International Journal of Medical Informatics 219 (2026) 106538

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Why almost all ML models for medicine are wrong-and what we need for evidence-based medical AI

Federico Cabitza[a, b, *], Giuseppe Jurman[c, b], Filippo Molinari[d], Riccardo Bellazzi[e]

[a] Department of Informatics, Systems and Communication, University of Milano-Bicocca, Viale Sarca 336, 20126, Milano, Italy
[b] Digital Health and Wellbeing Center, Fondazione Bruno Kessler (FBK), Via Sommarive 18, 38123, Trento, Italy
[c] Department of Biomedical Sciences, Humanitas University, Via Rita Levi Montalcini 4, 20072, Pieve Emanuele, Italy
[d] Department of Electronics and Telecommunications, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129, Torino, Italy
[e] Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Via Ferrata 5, 27100, Pavia, Italy

HIGHLIGHTS

- Most medical ML models rest on fragile evidence from noisy labels, poor thresholds, unstable metrics, and weak validation.
- Evidence-based medical AI demands rigorous ground truthing, calibration, uncertainty reporting, and external validation.
- Clinical utility assessment and post-deployment monitoring are essential but largely neglected in current ML pipelines.
- A call to researchers, reviewers, users, and vendors to debate and raise methodological standards in medical AI.

ARTICLE INFO

Keywords:
Evidence-based AI
Medical machine learning
Robustness
Calibration
External validation

ABSTRACT

Machine learning (ML) models are increasingly proposed to support clinical decision-making, yet their evidentiary basis remains weaker than their publication volume suggests. This editorial argues that the problem is not only translational, regulatory, or infrastructural, but methodological. Many medical ML pipelines rely on uncertain ground truths, optimize performance around clinically irrelevant thresholds, report unstable or prevalence-dependent metrics, neglect calibration and uncertainty, and lack rigorous external and temporal validation. These weaknesses produce optimistic estimates that do not reliably anticipate performance in heterogeneous clinical settings. We call for an evidence-based medical AI grounded in more reliable annotation practices, explicit modeling of uncertainty, clinically meaningful threshold selection, calibration and decision-utility analyses, robustness testing, external validation on independent datasets, and post-deployment monitoring. The editorial also invites authors, reviewers, users, and vendors to adopt stricter standards so that predictive models can become credible, accountable, and clinically useful tools in everyday practice, rather than merely publishable artifacts.

Over the last decade, thousands of Machine Learning (ML) models have been developed to support medical decision- making. Many of these studies have been published in high-impact journals, often accompanied by claims of unprecedented accuracy, and potential clinical utility [1,2]. Yet, very few of these models have been prospectively validated, externally replicated, or integrated into routine clinical workflows with measurable impact on patient outcomes, costs, or satisfaction [3,4]. The problem is not merely a matter of translation, regulation, or infrastructure. It lies at the heart of the current ML development pipeline itself, which is based on methodological assumptions that are rarely met in medicine. As a result, most ML models for medicine are methodologically fragile: they are trained and evaluated on unreliable ground truths, optimized on inappropriate decision thresholds, assessed with inadequate metrics, and validated on datasets that do not represent the diversity and uncertainty of real clinical settings. In short, they are often wrong in ways that matter, and their clinical application is therefore unreliable.

The first and perhaps most fundamental flaw concerns ground truthing. Supervised ML pipelines assume that correct and objective labels are available, against which models can be trained and predictions

* Corresponding author at: Department of Informatics, Systems and Communication, University of Milano-Bicocca, Viale Sarca 336, 20126, Milano, Italy.
Email address: federico.cabitza@unimib.it (F. Cabitza).

https://doi.org/10.1016/j.ijmedinf.2026.106538
Received 6 June 2026; Accepted 7 June 2026
Available online 11 June 2026

can be evaluated. This assumption is largely untenable in medicine [5,6]. Diagnostic categories are often fuzzy, operational, or provisional; clinical gold standards frequently rely on subjective expert interpretation; even laboratory and imaging findings, usually considered objective, are affected by pre-analytical, analytical, and peri-analytical variability [7,8]. Inter-observer variability among specialists is well documented across domains—from radiology to pathology to clinical examination—often reaching levels that would be unacceptable in other scientific fields [9,10]. Reliance on single-expert labels, or on retrospective diagnoses of uncertain accuracy, introduces irreducible noise and systematic biases into the training data. Addressing this requires deliberate methodological choices: involving multiple expert annotators with appropriate aggregation schemes, using post-mortem diagnoses when available, and explicitly modeling the intrinsic uncertainty of medical data instead of hiding it behind a veneer of categorical labels.

A second critical issue is threshold optimization and decision preferences. Current ML pipelines tend to optimize global metrics such as the Area Under the Receiver Operating Characteristic Curve (AUC), which implicitly assumes that all decision thresholds are equally relevant [11,12]. However, physicians never treat thresholds as neutral. Different clinical contexts reflect different trade-offs between false positives and false negatives, often in non-linear and context-dependent ways. For many tasks, clinicians operate within ranges of acceptable thresholds, reflecting perceptions of severity, risk, and downstream consequences rather than point estimates [13,14]. In other words, decision thresholds reflect values and have ethical implications [15]. Consequently, evaluating models on global AUC is misleading. Models should be selected based on local AUROC values or average net benefit scores computed within clinically meaningful threshold ranges. Performance reporting should also reflect these ranges, rather than abstract averages across all possible thresholds. Articles should explicitly state which clinical decision-making trade-offs are implicated by the use of the ML model, and how these trade-offs informed model development, threshold selection, and performance reporting.

Third, current evaluation practices rely heavily on threshold-dependent (e.g., sensitivity, specificity, balanced accuracy) and/or prevalence-dependent metrics (e.g., positive predictive value, accuracy, F1 score, MCC). These metrics are unstable: they can change dramatically across settings with different disease prevalence, case mix, or workflows [16]. For instance, accuracy can be artificially inflated under class imbalance, PPV can drop sharply in low-prevalence contexts, and MCC varies systematically with the base rate despite its reputation for robustness. Moreover, threshold-dependent metrics are often computed at an arbitrary cut-off of 0.5, which rarely corresponds to actual decision trade-offs. In addition, the prevalence in the evaluation set may not reflect the real-world event rate, or may even be unknown or unpredictable ex ante, limiting the practical relevance of these metrics for anticipating real-world performance. Reporting net benefit and accuracy at clinically relevant decision thresholds, and across credible context-specific prevalence rates derived from sensitivity and specificity estimates obtained from representative evaluation sets, would be an obvious and reasonable practice; yet it remains rarely adopted. Likewise, calibration—the agreement between predicted probabilities and observed event frequencies—constitutes a more robust and transportable property [17]. Calibration should not be assessed globally alone, but locally, around the decision thresholds that matter for clinical practice. Models should also be evaluated for clinical utility, through decision-analytic frameworks, and for robustness, i.e., their ability to maintain performance on naturally variable, heterogeneous, or out-of-distribution (OOD) data [18–20].

A fourth methodological weakness lies in the common practice of reporting performance metrics as point estimates only, thereby overlooking the statistical uncertainty associated with finite, and often limited, evaluation samples [21]. Reported values for accuracy, AUC, sensitivity, or calibration are usually presented as single numbers with excessive apparent precision, often to two or three decimal places, without confidence intervals or other measures of variability. Comparisons between models are then made on the basis of these point estimates, as if they were exact quantities rather than uncertain statistics derived from finite samples [22]. This practice is particularly problematic because many evaluation datasets are both clinically homogeneous, as discussed above, and too small to provide reliable estimates of model performance. Without appropriate uncertainty quantification, such as confidence or credible intervals and statistical tests for paired comparisons, claims of superiority between models are at best overstated and at worst misleading. Recognizing and transparently reporting the inherent uncertainty of performance estimates is essential to avoid spurious conclusions and to support evidence-based model selection.

Fifth, external validation is rarely conducted properly. Many studies test their models on held-out splits from the same population used for training, or on temporally separated but otherwise similar cohorts [23]. Such validation strategies provide little evidence about model performance in different hospitals, health systems, or populations. External validation—performed on datasets collected independently, under different conditions, ideally prospectively—is essential to assess generalizability and to avoid self-referential performance inflation [24,25]. Without rigorous external validation, reported metrics are best interpreted as upper bounds under idealized conditions, not as reliable estimates for deployment.

A further striking but often overlooked weakness is the disregard for the temporal dimension of clinical data and the inevitable performance degradation over time. Most models are developed and evaluated on historical datasets as if data distributions were static. In reality, clinical practice, diagnostic criteria, population characteristics, laboratory technologies, and disease prevalence all evolve. These temporal shifts, well documented in empirical studies [26], can rapidly erode model performance after deployment. Yet most development pipelines do not include temporal validation, prospective evaluation, or post-deployment monitoring, implicitly treating accuracy as an intrinsic and stable property of the model rather than a contingent and time-sensitive one. Incorporating temporal validation strategies and mechanisms for recalibration and continuous monitoring is both feasible and essential to produce models that remain reliable over time. In other words, Machine Learning Operations (MLOps) are an essential requirement for the successful deployment of any machine-learning model in real-world clinical settings [27].

Taken together, these flaws undermine the credibility, replicability, and practical relevance of a large share of the current literature on ML for medicine. They lead to optimistic but fragile findings, overfitted to methodological conventions rather than clinical needs. Worse, they waste time, data, and human effort, contributing to a literature that accumulates more claims than evidence.

A core but often underappreciated implication follows from the widely endorsed call to train models on real-world clinical data. If supervised ML is expected to make predictions from measurements and records as they are produced in routine practice—with all their operational constraints, heterogeneous workflows, and imperfect documentation—then the training signal should also be understood as intrinsically imperfect. In medicine, this imperfection is not an exception but a structural feature: measurements embed random and systematic errors; labels are frequently provisional, ambiguous, or contingent on evolving diagnostic criteria; and the same clinical state can legitimately be associated with different codes depending on context and expertise. Paradoxically, many current pipelines seek clinical realism in the inputs while implicitly assuming quasi-experimental cleanliness in the targets. This mismatch motivates a more explicit methodological research agenda in medical ML, including: (i) robust learning under noisy, incomplete, or corrupted data; (ii) learning under dataset shift, including distributional and covariate shift adaptation; (iii) uncertainty quantification and propagation in supervised and unsupervised learning; (iv)

learning from noisy, soft, probabilistic, or partially labeled data; (v) principled modeling of annotator disagreement and label aggregation; (vi) calibration, reliability analysis, and evaluation metrics aligned with clinical decision-making; and (vii) out-of-distribution detection and robust generalization. Without sustained progress on these fronts, real-world datasets risk becoming a rhetorical requirement rather than serving as a foundation for reliable, transportable, and clinically accountable evidence.

If medical AI is to deliver on its promises, we need to rethink the pipeline. This involves: (i) improving ground truthing through multiple expert annotations [10] and the explicit modeling of uncertainty [28]; (ii) aligning optimization objectives with clinically meaningful threshold ranges; (iii) shifting evaluation from threshold-dependent accuracy metrics to calibration, utility, reliability, and robustness [19,29]; (iv) explicitly reporting the decision-making trade-offs associated with the clinical application of the learned model, including their implications for threshold selection; (v) mandating rigorous external validation [24]; and (vi) transparently reporting all these aspects [30]. But this is not merely a technical upgrade to adopt. Scientific communities and their main venues of dissemination—especially leading journals in medical informatics and biomedical engineering—must assume responsibility for increasing methodological rigor. Specifically, we argue that top-tier journals should require stricter acceptance criteria, mandating external validation on adequately large independent datasets, robustness analyses across diverse clinical settings, and transparent reporting of calibration and uncertainty. At a minimum, studies should include external validation on sufficiently large independent datasets, robustness analyses across heterogeneous subgroups and settings, and transparent reporting of calibration and uncertainty [29]. These requirements would help shift the field from the proliferation of unverified claims to the consolidation of reliable evidence, enabling the scientific community to identify truly promising models worthy of further clinical investigation and adoption.

In short, we need to move from a publication-driven model development culture to an evidence-based medical AI [31], one that accumulates credible and actionable findings to inform design, regulation, and adoption decisions. This editorial is also intended as an invitation to research groups that share these concerns, in medical AI as well as in machine learning applied to other critical domains, to promote initiatives that raise awareness among colleagues, reviewers, users of predictive models, and institutional decision-makers; to encourage more rigorous development and validation practices; to demand stricter methodological standards in the review of specialized articles; and to request greater transparency from vendors that embed predictive models in decision-support tools. Only through explicit scholarly positions and candid debate can we improve the quality of our work and its impact on the everyday practices of professionals who make critical decisions, increasingly also in light of the outputs of AI-enabled systems. This is a necessary agenda to produce ML systems that actually improve outcomes, reduce costs, and increase the satisfaction of clinicians, patients, and health systems alike.

CRediT authorship contribution statement

Federico Cabitza: Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Giuseppe Jurman:** Methodology, Validation, Writing – review & editing. **Filippo Molinari:** Methodology, Validation, Writing – review & editing. **Riccardo Bellazzi:** Methodology, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] E.J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nat. Med. 25 (1) (2019) 44–56.

[2] M.P. Salinas, J. Sepúlveda, L. Hidalgo, D. Peirano, M. Morel, P. Uribe, V. Rotemberg, J. Briones, D. Mery, C. Navarrete-Dechent, A systematic review and meta-analysis of artificial intelligence versus clinicians for skin cancer diagnosis, NPJ Digit. Med. 7 (1) (2024) 125.

[3] M. Nagendran, Y. Chen, C.A. Lovejoy, A.C. Gordon, M. Komorowski, H. Harvey, E.J. Topol, J.P. Ioannidis, G.S. Collins, M. Maruthappu, Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies, BMJ 368 (2020) m689.

[4] D. Ben-Israel, W.B. Jacobs, S. Casha, S. Lang, W.H.A. Ryu, M. de Lotbiniere-Bassett, D.W. Cadotte, The impact of machine learning on patient care: a systematic review, Artif. Intell. Med. 103 (2020) 101785.

[5] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, C. Ré, Hidden stratification causes clinically meaningful failures in machine learning for medical imaging, in: Proceedings of the ACM Conference on Health, Inference, and Learning, 2020, pp. 151–159.

[6] F. Cabitza, A. Campagner, L.M. Sconfienza, As if sand were stone. New concepts and metrics to probe the ground on which to build trustable AI, BMC Med. Inform. Decis. Mak. 20 (1) (2020) 219.

[7] G. Lippi, A.-M. Simundic, The EFLM strategy for harmonization of the preanalytical phase. A joint recommendation by the EFLM working group for preanalytical phase (WG-PRE) and the latin American working group for preanalytical phase (WG-PRE-LATAM), Clin. Chem. Lab. Med. 55 (8) (2017) 1047–1051.

[8] A. Padoan, J. Cadamuro, G. Frans, F. Cabitza, A. Tolios, S. De Bruyne, W. van Doorn, J. Elias, Z. Debeljak, S.M. Perez, et al., Data flow in clinical laboratories: could metadata and peridata bridge the gap to new AI-based applications? An investigation on behalf of the European federation of clinical chemistry and laboratory medicine (EFLM) working group on artificial intelligence (WG-AI), Clin. Chem. Lab. Med. (CCLM) 63 (4) (2025) 684–691.

[9] J.G. Elmore, G.M. Longton, P.A. Carney, B.M. Geller, T. Onega, A.N. Tosteson, H.D. Nelson, M.S. Pepe, K.H. Allison, S.J. Schnitt, et al., Diagnostic concordance among pathologists interpreting breast biopsy specimens, JAMA 313 (11) (2015) 1122–1132.

[10] F. Cabitza, A. Campagner, D. Albano, A. Aliprandi, A. Bruno, V. Chianca, A. Corazza, F. Di Pietto, A. Gambino, S. Gitto, et al., The elephant in the machine: proposing a new metric of data reliability and its application to a medical case to assess classification reliability, Appl. Sci. 10 (11) (2020) 4014.

[11] A.J. Vickers, B. Van Calster, E.W. Steyerberg, Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests, BMJ 352 (2016) i6.

[12] S. Halligan, D.G. Altman, S. Mallett, Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach, European radiology 25 (4) (2015) 932–939.

[13] S.G. Pauker, J.P. Kassirer, The threshold approach to clinical decision making, New England Journal of Medicine 302 (20) (1980) 1109–1117.

[14] B. Van Calster, L. Wynants, J.F. Verbeek, J.Y. Verbakel, E. Christodoulou, A.J. Vickers, M.J. Roobol, E.W. Steyerberg, Reporting and interpreting decision curve analysis: a guide for investigators, Eur. Urol. 74 (6) (2018) 796–804.

[15] I.S. Kohane, A.K. Manrai, The missing value of medical artificial intelligence, Nat. Med. 12 (10) (2025) 3962–3963, https://doi.org/10.1038/s41591-025-04050-6

[16] M.M. Leeflang, A.W. Rutjes, J.B. Reitsma, L. Hooft, P.M. Bossuyt, Variation of a test's sensitivity and specificity with disease prevalence, CMAJ 185 (11) (2013) E537–E544.

[17] E.W. Steyerberg, Y. Vergouwe, Towards better clinical prediction models: seven steps for development and an ABCD for validation, Eur. Heart J. 35 (29) (2014) 1925–1931.

[18] A.J. Vickers, E.B. Elkin, Decision curve analysis: a novel method for evaluating prediction models, Med. Decis. Mak. 26 (6) (2006) 565–574.

[19] G. Nicora, M. Rios, A. Abu-Hanna, R. Bellazzi, Evaluating pointwise reliability of machine learning prediction, J. Biomed. Inform. (2022), https://doi.org/10.1016/j.jbi.2022.103996

[20] L. Peracchio, G. Nicora, E. Parimbelli, T.M. Buonocore, E. Tavazzi, R. Bergamaschi, A. Dagliati, R. Bellazzi, RelAI: an automated approach to judge pointwise ML prediction reliability, Int. J. Med. Inform. 197 (2025) 105857.

[21] M. Pavlou, G. Ambler, C. Qu, S.R. Seaman, I.R. White, R.Z. Omar, An evaluation of sample size requirements for developing risk prediction models with binary outcomes, BMC Med. Res. Methodol. 24 (1) (2024) 146.

[22] G.S. Collins, J.B. Reitsma, D.G. Altman, K.G. Moons, Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement, BMJ 350 (2015) g7594.

[23] L. Marconi, F. Cabitza, Show and tell: a critical review on robustness and uncertainty for a more responsible medical AI, Int. J. Med. Inform. (2025) 105970.

[24] F. Cabitza, A. Campagner, F. Soares, L.G. de Guadiana-Romualdo, F. Challa, A. Sulejmani, M. Seghezzi, A. Carobene, The importance of being external. Methodological insights for the external validation of machine learning models in medicine, Comput. Methods Programs Biomed. 208 (2021) 106288.

[25] L. Wynants, B. Van Calster, G.S. Collins, R.D. Riley, G. Heinze, E. Schuit, M.M. Bonten, D.L. Dahly, J.A. Damen, T.P. Debray, et al., Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal, BMJ 369 (2020) m1328.

[26] S.E. Davis, T.A. Lasko, G. Chen, E.D. Siew, M.E. Matheny, Calibration drift in regression and machine learning models for acute kidney injury, J. Am. Med. Inform. Assoc. 24 (6) (2017) 1052–1061.
[27] Y. Li, J. Tian, A. Xu, R. Greiner, J. Hayward, R. Greenshaw, B. Cao, Maturity framework for operationalizing machine learning applications in health care: scoping review, J. Med. Internet Res. (2025), https://doi.org/10.2196/66559
[28] F. Cabitza, A. Campagner, V. Basile, Toward a perspectivist turn in ground truthing for predictive computing, in: Proceedings of the AAAI Conference on Artificial Intelligence, 37, 2023, pp. 6860–6868.
[29] F. Cabitza, A. Campagner, The need to separate the wheat from the chaff in medical informatics: introducing a comprehensive checklist for the (self)-assessment of medical AI studies, Int. J. Med. Inform. 153 (2021) 104510.
[30] J.F. Cohen, P.M. Bossuyt, TRIPOD + AI: an updated reporting guideline for clinical prediction models, BMJ 385 (2024).
[31] L. Famiglini, A. Campagner, M. Barandas, G. La Maida, E. Gallazzi, F. Cabitza, et al., Evidence-based XAI: an empirical approach to design more effective and explainable decision support systems, Comput. Biol. Med. 170 (March 2024) (2024).