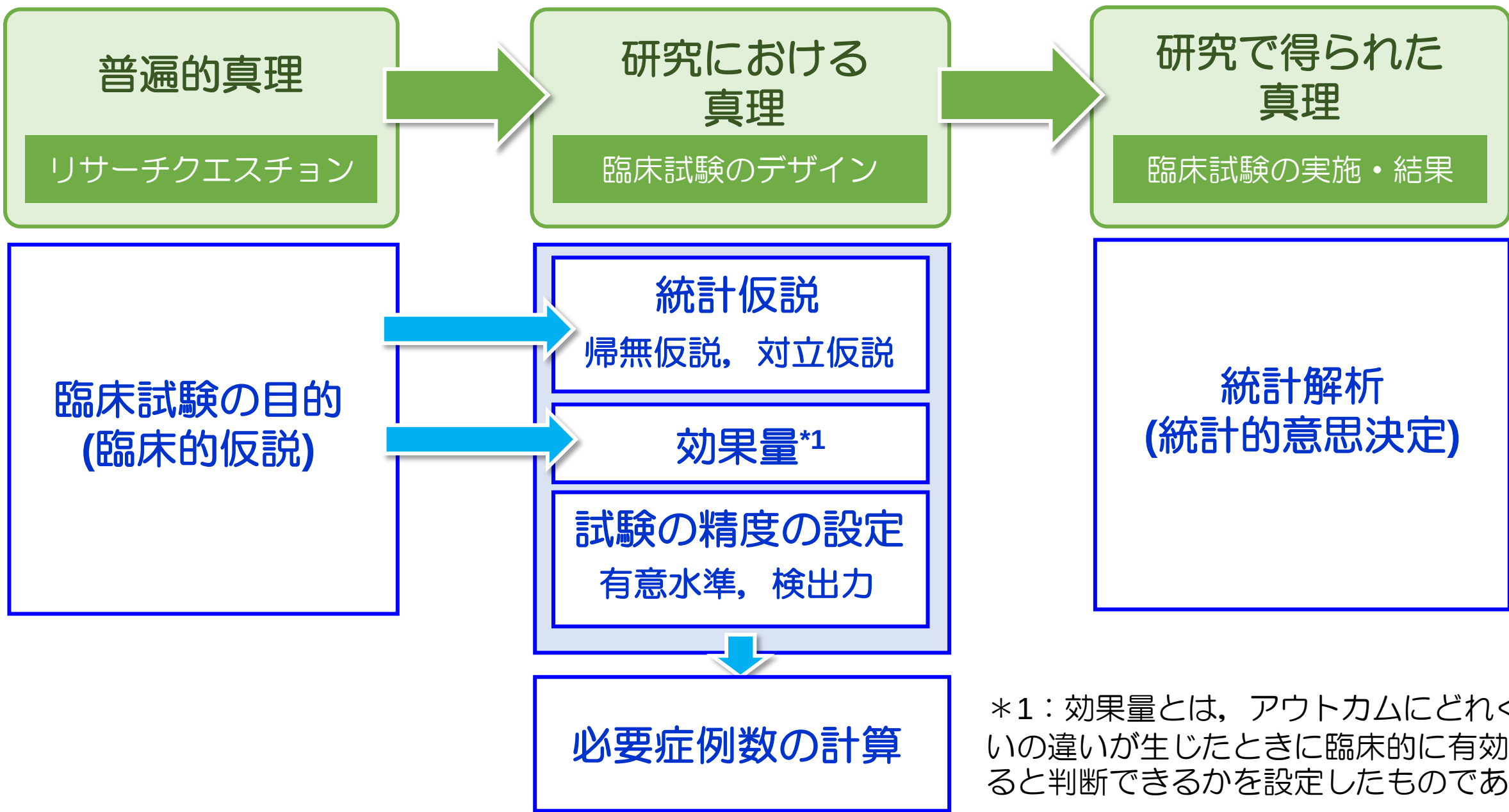
全く同じ奏効割合(0.50)であったとしても、症例数の違いから、Study Aではnegative studyとなり、Study Bではpositive studyとなる。

Study AがPositive studyになるには奏効割合がどれぐらい必要だったか？

Study AとStudy Bで有意になる奏効割合



実際の試験デザインのロジック



*1: 効果量とは、アウトカムにどれぐらいの違いが生じたときに臨床的に有効であると判断できるかを設定したものである。

症例数設定の一例

論文での記述例

Phase II study of trastuzumab in combination with S-1 plus cisplatin in HER2-positive gastric cancer (HERBIS-1)

The required sample size was estimated based on a threshold RR of 35% and an expected RR of 50%, 80% power, and an alpha value of 0.1 (one-sided) using the binomial test. Given 2% of ineligible patients, the target sample size was determined to be at least 50 patients.

(Kurokawa et al., British Journal of Cancer, 110, 1163-68, 2014)

■ 上記記載の要約

試験デザイン：Phase II Study (単アーム)

主要評価項目：奏効割合 (RR: Response Rate)

閾値奏効率：35%， 期待奏効率：50%

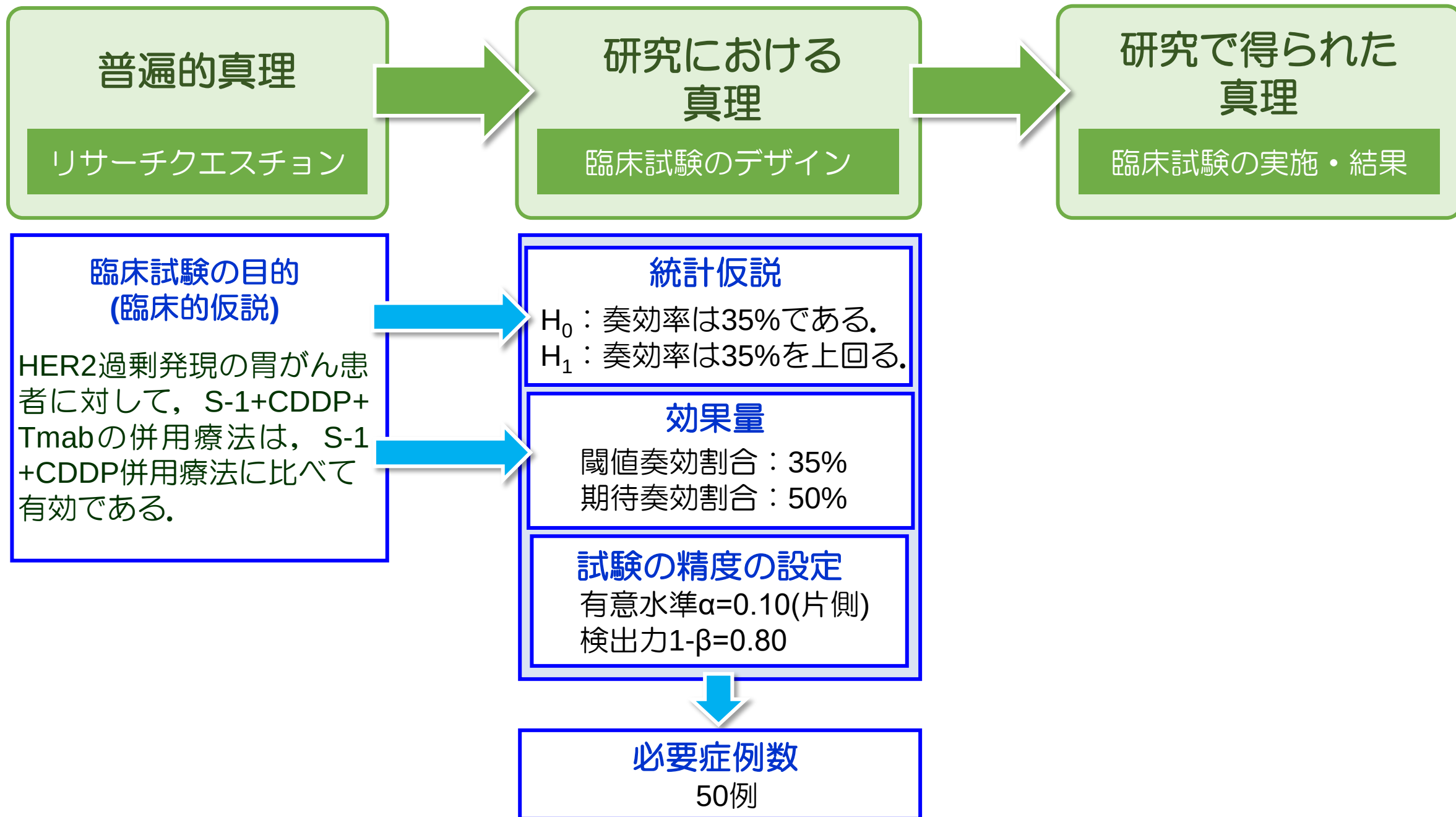
臨床的判断

第1種の過誤(type I error, 有意水準)： $\alpha = 0.1$

1－第2種の過誤(1-type II error, 検出力)： $1 - \beta = 0.8$ (80%)

統計的判断
必要症例数 = 50

試験デザインのロジックにトレースする



有意水準 α と検出力 $1-\beta$ の意味

		検定により評価		標本サイズで規定	
		真の結論			
		H_0 が正しい [H_1 ：偽]		H_1 が正しい [H_0 ：偽]	
試験の結果	H_0 を棄却 [H_1 を受容]	第1種の過誤 α (H_0 が正しいのに棄却) 有意水準		検出力 $1-\beta$ (H_0 が正しいので受容) 検出力	
	H_0 を受容 [H_1 を受容]	正しい判断 $1-\alpha$ (H_1 が正しいので棄却)		第2種の過誤 β (H_1 が正しいのに受容)	

- 第2種の過誤が一定水準未満(β 未満)になるように、標本サイズを規定する
(本当はpositiveなのに本試験を通じてnegativeと判断する可能性を低くする).
- 統計解析の結果、第1種の過誤が一定水準未満(α 未満)であるか否かを確認する
(本当はnegativeなのに本試験結果を通じてpositiveと結論付けるエラーがかなり低いことを確認する).

必要症例数設計のプロセス

STEP.1：主要評価項目(主たるアウトカム)を考える

- 本研究におけるクリニカルクエスチョンに答えられるものか(妥当性)?
- 当該分野においてAgreeされるものか(信頼性)?
- エンドポイントを取得できるか(実施可能性)?
一例えば、抗がん剤治療において標的病変がなければ奏効割合が測定できない等

STEP.2：仮説・エフェクトサイズ等を決定する

- 無作為化比較試験(RCT)の場合には非劣勢・優越性・同等性を検討する.
- 単アーム試験の場合には閾値を文献等から決定する.
- エフェクトサイズ(効果量)を検討する.

量的変数の場合：平均値(RCTの場合には平均値の差), 標準偏差(RCTの場合には共通)

2値変数の場合：比率(RCTの場合には比率の差)

生存時間の場合：MST or 年次生存割合 (RCTの場合にはHR), 登録期間・フォローアップ期間
(打ち切り割合を計算するため)

STEP.3：有意水準・検出力を決定する

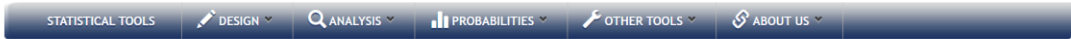
単群試験： $\alpha=0.05$ or 0.10 (one sided), $1-\beta = 0.8$

無作為化比較第II相試験(screening design)：単群試験と同程度

無作為化比較第III相試験： $\alpha=0.05$ (two sided), $1-\beta = 0.80$ or 0.90

サンプルサイズ計算・統計解析のためのフリーソフト

必要症例数を検討するWEB SITE (1/2)



Statistical Tools

Design	Analysis
One Arm Binomial	Frequency Table
One Arm Normal	Binomial Confidence Interval
One Arm Survival	Probabilities
Two Stage	Binomial
Two Arm Binomial	Normal
Two Arm Normal	Poisson
Two Arm Survival	Chi-Square
Binomial Interaction	Probability of Observing a Rare Event
Survival Interaction	Other Tools
Binomial Noninferiority	Prognostic Mixture
Survival Noninferiority	
Expected Events (1 Arm)	
Expected Events (2 Arm)	
Continuous Marker Prognostic Power	

<https://stattools.crab.org/>

単アーム試験

- ー母比率 (One Arm Binomial)
- ー母平均 (One Arm Normal)
- ー生存時間 (One Arm Survival)
- ー2段階デザイン(母比率のみ, Two Stage)

2アーム試験 (無作為化比較試験)

- ー母比率 (Two Arm Binomial)
- ー母平均 (Two Arm Normal)
- ー生存時間 (Two Arm Survival)

その他の試験

- ー母比率に対する交互作用 (Binomial Interaction)
- ー生存時間に対する交互作用 (Survival Interaction)
- ー母比率での非劣勢 (Binomial Noninferiority)
- ー生存時間での非劣勢 (Survival Noninferiority)
- ー期待死亡数 (Expected Event (1Arm) / (2Arm))

必要症例数を検討するWEB SITE (2/2)

The Chinese University of Hong Kong

https://www2.ccrb.cuhk.edu.hk/stat/

Phase II Clinical Trial

Introduction: The main goal of Phase II clinical trials is to identify the therapeutic efficacy of new treatments. They are usually single-arm studies, but may take the form of multiple-arm trials. Multiple-arm trials can be randomized or non-randomized with or without control arms.

Phase II clinical trials are generally single-arm studies aimed at estimating the activity of a new treatment. These “pilot” studies are commonly applied to anticancer drugs to assess the therapeutic efficacy and toxicity of new treatment regimens. Phase II trials decide whether the new treatment is promising and warrants further investigation in a large-scale randomized Phase III clinical trial based on an observed response rate that appears to be an improvement over the standard treatment or other experimental treatments. Because the sample size is small (generally less than 50 patients), Phase II clinical trials are only able to detect a large treatment improvement, e.g. greater than 10%. To detect a small difference in treatment, e.g. less than 5%, one would require a much larger sample size, which is not possible in Phase II studies due to the limited number of subjects eligible for the study and the large number of treatments awaiting study. Phase II studies are prominent in cancer therapeutics because new treatments frequently arise from combinations of existing therapies or by varying dose or radiation schedules.

There are three main objectives in treating patients in Phase II clinical trials. The primary objective is to test whether the therapeutic intervention benefits the patient. The second objective is to screen the experimental treatment for the response activity in a given type of cancer. The final objective is to extend our knowledge of the toxicology and pharmacology of the treatment.

An important characteristic of some Phase II trial designs is the use of early stopping rules. If there is sufficient evidence that one of the treatments under study has a positive treatment effect, then patient accrual is terminated and this treatment is declared promising. Also, if a treatment is sufficiently shown not to have a desirable effect, then patient accrual is terminated and this treatment is declared not promising.

This program helps users to determine sample sizes or decision rules for a Phase II clinical trial so that the sample size needed for a standard Phase II clinical trial can be reduced.

If you don’t know which program to use, please feel free to check : [User guidance, definition and terminology](#)

Means	Proportions	Survival Analysis	Phase II Clinical Trials	Epidemiological Studies	Confidence Intervals	Others
Phase II Clinical Trials						
1. Fleming's Phase II Procedure						
2. Bayesian Phase II Design						
3. Simon’s Randomized Phase II Design						

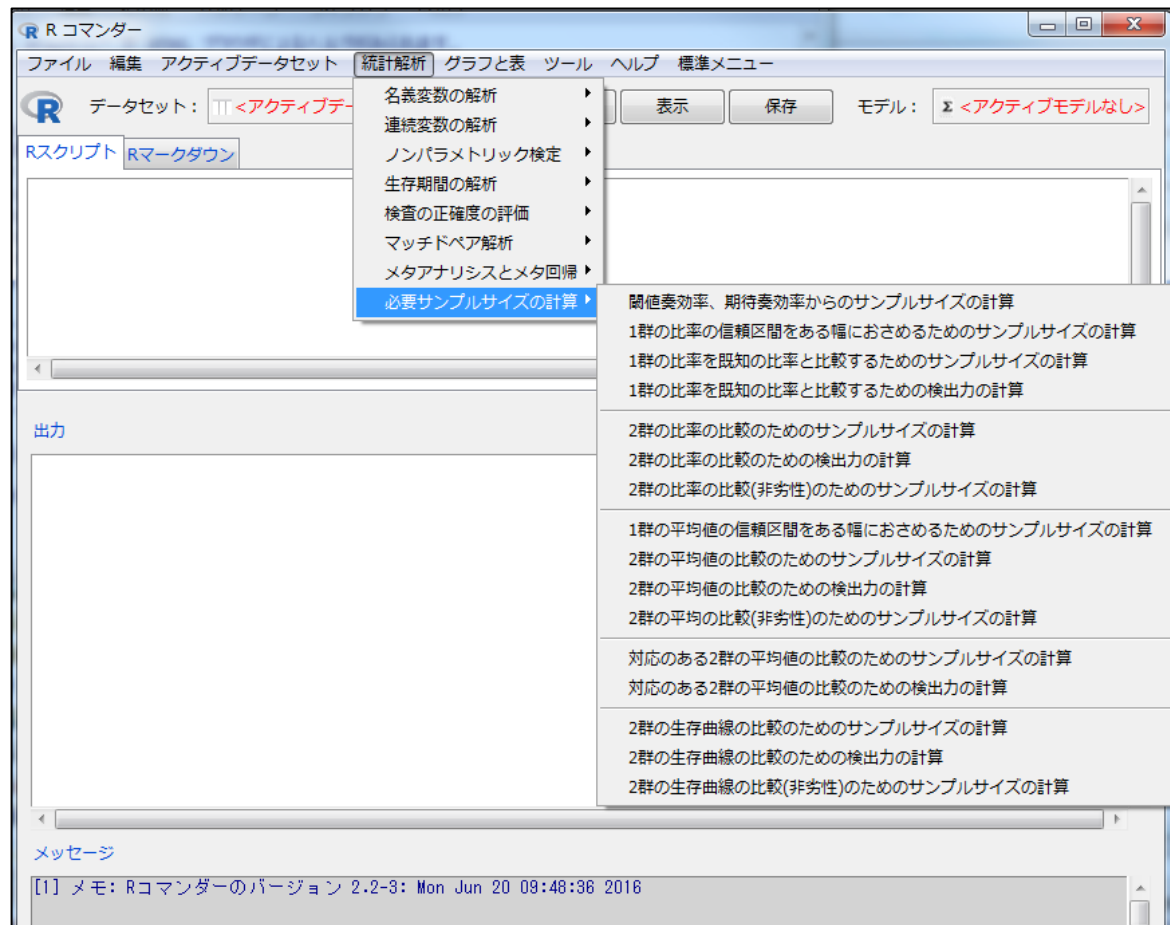
Further Information:

Difference between Fleming’s procedure and Bayesian design of Phase II clinical trials

This section describes both the hypotheses and design for Fleming’s and the Bayesian approach to single-arm Phase II clinical trials. Both designs are used for Phase II clinical trials with binary outcomes and continuous monitoring. The fundamental difference between the two designs is the frequentist basis for Fleming’s procedure only depends on the observed results whereas the Bayesian approach uses prior information. The testing procedure for Fleming’s procedure is based on the normal approximation to the binomial distribution of the observed number of treatment responses. The resulting decision boundaries, r_g and a_g , are solved analytically. The Bayesian design incorporates prior information about the treatment being investigated with the observed results to yield revised beliefs about the treatment. The testing procedure is based on the posterior probability of the experimental treatment given the observed data. The posterior probability is a conditional probability computed from a beta distribution which results in the upper and lower decision boundaries, U_n and L_n , which are evaluated using numerical integration, namely “Simpson’s Composite Algorithm”. Another difference between the two designs is that Fleming’s procedure has only two outcomes at the final recruitment stage, reject or accept H_0 , whereas the Bayesian design traditionally allows for an inconclusive trial at the final stage.

あまり知られていないか，比較的豊富な試験デザインが含まれている。

必要症例数を検討するフリーソフト

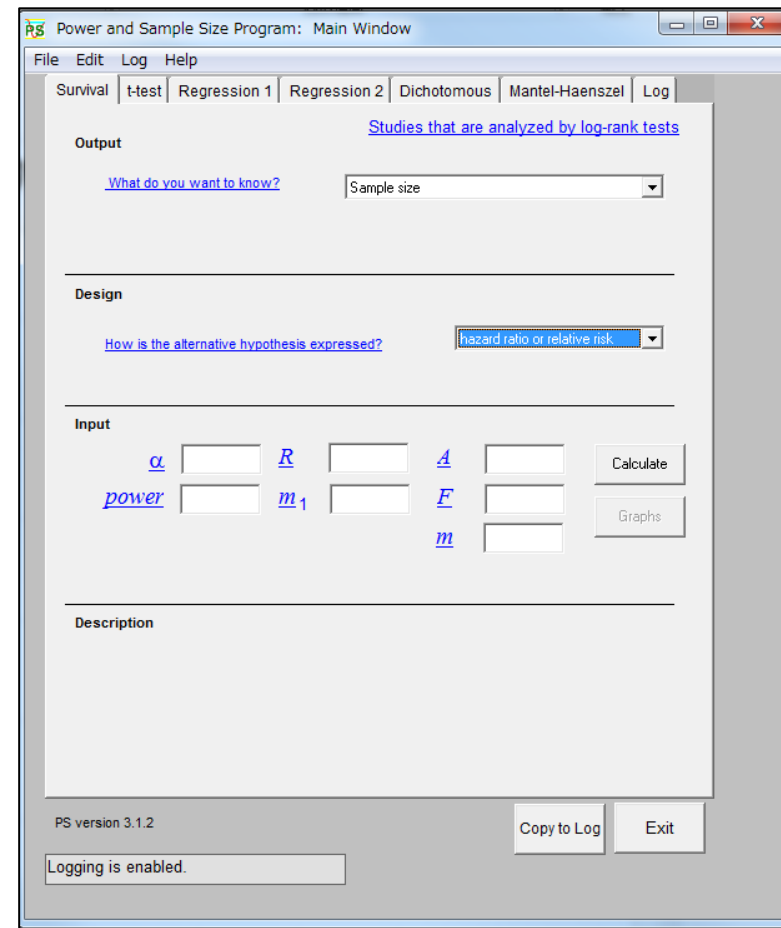


統計ソフトウェアR(統計学分野のデファクトスタンダード)をR言語が使えない利用者のために修正した統計ツール

<http://www.jichi.ac.jp/saitama-sct/SaitamaHP.files/statmed.html>



Power and sample size Program: PS



ヴァンダービルド大学が開発した標本サイズを計算するためのソフトウェア

<http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSize>

統計解析のためのフリーソフトウェアEZR

自治医科大学附属さいたま医療センター
血液科のホームページへようこそ。

自治医科大学附属さいたま医療センター
Jichi Medical University Saitama Medical Center



[ホーム](#) [施設紹介](#) [移植診療](#) [受診の案内](#) [研修・研究](#) [統計ソフトEZR](#)

[フリー統計ソフトEZR](#)

[EZRの解析機能](#)

[EZRの使い方、変更履歴](#)

[ダウンロード \(Windows標準版\)](#)

[ダウンロード \(MacOS X版\)](#)

[ダウンロード \(LINUX版\)](#)

[よくあるご質問 \(FAQ\)](#)

[English version](#)

***** [Click here for English version](#) *****

641769
(2012年6月11日のページ改訂後のアクセス数です。)

2019年12月1日 EZR version 1.41公開

2014年11月初心者向けマニュアル刊行

2015年4月EZRMニュアル第2版刊行
(↑クリックすると立ち読みすることができます)

2016年10月みんなのEBMと臨床研究刊行
(↑EBMの基礎、新倫理指針に適用した臨床研究の進め方、論文の書き方をまとめています)

1.23からサンプルの背景データのサマリー表(Table 1)を自動作成する新機能を搭載!!

1.40から傾向スコアマッチングなので必要なキャリアマッチングに対応!!!

動作確認済OS Windows XP~10、Mac OS X Snow Leopard~Mojave、Ubuntu 11.10~15.10

EZRを使用した学術論文を発表される場合は
[Bone Marrow Transplantation 2013; 48, 452-458](#)
を参考文献として引用くださいますようお願いいたします。
(「EZRの使い方、変更履歴」を参照。2019年11月の時点で英文論文での引用回数が3000回を超えました)

国際的に有名な[Softpedia software database](#)に
"100% CLEAN" Softpedia awardで掲載されました。

EBMと臨床研究の解説書

amazon.co.jp



ゼロから始めて
一冊でわかる...

¥ 3,960 

 [今すぐ購入](#)

EZRと統計解析の解説書

amazon.co.jp



初心者でもすぐ
にできるフリ...

¥ 4,180 

 [今すぐ購入](#)

↑ 初級~中級者用

amazon.co.jp



自治医科大学附属さいたま医療センター 血液科の
神田善伸教授が統計解析環境RのGUI環境であるR
コマンダーをカスタマイズしたものがEZRである。
EZRでは、医学統計学における諸種の統計的方法
をマウスのみで実行できる。

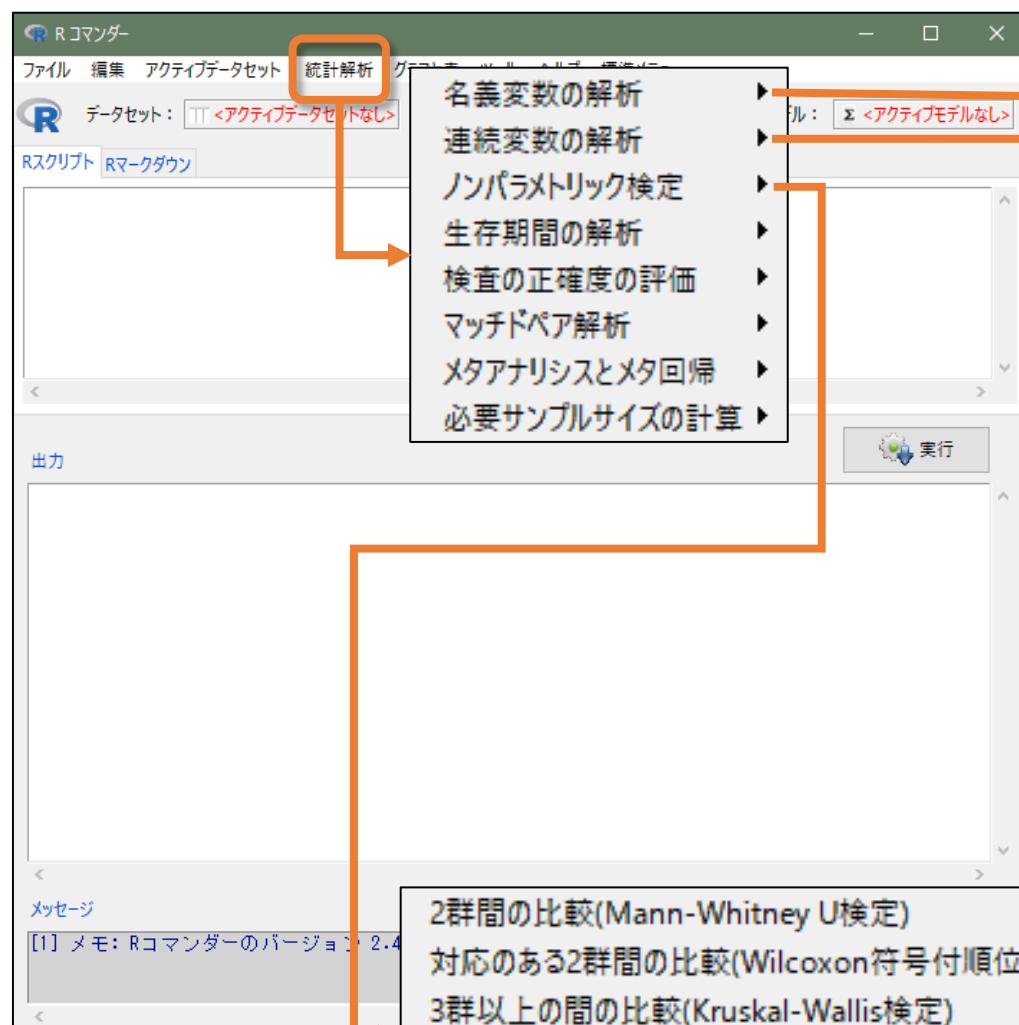
ダウンロードは、ブラウザで「EZR」と入力すれ
ば、左図のホームページがトップにくる。

Warning!!

MacOSでは、EZRのもととなるRcommanderを動かすために、X11というプラグインが必要になる。
そのため、EZRをインストールする前に、
<https://www.xquartz.org/>
から、X11をダウンロードおよびインストールする
必要がある。

<http://www.jichi.ac.jp/saitama-sct/SaitamaHP.files/statmed.html>

どのような統計解析ができるか(1/2)?



頻度分布
比率の信頼区間の計算
1標本の比率の検定
2群の比率の差の信頼区間の計算
2群の比率の比の信頼区間の計算

分割表の直接入力と解析
分割表の作成と群間の比率の比較(Fisherの正確検定)
対応のある比率の比較(二分割表の対称性の検定、McNemar検定)
対応のある3群以上の比率の比較(Cochran Q検定)
比率の傾向の検定(Cochran-Armitage検定)
二値変数に対する多変量解析(ロジスティック回帰)

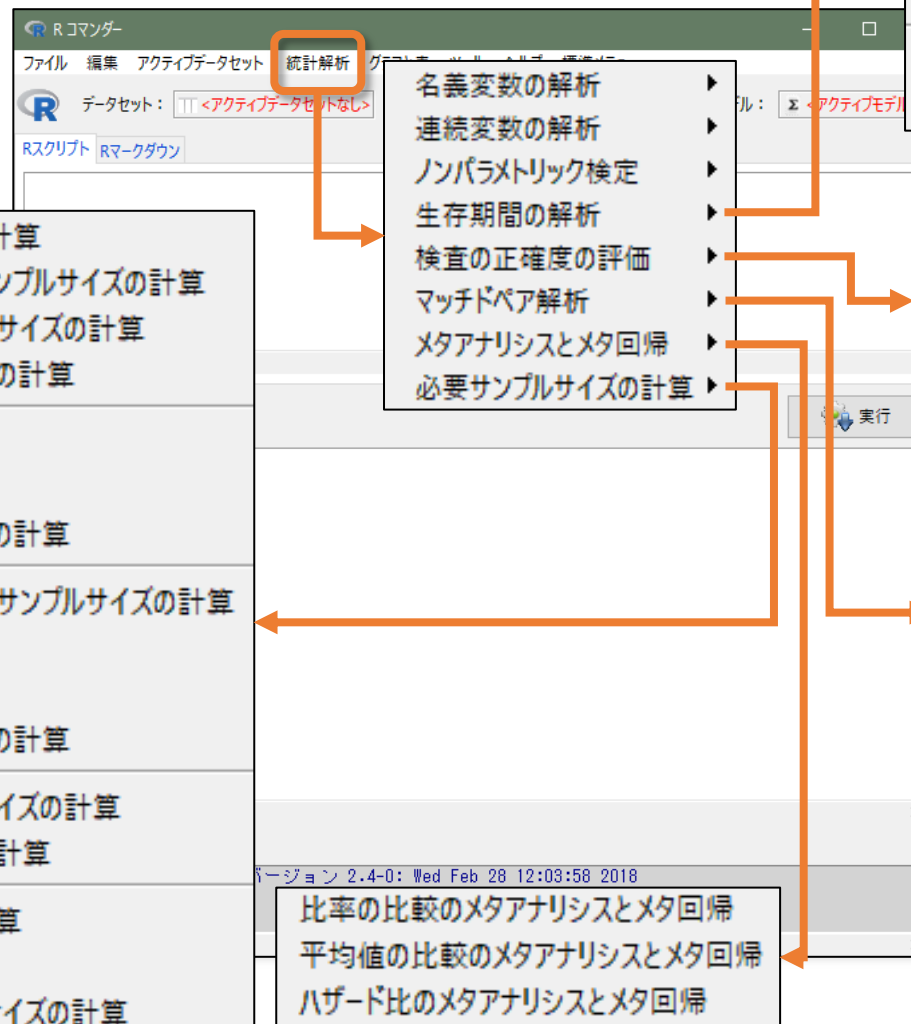
連続変数の要約
外れ値の検定と除外(Smirnov-Grubbs検定)
正規性の検定(Kolmogorov-Smirnov検定)
平均値の信頼区間の計算
1標本の平均値のt検定

2群の等分散性の検定(F検定)
2群間の平均値の比較(t検定)
対応のある2群間の平均値の比較(paired t検定)
3群以上の等分散性の検定(Bartlett検定)
3群以上の間の平均値の比較(一元配置分散分析one-way ANOVA)
対応のある2群以上の間の平均値の比較(反復[経時]測定分散分析)
複数の因子での平均値の比較(多元配置分散分析multi-way ANOVA)
連続変数で補正した2群以上の間の平均値の比較(共分散分析ANCOVA)

相関係数の検定(Pearsonの積率相関係数)
線形回帰(単回帰、重回帰)

2群間の比較(Mann-Whitney U検定)
対応のある2群間の比較(Wilcoxon符号付順位和検定)
3群以上の間の比較(Kruskal-Wallis検定)
対応のある3群以上の間の比較(Friedman検定)
連続変数の傾向の検定(Jonckheere-Terpstra検定)
相関係数の検定(Spearmanの順位相関係数)

どのような統計解析ができるか(2/2)?



Thank you for kind attention



shimokaw@wakayama-med.ac.jp

バックアップ：諸種のデータに対する統計的検定

医学研究においてとり扱われるデータ

■ 量的データ (numerical data)

血圧，腫瘍径といった数値に単位があるようなデータを指す．言い方を変えれば，平均値(中央値)が意味をもつデータを指す．

位置を表す指標：平均値，中央値

バラツキを表す指標：標準偏差(分散)，四分位範囲(最小-最大)

■ カウント・データ (count data)

ポリープの個数のように「個数」を意味するデータを指す．

位置を表す指標：中央値

バラツキを表す指標：四分位範囲(最小-最大)

■ 2値データ (binary data)

結果が(0,1), (成功, 失敗), (罹患, 非罹患)のように2パターンであらわされるデータをさす．例えば，奏効の有無，疾患の有無などがある．

要約指標：クロス集計表，関心がある事象の比率，オッズ比
ただし，患者背景を表す場合には個々のカテゴリ数を記述

■ カテゴリカル・データ (ordered categorical data)

順位がある(順序データ)：疾患のグレード(軽度, 中程度, 高度)など

順位がない(名義データ)：血液型や疾患の種類など

要約指標：クロス集計表, それぞれのカテゴリの比率
ただし, 患者背景を表す場合には個々のカテゴリ数を記述

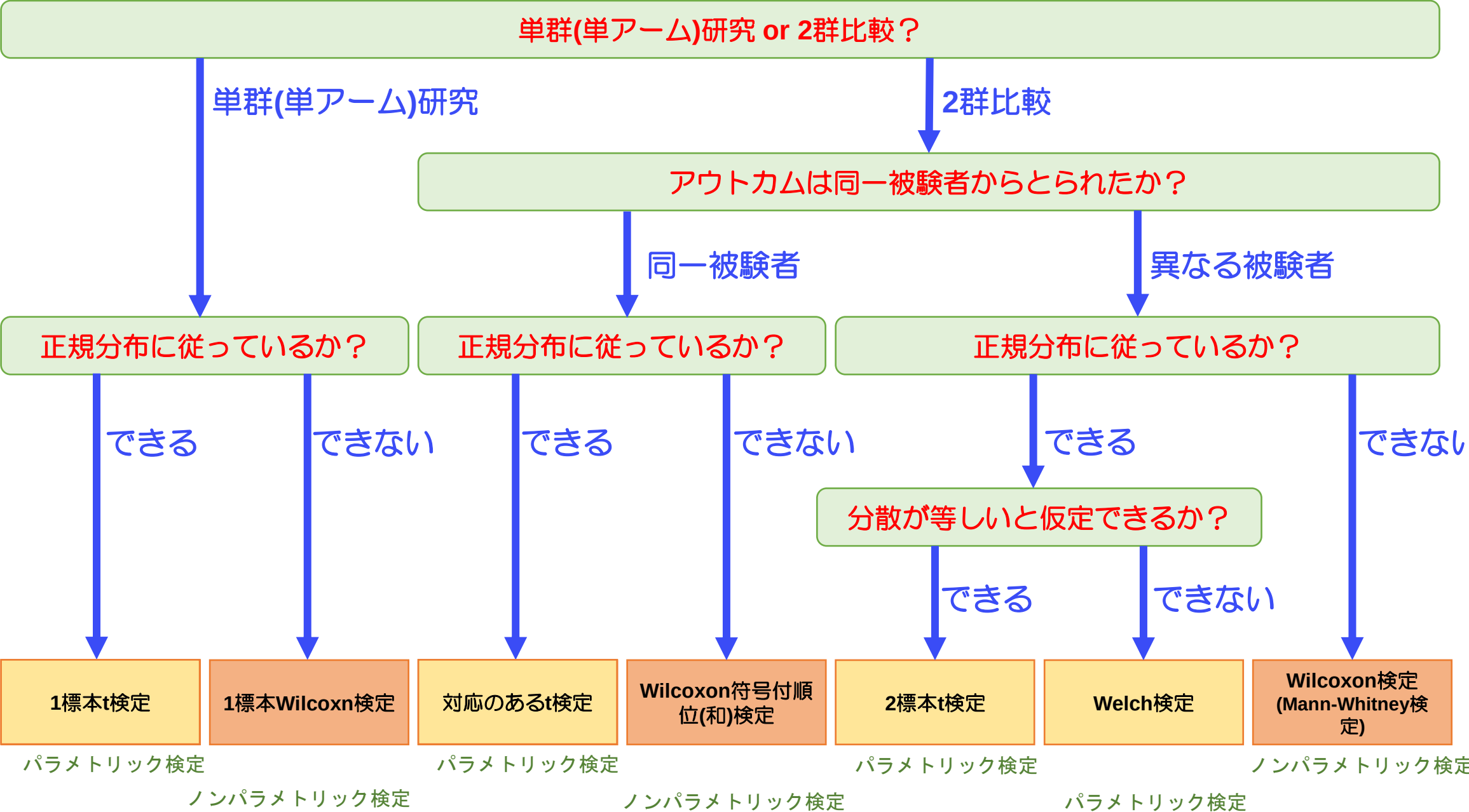
■ 生存時間データ (survival data)

がんの第III相試験などで主としてとり扱われるデータである。生存期間と中途打ち切(censoring)指標がペアでとり扱われる。Kaplan-Meierプロットを利用することが多い

バックアップ：諸種のデータに対する統計的検定

連続変数の場合

連続変数における統計的検定の取捨選択

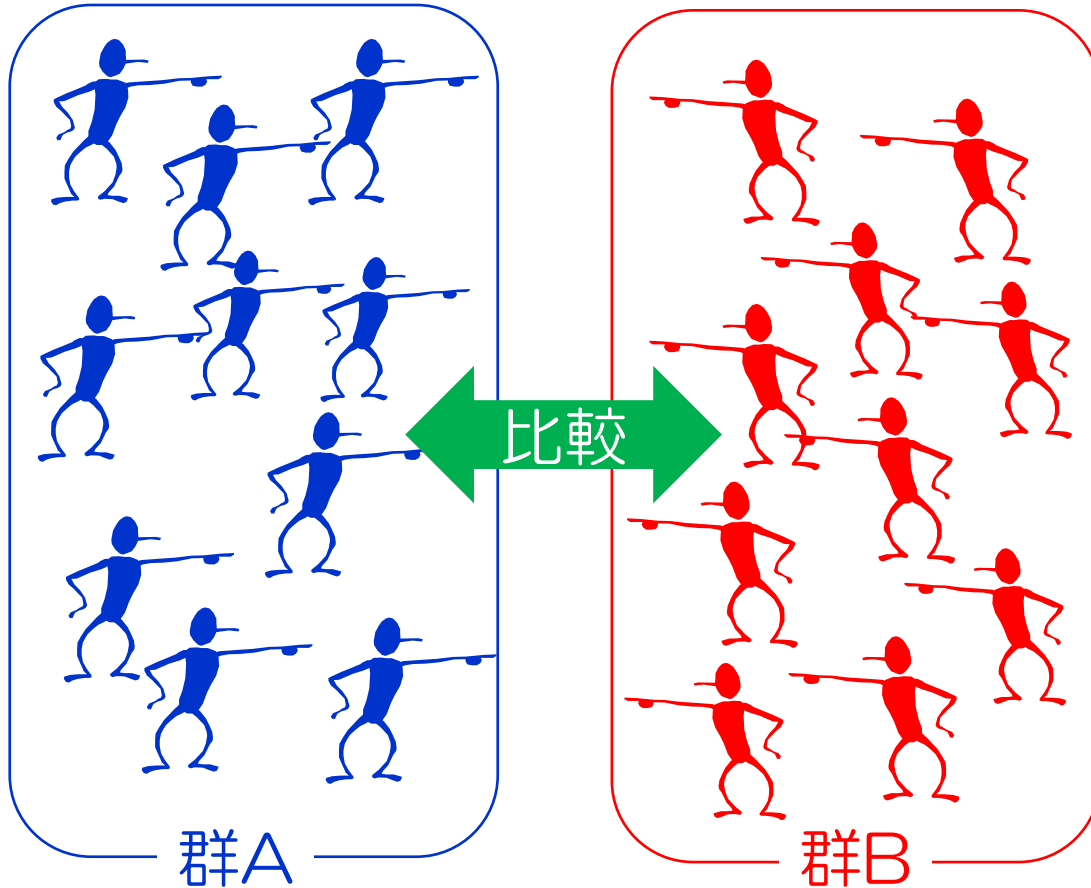


「対応がある／ない」とは何か？

対応のないデータ(独立2標本)

■例：無作為化比較試験

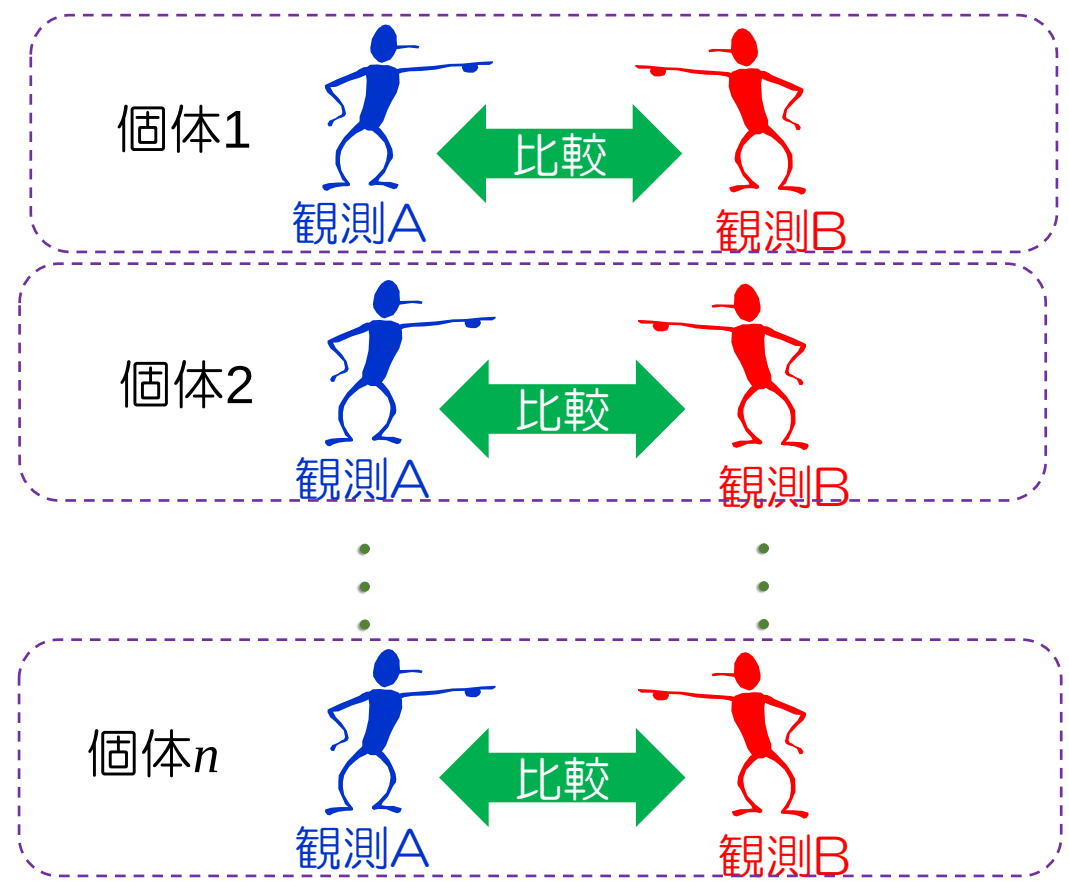
治療Aと治療Bを異なる被験者に実施して有効性を比較する。



対応のある標本

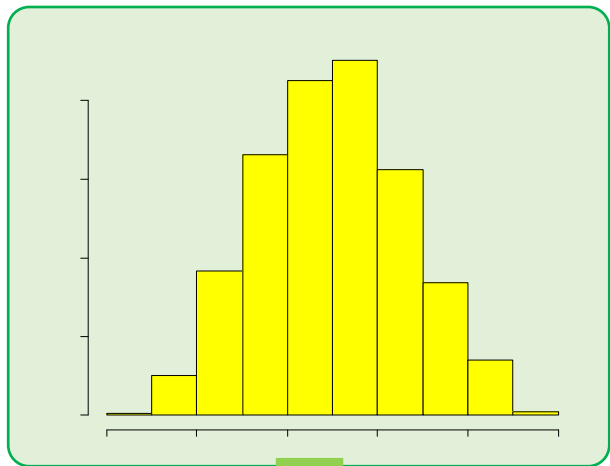
■例：クロスオーバー試験

同一被験者に対して、治療Aと治療Bに実施する。



2標本問題におけるパラメトリック検定とノンパラメトリック検定の取捨選択

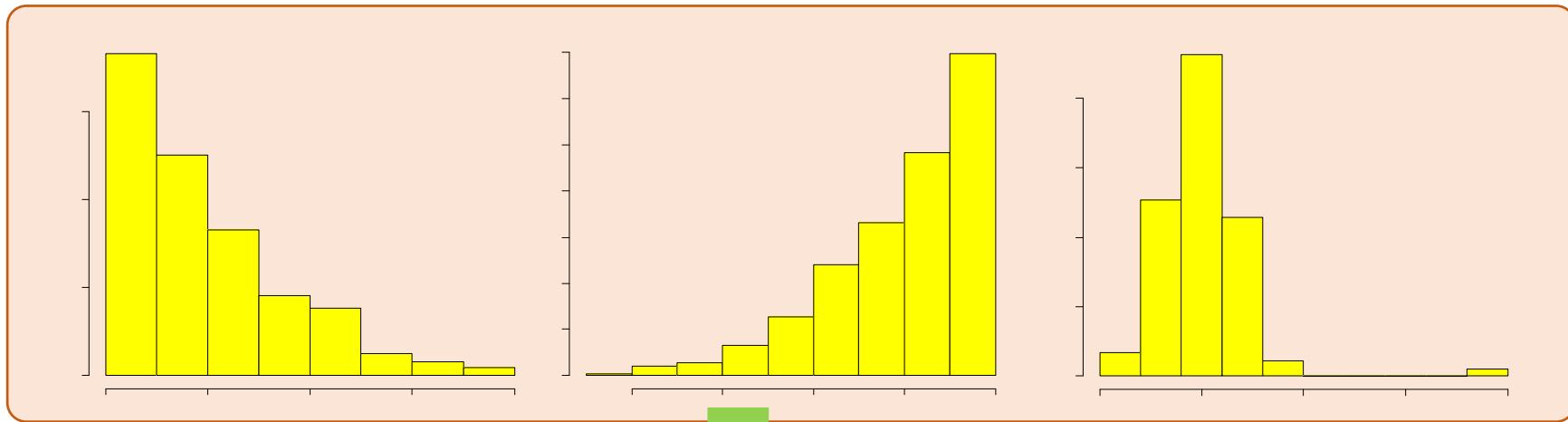
パラメトリック検定



2標本の場合

2標本t検定 (等分散)
or
~~Welch検定 (不等分散)~~

ノンパラメトリック検定



2標本の場合

Mann-Whitney-Wilcoxon検定
(Mann-Whitney検定, Wilcoxon検定)

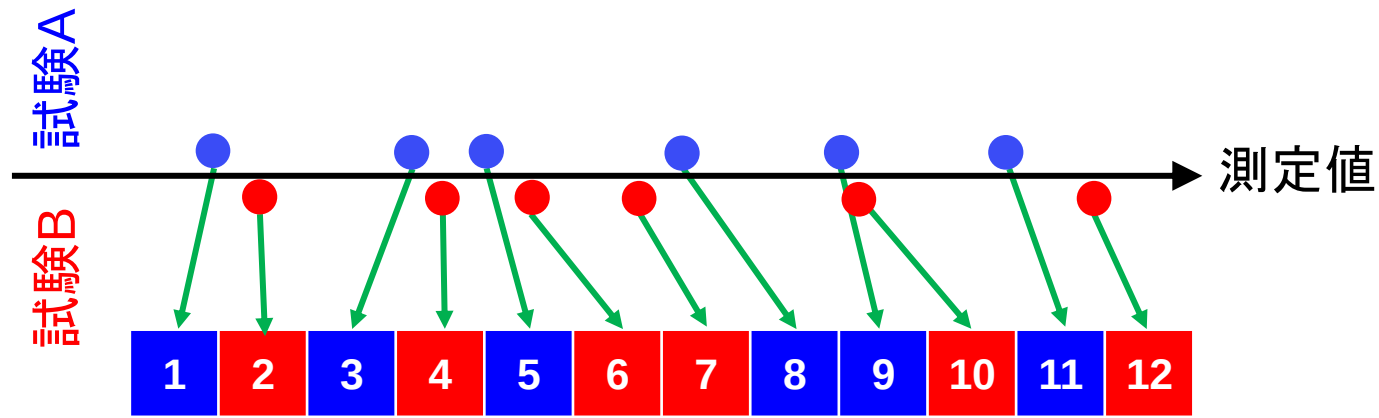
Welch検定に対する批判

2標本の検定の関心は、(1)母集団の違いに差があるのか、(2)平均値の差にあるのか、に大別される。関心の対象が(1)である場合には、不等分散であることを示すことができれば(等分散性の検定)、Welch検定を用いる必要は必ずしも存在しない。関心の対象が(2)である場合においても、試験結果の分散(標準偏差)に明らかな違いがなければ、2標本t検定で十分であることがいくつかの文献で指摘されている。また、母集団の分散が明らかに異なる場合には、母集団が正規分布に従っていないことが想定されるため、Mann-Whitney検定(Wilcoxon検定)などのノンパラメトリック検定を用いることが推奨される

Wilcoxon検定の模式図

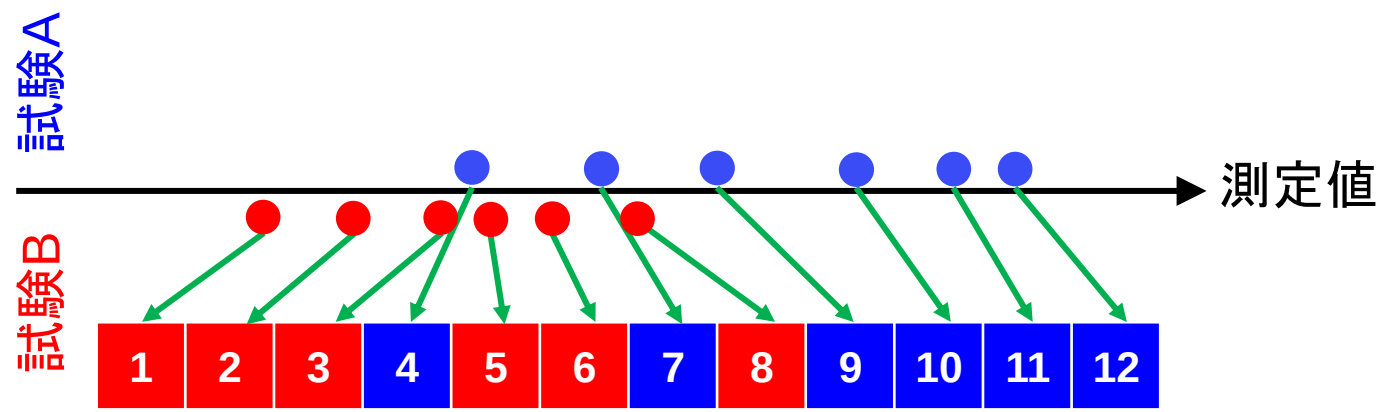
有意である状況

小さい順に並べ替えたときに、試験Aと試験Bの観測値が交互に並ぶ。



有意でない状況

小さい順に並べ替えたときに、試験Bが先にならび、その後で、試験Aの観測値が並ぶ。

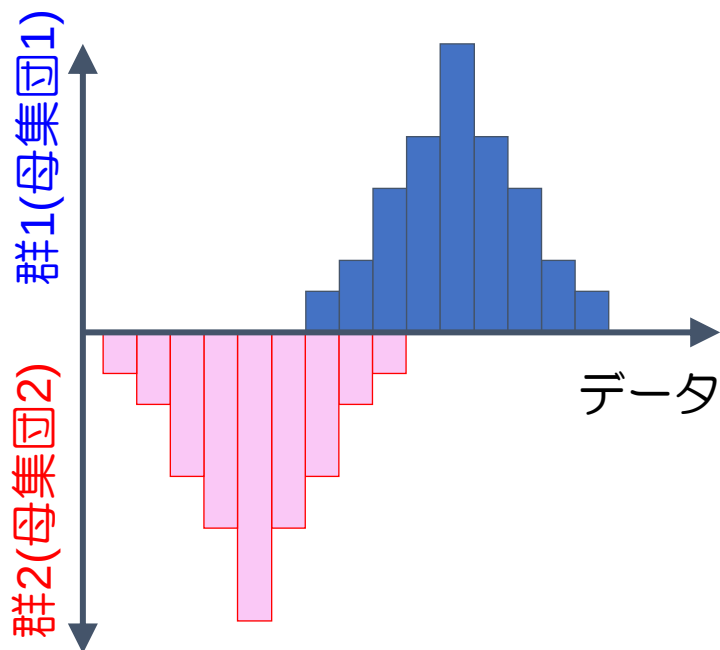


Wilcoxon検定では、小さい順に並べ替えたときのいずれかの群の順位の和に基づいて検定を行う。そのため、外れ値や分布の歪みの影響を受けることはない。

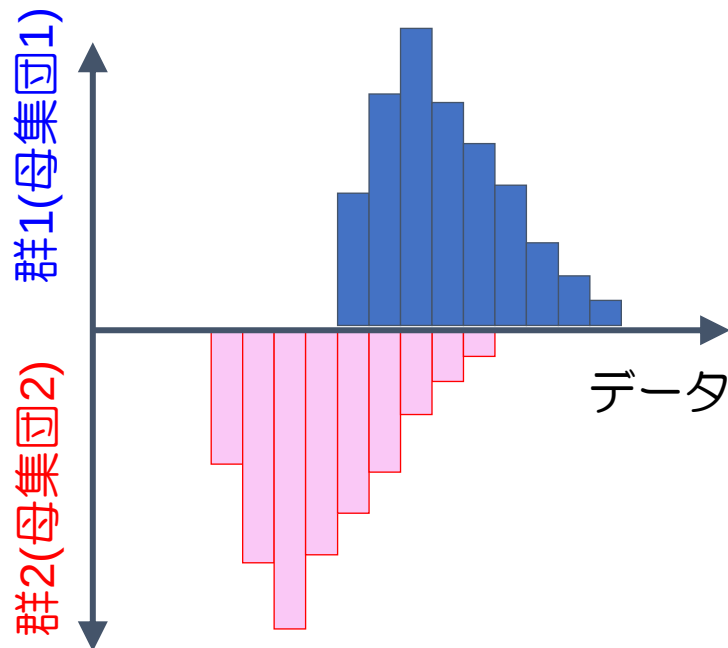
Warning!! Wilcoxon検定では相対的な位置関係を評価する

相対的な位置関係には様々な状況が考えられる。ここでは母集団1が母集団2よりも大きな状況を考え、ヒストグラムで表現する。

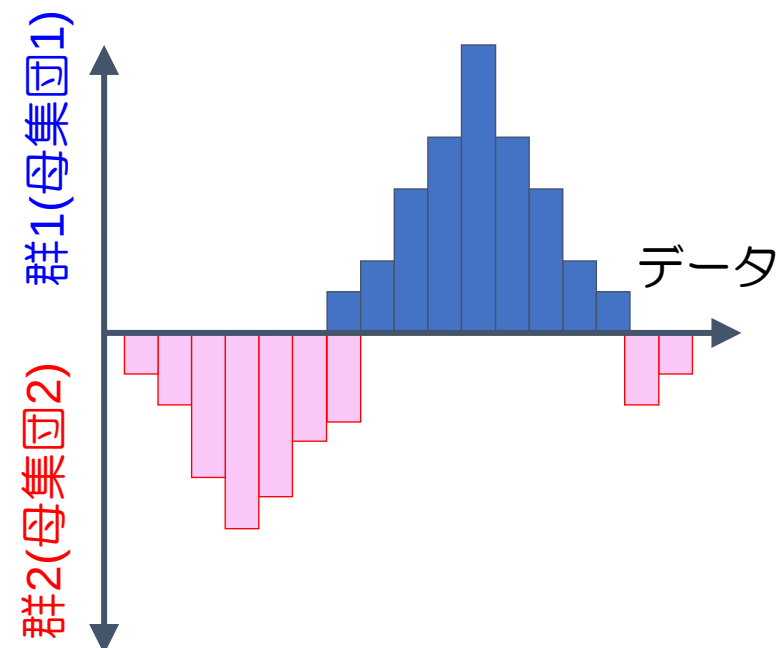
両群が対称な場合
(t検定による検定可能)



母集団が
歪んでいる場合



外れ値が
存在する場合



Wilcoxon検定では様々な母集団分布の状況で適用可能である。一方で「何を評価しているのかよくわからない」という状況になるため、t検定が使えない状況での代替と考えるべきである。

対応のある場合の比較：対応のあるt検定

各個体(被験者)に対して、「治療後－治療前」を計算し、その平均変化量に対して、0であるか否かを検定する。変化量が正規分布に従うとするならば、帰無仮説での平均変化量 $\mu_0=0$ としたときの、1標本t検定と計算していることと同じである。

H_0 ：治療後前後での検査値の平均変化量に違いはない。

➡ 変化量(治療後－治療前)の母平均 μ は0である。

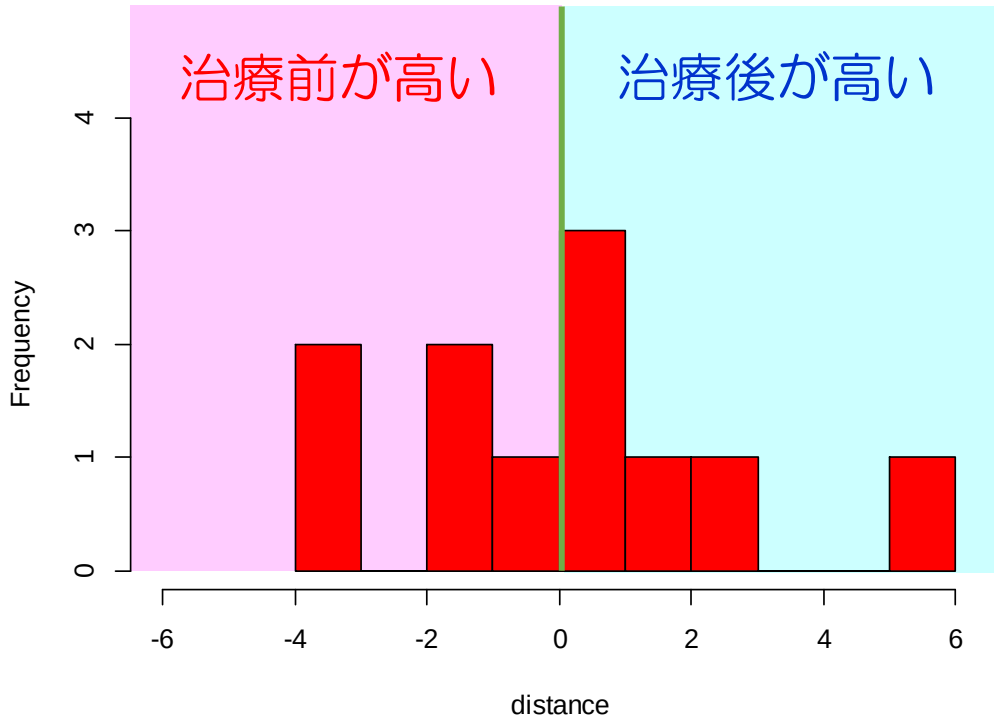
H_1 ：治療後の検査値のほうが治療前よりも大きい。

➡ 変化量(治療後－治療前)の母平均 μ は0より大きい。

正規分布に従わない場合の対応のあるデータの解析： Wilcoxon符号付順位検定

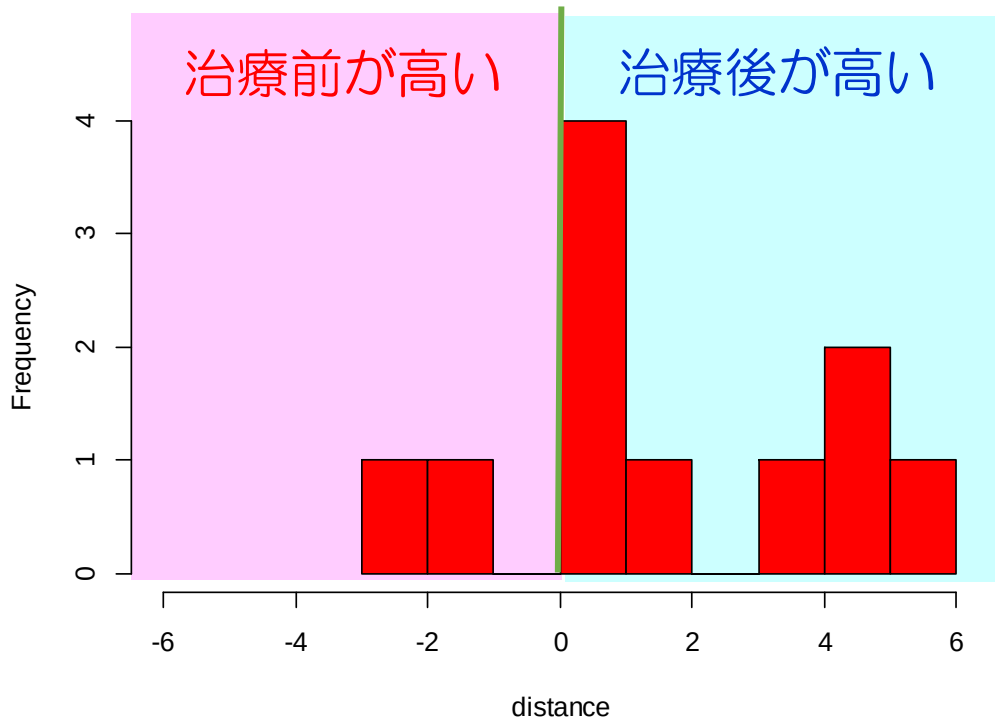
有意でない状況

No	前	後	差	No	前	後	差
1	5	4	-1	7	1	2	1
2	1	2	1	8	9	6	-3
3	7	8	1	9	4	3	-1
4	9	5	-4	10	2	2	0
5	3	9	6	11	3	6	3
6	2	4	2				



有意である状況

No	前	後	差	No	前	後	差
1	3	4	1	7	1	2	1
2	1	2	1	8	2	6	4
3	3	8	5	9	4	3	-1
4	4	5	1	10	5	2	-3
5	3	9	6	11	1	6	5
6	2	4	2				



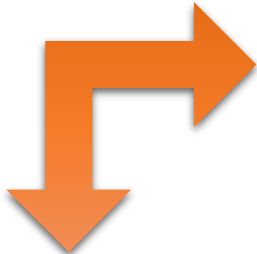
Wilcoxon符号付順位(和)検定では、青色の領域の数と赤色の領域の数の偏りを検定している。

分散分析の模式図

分散分析における仮説

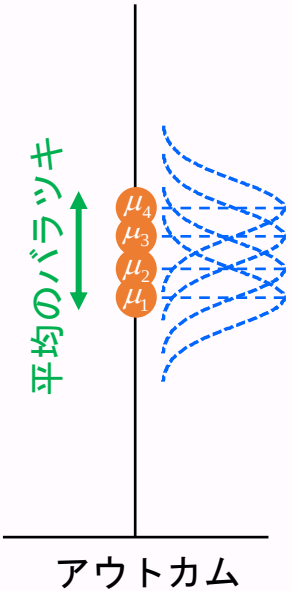
帰無仮説 H_0 : すべての群の母平均が等しい

対立仮説 H_1 : 帰無仮説 H_0 ではない



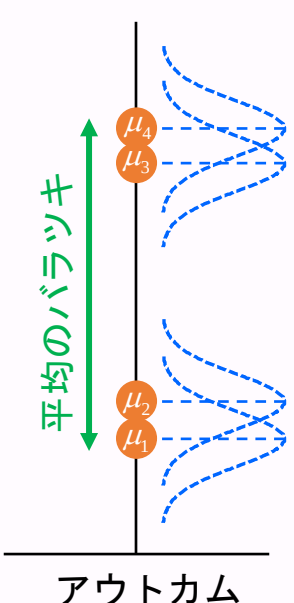
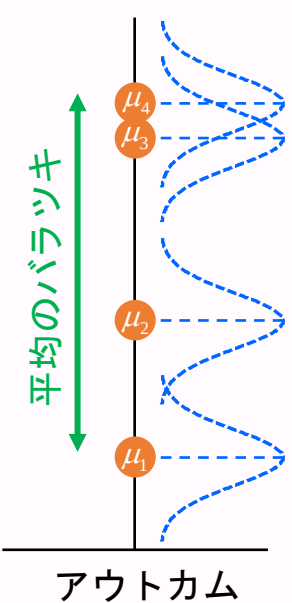
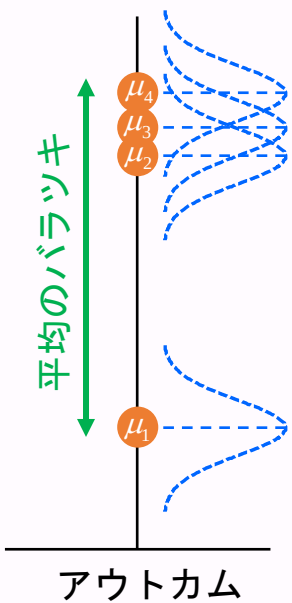
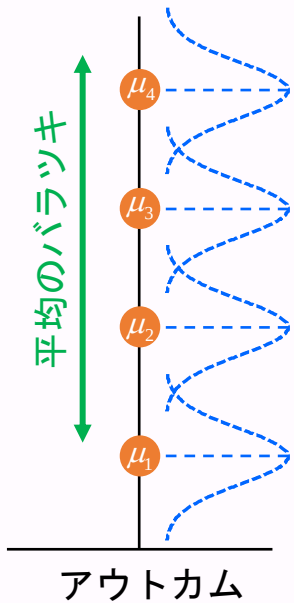
帰無仮説 H_0 が棄却できない(有意でない) 状況

平均値がほぼ同じ場合には、平均のバラツキ(=分散)が小さく、すべて同じであれば、0になる。



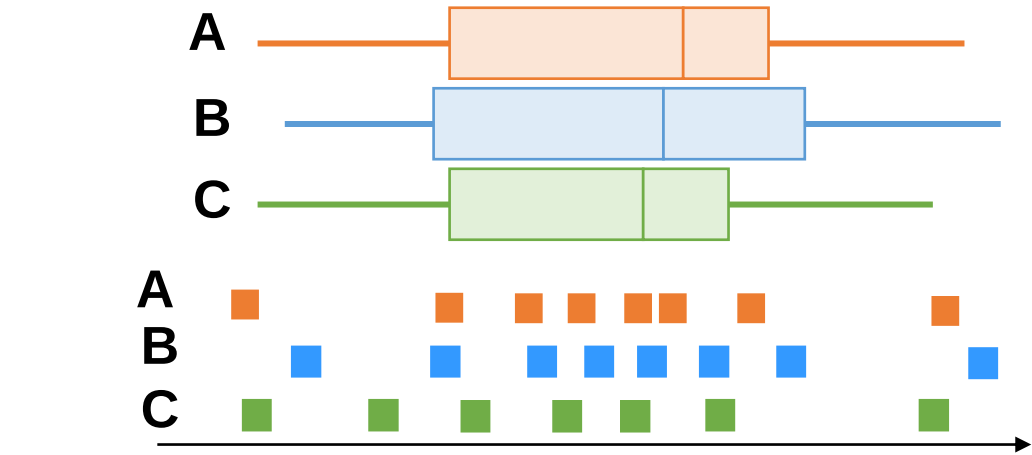
対立仮説 H_1 が正しい(有意)な状況

平均値が異なる場合には、平均のバラツキ(=分散)が大きい。

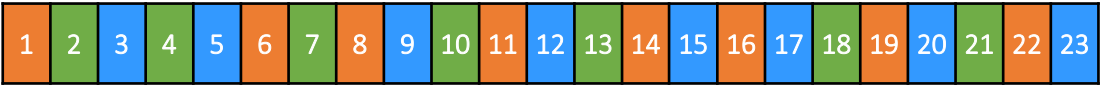


Kruskal-Wallis検定の概念図

□ 有意でない状況

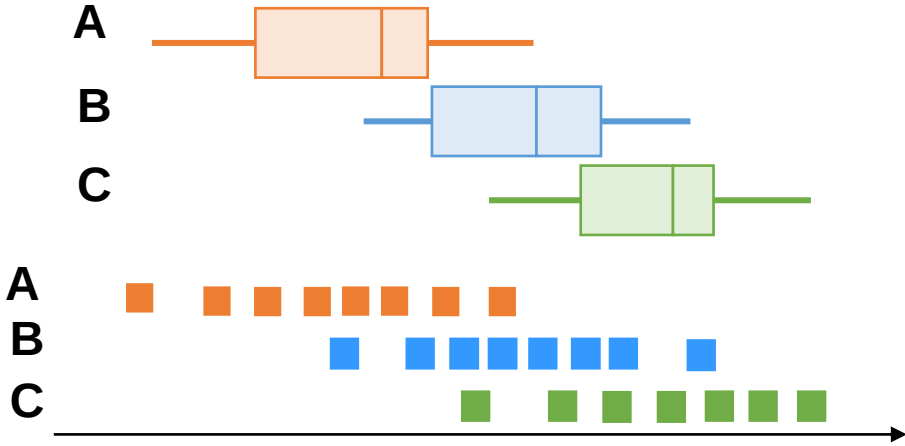


小さい順に並べると

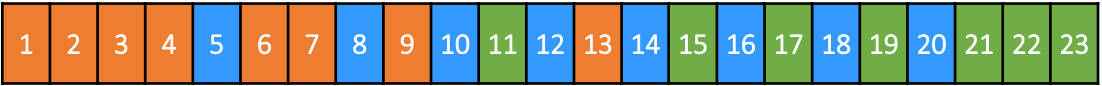


データをプールして順位をつける。そしてその
順位の平均値をとる。差がなければ、**順位の平均
値が3群ともほぼ同じになるはず**である。

□ 有意である状況



小さい順に並べると



データをプールして順位をつける。そしてその
順位の平均値をとる。差があれば、**順位の平均
値がいずれの群で違いがでるはず**である。

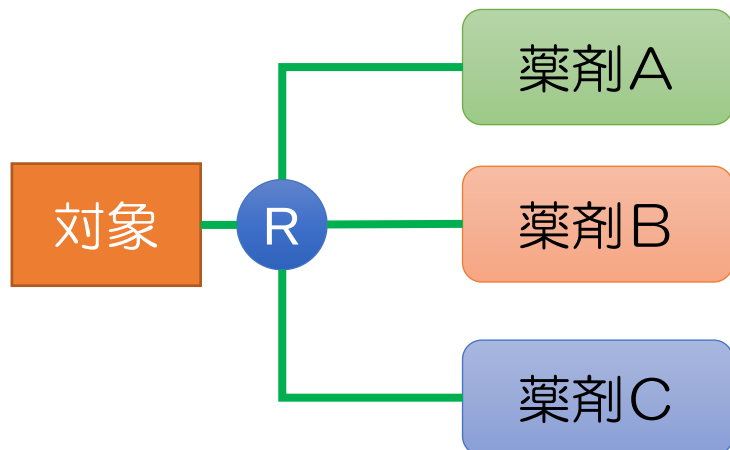
プールしたデータで順位をつけたときの群毎の平均値(順位和)を利用したのがKruskal-Wallis検定である。

分散分析の種類(1/4)：多標本データの比較



Clinical Question

神経障害性疼痛患者がランダムに割り付けられ、3種類の除痛薬(A, B, C)のいずれかが投与された。



Data structure

P：被験者

薬剤A

P P P P P P P P P P P P

薬剤B

P P P P P P P P P P P P

薬剤C

P P P P P P P P P P P P

3薬のVAS減少量に違いがあるか？

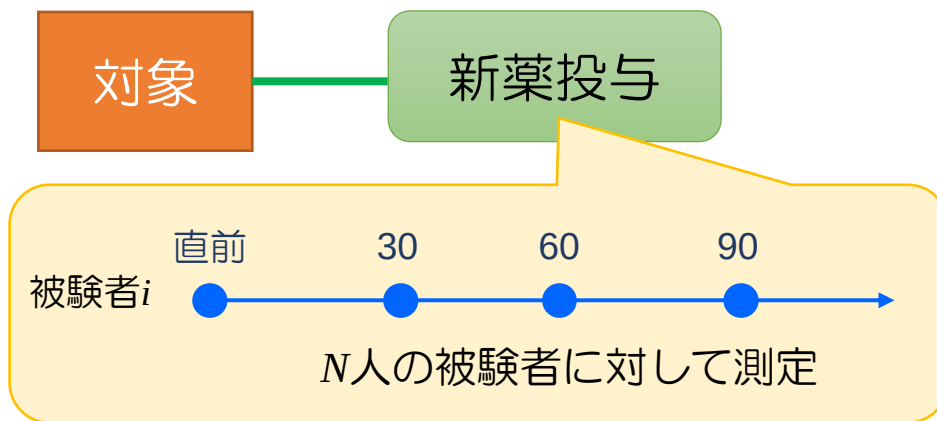
- パラメトリック検定：一元配置の分散分析
- ノンパラメトリック検定：Kruskal-Wallis検定

分散分析の種類(2/4)：繰り返し測定データの解析(1)

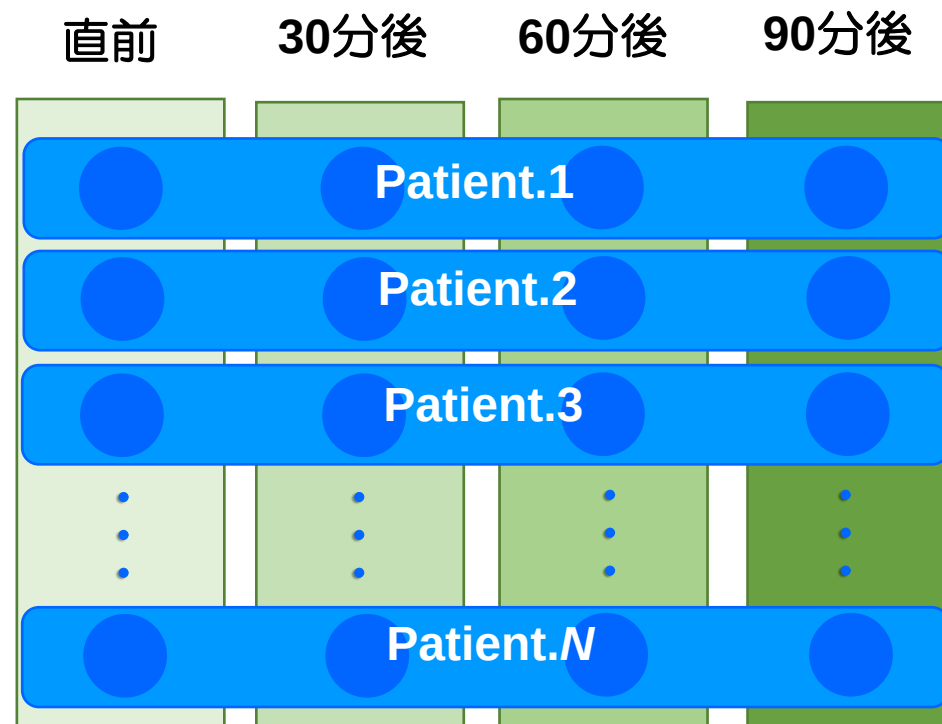


Clinical Question

N 名の神経障害性疼痛患者に対して、新薬の除痛薬を投与し、投与直前、30分後、60分後、90分後のVASの変化を調査した。



Data structure



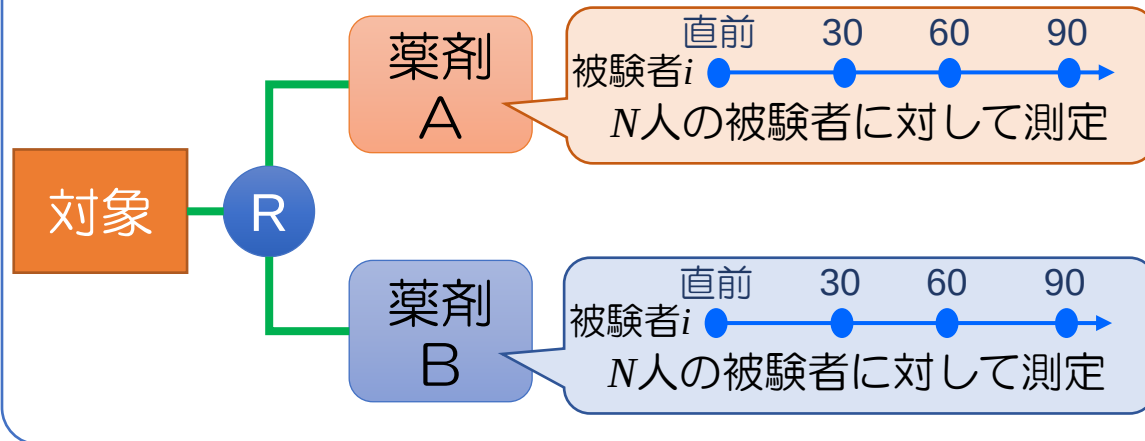
- パラメトリック検定：繰り返し測定分散分析
- ノンパラメトリック検定：Friedman検定

分散分析の種類(3)：繰り返し測定データの解析(2)

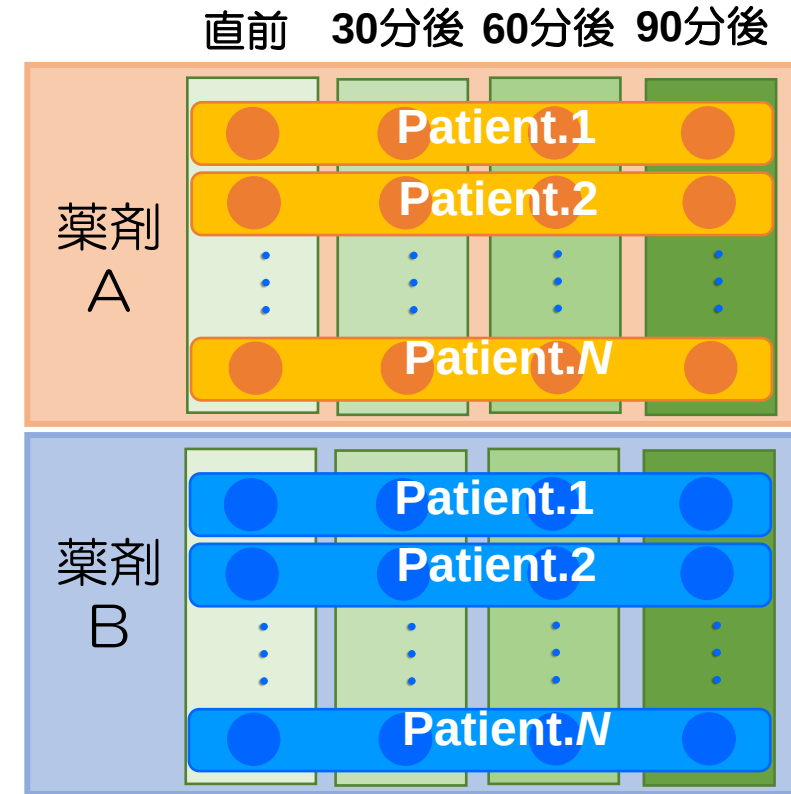


Clinical Question

神経障害性疼痛患者がランダムに割り付けられ，3種類の除痛薬(A,B)のいずれかが投与され，投与直前，30分後，60分後，90分後のVASの変化を調査した。



Data structure



- 2剤のVASに違いがあるか？
- VASに経時的変化があるか？
- (時間×薬剤の交互作用があるか？)

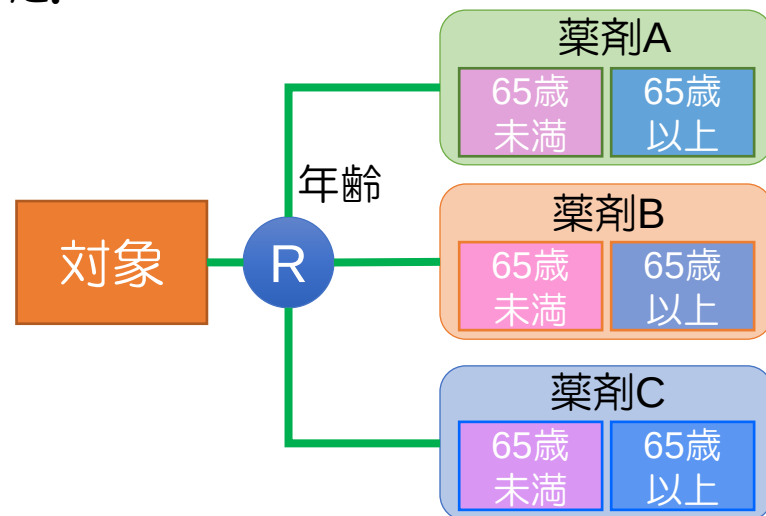
- パラメトリック検定：繰り返し測定分散分析
- ノンパラメトリック検定：存在しない

分散分析の種類(4)：多元配置の分散分析



Clinical Question

神経障害性疼痛患者がランダムに割り付けられ、3種類の除痛薬(A, B, C)のいずれかが投与された。この試験では、年齢(65歳以上, 65歳未満)を割付調整因子とした。



Data structure

P：被験者

	65歳未満	65歳以上
薬剤A	P P P P P P P P P P	P P P P P P P P P P P
薬剤B	P P P P P P P P P P P P	P P P P P P P P P P P
薬剤C	P P P P P P P P P P	P P P P P P P P P

- 年齢で調整(adjust)したときの3薬のVAS減少量に違いがあるか？
- (年齢×薬剤の交互作用効果があるか？)

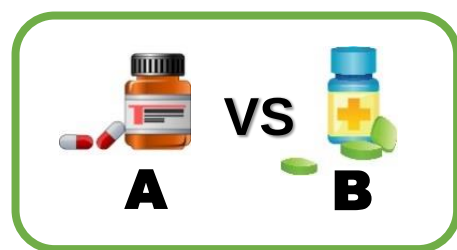
■ パラメトリック検定：多元配置の分散分析

■ ノンパラメトリック検定：存在しない

多重比較について

3群以上の比較を行う場合などに、「多重比較(multiple comparison)」を行わなければならない。
いま、3種類の薬剤の有効性を考える。

有意水準とは「有意差がないのに有意差があると誤る確率(第1種の過誤)」をあらわす



検定

有意水準
 $\alpha=0.05$

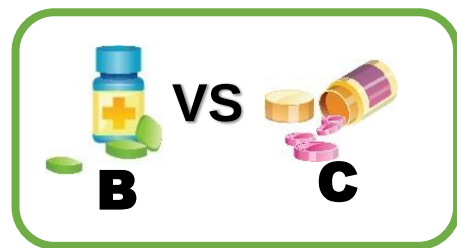
有意差がないことを正しく判断できる確率
 $1-\alpha=1-0.05=0.95$



検定

有意水準
 $\alpha=0.05$

有意差がないことを正しく判断できる確率
 $1-\alpha=1-0.05=0.95$



検定

有意水準
 $\alpha=0.05$

有意差がないことを正しく判断できる確率
 $1-\alpha=1-0.05=0.95$

3回の検定の中で1度でも有意差がないのに有意差があると誤る確率

$$1-0.95^3 = 0.142$$

誤りの確率が増大してしまう。これを補正するのが多重比較法である

バックアップ：諸種のデータに対する統計的検定

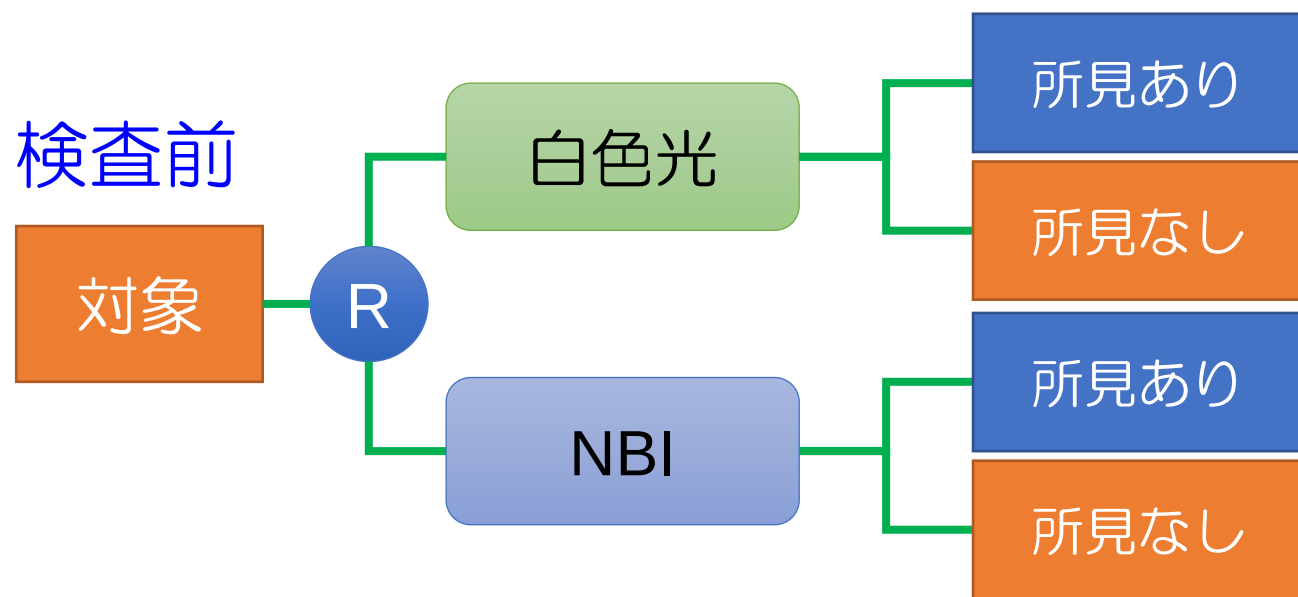
2値データ／カテゴリーカルデータの場合

カイ2乗検定およびFisherの正確検定の動機



Clinical Question

大腸内視鏡検査において， NBI検査を実施した被験者群と白色光検査を実施した被験者群で無作為化比較試験を実施し， 1個以上の鋸歯状病変の検出割合を比較した．



	所見有り	所見無し	計
NBI	204 (51.1%)	195 (48.9%)	399
白色光	158 (39.4%)	243 (60.6%)	401
計	362 (45.2%)	438 (54.2%)	800

統計的方法

検定：カイ2乗検定(標本サイズ多数)
Fisherの正確検定(標本サイズ少数)

要約：各群の所見有りの割合
オッズ比及び95%信頼区間

カイ2乗検定の意味

消化器癌(上部消化器, 下部消化器)に対する外科治療に対する縫合術(真皮縫合術, スキンステープラー)の比較の結果を以下に示す. 評価項目は, 創合併症の発現である(Tsujinaka et al., Lancet, 2013).

カイ2乗検定とは, 縦の項目(変数)と横の項目(変数)に関連性がない状況(縫合術によって創合併症が違いがない)でのクロス集計表からの乖離を評価する.

観測データ			期待度数(関連性がない状況)		
	創合併症あり	創合併症なし	創合併症あり	創合併症なし	
真皮縫合術	47	511	$\frac{106 \times 558}{1072} = 55.2$	$\frac{966 \times 558}{1072} = 502.8$	乖離 ↓
スキンステープラー	59	455	$\frac{106 \times 514}{1072} = 50.8$	$\frac{966 \times 514}{1072} = 463.2$	

	創合併症あり	創合併症なし
真皮縫合術	$47 - 55.2 = -8.2$ $(-8.2)^2 / 55.2 = 1.218$	$511 - 502.8 = 8.2$ $(8.2)^2 / 502.8 = 0.134$
スキンステープラー	$59 - 50.8 = 8.2$ $(8.2)^2 / 50.8 = 1.324$	$455 - 463.2 = -8.2$ $(-8.2)^2 / 463.2 = 0.145$

合計はカイ2乗値と呼ばれ, カイ2乗値を用いることでp値を計算することになる.

Fisherの正確検定の意味 (1/2)

2種類の抗癌剤に対するパイロット試験のクロス集計表

	奏効	非奏効	計
新規抗癌剤	12 (54.5%)	10 (45.5%)	22
既存抗癌剤	3 (20.0%)	12 (80.0%)	15
計	15 (40.5%)	22 (59.5%)	37

Fisherの正確検定では、周辺度数(赤色の部分)を固定したときに、現在のクロス集計表の状況が帰無仮説のもとでどれぐらいの確率で起こり得るかを評価する。

	抗腫瘍効果		計
	奏効	非奏効	
新薬	O_{11}	$O_{12}=n_1-O_{11}$	n_1
既存薬	$O_{21}=m_1-O_{11}$	$O_{22}=N-O_{11}-O_{21}-O_{12}$	n_2
計	m_1	m_2	N

周辺度数が固定されると、一つの度数(O_{11})が決まれば、残りの度数が定まる。

Fisherの正確検定の意味 (2/2)

確率 = 0.0000			確率 = 0.0000			確率 = 0.0000			確率 = 0.0001		
	奏効	非奏効		奏効	非奏効		奏効	非奏効		奏効	非奏効
新規	0	22	新規	1	21	新規	2	20	新規	3	19
既存	15	0	既存	14	1	既存	13	2	既存	12	3
確率 = 0.0011			確率 = 0.0084			確率 = 0.0399			確率 = 0.1172		
	奏効	非奏効		奏効	非奏効		奏効	非奏効		奏効	非奏効
新規	4	18	新規	5	17	新規	6	16	新規	7	15
既存	11	4	既存	10	5	既存	9	6	既存	8	7
確率 = 0.2197			確率 = 0.2659			確率 = 0.2074			確率 = 0.1028		
	奏効	非奏効		奏効	非奏効		奏効	非奏効		奏効	非奏効
新規	8	14	新規	9	13	新規	10	12	新規	11	11
既存	7	8	既存	6	9	既存	5	10	既存	4	11
確率 = 0.0314			確率 = 0.0056			確率 = 0.0005			確率 = 0.0000		
	奏効	非奏効		奏効	非奏効		奏効	非奏効		奏効	非奏効
新規	12	10	新規	13	9	新規	10	12	新規	11	11
既存	3	12	既存	2	13	既存	5	10	既存	4	11

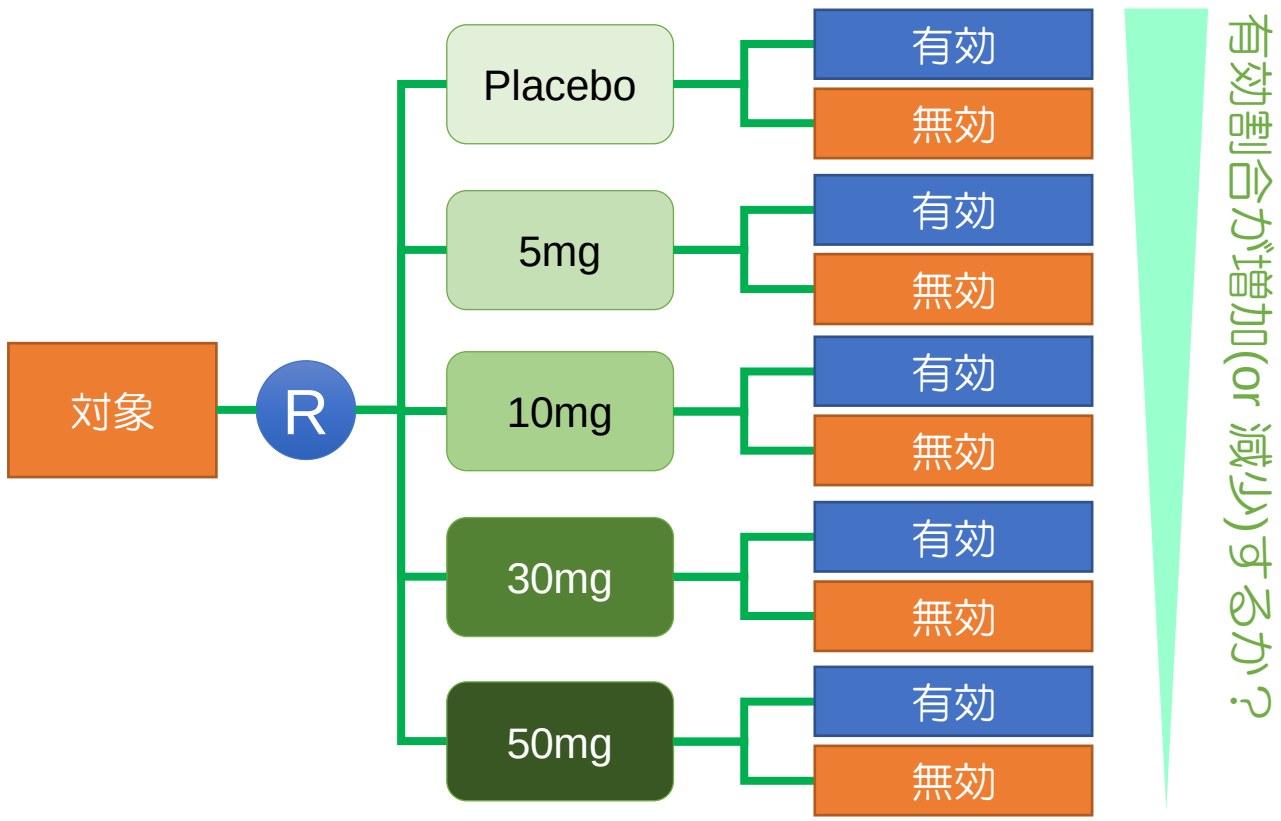
p値=0.0000 + 0.0000 + 0.0000 + 0.0001+ 0.0011 + 0.0084 + 0.0314 + 0.0056 + 0.0005 + 0.0000 = 0.0471

Cochran-Armitage検定の動機

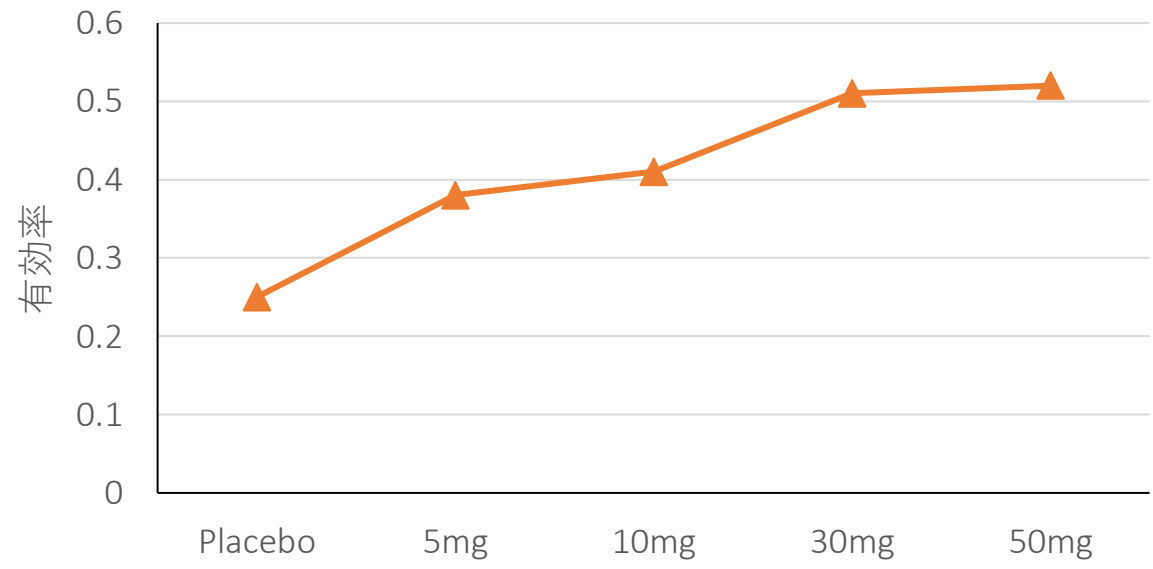


Clinical Question

ある薬剤の用量反応関係を調査するために， Placebo, 5mg, 10mg, 30mg, 50mgの投与群とそのときの効果の有無を調べた。



	Placebo	5mg	10mg	30mg	50mg
有効	79	81	169	318	379
無効	235	132	243	305	204



統計的方法

検定：Cochran-Armitage検定

要約：有効割合

Cochran-Armitage検定の略説

2種類の薬剤の投与量と有効性を評価したときのクロス集計表を下表に示す。

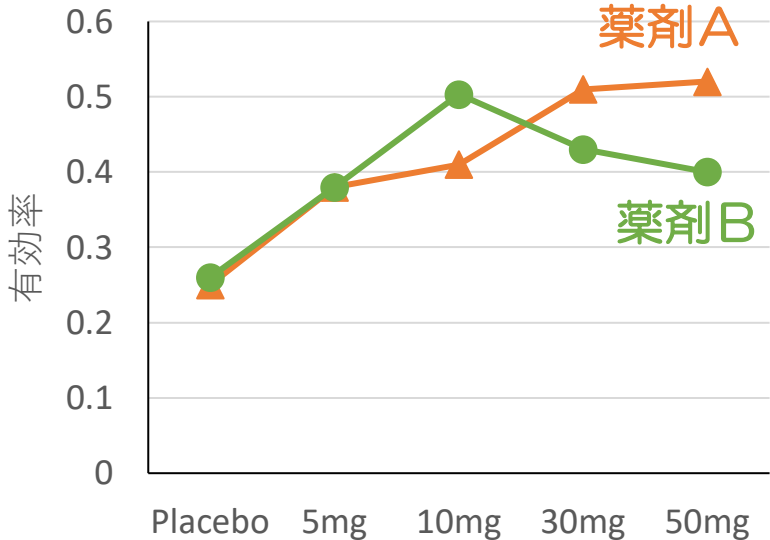
薬剤A

	Placebo	5mg	10mg	30mg	50mg
有効	79	81	169	318	379
無効	235	132	243	305	204

薬剤B

	Placebo	5mg	10mg	30mg	50mg
有効	82	81	220	268	233
無効	232	132	214	355	350

いずれの薬剤でもカイ2乗検定では有意($p<0.001$)である。すなわち、投与量によって有効性に違いが認められる。



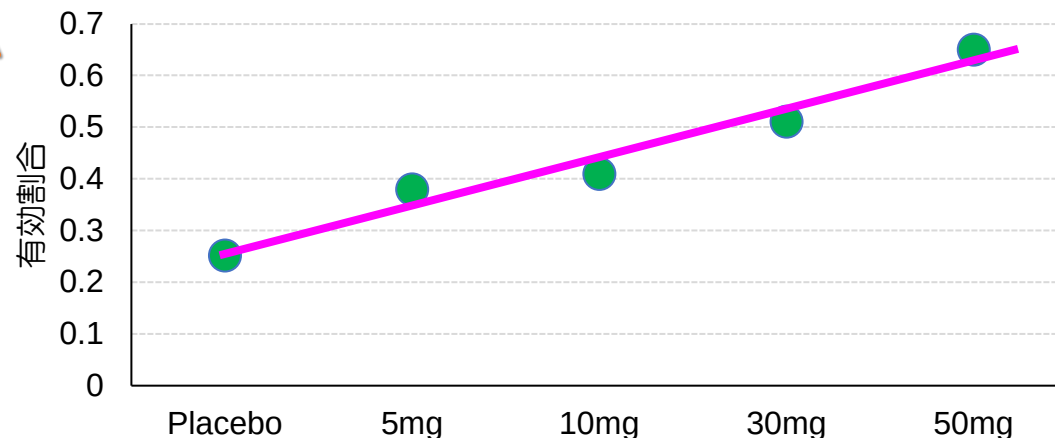
薬剤Aの場合には投与量が増加するにつれて有効性が増加しているものの、薬剤Bでは減少している。



このような研究を計画する理由としては、投与量による有効性の増加傾向を評価することにある。

Cochran-Armitage検定

薬剤A



有効割合

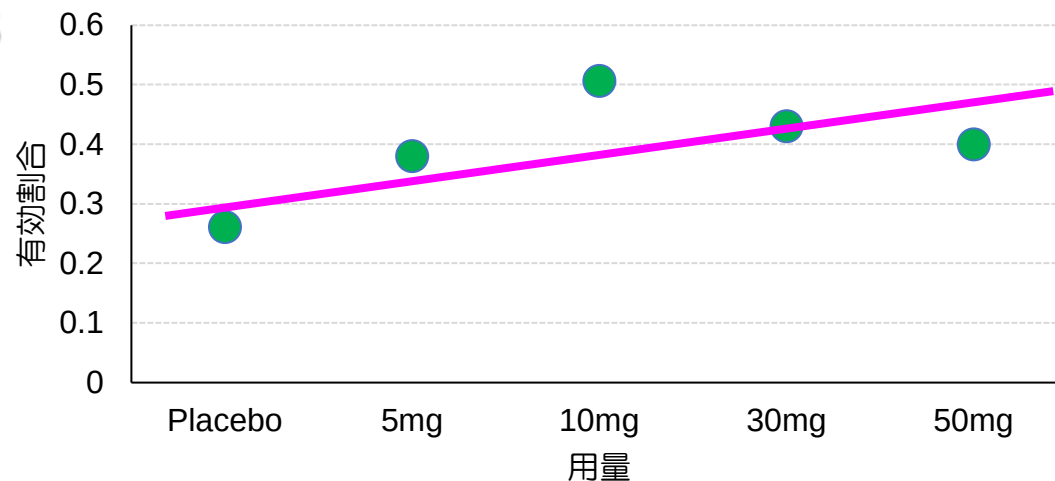
0.252 0.380 0.410 0.510 0.650

p値は

< 0.0001

であることから、投与量によって有効率の上昇傾向が認められた

薬剤B



有効割合

0.261 0.380 0.507 0.430 0.400

p値は

0.1796

であることから、投与量によって有効率の上昇傾向が認められなかった



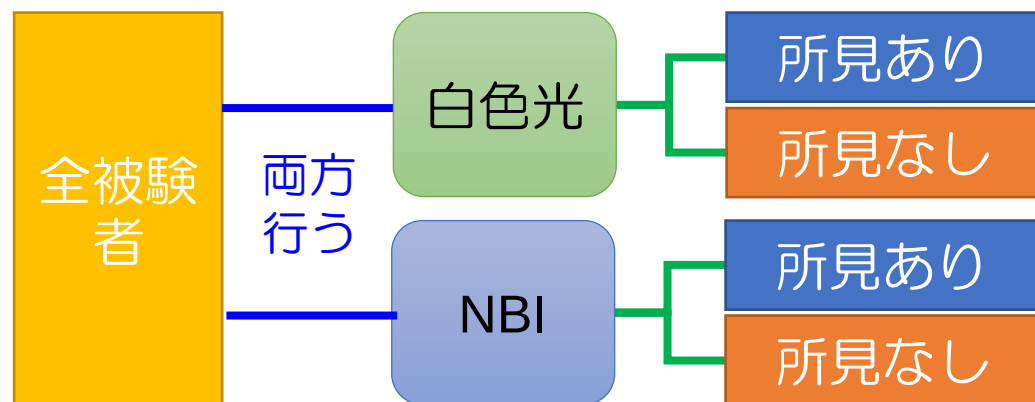
Cochran-Armitage検定では、直線を当てはめた上で、その直線に意味があるか否か(つまり傾向変化があるか)を検定する。

対応のあるクロス集計表の動機



Clinical Question

大腸内視鏡検査において，各被験者に対してNBI検査と白色光検査を実施し，1個以上の鋸歯状病変の検出割合を比較した．



		NBI		合計
		あり	なし	
白色光	あり	85	10	95
	なし	18	61	79
合計		103	71	174

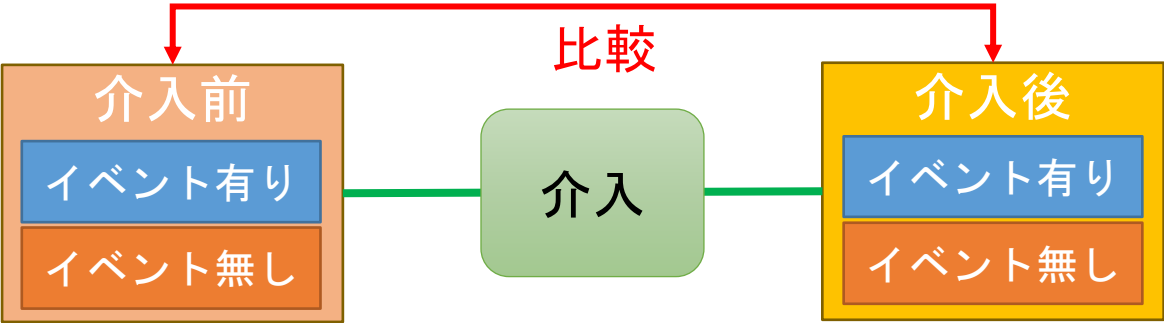
統計的方法

検定：McNemar検定

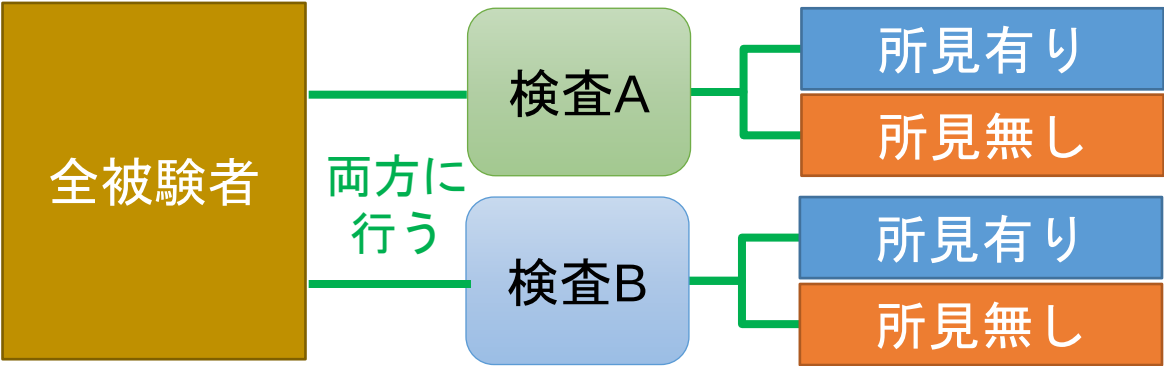
要約：対応のあるクロス集計表
全体パーセント

対応のあるクロス集計表を用いる幾つかのシチュエーション

Scenario.1
介入前後でのイベントの有無を評価する場合



Scenario.2
同一被験者に検査を実施する場合



Scenario.3
クロスオーバー試験の場合



McNemar検定の略説

白色光とNBIとの診断能を比較するために患者166例について二つの検査法を実施した

		白色光		合計
		所見あり	所見なし	
NBI	所見あり	85 無	18 NBI(良)	103
	所見なし	10 白色光(良)	61 無	71
合計		95	79	174

白色光で所見なしでNBIで所見ありの割合 (■のセルの割合)

$$\text{割合} = \frac{18}{174} = 0.1034$$


白色光で所見ありでNBIで所見なしの割合 (■のセルの割合)

$$\text{割合} = \frac{10}{174} = 0.0575$$

😊 二つの割合を比較することで評価するのがMcNemar検定

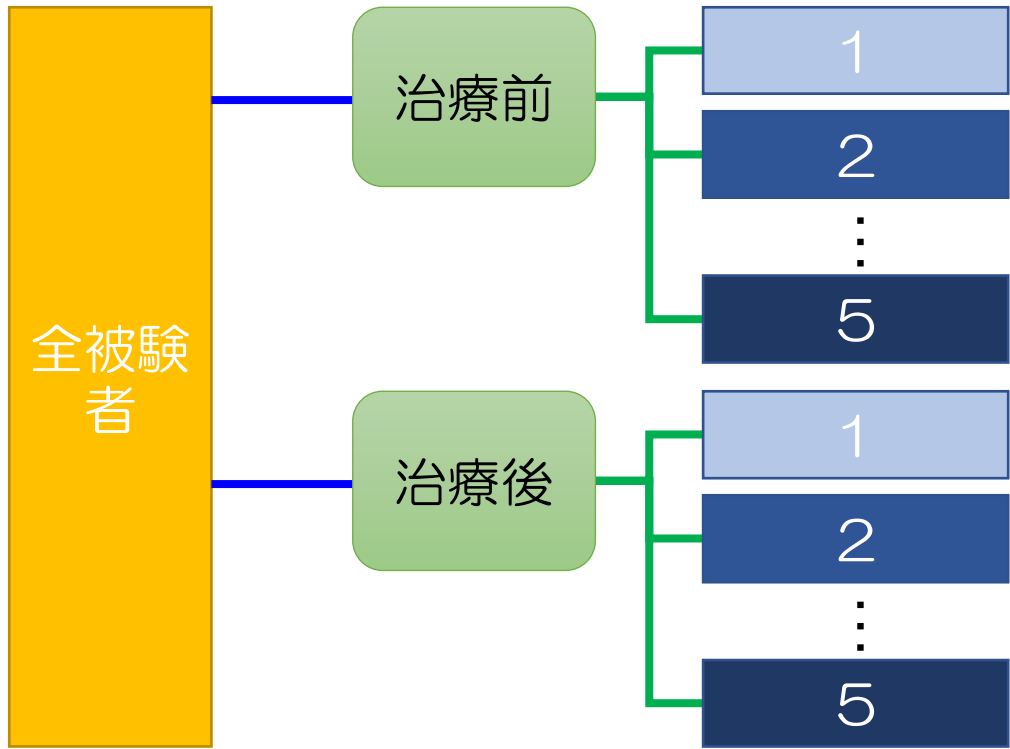
その結果、**p値は0.185**なので、有意でなかった。つまり、内視鏡検査の違いによる診断に差異があるとはいえなかった。

補足：McNemar検定が用いられるもう一つの状況



Clinical Question

あるペインクリニックにおいて、新たな治療法の効果が検討された。結果は、治療前後での5段階評価での評点によって実施された。



		治療後					計
		5	4	3	2	1	
治療前	5	8	3	1	0	1	13
	4	9	10	8	1	1	29
	3	3	16	24	3	1	47
	2	2	5	10	9	2	28
	1	0	3	6	9	5	23
計		22	37	49	22	10	140

統計的方法

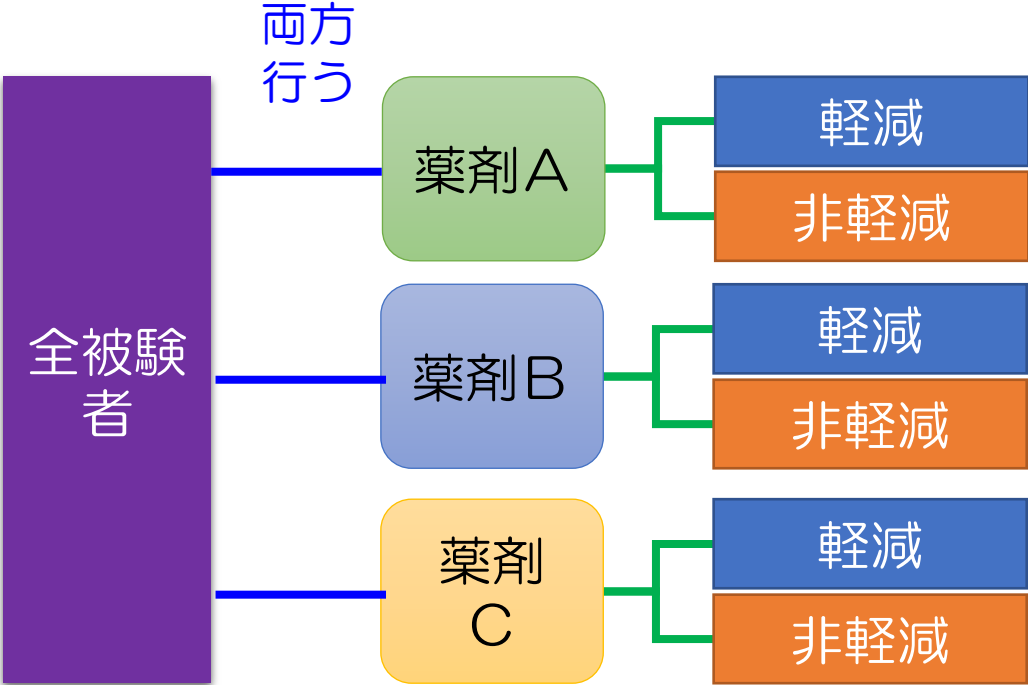
検定：McNemar検定
要約：対応のあるクロス集計表
全体パーセント

3群以上の比較の場合にはどうなるか？：CochranのQ検定



Clinical Question

神経障害性疼痛患者に対して，3種類の除痛薬 (A,B,C)を投与したときの痛みの軽減の有無を調べるためのチェンジオーバー試験を実施した。



		B	
		あり	なし
A	あり	5	1
	なし	4	2

		C	
		あり	なし
A	あり	2	4
	なし	1	5

		C	
		あり	なし
B	あり	3	6
	なし	0	3

統計的方法

検定：CochranのQ検定， McNemar検定の多重比較
要約：対応のあるクロス集計表
全体パーセント

バックアップ：諸種のデータに対する統計的検定

生存期間データの場合

生存期間を取り扱ううえで重要な測度(関数)

■ 生存時間関数

個体が少なくとも時点 t で生存する確率

時点 t まで生存した割合。年次生存割合などは、生存時間関数を用いて計算する

■ ハザード関数

個体が時点 t まで生存しているという仮定のもとで微小区間 $t \leq T < t + \Delta t$ で死亡する確率

時点 t における死亡リスク。症例数設計などは、ハザード関数を用いて計算する

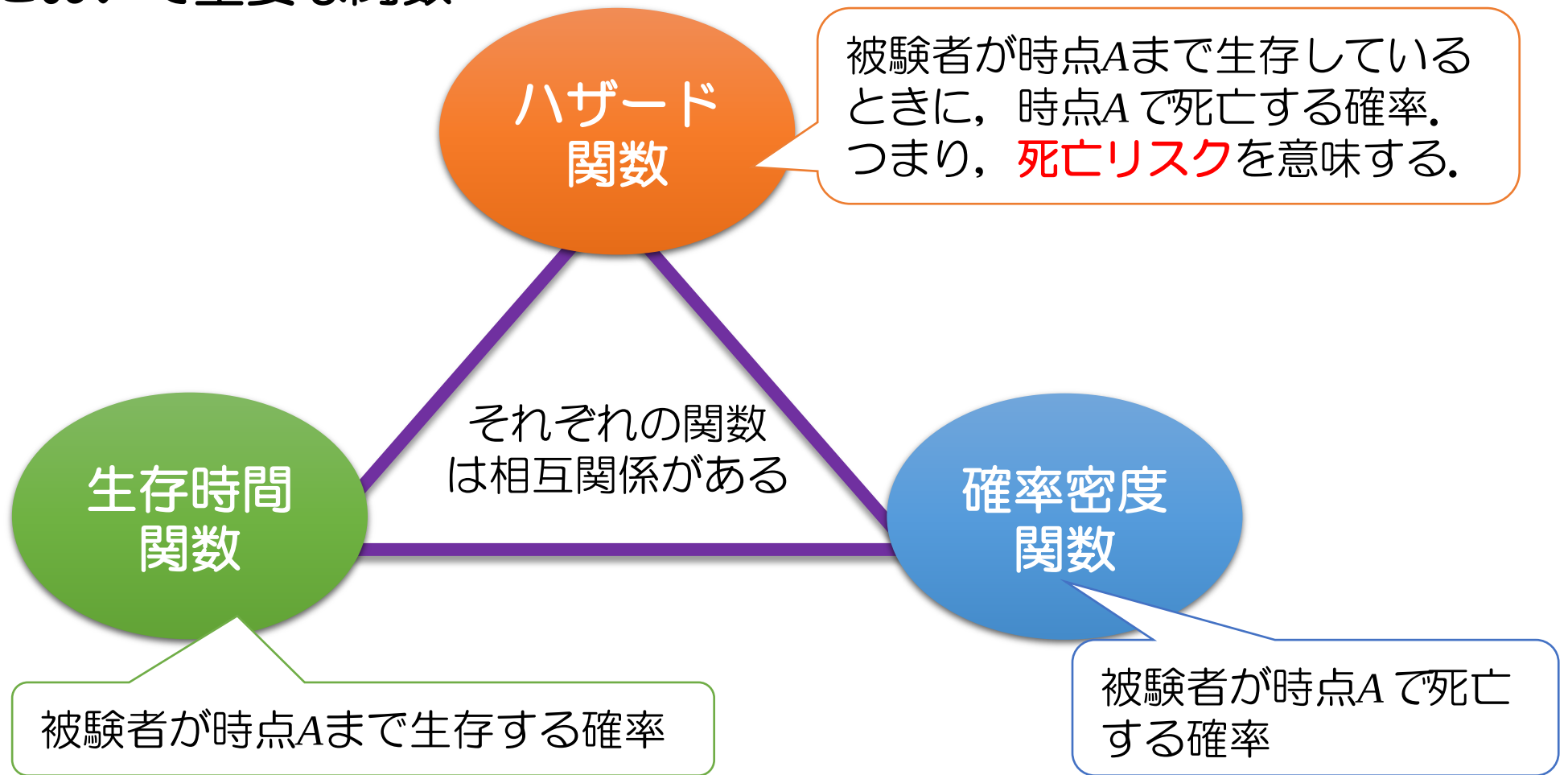
■ 確率密度関数 (probability density function)

個体が微小区間 $t \leq T < t + \Delta t$ 内で死亡する確率

■ 累積ハザード関数 (cumulative hazard function)

ハザード関数の累積値

生存時間解析において重要な関数

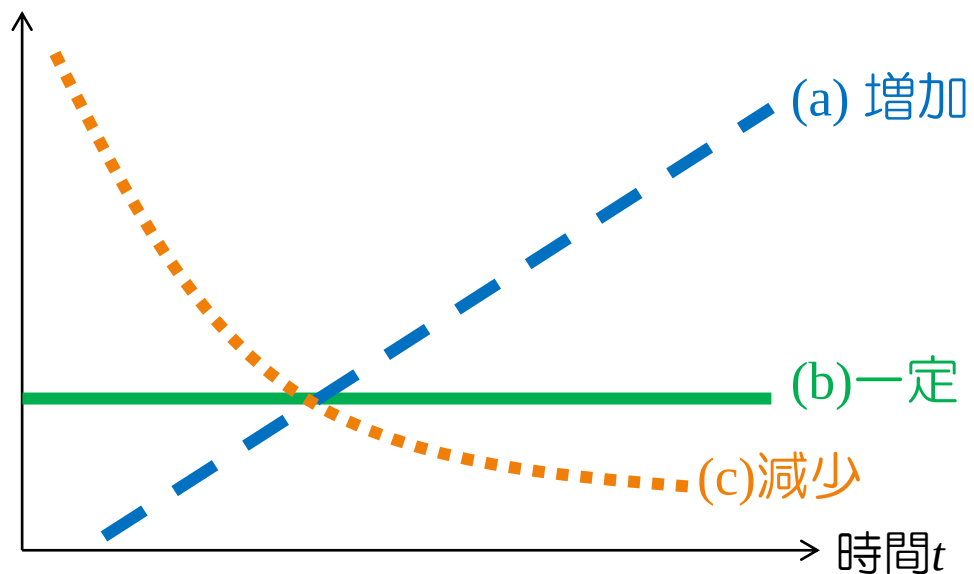


確率密度関数，確率密度関数，生存時間関数は相互関係があり，どれか1個が決まれば，すべてが決定できる。

ログランク検定ではハザードを評価しているのに，生存時間曲線を描写してグラフを描写している。これは，3種類の関数は相互関係があるため。

ハザードの意味

ハザード関数 $h(t)$



(a) の状況

登録時点では死亡(あるいはイベント)のリスクは低いものの時間経過とともに増大することを意味する。

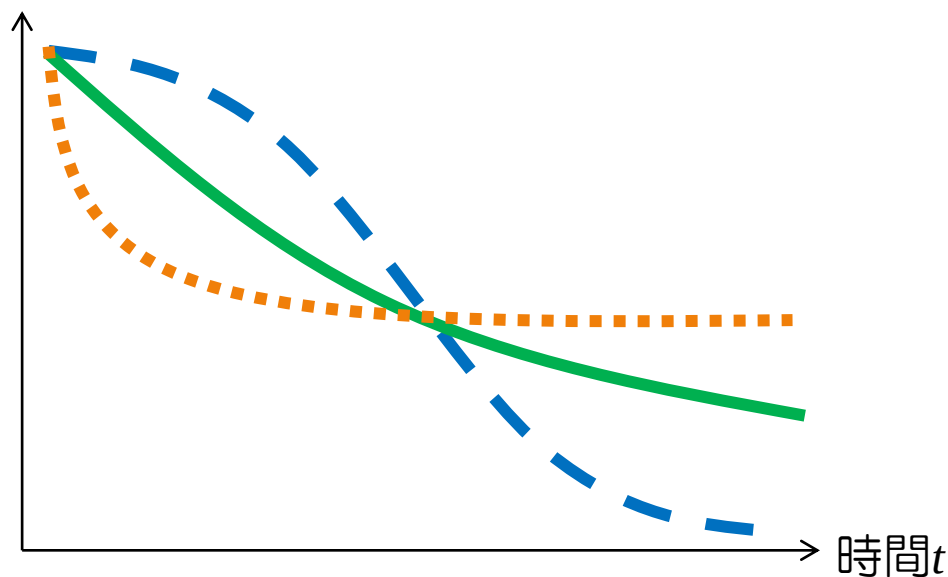
(b) の状況

全時間上でハザード(すなわちリスク)が一定であることを意味する。

(c) の状況

登録時点では死亡(あるいはイベント)のリスクは高いものの時間経過とともに減少することを意味する。一般に、進行がんに対して見られる。

生存確率 $S(t)$

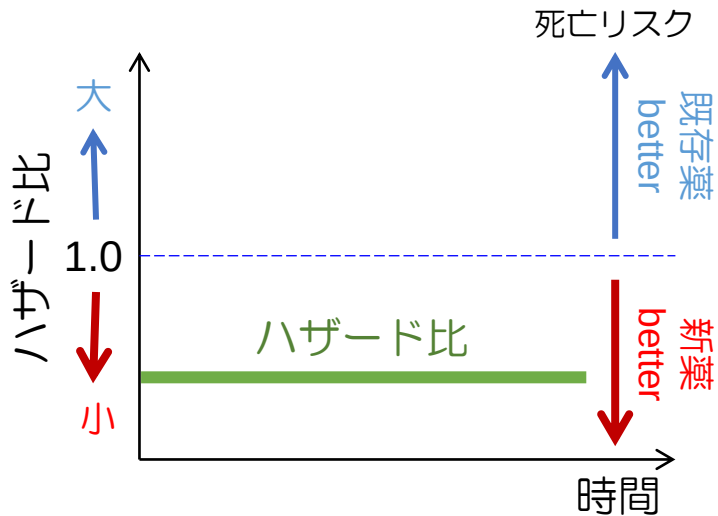


生存曲線の比較において重要な指標：ハザード比

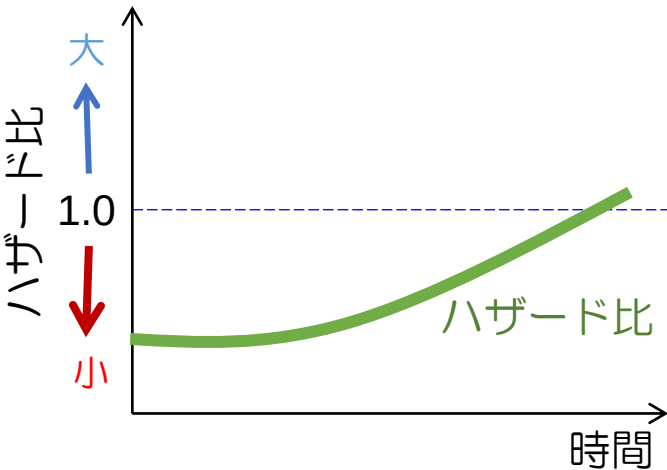
ハザード(危険度) $\lambda(t)$ とは「時間 t まで患者が生存しているという条件のもとで、時間 t で死亡する確率」を表す。つまり、時間 t での死亡リスクである。

ハザード比(相対危険度)とはハザードの比である。例えば、既存薬のハザードに対する新薬のハザードでのハザード比は

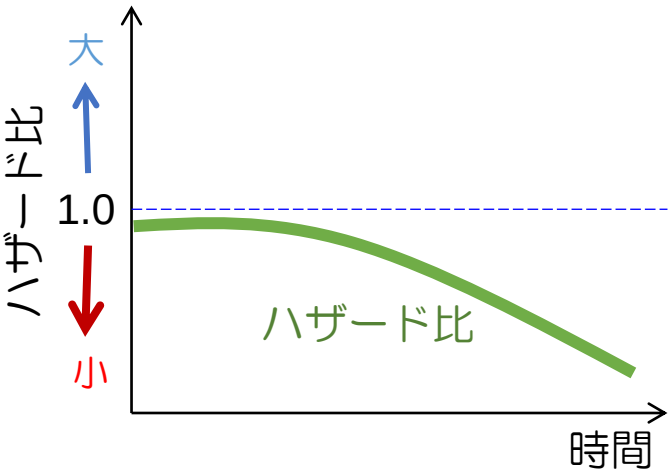
$$\text{ハザード比 } HR(t) = \frac{\text{新薬でのハザード } \lambda_1(t)}{\text{既存薬でのハザード } \lambda_0(t)}$$



ハザード比が生存時間に依らず一定。これは、新薬のほうが死亡リスクが低い。この状況が一般的に想定される(比例ハザード性の仮定)。

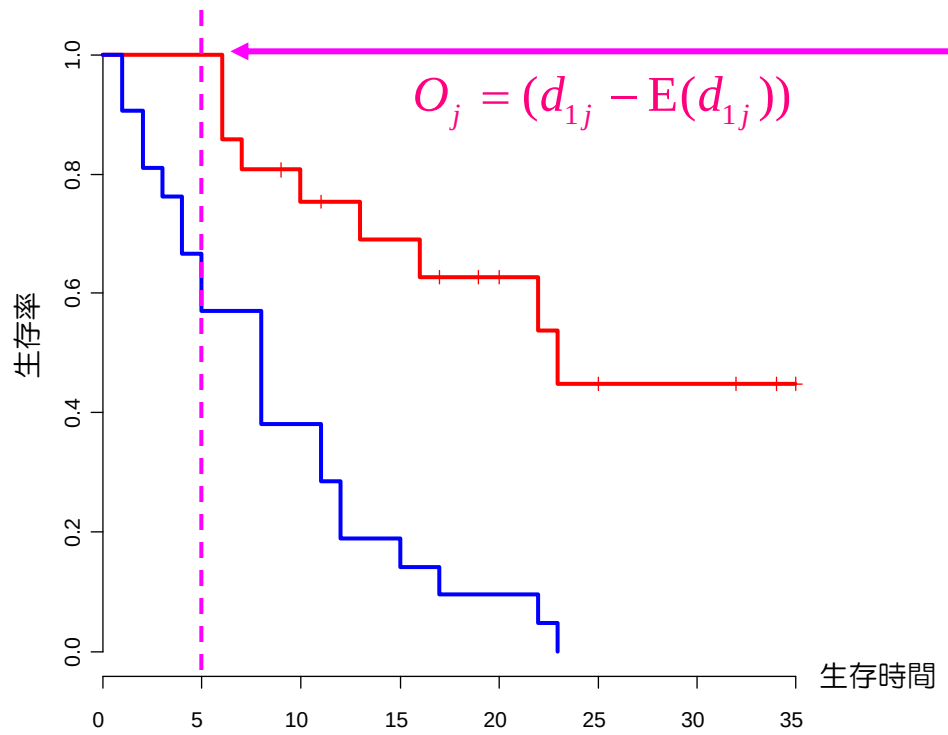


ハザード比が生存時間とともに上昇。これは、新薬での死亡リスクが生存時間の短い患者では低いものの、その有効性は長期生存症例になるにつれて薄れる。



ハザード比が生存時間とともに下降。これは、新薬での死亡リスクが生存時間の短い患者では既存薬と大差がないものの、その有効性は長期生存症例になるにつれて強くなる。

ロジック検定とは何をしているか？



ある時点 t_j における死亡・生存に対するクロス集計表

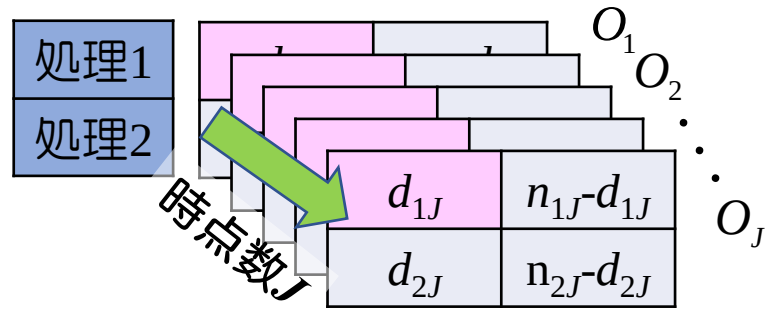
	死亡	生存	合計
処理 1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
処理 2	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
合計	d_j	$n_j - d_j$	n_j

もし、処理1と処理2で死亡数が同じであれば(つまり処理効果がなければ), の期待度数は、カイ2乗検定と同じ理屈から

$$E(d_{1j}) = \frac{d_j n_{1j}}{n_j}$$

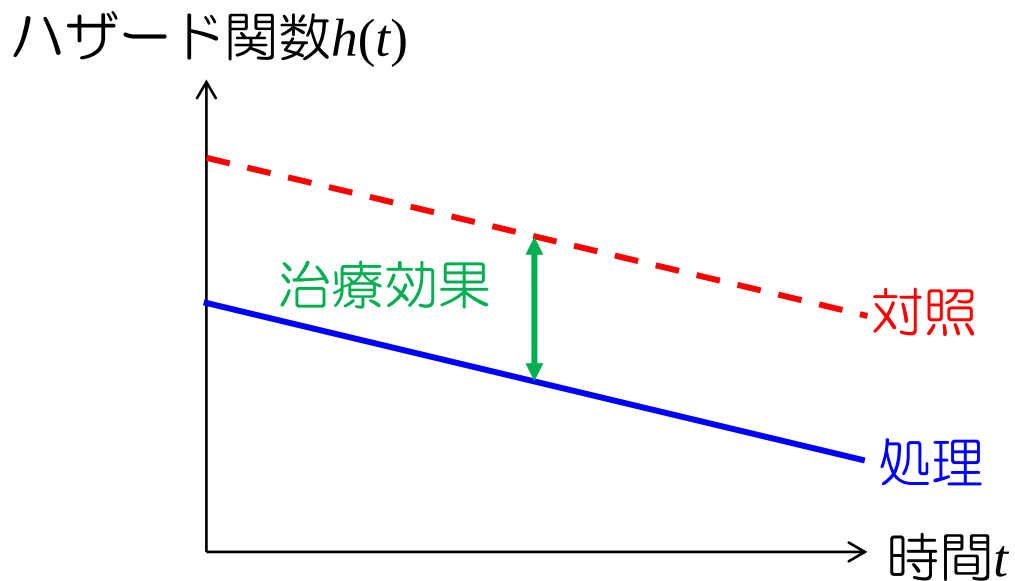
である.

上記の2×2クロス集計表の 部分をすべての時点において計算し、その総和をとる $O_1 + O_2 + \dots + O_J$ ことで構成されたのが、ロジック検定である.



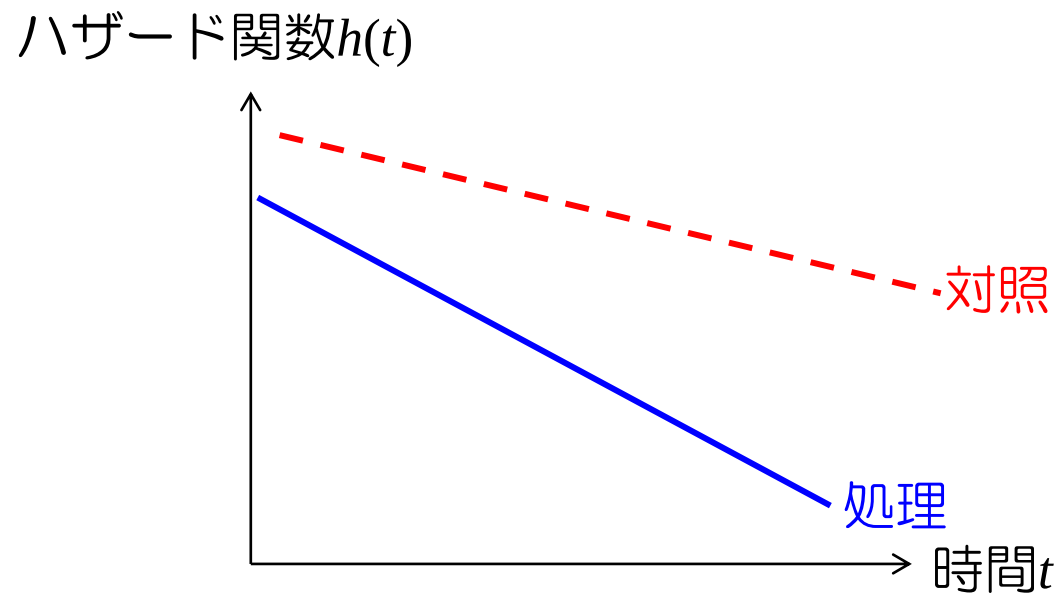
ログランク検定における仮定：比例ハザード性

○ 比例ハザード性



時間 t に関係なく治療効果の差が一致である。

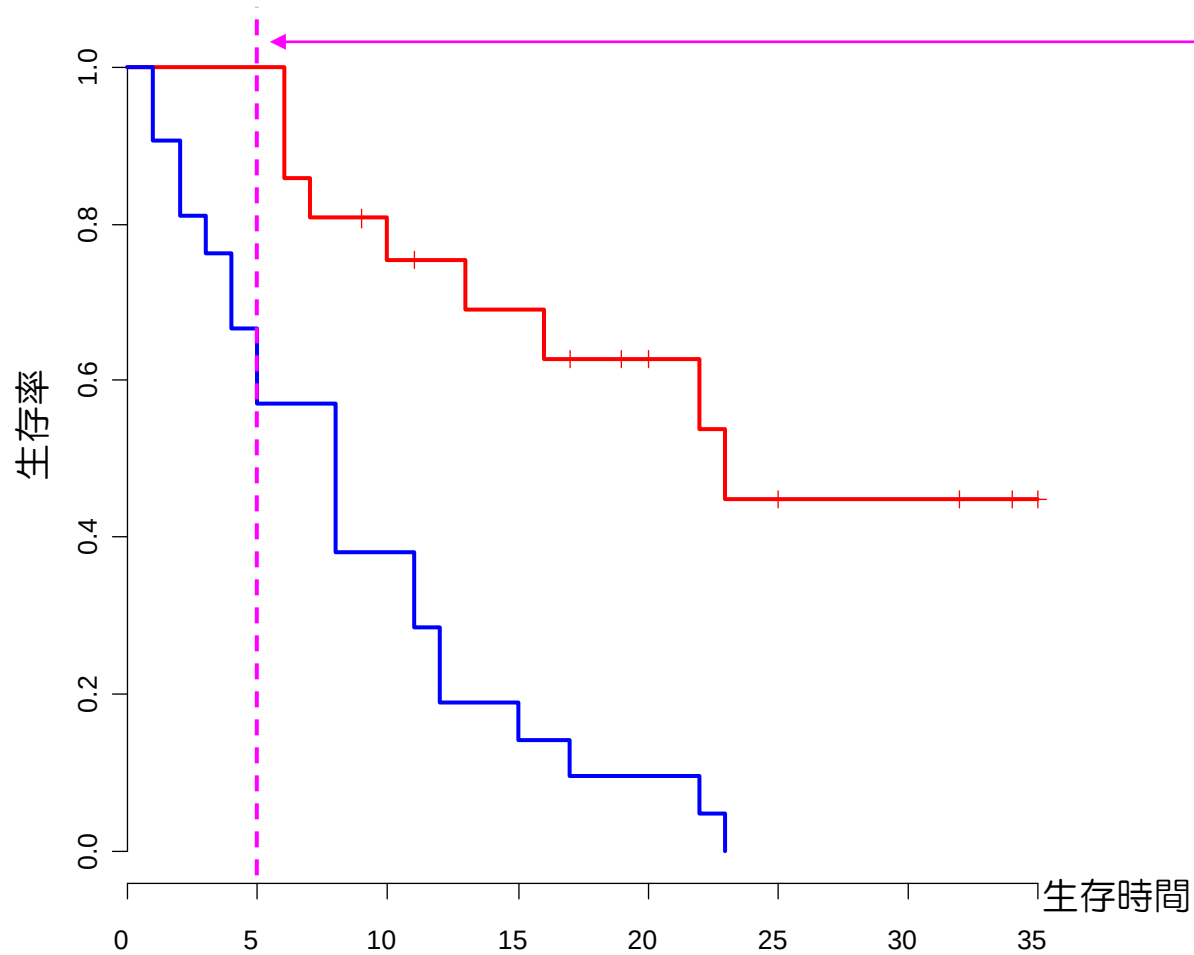
× 比例ハザード性



時間 t に関係なく治療効果の差が一致でない。

比例ハザード性の仮定は、ログランク検定だけでなく比例ハザードモデルにおいてもおかれている。

もうひとつの検定方法：一般化Wilcoxon検定



ログランク検定 ($d_{1j} - E(d_{1j})$)

一般化Wilcoxon検定

$$\hat{S}(t_j) \times (d_{1j} - E(d_{1j}))$$

一般化Wilcoxon検定では、ログランク検定に対して、生存率による重み付けを行っている(生存率の推定がKaplan-Meier法の場合にはPeto-Petoの一般化Wilcoxon検定、Prentice法の場合にはPrenticeの一般化Wilcoxon検定と呼ばれる)。

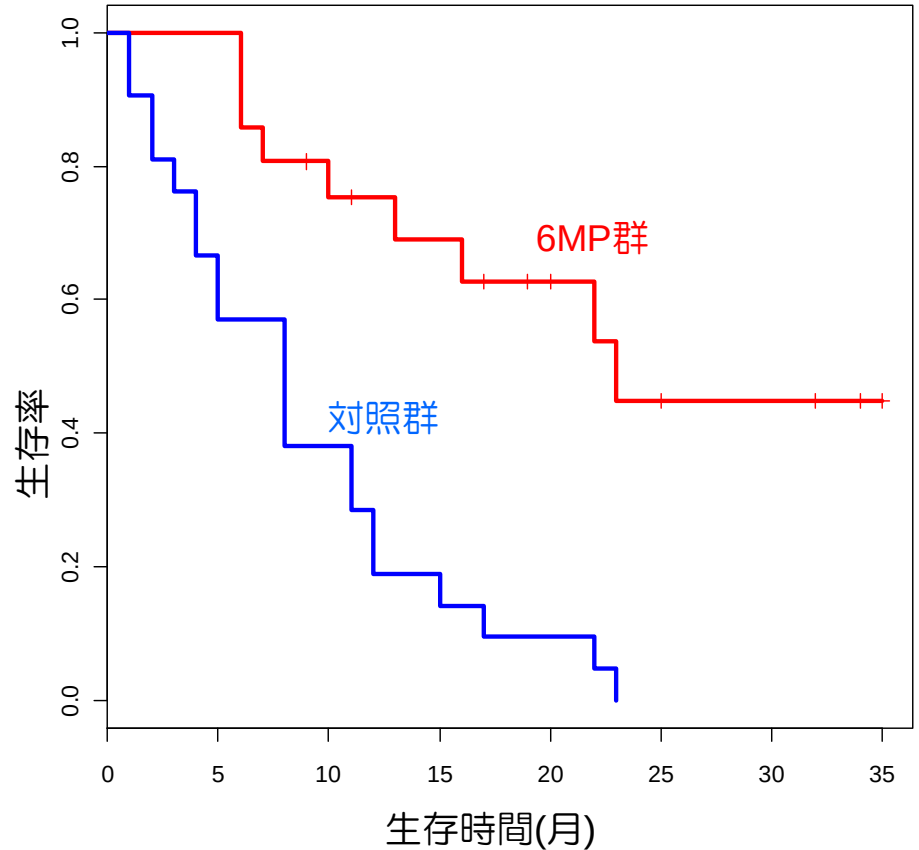
時間 t が小さい場合には、 $\hat{S}(t_j)$ が1に近い(重みが大い)ことから、一般化Wilcoxon検定では、早期の2群間の差に敏感である。

2標本の比較の例示(1)

+は中途打ち切り，単位は月

小児急性白血病データ
(Freireich *et al.*, 1963)

6MP 投与群	6	6	6	6+	7	9+	10	10+
	11+	13	16	17+	19+	20+	22	23
	25+	32+	32+	34+	35+			
対照群	1	1	2	2	3	4	4	5
	5	8	8	8	8	11	11	12
	12	15	17	22	23			



諸種の検定におけるP値

□ グランク検定 $p < 0.0001$
一般化Wilcoxon検定 $p = 0.0002$

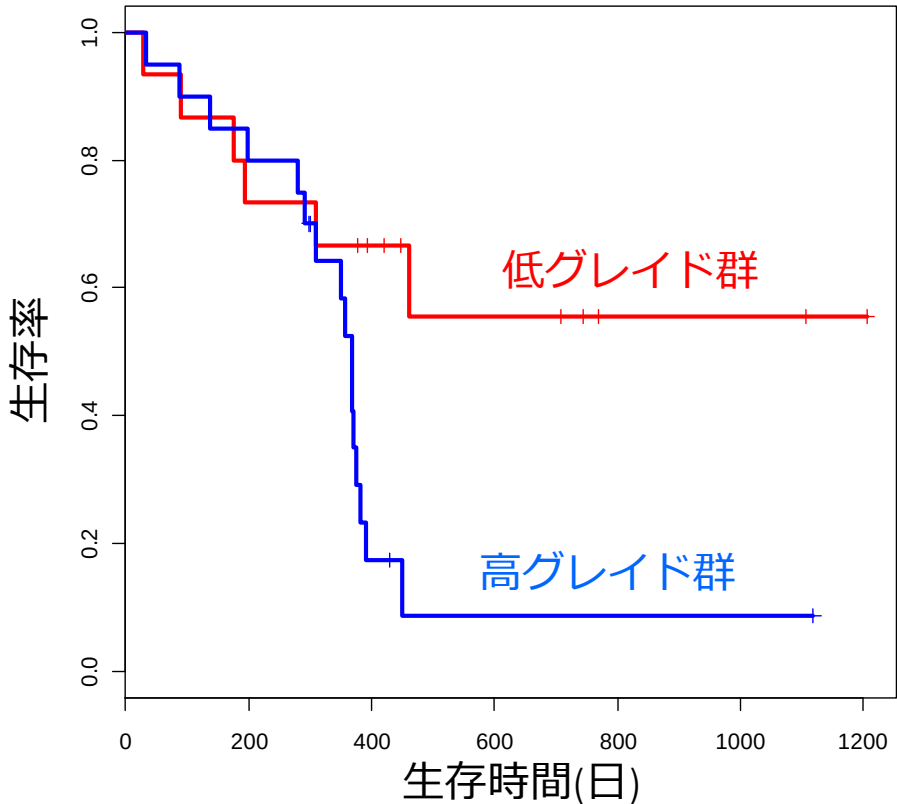
いずれの検定法でも6MP群と対照群のあいだで有意であることがわかる。

2標本の比較の例示(2)

卵巣がん患者のグレード毎での病気進行期間(Fleming *et al.*, 1980)

低グレード群	28	89	175	195	309	377+	393+	421+
	447+	462	709+	744+	770+	1106+	1206+	
高グレード群	34	88	137	199	280	291	299+	300+
	309	351	358	369	369	370	375	382
	392	429+	451	1119+				

+は中途打ち切り, 単位は日



諸種の検定におけるP値

□ グランク検定 $p=0.0183$

Gehanの一般化Wilcoxon検定 $p=0.0995$

比例性は満たされていないと指摘されているデータではあったものの、ログランク検定のみ有意差が認められた。初期の生存時間で差が認められなかったことが寄与していると考えられる。