Letter to Editor

Eurasian Journal of Critical Care

Beyond the AUROC: Data Leakage, Random Splitting, and Neglected Calibration in Critical Care Machine Learning

Ömer Faruk Çakıroğlu[1]

1 Department of Emergency Medicine, Kartal Doktor Lutfi Kirdar City Hospital, Istanbul, Turkey

Dear Editor,

Machine learning (ML) is increasingly proposed for clinical risk prediction in emergency and critical care, including the triage of undifferentiated patients,[1] yet reported performance may be inflated by avoidable methodological weaknesses. A systematic review found that studies of supervised ML models were frequently poorly reported, methodologically weak, and at high risk of bias.[2] Overfitting, in which a model captures sample-specific noise rather than reproducible signal, is a recurrent cause of this over-optimism.[3,4] Of 606 COVID-19 prognostic models in a living systematic review, 545 were at high risk of bias, inadequate sample size being the commonest cause (n=408, 67%),[5] a concern reiterated in recent guidance on sample size for artificial intelligence-based models.[6] High reported discrimination should therefore be read with caution.

A frequent, under-recognised source of bias is data leakage. When class-imbalance correction, feature selection, or normalisation is applied to the whole dataset before cross-validation rather than within each fold, the held-out observations contaminate model development. In radiomics, oversampling before cross-validation produced a positive bias of up to 0.34 in the area under the receiver operating characteristic curve (AUROC), attributable purely to this leakage.[7] All preprocessing, feature selection, resampling, and hyperparameter tuning should therefore be confined to the training data and repeated within each resampling iteration, as the signalling questions of PROBAST+AI now make explicit.[3,4,8] If a hold-out set is nonetheless used, partitioning should be at the patient level, so that all records from one individual, such as repeated images or admissions, stay within a single partition; otherwise information leaks across sets and performance is again overestimated[7].

Randomly splitting a single dataset into training and test subsets does not constitute external validation, because both subsets share one data source, and it is statistically inefficient, particularly in small or moderate-sized datasets.[3,9] Splitting removes data from development, increases overfitting, and leaves a test set too small for reliable estimation; it can also be repeated until favourable results emerge, an approach analogous to p-hacking.[3] These problems are magnified in small samples, where predictions become unstable and the estimated risk for one individual may span almost the entire probability scale.[6] Where feasible, bootstrapping, with every modelling step repeated in each bootstrap sample, or properly nested repeated cross-validation uses the available data more efficiently.[3,4]

Evaluation based solely on AUROC or accuracy is inadequate. A model may discriminate well yet be badly miscalibrated, systematically overestimating or underestimating risk and, in critical care, prompting inappropriate escalation or withholding of care; calibration is nevertheless assessed far less often than discrimination.[10] The Hosmer–Lemeshow test is sensitive to arbitrary grouping and sample size and does not describe the extent or direction of miscalibration; calibration curves are therefore preferable.[10,11] Authors should instead report a flexible calibration curve, calibration-in-the-large, and the calibration slope.[10,11] The Brier score may be added as a mea-

Corresponding author: Ömer Faruk Çakıroğlu
e-mail: omerfaruk1812@icloud.com
• **Received:** 13.07.2026 • **Revision:** 06.08.2026 • **Accepted:** 06.08.2026
DOI: 10.55994/ejcc.1993408

Cite this article as: Ömer Faruk Çakıroğlu. Beyond the AUROC: Data Leakage, Random Splitting, and Neglected Calibration in Critical Care Machine Learning. Eurasian Journal of Critical Care. 2026;8(2): 113-114

ORCID:
Ömer Faruk Çakıroğlu: orcid.org/ 0000-0001-8254-6821

sure of overall prediction error, but because it captures discrimination and calibration jointly, it should not be interpreted as a pure measure of calibration.[4,12]

Accuracy is particularly misleading under class imbalance, which is common for outcomes such as in-hospital cardiac arrest. Although imbalance corrections are routinely imported into ML pipelines, evidence from standard and ridge logistic regression shows that random undersampling, random oversampling, and SMOTE can substantially worsen calibration without improving discrimination, thereby reducing clinical utility.[12] Clinical usefulness should additionally be assessed through decision curve analysis.[4,10]

Model complexity should be justified rather than assumed. A systematic review of direct comparisons found no consistent evidence that ML outperforms logistic regression for clinical prediction,[13] while flexible algorithms have greater sample-size requirements to achieve comparable stability and calibration.[6] Where strategies perform similarly, the simpler model is preferable at the bedside.[4] Comparisons must also be fair: setting a new model's internally validated performance against an established score's externally validated performance is inappropriate, and head-to-head comparison should be performed in a dataset independent of both development cohorts.[9]

We therefore urge editors and reviewers to promote consistent use of two complementary resources: TRIPOD+AI for transparent reporting,[14] and PROBAST+AI for assessing methodological quality, risk of bias, and applicability.[8] Claims of clinical readiness should be constrained until a model has been externally validated, ideally by independent investigators, and, where feasible, prospectively evaluated for clinical impact.[3,9] External validation has historically been uncommon: in a PubMed search covering 1961 to 2019, only 5% of prediction model studies mentioned it in the title or abstract.[9] One performed within the development programme should not preclude subsequent independent assessment.[3,9] Applying these standards would help ensure that reported performance reflects genuine clinical value for critically ill patients rather than an artefact of study design.

References

1. Miles J, Turner J, Jacques R, Williams J, Mason S. Using machine-learning risk prediction models to triage the acuity of undifferentiated patients entering the emergency care system: a systematic review. *Diagn Progn Res*. 2020;4:16.
2. Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ*. 2021;375:n2281.
3. Collins GS, Dhiman P, Ma J, Schlussel MM, Archer L, Van Calster B, et al. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ*. 2024;384:e074819.
4. Efthimiou O, Seo M, Chalkou K, Debray T, Egger M, Salanti G. Developing clinical prediction models: a step-by-step guide. *BMJ*. 2024;386:e078276.
5. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ*. 2020;369:m1328.
6. Riley RD, Ensor J, Snell KIE, Archer L, Whittle R, Dhiman P, et al. Importance of sample size on the quality and utility of AI-based prediction models for healthcare. *Lancet Digit Health*. 2025;7(6):100857.
7. Demircioğlu A. Applying oversampling before cross-validation will lead to high bias in radiomics. *Sci Rep*. 2024;14(1):11563.
8. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*. 2025;388:e082505.
9. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? *Clin Kidney J*. 2021;14(1):49-58.
10. 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.
11. Riley RD, Archer L, Snell KIE, Ensor J, Dhiman P, Martin GP, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *BMJ*. 2024;384:e074820.
12. van den Goorbergh R, van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *J Am Med Inform Assoc*. 2022;29(9):1525-1534.
13. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110:12-22.
14. 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.