東京大学大学院
情報学環・学際情報学府
The University of Tokyo III/GSII

ベイズを用いた臨床試験と
そこから得られる結果の解釈

東京大学大学院　情報学環
大庭　幸治
obakoji@iii.u-tokyo.ac.jp

Use of Bayesian Methodology in Clinical Trials of Drug and Biological Products

Guidance for Industry

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

Comments and suggestions regarding this draft document should be submitted within 60 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance. Submit electronic comments to https://www.regulations.gov. Submit written comments to the Dockets Management Staff (HFA-305), Food and Drug Administration, 5630 Fishers Lane, Rm. 1061, Rockville, MD 20852. All comments should be identified with the docket number listed in the notice of availability that publishes in the *Federal Register*.

For questions regarding this draft document, contact (CDER) Scott Goldie at Scott.Goldie@fda.hhs.gov, or (CBER) Office of Communication, Outreach and Development, 800-835-4709 or 240-402-8010.

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)

January 2026
Biostatistics

https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-bayesian-methodology-clinical-trials-drug-and-biological-products

The BMJ
Cite this as: *BMJ* 2026;392:s180
http://doi.org/10.1136/bmj.s180
Published: 28 January 2026

FDA is now "open to bayesian statistics": transformational change or new Pandora's box?

Peter Doshi

The US Food and Drug Administration (FDA) is now "open to bayesian statistics," contrasting this with the frequentist approach that the agency and the drug industry have historically relied on for statistical analysis.

In a video posted to X on 12 January the FDA commissioner, Marty Makary, said that the agency had published a new guidance document "to encourage the use of bayesian statistics in clinical trial design and the readout of results" in new drug and biologic applications.[1]

The 29 page document focuses on incorporating bayesian methods into one of the FDA's most crucial functions: "primary inference in clinical trials intended to support the effectiveness and safety of drugs."[2]

Describing the change as "a leap forward beyond the frequentist model of analysing data," Makary said that the approach would improve the FDA's processes in designing and analysing clinical trials, lowering drug development costs and shortening the timeline for getting new treatments to market.

Publishing draft guidance on this topic was a formal commitment the FDA made with the industry during the Biden administration in 2022, as part of the Prescription Drug User Fee Act VII.[3] One major way the change could influence drug and biologic approval is the leveraging of phase 2 trial results in phase 3 study results.

Richard Lilford, professor of public health at the University of Birmingham, UK, has long called for greater adoption of bayesian approaches, such as in drug development for rare diseases, and was excited by the new guidance.

"It's good that after years of prompting, a decision body has decided to accept 'grown up' statistics," Lilford told *The BMJ*. While emphasising the need for countering companies with strong vested interests, he said that the FDA's announcement represented a "potentially transformational change."

But the move garnered a far more sceptical response from Sander Greenland, emeritus professor of epidemiology and statistics at University of California Los Angeles, who has studied statistical methods. "All this talk about frequentist versus bayesian is misdirected," he said, adding that almost any bayesian analysis could be duplicated numerically in a frequentist mode. But he warned, "By calling it bayesian, you now mystify it and you open a door for abuse."

At a glance: What are bayesian statistics?

A bayesian approach to statistical analysis combines collected study data with external sources of information—such as pharmacokinetic or pharmacodynamic data, other clinical trials, observational data, or expert opinion—to determine an outcome. It differs from how frequentist statistical approaches are commonly applied, in which only study data are assessed. It is named after the 18th century mathematician and theologian Thomas Bayes.

The perceived problem

Historically, industry sponsored clinical drug trials have been analysed using a set of statistical approaches classified as "frequentist." One prominent practice associated with frequentist methods is null hypothesis significance testing, especially at the 0.05 significance cut-off for the corresponding P value.

Such conventions have aided "go/no-go"-type decisions based on whether a result is—or is not—deemed "statistically significant" (meaning P≤0.05). But for decades statisticians and other scholars have decried the tyranny of "statistical significance" and P values, criticising them for encouraging mistaken decisions regarding efficacy and safety.[4]

Small P values were often incorrectly interpreted as indicating clinical or practical significance. And conversely, P values greater than 0.05 have been used to incorrectly conclude "no effect" when one exists.

Despite the criticism, significance testing has remained a fundamental part of drug regulation. "It seems frequentism becoming the dominant approach led to a certain sloth in applying it," Greenland said, calling approaches such as significance testing "unimaginative, oversimplified, [and] automatic."

Bayesian techniques, by contrast, seemed to offer an innovative alternative. Whereas, by frequentist convention, clinical trials are analysed in isolation, in a bayesian analysis the study data are combined with other sources of information.

The idea of synthesising information across experiments to draw new probability statements about a hypothesis, such as whether a drug is effective, appeals to many. "It means that analyses can take account of all we know, not just the data on one trial in isolation," says Lilford—something that he has argued can help bring new treatments for rare diseases to market.

Industry figures are also likely to welcome the news. In 2023, influential voices in the biopharmaceutical space writing in *Nature Reviews Drug Discovery* called on regulators to go bayesian.[5] They said, "We believe

Doshi P. BMJ. 2026;392:s180.

2026/02/27

産業界向けガイダンス

医薬品および生物学的製剤の臨床試験におけるベイズ流方法論の利用：ドラフトガイダンス

米国保健福祉省　食品医薬品局 (FDA)

医薬品評価研究センター（CDER）

生物製剤評価研究センター（CBER）

2026 年 1 月　生物統計学

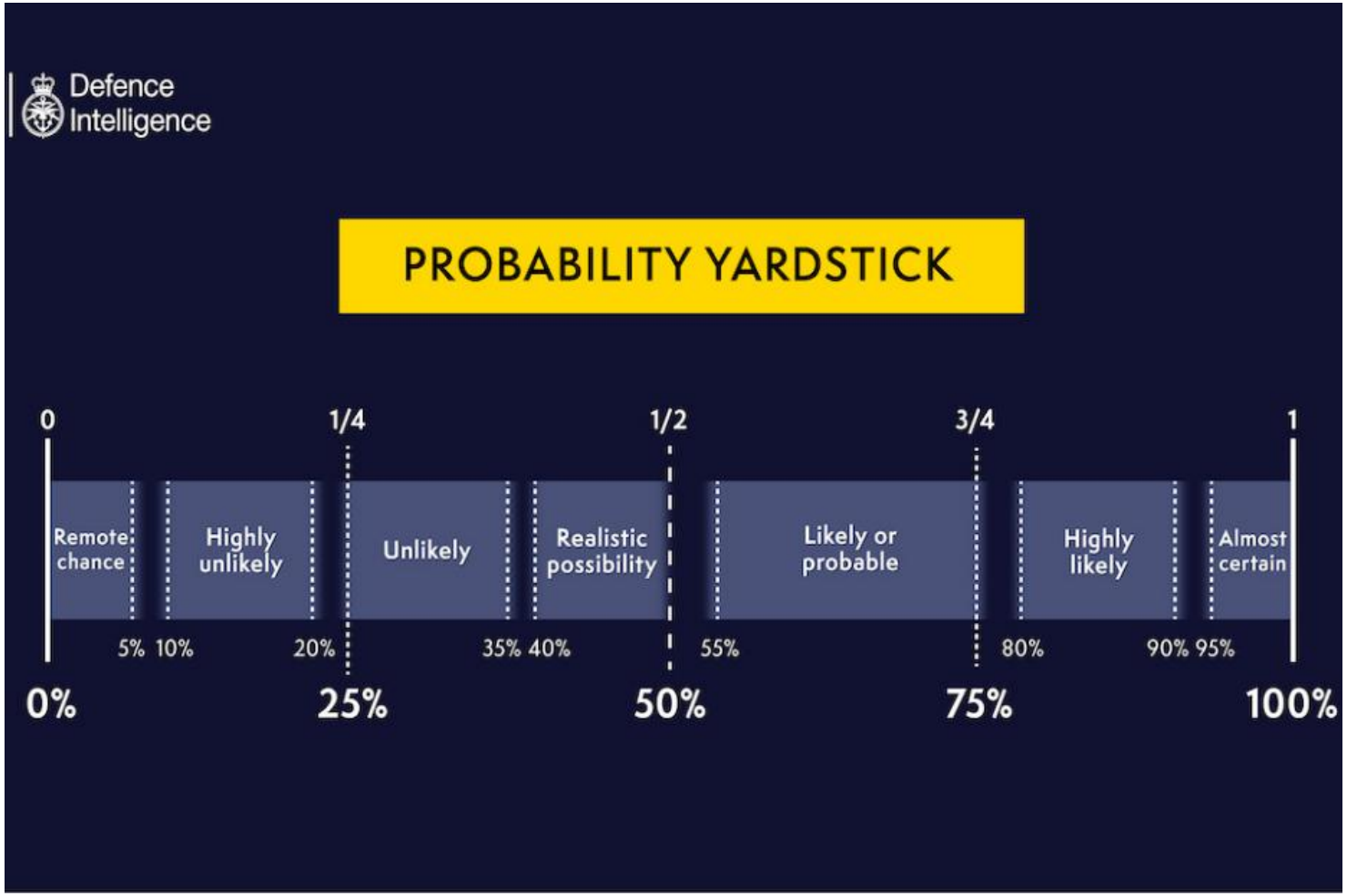
訳：三島 遼、手良向 聡

（京都府立医科大学 大学院医学研究科生物統計学／附属病院臨床研究推進センター）

ベイズを用いる理由：不確実性の取り扱い

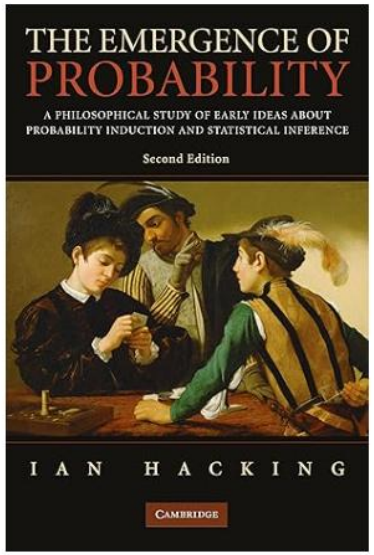
- 不確実性（Uncertainty）
  - 知らないということをはっきり認識していること
  - conscious awareness of ignorance
- 不確実性の程度を「言葉」で表現
  - 言葉が指す意味に共通認識はない
- 不確実性の程度を「数値」で表現
  - ‘likely’とは？

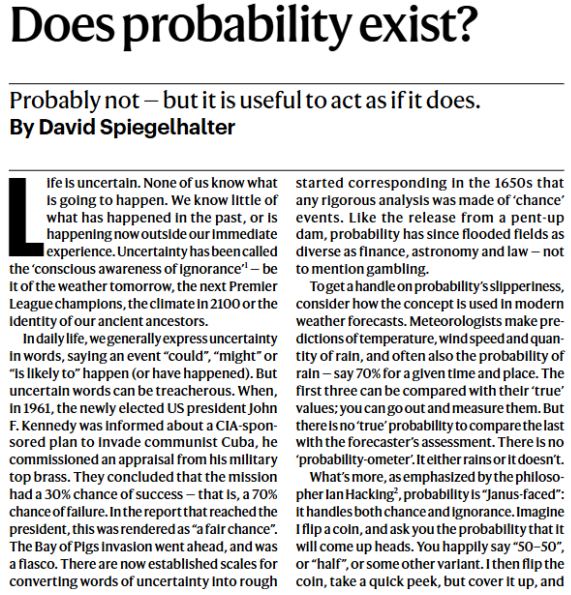
英国防情報局における確率の物差し



確率の2つの概念

- Aleatory Probability (Chance)
  - 純粋にランダムな現象から生じる不確実性の程度の大きさ
- Epistemic Probability (Ignorance)
  - 知識が不十分であることから生じる不確実性の程度の大きさ





Does probability exist?

Probably not – but it is useful to act as if it does.
By David Spiegelhalter

Life is uncertain. None of us know what is going to happen. We know little of what has happened in the past, or is happening now outside our immediate experience. Uncertainty has been called the 'conscious awareness of ignorance'[1] – be it of the weather tomorrow, the next Premier League champions, the climate in 2100 or the identity of our ancient ancestors.

In daily life, we generally express uncertainty in words, saying an event "could", "might" or "is likely to" happen (or have happened). But uncertain words can be treacherous. When, in 1961, the newly elected US president John F. Kennedy was informed about a CIA-sponsored plan to invade communist Cuba, he commissioned an appraisal from his military top brass. They concluded that the mission had a 30% chance of success – that is, a 70% chance of failure. In the report that reached the president, this was rendered as "a fair chance". The Bay of Pigs invasion went ahead, and was a fiasco. There are now established scales for converting words of uncertainty into rough started corresponding in the 1650s that any rigorous analysis was made of 'chance' events. Like the release from a pent-up dam, probability has since flooded fields as diverse as finance, astronomy and law – not to mention gambling.

To get a handle on probability's slipperiness, consider how the concept is used in modern weather forecasts. Meteorologists make predictions of temperature, wind speed and quantity of rain, and often also the probability of rain – say 70% for a given time and place. The first three can be compared with their 'true' values; you can go out and measure them. But there is no 'true' probability to compare the last with the forecaster's assessment. There is no 'probability-ometer'. It either rains or it doesn't.

What's more, as emphasized by the philosopher Ian Hacking[2], probability is "Janus-faced": it handles both chance and ignorance. Imagine I flip a coin, and ask you the probability that it will come up heads. You happily say "50–50", or "half", or some other variant. I then flip the coin, take a quick peek, but cover it up, and

Spiegelhalter D. Nature. 2024;636:560-563.

頻度論の確率、ベイズ論の確率

- 頻度論による確率（Aleatory）
  - 現象のランダムネスによる不確実性
  - ランダムな現象を繰り返し観察した際に得られる結果の相対的頻度の極限
    - 治療効果の大きさ：定数（固定値）
    - 信頼区間：解釈が難しい（真値が含まれる確率？、信頼区間の端の値は起きにくい？）
- ベイズ論による確率（Epistemic）
  - 知識の不確実性
  - 事前知識を取り込んだ形でデータ取得後の結果に対する事後確率（ベイズの定理）
    - 治療効果の大きさ：確率的に分布（事前確率から事後確率への更新）
    - 信用区間：解釈が容易（真値が含まれる確率、信用区間の端の値は起きにくい）
    - 結果の確率的表現：Pr(Benefit) = 0.85

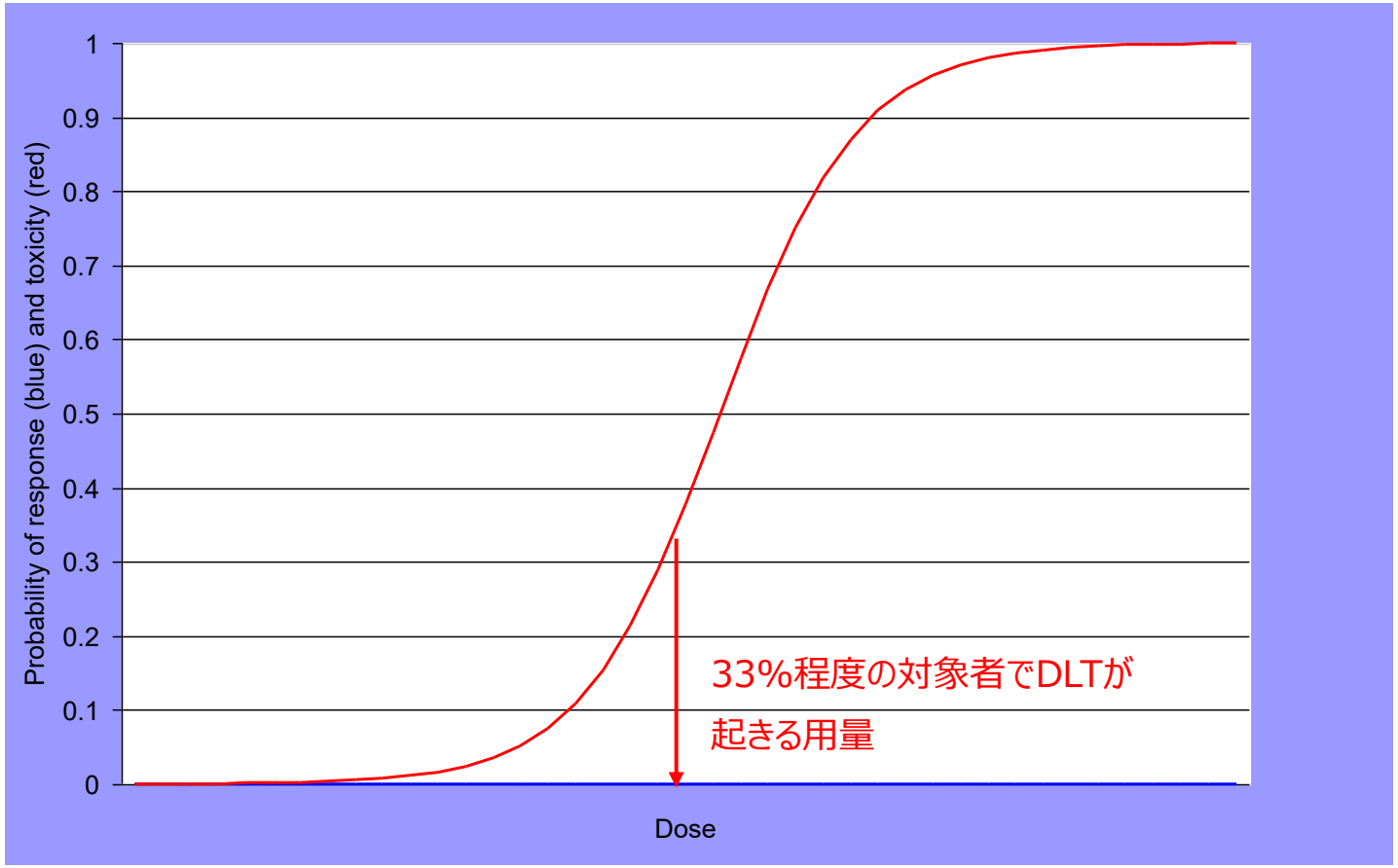
ベイズを用いる理由：より柔軟で効率的

- 既存の利用可能な試験あるいは単一試験内の集団に渡る情報の借用
  - がん領域における用量設定試験
    - 事前の患者の毒性状況で用量反応曲線を随時更新
  - 適応的デザインにおける途中までの情報を利用した試験計画変更
  - 患者集団間での情報借用
    - 成人の試験結果を小児の研究に利用
  - 先行臨床試験の情報借用
  - などなど

実際の事例を見てみよう

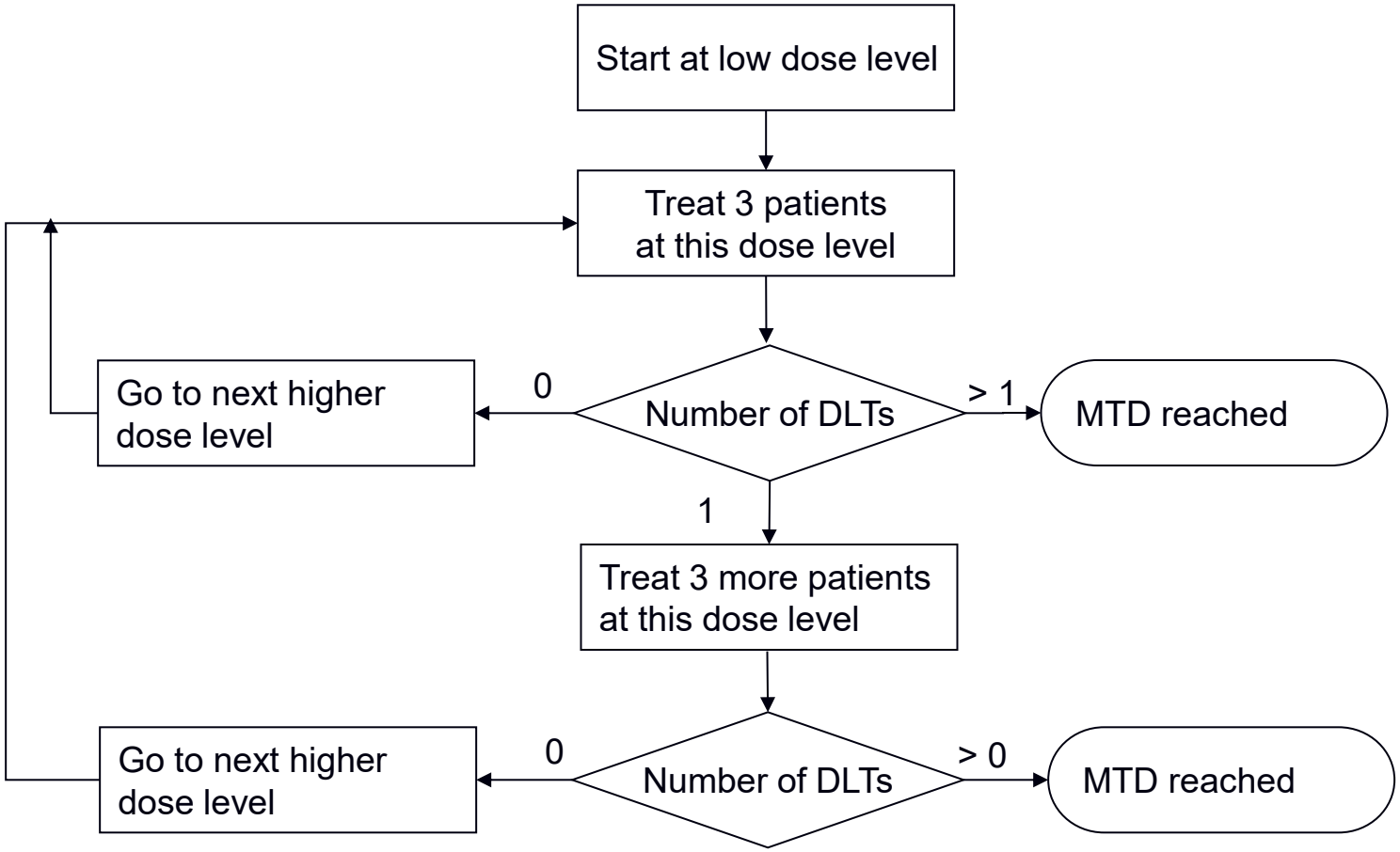
事例1：がん領域における用量設定試験

- 抗がん剤開発 第1相試験の目的
  - 最大耐用量（MTD：Maximum Tolerated Dose）の決定



古典的なルールベースのアプローチ

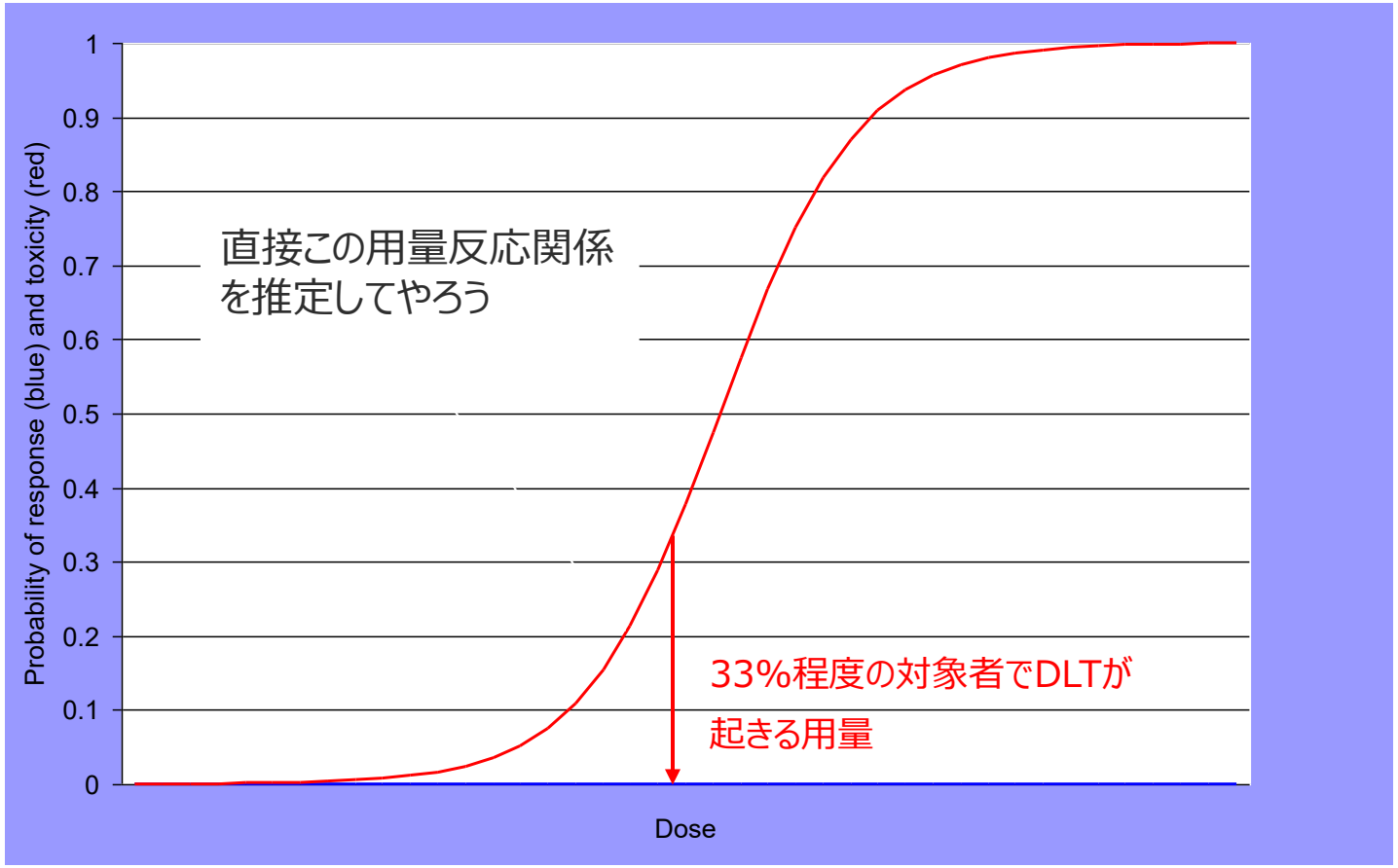
- 33%を決めるための3例コホート法



3例コホート法の問題点

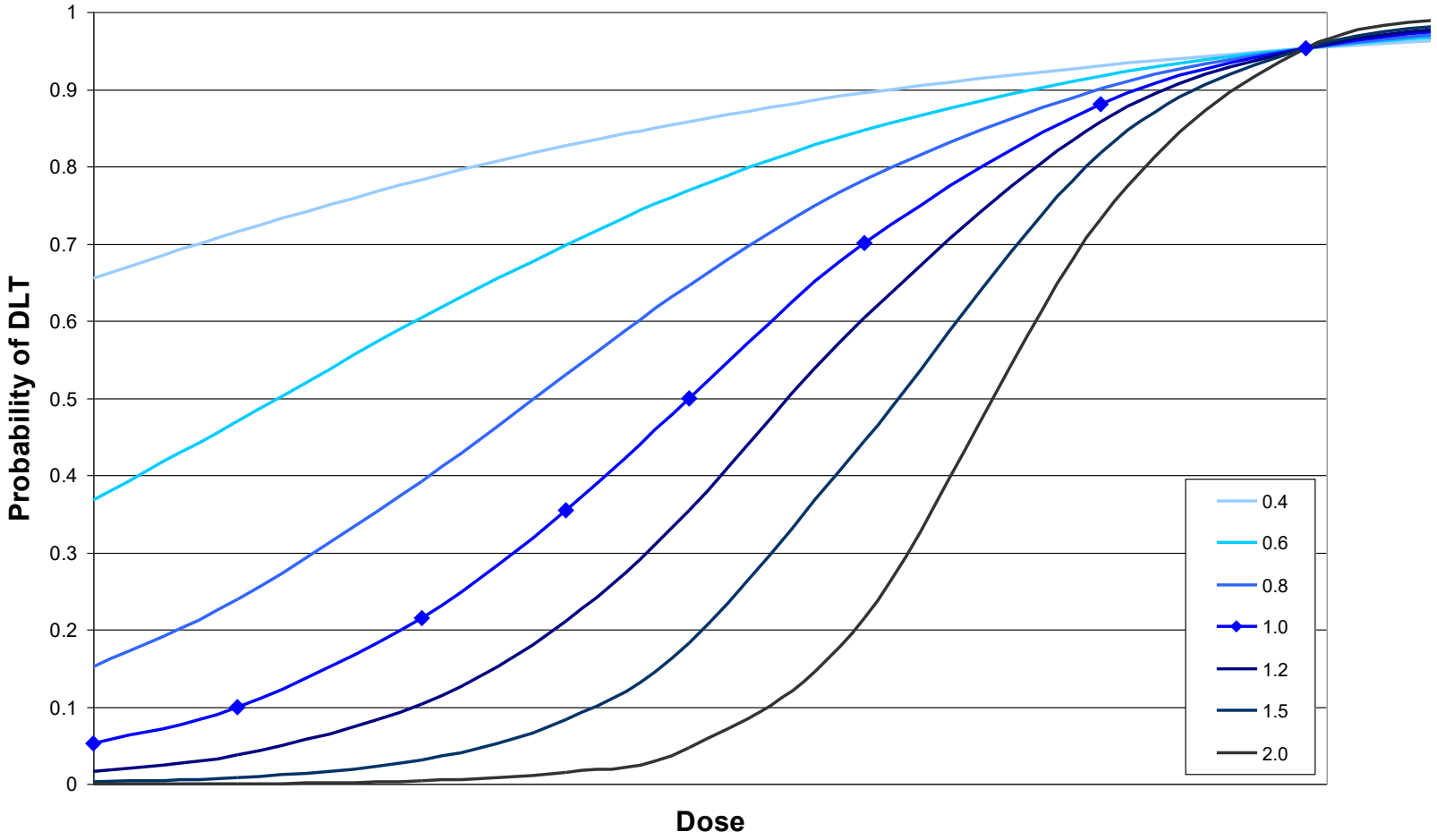
- 統計的な考えを全く利用しないこともあり、実施が簡単
  - 仕組みも理解しやすい？
- 統計的な考えを利用しないので、良くも悪くも偶然の結果に簡単に影響を受けてしまう

CRM(Continual Reassessment Method)



O'Quigley J, Pepe M, Fisher L. Biometrics. 1990;46(1):33-48.

用量反応関係に用いるモデル

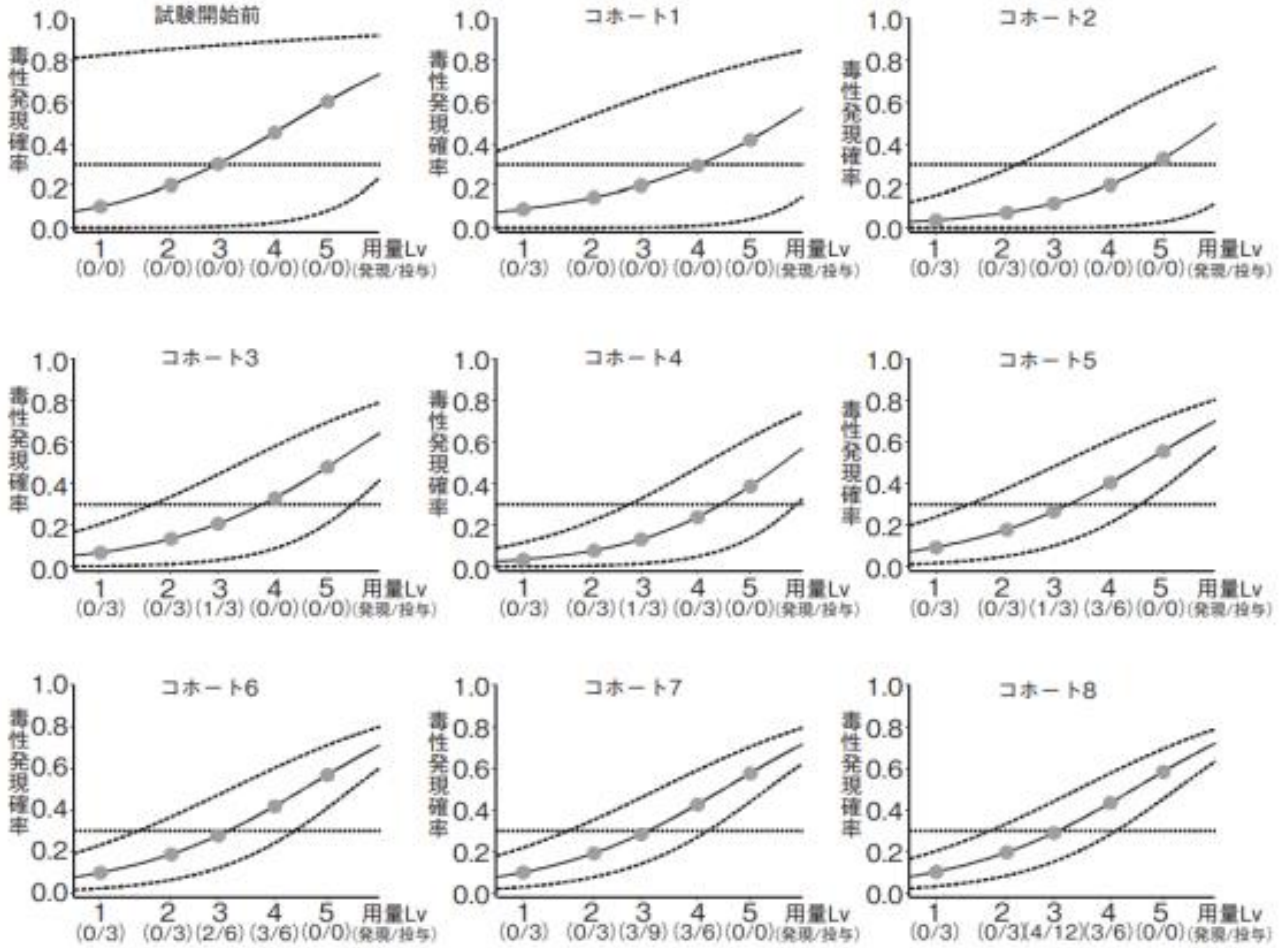


$$\psi_a(x_i) = \exp(3+ax_i)/[1+\exp(3+ax_i)]$$

CRMの基本的な手順

- 患者にその時点で最もDLT発現確率が33%に近い用量を投与する
- その患者でのDLT発現有無を評価し、その結果に応じて用量反応関係のモデルを更新する
  - ベイズ的アプローチ
- 統計的に情報が十分であるという基準に達した時点で終了する
  - 開始は低用量から、変更は1用量レベルまで、少なくとも同じ用量レベルに2人は入る、といった工夫は色々可能

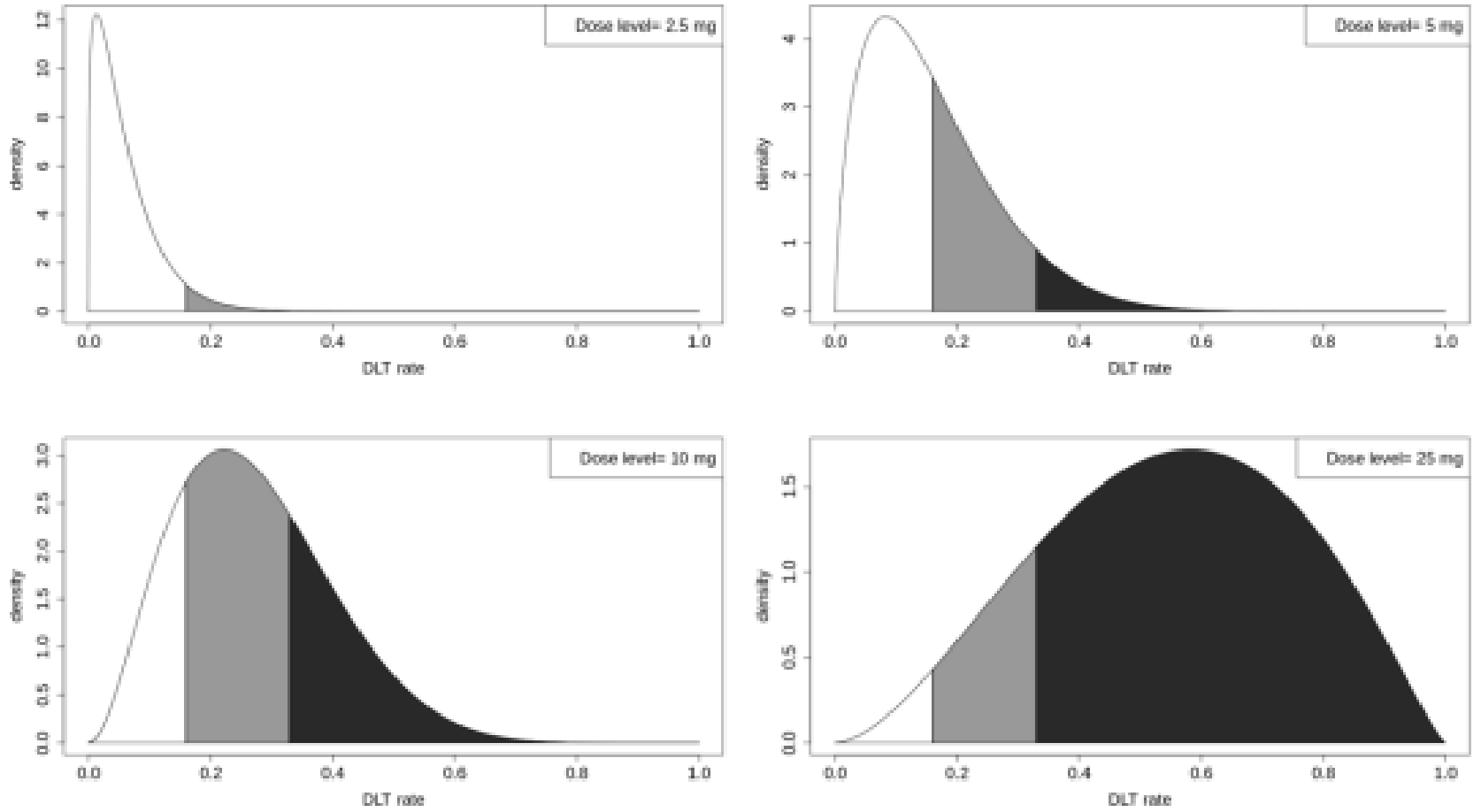
用量反応曲線の推移



実線：毒性発現確率の平均値；点線：90%信用区間；破線：標的毒性発現確率

矢田 真城, 浜田 知久馬. 臨床薬理. 2015;46:11-19.

各用量での毒性発現確率の事後分布



過少用量区間（白色）：DLT 発現確率が[0, 0.16]、標的用量区間（灰色）：(0.16, 0.33]、過大用量区間（黒色）：(0.33, 1.0]、DLT 発現確率が各区間範囲である確率は各区間の曲線下面積に対応

様々な発展：モデル支援型デザイン

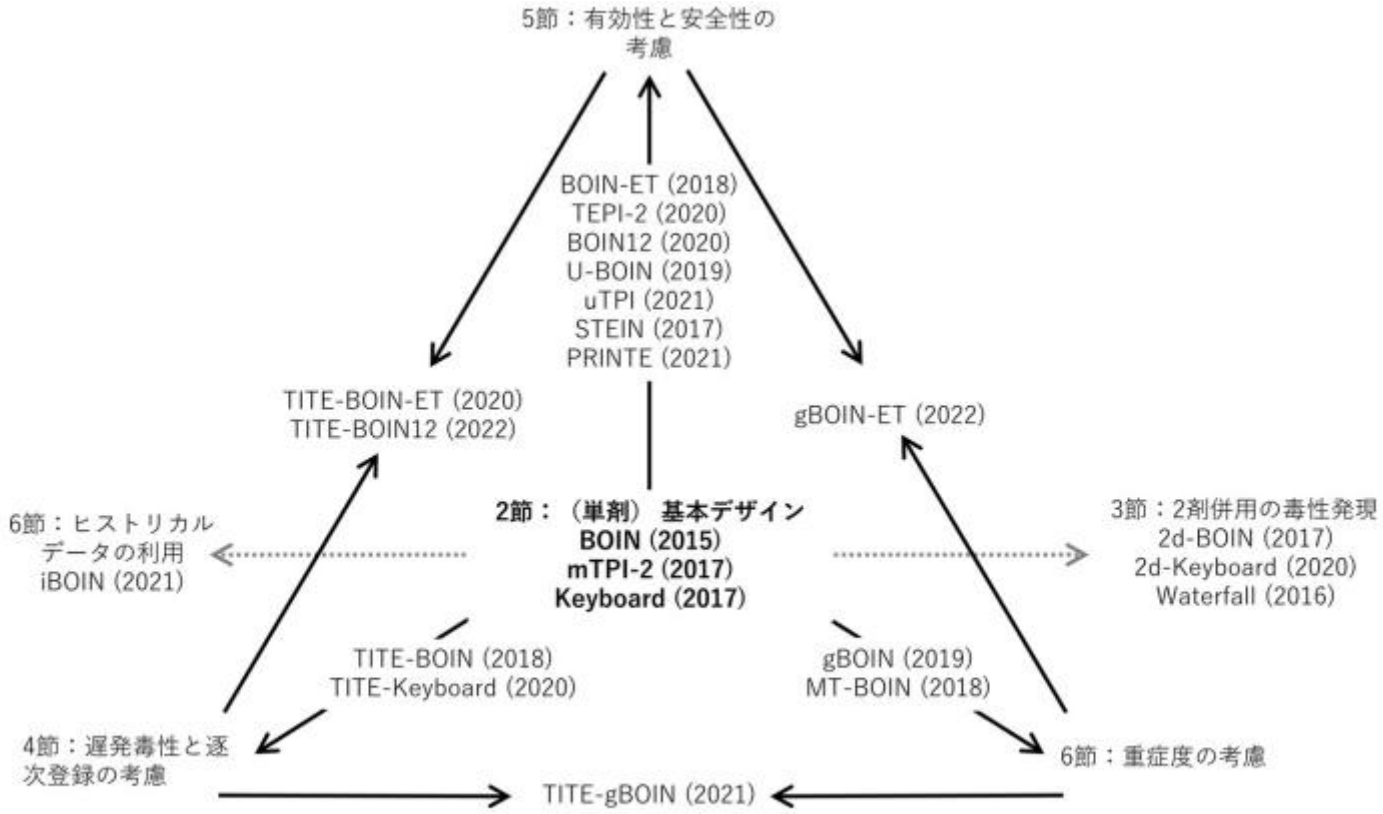
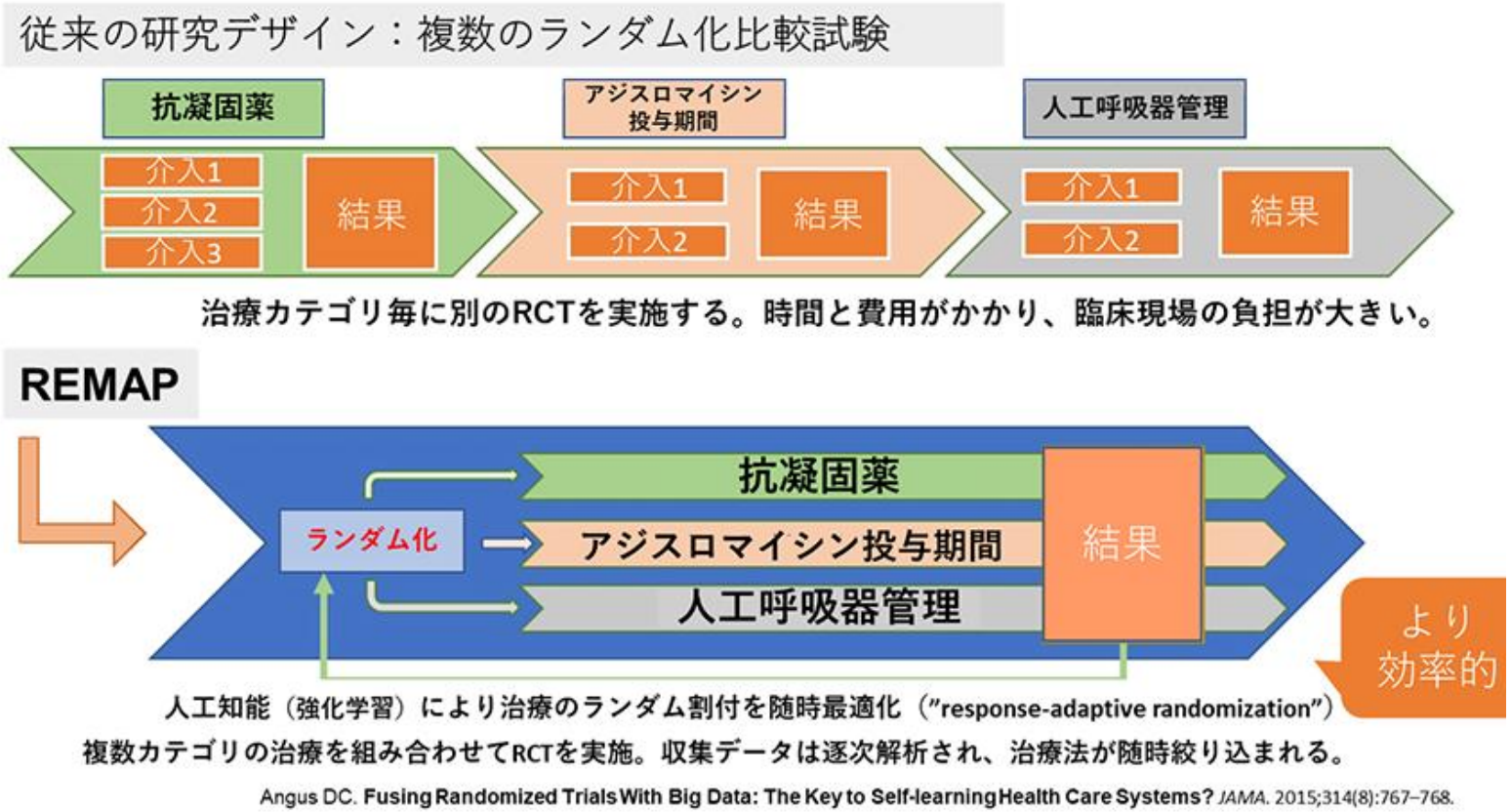


図 1. モデル支援型デザインの発展の変遷

橋詰公一, 他. 計量生物学. 2022;43:3-36.

事例2：適応的デザインにおける利用

- REMAP-CAP Studv

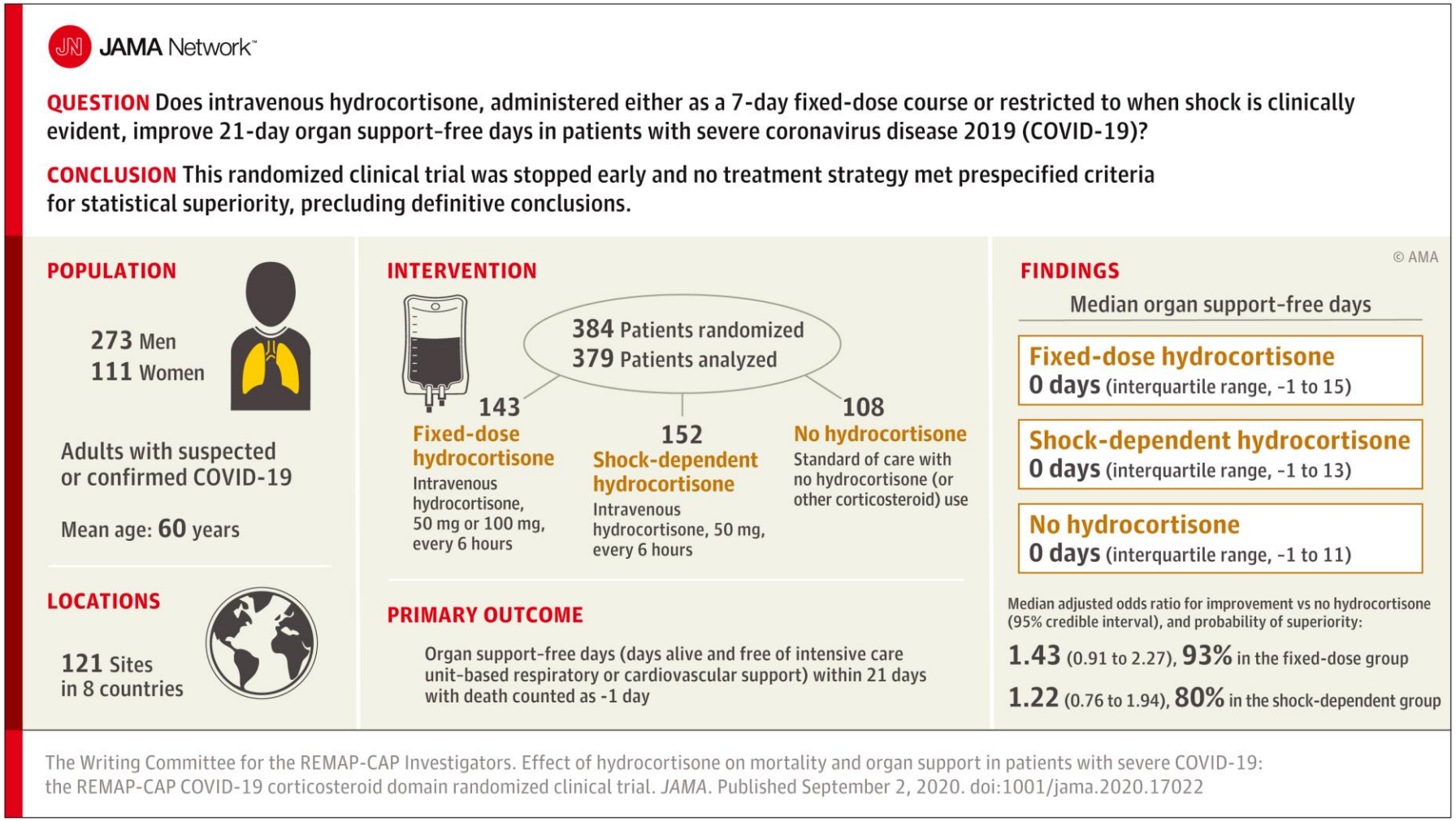


Angus DC. **Fusing Randomized Trials With Big Data: The Key to Self-learning Health Care Systems?** *JAMA*. 2015;314(8):767–768.

Table 1.1: List of all interventions to which a patient may be allocated.

Code	Intervention
Antibiotic	
A1	Ceftriaxone + Macrolide
A2	Moxifloxacin or Levofloxacin
A3	Piperacillin-Tazobactam + Macrolide
A4	Ceftaroline + Macrolide
A5	Amoxicillin-Clavulanate + Macrolide
Macrolide Duration	
M1	Standard course (3 to 5 days)
M2	Extended course (14 days)
Corticosteroid	
C1	No corticosteroids
C2	Hydrocortisone (50mg)
C3	Shock dependent hydrocortisone
C4	High-dose hydrocortisone (100mg)
Antiviral	
I1	No antiviral
I2	Oseltamivir 5 days
I3	Oseltamivir 10 days
COVID-19 Antiviral	
X1	No antiviral for COVID-19
X2	Lopinavir/ritonavir
X3	Hydroxychloroquine
X4	Hydroxychloroquine + lopinavir/ritonavir
COVID-19 Immune Modulation	
Y1	No immune modulation for COVID-19
Y2	Interferon-Beta-1a
Y3	Anakinra
Y4	Tocilizumab
Y5	Sarilumab
COVID-19 Immunoglobulin	
P1	No Immunoglobulin against COVID-19
P2	Convalescent plasma
P3	Delayed convalescent plasma
COVID-19 Therapeutic Anticoagulation	
H1	Standard practice thromboprophylaxis
H2	Therapeutic anticoagulation
Vitamin C	
L1	No vitamin C
L2	Vitamin C

REMAP CAP Corticosteroidドメイン



REMAP CAPの適応的要素

- Response-adaptive ランダム化
  - 試験が進むにつれて、成績が良いと推定される群への割付確率を高める
    - 各治療法の有効性に関する事後分布を更新し、割付確率に反映
- 逐次的に優越性・無益性・同等性の判定
  - 優越性：ある治療が最善である事後確率が 99%以上に達した場合
  - 無益性：対照群より優れている（オッズ比が1を超える）事後確率が5%未満に低下した場合
  - 同等性：事前に定義された「同等範囲」に収まる事後確率が 90%以上に達した場合

メインの結果の提示

- エンドポイント：順序尺度で評価されるOrgan Support-Free Days
  - 22段階：死亡は-1で、それ以外は入院後20日までの日数
- 解析モデル：共変量を調整した比例オッズモデル
- 事前分布：22次元のディリクレ分布
- 事後分布：MCMC法による事後分布推定

事前分布

OSFD	Control	Hydrocortisone	Total	Prior parameters	Cumulative odds ratios
−1	33	41	74	0.295	1.14
0	22	29	51	0.225	1.14
1	2	2	4	0.015	1.17
2	1	0	1	0.015	1.22
3	4	1	5	0.015	1.39
4	1	2	3	0.015	1.37
5	2	1	3	0.015	1.45
6	4	1	5	0.015	1.68
7	1	3	4	0.015	1.61
8	1	3	4	0.015	1.54
9	2	2	4	0.015	1.59
10	1	3	4	0.015	1.53
11	5	1	6	0.030	1.94
12	8	1	9	0.030	3.25
13	1	4	5	0.030	3.10
14	1	5	6	0.030	2.85
15	1	5	6	0.030	2.60
16	3	3	6	0.030	3.26
17	2	12	14	0.030	2.39
18	0	6	6	0.030	1.52
19	4	7	11	0.030	1.88
20	2	5	7	0.060	–
total	101	137	238	1	–

Note: The assumed prior parameters specify a 22-dimensional Dirichlet distribution which partly defines the prior distribution for the Bayesian proportional odds model.

解析方法は非常に複雑

of the adaptive design rules. If both hydrocortisone groups had effect sizes (odds ratios) of 1.75 compared with the no hydrocortisone group, there would be 90% power to determine whether either group was superior to the no hydrocortisone group with a sample size of 500 patients. If the effect was 1.5, there would be 90% power with a sample size of 1000 patients.

Statistical Analysis

The SAP for the COVID-19 corticosteroid domain was written by blinded steering committee members, posted online (https://www.remapcap.org/) before data lock and analysis, and appears in Supplement 1). The primary analysis was generated from a bayesian cumulative logistic model, which estimated posterior probability distributions of the 21-day organ support-free days (primary outcome) based on the evidence accumulated in the trial in terms of the observed primary outcome and assumed prior knowledge in the form of a prior distribution. Data from the United Kingdom national clinical audit on all COVID-19 ICU admissions (provided by Intensive Care National Audit & Research Centre, London, United Kingdom) were used to inform prior distributions, necessary for bayesian analyses, including initial estimates of the effect of age on outcome. Prior distributions for treatment effects were neutral.

The primary model adjusted for location (site, nested within country), age (categorized into 6 groups), sex, and time period (2-week epochs). The model estimated treatment effects for each intervention within each domain and prespecified treatment-by-treatment interactions across domains. The primary analysis was conducted on all randomized patients who met severe COVID-19 criteria as of June 17, 2020, and not just those randomized within the corticosteroid domain. This approach allowed maximal incorporation of all information, providing the most robust estimation of the coefficients of all included covariates. Not all patients were eligible for all domains nor for all interventions within each domain (depending on site participation, baseline entry criteria, and patient or surrogate preference). Therefore, the model included covariate terms reflecting each patient's intervention and domain eligibility, such that the estimate of an intervention's effectiveness relative to any other intervention within that domain was generated from those patients who might have been randomized to either.

Because the primary model included information about assignment to interventions within domains whose evaluation is ongoing, it was run by the fully unblinded statistical analysis committee (Supplement 1), which conducts all protocol-specified trial update analyses and reports those results to the data and safety monitoring board. For the primary analysis, the 2 fixed-dose hydrocortisone groups were combined, such that there were 3 groups: fixed-dose, shock-dependent, and no hydrocortisone. The cumulative log odds for the primary end point was modeled such that a parameter greater than 0 reflects an increase in the cumulative odds for the organ support-free day outcome, which implies benefit. The model assumed proportional effects across the ordinal organ support-free days scale. This assumption was assessed by inspection of the distribution for clinically important deviations. Patients missing the primary end point (n = 5) were ignored; there was no imputation of missing primary (or secondary) end point values. A patient who survived to hospital discharge was assumed to be free of organ support through 21 days (last status carried forward).

The model was fit using a Markov Chain Monte Carlo algorithm that drew iteratively (10 000 draws) from the joint posterior distribution, allowing calculation of odds ratios with their 95% credible intervals (CrIs) and the probability that each corticosteroid domain intervention (including the no hydrocortisone group) was optimal, that either hydrocortisone group was superior to no hydrocortisone, and that the fixed-dose and shock-dependent hydrocortisone groups were equivalent. An odds ratio greater than 1 represents more survival and more days free from ICU organ support. Although this analysis was conducted in response to the disclosure of the RECOVERY trial results, it was also the first interim analysis of the COVID-19 patient cohort, which had preexisting internal statistical triggers for trial conclusions and disclosure of results (99% probability of superiority or inferiority, defined as odds ratio >1 and <1, respectively, and 90% probability for equivalence, defined as an odds ratio between 1/1.2 and 1.2).

Analysis of the primary outcome was then repeated in a second model using only data from those patients enrolled in the corticosteroid domain with no adjustment for assignment to interventions in other domains. Although using less information, this analysis is more typical for an RCT. Further secondary analyses explored the effects of excluding patients who were ruled out for COVID-19 (defined as documented negative test results for SARS-CoV-2 infection and no positive test results), of excluding adjustment for site and time epoch, and of combining the fixed-dose and shock-dependent hydrocortisone groups.

Identical analyses were conducted to estimate the effect on mortality, except the outcome was dichotomous (alive or dead at hospital discharge). There were also 7 secondary outcome analyses (all using the corticosteroid domain cohort): time to death, respiratory support-free days, cardiovascular organ support-free days, length of ICU stay, length of hospital stay, the WHO ordinal scale at 14 days, and progression to invasive mechanical ventilation, ECMO, or death in those not receiving invasive mechanical ventilation at enrollment. The time-to-death and length-of-stay outcomes were time-to-event analyses with results expressed as hazard ratios. The proportional hazards assumption was assessed by testing whether scaled Schoenfeld residuals and time were independent ($P > .05$) for each covariate. All 3 models met the assumption. The primary safety analysis compared the proportion of patients who developed 1 or more serious adverse events across groups. All analyses were prespecified and are listed in section 15 of the COVID-19 Corticosteroid Domain SAP (pp 391-431) in Supplement 1. Data management and summaries were created using R version 3.5.2, and the primary analysis was computed in R version 4.0.0 using the rstan package version 2.19.3 (R Foundation). Additional data

7.2. *Probability of Optimal Intervention*

While $O_{g_k}(r)$ tracks the posterior probability that a regimen is optimal, we also track the probability that an individual intervention is in the optimal regimen. We refer to the posterior probability an intervention *j*, from domain *d*, is in the optimal regimen for group g_k as $\Lambda_{g_k}(d_j)$:

$$\Lambda_{g_k}(d_j) = \frac{1}{M}\sum_{m=1}^{M} I\left[d_j \in r | \pi_{r,g_k} < \pi_{q,g_k} \text{ for all } q \neq r\right].$$

7.3. *Probability of Superiority/Harm Compared to Another Intervention*

For domains that include a standard-of-care arm, it may be of interest to compare the relative effectiveness of an intervention to the standard of care to evaluate effectiveness or harm. We refer to the posterior probability an intervention *j*, from domain *d*, is superior to intervention *i* in group g_k as:

$$\Gamma_{g_k}(d_j, d_i) = \frac{1}{M}\sum_{m=1}^{M} I[\theta_{d_j,g_k} > \theta_{d_i,g_k}]$$

where $\beta_{x,y}$ refers to the effect of intervention *x* in group *y*. We refer to the posterior probability an intervention *j*, from domain *d*, is harmful compared to intervention *i* in group g_k as:

$$\Gamma_{g_k}(d_j, d_i) = \frac{1}{M}\sum_{m=1}^{M} I[\theta_{d_j,g_k} < \theta_{d_i,g_k}]$$

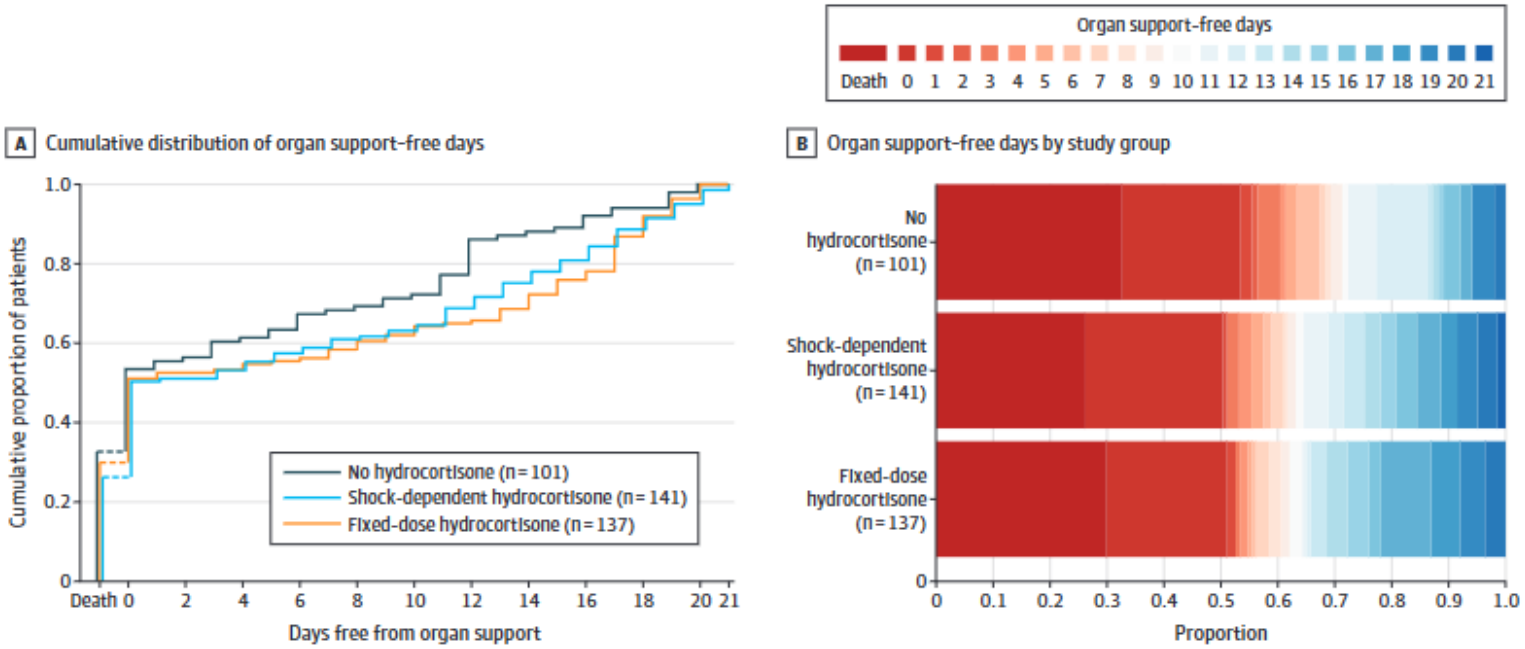
The posterior probability of harm between two interventions is equal to 1 minus the posterior probability of superiority.

7.1. *Probability of Futility/Equivalence/Non-Inferiority Compared to Another Intervention*

In addition to looking at the probability of superiority/harm above which calculate the probability that the difference between two interventions' effects is above/below zero, the posterior probability that the difference between effects falls into specific regions of interest may be computed to evaluate futility, non-inferiority, or equivalence of two interventions. We refer to the posterior probability an intervention *j*, from domain *d*, is **equivalent** to intervention *i* in group g_k as:

メインの結果

Figure 2. Organ Support–Free Days



A, Distributions of organ support–free days (see the Methods section for definition) by study group as the cumulative proportion (y-axis) for each study group by day (x-axis), with death listed first. Curves that rise more slowly are more favorable. B, Organ support–free days as horizontally stacked proportions by study group. Red represents worse values and blue represents better values. The median adjusted odds ratios from the primary analysis, using a bayesian cumulative logistic model, were 1.43 (95% credible interval, 0.91-2.27) and 1.22 (95% credible interval, 0.76-1.94) for the fixed-dose and shock-dependent hydrocortisone groups compared with the no hydrocortisone group, yielding 93% and 80% probabilities of superiority over the no hydrocortisone group, respectively.

Table 2. Primary Outcome

Outcome/analysis[a]	Fixed-dose hydrocortisone (n = 137)	Shock-dependent hydrocortisone (n = 141)	No hydrocortisone (n = 101)
Primary outcome, organ support–free days			
Median (IQR)	0 (-1 to 15)	0 (-1 to 13)	0 (-1 to 11)
Subcomponents of organ support–free days			
In-hospital deaths, No. (%)	41 (30)	37 (26)	33 (33)
Organ support–free days among survivors, median (IQR)	11.5 (0 to 17)	9.5 (0 to 16)	6 (0 to 12)
Primary analysis of the primary outcome, using covariate data from all severe-state participants with COVID-19 (n = 576)[b]			
Adjusted odds ratio			
Mean (SD)	1.47 (0.35)	1.26 (0.31)	1 [Reference]
Median (95% CrI)	1.43 (0.91 to 2.27)	1.22 (0.76 to 1.94)	1 [Reference]
Probability of superiority to no hydrocortisone, %	93	80	
Secondary analysis of the primary outcome, restricted to corticosteroid domain participants (n = 379) with no adjustment for intervention assignment in other domains[c]			
Adjusted odds ratio			
Mean (SD)	1.49 (0.35)	8 (0.30)	1 [Reference]
Median (95% CrI)	1.45 (0.93 to 2.30)	4 (0.80 to 1.95)	1 [Reference]
Probability of superiority to no hydrocortisone, %	95		

Abbreviations: COVID-19, coronavirus disease 2019; IQR, interquartile range; CrI, credible interval.

[a] Definitions of organ support–free days and other outcomes are provided in the Methods section and the study protocol (Supplement 1). Models are structured such that a higher odds ratio is favorable. Other sensitivity analyses are described in the Results section and provided in eTables 1 and 2 and eAppendices 3 and 4 in Supplement 2.

[b] The primary analysis used data from all participants enrolled in the trial who

severe state criteria and were randomized within at least 1
76), adjusting for age, sex, time period, site, region, domain and
ligibility, and intervention assignment (see COVID-19
Domain statistical analysis plan in Supplement 1 and full report
stical analysis committee in eAppendix 3 in Supplement 2).

sis was restricted to participants enrolled in the
main (n = 379) and did not include information on
interventions other than hydrocortisone.

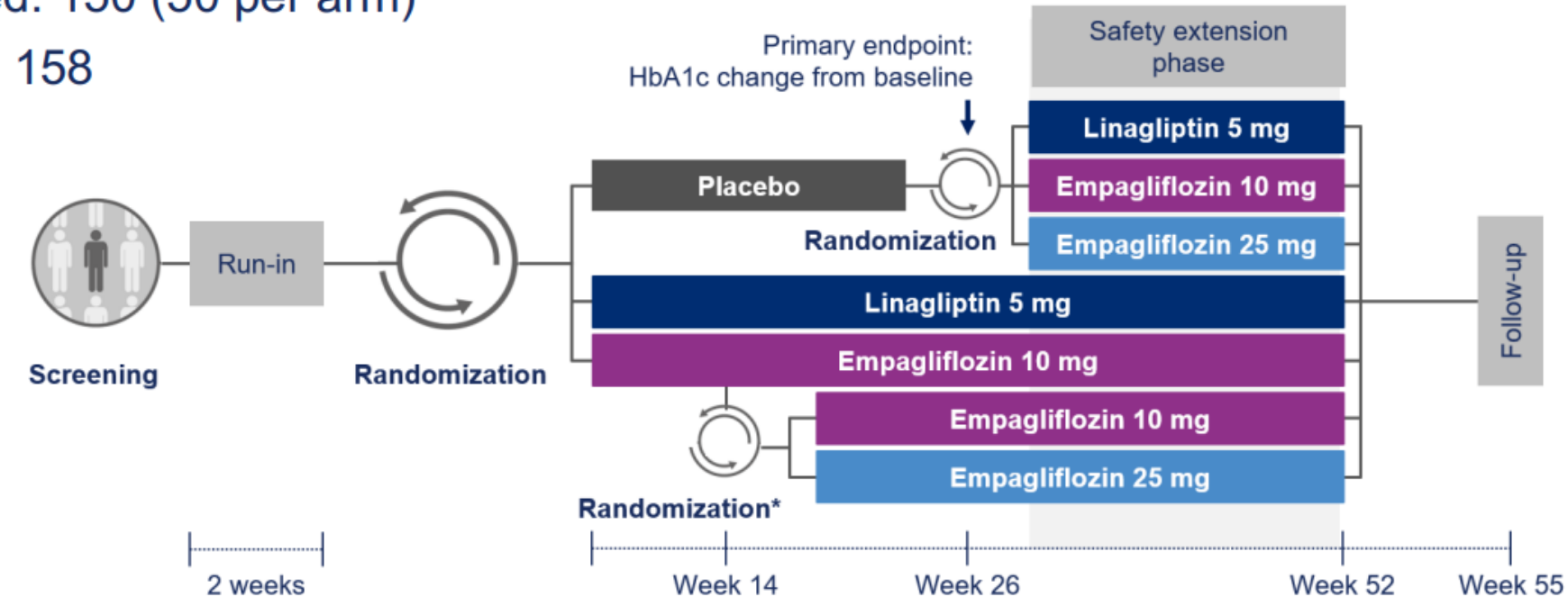
メディアン調整事後オッズ比
1.43（95%信用区間 0.91-2.27）
Pr(Benefit) = 93%

the trial was stopped early and no treatment strategy met prespecified criteria for statistical superiority, precluding definitive conclusions.

事例3：2型糖尿病 小児集団への外挿

DINAMO study design

- N planned: 150 (50 per arm)
- N actual: 158



* Re-randomization at week 14 for participants not achieving HbA1c <7% at week 12

Laffel LM, et al. Lancet Diabetes Endocrinol. 2023;11(3):169-181.

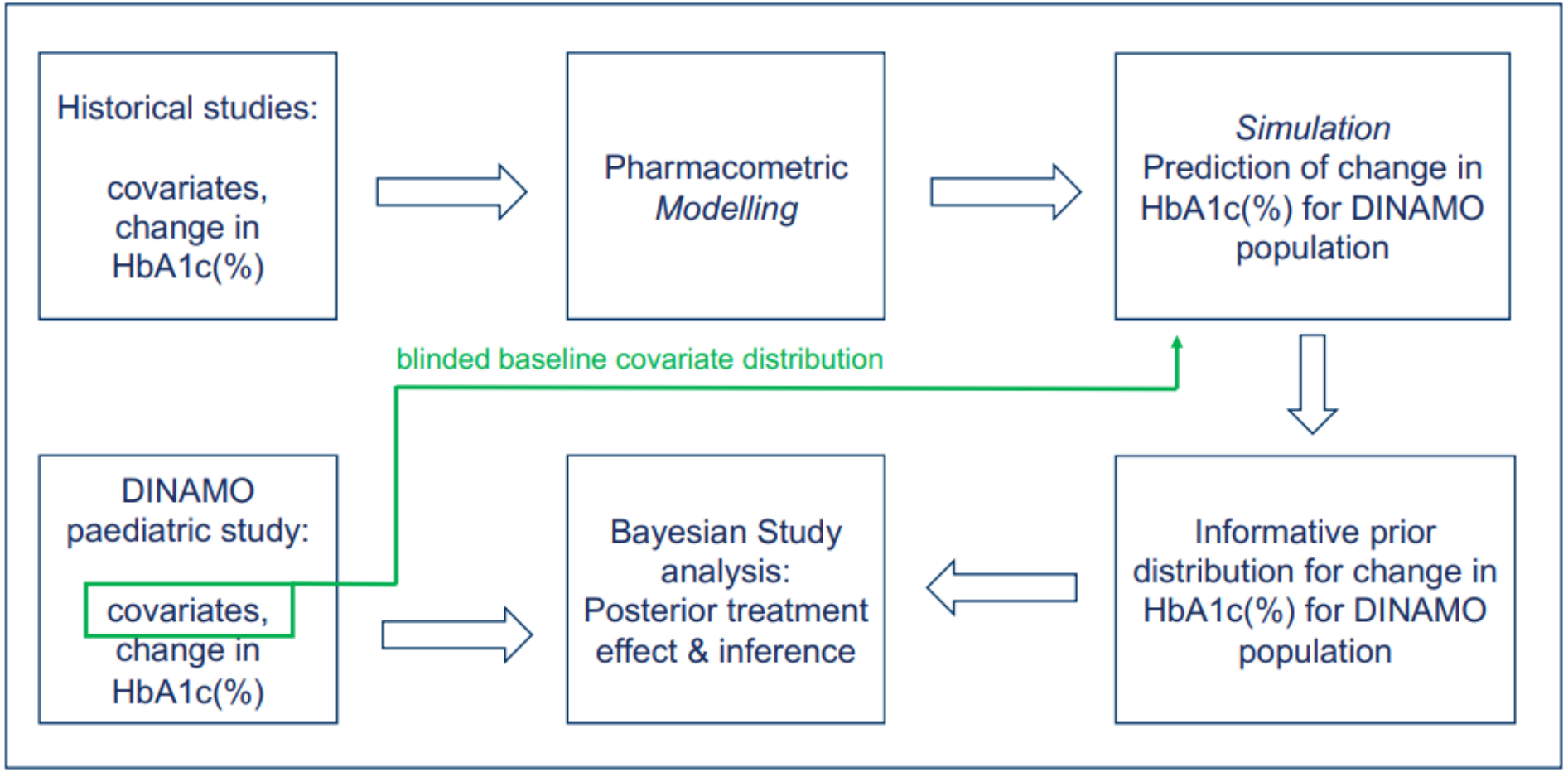
予定されていた解析

- 主要評価項目：Change in HbA1c from baseline to week 26
  - Pooled empagliflozin vs placebo
  - linagliptin vs placebo
- 解析モデル：ANCOVA model with baseline HbA1c as a continuous covariate, and with categorical covariates for treatment and age group
- 症例数設計
  - 両側有意水準5%で検出力85%

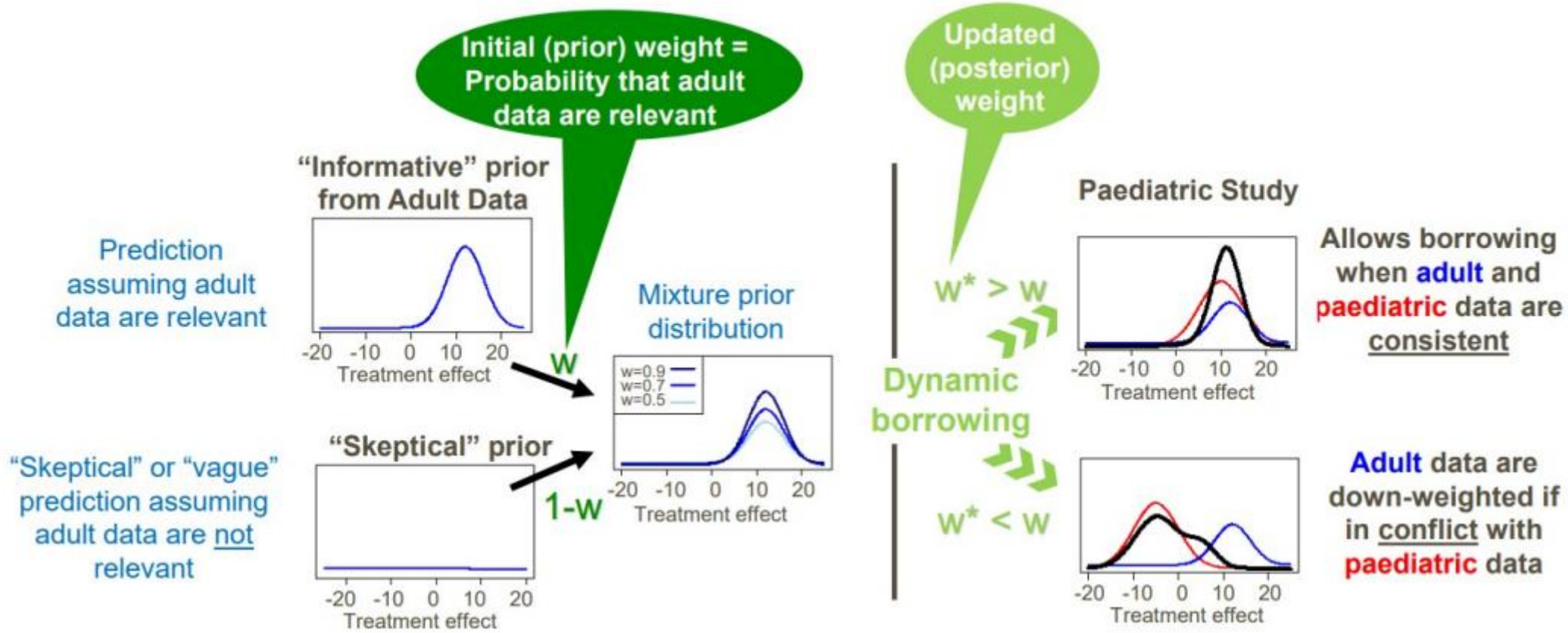
補足的解析にベイズ的アプローチを利用

- 登録完了後、初期の盲検化データにおいて標準偏差が高いことが観察
  - 検出力（パワー）の低下の可能性に対処する必要性
- 被験者登録の再開は最適な選択肢とは見なされなかった
  - 運用上の実現可能性の問題
  - サンプルサイズの大幅な増加が必要
  - 試験結果の取得（read-out）の大幅な遅延
- 成人データからの部分的な外挿にチャレンジ
  - 補足的な解析として事前にFDAと協議

全体像


Historical studies:
covariates, change in HbA1c(%)
Pharmacometric *Modelling*
Simulation Prediction of change in HbA1c(%) for DINAMO population
blinded baseline covariate distribution
DINAMO paediatric study:
covariates, change in HbA1c(%)
Bayesian Study analysis: Posterior treatment effect & inference
Informative prior distribution for change in HbA1c(%) for DINAMO population

事前分布


Initial (prior) weight = Probability that adult data are relevant
Updated (posterior) weight
"Informative" prior from Adult Data
Prediction assuming adult data are relevant
Mixture prior distribution
w=0.9
w=0.7
w=0.5
w
1-w
"Skeptical" prior
"Skeptical" or "vague" prediction assuming adult data are not relevant
-20 -10 0 10 20
Treatment effect
Dynamic borrowing
w* > w
w* < w
Paediatric Study
Allows borrowing when adult and paediatric data are consistent
Adult data are down-weighted if in conflict with paediatric data

事前分布の情報（有効サンプルサイズ）

- 混合事前分布

$$p_I(\theta_I) = w_I Norm(\mu_I, v_I^*) + (1 - w_I) Norm(\mu_I, \sigma_I^2)$$

情報がある分布　　情報がない分布

混合させるときの重み

- 専門家による評価
- FDAとの議論
  - w_I=65%　➡ 有効サンプルサイズ＝51人
  - empagliflozin、placeboの登録105人の半分程度の事前情報

頻度論的解析

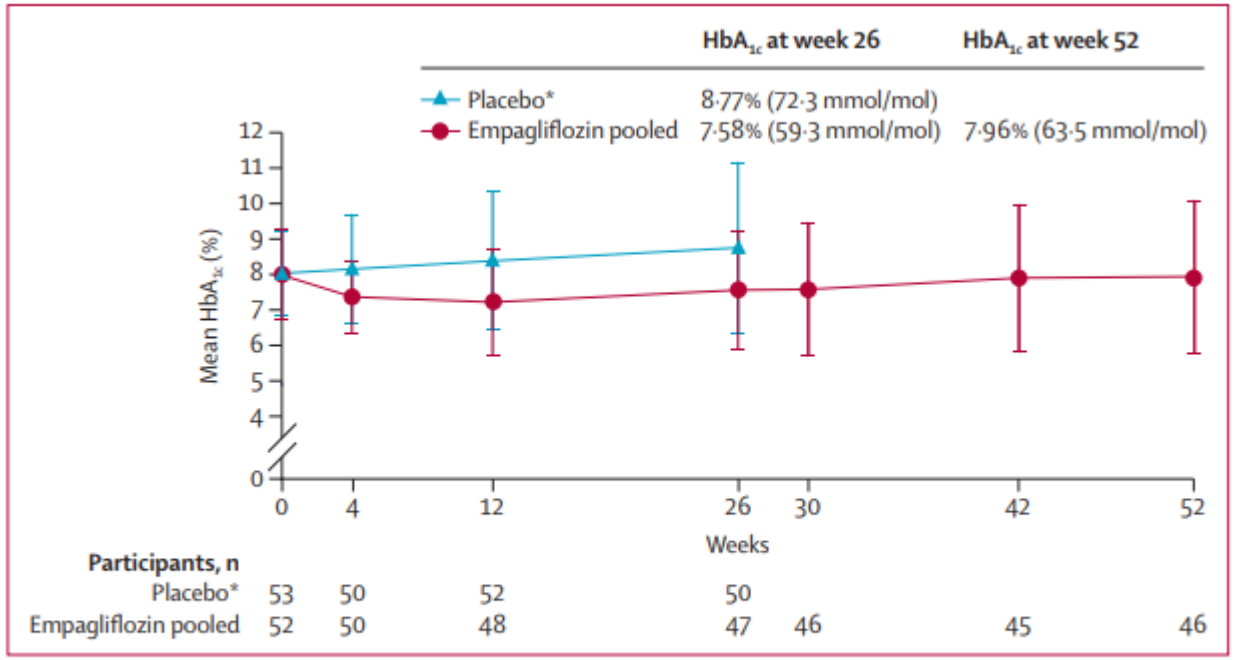


Figure 2: **Change in HbA_{1c} from baseline to week 26**
Descriptive data reflecting mean HbA_{1c} over time from baseline to week 52 for empagliflozin versus placebo in the modified intention-to-treat population. Error bars denote SDs. *Placebo treatment stopped at week 26.

	Participants analysed, n	Baseline, mean (SD)	Change from baseline, adjusted mean (95% CI)	Comparison *vs* placebo, adjusted mean (95% CI)
Primary hypotheses (empagliflozin pooled *vs* placebo and linagliptin *vs* placebo)				
Placebo	53	8·05 (1·23)	0·68 (0·23 to 1·13)	..
Empagliflozin pooled	52	8·00 (1·29)	−0·17 (−0·64 to 0·31)	−0·84 (−1·50 to −0·19)*
Linagliptin 5 mg	52	8·05 (1·11)	0·33 (−0·13 to 0·79)	−0·34 (−0·99 to 0·30)†
Secondary hypothesis (empagliflozin responders on 10 mg plus empagliflozin non-responders randomly reassigned to empagliflozin 25 mg *vs* placebo)				
Placebo	53	8·05 (1·23)	0·66 (0·12 to 1·21)	..
Empagliflozin 10/25 mg	41	7·80 (1·26)	0·14 (−0·42 to 0·71)	−0·52 (−1·31 to 0·27)‡
Secondary hypothesis (empagliflozin responders on 10 mg plus empagliflozin non-responders randomly reassigned to empagliflozin 10 mg *vs* placebo)				
Placebo	53	8·05 (1·23)	0·68 (0·19 to 1·17)	..
Empagliflozin 10/10 mg	39	7·92 (1·36)	−0·49 (−1·03 to 0·04)	−1·18 (−1·90 to −0·45)§

Primary endpoint was HbA_{1c} percentage change from baseline at week 26 in the modified intention-to-treat set. Analyses included all available HbA_{1c} data on treatment, after start of rescue medication, and after premature treatment discontinuation. Missing data were multiply imputed with wash-out approach, using ANCOVA with baseline HbA_{1c} as a linear covariate, treatment and age as categorical covariates, and applying Rubin's rules to combine multiple imputations. Mean change from baseline was adjusted for baseline HbA_{1c} and age category for all hypotheses and additionally weighted for secondary hypotheses. Baseline means were not weighted. To convert the values for HbA1c percentage to mmol/mol, subtract 2·15 and multiply the result by 10·929. *p=0·012. †p=0·29. ‡p=0·19. §p=0·0015.

Table 2: **Primary outcomes**

Empa -0.84 (95%信頼区間 -1.50- -0.19)
Lina -0.34 (95%信頼区間 -0.99- 0.30)

Laffel LM, et al. Lancet Diabetes Endocrinol. 2023;11(3):169-181.

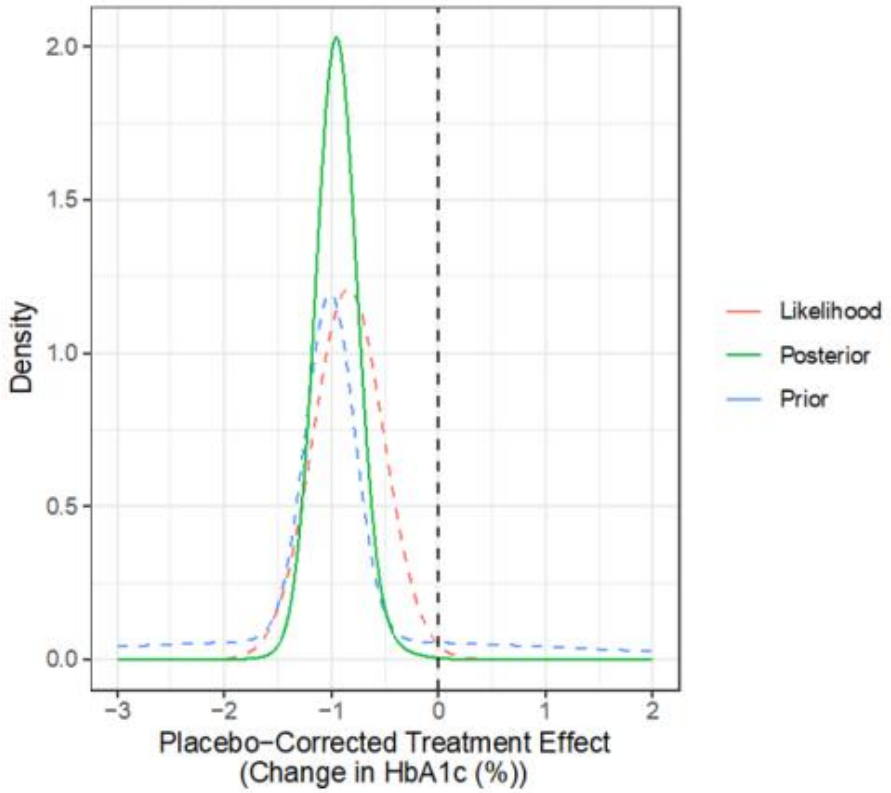
ベイズ流の解析結果 (Empagliflozin)

- DINAMO試験の結果は、情報のある事前分布とかなり近い結果
  - 情報のない事前分布の重みがほぼ0

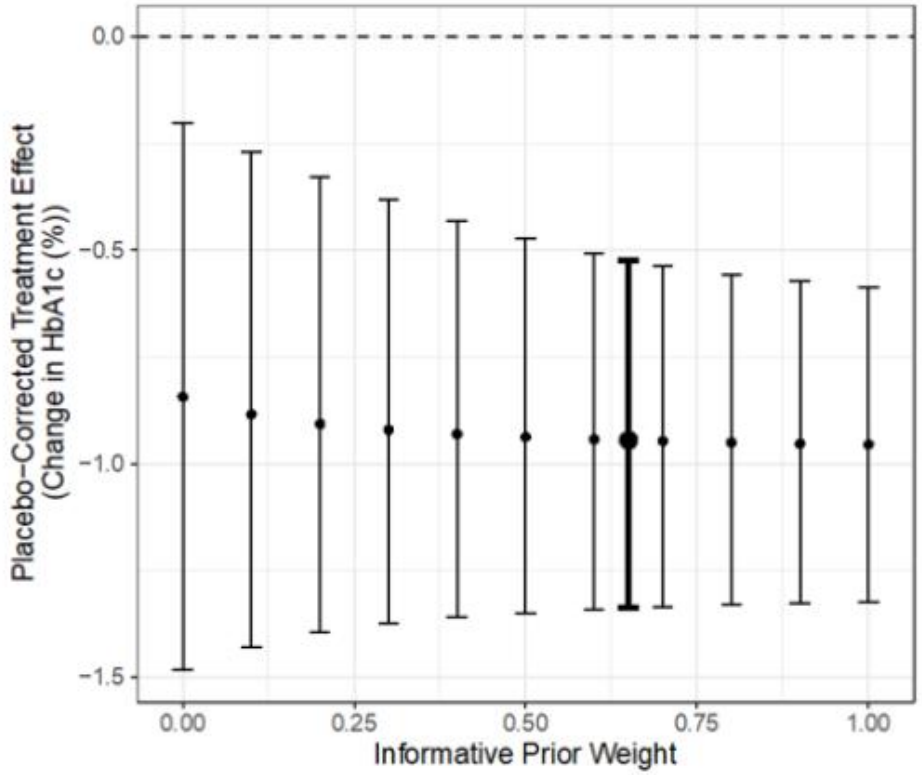
	Mean	SD	P2.5%	P5%	Median	P95%	P97.5%	Prob. superiority
Prior (exposure-response based)	-1.01	1.37	-4.37	-3.46	-1.01	1.43	2.34	0.885
Likelihood (DINAMO data)+	-0.84	0.33	-1.50	-	-	-	-0.19	-
Posterior distribution	-0.945	0.207	-1.34	-1.27	-0.949	-0.605	-0.524	>0.999

- Pr(Superior) >0.999
- 群間差 -0.945, 95% credible interval (-1.34, -0.524)

事前分布の影響を評価：Tipping point analysis



Assessment of prior-data conflict
Prior/Posterior ESS_{ELIR}: 55/138



Sensitivity tipping point analysis

ベイズ流の解析結果 (Linagliptin)

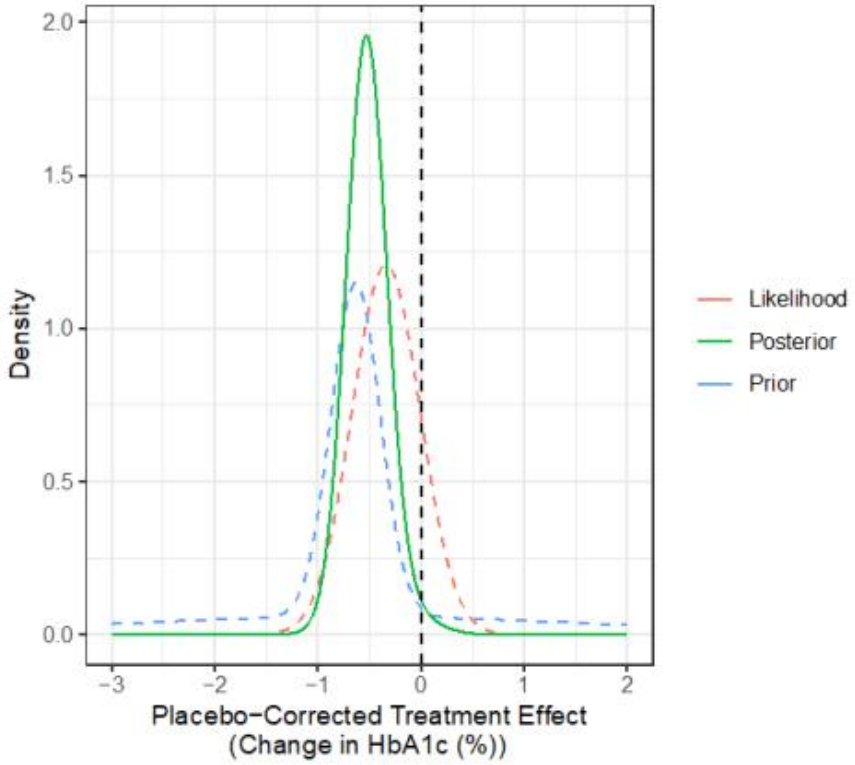
- こちらも情報のある事前分布とかなり近い結果

	Mean	SD	P2.5%	P5%	Median	P95%	P97.5%	Prob. superiority
Prior (exposure-response based)	-0.635	1.42	-4.12	-3.18	-0.635	1.91	2.85	0.859
Likelihood (DINAMO data)*	-0.34	0.33	-0.99	-	-	-	0.30	-
Posterior distribution	-0.514	0.219	-0.919	-0.854	-0.523	-0.151	-0.052	0.982

* From DINAMO primary analysis, adjusted mean, SE and 95% confidence interval (p=0.2935)

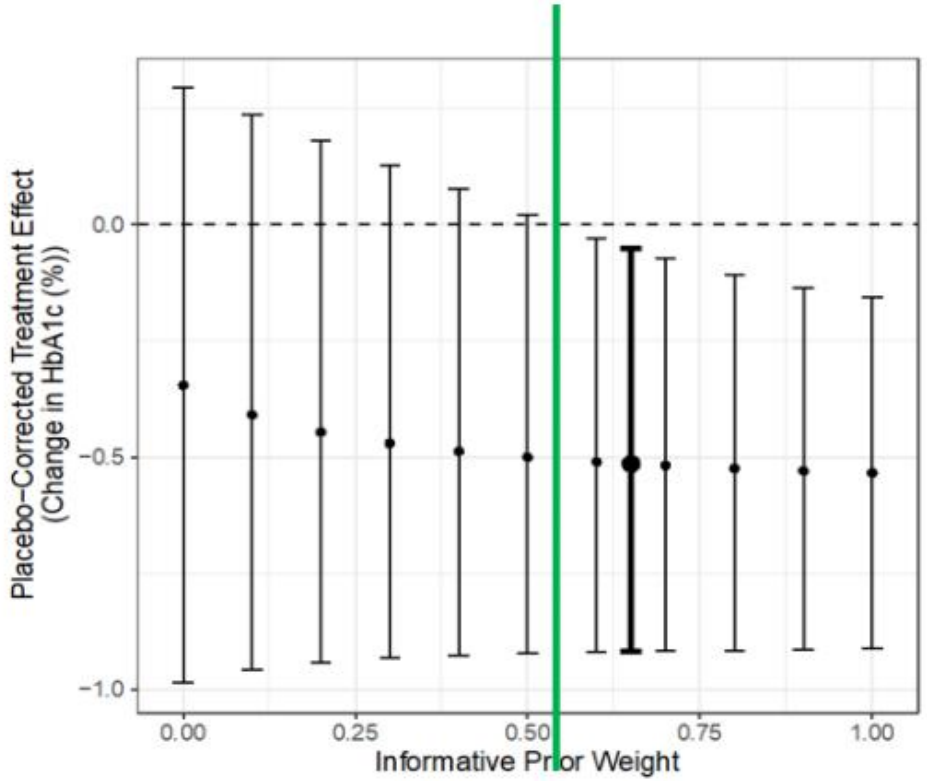
- Pr(Superior) ＝0.982
- 群間差 -0.514, 95% credible interval (-0.919, -0.052)

事前分布の影響を評価：Tipping point analysis



Assessment of prior-data conflict

Prior/Posterior ESS_{ELIR}: 51/128



Sensitivity tipping point analysis

Tipping point w=0.542

FDA comments

- 小児と成人の2型糖尿病集団間で疾患進行に違いがあるものの、小児の病態生理は成人におけるそれと類似しており、よってそれらの情報は関連があり、借用は正当化されると結論
- Based on discussion and feedback obtained from the clinical team, it appears reasonable to assume at least 54% weight on the relevance of the adult information to the pediatric population and we can therefore conclude that there is at least 97.5% posterior probability that it has a positive treatment effect in pediatric subjects

ベイズ流アプローチを利用する際の論点

- ベイズ流アプローチを利用する理由
- 事前情報の選択
- 治療効果に関する閾値と達成すべき事後確率の設定
- 症例数設計と動作特性評価

平川晃弘、他．希少疾患領域の治験におけるベイズ流アプローチの利用可能性と留意事項．
薬理と治療.2021;49(suppl1):s72-s79.

話は戻って…

NEWS

Check for updates

The BMJ

Cite this as: *BMJ* 2026;392:s180
http://doi.org/10.1136/bmj.s180
Published: 28 January 2026

FDA is now "open to bayesian statistics": transformational change or new Pandora's box?

Peter Doshi

The US Food and Drug Administration (FDA) is now "open to bayesian statistics," contrasting this with the frequentist approach that the agency and the drug industry have historically relied on for statistical analysis.

In a video posted to X on 12 January the FDA commissioner, Marty Makary, said that the agency had published a new guidance document "to encourage the use of bayesian statistics in clinical trial design and the readout of results" in new drug and biologic applications.[1]

At a glance: What are bayesian statistics?

A bayesian approach to statistical analysis combines collected study data with external sources of information—such as pharmacokinetic or pharmacodynamic data, other clinical trials, observational data, or expert opinion—to determine an outcome. It differs from how frequentist statistical approaches are commonly applied, in which only study data are assessed. It is named after the 18th century mathematician and theologian Thomas Bayes.

The perceived problem

ustry sponsored clinical drug trials
sed using a set of statistical
ified as "frequentist." One prominent
ted with frequentist methods is null
ficance testing, especially at the 0.05
-off for the corresponding P value.

ns have aided "go/no-go"-type
on whether a result is—or is
tatistically significant" (meaning
decades statisticians and other
ecried the tyranny of "statistical
d P values, criticising them for
staken decisions regarding efficacy

were often incorrectly interpreted as
al or practical significance. And
ues greater than 0.05 have been used
nclude "no effect" when one exists.

cism, significance testing has
lamental part of drug regulation. "It
am becoming the dominant approach
loth in applying it," Greenland said,
ies such as significance testing
, oversimplified, [and] automatic."

ques, by contrast, seemed to offer an
native. Whereas, by frequentist
ical trials are analysed in isolation,
alysis the study data are combined
es of information.

hesising information across
draw new probability statements
sis, such as whether a drug is
ls to many. "It means that analyses
t of all we know, not just the data on
tion," says Lilford—something that
n help bring new treatments for rare
diseases to market.

Industry figures are also likely to welcome the news. In 2023, influential voices in the biopharmaceutical space writing in *Nature Reviews Drug Discovery* called on regulators to go bayesian.[5] They said, "We believe

in a frequentist mode. But he warned, "By calling it bayesian, you now mystify it and you open a door for abuse."

Sander Greenland

Greenland in 2018

Born January 16, 1951 (age 75)

- 従来の頻度論的アプローチは確かに問題がある
- しかし、ベイズ流アプローチも新たな不正の温床を開く可能性がある （Sander Greenland）

- 事前分布（prior）と呼ばれる外部情報の信頼性と適切性が極めて重要
  - 「弱情報的」とされた事前分布が、実際には「大規模試験に匹敵する影響」を与えることもある
- ベイズ流アプローチの前に頻度主義的結果を提示し、事前分布が結果に与える影響を可視化すべき
  - Tipping point analysisのような影響度分析

頻度論に基づいた結果（共変量は未調整）

- 比例オッズモデルのあてはめ
- オッズ比 1.44
- 95%信頼区間 0.91-2.25

Table 2. Primary Outcome

Outcome/analysis[a]	Fixed-dose hydrocortisone (n = 137)	Shock-dependent hydrocortisone (n = 141)	No hydrocortisone (n = 101)
Primary outcome, organ support–free days			
Median (IQR)	0 (-1 to 15)	0 (-1 to 13)	0 (-1 to 11)
Subcomponents of organ support–free days			
In-hospital deaths, No. (%)	41 (30)	37 (26)	33 (33)
Organ support–free days among survivors, median (IQR)	11.5 (0 to 17)	9.5 (0 to 16)	6 (0 to 12)
Primary analysis of the primary outcome, using covariate data from all severe-state participants with COVID-19 (n = 576)[b]			
Adjusted odds ratio			
Mean (SD)	1.47 (0.35)	1.26 (0.31)	1 [Reference]
Median (95% CrI)	1.43 (0.91 to 2.27)	1.22 (0.76 to 1.94)	1 [Reference]
Probability of superiority to no hydrocortisone, %	93	80	
Secondary analysis of the primary outcome, restricted to corticosteroid domain participants (n = 379) with no adjustment for intervention assignment in other domains[c]			
Adjusted odds ratio			
Mean (SD)	1.49 (0.35)	8 (0.30)	1 [Reference]
Median (95% CrI)	1.45 (0.93 to 2.30)	4 (0.80 to 1.95)	1 [Reference]
Probability of superiority to no hydrocortisone, %	95		

Abbreviations: COVID-19, coronavirus disease 2019; IQR, interquartile range; CrI, credible interval.

[a] Definitions of organ support–free days and other outcomes are provided in the Methods section and the study protocol (Supplement 1). Models are structured such that a higher odds ratio is favorable. Other sensitivity analyses are described in the Results section and provided in eTables 1 and 2 and eAppendices 3 and 4 in Supplement 2.

[b] The primary analysis used data from all participants enrolled in the trial who severe state criteria and were randomized within at least 1 76), adjusting for age, sex, time period, site, region, domain and ligibility, and intervention assignment (see COVID-19 Domain statistical analysis plan in Supplement 1 and full report stical analysis committee in eAppendix 3 in Supplement 2).

sis was restricted to participants enrolled in the main (n = 379) and did not include information on interventions other than hydrocortisone.

ほとんど同じ…

メディアン調整事後オッズ比
1.43（95%信用区間 0.91-2.27）
Pr(Benefit) = 93%

じゃあ、解釈だけベイズ的にやったら良いのでは？

- 信頼分布 Confidence Distribution
  - 頻度論的推論の枠組みの中で、知識の不確実性と、偶然の確実性の両方が併存することを可能にする方法

4. Interval estimation. Much controversy has centred on the distinction between fiducial and confidence estimation. Here follow five remarks, not about the mathematics, but about the general aims of the two methods.

(i) The fiducial approach leads to a distribution for the unknown parameter, whereas the method of confidence intervals, as usually formulated, gives only one interval at some preselected level of probability. This seems at first sight a distinct point in favour of the fiducial method. For when we write down the confidence interval $(\bar{x} - 1.96\,\sigma/\sqrt{n}, \bar{x} + 1.96\,\sigma/\sqrt{n})$ for a completely unknown normal mean, there is certainly a sense in which the unknown mean θ is likely to lie near the centre of the interval, and rather unlikely to lie near the ends and in which, in this case, even if θ does lie outside the interval, it is probably not far outside. The usual theory of confidence intervals gives no direct expression of these facts.

Yet this seems to a large extent a matter of presentation; in the common simple cases, where the upper α limit for θ is monotone in α, there seems no reason why we should not work with confidence distributions for the unknown parameter. These can either be defined directly, or can be introduced in terms of the set of all confidence intervals at different levels of probability. Statements made on the basis of this distribution, provided we are careful about their form, have a direct frequency interpretation. In applications it will often be enough to specify the confidence distribution, by for example a pair of intervals, and this corresponds to the common practice of quoting say both the 95 per cent and the 99 per cent confidence intervals.

Confidence interval as probability statements

…when we write down the confidence interval for a completely unknown normal mean, there is certainly a sense in which the unknown mean is likely to lie near the centre of the interval, and rather unlikely to lie near the ends and in which, in this case, even if does lie outside-the interval, it is probably not far outside. The usual theory of confidence intervals gives no direct expression of these facts.

Proposal of confidence distribution

…there seems no reason why we should not work with confidence distributions for the unknown parameter. These can either be defined directly, or can be introduced in terms of the set of all confidence intervals at different levels of probability.

Cox DR. Ann Math Stat. 1958;29(2):357-372.

信頼分布

Received: 17 January 2023 | Revised: 24 August 2023 | Accepted: 22 December 2023
DOI: 10.1002/sim.10000

TUTORIAL IN BIOSTATISTICS

Statistics in Medicine WILEY

Confidence distributions for treatment effects in clinical trials: Posteriors without priors

Ian C. Marschner

NHMRC Clinical Trials Centre, The University of Sydney, Camperdown, Australia

Correspondence
Ian C. Marschner, NHMRC Clinical Trials Centre, The University of Sydney, Locked Bag 77, Camperdown, NSW 1450, Australia.
ian.marschner@sydney.edu.au

Funding information
National Health and Medical Research Council, Grant/Award Numbers: 1150467, 1171422

An attractive feature of using a Bayesian analysis for a clinical trial is that knowledge and uncertainty about the treatment effect is summarized in a posterior probability distribution. Researchers often find probability statements about treatment effects highly intuitive and the fact that this is not accommodated in frequentist inference is a disadvantage. At the same time, the requirement to specify a prior distribution in order to obtain a posterior distribution is sometimes an artificial process that may introduce subjectivity or complexity into the analysis. This paper considers a compromise involving confidence distributions, which are probability distributions that summarize uncertainty about the treatment effect without the need for a prior distribution and in a way that is fully compatible with frequentist inference. The concept of a confidence distribution provides a posterior–like probability distribution that is distinct from, but exists in tandem with, the relative frequency interpretation of probability used in frequentist inference. Although they have been discussed for decades, confidence distributions are not well known among clinical trial statisticians and the goal of this paper is to discuss their use in analyzing treatment effects from randomized trials. As well as providing an introduction to confidence distributions, some illustrative examples relevant to clinical trials are presented, along with various case studies based on real clinical trials. It is recommended that trial statisticians consider presenting confidence distributions for treatment effects when reporting analyses of clinical trials.

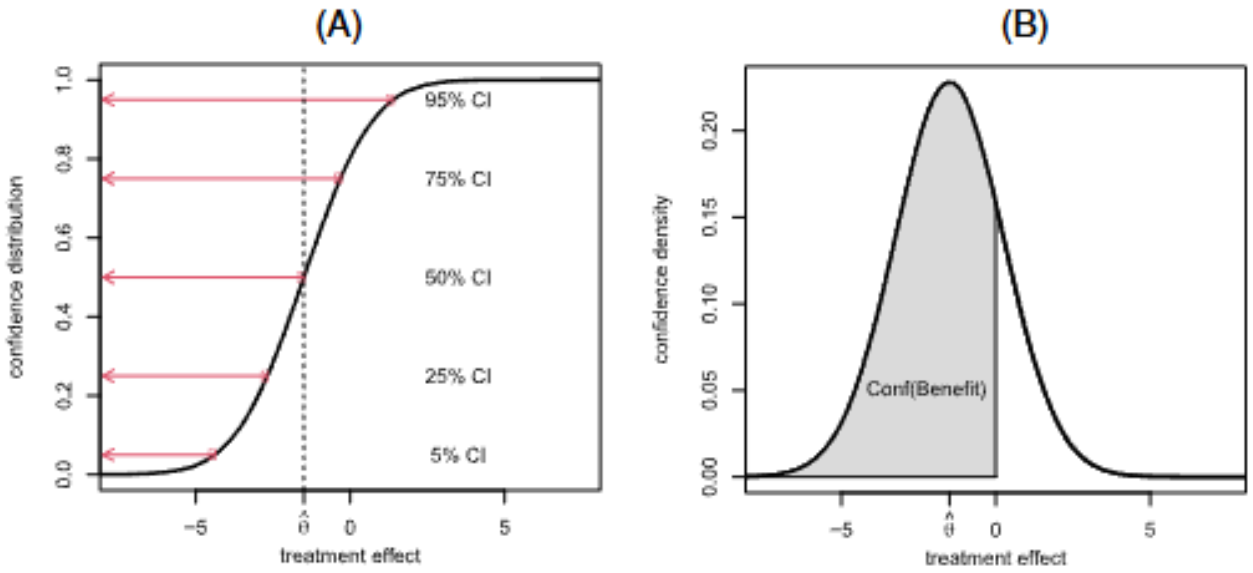
KEYWORDS
clinical trial, confidence distribution, posterior distribution, probability, treatment effect

1 | INTRODUCTION

Bayesian inference has long been advocated for clinical trials.[1] It allows the incorporation of prior information such as historical controls or evidence from related populations,[2] which is particularly applicable for early phase studies or confirmatory studies in rare diseases and small populations. The wider use of adaptive designs has also led to increased use of Bayesian inference in clinical trials.[3,4]

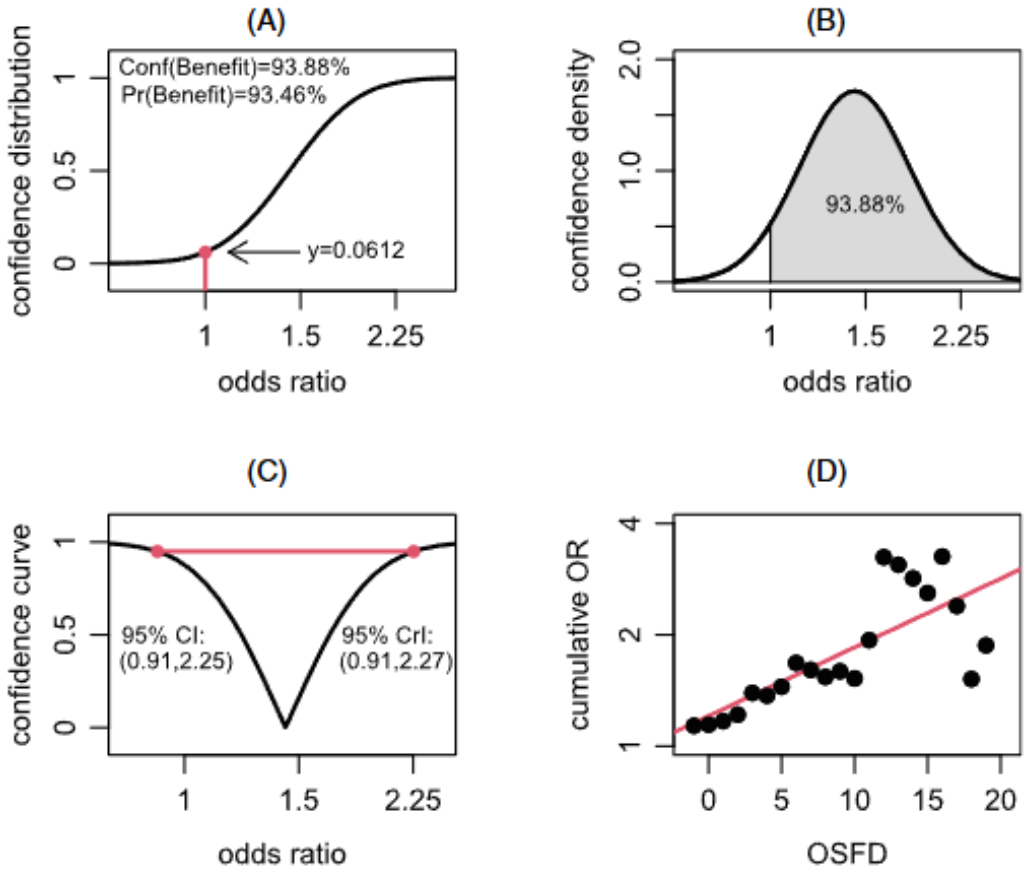
One of the most appealing features of a Bayesian analysis is that uncertainty about the treatment effect can be expressed using a probability distribution – the posterior distribution. This allows probability statements to be made about the treatment effect. Thus, for example, a Bayesian analysis of a recent clinical trial was able to summarize the

- 治療効果に関する不確実性を要約する分布
  - 事前分布は不要
  - 治療効果の事後分布的解釈が可能
- 治療効果に対する片側(1-α)%信頼限界の集まり
  - 信頼分布関数、信頼密度関数として表現



Marschner IC. Stat Med. 2024;43(6):1271-1289.

REMAP CAPの再評価



オッズ比が1より大きい
信頼確率＝93.88%

FIGURE 4 Confidence distribution for the treatment effect in the REMAP-CAP trial data from Table 3. Panel A provides the confidence distribution function with Conf(BENEFIT) being the confidence that fixed duration hydrocortisone is beneficial relative to no hydrocortisone and Pr(BENEFIT) being the corresponding Bayesian posterior probability of benefit. Panel B displays the confidence density and Panel C displays the confidence curve, together with the 95% confidence interval (CI) and Bayesian credible interval (CrI). Panel D displays the crude cumulative odds ratios. Odds ratios (OR) are plotted on the log scale.

まとめ

- ベイズを用いた臨床試験は既に多く実施されている
  - FDAもガイダンスを出して、いよいよベイズの時代か？
- ベイズは、確率の自然な解釈を可能にし、説得力のある事前の情報がある場合には効率的に試験を進めることが可能になる
- 一方で、事前分布の影響は無視できない他、これまでに積み上げられてきた頻度論に基づく方法とは、異なる留意点が出てくる
  - 多くが納得できる状況での適用と経験の蓄積も必要
- 一足飛びにベイズに行くのは難しいかもしれないが、解釈だけベイズ的に行うということは良いかもしれない（信頼分布の利用）