A Practical Guide to Interpret a Randomized Controlled Trial: Underpowered ≠ Inconclusive ≠ Negative ≠ Neutral

Ibrahim Halil Tanboga, MD, PhD
Department of Biostatistics and Cardiology, Nisantasi University Medical School, Istan
haliltanboga@gmail.com

Abstract

The most dangerous error in clinical trial interpretation is equating $p > 0.05$ with "no effect." This review provides a practical, algorithm-based framework for classifying randomized controlled trial (RCT) results into six distinct categories—positive, imprecise (+), neutral, inconclusive, negative, and harmful—using confidence interval (CI) position relative to the minimal clinically important difference (MCID) as the primary tool, augmented by Bayesian posterior probabilities. We demonstrate that the same $p > 0.05$ result can represent three fundamentally different conclusions (inconclusive, negative, or neutral), show how Bayesian reanalysis can rescue benefit signals missed by frequentist thresholds, and illustrate the framework with real-world examples from critical care and cardiology trials. The framework synthesizes guidance from Altman, Harrell, Pocock, Zampieri, the ASA, and ICH E9 into a single coherent decision algorithm.

Contents

1 Introduction: Why p-values Alone Are Insufficient

The single most dangerous error in clinical trial interpretation is equating $p > 0.05$ with "no effect." Altman and Bland formalized this in *BMJ* (1995) [Altman and Bland, 1995]: "Absence of evidence is not evidence of absence."

The American Statistical Association (ASA, 2016) stated unequivocally that scientific conclusions should not be based solely on whether a p-value crosses a specific threshold [Wasserstein and Lazar, 2016]. A small p-value does not necessarily indicate a large or important effect, and a large p-value does not mean the effect is absent.

Harrell argues the terms "positive" and "negative" should be abandoned entirely, replaced by continuous probability statements [Harrell, 2017]. His blog epigraph captures this: "To avoid 'false positives' do away with 'positive.'" On the Datamethods forum (2018), he proposed increasingly honest formulations for non-significant results:

> "The most brutally honest summary of an underpowered study: 'The money was spent.' Less brutal: 'The presumption of no difference is not yet overcome by data.'"

Critically, he acknowledged that even this language was "still too 'dichotomous' in that it was written assuming we would have different language for 'positive' vs. 'negative' studies. This implies a threshold for 'positive' which is what we're trying to get away from."

On underpowered trials specifically, Harrell is withering: "Underpowered trials are worse than no trials. Because used as evidence they can mislead especially those who aren't cognizant of power & errors."

The only reliable way to classify an RCT result is to examine the position of the 95% confidence interval (CI) relative to the pre-specified MCID (δ) and the null value. The p-value alone cannot distinguish positive, negative, neutral, or inconclusive outcomes. When available, Bayesian posterior probabilities resolve the remaining ambiguity that even CI + MCID cannot fully address.

2 The Complete Decision Algorithm

This algorithm has two parallel tracks: a **frequentist track** (CI + MCID) and a **Bayesian track** (posterior probabilities). The frequentist track is sufficient for most cases; the Bayesian track adds decisive value when the p-value falls near 0.05 or when the neutral vs. negative distinction matters clinically.

2.1 Track A — Frequentist (CI + MCID)

Step 1 — Define effect measure, null value, and MCID (δ). The effect measure may be HR, RR, OR, mean difference, or ARD. Null value: HR/RR/OR = 1.0, mean difference/ARD = 0.

What is the MCID? The Minimal Clinically Important Difference (MCID) is the smallest treatment effect that would be considered clinically meaningful to patients and clinicians—the threshold below which a statistically significant result would not change clinical practice. The MCID is *not* a post-hoc concept: it must be pre-specified in the

trial's statistical analysis plan, typically in the "Statistical Methods" section of the protocol, *before* data collection begins. It is determined by a combination of clinical judgment, prior evidence, and patient-centered outcome data [McGlothlin, 2013].

For example, in a cardiovascular mortality trial, a hazard ratio (HR) of 0.95 (5% relative reduction) would generally not change practice, whereas HR ≤ 0.80 ($\geq$20% relative reduction) would be considered a clinically meaningful benefit. Thus, the MCID-benefit threshold would be set at $\delta = 0.80$. Similarly, a harm threshold might be set at HR ≥ 1.25 (25% relative increase in the endpoint). In heart failure trials, where baseline event rates are high, even a 15% relative reduction (HR ≤ 0.85) may qualify as the MCID. In oncology, where therapies carry significant toxicity, MCID thresholds tend to be more demanding. The key point is that the MCID anchors the entire classification framework: without it, a CI can tell you about precision but cannot tell you about clinical relevance.

Step 2 — Does the 95% CI exclude the null? If **YES** ($p < 0.05$):

2a. Entire CI beyond MCID-benefit → POSITIVE

2b. CI crosses MCID (includes effects both above and below δ) → IMPRECISE (+) — significant but magnitude uncertain

2c. Entire CI in harm zone beyond MCID-harm → HARMFUL

If **NO** ($p \geq 0.05$) → go to Step 3.

Step 3 — How wide is the CI relative to MCID zones?

- Narrow CI within $[-\delta, +\delta]$ indifference zone — both MCID-benefit and MCID-harm excluded → NEUTRAL (precise null)
- Narrow CI excludes MCID-benefit but *not* MCID-harm — clinically meaningful benefit ruled out → NEGATIVE (benefit excluded)
- Wide CI crosses MCID-benefit and/or MCID-harm — includes both clinically meaningful benefit and harm → INCONCLUSIVE

Step 4 — NEVER compute post-hoc power. CI already contains all precision information. Post-hoc power = $f(p\text{-value})$ = zero additional information.

2.2 Track B — Bayesian (Zampieri/Harrell Framework)

Step 5 — Define priors (minimum 3 per Zampieri et al. AJRCCM 2021). Select belief strength (weak/moderate/strong) based on prior evidence per Zampieri Table 1 [Zampieri et al., 2021]. See Table 1.

Step 6 — Compute 3 posterior metrics for each prior. See Table 3 for metric definitions.

Step 7 — Classify by dominant metric. See Table 5 for classification rules.

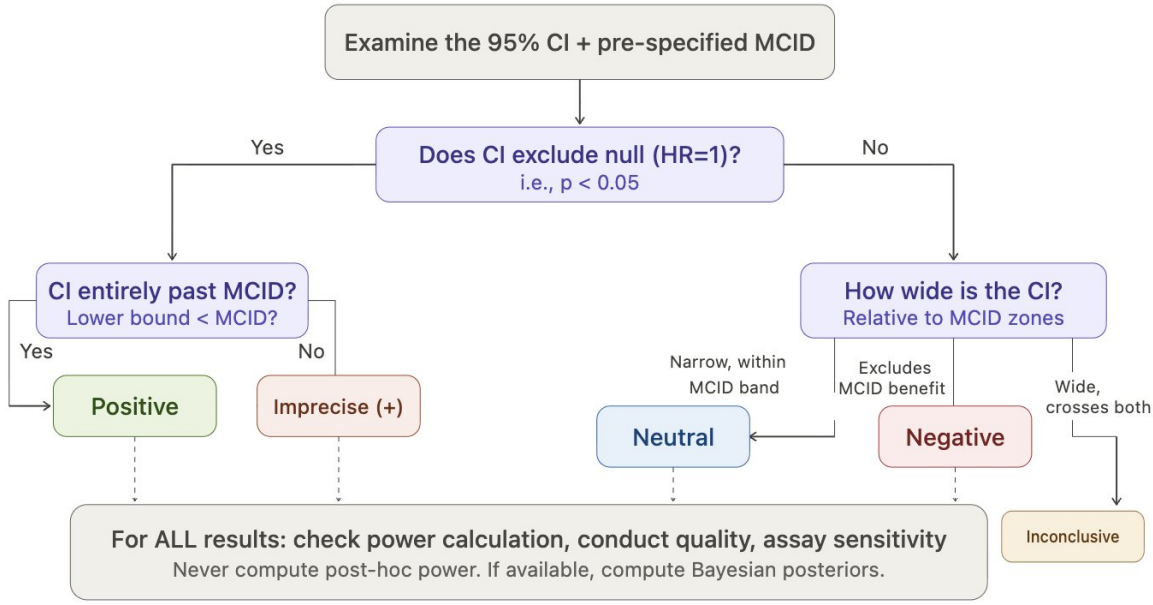


Figure 1: **CI + MCID Decision Flowchart.** The frequentist classification algorithm. First determine whether the 95% CI excludes the null; then classify by CI width relative to MCID zones. Post-hoc power should never be computed.

Table 1: **Prior specifications for Bayesian reanalysis** (adapted from Zampieri et al. 2021).

Prior type	**Center**	**Distribution**	**Constraint**
Skeptical (null-centered)	OR = 1	$\mathcal{N}(0,\ 0.355)$	Moderate strength
Optimistic (benefit-centered)	Expected benefit	$\mathcal{N}(-\delta,\ \sigma)$	$\Pr(\text{harm}) \geq 15\%$
Pessimistic (harm-centered)	Expected harm	$\mathcal{N}(+\delta,\ \sigma)$	$\Pr(\text{benefit}) \geq 15\%$
Data-derived (meta-analysis)	Pooled estimate	$\mathcal{N}(\hat{\mu},\ \hat{\sigma})$	Optional

Step 8 — Sensitivity check. Compute I^2 across prior results. If $I^2 < 0.20$, conclusion is robust to prior choice — the data dominate.

Key principle: if the verdict is the same across skeptical, optimistic, and pessimistic priors → the data are talking, not the prior.

3 The CI + MCID Classification

Figure 2 presents a forest plot illustrating the six possible trial classifications based on CI position relative to the null and MCID thresholds.

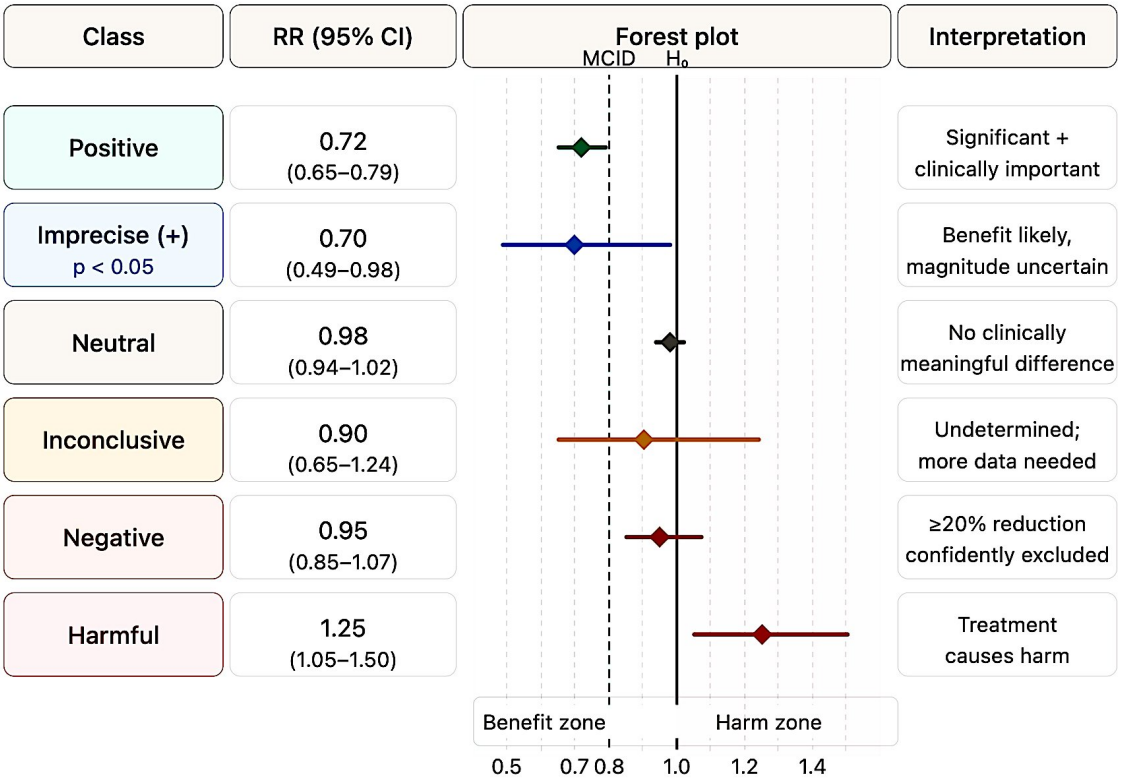


Figure 2: **Six-class forest plot.** Each row illustrates a prototypical CI position. The dashed lines mark the MCID for benefit (left) and harm (right); the solid line marks the null (HR = 1.0). Classification depends on CI position relative to these thresholds, not on the p-value alone.

3.1 How Leading Authorities' Terminologies Compare and Conflict

The terms "positive," "negative," "neutral," and "inconclusive" are used inconsistently across the literature. Table 2 maps each authority's preferred terms onto the six categories in this framework.

Table 2: **Terminology comparison across leading authorities.** Each cell shows how a given authority defines or uses the corresponding term. C = Classical/frequentist; B = Bayesian; F = Framework/regulatory.

Authority	**Positive**	**Negative**	**Inconclusive**	**Under-powered**	**Neutral**
Altman & Bland (C)	Not addressed	Narrow CI excluding meaningful effects; label should be abandoned	Default for most non-significant results	Design deficiency leading to inconclusiveness	Not distinguished from negative
Freiman et al. (C)	Not addressed	Requires ≥90% power + CI excluding meaningful effects	Implicit — trials haven't received 'a fair test'	94% of 'negative' trials were underpowered	Not used
Harrell (B)	Rejects term — use Pr(benefit)	'Absence of evidence ≠ evidence of absence'	Prefers 'uninformative' or 'the money was spent'	'Worse than no trials'	Compute Pr(similarity) via Bayesian model
Pocock & Stone (F)	Addressed in companion paper	12-question evaluation required	Explicitly used (e.g., PRoactive)	Question #2 in 12-question framework	Not a primary term
Hawkins & Samuels (C)	Not addressed	Interval A: CI within MCID boundaries	Interval B or C (spanning both)	Post-hoc power 'irrelevant once completed'	Subsumed under 'negative or neutral'
Zampieri et al. (B)	High Pr(benefit) across priors	High Pr(harm) across priors	'Indeterminate' — sensitive to prior choice	Bypassed by Bayesian framework	High ROPE probability
Hoenig & Heisey (C)	Not addressed	Requires equivalence testing (TOST)	Wide CI including null and meaningful effects	Post-hoc power is a fallacy	Demonstrated via equivalence testing
Gelman & Carlin (B)	May be Type S/M corrupted	Not addressed directly	Low-power = low information	Corrupts both 'positive' AND 'negative'	Not addressed

Continued on next page

Authority	Positive	Negative	Inconclusive	Under-powered	Neutral
ASA (2016, 2019) (F)	'Don't say statistically significant'	'Large p does not imply evidence for null'	Implicit — 'a conclusion does not address'	Not specifically addressed	Not addressed
Goodman (B)	$p < 0.05$ gives $\leq$13% posterior null prob.	'Nonsignificant $\neq$ no effect'	Default state — external evidence required	Not specifically used	Not used
Senn (C)	Defends p-values as evidence gradients	'Not found good evidence for a real effect'	Default for non-significant results	Depends on pre-specified δ	Not used as distinct term
ICH E9 (F)	Not classified by terminology	Cannot conclude from non-significant superiority	Implicit — non-significant $\neq$ equivalent	Pre-specified power requirements	Requires formal equivalence testing
Campbell & Gustafson CET (F)	NHST rejects null	Equivalence test shows effect within δ	Neither test rejects — explicit third category	Design property leading to inconclusive	Subsumed under 'negative' if within δ

Three points of notable divergence:

"Inconclusive" vs "indeterminate": Harrell and Zampieri's group prefer "indeterminate" or "uninformative" — reflecting their view that even "inconclusive" implies more structure than the data warrant.

"Neutral" is the most inconsistently used term: Hawkins and Samuels bundle it with "negative," Zampieri et al. map it to high ROPE probability, and most others avoid it entirely. The unique contribution of Bayesian analysis is that only Pr(ROPE) can formally operationalize "neutral."

"Underpowered" corrupts all categories: Altman, Bland, Hawkins, and Samuels treat underpowering as a *design* property that causes inconclusiveness (a *result* property). Gelman and Carlin go further: underpowering corrupts *all* result categories — including ostensibly "positive" findings via Type M/S errors.

Despite terminological variation, every authority converges on a single core principle: **a non-significant p-value cannot, by itself, classify a trial as "negative."**

4 Term Definitions and Diagnostic Criteria

4.1 Underpowered

Definition: The study's sample size or event count was insufficient to reliably detect the pre-specified MCID at >80% power. Applies to both $p < 0.05$ and $p > 0.05$ results.

Diagnostic criteria:

- CI includes both the null value and the MCID (wide CI)
- Protocol power calculation target (effect size or event count) was not achieved
- If $p < 0.05$: high risk of Type M (magnitude exaggeration) and Type S (sign error) per Gelman & Carlin (2014) [Gelman and Carlin, 2014]

The core insight — same drug, same truth, different sample sizes: Two trials can find the same point estimate (e.g., HR 0.82) but tell completely different stories:

- $n = 500$ (underpowered): CI 0.55–1.18 → spans 48% benefit to 18% harm. Compatible with everything. Harrell: "Almost nothing was learned." → INCONCLUSIVE
- $n = 4{,}500$ (adequate): CI 0.74–0.91 → entirely in benefit zone, past MCID. → POSITIVE

CI width is proportional to $1/\sqrt{n}$. To halve the CI width, you need approximately 4× the sample size (or events).

The most practical way to detect underpowering: look at the CI. If it includes both the MCID and the null — regardless of the p-value — the trial is underpowered and the result is inconclusive.

The post-hoc power fallacy. Post-hoc power = $f(p\text{-value})$; zero additional information. When $p = \alpha$, observed power is exactly 50% regardless of sample size. The "Power Approach Paradox" [Hoenig and Heisey, 2001]: experiments closer to significance have *higher* observed power, paradoxically implying they better support the null. CONSORT 2010 and Christensen et al. (2024): "Statistical power is exclusively a pre-trial concept."

4.2 Winner's Curse: Why Underpowered "Positive" Results Exaggerate

The mechanism: In a low-powered study, the true effect is small, but estimates are very noisy. To pass the $p < 0.05$ threshold, the observed effect must randomly deviate far from truth. Most estimates cluster near the true (small) effect and fail to reach significance. Only the rare, grossly overestimated values cross the threshold. These are the ones that get published — and they are "cursed" with an inflated effect.

Gelman & Carlin (2014) simulation — study with 6% power [Gelman and Carlin, 2014]:

- Type M error: 9.7× — the published effect is nearly 10 times the true effect
- Type S error: 24% — 1 in 4 significant results has the wrong sign entirely

Practical chain reaction: Researcher takes the inflated effect (e.g., HR 0.65) and uses it for the confirmatory trial's power calculation → "500 patients will suffice" → but the true effect is HR 0.92, so 500 patients are grossly insufficient → confirmatory trial is "negative" → "first result failed to replicate" → the problem was never the drug, it was the inflated initial estimate.

Empirical evidence:

- Button et al. (*Nat Rev Neurosci*, 2013) [Button et al., 2013]: Underpowered studies (median power 8–31%) inflate initial effect estimates by 25–50%
- Ioannidis (*PLoS Med*, 2005) [Ioannidis, 2005]: Low power + modest prior probability → a statistically significant result is more likely false than true (PPV $< 50\%$)
- Sidebotham & Barlow (*Anaesthesia*, 2024) [Sidebotham and Barlow, 2024]: Directly warned against using small-trial effects for confirmatory power calculations

4.3 Inconclusive

Definition: The CI is too wide to permit a definitive classification — it spans both the null and the MCID, or spans from clinically meaningful benefit to clinically meaningful harm.

Diagnostic criteria: $p > 0.05$ AND the 95% CI crosses both the null and the MCID thresholds. Example: HR = 0.90 (CI 0.65–1.24) → both 35% benefit and 24% harm are possible.

Freiman et al. (*NEJM*, 1978) [Freiman et al., 1978]: Of 71 "negative" RCTs, only 6% had $\geq$90% power to detect a 25% improvement.

Bayesian signature: No probability dominates — Pr(benefit) $\sim$ 38%, Pr(equivalence) $\sim$ 35%, Pr(harm) $\sim$ 18%. The posterior is spread across all possibilities almost uniformly. This is the mathematical definition of "the data haven't told us anything decisive."

Conditional Equivalence Testing (CET). Campbell & Gustafson (*PLoS ONE* 2018) [Campbell and Gustafson, 2018]: CET is the only framework that explicitly and formally defines all three outcome categories as part of a single testing procedure:

- **Positive:** Standard NHST rejects the null (significant difference found)
- **Negative:** Equivalence test (TOST) demonstrates the effect lies within a pre-specified equivalence margin δ
- **Inconclusive:** Neither test rejects — insufficient evidence to claim either a difference or equivalence

The two-stage procedure works sequentially: standard NHST is performed first; if it fails to reject ($P > \alpha$), an equivalence test is conducted. CET requires pre-specification of an equivalence margin (δ), analogous to the MCID. Simulation studies showed CET and Bayesian JZS Bayes Factor testing reach similar conclusions in many scenarios.

Important: 54–56% of non-significant RCTs use "no difference" or "no benefit" in the abstract (Gates et al., 2019). The correct approach: emphasize uncertainty + report CI.

4.4 Negative vs Neutral: The Critical Distinction

These two terms are the most commonly confused. The distinction is precise:

Negative says: "This treatment does NOT provide meaningful benefit." But it leaves an open question — the CI may extend toward harm. Only the benefit side is excluded by MCID.
Example: HR 0.95 (CI 0.85–1.07), MCID = 0.80 → Entire CI above 0.80, ≥20% benefit excluded. But upper bound 1.07 means 7% harm still possible.
Bayesian: Pr(MCID benefit) ≈ 3%, Pr(ROPE) ≈ 88%, Pr(harm) ≈ 20%.

Neutral says: "The two treatments are essentially THE SAME." The CI is so narrow that both benefit and harm are excluded.
Example: HR 0.98 (CI 0.94–1.02) → Neither 20% benefit nor 25% harm is possible.
Bayesian: Pr(MCID benefit) < 1%, Pr(ROPE) ≈ 97%, Pr(harm) ≈ 5%.

Key formula:

$$\text{Negative} = \text{benefit excluded (one-sided)} \rightarrow \text{"doesn't work"}$$
$$\text{Neutral} = \text{benefit AND harm excluded (two-sided)} \rightarrow \text{"they're the same"}$$

Neutral is a much stronger statement. It requires a narrow CI and ideally formal confirmation through an equivalence test or Bayesian Pr(ROPE). The unique power of Bayesian analysis is that only it can formally quantify Pr(equivalence) — the single number that defines neutral.

4.5 Positive

Definition: CI entirely on benefit side and beyond MCID — both statistical and clinical significance. Verify adequate power; if low, suspect Type M error (Gelman & Carlin).

4.6 Harmful

Definition: CI entirely in harm zone beyond MCID-harm. Bayesian confirmation: Pr(severe harm) > 40% across all priors including optimistic. The ART trial (JAMA 2017) is the paradigmatic example — see Section 7.3.

5 The Three Faces of $p > 0.05$

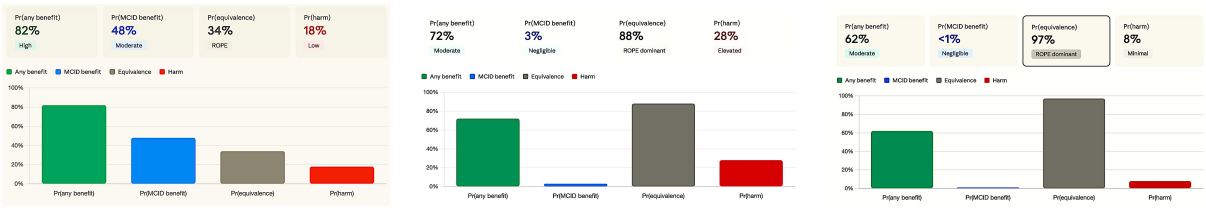
This is the most important conceptual illustration in the entire framework. Three trials, all $p > 0.05$, all "non-significant" — yet they mean completely different things:

Scenario A: Underpowered ($n = 400$, $p = 0.18$). HR 0.78 (CI 0.52–1.18). CI spans 48% benefit to 18% harm. Compatible with everything.
→ INCONCLUSIVE — "Almost nothing was learned" (Harrell).

Scenario B: Negative ($n = 8{,}000$, $p = 0.22$). HR 0.95 (CI 0.87–1.04). Narrow CI near null. Entire CI above MCID. Best case: only 13% benefit.
→ NEGATIVE — "Clinically meaningful benefit is ruled out."

Scenario C: Neutral ($n = 12{,}000$, $p = 0.48$). HR 0.98 (CI 0.93–1.03). Very narrow CI hugs null. Both benefit and harm excluded.
→ NEUTRAL — "No meaningful difference in either direction."

The tragedy: Most papers report all three as "no significant difference" — a single phrase hiding three fundamentally different conclusions. This is the error Altman warned about in 1995 and that persists in over half of published RCTs today.



(a) Inconclusive (b) Negative (c) Neutral

Figure 3: **The three faces of $p > 0.05$.** Three trials with non-significant p-values produce completely different Bayesian posterior probability profiles. (a) Inconclusive: no metric dominates. (b) Negative: ROPE dominates but MCID benefit ≈ 0. (c) Neutral: ROPE overwhelmingly dominant (>97%).

6 Bayesian Analysis for All 6 Classifications

6.1 Why Bayesian?

The frequentist framework answers: "Can we reject the null?" (Yes/No — same answer for all three $p > 0.05$ scenarios above). The Bayesian framework answers four separate questions: "What is the probability of benefit? Of meaningful benefit? Of equivalence? Of harm?" — producing four numbers that create completely different profiles for each class.

6.2 Three Bayesian Metrics from Zampieri et al. (AJRCCM 2021)

Zampieri, Casey, Shankar-Hari, Harrell, and Harhay formalized a standardized framework for Bayesian reanalysis of clinical trials [Zampieri et al., 2021]:

Table 3: **Bayesian posterior metrics** for trial classification (per Zampieri et al. 2021).

Metric	Definition	Maps to
Pr(outstanding benefit)	Pr(OR < MCID)	Positive end
Pr(ROPE)	Pr(1/1.1 < OR < 1.1)	Neutral (equivalence)
Pr(severe harm)	Pr(OR > MCID-harm)	Harmful end
Supplementary metrics:		
Pr(any benefit)	Pr(OR < 1.0)	
Posterior median OR	Median + 95% CrI	
Absolute risk reduction	ARR = CER × (1 − OR)	

These three metrics together produce a unique signature for each of the 6 classifications. The ROPE concept [Kruschke, 2018] provides the Bayesian analog to MCID — if

the Highest Density Interval falls entirely within ROPE → equivalence; outside → effect present; partial overlap → inconclusive.

6.3 Complete Bayesian Fingerprints

Table 4: **Complete Bayesian fingerprints** for all six trial classifications. Each class produces a distinctive posterior probability profile.

Class	Pr(any benefit)	Pr(MCID benefit)	Pr(ROPE)	Pr(harm)	Dominant signal
Positive	>99%	>90%	<1%	<1%	MCID benefit overwhe
Imprecise (+)	∼97%	50–70%	∼8%	∼3%	Gap: any benefit MCID benefit ↓
Neutral	∼62%	<1%	>90%	∼5%	ROPE (equivalence) d nates
Inconclusive	∼74%	∼38%	∼35%	∼18%	Nothing dominates
Negative	∼72%	∼3%	∼88%	∼20%	MCID benefit ≈ 0 + ROPE
Harmful	<1%	<1%	∼4%	>95%	Harm dominates

Table 5: **Bayesian classification criteria.** Primary and supporting conditions for assigning each verdict.

Verdict	Primary signal	Supporting condition
Positive	Pr(outstanding benefit) > 80%; Pr(ROPE) = 0; Pr(harm) = 0	Harm posterior near zero across all prior specs
Imprecise (+)	Pr(any benefit) high; Pr(outstanding benefit) 40–70%	Gap between "any benefit" and "outstanding benefit" = diagnostic imprecision
Neutral	Pr(ROPE) > 90%; Pr(outstanding benefit) = 0	Pr(severe harm) < 10%; benefit effectively excluded
Inconclusive	All posteriors < 50%; no single hypothesis wins	Posterior mass spread across benefit, ROPE, harm simultaneously
Negative	Pr(outstanding benefit) = 0; Pr(ROPE) > 80%	Pr(severe harm) < 20%; harm not dominant driver
Harmful	Pr(severe harm) > 40%; robust across prior specs	Pr(benefit) = 0; harm signal prior-independent

7 Real-World Examples: Bayesian Reanalysis in Action

7.1 EOLIA — ECMO for Severe ARDS (NEJM 2018) — Bayesian Rescues Benefit

Frequentist verdict: "NEGATIVE" — 60-day mortality: ECMO 35% vs control 46%, RR 0.76 (95% CI 0.55–1.04), $p = 0.09$.

Conclusion in the paper: "Early ECMO was not associated with mortality that was significantly lower" [Combes et al., 2018].

Bayesian reanalysis (Goligher et al., *JAMA* 2018) [Goligher et al., 2018]: Even under strong skepticism (equivalent to a hypothetical 264-patient trial finding zero effect), there is an 88% probability ECMO reduces mortality. With prior studies incorporated, Pr(benefit) reaches 99%.

The $p = 0.09$ label of "negative" obscured what was actually strong evidence of benefit — an 11% absolute mortality reduction with 96% posterior probability.

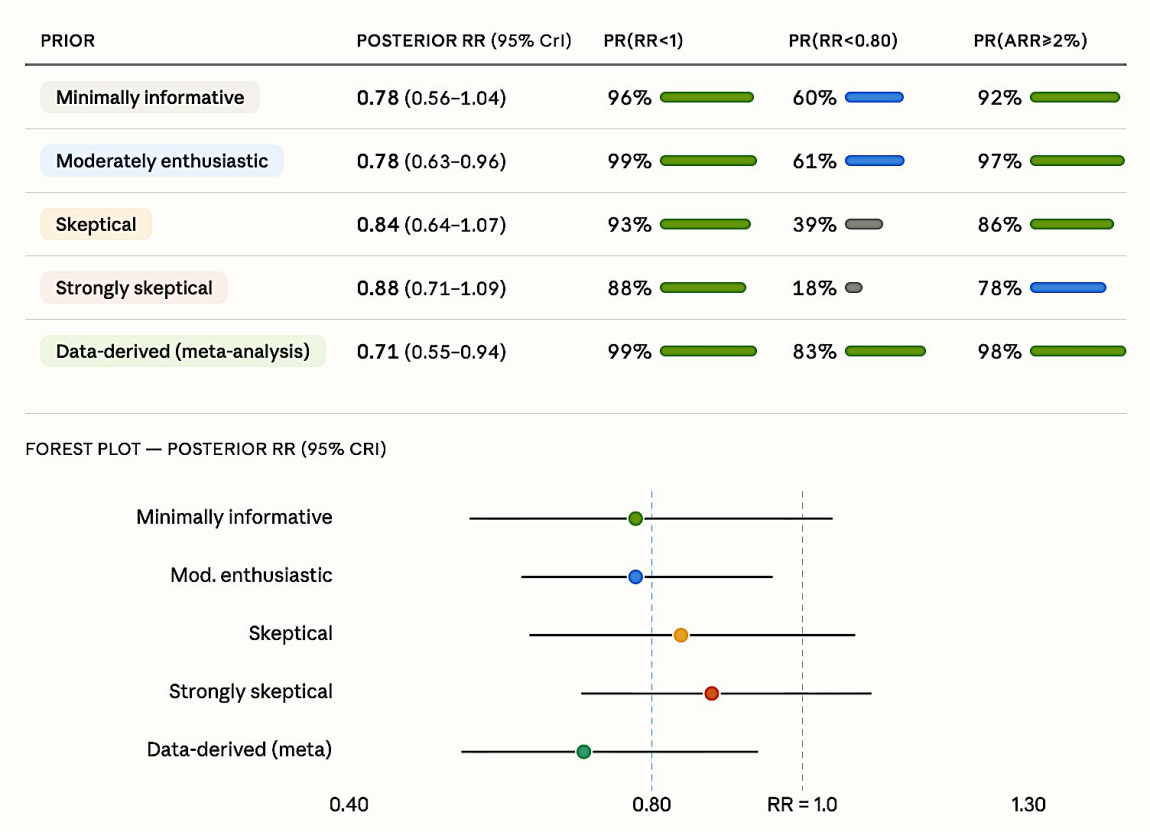


Figure 4: **EOLIA Bayesian reanalysis.** Posterior RR estimates across five prior specifications. Even the strongly skeptical prior yields 88% probability of benefit. $\Pr(\text{RR} < 0.80) = 18$–$83\%$ depending on prior; $\Pr(\text{ARR} > 2\%) = 78$–$98\%$.

7.2 ANDROMEDA-SHOCK — CRT vs Lactate-Guided Resuscitation (JAMA 2019) — Bayesian Rescues Benefit

Frequentist verdict: "NEGATIVE" — 28-day mortality: CRT-guided 34.9% vs lactate-guided 43.4%, HR 0.75 (95% CI 0.55–1.02), $p = 0.06$ [Hernández et al., 2019].

Bayesian reanalysis (Zampieri et al., *AJRCCM* 2020) [Zampieri et al., 2020]: Pr(benefit) exceeds 90% under ALL four priors — including the pessimistic one that assumes the treatment is harmful.

A critical finding: simply switching from Cox to logistic regression yields $p = 0.022$ — the "negative" label was an artifact of the statistical model choice, not the data. The Bayesian approach is immune to this model dependency.

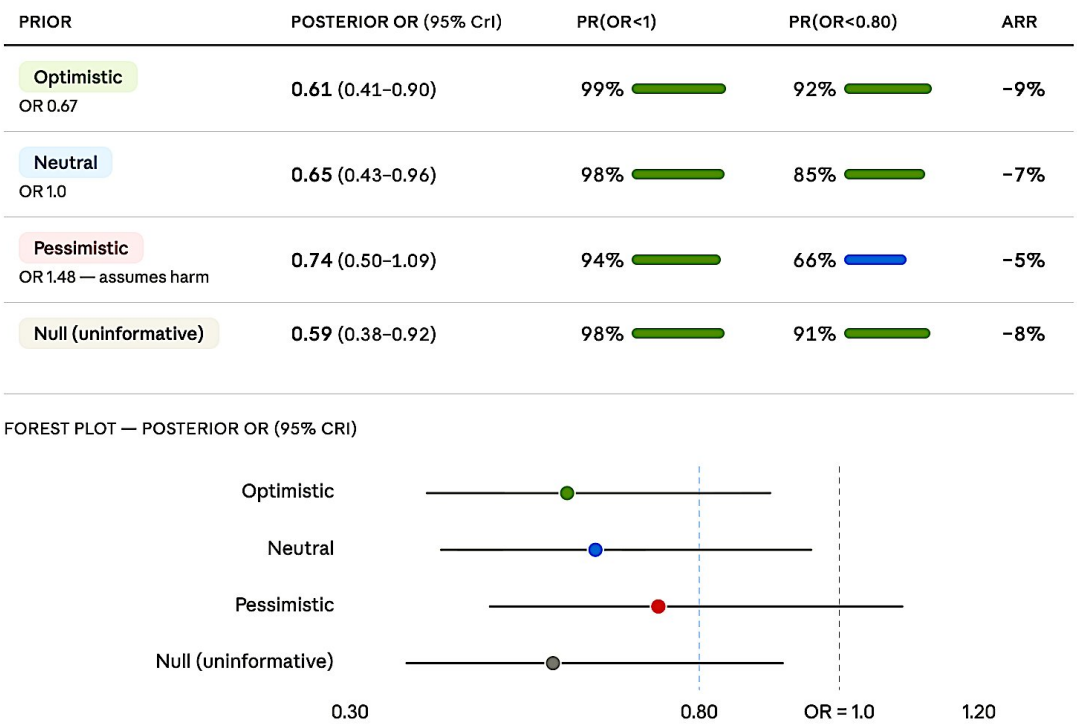


Figure 5: **ANDROMEDA-SHOCK Bayesian reanalysis.** Posterior OR estimates under four priors. Pr(OR < 1) ranges from 94% (pessimistic) to 99% (optimistic). Even the pessimistic prior yields 66% probability of Pr(OR < 0.80).

7.3 ART — Open-Lung Ventilation in ARDS (JAMA 2017) — Bayesian Confirms Harm

Frequentist: borderline — OR 1.27 (95% CI 0.99–1.63), $p = 0.057$ [Cavalcanti et al., 2017].

Bayesian reanalysis (Zampieri et al., *AJRCCM* 2021) [Zampieri et al., 2021]: Even the optimistic prior (which assumes the treatment is beneficial!) yields 93.6% probability of harm and 34.8% probability of severe harm. Pr(benefit) $\approx 0\%$ under ALL priors. Prior sensitivity $I^2 = 0.11$ — priors barely matter, the data overwhelm them.

This is what a decisive HARMFUL classification looks like in Bayesian language — the mirror image of EOLIA and ANDROMEDA-SHOCK.

7.4 The Pattern Across All Three Reanalyses

Key principle: When the conclusion is robust across skeptical-to-enthusiastic priors, it's the data talking, not the prior. Bayesian analysis works in all directions — rescues benefit when it exists, confirms harm when it exists, and formally quantifies inconclusive when data are insufficient.

8 Cardiology RCT Examples

Below are analyses of selected cardiology RCTs demonstrating each of the six classifications.

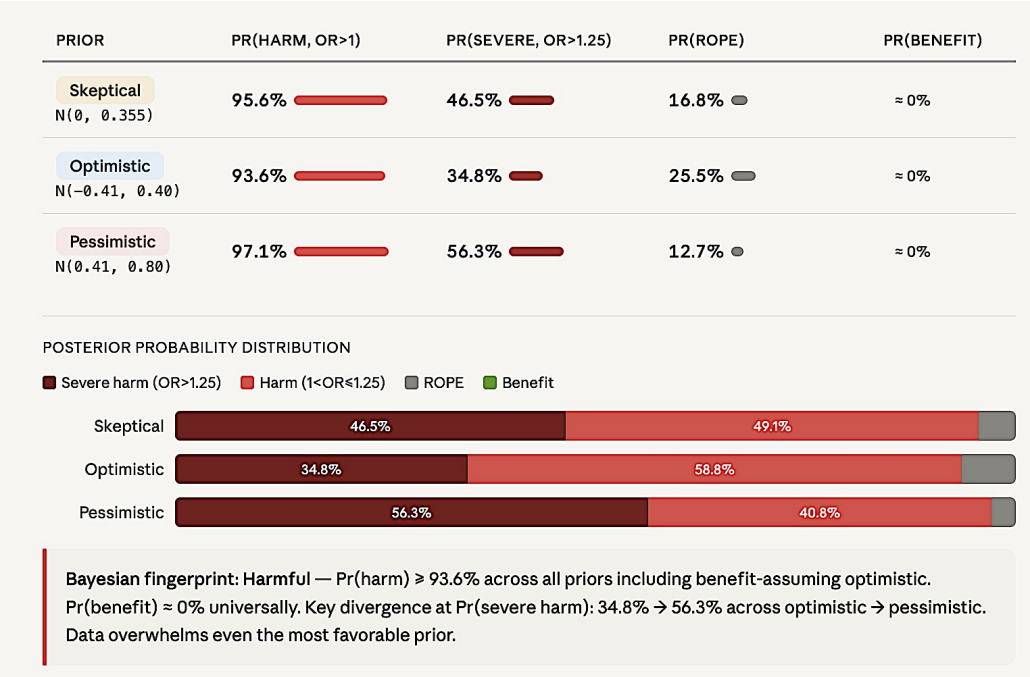


Figure 6: **ART Bayesian reanalysis.** Posterior probability distributions under three priors. Pr(harm) $>$ 93.6% across all priors including benefit-assuming optimistic. Pr(benefit) = 0% universally.

9 Most Common Errors

Error #1: $p > 0.05$ = "No difference." $p > 0.05$ only means the null cannot be rejected. Among 423 "negative" oncology RCTs: only 10.6% had adequate power.

Error #2: Post-hoc power. Post-hoc power = $f(p\text{-value})$ = zero information. CI tells you everything.

Error #3: "Neutral" for inconclusive. Neutral = narrow CI + MCID excluded. Wide CI = inconclusive.

Error #4: Taking underpowered "positive" at face value. Winner's curse: effect inflated 5–10×.

Error #5: Labeling "negative" when Bayesian shows strong benefit. EOLIA ($p = 0.09$, Pr(benefit) = 96%) and ANDROMEDA-SHOCK ($p = 0.06$, Pr(benefit) = 98%) — frequentist "non-significance" masked decisive evidence.

10 Reporting Templates

11 Non-Inferiority and Equivalence

Non-inferiority: One-sided threshold Δ. CI harm-side within $\Delta \rightarrow$ "non-inferior." Exceeds $\rightarrow$ "inferior." Crosses $\rightarrow$ inconclusive. FDA NI guidance (2016): M1/M2 margin framework.

Equivalence: Two-sided $\pm\Delta$. CI within band $\rightarrow$ "equivalent." Crosses $\rightarrow$ inconclusive.

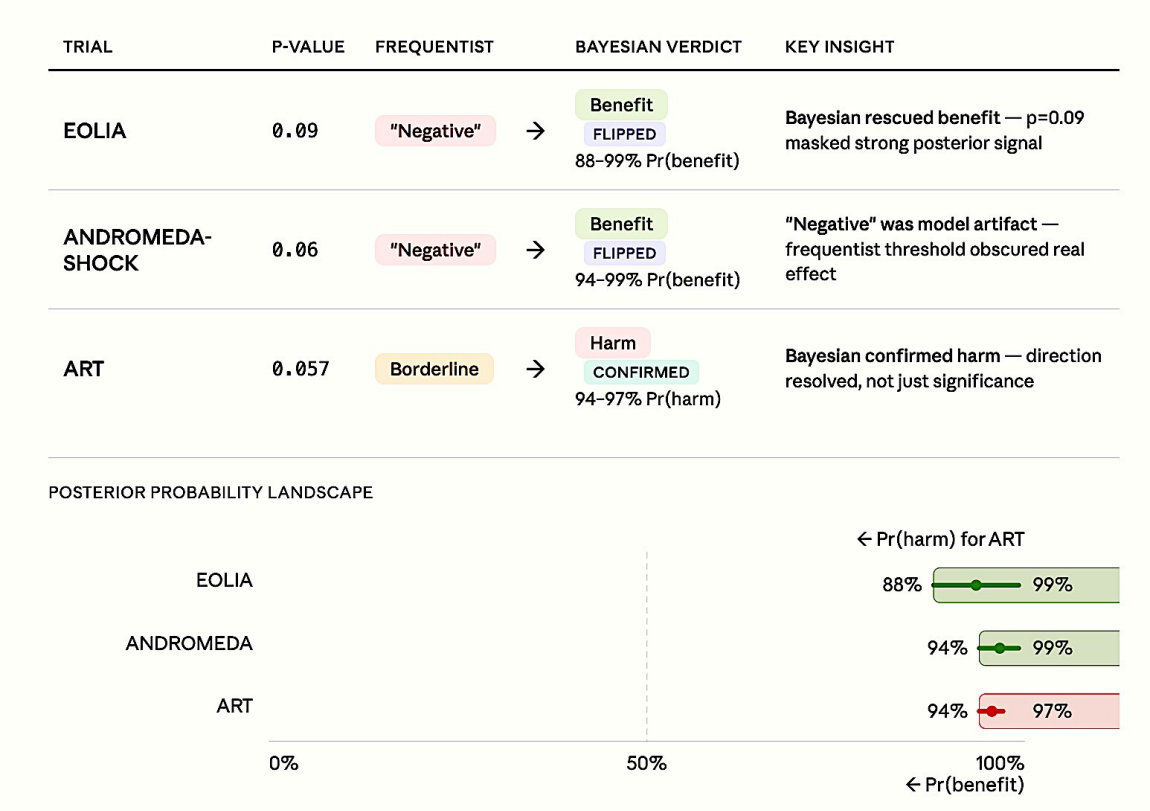
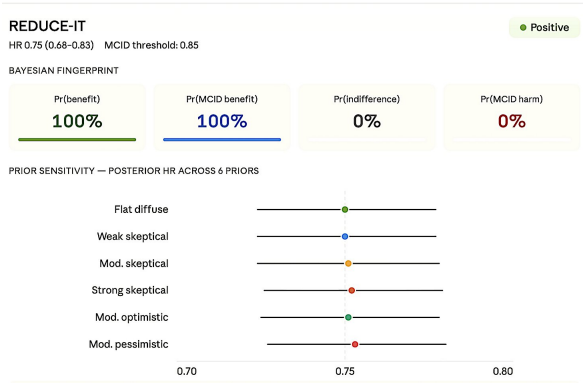
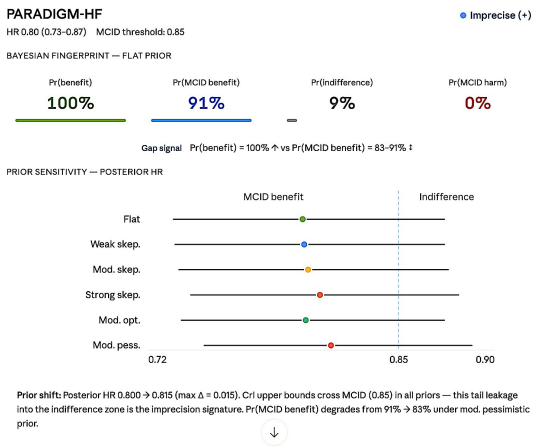


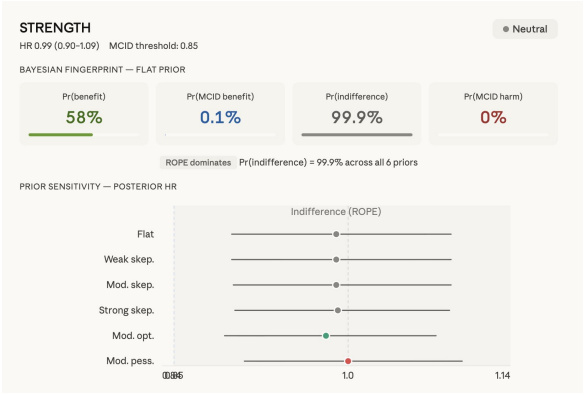
Figure 7: **Pattern across three landmark reanalyses.** EOLIA and ANDROMEDA-SHOCK: frequentist "negative" labels flipped to benefit (88–99% posterior probability). ART: borderline frequentist result confirmed as harmful (94–97% posterior probability). In all three cases, p-values near 0.05 obscured decisive evidence that Bayesian analysis revealed.



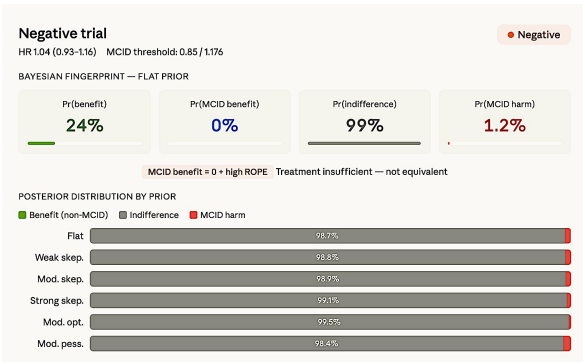
(a) REDUCE-IT — Positive: HR 0.75 (0.68–0.83), MCID 0.85



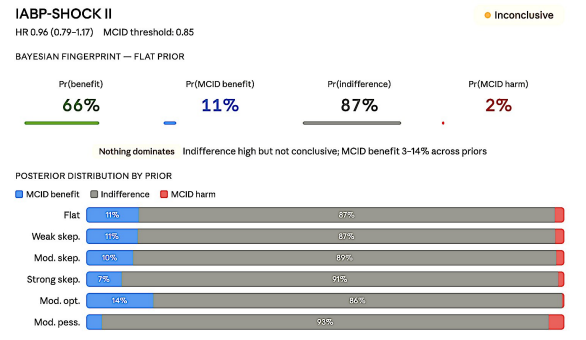
(b) PARADIGM-HF — Imprecise (+): HR 0.80 (0.73–0.87), MCID 0.85



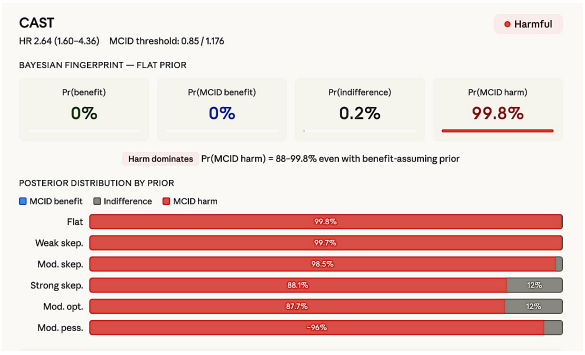
(c) STRENGTH — Neutral: HR 0.99 (0.90–1.09), MCID 0.85



(d) dal-OUTCOMES — Negative: HR 1.04 (0.93–1.16), MCID 0.85



(e) IABP-SHOCK II — Inconclusive: HR 0.96 (0.79–1.17), MCID 0.85



(f) CAST — Harmful: HR 2.64 (1.60–4.36)

Figure 8: **Cardiology RCT examples** illustrating all six classifications with Bayesian fingerprints and prior sensitivity analyses.

Table 6: **Recommended reporting templates** for each trial classification.

Verdict	**Example**	**Reporting template**
Positive	HR = 0.74 (0.65–0.85)	>20% reduction threshold exceeded. Bayesian Pr(MCID benefit) = 97%. MCID benefit dominant; CI excludes null; high precision.
Imprecise (+)	HR = 0.70 (0.49–0.98)	Benefit likely but magnitude uncertain. Bayesian Pr(MCID benefit) = 62%. Any-benefit high, MCID uncertain; wide CI; needs larger N.
Neutral	RR = 0.98 (0.94–1.02)	Clinically meaningful difference excluded. Bayesian Pr(ROPE) = 97%. Equivalence established; tight CI around null; high precision.
Inconclusive	HR = 0.90 (0.65–1.24)	Neither benefit nor harm excluded. Bayesian: no Pr exceeds 50%. Nothing dominates; CI spans null widely; uninformative.
Negative	RR = 0.95 (0.85–1.07)	>20% reduction excluded. Bayesian Pr(MCID benefit) = 3%. MCID benefit $\approx$ 0; high ROPE; treatment insufficient.
Harmful	OR = 1.27 (0.99–1.63)	Bayesian Pr(severe harm) = 47% even with optimistic prior. Harm dominates; prior-resistant signal; safety concern.

Conclusion

"Never interpret $p > 0.05$ as 'no effect.' Always report CI + effect size + clinical significance. When p is near 0.05, compute Bayesian posteriors before labeling the trial." — Harrell, Pocock, Zampieri, ASA, ICH E9.

References

Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ*. 1995;311:485.

Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. *Am Stat*. 2016;70:129–133.

Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "$p < 0.05$". *Am Stat*. 2019;73(sup1):1–19.

Harrell FE. Statistical errors in the medical literature. fharrell.com/post/errmed. 2017.

Pocock SJ, Stone GW. The primary outcome fails — what next? *NEJM*. 2016;375:861–870.

Hawkins AT, Samuels JD. How to interpret a clinical trial that did not meet its primary outcome. *JAMA*. 2021;326:1875–1876.

Zampieri FG, Casey JD, Shankar-Hari M, Harrell FE, Harhay MO. Using Bayesian methods to augment the interpretation of critical care trials. *Am J Respir Crit Care Med*. 2021;203:543–552.

Goligher EC, Tomlinson G, Hajage D, et al. Extracorporeal membrane oxygenation for severe acute respiratory distress syndrome and posterior probability of mortality benefit in a post hoc Bayesian analysis of a randomized clinical trial. *JAMA*. 2018;320:2251–2259.

Zampieri FG, Damiani LP, Bakker J, et al. Effects of a resuscitation strategy targeting peripheral perfusion status versus serum lactate levels among patients with septic shock: a Bayesian reanalysis of the ANDROMEDA-SHOCK trial. *Am J Respir Crit Care Med*. 2020;201:423–429.

Kruschke JK. Rejecting or accepting parameter values in Bayesian estimation. *Adv Methods Pract Psychol Sci*. 2018;1:270–280.

Combes A, Hajage D, Capellier G, et al. Extracorporeal membrane oxygenation for severe acute respiratory distress syndrome. *NEJM*. 2018;378:1965–1975.

Hernández G, Ospina-Tascón GA, Damiani LP, et al. Effect of a resuscitation strategy targeting peripheral perfusion status vs serum lactate levels on 28-day mortality among patients with septic shock: the ANDROMEDA-SHOCK randomized clinical trial. *JAMA*. 2019;321:654–664.

Cavalcanti AB, Suzumura ÉA, Laranjeira LN, et al. Effect of lung recruitment and titrated positive end-expiratory pressure (PEEP) vs low PEEP on mortality in patients with acute respiratory distress syndrome: a randomized clinical trial. *JAMA*. 2017;318:1335–1345.

Gelman A, Carlin J. Beyond power calculations: assessing Type S (sign) and Type M (magnitude) errors. *Perspect Psychol Sci*. 2014;9:641–651.

Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. *NEJM*. 1978;299:690–694.

Button KS, Ioannidis JPA, Mokrysz C, et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci*. 2013;14:365–376.

Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. *Am Stat*. 2001;55:19–24.

Ioannidis JPA. Why most published research findings are false. *PLoS Med*. 2005;2:e124.

Sidebotham D, Barlow J. Effect sizes from small discovery trials should not be used for confirmatory power calculations. *Anaesthesia*. 2024.

Campbell H, Gustafson P. Conditional equivalence testing: an alternative remedy for publication bias. *PLoS ONE*. 2018;13:e0195145.

McGlothlin AE, Lewis RJ. Minimal clinically important difference: defining what really matters to patients. *JAMA*. 2014;312:1342–1343.

Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Wiley; 2004.