Pooled validation found nothing. Stratified by hospital size, it did.

An independent adversarial validation system, pointed at sepsis prediction models retrained on four public ICU datasets and 286,510 admissions. Thresholds locked on OSF before data access. Sepsis is the test case. The result is about how clinical AI gets validated.

Adam Dickens
RocSite, Inc. / adam@rocsite.com / rocsite.com

Pre-registered OSF, 11 March 2026 / **Primary runs** 14 March 2026
Preprint medRxiv 2026.03.17.26348414v2 / **Code** Apache-2.0, public

01 The system, and the test case

RocSite is an independent adversarial validation layer for clinical AI. It does not build the model. It attacks the model on terms fixed before the data is opened, and records the run so a third party can check the attack was real.

The demonstration needed a well studied target with a contested failure mode. Sepsis prediction qualifies: the standing concern is that these models learn **care-process intensity** rather than biological deterioration.

Models under test are of the type published in the sepsis prediction literature, retrained here on each dataset. No vendor or deployed system was audited. **Cohorts**: MIMIC-IV 65,241 / MIMIC-III 44,091 / eICU-CRD 136,864 / Challenge 2019 40,314, total **286,510**. XGBoost 300 trees, depth 6, 5-fold stratified CV, seed 42. The four locked rules are printed in Figure 1.

STEP 01
PRE-REGISTER
Hypothesis, plan and numeric thresholds locked and timestamped on a public registry before any data access.
OSF, 11 March 2026

STEP 02
PROBE, DO NOT BENCHMARK
Adversarial checks, not leaderboards. Label instability, care-process confounding, feature ablation, synthetic discrimination.
4 locked tests, 4 datasets

STEP 03
RECORD EVERYTHING
Each run emits a manifest with seed, library versions and platform, per-phase JSON, a warnings file and a summary.
run_20260314_153526_9ac64a5a

STEP 04
ATTEST AND PUBLISH
Ed25519 signing. Credentialed data only. Results published in whichever direction they fall.
Apache-2.0, public code

02 Confirmatory results

PRE-REGISTERED PHASE AND ITS LOCKED RULE	MIMIC-IV v3.1 · n = 65,241	MIMIC-III v1.4 · n = 44,091	eICU-CRD v2.0 · n = 136,864	Challenge 2019 · n = 40,314
PHASE 1 Ground truth stability confirms if mean Jaccard < 0.50	0.5124 NOT CONFIRMED	0.4281 CONFIRMED	0.4482 CONFIRMED	0.3042* ARTIFACT, NOT COUNTED
PHASE 2 Feature dependence confirms if AUROC drop > 0.15	0.0027 NOT CONFIRMED	0.0137 NOT CONFIRMED	0.0760 NOT CONFIRMED	0.0292 NOT CONFIRMED
PHASE 3 Care-intensity universality confirms if care-only AUROC > 0.70	0.6603 NOT CONFIRMED	0.6087 NOT CONFIRMED	0.6589 NOT CONFIRMED	0.6939 NOT CONFIRMED
PHASE 4 Synthetic validation confirms if discriminator AUROC < 0.60	0.6330 NOT CONFIRMED	0.7864 NOT CONFIRMED	0.6728 NOT CONFIRMED	0.4598 CONFIRMED

Figure 1. Every pre-registered phase run independently against every dataset. * Challenge 2019 carries no ICD billing codes, so two of its three pairwise Jaccard terms are structurally zero. That Phase 1 confirmation is an artifact and we do not count it. Values are the 14 March primary runs; the v2 preprint reports MIMIC-IV Phase 1 as 0.5122 after a GCS fix we found and published ourselves, and no verdict changed.

4 of 16 pre-registered tests confirmed, or **3 of 16** once the Challenge 2019 artifact is removed. Ground truth instability confirmed in **2 of the 3** cohorts carrying billing labels. On the confirmatory analysis the care-process shortcut hypothesis is **not supported**.

03 Exploratory findings

POST HOC, NOT PRE-REGISTERED

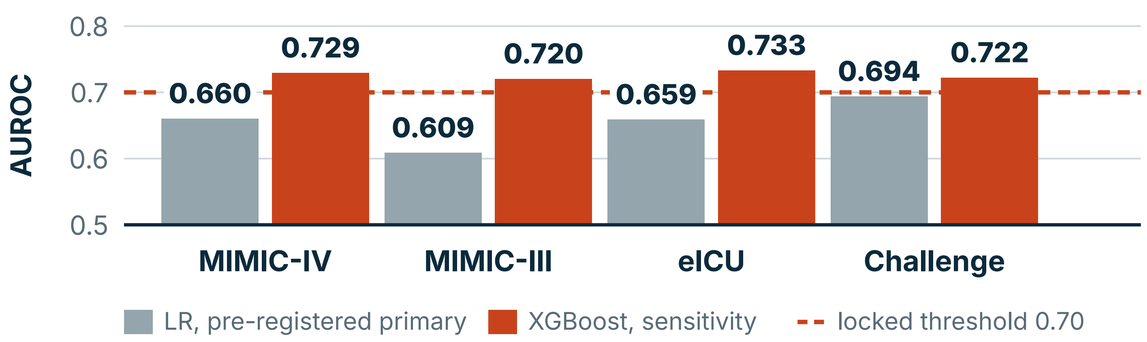

Figure 2. Phase 3 by estimator. The pre-registered primary, logistic regression, fell short of 0.70 in all four datasets. The XGBoost sensitivity analysis cleared it in all four: care-intensity features alone reach 0.72 to 0.73.

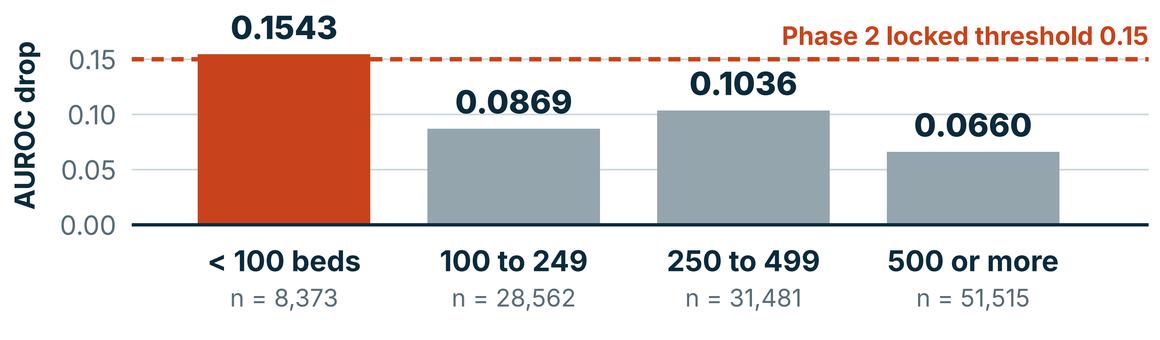

Figure 3. Phase 2 recomputed within eICU strata by hospital bed count. Under 100 beds the drop is **0.1543**, exceeding the locked 0.15 threshold; pooled it is 0.0760. The trend is not monotonic (rho = -0.80, p = 0.20, 4 bins).

eICU COMMUNITY HOSPITALS, SHAP ATTRIBUTION, n = 101,278
physician_order_rate is the most important feature in the model. Mean absolute SHAP 1.129, **4.6 times** the next feature, and the top feature for **51.6%** of patients. Dependence also rises as hospitals shrink: drop 0.0489 academic, 0.0572 teaching, 0.0853 community.

HOW TO READ SECTION 03
None of this was pre-registered. It ran after the confirmatory verdict was recorded, it is hypothesis generating, and it does not change that verdict. It suggests the 0.15 threshold was set too high and the pooled design too coarse, and that a dependence invisible at cohort level may concentrate in small community hospitals. That is a study to pre-register, not a claim to make here.

04 Discussion, limitations and conclusion

Discussion. The pooled pre-registered test found nothing. A stratification by hospital size found a dependence exceeding the same threshold. Pooling is the default in clinical AI validation, and pooling is exactly where this would have been missed.

Limitations. Retrospective secondary analysis. Jaccard is intersection over union, not percent agreement. Features are whole-stay aggregates, which inflates AUROC. **The author is an AI systems architect, not a clinician; clinical interpretation needs collaborators we are seeking.**

Conclusion. A validation procedure that cannot return NOT CONFIRMED is not validating anything. A procedure that only ever reports pooled results can fail to fail. Both are properties of the procedure, not of sepsis.

References. 1. Dickens A. medRxiv 2026.03.17.26348414v2. 2. OSF pre-registration osf.io/9tbjm, 11 March 2026. 3. Johnson AEW et al. MIMIC-IV v3.1, MIMIC-III v1.4. PhysioNet. 4. Pollard TJ et al. eICU-CRD. Sci Data 2018. 5. Reyna MA et al. PhysioNet/CinC Challenge 2019. 6. Singer M et al. JAMA 2016.

Funding and competing interests. No external funding. A.D. is the founder of RocSite, Inc., which develops independent validation tooling for clinical AI and therefore has a commercial interest in the methodology reported here. Code is public under Apache-2.0. All work was under PhysioNet credentialed access and no patient-level data left its host environment. RocSite also builds AI Governor, ICH Triage, Caliper and Discovery.

PREPRINT PRE-REG CODE