Center for Tuberculosis
UCSF
University of California San Francisco

Limitations of Non-Inferiority Designs and Pragmatic Trials as Alternatives

20th May 2026

Patrick Phillips, PhD

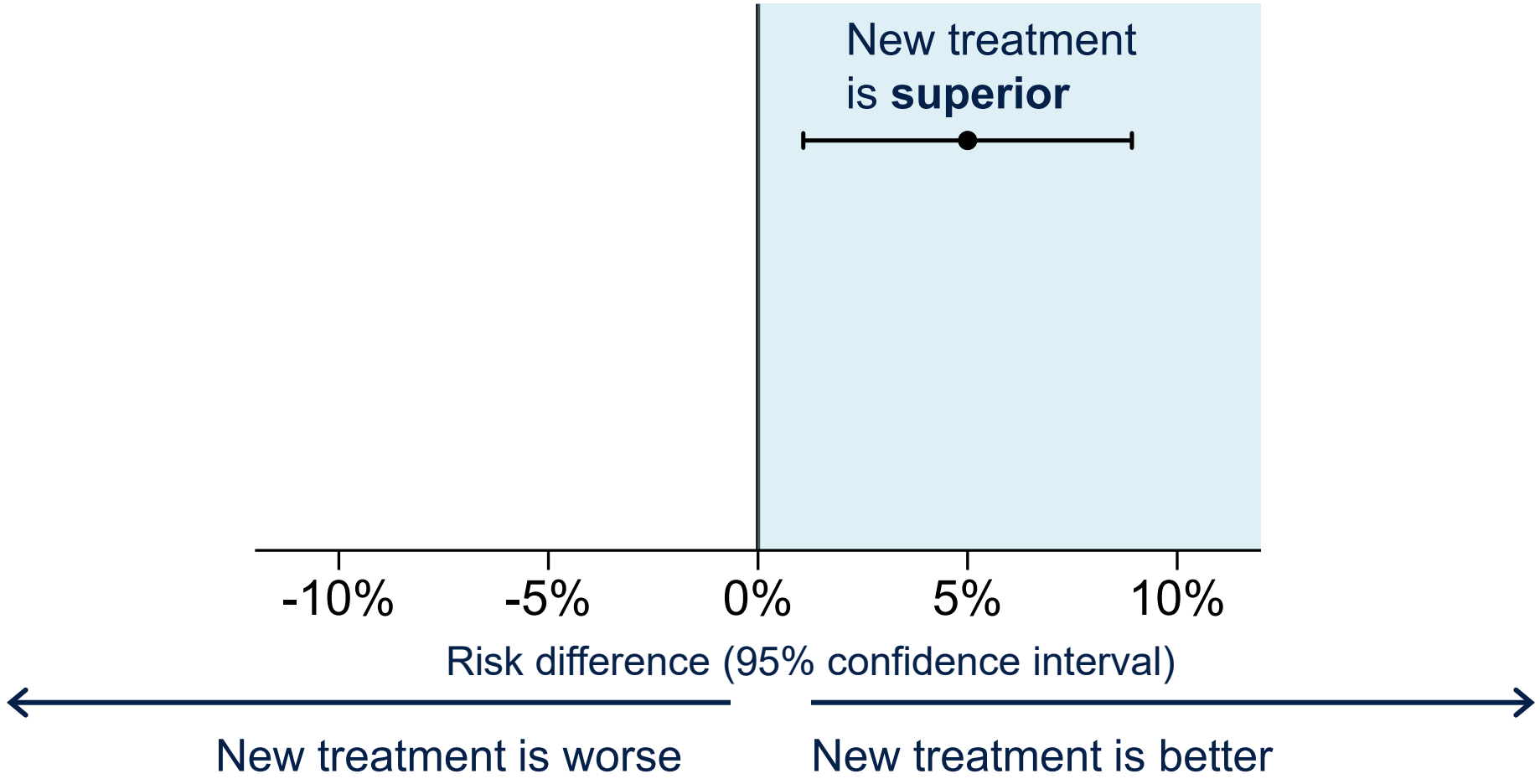
Associate Professor in Residence
Departments of Medicine, and Epidemiology & Biostatistics
Patrick.Phillips@ucsf.edu

What would you choose?

Outline

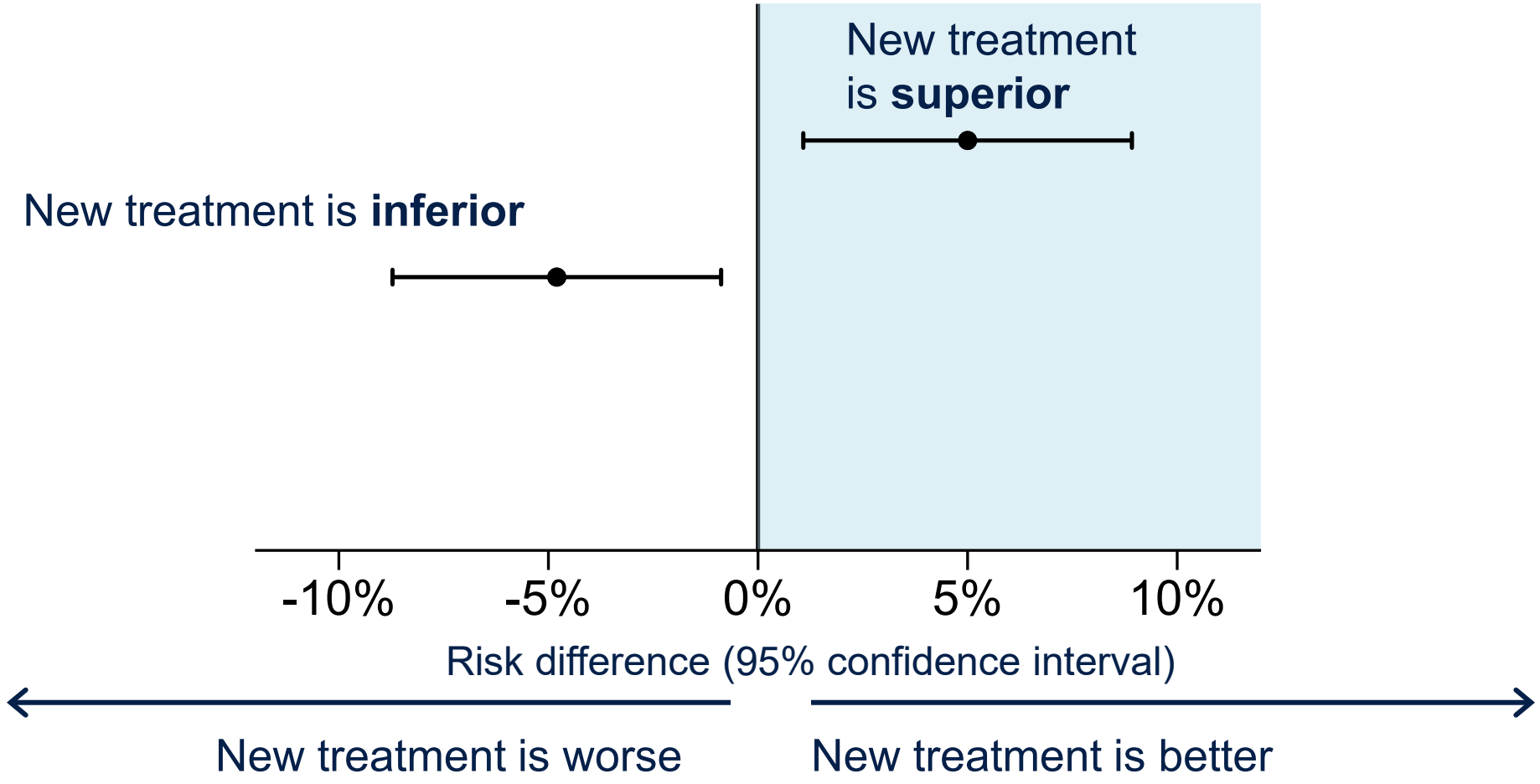
- What is non-inferiority?
- When might it be appropriate?
- What are the limitations?
- What are the alternatives?
  1. Pragmatic Trials
  2. Averted Infections Ratio (Dave Glidden)
  3. Duration-Response Randomized Design (Suzanne Dufault)
  4. Net Benefit and Hierarchical Outcomes (Johan Verbeeck)

What is non-inferiority?



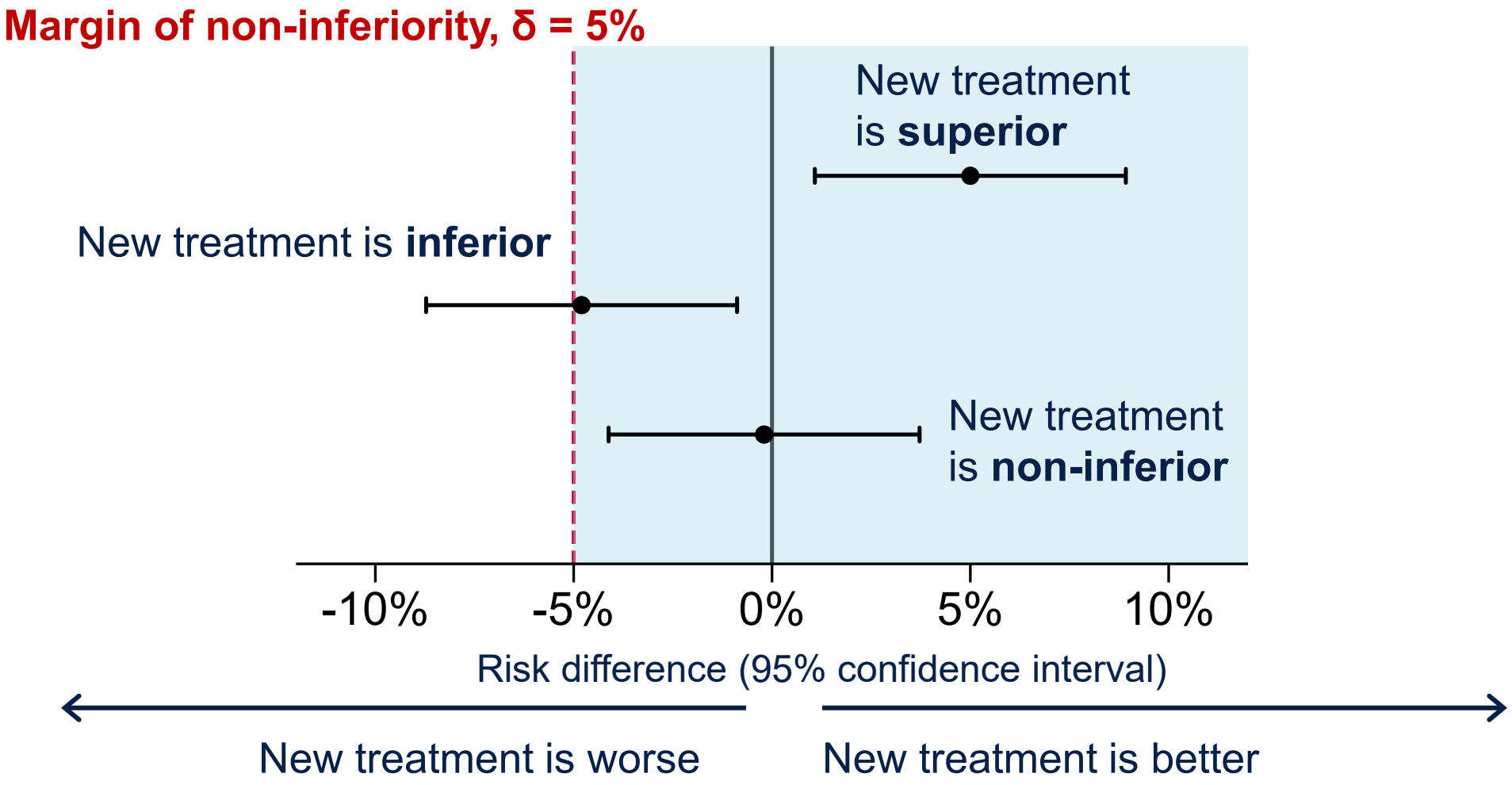
UCSF

What is non-inferiority?



UCSF

What is non-inferiority?



Comparison of hypotheses (null, H_0, and alternative, H_1)

Superiority trial

- H_0: **d = 0**
- H_1: **d > 0**

- Assume **d = δ** under H_1 in the sample size calculation.

Non-inferiority trial

- H_0: **d = -δ**
- H_1: **d > -δ**

- Assume **d = 0** under H_1 in the sample size calculation.

Just a simple translation of hypotheses by **δ**?

Comparison of hypotheses (null, H_0, and alternative, H_1)

Non-inferiority trial

- H_0: **$d = -\delta$**
- H_1: **$d > -\delta$**

Superiority trial

- H_0: **$d = 0$**
- H_1: **$d > 0$**

- This is a **conceptual change**, not just a **numerical shift**
  - A superiority trial asks whether there is evidence **for any difference between arms**
  - A non-inferiority trial asks whether there is evidence **that the intervention is not unacceptably worse than the control**.

Non-inferiority is not new

- **1950s-1970s:** *Equivalence* and *non-inferiority* trials conducted
- **1980s**: First formal statistical development of methods
- **2010**: FDA draft guidance for non-inferiority trials (finalized **2016**)
- Incidence of non-inferiority has grown dramatically:
  - **483** indexed in PubMed in **2023**
  - **168** published in major general medical journals (IF > 10) in **2010-2015**

Rehal S, Morris TP, Fielding K, Carpenter JR, Phillips PP. Non-inferiority trials: are they inferior? A systematic review of reporting in major medical journals. BMJ Open. 2016;6(10):e012594.

When might non-inferiority be appropriate?

- Invoked when there is already an **established standard of care**.

- Typically considered appropriate when there are benefits with the new treatment:
  - Improved tolerability or quality of life, shorter duration, higher barrier to drug resistance, more acceptable to patients, easier administration,..
- However, can be used '*to support the conclusion that the new test drug is… effective.*' (FDA guidance, 2016)
  - Evaluate a new treatment in an already crowded market.

What are the limitations?

Controversies

Non-inferiority trials are unethical because they disregard patients' interests

Silvio Garattini, Vittorio Bertele'

Lancet 2007; 370: 1875–77

Published Online
October 23, 2007
DOI:10.1016/S0140-6736(07)61604-3

Mario Negri Institute for Pharmacological Research, Milan, Italy (Prof S Garattini MD, V Bertele' MD)

Correspondence to:
Dr Vittorio Bertele', Mario Negri Institute for Pharmacological Research, Milan, 20156, Italy
bertele@marionegri.it

Equivalence trials[1] have been widely used to assess new drugs, but have recently lost ground to a non-inferiority design. This type of trial is usually accepted by regulatory authorities for approval of new drugs or new indications,

but not to the extent that it is recognised as such. For example, if the non-inferiority limit is set at 7·5%, an increase in the incidence of serious events or deaths—say 7% instead of the 5% currently established

*"We argue that the scientific community should **ban non-inferiority trials because they are unethical**…"*

The ethics of non-inferiority trials

Silvio Garattini and Vittorio Bertele'[1] rightly caution about the traps to be avoided in doing non-inferiority trials, but go too far in suggesting t
trials should be uniformly
"because they are unethical".
example of the appropriate
a non-inferiority trial is the
research programmes to deve
drugs for treatment of drug-s
tuberculosis.

Internationally recommended regimens are highly effective, curing 95% or more of patients in clinical trials in a wide variety of settings.[2,3] However, they require a minimum of three drugs which have significant side-effects and need to be given for at least 6 months. Improving on such high cure rates is almost impossible, but shortening treatment duration would improve completion rates and reduce both the time that patients are exposed to potentially toxic drugs and the cost of delivering tuberculosis

situation in which patients can justifiably be entered into

(Dec 1, p 1875)[1] argue that
inferiority trials "have no et
justification, since they do not offe
possible advantage...to patients". T
conclusion is based on the traditi
ethic of physicians whereby advo
for each patient's best interest
supersede all other considerati
However, this ethic only applies
world where resources for health
are endless and hence do not ma

een as large enough to
he new and the control
herefore be considered
even if in 1000 patients
could be 20 more deaths

These arguments also apply to equivalence trials, which aim to prove similarity of a new drug to the comparator,

Garattini, S. and V. Bertele (2007). "Non-inferiority trials are unethical because they disregard patients' interests." Lancet **370**(9602): 1875-1877.

UCSF

1. Not person-centered

- Patients want treatments that are superior, not ‘slightly worse’.
- Putative benefits of treatment are rarely included in the primary outcome.
  - Improved tolerability or quality of life, shorter duration, higher barrier to drug resistance, more acceptable to patients, easier administration,..

2. A negotiated compromise

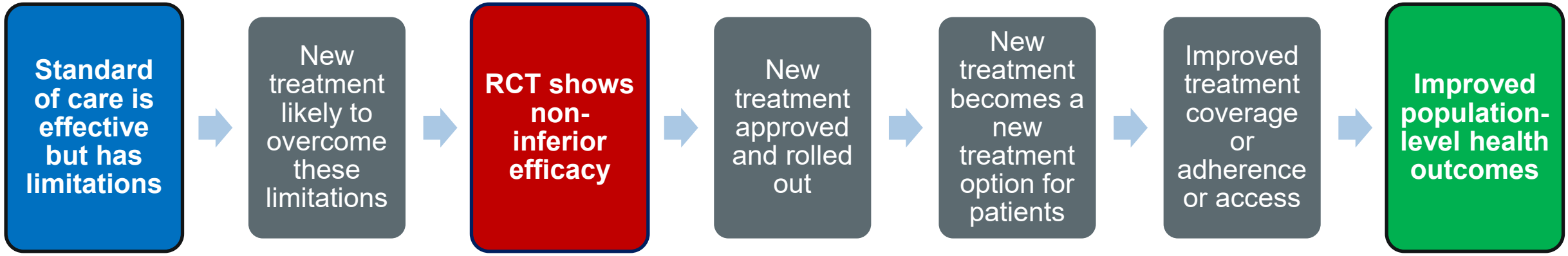
- The margin of non-inferiority is essentially arbitrary.
  - Different stakeholders will value the risk-benefit balance differently.
  - In practice, the margin of non-inferiority is the largest reduction in efficacy that funders/reviewers/regulators will accept.

3. Historically contingent

- FDA guidance requires that margin justification depend explicitly on previous trials.
- A non-inferiority trial must be similar in all respects to previous trials of the active control (**constancy assumption**)
  - This is highly implausible.
- Trial quality must be compared against external standards of 'what is good enough' (**assay sensitivity**) rather than letting the results stand alone.

What are the alternatives?
1. Pragmatic Trials

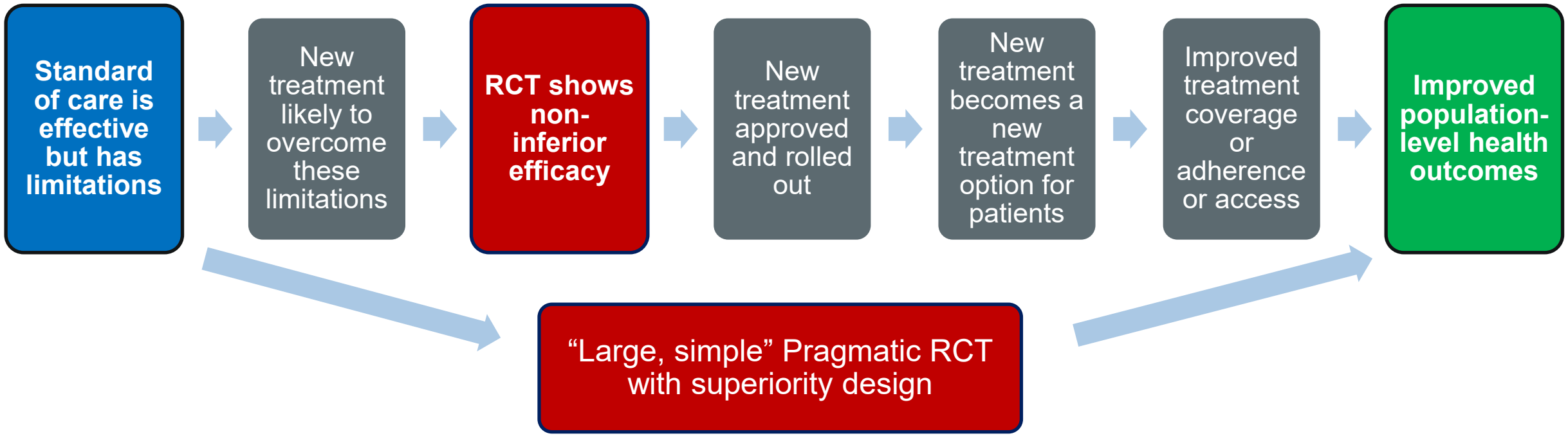
The conceptual paradigm for non-inferiority



UCSF

The conceptual paradigm for non-inferiority

Reframing as a **pragmatic superiority trial**



UCSF

Case study 1. The BLISTER trial

- Bullous pemphigoid (BP) is a rare autoimmune skin disease, occurring mainly in the elderly
- Standard of care: Prednisolone (oral steroids)
  - Highly effective
  - Many significant long-term side effects
  - Increased mortality
- Intervention: Doxycycline
  - Likely less effective
  - Much safer
- Primary outcome: No more than three blisters after 6 weeks

Case study 1. The BLISTER trial

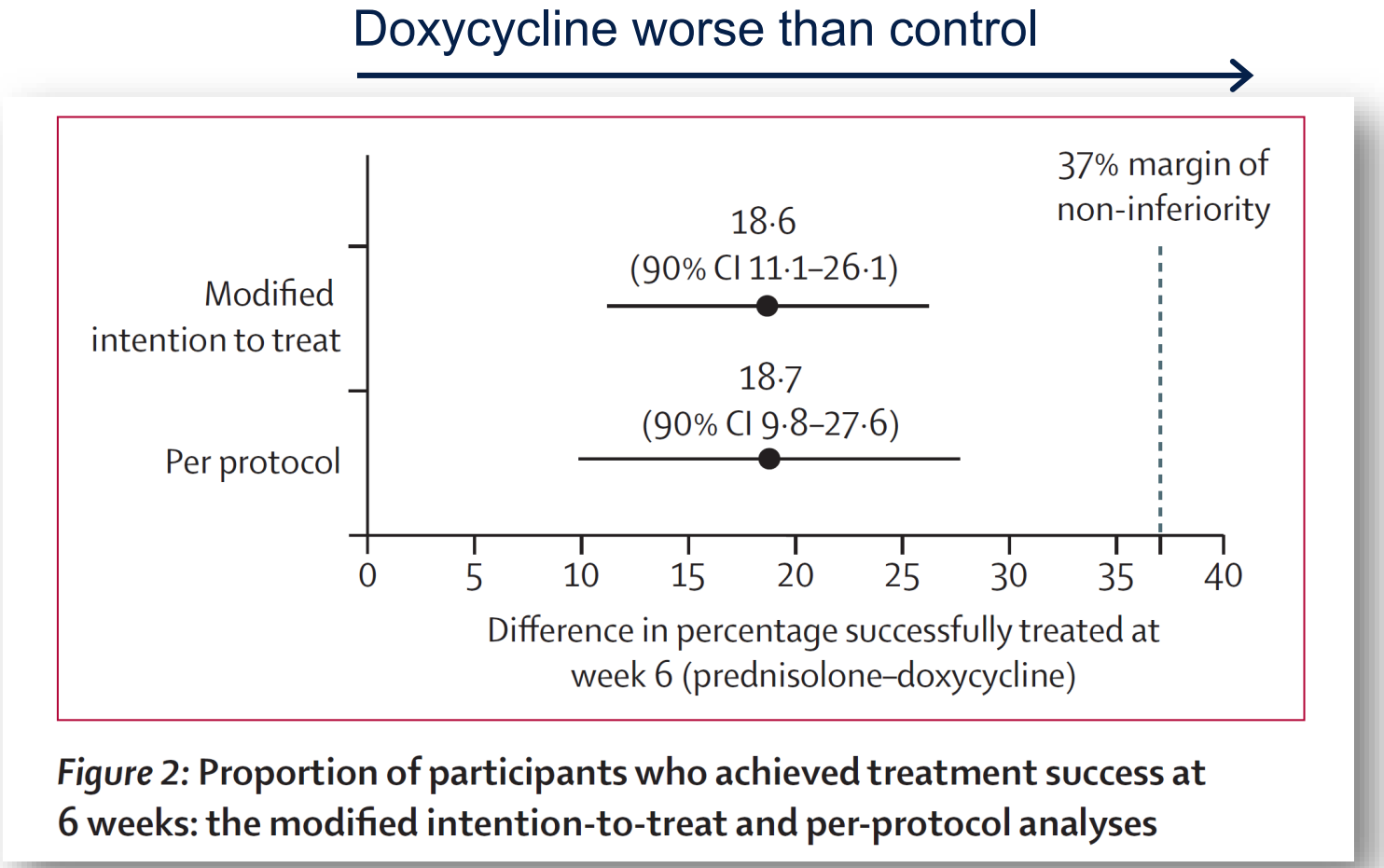


Figure 2: **Proportion of participants who achieved treatment success at 6 weeks: the modified intention-to-treat and per-protocol analyses**

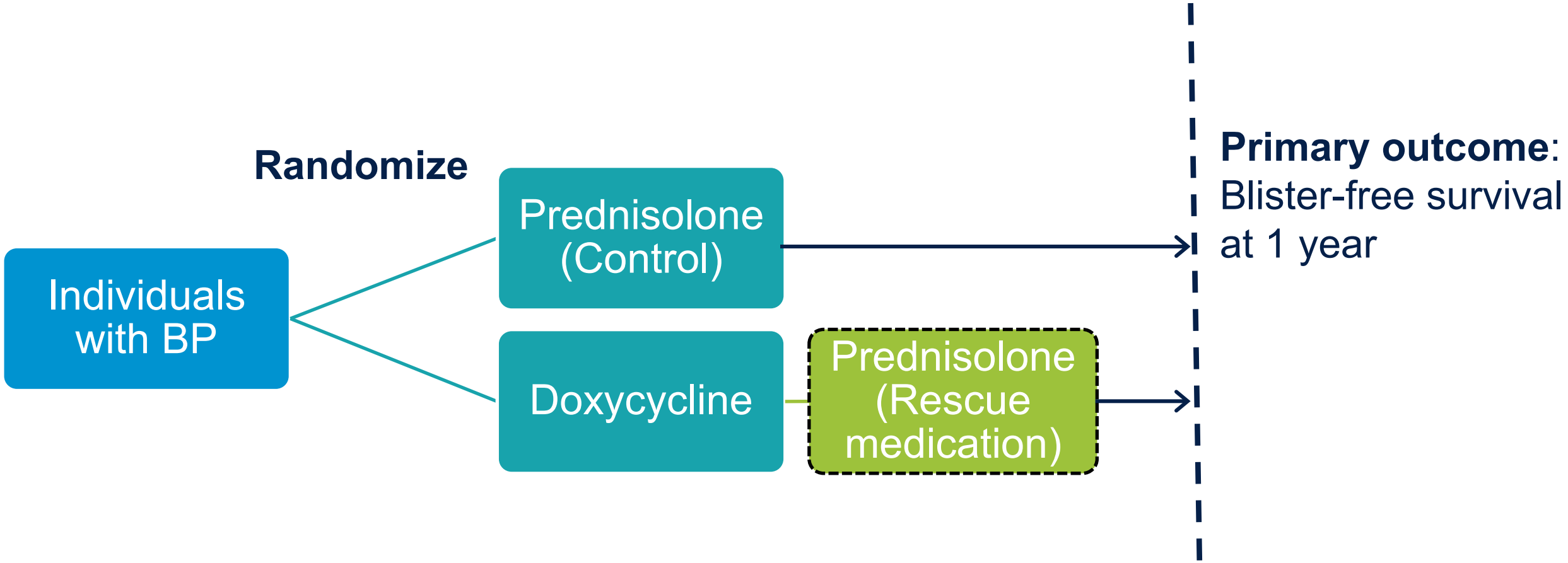
Interpretation Starting patients on doxycycline is non-inferior to standard treatment with oral prednisolone for short-term blister control in bullous pemphigoid and significantly safer in the long-term.

Williams HC, Wojnarowska F, Kirtschig G, Mason J, Godec TR, Schmidt E, et al. Doxycycline versus prednisolone as an initial treatment strategy for bullous pemphigoid: a pragmatic, non-inferiority, randomised controlled trial. Lancet. 2017;389(10079):1630-8.

UCSF

Case study 1. The BLISTER trial

Reframing as a **pragmatic superiority strategy trial**

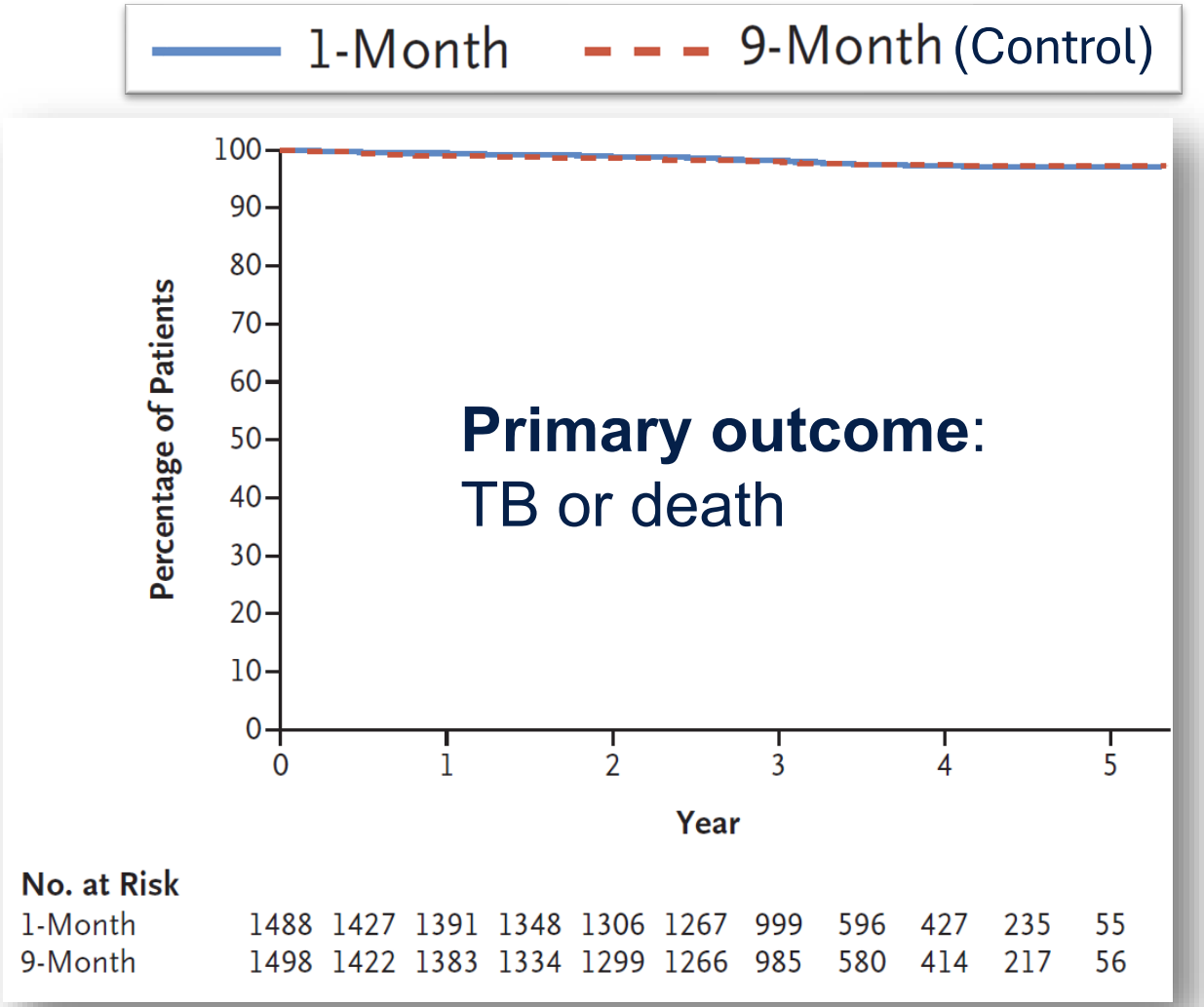


Case study 2. BRIEF-TB: Tuberculosis (TB) Prevention

- BRIEF-TB evaluated a new 1-month regimen compared to the 9-month control with a non-inferiority design.

CONCLUSIONS

A 1-month regimen of rifapentine plus isoniazid was noninferior to 9 months of isoniazid alone for preventing tuberculosis in HIV-infected patients.



Swindells S, Ramchandani R, Gupta A, Benson CA, Leon-Cruz J, Mwelase N, et al. One Month of Rifapentine plus Isoniazid to Prevent HIV-Related Tuberculosis. N Engl J Med. 2019;380(11):1001–11.

UCSF

Case study 2. BRIEF-TB: Tuberculosis (TB) Prevention

Reframing as a **pragmatic superiority cluster-randomized trial**

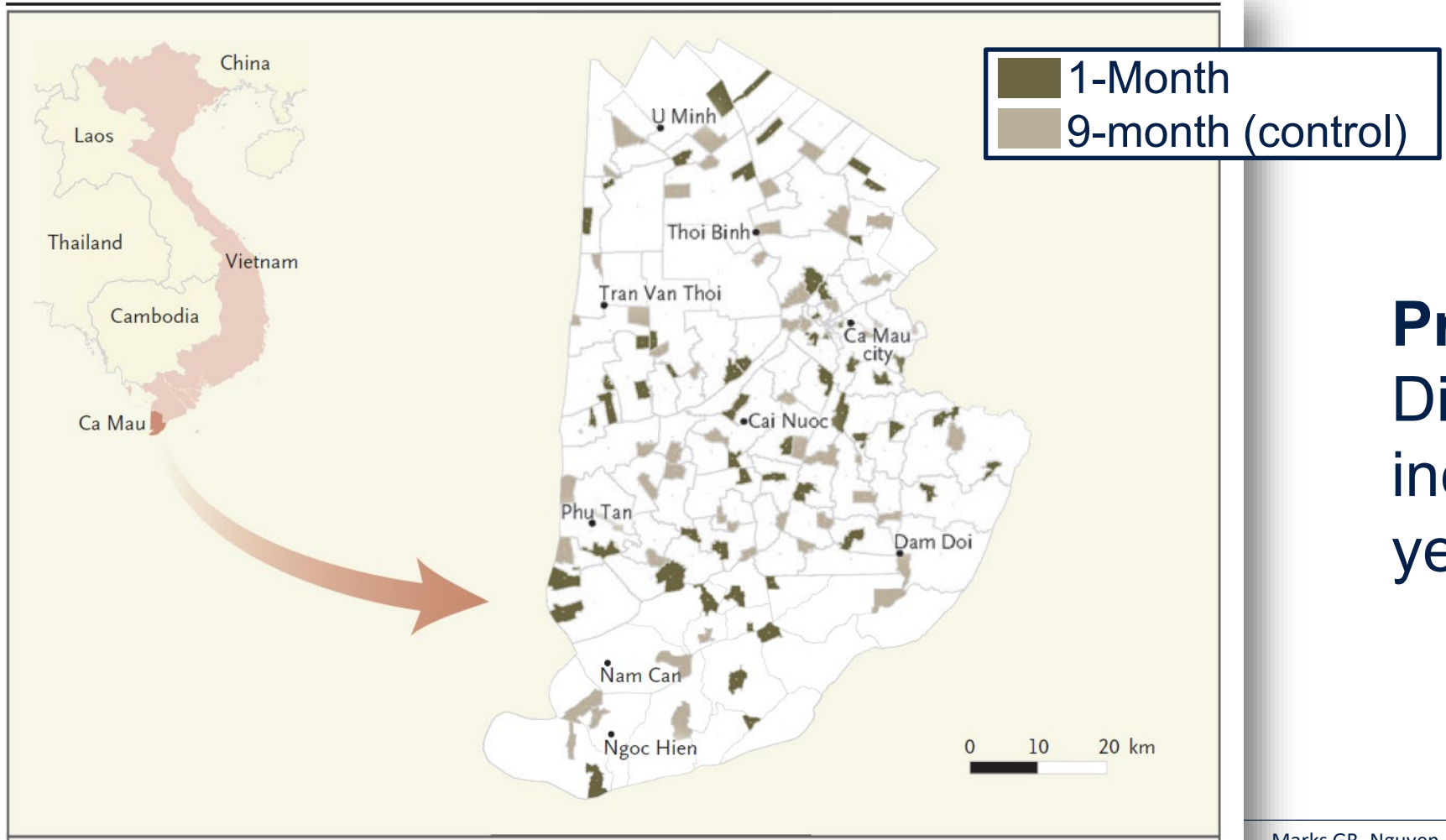


Figure 1. Map of the 9 Districts and 949 Subcommunes in Ca Mau Province, Vietnam.

Primary outcome: District-level TB incidence after 5 years

Marks GB, Nguyen NV, Nguyen PTB, Nguyen TA, Nguyen HB, Tran KH, et al. Community-wide Screening for Tuberculosis in a High-Prevalence Setting. N Engl J Med. 2019;381(14):1347–57.

UCSF

Conclusions

- Non-inferiority trials yield results that are arbitrary, obscure, and unreliable.
- Avoid at all costs, except when necessary!
- There are alternatives…

Non Inferiority Based on the Averted Events Ratio: A Bayesian Framework

David V. Glidden, Ph.D.
Professor of Biostatistics
University of California, San Francisco
david.glidden@ucsf.edu

Landscape HIV Prevention

- Oral PrEP (TDF/FTC, truvada) is effective, safe
  PrEP has not yet dented HIV incidence
  adherence is variable and determines effectiveness

- Alternative agents remain clinically valuable
  want effective <u>alternatives</u>, not a <u>replacement</u>

- Trial Challenge:
  low incidence→low # events→difficult interpretation
  NI margins using 95/95 suggest very large sample sizes

- How to advance a robust pipeline of products?

DISCOVER Results

Mayer et al 2020

Arm	Pt Years FU	Post-Enroll HIV+	HIV Rate (100 PY)
TAF/FTC	4370	6	0.14
TDF/FTC	4386	11	0.25

Very low HIV rates: Good adherence or low risk population?

1 HIV switch would fail NI criterion

Rel. Risk = 0.55, 0.95 CI (0.20.1.48)

< 1.62 non-inferiority margin

Fails margin of 1.23 by another similar trial

Counterfactual Placebo

Population	Randomized	Followed for Safety and Incident HIV acquisition	Outcome
Eligible Consenting	Experimental	TDF/FTC Placebo - Daily TAF/FTC - Daily	λ_E
	Control	TDF/FTC - Daily TAF/FTC Placebo - Daily	λ_C
Comparable population followed in same way Unobserved	No Rx	Hypothetical Placebo	λ_0

No treatment "background" HIV rate (bHIV)

Gao, et al 2021 Glidden, et al 2021 Zhu, et al 2024

Revisiting DISCOVER

Under the assumption of bHIV 3.1 per 100 PY

Arm	Pt Years FU	Post-Enroll HIV+	HIV Rate (100 PY)	Averted HIV Events	Effectiveness
TAF/FTC	4370	6	0.14	129	96%
TDF/FTC	4386	11	0.25	124	92%
bHIV	4378	135	3.1	—	—

TAF/FTC prevented 1.03 (129/124) times as many infections

Averted Event Ratio (AER)

Large number of averted infections: AER estimate stable

Averted Event Ratio (AER)

Interpretable meta estimand comparing prevented events

Requires an estimate of λ_0

$$\text{Effectiveness}_E := \theta_E = 1 - \lambda_E/\lambda_0 \qquad \text{Effectiveness}_C := \theta_C = 1 - \lambda_C/\lambda_0$$

Effectiveness is proportion of events prevented by a regimen

$$\text{AER} = \Psi = \frac{\lambda_E - \lambda_0}{\lambda_C - \lambda_0} \qquad \Psi = \theta_E/\theta_C$$

Comparison of prevented events: index effect preservation

Dunn et al 2026, 2023, 2021, 2019, 2018

Estimating bHIV: Adherence

Incidence TDF/ FTC arm

Aggregate Risk Reduction Trial

bHIV TDF/ FTC arm

$$\lambda_C = \hat{RR} \quad \lambda_0 \qquad \hat{\lambda}_0 = \frac{\hat{\lambda}_C}{\hat{RR}}$$

- Uncertainty in $\hat{\lambda}_c$, $\hat{\lambda}_E$
- Uncertainty in adherence measurement
- Uncertainty in historical protection measurement

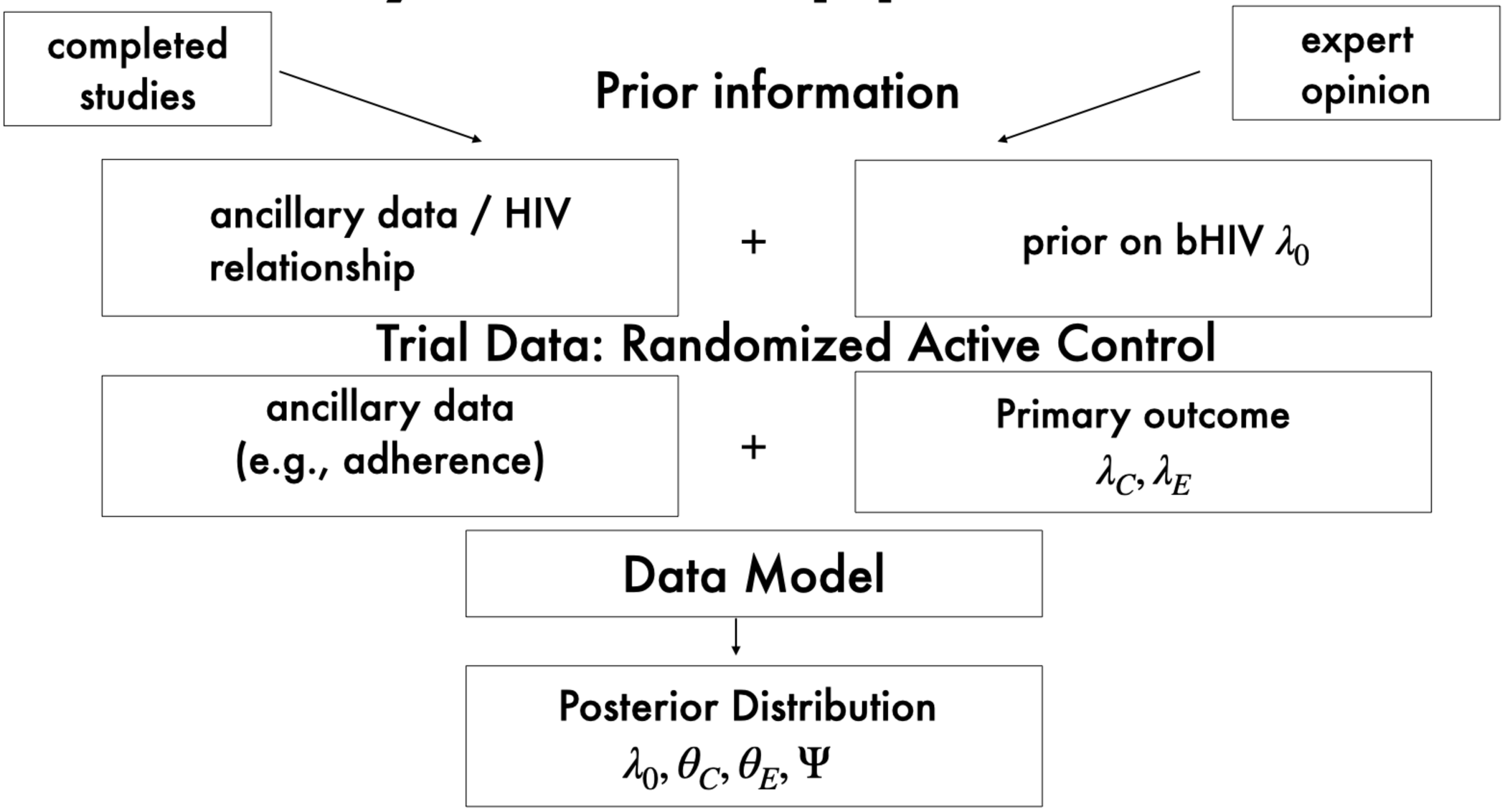
Effectiveness TAF/FTC =

$$1 - \frac{\hat{\lambda}_E}{\hat{\lambda}_P} * \hat{RR}$$

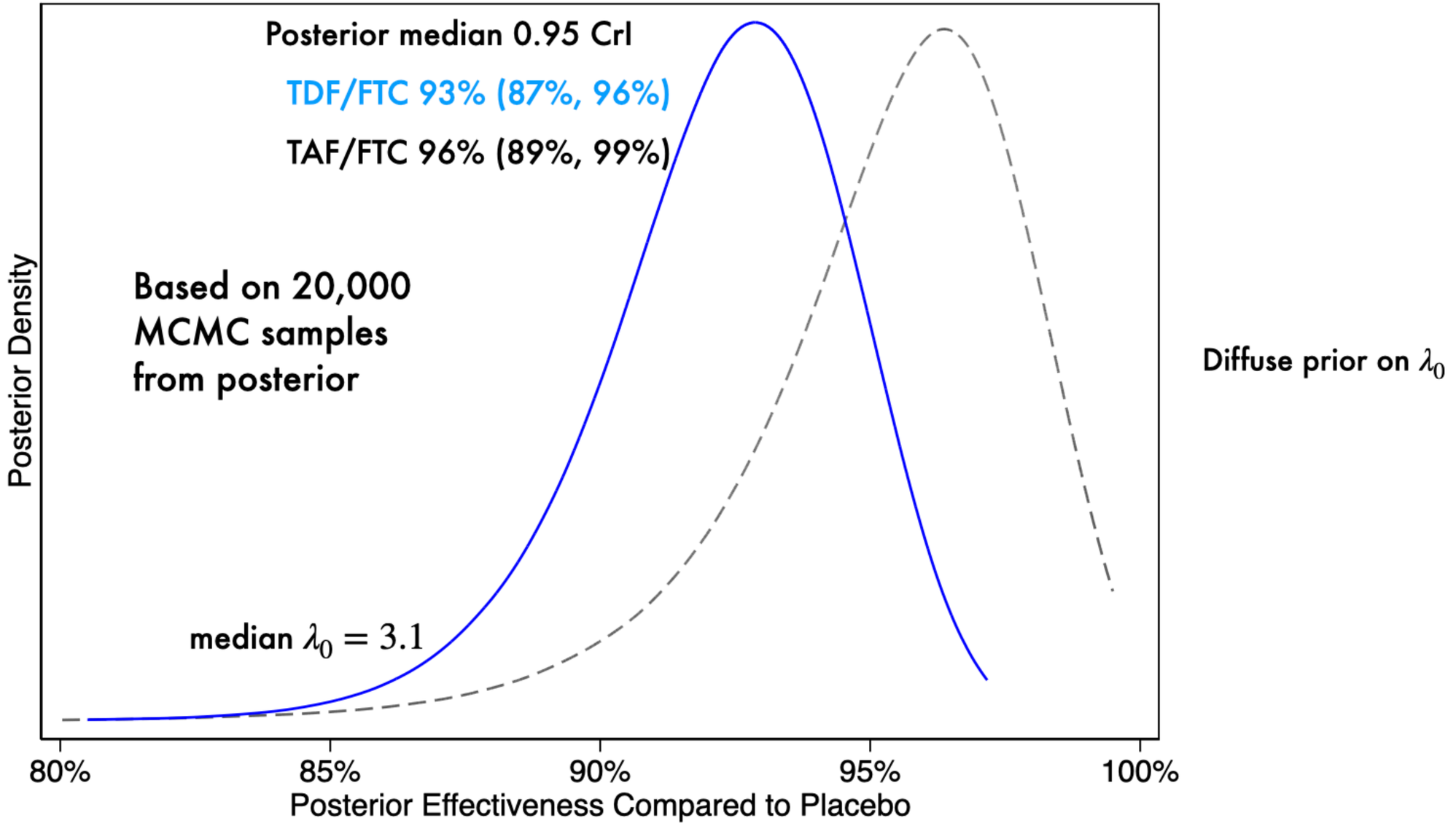
Glidden et al, 2021

Want to synthesize fully uncertainty

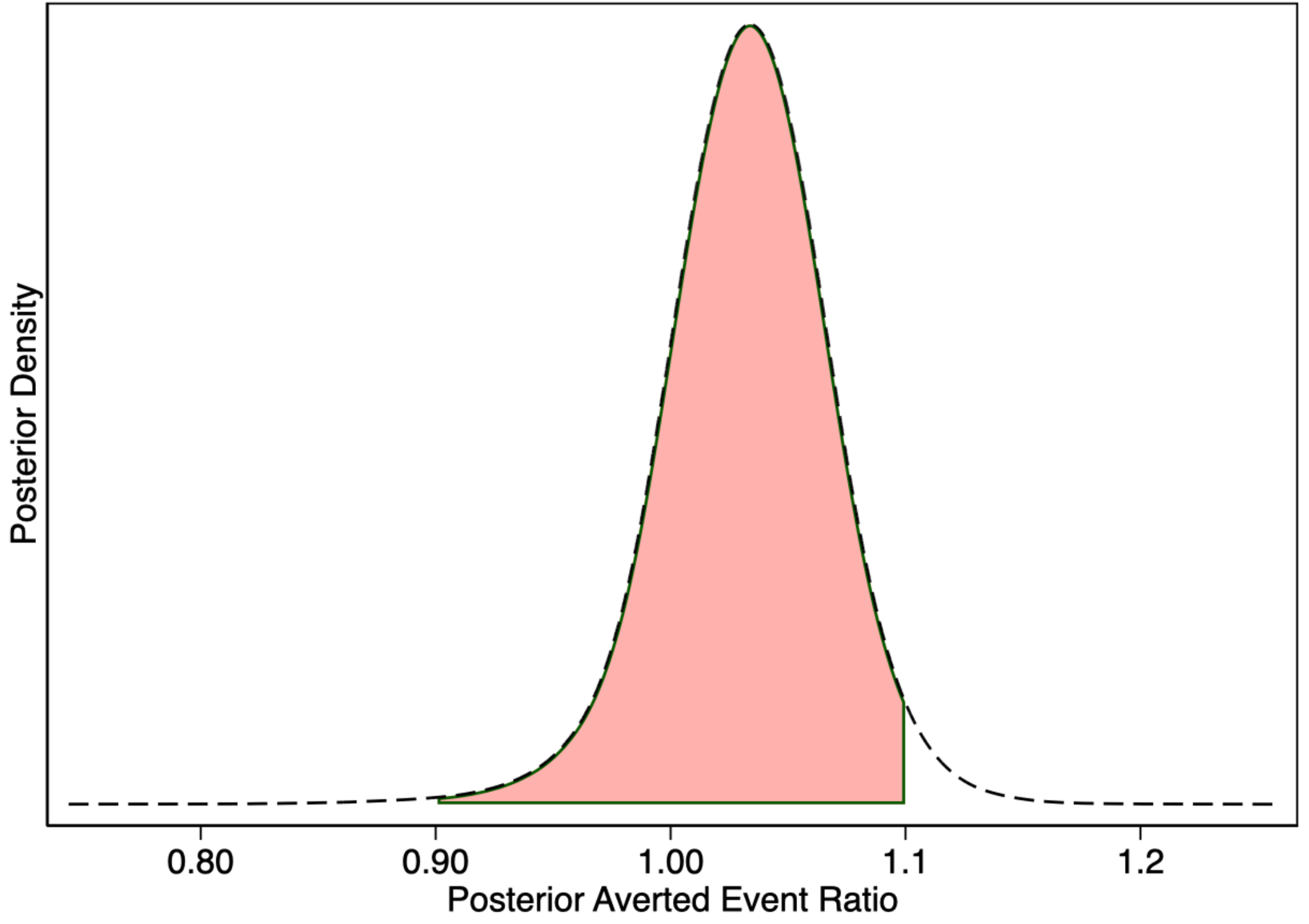
Bayesian Approach



Posterior Effectiveness



Averted Event Ratio Ψ



Posterior median AER (0.95 CrI)

1.03 (0.97, 1.09)

$\Pr(\Psi \geq 1) = 0.89$

$\Pr(\Psi \in [0.9,1.1]) = 0.98$

Based on 20,000 MCMC samples from posterior

Diffuse prior on λ_0

Framework

- Turned the comparison into an estimation problem
  Estimand: AER: interpretable quantity
  requires an estimate of bHIV (no treatment)

- Use data, prior opinion to inform bHIV
  diffuse prior similar to synthesis method
  skeptical prior provides sensitivity analysis

- Explicit assumptions, prior data, sensitivity analyses

Many Thanks

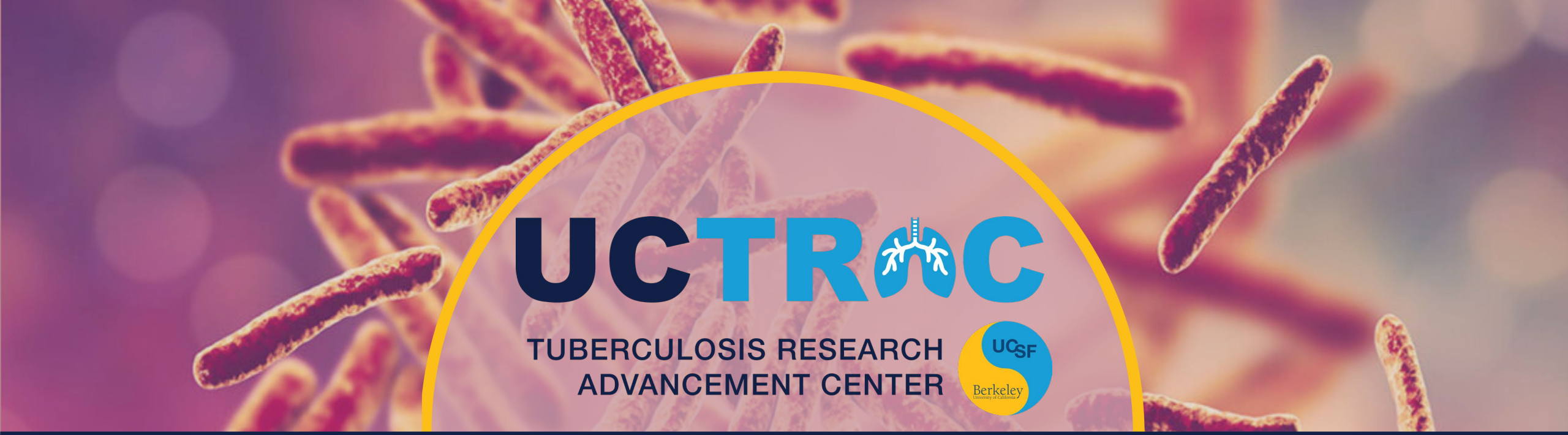
E. Brown, H. Janes,
F. Gao, D. Donnell

S. McCormack, D. Dunn,
O. Stirrup

P. Phillips
L. Dodd

M. Das,
J. Baeten

P. Anderson



The Duration-Response Randomized Design

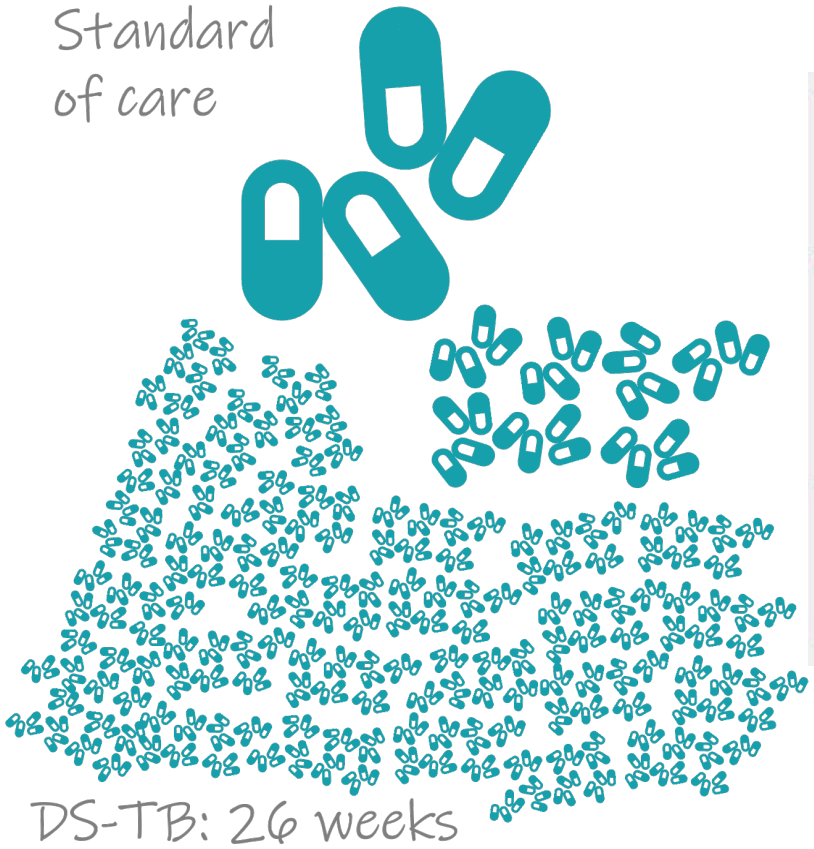
Suzanne Dufault, PhD
Division of Biostatistics, UCSF

Center for Tuberculosis
UCSF

*Funding is provided by NIH/NIAID awards **P30AI168440** and **1R25AI147375***

TB RAMP
TB Research and Mentorship Program

TB treatment is effective, but long



4-month regimen (S31/A5349)
Rpt Inh Pza Mfx

4-month regimen (SHINE)
Rif Inh Pza +/- Emb
(only for children with minimal TB)

6-month regimen
Rif Inh Pza Emb

DS-TB
no resistance

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Four-Month Rifapentine Regimens with or without Moxifloxacin for Tuberculosis

S.E. Dorman, P. Nahid, E.V. Kurbatova, P.P.J. Phillips, K. Bryant, K.E. Dooley, M. Engle, S.V. Goldberg, H.T.T. Phan, J. Hakim, J.L. Johnson, M. Lourens, N.A. Martinson, G. Muzanyi, K. Narunsky, S. Nerette, N.V. Nguyen, T.H. Pham, S. Pierre, A.E. Purfield, W. Samaneka, R.M. Savic, I. Sanne, N.A. Scott, J. Shenje, E. Sizemore, A. Vernon, Z. Waja, M. Weiner, S. Swindells, and R.E. Chaisson, for the AIDS Clinical Trials Group and the Tuberculosis Trials Consortium

2021 – Study 31 reported a successful Phase 3 study that found a 4-month regimen non-inferior to the standard of care 6-month regimen.

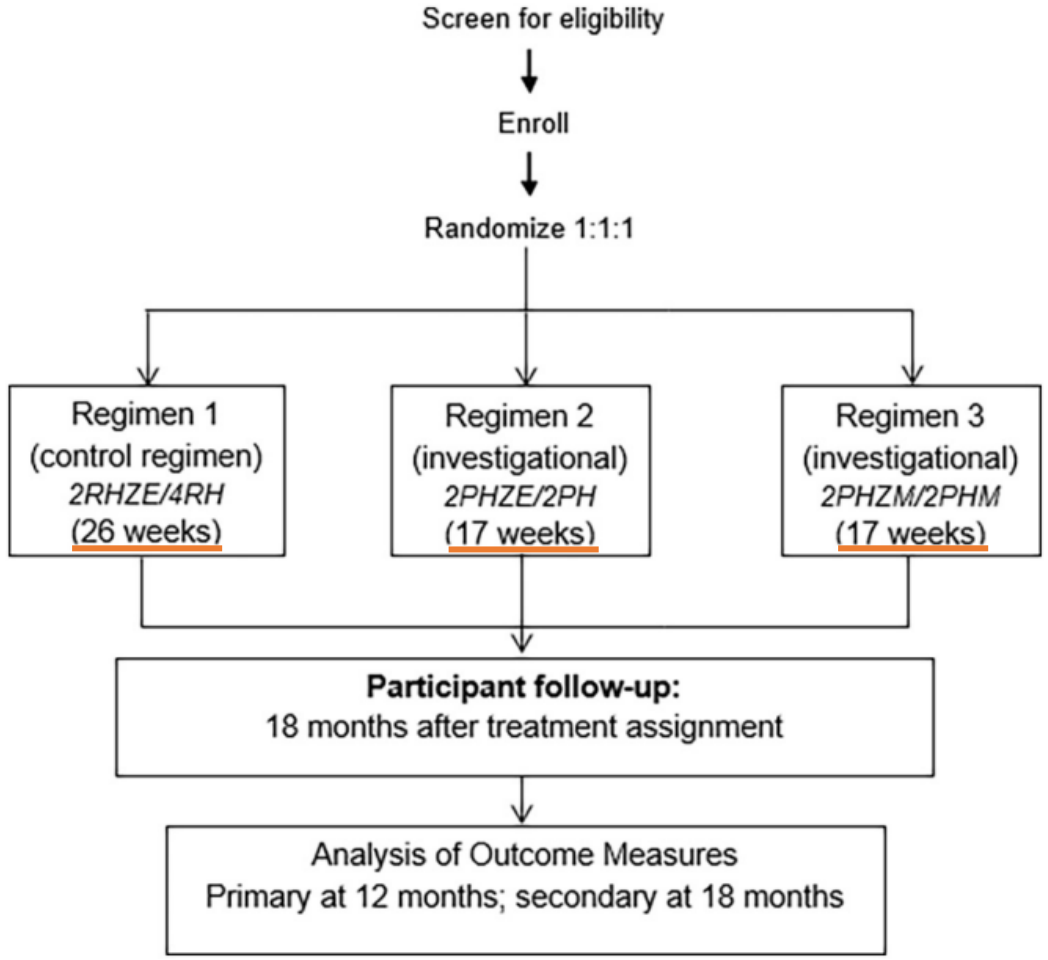


Fig. 1. Study schematic.
Study schematic for the trial "Rifapentine-containing treatment shortening regimens for pulmonary tuberculosis: a randomized, open-label, controlled phase 3 clinical trial (S31/A5349)".
Note. R = rifampin, H = isoniazid, Z = pyrazinamide, E = ethambutol, P = rifapentine, M = moxifloxacin.

Protocol: Dorman SE, Nahid P, Kurbatova EV, et al. High-dose rifapentine with or without moxifloxacin for shortening treatment of pulmonary tuberculosis: Study protocol for TBTC study 31/ACTG A5349 phase 3 clinical trial. *Contemporary Clinical Trials*. 2020;90:105938. doi:10.1016/j.cct.2020.105938

How can we do earlier, more efficient duration-ranging studies?

Duration-response design

Standard: Parallel Arm

Traditionally:

- Randomization to 2+ arms
- Non-inferiority trials

Key questions:

1. How do we pick the shortened durations for comparison?
2. Should we treat different durations of the same treatment as independent?

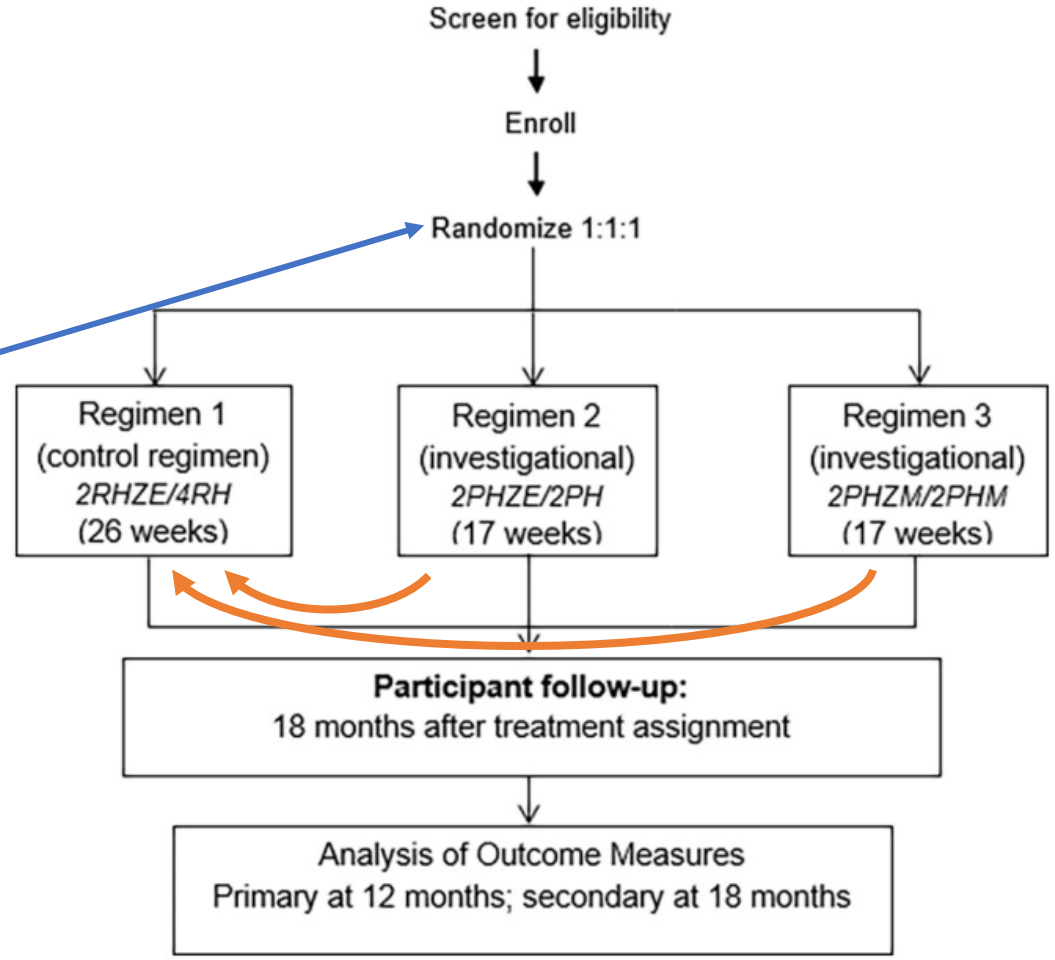
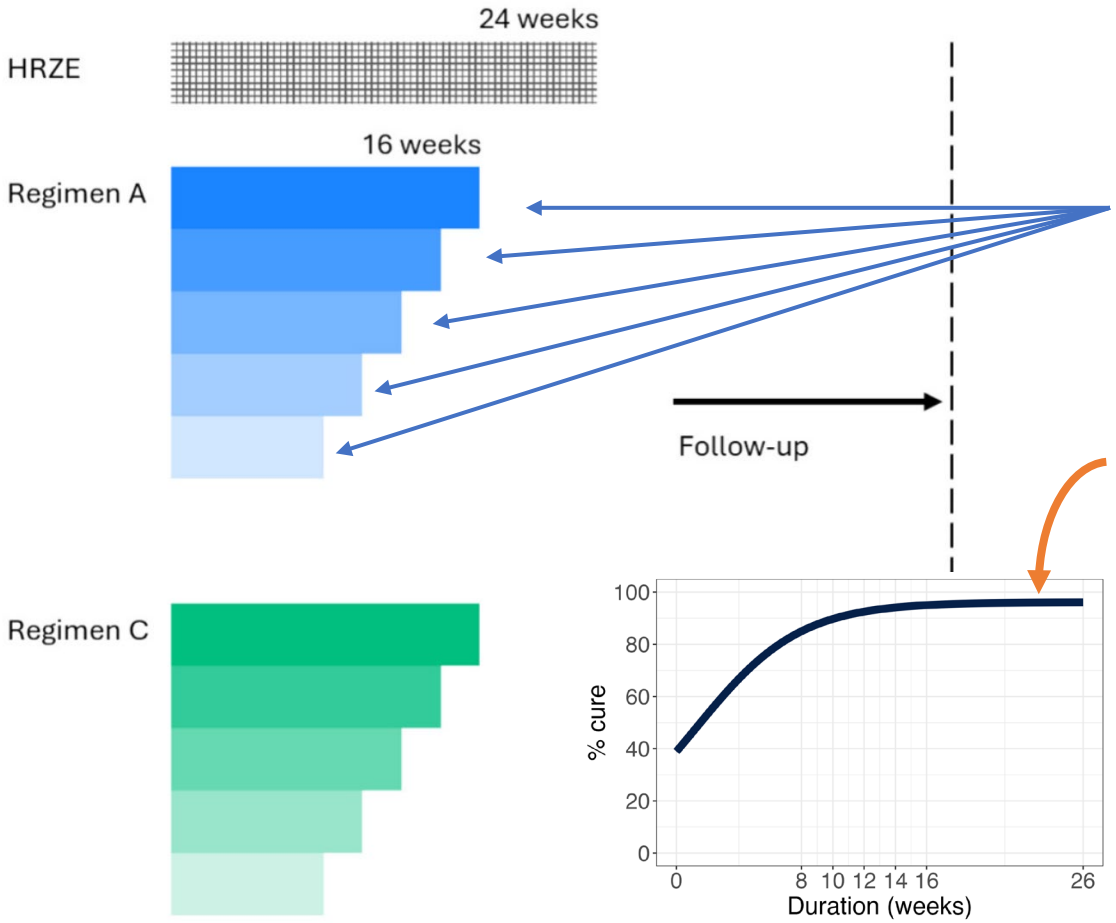


Fig. 1. Study schematic.
Study schematic for the trial "Rifapentine-containing treatment shortening regimens for pulmonary tuberculosis: a randomized, open-label, controlled phase 3 clinical trial (S31/A5349)".
Note. R = rifampin, H = isoniazid, Z = pyrazinamide, E = ethambutol, P = rifapentine, M = moxifloxacin.

Duration-response design



Response over Continuous Intervention (ROCI)

Proposed:

- Randomization to many arms within a reasonable range of treatment durations
- Fit a model across the multiple durations and the outcome response
  - Can still use NI testing

Source: Pham TM, Crook AM, Rolfe K, Phillips PPJ, Dufault SM, Quartagno M. Designing a response-over-continuous-intervention (ROCI) randomised trial: Implementation in the Phase 2C part (duration ranging) of the PARADIGM4TB trial. *Contemporary Clinical Trials*. 2025;155:108002. doi:10.1016/j.cct.2025.108002

Duration-response design

Standard: Parallel Arm

Traditionally:

- Randomization to 2+ arms
- Non-inferiority trials

Questions:

1. How do we pick the shortened durations for comparison?
2. Should we treat different durations of the same treatment as independent?

Response over Continuous Intervention (ROCI)

Proposed:

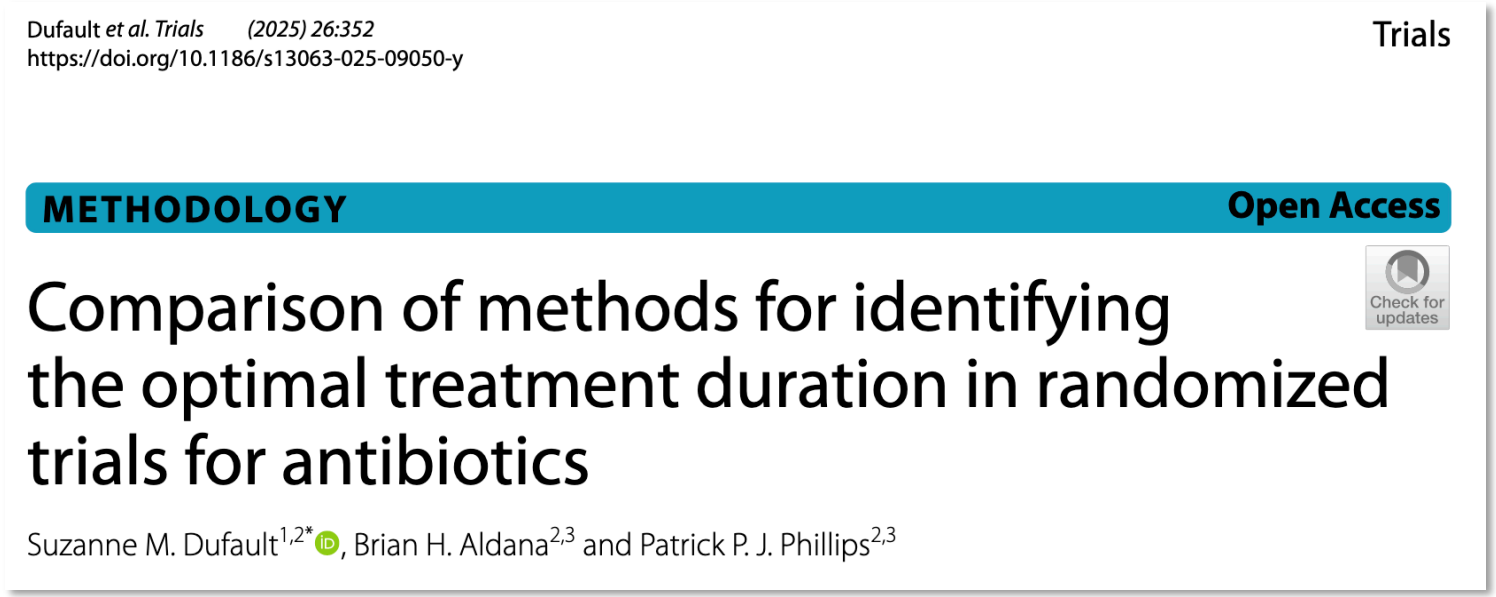
- Randomization to many arms within a reasonable range of treatment durations
- Fit a model across the multiple durations and the outcome response
  - Can still use NI testing

Easy answers!

Case Study 1: Is ROCI more accurate?

Aims

1. Adapt candidate model-based dose-ranging methodologies for the task of duration-ranging.
2. Compare through simulation study model-based duration-ranging methodologies against standard qualitative methods



Dufault *et al. Trials* (2025) 26:352
https://doi.org/10.1186/s13063-025-09050-y

Trials

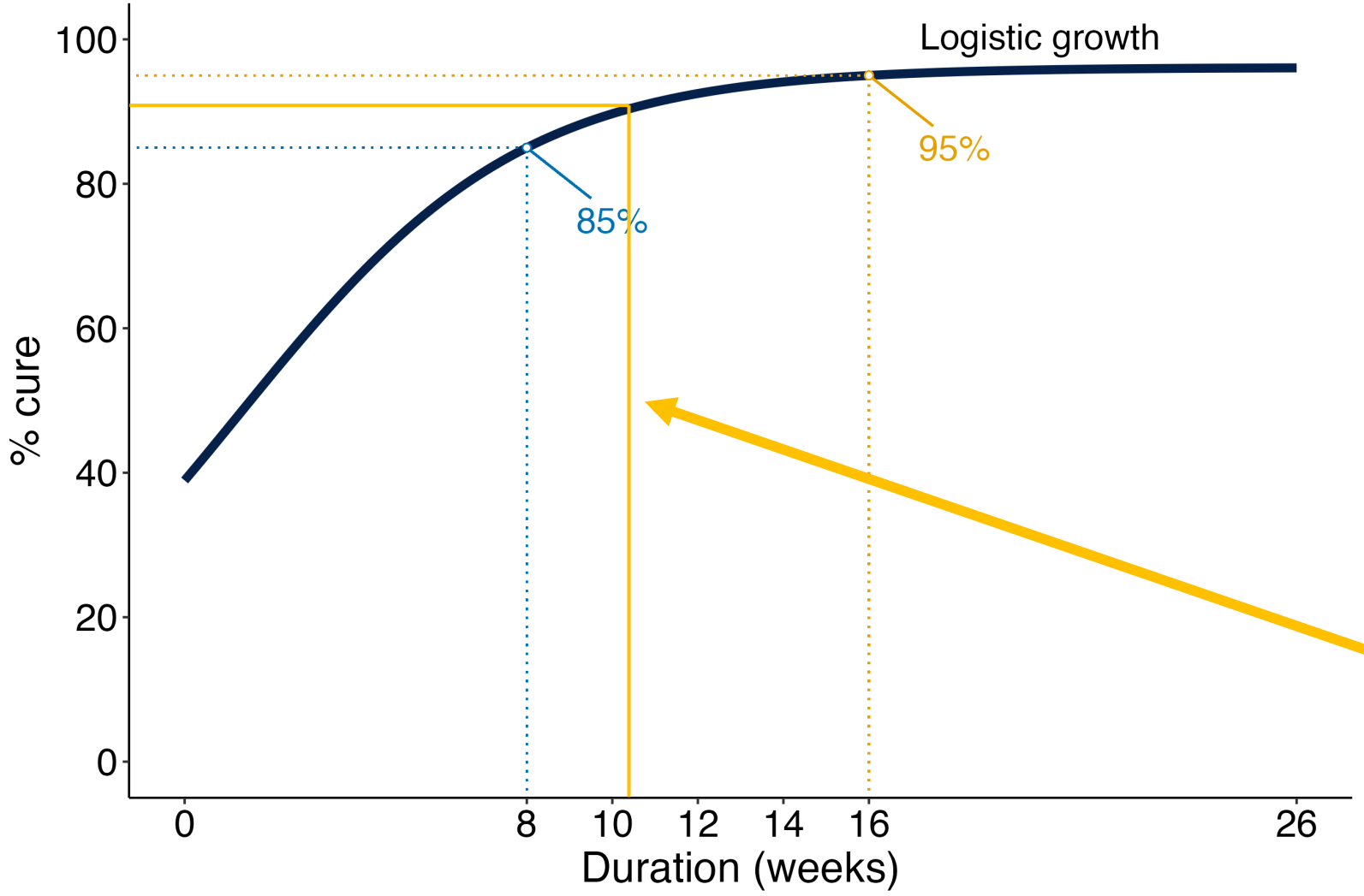
METHODOLOGY Open Access

Comparison of methods for identifying the optimal treatment duration in randomized trials for antibiotics

Suzanne M. Dufault[1,2*], Brian H. Aldana[2,3] and Patrick P. J. Phillips[2,3]

Case Study 1: Is ROCI more accurate?

One illustrative scenario

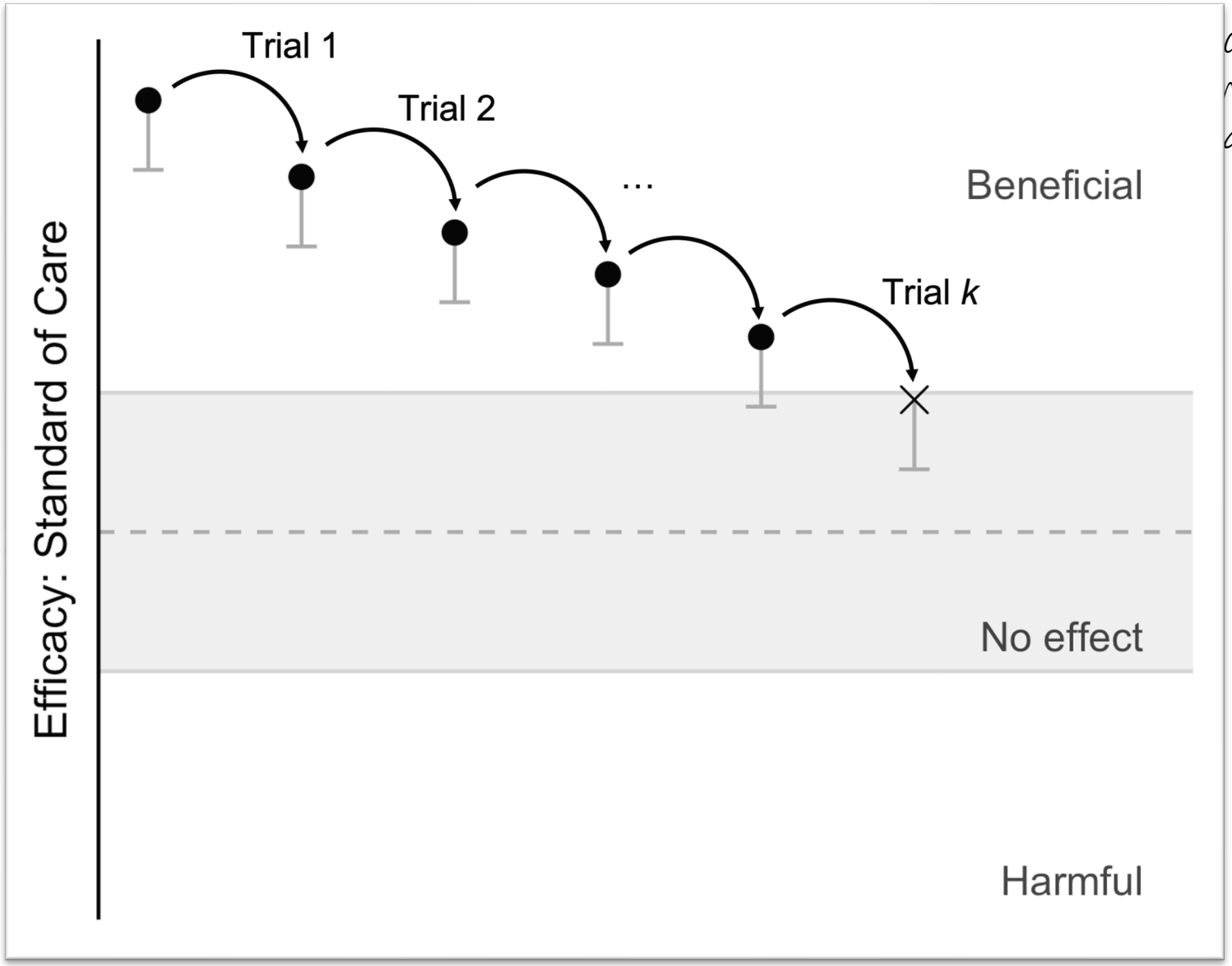


True minimum effective duration?

10.2 weeks, which corresponds to a response rate of 90%.

Note: this work is performed in a Phase II context.

Case Study 1: Is ROCI more accurate?



ccessfully
num effective
appropriate

Adapted methods are far more accurate at estimating the minimum effective duration.

…and estimation avoids the non-inferiority risk of "biocreep"

Case Study 2: Is ROCI more efficient?

Designed a simulation study to **identify the best design** to assess whether any regimen can achieve a similar efficacy to the standard of care HRZE of 26 weeks, without being as long.

Table 1
Design parameters used in simulation.

Design parameter	Level
Control (HRZE) event risks	5 % 10 %
Durations of novel treatment regimens	5 durations: 8, 10, 12, 14, 16 weeks 3 durations: 8, 12, 16 weeks
Sample sizes to be randomised to all durations of novel treatment regimens (as well as correspondong sample sizes per arm, and overall sample sizes for the trial)	$N_{tot\_dur} = 200$ (e.g. for a 5-duration design with 1 control arm, $N_{arm} = 40$ and $N_{overall} = 240$), $N_{tot\_dur} = 300$
Shapes of duration-response curve	Flat: Same event risk as the control arm for all duration arms (5 %, 10 %) Non-flat: same event risk as the control arm for longest duration arms, then event risk increases for shorter durations (Fig. 2, Fig. S1)
Non-inferiority margins	12 % 15 %

Source: Pham TM, Crook AM, Rolfe K, Phillips PPJ, Dufault SM, Quartagno M. Designing a response-over-continuous-intervention (ROCI) randomised trial: Implementation in the Phase 2C part (duration ranging) of the PARADIGM4TB trial. *Contemporary Clinical Trials*. 2025;155:108002. doi:10.1016/j.cct.2025.108002

Case Study 2: Is ROCI more efficient?

Sample size N_{tot_dur}	HRZE control event risk	Non-inferiority margin	Power (%)			
	(%)	(%)	No model	Quadratic	FP1	FP2
200 (40 per arm)	5	12	80.1	84.7	85.7	81.3
	5	15	90.3	95.3	96.9	92.9
	10	12				
	10	15				

4.2. Is modelling the duration-response curve necessary for improving the trial's operating characteristics?

As seen in Table 2 and Fig. 3, power is always higher when using models to estimate the confidence intervals around the risk differences. In some cases the differences are substantial (i.e. up to 10 percentage points) across scenarios when using FP1 or a quadratic model.

Source: Pham TM, Crook AM, Rolfe K, Phillips PPJ, Dufault SM, Quartagno M. Designing a response-over-continuous-intervention (ROCI) randomised trial: Implementation in the Phase 2C part (duration ranging) of the PARADIGM4TB trial. *Contemporary Clinical Trials*. 2025;155:108002. doi:10.1016/j.cct.2025.108002

How can we do earlier, more efficient duration-ranging studies?

Trial Context

Clinical Development

Phase I

Is it safe?
At what dose?

Phase II

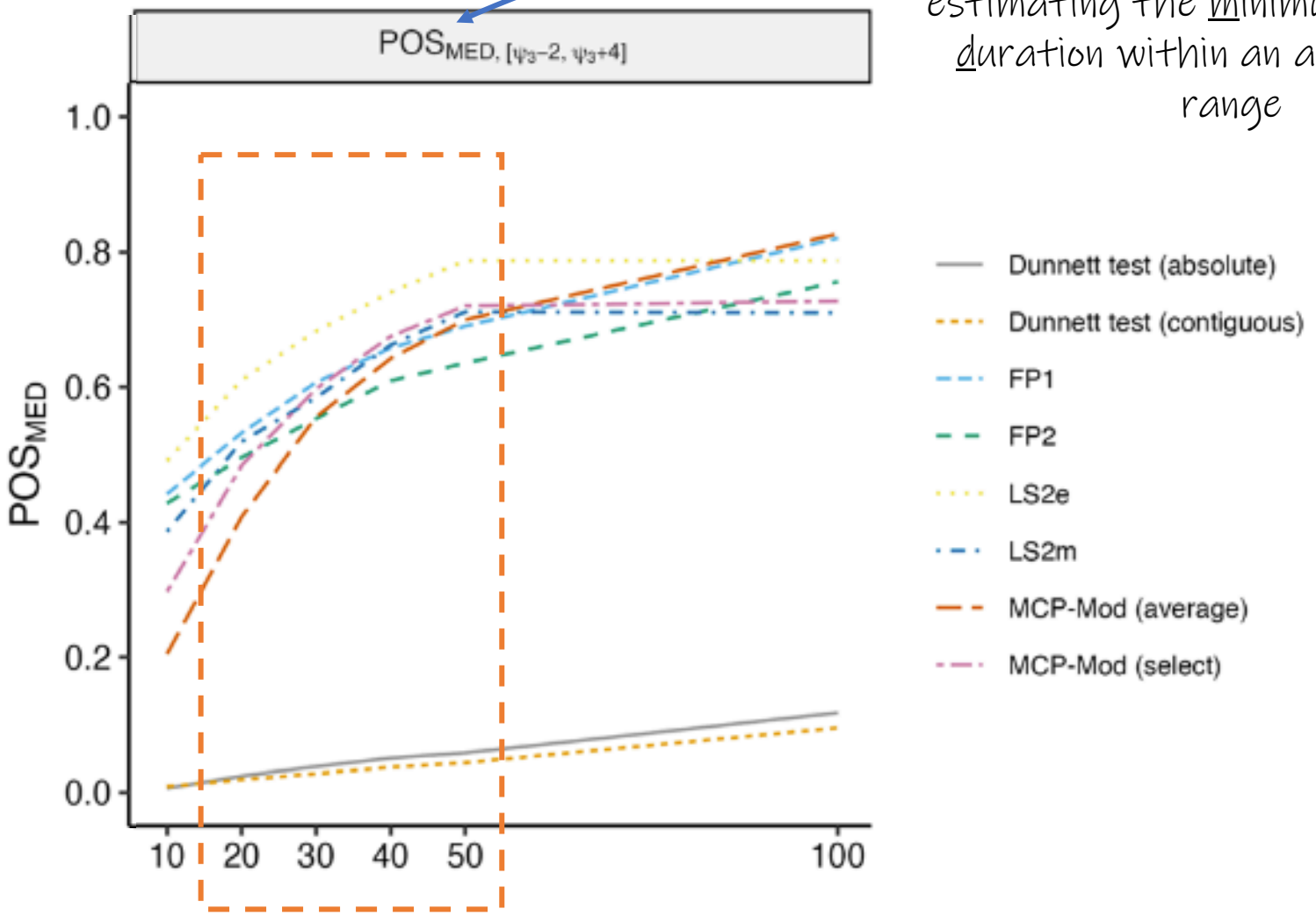
Does it work?

Phase III

How does it compare to standard of care?

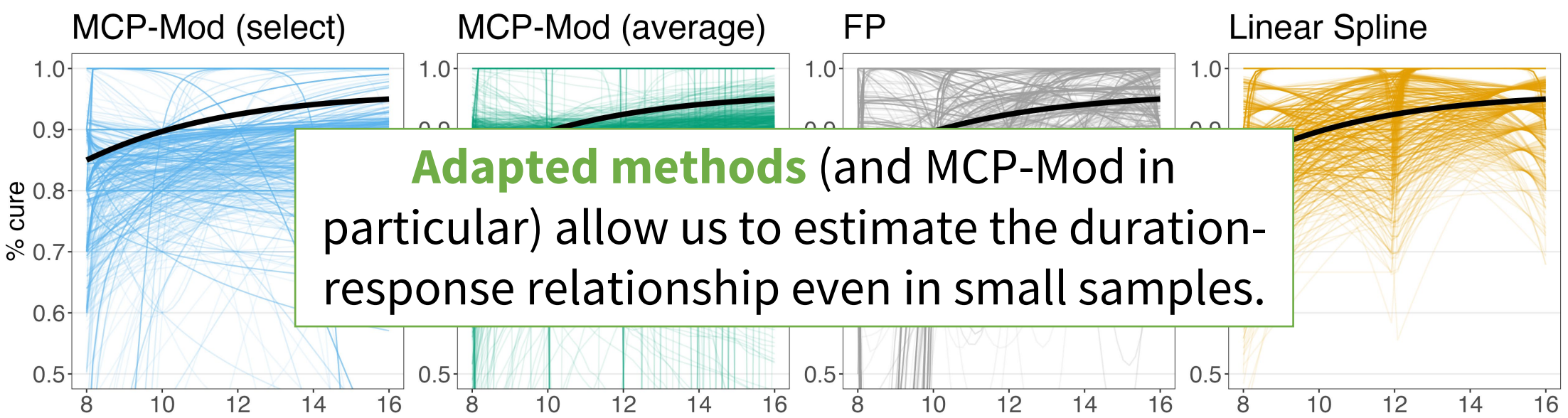
Case Study 2+: Can ROCI be done earlier?

Probability of successfully estimating the minimum effective duration within an appropriate range



What is the duration-response relationship?

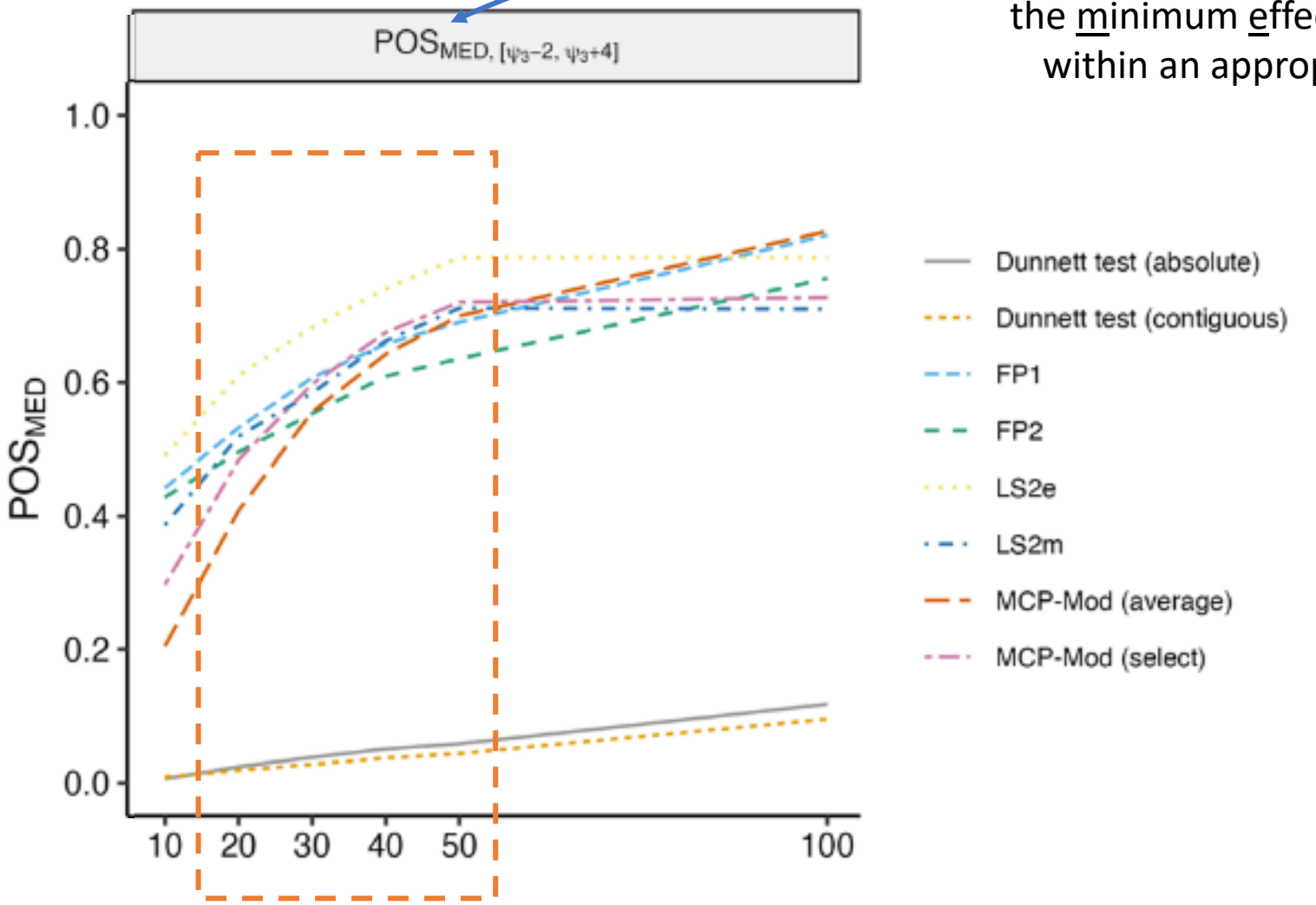
$n_d = 10$



Adapted methods (and MCP-Mod in particular) allow us to estimate the duration-response relationship even in small samples.

Case Study 3: Can ROCI be done earlier?

Probability of successfully estimating the minimum effective duration within an appropriate range



In our new work, we've explored modeling methods to handle small sample size challenges and the handling of intercurrent events.

Ongoing and upcoming duration-response TB trials

Ongoing and Upcoming ROCI Trials

- PARADIGM4TB (NCT06114628)
  - Ongoing
  - Initial Phase 2B regimen selection from 10+ novel multidrug combinations
  - “Winners” proceed to Phase 2C duration optimization step (5 durations each)
- DRAMATIC (NCT03828201)
  - Four arm Phase 2
  - All-oral regimen of bedaquiline, delamanid, levofloxacin, linezolid, and clofazimine for treatment of MDR-TB participants
  - Randomized between durations of 16, 24, 32, or 40 weeks
- SPECTRA TB (ACTG A5414)
  - Stratified medicine approach to shortening isoniazid, rifapentine, moxifloxacin, and pyrazinamide for DS-TB

"[S]election of dose for phase III is an estimation problem and should not be addressed via hypothesis testing"

Source: EMA/EFPIA workshop on dose-finding (2014)

Summary

Thank you

Thank you to the session chairs for the opportunity to present in this SCT symposium.

I want to acknowledge the contribution of all individuals who have provided input to this work. I gratefully acknowledge funding from the UCSF Center for TB and TB RAMP scholar program (NIH/NIAID R25AI147375).

Contact Info
suzanne.dufault@ucsf.edu
suzanne-dufault-phd.netlify.com

Work performed in collaboration with:

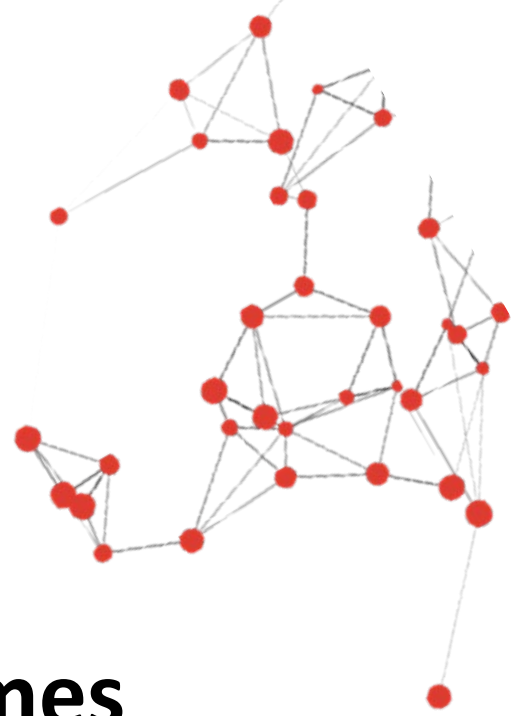
Patrick P. J. Phillips, Brian Aldana (UCSF)

Tra My Pham, Matteo Quartagno, Angela Crook (UCL)

Linda Harrison (Harvard SDAC)

Katie Rolfe (GSK) … and many more!

Please contact me with any questions, comments, or feedback.



Superiority trials with Hierarchical Outcomes

Johan Verbeeck PhD

johan.verbeeck@uhasselt.be

Data Science Institute

UHasselt - Belgium

DSI DATA SCIENCE INSTITUTE

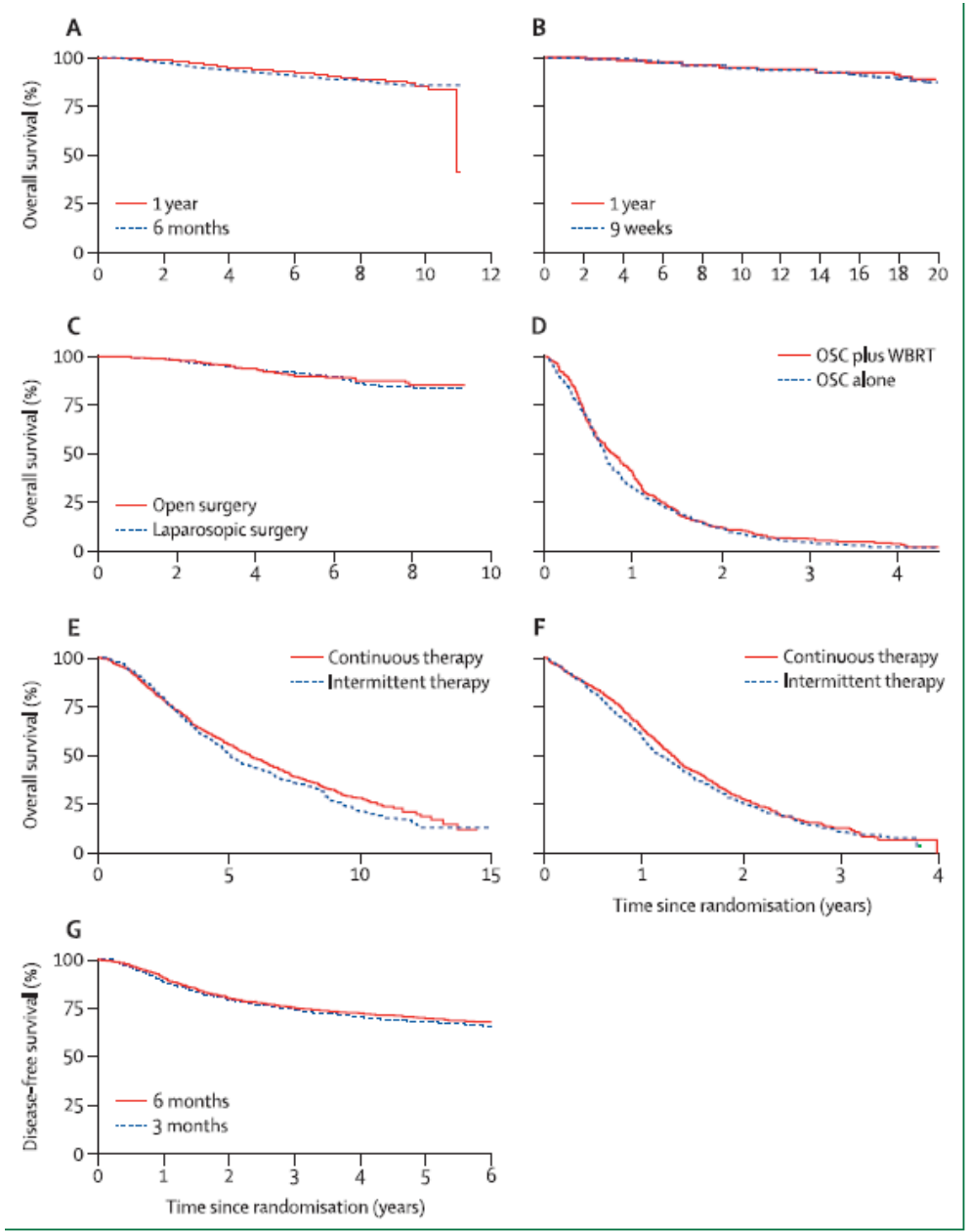
UHASSELT

WWW.UHASSELT.BE/DSI

Motivation: Dose-reduction trials

- Conventional dose-finding trial designs in **oncology** are based on the concept of **maximum tolerated dose**
- May identify a dose or treatment schedule that is **more intense** than the **most effective dose**
- Resulting in **unnecessary toxicity**, often due to off-target effects
- Alternative dose and schedules are **limited** by the perception that they must be **non-inferior** to approved regimens

Non-inferiority trials are large, expensive and with many flaws



Fail to show non-inferiority!

- Due to inadequately justified and overly **large non-inferiority margins**, which oftentimes exceed benefits observed in superiority trials
- But chosen to reduce the required sample size

Tannock et al. *Lancet Oncology* (2024)

Additional issues

- Non-inferiority trials do not answer the most important clinical question:

  "**Under which therapy are patients are better off**"

- Non-inferiority designs aim to show that a novel therapy is **not worse** than a standard of care by more than a pre-specified non-inferiority margin on an **efficacy outcome**.

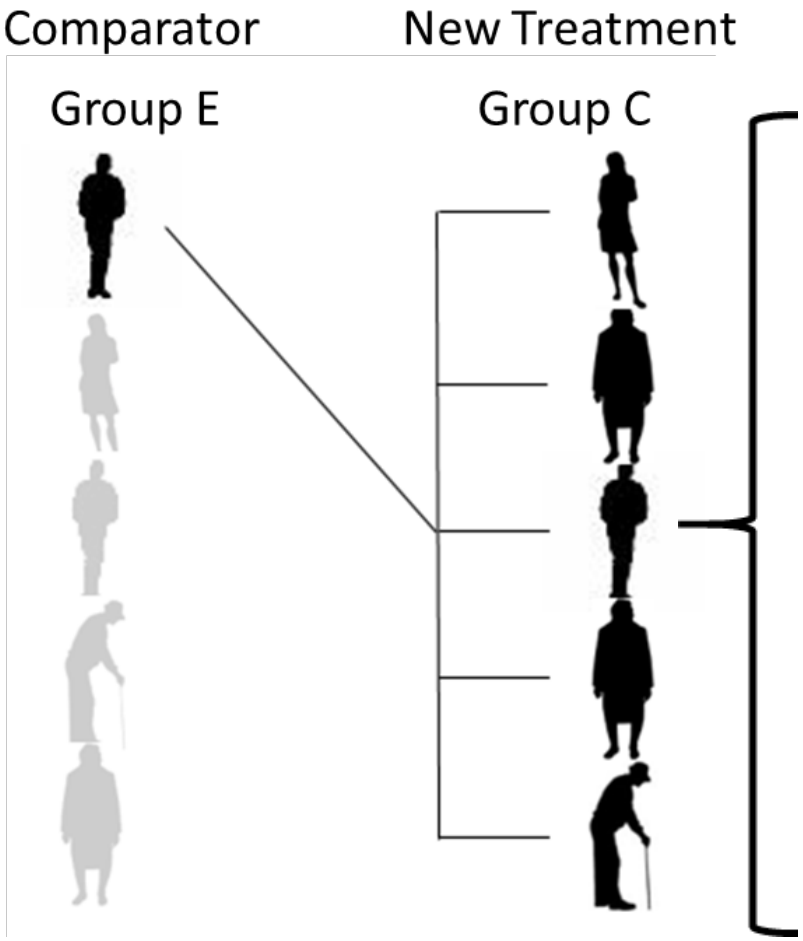
  - The non-inferiority margin = acceptable loss on efficacy
  - **Justified by a putative advantage**: improved safety, better quality of life, more convenient administration, lower cost,...

Alternative for non-inferiority trial

- Given that a definite **advantage** from the novel therapy is **expected**
- **Balance efficacy and benefits** of the novel therapy as compared with a standard of care into a **patient-centric superiority analysis** of several outcomes
- Showing a significant net benefit of the novel therapy would be at once **easier from a statistical point of view** and **more relevant from a clinical point of view**

Generalized pairwise comparisons (GPC) allow for superiority designs of multiple outcomes prioritized in a patient-centric way

Generalized Pairwise Comparisons (GPC)



Highest ranked outcome (Efficiency)

tie

Middle ranked outcome (Benefit)

tie

Lowest ranked outcome (Benefit)

Note: priorities may be patient-centric

1. Perform pairwise comparisons between all elements of Y^E and Y^C

2. Calculate $U_{ij} = \begin{cases} 1 & if\ Y_i^E \succ Y_j^C \\ -1 & if\ Y_i^E \prec Y_j^C \\ 0 & if\ Y_i^E \asymp Y_j^C \end{cases}$

3. The statistic $\widehat{\Delta} = \frac{1}{n^E n^C} \sum_{i=1}^{n^E} \sum_{j=1}^{n^C} U_{ij}$ has a known distribution under H_0

Buyse. Stat Med (2010)
Pocock et al. Eur Heart J (2012)
Buyse et al. Chapmann & Hall/CRC (2025)

GPC – threshold of clinical similarity

Quantify acceptable changes in efficacy outcomes in relation to the provided benefits.

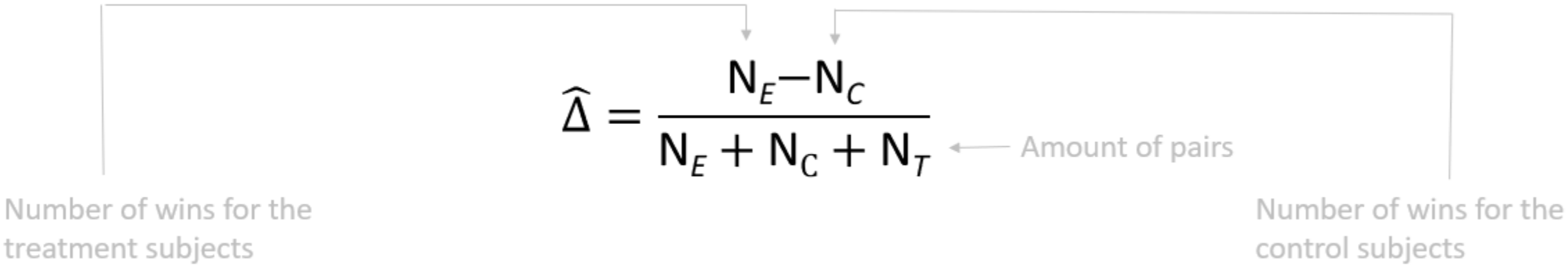
1. Perform pairwise comparisons between all elements of Y^E and Y^C

2. Calculate $U_{ij} = \begin{cases} 1 & if\ Y_i^E > Y_j^C + \tau \\ -1 & if\ Y_i^E < Y_j^C + \tau \\ 0 & otherwise \end{cases}$

3. The statistic $\widehat{\Delta}_\tau = \frac{1}{n^E n^C} \sum_{i=1}^{n^E} \sum_{j=1}^{n^C} U_{ij}$ has a known distribution under H_0

Buyse. *Stat Med* (2010)
Buyse et al. *Chapmann & Hall/CRC* (2025)

Default GPC statistic



Net treatment benefit (NTB)

NTB ranges from -1 to +1, with 0 indicating no overall treatment effect

NTB is the **net probability** of a better outcome in one treatment group than in the other

$\Delta = P(Y_E > Y_C) - P(Y_E < Y_C)$

Buyse. Stat Med (2010)
Hoeffding. Ann Math Stat (1948)

Alternative GPC statistics

$$\widehat{\Delta} = \frac{N_E - N_C}{N_E + N_C + N_T}$$ ← Amount of pairs

Number of wins for the treatment subjects

Number of wins for the control subjects

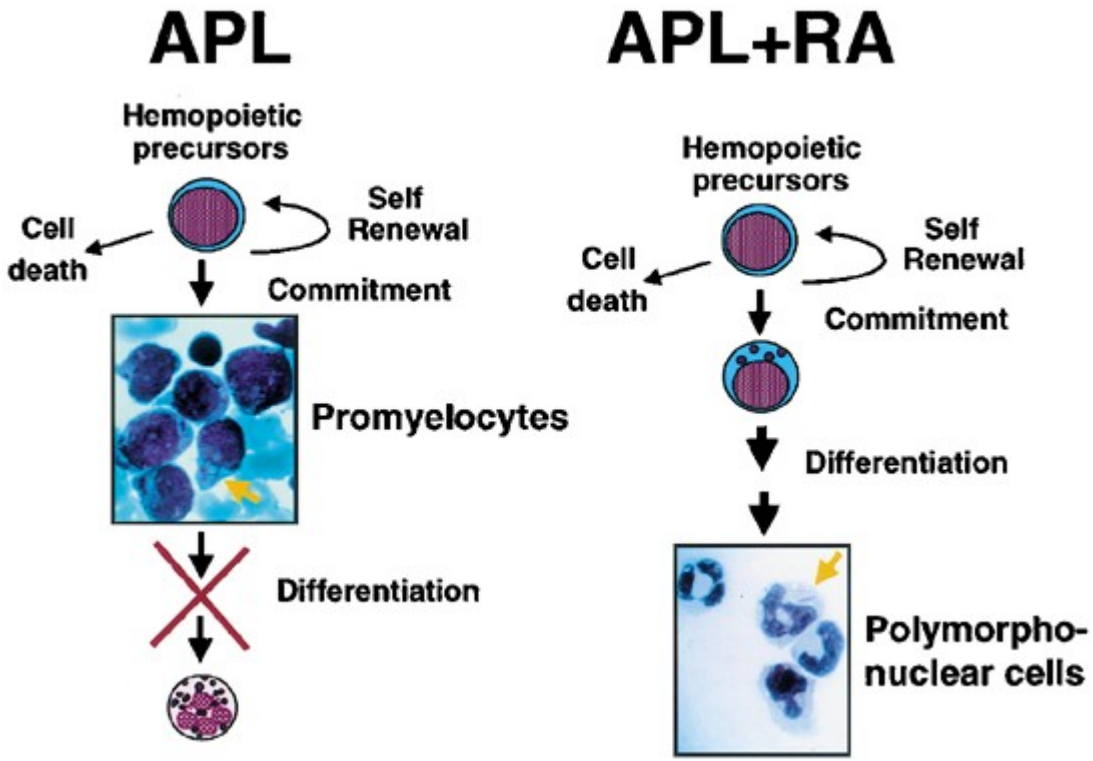
$$\text{Success Odds} = \frac{N_E + {}^1/_2 N_T}{N_C + {}^1/_2 N_T}$$ ← Number of ties

$$\text{Win Ratio} = \frac{N_E}{N_C}$$

Note : NTB = (SO-1)/(SO+1)

Buyse. Stat Med (2010)
Pocock et al. Eur Heart J (2012)
Dong et al. Stat Biopharm Res (2019)
Brunner et al. Stat Med (2021)

Example: Acute promyelocytic leukemia



- Standard of care: high dose of ATRA (all-trans retinoic acid)
- Toxicity of ATRA remains a problem
- Real-world data suggest that a reduced dose may provide similar efficacy and less toxicities

“Does benefit-risk balance favor reduced dose?”

De Backer et al. *Clinical Trials* (2024)
Buyse et al. Chapmann & Hall/CRC (2025)

Design

Traditionally :

- **Non-inferiority** on Event-Free Survival (EFS) at 2 years

  -> Expensive: large sample size required to meet a ‘narrow’ non-inferiority margin

  -> Inefficient: key **toxicity outcomes** are relegated to **secondary analyses**

  Does not answer the clinical question: “which dose patients are better off”

GPC:

Superiority trial with prioritized outcomes:

1. EFS at 2 years of follow-up (alive and disease-free vs not)
2. Grade 3/4 documented infections (no vs yes)
3. Grade 3/4 differentiation syndrome (no vs yes)
4. Grade 3/4 hepatotoxicity (no vs yes)
5. Grade 3/4 neuropathy (no vs yes)

Gains on toxicities are acceptable ***only if*** EFS results are similar.

De Backer, Clin Trials (2024)

Trial design for a benefit-risk question

GPC accumulates evidence across all risks and benefits
→ smaller studies than to show non-inferiority on efficiency

Power	Traditional design (Non-Inferiority)	GPC design (Superiority)
80%	~750 patients	~310 patients
90%	~1010 patients	~410 patients

De Backer, Clin Trials 2024

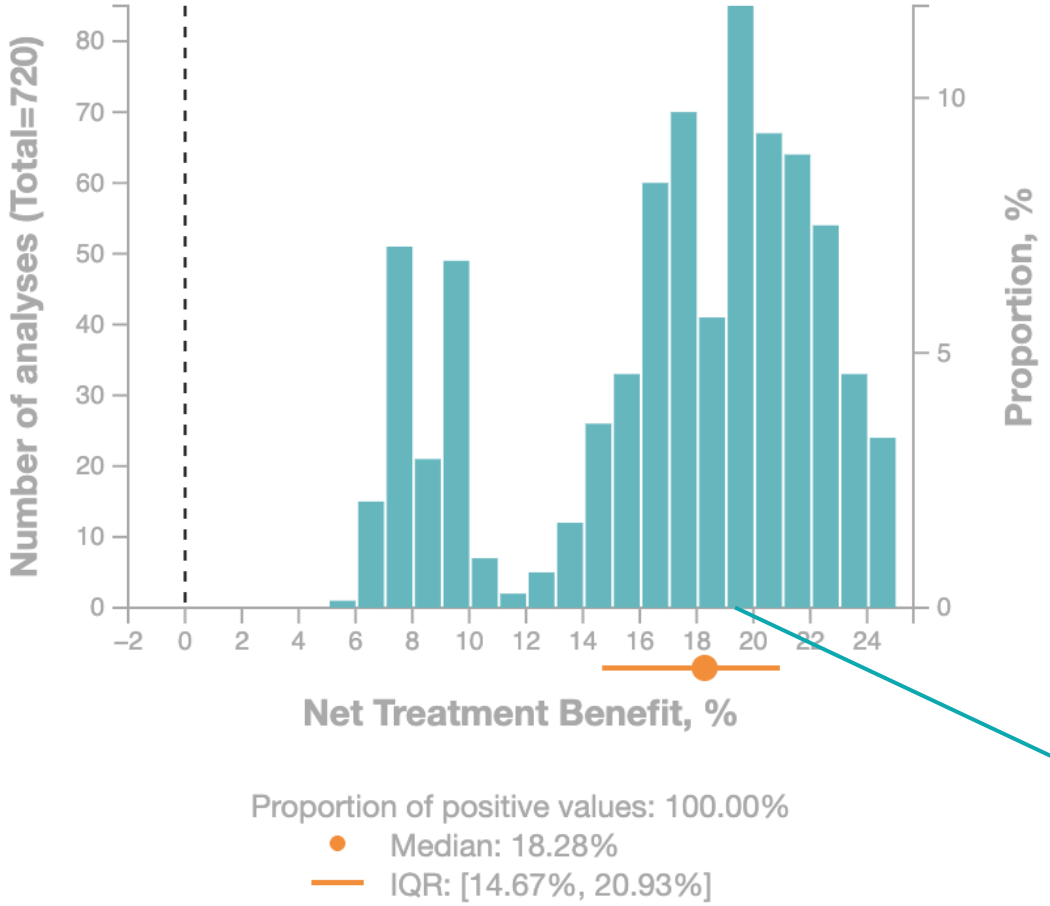
GPC superiority benefit-risk trial

Outcome	Pairs	Favor less intensive regimen (wins)	Favor neither (ties)	Favor standard of care (losses)	Contribution to NTB	NTB	*P* value (median)
Alive and event-free at 2 years	19,600	0.075	0.809	0.116	−0.040	−0.040	0.276
Documented infections	15,852	0.284	0.410	0.114	0.170	0.129	0.040
Differentiation syndrome	8037	0.057	0.327	0.025	0.032	0.161	0.013
Hepatotoxicity	6417	0.029	0.283	0.015	0.014	0.175	0.008
Neuropathy	5553	0.023	0.250	0.011	0.012	0.186	0.005
All outcomes	19,600	0.468	0.250	0.282	0.186	0.186	0.005

De Backer, Clin Trials (2024)

Sensitivity analysis: permutation of outcomes

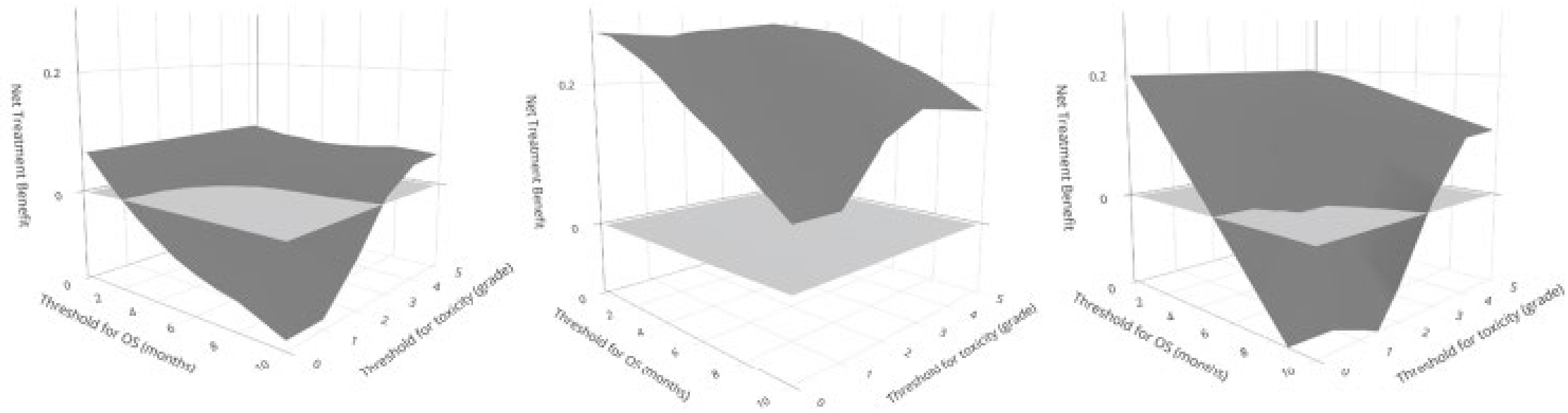
All possible permutations of the selected outcomes



- Allows to include efficacy and safety outcomes
- Allows to start with safety outcomes in some analyses
- Allows to understand the overall medical value of a treatment

NTB always favors the low dose

Sensitivity analysis: varying thresholds



Buyse et al. Chapmann & Hall/CRC (2025)

Conclusions

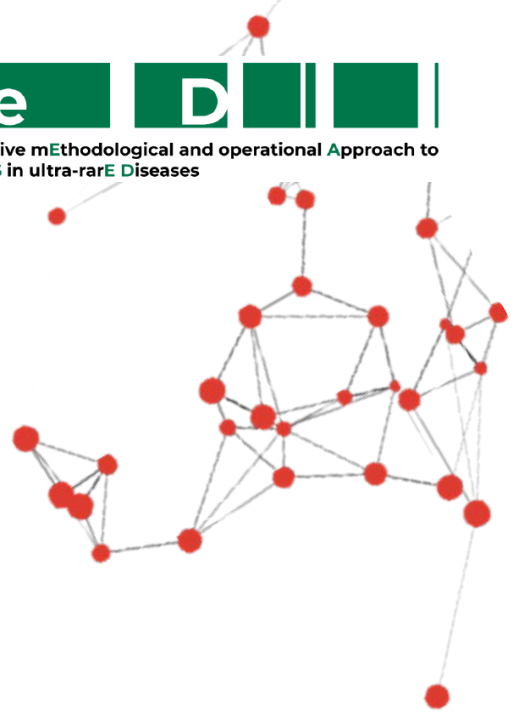
- **Patients** prioritize therapies that are **superior** in terms of safety, convenience, and cost
- Need for **superiority comparative trials** over non-inferiority designs
- **GPC** allows for these principled patient-centric assessments by **hierarchical endpoints**, balancing efficiency and benefits
- **Threshold of clinical relevance** quantify acceptable changes in efficacy outcomes in relation to the provided benefits
- **More relevant from a clinical point of view, easier from a statistical point of view and more efficient**

A ‘mature’ framework that can handle many of the usual complications

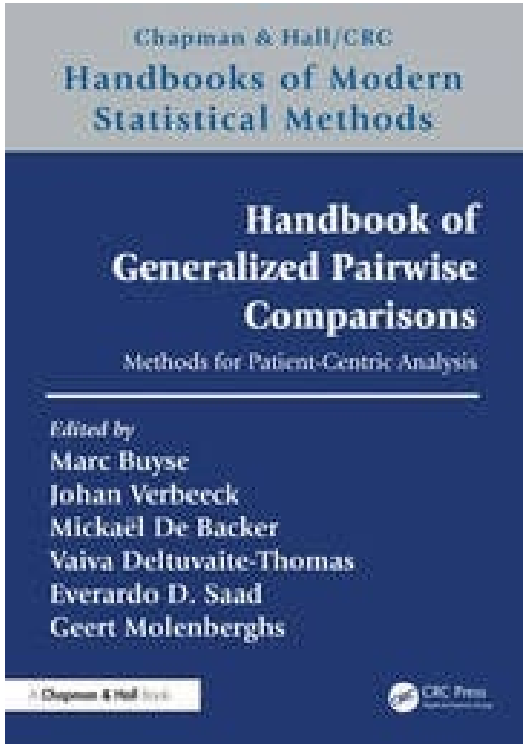
- right-censoring, competing risks, covariates, multiple testing
- The  package BuyseTest attempts to provide a convenient & transparent interface

ERDERA
EUROPEAN RARE DISEASES RESEARCH ALLIANCE

Realise D
compRehensive mEthodological and operational Approach to cLinical trIalS in ultra-rarE Diseases



Questions ?



Johan Verbeeck PhD
johan.verbeeck@uhasselt.be
Data Science Institute
UHasselt - Belgium

GPC statistics - differences

	Wins N_E (%)	Losses N_C (%)	Ties N_T (%)	NTB $\frac{N_E - N_C}{N_E + N_C + N_T}$	SO $\frac{N_E + 0.5N_T}{N_C + 0.5N_T}$	WR $\frac{N_E}{N_C}$
Trial 1	3 (0.06%)	1 (0.02%)	4,996 (99.92%)	0.0004	1.0008	3.00
Trial 2	3,000 (60%)	1,000 (20%)	1,000 (20%)	0.40	2.33	3.00

The WR ignores the ties or redistributes the ties according to the observed win/loss proportions -> overestimation of effect

Consequences of redistributing ties by WR

Continuous

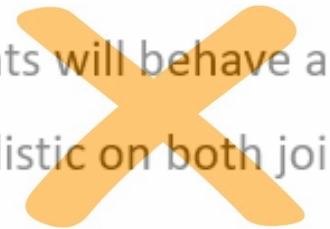
- f.e. relative change in NT-proBNP (PARACHUTE-HF)
- Chance of a tie is negligible; unless you use a threshold

Discrete

- f.e. count (# of hospitalisations), categorical (Yes/No, QoL, 6MWT improvement,..)
- Chance of a tie is very high
- WR redistributes ties according to win proportions of not-tied pairs

Survival

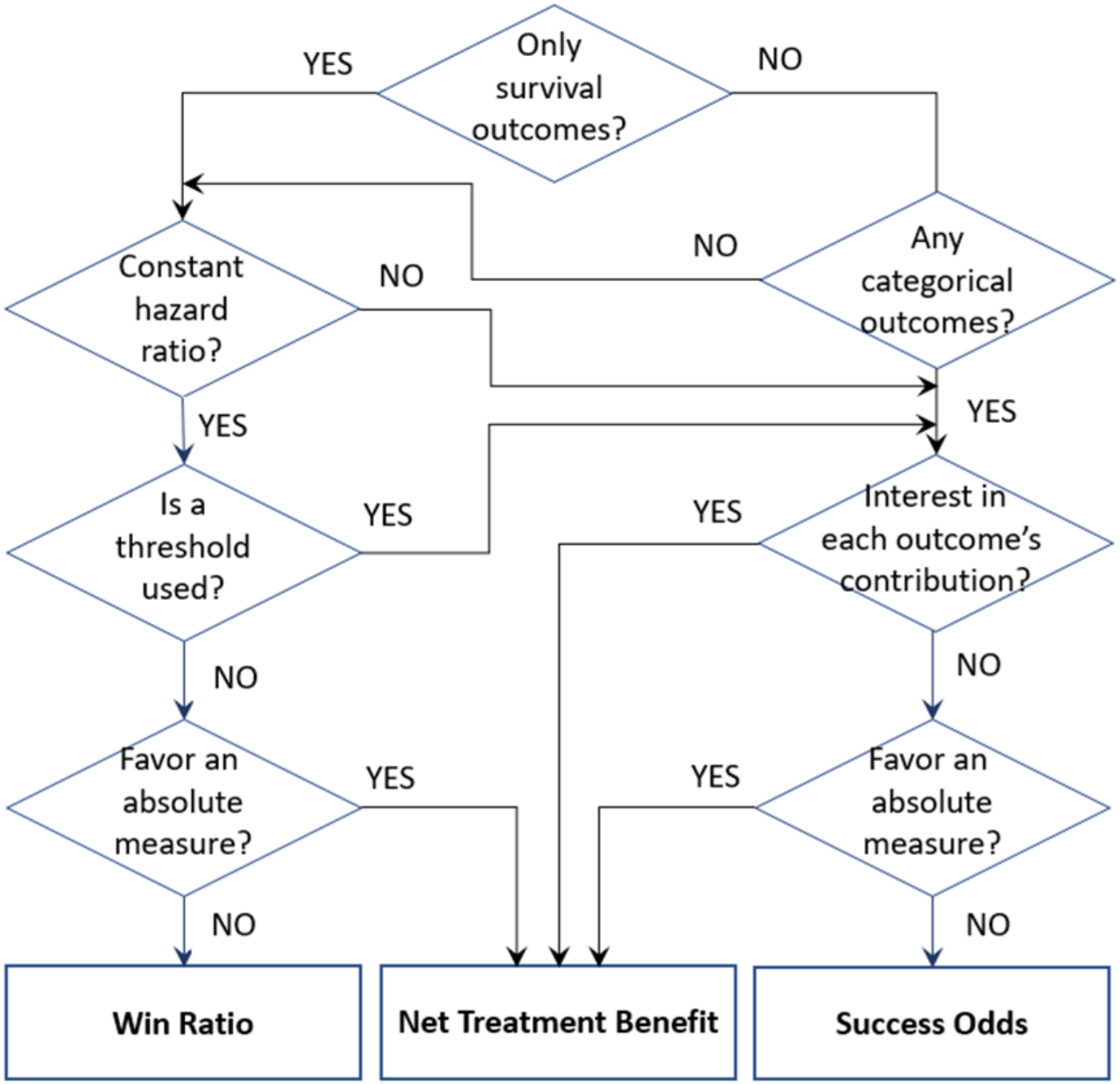
- f.e. equal time to death or censored death time
- Chance of a tie due to censored event time is very common; due to equal event time uncommon
- WR assumes that censored events will behave as observed events = ok under proportional hazards, but unrealistic on both joint and individual survival outcome



Additive decomposition of NTB

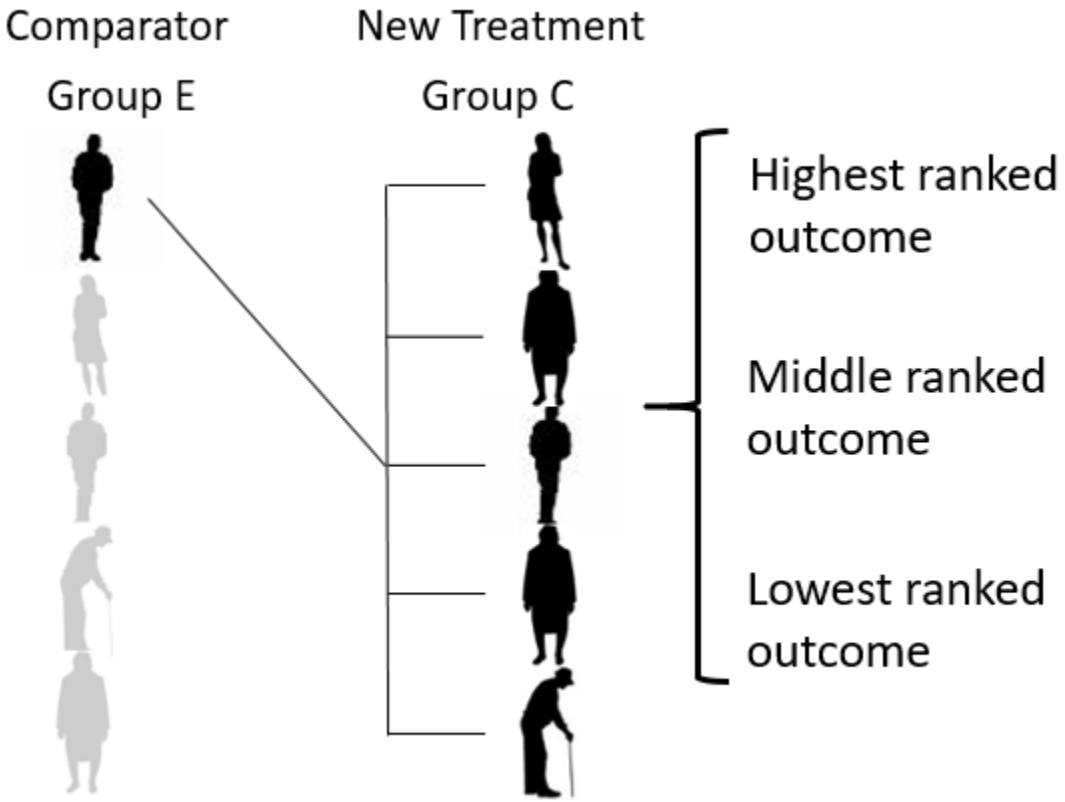
365 days	% wins	% losses	% ties	NTB (95%CI)	SO (95%CI)	WR (95%CI)
Death	4.31	3.62	92.07	0.0069	1.01	1.19
Hemor. Stroke	0.05	0.07	91.95	-0.0002	1.00	0.67
Isch. Stroke	0.41	0.29	91.25	0.0011	1.00	1.39
MI	9.70	8.90	72.65	0.0080	1.02	1.09
Total MACE	14.47	12.88	72.65	0.016 (0.000-0.031)	1.03 (1.00-1.06)	1.12 (1.00-1.26)
p-value MACE				0.0413	0.0413	0.0414

Guidance of GPC measures in clinical trials



Verbeeck et al. JACC (2023)

GPC – multiple weighted outcomes



1. Perform pairwise comparisons between all elements of Y^E and Y^C for each of the k =(1,..,d) outcomes

2. Calculate $U_{ij}(k) = \begin{cases} 1 & if\ Y_i^E > Y_j^C \\ -1 & if\ Y_i^E < Y_j^C \\ 0 & if\ Y_i^E = Y_j^C \end{cases}$

3. The statistic $\widehat{\Delta}^O = \frac{1}{n^E n^C} \sum_{i=1}^{n^E} \sum_{j=1}^{n^C} \sum_{k=1}^{d} w(k)\, U_{ij}(k)$ has a known distribution under H_0

Note: weights $w(k)$ are arbitrary, usually chosen so that $\sum_{k=1}^{d} w(k) = 1$

O'Brien. Biometrics (1984)

GPC – Links with conventional effect size measures for univariate outcomes

Binary endpoint (denote success by 1 and failure by 0)

$$\Delta = P_E - P_C \qquad \text{WR} = \frac{P_E/(1-P_E)}{P_C/(1-P_C)}$$

Continuous endpoint

$\Delta = 2\Phi\left(\frac{d}{\sqrt{2}}\right) - 1,$ with d = Cohen's d

Survival endpoint (denote $\delta^E = \delta^C = 0$ as censored and $\delta^E = \delta^C = 1$ as observed event)

$$U_{ij} = \begin{cases} 1, & \text{if } Y_i^E > Y_j^C, \text{ and } \delta_j^C = 1 \\ -1, & \text{if } Y_i^E < Y_j^C, \text{ and } \delta_i^E = 1 \\ 0, & \text{if } Y_i^E = Y_j^C, \text{ and } \delta_i^E = \delta_j^C = 1 \\ 0, & \text{otherwise.} \end{cases} \qquad \Delta = \frac{1-HR}{1+HR} \qquad \text{WR} = \frac{1}{HR}$$

Gehan. Biometrika (1965)

Limitations of Non-Inferiority Designs

The central problem

"Non-inferior" means "not worse than standard care by more than an acceptable amount" — but defining and interpreting that acceptable amount is often the hardest part.

1. The margin is hard to justify

The non-inferiority margin can feel arbitrary: too wide makes weak treatments look acceptable; too narrow may make useful alternatives fail.

2. Effect preservation

The design assumes the active control would still beat placebo today and that enough of its historical effect is preserved — assumptions that cannot be directly verified.

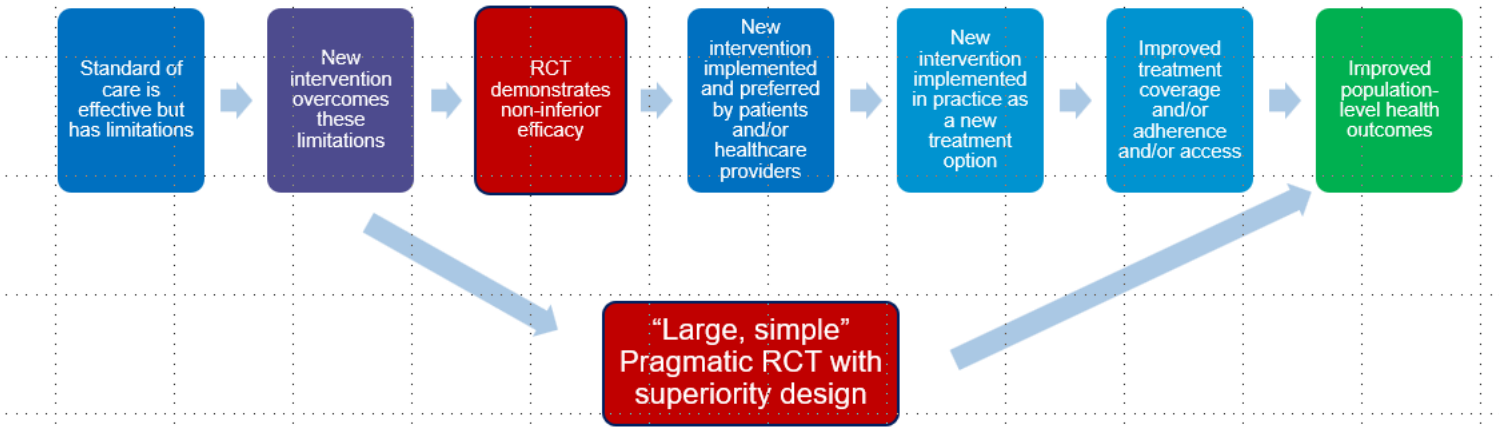
3. Low event rates can distort inference

When few outcome events occur, estimates become unstable and conclusions may depend heavily on the control-arm event rate rather than the true treatment effect.

4. The logic feels backwards

Researchers are asked to rule out "too much worse," instead of showing that the new intervention is better on outcomes patients actually care about.

The conceptual paradigm for the NI design



Pragmatic trials with a superiority design are one solution.

How pragmatic?

What data are needed before launching a pragmatic trial?

--dosing, duration?

--phase II activity data?

Risk of getting an answer to the wrong question if knowledge about treatment is not adequate.

Reframing the trial design – a good alternative

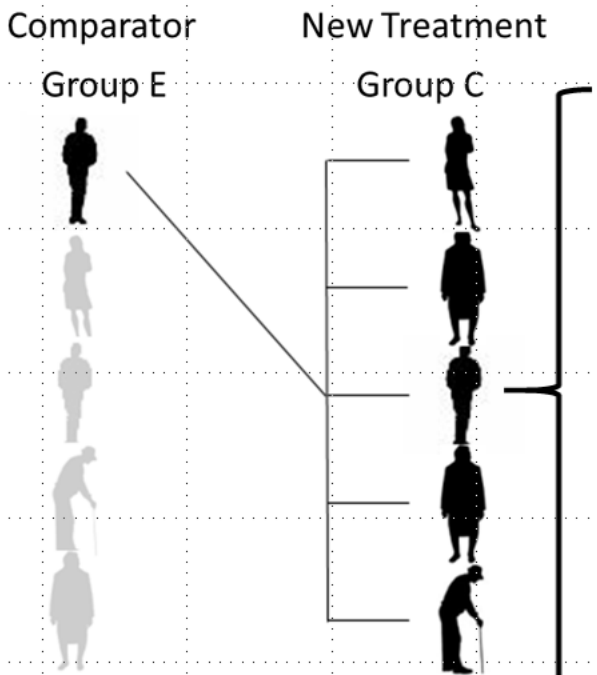
- Blister trial: **strategy superiority trial**:
  - Participants randomized to
    1. First-line therapy with doxycycline with steroids as rescue medication if no clearance of rash
    2. Steroid as first-line therapy.
- Primary outcome could be rash-free survival at 3 months.

TB Prevention: **superiority cluster-randomized trial**:

- Randomized districts:
  1. Treatment with new regimen implemented into health systems
  2. Treatment with standard of care existing approach
- Primary outcome (idealized):
  - District-level TB incidence over 5-10 years.

Are these better questions?
Are sample sizes feasible?

Generalized Pairwise Comparisons (GPC)



Highest ranked outcome (Efficiency)

tie

Middle ranked outcome (Benefit)

tie

Lowest ranked outcome (Benefit)

Note: priorities may be patient-centric

1. Perform pairwise comparisons between all elements of Y^E and Y^C

2. Calculate $U_{ij} = \begin{cases} 1 & if\ Y_i^E \succ Y_j^C \\ -1 & if\ Y_i^E \prec Y_j^C \\ 0 & if\ Y_i^E \asymp Y_j^C \end{cases}$

3. The statistic $\widehat{\Delta} = \frac{1}{n^E n^C} \sum_{i=1}^{n^E} \sum_{j=1}^{n^C} U_{ij}$ has a known distribution under H_0

Buyse. Stat Med (2010)
Pocock et al. Eur Heart J (2012)
Buyse et al. Chapmann & Hall/CRC (2025)

GPC:

Superiority trial with prioritized outcomes:

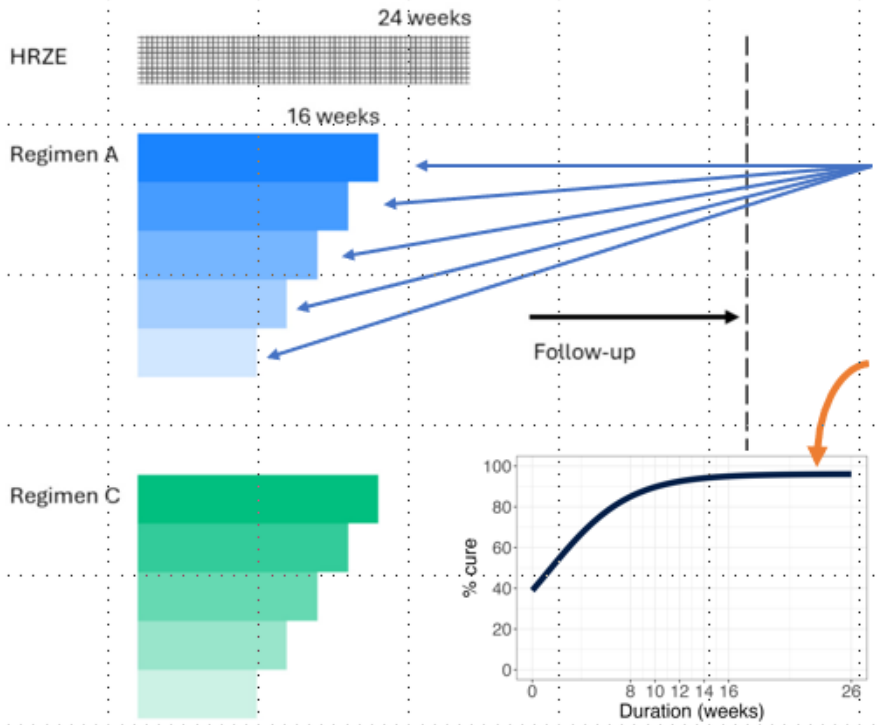
1. EFS at 2 years of follow-up (alive and disease-free vs not)
2. Grade 3/4 documented infections (no vs yes)
3. Grade 3/4 differentiation syndrome (no vs yes)
4. Grade 3/4 hepatotoxicity (no vs yes)
5. Grade 3/4 neuropathy (no vs yes)

Gains on toxicities are acceptable ***only if*** EFS results are similar.

De Backer, Clin Trials (2024) 11

- **Hard to explain intuitively:** The estimand can be less familiar than a risk difference, rate ratio, or mean difference.
- **Depends on outcome hierarchy:** Results may change depending on which outcomes are prioritized first.
- **Can hide tradeoffs:** A favorable net benefit may mask harm on an important safety outcome or modest benefit across many minor outcomes.
- **Subjective clinical judgments:** Weights, thresholds, and definitions of “win,” “loss,” or “tie” may be debatable.

Duration-response design



Response over Continuous Intervention (ROCI)

Proposed:

- Randomization to many arms within a reasonable range of treatment durations
- Fit a model across the multiple durations and the outcome response
  - Can still use NI testing

Source: Pham TM, Crook AM, Rolfe K, Phillips PPJ, Dufault SM, Quartagno M. Designing a response-over-continuous-intervention (ROCI) randomised trial: Implementation in the Phase 2C part (duration ranging) of the PARADIGM4TB trial. *Contemporary Clinical Trials*. 2025;155:108002. doi:10.1016/j.cct.2025.108002

Case Study 2: Is ROCI more efficient?

Sample size N_{tot_dur}	HRZE control event risk	Non-inferiority margin	Power (%)			
	(%)	(%)	No model	Quadratic	FP1	FP2
200 (40 per arm)	5	12	80.1	84.7	85.7	81.3
	5	15	90.3	95.3	96.9	92.9
	10	12				
	10	15				

4.2. Is modelling the duration-response curve necessary for improving the trial's operating characteristics?

As seen in Table 2 and Fig. 3, power is always higher when using models to estimate the confidence intervals around the risk differences. In some cases the differences are substantial (i.e. up to 10 percentage points) across scenarios when using FP1 or a quadratic model.

Source: Pham TM, Crook AM, Rolfe K, Phillips PPJ, Dufault SM, Quartagno M. Designing a response-over-continuous-intervention (ROCI) randomised trial: Implementation in the Phase 2C part (duration ranging) of the PARADIGM4TB trial. *Contemporary Clinical Trials*. 2025;155:108002. doi:10.1016/j.cct.2025.108002

ROCI

- Appropriate for early phase designs in TB.
- Model dependence: conclusions may hinge on the assumed duration-response shape, especially if the true curve is flat, irregular, or has a sharp threshold.
  - How does the method perform in the absence of efficacy at any dose/duration?
- Small-sample vulnerability: limited information at each duration, making uncertainty and convergence important.
- Intercurrent events and adherence matter: treatment interruption, regimen switches, loss to follow-up, and rescue therapy can complicate the estimand.
- Not an escape from non-inferiority: ROCI can reduce reliance on binary NI comparisons, but confirmatory use may still require margins or thresholds.

What problem is AER trying to solve?

First-generation PrEP non-inferiority trials can be hard to interpret when the active comparator is very effective and observed HIV events are rare.

Non-inferiority margins are opaque and inconsistent across trials.

Very low event rates make the usual risk-ratio comparison unstable and fragile.

The scientific goal is not merely “not much worse”; it is to evaluate whether a new PrEP option prevents enough infections to justify use.

The presentation reframes the question from a binary test to estimation with uncertainty.

What would make a result clinically convincing when both arms have very few infections?

Averted Events Ratio + Bayesian synthesis

Define a counterfactual background rate: what would HIV incidence have been with no PrEP in the same trial population?

Estimate how many infections each regimen averted relative to that background rate.

Compare regimens using the Averted Events Ratio: the ratio of infections prevented by experimental vs control.

Use Bayesian modeling to combine trial data, prior evidence, and uncertainty about the counterfactual rate.

Key appeal: more interpretable for prevention — "how many infections were prevented?"

Key cost: it depends on assumptions and auxiliary evidence for the no-PrEP rate.

Where is this strongest — and where is it vulnerable?

How credible is the estimated background HIV incidence, and which auxiliary sources should count most?

How should skeptical priors and sensitivity analyses be presented?

Does AER solve the low-event-rate problem, or mainly shift the uncertainty into the counterfactual rate?

Would clinicians, regulators, and communities find “infections averted” more decision-relevant than a non-inferiority margin?

How should safety, acceptability, adherence, cost, and delivery burden be incorporated alongside AER?

A promising estimand for prevention trials with low event rates; credibility depends on transparent assumptions.

Questions for discussion

- What do you recommend for a regulatory setting?
- Can patient-reported outcomes be utilized to address limitations of NI in some disease settings?
  - Patient-centered, measurable but not without limitations
- Are we comfortable letting a non-inferiority margin define what patients should accept, or should trials directly estimate the benefit–harm tradeoff that would make the new intervention preferable?