A Patient-Centric Paradigm for Clinical Research: The DOOR is Open

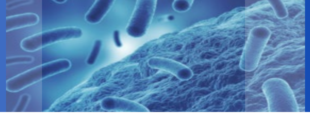
Toshi Hamasaki, PhD and Scott Evans, PhD
The Biostatistics Center
Department of Biostatistics and Bioinformatics
Milken Institute School of Public Health
George Washington University

Medical University of Vienna
Biometric Colloquium
April 10, 2025

Council for International Organizations of Medical Sciences (CIOMS)

- Jointly established by the WHO and UNESCO in 1949
- Convened a working group of regulators (including FDA (e.g., Richard Forshee, Hong Yang), PMDA, EMA), industry representatives, and academics
- Publishing “Benefit-risk balance for medicinal products” in 2025
- Two new points of emphasis:

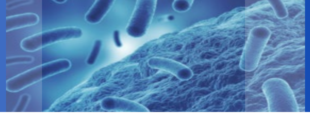
  1. Transitioning benefit-risk evaluation from a post-hoc exercise to an a comprehensively integrated element of clinical trial design and conduct, and
  2. A pragmatic patient-centric approach to benefit-risk assessment to ensure proper reflection and evaluation of the benefits and harms as experienced by patients.



Randomized clinical trials are the gold standard of evidence for evaluating the benefits and harms of medical and public health interventions.

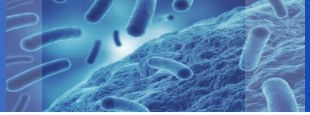
Most trials fail to provide the evidence needed to inform medical decision-making. The serious implications of this deficit are largely absent from public discourse.

DeMets and Califf, JAMA, 2011



Pragmatism and “Real World” Evidence

- Noble motivation
- Defines as obtaining the evidence that is the most useful for informing clinical practice
- However the terms are now generally defined by the data source
- The term “real world” is misleading, seemingly implying that trials that do not use associated data sources do not provide real world evidence
- Furthermore, true pragmatism requires going beyond the data source
- It involves asking the right questions, and implementing methods for trial design and analyses that are focused on patient-centric effectiveness



The good physician treats the disease;
the great physician treats the patient who has the disease.

Sir William Osler

Take care of your patient, not their organ.

Arun Sanyal

A Leaky Roof…

- Created a water bubble in my wall
- I need a new roof and I had to re-paper the wall
- I asked my neighbor, who recently papered a similarly sized room:

“How much paper did you buy?”

- He replied: “Six rolls.”

Upon finishing the papering of the wall…

- I had only used only 4 rolls
- I told my neighbor that I had 2 rolls left
- He replied:

“Oh. That happened to you too?”

It's a healthy thing now and then to hang a question mark on the things you have long taken for granted.

Bertrand Russell

Let us check the clinical trial arithmetic:

Challenges in Benefit:risk Evaluation: Totality of Evidence

- Typical benefit:risk analyses
  - Compare interventions for each efficacy and safety outcome
  - Combine these effects

- These analyses
  - Fail to incorporate associations between outcomes
  - Fail to recognize the cumulative nature of outcomes on individuals
  - Suffer from competing risk complexities during interpretation of individual outcomes, and
  - Since efficacy and safety analyses are often conducted on different populations, generalizability is unclear.

Question 1

- We define analysis populations
  - Efficacy: ITT population
  - Safety: safety population
- Efficacy population ≠ safety population
- We combine these analyses into benefit:risk analyses
- To whom does this analysis apply?
- What is the estimand?

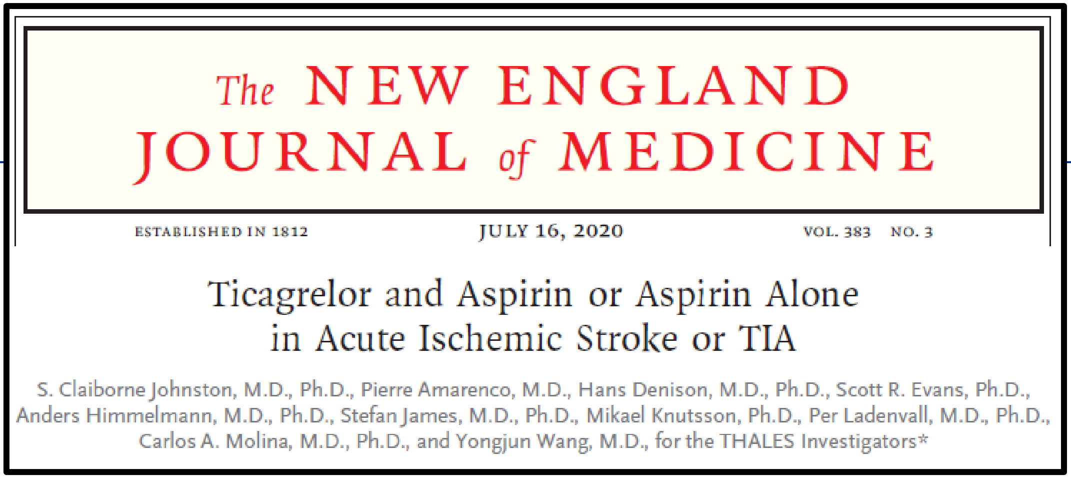
Question 2

- Suppose we measure the duration of hospitalization
- Shorter duration is better … or is it?
- The faster the patient dies, the shorter the duration
- Interpretation of an outcome needs context of other outcomes for the same patient

Question 3

- Suppose risk of death increases from 1 in 10 to 2 in 10
- RR=2. Very important.
- Suppose risk of death increases from 1 in 100,000 to 2 in 100,000
- RR=2. Nearly irrelevant.
- Are relative risks and ratios what we want?
- Additional challenges arise when interpreting multiple relative risks simultaneously…

THALES



The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 JULY 16, 2020 VOL. 383 NO. 3

Ticagrelor and Aspirin or Aspirin Alone in Acute Ischemic Stroke or TIA

S. Claiborne Johnston, M.D., Ph.D., Pierre Amarenco, M.D., Hans Denison, M.D., Ph.D., Scott R. Evans, Ph.D., Anders Himmelmann, M.D., Ph.D., Stefan James, M.D., Ph.D., Mikael Knutsson, Ph.D., Per Ladenvall, M.D., Ph.D., Carlos A. Molina, M.D., Ph.D., and Yongjun Wang, M.D., for the THALES Investigators*

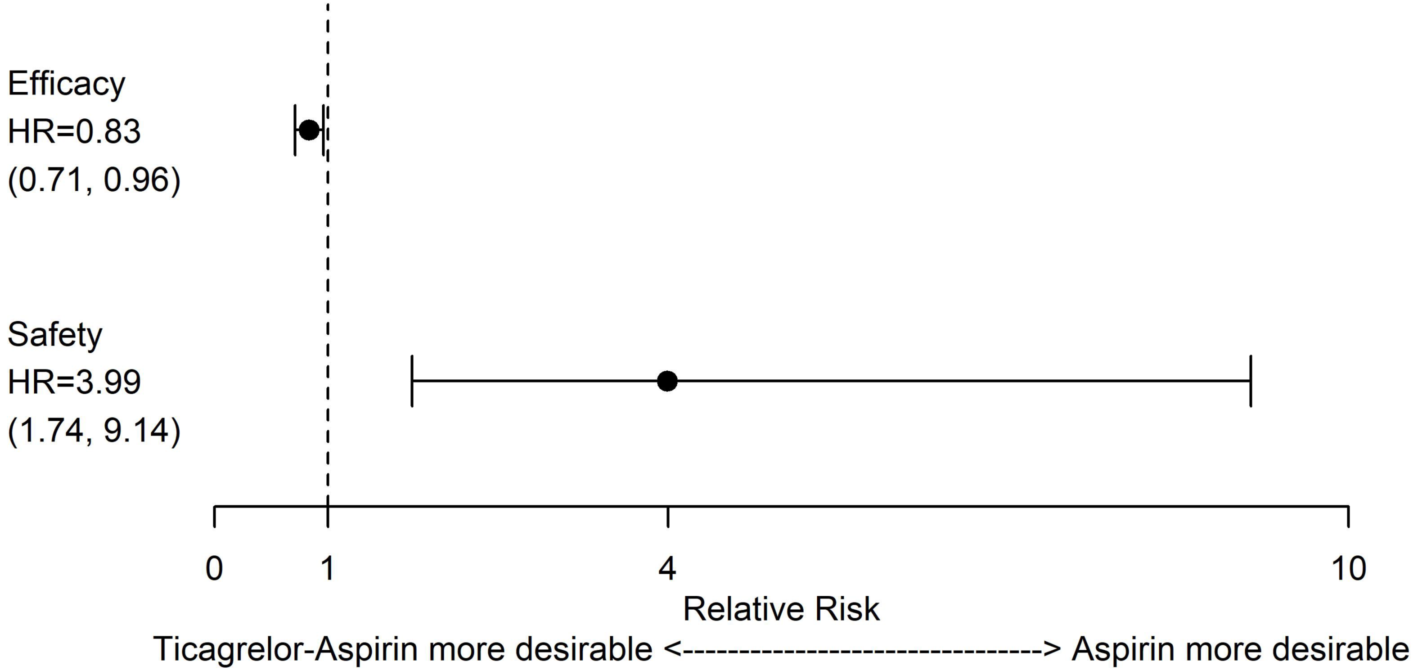
- Randomized, double-blinded, placebo-controlled trial (N=11,016)

- Primary outcome: time to stroke or death at 30 days
  - **HR = 0.83**, 95% CI = (0.71, 0.96), p=0.015

- Primary safety outcome: time to severe bleeding by 30 days
  - **HR = 3.99** 95% CI = (1.74, 9.14), p=0.001

- Too much bleeding?

Safety Problem > Efficacy Gain?



THALES

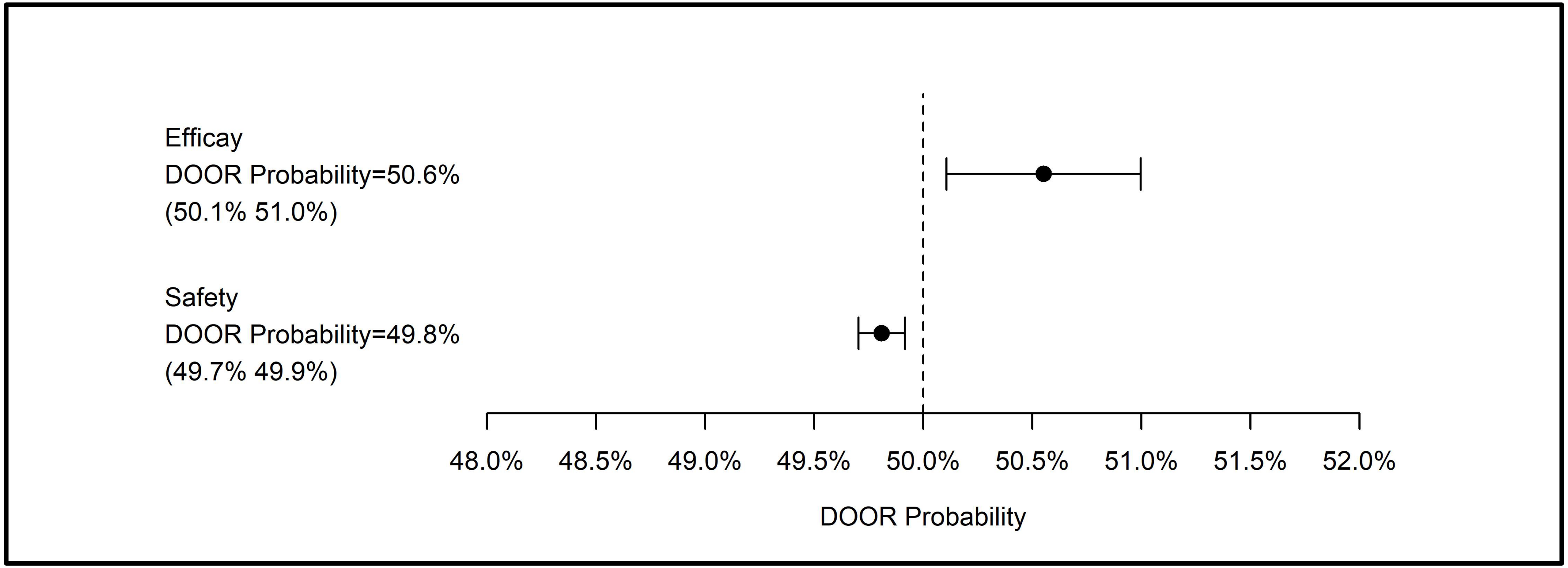
- Primary outcome: time to stroke or death at 30 days
  - Ticagrelor: 303 events (5.5%); Placebo: 362 events (6.6%)
  - Saved 59 events

- Primary safety outcome: time to severe bleeding by 30 days
  - Ticagrelor: 28 events (0.5%); Placebo: 7 events (0.1%)
  - Cost: 21 events

- Total savings: 38 events

- The benefit:risk community has known for a long time that evaluations must be on the absolute risk scale

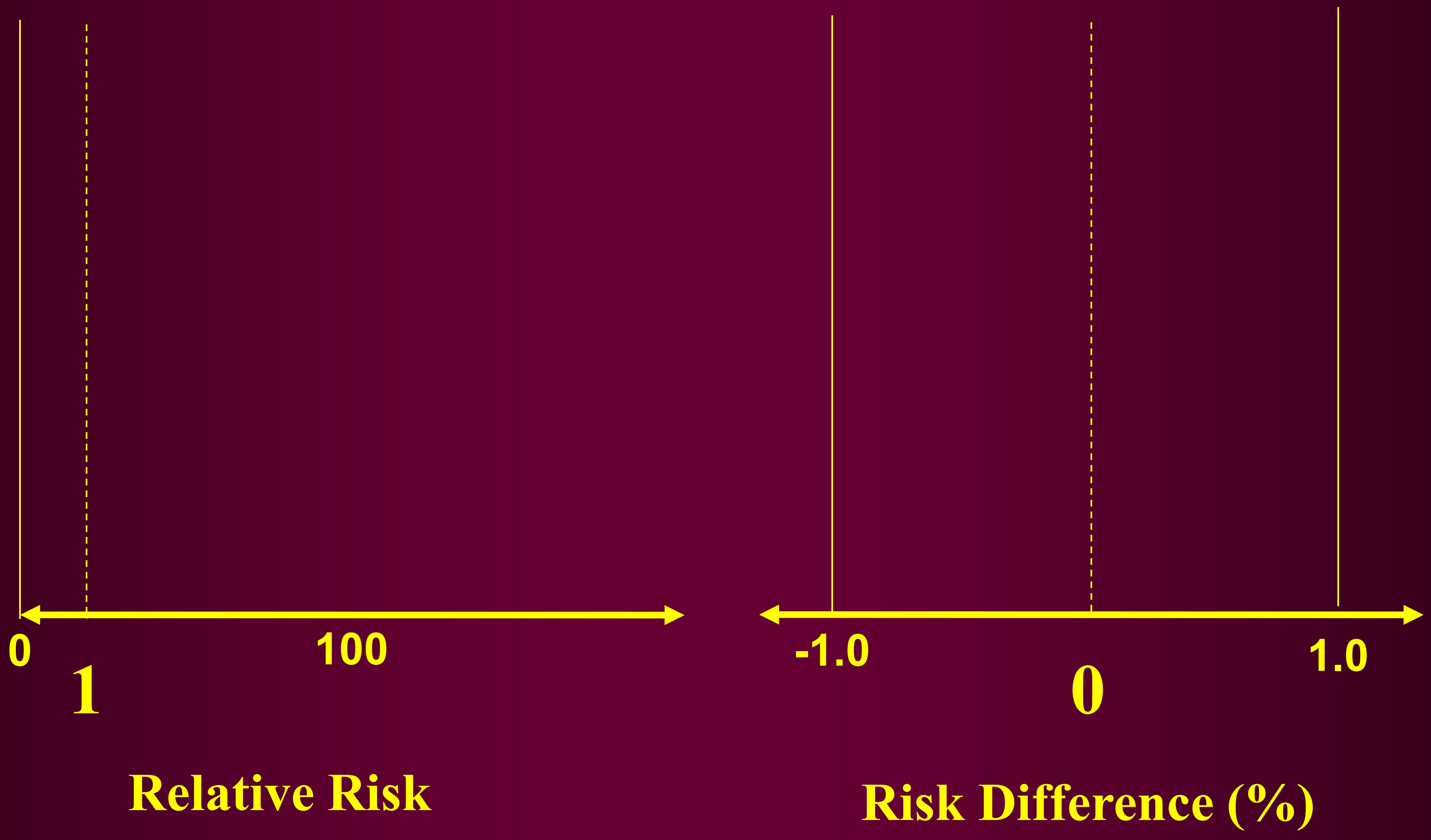
Now on an interpretable /common scale: safety issue < efficacy gain; uncertainty properly reflected



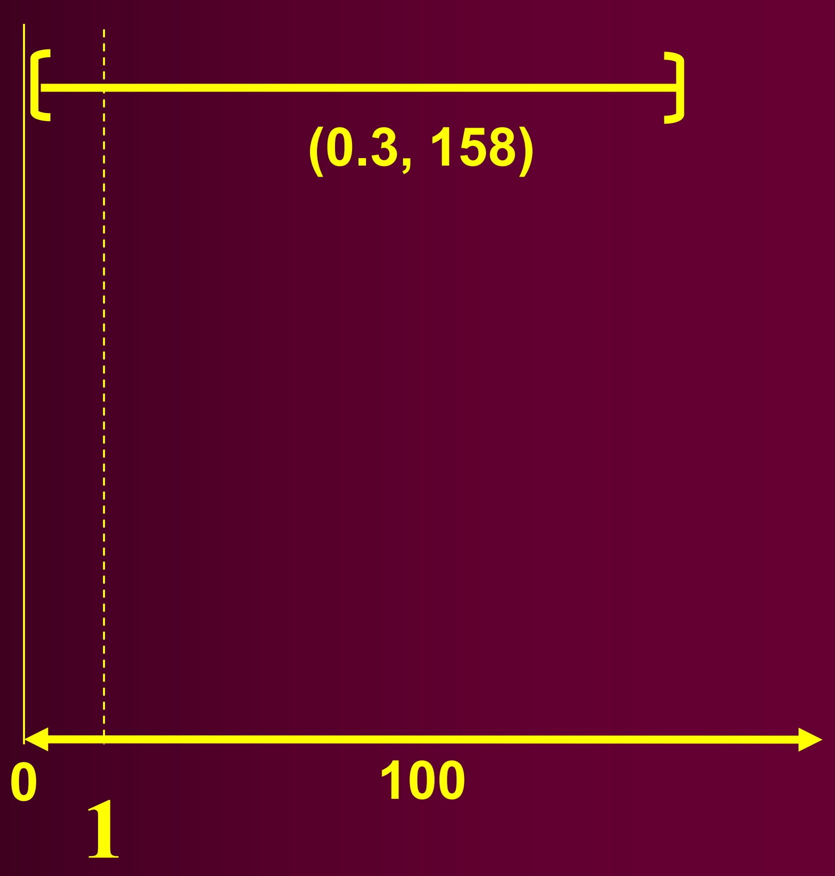
← Control more desirable | Ticagrelor more desirable →

Are we being efficient with the information?

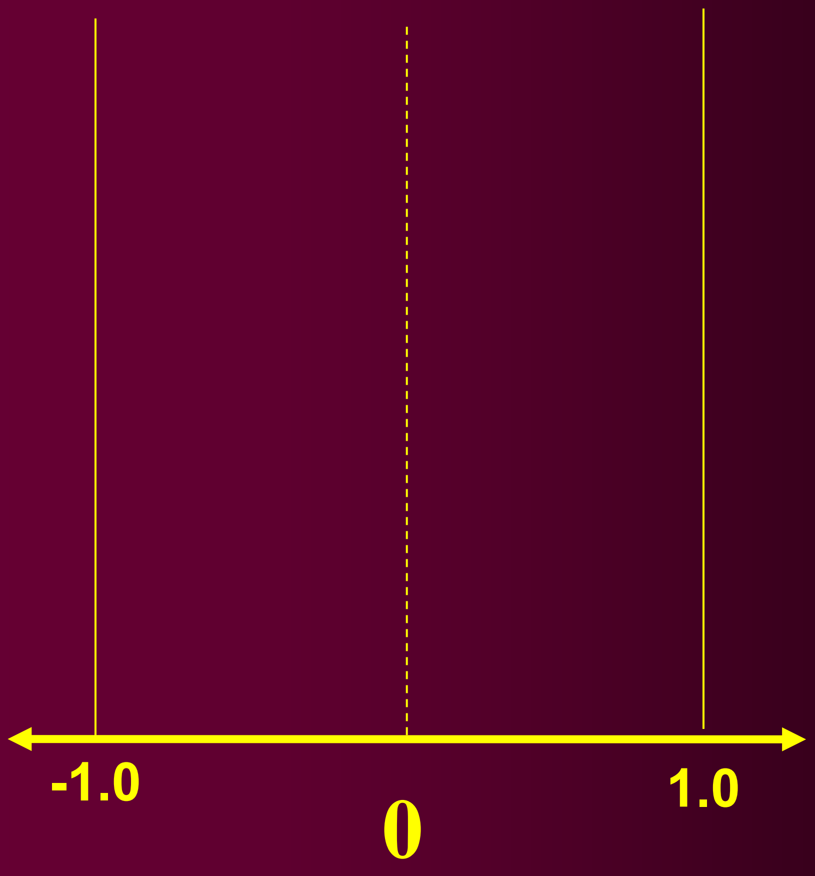
Trial 1: A (1/500) vs. B (3/500)



Trial 1: A (1/500) vs. B (3/500)

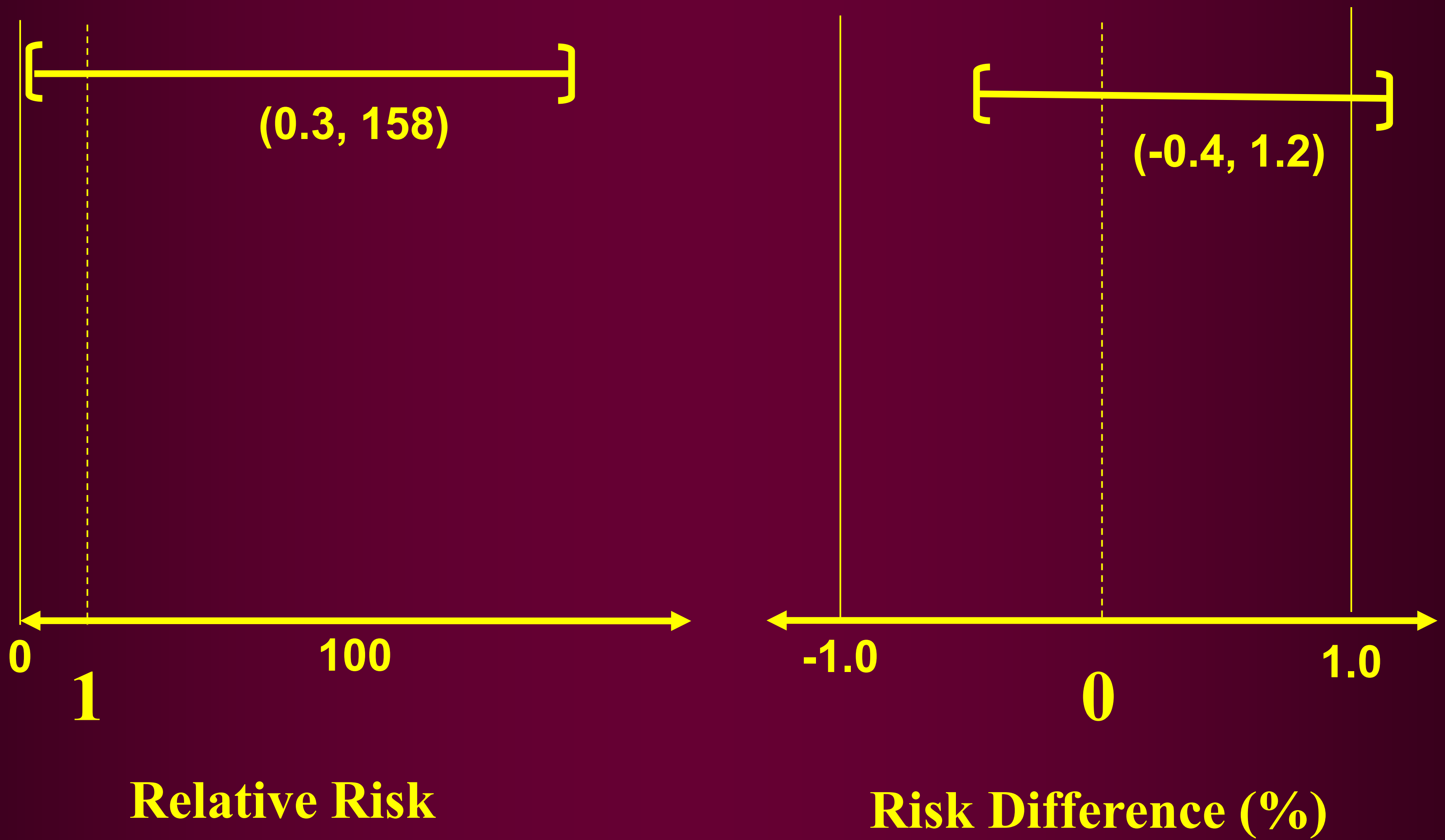


Relative Risk

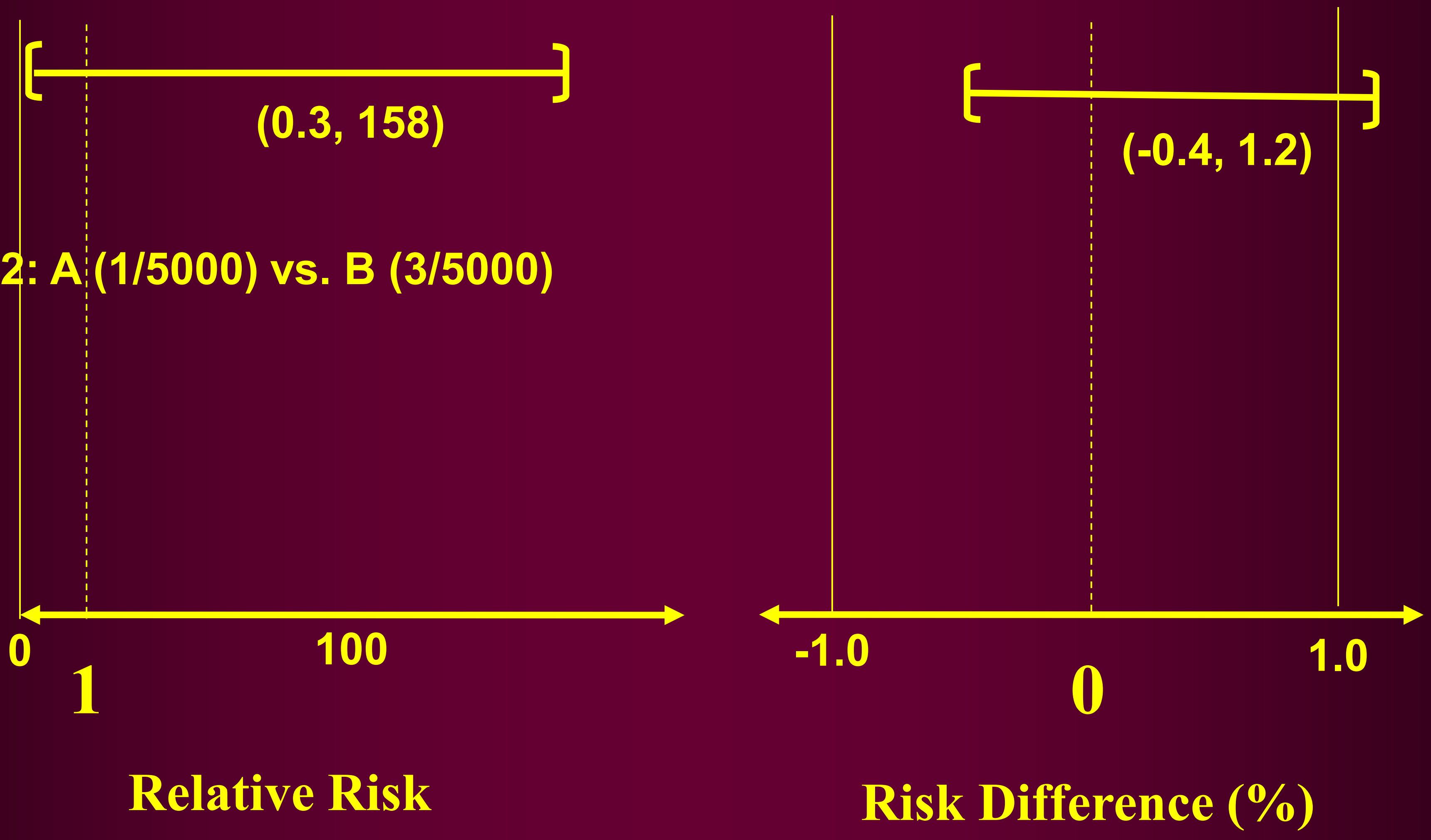


Risk Difference (%)

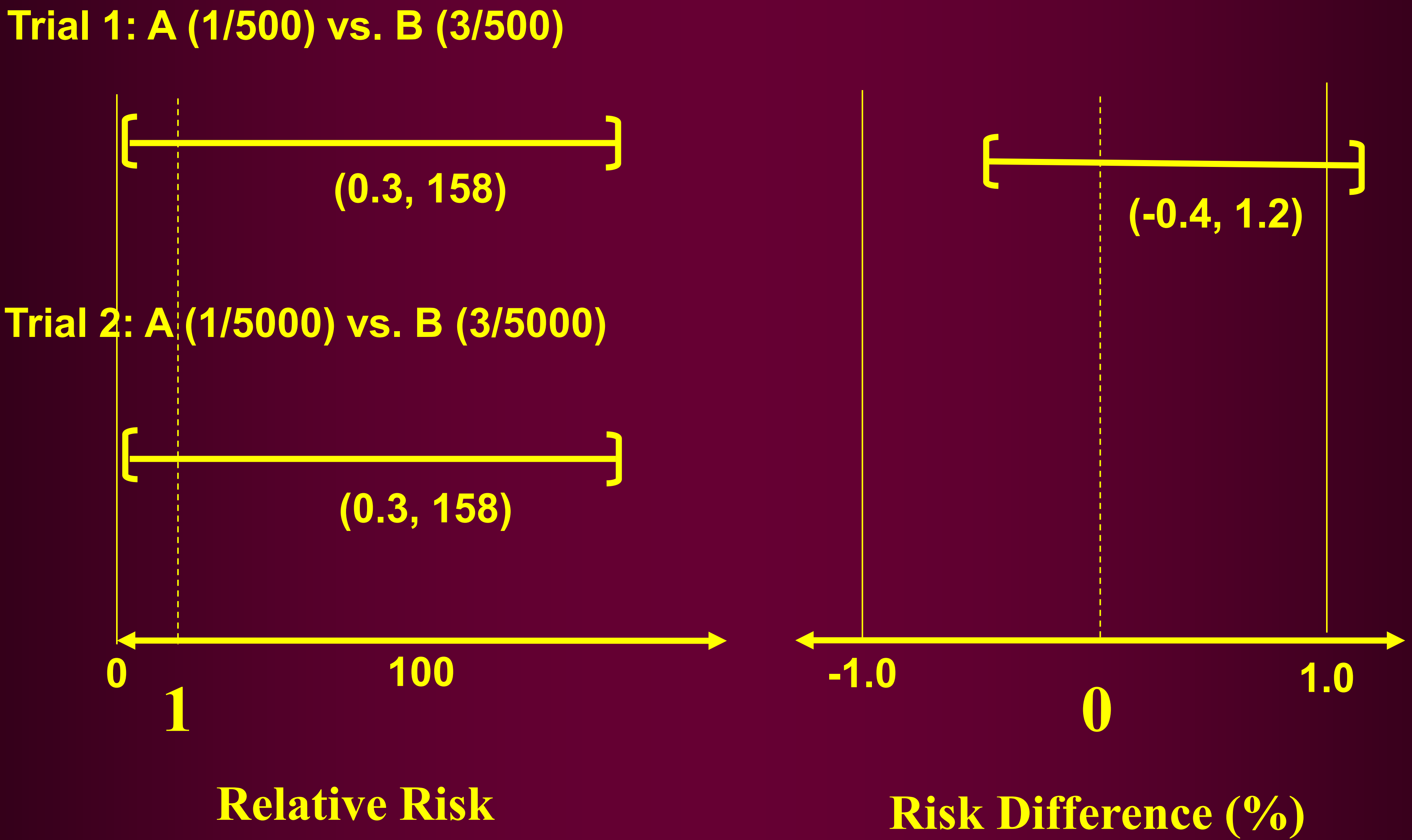
Trial 1: A (1/500) vs. B (3/500)



Trial 1: A (1/500) vs. B (3/500)



Trial 2: A (1/5000) vs. B (3/5000)

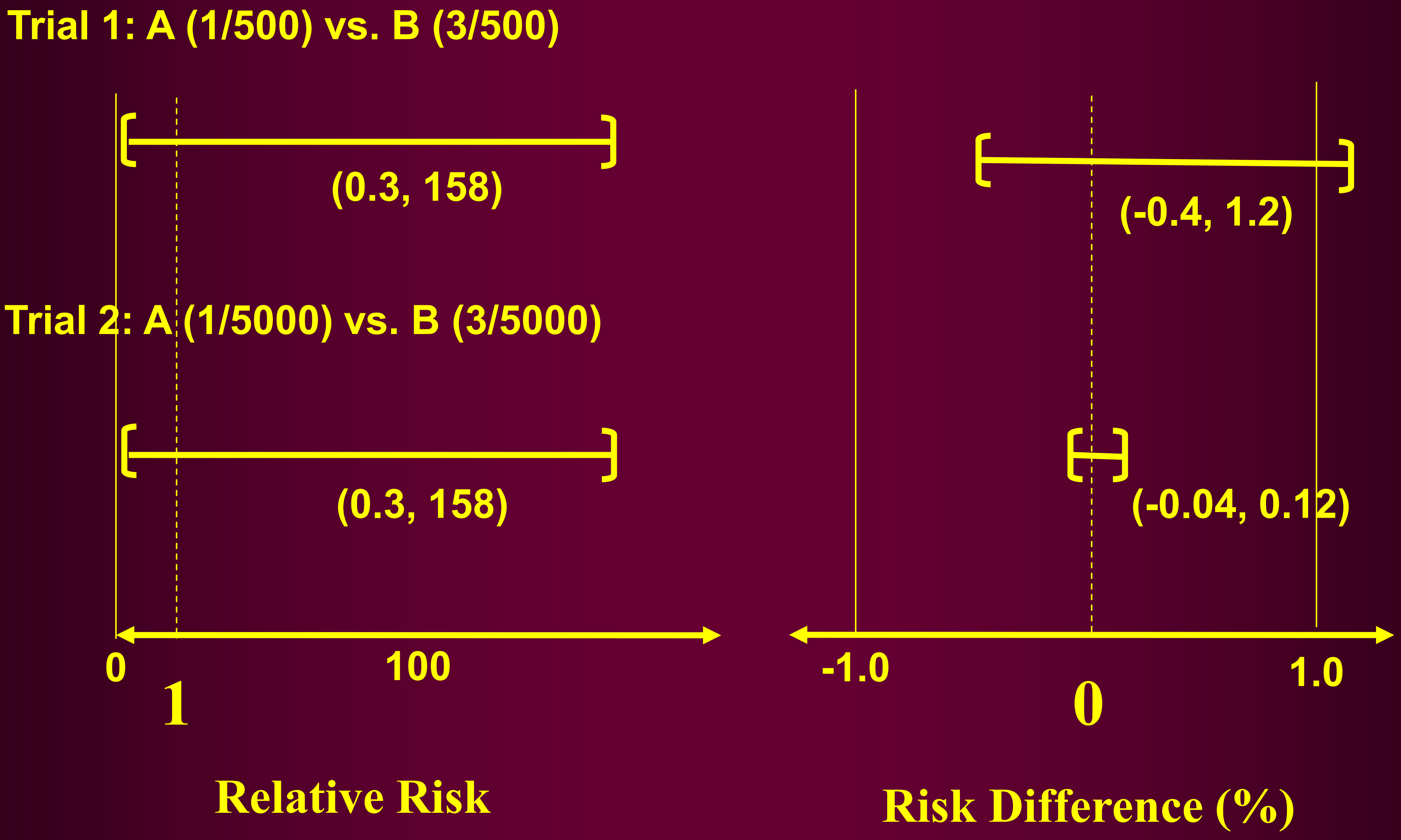


Trial 1: A (1/500) vs. B (3/500)

Trial 2: A (1/5000) vs. B (3/5000)

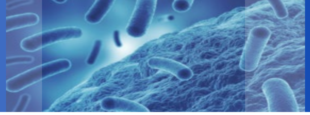
"No additional information."

We just observed an additional 9000 patients. This is no information?



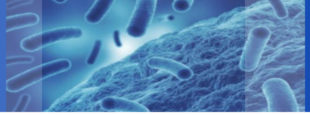
Trial 1: A (1/500) vs. B (3/500)

Trial 2: A (1/5000) vs. B (3/5000)



Question 4

- Suppose a loved one is diagnosed with a serious disease
- You are selecting treatment
- 3 treatment options: A, B, and C
- 2 outcomes, equally important
  - Treatment success: yes/no
  - Safety event: yes/no

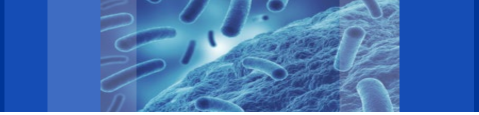


RCT Comparing A, B, and C
Analysis of Outcomes

A (N=100) B (N=100) C (N=100)

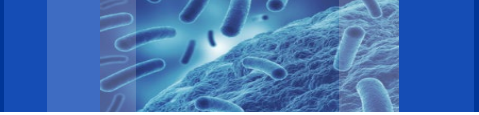
RCT Comparing A, B, and C
Analysis of Outcomes

A (N=100)	B (N=100)	C (N=100)
Success: 50%	Success: 50%	Success: 50%



RCT Comparing A, B, and C
Analysis of Outcomes

A (N=100)	B (N=100)	C (N=100)
Success: 50%	Success: 50%	Success: 50%
Safety event: 30%	Safety event: 50%	Safety event: 50%



RCT Comparing A, B, and C
Analysis of Outcomes

A (N=100)

Success: 50%

Safety event: 30%

B (N=100)

Success: 50%

Safety event: 50%

C (N=100)

Success: 50%

Safety event: 50%

Which treatment would you choose?

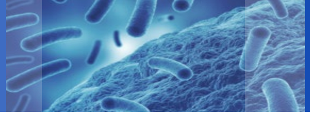
RCT Comparing A, B, and C

Analysis of Outcomes

A (N=100)	B (N=100)	C (N=100)
Success: 50%	Success: 50%	Success: 50%
Safety event: 30%	Safety event: 50%	Safety event: 50%

Which treatment would you choose?

They all have the same success rate.



RCT Comparing A, B, and C

Analysis of Outcomes

A (N=100)	B (N=100)	C (N=100)
Success: 50%	Success: 50%	Success: 50%
Safety event: 30%	Safety event: 50%	Safety event: 50%

Which treatment would you choose?

They all have the same success rate.

A has the lowest safety event rate.

RCT Comparing A, B, and C
Analysis of Outcomes

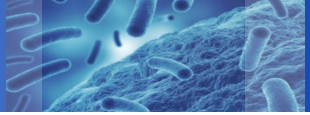
A (N=100)	B (N=100)	C (N=100)
Success: 50%	Success: 50%	Success: 50%
Safety event: 30%	Safety event: 50%	Safety event: 50%

Which treatment would you choose?

They all have the same success rate.

A has the lowest safety event rate.

B and C are indistinguishable.



RCT Comparing A, B, and C
Analysis of Outcomes

A (N=100)	B (N=100)	C (N=100)
Success: 50%	Success: 50%	Success: 50%
Safety event: 30%	Safety event: 50%	Safety event: 50%

Which treatment would you choose?

They all have the same success rate.

A has the lowest safety event rate.

B and C are indistinguishable.

Choose A…right?

Our culture is to use patients to analyze the outcomes.

Shouldn’t we use outcomes to analyze the patients?

Analysis of Patients: 4 Possible Outcomes

A (N=100)

Success: 50%

Safety event: 30%

		Success	
		+	-
SE	+	15	15
	-	35	35

B (N=100)

Success: 50%

Safety event: 50%

Success	
+	-
50	0
0	50

C (N=100)

Success: 50%

Safety event: 50%

Success	
+	-
0	50
50	0

Analysis of Patients: 4 Possible Outcomes

A (N=100)

Success: 50%

Safety event: 30%

		Success	
		+	-
SE	+	15	15
	-	35	35

B (N=100)

Success: 50%

Safety event: 50%

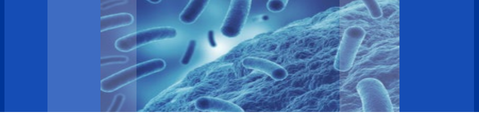
Success	
+	-
50	0
0	50

C (N=100)

Success: 50%

Safety event: 50%

Success	
+	-
0	50
50	0



Analysis of Patients: 4 Possible Outcomes

A (N=100)

Success: 50%

Safety event: 30%

		Success	
		+	-
SE	+	15	15
	-	35	35

B (N=100)

Success: 50%

Safety event: 50%

Success	
+	-
50	0
0	50

C (N=100)

Success: 50%

Safety event: 50%

Success	
+	-
0	50
50	0

Analysis of Patients: 4 Possible Outcomes

A (N=100)

Success: 50%

Safety event: 30%

		Success	
		+	-
SE	+	15	15
	-	35	35

B (N=100)

Success: 50%

Safety event: 50%

		Success	
		+	-
	+	50	0
	-	0	50

C (N=100)

Success: 50%

Safety event: 50%

		Success	
		+	-
	+	0	50
	-	50	0

PRagmatic-Explanatory Continuum Indicator Summary 2 (PRECIS-2) Wheel

Pragmatism implies…

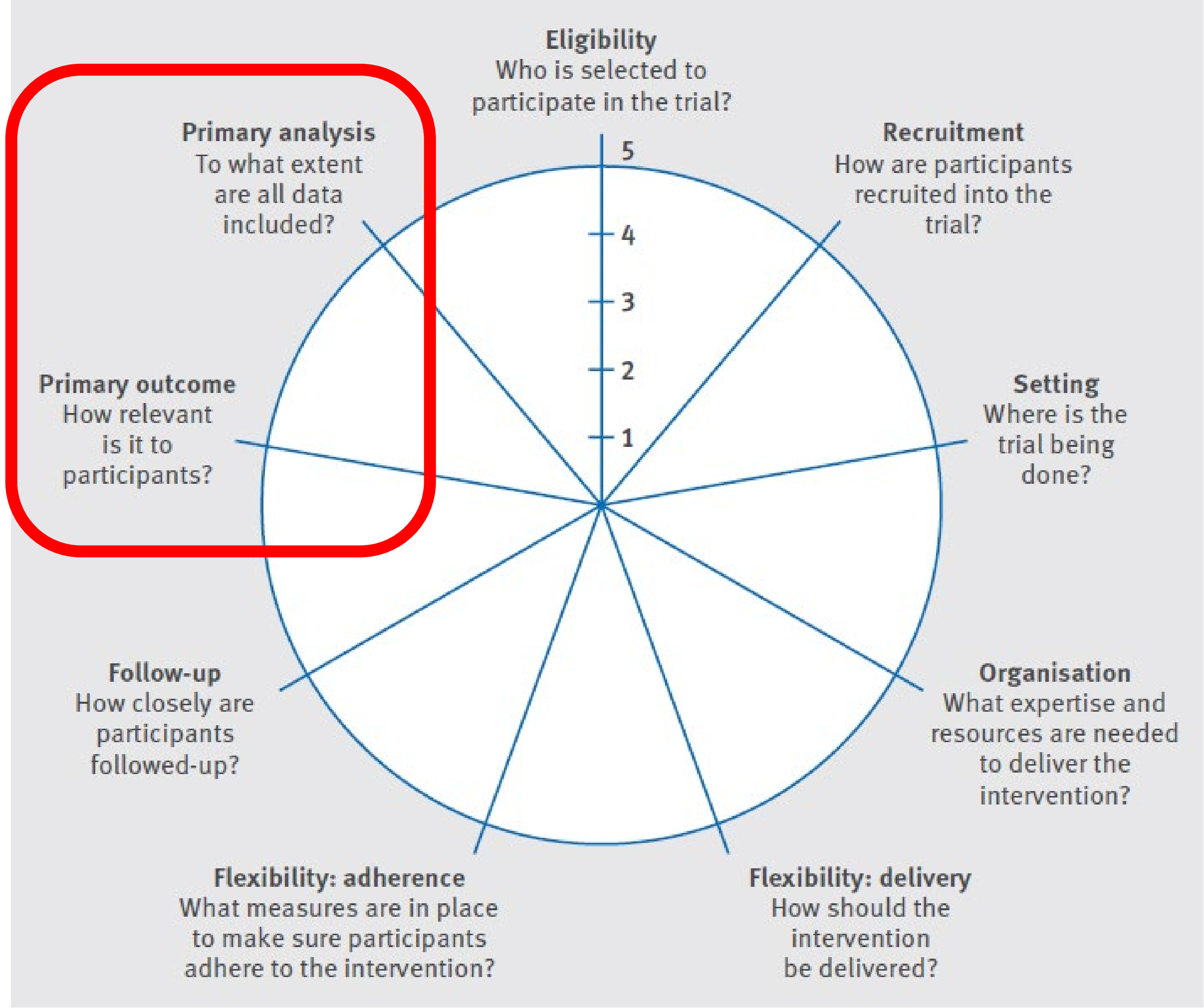
Patient-centricity

The outcome:

Must not only be relevant to patients…

But rather a holistic assessment of the patient…

THE patient outcome


Eligibility
Who is selected to participate in the trial?
Recruitment
How are participants recruited into the trial?
Setting
Where is the trial being done?
Organisation
What expertise and resources are needed to deliver the intervention?
Flexibility: delivery
How should the intervention be delivered?
Flexibility: adherence
What measures are in place to make sure participants adhere to the intervention?
Follow-up
How closely are participants followed-up?
Primary outcome
How relevant is it to participants?
Primary analysis
To what extent are all data included?
5
4
3
2
1

The analyses:

Must go beyond including all data…

But must analyze the patient…

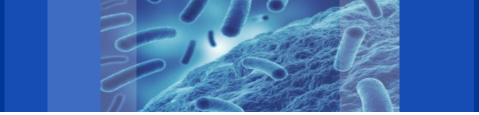
Rather than siloed elements of the patient

Question 5

- Many trials evaluate time-to-first event (e.g., death, MI, stroke, hospitalization)

- Fails to recognize multiple events

- Fail to distinguish differential importance of events
  - Death > non-fatal event
  - Disabling > non-disabling event
  - Permanent sequelae > transient sequelae



Question 5

- Many trials use binary endpoints

- Consider “clinical failure” e.g., death or failure of symptom improvement
  - One fails because they die
  - Another fails because of a failure of symptom improvement
  - Primary analyses treats these outcomes equivalently

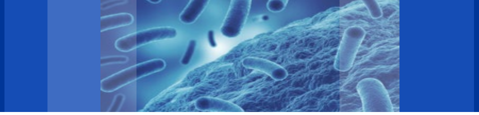
- Fails to recognize multiple events
  - More bad events worse than fewer

Question 5

- Mortality: a simpler objective clinically important “non-composite” binary endpoint

- Its non-composite nature does not imply homogeneity of response within survival

- E.g., the response of a patient that survives without morbidity, is classified equivalently to the response of one that survives with organ support e.g., ECMO, dialysis, … or both



Question 5

There is often a reluctance to use ordinal outcomes …
perhaps due to uncertainty about how to compose, layer, or grade outcomes.

It is often believed that binary endpoints avoid this issue…
though possibly unintentional or unwittingly,
composing, layering, and grading are already present,
resulting in incidental equivalent grading.

These gradations of responses are important…
can we do better than all or nothing?

Can we design and analyze trials in recognition of this?

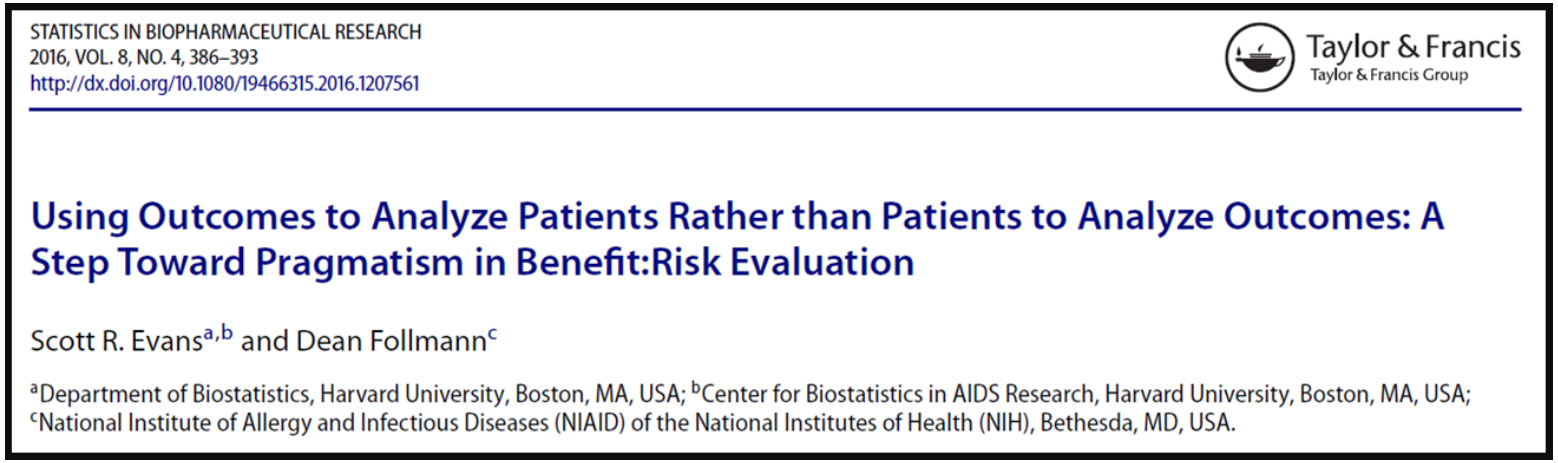
Question 6

- A negative trial does not mean a uniformly worthless intervention
- A positive trial does not mean the intervention works for everyone
- Predictive markers typically focus on a single efficacy/safety endpoint
- How do we identify the subgroup of patients we want to treat?
- Should “personalized medicine” be based on benefit:risk?

Question 7

- The effects of interventions are multi-dimensional with resulting cumulative effects on individual patients.

- Why are we not evaluating these cumulative effects?



STATISTICS IN BIOPHARMACEUTICAL RESEARCH
2016, VOL. 8, NO. 4, 386–393
http://dx.doi.org/10.1080/19466315.2016.1207561

Taylor & Francis
Taylor & Francis Group

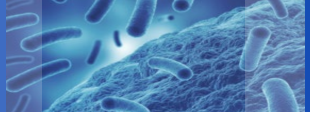
Using Outcomes to Analyze Patients Rather than Patients to Analyze Outcomes: A Step Toward Pragmatism in Benefit:Risk Evaluation

Scott R. Evans[a,b] and Dean Follmann[c]

[a]Department of Biostatistics, Harvard University, Boston, MA, USA; [b]Center for Biostatistics in AIDS Research, Harvard University, Boston, MA, USA; [c]National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health (NIH), Bethesda, MD, USA.

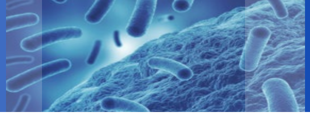
Scott's father (a math teacher) to his confused son many years ago:

"The order of operations is important…"



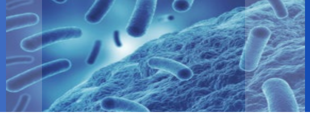
What people think of as the moment of discovery …
is really the discovery of the question.

Jonas Salk



Provenance and Philosophy of the DOOR Paradigm

- Optimal pragmatism requires a patient-centric outcome, composing typical component outcomes to represent the patient experience
  - Not all composites need to be binary
  - Not all components need to be treated as equivalent
    - E.g., death is more important than other events
- Composite endpoint have challenges requiring attention
  - All components should be evaluated individually to see if effects go in similar vs opposing directions, and to elucidate components driving the overall response
  - Cumulative evaluation of ordinal composite
- The multiple outcome nature of the composite and associated components necessitates an absolute risk scale and avoidance of relative risk / ratio measures
- Design and analyses should prioritize robustness, objectivity, error control, and transparency, and avoidance of concessions of these principles



Desirability Of Outcome Ranking (DOOR)

- A patient-centric paradigm for the design, monitoring, analyses and reporting of clinical trials based on benefit:risk

- Uses outcomes to analyze patients rather than patients to analyze outcomes
  - Representing a closer reflection of the effects on patients

- Addresses noted challenges

Before we analyze several hundred patients,
we must understand how to analyze one.

- Construct patient-centric DOOR outcome based on the patient journey

DOOR Analyses

- Guiding principles for replicability, pragmatism, and robustness

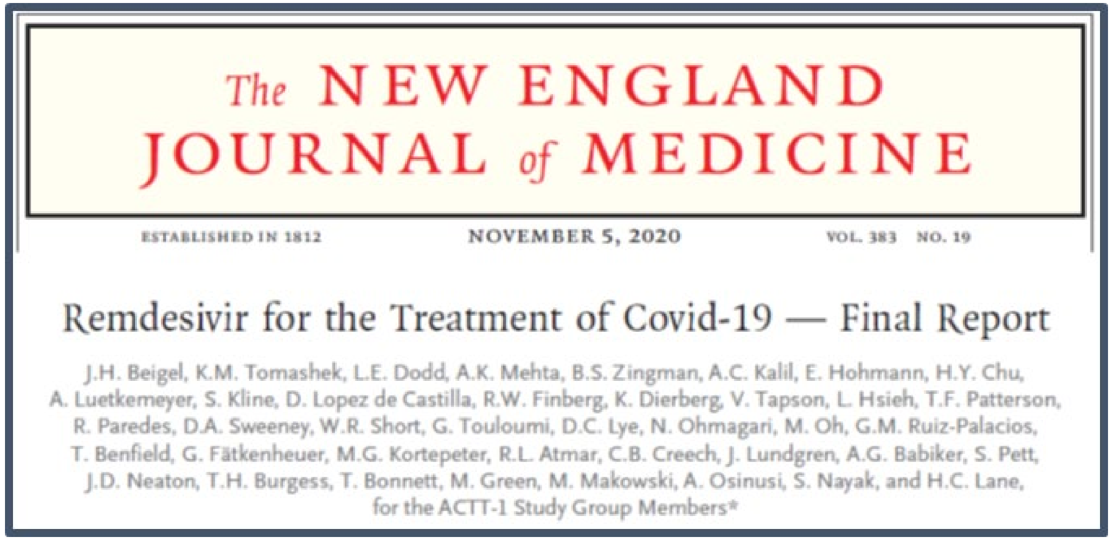
- Two complimentary analyses
    1. Rank-based
        - Estimate the DOOR probability: the probability that a patient from treatment has a more desirable outcome than a patient on control
            - Equivalent distributions imply 50%
            - Unconventional though … intuitively attractive
    2. Partial credit (grade-based analyses)

- Analyses of component outcomes is an integrated part of the evaluation

Adaptive Covid-19 Treatment Trial (ACTT-1)

- No known efficacious treatments for COVID-19 at the time

- ACTT-1
  - Randomized double-blind placebo-controlled trial of IV remdesivir in hospitalized adult COVID-19 patients w/ LRTI
  - N=1062



The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 NOVEMBER 5, 2020 VOL. 383 NO. 19

Remdesivir for the Treatment of Covid-19 — Final Report

J.H. Beigel, K.M. Tomashek, L.E. Dodd, A.K. Mehta, B.S. Zingman, A.C. Kalil, E. Hohmann, H.Y. Chu, A. Luetkemeyer, S. Kline, D. Lopez de Castilla, R.W. Finberg, K. Dierberg, V. Tapson, L. Hsieh, T.F. Patterson, R. Paredes, D.A. Sweeney, W.R. Short, G. Touloumi, D.C. Lye, N. Ohmagari, M. Oh, G.M. Ruiz-Palacios, T. Benfield, G. Fätkenheuer, M.G. Kortepeter, R.L. Atmar, C.B. Creech, J. Lundgren, A.G. Babiker, S. Pett, J.D. Neaton, T.H. Burgess, T. Bonnett, M. Green, M. Makowski, A. Osinusi, S. Nayak, and H.C. Lane, for the ACTT-1 Study Group Members*

ACTT-1

- Important events
  - Death
  - Hospitalized with invasive mechanical ventilation / ECMO
  - SAE that is not resolved or resolved with sequelae

	Treatment	
DOOR (Day 29)	**Remdesivir (N=541)**	**Placebo (N=521)**
Alive: 0 of the other events above		382 (73.3%)
Alive: 1 of the other events above		57 (10.9%)
Alive: both of the other events above		6 (1.2%)
Death		76 (14.6%)

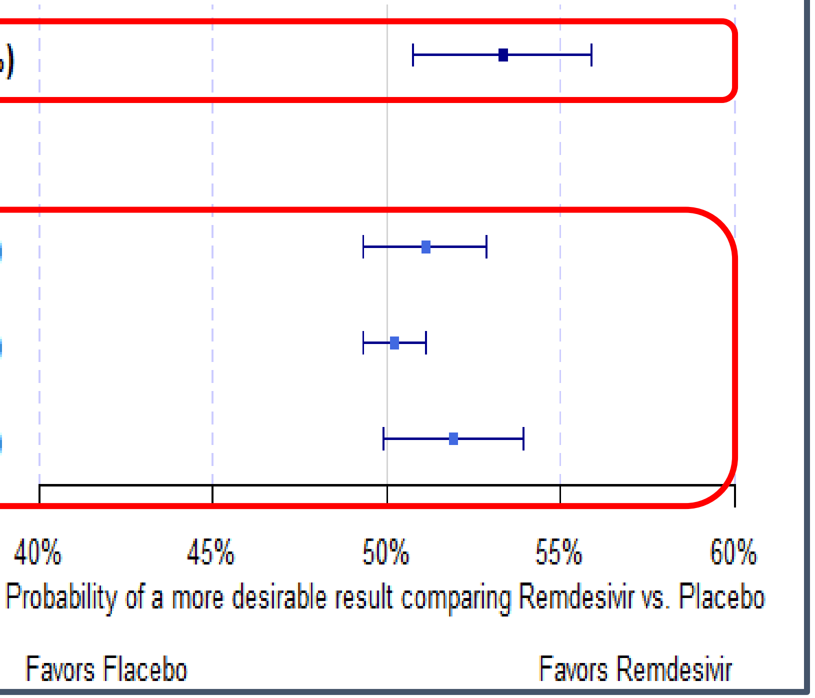
A northward migration to more desirable categories with treatment?

ACTT-1

- Important events
  - Death
  - Hospitalized with invasive mechanical ventilation / ECMO
  - SAE that is not resolved or resolved with sequelae

	Treatment	
DOOR (Day 29)	**Remdesivir (N=541)**	**Placebo (N=521)**
Alive: 0 of the other events above	433 (80.0%)	382 (73.3%)
Alive: 1 of the other events above	42 (7.8%)	57 (10.9%)
Alive: both of the other events above	8 (1.5%)	6 (1.2%)
Death	58 (10.7%)	76 (14.6%)

	Placebo (N=521) **n(%)**	**Remdesivir (N=541)** **n(%)**	**DOOR probability** **(95% CI)**
Primary DOOR			**53.3% (50.8%, 55.9%)**
DOOR components			
Hosp. w/ invasive mechanical ventilation / ECMO	55(10.6%)	45(8.3%)	51.1% (49.4%, 52.9%)
SAE	13(2.5%)	11(2.0%)	50.2% (49.3%, 51.1%)
Death	76(14.6%)	58(10.7%)	51.9% (49.9%, 53.9%)



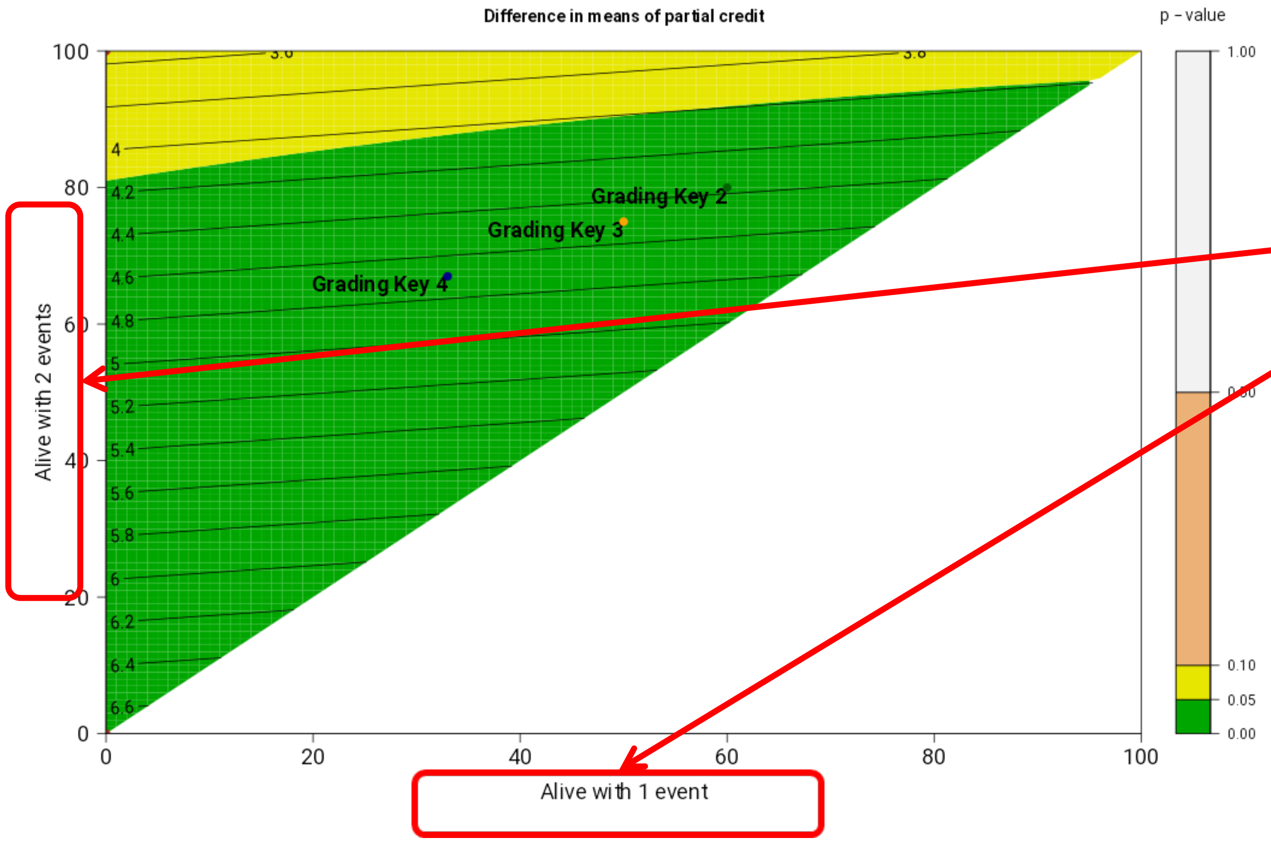
Partial Credit

	Score
1. Alive: 0 of the events	100
2. Alive: 1 of the events	Partial credit
3. Alive: both of the events	Partial credit
4. Death	0

Partial credit can be used to account for:

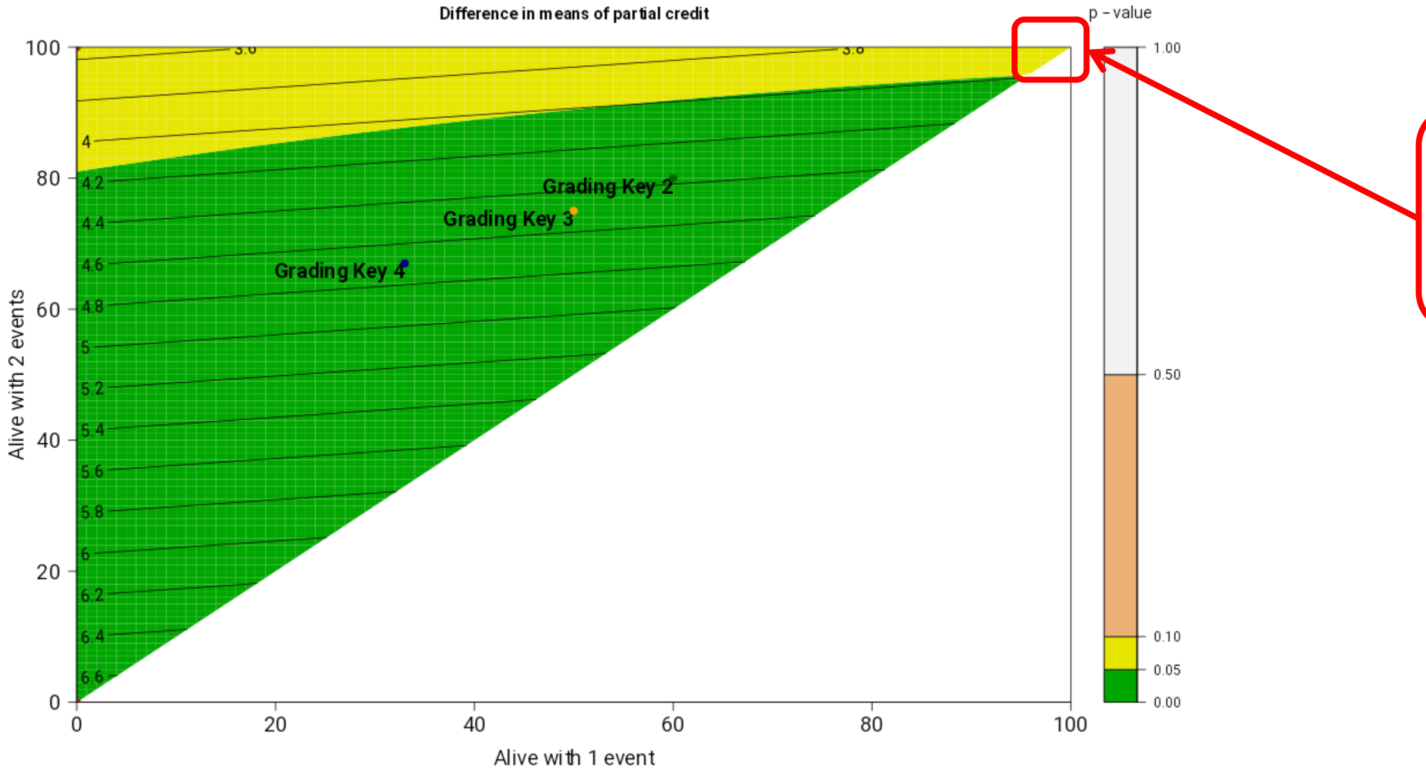
1. Strategic distancing between steps in a calculated way
2. Personalized perspectives among patients / clinicians regarding the desirability of the categories
3. Robustness analyses

Contours of Effects as Partial Credit Varies



Category	Credit
Alive; 0 events	100
Alive; 1 event	Partial credit
Alive; both events	Partial credit
Death	0

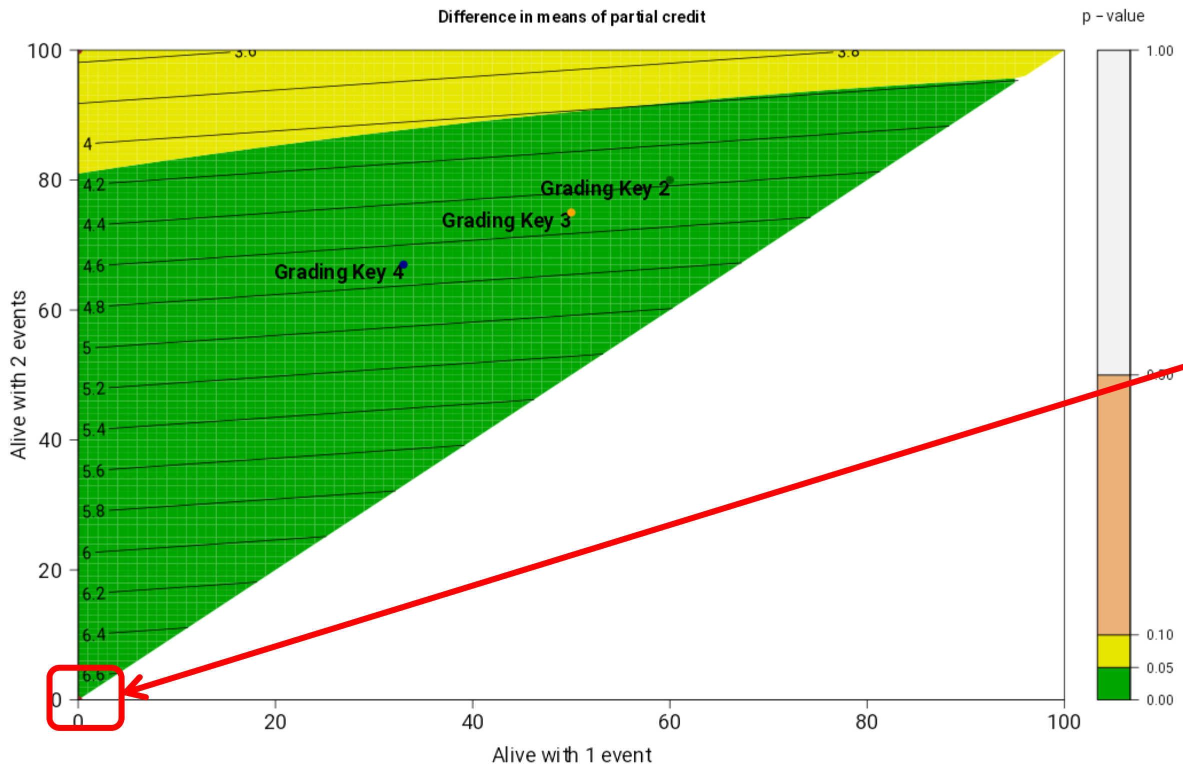
The Easy Grader



Category	Credit
Alive; 0 events	100
Alive; 1 event	100
Alive; both events	100
Death	0

Remdesivir Advantage ≈ 3.9%

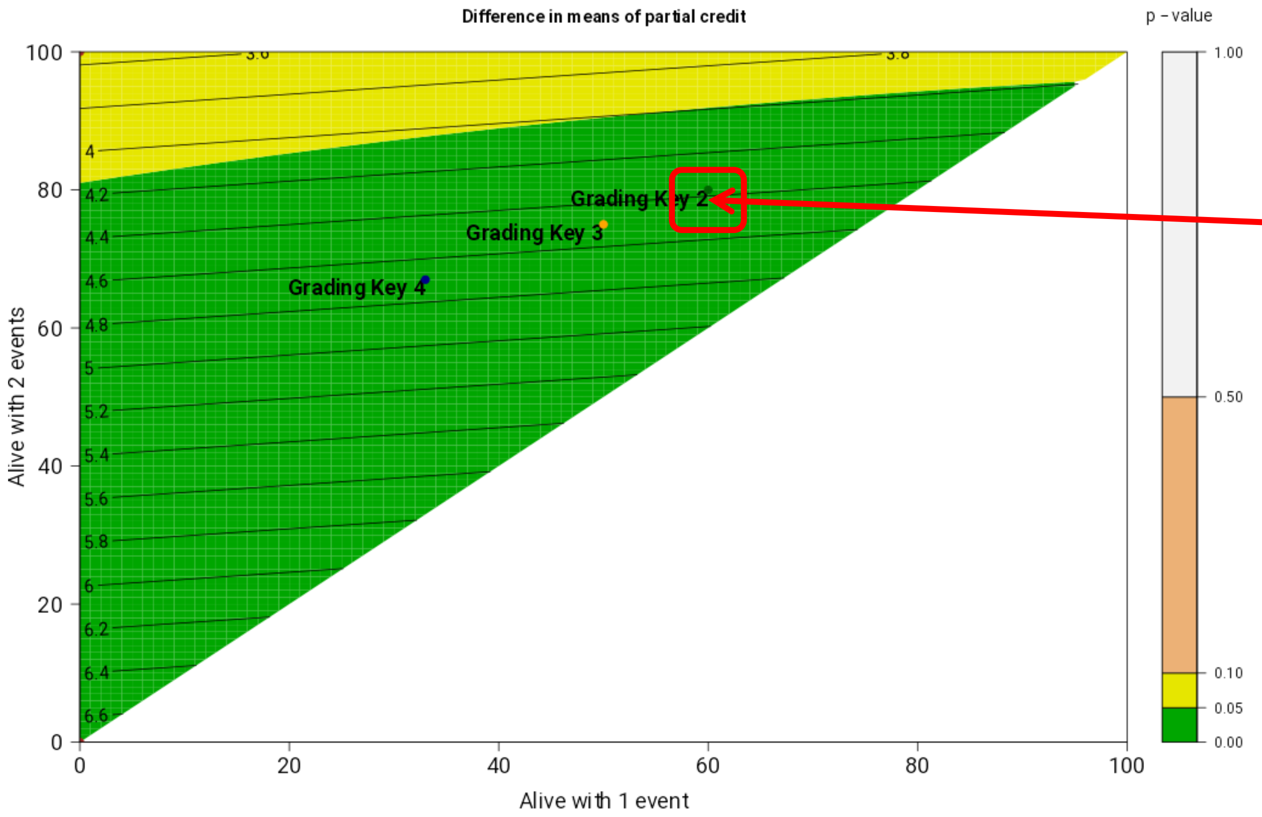
The Curmudgeon Professor



Category	Credit
Alive; 0 events	100
Alive; 1 event	0
Alive; both events	0
Death	0

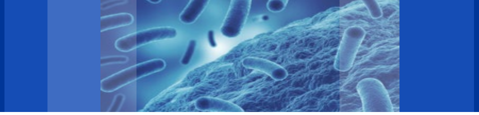
Remdesivir Advantage ≈ 6.7%

The Layered Compromise



Category	Credit
Alive; 0 events	100
Alive; 1 event	80
Alive; both events	60
Death	0

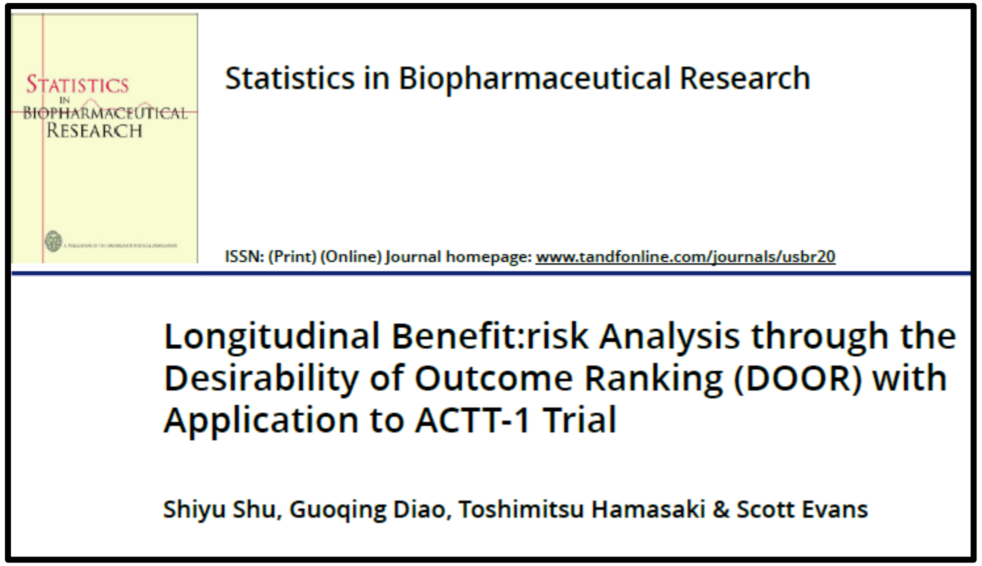
Remdesivir Advantage ≈ 4.4%



Sliding DOOR

- View DOOR outcome as a longitudinal patient state
- Methods for:
  - Repeated measures
    - Simultaneous confidence bands
  - Monotone progression
    - Multidimensional RMST
  - Non-monotone levels
    - Anthology of Patient Stories

Sliding DOOR



Statistics in Biopharmaceutical Research

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/usbr20

Longitudinal Benefit:risk Analysis through the Desirability of Outcome Ranking (DOOR) with Application to ACTT-1 Trial

Shiyu Shu, Guoqing Diao, Toshimitsu Hamasaki & Scott Evans

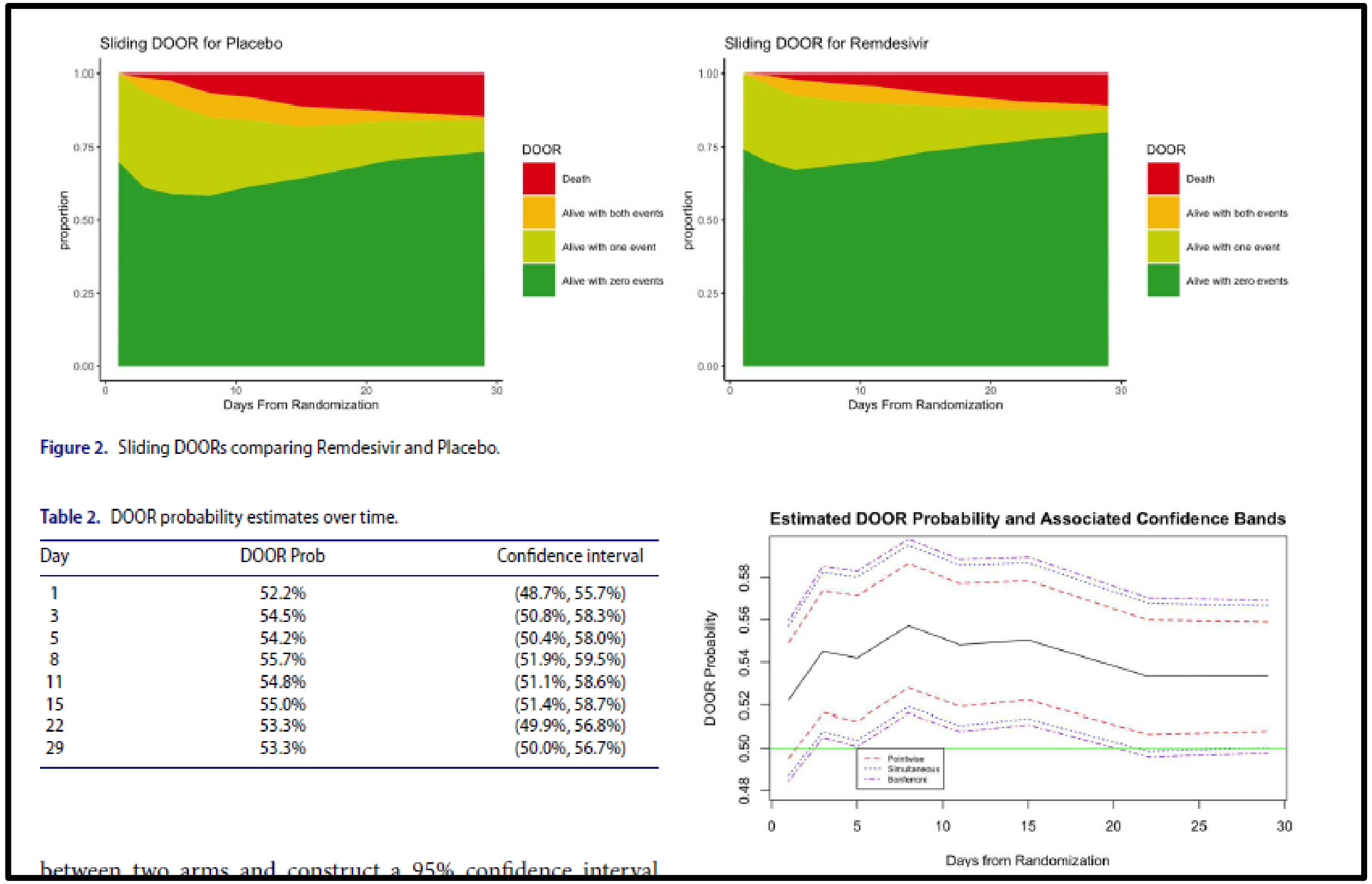


Figure 2. Sliding DOORs comparing Remdesivir and Placebo.

Table 2. DOOR probability estimates over time.

Day	DOOR Prob	Confidence interval
1	52.2%	(48.7%, 55.7%)
3	54.5%	(50.8%, 58.3%)
5	54.2%	(50.4%, 58.0%)
8	55.7%	(51.9%, 59.5%)
11	54.8%	(51.1%, 58.6%)
15	55.0%	(51.4%, 58.7%)
22	53.3%	(49.9%, 56.8%)
29	53.3%	(50.0%, 56.7%)

between two arms and construct a 95% confidence interval

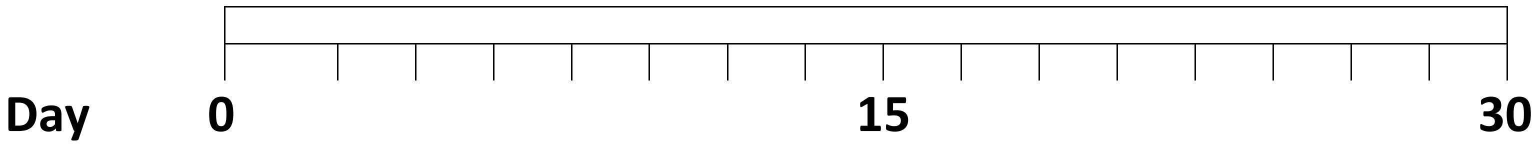
Anthology of Patient Stories

- Recognize the non-monotone nature of the patient status
  - Status could improve or decline

- Recognize
  - When events occur and for how long
  - If and when they recover
  - If and when they relapse

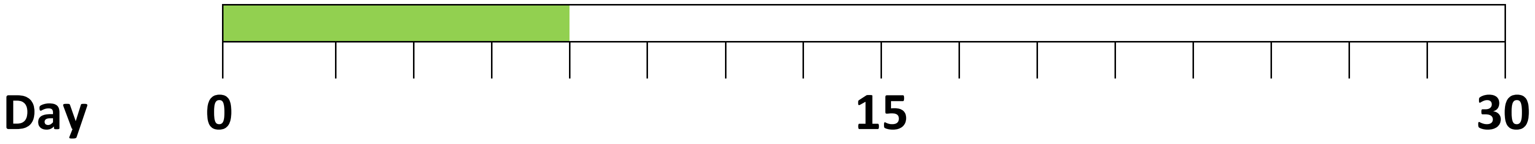
- Retain patient-centricity

The Patient Story



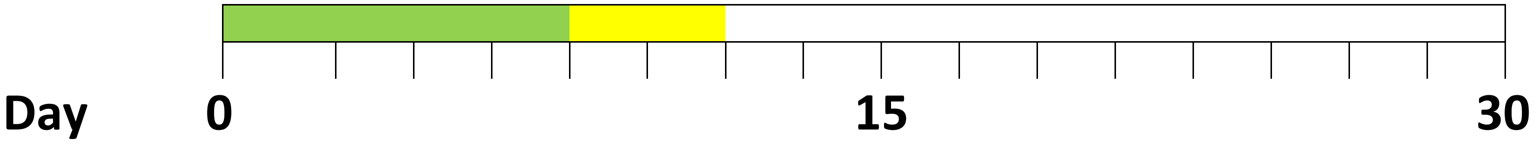
Alive with 0 events	🟩
Alive with 1 event	🟨
Alive with 2 events	🟥
Death	⬛

The Patient Story



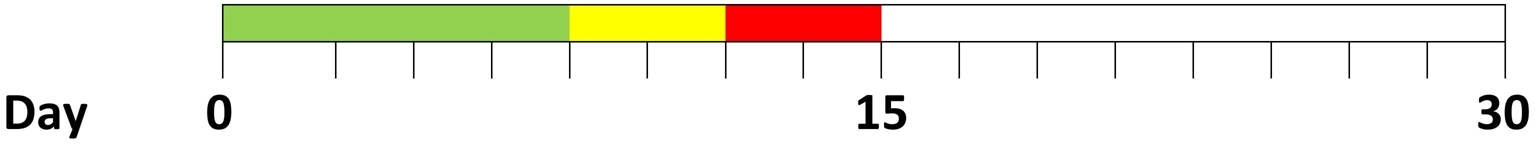
Alive with 0 events	🟩
Alive with 1 event	🟨
Alive with 2 events	🟥
Death	⬛

The Patient Story



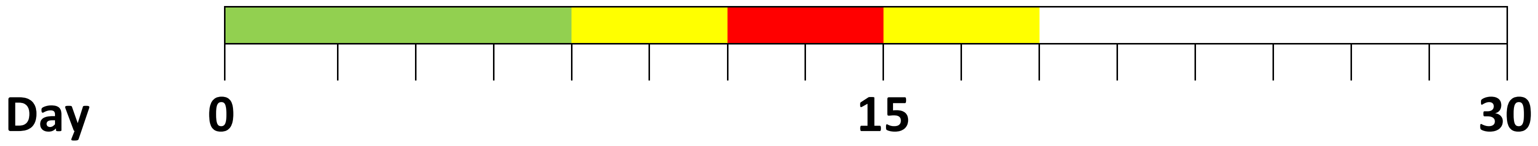
Alive with 0 events	🟩
Alive with 1 event	🟨
Alive with 2 events	🟥
Death	⬛

The Patient Story



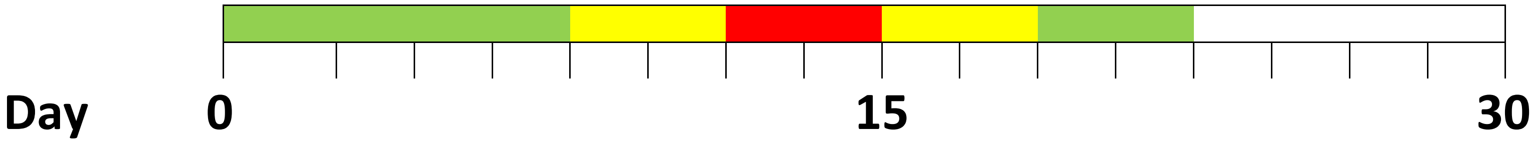
Alive with 0 events	🟩
Alive with 1 event	🟨
Alive with 2 events	🟥
Death	⬛

The Patient Story



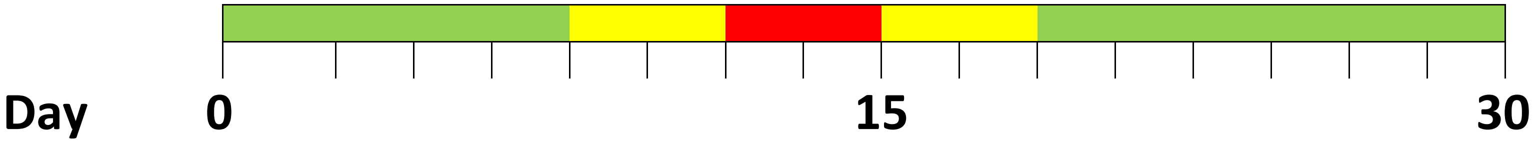
Alive with 0 events	🟩
Alive with 1 event	🟨
Alive with 2 events	🟥
Death	⬛

The Patient Story



Alive with 0 events	🟩
Alive with 1 event	🟨
Alive with 2 events	🟥
Death	⬛

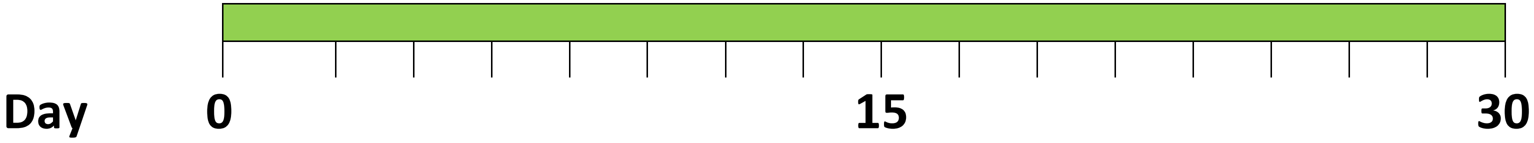
The Patient Story



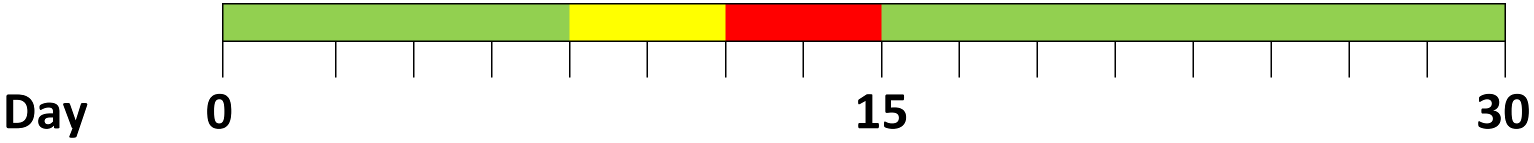
Alive with 0 events	🟩
Alive with 1 event	🟨
Alive with 2 events	🟥
Death	⬛

The Trial Anthology: A Collection of Patient Stories

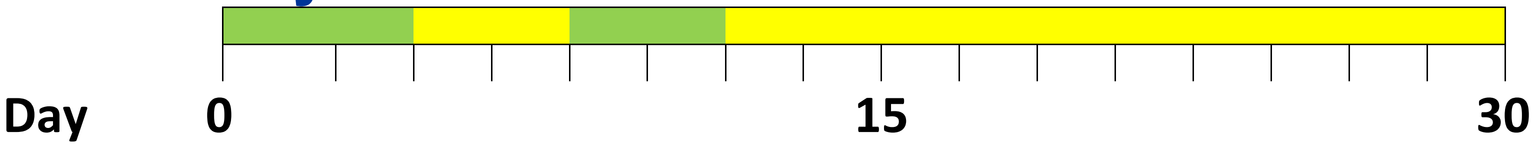
Saul Goodman



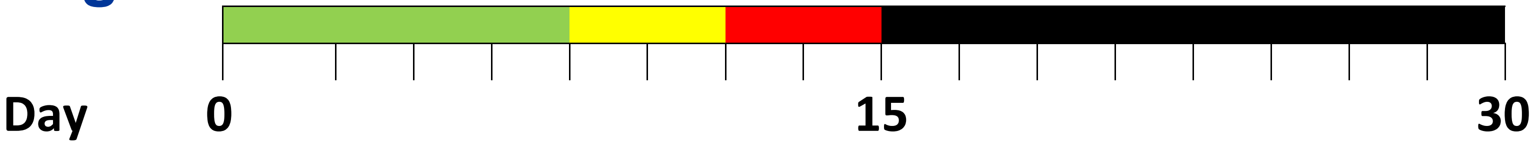
Ita Lendswell



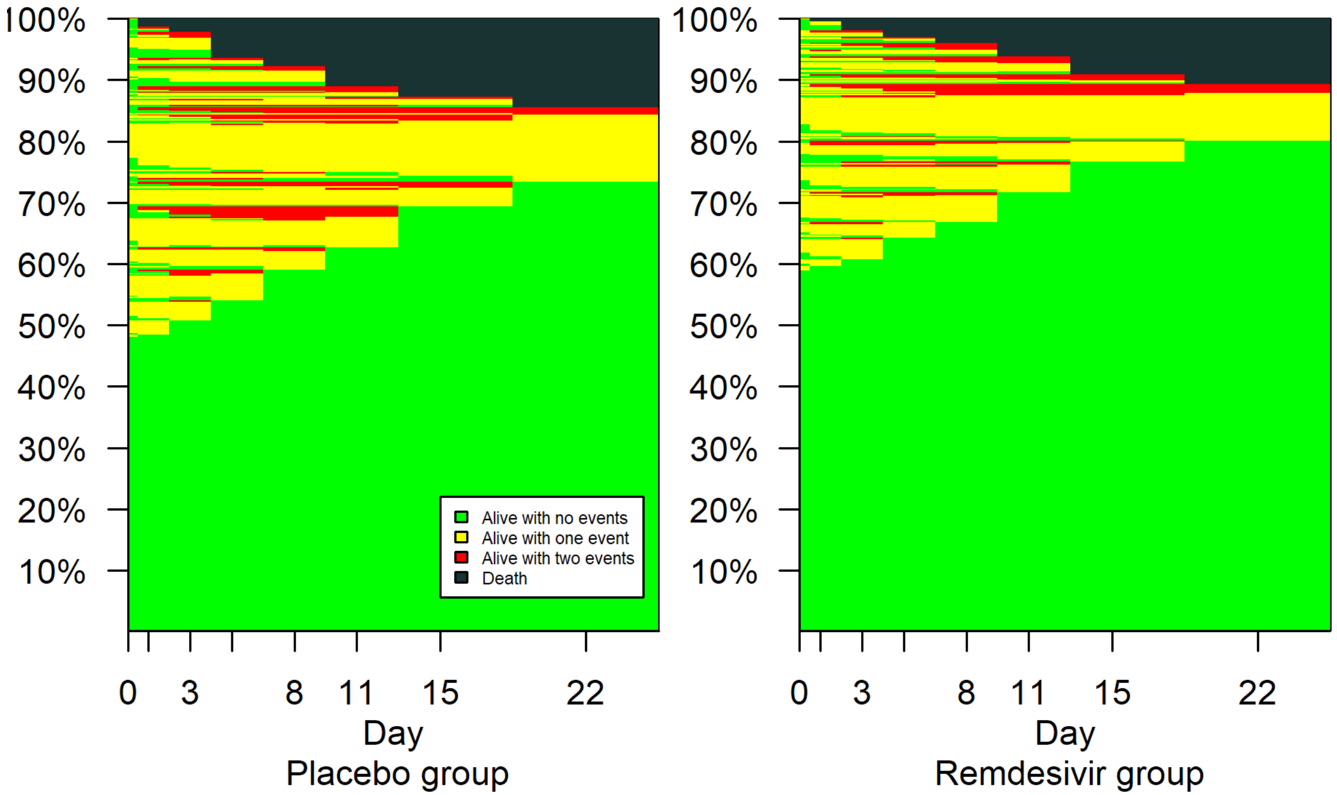
Nori Kovery



Marge N. O’Vera



The Trial Anthology of Patient Stories



- Mortality at Day 29: 14.6% in placebo; 10.7% in Remdesivir
- No events at Day 29: 73.3% in placebo; 80% in Remdesivir
- No events in all time intervals: 48% in placebo; 58.8% in Remdesivir

Sliding DOOR Analyses

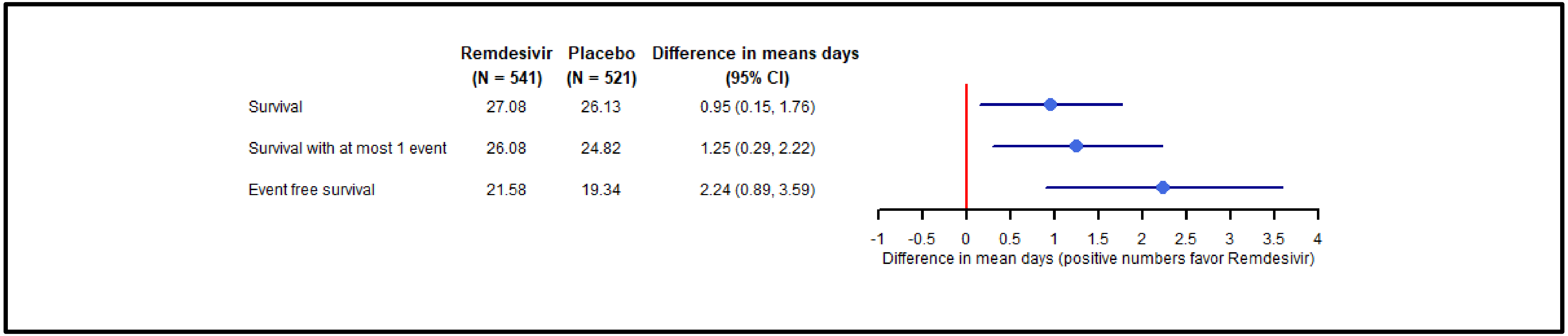
- Ranking strategy
  - First priority: DOOR at day 29
  - For survivors: further rank using average events per day (time-weighted) … larger average is worse

Estimate of DOOR probability of a more desirable result with Remdesivir vs. Placebo

55.9%

(95% CI: 52.7% - 59.1%)

ACTT-1 DOOR RMST: Cumulative Days Gained / Lost



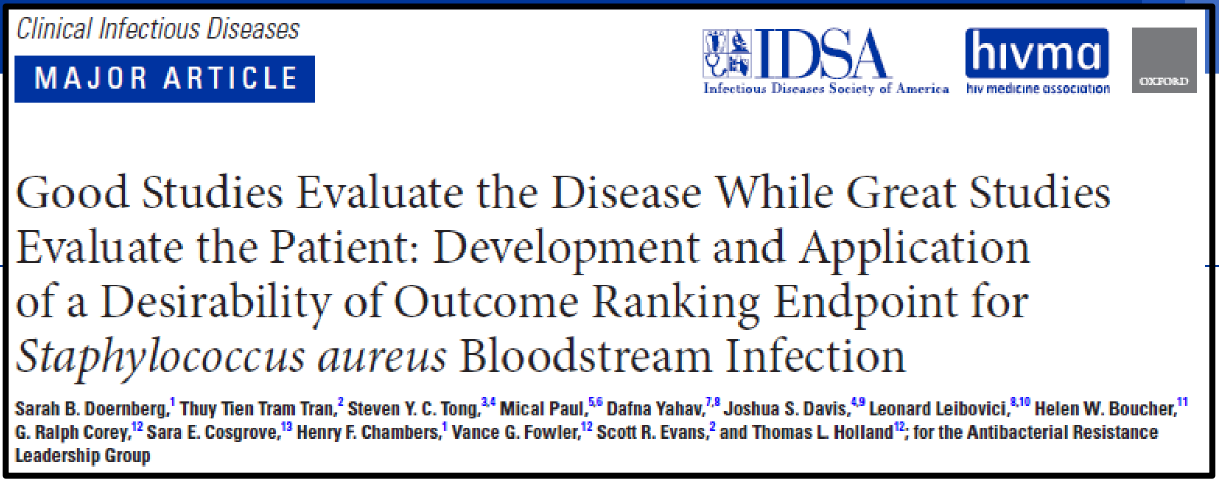
	Remdesivir (N = 541)	Placebo (N = 521)	Difference in means days (95% CI)
Survival	27.08	26.13	0.95 (0.15, 1.76)
Survival with at most 1 event	26.08	24.82	1.25 (0.29, 2.22)
Event free survival	21.58	19.34	2.24 (0.89, 3.59)

Defining the DOOR Outcome

- Goals:
  - Define gradations of patient response to recognize importantly different outcomes
  - Ensure gradations of response are clinically meaningful
  - Simplicity

- Evaluating the tradeoffs among individual outcomes, and the cumulative nature of benefits and harms on patients

- Methods that have been used to guide construction include
  - Conjoint analyses
  - Delphi analyses
  - Surveys of expert clinicians and patients
  - Discussion with regulators



Clinical Infectious Diseases

MAJOR ARTICLE

IDSA Infectious Diseases Society of America

hivma hiv medicine association

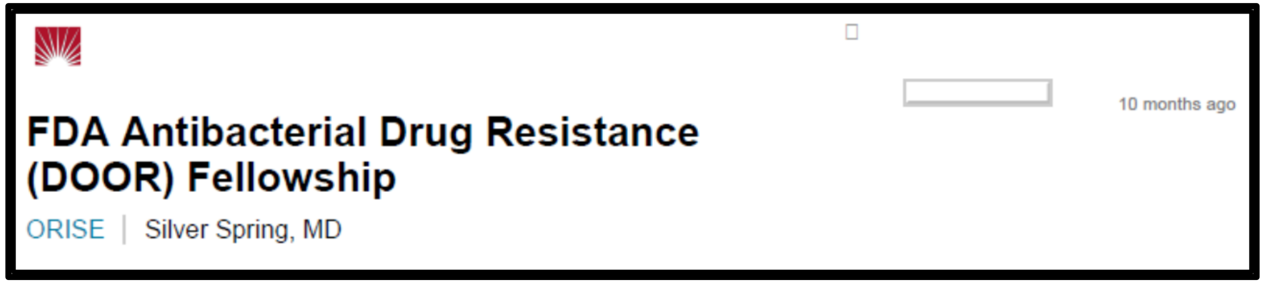
OXFORD

Good Studies Evaluate the Disease While Great Studies Evaluate the Patient: Development and Application of a Desirability of Outcome Ranking Endpoint for *Staphylococcus aureus* Bloodstream Infection

Sarah B. Doernberg,[1] Thuy Tien Tram Tran,[2] Steven Y. C. Tong,[3,4] Mical Paul,[5,6] Dafna Yahav,[7,8] Joshua S. Davis,[4,9] Leonard Leibovici,[8,10] Helen W. Boucher,[11] G. Ralph Corey,[12] Sara E. Cosgrove,[13] Henry F. Chambers,[1] Vance G. Fowler,[12] Scott R. Evans,[2] and Thomas L Holland[12]; for the Antibacterial Resistance Leadership Group

- ARLG conducted a pre-trial sub-study to develop DOOR in *Staphylococcus aureus* bacteremia
- 20 representative patient profiles (benefits, harms) constructed based on experiences observed in prior trials
- Profiles sent to 43 expert clinicians. They were asked to rank the patient profiles by desirability of outcome.
- Examined clinician consensus and component outcomes that drive clinician rankings

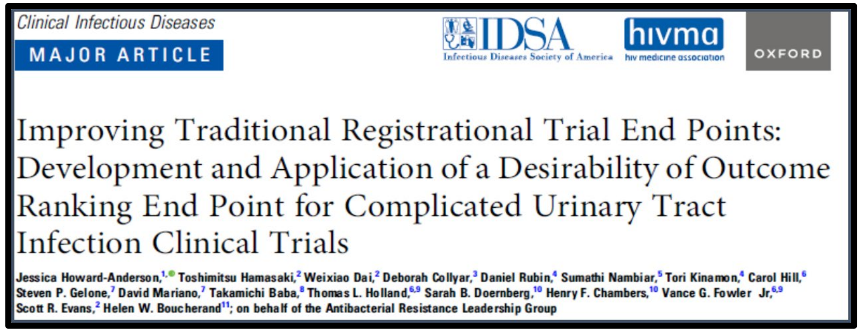
Regulator, Academic, Industry Collaboration



10 months ago

FDA Antibacterial Drug Resistance (DOOR) Fellowship

ORISE | Silver Spring, MD

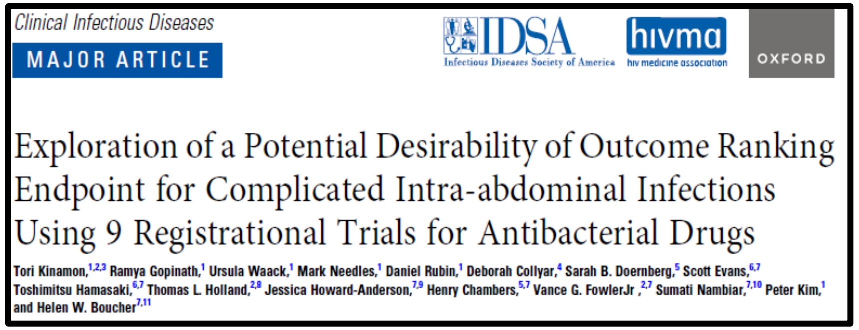


Clinical Infectious Diseases

MAJOR ARTICLE

Improving Traditional Registrational Trial End Points: Development and Application of a Desirability of Outcome Ranking End Point for Complicated Urinary Tract Infection Clinical Trials

Jessica Howard-Anderson,[1] Toshimitsu Hamasaki,[2] Weixiao Dai,[2] Deborah Collyar,[3] Daniel Rubin,[4] Sumathi Nambiar,[5] Tori Kinamon,[4] Carol Hill,[6] Steven P. Gelone,[7] David Mariano,[7] Takamichi Baba,[8] Thomas L. Holland,[6,9] Sarah B. Doernberg,[10] Henry F. Chambers,[10] Vance G. Fowler Jr,[6,9] Scott R. Evans,[2] Helen W. Boucherand[11]; on behalf of the Antibacterial Resistance Leadership Group

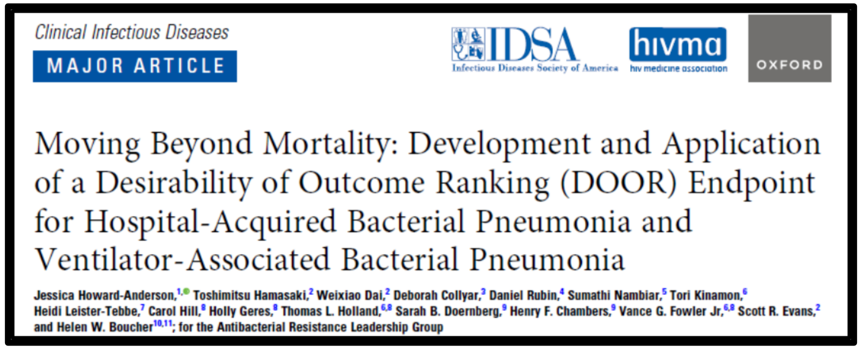


Clinical Infectious Diseases

MAJOR ARTICLE

Exploration of a Potential Desirability of Outcome Ranking Endpoint for Complicated Intra-abdominal Infections Using 9 Registrational Trials for Antibacterial Drugs

Tori Kinamon,[1,2,3] Ramya Gopinath,[1] Ursula Waack,[1] Mark Needles,[1] Daniel Rubin,[1] Deborah Collyar,[4] Sarah B. Doernberg,[5] Scott Evans,[6,7] Toshimitsu Hamasaki,[6,7] Thomas L. Holland,[2,8] Jessica Howard-Anderson,[7,9] Henry Chambers,[5,7] Vance G. FowlerJr ,[2,7] Sumati Nambiar,[7,10] Peter Kim,[1] and Helen W. Boucher[7,11]

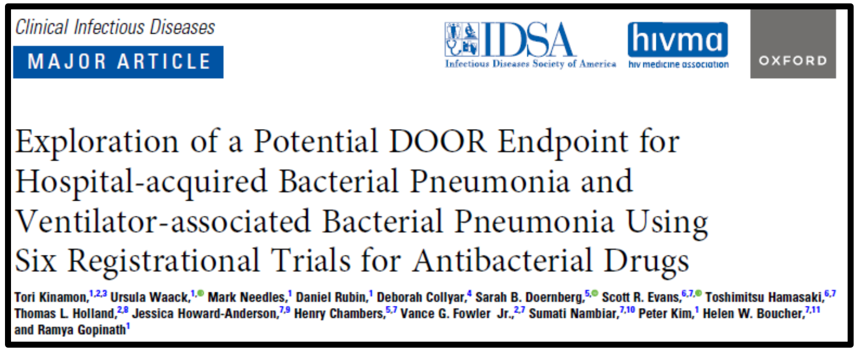


Clinical Infectious Diseases

MAJOR ARTICLE

Moving Beyond Mortality: Development and Application of a Desirability of Outcome Ranking (DOOR) Endpoint for Hospital-Acquired Bacterial Pneumonia and Ventilator-Associated Bacterial Pneumonia

Jessica Howard-Anderson,[1] Toshimitsu Hamasaki,[2] Weixiao Dai,[2] Deborah Collyar,[3] Daniel Rubin,[4] Sumathi Nambiar,[5] Tori Kinamon,[6] Heidi Leister-Tebbe,[7] Carol Hill,[8] Holly Geres,[8] Thomas L. Holland,[6,8] Sarah B. Doernberg,[9] Henry F. Chambers,[9] Vance G. Fowler Jr,[6,8] Scott R. Evans,[2] and Helen W. Boucher[10,11]; for the Antibacterial Resistance Leadership Group



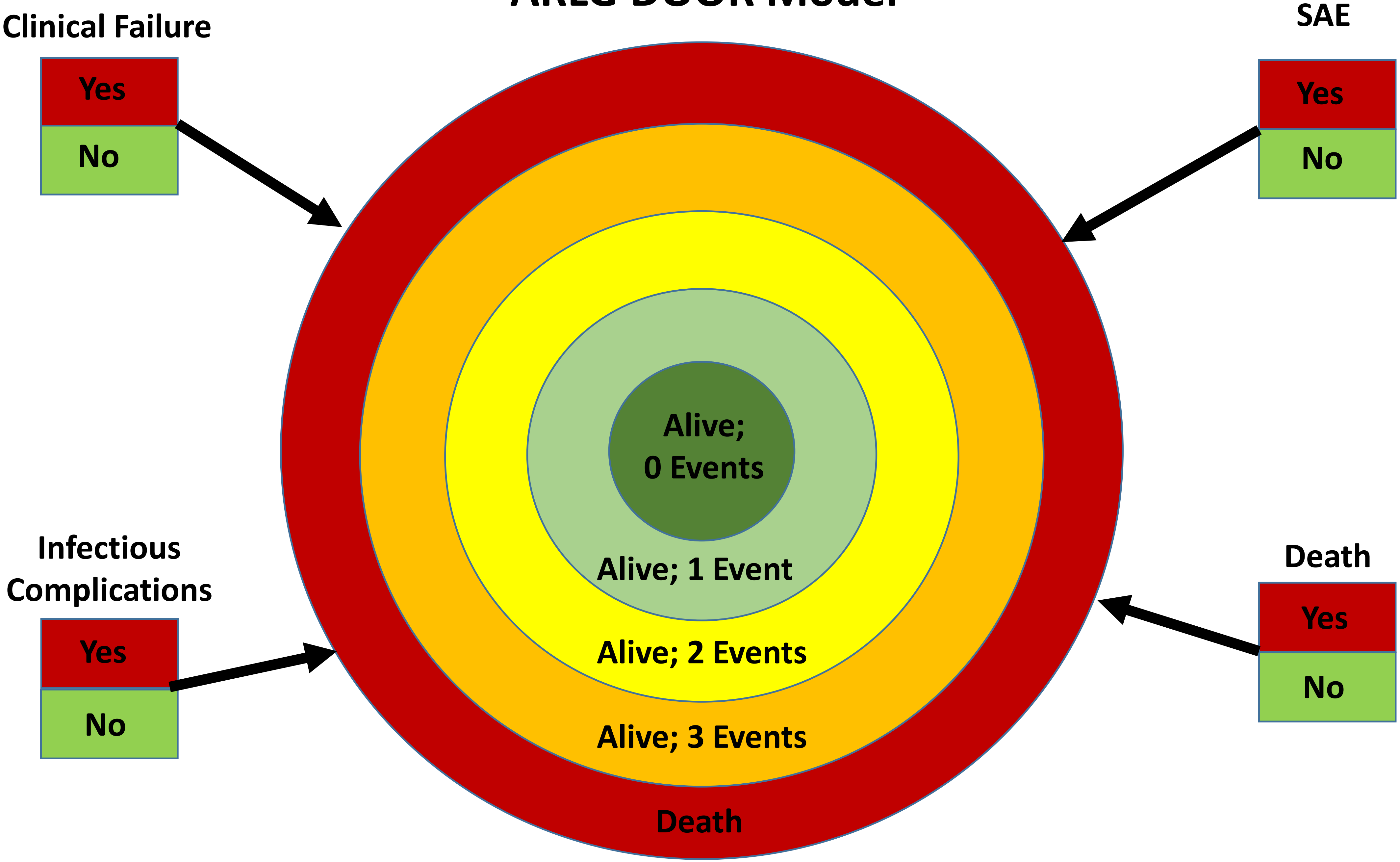
Clinical Infectious Diseases

MAJOR ARTICLE

Exploration of a Potential DOOR Endpoint for Hospital-acquired Bacterial Pneumonia and Ventilator-associated Bacterial Pneumonia Using Six Registrational Trials for Antibacterial Drugs

Tori Kinamon,[1,2,3] Ursula Waack,[1] Mark Needles,[1] Daniel Rubin,[1] Deborah Collyar,[4] Sarah B. Doernberg,[5] Scott R. Evans,[6,7] Toshimitsu Hamasaki,[6,7] Thomas L. Holland,[2,8] Jessica Howard-Anderson,[7,9] Henry Chambers,[5,7] Vance G. Fowler Jr.,[2,7] Sumati Nambiar,[7,10] Peter Kim,[1] Helen W. Boucher,[7,11] and Ramya Gopinath[1]

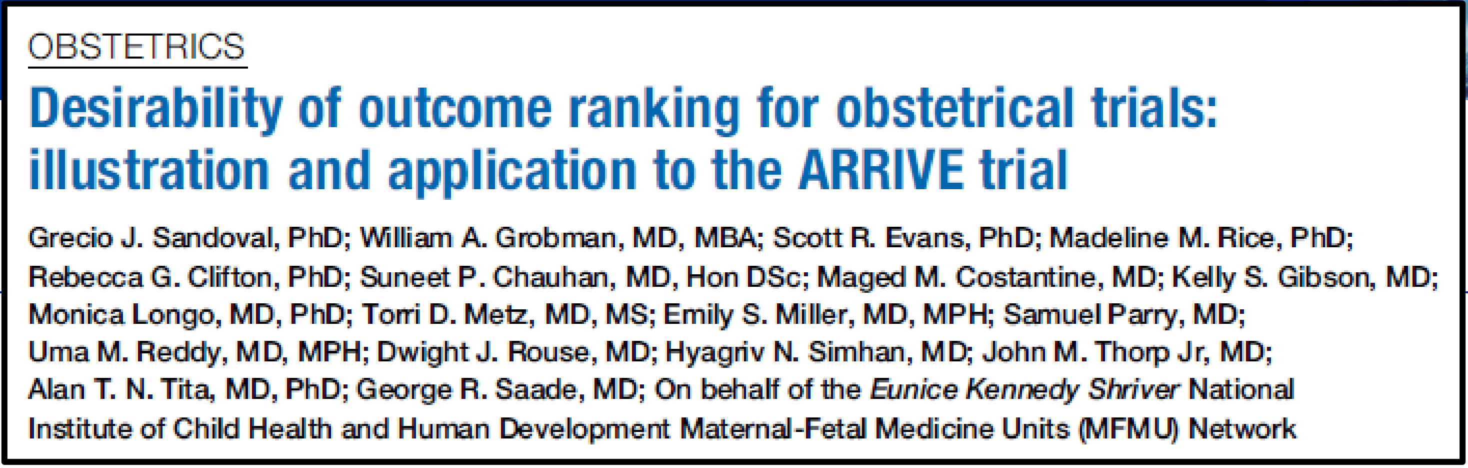
ARLG DOOR Model



Cardiovascular Prevention DOOR (CAR-DOOR?)

- Events of interest
  - Death
  - Stroke
  - MI
  - Major bleeding

- DOOR
  - Alive with no events
  - Alive with 1 non-disabling event
  - Alive with >1 non-disabling event
  - Alive with disabling event
  - Death



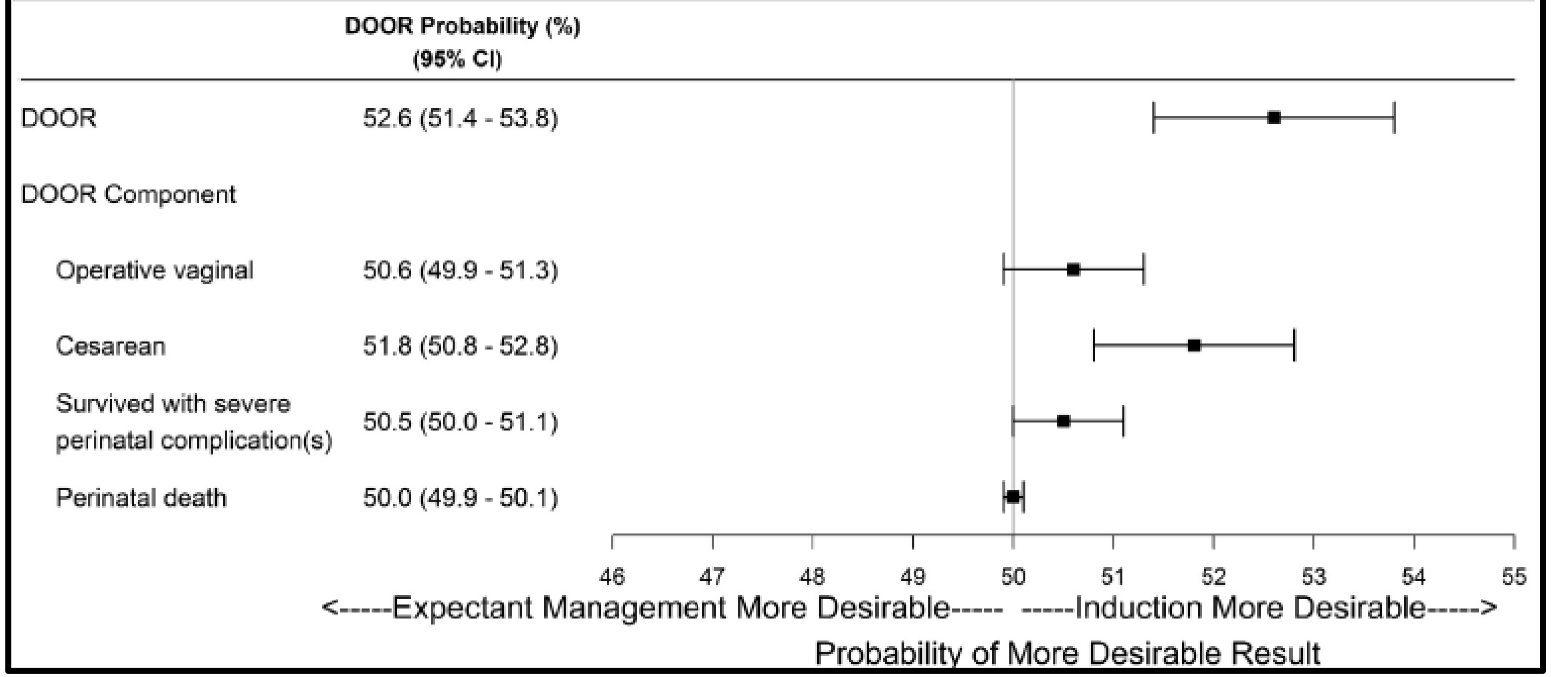
OBSTETRICS

Desirability of outcome ranking for obstetrical trials: illustration and application to the ARRIVE trial

Grecio J. Sandoval, PhD; William A. Grobman, MD, MBA; Scott R. Evans, PhD; Madeline M. Rice, PhD; Rebecca G. Clifton, PhD; Suneet P. Chauhan, MD, Hon DSc; Maged M. Costantine, MD; Kelly S. Gibson, MD; Monica Longo, MD, PhD; Torri D. Metz, MD, MS; Emily S. Miller, MD, MPH; Samuel Parry, MD; Uma M. Reddy, MD, MPH; Dwight J. Rouse, MD; Hyagriv N. Simhan, MD; John M. Thorp Jr, MD; Alan T. N. Tita, MD, PhD; George R. Saade, MD; On behalf of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development Maternal-Fetal Medicine Units (MFMU) Network

FIGURE 1

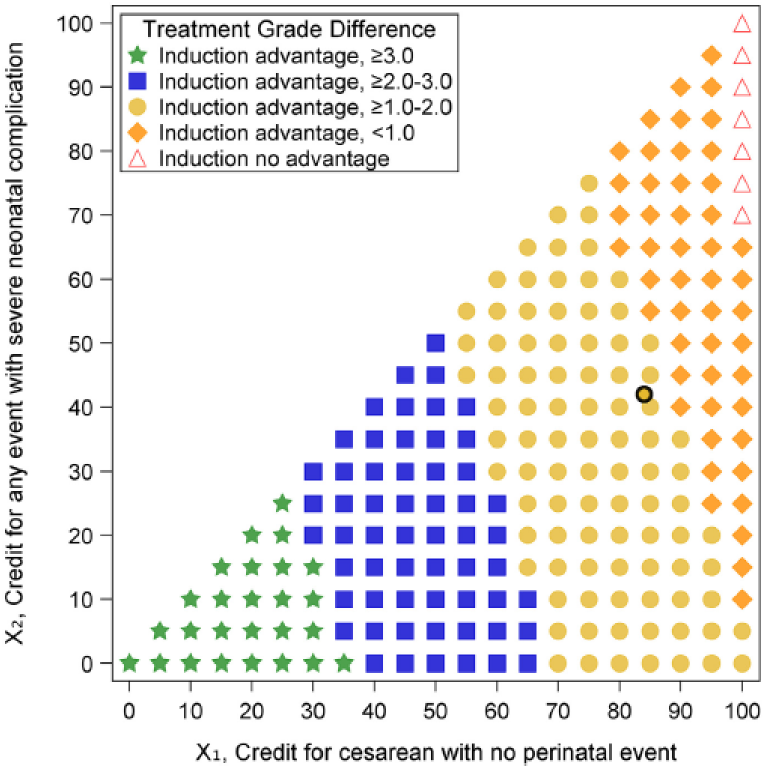
DOOR forest plot for induction vs expectant management from the ARRIVE trial

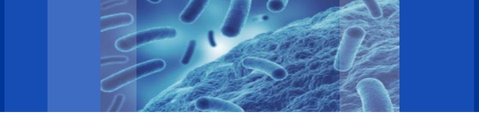


	DOOR Probability (%) (95% CI)
DOOR	52.6 (51.4 - 53.8)
DOOR Component	
Operative vaginal	50.6 (49.9 - 51.3)
Cesarean	51.8 (50.8 - 52.8)
Survived with severe perinatal complication(s)	50.5 (50.0 - 51.1)
Perinatal death	50.0 (49.9 - 50.1)

FIGURE 2

Treatment effects for partial credit grades from the ARRIVE trial





Metabolic Health

- Diet-induced adiposity causes metabolic stress, systemic inflammation and fibrosis
- This affects the:
  - Arteries (hypertension, CVD, CAD, PVD)
  - Heart (HFPEF)
  - Liver (NAFDL)
  - Pancreas (T2D)
  - Kidney (CKD)
  - Brain (cognitive decline)
  - and other organs
- These afflictions have shared biology and occur in the same individual patients
- Yet evaluation of treatments is organ-specific
- However, the benefits of treatment may be broader, affecting multiple organs, resulting in greater overall benefits to the patient than organ-specific evaluations would uncover

Rigor, Robustness, and Discipline

“Clinical trials are the best medical invention in history.

They don’t make them like they used to.”

The current President of the Society for Clinical Trials

Clinical Trials

Strengths, weaknesses, opportunities, and threats (SWOT) analyses

Strengths

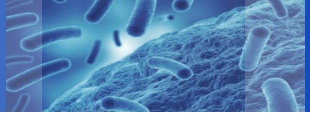
- Randomization (the foundation for statistical inference)
- Blinding
- Control groups
- Prospective observation
- ITT (protects the benefits of randomization; assesses pragmatic questions)
- Standardization of measurement and procedures
- A comprehensive protocol outlining scientific strategy and operational approaches
- Pre-specification of endpoints and hypotheses providing multiplicity context and a framework by which to control errors and provide the correct coverage probabilities
- Protection of trial participants, society, and trial integrity via independent monitoring of benefits and harms by DSMBs
- Registration which increases transparency and helps to curtail selective reporting

Weaknesses

- Expensive and resource intensive
- Time-consuming
- May lack generalizability and clinical applicability if not pragmatic, for example with use of restrictive entry criteria, surrogate rather than clinical endpoints, other than ITT analysis sets, and marginal analyses of endpoints rather than patient-centric evaluation

Opportunities

- Greater pragmatism: more relevant questions and answers for clinical decision-making
- Emerging technologies to timely obtain important data
- Improving clinical trials education with emphasis on fundamentals of the scientific principles and operations
- Improvement to DSMB processes



Threats

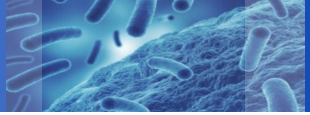
- Innate desire to do things faster and cheaper which can threaten objectivity, and result in studies with low replicability, integrity, and applicability

- Insufficient education regarding the role of clinical trials as a scientific instrument rather than a commercial tool
  - A “successful trial” has been perverted to imply a positive trial, rather than a trial that addresses important questions and gets robust answers to those questions regardless of the directionality and magnitude of the treatment effects
  - We should be objective about the objective, i.e., striving to correctly “determine whether” an effect exists rather than “to establish” that one does

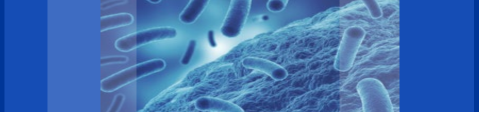
- A decline of academic leadership in clinical trials. David DeMets and FDA Commissioner Rob Califf wrote “where have the academics gone?”

Threats

- Misinformation, disinformation, and incomplete information regarding the merits of trending methods and technologies
  - Some approaches are labeled as innovative, presented with a degree of commercialism rather than scientific objectivity. Closer evaluation reveals that they are fancy ways of lowering the usual integrity and evidentiary standards and introduce greater uncertainty through concessions of:
    - (i) randomized evidence for non-randomized evidence;
    - (ii) controlled evidence for uncontrolled evidence’
    - (iii) robustness via greater reliance upon assumptions,
    - (iv) objectivity via the incorporation of beliefs,
    - (v) transparency relenting to black box approaches, and
    - (vi) the theoretical foundation for statistical inference.
  - See efforts to protect the scientific community from compromises in scientific rigor and the decline in integrity in e.g., Emerson and Fleming telling "the rest of the story" and Collins, Bowman, and Landray, and Peto’s “The magic of randomization versus the myth of real-world evidence”.



We share a duty in protecting the ideals.



Guiding principles for maximizing pragmatism, robustness, replicability, objectivity, and transparency

Patient-centricity and pragmatism preservation
Analyze the patient story / journey; recognize the cumulative nature of effects
Distinguish important gradations of patient response
Best Practices
Composite endpoints (integrated presentation of component analyses)
Multi-outcome / benefit:risk analyses (analyses based on the absolute risk scale providing a common scale for simultaneous interpretation of multiple outcomes)
Ordinal outcomes (cumulative analyses)
Statistical Integrity and Discipline
Robustness: avoid / minimize reliance upon assumptions e.g., common odds, distribution of treatment effects, specification of model form, for analysis validity
Incorporate competing risks
Intention-to-treat principle with full analysis set for all outcomes (clarity of generalizability; known applicability at the time of treatment initiation)
Objectivity: free from subjective beliefs
Defined population parameters and estimands
Theoretical foundation for the confirmatory evidence standard • Unbiased estimates of treatment effects • Correct coverage probability for confidence interval estimation • Error control in hypothesis testing
Incorporation of ties into rank-based statistics utilizing pair-wise comparisons
Implementation of rank-based and grade-based analyses of treatment contrast
Evaluation of robustness of grade-based analyses

The DOOR paradigm was designed to meet these principles.

A comprehensive statistical analysis plan was developed in line with the principles.

Comprehensive Statistical Analysis Plan

- To optimally reflect the experience of patients we must change the order of operations, composing information within patient

- Composite endpoints have challenges
  - Reporting data on all components is advised to reveal and understand the full story
  - Comprehensive understanding requires careful evaluation the relative importance of the components, which components are driving the observed effects, and whether the effects on the components go in similar vs. opposing directions

- An absolute risk scale is required to interpret multiple outcomes together

Freely available online software was designed to produce all of the output with the statistical analysis plan.

The software will be submitted as a regulatory science tool.

Milken Institute School of Public Health
THE GEORGE WASHINGTON UNIVERSITY

THE BIOSTATISTICS CENTER

A Patient-Centric Paradigm for Clinical Research: The DOOR is Open

Design and analysis of clinical trials and other research using the DOOR methodology

Toshimitsu Hamasaki, PhD, MS, Pstat®
Scott R. Evans, PhD, MS

The Biostatistics Center | Department of Biostatistics and Bioinformatics
Milken Institute School of Public Heath | The George Washington University

10 April 2025 at 9:30-12:00 @ Medical University of Vienna
Wiener Biometrische Sektion der Internationalen Biometrischen Gesellschaft Region Österreich – Schweiz

A Patient-Centric Paradigm for Clinical Research: DOOR is Open

Objectives

- To discuss statistical methods for design and analysis of the DOOR methodology in clinical trials and other clinical trials, that have been adapted in the web-based tools
  - ❑ DOOR analyses: Rank-based and grade-based analyses
  - ❑ Sample size and power assessment

- To demonstrate, using web-based tools

Application of the DOOR

DORI-05: doripenem vs. levofloxacin

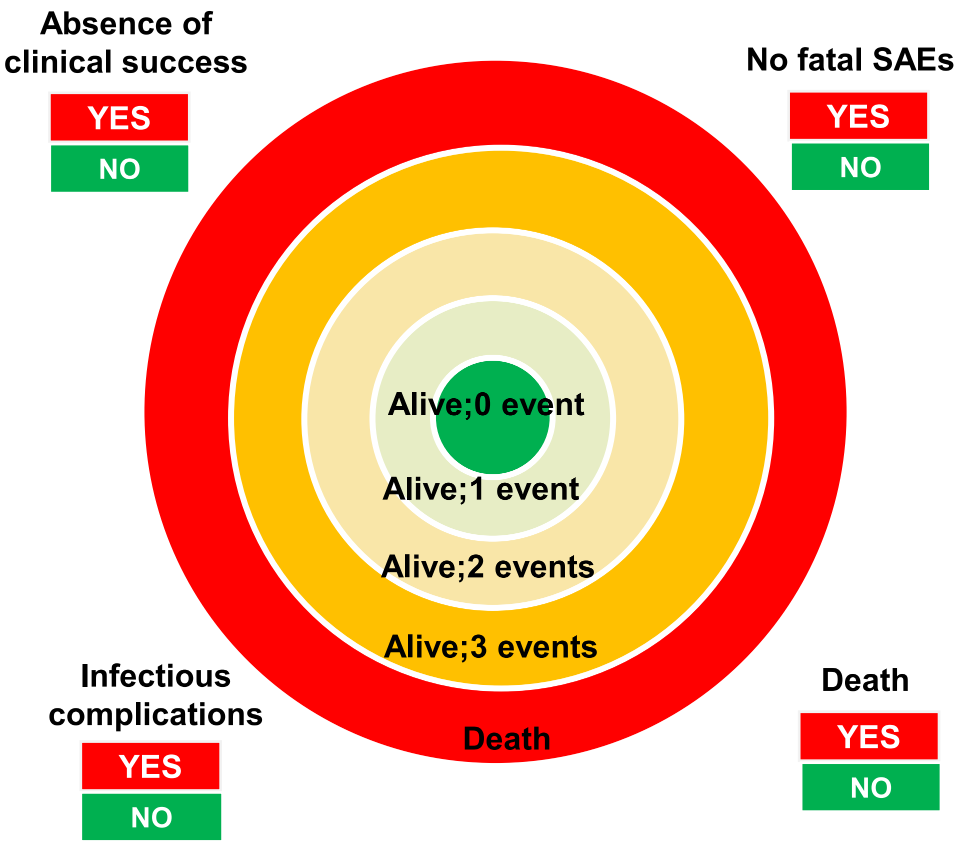
How to analyze the DOOR outcome?

DOOR rank category	Doripenem		Levofloxacin	
	Freq	Prop (%)	Freq	Prop (%)
Alive with no events	263	70.3	253	67.6
Alive with 1 event	93	24.9	111	29.7
Alive with 2 events	16	4.3	9	2.4
Alive with 3 events	1	0.3	1	0.3
Death	1	0.3	0	0
Total	374	100	374	100

From a randomized double-blind clinical trial that evaluated whether intravenous (IV) administration of doripenem (DORI) was inferior to IV administration of levofloxacin in patients with cUTI (complicated urinary tract infection) (Naber KG et al. Antimicrob Agents Chemother 2009; 53:3782-3792)
(Howard-Anderson J at. Clin Infect Dis. 2023; 76:e1157-e1165)

ARLG
Antibacterial Resistance Leadership Group

DOOR for cUTI



Renal or intraabdominal abscess; Septic shock; Bacteremia; Recurrent UTI or pyelonephritis; *C. difficile*

Common Statistical Concerns

Ordinal outcomes analysis

Analysis		Concern
Responder analysis	● Dichotomize ordinal outcome to a binary outcome i.e., responder vs non-responder. ● Estimate odds ratio of responder vs non-responder between groups; conduct associated hypothesis testing for the odds ratio	● Loss of information and potentially power by ignoring finer but important gradations of patient status ● Robustness: assumption-reliant (e.g., the model linearity) ● Interpretations are not intuitive
Proportional odds regression	● Estimate common odds ratio by proportional odds regression; Conduct associated methods for hypothesis testing and interval estimation	● Robustness: assumption-reliant (e.g., the model linearity, proportional odds) ● Interpretations are not intuitive

The DOOR Methodology

Key principles

Sensitivity

❑ Simple tools for assessing the robustness of results

Robustness

❑ Avoidance of reliance on strong and unconfirmable assumptions

Unbiased estimators

❑ Avoidance of over or under-estimation of treatment effects

Simple implementation

❑ No advanced mathematical skills or programming knowledge required

KEY PRINCIPLES

Error controls

❑ Avoidance of false positive or false negative results

Generalizability

❑ Clearer estimands and population

Intuitive reporting & presentation

❑ Enhanced understanding, informed decision-making, better communication with patients

Intuitive measures

❑ A clear and easily interpretable summary measure

Analytical Methods

Analyses of DOOR outcomes

DOOR probability-based analysis

❑ **Rank-based analysis**: Distribution-free

❑ Pairwise comparisons at individual patient level

❑ The DOOR probability

- ➢ A patient randomly selected from one group has a more desirable outcome than a patient randomly selected from the other group
- ➢ 50% if two DOOR outcomes are identical between groups
- ➢ Estimated by Wilcoxon-Mann-Whitney (WMW) statistic

Partial credit analysis

❑ **Grade-based analysis**

- ➢ Evaluation of the relative importance of DOOR outcome categories: robustness analyses

❑ Methods for continuous outcomes as if they were continuous after assigning grading keys

- ➢ Welch t-statistic based method

The DOOR Probability

The DOOR probability-based analysis

- The DOOR probability $\pi_{\mathrm{E}\geq\mathrm{C}} = \mathbb{P}\left[Y^{E} > Y^{\mathrm{C}}\right] + 1/2\mathbb{P}\left[Y^{\mathrm{E}} = Y^{\mathrm{C}}\right] = 1 - \pi_{\mathrm{C}\geq\mathrm{E}}$
  - A probability that a participant’s outcome in Experimental (E) Y^{E} is more desirable than that of a participant in Control (C) Y^{C} (Evans et al. 2015; Evans and Follmann 2016)
  - Unbiased, estimated by WMW statistic (Wilcoxon 1945; Mann and Whitney 1947) corrected for ties
  - $\pi_{\mathrm{E}\geq\mathrm{C}} = 0.5$ if Y^{E} and Y^{C} are identically distributed, but $\pi_{\mathrm{E}\geq\mathrm{C}} = 0.5$ does not support this
  - Does not depend on the specific potential outcome pairings; Population causal effect, not individual causal effect (Fay et al 2018)
  - Can be applicable to continuous, binary, and time-to-event outcomes as well

Simonoff JS et al. Biometrics 1986; 42:895-907. Fay MP et al. Stat Med 2018;37:2923-2937

The DOOR Probability: Confidence Intervals (CIs)

The DOOR probability-based analysis

Methods	Feature	Reference
Wald-type	● Easy to construct ● Symmetric CIs; the lower or upper limit may fall outside of [0, 1] in extreme cases.	Ryu, Agresti. Stat Med 2008; 27:1703-1717.
Halperin et al (1989)	● Easy to construct ● Asymmetric CI; given by solving the quadratic inequity (chi-square statistic)	Halperin M. et al. Biometrics 1989; 45:509-521.
tahn^{-1} transformation	● Easy to construct ● Asymmetric CI; Construct a Wal type CIs on a logit scale and then back-transform it to the original scale	Edwardes M. Biometrics 1995; 51:571-578
Score/Pseudo-Score; Likelihood	● Require an iterative procedure ● Asymmetric CIs; Need to restricted maximum likelihood estimates of category proportions given a value of $\pi_{\mathrm{E}\geq\mathrm{C}}$	Ryu, Agresti. Stat Med 2008; 27:1703-1717.
Bootstrap percentile	● Easy implementation, but computationally intensive compared to other methods ● Asymmetric CIs	van Duin D et al CID 2018; 66:163-171.

The DOOR Probability: Recommendations on CIs

The DOOR probability-based analysis

- **Recommend using the Halperin et al. method (or $\tanh^{-1}$ transformation based method)**
  - ❑ Works well (in terms of coverage probability); Easy to construct with the closed form, without the need for an iterative procedure
  - ❑ Tends to be "liberal" when the group sample size is extremely unbalanced (4:1 allocation ratio or higher) and sample size is smaller; Use the Halperin et al. (1989) with the pseudo-score approach
- **Do not use Wald-type CIs.**
  - ❑ very liberal, failing to control for the coverage probability at the desired confidence level. particularly in small sample sizes and/or unbalanced sample sizes between groups
  - ❑ The lower or upper limit of these intervals may fall outside of [0, 1], in near-extreme cases; the DOOR probability distributions is not symmetric
- **Occasionally can use score, pseudo-score, and likelihood CIs**
  - ❑ Work well, but generally "conservative" with small sample sizes
  - ❑ Require an iterative procedure; Fail to find the restricted likelihood estimates with small sample sizes
  - ❑ Fails to find the restricted likelihood estimates with small sample sizes.
- **Better to avoid using Bootstrap percentile CIs**
  - ❑ Generally "liberal", better than Wald-type CIs but never better than CIs by Halperin et al.

The DOOR Probability: Other CIs

The DOOR probability-based analysis

- CI based on the statistic discussed in Brunner and Munzel (2000)
  - ❑ Symmetric CI , centered at $\hat{\pi}_{\mathrm{E}\geq\mathrm{C}}$: the same disadvantages of the Wald-type CIs
    - ➢ Ex. The lower or upper limit may fall outside of [0, 1] in extreme cases
  - ❑ Performs better than the Wald-type CIs, but not as well as the Halperin, score, or pseudo-score CIs (Ryu and Agresti, 2008)

Brunner E, Munzel U. Biometrical Journal 2000; 42; 17-25

DOOR Probability: Hypothesis Tests

The DOOR probability-based analysis

- **Hypothesis (Superiority)**

One-sided $\begin{cases} \mathrm{H}_0 : \pi_{\mathrm{E} \geq \mathrm{C}} \leq \delta_0 \\ \mathrm{H}_1 : \pi_{\mathrm{E} \geq \mathrm{C}} > \delta_0 \end{cases}$ two-sided $\begin{cases} \mathrm{H}_0 : \pi_{\mathrm{E} \geq \mathrm{C}} = \delta_0 \\ \mathrm{H}_1 : \pi_{\mathrm{E} \geq \mathrm{C}} \neq \delta_0 \end{cases}$

δ_0 : 0.5 generally chosen

- Natural to implement the WMW test for the hypothesis test for DOOR probability as the DOOR probability is equivalent to the WMW statistic
- A concern: The normal approximation to the WMW statistic to calculate p-values may be often inaccurate when the outcomes are heavily tied?
  - ❑ **Correction for continuity** by shifting the value of the statistic (e.g., see Lehman and D'Abrera (1975))
  - ❑ **the t-approximation** for the p-value calculation to improve the normal approximation.
  - ❑ However, these p-values generally tend to be "conservative", less than the desired significance level

Lehmann E, D'Abrera H (1975). Statistical Methods Based on Ranks. Holden-Day.

DOOR Probability: Recommendations on Hypothesis Test

The DOOR probability-based analysis

- **Recommend using the normal approximation WMW test without continuity correction**
  - works well, controlling the Type I error probability at the desired significance level
  - tends to be conservative when the total sample size is 50 and the group sample size unbalanced, such as 7:3 allocation ratio or higher
  - p-values from using continuity correction and/or t-approximation generally tend to be “conservative”, less than the desired significance level
- **Do not use O’Brien–Castelloe method or log Win odds-based method**
  - fails to control the Type I error probability at the desired significance level appropriately: the Type I error probability is inflated

The DOOR Probability-based Analysis

Limitations

- The DOOR probability may not provide the appropriate amount of influence to each specific rank category.
  - An individual pairwise comparison is labeled a “more desirable”, “less desirable”, or “tie”, depending on whether the experimental group patient had a more desirable, less desirable, or tied DOOR outcome.
  - Ex. “Alive with no event vs. Alive with 1 event” = “Alive with no event vs. Death”

	Doripenem	Levofloxacin
DOOR rank category	Freq	Freq
Alive with no events	263	253
Alive with 1 event	93	111
Alive with 2 events	16	9
Alive with 3 events	1	1
Death	1	0
Total	374	374

Grade-based Analysis: Partial Credit Analysis

Robustness of the DOOR probability-based analysis

- Assigns the relative importance to each category directly.

DOOR rank category	Score
Alive with no events	100
Alive with 1 event	0≤Partial Score 1≤100
Alive with 2 event	0≤Partial Score 2≤100
Alive with 3 events	0≤Partial Score 3≤100
Death	0

Partial Score 1
≥
Partial Score 2
≥
Partial Score 3

- Conduct an analyses consist of estimating the between-group difference in mean scores on a 100-point scale as if the outcome was “**continuous**”
  - Welch’s t-statistic based approach for two-group comparisons.
    - The range of scores is limited from 0 to 100, and the possible values of scores are limited: Using the Welch t-statistic can at least protect against false positive conclusion.
  - Can still implement the rank-based method
    - Assigning the same partial credit score to different adjacent rank categories results in those categories being combined into one category.

Analytical methods

ARLG recommendations

Analysis	Outcome	Statistical method
Descriptive analysis	● DOOR ● Components	● Summary distribution table by intervention group ● Bar-chart by intervention group
	● DOOR and Components	● Anthology of Patient Stories (APS) plot
Rank-based analysis: DOOR probability	● DOOR ● Components	● Forest Plot of estimates of the DOOR probability for the DOOR outcome and respective components
	● DOOR	● Forest plot of the estimates for the cumulative DOOR probability based on sequential dichotomization of the DOOR outcome
Grade-based Analysis: Partial Credit	● DOOR	● Welch’s t-statistic based analysis ● Scatter plot of the differences in mean partial credit between interventions vs the corresponding DOOR probabilities

Online tools for implementing DOOR analyses

DOOR analysis apps

	Standard Edition	Professional Edition
Data Input	Summary table by group	Individual patient-level data
Analysis		
1. Descriptive analysis Summary table Bar-chart Anthology of patient stories plot		

Power-based approach: Superiority clinical trials

Closed form sample size

Methods	Feature	Reference
Tang (2011)	● Requires category proportions ● Assumes in $n/(n-1) \rightarrow 1$ for large sample size (this approximation works well if the sample size is extremely small, e.g., 10)	Tang Y. Stat Med 2011; 30:3461-3470.
Zhang et al (2008)	● Requires category proportions ● Assumes $\sigma_A \approx \sigma_0$: the approximation becomes inaccurate when the DOOR probability to be detected is far from 50%.	Zhao YD et al. Statistics in Medicine, 27, 462–468
Noether (1987)	● Requires DOOR Probability to be detected, not category proportions ● Ignores ties → larger variance → larger sample size ● Available in commercial software (nQuery, SAS etc)	Noether GE. J Amer Stat Assoc, 1987; 82:645–647
$\tanh^{-1}$ transformation based method/O'Brien and Castelloe (2006)	● Requires category proportions ● Assumes in $n/(n-1) \rightarrow 1$ for large sample size ● Available in commercial software (SAS) ● Related to Win Odds	O'Brien RG, Castelloe JM. SAS Users Group Int Conf, SAS Institute 2006

The DOOR Probability: Other methods for sample size calculations

The DOOR probability-based analysis

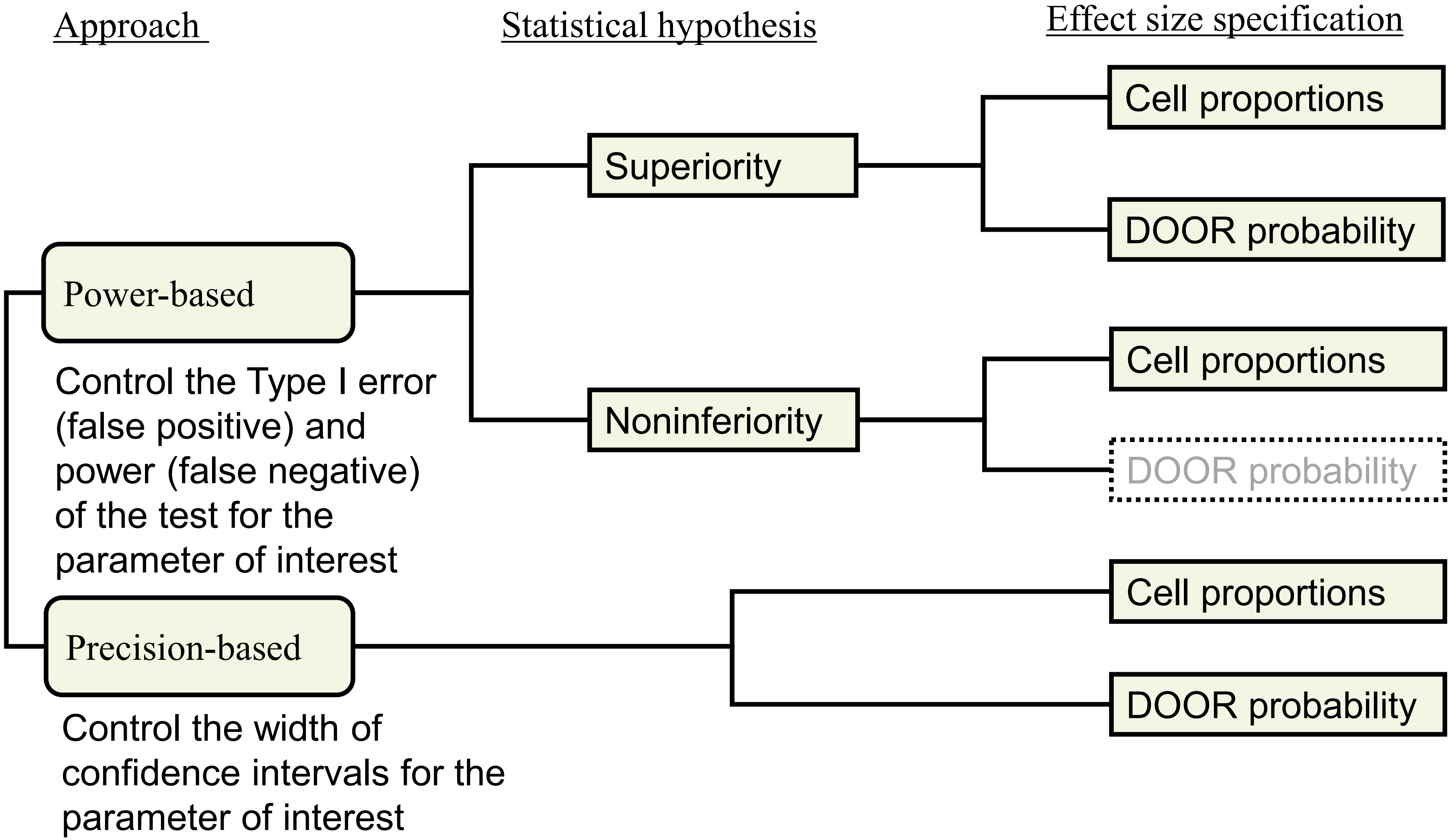
- Kolassa (1995)
  - Introduced a method for calculating power by approximating the unconditional distribution of the WMW statistic under the null and alternative hypotheses using the first four moments of the distribution.
    - Lacks a closed-form solution for sample size calculation- need for an iterative procedure
    - Works well for equally sized groups but performs poorly when the intervention groups are of unequal sizes or when the sample size is small (Tang 2015).
- Wellek (2017); Zho, Zou, and Choi (2022)
  - Discussed WMW-based methods for sample size calculation in clinical trials with ordinal outcomes for superiority and/or noninferiority trial designs.
  - Proposed closed-form solutions
  - Were unable to replicate their results- so did not adapt them in the web tools

Kolassa JE. Stat Med 1995; 14:1577–1581. Tang Y. Commun Stat-Simul Comput 2015; 45:240–251. Wellek S. Stat Med 2017; 36:799-812. Zou G et al. Stroke 2022; 53: 3025-3031.

Designing clinical trials with the DOOR methodology

Power and sample size determination

WMH statistic-based approach- Tang (2011), Noether (1987) and their extensions



Approach

Statistical hypothesis

Effect size specification

Power-based

Control the Type I error (false positive) and power (false negative) of the test for the parameter of interest

Superiority

Cell proportions

DOOR probability

Noninferiority

Cell proportions

DOOR probability

Precision-based

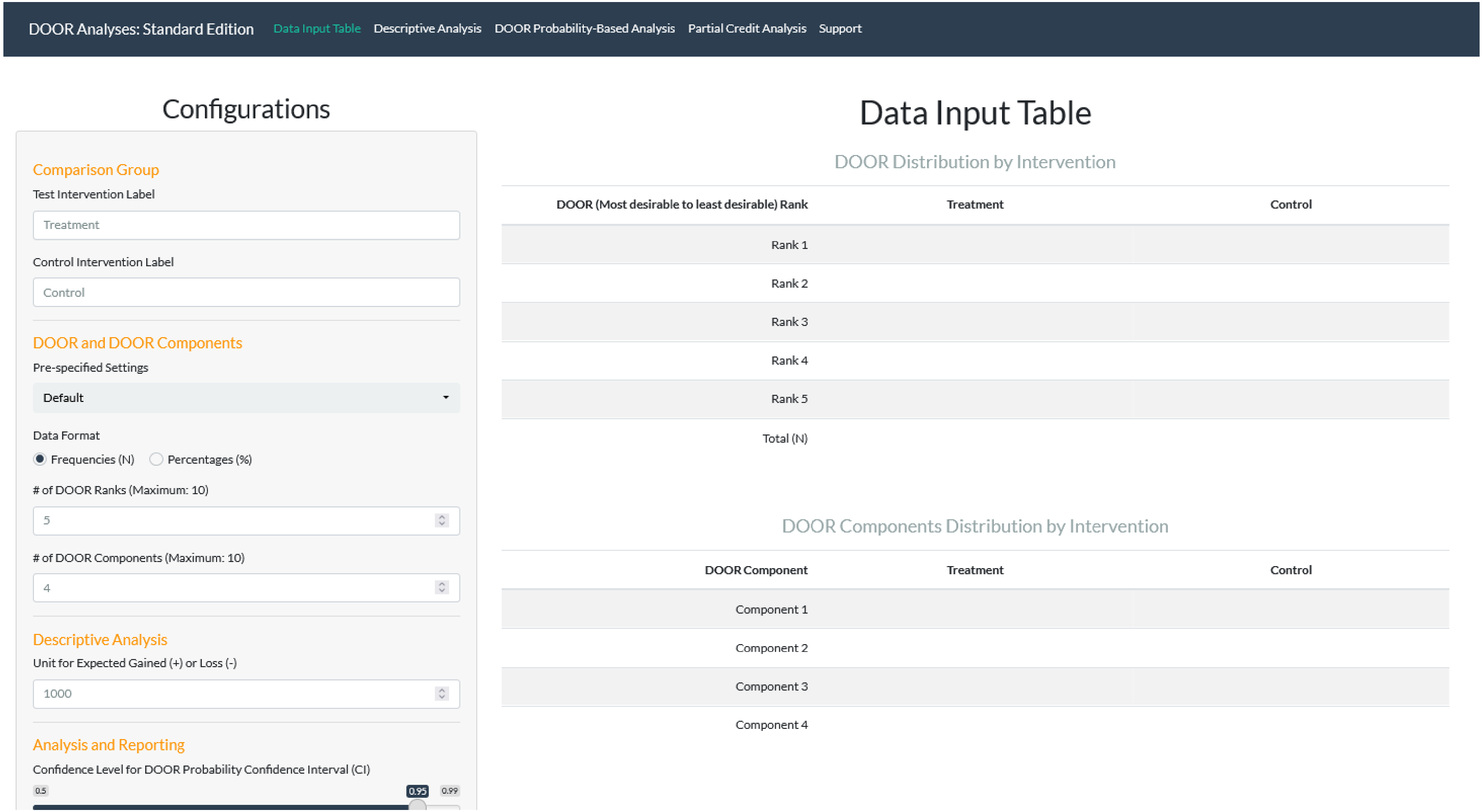
Control the width of confidence intervals for the parameter of interest

Cell proportions

DOOR probability

Online tools for implementing DOOR analyses

Analysis app standard edition demonstration



Configurations

Comparison Group

Test Intervention Label

Treatment

Control Intervention Label

Control

DOOR and DOOR Components

Pre-specified Settings

Default

Data Format

● Frequencies (N) ○ Percentages (%)

# of DOOR Ranks (Maximum: 10)

5

# of DOOR Components (Maximum: 10)

4

Descriptive Analysis

Unit for Expected Gained (+) or Loss (-)

1000

Analysis and Reporting

Confidence Level for DOOR Probability Confidence Interval (CI)

0.5 0.95 0.99

Data Input Table

DOOR Distribution by Intervention

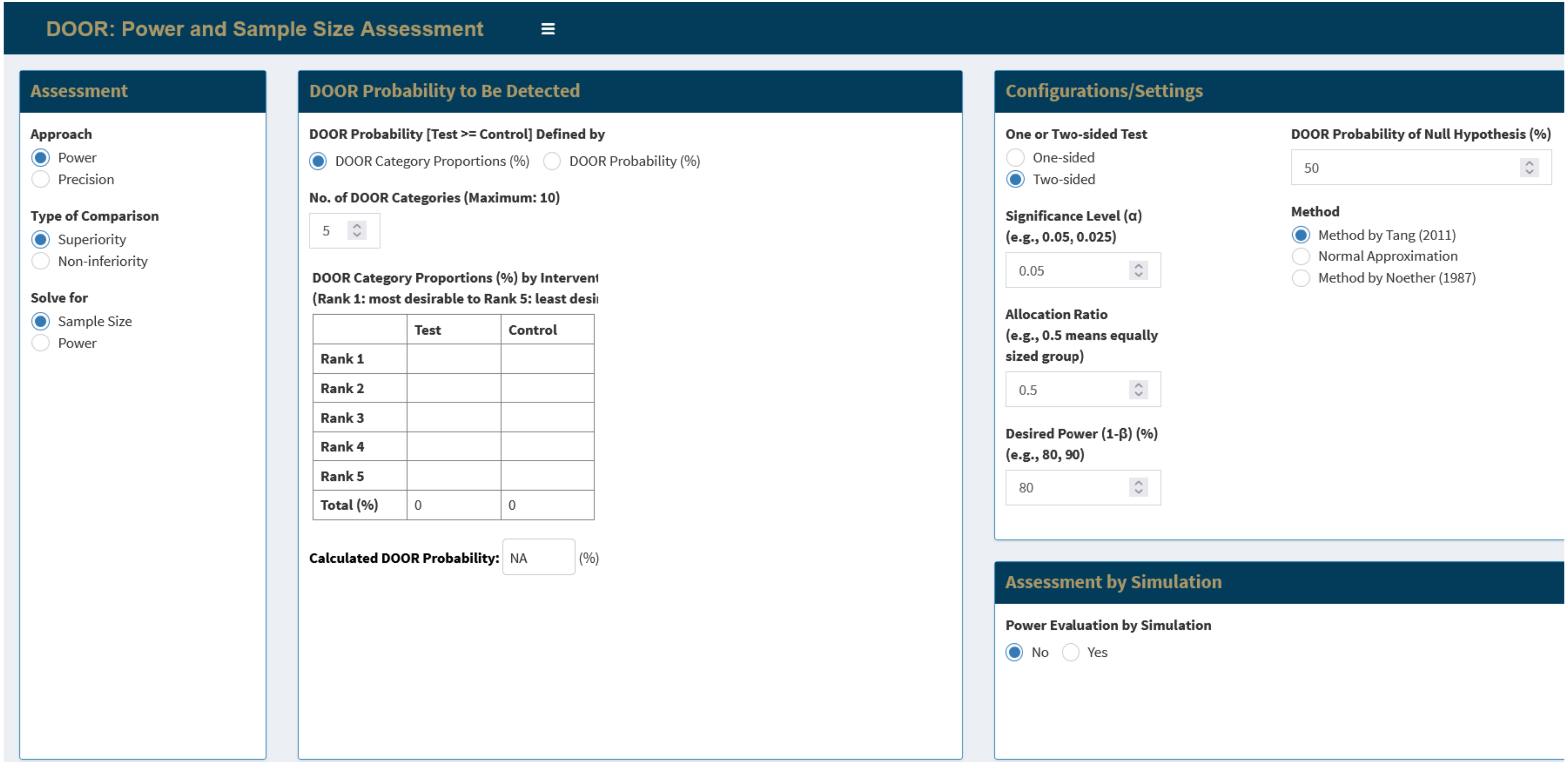
DOOR (Most desirable to least desirable) Rank	Treatment	Control
Rank 1		
Rank 2		
Rank 3		
Rank 4		
Rank 5		
Total (N)		

DOOR Components Distribution by Intervention

DOOR Component	Treatment	Control
Component 1		
Component 2		
Component 3		
Component 4		

Online tools for designing clinical trials with the DOOR methodology

Power and sample size determination demonstration



DOOR: Power and Sample Size Assessment ≡

Assessment

Approach

● Power
○ Precision

Type of Comparison

● Superiority
○ Non-inferiority

Solve for

● Sample Size
○ Power

DOOR Probability to Be Detected

DOOR Probability [Test >= Control] Defined by

● DOOR Category Proportions (%) ○ DOOR Probability (%)

No. of DOOR Categories (Maximum: 10)

5

DOOR Category Proportions (%) by Intervent
(Rank 1: most desirable to Rank 5: least desi

	Test	Control
Rank 1		
Rank 2		
Rank 3		
Rank 4		
Rank 5		
Total (%)	0	0

Calculated DOOR Probability: NA (%)

Configurations/Settings

One or Two-sided Test

○ One-sided
● Two-sided

Significance Level (α) (e.g., 0.05, 0.025)

0.05

Allocation Ratio (e.g., 0.5 means equally sized group)

0.5

Desired Power (1-β) (%) (e.g., 80, 90)

80

DOOR Probability of Null Hypothesis (%)

50

Method

● Method by Tang (2011)
○ Normal Approximation
○ Method by Noether (1987)

Assessment by Simulation

Power Evaluation by Simulation

● No ○ Yes

More to come!

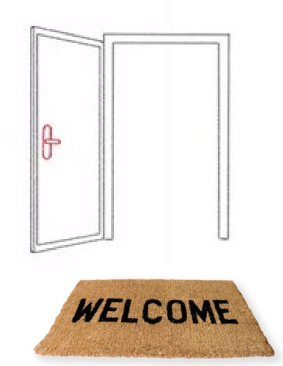
Online tools for the DOOR Methodology

- □ Monitoring of clinical trials, including group-sequential and adaptive designs
- □ Covariate-adjusted analysis: stratified analysis
- □ Subgroup analysis
- □ Integrated analyses: meta-analysis
- □ Longitudinal time-to-event type DOOR outcomes (Shu S et al 2024)

Yijie He
4th year student in PhD Program
(Applied Biostatistics Track)

Qihang Wu
2nd year student in PhD Program
(Applied Biostatistics Track)

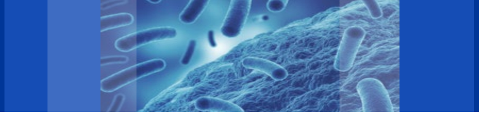
Shu S et al. Stat Biopharm Res 2024, 1–8 (First published online on 02 December 2024 as doi: 10.1080/19466315.2024.2413059)



The DOOR is Open!

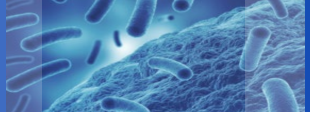
THANK YOU

https://methods.bsc.gwu.edu/



If I had one hour to solve a problem…
I would spend 55 minutes defining and understanding the problem…
and 5 minutes solving it.

Albert Einstein



Summary

- Place interest on pragmatic questions to match their clinical importance
  - Implies a patient-centric benefit:risk focus

- Elevate patient-centric benefit:risk from a post-hoc exercise into trial design and conduct

- DOOR
  - Patient-centric paradigm for the design, data monitoring, analysis, interpretation, and reporting of clinical trials and other studies based on benefit-risk evaluation
  - Uses outcomes to analyze patients for a closer reflection of the effects on patients
  - Robust analyses

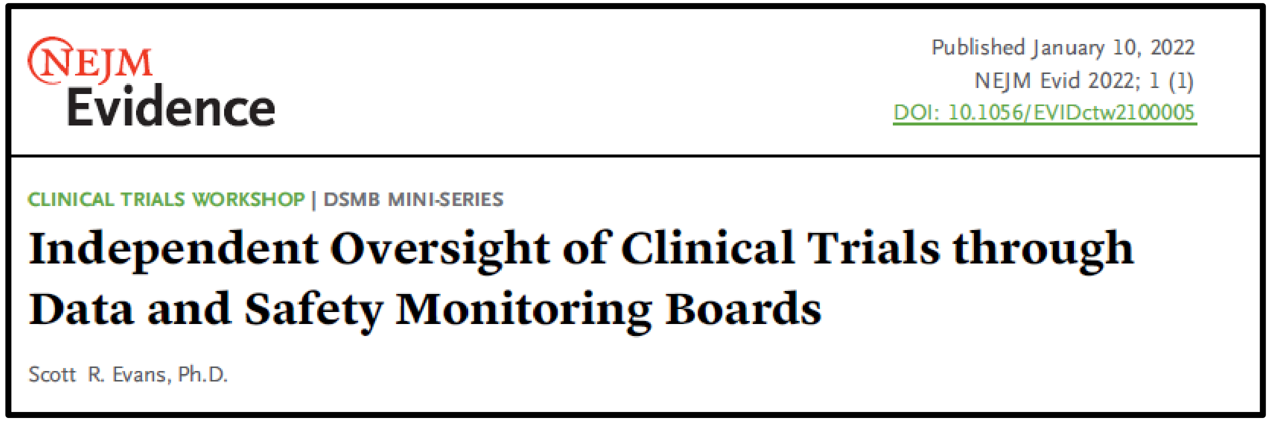
Summary

- Ongoing DOOR work
  - Regulatory science tools for design and conduct
  - DOOR outcome development with FDA and academic colleagues
  - Meta-analyses (for FDA)
  - Subgroup evaluation
  - Interim monitoring (group sequential and adaptive designs)
  - Covariate-adjusted analysis / stratified analysis
  - Longitudinal time-to-event type DOOR outcomes
  - Cluster randomization

Significant Contributors (p<0.001)

- Faculty: Guoqing Diao, Greg Sandoval
- NIH: Dean Follmann, Colin Wu
- FDA: Dan Rubin, Gene Pennello
- PhD students: Weixiao Dai, Yijie He, Richard Shu, Lizhao Ge, Shanshan Zhang, Lijuan Zeng, Wanying Shao, Yike Wang, Qihang Wu
- The Antibacterial Resistance Leadership Group
- FDA ORISE DOOR Fellowship team
- ACTT-1 research team
- George Saade, MD
- Arun Sanyal, MD
- Chip Chambers, MD

DMCs



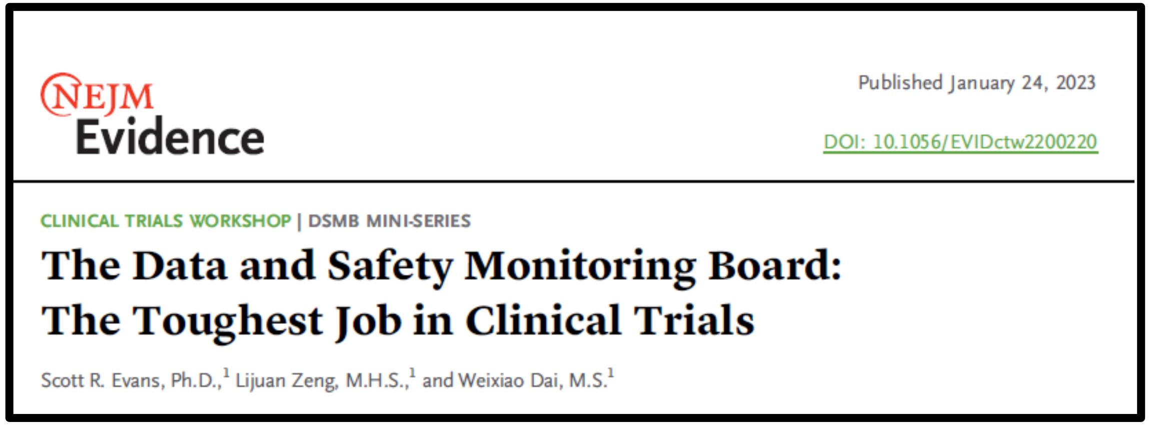
NEJM Evidence

Published January 10, 2022
NEJM Evid 2022; 1 (1)
DOI: 10.1056/EVIDctw2100005

CLINICAL TRIALS WORKSHOP | DSMB MINI-SERIES

Independent Oversight of Clinical Trials through Data and Safety Monitoring Boards

Scott R. Evans, Ph.D.



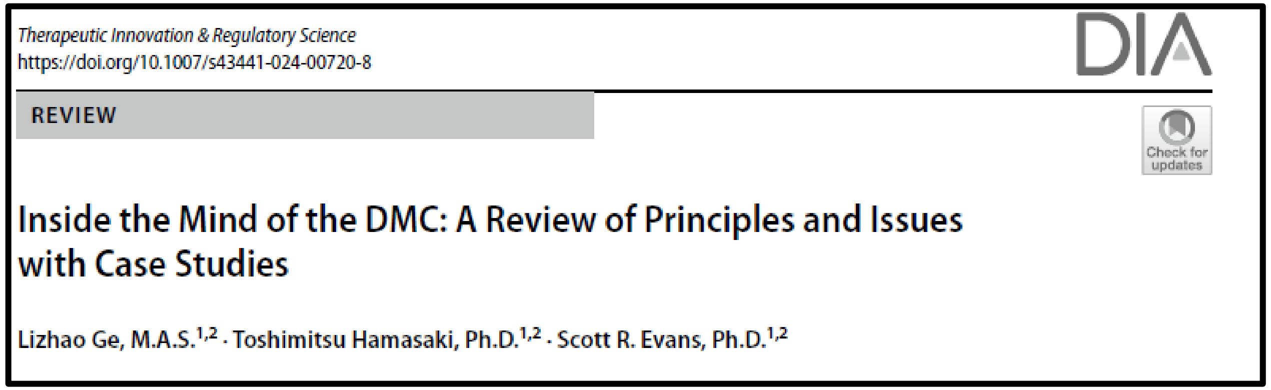
NEJM Evidence

Published January 24, 2023
DOI: 10.1056/EVIDctw2200220

CLINICAL TRIALS WORKSHOP | DSMB MINI-SERIES

The Data and Safety Monitoring Board: The Toughest Job in Clinical Trials

Scott R. Evans, Ph.D.,[1] Lijuan Zeng, M.H.S.,[1] and Weixiao Dai, M.S.[1]



Therapeutic Innovation & Regulatory Science
https://doi.org/10.1007/s43441-024-00720-8

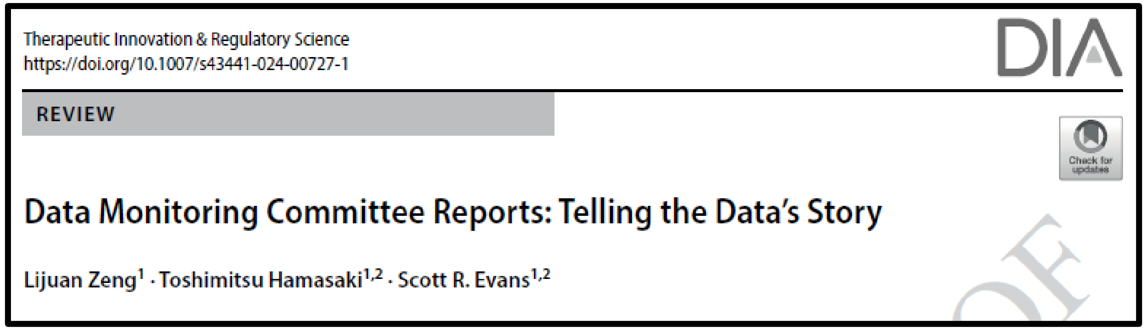
DIA

REVIEW

Check for updates

Inside the Mind of the DMC: A Review of Principles and Issues with Case Studies

Lizhao Ge, M.A.S.[1,2] · Toshimitsu Hamasaki, Ph.D.[1,2] · Scott R. Evans, Ph.D.[1,2]



Therapeutic Innovation & Regulatory Science
https://doi.org/10.1007/s43441-024-00727-1

DIA

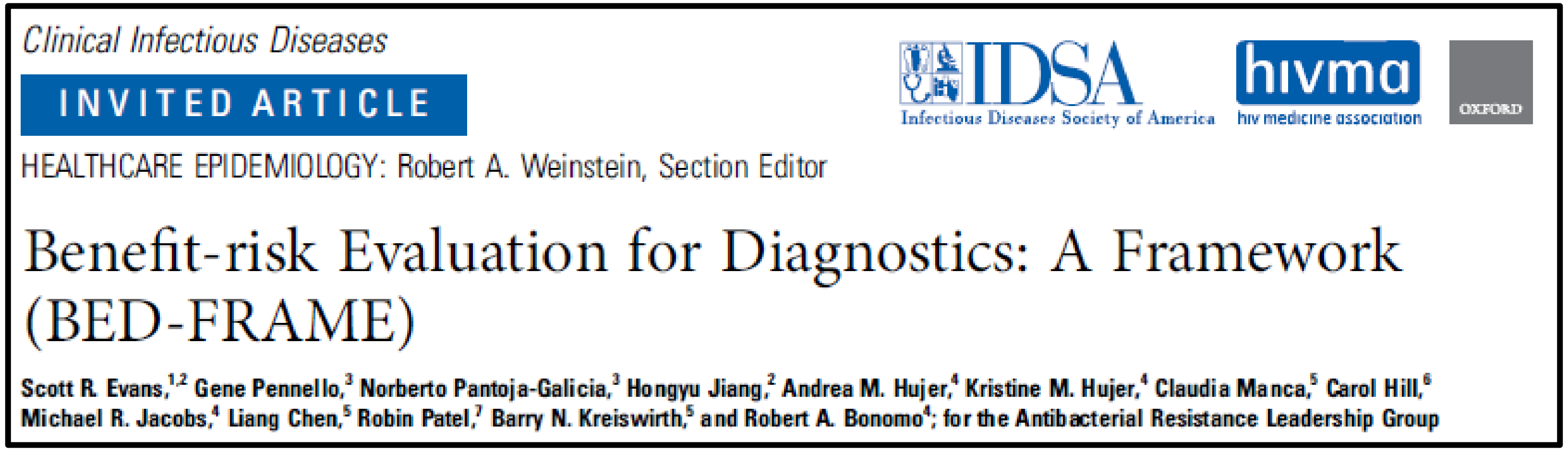
REVIEW

Check for updates

Data Monitoring Committee Reports: Telling the Data's Story

Lijuan Zeng[1] · Toshimitsu Hamasaki[1,2] · Scott R. Evans[1,2]

Pragmatic Diagnostic Evaluation



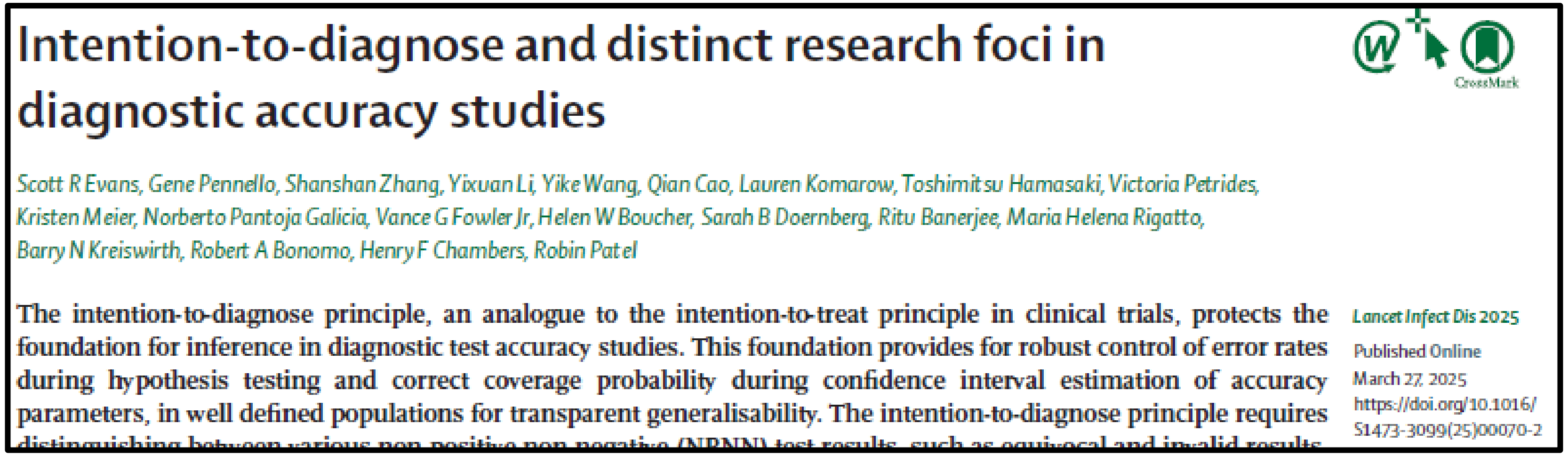
Clinical Infectious Diseases

INVITED ARTICLE

HEALTHCARE EPIDEMIOLOGY: Robert A. Weinstein, Section Editor

Benefit-risk Evaluation for Diagnostics: A Framework (BED-FRAME)

Scott R. Evans,[1,2] Gene Pennello,[3] Norberto Pantoja-Galicia,[3] Hongyu Jiang,[2] Andrea M. Hujer,[4] Kristine M. Hujer,[4] Claudia Manca,[5] Carol Hill,[6] Michael R. Jacobs,[4] Liang Chen,[5] Robin Patel,[7] Barry N. Kreiswirth,[5] and Robert A. Bonomo[4]; for the Antibacterial Resistance Leadership Group

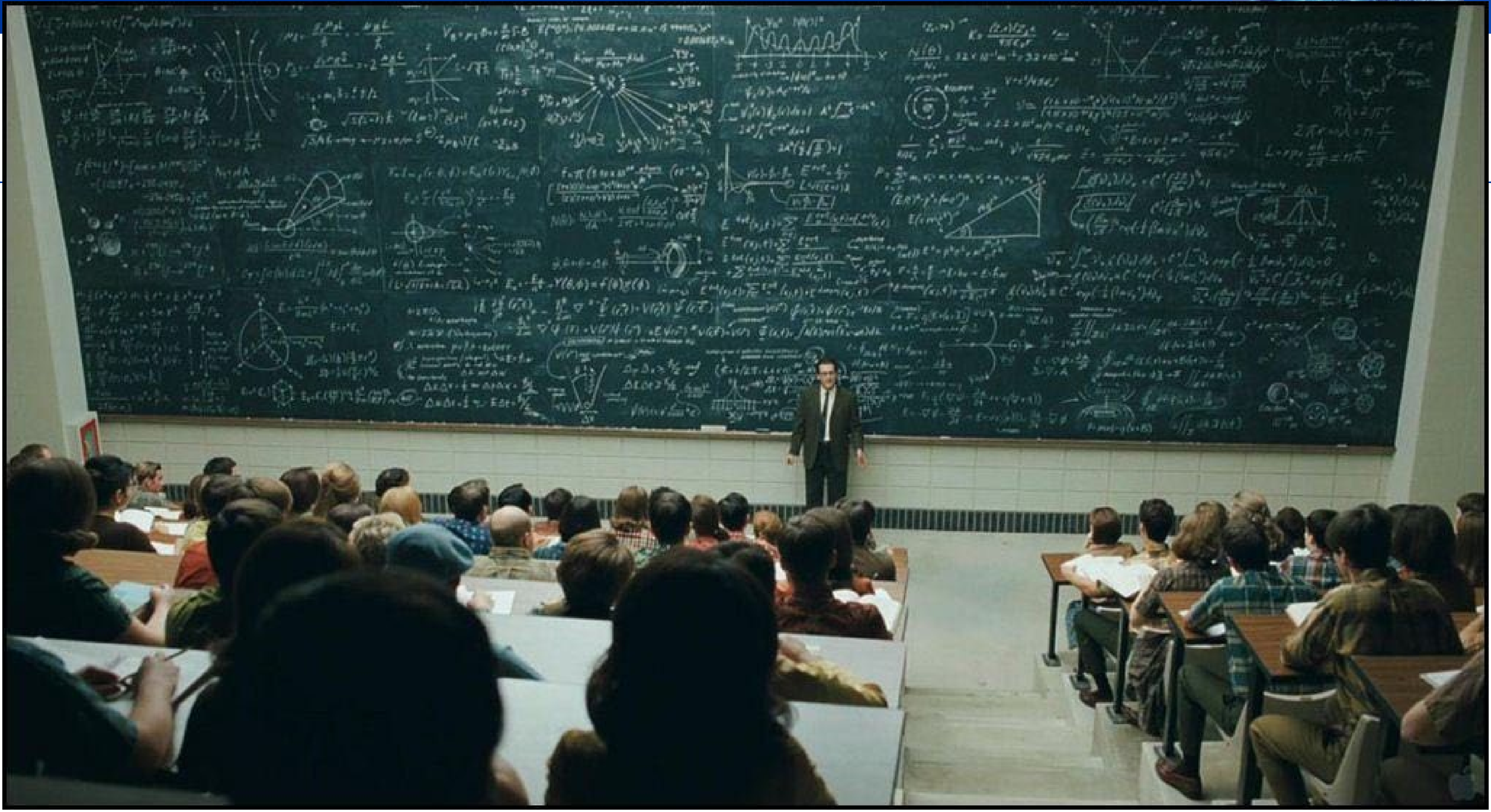


Intention-to-diagnose and distinct research foci in diagnostic accuracy studies

Scott R Evans, Gene Pennello, Shanshan Zhang, Yixuan Li, Yike Wang, Qian Cao, Lauren Komarow, Toshimitsu Hamasaki, Victoria Petrides, Kristen Meier, Norberto Pantoja Galicia, Vance G Fowler Jr, Helen W Boucher, Sarah B Doernberg, Ritu Banerjee, Maria Helena Rigatto, Barry N Kreiswirth, Robert A Bonomo, Henry F Chambers, Robin Patel

The intention-to-diagnose principle, an analogue to the intention-to-treat principle in clinical trials, protects the foundation for inference in diagnostic test accuracy studies. This foundation provides for robust control of error rates during hypothesis testing and correct coverage probability during confidence interval estimation of accuracy parameters, in well defined populations for transparent generalisability. The intention-to-diagnose principle requires distinguishing between various non-positive non-negative (NPNN) test results, such as equivocal and invalid results

Lancet Infect Dis 2025
Published Online
March 27, 2025
https://doi.org/10.1016/S1473-3099(25)00070-2



We have no doubt that you will enthusiastically applaud now …

because you are so relieved that it is over.

Thank you.