

Significance in ePCTs: *P* Values vs Decision-Maker Perspectives



**NIH PRAGMATIC TRIALS
COLLABORATORY**

Rethinking Clinical Trials®

In this session

Greg Simon

Kaiser Permanente Washington, Health Care Systems Interactions Core

Susan Huang

UC Irvine, ABATE Infection trial

Liz Turner

Duke University, Biostatistics and Study Design Core

“H1-H0, H1-H0” is a song we seldom hear in the real world



Gregory Simon

March 3, 2020

MHRN Blog



Arne Beck and I were recently revising the description of one of our Mental Health Research Network projects. We really tried to use the traditional scientific format, specifying H_1 (our hypothesis) and H_0 (the null hypothesis). But our research just didn't fit into that mold. We eventually gave up and just used a plain-language description of our question: For women at risk for relapse of depression in pregnancy, how do the benefits and costs of a peer coaching program compare to those of coaching from traditional clinicians?

Where did this $p < .05$ thing start?



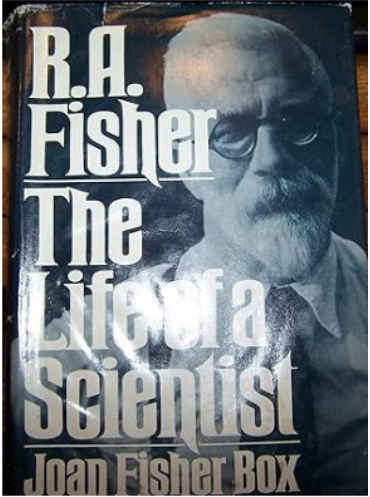
Statistical Methods for Medical Workers

R A Fisher 1925

“It is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not.”



The real origin story



Fisher's colleague, Muriel Bristol, insisted that she could taste the difference when milk was poured into tea vs. tea poured into milk.

Fisher didn't believe it and proposed a test: correctly classifying 8 teacups (4 milk-first, 4 tea-first) presented in random order.

Bristol correctly classified all 8 cups!

Fisher calculated that random guessing could yield that result 1 time out of 70.

And he decided: That's convincing enough for me!



What's wrong with “H1-H0, H1-H0”?

- Why a single threshold for “statistical significance”?
- Real-world decisions are usually multi-dimensional.
- Healthcare decision makers care about subgroups.

BUT – Is this just a slippery slope to data-dredging and post-hoc chicanery?

A SHOW OF CONFIDENCE

MANY medical researchers believe that it would be fruitless to submit for publication any paper that fails to show statistical tests of significance. Their belief is well founded: editors and referees commonly reject papers that lack a show of significance as indicators of meaningless statistics. The word "significant" means to assure confidence in the results of a study. The preoccupation with significance is embodied in the arbitrary decision that a P value is less than 0.05; if the P value is greater than 0.05, the results are "not significant." The question of whether or not the P value is less than 0.05 is a matter of random variation. Dr. Rennie's editorial of

which two groups differ or two variables are associated. Highly "significant" P values can accompany negligible differences (if the study is large), and unimpressive P values can accompany strong associations (if the study is small). P values, therefore, are not good measures of the strength of the relation between study variables. P values serve poorly as descriptive statistics.

In medical research, decisions are rarely made on the basis of a single trial or experiment. Without a need to classify the results from a single study into a "yes" or "no" decision, employing a test of statistical significance to the point of degrading the findings into such a dichotomy is counterproductive and potentially misleading. In a recent paper, Dr. Rennie reviewed the findings of 71 clinical trials that compared treatment with placebo. In 10 trials, the new treatment was significantly better than placebo; in 10 trials, the new treatment was significantly worse than placebo; and in 51 trials, the results were consistent with a null hypothesis of no effect. The misinterpretations of the results of these trials are a direct consequence of the investigators using significance testing as their primary statistical analysis, rather than a more descriptive and informative analysis.

If significance testing is misleading, how should results be presented? By choosing a measure that quantifies the degree of association or effect in the data and then calculating a confidence interval, researchers can summarize the strength of association in their data and allow for random variation in a sim-


"P Values serve poorly as descriptive statistics"

Decolonization in Hospitals

What is it?

The use of topical antiseptic soaps and nasal ointments to reduce the body's surface bacteria during high-risk times for infection

Context

- Decolonization reduces infections from wounds, surgery, devices
- Penalty for hospitals with high bloodstream infection rates
- ICU decolonization trials  bloodstream infection 30-44% → ICU standard-of-care
- Can the benefit extend to general medical and surgical units?
- Devices more common in ICUs vs. general units, but more total devices outside ICUs
- Hospitals were adopting decolonization for patients with devices outside the ICU

Decolonization in General Medical/Surgical Units

- Pragmatic cluster randomized trial of universal decolonization vs routine care

- **ABATE Infection Trial**

- 53 hospitals, 194 non-ICUs, 340,000 adult patients
- Decolonization not effective for all non-ICU patients
 - Reduced combined methicillin resistant *S. aureus* (MRSA) & vancomycin resistant enterococcus (VRE) by 9%, $p=NS$
 - Reduced bloodstream by 6%, $p=NS$
- Highly effective in ***post-hoc analysis*** of those with medical device
 - Reduced MRSA & VRE by 37%, $p<0.05$
 - Reduced bloodstream by 32%, $p<0.05$

THE LANCET


Chlorhexidine versus routine bathing to prevent multidrug-resistant organisms and all-cause bloodstream infections in general medical and surgical units (ABATE Infection trial): a cluster-randomised trial

Huang SS et al. Lancet 2019;393(10177):1205-1215

Susan S Huang, Edward Septimus, Ken Kleinman, Julia Moody, Jason Hickok, Lauren Heim, Adrijana Gombosev, Taliser R Avery, Katherine Haffner-Reffer, Lauren Shimelman, Mary K Hayden, Robert A Weinstein, Caren Spencer-Smith, Rebecca E Kaganov, Michael V Murphy, Tyler Forehand, Julie Lankiewicz, Micaela H Coady, Lena Portillo, Jalpa Sarup-Patel, John A Jernigan, Jonathan B Perl, Richard Platt, for the ABATE Infection trial team

ABATE Infection Trial National Toolkit

“Patients with specific medical devices... who received decolonization had a 30 percent reduction in bloodstream infection.”

**Agency for Healthcare
Research and Quality**

Topics ▾ Programs ▾ Research ▾ Data & Analytics ▾ Tools ▾ Funding & Grants ▾ News ▾ About AHRQ ▾

Home > Healthcare-Associated Infections Program > Decolonization

Healthcare-Associated Infections Program


Combating Antibiotic-Resistant Bacteria

About the CUSP Method

Decolonization

Tools

Toolkit for Decolonization of Non-ICU Patients With Devices



Decolonization is an infection prevention practice that removes germs from the skin. It is often used in healthcare facilities when patients carry methicillin-resistant *Staphylococcus aureus* (MRSA) or other dangerous germs on their bodies. It can keep them from developing an infection themselves or passing the germs on to others.

Decolonization includes an antiseptic bathing routine using a special soap on the skin and applying an antiseptic or antibiotic product in the nose. This practice has been shown to reduce the amount of germs on the body and reduce infections overall, including hard-to-treat MRSA infections.

ABATE Trial (Active Bathing to Eliminate Infection)

This cluster-randomized trial of 53 hospitals investigated the use of decolonization to prevent infection in non-intensive care unit patients. Patients with specific medical devices (central venous catheters, midline catheters, and lumbar drains) who received decolonization had a 30 percent reduction in bloodstream infections.

<https://www.ahrq.gov/hai/universal-icu-decolonization/index.html>

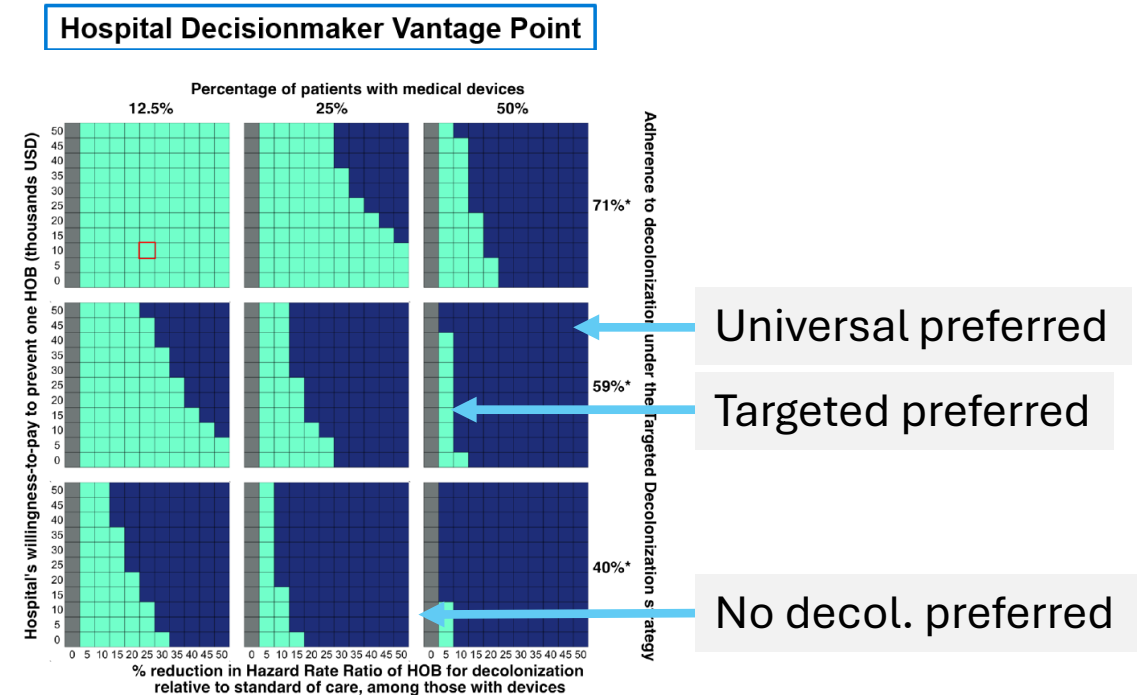
<https://www.ahrq.gov/hai/tools/abate/index.html>

Decision-Making for Decolonization in non-ICUs

- Cost-effectiveness analysis evaluating universal, targeted, or no decolonization for patients with central venous lines and other devices

➤ ABATE Infection Trial

- Depends on
 - ✓ Prevalence of device use
 - ✓ Adherence to targeted decolonization
 - ✓ Financial penalties for bloodstream infection

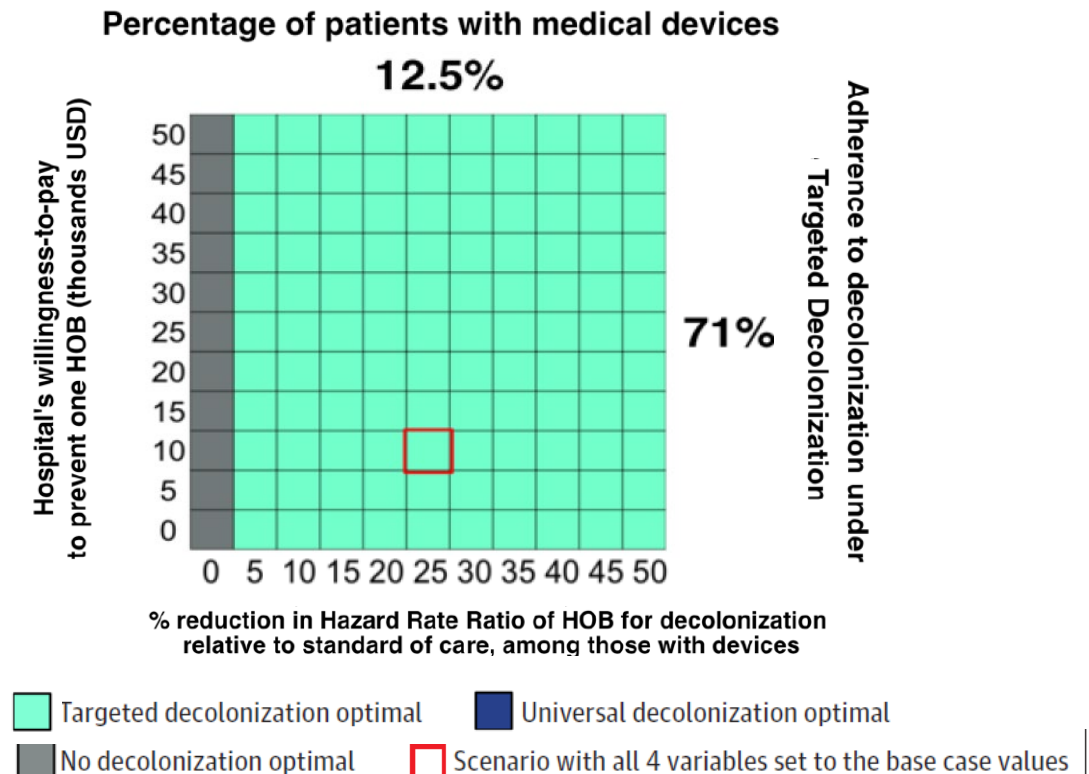


Decision-Making for Decolonization in non-ICUs

- Cost-effectiveness analysis evaluating universal, targeted, or no decolonization for patients with central venous lines and other devices

➤ ABATE Infection Trial

- Depends on
 - ✓ Prevalence of device use
 - ✓ Adherence to targeted decolonization
 - ✓ Financial penalties for bloodstream infection

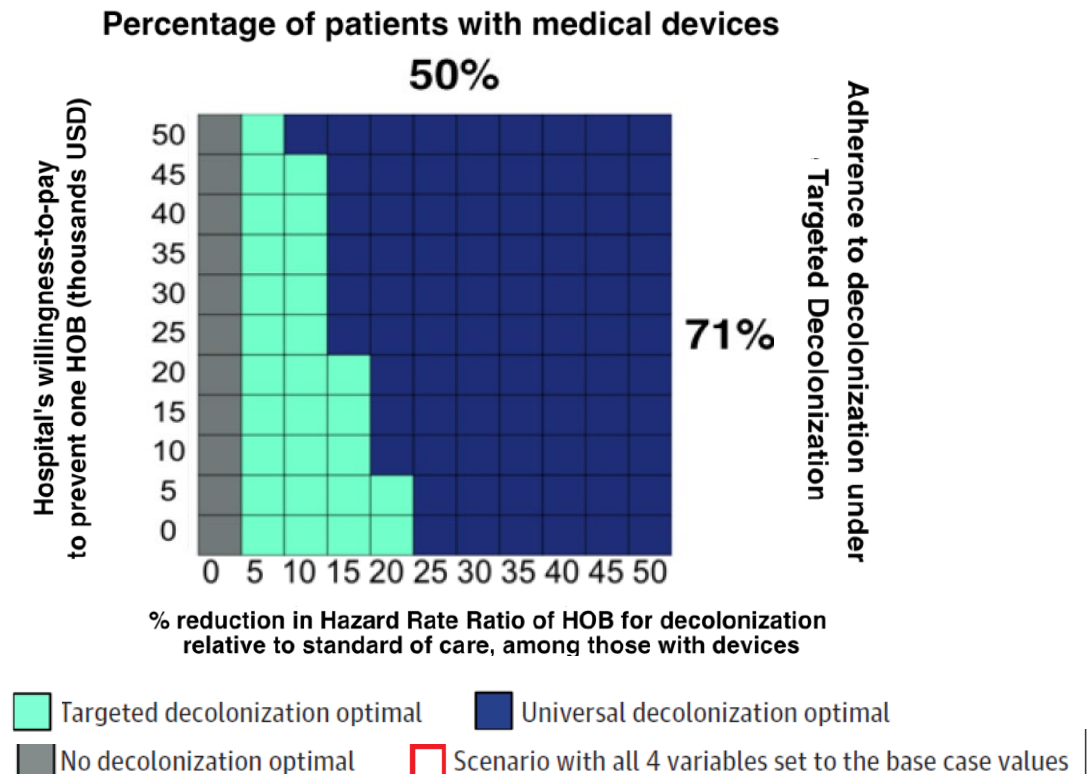


Decision-Making for Decolonization in non-ICUs

- Cost-effectiveness analysis evaluating universal, targeted, or no decolonization for patients with central venous lines and other devices

➤ ABATE Infection Trial

- Depends on
 - ✓ Prevalence of device use
 - ✓ Adherence to targeted decolonization
 - ✓ Financial penalties for bloodstream infection



The Dogma of the P-Value in Driving Care

Academic Rules

- **Goal:** improve, define clinical care
- Guidance requires high certainty
- Scientific rigor demands rules, threshold
- Cost doesn't affect threshold ($P=0.05$)
- Experience in other populations doesn't affect threshold ($P=0.05$)
- Decision-making doesn't include probability of benefit at given p-value. Circumstances don't affect conclusions.

Pragmatic Use in Hospitals

- **Goal:** improve, define clinical care
- Decisions based on little data, short time
- Clinical decisions rarely based on certainty
- Low cost, possible benefit can fly
- Experience in other populations commonly used to infer benefit
- Decision-making needs to include probability of benefit at given p-value. Circumstances may warrant adoption.

Significance in ePCTs: *P* Values vs Decision-Maker Perspectives

Liz Turner, PhD
Duke University; Co-Lead, Biostatistics &
Study Design Core, NIH Collaboratory



**NIH PRAGMATIC TRIALS
COLLABORATORY**

Rethinking Clinical Trials®

Main points: Decision-making & Pragmatic Trials

- Decision-making is complex and multidimensional
- What is important depends on:
 - Context, audience, many other factors
- P-values can be a useful part of decision-making
 - Certainly not the only one!

P-Values: What?

- P-values are part of statistical process of hypothesis/significance testing:
 - Data evidence for degree of ‘surprise’ of a finding
 - Compared to null hypothesis (e.g. current practice)
 - Testing typically dichotomous ($p < 0.05$ vs. $p \geq 0.05$)
 - An attempt to minimize risk of wrong decision (Type I error)

P-Values: Lots has been said



AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®

732 North Washington Street, Alexandria, VA 22314 • (703) 684-1221 • Toll Free: (888) 231-3473 • www.amsta.org • [www.twitter.com/AmstatNews](https://twitter.com/AmstatNews)

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND P-VALUES

*Provides Principles to Improve the Conduct and Interpretation of Quantitative
Science*

March 7, 2016



The *p*-value statement, five years on

The American Statistical Association's 2016 *p*-value statement generated debates and disagreements, editorials and symposia, and a plethora of ideas for how science could be changed for the better. Now, five years on, Robert Matthews asks what, if anything, has the statement achieved?

JAMA

In the reporting of results, when possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty, such as confidence intervals (see **Reporting Standards and Data Presentation**). Avoid relying solely on statistical hypothesis testing, such as the use of *P* values, which fails to convey important quantitative information.

16 | SIGNIFICANCE | April 2021

Ref: <https://jamanetwork.com/journals/jama/pages/instructions-for-authors#SecReportingStandardsandDataPresentation>

P-Values: Why Now?

- NIH Collaboratory experiences with pragmatic trials
- Decisions made based on pragmatic trials different to decisions based on traditional explanatory trial, e.g. FDA decision re approval of new drug
- What kind of information do decision-makers value for pragmatic trials?

Pragmatic Trials: What Information is Valuable?

- How large?
 - Is intervention effect size meaningful?
- For what?
 - Which outcomes? Multiple outcomes?
- For whom?
 - Sub-groups
 - Generalizability (e.g. multicenter study)
- How? Why?
 - Implementation questions on fidelity, reach etc.
 - Is the intervention already in place? Easy to apply?
 - Reasonably priced?

Decision-making: Three quantitative practices

- Hypothesis testing with P-value benchmark as a tool
- Bayesian evidence synthesis
- Process to elicit what matters to decision-makers
 - Before trial start
 - During trial, before examining results

Practice 1: Decision-maker & P-values

STATISTICS IN BIOPHARMACEUTICAL RESEARCH
2021, VOL. 13, NO. 1, 57–58
<https://doi.org/10.1080/19466315.2021.1886164>



Taylor & Francis
Taylor & Francis Group



Statement on *P*-values

Thomas Gwise, Mark D. Rothmann, H.M. James Hung, Anup Amatya, Rebecca Rothwell, Sue Jane Wang, Yute Wu, Fraser Smith, Yu-Ting Weng, Eugenio Andraca-Carrera, Stella Grosser, Somesh Chattopadhyay, and Sylva H. Collins

FDA Center for Drug Evaluation and Research, Office of Translational Science, Office of Biostatistics, Silver Spring, MD

Practice 1: Decision-maker & P-values

STATISTICS IN BIOPHARMACEUTICAL RESEARCH
2021, VOL. 13, NO. 1, 57–58
<https://doi.org/10.1080/19466315.2021.1886164>



Statement on *P*-values

Thomas Gwise, Mark D. Rothmann, H.M. James Hung, Anup Amatya, Rebecca Rothwell, Sue Jane Wang, Yute Wu, Fraser Smith, Yu-Ting Weng, Eugenio Andraca-Carrera, Stella Grosser, Somesh Chattopadhyay, and Sylva H. Collins

FDA Center for Drug Evaluation and Research, Office of Translational Science, Office of Biostatistics, Silver Spring, MD

“The convention of typically limiting type I error probability to two-sided 0.05, the usual *p*-value benchmark, provides a common reference point for study design.

Practice 1: Decision-maker & P-values

STATISTICS IN BIOPHARMACEUTICAL RESEARCH
2021, VOL. 13, NO. 1, 57–58
<https://doi.org/10.1080/19466315.2021.1886164>



Statement on *P*-values

Thomas Gwise, Mark D. Rothmann, H.M. James Hung, Anup Amatya, Rebecca Rothwell, Sue Jane Wang, Yute Wu, Fraser Smith, Yu-Ting Weng, Eugenio Andraca-Carrera, Stella Grosser, Somesh Chattopadhyay, and Sylva H. Collins

FDA Center for Drug Evaluation and Research, Office of Translational Science, Office of Biostatistics, Silver Spring, MD

“The convention of typically limiting type I error probability to two-sided 0.05, the usual *p*-value benchmark, provides a common reference point for study design. Without common benchmarks, drug developers would face added uncertainty in designing studies and regulators would need to spend more time independently establishing expectations for every study reviewed.

Practice 1: Decision-maker & P-values

STATISTICS IN BIOPHARMACEUTICAL RESEARCH
2021, VOL. 13, NO. 1, 57–58
<https://doi.org/10.1080/19466315.2021.1886164>



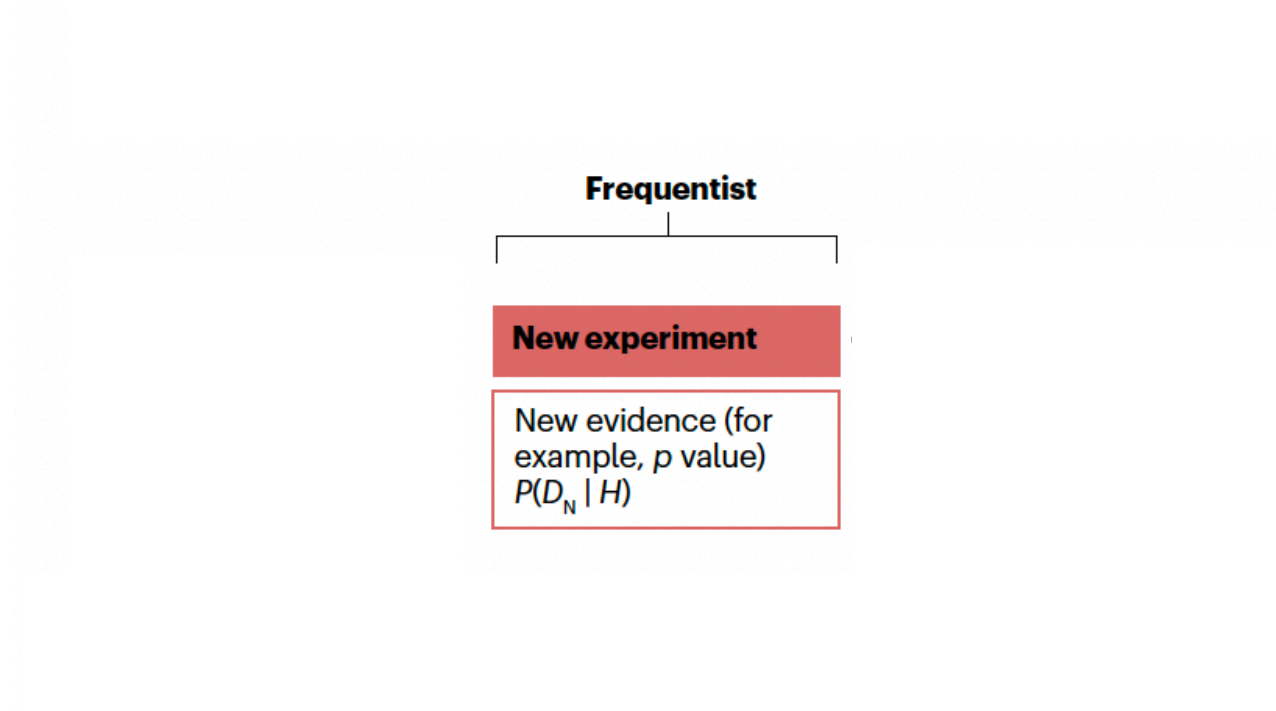
Statement on *P*-values

Thomas Gwise, Mark D. Rothmann, H.M. James Hung, Anup Amatya, Rebecca Rothwell, Sue Jane Wang, Yute Wu, Fraser Smith, Yu-Ting Weng, Eugenio Andraca-Carrera, Stella Grosser, Somesh Chattopadhyay, and Sylva H. Collins

FDA Center for Drug Evaluation and Research, Office of Translational Science, Office of Biostatistics, Silver Spring, MD

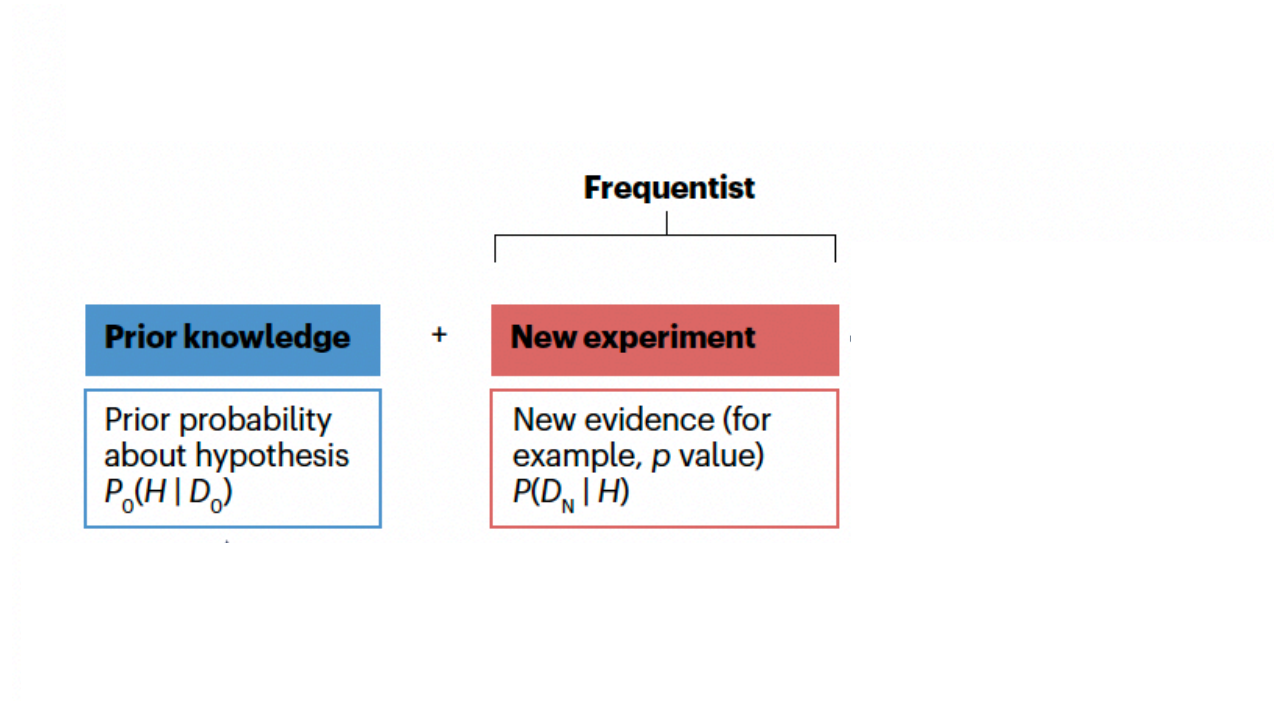
“The convention of typically limiting type I error probability to two-sided 0.05, the usual *p*-value benchmark, provides a common reference point for study design. Without common benchmarks, drug developers would face added uncertainty in designing studies and regulators would need to spend more time independently establishing expectations for every study reviewed. CDER alone reviews over 2,000 analyses per year. Maintaining consistency, fairness and transparency while reviewing products in a timely manner would be challenging without a conventional, if arbitrary, benchmark

Practice 2: Bayesian evidence synthesis



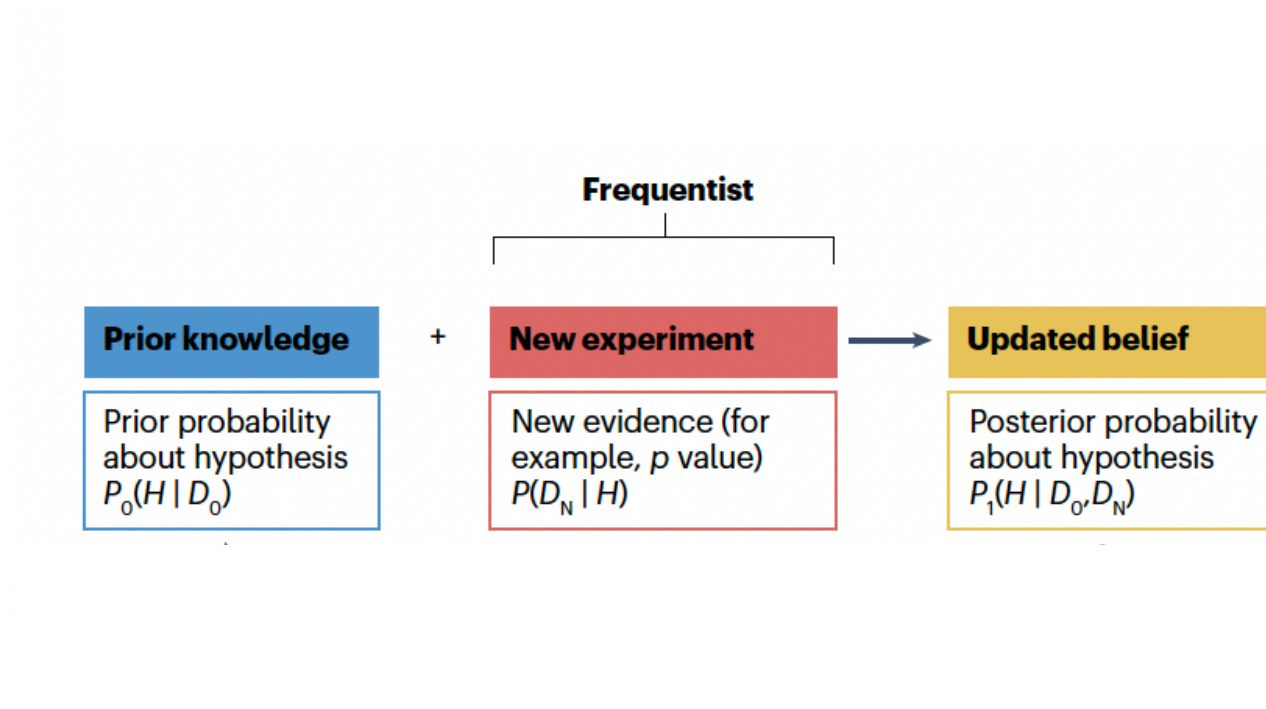
From: Ruberg *et al*, Nature Reviews Drug Discovery 2023

Practice 2: Bayesian evidence synthesis



From: Ruberg *et al*, Nature Reviews Drug Discovery 2023

Practice 2: Bayesian evidence synthesis



From: Ruberg *et al*, Nature Reviews Drug Discovery 2023

Practice 2: Bayesian evidence synthesis

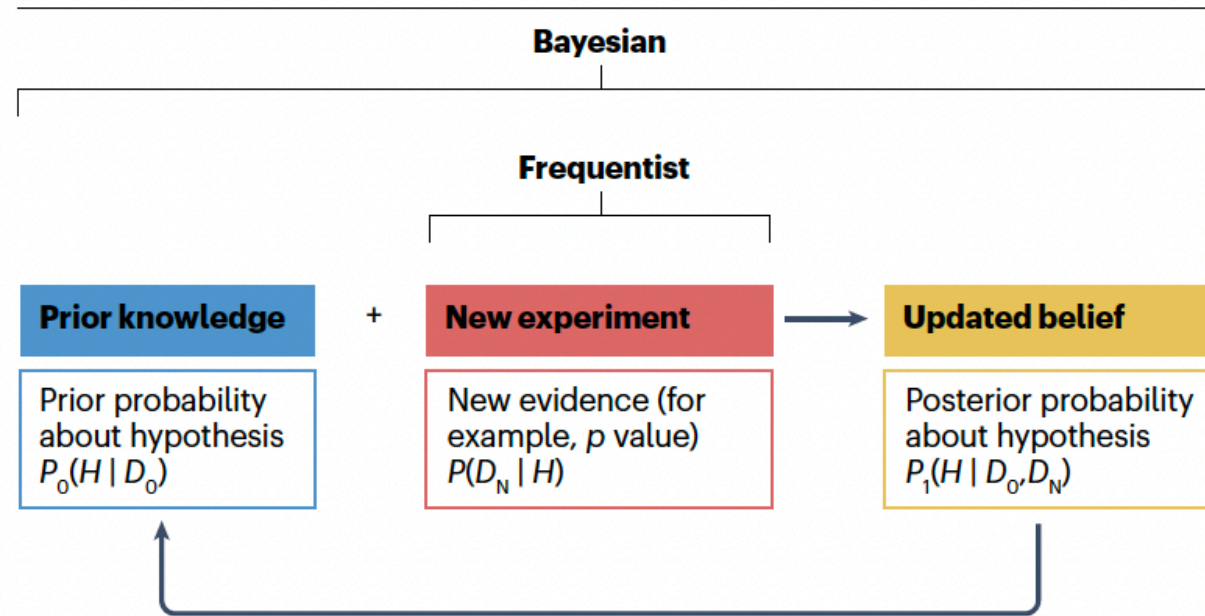


Fig. 1 | Comparison between Bayesian and frequentist approaches.

From: Ruberg *et al*, Nature Reviews Drug Discovery 2023

Practice 2: Bayesian evidence synthesis

Pfizer and BioNTech sponsored a trial of their BNT162b2 mRNA vaccine for prevention of COVID-19. For the phase III portion of their clinical programme, the primary efficacy analysis was based on the Bayesian posterior probability that vaccine efficacy was $>30\%$. The success

From: Ruberg *et al*, Nature Reviews Drug Discovery 2023

Practice 2: Bayesian evidence synthesis

The success
criterion was explicitly defined as $P(\text{vaccine efficacy} > 30\%) > 98.6\%$.

From: Ruberg *et al*, Nature Reviews Drug Discovery 2023

Practice 2: Bayesian evidence synthesis

The success criterion was explicitly defined as $P(\text{vaccine efficacy} > 30\%) > 98.6\%$. That is, regardless of any p value calculation, the study success criterion was a 98.6% probability that the true vaccine efficacy (VE) was greater than the public health minimum requirement of 30%.

From: Ruberg *et al*, Nature Reviews Drug Discovery 2023

Practice 3: Elicitation of “what matters?”

Example: Pre-Trial

What are the characteristics of a trial that will lead to high-quality evidence that would help you change your decision-making or change practice?

Practice 3: Elicitation of “what matters?”

PLOS MEDICINE

RESEARCH ARTICLE

Effectiveness of a primary care-based integrated mobile health intervention for stroke management in rural China (SINEMA): A cluster-randomized controlled trial

Lijing L. Yan^{1,2,3,4,5*}, Enying Gong^{1,6}, Wanbing Gu^{1,7}, Elizabeth L. Turner^{2,8}, John A. Gallis^{2,8}, Yun Zhou⁹, Zixiao Li⁹, Kara E. McCormack⁸, Li-Qun Xu¹⁰, Janet P. Bettger^{2,11}, Shenglan Tang², Yilong Wang⁹, Brian Oldenburg⁶

Example: During Trial & Pre-Analysis

- Trial powered for between-arm mean diff. of 5mmHg
- Team meeting prior to analysis: what is meaningful effect?

Results: Between-arm mean diff. in 1-year change in SBP:
2.8 mmHg (95% CI: 0.9, 4.8, $p=0.005$)

Practice 3: Elicitation of “what matters?”

PLOS MEDICINE

RESEARCH ARTICLE

Effectiveness of a primary care-based integrated mobile health intervention for stroke management in rural China (SINEMA): A cluster-randomized controlled trial

Lijing L. Yan^{1,2,3,4,5*}, Enying Gong^{1,6}, Wanbing Gu^{1,7}, Elizabeth L. Turner^{2,8}, John A. Gallis^{2,8}, Yun Zhou⁹, Zixiao Li⁹, Kara E. McCormack⁸, Li-Qun Xu¹⁰, Janet P. Bettger^{2,11}, Shenglan Tang², Yilong Wang⁹, Brian Oldenburg⁶

“.....BP control was significantly improved [...]. The intervention also improved 6 out of 7 prespecified secondary outcomes and all exploratory outcomes on stroke recurrence, hospitalization, disability, and mortality, at an annual cost of less than US\$24 per patient.”

Yan et al. (2021), PLoS Medicine 18(4): e1003582

Main points: Decision-making & Pragmatic Trials

- Decision-making is complex and multidimensional
- What is important depends on context & the audience
- P-values can be a useful part of decision-making
 - Certainly not the only one!
- Session title: P-Values vs. Decision-Maker Perspectives
 - Instead: P-Values as part of Decision-Maker Perspectives