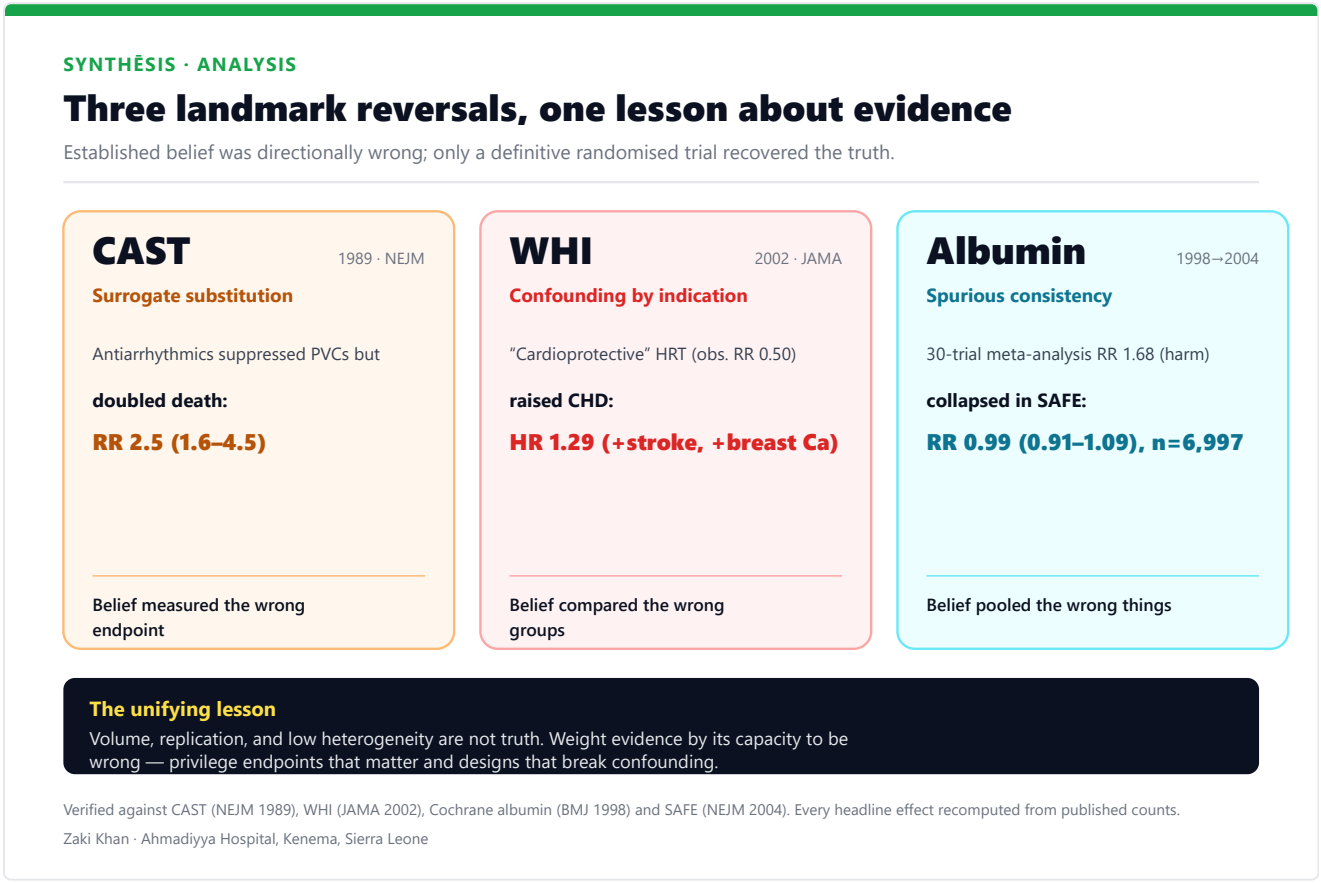
SYNTHĒSIS · ANALYSIS

Three reversals, one lesson: surrogate substitution, confounding, and spurious consistency in the evidence that misled medicine

Zaki Khan

Ahmadiyya Hospital, Kenema, Sierra Leone

Corresponding author: zakikhan101@gmail.com



Visual abstract. Three landmark reversals — CAST, WHI and the albumin controversy — in which established belief was directionally wrong and only a definitive randomised trial recovered the true effect, each through a distinct mechanism.

ABSTRACT

Background. Clinical practice is periodically built on evidence that proves not merely incomplete but directionally wrong. We ask whether such reversals share a common structure.

Approach. We re-examine three landmark evidence reversals — the Cardiac Arrhythmia Suppression Trial (CAST, 1989), the Women's Health Initiative oestrogen-plus-progestin trial (WHI, 2002), and the albumin controversy (Cochrane meta-analysis 1998 versus the SAFE trial, 2004) — recomputing each headline effect from the published counts and reading them not as three case reports but as three instances of one phenomenon: truth recovery, in which a definitive randomised trial overturns an established belief.

Findings. In each case the prevailing belief was confidently held and wrong in direction. CAST: antiarrhythmics that suppressed the surrogate (ventricular ectopy) increased all-cause death or cardiac arrest, 56/730 versus 22/725, relative risk 2.5 (95% CI 1.6–4.5). WHI: hormone therapy, judged "cardioprotective" across four decades of observational data (pooled observational RR ≈0.50), instead raised coronary heart disease (HR 1.29), stroke (1.41) and breast cancer (1.26). Albumin: a pooled relative risk of death of 1.68 (1.26–2.23) across 30 small trials collapsed to 0.99 (0.91–1.09) in the 6,997-patient SAFE trial.

The three errors have distinct mechanisms — surrogate substitution, confounding by indication, and spurious consistency among underpowered trials sharing biases.

Conclusion. The reversals are not anomalies but a taxonomy. They argue for weighting evidence by its capacity to be wrong — privileging endpoints that matter, designs that break confounding, and adequate power over mere agreement — and for treating consistency in synthesis as something shared bias can manufacture, not a guarantee of truth.

Central claim. Three famous evidence reversals are one phenomenon — truth recovery by a definitive RCT — acting through three distinct failure mechanisms (surrogate substitution, confounding by indication, spurious consistency), which together form a usable taxonomy for weighting evidence.

Introduction: three reversals as one phenomenon

Medicine advances by correction, but some corrections are of a special kind: not refinements at the margin but reversals of direction. A treatment believed to save lives is found to take them; a hormone believed to protect the heart is found to attack it; a fluid believed to harm turns out to be neutral. These are not cases of "we did not yet know." The field knew, taught, and prescribed on a belief that was wrong. We call such episodes truth-recovery events: moments when a single definitive randomised controlled trial (RCT) recovered an effect that the accumulated prior evidence had systematically obscured.

Three of these events — CAST, WHI, and the albumin reversal — are usually told in isolation, each a self-contained cautionary tale. Told together they offer something more useful than three warnings: a taxonomy of how evidence misleads. Each reversal failed at a structurally different point. CAST is a failure of what was measured; WHI is a failure of who was compared; albumin is a failure of how the evidence was pooled. Read as a set, they answer a practical question for anyone who weighs evidence — which features of a body of evidence should raise our suspicion that we are about to be wrong?

Case 1 — CAST: when the surrogate lied

Premature ventricular contractions (PVCs) after myocardial infarction predict sudden death, so suppressing them seemed self-evidently beneficial. The class IC antiarrhythmics encainide and flecainide suppressed PVCs reliably, and a whole prescribing practice followed from that surrogate. CAST tested the assumption against the endpoint that actually matters. In the 1989 preliminary report, 56 of 730 patients on active drug died or had a cardiac arrest, versus 22 of 725 on placebo; recomputing from these counts gives a relative risk of 2.53, matching the trial's reported 2.5 (95% CI 1.6–4.5). The drugs worked perfectly on the surrogate and roughly doubled mortality.

The mechanism of error is surrogate substitution: a biomarker on the causal pathway was treated as though it were the outcome, but the intervention moved the marker and the outcome in opposite directions. The lesson is not that surrogates are useless. It is that a surrogate validated as a prognostic marker is not thereby validated as a treatment target — whether an intervention's effect on the marker transmits to the outcome is a separate, and itself randomised, question.

Case 2 — WHI: when the comparison lied

For forty years, observational cohorts reported that postmenopausal women taking hormone therapy had less coronary disease. The signal was strong and repeatedly replicated; a 1991 quantitative review put the pooled relative risk near 0.50 and judged the protective effect "unlikely to be explained by confounding." WHI randomised 16,608 women to combined oestrogen-plus-progestin or placebo and found the

opposite: hazard ratios of 1.29 for coronary heart disease, 1.41 for stroke, and 1.26 for invasive breast cancer.

The observational signal was not noise; it was bias with a direction. Women who took hormones were systematically healthier, wealthier and more health-engaged than those who did not — the “healthy-user” form of confounding by indication. Replication across cohorts did not cure this; it propagated it, because every cohort shared the same confounding structure. The treated and untreated groups were never exchangeable, and no statistical adjustment could recover a comparison the data did not contain. Randomisation could, by construction — and that is precisely the difference the trial revealed.

Case 3 — Albumin: when consistency lied

By 1998, dozens of small trials had compared albumin with crystalloid in critically ill patients. The Cochrane Injuries Group pooled 30 of them (1,419 patients) and found a pooled relative risk of death of 1.68 (95% CI 1.26–2.23) — apparently consistent, statistically significant evidence of harm, from a formal meta-analysis. Six years later the SAFE trial randomised 6,997 patients and found a relative risk of 0.99 (95% CI 0.91–1.09): no effect. Recomputing from SAFE's counts (726/3,473 versus 729/3,460) returns 0.99. The meta-analytic estimate was not merely imprecise; it was centred in the wrong place.

The mechanism is spurious consistency. Many small trials that share the same selection and reporting biases produce estimates that agree with one another — low heterogeneity, a tight pooled confidence interval — while all pointing, in concert, away from the truth. A low I^2 among underpowered studies with correlated biases is agreement, not validity; meta-analysis then propagates the shared bias under a veneer of precision. Consistency is reassuring only when the contributing studies could have disagreed for independent reasons.

Synthesis: a taxonomy of misleading evidence

The three cases form a complementary set rather than a list (Figures 1 and 2). They break at three different joints of the inferential chain: the outcome (CAST measured the wrong thing), the comparison (WHI compared the wrong groups), and the aggregation (albumin combined biased things). Each prior belief carried a credential that, in hindsight, was worthless as a guarantee of truth — a validated surrogate, four decades of replication, and a significant pooled estimate with low heterogeneity. None of these credentials can be wrong in the way a definitive trial can be wrong, and that is the point. The common repair is the same in all three: an RCT large enough to settle the question recovers the truth because randomisation makes the groups exchangeable and because it can be powered on outcomes that matter rather than on markers or on the reconciliation of biased fragments.

This has operational consequences for evidence synthesis. First, weight evidence by its capacity to be falsified: a body of evidence that cannot, by design, distinguish a true effect from a confounded one deserves less confidence regardless of its volume or its internal agreement. Second, treat consistency as diagnostic only when the contributing studies do not share the bias in question — heterogeneity statistics measure agreement, not freedom from common-mode error. Third, keep the surrogate-to-outcome gap explicit, since a surrogate earns the right to substitute for an outcome only once an intervention's effect has been shown to carry across. GRADE already encodes parts of this through its downgrades for indirectness, risk of bias and imprecision; the three reversals warn that the most dangerous case is the one in which those signals are quietly green.

Limitations

This is an interpretive synthesis of three deliberately selected, famous reversals, not a systematic sample; it can illustrate how prior evidence misleads but cannot estimate how often. Survivorship cuts the other way too: definitive trials occasionally overturn beliefs that were in fact correct, or are themselves later qualified — the age-stratified re-analyses of WHI are one example — and a single RCT is not infallible. Our recomputations confirm the published headline effects but inherit each trial's own analytic choices. The taxonomy is a lens, not a law.

Conclusion

CAST, WHI, and albumin are usually remembered as three embarrassments. They are better remembered as one instrument: a way of asking, before we trust a body of evidence, how it could be systematically wrong — through the surrogate we substituted, the comparison we never actually made, or the consistency we mistook for proof. The definitive randomised trial sits atop the evidence hierarchy not by fashion but because it is the design most able to be wrong, and therefore the one whose agreement with reality counts for most.

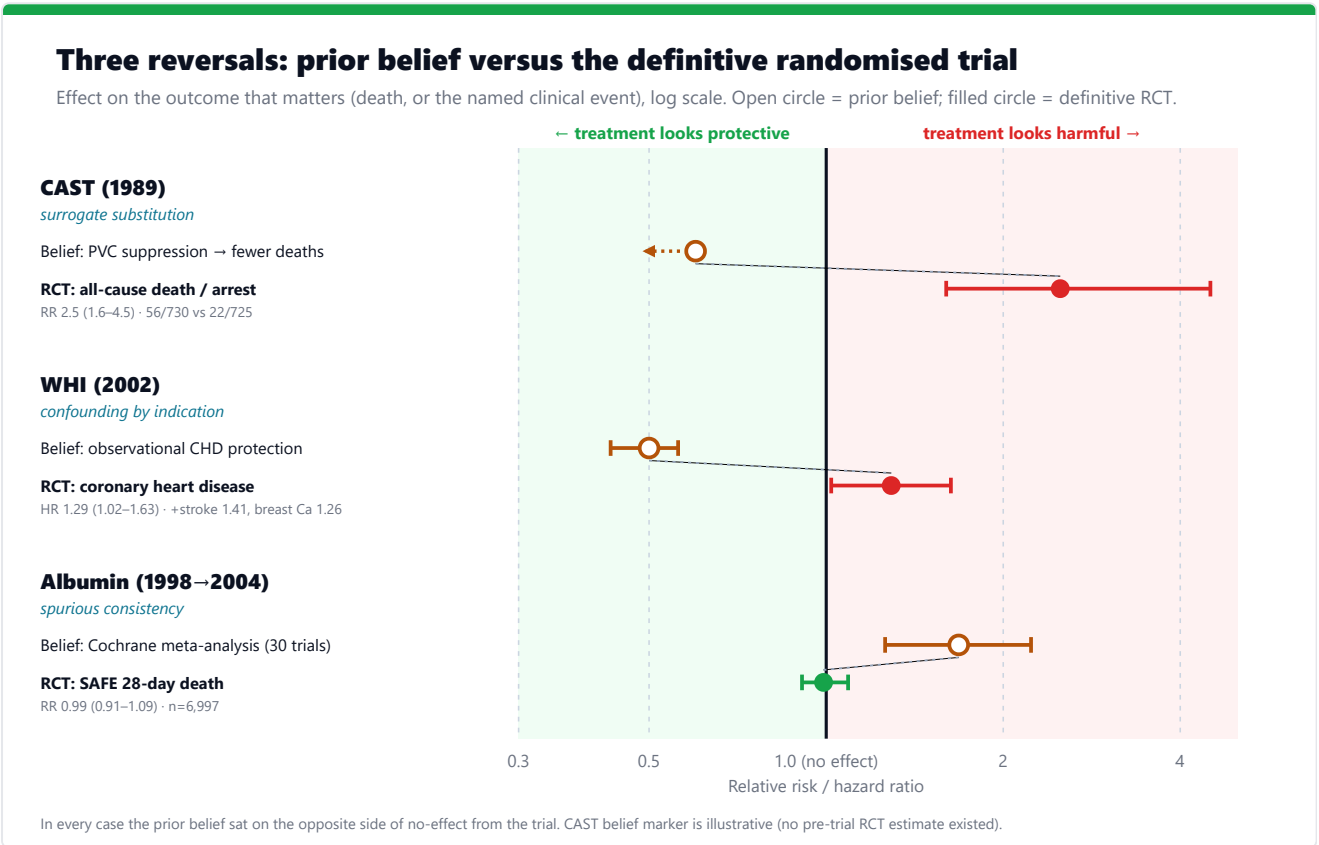


Figure 1. For each reversal, the prior belief (open circle) and the definitive randomised trial (filled circle) sit on opposite sides of no effect. CAST: all-cause death/cardiac arrest, RR 2.5 (1.6–4.5); the belief marker is illustrative because no pre-trial RCT estimate existed. WHI: coronary heart disease, prior pooled observational RR 0.50 (0.43–0.56) versus trial HR 1.29 (1.02–1.63). Albumin: Cochrane meta-analysis RR 1.68 (1.26–2.23) versus SAFE RR 0.99 (0.91–1.09).

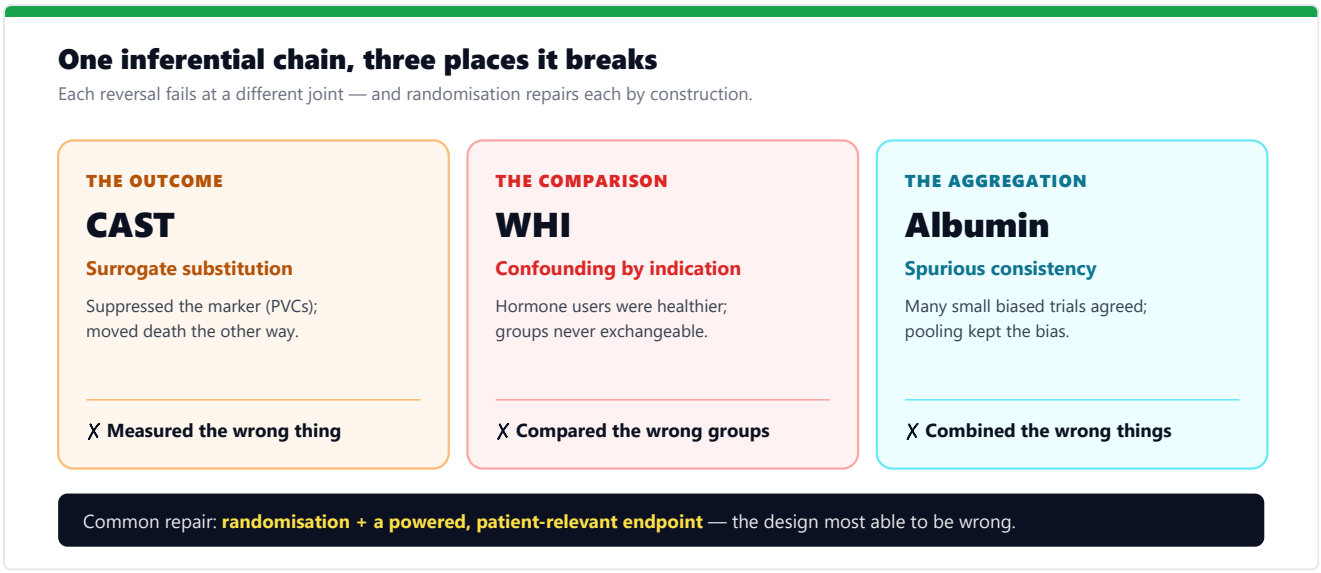


Figure 2. The same inferential chain breaks at three different joints — the outcome (CAST), the comparison (WHI) and the aggregation (albumin) — each repaired by randomisation with a powered, patient-relevant endpoint.

Verification table (literature versus used)

Trial (source)	Headline, as published	Used here	Recompute / status
CAST — preliminary report (NEJM 1989;321:406–412)	All-cause death/cardiac arrest 56/730 (7.7%) drug vs 22/725 (3.0%) placebo; RR 2.5 (95% CI 1.6–4.5)	56/730 vs 22/725; RR 2.5 (1.6–4.5)	**MATCH.** 56/730 ÷ 22/725 = 2.53 ≈ 2.5. **Flag:** the brief's denominators 755/743 are the 1991 *final*-report enrolment; the 56/22 counts and RR 2.5 (1.6–4.5) belong to the 1989 preliminary report — we cite and use the 1989 figures.
WHI E+P (JAMA 2002;288:321–333)	n=16,608; CHD HR 1.29 (1.02–1.63), stroke 1.41 (1.07–1.85), invasive breast cancer 1.26 (1.00–1.59); 5.2 yr	+29% CHD, +41% stroke, +26% breast Ca; n=16,608	**MATCH.** HRs and n confirmed; CIs are nominal (unadjusted) per the report.
Observational HRT prior (Stampfer & Colditz, Prev Med 1991;20:47–63)	Pooled observational CHD RR ≈0.50 (0.43–0.56), high-quality prospective/angiographic studies	Prior-belief RR ≈0.50 (Figure 1)	**MATCH.** Used only to anchor the "cardioprotective" prior reversed by WHI.
Cochrane albumin (BMJ 1998;317:235–240)	30 trials with mortality data (32 met inclusion), 1,419 patients; pooled RR death 1.68 (1.26–2.23)	30 trials, 1,419 patients; RR 1.68 (1.26–2.23)	**RR MATCH. Flag:** the brief said "28-trial"; the published count is **30** trials with mortality data — corrected here.
SAFE (NEJM 2004;350:2247–2256)	n=6,997; 28-day death 726/3,473 (20.9%) albumin vs 729/3,460 (21.1%) saline; RR 0.99 (0.91–1.09)	n=6,997; RR 0.99 (0.91–1.09)	**MATCH.** 726/3,473 ÷ 729/3,460 = 0.99.

References

1. Cardiac Arrhythmia Suppression Trial (CAST) Investigators. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. N Engl J Med 1989;321:406–412.
2. Echt DS, Liebson PR, Mitchell LB, et al. Mortality and morbidity in patients receiving encainide, flecainide, or placebo: the Cardiac Arrhythmia Suppression Trial. N Engl J Med 1991;324:781–788.
3. Writing Group for the Women's Health Initiative Investigators. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial. JAMA 2002;288:321–333.
4. Stampfer MJ, Colditz GA. Estrogen replacement therapy and coronary heart disease: a quantitative assessment of the epidemiologic evidence. Prev Med 1991;20:47–63.

5. Cochrane Injuries Group Albumin Reviewers. Human albumin administration in critically ill patients: systematic review of randomised controlled trials. BMJ 1998;317:235–240.
6. Finfer S, Bellomo R, Boyce N, et al. (SAFE Study Investigators). A comparison of albumin and saline for fluid resuscitation in the intensive care unit. N Engl J Med 2004;350:2247–2256.
7. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336:924–926.

Article metadata

Author. Zaki Khan, Ahmadiyya Hospital, Kenema, Sierra Leone (zakikhan101@gmail.com). No ORCID.
Funding. None.
Competing interests. None declared.
Article type. Analysis (narrative synthesis of three landmark trials).
Data availability. All primary sources are public, peer-reviewed publications cited in the reference list; every headline effect was recomputed from the published counts (see verification table).
Ethics. Not required; analysis of published, aggregate trial results with no human participants and no patient-identifiable data.

Synthēsis · Analysis. Body ≈1173 words. Galley generated 2026-06-23.