Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Assessing the replicability of RCTs in RWE emulations

HTUN TEZA

6738627 RADI-D

Doctor Of Philosophy

(Data Science For Health Care And Clinical Informatics)

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Randomized Controlled Trials

Gold Standard

- RCTs offer high internal validity through randomization

- Funding – LMIC populations are under-represented
- Time – Long-term effects are less studied
- Population – Vulnerable population (older, multi-morbid, and pregnant patients) are excluded

- Therefore: Low external validity
- Does the drug work *in practice*, not just in the trial?

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Randomized Controlled Trials

Types of trials

- Superiority Trials: null = “no difference”; both effects should have HR < 1; direction is clear
- Non-Inferiority Trials: null = “new drug is worse by more than margin δ”; what matters is position relative to δ, not HR = 1
- The direction of the effect could be:
    - HR < 1 (new drug actually superior)
    - HR between 1 and δ (new drug non-inferior but not superior)
    - HR > δ (new drug fails non-inferiority)

- where δ > 1 (e.g., δ = 1.3 means you will tolerate up to 30% more events with the new drug in exchange for its other benefits).

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

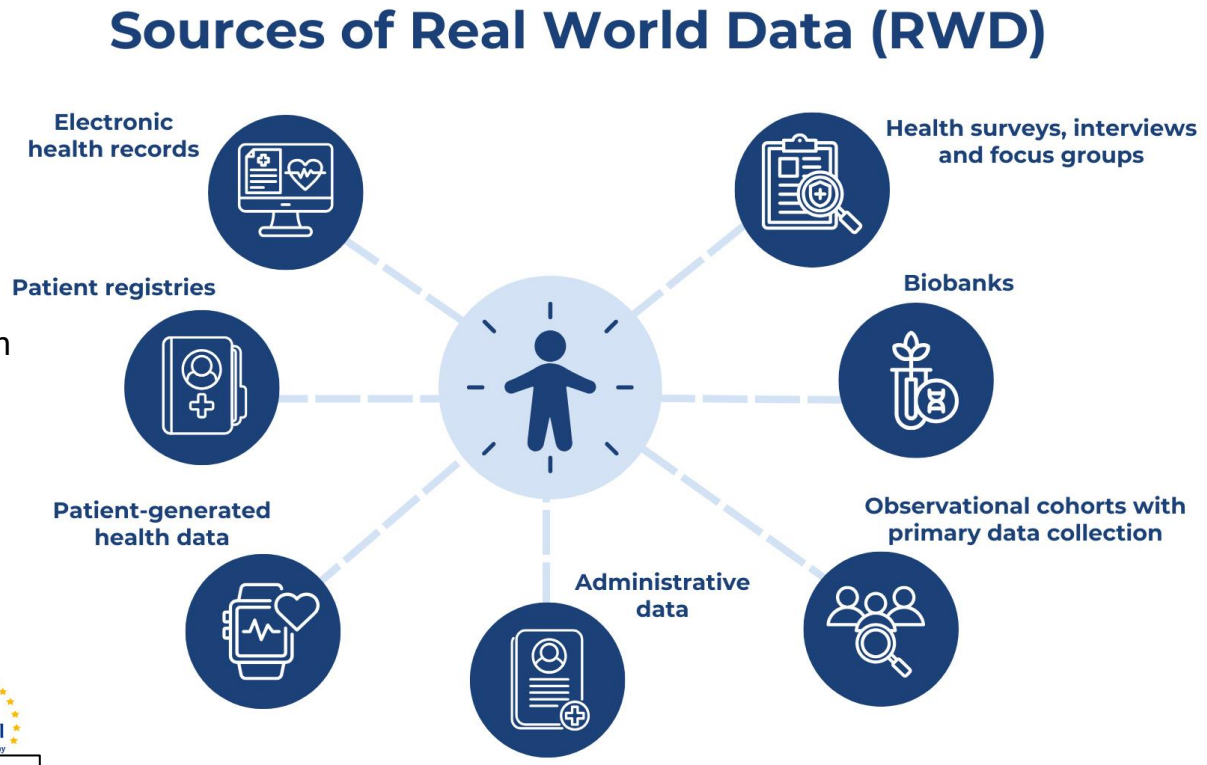
Real World Evidence

The Bridge

- Clinical evidence derived from routine healthcare data (claims, EHR), not from trials

- RWE is commonly used for post-market analysis of treatments in actual real-world care to ensure patient safety

- To address the efficacy–effectiveness gap
  - Efficacy: "Can it work?" (ideal trial conditions)
  - Effectiveness: "Does it work?" (routine clinical care)



Eichler HG, Abadie E, Breckenridge A, Flamion B, Gustafsson LL, Leufkens H, et al. Bridging the efficacy-effectiveness gap: a regulator's perspective on addressing variability of drug response. Nat Rev Drug Discov. 2011;10(7):495–506. https:// doi. org/ 10. 1038/ nrd35 01.

"Real-world data" is technically a misnomer — all study data come from the real world. RWD refers specifically to data collected outside a trial protocol.

RCT DUPLICATE

Randomized Controlled Trials Duplicated Using Prospective Longitudinal Insurance Claims: Applying Techniques of Epidemiology

- Division of Pharmacoepidemiology and Pharmacoeconomics at the Brigham and Women's Hospital
- If and under which circumstances RWD studies explicitly designed to emulate RCTs are useful to draw the same causal conclusions
- 32 RCT-RWE emulation pairs published
- Using United States Data
  - Optum Clinformatics Data Mart (01/2004–03/2019),
  - IBM MarketScan Commercial Claims and Encounters Database (2003–2017), and
  - Medicare Parts A, B, and D (2009–2017).

Franklin JM, Patorno E, Desai RJ, Glynn RJ, Martin D, Quinto K, et al. Emulating Randomized Clinical Trials With Nonrandomized Real-World Evidence Studies. Circulation. 2021;143(10):1002–13. https:// doi. org/ 10.1161/ CIRCU LATIO NAHA. 120. 051718.

Wang SV, Schneeweiss S, Initiative RD. Emulation of Randomized Clinical Trials With Nonrandomized Database Analyses: Results of 32 Clinical Trials. JAMA. 2023;329(16):1376–85. https:// doi. org/ 10. 1001/ jama. 2023.4221.

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Replication studies

Treatment Effectiveness

- For regulatory approval, FDA (used to) require two trials – "at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness."
- In February 2026, FDA ended this requirement: so can we use RWE as the second source of evidence?

	Direct Replication	Conceptual Replication
Example	Confirmatory RCT, multicenter trial	RWE emulation
Population	Same protocol, same eligibility	Routine care — broader, older, sicker
Confounding control	Randomization	Propensity score matching
Adherence	Protocol-defined	Real-world persistence
Outcomes	Protocol-defined	Coded proxies
Effect size differences	Due to chance	Due to structural design differences

Held L. Beyond the two-trials rule. Stat Med. 2024;43(26):5023–42. https://doi. org/ 10. 1002/ sim. 10055.

Meetal Jotangia. (2026, February 20). FDA ends "two-trial dogma" in historic shift. European Medical Journal; EMJ. https://www.emjreviews.com/emj-gold/news/da-ends-two-trial-dogma-in-historic-shift/

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Replication success

Treatment Effectiveness

- Replication is not simply "getting a significant result again" — two studies can both be significant yet tell very different stories about the size of the effect
- A treatment might show HR = 0.60 in the RCT and HR = 0.95 in the emulation — both $p < 0.025$, yet clinically worlds apart
- We need a criterion that captures not just whether both studies passed a threshold, but how consistently they estimated the same underlying effect
- Three candidate approaches exist: binary direction agreement, the Two Trials Rule, and the sceptical p-value — each making progressively stronger demands on the evidence

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Two Trials Rule

FDA standard procedure

- If both studies are significant ($p \leq \alpha / 0.025$) in the same direction

$$p_{TTR} = max(p_{RCT}, p_{RWE}) \leq \alpha$$

- Overall Type I error rate = $\alpha^2 = 0.000625$
- The TTR was designed for two confirmatory RCTs of similar size. Importing it unchanged into an RCT vs RWE comparison breaks both its logic and its fairness.

Why it is not appropriate for RWE replication

- Sample size blind: A massive RWE that barely crosses $p = 0.025$ is treated identically to a convincing one
- Asymmetry ignored: RCT and RWE are not interchangeable; the TTR treats them as if they were

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Meta-Analysis

Alternative?

- Meta-analysis pools both effect sizes into a single estimate and tests significance
- Intuitive, familiar, and statistically valid if studies are interchangeable
- The core problem: meta-analysis assumes the RCT and RWE are exchangeable units studying the same question in the same population
- An RCT and its emulation are by definition not the same study different population (trial-eligible vs routine care), different confounding control, different outcome capture

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Replication criterion

Considerations

- Asymmetry-aware: the RCT and RWE are not the same type of study and should not be treated as interchangeable units
- Effect size-sensitive: a replication that confirms direction but finds a much smaller effect should be penalised, not rewarded equally
- Sample size-aware: a massive database achieving borderline significance should be held to a higher standard than a smaller study achieving the same p-value
- Consistent across trial types: the framework should apply the same logic to both superiority and non-inferiority designs without requiring ad hoc adaptations
- Statistically controlled: the overall false positive rate should be bounded at α^2 regardless of how unequal the two study sizes are

C&B

Sceptical P-Value

Proposed

- Given what the RCT showed, would a rational sceptic "someone who genuinely disbelieves the treatment effect" be convinced by the RWE?
- Two components

1. Matthews' Analysis of Credibility: Build a prior distribution representing the most extreme disbelief the RCT result can still survive.
    - Narrow prior = the RCT was convincing and the sceptic has little room to doubt.
    - Wide prior = the RCT was weak.
2. Box's Prior-Data Conflict Test: Does the RWE data significantly conflict with the sceptic's prior? If yes — replication success.

- p_S is one-sided; replication can only succeed if the RWE effect is in the same direction as the RCT.

Matthews RAJ. Methods for assessing the credibility of clinical trial outcomes. Drug Inf J. 2001;35:1469–78.

Box GEP. Sampling and Bayes' Inference in Scientific Modelling and Robustness (with discussion). J R Stat Soc Ser A. 1980;143:383–430.

Example Trial

Sceptical P-value

- RCT : Trial to Assess Improvement in Therapeutic Outcomes by Optimizing Platelet Inhibition with Prasugrel–Thrombolysis in Myocardial Infarction: TRITON-TIMI 38 (2007)
- Goal : To compare prasugrel (New) vs clopidogrel (Ref)
- Selected outcome : 3-point MACE (Cardiovascular Death, Non-fatal MI, Non-fatal Stroke)

	Hazard Ratio	Log HR (θ)	SE (σ)	Z-score (θ/σ)	P-value CDF(Z)
RCT	0.81	-0.21	0.053	-3.96	<0.0001
RWE (pooled)	0.88	-0.128	0.052	-2.46	0.007

Wiviott, S. D., Braunwald, E., et al.. (2007). Prasugrel versus Clopidogrel in Patients with Acute Coronary Syndromes. New England Journal of Medicine, 357(20), 2001–2015. https://doi.org/10.1056/nejmoa0706482

Wang SV, Schneeweiss S, Initiative RD. Emulation of Randomized Clinical Trials With Nonrandomized Database Analyses: Results of 32 Clinical Trials. JAMA. 2023;329(16):1376–85. https:// doi. org/ 10. 1001/ jama. 2023.4221.

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Who is the Sceptic?

Sceptical P-value

- Imagine a regulator who saw the RCT result but refuses to believe it — they think the drug doesn't work
- They're not irrational: they represent the most extreme disbelief the RCT data can still survive
- Their disbelief is encoded as a probability distribution (a "prior") centred at zero (no effect)
- The question is: can the RWE force this person to change their mind?
- If yes → replication success

TRITON-TIMI 38	Hazard Ratio	Log HR (θ)	SE (σ)	Z-score (θ/σ)	P-value CDF(Z)
RCT	0.81	-0.21	0.053	-3.96	<0.0001
RWE (pooled)	0.88	-0.128	0.052	-2.46	0.007

Matthews RAJ. Methods for assessing the credibility of clinical trial outcomes. Drug Inf J. 2001;35:1469–78.

Box GEP. Sampling and Bayes' Inference in Scientific Modelling and Robustness (with discussion). J R Stat Soc Ser A. 1980;143:383–430.

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

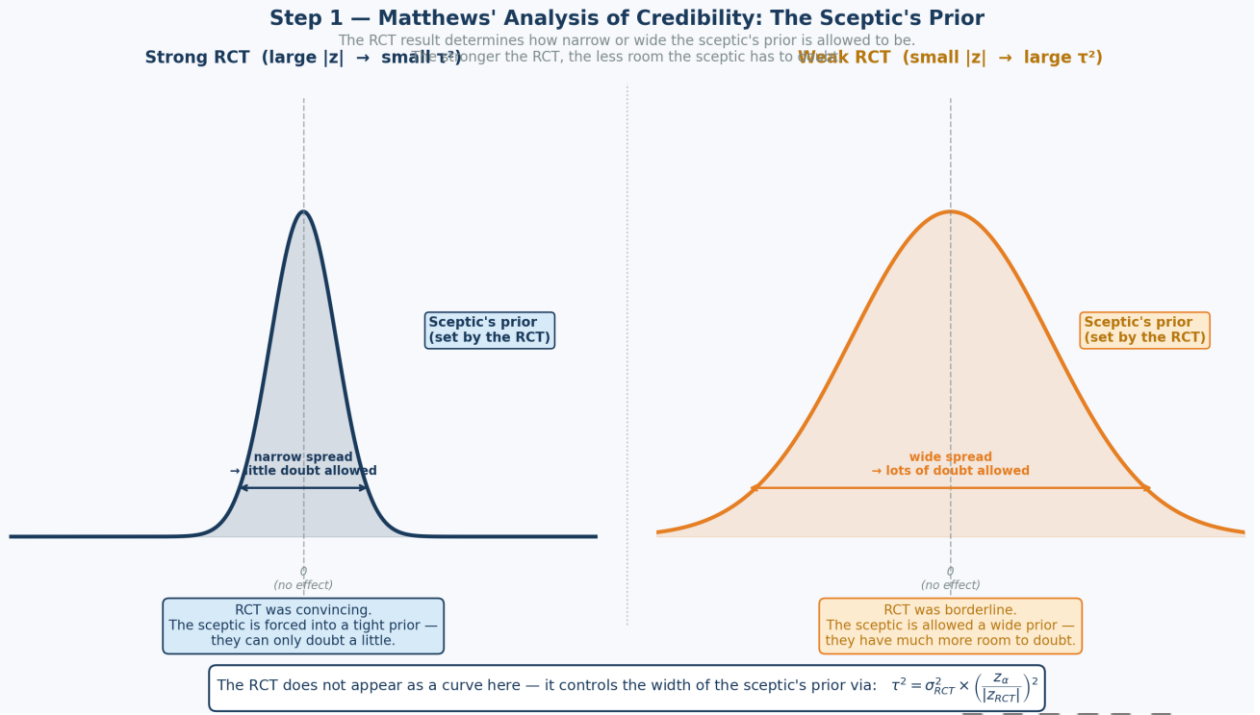
Matthews' Analysis of Credibility

Sceptical P-value

- Matthews (2001)
- Bayes : "given a prior, what does the data tell us?"
- Reverse-Bayes : "given the data, what is the most extreme prior disbelief that the RCT result can still survive?"

$$\tau^2 = \sigma^2_{RCT} \times \left(\frac{z_\alpha}{|z_{RCT}|}\right)^2$$

- Smaller σ (i.e. Larger Z score) mean larger/more precise study (RCT), therefore it is harder to dismiss, and the sceptic is forced to adopt a lower Sceptical Prior



The sceptic's expected distribution of RWE result

- Centered at 0 (Believes No effect)
- Spread by τ^2 (Prior/Sceptic's doubt) + σ^2_{RWE} (Variance/Uncertainty in RWE)

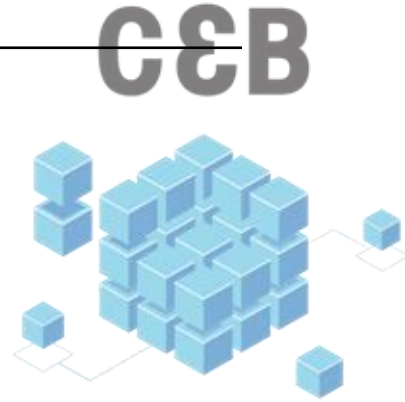
Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Setting up the sceptic

TRITON-TIMI 38 — prasugrel vs clopidogrel (3-point MACE)

- Initial significance (α) set at 0.025, z_α is 1.96
- Sceptical prior $\tau^2 = \sigma_{RCT}^2 \times (\frac{z_\alpha}{|z_{RCT}|})^2$ = 0.00281 × (1.96/3.96)² = 0.000688

	Hazard Ratio	Log HR (θ)	SE (σ)	Z-score (θ/σ)	P-value CDF(Z)
RCT	0.81	-0.21	0.053	-3.96	<0.0001
RWE (pooled)	0.88	-0.128	0.052	-2.46	0.007



Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Box's Prior-Data Conflict Test

Sceptical P-value

- Box (1980)
- How likely is this data under the prior-predictive distribution?
- i.e. How likely is that the sceptic is correct?

$$t_{Box} = \frac{\theta_{RWE}}{\sqrt{\tau^2 + \sigma^2_{RWE}}}$$

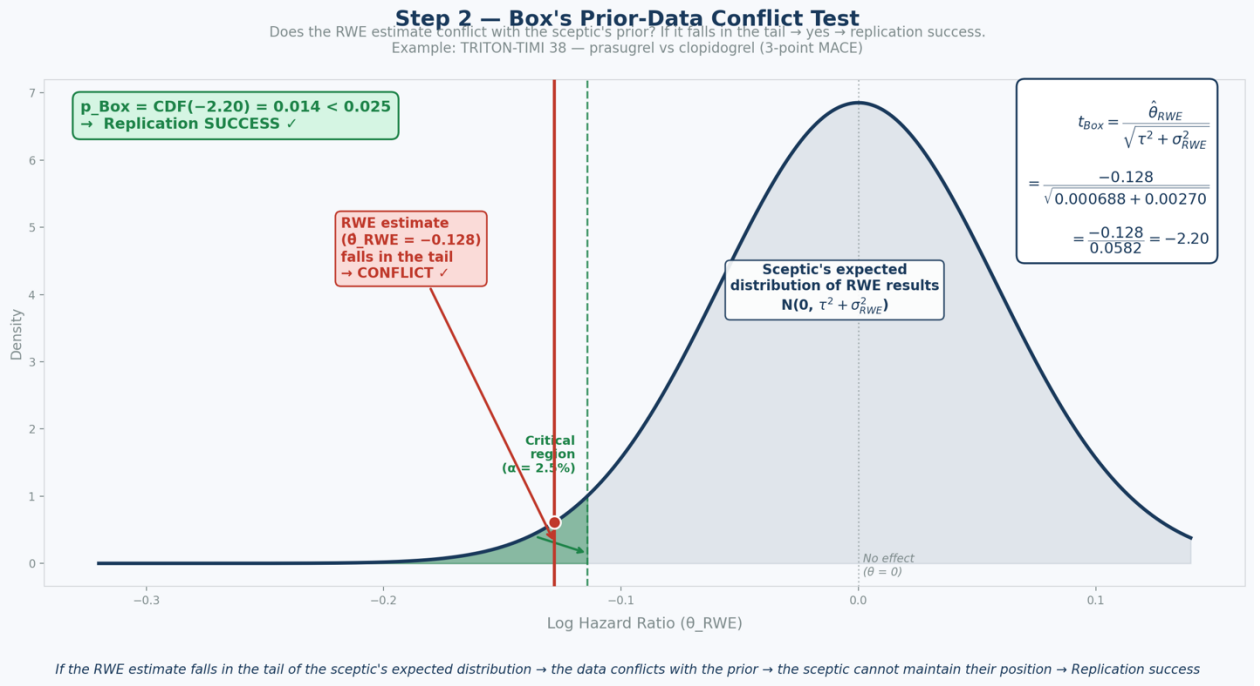
- Direction of t_Box : comes entirely from θ_RWE.
  - If the RWE effect is negative (treatment works), t_Box is negative, and we land in the left tail.
  - If the RWE found an effect in the wrong direction, t_Box would be positive, and we'd land in the right tail — far from replication success.
- Size of t_Box :
  - The Sceptical Prior pads the denominator as a penalty for the reliability of the original study reducing the size.
  - Larger Variance σ^2 is less reliable; to the same effect of reducing the size of t_Box.

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Does the RWE defeat the sceptic?

TRITON-TIMI 38 — prasugrel vs clopidogrel (3-point MACE)

- Box conflict test t_{Box} = −0.128 / √(0.000688 + 0.00270) = −0.128 / 0.0582 = −2.20
- p_{Box} is its one sided p value (left tailed because it is a reduce risk scenario)
- If the sceptic is right, how likely is it to get a result this extreme or more extreme? (further into whichever tail you care about)
- p_{Box} = CDF (t_{Box}) = CDF (−2.20) = 0.014 < 0.025
- This means the RWE data significantly conflicts with the sceptic's disbelief. The sceptic cannot maintain their position.



	Hazard Ratio	Log HR (θ)	SE (σ)	Z-score (θ/σ)	P-value CDF(Z)
RCT	0.81	-0.21	0.053	-3.96	<0.0001
RWE (pooled)	0.88	-0.128	0.052	-2.46	0.007

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Sceptical P-Value

TRITON-TIMI 38 — prasugrel vs clopidogrel (3-point MACE)

$$\tau^2 = \sigma_{RCT}^2 \times \left(\frac{z_\alpha}{|z_{RCT}|}\right)^2$$

$$t_{Box} = \frac{\theta_{RWE}}{\sqrt{\tau^2 + \sigma_{RWE}^2}}$$

- The sceptical prior τ^2 is essentially a penalty term that reflects how much doubt the RCT deserves.
  - A reliable RCT (small σ_{RCT}, large z) → small τ^2 → sceptic gets a narrow prior → little room to doubt → denominator of t_{Box} stays small → easier for RWE to generate conflict → lower p_{Box}
  - An unreliable RCT (large σ_{RCT}, small z) → large τ^2 → sceptic gets a wide prior → lots of room to doubt → denominator of t_{Box} inflates → harder for RWE to overcome → higher p_{Box}
- And then the RWE side mirrors it
  - Reliable RWE (small σ_{RWE}) → denominator shrinks further → t_{Box} gets larger → p_{Box} smaller → easier to achieve replication success
- p_{Box} tells you whether replication success is achieved at a specific α level (0.025 in our case).
- But its value depends on whichever α you chose; change α, and p_{Box} changes too.

C&B

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Sceptical P-Value

TRITON-TIMI 38 — prasugrel vs clopidogrel (3-point MACE)

$$\tau^2 = \sigma_{RCT}^2 \times \left(\frac{z_\alpha}{|z_{RCT}|}\right)^2$$

$$t_{Box} = \frac{\theta_{RWE}}{\sqrt{\tau^2 + \sigma_{RWE}^2}}$$

Full Chain

- Starting from α = 0.025 where p_{Box} = 0.0068, p_{Box} is currently smaller than α.
- α ↓ → z_α ↑ → τ^2 ↑ → t_{Box} ↓ → p_{Box} ↑
- When α is lowered which drives p_{Box} upward through the chain above, they meet at a value: and it is **Sceptical p-value**
- p_{Box} tells you pass or fail at a chosen α.
- $p_{Sceptical}$ is the α at which you're right on the boundary, the most demanding threshold the RWE can still survive. It's a single number you can report, like an ordinary p-value.

C&B

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

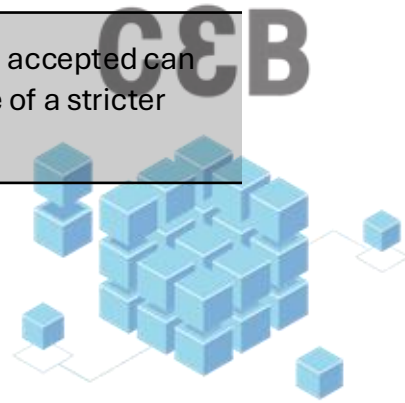
Sceptical P-Value

- a joint credibility test of both studies together, not just two separate significance tests stapled together.

Two Trial Rule	Sceptic P-value
Did the worse study pass?	Given what the RCT showed, did the RWE defeat the sceptic?

How Sceptic P-value related to Two Trial Rule

$p_{Sceptical} \leq \alpha$ always implies $p_{TTR} \leq \alpha$	False positive rate falls below α^2	Some valid replications are rejected
But not vice versa — every sceptical success is a TTR success, not the other way around	TTR sits exactly at α^2. $p_{Sceptical}$ is strictly more conservative, lower Type-I error	Replications TTR would have accepted can still fail $p_{Sceptical}$ — the price of a stricter standard



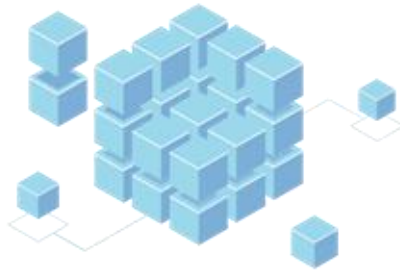
Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Non-Inferiority trials

D5896 example

- FDA-mandated postmarketing study comparing budesonide-formoterol (inhaled corticosteroid + LABA) versus budesonide alone (inhaled corticosteroid only) in asthma patients
- NI margin: Upper CI limit must be < 2.0 (i.e., the combination cannot cause more than double the risk of serious events)
- The TTR (checking same side of HR = 1) says: both null, both on same side — agreement
- The correct question: are both estimates within the NI margin? — disagreement

	Hazard Ratio	95% CI	Conclusion
RCT	1.07	0.70 – 1.65	Non-inferior ✓
RWE (pooled)	1.38	0.90 – 2.13	Non-inferior ✗



Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Sceptical P-Value

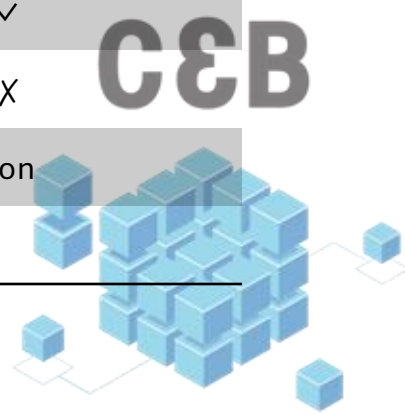
Non-Inferiority trials

- In NI trials, the sceptical prior is re-centred at the NI margin δ (not zero)

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

- The Box conflict test now asks: does the RWE effect significantly depart from δ in the favourable direction?
- This correctly answers the right question — position relative to δ, not relative to HR = 1

D5896	Hazard Ratio	95% CI	P value	Conclusion
RCT	1.07	0.70 – 1.65	0.002	Non-inferior ✓
RWE (pooled)	1.38	0.90 – 2.13	0.046	Non-inferior ✗
Two Trials Rule			0.046	Same Direction
Sceptical P Value			0.030	Re-centered



Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

RCT DUPLICATE

Randomized Controlled Trials Duplicated Using Prospective Longitudinal Insurance Claims: Applying Techniques of Epidemiology

- The sceptical p-value has been developed and validated theoretically, the next question is whether it changes conclusions in practice
- The RCT DUPLICATE initiative provides an ideal testbed: 32 pre-specified RCT emulations, conducted prospectively, using real US claims data
- Two applied questions drive the analysis: does $p_{Sceptical}$ agree with TTR on which emulations succeed, and where it disagrees, which verdict is more defensible?
- A secondary question: does $p_{Sceptical}$ offer meaningful gains in power over TTR, or does the stricter standard come at too high a cost?

C&B

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

RCT DUPLICATE

Randomized Controlled Trials Duplicated Using Prospective Longitudinal Insurance Claims: Applying Techniques of Epidemiology

- From 32 studies, 3 are excluded. (LEAD2 was excluded because continuous outcome. ISAR-REACT5 and VERO were excluded because of heterogeneity within RWE.)
- 29 studies evaluated here; pooled estimates across three RWD (11 superiority and 18 non-inferiority)
- 26/29 same direction and significant
- 28/29 same direction
- 1/29 opposite direction
- 3/29 non-significant original RCTs

C&B

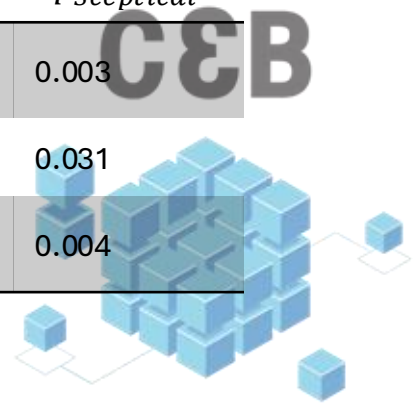
Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Results

Same verdict, stronger evidence

- 20/29 (69%) emulations declared replication success — identical conclusion with both methods
- $p_{Sceptical} \leq p_{TTR}$ in 27/29 studies — consistent advantage
- $p_{Sceptical} < p_{RWE}$ in 7/29 studies — mathematically impossible for p_{TTR} $[max(p_{RCT}, p_{RWE})]$
  - Very convincing RCT (Small σ_{RCT}) forces narrow Sceptic's prior (Small τ^2)
  - Even moderately sized RWE can conflict the sceptic
  - Joint credibility of two studies together is stronger than RWE alone.

Trial	Name	p_{RCT}	p_{RWE}	p_{TTR}	$p_{Sceptical}$
TRITON-TIMI	Trial to Assess Improvement in Therapeutic Outcomes by Optimizing Platelet Inhibition with Prasugrel–Thrombolysis in Myocardial Infarction	< 0.0001	0.007	0.007	0.003
PLATO	PLATelet inhibition and patient Outcomes	< 0.0001	0.056	0.056	0.031
P04334	Effects of Mometasone Furoate/Formoterol Combination Versus Mometasone Furoate Alone in Persistent Asthmatics	< 0.0001	0.015	0.015	0.004



Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Results

Same verdict, stronger evidence

- CARMELINA, TRANSCEND, INSPIRE: $p_{Sceptical} < p_{RCT}$
  - Weak RCT (Large σ_{RCT}) gives wide Sceptic's prior (Small τ^2)
  - But Reliable RWE (small σ_{RWE}) still conflict with the sceptic
  - Joint credibility of two studies together is stronger than RCT alone.

Trial	Name	p_{RCT}	p_{RWE}	p_{TTR}	$p_{Sceptical}$
CARMELINA	Cardiovascular and Renal Microvascular Outcome Study with Linagliptin in Patients with Type 2 Diabetes Mellitus	0.0003	< 0.0001	0.0003	< 0.0001
TRANSCEND	Telmisartan Randomised Assessment Study in ACE Intolerant Subjects With Cardiovascular Disease	0.10	0.002	0.10	0.0064
INSPIRE	Investigation of the Safety and Pharmacology of Dry Powder Inhalation of Treprostinil	0.34	< 0.0001	0.34	0.25

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Results

Power

- Power = probability of correctly declaring replication success when the drug truly works
- Low power = we miss real replications even when the drug genuinely works
- Conditional power = assumes RCT estimate is exactly the true effect (optimistic)
- Predictive power = averages across uncertainty in the RCT estimate (conservative, always lower)

	Conditional power		Predictive power	
	Two trials Rule	**Sceptical P value**	**Two trials Rule**	**Sceptical P value**
Overall	85.9	87.0	83.5	85.0
TRITON-TIMI 38	98.1	99.2	92.6	95.3
D5896	81.2	85.7	73.5	77.5

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Results

Power

- Across all 29 trials, $p_{Sceptical}$ consistently outperforms TTR on both measures
- TTR asks one fixed question: did RWE cross p < 0.025?, the bar never changes regardless of how strong the RCT was
- $p_{Sceptical}$ adapts: a stronger RCT forces a narrower Sceptical prior, which is easier for the RWE to conflict with
- So even though $p_{Sceptical}$ is stricter in principle, it gives the RWE a lower effective bar when the RCT earned it

	Conditional power		Predictive power	
	Two trials Rule	**Sceptical P value**	**Two trials Rule**	**Sceptical P value**
Overall	85.9	87.0	83.5	85.0
TRITON-TIMI 38	98.1	99.2	92.6	95.3
D5896	81.2	85.7	73.5	77.5

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Summary

	Binary Direction	Two Trials Rule	Sceptical P Value
Uses correct reference for NI (δ, not HR=1)	✗	Inconsistent	✓
Sensitive to effect size distance	✗	✗	✓ (Variance f both studies incorporated)
Sensitive to sample size differences	✗	✗	✓
Consistent across superiority and NI	✗	Inconsistent	✓
Same Type I error rate as TTR		Reference	✓ (lower; equal in controlled)
Higher power than TTR		Reference	✓

C&B