VALIDATING CAUSAL BENCHMARKING IN CLINICAL REGISTRIES: ANTIMICROBIAL TIMING IN SEPTIC SHOCK

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative causal benchmarks offer a way to compare treatment-effect estimators when the true effects in observational data are unknown. Their appeal is to combine known, model-defined effects with data that resemble the original study. However, a benchmark can pass verification simply because the tests cannot detect its errors. We investigate this problem through controlled experiments and the Cooperative Antimicrobial Therapy of Septic Shock (CATSS) registry of 11,187 patients from 32 hospitals, where we study antimicrobial timing and mortality. Septic shock is a time-critical condition with substantial mortality, and effective antimicrobial therapy is often initiated under urgent clinical uncertainty. This makes treatment timing a consequential setting in which reliable causal evaluation is particularly important. We find that outcome type sharply changes how added covariates affect verification power. Under severe generator error at $n = 2{,}000$, adding just five covariates reduces the energy test's detection rate from 100% to 0% for binary outcomes. In the corresponding continuous-outcome setting, detection remains 100% through 50 covariates. A similar failure appears in CATSS: energy and maximum mean discrepancy detect a deliberately broken generator using treatment and outcome alone, but miss it once patient covariates are included. A classifier two-sample test detects the failure in both representations. On benchmarks generated by two model families, estimator comparisons further reveal robustness differences obscured by full-model evaluations: the doubly robust estimators remain comparatively stable when either the treatment or outcome model alone is restricted. These comparisons inform our clinical analysis, where effective antimicrobial therapy within one hour of shock onset is associated with a 14.4-percentage-point lower adjusted mortality risk. These findings show that apparent agreement between generated and observed data is not enough. Before using a benchmark to judge causal estimators, we should establish which generator failures its verification procedure can actually detect.

1 INTRODUCTION

Choosing a causal estimator is difficult precisely where a reliable estimate is needed: in an observational study whose treatment effect is unknown. Consider antimicrobial timing in septic shock. Earlier effective therapy has been associated with improved survival (Kumar et al., 2006; Seymour et al., 2017), but a patient's record cannot reveal what would have happened had treatment started at a different time (Holland, 1986). Unlike a predictive model, a treatment-effect estimator therefore cannot be scored directly against observed outcomes. We need a way to compare estimators that reflects the clinical study without assuming that we already know its answer.

Generative causal benchmarks offer one way to make that comparison. RealCause (Neal et al., 2021), for example, retains observed patient characteristics and fits models that generate treatment assignments and outcomes. The fitted mechanisms define a known treatment-effect target against which estimators can be evaluated. But knowing the benchmark's answer does not tell us whether it poses the right problem. Its relevance to the observational study depends on reproducing important features of that study (Curth et al., 2021). RealCause therefore compares generated and observed records using two-sample tests, a step we refer to as *benchmark verification*.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Verification introduces a second evaluation problem. A test may fail to distinguish generated from observed data because the generator fits well, or because the test cannot detect its errors. Distance- and kernel-based tests can lose power as dimensionality increases (Ramdas et al., 2015). For benchmarks that retain observed covariates, this creates a practical tension: omitting those covariates leaves their relationships with treatment and outcome unchecked, while including them adds dimensions shared by real and generated records. Does this loss of sensitivity affect binary and continuous outcomes in the same way? And can a verification procedure miss a known generator error even when only a few covariates are included?

We investigate these questions through controlled experiments and the Cooperative Antimicrobial Therapy of Septic Shock (CATSS) registry of 11,187 patients from 32 hospitals. Known-error generators test sensitivity to specified failures, while real-versus-real comparisons provide a null reference. The controlled experiments reveal a sharp contrast between outcome types. Under severe generator error at $n = 2{,}000$, adding five covariates reduces the energy test's observed detection rate from 100% to 0% for binary outcomes. In the corresponding continuous-outcome setting, detection remains 100% through 50 covariates. The contrast shows why covariate dimensionality alone does not explain when verification fails.

The same failure appears in CATSS, which records binary mortality outcomes alongside 49 patient covariates. Alongside eight fitted generator families, we evaluate a deliberately broken generator that ignores dependence on these covariates. Energy and maximum mean discrepancy (MMD) detect its error using treatment and outcome alone, but fail when covariates are included. A classifier two-sample test (C2ST), which learns to distinguish real from generated records, detects the failure in both representations; its accuracy on the fitted generators evaluated with C2ST remains near chance. The error is detectable in the full representation, but not by every test. This contrast supports assessing sensitivity to known failures in the outcome and representation setting where verification will be used.

We then examine what the generated benchmarks reveal about estimator choice. On the TARNet benchmark (Shalit et al., 2017) at the three-hour treatment window, six causal estimators recover the average-effect target with similar accuracy when fitted with the full covariate set. Restricting the treatment or outcome model to age and sex changes this comparison. Errors increase for estimators relying on the restricted component, whereas augmented inverse probability weighting (AIPW) (Robins et al., 1994) and targeted maximum likelihood estimation (TMLE) (van der Laan & Rubin, 2006) remain close to their full-model performance when only one component is restricted. This advantage disappears when both are restricted. The same qualitative response to model restriction is reproduced with a TabPFN-3 generator fitted to the same registry. These stress tests distinguish robustness properties that the full-covariate comparison leaves largely hidden.

These comparisons guide our use of AIPW, with TMLE as a robustness check, in the CATSS analysis. In the full cohort, effective therapy received by one hour after shock onset is associated with a 14.4-percentage-point lower adjusted mortality risk. This comparison includes patients already receiving therapy at shock onset; excluding them increases the estimated difference. Benchmarking helps choose the estimator, but does not make the clinical effect known: the estimates remain observational and susceptible to unmeasured confounding and immortal-time bias.

Contributions.

- **Outcome-dependent loss of verification power.** We identify controlled settings in which adding a few shared covariates prevents detection of binary-outcome generator errors, while corresponding continuous-outcome errors remain detectable.
- **Classifier-based detection of a known failure.** In CATSS, C2ST detects a deliberately broken generator in the full covariate representation where the evaluated distance- and kernel-based procedures fail.
- **Estimator robustness linked to clinical analysis.** TARNet and TabPFN-3 benchmarks reveal differences under treatment- and outcome-model restriction that inform estimator choice for the CATSS antimicrobial-timing study.

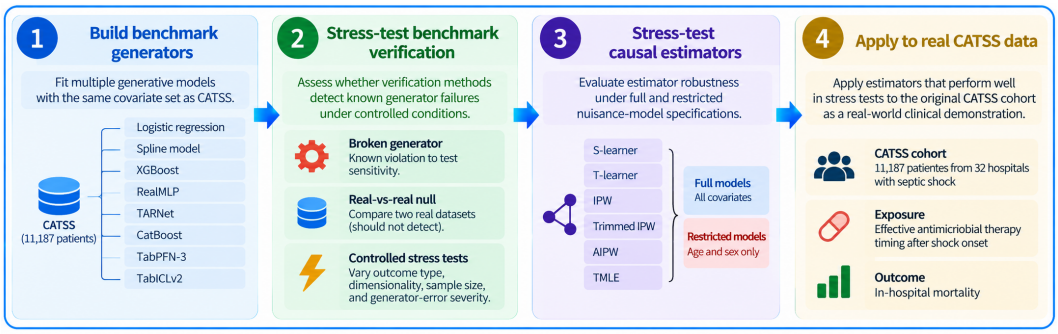

Figure 1: **Verification-first study workflow.** Benchmark generation, verification, estimator evaluation, and clinical application.

2 Background and Related Work

This work draws on causal benchmarking, two-sample testing, causal estimation, and prior studies of antimicrobial timing in septic shock.

Causal benchmarks. Because individual causal effects cannot be directly observed (Holland, 1986), causal methods are commonly evaluated using semi-synthetic benchmarks such as IHDP (Hill, 2011), ACIC (Dorie et al., 2019), Twins (Louizos et al., 2017), and Jobs (LaLonde, 1986). These datasets preserve some characteristics of real data while generating outcomes for which the treatment effect is known. Their usefulness, however, depends in part on how well the generated data reflect the setting being studied (Curth et al., 2021; Curth & van der Schaar, 2021).

RealCause and verification. RealCause (Neal et al., 2021) instead fits generative models to observed data and tests whether real and generated samples can be distinguished before benchmarking causal estimators. We focus on this verification step and ask whether it can reliably detect a generator known to be incorrect, particularly for binary outcomes and higher-dimensional clinical data.

Causal estimators and two-sample tests. The estimators considered here include meta-learners (Künzel et al., 2019), inverse probability weighting (Horvitz & Thompson, 1952), augmented inverse probability weighting (AIPW) (Robins et al., 1994), and targeted maximum likelihood estimation (TMLE) (van der Laan & Rubin, 2006). AIPW and TMLE are doubly robust and remain consistent when either the treatment or outcome model is correctly specified (Bang & Robins, 2005). For verification, we use energy (Székely & Rizzo, 2013), Wasserstein, nearest-neighbour statistics (Friedman & Rafsky, 1979; Schilling, 1986), MMD, and classifier-based two-sample testing. The power of distance-based tests can decrease as dimensionality increases (Ramdas et al., 2015), motivating our evaluation across different outcome types and covariate dimensions.

Antimicrobial timing in septic shock. Earlier effective antimicrobial therapy has repeatedly been associated with improved survival in septic shock (Kumar et al., 2006; Ferrer et al., 2014; Seymour et al., 2017; Liu et al., 2017). However, the most important treatment window remains uncertain. Current sepsis guidelines emphasize rapid treatment (Evans et al., 2021), while more recent evidence suggests that the strength of the association varies across treatment windows and patient populations (Leung et al., 2024). We revisit this question using CATSS while accounting for patients already receiving therapy, immortal-time bias, changing treatment-group composition, and variation across infection sources.

3 Method

Our study follows a verification-first workflow: we fit generators to CATSS, stress-test verification against known failures, compare causal estimators under nuisance-model restrictions, and apply selected estimators to CATSS (Figure 1).

Study data and notation. The Cooperative Antimicrobial Therapy of Septic Shock (CATSS) registry contains 11,187 patients with septic shock from 32 hospitals (Kumar et al., 2006). We denote

the observed patient characteristics by W, treatment by T, and outcome by Y. Specifically, W contains 49 covariates available at or before hypotension onset, T is a binary indicator of whether effective antimicrobial therapy is received within a specified treatment window, and Y indicates in-hospital mortality.

For development and evaluation of the generative models, the cohort is divided into training, validation, and test sets using a 50/10/40 split (5,593/1,119/4,475 patients). The subsequent observational causal analysis is performed on the full cohort of 11,187 patients.

Benchmark generation. Following the RealCause framework (Neal et al., 2021), we retain the observed covariates W and model two components of the data-generating process. The treatment mechanism is

$$e(W) = P(T = 1 \mid W),$$

which represents the probability that a patient receives treatment given their observed characteristics. The treatment-specific outcome mechanisms are

$$\mu_t(W) = P(Y = 1 \mid T = t, W), \qquad t \in \{0, 1\},$$

which represent the predicted probability of in-hospital mortality under the treated and untreated conditions.

For each observed covariate profile W_i, a generated treatment is sampled as

$$T_i^* \sim \text{Bernoulli}\{e(W_i)\},$$

and the generated outcome is then sampled from the corresponding treatment-specific outcome model,

$$Y_i^* \sim \text{Bernoulli}\{\mu_{T_i^*}(W_i)\}.$$

Thus, for each patient, we keep the observed characteristics W_i unchanged and generate new treatment and outcome values, (T_i^*, Y_i^*).

We evaluate eight generator families spanning different modeling approaches: logistic regression, restricted cubic splines, TARNet (Shalit et al., 2017; Shi et al., 2019), XGBoost, CatBoost, RealMLP, TabPFN-3, and TabICLv2. Using multiple generator families allows us to examine whether verification behavior is specific to one modeling architecture or persists across substantially different generators. Generator definitions, development procedures, and held-out fit diagnostics are provided in Appendix A.

Verification framework. Verification is the central methodological focus of the study. A generated benchmark may be difficult to distinguish from observed data because it reproduces the relevant data structure well, but the same result may occur when the verification test lacks sufficient power to detect generator error. We therefore do not interpret a non-significant two-sample test as evidence of benchmark reliability by itself. Instead, we evaluate whether the verification procedure can detect failures whose presence is known in advance.

Observed and generated samples are compared in two representations: (T, Y), which contains only treatment and outcome, and (W, T, Y), which also contains the full covariate representation. Verification is evaluated across all eight fitted generator families.

We use five complementary two-sample procedures.

Energy test. Energy compares within- and between-sample pairwise distances to identify differences in distributional structure.

Wasserstein-1 distance. Wasserstein-1 measures the transport cost required to transform one empirical distribution into the other.

k-nearest-neighbour test. The kNN test evaluates whether observed and generated records mix within local neighborhoods, as expected when the two distributions are similar.

Maximum mean discrepancy. Maximum mean discrepancy (MMD) uses kernel representations to detect distributional differences without explicitly estimating probability densities.

Classifier two-sample test. C2ST trains a classifier to distinguish observed from generated records. Accuracy near chance indicates that the samples are difficult to separate, whereas above-chance accuracy provides evidence of distributional differences. Unlike fixed-distance methods, C2ST can learn which features are most informative for distinguishing the samples.

Because observed profile W_i and its generated counterpart share the same covariates, C2ST uses grouped splits that keep paired observed/generated profiles in the same partition, preventing overlap between training and test records.

Positive and null verification controls. To determine whether the verification procedures are informative, we introduce both positive and null controls.

A deliberately broken generator serves as the positive control. It removes the dependence of generated treatment and outcome on patient covariates W. Because this generator is known by construction to violate the observed conditional structure, a useful verification procedure should distinguish its samples from the observed data.

Real-versus-real comparisons provide the null reference. Two samples drawn from the observed data should not contain a systematic generator-induced difference, so these comparisons characterize verification behavior when no such failure has been introduced.

Together, these controls characterize sensitivity to specified generator errors and behavior under the null, providing context for interpreting non-rejection.

Controlled verification stress tests. We next study verification under controlled conditions in which the presence and severity of generator error are known. These experiments vary four factors: outcome type (binary or continuous), sample size ($n = 250$–$2,000$), test dimensionality (0–50 covariates), and generator-error severity (none, mild, or severe).

The purpose of these experiments is not to compare generator architectures. Instead, they isolate conditions under which a verification procedure gains or loses power. In particular, varying outcome type and dimensionality separately allows us to test whether deterioration in verification is attributable simply to additional dimensions or whether the behavior differs for binary and continuous outcomes.

We refer to these experiments as *verification stress tests*: controlled errors are introduced into the data-generating process and the question is whether the verification procedure detects them.

Estimator benchmarking. After evaluating verification, we use generated benchmarks to compare causal estimators against a generator-defined treatment-effect target. TARNet serves as the primary benchmark, with an independently trained TabPFN-3 model used for replication.

For a fitted generator, the benchmark average treatment effect is

$$\tau_{\text{gen}} = \frac{1}{n} \sum_{i=1}^{n} \left[\mu_1(W_i) - \mu_0(W_i) \right].$$

Because both treatment-specific outcome mechanisms are known, this is a *generator-defined benchmark target*, not the unknown causal effect in the original clinical population.

We compare six causal estimators. The S-learner fits one outcome model that includes treatment as an input, whereas the T-learner fits separate outcome models for treated and untreated patients (Künzel et al., 2019). Inverse probability weighting (IPW) uses the estimated treatment probabilities to reweight patients and create a pseudo-population with improved treatment-group comparability (Horvitz & Thompson, 1952); we evaluate IPW both with and without trimming of extreme weights. Augmented inverse probability weighting (AIPW) combines treatment and outcome models (Robins et al., 1994). Targeted maximum likelihood estimation (TMLE) similarly combines both compo-

nents and updates the initial outcome model to target the treatment-effect estimand (van der Laan & Rubin, 2006).

The treatment and outcome models used by these estimators are referred to as *nuisance models*: they are intermediate models required for treatment effect estimation rather than the final estimand itself.

Estimator performance is measured using root mean squared error (RMSE) relative to τ_{gen}. We first evaluate all estimators using the full nuisance models. We then deliberately restrict the treatment model to age and sex, restrict the outcome model to age and sex, or restrict both.

These experiments are distinct from the verification stress tests above. Verification stress tests introduce errors into the *generator* to test whether a verification procedure can detect them. Estimator stress tests instead restrict the *nuisance models used by the causal estimator* to evaluate robustness when treatment or outcome modelling is incomplete.

Controlled treatment-effect heterogeneity. Conditional-effect recovery is evaluated separately using the TARNet benchmark. Treatment-effect heterogeneity is varied in a controlled manner, providing a known setting in which to examine whether the downstream subgroup analysis can detect increasing effect variation.

This experiment serves as a self-audit of the heterogeneity analysis rather than as an additional comparison of generator families. We evaluate both null calibration, where no additional heterogeneity is introduced, and sensitivity as the imposed treatment-effect spread increases. Additional details are provided in Appendix C.

Clinical application. We apply AIPW, with TMLE as a robustness check, to in-hospital mortality in the original CATSS cohort using all 49 baseline covariates. At landmarks $L \in \{1, 3, 6, 12\}$ hours, patients are classified as treated if effective therapy was active at t_0 or initiated by L, and untreated otherwise.

Because treatment classification uses post-t_0 information, these analyses remain susceptible to immortal-time bias (Suissa, 2008; Hernán et al., 2016; Cain et al., 2010). Hour-level death timing was unavailable, so we assess sensitivity rather than claiming elimination of this bias. We also separately examine the 2,621 patients already receiving therapy at shock onset.

The 1- and 3-hour analyses are primary; 6 and 12 hours are secondary because group composition changes at later landmarks. A separate spline analysis describes the association between post-shock treatment delay and mortality. Additional details are in Appendix D.

4 Verification Results

Verification power depends on outcome type and representation. The controlled experiments showed that dimensionality alone does not explain verification failure. Under severe generator error at $n = 2{,}000$, the energy test detected the error in 100% of binary-outcome replicates when no covariates were included. Detection fell to approximately 50% with two covariates and to 0% with five or more. In contrast, for continuous outcomes, energy retained 100% detection through 50 covariates. This outcome-dependent loss of verification power is summarized in Figure 2A.

Milder errors were substantially harder to detect for binary outcomes. With covariates excluded, detection rates were 9% for energy, 19% for kNN, and 0% for Wasserstein-1, whereas the corresponding continuous-outcome experiments achieved approximately 97–100% detection. Across 96 no-error conditions, false-positive rates were 0% for energy, 0% for Wasserstein-1, and 0.5% for kNN.

A known generator failure can pass several verification procedures. In CATSS, energy and MMD detected the deliberately broken generator when verification used treatment and outcome alone, (T, Y), but failed when the shared patient covariates were included, (W, T, Y). Wasserstein-1 did not detect the broken generator in either representation. The kNN result was unstable in (T, Y) and did not detect the broken generator in (W, T, Y).

C2ST behaved differently. With grouped patient-level splitting, it detected the broken generator in both (T, Y) and (W, T, Y), while classification accuracy for the fitted generators evaluated with C2ST remained near chance. Thus, the failure of the distance- and kernel-based procedures does

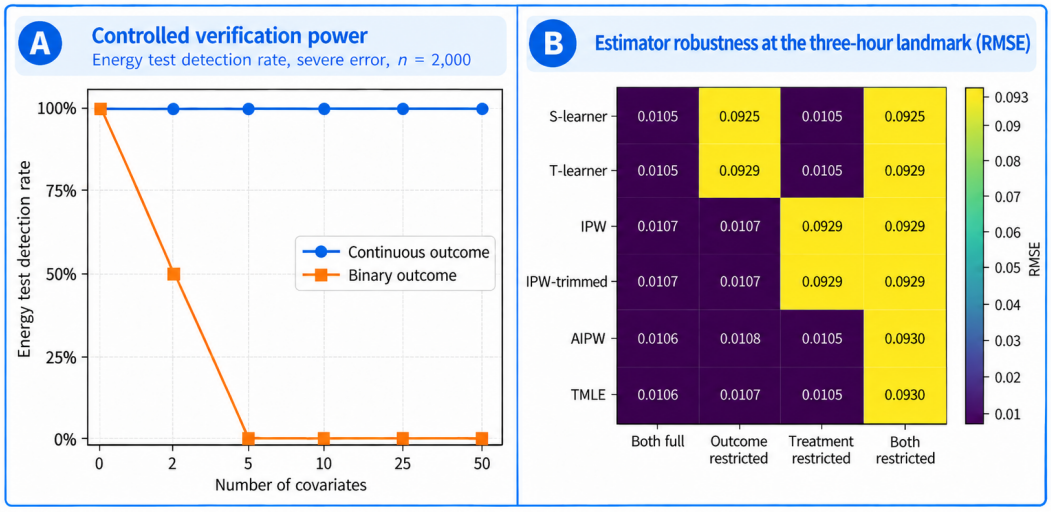

Figure 2: **Verification power and estimator robustness.** **(A)** Energy-test detection under severe generator error across covariate dimensions. **(B)** Estimator RMSE under full and restricted nuisance models.

not imply that verification with covariates is inherently impossible; sensitivity depends on the verification procedure and representation.

The verification limitation is not specific to one generator family. The eight fitted generator families showed similar treatment-model discrimination, with held-out AUCs ranging from 0.636 to 0.652. The main verification limitation persisted across statistical, boosted-tree, neural-network, and tabular foundation-model generators.

Together, these results show that a non-significant verification result should be interpreted as failure to detect a discrepancy rather than as evidence that the generated benchmark is reliable. Additional verification analyses, including covariate compression, grouped C2ST implementation, and kNN stability, are reported in Appendix B.

5 Estimator Benchmark

Estimators perform similarly with full nuisance models. When both nuisance models used the full covariate representation, all six estimators recovered the generator-defined treatment-effect target with similar accuracy. At the three-hour landmark, where the benchmark target was approximately -0.210, RMSE ranged from 0.0105 to 0.0107.

Nuisance-model restriction reveals robustness differences. The comparison changed substantially when one or both nuisance models were restricted to age and sex. The S-learner and T-learner showed much larger error when the outcome model was restricted, whereas IPW showed much larger error when the treatment model was restricted. AIPW and TMLE remained close to their full-model performance when either nuisance model alone was restricted, but this advantage disappeared when both nuisance models were restricted.

The corresponding verification-power and estimator-robustness patterns are shown in Figure 2; detailed RMSE values are reported in Appendix C.

The pattern replicates with TabPFN-3. The estimator stress test was independently repeated using a TabPFN-3 generator. The same qualitative pattern was reproduced: outcome-model estimators were primarily sensitive to outcome-model restriction, weighting estimators to treatment-model restriction, and AIPW and TMLE remained stable under a single nuisance-model restriction.

Additional conditional-effect and heterogeneity analyses are reported in Appendix C.

Table 1: **Clinical estimates across treatment windows.** Adjusted mortality risk differences; negative values favor earlier treatment.

Window	AIPW (95% CI)	TMLE
1 h	-0.144 (-0.161, -0.127)	-0.144
3 h	-0.216 (-0.232, -0.199)	-0.216
6 h	-0.292 (-0.310, -0.275)	-0.292
12 h	-0.360 (-0.380, -0.339)	-0.360

6 Application: Antimicrobial Timing in Septic Shock

The clinical analysis serves as a downstream demonstration of the verification-aware workflow. We carry AIPW forward as the primary estimator, with TMLE as a robustness check, in the original CATSS cohort. Table 1 summarizes the main estimates. The 1- and 3-hour windows are primary analyses; the 6- and 12-hour windows are secondary because treatment-group composition changes increasingly at later landmarks. Additional positivity diagnostics, sensitivity analyses, and subgroup results are provided in Appendix D.

Earlier effective therapy was associated with lower adjusted mortality. In the primary analyses, effective therapy within one hour was associated with an adjusted mortality-risk difference of -0.144 (95% CI, -0.161 to -0.127), and therapy within three hours with a difference of -0.216 (95% CI, -0.232 to -0.199). TMLE produced nearly identical point estimates.

Interpretation of these estimates requires caution. A total of 2,621 patients (23.4%) were already receiving effective therapy when shock began, representing a clinically different exposure state from rapid post-shock initiation. The fixed-window analyses also remain susceptible to immortal-time bias because treatment classification uses information observed after shock onset. In addition, the composition of the treated and untreated groups changes as the treatment window increases; the 6- and 12-hour analyses are therefore treated as secondary.

Additional analyses of pretreatment, potential immortal-time bias, changing group composition, continuous post-shock treatment delay, positivity, and infection-source heterogeneity are reported in Appendix D. The clinical estimates remain observational adjusted associations rather than experimentally established treatment effects.

7 Discussion

What the verification step can support. The main finding is that passing a causal-benchmark verification test does not establish benchmark reliability: generated data may appear consistent with observed data either because relevant structure is reproduced or because the test lacks power to detect error.

Our controlled experiments show that this problem is not explained by dimensionality alone; verification depends on outcome type, representation, and the test itself. Under severe error, energy retained high sensitivity for continuous outcomes but declined rapidly for binary outcomes, while mild binary-outcome errors were difficult to detect even without covariates.

In CATSS, energy and MMD detected the broken generator using (T, Y) but failed with the full covariate representation. In contrast, C2ST detected it in both (T, Y) and (W, T, Y) while remaining near chance for fitted generators. Thus, adding covariates is not inherently uninformative; performance depends on how distributional differences are measured.

These findings support a verification-first workflow: a broken generator tests sensitivity to known failure, while real-versus-real comparisons provide a null reference. A large verification p-value therefore indicates failure to detect a discrepancy, not proof of benchmark realism.

Relation to prior work. Distance-based two-sample tests are known to lose power as dimensionality increases (Ramdas et al., 2015; Neal et al., 2021). Our results refine this observation by showing that outcome type and representation also affect verification power: binary-outcome verification deteriorated rapidly for several procedures, whereas continuous-outcome verification remained more

sensitive. C2ST further shows that the problem is not simply the inclusion of covariates. More broadly, our findings complement concerns about causal benchmark realism (Curth et al., 2021) by showing that the verification procedure itself should first demonstrate sensitivity to known generator failure.

Implications for downstream causal evaluation. The estimator benchmark illustrates why verification should precede estimator comparison. With full nuisance models, all six estimators performed similarly. Under deliberate restriction, estimators depending on the restricted component showed much larger error, whereas AIPW and TMLE remained comparatively stable when either nuisance model alone was restricted. The same qualitative pattern was reproduced with TabPFN-3.

This comparison is a downstream use of the verification-aware benchmark, not the primary contribution. Conditional-effect analyses also showed imperfect subgroup recovery and calibration; details are provided in Appendix C.

Clinical application. The CATSS analysis serves as a downstream demonstration. Earlier effective antimicrobial therapy was associated with lower adjusted mortality, with similar AIPW and TMLE estimates. The analysis also highlights limitations that benchmark verification cannot remove. Pretreatment at shock onset affected the early contrast, fixed-window definitions remained susceptible to immortal-time bias, and untreated-group composition changed as the landmark widened. The 6- and 12-hour estimates therefore should not be interpreted as a simple dose-response relationship. The spline analysis showed a nonlinear association between recorded treatment delay and mortality, with the steepest increase during the first six hours. This analysis is descriptive and does not represent the same estimand as the fixed-window comparisons.

7.1 Limitations

The empirical verification analysis uses a single clinical registry, so additional datasets are needed to assess generalizability. The deliberately broken generator also represents a substantial violation; milder errors, particularly for binary outcomes, may remain difficult to detect.

Verification is additionally representation-dependent. Agreement in (T, Y) does not establish that relationships involving W or subgroup structure have been reproduced, and no single two-sample test can establish realism for every downstream estimand.

Conditional-effect evaluation showed imperfect calibration: under the null, the heterogeneity procedure rejected in 8.5% of 200 replicates (95% CI, 5.2–13.0%) at a nominal 5% level. In the TARNet-based benchmark, even the strongest planted heterogeneity setting, six times the fitted level, produced a maximum source-level treatment-effect spread of 0.161, compared with approximately 0.24 observed in CATSS.

Finally, the clinical analysis remains observational and may be affected by unmeasured confounding and immortal-time bias. CATSS was collected between 1989 and 2004, which may also limit transportability to current practice.

7.2 Conclusion

Causal benchmarks are useful when counterfactual outcomes are unavailable, but their value depends on reliable verification. Our results show that outcome type, representation, and the verification procedure itself all affect sensitivity to generator error.

A benchmark should therefore not be considered reliable simply because a two-sample test fails to reject it. Verification should first demonstrate sensitivity to known failure and appropriate null behavior, with downstream structures such as subgroup effects evaluated separately when relevant.

The estimator benchmark and CATSS application illustrate this verification-first approach, but the central message is: *before using a generated benchmark to judge causal estimators, we should first establish that the verification procedure can detect when the benchmark is wrong.*

AI USE STATEMENT

We used generative AI tools to assist with brainstorming, analysis of related work, refinement of the manuscript's writing for clarity and readability, and generation of figures used in the paper. The tools were used to help organize ideas, summarize and compare relevant literature, improve the wording of text originally prepared by the authors, and support the creation or refinement of visual figures.

All AI-assisted content was reviewed and revised by the authors. References and claims arising from the related-work analysis were checked against the original sources, and the authors independently evaluated the relevance and accuracy of the suggested literature and interpretations. Figures produced with AI assistance were reviewed and edited by the authors to ensure that they accurately represent the intended content. Generative AI was not used to fabricate data, generate experimental results, or make final scientific conclusions. The authors take full responsibility for the final content of this work, including all text, claims, figures, analyses, and citations produced with the assistance of generative AI.

ETHICS STATEMENT

This study is a secondary analysis of the de-identified Cooperative Antimicrobial Therapy of Septic Shock (CATSS) multicentre registry, which was originally collected with ethics approval at participating institutions (Kumar et al., 2006). No new patient data were collected and no patient contact occurred. All analyses were performed using de-identified registry data.

REPRODUCIBILITY STATEMENT

Analyses were repeated on two machines using fixed random seeds, with consistent results except for the k-nearest-neighbour test discussed in Appendix B. Section 3 describes the main methods, while the appendices provide details of cohort construction, missing-data handling, and additional results. Code and supporting materials will be provided as supplementary material.

REFERENCES

Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Lauren E. Cain, James M. Robins, Emilie Lanoy, Roger Logan, Dominique Costagliola, and Miguel A. Hernán. When to start treatment? a systematic approach to the comparison of dynamic regimes using observational data. *The International Journal of Biostatistics*, 6(2):Article 18, 2010. doi: 10.2202/1557-4679.1212.

Alicia Curth and Mihaela van der Schaar. On inductive biases for heterogeneous treatment effect estimation. *Advances in Neural Information Processing Systems*, 34:15883–15894, 2021.

Alicia Curth, David Svensson, James Weatherall, and Mihaela van der Schaar. Really doing great at estimating cate? a critical look at ml benchmarking practices in treatment effect estimation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks)*, 2021. URL `https://openreview.net/forum?id=FQLzQqGEAH`.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019. doi: 10.1214/18-STS667.

Laura Evans, Andrew Rhodes, Waleed Alhazzani, et al. Surviving sepsis campaign: International guidelines for management of sepsis and septic shock 2021. *Critical Care Medicine*, 49(11): e1063–e1143, 2021. doi: 10.1097/CCM.0000000000005337.

Ricard Ferrer, Ignacio Martin-Loeches, Gary Phillips, Tiffany M. Osborn, Sean Townsend, R. Phillip Dellinger, Antonio Artigas, Christa Schorr, and Mitchell M. Levy. Empiric antibiotic treatment reduces mortality in severe sepsis and septic shock from the first hour: Results from a guideline-based performance improvement program. *Critical Care Medicine*, 42(8):1749–1755, 2014. doi: 10.1097/CCM.0000000000000330.

Jerome H. Friedman and Lawrence C. Rafsky. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics*, 7(4):697–717, 1979.

Miguel A. Hernán, Brian C. Sauer, Sonia Hernández-Díaz, Robert Platt, and Ian Shrier. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *Journal of Clinical Epidemiology*, 79:70–75, 2016. doi: 10.1016/j.jclinepi.2016.04.014.

Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011. doi: 10.1198/jcgs.2010.08162.

Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986. doi: 10.2307/2289064. URL https://doi.org/10.2307/2289064.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

Anand Kumar, Daniel Roberts, Kenneth E. Wood, Barbara Light, Joseph E. Parrillo, Satendra Sharma, Robert Suppes, Daniel Feinstein, Steven Zanotti, Laura Taiberg, David Gurka, Ajay Kumar, and Mary Cheang. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Critical Care Medicine*, 34 (6):1589–1596, June 2006. doi: 10.1097/01.CCM.0000217961.75225.E9.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019. doi: 10.1073/pnas.1804597116.

Robert J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4):604–620, 1986.

Ling Yan Leung, Hsi-Lan Huang, Kevin K. C. Hung, Chi Yan Leung, Cherry C. Y. Lam, Ronson S. L. Lo, Chun Yu Yeung, Peter Joseph Tsoi, Michael Lai, Mikkel Brabrand, Joseph H. Walline, and Colin A. Graham. Door-to-antibiotic time and mortality in patients with sepsis: Systematic review and meta-analysis. *European Journal of Internal Medicine*, 129:48–61, 2024. doi: 10.1016/j.ejim.2024.06.015.

Vincent X. Liu, Vikram Fielding-Singh, John D. Greene, Jennifer M. Baker, Theodore J. Iwashyna, Jay Bhattacharya, and Gabriel J. Escobar. The timing of early antibiotics and hospital mortality in sepsis. *American Journal of Respiratory and Critical Care Medicine*, 196(7):856–863, 2017. doi: 10.1164/rccm.201609-1848OC.

Christos Louizos, Uri Shalit, Joris M. Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems 30*, pp. 6446–6456, 2017.

Brady Neal, Chin-Wei Huang, and Sunand Raghupathi. Realcause: Realistic causal inference benchmarking, 2021. URL https://arxiv.org/abs/2011.15007.

Aaditya Ramdas, Sashank J. Reddi, Barnabás Póczos, Aarti Singh, and Larry Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 3571–3577, 2015.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

Mark F. Schilling. Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, 81(395):799–806, 1986.

Christopher W. Seymour, Foster Gesten, Hallie C. Prescott, Marcus E. Friedrich, Theodore J. Iwashyna, Gary S. Phillips, Stanley Lemeshow, Tiffany Osborn, Kathleen M. Terry, and Mitchell M. Levy. Time to treatment and mortality during mandated emergency care for sepsis. *New England Journal of Medicine*, 376(23):2235–2244, 2017. doi: 10.1056/NEJMoa1703058.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *International Conference on Machine Learning*, pp. 3076–3085, 2017.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*, pp. 2507–2517, 2019.

Samy Suissa. Immortal time bias in pharmaco-epidemiology. *American Journal of Epidemiology*, 167(4):492–499, 2008. doi: 10.1093/aje/kwm324.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

Mark J. van der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006. doi: 10.2202/1557-4679.1043.

A Implementation Details

This appendix provides additional implementation details for cohort construction, preprocessing, benchmark generation, verification, controlled stress testing, and causal-estimator evaluation.

A.1 CATSS Cohort Construction

The Cooperative Antimicrobial Therapy of Septic Shock (CATSS) registry contained 11,188 patient records. One patient without a recorded in-hospital mortality outcome was excluded, resulting in a final cohort of

$$n = 11{,}187.$$

The clinical time origin, denoted by t_0, is the onset of recurrent or persistent hypotension associated with septic shock.

We denote the observed baseline patient characteristics by W, treatment by T, and outcome by Y. The final covariate vector W contains 49 variables available at or before t_0. Treatment T is a binary indicator of whether effective antimicrobial therapy was received within the treatment window under consideration, and Y is a binary indicator of in-hospital mortality.

The same baseline covariate set was used for benchmark construction and for adjustment in the clinical analyses.

A.2 Covariate Selection

Only variables available at or before hypotension onset were eligible for inclusion in W. Variables measured after t_0 were excluded so that the adjustment set did not contain information arising after treatment initiation.

The covariate set included demographic characteristics such as age and sex, severity of illness measured by APACHE II, 28 comorbidity indicators, baseline laboratory measurements, weight, mode of infection acquisition (community versus nosocomial), bacteraemia, and whether the infection required source control.

Because CATSS spans multiple calendar years during which sepsis management changed, a tertile-based calendar-era variable was also included.

After variable selection and preprocessing, the final covariate vector contained 49 variables.

A.3 Missing-Data Handling

Missing-data handling depended on the amount and pattern of missingness rather than applying the same procedure to every variable.

Highly missing laboratory variables. Lactate, platelet count, and international normalized ratio (INR) were missing in approximately 52–78% of patients. Missingness in these variables was associated with both treatment timing and outcome, suggesting that whether a test was ordered itself contained clinical information. We therefore retained a binary indicator of whether each laboratory test had been performed but did not use the measured value itself.

Excluded variables. Sodium, chloride, height, and body mass index were not retained in the final adjustment set. Their missingness was strongly related to calendar era and they were not retained as key confounders for the antimicrobial-timing analysis.

Iteratively imputed variables. Five retained variables contained moderate levels of missingness:

- APACHE II score,
- creatinine,
- bilirubin,
- weight, and
- bicarbonate.

Missingness among these variables ranged from approximately 5.9% to 24.0%. They were imputed using multivariate iterative imputation.

The imputation procedure was fitted using the training partition and subsequently applied to the validation and test partitions. A binary missingness indicator was retained for each imputed variable.

Approximately 75 physiologically implausible recorded values were first treated as missing before imputation.

A.4 Infection-Source Definition

Infection source was derived from the recorded site of infection.

For the more detailed subgroup analysis, infection sources were grouped into six categories:

1. pneumonia,
2. gastrointestinal,
3. urinary tract,
4. skin and soft tissue,
5. bacteraemia, and
6. other.

The "other" category combined less frequent infection sites that were too small to support stable individual estimates.

A second, coarser grouping consisted of respiratory, intra-abdominal, genitourinary, and other infections. Both representations were examined so that conclusions about infection-source heterogeneity were not dependent on a single categorization.

A.5 Training, Validation, and Test Splits

For development and evaluation of the generative models, the cohort was divided using a 50/10/40 training, validation, and test split:

$$n_{\text{train}} = 5{,}593, \qquad n_{\text{val}} = 1{,}119, \qquad n_{\text{test}} = 4{,}475.$$

Preprocessing procedures requiring parameter estimation, including multivariate imputation, were fitted using the training partition and then applied to the validation and test partitions.

These partitions were used for generative-model development and evaluation. The subsequent observational clinical analysis used the complete CATSS cohort of 11,187 patients.

A.6 Generator Implementations

Each generator models the treatment mechanism

$$e(W) = P(T = 1 \mid W)$$

and the treatment-specific outcome mechanisms

$$\mu_t(W) = P(Y = 1 \mid T = t, W), \qquad t \in \{0, 1\}.$$

Thus, $\mu_1(W)$ represents the predicted probability of in-hospital mortality under treatment, while $\mu_0(W)$ represents the corresponding probability under no treatment.

We evaluated eight generator families chosen to span substantially different modelling approaches.

Logistic regression. Logistic regression provides a low-complexity parametric baseline. Separate models were used for the treatment mechanism and treatment-specific outcome mechanisms.

Restricted cubic splines. The spline generator extends the parametric baseline by allowing nonlinear covariate effects while retaining an additive statistical model.

TARNet. TARNet uses a shared neural representation of the covariates together with treatment-specific outcome heads (Shalit et al., 2017; Shi et al., 2019). TARNet was also used as the primary generator for the causal-estimator benchmark.

XGBoost. XGBoost provides a gradient-boosted tree generator in which nonlinearities and interactions are represented through recursive tree splits.

CatBoost. CatBoost provides a second boosted-tree architecture with a different optimization and feature-handling strategy.

RealMLP. RealMLP provides a general-purpose neural-network model for tabular data. The implementation used its tuned-default configuration.

TabPFN-3. TabPFN-3 is a pretrained tabular foundation model that performs prediction through in-context inference rather than conventional task-specific neural network training. It was also used as an independent replication generator for estimator benchmarking.

TabICLv2. TabICLv2 provides a second tabular foundation-model family and extends the generator comparison beyond conventional statistical, tree-based, and neural models.

All eight generator families were included in the verification analyses. Estimator benchmarking was more narrowly scoped: TARNet served as the primary benchmark generator and TabPFN-3 as an independent replication.

A.7 Generator Fit Evaluation

Generator fit was evaluated on the held-out test partition. Treatment-model AUCs ranged from 0.636 to 0.652. Outcome-model AUCs ranged from 0.758 to 0.780 for untreated patients and from 0.786 to 0.798 for treated patients.

Predictive fit was used to characterize the generators but was not treated as evidence that a generator was adequately verified. Verification was evaluated separately using the procedures described below.

A.8 Benchmark Generation

For each patient i, the observed covariate profile W_i was retained.

The fitted treatment mechanism provides

Table 2: **Held-out fit of the eight generator families.** Treatment AUC evaluates $P(T = 1 \mid W)$; outcome AUCs evaluate the two treatment-specific mortality mechanisms.

Generator	Treatment AUC	Outcome $T = 0$	Outcome $T = 1$	Implied effect	Fit time
TabICLv2	0.652	0.779	0.798	-0.225	15 s
TabPFN-3	0.650	0.780	0.797	-0.226	59 s
CatBoost	0.646	0.766	0.792	-0.222	4 s
TARNet	0.643	0.776	0.791	-0.209	$\sim$60 s
RealMLP	0.642	0.758	0.789	-0.222	90 s
Logistic	0.640	0.776	0.786	-0.219	0.03 s
Spline	0.640	0.777	0.790	-0.223	$\sim$1 s
XGBoost	0.636	0.760	0.787	-0.225	1.5 s

$$e(W_i) = P(T = 1 \mid W_i),$$

and a generated treatment assignment is sampled according to

$$T_i^* \sim \text{Bernoulli}\left\{e(W_i)\right\}.$$

The fitted outcome mechanisms provide

$$\mu_1(W_i) = P(Y = 1 \mid T = 1, W_i)$$

and

$$\mu_0(W_i) = P(Y = 1 \mid T = 0, W_i).$$

After T_i^* is sampled, the generated outcome is drawn from the corresponding treatment-specific mechanism:

$$Y_i^* \sim \text{Bernoulli}\left\{\mu_{T_i^*}(W_i)\right\}.$$

Each generated record therefore has the form

$$(W_i, T_i^*, Y_i^*),$$

where the observed patient characteristics are retained while treatment and outcome are generated.

The generator-defined average treatment-effect target is

$$\tau_{\text{gen}} = \frac{1}{n}\sum_{i=1}^{n}\left[\mu_1(W_i) - \mu_0(W_i)\right].$$

This is a model-defined benchmark target and is not interpreted as the unknown true causal effect in the original CATSS population.

A.9 Verification Representations

Observed and generated samples were compared using two representations.

The first uses treatment and outcome only,

$$(T, Y),$$

while the second includes the complete covariate representation,

$$(W, T, Y).$$

Because each generated observation retains the original covariate profile W_i, observed and generated records share the same W by construction. The controlled experiments therefore specifically examined how inclusion of these shared covariates affects verification power.

A.10 Verification Tests

Five complementary two-sample procedures were evaluated.

Energy test. The energy two-sample test compares pairwise distances within and between the observed and generated samples.

Wasserstein-1 distance. Wasserstein-1 measures the transport cost required to move one empirical distribution onto another.

k-nearest-neighbour test. The kNN test examines whether local neighbourhoods contain the expected mixture of observed and generated records. Exact ties were additionally examined because the binary (T, Y) representation contains only four possible states.

Maximum mean discrepancy. Maximum mean discrepancy (MMD) compares distributions through kernel mean embeddings. We used a Gaussian kernel, the median heuristic applied to squared Euclidean distances for bandwidth selection, and an unbiased estimate of MMD^2. Statistical significance was evaluated by permutation.

Classifier two-sample test. Classifier two-sample testing (C2ST) labels observations as real or generated and trains a classifier to distinguish the two groups. Classification accuracy near 0.5 indicates that the samples are difficult to distinguish.

Unlike fixed-distance procedures, C2ST can learn which variables are informative for separating the two samples.

Because each real record and its generated counterpart share W_i, the final C2ST implementation used grouped patient-level splits. Paired real/generated records associated with the same patient were always placed in the same partition. Permutation testing preserved this paired structure.

A.11 Verification Controls

Two controls were used to determine whether a verification procedure was informative.

Broken-generator positive control. A deliberately broken generator drew treatment and outcome from marginal rates while ignoring patient covariates W. Because this generator is known by construction to remove important dependence on patient characteristics, an informative verification procedure should distinguish it from observed data.

Real-versus-real null reference. Two samples from the observed data were compared to provide a reference for test behavior when no deliberate generator discrepancy was introduced.

These controls distinguish a generator that is difficult to distinguish from a test that is simply unable to detect a known failure.

A.12 Controlled Verification Stress Tests

Controlled experiments varied four factors:

- **Outcome type:** binary or continuous;

- **Sample size:** $n = 250$–$2{,}000$;
- **Covariate dimensionality:** 0–50 covariates;
- **Generator-error severity:** none, mild, or severe.

Across all conditions, 3,456 verification tests were performed.

These experiments are referred to as *verification stress tests*. They perturb the data-generating process and ask whether the verification procedure detects the resulting discrepancy.

A.13 Causal Estimator Implementation

Six estimators were evaluated.

S-learner. The S-learner fits one outcome model including treatment and covariates and evaluates that model under both treatment states.

T-learner. The T-learner fits separate outcome models for treated and untreated patients.

Inverse probability weighting. IPW uses the estimated propensity score $e(W)$ to reweight observations according to the inverse probability of their observed treatment.

Trimmed IPW. The trimmed IPW analysis excluded observations with estimated propensity scores outside

$$[0.01, 0.99]$$

to limit the influence of extreme weights.

Augmented inverse probability weighting. AIPW combines treatment and outcome models through a doubly robust estimating equation.

Targeted maximum likelihood estimation. TMLE updates an initial outcome model using treatment-model information so that the resulting estimate is targeted toward the treatment-effect estimand.

The treatment and outcome models used by these procedures are referred to as *nuisance models*: they are intermediate models required to estimate the effect rather than the final estimand itself.

A.14 Estimator Stress-Test Design

Estimator accuracy was evaluated relative to τ_{gen} using root mean squared error (RMSE).

Four nuisance-model settings were considered:

1. full treatment model and full outcome model;
2. full treatment model and outcome model restricted to age and sex;
3. treatment model restricted to age and sex and full outcome model;
4. both models restricted to age and sex.

These are referred to as *estimator stress tests*. They are distinct from the verification stress tests: verification stress tests perturb the generator, whereas estimator stress tests keep the benchmark target fixed and restrict the models used by the estimator.

A.15 Controlled Treatment-Effect Heterogeneity

Conditional-effect recovery was examined separately using the TARNet benchmark. Treatment-effect heterogeneity was increased in a controlled manner to evaluate whether the subgroup procedure detected increasingly strong effect variation.

A null condition with no added heterogeneity was used to assess calibration. The null experiment consisted of 200 replicates.

A.16 Relationship Between Verification and Estimator Evaluation

Verification and estimator benchmarking answer different questions.

Verification asks whether a generated sample can be distinguished from observed data and whether the verification procedure itself can detect a generator known to be incorrect.

Estimator benchmarking instead assumes a generator-defined target and asks how accurately different causal estimators recover that target.

A large verification p-value is therefore interpreted only as failure to detect a difference, not as proof that the generator is clinically correct or that its model-defined treatment effect equals the true CATSS treatment effect.

B Additional Verification Results

B.1 Broken-Generator Positive Control

The deliberately broken generator provides a direct test of whether each verification procedure can detect a known failure.

Table 3: **Verification of the deliberately broken generator.** Entries are p-values from the final verification analyses, except where the kNN result was unstable across computing environments.

Test	(T, Y)	(W, T, Y)
Energy	0.004	1.000
Wasserstein-1	0.571	1.000
kNN	unstable	1.000
MMD	0.005	1.000
C2ST	0.005	0.005

Energy and MMD detected the broken generator when verification was restricted to (T, Y) but failed after the shared covariates W were included. Wasserstein-1 did not detect the broken generator in either representation. The kNN result was unstable in (T, Y) and did not detect the broken generator in (W, T, Y).

C2ST behaved differently. It detected the deliberately broken generator both on (T, Y) and on (W, T, Y). Classification accuracy was 0.585 using (T, Y) and 0.608 using (W, T, Y), whereas accuracies for the fitted generators remained close to chance, between 0.497 and 0.513.

B.2 Importance of Grouped C2ST Splits

An initial row-level C2ST split produced misleading below-chance classification accuracies because real and generated observations from the same patient could appear in different partitions while sharing identical covariates.

The final C2ST therefore split data at the patient level, keeping paired real and generated records in the same partition. All C2ST results reported in the manuscript use this grouped implementation.

B.3 Effect of Covariate Compression

Reducing the covariate representation with principal components did not restore useful power for the energy test against the broken generator.

Table 4: **Energy-test p-value against the broken generator after covariate compression.**

Covariate representation	Energy p-value
No covariates	0.004
2 principal components	0.356
5 principal components	0.970
10 or more principal components	1.000

Thus, even a small number of added covariate dimensions substantially reduced the sensitivity of this distance-based verification procedure.

B.4 Controlled Verification Experiments

Under severe generator error and $n = 2{,}000$, the energy test detected the broken generator in 100% of binary-outcome replicates when no covariates were included. Detection fell to approximately 50% with two covariates and to 0% with five or more.

For continuous outcomes, energy retained 100% detection through 50 included covariates under the corresponding severe-error setting.

Mild generator errors were substantially harder to identify for binary outcomes. With covariates excluded, detection rates were 9% for energy, 19% for kNN, and 0% for Wasserstein-1, whereas the corresponding continuous-outcome experiments achieved approximately 97–100% detection.

Across 96 no-error conditions, false-positive rates were 0% for energy, 0% for Wasserstein-1, and 0.5% for kNN.

B.5 kNN Stability

The kNN statistic was sensitive to exact nearest-neighbour ties in the low-cardinality binary (T, Y) representation. Repeated analyses of the same comparison produced substantial variation under different tie resolutions.

For the same dataset and nominal random seed, the kNN p-value differed from 0.007 to 1.000 across computing environments. Energy and Wasserstein-1 were substantially more stable under the same reproducibility checks.

For this reason, kNN results were treated as unstable rather than as a reliable basis for benchmark acceptance.

B.6 Verification Across Generator Families

The eight fitted generator families showed similar treatment-model discrimination, with held-out AUCs ranging from 0.636 to 0.652.

More importantly, the main verification limitation was not specific to one generator family. Distance-based joint verification remained weak across statistical, boosted-tree, neural-network, and tabular foundation-model generators.

These results motivate interpreting a non-significant verification result only after the same procedure has demonstrated sensitivity to a known positive control in the relevant outcome and representation setting.

C Additional Estimator Benchmark Results

C.1 Average-Effect Recovery

When both nuisance models used the full covariate representation, the six estimators recovered the generator-defined treatment-effect target with similar accuracy. At the three-hour landmark, where the benchmark target was approximately -0.210, RMSE ranged from 0.0105 to 0.0107.

Table 5 provides the detailed RMSE values underlying the estimator-robustness results summarized in Figure 2B.

Table 5: **Estimator RMSE at the three-hour landmark under nuisance-model restrictions.** "Outcome restricted" and "treatment restricted" indicate that the corresponding nuisance model used age and sex only.

Estimator	Both full	Outcome restricted	Treatment restricted	Both restricted
S-learner	0.0105	0.0925	0.0105	0.0925
T-learner	0.0105	0.0929	0.0105	0.0929
IPW	0.0107	0.0107	0.0929	0.0929
Trimmed IPW	0.0107	0.0107	0.0929	0.0929
AIPW	0.0106	0.0108	0.0105	0.0930
TMLE	0.0106	0.0107	0.0105	0.0930

The detailed results show the expected estimator-specific pattern: S- and T-learners showed substantially larger error when the outcome model was restricted, whereas IPW showed substantially larger error when the treatment model was restricted. AIPW and TMLE remained close to their full-model performance when either nuisance model alone was restricted, but their error increased when both nuisance models were restricted.

C.2 Replication With TabPFN-3

The estimator stress test was independently repeated using a TabPFN-3 generator. The same qualitative pattern was reproduced: outcome-model estimators were sensitive primarily to outcome-model restriction, weighting estimators to treatment-model restriction, and AIPW and TMLE remained stable under a single nuisance-model restriction.

The replication was used to examine whether the estimator-stress-test pattern was specific to the primary benchmark generator rather than to claim that TabPFN-3 represented clinical ground truth.

C.3 Conditional-Effect Recovery

Conditional treatment effects were harder to recover than the average treatment effect.

Table 6: **Precision in estimation of heterogeneous effects (PEHE) in the TARNet benchmark.**

Estimator	1 h	3 h
S-learner	0.047	0.094
T-learner	0.063	0.081

This motivated a separate calibration analysis of the subgroup heterogeneity procedure.

C.4 Heterogeneity Calibration

Under the null condition with no additional planted heterogeneity, the subgroup procedure rejected in 8.5% of 200 replicates. The corresponding 95% confidence interval was 5.2–13.0%, compared with a nominal 5% rejection level.

Detection increased as stronger treatment-effect heterogeneity was introduced. At the strongest planted setting, six times the fitted heterogeneity, the maximum source-level treatment-effect spread

was 0.161, compared with approximately 0.24 observed in CATSS. Because the null rejection rate was somewhat above its nominal level, clinical subgroup p-values are interpreted cautiously.

D Clinical Sensitivity and Subgroup Analyses

This appendix provides additional details for the observational CATSS application, including treatment-window implementation, positivity, pretreatment, potential immortal-time bias, changing treatment-group composition, the continuous-delay analysis, and infection-source heterogeneity.

D.1 Treatment-Window Definition

For each landmark

$$L \in \{1, 3, 6, 12\} \text{ hours},$$

a patient was classified as treated if effective antimicrobial therapy was already active at t_0 or began by hour L. Patients whose therapy began after the landmark, together with patients who never received effective therapy, were classified as untreated at that landmark.

No patient was removed simply because treatment had not occurred by the landmark. Follow-up began at t_0 for all patients.

The 1- and 3-hour analyses were treated as primary. The 6- and 12-hour comparisons were secondary because treatment-group composition changed substantially at later landmarks.

AIPW was used for the primary adjusted estimates, with TMLE as a robustness check.

D.2 Landmark Estimates

Table 7: **Mortality-risk differences by treatment window.** Negative values indicate lower adjusted in-hospital mortality among patients classified as treated within the corresponding window.

Window	Crude difference	AIPW (95% CI)	TMLE
1 h	-0.210	-0.144 (-0.161, -0.127)	-0.144
3 h	-0.309	-0.216 (-0.232, -0.199)	-0.216
6 h	-0.396	-0.292 (-0.310, -0.275)	-0.292
12 h	-0.450	-0.360 (-0.380, -0.339)	-0.360

These estimates are observational adjusted risk differences and are not interpreted as experimentally established treatment effects.

D.3 Positivity Diagnostics

Across 28 evaluated strata, none had more than 5% of patients with extreme estimated treatment probabilities. No patient had an estimated propensity score outside

$$0.01 \leq \hat{e}(W) \leq 0.99.$$

These diagnostics did not identify a major empirical overlap problem for the prespecified landmark comparisons.

D.4 Patients Already Receiving Therapy at Shock Onset

A total of 2,621 patients (23.4%) were already receiving effective antimicrobial therapy when shock began.

Mortality in this group was 43.3%, compared with 23.0% among patients whose effective therapy began within the first hour after shock onset.

The landmark analyses were therefore repeated after excluding patients already receiving effective therapy at t_0.

Table 8: **AIPW estimates after excluding patients already receiving effective therapy at shock onset.**

Window	Full cohort	Pre-shock-treated excluded
1 h	-0.144	-0.266
3 h	-0.216	-0.280
6 h	-0.292	-0.329
12 h	-0.360	-0.372

D.5 Sensitivity to Potential Immortal-Time Bias

Because treatment status is determined using information after t_0, the fixed-window analyses remain susceptible to immortal-time bias.

A full hour-level clone–censor–weight analysis could not be implemented because mortality timing was recorded too coarsely for the 1–12 hour grace periods.

We therefore performed an intentionally extreme sensitivity analysis rather than claiming that the bias had been eliminated. A total of 388 patients (3.5% of the cohort) were recorded as dying within hospital day 1. First-day deaths classified as untreated were reassigned to the treated group.

Table 9: **Extreme sensitivity analysis for potential immortal-time bias.**

Window	Observed	Extreme scenario	Shift	Upper 95% CI endpoint
1 h	-0.144	-0.112	$+0.032$	-0.096
3 h	-0.216	-0.192	$+0.023$	-0.176
6 h	-0.292	-0.277	$+0.015$	-0.260
12 h	-0.360	-0.349	$+0.010$	-0.329

This is an extreme sensitivity scenario rather than a formal identification bound or correction for immortal-time bias.

D.6 Changing Treatment-Group Composition

The increasing magnitude of the landmark contrasts should not be interpreted as a simple dose-response relationship because the composition of both groups changes as L increases.

Table 10: **Composition of treated and untreated groups across treatment windows.**

Window	Untreated n	Never treated (%)	Untreated mortality (%)	Treated n	Pre-shock treated (%)	Treated mortality (%)
1 h	7,289	8.0	57.6	3,898	67.2	36.7
3 h	5,565	10.5	65.9	5,622	46.6	35.0
6 h	3,966	14.7	75.9	7,221	36.3	36.3
12 h	2,537	23.0	85.1	8,650	30.3	40.1

The 583 patients who never received effective therapy had 97.4% mortality and remained untreated at every landmark. Their share of the untreated group increased from 8.0% at one hour to 23.0% at twelve hours.

Even after excluding never-treated patients, mortality among patients remaining untreated increased from 54.2% to 81.4%.

Patients already receiving treatment at shock onset represented 67.2% of the treated group at one hour but only 30.3% at twelve hours.

These compositional changes are why the 6- and 12-hour analyses are treated as secondary.

D.7 Continuous Post-Shock Treatment Delay

A separate restricted-cubic-spline analysis described the adjusted association between observed post-shock treatment delay and mortality.

The analysis included 7,983 patients with a recorded post-shock delay and excluded patients already receiving effective therapy at onset and those who never received effective therapy.

The delay–mortality association was strongly nonlinear,

$$\chi^2 = 244.8, \qquad df = 3, \qquad p < 10^{-6}.$$

Adjusted mortality was 25.5% at 0.5 hours, 46.4% at six hours, and 83.0% at 24 hours.

Table 11: **Approximate slopes of the fitted delay–mortality curve.**

Delay interval	Percentage-point increase per hour
0.5–1 h	3.51
1–3 h	3.78
3–6 h	3.85
6–12 h	3.05
12–24 h	1.53
24–48 h	0.32

The spline and landmark analyses answer different questions. The landmark analyses compare groups whose membership changes with L, whereas the spline describes the association between recorded delay and mortality among patients with observed post-shock treatment delay. The spline is therefore interpreted descriptively rather than as a causal treatment-time response curve.

D.8 Infection-Source Heterogeneity

In the full cohort, evidence for one-hour heterogeneity was limited (Cochran's Q, $p = 0.34$; $I^2 = 11.7\%$).

After excluding patients already receiving effective therapy at shock onset, heterogeneity increased. Under the six-category infection-source definition,

$$I^2 = 79.2\%, \qquad Q\ p = 0.0002$$

at one hour, and

$$I^2 = 66.2\%, \qquad Q\ p = 0.011$$

at three hours.

Under the four-category infection-source representation, one-hour heterogeneity was

$$I^2 = 73.1\%, \qquad Q\ p = 0.011.$$

Because never-treated patients were unevenly distributed across infection sources and had very high mortality, the heterogeneity analysis was also repeated after excluding them. At one hour,

$$I^2 = 75.2\%, \qquad Q\ p = 0.0012.$$

These subgroup results are interpreted cautiously because the controlled heterogeneity analysis showed a null rejection rate of 8.5% at a nominal 5% level.

D.9 Infection Source and Delay-Curve Shape

We separately tested whether the shape of the delay–mortality association differed across infection sources.

The infection-source-by-delay interaction was not statistically supported:

$$\chi^2 = 5.44, \qquad df = 12, \qquad p = 0.942.$$

Thus, the available data provided stronger evidence for differences in the magnitude of the early adjusted mortality contrast than for differences in the shape of the observed delay–mortality association.

D.10 Interpretation of the Clinical Analyses

The CATSS analyses remain observational and depend on adjustment for measured baseline characteristics. Residual or unmeasured confounding cannot be ruled out.

The generated-data benchmark provides a controlled environment in which the target is model-defined and known. It does not establish that the corresponding generator-defined treatment effect is the true treatment effect in CATSS.

Similarly, the immortal-time sensitivity analysis evaluates how estimates change under an extreme reassignment of early deaths but does not eliminate immortal-time bias.

For these reasons, the clinical findings are described as adjusted mortality-risk differences and observational associations rather than as experimentally established causal effects.