COMMENTARY

BUNJAMES NGETH

17th APRIL 2026

OUTLINE

1. INTRODUCTION
2. PAPER COMMENTARY
3. CONCLUSION
4. FUTURE IMPLICATIONS

INTRODUCTION

RCT vs. RWE

Randomized Control Trial

Randomized controlled trials (RCTs) are widely regarded as the gold standard for estimating treatment effects in clinical research, but their high cost and operational complexity often limitfeasibility for many clinically important questions.

Real World Evidence

RWE is given as the clinical evidence based on so-called real-world data (RWD), i.e. data outside of RCTs given as data relating to patient health status or data routinely collected from different sources.

Why Emulate RCTs?

Gold Standard: Randomized Controlled Trials (RCTs) eliminate confounding through controlled randomization.

Cost & Feasibility: Highly expensive, slow to execute, and often feature restrictive inclusion criteria that do not represent everyday patients.

Representativeness: Often exclude complex, heterogenous real-world population.

Rigid Constraints: Strict regulatory mandates, often discard cumulative evidence through double-dichotomization.

Target Trial Emulation (TTE)

- Designed to improve the quality and interpretability of observational research in several ways:

  - **Avoiding Common Biases:** Mimicked RCT design, researchers can prevent structural errors that plagued traditional observational analyses.
  - **Actionable Causal Inference:** forces researchers to ask specific questions about interventions rather than just looking at associations with biomarkers, leading to findings that directly inform clinical decision-making.
  - **Identifying Necessary Data:** Specify target trial clarifies which confounders must be adjusted for and what specific data points are required for a valid analysis.

Source: https://www.neurology.org/doi/10.1212/WNL.0000000000214269

Target Trial Emulation (TTE)

Eligibility Criteria

- Clear inclusion and exclusion criteria

Treatment Strategies

- Well-defined intervention and comparisons groups to avoid ill-defined association
- Specify point (intention to treat) and/or sustained (static or dynamic) interventions

Treatment Assignment

- Assess what confounders can be included
- Assess for unmeasured confounders and variable bias
- Consider treatment-confounder feedback

Follow-up

- Define outcomes of interest
- Ensure that eligibility assessment, treatment allocation, and the moment starting when outcomes begin to be counted align (or else use appropriate analytic methods to account for informative censorship)

Source: https://www.neurology.org/doi/10.1212/WNL.0000000000214269

Target Trial Emulation (TTE)

- **Limitation**
  - Validity depends heavily on data fitness. Whether the source material contains enough detail to support the required trial components.
    - Certain clinical judgments used in trial eligibility may not be recorded in structured data.
  - TTE cannot entirely eliminate unmeasured confounding, as it relies on the assumption that all factors influencing treatment assignment are captured and adjusted for in the data.

Two-Trial Rule (TTR)

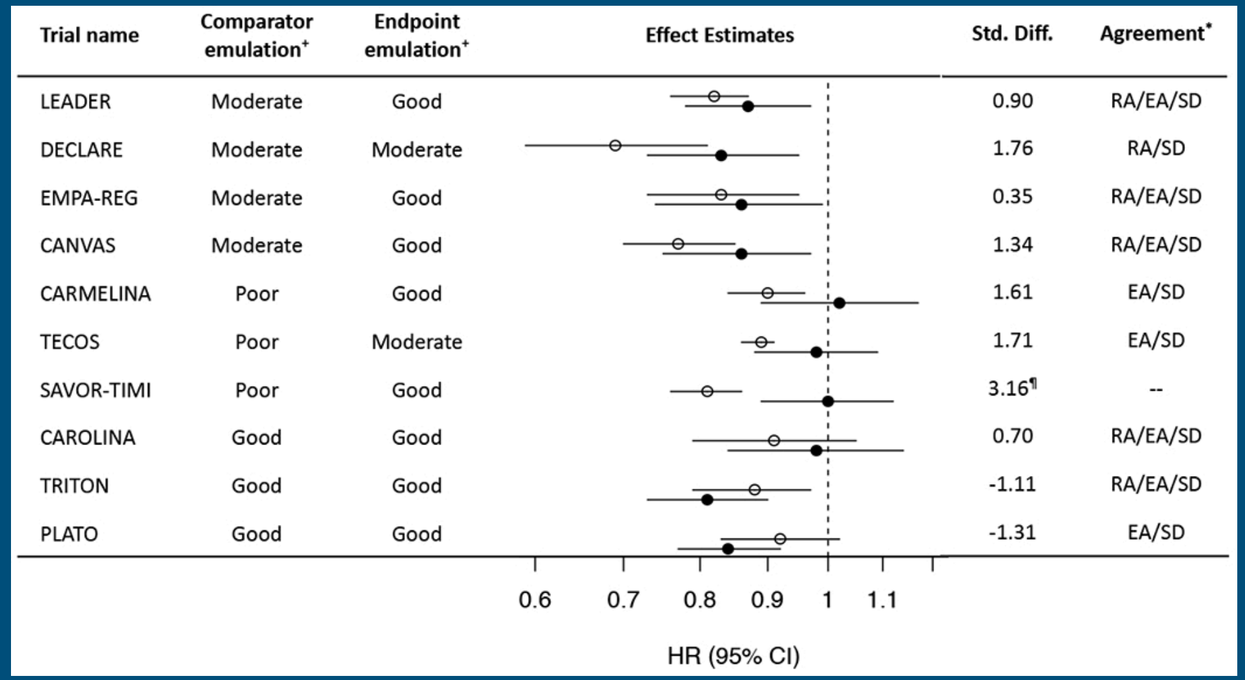
- Universally understood, simple to calculate, deeply entrenched in medical research field.
- A replication is deemed successful if both the original study and the replication study achieve a statistically significant p-value in the same direction.
- The two-trials rule for drug approval requires **“at least two adequate andwell-controlled studies, each convincing on its own, to establish effectiveness.”**
  - The rule is usually implemented by requiring two independent studies to be significant at the standard (one-sided) α = 0.025 level
  - *P(original) < 0.05 AND P(replication) < 0.05*

Source: https://pmc.ncbi.nlm.nih.gov/articles/PMC10087480/

PAPER COMMENTARY

RCT DUPLICATE Initiative

- **Franklin et al. (2020)** reported the first 10 emulations from the **RCT DUPLICATE initiative**.
- Key Finding: Agreement between RCT and RWE varied depending on the metric used:
  - **Regulatory Agreement (Two-Trials Rule):** Achieved in 6/10 emulations.
  - **Estimate Agreement (RWE estimate within RCT 95% CI):** Achieved in 8/10 emulations.
  - **Standardized Difference < 2:** Achieved in 9/10 emulations.

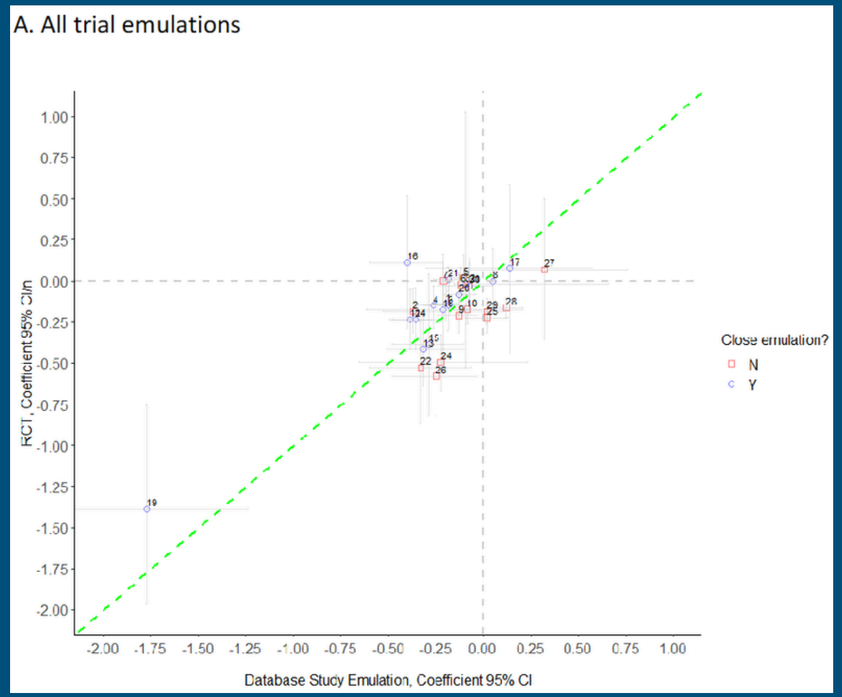


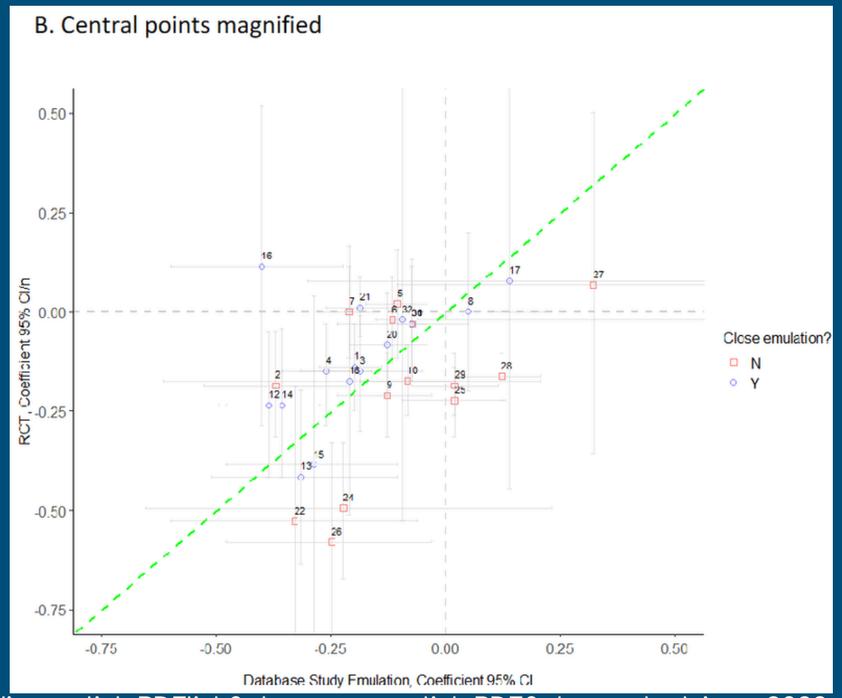
- **Outcomes:** Analysis of outcomes with known null effects helped diagnose residual confounding in specific emulations.

Source: https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.120.051718

RCT DUPLICATE Initiative

- **Wang et al. (2023)** systematically compare **32 RCTs** with nonrandomized database emulations using US insurance claims (Optum, MarketScan, Medicare).
- **Key Finding:** Overall Pearson correlation r = 0.82 (95% CI, 0.64–0.91) between RCT and RWE effect estimates.
- **Binary Agreement Metrics (Prespecified):**
  - **Statistical Significance Agreement:** 75% (56% full, 19% partial)
  - **Estimate Agreement:** 66% (RWE estimate within RCT 95% CI)
  - **Standardized Difference Agreement:** 75%
- **Critical Stratification (Post Hoc):**
  - Close Emulation (n=16): Pearson **r = 0.93**; 94% statistical significance agreement.
  - Not Close Emulation (n=16): Pearson **r = 0.53**; 56% statistical significance agreement.





Source: https://jamanetwork.com/journals/jama/fullarticle/2804067?utm_campaign=articlePDF&utm_medium=articlePDFlink&utm_source=articlePDF&utm_content=jama.2023.4221

Table 1. Effect Estimates and Agreement Metrics

Study No.	Trial name	Effect estimates (95% CI)			Standardized difference[c]	Agreement		
		RCT	Database studyAdjusted[a,b]	Database study crude[a,b]		Statistical significance	Estimate	Stand differ
1	LEADER	0.87 (0.78 to 0.97)	0.82 (0.76 to 0.87)	0.57 (0.54 to 0.61)	0.90	SA	EA	SD
2	DECLARE-TIMI58	0.83 (0.73 to 0.95)	0.69 (0.59 to 0.81)	0.47 (0.41 to 0.53)	1.76	SA		SD
3	EMPA-REG	0.86 (0.74 to 0.99)	0.83 (0.73 to 0.95)	0.63 (0.57 to 0.70)	0.35	SA	EA	SD
4	CANVAS	0.86 (0.75 to 0.97)	0.77 (0.70 to 0.85)	0.58 (0.54 to 0.62)	1.34	SA	EA	SD
5	CARMELINA	1.02 (0.89 to 1.17)	0.90 (0.84 to 0.96)	0.90 (0.86 to 0.95)	1.61	SAP	EA	SD
6	TECOS	0.98 (0.88 to 1.09)	0.89 (0.86 to 0.91)	0.81 (0.79 to 0.84)	1.71	SAP	EA	SD
7	SAVOR-TIMI	1.00 (0.89 to 1.12)	0.81 (0.76 to 0.86)	0.65 (0.62 to 0.69)	3.16	SAP		
8	LEAD-2	0 (−0.20 to 0.20)	0.05 (−0.11 to 0.22)	0.01 (−0.11 to 0.13)	−0.37	SA	EA	SD
9	TRITON-TIMI	0.81 (0.73 to 0.90)	0.88 (0.79 to 0.97)	0.70 (0.65 to 0.76)	−1.11	SA	EA	SD
10	PLATO	0.84 (0.77 to 0.92)	0.92 (0.83 to 1.02)	0.84 (0.78 to 0.91)	−1.31		EA	SD
11	ISAR-REACT 5	1.36 (1.09 to 1.70)	NA[d]	NA[d]	NA[d]	NA[d]		
12	ARISTOTLE	0.79 (0.66 to 0.95)	0.68 (0.61 to 0.76)	0.66 (0.62 to 0.71)	1.36	SA	EA	SD
13	RE-LY	0.66 (0.53 to 0.82)	0.73 (0.60 to 0.90)	0.67 (0.58 to 0.78)	−0.66	SA	EA	SD
14	ROCKET AF	0.79 (0.66 to 0.96)	0.70 (0.62 to 0.80)	0.76 (0.69 to 0.84)	1.00	SA	EA	SD
15	EINSTEIN DVT	0.68 (0.44 to 1.04)	0.75 (0.62 to 0.90)	0.85 (0.76 to 0.95)	−0.42	SAP	EA	SD
16	EINSTEIN PE	1.12 (0.75 to 1.68)	0.67 (0.55 to 0.80)	0.73 (0.64 to 0.83)	2.28	SAP		
17	RE-COVER II	1.08 (0.64 to 1.80)	1.15 (0.74 to 1.78)	1.48 (1.09 to 2.00)	−0.18	SA	EA	SD
18	AMPLIFY	0.84 (0.60 to 1.18)	0.81 (0.54 to 1.23)	0.64 (0.50 to 0.82)	0.13	SA	EA	SD
19	RECORD1	0.25 (0.14 to 0.47)	0.17 (0.10 to 0.29)	0.25 (0.18 to 0.34)	0.63	SA	EA	SD
20	TRANSCEND	0.92 (0.81 to 1.05)	0.88 (0.81 to 0.96)	0.80 (0.74 to 0.85)	0.55		EA	SD
21	ONTARGET	1.01 (0.94 to 1.09)	0.83 (0.77 to 0.90)	0.68 (0.64 to 0.72)	3.46	SAP		
22	HORIZON-PFT	0.59 (0.42 to 0.83)	0.72 (0.55 to 0.94)	1.08 (0.86 to 1.35)	−0.90	SA	EA	SD
23	VERO	0.44 (0.29 to 0.68)	NA[d]	NA[d]	NA[d]	NA[d]	NA[d]	NA[d]
24	DAPA-CKD	0.61 (0.51 to 0.72)	0.80 (0.52 to 1.26)	0.41 (0.29 to 0.58)	−1.10	SD		SD
25	PARADIGM-HF	0.80 (0.73 to 0.87)	1.02 (0.91 to 1.14)	0.95 (0.90 to 1.02)	−3.42			
26	P04334[e,f]	0.56 (0.44 to 0.72)	0.78 (0.62 to 0.97)	0.87 (0.76 to 0.99)	−1.95	SA	SD	SD
27	D5896	1.07 (0.70 to 1.65)	1.38 (0.90 to 2.13)	1.41 (1.00 to 1.98)	−0.81	SA	EA	SD
28	IMPACT[e,g]	0.85 (0.80 to 0.90)	1.13 (1.04 to 1.23)	1.22 (1.15 to 1.30)	−5.46			
29	POET-COPD	0.83 (0.77 to 0.90)	1.02 (0.93 to 1.12)	1.05 (0.99 to 1.12)	−3.27			
30	INSPIRE[h]	0.97 (0.84 to 1.12)	0.93 (0.90 to 0.96)	0.83 (0.81 to 0.85)	0.56	SA	EA	SD
31	CAROLINA[i]	0.98 (0.84 to 1.14)	0.91 (0.79 to 1.05)	0.92 (0.83-1.01)	0.70	SA	EA	SD
32	PRONOUNCE[i]	1.28 (0.59 to 2.79)	1.35 (0.94 to 1.93)	1.70 (1.30 to 2.21)	−0.12	SA	EA	SD

Source: https://jamanetwork.com/journals/jama/fullarticle/2804067?utm_campaign=articlePDF&utm_medium=articlePDFlink&utm_source=articlePDF&utm_content=jama.2023.4221

No.	Trial name	Comparator emulation[a]	Outcome emulation[b]	Age distribution, mean difference, y	Sex distribution, difference in % female	Run-in window[c]	Placebo control	In-hospital start of medication	Dose titration during follow-up	Discontinuation of maintenance therapy at randomization	Delayed effect[d]	Close emulation[e]
1	LEADER	Moderate	Good	-3.4	-17.8	Yes, placebo	Yes	No	Yes	No	No	Yes
2	DECLARE	Moderate	Moderate	1.4	-4.9	Yes, placebo	Yes	No	No	No	No	No
3	EMPA-REG	Moderate	Good	1.2	-11.9	Yes, placebo	Yes	No	No	No	No	Yes
4	CANVAS	Moderate	Good	-2.0	-10.5	Yes, placebo	Yes	No	No	No	No	Yes
5	CARMELINA	Poor	Good	-6.4	-16.2	No	Yes	No	No	No	No	No
6	TECOS	Poor	Moderate	-6.8	-18.1	No	Yes	No	No	No	No	No
7	SAVOR-TIMI	Poor	Good	-3.8	-13.7	No	Yes	No	No	No	No	No
8	LEAD-2	Good	Moderate	-2.0	-6.0	Yes, both groups	No	No	Yes	No	No	Yes
9	TRITON-TIMI 38	Good	Good	3.4[f]	4.9	No	No	Yes	Yes	No	No	No
10	PLATO	Good	Good	-3.3[f]	-4.1	No	No	Yes	Yes	No	No	No
11	ISAR-REACT 5	Good	Good	5.6	0.9	No	No	Yes	Yes	No	No	No
12	ARISTOTLE	Good	Good	-6.1	-16.7	No	No	No	No	Yes[g]	No	Yes
13	RE-LY	Good	Good	-4.7	-5.9	No	No	No	No	Yes[g]	No	Yes
14	ROCKET-AF	Good	Good	-4.5	-14.9	No	No	No	No	Yes[g]	No	Yes
15	EINSTEIN DVT	Good	Moderate	-14.7	-17.0	No	No	No	Yes	No	No	Yes
16	EINSTEIN PE	Good	Moderate	-8.2	-4.9	No	No	No	Yes	No	No	Yes
17	RE-COVER II	Good	Moderate	-13.5	-16.4	No	No	No	No	No	No	Yes
18	AMPLIFY	Good	Moderate	-0.6	-10.1	No	No	No	Yes	No	No	Yes
19	RECORD1	Good	Good	1.0	1.6	No	No	No	No	No	No	Yes
20	TRANSCEND	Moderate	Good	-4.0	-14.1	Yes, both groups	Yes	No	No	No	No	Yes
21	ON TARGET	Good	Good	-2.4	-27.2	Yes, both groups	No	No	Yes	No	No	Yes
22	HORIZON PFT	Moderate	Good	-1.0	0	No	Yes	No	No	No	Yes	No
23	VERO	Good	Moderate	1.1	0	No	No	No	No	No	Yes	No
24	DAPA-CKD	Moderate	Moderate	-5.5	-11.4	No	Yes	No	No	No	Yes	No
25	PARADIGM-HF	Moderate	Moderate	-4.7	-6.2	Yes, both groups	No	No	No	Yes	No	No
26	P04334	Good	Good	-11.2	1.9	Yes, 1 class	No	No	No	Yes	No	No
27	D5896	Good	Good	-3.3	-1.8	No	No	No	No	Yes	No	No
28	IMPACT	Good	Good	-4.0	-25.5	Yes, baseline prescription	No	No	No	Yes	No	No
29	POET-COPD	Good	Good	-7.5	-28.3	Yes, mixed	No	No	No	Yes	No	No
30	INSPIRE[h]	Good	Moderate	-1.5	-44.6	Yes, 1 class	No	No	No	Yes	No	No
31	CAROLINA[i]	Good	Good	-6.3	-12.3	Yes, placebo	No	No	Yes	No	No	Yes
32	PRONOUNCE1[i]	Good	Good	-3.0	0	No	No	No	Yes	No	No	Yes

Key Insight from 2023 Paper

- **Factors that compromised emulation:**
  - Using active comparators as placebo proxies introduced bias.
  - Claims data miss in-hospital medication starts (e.g., antiplatelet trials).
  - Trials that selected "responders" could not be emulated.
  - Mixing randomization effects with withdrawal effects.
  - Short RWE follow-up missed late-emerging benefits (e.g., HORIZON-PIVOTAL).
  - 83% of negative controls confirmed expected null effects, providing partial reassurance against residual confounding.

Why Binary Metrics Fall Short

- **Binary Metrics (Pass/Fail):**
  - Treat all "successes" as equal, regardless of sample size or effect magnitude.
  - A borderline significant result ($p=0.049$) in a massive RWE study is treated the same as a highly significant result ($p=0.001$) in a small study.
- **The Variance Ratio Problem:**
  - RWE studies often have much larger sample sizes than the original RCTs.
  - Binary metrics ignore this imbalance, potentially overstating replication success.
- **The Regulatory Need:**
  - The Two-Trials Rule (both studies $p<0.05$) is the legal standard, but it is sample-size agnostic.

Methodological Extension: Non-Inferority Trial

Köppe et al. adapt the model for non-inferiority trials by shifting the sceptical prior center from zero to the log hazard ratio margin.

A major mathematical advantage over binary TTR metrics, which struggle to incorporate prespecified margins across differing trial designs.

Assumes proportional hazards remains consistent across cohorts.

$$t_{\text{Box}} = \frac{\hat{\theta}_{\text{RWE}} - \delta}{\sqrt{\tau^2 + \sigma^2_{\text{RWE}}}}$$

Log hazard ratio margin (δ)

Sceptical P-Value

A p-value is a number describing how likely it is that your data would have occurred by random chance (i.e., that the null hypothesis is true).

- The level of statistical significance is often expressed as a p-value between ***0*** and ***1***.
- The smaller the p-value, the less likely the results occurred by random chance, and the stronger the evidence that you should reject the null hypothesis.

Sceptical P-Value

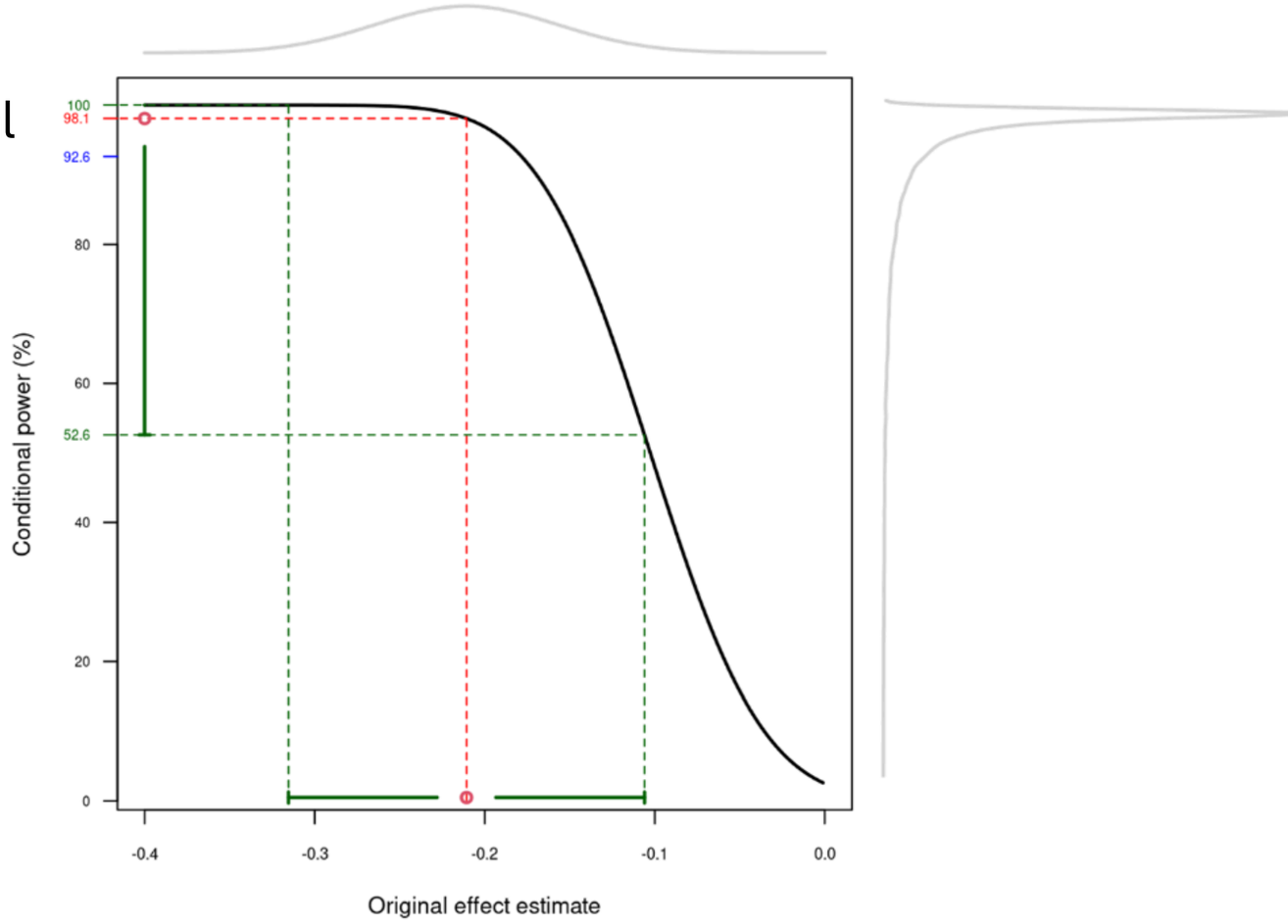
- **How it Penalizes Big Data:** If an RWE study has a massive sample size but only shows a tiny effect size, the sceptical p-value will increase, rejecting the replication.
- **The Requirement:** Massive RWE studies must now prove a mathematically convincing effect size, not just rely on statistical power.

Conditional vs. Predictive Power

The sceptical power consistently yielded higher predictive power (85.0%) compared to TTR **(83.5%)** on average for borderline original p-values.

Conditional Power Flaws
Calculating replication power based on fixed RCT point estimate ignores the inherent uncertainty of the original trial.

Predictive Power Integration
Averages conditional power across the entire probability distribution of the original effect estimate.



Insights from the RCT DUPLICATE Data

- **Data Source Impact:**
  - Replication success **84% with Medicare** vs **50% without**.
- **No Effect Shrinkage:**
  - RWE effect estimates in this dataset did not systematically shrink toward zero. **(28/29 in intended direction)**.
- **External Validity Signal:**
  - Successful sceptical p-value suggests RCT result generalizes to a broader population.
  - Failure may reflect limited generalizability of the RCT, not a failure of the RWE.

Strength & Limitation

- Outperforms the traditional **"Two-Trial rule"** and binary agreement metrics.
- Integrates p-values, effect sizes, and sample sizes from both RCT and RWE emulation.
- Relies on variance ratio, demanding larger sample sizes for replication success.
- Offers better conditional power compared to TTR approach.
- Ensures rigorous error control across studies for any variance ratio.
- Applicable to non-inferiority trials, considering specific margins.
- Provides robust one-sided confidence intervals, unlike standard meta-analysis.

- Lack of formalism for assessing replication success of null effects in superiority trials.
- **Sceptical p-value** may not be able to diagnose reasons for replication failure, such as design or model differences.
- Requires large sample sizes for replication if original trial was non-significant.
- Potential for increased Type-I error if original study's power is low.
- Design differences between RCTs and RWE emulations may introduce biases.

Conclusion

- **Köppe et al. (2025)** demonstrate that the sceptical p-value can solves the sample size paradox, providing a robust path forward for regulatory science.
- Traditional **Two-Trials Rule** is fundamentally flawed when applied to massive RWE datasets.
- Target Trial Emulation is the future of observational research, but requires further validation.

Future Implication

- Mathematical complexity of the sceptical p-value may hinder its adoption compared to the simple Two-Trials Rule.
- Regulators need advanced tools like this to safely accept RWE without endangering public health.

THANK YOU

REFERENCES

1. Assessing the replicability of RCTs in RWE emulations
2. Emulating Randomized Clinical Trials With Nonrandomized Real-World Evidence Studies: First Results From the RCT DUPLICATE Initiative
3. Emulation of Randomized Clinical Trials With Nonrandomized Database Analyses
4. Target trial emulation: applying principles of randomised trials to observational studies
5. Target Trial Emulation to Improve Causal Inference from Observational Data: What, Why, and How?