

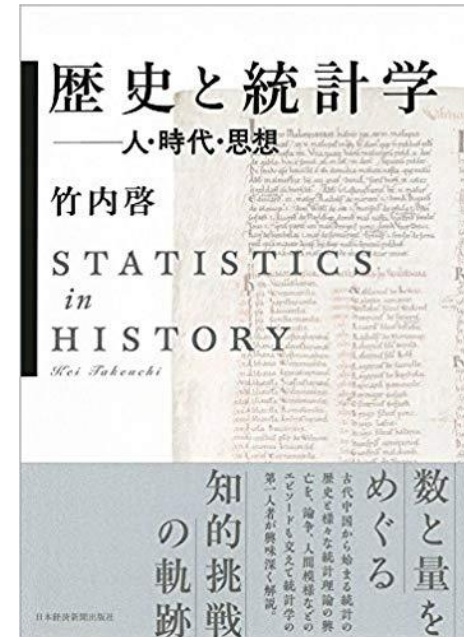


研究立案時における統計の視点

大阪大学大学院医学系研究科
医療データ科学共同研究講座
飛田 英祐

統計学(Statistics)とは

- 「統計」の広い解釈
 - 何らかの客観的な対象に関する数字データが集計され、解釈されたもの
 - 3つの要素が含まれる
 - 客観的な対象
 - 数字データの作成者
 - データを「統計」として解釈する者



「木を見て森を見ず」でも「森を見て木を見ず」でもなく、
「木も見て森も見て」の視野を持つことが理想的

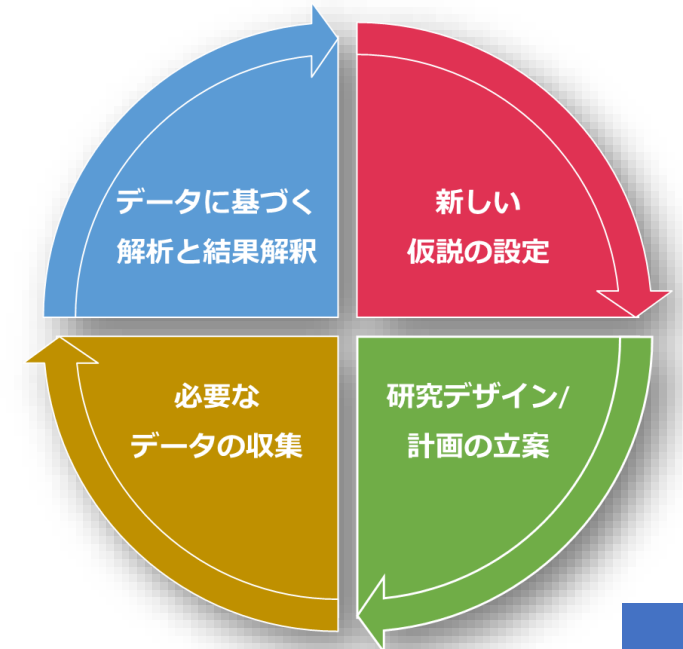
医学研究におけるサイクル

- 医学研究(臨床試験)は本質的に実験科学
 - 生体システムや、それを取り巻く環境における因果関係の究明を目的
 - 疾患の原因、リスク因子を調べることで疾患の予防に繋げる
 - 疾患の診断法や治療法の効果を調べることで早期発見や治療に繋げる



臨床研究によってエビデンスを作る

- 目的に応じたデータが収集できる**計画を立て、**
- 計画通りに**データを収集し、**
- データの特性に応じた**要約、解析**を行い、
- 定量的な**結果を導き、意思決定に関与する**

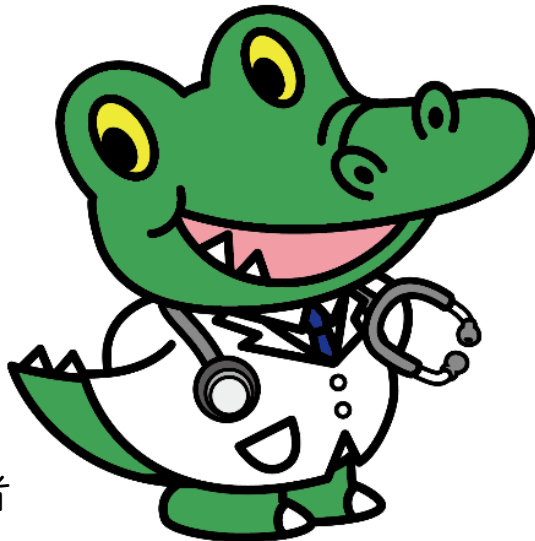


統計コンサルテーションで・・・

Clinical Question

急性期心筋梗塞でPCI後の心機能の低下した患者にEPOを投与することは梗塞サイズの縮小につながるかを調べたい!
効果が期待できる最適なEPOの用量は?

医師
研究者



さて、困ったぞ～

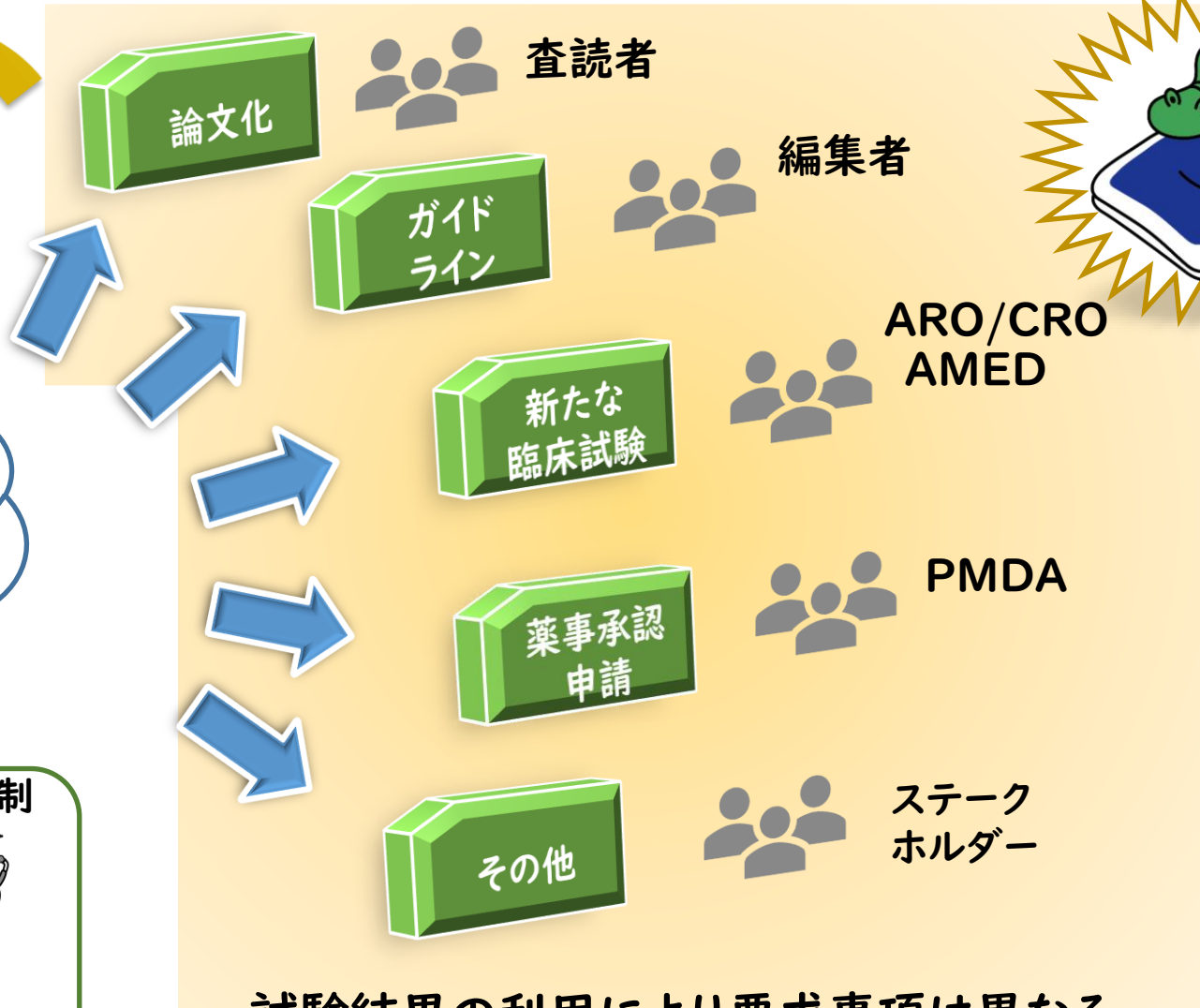
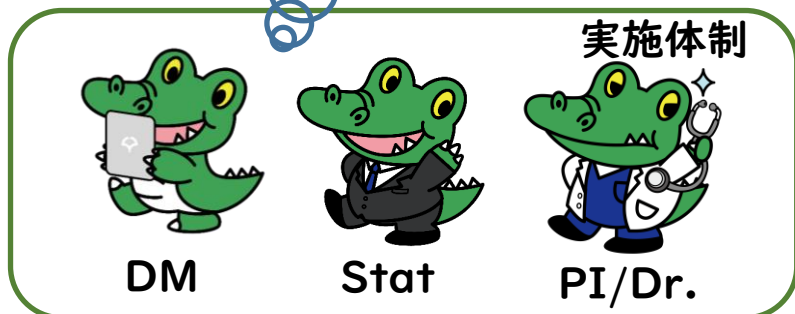
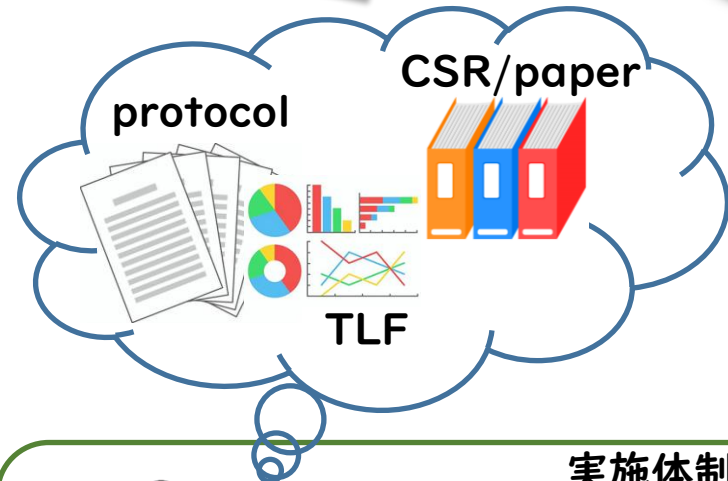
このCQに対するResearch Question
や試験の目的は?
この研究全体のゴールは?
薬事承認?論文化?ガイドライン作成?
その当たりからの整理が必要!!

統計担当者



臨床試験の先には・・・

眼の前の1試験だけでなく
その先の研究全体のゴール
要求される水準を
視野に入れた過不足のない
試験計画を!



試験結果の利用により要求事項は異なる

研究全体のゴールは？

研究全体のロードマップの整理

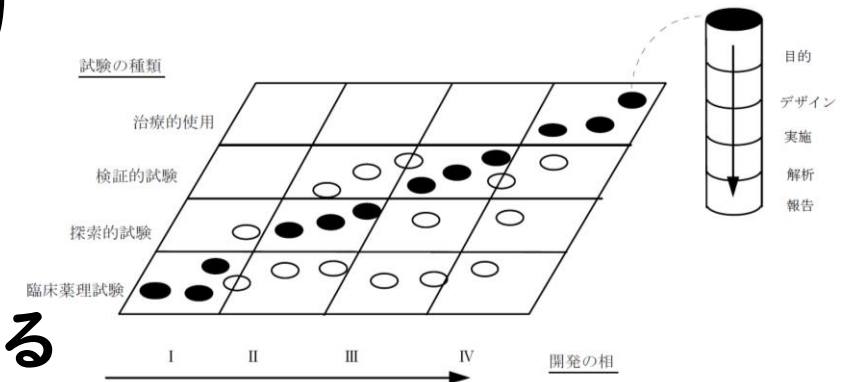
– 一つの臨床試験で証明できることは通常一つだけ

- 試験の目的/仮説に基づいて目標症例数を算出
- 複数の試験の積み重ねにより、研究全体のゴールを達成

• 研究全体における実施予定の試験の役割/目的

– 一つの臨床試験に複数の目的を含めることは試験が煩雑になり、ミスやエラーの増加に繋がる

• どのデータが重要？全て？...



ICH E8 臨床試験の一般指針

<https://www.pmda.go.jp/files/000156372.pdf>

臨床試験における科学的視点

- 例えば、
「統計学的に有意な結果が示された」として…
 - そもそも、取得したデータで適切にかつ十分に新しい治療の真の効果を表しているのか？
 - データの取得状況や解析方法など、何らかの理由で、真の効果からズレた(より良い)値を見せられているのでは？

• 検討対象となるデータを適切に取得し、科学的に妥当な信頼できる結果を得るためには適切な事前の計画が重要

臨床試験における科学的視点

- 研究対象者（被験者）のリスクを最小にして、科学的な議論が可能となるデータを取得したい

– ヒトが対象となる臨床試験では

- **倫理的な制約**

- 研究対象者のリスクを軽減するために、できる限り対象者の数を少なくしたい

⇒ 統計学的な観点を含めた
Sample sizeの決定

- **個体差の問題**

- バラツキ（個体間/内変動）の存在下で如何に科学的な議論を可能とするか？

⇒ データ収集プロセス管理
確率・統計学的な議論

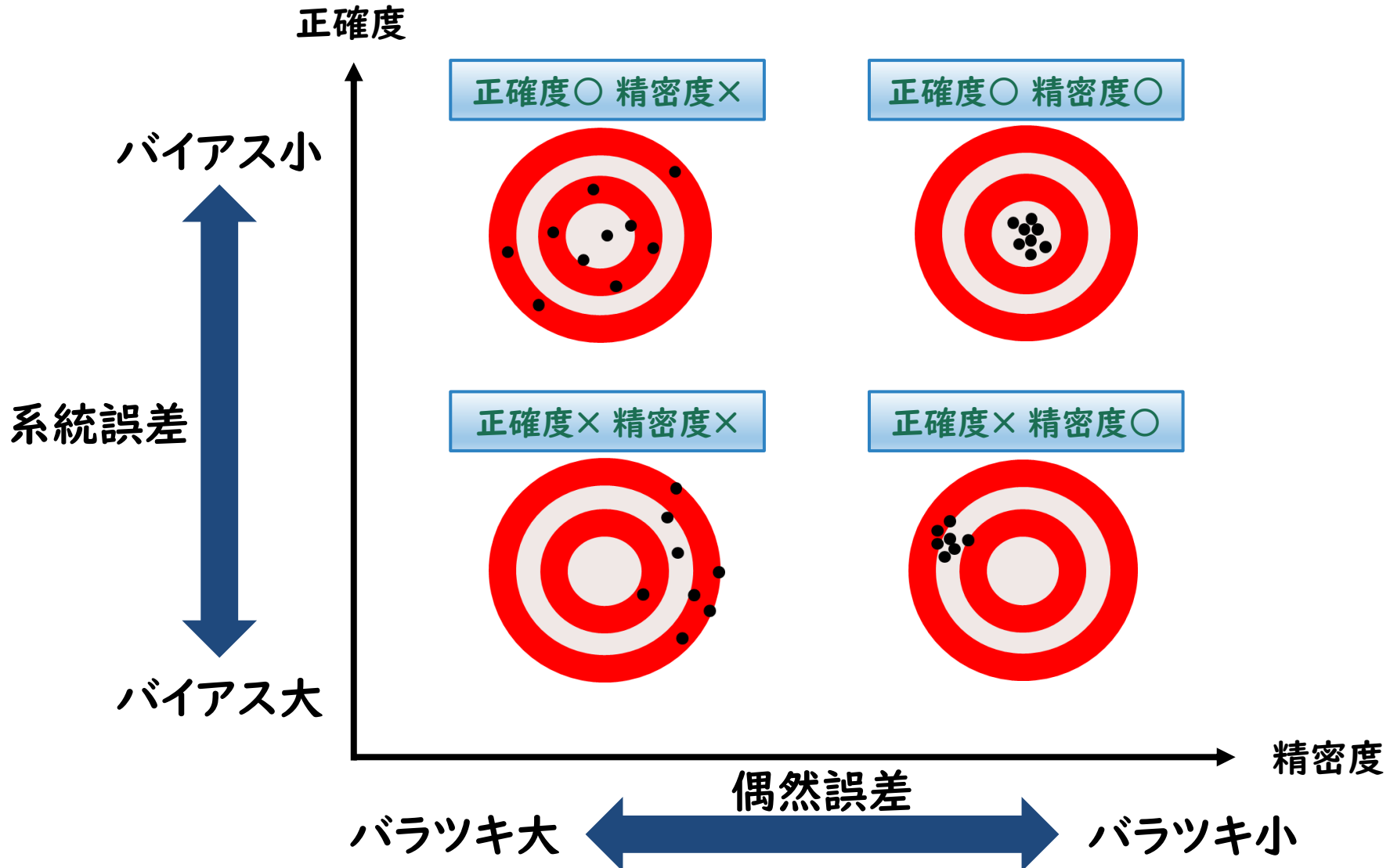
- **実診療下での研究**

- 偏りのあるサンプリングへの対処

⇒ デザインや統計モデル

- 臨床試験で生じる偏り(バイアス)を最小化
 - 系統的誤差(バイアス: bias)
 - 偏った結果に導く「悪性」の誤差
 - 臨床試験の計画、実施、解析及び結果の評価と関連した因子の影響により、試験治療の効果の推定値と真の値に系統的な差が生じる
- 臨床試験の精度を最大化 ⇔ 偶然誤差を最小化
 - 偶然誤差(ランダムエラー: random error)
 - ある程度に制御可能な「良性」の誤差
 - ミス/エラーを0%にすることは難しいが、手順書などにより一定の許容されるレベル以下に抑える、研究対象者の数を増やす、ことなどで制御可能

バラツキとバイアス



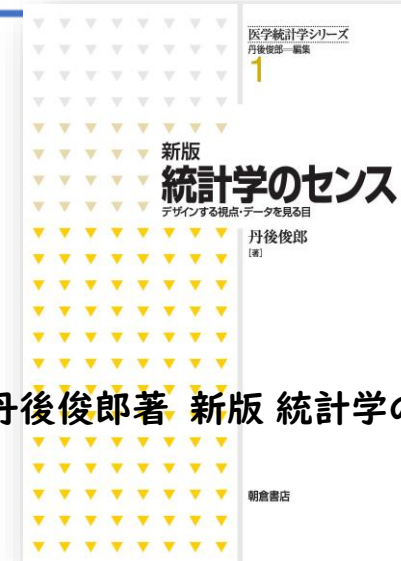
臨床試験における証拠能力の高さ

- 試験結果から主張できる証拠能力の高さは、基本的に試験デザインと試験の質によって決まる
- 試験デザイン ⇔ バイアスの最小化
 - 事前に設定する広義の試験デザイン
- 試験の質 ⇔ 精度の最大化 / ランダムエラーの最小化
 - データの品質 / 試験の品質
 - データ収集プロセスの管理/モニタリング
 - Quality Management System(QMS) / Risk Based Approach(RBA) / Quality by Design

事前に作り込まれた研究計画が重要

臨床試験の計画・プロトコル

- 臨床試験を成功させるには
 - エンドポイント: endpoint
 - ランダム化: randomization
 - 交絡因子: confounding factor
 - バイアス: bias



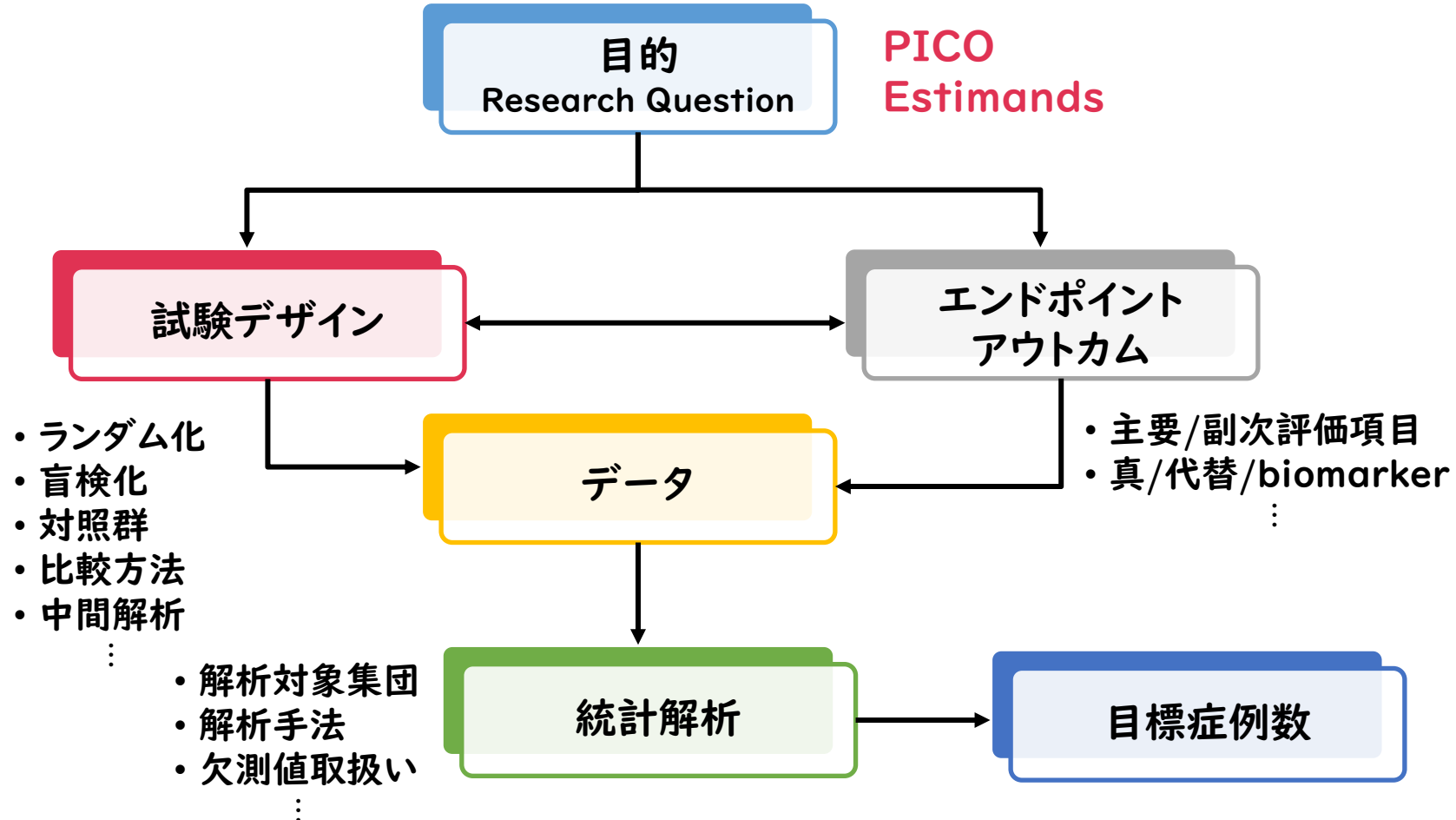
(丹後俊郎著 新版 統計学のセンス)

を十分に考慮した**センス**のあるデザイン・計画が必須条件
及びデータマネージャ

- 計画段階から生物統計家の協力・参加が必要
 - 研究者など統計専門以外の担当者の統計学的な知識の向上
 - 統計家の臨床的な知識の向上

統計家視点のプロトコル

・ 広義の試験デザイン

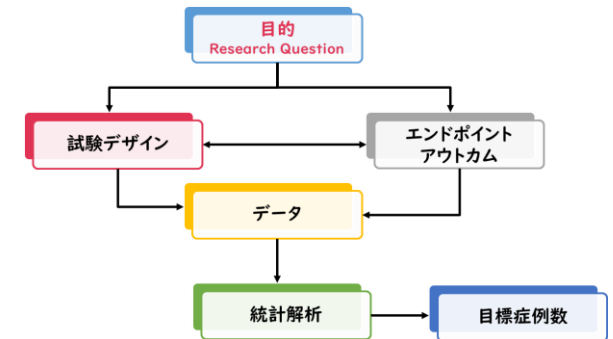


Clinical QuestionからResearch Questionに

- 試験目的を明確にするためのClinical Questionの定式化
PICO (介入研究) / PECO (観察研究)

最も重要な試験の根幹が「試験の目的」

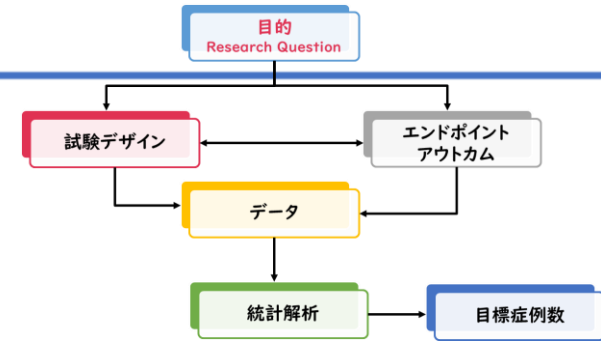
- P (patients) : 誰を (どんな患者を) 対象として
- I (Intervention) : どんな介入を行い
- E (Exposure) : どんな曝露/リスクが
- C (Comparison/Control) : 何と比べて
- O (Outcome) : どのような結論を得たいのか



- 例) 風邪に罹患した患者(P)を対象として、かぜ薬Z(I)はプラセボ(C)よりも投与3日後の治癒割合(O)が有意に高いことを検証する

さて、EPO試験のPICOは？

心筋梗塞患者に対するエポエチンベータ 投与による心機能改善効果に関する研究-II



- P (対象): 初回発症のST上昇型心筋梗塞患者で発症から
12時間以内に再灌流に成功した患者
- I (介入): エポエチンベータ2用量(6000単位, 12000単位)
- C (比較): プラセボ
 - 試験デザイン: 多施設共同二重盲検無作為化並行群試験
- O (結果): 左室駆出率が維持される
 - 評価項目: 左室駆出率(LVEF)の急性期から6カ月後までの改善度

PICO -> RQ

53

初回発症のST上昇型心筋梗塞患者で発症から
12時間以内に再灌流に成功した低心機能患者を
対象として、PCI後6時間以内に低用量のEPOを
単回投与すると、プラセボと比較して、
左室駆出率が維持されるか

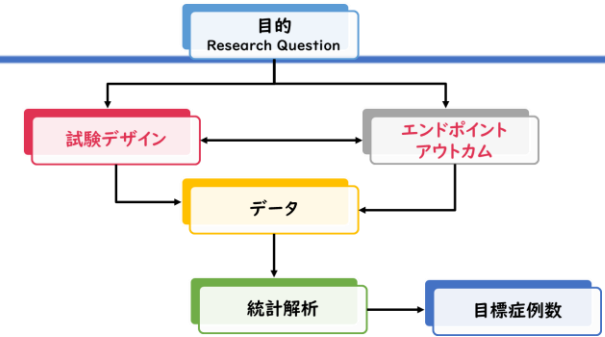
医学研究の基本は比較

• ワニ博士が風邪をひいた

- 12月9日: 風邪をひいたので、「かぜ薬Z」を飲んだ
- 12月12日: すっかり風邪が治った
- ワニ博士:「かぜ薬Zは有効だ!!」と考えた



Clinical Question



• 正しい因果関係(原因と結果)を示している?

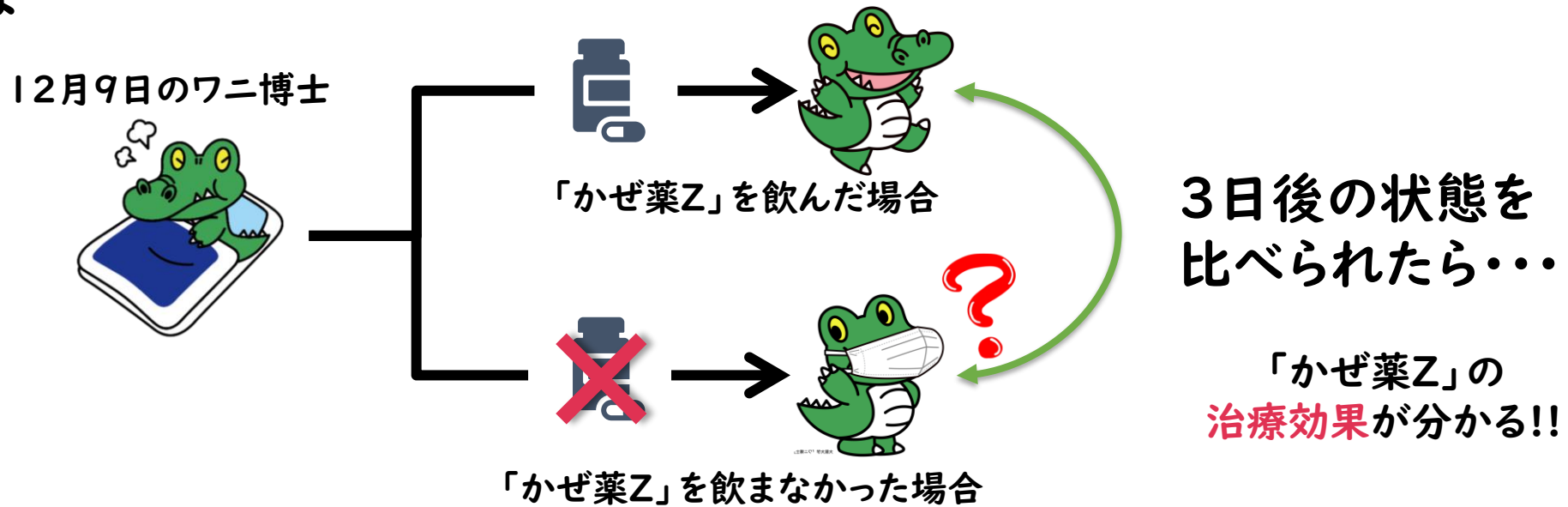
- 風邪薬を飲まなかったら風邪は治らない?
- 飲まなくても十分に静養すればすっかり治るかも
- 正しい場合もあるかもしれないが、多くは偶然...



薬を「使った、治った、だから効いた」→サンタ(三た)論法

医学研究の基本は比較

- 「かぜ薬Z」の服用以外の条件（風邪の原因、症状、環境等）が全く同じであれば・・・



- 同じワニ/人で、同じ(時点での)病気に違う治療は試せない
 - 似たようなワニ/人を連れてきて比較すればよい？
 - ⇒ 現実的には不可能な場合が多い

医学研究の基本は比較

- 様々な特徴が類似した集団間で比較すればよいのでは？



- 似たような人を見つけられなくても、
 - 集団として年齢、性別、重症度の他に様々な背景の分布が類似していれば、「かぜ薬Z」の治療効果が調べられるだろう

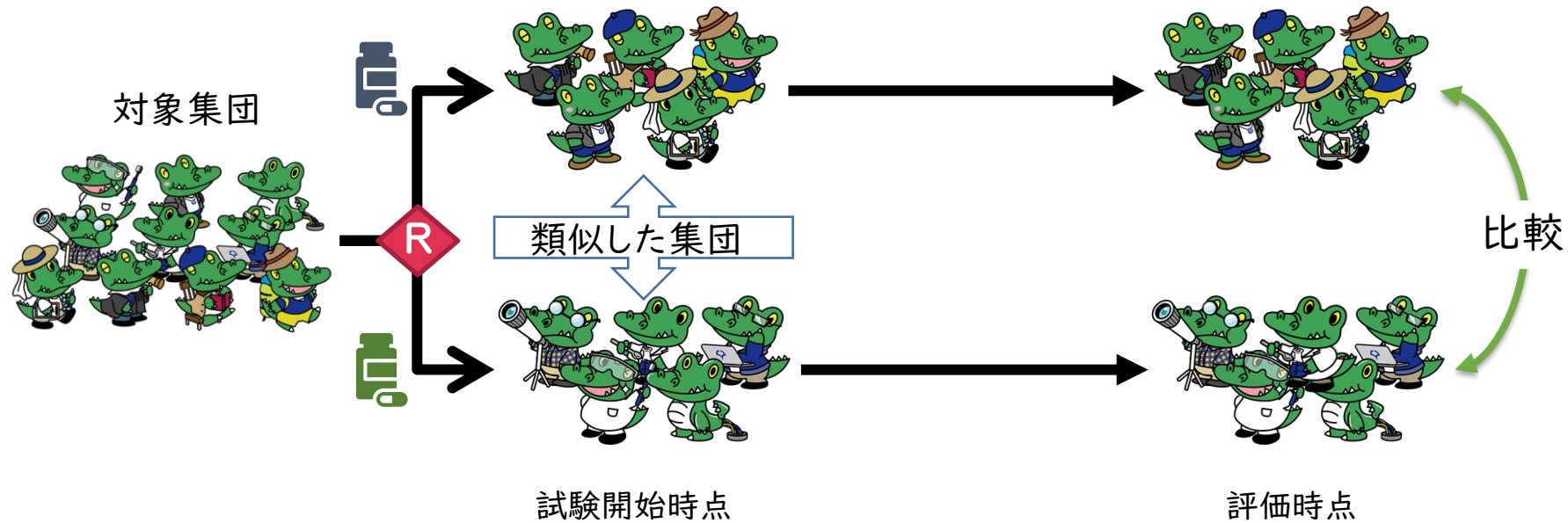
類似した集団を作る操作が「ランダム化/無作為化」

ランダム化を行う意義

- 偶然の要素だけで治療群に分けることで、
 - 治療群間の特徴が平均的に一致（背景因子の均一化）
 - 治療群間の比較可能性（内的妥当性）の保証
 - 年齢、性別、重症度など既知の観測される因子だけでなく、未観測や未知の因子も理論的に均一化
 - 交絡によるバイアスの除去
 - 治療効果の定量的な評価のための統計的根拠を与える
 - さまざまな変動の誤差を確率化
- ITT(Intention-to-treat)の原則は、ランダム化による類似した集団をできる限り保持しようとする考え方

ランダム化比較試験 (RCT; Randomized Control Trial)

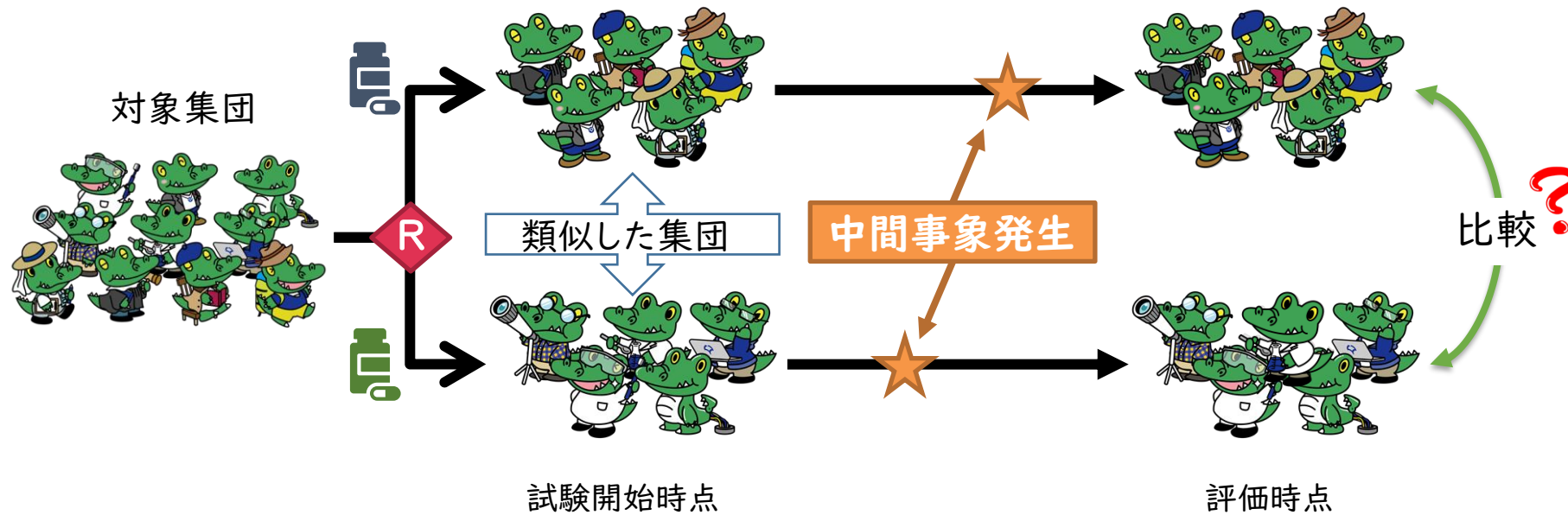
- 最も一般的な比較試験デザイン
 - 交絡によるバイアスを除去し、比較可能性を担保



- EPOの試験では、登録された対象者を6000IU、12000IU、プラセボの3群にランダムに割り付けて、治療効果を比較する試験デザイン

ランダム化比較試験 (RCT; Randomized Control Trial)

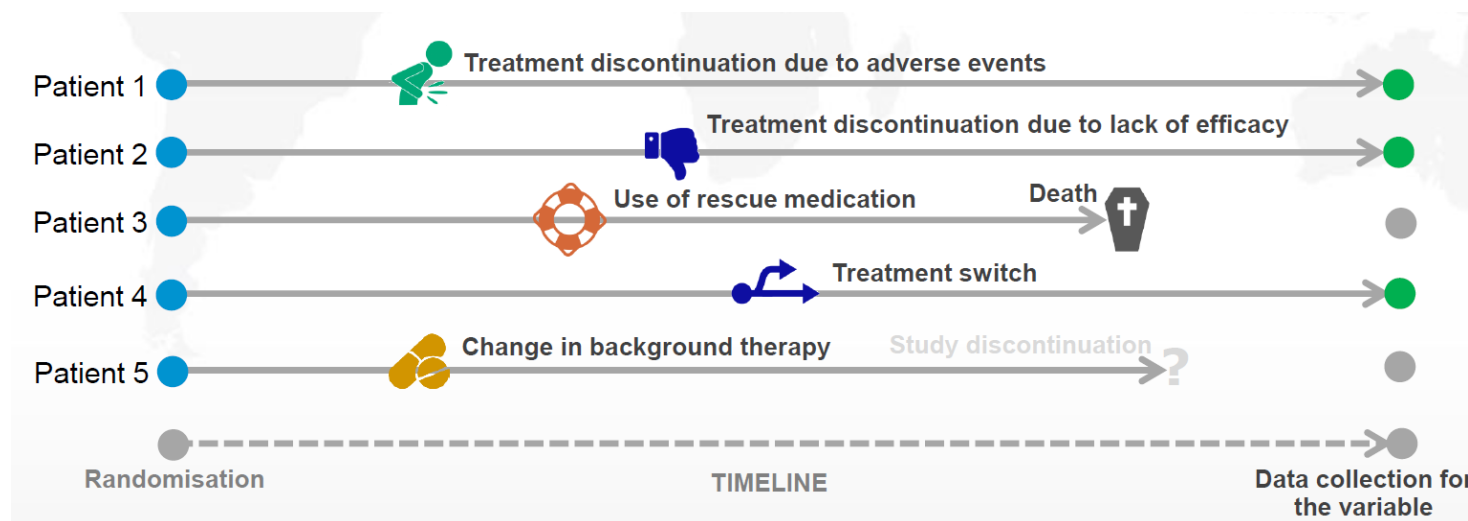
- 最も一般的な比較試験デザイン
 - 交絡によるバイアスを除去し、比較可能性を担保



- 試験実施中に**中間事象**が発生すると、ランダム化が維持できない、解析に利用するデータが欠測するなど、治療効果の評価が困難になる

中間事象 (intercurrent event)

- 治療開始後に発現し、変数を観測できなくする又は変数の解釈に影響を与える事象
 - 治療不遵守、治療の中止(有害事象、有効性不足)、代替治療の使用(レスキュー治療など)、転居、死亡などの終末事象、など



欠測とは異なり、中間事象によっては以降の評価時点でデータの測定が可能
 (そのデータは解析に利用する?/利用しない?)
 死亡といった中間事象の後では、データの測定は不可能

EPO試験では・・・

研究実施体制の各担当で議論して 作成した観察スケジュール

検査・観察スケジュール

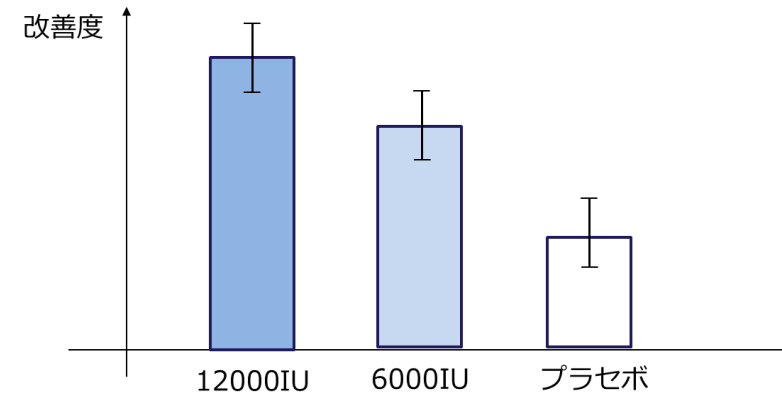
検査内容	PCI後～ 登録時	投与後 1時間	投与後 4日	投与後 7日	投与後 35日	投与後 6ヵ月
				+2	±7	±28
同意取得	○					
患者背景	○					
一般所見（自他覚症状）	○	○	○	○	○	○
臨床所見（有害事象等）	○	○	○	○	○	○
バイタルサイン	○	○	○	○	○	○
血液学的検査	○ *1		○	○ *1	○	○
生化学的検査	○ *2			○	○	○
生化学的検査（NT-ProBNP）						○
心電図・胸部 X 線・心エコー	○					
安静時心筋シンチ			○ *3			
アデノシン負荷心筋シンチ						○ *4

プロトコルを基にEDCを構築し、
EPO試験で必要なデータを収集する

得られたデータでどのような集計や解析、
作図をするかも想定しておくことが重要

主要評価項目: LVEFの改善度

	12000IU群	6000IU群	プラセボ群
症例数	xxx	xxx	xxx
平均値	x.xx	x.xx	x.xx
95%信頼区間	[x.xx-x.xx]	[x.xx-x.xx]	[x.xx-x.xx]
Williams検定	p=0.xxx	p=0.xxx	-



ところが、欠測や中間事象が発生した対象者のデータはどうやって解析や図に使い、どう評価されるの？
中間事象発生後のデータ収集は不要？
他にどんなデータ/情報が必要？

過去の実例では

<http://www.pmda.go.jp/PmdaSearch/iyakuDetail/GeneralList/1179409G1>

- **パリペリドンパルミチン酸エステル (ゼプリオン)**

- **効能・効果：統合失調症**

- **アジア共同臨床第III相試験 (PALM-JPN-4)**

- **主要目的:** 統合失調症患者を対象とし、試験薬の有効性をプラセボと比較 (優越性試験)
- **対象:** 急性期症状を有する統合失調症患者
- **デザイン:** ランダム化プラセボ対照二重盲検並行群間比較試験
- **主要評価項目:** PANSS総スコアのベースラインから最終評価時 (13週又は二重盲検期中止時) までの変化量
- **目標被験者数:** 308例 (154例/群)

過去の実例では

<http://www.pmda.go.jp/PmdaSearch/iyakuDetail/GeneralList/1179409G1>

• 主要評価項目の解析結果

表 17 PANSS 総スコア変化量 (FAS、LOCF)

	評価例数	PANSS 総スコア		変化量	プラセボ群との対比較 ^{a)}	
		ベースライン	最終評価時		群間差 [95%信頼区間] ^{b)}	p 値
プラセボ群	164	83.5 ± 15.18	90.3 ± 22.35	6.9 ± 19.13	-9.7 [-14.0, -5.4]	< 0.0001
本剤群	159	85.7 ± 14.57	82.4 ± 23.52	-3.1 ± 20.32		

平均値 ± 標準偏差
 a) 投与群及び国を因子、ベースラインの PANSS 総スコアを共変量とした共分散分析モデル
 b) 調整済み平均値

ところが

被験者の内訳及び中止被験者の中止理由

	プラセボ		PALI palm		合計	
	被験者数	(%)	被験者数	(%)	被験者数	(%)
ランダム化された被験者	164		160		324	
治験薬を投与された被験者	164	(100.0)	159	(99.4)	323	(99.7)
治験薬投与前に中止した被験者	0		1	(0.6)	1	(0.3)
二重盲検期を完了した被験者	55	(33.5)	95	(59.4)	150	(46.3)
二重盲検期中止した被験者	109	(66.5)	64	(40.0)	173	(53.4)
中止理由						
安全上の理由 (有害事象等)	45	(27.4)	27	(16.9)	72	(22.2)
効果不十分	45	(27.4)	21	(13.1)	66	(20.4)
同意撤回	16	(9.8)	13	(8.1)	29	(9.0)
追跡不能	2	(1.2)	1	(0.6)	3	(0.9)
治験開始後に「治験対象集団」の基準に合致しないことが判明	0		1	(0.6)	1	(0.3)
その他	1	(0.6)	1	(0.6)	2	(0.6)
後観察期を完了した被験者	144	(87.8)	149	(93.1)	293	(90.4)
後観察期中止した被験者	12	(7.3)	6	(3.8)	18	(5.6)

中間事象

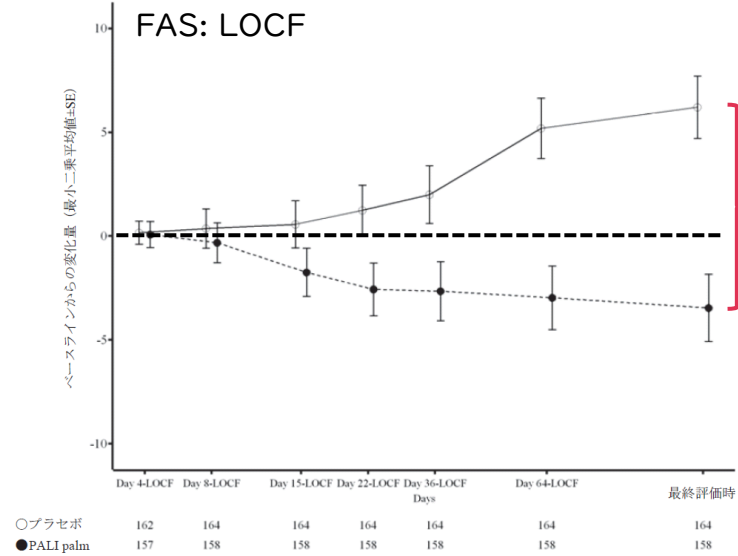
過去の実例では

<http://www.pmda.go.jp/PmdaSearch/iyakuDetail/GeneralList/1179409G1>

• 主要評価項目の解析結果

LOCF (Last Observation Carried Forward)

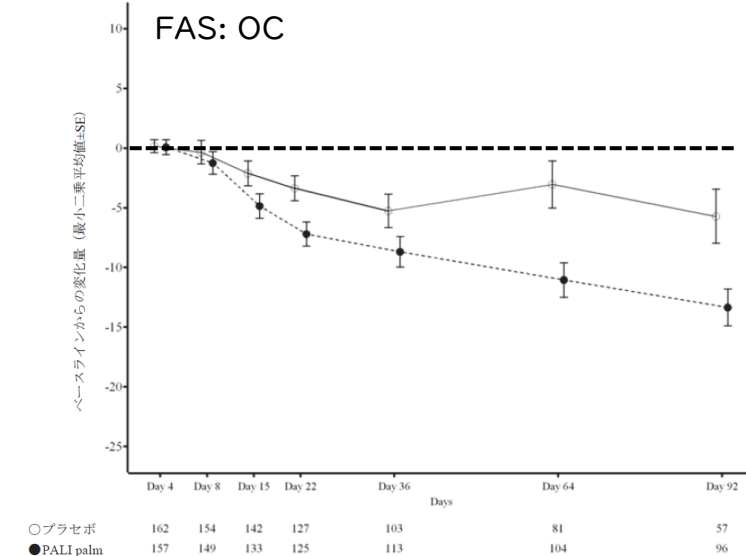
中止例は中止以前に測定されたスコアで
中止後の測定値を補完



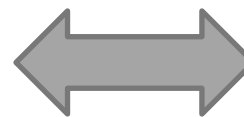
中間事象による中止例の効果は中止前の
最終観測時点から変わらないと仮定した
治療効果

OC (Observed Case)

各評価時点で測定された対象のみのスコア



中間事象の有無に関わらず、投与開始
13週時点で測定できた対象のみの治療効果



本当に知りたい治療効果は？
計画時に想定していた治療効果が得られているの？

欠測値による問題点

• バイアスの問題

- 群間での欠測割合、欠測理由が異なる場合に、ランダム化で保証された比較可能性が崩れ、治療効果の比較にバイアスが生じる

• 推定の精度低下の問題

- データ数が減少するため、治療効果の推定精度や検出力が低下する

• 研究の質に対する問題

- 欠測値が多い程、研究の質が低下

- 欠測値は臨床試験で偏りを起こし得る代表的な原因
 - データ収集及びデータマネージメントに関しては、プロトコルに記載された項目を全て収集するように努力すべき
 - 欠測データを扱う適切な方法が事前にプロトコルに記載されていれば、妥当とみなせる


- 欠測値に対処する方法
 - 普遍的に適用可能と薦められる方法はない
 - 対処する方法により解析結果がどの程度変わり易いかを検討すべき
⇒ 感度解析

- 「中間事象」とは記載されていないが、、
 - 解析を損なうと予想される実施中の変則的な事例のすべて (様々な種類のプロトコル違反、試験治療の中止、欠測値を含む) を最小にするための手続きも明記すべきであり、そのような問題の発生頻度を少なくする方法、データ解析中に起こる問題への対処方法を考慮しておくべき
- ITT (Intention-to-treat) の原則で対応できるだろう、、
 - ランダム化が行われた全被験者を主要な解析に含めるべき
 - 結果を得るためにランダム化が行われた全被験者を完全に追跡することを必要とする
 - 実際には難しいので、「最大の解析対象集団」というITTの原則に可能な限り完全に近づけた解析対象集団を表す

ICH E9 (R1)

臨床試験におけるestimandと感度分析

- データ収集や欠測値を取り扱う解析方法などによって、治療効果を推定する/できる方法が複数想定される
 - データ収集方法や欠測データを含む解析手法などによって、推定される治療効果が決められることがあり、試験の目的と推定される治療効果の間に潜在的な不整合が生じることがあった

- 
- E9 (R1) ガイドライン
 - 試験の目的と推定される治療効果との整合を明確化
 - 臨床試験で知りたい治療効果 (**estimand**) を設定し、どのように解析していくか、**フレームワーク**を提案

<https://www.pmda.go.jp/int-activities/int-harmony/ich/0031.html>

フレームワークの順に従った試験計画

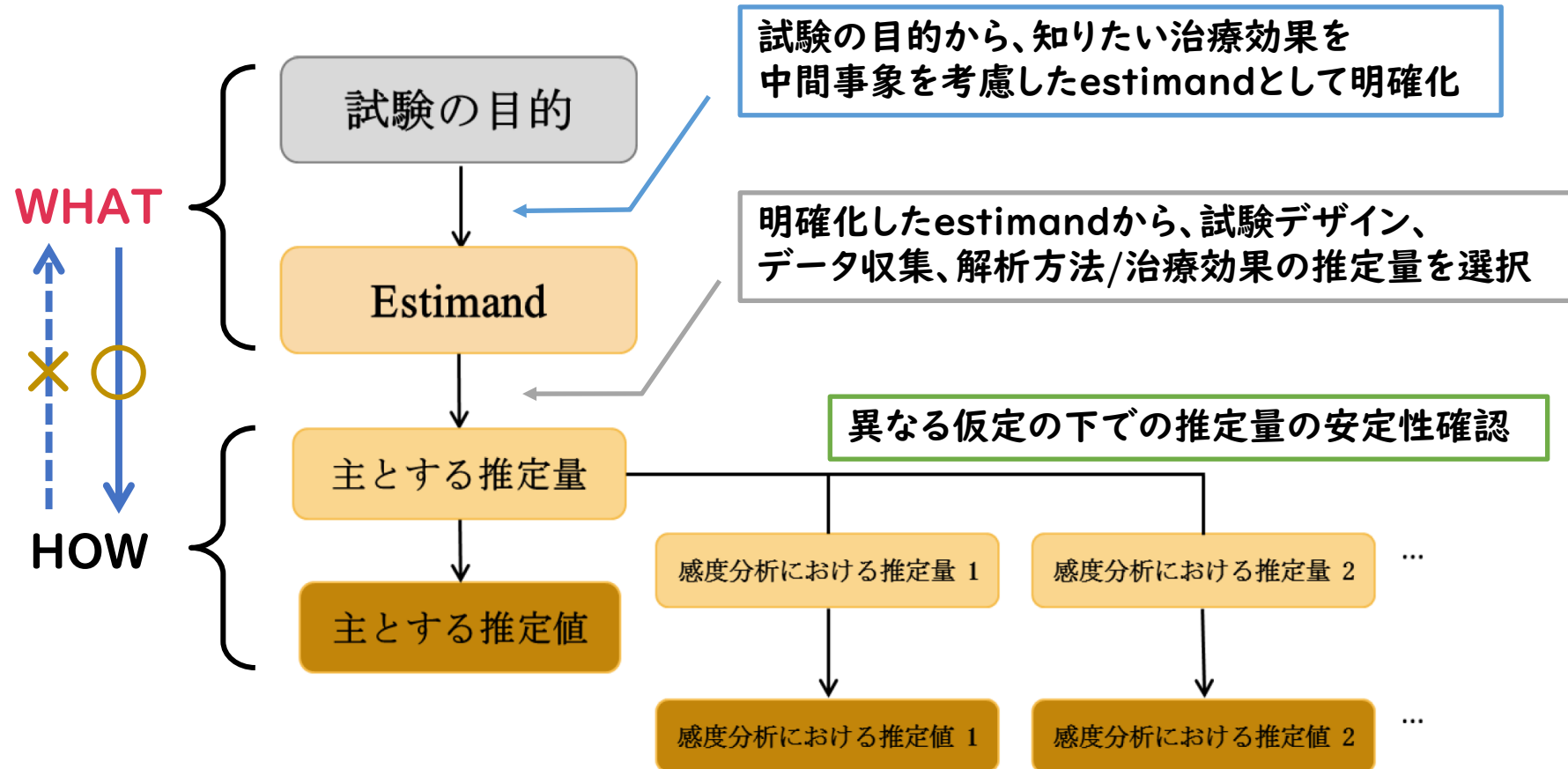


図 1: 定められた試験の目的に対して、推定の対象、推定の方法及び感度分析を整合させる

試験の目的によってestimandが決まる ⇒ How(どのように)からWHAT(何を)に重点を

Estimandによって規定される治療効果

- 中間事象に対応するストラテジーを考慮し、4つの要素からestimandを構成 ⇒ その試験で知りたい/推定したい治療効果を定義



A) 対象集団: 対象となる患者

B) 評価項目: 得るべき変数 (測定値/臨床結果など)

C) 中間事象: 中間変数をどのように考慮するか

D) 要約: 必要に応じて治療条件間の比較のための基礎となる、
集団レベルでの変数の要約指標

estimand

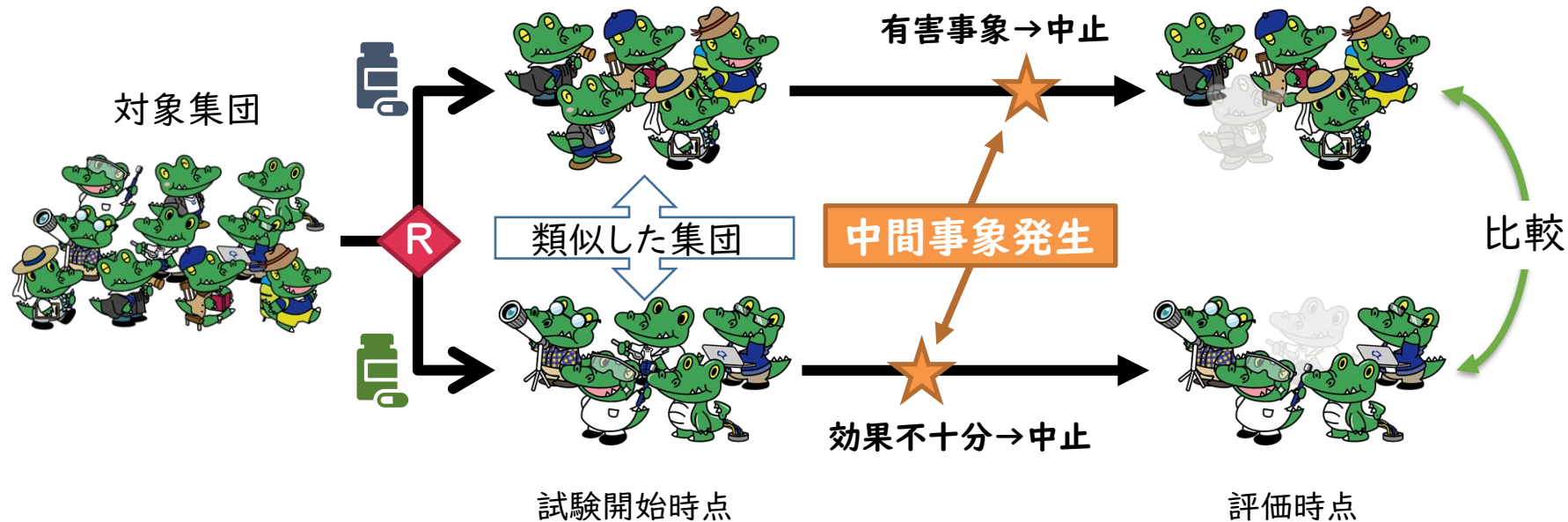
つまり、どのような患者に対して、どのような尺度で、
どのような条件下での治療効果を、どのように要約するか

中間事象に対応するストラテジー

- 治療方針ストラテジー (ITTの原則)
 - 中間事象に関わらず、関心のある変数の値を用いる
- 複合ストラテジー
 - 中間事象の発現を変数の構成要素とし、他の臨床結果の指標と統合して関心のある変数とする
- 仮想ストラテジー
 - 中間事象が発現しなかったという仮想的な状況を想定する
- 主要層ストラテジー
 - 対象集団を、中間事象が起こらないであろう主要層とする
- 治療下ストラテジー
 - 中間事象の発現前までの治療に対する反応を関心の対象とする

中間事象に対応するストラテジー

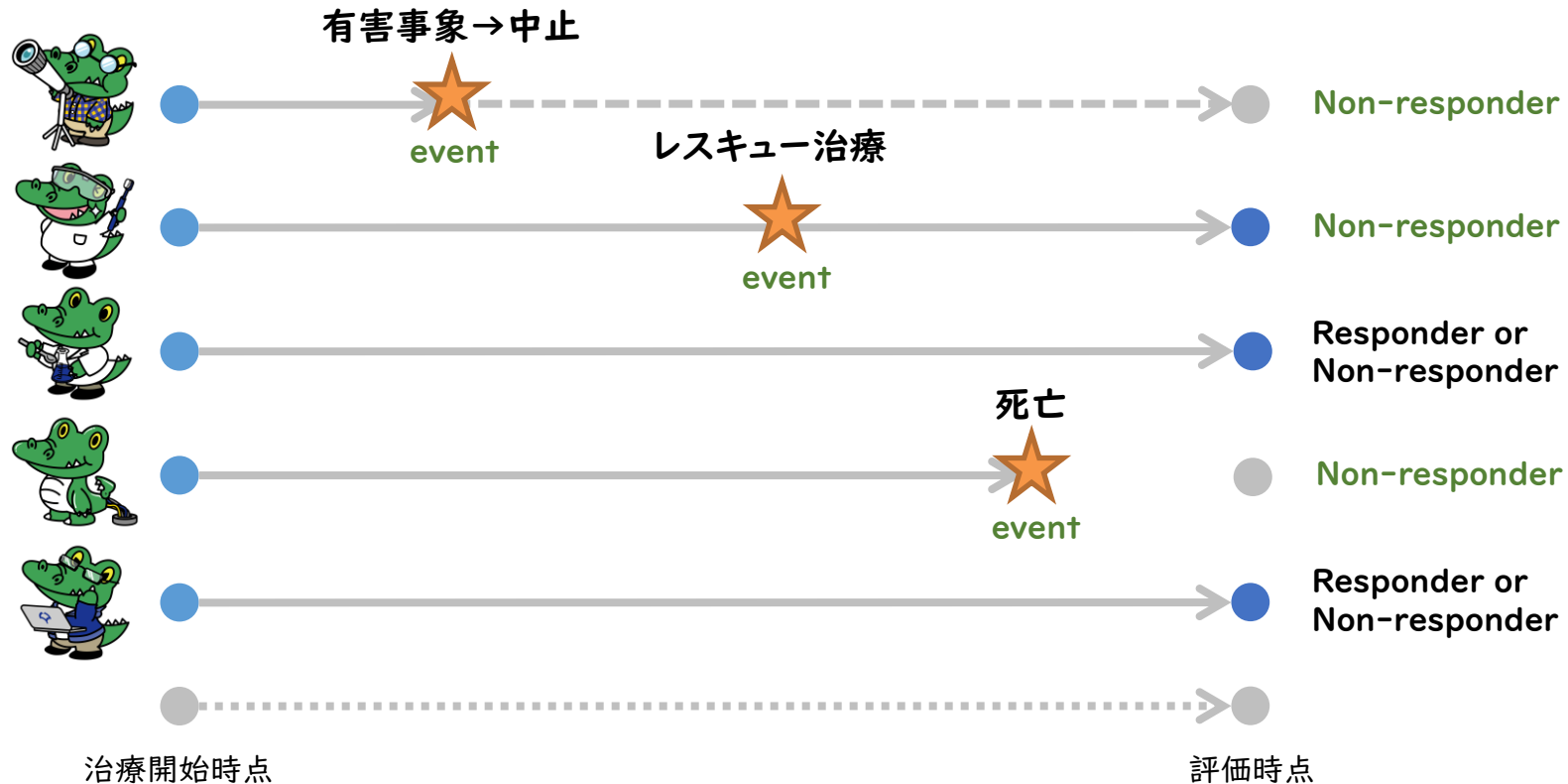
治療方針ストラテジー (ITTの原則)



- 治療継続の有無に関わらずデータを収集し治療効果として評価する
 - ITTの原則に従った治療効果は、実際に受けた治療や遵守状況等に関わらず割付られた治療のまま追跡、評価、解析される「治療方針」の効果

中間事象に対応するストラテジー

複合ストラテジー



- 中間事象の発現例は非反応例とするresponder解析で評価、あるいは、中間事象を含む複合エンドポイントでイベント発生例として評価する など

中間事象に対応するストラテジー

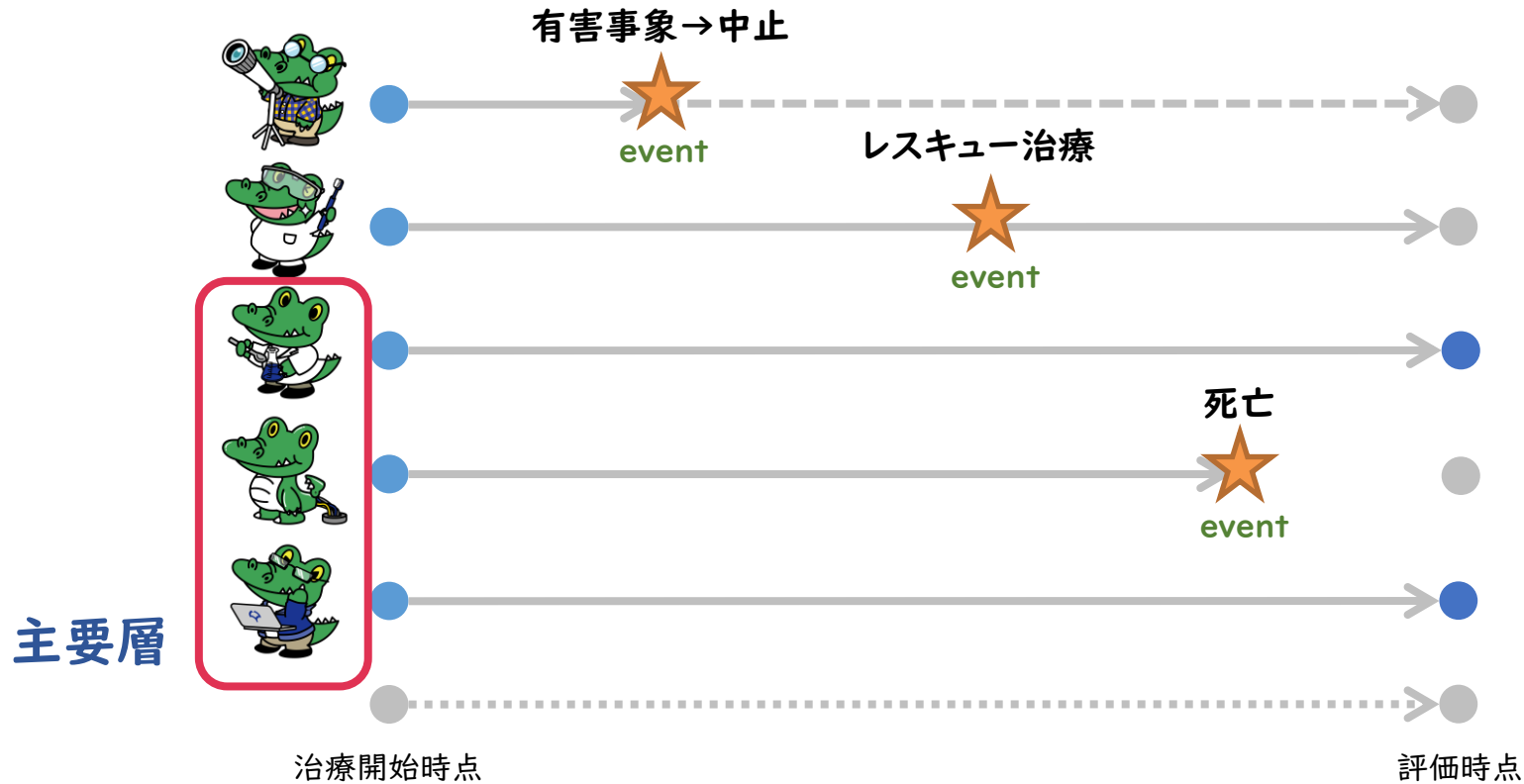
- 仮想ストラテジー：中間事象が発現しなかった状況を想定



- 例) MMRM(mixed effect model for repeated measurements): 中間事象非発現例のデータを利用して群全体の治療効果を推定する方法
 - 中間事象が発現していなければ非発現例と同様な経時変化になるだろうと想定

中間事象に対応するストラテジー

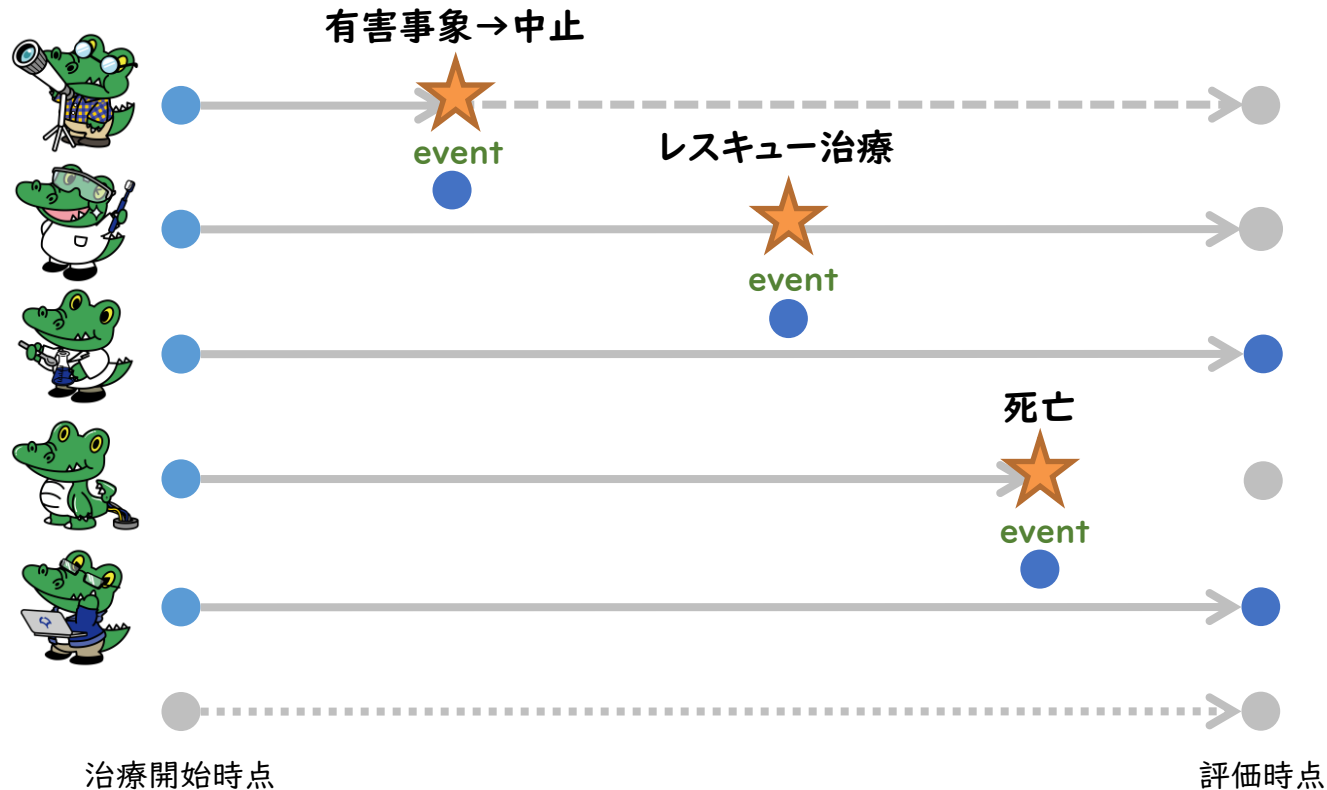
・ 主要層ストラテジー



- PPSや部分集団解析とは異なるもので、試験に先立って被験者を特定できない
- 例) 治療計画を遵守できるであろう患者

中間事象に対応するストラテジー

・ 治療下ストラテジー



- 中間事象の発現時点までのデータで評価
- 例) LOCF/BOCFなど

中間事象後のデータ収集の要否は？

- 中間事象後のデータが必要
 - 治療方針ストラテジー (= ITTの原則)
 - 中間事象後データは必要だが、中間事象による中止、脱落でデータが欠測となることもある
- 中間事象後のデータを利用しない
 - 複合ストラテジー: 中間事象の有無を評価項目に含めた治療効果を評価
 - 仮想ストラテジー: 中間事象後は全て欠測値として取り扱う
 - 主要層ストラテジー: 中間事象が発生しない層だけで治療効果を評価
 - 治療下ストラテジー: 中間事象までのデータで治療効果を評価

とは言え、主要/副次評価項目や安全性データは、試験結果や解釈、意思決定に利用されるため (**Critical data**)、中間事象が発生した理由を含めて、中間事象後のデータも可能な限り収集しておく必要がある

試験の目的とestimand

- 例) 風邪に罹患した患者(P)を対象として、カゼナオール(I)の投与3日後の治癒割合がプラセボ(C)よりも有意に高いこと (O)を検証する
 - 対象集団: 風邪に罹患した患者
 - 診断ガイドライン、適切な選択・除外基準により定義される集団
 - 評価項目: 投与3日後における治癒の有無
 - 風邪症状(体温、のど・せき・はなみず症状スコア)の合計スコアが3点以下を治癒と定義
 - 中間事象: 治療中止、レスキュー薬の使用など
 - 中間事象の有無にかかわらず、投与3日後での治癒の有無を評価する
 - 全ての被験者が計画通りに投与を完了したと想定、など
 - 要約
 - 評価項目である治癒に至った各群での患者の割合の差

つまり、どのような患者に対して、どのような尺度で、
どのような条件下での治療効果を、どのように要約するか

EPO試験の目的とestimandは？

- 初回発症のST上昇型心筋梗塞患者で発症から12時間以内に再灌流に成功した低心機能患者(P)を対象として、PCI後6時間以内にエポエチンベータ(I)2用量(6000単位, 12000単位)を単回投与すると、プラセボ(C)と比較して、左室駆出率が維持される(O)かを検討する

- 対象集団: PICOのPであり、適格基準を満たす集団
 - 評価項目: 左室駆出率(LVEF)の急性期から6カ月後までの改善度
 - 中間事象: 逸脱・中止、心筋シンチ未実施、死亡など
 - 2時点でしか実施されない薬剤負荷シンチ/安静時シンチ検査
 - ストラテジーは治療方針?治療下?仮想?
 - 要約: LVEFの急性期から6カ月後までの平均変化量
 - 用量反応性(Williamsの方法): ①プラセボ vs 12000単位 → ①が有意な場合に限り、②プラセボ vs 6000単位
- 起こりえる中間事象を想定し、Estimandに即したデータを収集できる計画を!!

統計解析と症例取扱い

• 症例の採否/データ取扱い規準

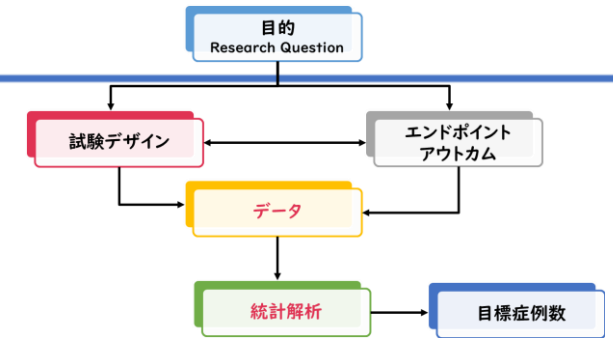
- 集積された症例をあらかじめプロトコルで計画された解析をする上で、どのように取り扱うかを事前に規定

• 症例検討会

- 症例の採否/データの取扱い規準を満たすか否かの確認
- 計画時には想定していなかった逸脱例や違反例や中間事象の取扱いを決定

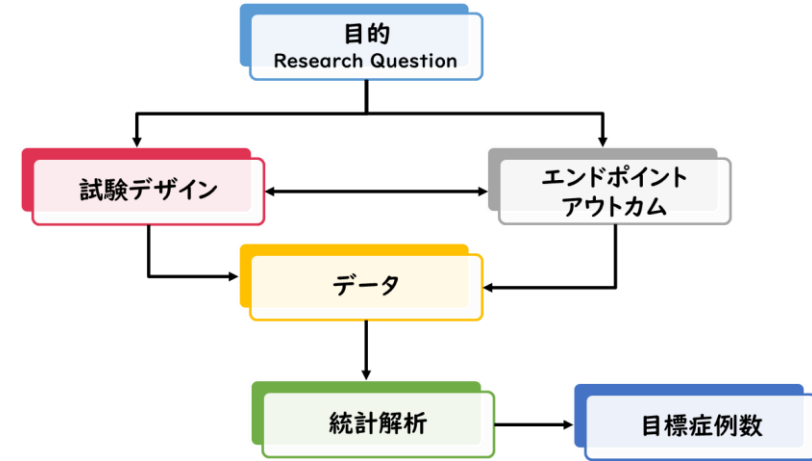
• 例) COVID-19感染拡大による想定されなかった中間事象

- 感冒症状のため、院内感染対策に則り規定の来院ができず、規定されたvisitでの試験薬の投与ができず、各種検査も欠測となった症例の採否は？



まとめ

- 臨床試験を成功させるには
 - エンドポイント: endpoint
 - ランダム化: randomization
 - 交絡因子: confounding factor
 - バイアス: bias



を十分に考慮したセンスのある広義の試験デザインの計画が必須条件

- 計画段階から生物統計家だけでなく、データマネージャの協力・参加が必要
- 眼の前の1つの臨床試験だけではなく、研究全体のゴール、ステークホルダーの要求水準やestimandなどを見据え、統計学的な目及び広い視野をもって研究の計画・実施・まとめができるDM活動を!!

参考資料

- 今後、統計担当者とより深いコミュニケーションをとり、臨床試験を成功させるために知っておきたい統計の知識として
 - ICH E9: 臨床試験のための統計的原則
 - ICH E10: 臨床試験における対照群の選択とそれに関連する諸問題
 - 統計学のセンス (新版). 丹後俊郎著 (朝倉出版)
 - 無作為化比較試験 (新版). 丹後俊郎著 (朝倉出版)
 - 臨床研究デザイン. 折笠秀樹著 (真興交易医書出版部)
 - これからの臨床試験. 椿広計・藤田利治・佐藤俊哉編集 (朝倉書店)
 - 臨床試験のデザインと解析. 角間辰之・服部聡著 (近代科学社)