Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Digital twins as synthetic control arms: a transportability and calibration credibility check on real critically-ill survival data

Ahmed Fawaz

ahmedfawa879@gmail.com

Port Said University

Research Article

Keywords: digital twin, synthetic control arm, external control, prognostic covariate adjustment, calibration, transportability, survival analysis, in-silico trial

Posted Date: September 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10940313/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background.

Digital-twin synthetic control arms generate per-patient counterfactual outcomes from a model trained on a reference population, in order to reduce or partially replace concurrent control recruitment in clinical trials. The per-patient counterfactual is, by construction, never directly observable; however, the transportability assumption on which the method depends — that a model of the covariate-to-outcome relationship learned in one population predicts a genuinely different population's real, observed outcomes — is directly testable on real data.

Methods.

Using SUPPORT2 (9,105 critically ill adults across five US academic medical centres, with real right-censored survival), we defined two clinically distinct, non-randomised populations by resuscitation status (full-code versus do-not-resuscitate [DNR]). We trained digital-twin survival models at two complexity tiers — Cox proportional hazards (Tier A) and a discrete-time DeepHit-style neural model (Tier B) — on 14 baseline clinical covariates, using patient-level, leakage-checked reference/holdout splits with all preprocessing fit on reference data only. Aggregate calibration was pre-specified at 30, 60, 90 and 180 days as whether the mean predicted survival in the holdout fell within a bootstrap 95% confidence interval (1,000 resamples) of the holdout's real Kaplan-Meier survival. We ran a within-population sanity check (full-code reference to full-code holdout) and the cross-population synthetic-control scenario (DNR reference to full-code holdout).

Results.

Within-population, mean predicted survival fell inside the real Kaplan-Meier 95% CI at 1 of 8 checkpoint-tier combinations (Tier A, 1 of 4, at 180 days only; Tier B, 0 of 4), but every within-population prediction was within 4.2 percentage points of the truth — a stringency artefact of narrow confidence intervals rather than meaningful miscalibration. In the cross-population synthetic-control scenario, coverage was 0 of 8 and the twin underestimated survival by 51 to 67 percentage points at every horizon (for example, predicting 17% versus a real 73% at 180 days).

Conclusions.

In this cohort, a digital twin that was well calibrated within its own training population became severely miscalibrated when transported to a different population, and was therefore not a trustworthy synthetic control. The transportability assumption underlying synthetic-control substitution should be empirically

demonstrated, per population and per time horizon, before such twins are permitted to inform comparative inference. We provide a small, fully reproducible harness for performing this check.

Background

Randomised controlled trials remain the reference standard for causal inference, but recruiting an adequately sized concurrent control arm is often the most expensive, slow, and ethically fraught part of a study. This has driven sustained interest in external, synthetic, and “digital-twin” control arms, in which historical or model-generated data stand in for some or all of the concurrent controls [8, 9]. In the digital-twin formulation, a generative or prognostic model is trained on a reference population and then used to produce, for each enrolled trial participant, a personalised prediction of the outcome that participant would have experienced under standard of care — a per-patient counterfactual that can supplement or partially replace the real control arm.

The epistemic difficulty is fundamental rather than technical. A counterfactual outcome — what a specific enrolled patient would have experienced in the arm they were not assigned to — is never observed, so a digital twin's per-patient prediction cannot be checked against ground truth directly. What can be checked, on real data, is the assumption the entire construction rests upon: does a model of the covariate-to-outcome relationship, learned in one population, correctly predict a genuinely different, held-out population's real, observed outcomes? If it does not, the synthetic-control substitution is not trustworthy for that population, irrespective of how the twin was built.

This verification-versus-validation distinction is made explicit in the credibility framework used to assess computational models of medical devices, where verification asks whether the code correctly implements the intended model and validation asks whether the model accurately represents reality [10]. Applied to a digital-twin survival model, verification is a matter of numerical correctness, whereas validation is precisely the transportability question above.

We therefore set out to provide a minimal, transparent, and fully reproducible transportability-and-calibration check for digital-twin synthetic controls, and to demonstrate it on a real, openly available cohort — reporting, deliberately, whatever the data showed, including a negative result. We emphasise at the outset that this is a cautionary methodological demonstration on a single observational dataset, not an efficacy or treatment-effect study, and we draw no causal conclusion from the data.

Methods

Data source

We used the SUPPORT2 dataset (Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments), comprising 9,105 critically ill adults across five US academic medical centres, with real right-censored survival outcomes over a 180-day study horizon [1, 2]. (SUPPORT phase I/II accrual 1989–1994; 180-day follow-up horizon). The dataset is publicly available without credentialing through

the UCI Machine Learning Repository (and a mirrored Kaggle copy) [12]. Fourteen baseline clinical covariates were used as model inputs: age, sex, number of comorbidities, mean arterial blood pressure, heart rate, respiratory rate, temperature, white blood cell count, serum sodium, creatinine, albumin, bilirubin, blood urea nitrogen, and glucose. The outcome was death (event indicator) with observed or censored duration. Missing covariates were imputed with medians and all covariates standardised; critically, medians and standardisation parameters were fitted on the reference cohort only and then applied unchanged to the holdout, so that no holdout information influenced preprocessing.

Two real, clinically distinct populations

To obtain two genuinely different populations from a single dataset — rather than an artificial random partition — we split patients by resuscitation status into a full-code group (no DNR order) and a DNR group (any DNR order). This is an extensively studied, clinically meaningful contrast in the original SUPPORT literature, driven by real clinical decisions rather than randomisation. We stress that SUPPORT2 is observational: the two groups are not randomised arms, the analysis draws no causal treatment-effect conclusion, and DNR status is itself strongly associated with prognosis. The groups serve here only as two authentically different populations across which to test transportability — which is exactly the situation a real-world synthetic control must survive.

Digital-twin survival models (a complexity gradient)

To test whether any conclusion depended on model complexity, we implemented two tiers. Tier A was a Cox proportional-hazards model [3], the interpretable field standard. Tier B was a small discrete-time neural survival model in the DeepHit family [5] (20 time bins over the 180-day horizon, hidden dimension 64, trained for 60 epochs with a discrete-time negative-log-likelihood loss). Each model was trained on a reference population and used to generate, for every holdout patient, a predicted survival function from baseline covariates alone — the digital-twin prediction.

Reference/holdout split and leakage control

Within each experiment, the reference population was split from the holdout at the patient level (70%/30%; fixed random seed) with an explicit assertion that no patient identifier appeared in both sets. All model fitting and preprocessing used the reference set only; the holdout was used solely for evaluation.

Pre-specified calibration endpoint

The primary endpoint was aggregate calibration of the digital twin against the holdout population's real outcomes. At each of four checkpoints (30, 60, 90 and 180 days) we computed the mean predicted survival across the holdout and compared it against a bootstrap 95% confidence interval (1,000

resamples) of the holdout's real Kaplan-Meier survival [4] at that checkpoint. The per-checkpoint result was a binary indicator of whether the mean prediction fell within the real KM 95% CI. Coverage was summarised across the four checkpoints and two tiers (eight combinations per experiment). This is a first-order, necessary-not-sufficient calibration criterion, chosen for transparency and because a synthetic control whose mean prediction lies outside the real interval is, by definition, miscalibrated for that population at that horizon.

Two experiments

Experiment 1 (within-population sanity check)

the full-code reference set trained the twin, which was then evaluated on the full-code holdout. A synthetic-control method must, at minimum, pass this before any cross-population use.

Experiment 2 (cross-population stress test)

the DNR reference set trained the twin, which was then evaluated on the full-code holdout — the actual digital-twin-as-synthetic-control scenario, in which a model built on one population is asked to stand in for a different one.

Verification and reproducibility

Verification comprised unit tests confirming that the discrete-time model's predicted probability mass summed to one across bins, that the neural loss agreed with the Cox baseline on toy data, that the calibration outputs had correct shapes, and that reference and holdout sets never shared a patient. Analyses used Python with the lifelines library [11] and a fixed global random seed, and are reported in line with the TRIPOD + AI reporting guideline for prediction-model studies [6]. All code is openly available under an MIT licence (see Availability of data and materials).

Results

Cohort

Of 9,105 patients, 30 were removed for missing event, duration, or resuscitation status, leaving 9,075 analysable patients with an overall observed mortality of 68.1% over the 180-day horizon. The full-code group comprised 5,880 patients (reference 4,116; holdout 1,764) and the DNR group 3,195 patients (reference 2,236; holdout 959); no patient identifier appeared in both a reference and a holdout set (Table 1). Both survival models were fit on the reference set of the relevant experiment using the 14 baseline covariates.

Within-population calibration (Experiment 1)

When the twin was trained and evaluated within the full-code population, mean predicted survival fell inside the real Kaplan-Meier 95% CI at only 1 of 8 checkpoint-tier combinations: the Cox model met the criterion at 1 of 4 checkpoints (at 180 days only), and the neural model at 0 of 4 (Table 2; Fig. 1, left panel). Critically, however, the misses were of negligible clinical magnitude: across all eight within-population combinations the mean prediction was within 0.0 to 4.2 percentage points of the real survival (Cox: 1.3, 1.8, 1.9 and 0.0 points at 30/60/90/180 days; neural: 2.1, 2.6, 4.1 and 4.2 points). These misses reflect the stringency of a bootstrap confidence interval only about 2 to 4 percentage points wide, not a clinically meaningful calibration error — within-population, the twin was in fact well calibrated.

Cross-population calibration (Experiment 2)

When the twin was trained on the DNR population and asked to serve as a synthetic control for full-code patients — the operational synthetic-control scenario — mean predicted survival fell inside the real KM 95% CI at 0 of 8 checkpoint-tier combinations (both tiers, all horizons; Table 2; Fig. 1, right panel). Here the failure was not a matter of stringency but of magnitude: the DNR-trained twin predicted 180-day survival of 16.7% (Cox) and 11.4% (neural) for full-code patients whose real survival was 73.0%, and across horizons the mean predictions were wrong by 51 to 67 percentage points (Table 2). A twin built on the sicker DNR population thus massively underestimated the survival of the healthier full-code population — direct, real-data evidence that the DNR reference did not transport as a synthetic-control source for full-code patients in this cohort.

Table 1
Cohort derivation and reference/holdout allocation. Of 9,105 SUPPORT2 patients, 30 with missing event, duration, or resuscitation status were excluded, leaving 9,075. Within each arm the reference/holdout split was 70/30 at the patient level, with no patient in both sets. Follow-up horizon 180 days.

Analysed patients, n	**Full-code**	**DNR**	**Total**
	5,880	**3,195**	**9,075**
Reference set, n	4,116	2,236	**6,352**
Holdout set, n	1,764	959	**2,723**
Overall mortality, %	—	—	**68.1**

Table 2

Per-horizon calibration of the digital-twin survival models against the real full-code holdout. Predicted = holdout mean predicted survival; Real KM = holdout Kaplan-Meier survival with bootstrap 95% CI (1,000 resamples); Gap = absolute difference (percentage points); In CI = prediction within the CI. Coverage: within-population 1/8, cross-population 0/8. Within-population gaps are ≤ 4.2 points; cross-population gaps are 51–67 points.

Exp	Tier	Day	Pred. %	Real KM % (95% CI)	Gap	In CI
1	Cox PH	30	90.4	91.7 (90.5–92.9)	1.3	No
1	Cox PH	60	85.2	87.0 (85.5–88.6)	1.8	No
1	Cox PH	90	81.2	83.2 (81.3–84.9)	1.9	No
1	Cox PH	180	72.9	73.0 (70.9–74.9)	0.0	**Yes**
1	DeepHit-style	30	89.6	91.7 (90.5–92.9)	2.1	No
1	DeepHit-style	60	84.4	87.0 (85.5–88.6)	2.6	No
1	DeepHit-style	90	79.1	83.2 (81.3–84.9)	4.1	No
1	DeepHit-style	180	68.8	73.0 (70.9–74.9)	4.2	No
2	Cox PH	30	40.6	91.7 (90.5–92.9)	51.1	No
2	Cox PH	60	27.9	87.0 (85.5–88.6)	59.1	No
2	Cox PH	90	22.9	83.2 (81.3–84.9)	60.2	No
2	Cox PH	180	16.7	73.0 (70.9–74.9)	56.3	No
2	DeepHit-style	30	38.6	91.7 (90.5–92.9)	53.0	No
2	DeepHit-style	60	22.4	87.0 (85.5–88.6)	64.6	No
2	DeepHit-style	90	16.5	83.2 (81.3–84.9)	66.6	No
2	DeepHit-style	180	11.4	73.0 (70.9–74.9)	61.6	No

Exp 1 = within-population (full-code reference → full-code holdout); Exp 2 = cross-population (DNR reference → full-code holdout).

Discussion

On a real cohort of critically ill adults, a digital-twin survival model trained on one clinically-defined population was not a calibrated synthetic control for a different population. The signal was unambiguous and large: trained on the sicker DNR population, the twin underestimated full-code survival by 51 to 67 percentage points at every horizon, predicting that roughly one in six full-code patients would survive to 180 days when in fact nearly three in four did. The failure was consistent across an interpretable statistical model and a neural one, indicating that it is not an artefact of a single modelling choice.

The within-population results calibrate how to read this. There, the same binary criterion recorded only 1 of 8 passes, yet every within-population prediction lay within 4.2 percentage points of the truth (Table 2) — clinically negligible errors that happened to fall outside a bootstrap confidence interval only 2 to 4 points wide. Within-population, in other words, the twin was well calibrated, and the binary criterion overstates its failure. That contrast is what makes the cross-population result decisive: it is not a stringency artefact but a genuine, order-of-magnitude transportability failure. A real synthetic-control deployment transports a model across populations that differ in exactly the confounded, prognosis-associated ways that full-code and DNR patients differ; our result is a concrete demonstration that such differences can turn a well-behaved twin into a catastrophically miscalibrated one.

These findings do not contradict the broader prognostic-covariate-adjustment literature. Methods that incorporate a twin-derived prognostic score as a covariate alongside a real, concurrent control arm can preserve type-I error by design, because the concurrent randomised comparison is retained [9]. Our caution applies specifically to the more aggressive use in which a twin's predictions substitute for real control outcomes; it is that substitution that depends on transportability, and it is transportability that we show cannot be assumed. A simple, pre-specified calibration check of the kind demonstrated here is a low-cost gate before such substitution is contemplated.

Strengths

The study uses real right-censored survival data rather than simulation; enforces patient-level, leakage-checked splits; fits all preprocessing on reference data only; pre-specifies a transparent binary calibration criterion; separates verification from validation explicitly; and is fully reproducible from public data and open code. It also reports a negative result plainly, without reframing it as a qualified success.

Limitations

Several limitations bound the claim. (i) This is a single observational cohort; the two populations differ by a clinical decision that is itself strongly outcome-associated, so the cross-population failure reflects a realistic but confounded population difference and does not isolate a single mechanism — it demonstrates failure under a representative real-world difference rather than under a controlled one. (ii) The endpoint is a first-order aggregate calibration check; it is necessary but not sufficient, and does not assess individual-level calibration across risk strata (for example, integrated calibration index or distribution calibration [7]) or discrimination — passing it would not by itself certify a twin. (iii) The binary criterion's stringency at large sample size means the absolute miscalibration magnitude, not only pass/fail, must be inspected; we report it in full (Table 2), and it is precisely what distinguishes the negligible within-population misses from the catastrophic cross-population ones. (iv) Only two model tiers with limited tuning were examined, and the neural tier may be under-optimised; however, the appropriately-specified Cox model also failed within-population at most horizons, so the finding is not merely an under-trained deep model. (v) Follow-up is capped at the 180-day SUPPORT2 horizon, leaving longer-horizon transportability untested.

Implications

Before a digital twin's predictions are permitted to substitute for real control outcomes, a transportability-and-calibration check should be performed and reported per target population and per time horizon, treating an adequately calibrated within-population result as a necessary precondition and an adequately calibrated cross-population result as the actual requirement. We provide a minimal, reusable harness to make this check routine.

Conclusions

A digital-twin survival model trained on one real population failed to serve as a calibrated synthetic control for a different real population, and failed the same aggregate-calibration criterion at most horizons even within-population. The counterfactual a digital twin promises cannot be validated directly, but the transportability it silently assumes can — and, in this cohort, it did not hold. Transportability should be demonstrated, not assumed, before digital-twin synthetic controls inform comparative clinical inference.

Declarations

Clinical trial number. Not applicable.

Ethics approval and consent to participate.

Not applicable; this is a secondary analysis of a fully de-identified, publicly available dataset (SUPPORT2). The original SUPPORT study obtained its own institutional approvals and consent [1, 2].

Consent for publication.

Not applicable, no patients or members of the public were involved in this study.

Competing interests.

The author declares no competing interests.

Funding.

This work received no specific funding.

Author Contribution

A.F. is the sole author and was solely responsible for all aspects of this work, including its conception and design, the analysis and interpretation of the data, the writing of the software and execution of the experiments, and the drafting and critical revision of the manuscript. The author read and approved the final manuscript.

Acknowledgements.

The author thanks the SUPPORT investigators for making the data available and the maintainers of the open-source lifelines library.

Data Availability

The dataset analysed in this study, SUPPORT2, is publicly available in the UCI Machine Learning Repository (dataset 880) at https://archive.ics.uci.edu/dataset/880/support2, and can be accessed programmatically via the ucimlrepo Python package (fetch_ucirepo(id=880)); no registration or credentialing is required. The dataset was originally described by Knaus et al. (Ann Intern Med. 1995;122:191–203). All code used to load the data, fit the models, and generate the results and figures in this manuscript is openly available in a public GitHub repository at https://github.com/ahmedfawaz879/digital-twin-synthetic-controls under an MIT licence. No new data were generated in this study.

References

1. Knaus WA, Harrell FE Jr, Lynn J, Goldman L, Phillips RS, Connors AF Jr, et al. The SUPPORT prognostic model: objective estimates of survival for seriously ill hospitalized adults. Ann Intern Med. 1995;122(3):191–203. 10.7326/0003-4819-122-3-199502010-00007.
2. The SUPPORT Principal Investigators. A controlled trial to improve care for seriously ill hospitalized patients (SUPPORT). JAMA. 1995;274(20):1591–8. 10.1001/jama.274.20.1591.
3. Cox DR. Regression models and life-tables. J R Stat Soc Ser B. 1972;34(2):187–202.
4. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53(282):457–81. 10.1080/01621459.1958.10501452.
5. Lee C, Zame WR, Yoon J, van der Schaar M. DeepHit: a deep learning approach to survival analysis with competing risks. Proc AAAI Conf Artif Intell. 2018;32(1):2314–21. 10.1609/aaai.v32i1.11842.
6. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. 10.1136/bmj-2023-078378.
7. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230. 10.1186/s12916-019-1466-7.

8. Fisher CK, Smith AM, Walsh JR. Coalition Against Major Diseases. Machine learning for comprehensive forecasting of Alzheimer's disease progression. Sci Rep. 2019;9(1):13622. 10.1038/s41598-019-49656-2.
9. Schuler A, Walsh D, Hall D, Walsh J, Fisher C, Alzheimer's Disease Neuroimaging Initiative. Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score. Int J Biostat. 2022;18(2):329–56. 10.1515/ijb-2021-0072.
10. American Society of Mechanical Engineers. ASME V&V 40-2018. Assessing credibility of computational modeling through verification and validation — application to medical devices. New York: ASME; 2018.
11. Davidson-Pilon C. lifelines: survival analysis in Python. J Open Source Softw. 2019;4(40):1317. 10.21105/joss.01317.
12. Harrell FE, Machine Learning UCI. Repository; 1995. Dataset 880. https://archive.ics.uci.edu/dataset/880/support2

Figures

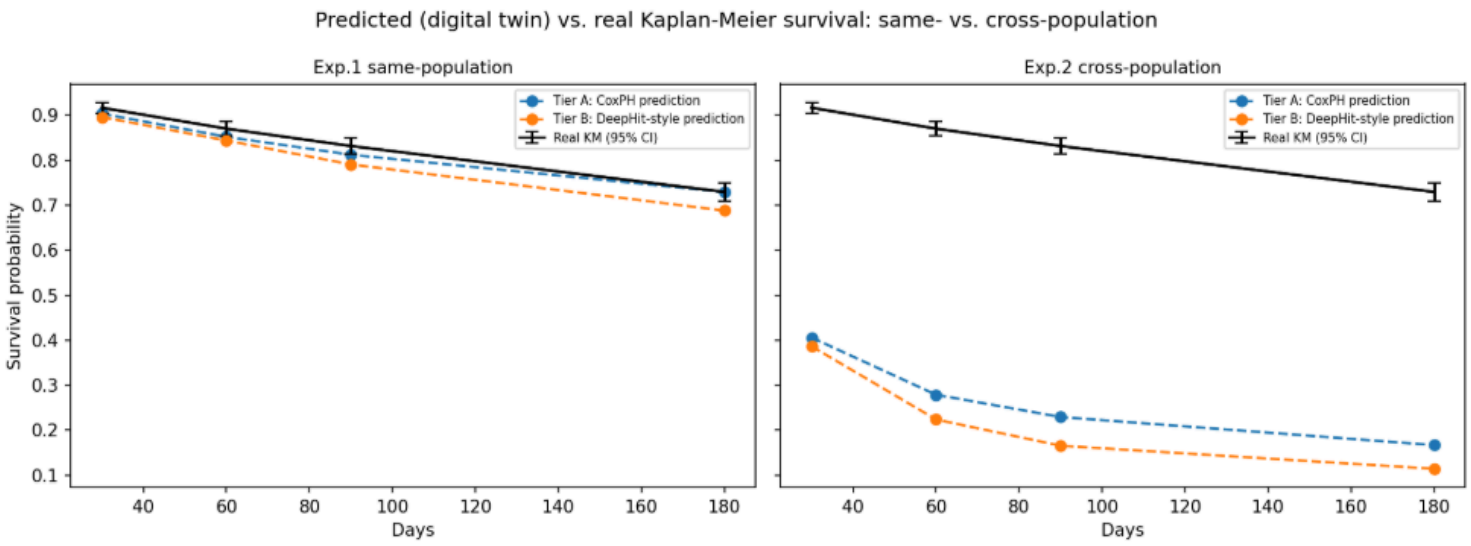


Figure 1

Digital-twin predicted survival versus the real full-code holdout Kaplan-Meier curve (black, with bootstrap 95% CI), at 30/60/90/180 days. Left: within-population training (Exp 1), where both tiers track the real curve closely. Right: cross-population training (Exp 2), where both tiers fall far below the real survival — the transportability failure. Same axes in both panels.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- DigitalTwinTRIPODAIChecklist.docx