Developing a Reproducible Transportability Framework for Real-World External Control Arms: An Adult Sepsis Example Using MIMIC-IV

MSR11

Nikhil Tiwari[1], Shivesh Gupta[1]

[1]Frekil, San Francisco, CA, USA

Background

- FDA accepts external control arms (ECAs) built from patient-level real-world data.[11] EU JCA calls them high risk and asks for full patient-level data and pre-set methods.[12]
- For other sponsors' trials, only the **published baseline table (Table 1)** and outcomes are public.[13] So an ECA can be checked against a past trial only through these.
- An ECA can replace a control arm only if real-world patients, made to resemble the trial's, fare as its controls did (**transportability**). Few studies test this.

Objective

To find which steps bring an EHR control arm close to a real trial control arm when only Table 1 is available, freeze them as one rule, and test that rule on trials it had not seen.

Methods

Data: MIMIC-IV v3.1[8] (one US academic centre, 2008–2022); adults, first ICU stay, Sepsis-3. Second database: eICU-CRD v2.0[9] (208 US hospitals, 2014–2015).

Framework, and why each step is there

1. **Copy the trial's entry rule, time zero (t0) and outcome day.**
   Why: starting follow-up at a different point compares patients at a different stage of illness and invites immortal-time bias.[14]
2. **Drop patients the trial could not enrol:** DNR/DNI orders before t0, and any heart surgery in the same stay.
   Why: trials rarely enrol them. An early DNR triples the odds of death in septic shock.[15] Shock after bypass is vasoplegia, not infection.[16]
3. **Start the clock 6 h after t0:** keep those still alive with no new DNR or comfort-care order; count deaths from then.
   Why: trials take hours to enrol, so patients who die or stop treatment early never join. 6 h set on learning trials; removes 3–9%. 0, 12, 24 h: see Sensitivity.
4. **Rebuild SOFA** from the worst raw values in 24 h before t0.
   Why: SOFA uses each day's worst value.[17] Charted SOFA uses another window, so weights would match a different score.
5. **Impute missing covariates** (M = 10; skewed labs on log scale; outcome not used in the model at all).
   Why: dropping incomplete records keeps a non-random subset.[18] No outcome in the model: design fixed first.[19]
6. **Stable balancing weights** to Table 1: each covariate within 0.05 SD, weights capped at 4.[1] Pooled by Rubin's rules.[3]
   Why: plain MAIC hit the same balance with weights up to 75. Stable weights give the largest ESS for that balance.[20] Cap set by bootstrap balance alone.[2]

An ECA "agrees" (compatible, not equivalent) if all four hold:

- **Enough patients count** — ESS ≥ 20%
- **No patient dominates** — largest weight ≤ 20
- **Groups match Table 1** — every |SMD| ≤ 0.1
- **Compatible with the trial** — 95% CIs overlap

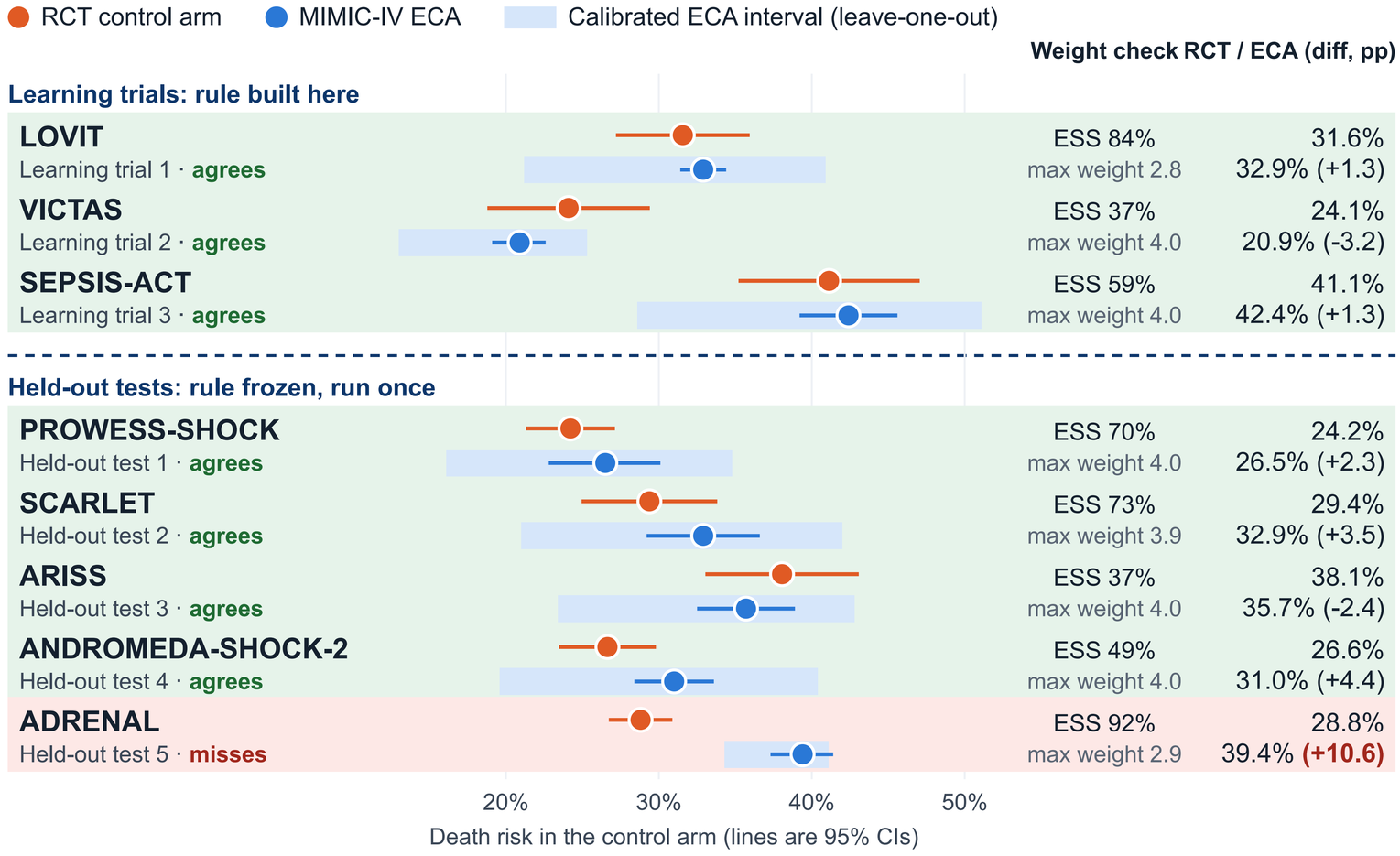
Benchmark control arms: learn, freeze, then test

Trial	Role	Day	Death	ECA n
LOVIT 2022	Learn	28	31.6%	1,927
VICTAS 2021	Learn	30	24.1%	7,544
SEPSIS-ACT 2019	Learn	90	41.1%	1,687
PROWESS-SHOCK 2012	**Test**	28	24.2%	1,619
SCARLET 2019	**Test**	28	29.4%	904
ARISS 2026	**Test**	28	38.1%	2,614
ANDROMEDA-SHOCK-2 2025	**Test**	28	26.6%	2,731
ADRENAL 2018	**Test**	90	28.8%	2,529

Learn: each step above was tried and kept on these 3 trials. Test: rule run once, unchanged. ECA n is before exclusions.

From each trial's published Table 1 alone, one fixed rule was compatible with the real control arm in 7 of 8 sepsis trials.

Figure 1. Control-arm death risk: trial vs MIMIC-IV ECA. All 8 trials pass the weight check; every |SMD| ≤ 0.05.



Sensitivity analyses

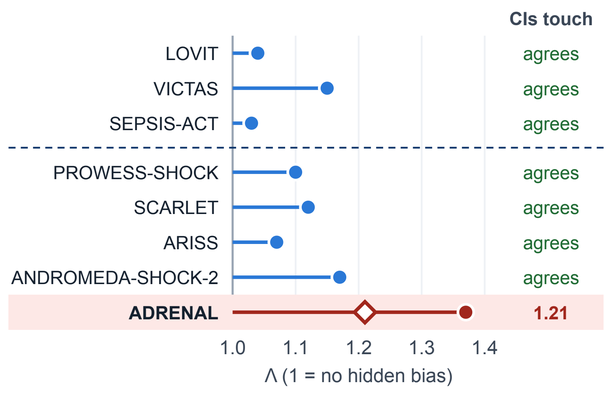
Weight cap 4, 10 or none, and **bias calibration**[4,5] (blue band; mean +2 pp): no verdict changes.

Start time. ADRENAL enrolled 21 h after shock onset.[10] 12 or 24 h: same 7 of 8; 0 h: 6 of 8 (ANDROMEDA-SHOCK-2 also misses).

Stricter agreement. RCT-DUPLICATE metrics[4] (|z| < 1.96, ECA inside trial CI): 6 of 8 (ANDROMEDA-SHOCK-2 also misses). ±5 pp equivalence: 0 of 8; trial CIs too wide.

Joint mix. Table 1 says how many patients are older and how many are ventilated, but not how many are both.[13] In MIMIC, we matched one half of patients to the other with and without these "both" counts: death rates moved ≤ 0.2 pp. Trials may still differ.

Figure 2. Hidden bias (Λ) per trial.[6]



Λ: how strong a hidden factor must be to explain the gap (1 = none). Dot: close it fully; ◇: make CIs overlap (ADRENAL only; E-value 1.71[7]).

Same rule in eICU (208 hospitals)

Adjustment. Trials count all deaths; eICU sees only those in hospital. We add the deaths after discharge seen in US sepsis cohorts: +7.7 pp by 30 d[21] and +10.2 pp by 90 d.[22] MIMIC, linked to state death records, shows the same order (+4, +10 pp). CIs exclude add-back uncertainty; any add-back from +3.1 to +8.4 pp gives the same 28/30 d verdicts.

Trial	RCT control	eICU ECA, adjusted	
PROWESS-SHOCK 28 d	24.2% (21.3–27.1)	28.7% (26.4–31.0)	agrees
SCARLET 28 d	29.4% (24.9–33.8)	33.9% (28.9–38.9)	agrees
LOVIT 28 d	31.6% (27.2–36.0)	30.2% (28.6–31.8)	agrees
VICTAS 30 d	24.1% (18.8–29.4)	24.4% (22.7–26.0)	agrees
SEPSIS-ACT 90 d	41.1% (35.2–47.1)	35.1% (32.4–37.8)	agrees
ADRENAL 90 d	28.8% (26.7–30.9)	39.4% (36.2–42.6)	misses

Death % (95% CI). ARISS, ANDROMEDA-SHOCK-2: not feasible.

Why ADRENAL misses

miss **MIMIC** 39.4% vs trial 28.8%. Main reason: patients put on DNR or comfort care in the first 48 h; trials rarely enrol them.[23] Drop them and non-septic shock: 33.0%, **agrees**.

miss **eICU** sees only deaths in hospital, so it can't test the 90-day gap. It can test the cause: dropping DNR/comfort-care patients lowers deaths in hospital by 4.8 pp in eICU (29.2% → 24.4%), as in MIMIC (30.8% → 25.4%, −5.4 pp).

Found after the rule was fixed; rule unchanged. Removing them lowers every trial 4–10 pp: likely, not proven.

Conclusions

- **Compatible from summaries alone.** 6 of 8 met RCT-DUPLICATE's metrics (75%), like its patient-level claims emulations (66–75%; theirs compared effects).[4]
- **Design matters more than weighting**: copy entry, t0 and who the trial could enrol.
- Frozen after 3 trials: compatible in 4 of 5 new ones (mean bias +2 pp). Same code, eICU: 5 of 6.
- Report past-trial bias with every ECA. ADRENAL's gap (Λ 1.21) may reflect care practice; no Table 1 fixes it.

Limitations: one US centre; Table 1 only; 8 trials (7 of 8: 95% CI 47–100%); eICU later deaths assumed.

1. Zubizarreta JR. JASA 2015;110:910-22.
2. Wang Y, et al. Biometrika 2020;107:93-105.
3. Rubin DB. Multiple Imputation. Wiley; 1987.
4. Wang SV, et al. JAMA 2023;329:1376-85.
5. Wang SV. BenchExCal; 2025. PMC12087693.
6. Asiaee A, et al. arXiv:2602.09595. 2026.
7. VanderWeele TJ. Ann Intern Med 2017.
8. Johnson AEW, et al. Sci Data 2023;10:1.
9. Pollard TJ, et al. Sci Data 2018;5:180178.
10. Trials: Lamontagne 2022; Sevransky 2021; Laterre 2019; Ranieri 2012; Vincent 2019; Sakr 2026; A-S2 2025; Venkatesh 2018.
11. FDA. Externally controlled trials; 2023.
12. EU HTA CG. Indirect comparisons; 2024.
13. Phillippo DM, et al. NICE DSU TSD 18; 2016.
14. Hernán MA. J Clin Epidemiol 2016;79:70-5.
15. Bradford MA. Crit Care Med 2014;42:2042.
16. Busse LW, et al. Crit Care 2020;24:36.
17. Vincent JL. Intensive Care Med 1996;22:707.
18. Sterne JAC, et al. BMJ 2009;338:b2393.
19. Rubin DB. Stat Med 2007;26:20-36.
20. Jackson D. Res Synth Methods 2021;12:333.
21. Prescott HC, et al. JAMA 2025;333:1724-6.
22. Govindan S, et al. Trials 2018;19:198.
23. Gordon AC, et al. N Engl J Med 2016;375:1638-48.

MIMIC-IV and eICU-CRD used under PhysioNet credentialed access; only summary numbers shown. Both authors are co-founders of Frekil. No external funding.