

Reporting machine learning studies in clinical research: Practical guidelines for transparent and reproducible analysis

Łukasz Lewandowski^{A,D–F}, Kinga Gostomska-Pampuch^{D–F}

Department of Biochemistry and Immunochemistry, Faculty of Medicine, Wrocław Medical University, Poland

A – research concept and design; B – collection and/or assembly of data; C – data analysis and interpretation; D – writing the article; E – critical revision of the article; F – final approval of the article

Advances in Clinical and Experimental Medicine, ISSN 1899–5276 (print), ISSN 2451–2680 (online)

Adv Clin Exp Med. 2026

Address for correspondence

Kinga Gostomska-Pampuch
E-mail: kinga.gostomska-pampuch@umw.edu.pl

Funding sources

None declared

Conflict of interest

None declared

Received on May 1, 2026

Reviewed on June 24, 2026

Accepted on June 30, 2026

Published online on October 15, 2026

Abstract

The expanding use of machine learning (ML) in clinical research has highlighted persistent methodological weaknesses that undermine the validity and interpretability of reported findings. Common issues include data leakage, inappropriate validation strategies, unstable model performance, and overinterpretation of predictive results. Importantly, strong predictive accuracy does not necessarily translate into clinical utility or biological insight, and poorly validated models may create a misleading impression of robustness.

This editorial proposes a concise set of reporting guidelines to improve transparency, reproducibility, and the critical appraisal of ML-based analyses. The framework emphasizes key elements required for evaluation, including study objectives, data preprocessing, validation design, model development, performance assessment, and interpretation. Additional considerations address domain-specific challenges such as deep learning, missing data, and time-dependent structures.

Rather than limiting methodological flexibility, these guidelines establish a minimal reporting standard that enables readers and reviewers to assess the credibility and generalizability of findings. By promoting clarity and completeness in reporting, they aim to reduce bias, improve reproducibility, and support the responsible integration of ML methods into clinical research.

Key words: machine learning, guidelines, reproducibility of results

Cite as

Lewandowski Ł, Gostomska-Pampuch K. Reporting machine learning studies in clinical research: Practical guidelines for transparent and reproducible analysis [published online as ahead of print on October 15, 2026]. *Adv Clin Exp Med*. 2026. doi:10.17219/acem/225290

DOI

10.17219/acem/225290

Copyright

Copyright by Author(s)

This is an article distributed under the terms of the Creative Commons Attribution 3.0 Unported (CC BY 3.0) (<https://creativecommons.org/licenses/by/3.0/>)

Highlights

- Machine learning models in clinical research are frequently affected by methodological flaws that compromise validity.
- Data leakage and inappropriate validation strategies remain major sources of bias.
- Predictive performance does not imply clinical usefulness or biological insight.
- Transparent reporting is essential for reproducibility and critical appraisal.
- These guidelines define a minimal standard for evaluating ML-based analyses in clinical studies.

Introduction

With the increasing availability of clinical data and the growing use of machine learning (ML) methods, the correctness of analytical procedures and transparency of reporting have become critical. Machine learning methods can support prediction, pattern recognition, dimensionality reduction, and exploratory analyses. However, their growing popularity has also exposed recurring methodological weaknesses. Despite the widespread use of ML, numerous studies demonstrate deficiencies including improper validation, data leakage, unstable model performance, insufficient calibration assessment, and over-interpretation of model outputs.^{1,2}

These issues are not merely technical limitations. They directly affect the credibility of findings and their ability to generalize beyond the analyzed dataset. In clinical research, a model that performs well only because of flawed validation or hidden leakage does not offer useful prediction. It represents a methodological mirage: impressive from a distance, but disappearing when exposed to real-world data.

The problem is particularly important because ML outputs are often perceived as objective, sophisticated, or inherently superior to those of simpler statistical models. This perception can be misleading. A complex model can still be poorly validated, unstable, clinically irrelevant, or biologically uninterpretable. Conversely, a simpler model with transparent assumptions and appropriate validation may provide more reliable evidence.

The primary objective of these guidelines is to ensure that ML-based analyses are reported in a manner that allows independent verification and reproducibility. A sufficient standard of reporting requires that readers and reviewers can assess the validity of the analysis and, where necessary, reproduce the analytical process. These guidelines are not intended to replace methodological expertise. Rather, they define a minimal reporting standard required for transparent evaluation of ML-based studies.^{3,4}

The present recommendations should be interpreted as journal-specific methodological guidance and are intended to complement, rather than replace, established reporting guidelines appropriate for the study design (e.g., TRIPOD-AI, PROBAST-AI, CONSORT-AI, or STARD-AI,

where applicable). Authors should follow the relevant reporting framework and its AI/ML-specific extensions whenever appropriate.

Objectives

This editorial aims to define a practical and concise reporting framework for ML-based analyses in clinical and experimental research. The proposed guidelines focus on elements that are essential for critical appraisal: the purpose and scope of the model, data handling and preprocessing, validation strategy, model development, final model reporting, performance evaluation, model interpretation, deep learning-specific issues, missing data, time series, anomaly handling, and reproducibility.^{3,4}

The goal is not to restrict methodological innovation. Instead, it is to ensure that innovation is accompanied by enough transparency for readers to understand what was done, why it was done, and whether the conclusions are justified.

Defining the purpose and scope of the analysis

A fundamental starting point in any ML-based analysis is a clear definition of its purpose. Models may serve different roles, including unsupervised exploration, supervised exploratory analysis, or prediction. These objectives are not interchangeable and should not be treated as such.

For unsupervised methods, authors should clearly describe whether the aim was visualization, clustering, dimensionality reduction, or exploratory pattern detection. Such analyses may be useful for generating hypotheses or describing structure in the data, but their outputs should not be presented as predictive, causal, or directly generalizable findings.

Supervised exploratory analyses may identify associations within a specific dataset. However, dataset-specific associations should not be presented as generalizable discoveries unless appropriate validation supports that interpretation. Variables should only be interpreted as generalizable predictors when supported by adequate validation.

Even then, caution is required, as predictive performance alone does not guarantee stability, clinical relevance, or biological meaning.³

Data handling, preprocessing, and leakage

Data handling and preprocessing are among the most critical and error-prone stages of the ML pipeline. Authors should describe all preprocessing steps in sufficient detail, including transformations, scaling, encoding of categorical variables, imputation, feature selection, anomaly handling, and any other operation applied before modeling.

Improper data handling remains a major source of bias, particularly through data leakage. Data leakage can be defined as the unintended use of information not available at the time of prediction.⁵ It may occur in obvious forms, such as the direct inclusion of the outcome or outcome-derived variables, but also in subtler forms. Outcome leakage may arise when post-baseline variables, future measurements, treatment decisions, or features indirectly related to the outcome are included in the model without appropriate justification.^{1,5}

Leakage is dangerous because it can produce apparently excellent model performance while destroying real-world validity. A model may appear accurate not because it has learned a clinically useful pattern, but because it has been given information that would not be available in the intended prediction setting.

To prevent leakage, all preprocessing must be performed within training data during validation. This includes imputation, scaling, feature selection, anomaly detection, and model tuning. If preprocessing is performed before data splitting or using the entire dataset, information from the validation or test data can influence model development and inflate performance estimates.^{3,4}

The data splitting strategy must also be explicitly described. Authors should specify whether training, validation, and test sets were used, or whether cross-validation, nested validation, bootstrap validation, or external validation was applied. For classification tasks, stratification is recommended, especially when class imbalance is present.

Splitting must respect the structure of the data. Observations from the same unit, such as the same patient, visit series, family, center, hospital, or experimental batch, should not be distributed across training and evaluation datasets if this creates dependency. Overlap of related observations across splits can substantially inflate model performance and give a false impression of generalizability.⁴

Missing data

Handling of missing data is a critical component of ML-based analyses. Missingness is rarely neutral and may reflect underlying clinical processes, disease severity,

workflow decisions, access to care, or measurement practices. Therefore, missing data should be treated as part of the analytical problem rather than a technical nuisance.

Authors should report the extent and pattern of missing data and assess whether missingness may influence the results.³ Where possible, the likely mechanism for missing data should be discussed, including whether the missingness may be missing at random or not at random.

Fixed imputation methods, such as replacing missing values with a constant, should generally be avoided unless clearly justified. Data-driven imputation methods are preferred for missing-at-random data. Multiple imputation should be considered when appropriate, with practical implementations such as multiple imputation by chained equations (MICE) and with attention to the uncertainty introduced by the imputation process.⁶

Imputation must be performed within the training data during validation. Performing imputation on the full dataset before splitting can introduce leakage by allowing information from validation or test data to influence the imputed values. Results should be interpreted with consideration of imputation uncertainty.

Data cleaning and anomaly handling

Data cleaning and anomaly handling should be explicitly documented. Authors should describe anomaly detection methods and justify all data cleaning procedures. This includes decisions about outliers, implausible values, duplicated observations, batch effects, technical artifacts, and biologically extreme observations.

The handling of anomalies should be clearly stated. Authors should specify whether anomalies were removed, transformed, winsorized, imputed, corrected, or retained. These decisions may substantially influence model performance and interpretation.⁷

Anomaly detection procedures must also avoid leakage. If anomalies are identified using the entire dataset before validation, information from evaluation data may influence model development. As with imputation and feature selection, anomaly handling should be embedded within the training process when it affects model development.

Model development and final model reporting

Model development requires transparency in both conceptual and technical aspects. The choice of model should be justified in relation to the research question, dataset size, outcome type, predictor structure, and intended use. Authors should avoid presenting model choice as a purely computational decision detached from the clinical or biological problem.

Hyperparameters and the procedure used to select them must be reported. Feature selection methods should be clearly described, including whether selection was supervised or unsupervised, whether it was performed inside the validation loop, and whether the selected features were carried forward into the final model.

Consistency between feature selection and the final model is important. Feature selection based on one model class and final inference or prediction based on another model class can be problematic unless explicitly justified. For example, selecting variables using one algorithm and then interpreting them in a different final model may create a mismatch between the selection mechanism and the reported results.

Automated model selection procedures should not be used as a black box. If automated selection is applied, the full development procedure should be reported, including candidate models, tuning grids, evaluation criteria, and decision rules. Otherwise, the analytical pathway becomes opaque, and the risk of overfitting is difficult to assess.

A clear distinction must be maintained between model development, hyperparameter tuning, feature selection, validation, and final model reporting. Model evaluation must be based on data not used during model development. Performance metrics derived from the same data used for model training, tuning, or feature selection should not be presented as unbiased estimates of model performance.

Authors should specify whether the final model was refitted on the full dataset after validation. If the final model differs from the model evaluated during validation, this difference must be clearly described. The final set of variables, hyperparameters, preprocessing steps, and modeling procedures should be reported in sufficient detail.

Model evaluation and clinical usefulness

Model performance should be interpreted as an estimate rather than a fixed property. This distinction is essential in clinical datasets, especially when sample sizes are small or moderate. Performance estimates may be highly unstable and sensitive to the chosen data split, resampling strategy, or validation procedure.²

Single train-test splits are particularly vulnerable to instability. A favorable or unfavorable split may substantially affect the model's apparent performance. Therefore, single-split results should be interpreted with caution and, where feasible, supported by repeated resampling, cross-validation, bootstrap validation, nested validation, or external validation.

When class imbalance is present, it must be explicitly reported. Authors should describe its extent and any method used to address it, such as resampling, class weighting, threshold adjustment, or alternative performance metrics.

Accuracy alone is often misleading in imbalanced classification problems and should be interpreted with caution.

Whenever possible, ML models should be compared against simpler baseline models. A complex model is not automatically better because it is more sophisticated. If a simpler model demonstrates comparable performance, superior calibration, or greater interpretability, it should be preferred or presented alongside more complex models, as such information is clinically and scientifically relevant.

Model comparison should be based on common and clinically meaningful metrics. For classification, authors should report metrics such as area under the curve (AUC), area under the precision–recall curve (PR-AUC), sensitivity, specificity, balanced accuracy, and accuracy, with accuracy interpreted with caution when the class distribution is imbalanced. For regression, root mean square error (RMSE), mean absolute deviation (MAD), and other relevant error metrics should be reported.

Calibration should be assessed and reported, particularly when predicted probabilities are clinically meaningful. A model may discriminate well but still produce poorly calibrated risk estimates. When clinical use is implied, decision thresholds must be defined and justified. Thresholds should not be chosen post hoc only to make performance appear favorable. The stability of performance and feature importance should be assessed where feasible. This is especially important when models are developed in small datasets or when many predictors are considered. High predictive performance does not imply clinical usefulness.⁸ A model may perform well statistically but still fail to improve decision-making, workflow, patient outcomes, or biological understanding.

Prediction, explanation, and interpretation

Model interpretation requires particular caution. A key distinction in ML-based analyses is the distinction between prediction and explanation. A model can predict an outcome without revealing why the outcome occurs; predictive accuracy reflects associations captured by the model rather than underlying causal structure. High predictive accuracy does not imply that the model has identified meaningful biological mechanisms.

Feature importance measures, including SHapley Additive exPlanations (SHAP) and similar approaches, should be treated as exploratory interpretation tools. They should not be interpreted as evidence of causal, directional, or mechanistic effects.³ Importance reflects contribution to prediction within a given model and dataset; it does not establish that a feature causes the outcome or represents an independent biological pathway.

This distinction is often blurred in clinical ML studies. Predictive features are sometimes described as risk

factors, mechanisms, therapeutic targets, or biomarkers, despite insufficient evidence. Such interpretations should be avoided unless supported by an appropriate study design, causal framework, external validation, and domain-specific justification.

Interpretable models should be included where possible, particularly when the aim involves clinical understanding or decision support. For black-box models, limitations of interpretability should be clearly stated. Model performance does not validate biological mechanisms. It only indicates that, under the reported validation conditions, the model was able to predict the outcome to a certain extent.

Special considerations for deep learning

Deep learning models require additional transparency because of their flexibility, complexity, and high risk of overfitting.¹ Their use should be clearly justified in relation to dataset size, data type, and problem complexity. In small clinical datasets, deep learning may be inappropriate unless supported by transfer learning, strong regularization, adequate validation, or other safeguards.

Authors should describe the model architecture at a level sufficient for reproducibility. The training procedure should be reported, including data splitting, validation strategy, batch size, number of epochs, optimizer, learning rate, loss function, stopping criteria, and any tuning procedures.

Regularization methods should also be described, including dropout, weight decay, early stopping, data augmentation, or other techniques used to reduce overfitting.¹ Hyperparameter tuning and model selection must be performed without using test data. Final performance should be evaluated on data not used during training or model selection.

For image, signal, or sequence data, preprocessing steps and transformations applied to the input data must be described. This includes resizing, normalization, filtering, segmentation, augmentation, feature extraction, and any manual or automated preprocessing. Deep learning outputs should be interpreted with caution and not treated as evidence of causal relationships or biological mechanisms.

Time series and temporally dependent data

Time series and temporally dependent data require validation strategies that respect the temporal structure of the problem. Authors should describe relevant temporal components, including trends, seasonality, cyclic patterns, and time-related dependencies where applicable.

Random splitting of temporally ordered observations should be avoided unless explicitly justified. Such splitting may allow future information to influence model training and can lead to overly optimistic performance estimates. In prediction problems where the goal is to forecast future outcomes, validation should mimic the intended use case.

Appropriate approaches include temporal holdout validation, rolling-origin evaluation, and walk-forward validation. The key principle is simple: information from the future should not be used to predict earlier observations.

Reproducibility and transparency

Reproducibility is a fundamental requirement for scientific validity. In the context of ML, lack of transparency in preprocessing, model development, tuning, validation, or evaluation effectively prevents independent verification of results. Authors should report the software, programming language, and package versions used in the analysis. Code and preprocessing scripts should be provided where possible. Random seeds should be reported when relevant, although authors should also recognize that setting a seed does not by itself guarantee robustness. Data sources and access conditions should be clearly described. Repository links should be provided for public datasets. If code, data, trained models, or preprocessing scripts cannot be shared, authors should provide a clear reason. Transparent reporting does not require that every study use the same method. It requires that the method used be understandable, evaluable, and, where possible, reproducible.

Conclusions

Machine learning methods can contribute meaningfully to clinical and experimental research, but only when their development, validation, interpretation, and reporting are transparent. Apparent predictive performance is not enough. Without appropriate validation, leakage prevention, calibration assessment, careful handling of missing data, and reproducible reporting, ML models may produce results that are difficult to trust and impossible to reproduce.

These guidelines define a minimal reporting standard for ML-based analyses submitted to the journal. Their purpose is not to discourage methodological innovation, but to ensure that such innovation is accompanied by scientific accountability. Models should be evaluated not only by their complexity or the impressiveness of their performance, but also by whether their results are valid, reproducible, interpretable, and clinically meaningful.

A concise summary of the proposed reporting framework, including minimum reporting requirements, recommended good practices, and common methodological errors, is presented in Table 1.

Table 1. Checklist for reporting machine learning analyses

Reporting element	Minimum requirement	Recommended good practice	Common error
Purpose of analysis	Clearly define the clinical question, intended use, and whether the analysis is exploratory, predictive, supervised, or unsupervised.	Justify the clinical relevance of the task and specify the target population and decision context.	Treating exploratory pattern-finding as generalizable prediction or clinical inference.
Data splitting and validation	Describe the train/test split, cross-validation strategy, or external validation approach.	Use nested cross-validation or external validation where feasible.	Data leakage between training, validation, and test datasets.
Preprocessing	Report all preprocessing, normalization, transformation, and feature engineering procedures.	Perform preprocessing within each validation fold or training split.	Applying preprocessing to the full dataset before splitting.
Data structure	Ensure that splitting respects patient-, visit-, image-, specimen-, batch-, or center-level dependencies.	Assess clustering, batch effects, center effects, and repeated measurements where relevant.	Allowing related observations from the same patient or center to appear in both training and test sets.
Model development	Describe the model type, candidate predictors, feature selection strategy, and hyperparameter tuning.	Compare complex models against simpler baseline or clinically interpretable models.	Automated model selection without transparent reporting of the modeling process.
Model evaluation	Report metrics appropriate to the task, including discrimination, calibration, and clinical usefulness where relevant.	Assess uncertainty, model stability, and stability of selected predictors or feature importance.	Reporting only accuracy, area under the curve (AUC), or a single optimistic performance estimate.
Interpretation	Distinguish prediction, association, causality, and biological interpretation.	Use interpretable models or model explanation tools where appropriate, while acknowledging their limits.	Overinterpreting feature importance as causal or mechanistic evidence.
Class imbalance	Report outcome prevalence and any methods used to address imbalance.	Use appropriate resampling, weighting, threshold analysis, or clinically meaningful performance metrics.	Ignoring imbalance or relying on accuracy in rare-outcome settings.
Missing data	Report the amount, pattern, and handling of missing data.	Use appropriate imputation strategies and sensitivity analyses where feasible.	Fixed ad hoc imputation without reporting or assessing its impact.
Deep learning, where applicable	Describe architecture, input representation, training procedure, validation, and tuning strategy.	Report regularization, stopping criteria, augmentation, and computational reproducibility.	Using test data during architecture selection or hyperparameter tuning.
Time series, where applicable	Use temporally appropriate validation and report the temporal structure of the data.	Assess temporal drift, updating needs, and stability over time.	Random splitting of temporally ordered observations.
Reproducibility	Report software, package versions, random seeds, and data/code availability.	Provide analysis code, model specifications, or repository links when possible.	Insufficient methodological detail preventing replication or critical appraisal.

Use of AI and AI-assisted technologies

ChatGPT (OpenAI) was used to support language editing and drafting of the manuscript text. The authors reviewed, revised, and approved the final content and remain fully responsible for the accuracy, integrity, and interpretation of the manuscript.

ORCID iDs

Łukasz Lewandowski  <https://orcid.org/0000-0002-7624-631X>

Kinga Gostomska-Pampuch  <https://orcid.org/0000-0002-0771-3893>

References

- Roberts M, Driggs D, Thorpe M, et al; AIX-COVNET. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat Mach Intell.* 2021;3(3):199–217. doi:10.1038/s42256-021-00307-0
- Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. *BMJ.* 2020;369:m1328. doi:10.1136/bmj.m1328
- Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ.* 2024;385:e078378. doi:10.1136/bmj-2023-078378
- Moons KGM, Damen JAA, Kaul T, 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. doi:10.1136/bmj-2024-082505
- Kaufman S, Rosset S, Perlich C. Leakage in data mining: Formulation, detection, and avoidance. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Diego, USA: Association for Computing Machinery (ACM); 2011: 556–563. doi:10.1145/2020408.2020496
- Rubin DB. Multiple imputation after 18+ years. *J Am Stat Assoc.* 1996; 91(434):473–489. doi:10.1080/01621459.1996.10476908
- Nassif AB, Talib MA, Nasir Q, Dakalbab FM. Machine learning for anomaly detection: A systematic review. *IEEE Access.* 2021;9:78658–78700. doi:10.1109/ACCESS.2021.3083060
- Vasey B, Nagendran M, Campbell B, et al; DECIDE-AI expert group. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ.* 2022;377:e070904. doi:10.1136/bmj-2022-070904