Privacy-Preserving Multicenter Evaluation of ICU Clinical Decision Support: A Federated Framework Across Eight U.S. Health Systems

Vaishvik Chaudhari, MS[1], Kaveri Chhikara, MS[2], Hoda Materi, MS[1], Zewei Liao, MS[2], Nicholas E. Ingraham, MD, MS[5], Patrick G. Lyons, MD, MSc[3], Brenna Park-Egan, MS[3], Catherine A. Gao, MD, MS[4], Wan-Ting Liao, MS[4], Sivasubramanium Bhavani, MD[6], Aartik Sarma, MD[8], Navya Ramesh, MS[6], Chad H. Hochberg, MD, MHS[7], William F. Parker, MD, PhD[2], Kevin Buell, MBBS, MS[1], Juan C. Rojas, MD, MS[1]

[1]Rush University, Chicago, IL [2]University of Chicago, Chicago, IL [3]Oregon Health & Science University, Portland, OR [4]Northwestern University, Chicago, IL
[5]University of Minnesota, Minneapolis, MN [6]Emory University, Atlanta, GA [7]Johns Hopkins University, Baltimore, MD [8]University of California, San Francisco, CA

BACKGROUND

- AI-based clinical decision support tools trained at one institution often demonstrate poor external validity when deployed at other sites
- External validation typically requires sharing patient-level data, creating regulatory and privacy barriers
- Widely implemented commercial sepsis prediction models have shown poor performance in external validation [1]

Objective: Using inpatient mortality prediction as an initial use case, we compared three deployment strategies for ICU prediction models across 8 U.S. health systems

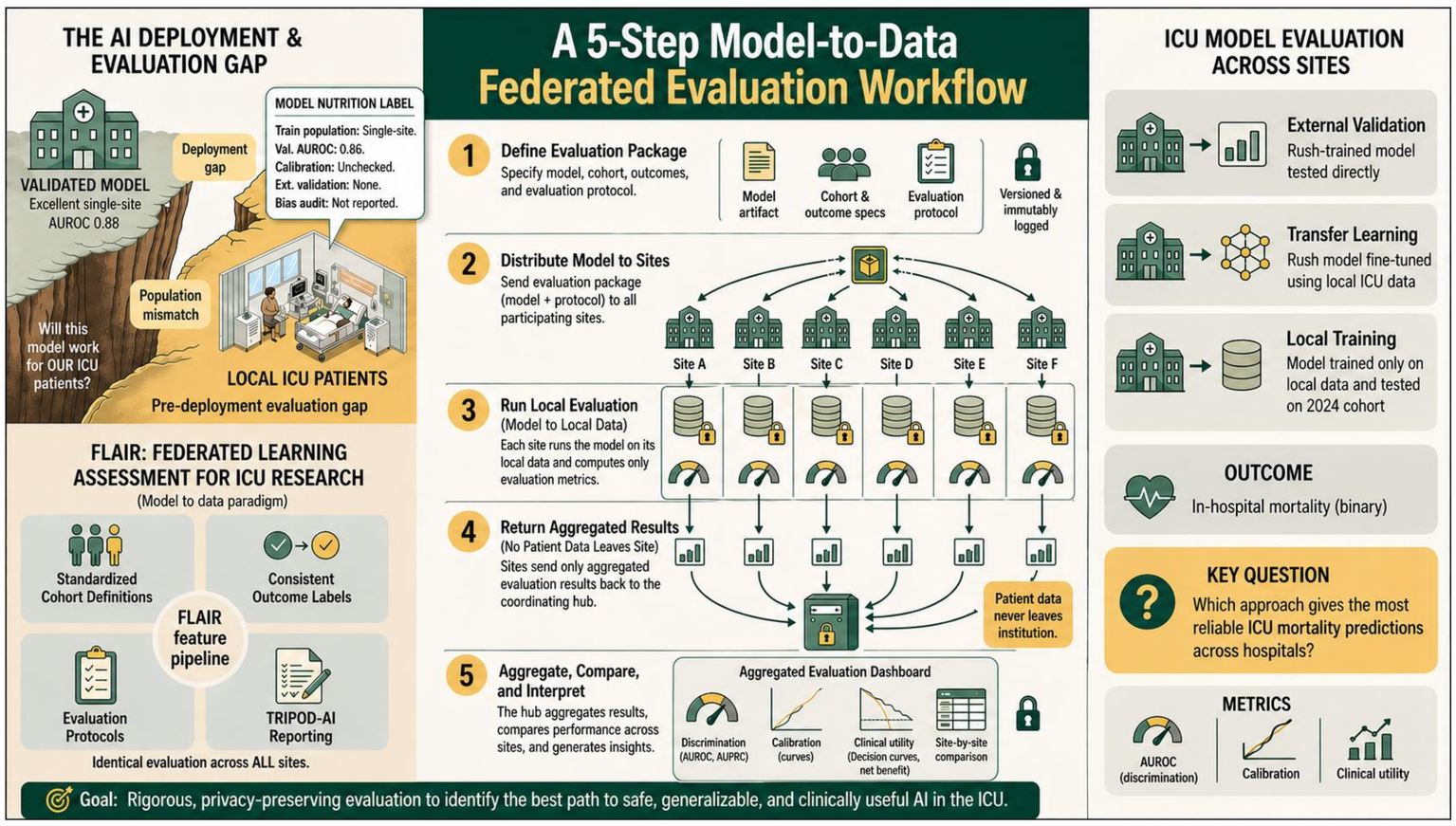

FIGURE 1. MODEL EVALUATION WORKFLOW

METHODS

Framework

FLAIR (Federated Learning Assessment for ICU Research): a privacy-preserving benchmarking framework. Trained model objects cross institutional firewalls; patient data remains local. Model to Data (MTD) paradigm [2].

Data Standard

Common Longitudinal ICU data Format (CLIF)[3] across 8 U.S. academic medical centers.

Study Population

- Adult ICU patients admitted 2018–2024
- Survived ≥24 hours post-ICU admission
- Demographics, vitals, labs, respiratory support extracted at 24 hours

Primary Outcome

In-hospital mortality (binary classification)

Models

- XGBoost and ElasticNet

Metric

- TRIPOD-AI Evaluation Metrics: AUROC, Brier Score, and Decision Curve Analysis

Deployment Strategies Compared

A. External Validation — Rush-trained model evaluated directly on external 2024 data

B. Transfer Learning — Rush model weights initialize site-specific fine-tuning on 2018–2023 data

C. Independent Training — models trained entirely on local 2018–2023 data, tested on 2024 data

KEY RESULTS

- **372,673 ICU encounters** across **8** U.S. health systems (mean age 62 years, 45% female)
- In-hospital mortality 11–18% across sites
- Demographic diversity: Black 2.4–59.1%; Hispanic 2.1–18.8%

XGBoost AUROC across strategies:

Cross-site validation: 0.85 median, range [0.78–0.89]

Transfer learning: 0.86 median, range [0.79–0.89]

Independent training: 0.87 median, range [0.85–0.89]

- Largest improvement: poorest cross-site AUC (0.78) rose to 0.85 with independent training
- *Calibration:* Brier scores improved from 0.15–0.19 (cross-site) to 0.13–0.16 (independent)
- *ElasticNet:* AUC improved from 0.77 to 0.82 with transfer learning at the poorest-performing site

FIGURE 2. MODEL PERFORMANCE ACROSS SITES AND DEPLOYMENT STRATEGY

CONCLUSIONS

- Site-specific model training outperformed Rush models for mortality prediction across all external sites
- CLIF addresses data harmonization; FLAIR enables federated model evaluation without sharing patient data
- Framework is extensible to additional ICU prediction tasks and available to any institution adopting CLIF
- MIMIC-IV in CLIF format provides an open-source entry point for broader adoption and benchmarking

CLIF

FLAME ICU REPO

REFERENCES

1. Wong A, et al. External validation of a widely implemented proprietary sepsis prediction model. **JAMA Intern Med 2021;181(8):1065–70.**
2. Guinney J, Saez-Rodriguez J. "Alternative models for sharing confidential biomedical data." **Nat Biotechnol 2018;36(5):391-2.**
3. Rojas JC, Lyons PG, et al. A common longitudinal ICU data format. **Intensive Care Med 2025;51(3):556–69.**
4. FLAIR: github.com/Common-Longitudinal-ICU-data-Format/**FLAIR**
5. FLAME-ICU: github.com/Common-Longitudinal-ICU-data-Format/**FLAME-ICU**
6. Liao, Zewei, et al. "MIMIC-IV-Ext-CLIF: MIMIC-IV in the Common Longitudinal ICU data Format (CLIF)" **(version 1.1.0). PhysioNet (2026). RRID:SCR_007345.**