International Journal of Medical Informatics 202 (2025) 105970

Contents lists available at ScienceDirect

International Journal of Medical Informatics

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

Show and tell: A critical review on robustness and uncertainty for a more responsible medical AI

Luca Marconi [a,*,1], Federico Cabitza [a,b,1]

[a] Department of Informatics, Systems and Communication, University of Milano-Bicocca, Viale Sarca 336, 20126, Milano, Italy
[b] IRCCS Ospedale Galeazzi-Sant'Ambrogio, Via Cristina Belgioioso, 173, Milano, Italy

ARTICLE INFO

Keywords:
External validation
Uncertainty quantification
Uncertainty evaluation
Reproducibility
Robustness
Machine learning
Medical informatics

ABSTRACT

This critical review explores two interrelated trends: the rapid increase in studies on machine learning (ML) applications within health informatics and the growing concerns about the reproducibility of these applications across different healthcare settings. Addressing these concerns necessitates acknowledging the uncertainty inherent in evaluating medical decision support systems. Therefore, we emphasize the importance of external validation and robustness assessment of the underlying ML models to better estimate their performance across diverse real-world scenarios.

To raise awareness among health practitioners and ML researchers, we advocate for the widespread adoption of external validation practices and uncertainty quantification techniques. Our survey of specialized literature reveals that fewer than 4% of studies published in high-impact medical informatics journals over the past 13 years have validated their systems using data from settings different from those that provided the training data. This low percentage is incompatible with responsible research, given the potential risks posed by unreliable ML models in healthcare.

Raising the standards for medical AI evaluation is crucial to improving practitioners' understanding of the potential and limitations of decision support systems in real-world settings. It is essential that uncertainty is not hidden in studies aimed at advancing knowledge in this field.

1. Introduction

The integration of machine learning (ML) into medical informatics has created transformative opportunities for clinical decision-making. However, it also presents significant challenges related to reproducibility and robustness across diverse healthcare environments [1,2]. Central to these challenges is the inherent uncertainty in medical decision support systems, as well as the critical need for external validation to ensure reliable model performance in real-world settings. These factors are essential for addressing the reproducibility crisis in medical AI, wherein many systems fail to generalize beyond their original development contexts [3,4].

Uncertainty, a fundamental concept in medical AI, arises from intrinsic data variability (aleatoric uncertainty) and gaps in model knowledge (epistemic uncertainty) [5]. Effectively quantifying and managing uncertainty is essential for enhancing model reliability and safety, particularly in high-stakes clinical decision-making [6]. External validation—by testing models on data from different institutions, regions, or timeframes—provides critical evidence of generalizability and ensures that models are not overfitted to their development environments [7].

Despite widespread recognition of these best practices, our review identifies significant gaps in their implementation. While prior research emphasizes the role of external validation in demonstrating model generalizability, it remains underutilized within the broader field of medical informatics. Fewer than 4% of ML studies published in high-impact medical informatics journals conduct external validation, and an even smaller proportion incorporate uncertainty quantification. This lack of rigor underscores the urgent need for improved standards in the development and evaluation of medical AI.

To address these challenges, this critical review synthesizes existing literature on external validation and uncertainty in medical AI, emphasizing their implications for medical AI practices. By analyzing methodological trends in high-impact studies, we identify significant gaps and misconceptions that impede the adoption of rigorous evaluation standards. Our findings highlight the urgent need for comprehen-

* Corresponding author.
E-mail address: luca.marconi@unimib.it (L. Marconi).
[1] Authors contributed equally to this work.

https://doi.org/10.1016/j.ijmedinf.2025.105970
Received 22 May 2024; Received in revised form 8 May 2025; Accepted 14 May 2025
Available online 19 May 2025

sive approaches that incorporate diverse datasets and robust statistical techniques to quantify uncertainty and enhance model reliability.

In this review, we aim to bridge the gap between current practices and ideal standards in medical AI evaluation. We propose actionable recommendations for researchers, practitioners, and journal editors. By establishing these practices as benchmarks for publication, we seek to elevate the quality and accountability of AI applications in healthcare, ensuring their safe and effective integration into clinical workflows.

2. Background and motivation

The application of decision aids in medicine presents significant challenges due to the inherent complexity and uncertainty involved. Medical phenomena, such as human phenotypes and adverse events, are influenced by multiple factors, many of which are either imprecisely measured or entirely unmeasurable. Ensuring the robustness, uncertainty quantification, and external validation of ML models in medical AI is essential to guarantee their reliability across diverse clinical settings. Without these safeguards, models that perform well in controlled environments may fail in real-world applications, potentially leading to harmful consequences in clinical decision-making.

This review critically examines the issues of reproducibility and robustness in medical AI, both of which are mandated by current regulatory frameworks. Reproducibility ensures research transparency by allowing results to be replicated across different settings, while robustness guarantees consistent AI system performance under varying conditions. The necessity of external validation is particularly emphasized, as it plays a fundamental role in assessing the generalizability of AI systems beyond the data used for model development.

Despite the importance of these factors, our review reveals a significant gap in their implementation within medical AI research. Studies rarely incorporate external validation or uncertainty quantification, raising concerns about the reliability and generalizability of reported results. To address this, our review adopts a rigorous query strategy, focusing exclusively on high-impact (Q1) medical informatics journals. While this approach may exclude some relevant contributions, it ensures a sharp focus on literature that meets the highest editorial standards, offering a refined perspective on external validation and uncertainty quantification.

To operationalize this perspective, we conducted a targeted search in the Scopus database in April 2024. Structured queries combined the terms *machine learning* or *deep learning* with *external validation uncertainty quantification uncertainty management* or *uncertainty evaluation*. Searches were restricted to nine journals ranked in the first quartile (Q1) of relevant categories, based on JCR and SJR rankings, and covered publications from 2010 to 2023.

For each journal, we executed four separate queries to identify: (i) all ML/DL-related articles; (ii) those mentioning external validation; (iii) those referring to uncertainty-related concepts; and (iv) those addressing both. This approach enabled us to estimate the relative prevalence of external validation and uncertainty quantification practices within top-tier medical informatics literature. Full query specifications are provided in the footnotes of the Findings section to ensure transparency and reproducibility.

We excluded articles from non-Q1 journals and those published outside the 2010–2023 timeframe. Although no formal filters were applied regarding document type (e.g., reviews, editorials, or conference proceedings), the corpus predominantly comprises peer-reviewed empirical studies. This likely reflects both the editorial standards of the selected journals and the practical orientation of the search terms.

Uncertainty in ML is commonly classified into aleatoric uncertainty, which arises from intrinsic variability in the data and cannot be reduced by acquiring more data, and epistemic uncertainty, which reflects a lack of knowledge about the model or dataset and can be mitigated through model refinement or data expansion [5]. In medical AI, effective uncertainty quantification is essential for supporting clinical decision-making [6], offering a structured means to assess and enhance model robustness.

Key sources of uncertainty include the quantity and heterogeneity of available data, which impact precision, as well as the complexity of mathematical models, which affects their capacity to accurately represent and extrapolate data. Surprisingly, despite its significance, uncertainty quantification remains underexplored in medical AI, particularly in prognostic and diagnostic applications. A structured representation of uncertainty, such as confidence intervals and conformal prediction sets, can enhance the interpretability and practical utility of decision-support systems.

While confidence intervals are increasingly used in medical research, their application in medical informatics remains limited, particularly in assessing performance metrics. Furthermore, these techniques primarily address variability due to data quantity but provide limited insights into data representativeness. Even when model performance is reported with accuracy measures and confidence intervals, assumptions about the representativeness of test data for future applications are often overly optimistic.

Among the most promising methodologies for addressing uncertainty, cautious classification and conformal prediction have demonstrated effectiveness in various clinical applications. Notable studies have applied these approaches to diagnostic pathology [8] and sepsis detection [9,10], highlighting their potential for enhancing robustness and reliability in medical AI.

To ensure that AI-driven medical decision-support systems are genuinely beneficial — improving error rates, resource utilization, and user satisfaction — robust evaluation practices must be implemented. Rather than concealing uncertainty from users, it should be quantified and transparently communicated.

This need is reinforced by two interrelated trends: the exponential growth in ML-based health informatics research (Fig. 1) and the widening "reproducibility crisis" in biomedical science [11], medical informatics [12], and AI-driven medical applications [13,14]. The latter issue underscores the risk that findings validated in one setting may not generalize to another. In a previous editorial [15], referencing Ioannidis [16], we provocatively argued that "most published accuracy scores are false" or, more bluntly, "most published studies applying ML techniques to medicine are simply not valid enough". Although we were not the first to make this claim (see e.g., [17,18]), we recognized it as the proverbial elephant in the room of medical informatics [19] that few were willing to acknowledge. Despite being cited nearly 150 times in the past three years, that editorial seems to have failed to bring about the necessary change in the status quo.

In this evolving landscape, uncertainty quantification in AI models is increasingly recognized as a critical factor. Key studies [20,21] have outlined key aspects of uncertainty in clinical AI. A recent review [22] offered an in-depth examination of the topic, covering several crucial aspects: the types of uncertainty, their implications for healthcare applications, and the benefits of uncertainty-aware models. However, the scarcity of studies that effectively integrate these methodologies underscores a significant research opportunity.

A major conceptual challenge lies in the differing definitions of validation across disciplines [23]. In medical research, particularly for medical devices, validation entails demonstrating that a system consistently meets intended-use requirements.[2] In contrast, ML practitioners often use "validation" to describe internal procedures such as hyperparameter tuning or comparisons based on models' performance on a subset of the available data (the so called validation data[3]) to identify the best one and report its performance. To clarify this distinction, we consider

[2] Artificial Intelligence Medical Devices (AIMD) Working Group of the International Medical Device Regulators Forum. Machine Learning-enabled Medical Devices: Key Terms and Definitions. IMDRF/AIMD WG/N67. 2022.

[3] Cf. definition 3.2.15 from the ISO/IEC FDIS 22989:2022.

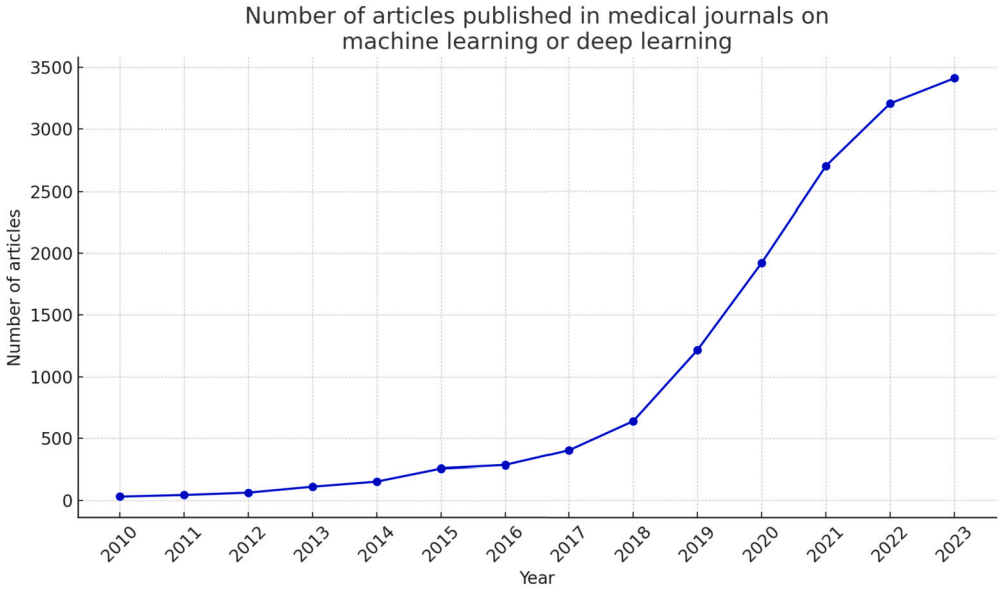

Fig. 1. Number of articles published in health-related journals on machine learning and deep learning over the last 13 years.
According to a search executed on 11/04/2024 on Scopus with the following query string: ((SRCTITLE (health) OR SRCTITLE (medical)) AND (TITLE-ABS-KEY ("machine learning") OR TITLE-ABS-KEY ("deep learning"))) AND (LIMIT-TO (DOCTYPE, "ar")) AND (LIMIT-TO (SUBJAREA, "MEDI") OR LIMIT-TO (SUBJAREA, "COMP") OR LIMIT-TO (SUBJAREA, "HEAL") OR LIMIT-TO (SUBJAREA, "ENGI")) AND (LIMIT-TO (PUBYEAR, 2023) OR LIMIT-TO (PUBYEAR, 2022) OR LIMIT-TO (PUBYEAR, 2021) OR LIMIT-TO (PUBYEAR, 2020) OR LIMIT-TO (PUBYEAR, 2019) OR LIMIT-TO (PUBYEAR, 2018) OR LIMIT-TO (PUBYEAR, 2017) OR LIMIT-TO (PUBYEAR, 2016) OR LIMIT-TO (PUBYEAR, 2015) OR LIMIT-TO (PUBYEAR, 2014) OR LIMIT-TO (PUBYEAR, 2013) OR LIMIT-TO (PUBYEAR, 2012) OR LIMIT-TO (PUBYEAR, 2011) OR LIMIT-TO (PUBYEAR, 2010)).

internal validation as model evaluation based on internal data (i.e. data used to develop and test the model), whereas external validation assesses model performance under varying conditions and on independent datasets.

True external validation involves providing statistically significant evidence — and indications of practical relevance — that a system will function reliably across anticipated scenarios within the scope of its *intended use*. This means the system should maintain similar accuracy when tested with naturally varied and distinct data, as opposed to data that has been artificially or maliciously altered, such as in adversarial attacks or other cybersecurity threats, compared to the data used for training and internal validation.

In the context of AI regulations, such as the European AI Act, validation primarily refers to evaluating trained AI systems to confirm expected performance before deployment. Here, our focus is on external validation, which requires independent test datasets from institutions distinct from those contributing to model training. This practice is essential for confirming real-world applicability and ensuring that AI-driven medical tools meet clinical standards.

Uncertainty quantification further enhances robustness by assessing the confidence level in collected data and model predictions. The ISO/IEC FDIS 22989:2022 standard defines robustness as "the ability [of a system] to maintain [its] level of performance, as intended by its developers, under any circumstances". Similarly, it refers to "the ability to perform comparably on atypical data, as opposed to the data expected in typical operations, or on inputs dissimilar to those on which it was trained" (see ISO/IEC TR 24029-1). Concept drift and label noise [7] can render internally validated models ineffective in real-world scenarios, highlighting the necessity of rigorous external testing.

A straightforward approach to achieving external validation involves testing models on data from multiple healthcare settings, such as different laboratories or hospitals [24], potentially spanning different regions or time periods. If performance metrics remain stable across internal and external data—analyzed via hypothesis testing (e.g., binomial tests for error rates, t-tests for regression metrics) or confidence intervals—then the model can be considered externally valid and robust. However, true reliability also necessitates that AI systems outperform unaided human decision-making within their intended clinical use.

In [7], we also noted that the similarity between internal and external data should be considered, with higher performance on more varied data indicating greater robustness. This robustness can also be assessed qualitatively using a diagram similar to the one we called External Performance Diagram (depicted in Fig. 2). Such a diagram can be currently generated with an online tool.[4] This diagram also considers other relevant quality dimensions, such as model calibration (the extent to which the model estimates probabilities correctly) and utility (the extent to which errors undermine the benefits of correct decisions). Although the scientific community currently regards evaluations based solely on error rates as decisive and informative, we believe that an evaluation aimed at determining the actual *value* of a predictive model (as any *validation* should, by definition) should not be limited to the discriminative dimension alone. Instead, it should assess the model's performance in a holistic and multidimensional manner [25], provided the evaluation is based on data that represents the intrinsic variability—population, biological, and instantial—of the medical phenotype under consideration.

3. Findings

Once we have clarified what we mean by robustness, uncertainty, and external validation, it becomes clear that this type of analysis is essential and should no longer be left to the discretion—and therefore the potential arbitrariness or thoroughness—of researchers developing ML models for prospective medical applications, nor to the idiosyncrasies and preferences of selected reviewers. What has this inaction led to?

To address this, let us consider how many articles in the medical informatics domain have included results from external validation and uncertainty quantification. The answer is surprisingly few. In the last 13 years, fewer than one in twenty-six articles (3.7%) have reported external validation results among those published in a selection of the most impactful journals in the medical informatics field (see Table 1).

[4] https://mudilab.github.io/dss-quality-assessment/.

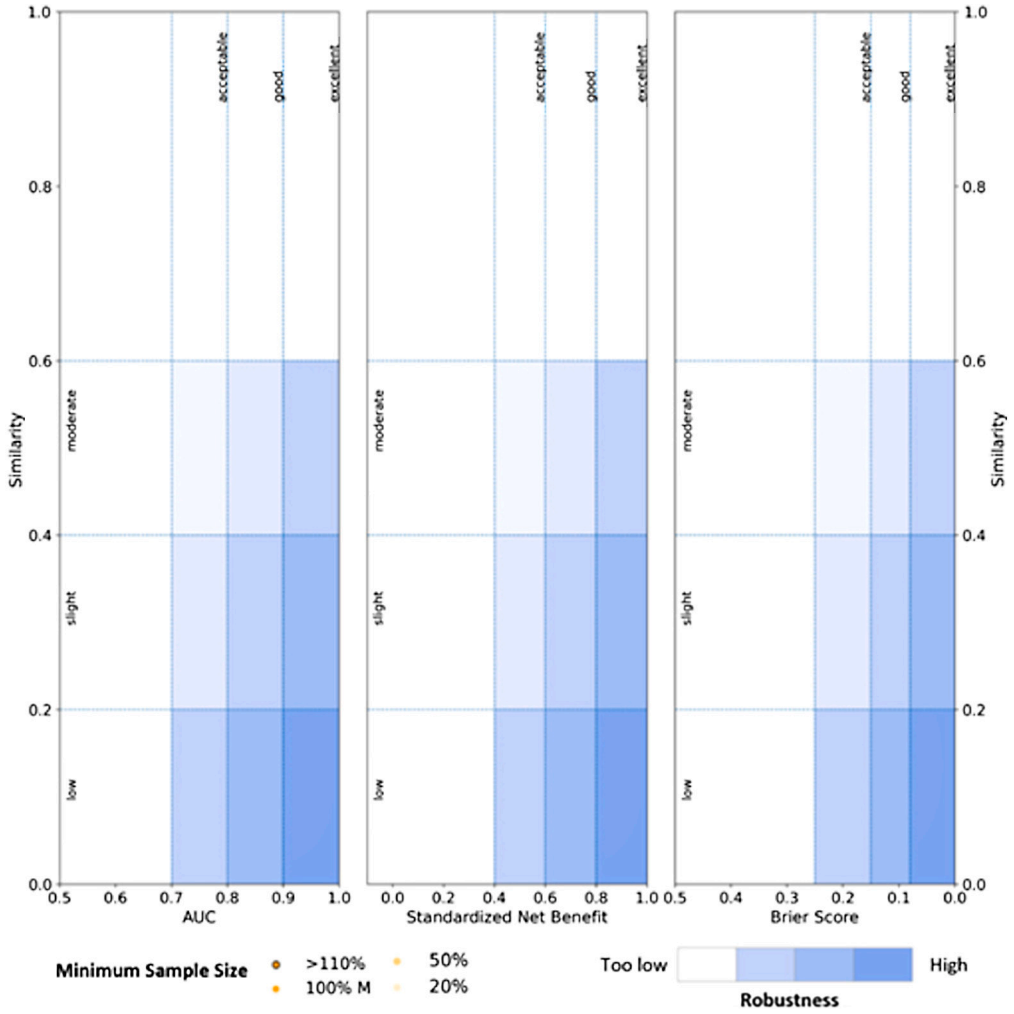

Fig. 2. An External Performance Diagram, as it was first proposed in [7]. In this kind of visualization, robustness is qualitatively rendered as a function of discriminative performance, as well as of calibration and utility scores. All dimensions are put in relation with the extent test and training datasets are similar, and with the sample size of the test sets.

Only four journals have a percentage higher than 5%, and in one journal, only 3 out of 322 papers report external validation (approximately 9‰). However, interest in external validation is undoubtedly increasing: of the 239 ML articles reporting external validation in medical settings, 232 were published in the last five years. Nevertheless, over the last 13 years, the prevalence of studies incorporating uncertainty quantification or uncertainty management in a selection of the most impactful journals in the medical informatics field remains remarkably low, with only 0.3% of ML articles addressing this aspect. In our search, the number of ML articles,[5] external validations,[6] uncertainty quantification (or uncertainty management or uncertainty evaluation) studies,[7] and studies including both external validations and uncertainty quantification (or management or evaluation)[8] were computed by querying Scopus. This query strategy was carefully designed to focus on high-impact journals (all Q1) in medical informatics, ensuring rigor and relevance. While this focus may inherently exclude certain methodologies or articles from other domains, it aligns with the study's objective of prioritizing robust and impactful literature.

When considering studies that address both external validation and uncertainty quantification or uncertainty management, the figure is even more strikingly low at just 0.01%, underscoring a critical gap in the robust application of ML in healthcare. Recent studies in our critical review have validated some key insights highlighted in a comprehensive review on uncertainty quantification in healthcare AI systems [22]. Specifically, the benefits of including uncertainty quantification in AI models were highlighted by improvements in diagnostic accuracy in our review findings [26,27], aligning with the emphasis on enhanced clinical decision-making. Additionally, our research corroborates the review's findings regarding the scarcity of models that effectively account for uncertainty.

[5] Number of articles retrieved on Scopus with the following query: (SRCTITLE("Journal Name") AND (TITLE-ABS-KEY("machine learning") OR TITLE-ABS-KEY("deep learning"))).

[6] Number of articles retrieved on Scopus with the following query: (SRCTITLE("Journal Name") AND (TITLE-ABS-KEY("machine learning") OR TITLE-ABS-KEY("deep learning")) AND (TITLE-ABS-KEY("external validation") OR TITLE-ABS-KEY("externally") OR TITLE-ABS-KEY("external dataset"))).

[7] (SRCTITLE("Journal Name") AND (TITLE-ABS-KEY("machine learning") OR TITLE-ABS-KEY("deep learning")) AND (TITLE-ABS-KEY("uncertainty quantification") OR TITLE-ABS-KEY("uncertainty management") OR TITLE-ABS-KEY("uncertainty evaluation"))).

[8] (SRCTITLE("Journal Name") AND ((TITLE-ABS-KEY("machine learning") OR TITLE-ABS-KEY("deep learning")) AND (TITLE-ABS-KEY("external validation") OR TITLE-ABS-KEY("externally") OR TITLE-ABS-KEY("external dataset")) AND (TITLE-ABS-KEY("uncertainty quantification") OR TITLE-ABS-KEY("uncertainty management") OR TITLE-ABS-KEY("uncertainty evaluation")))).

Table 1
Prevalence of (i) externally validated studies, (ii) uncertainty quantification or uncertainty management studies, and (iii) studies addressing both external validations and uncertainty quantification or uncertainty management in some of the top journals in medical informatics.

JCR Impact Factor 2023	SJR Quartile	Journal Name	ML articles	External Validations	Uncertainty Management or Uncertainty Evaluation	External Validations and Uncertainty Quantification (or Management or Evaluation)
30.8	Q1	Lancet Digital health	234	51 (21.8%)	0 (0%)	0 (0%)
15.2	Q1	NPJ Digital Medicine	364	28 (7.7%)	2 (0.5%)	0 (0%)
7.7	Q1	Computers in Biology and Medicine	2009	32 (1.6%)	9 (0.4%)	0 (0%)
7.4	Q1	Journal of Medical Internet Research	666	37 (5.5%)	0 (0%)	0 (0%)
6.4	Q1	Journal of the American Medical Informatics Association	494	23 (4.6%)	0 (0%)	0 (0%)
6.1	Q1	Computer Methods and Programs in Biomedicine	1278	25 (1.9%)	4 (0.3%)	0 (0%)
5.3	Q1	Journal of Medical Systems	322	3 (0.9%)	0 (0%)	0 (0%)
4.9	Q1	International Journal of Medical Informatics	324	25 (7.7%)	1 (0.3%)	1 (0.3%)
4.5	Q1	Journal of Biomedical Informatics	768	15 (1.9%)	1 (0.1%)	0 (0%)
			6459	239 (3.7%)	17 (0.3%)	1 (0.01%)

But should we care about this, and is it relevant? An increasing number of studies suggest that we should. For instance, the authors of [18,28,29] concluded that most ML models reported in the specialist literature on risk stratification, radiological discrimination, and cardiological tasks (respectively) perform poorly on external (i.e., real-world) data, or significantly worse than on internal data, potentially rendering them practically useless, if not harmful, to health practitioners and their patients. This conclusion has also been reported in other studies (e.g., [30,7,31]).

In what follows, we structured the findings section based on methodological trends identified in the reviewed literature. The categorization emerged organically from recurring approaches and practices observed across the studies. This organization avoids artificial separations and ensures that each category encapsulates distinct methodological principles while maintaining coherence with broader research themes.

3.1. Findings on external validation

Regarding external validation, several studies have utilized deep learning models for diagnostic purposes, underscoring the necessity of external validation. For example, Xue et al. (2023) employed multiple instance learning with two-stage attention for COPD identification using CT scans. Their model was validated using data from different clinical settings to ensure its robustness across various patient populations [32]. Similarly, Yoo et al. (2019) employed ML to identify candidates for corneal refractive surgery, validating their model with datasets from multiple sources to confirm its generalizability [33].

Risk prediction models have also highlighted the importance of external validation. Peng et al. (2020) used deep learning to predict the risk of late age-related macular degeneration, validating their model with external datasets to ensure its accuracy and applicability across different clinical environments [34]. Banda et al. (2019) applied ML to identify missed cases of familial hypercholesterolemia within health systems, demonstrating their model's effectiveness across various demographic groups by validating it with external data [35].

Clinical decision support systems (CDSS) represent another critical application area, with numerous studies focusing on their development and validation. Eng et al. (2021) developed an automated coronary calcium scoring system, validated across multiple centers, which underscored its reliability and generalizability [36]. Similarly, Shashikumar et al. (2021) created a sepsis prediction algorithm that indicated uncertainty in its predictions, validating it with data from different hospitals to ensure its applicability in varied clinical settings [10].

Methodological advancements also focused on enhancing model generalizability and robustness through external validation. Dou et al. (2021) employed federated learning to detect COVID-19 lung abnormalities from CT scans across multiple countries, ensuring data privacy and model generalizability [37]. Yang et al. (2022) used a 3D multi-scale residual fully convolutional neural network for large-sized kidney tumor segmentation, validating their model with diverse datasets to demonstrate its effectiveness in various clinical scenarios [38].

Integrating multiple data modalities to enhance diagnostic and predictive performance was another common theme, with external validation playing a crucial role. Chen et al. (2023) proposed an AI-enabled dynamic risk stratification model for emergency department patients, integrating ECG and CXR data, and validated the model using external datasets to ensure comprehensive patient assessments [39].

3.2. Findings on uncertainty quantification

Regarding uncertainty quantification, the studies collectively highlight its significance in enhancing the reliability and accuracy of medical diagnostic and prognostic tools.

Several studies employed Bayesian approaches to manage uncertainty in medical applications. Garifullin et al. (2021) used a deep Bayesian model to segment diabetic retinopathy lesions, improving lesion identification by incorporating uncertainty into the segmentation process [40]. Similarly, Abdar et al. (2021) applied Bayesian deep learning for skin cancer classification, using a three-way decision model that quantifies uncertainty to improve diagnostic accuracy [41]. Kabir et al. (2022) focused on aleatory uncertainty in transfer learning, employing Bayesian methods to enhance model generalizability in diagnostic tasks [42]. Goncalves et al. (2019) employed Bayesian multitask learning to analyze heterogeneous patient cohorts, demonstrating the method's ability to manage diverse clinical data with varying levels of uncertainty [43]. These studies share a common goal of improving model reliability by explicitly modeling uncertainty, enabling more informed clinical decisions. Similarly, recent approaches, such as Monte Carlo Dropout (MCD) [44] and Monte Carlo DropBlock Sampling (MCDbS) [44,45],

extend Bayesian methodologies to neural network models for segmentation and classification tasks. In cardiovascular modeling, Hamiltonian Monte Carlo (HMC) has been employed to quantify uncertainty in parameter estimation, providing robust posterior distributions in clinical decision-making [46].

Similarly, probabilistic approaches employ uncertainty quantification through the generation of probability distributions. Toledo-Cortés et al. (2022) employed probabilistic representations via density matrices to quantify uncertainty in ordinal regression tasks, such as diabetic retinopathy and prostate cancer grading [47]. By measuring the variance of predicted probability distributions, this approach provides an interpretable indicator of confidence, integrating uncertainty directly into the diagnostic workflow [47]. Jahmunah et al. (2023) DenseNet with Dirichlet priors have been used to quantify uncertainty in multi-class classification tasks [48]. In the context of myocardial infarction (MI) diagnosis using ECG signals, this approach effectively captured uncertainty [48]. Wang et al. (2023), introduced multi-tasking frameworks such as the use of alpha matte segmentation for boundary regions in medical images, where structural uncertainty is explicitly modeled [49]. The alpha matte approach leverages a probabilistic representation of transitional areas between tissues, capturing uncertainty in ambiguous or low-contrast regions [49].

Calibration was prominently featured in Buddenkotte et al. (2023), who used calibrated ensemble models for scalable uncertainty quantification in medical image segmentation [50]. This approach aggregates predictions from multiple models to provide a more robust and reliable estimate, accounting for uncertainty in the segmentation process. Similarly, Verhaeghe et al. (2023) developed calibrated ML models for atrial fibrillation risk prediction in ICU patients. They applied calibration methods to ensure that the predicted probabilities were well-aligned with actual event probabilities, which is crucial for making reliable clinical decisions. This calibration process adjusts the model outputs to accurately reflect the true uncertainty associated with the predictions [51]. Notably, this study is the only one identified that integrates both external validation and uncertainty quantification. This highlights a substantial open research area, underscoring the need for more studies that rigorously combine these critical validation strategies. This finding aligns with the overarching message of our review, which calls for increased attention to robustness and uncertainty management in the development of medical AI, ensuring reliable and responsible clinical applications.

Optimization-based methods also represent a growing field of interest. For instance, the Slime Mold Algorithm (SMA) has been utilized to quantify uncertainty in cardiovascular parameter estimation, demonstrating its ability to avoid local minima and provide significant solutions [52].

To acknowledge the importance of alternative methodologies, we highlight the potential of cautious classification and conformal prediction, as demonstrated in recent studies [8–10]. While these approaches were not captured by our review due to the specificity of the query design, they offer valuable contributions to uncertainty quantification, particularly in enhancing model interpretability and reliability in diverse clinical scenarios. This underscores the need for future studies to consider a broader methodological perspective, integrating these techniques to enrich the current understanding and practice of uncertainty quantification in medical AI.

Incorporating uncertainty quantification in medical applications has shown significant potential for improving the reliability, accuracy, and generalizability of diagnostic and prognostic tools. By explicitly modeling and managing uncertainty, these methods support more robust and informed clinical decision-making, ultimately enhancing patient outcomes.

4. Recommendations

To achieve reliable external validation and robust ML models in health informatics, we propose the following recommendations, addressing common misconceptions and informed by our insights. A detailed justification for each recommendation, supported by empirical evidence and prior research, is provided in Appendix A.1 and Appendix A.2.

1. Researchers should prioritize external validation using diverse datasets spanning multiple institutions, populations, and time periods to enhance model generalizability and robustness [34,35,37,10,53]
2. External validation should employ a multidimensional evaluation approach, considering multiple performance metrics, calibration, and clinical impact rather than relying on a single metric such as accuracy [37,10,50].
3. Meaningful external validation can still be achieved with limited data through well-curated small datasets and federated learning, which also mitigate privacy concerns [54–56].
4. Uncertainty quantification techniques should be integrated into validation processes to enhance trustworthiness and reliability in clinical applications [22,57,58].
5. Predictor selection should be guided by clinical relevance and external knowledge to ensure model robustness and adherence to data minimization principles [59–62].
6. Iterative external validation combined with uncertainty quantification strengthens model reliability, mitigates biases, and enhances generalizability across clinical settings [21,63,64].
7. Validation strategies should align with regulatory standards and incorporate real-world impact assessments to ensure AI models are clinically effective, ethically sound, and practically deployable [65–67].

Recent theoretical work has emphasized that methodological rigor in AI evaluation must be understood within a broader sociotechnical and normative framework. In particular, the notion of epistemic sustainability, introduced in [68], extends the idea of sustainability to include the long-term validity, interpretability, and trustworthiness of AI-generated knowledge. This perspective aligns closely with our focus on external validation and uncertainty quantification, highlighting that these are not merely technical refinements, but normative imperatives essential to maintaining public trust and the ethical deployment of AI in healthcare.

Complementing this, Shin's work [69] examines the politics of algorithmic trust and the centrality of epistemic uncertainty in human-AI systems. Shin argues that uncertainty should be surfaced, not hidden, to support user understanding and cognitive calibration—particularly in high-stakes contexts such as clinical decision-making. While our review underscores the scarcity of uncertainty quantification practices in medical AI research, grounding this critique in Shin's conceptualization of algorithmic opacity reinforces the idea that failing to quantify and communicate uncertainty meaningfully not only weakens model performance, but may also erode clinician confidence and compromise patient safety. Reframing uncertainty as an epistemological and ethical concern, rather than a purely technical one, strengthens the call for more responsible and transparent practices in AI development.

In light of this growing body of work and evidence, we concur with the authors in [70], who assert that most "models produce inaccurate predictions that, if applied in clinical settings, could lead to worse outcomes than simply an amusing and misplaced advertisement or finding the wrong movie to stream."

For all these reasons, we argue that the most effective way to contribute to the improvement of the scientific literature on ML, advance health informatics and enhance the safety of ML-based applications for clinical decision-making is for specialist journals to accept only studies

that include external validation and report all performance estimates with an uncertainty quantification technique, such as confidence intervals.

As observed in Table 1, journals with higher impact factors exhibit a clear trend of incorporating a greater proportion of studies that include external validation and uncertainty quantification. This trend highlights the association between rigorous methodological practices and the prestige of the publishing venue. These findings align with previous research emphasizing the necessity of robust evaluation methods to ensure the reliability and generalizability of ML models in medical applications.

Moreover, prior studies have underscored the importance of adhering to rigorous methodological standards in medical informatics. In this context, Cabitza and Campagner (2021) propose a practical checklist to help authors self-assess the quality of their contributions and assist reviewers in identifying and promoting high-quality medical ML studies. This checklist explicitly addresses key methodological aspects, such as external validation and uncertainty quantification, recognizing them as fundamental for distinguishing robust research from studies that merely apply ML techniques to medical data without sufficient methodological rigor [15].

We recognize that this decision may result in a significant reduction in the number of papers submitted to journals adopting such a policy, potentially affecting the journal's bibliometric indicators. However, vanity metrics should never interfere with a scientific journal's mission to disseminate only the best research and most reliable conclusions.

Furthermore, if leading journals adopt this policy, others may follow, recognizing the importance of what is at stake and the potential impact of quality research on healthcare professionals and patient health. A joint initiative of this kind would undoubtedly raise the quality standards for articles in medical informatics and send a strong signal to the broader Artificial Intelligence and ML community to renew their testing and validation practices.

This policy would require researchers in medical ML to take greater responsibility for their work and implement necessary actions to ensure the generalizability and robustness of the systems they develop and evaluate, taking into account the uncertainty affecting their estimates. We also hope that the public sharing of codes and training data, along with more complete descriptions of the ML training pipeline—as required by resources like the IJMEDI checklist for assessing the quality of ML articles[9]—will facilitate the replication of these pipelines by other researchers and support confirmatory studies [71,72]. This approach would ensure that evidence of robustness comes from independent researchers using data that truly represents the real challenges predictive models will face when (and if) they are implemented to deliver the promised benefits to users and patients.

One final note: The position expressed in this critical review is not intended as an end point, but rather as a reflection of what we believe to be an unprecedented phase of interest in a specific class of applications—classification-oriented decision support—and development techniques, including ML and deep learning, among healthcare professionals and practitioners. Rather, we see it as a necessary step in the continuous evolution of the medical informatics scientific community and all stakeholders. This involves raising awareness among study authors and readers about the need to base any conclusions regarding the value of a decision support system on robust evidence, collected empirically and evaluated using appropriate statistical techniques. Raising the bar will be a crucial step for more applications to achieve higher levels of maturity, such as advanced Technology Readiness Levels, and, once integrated into systems certified as medical devices, to provide tangible support to more physicians and patients. This will help us move closer to the ideal of a discipline grounded in solid and reliable evidence [73].

[9] https://zenodo.org/record/6451243.

5. Conclusions

This study provides a critical analysis of the state of external validation and uncertainty quantification in medical AI, addressing a significant gap in the current literature. While previous reviews in medical AI and health informatics have analyzed external validation [74,28] or uncertainty quantification [21,75] individually, to the best of our knowledge, no existing review has critically examined both dimensions together. By integrating these two crucial aspects, our work offers a unique contribution, bridging the divide between methodological rigor and clinical utility.

Our review highlights key limitations in prior research, particularly the lack of studies incorporating robust external validation and uncertainty quantification across diverse healthcare settings and datasets. Over the past 13 years, only 3.7% of studies published in leading medical informatics journals include external validation. More strikingly, a mere 0.01% of these studies combine external validation with uncertainty quantification or management, demonstrating a significant gap in the integration of these essential practices. This indicates that most research fails to address the variability inherent in real-world healthcare applications. Furthermore, while uncertainty quantification is widely recognized as crucial for evaluating the reliability of AI predictions, only 0.3% of studies explicitly address this aspect, underscoring a pervasive oversight in the methodological rigor of the field. These findings highlight a pressing need for more comprehensive and integrative approaches to validation and uncertainty management in medical AI research.

The contributions of this review are manifold. First, we offer a critical examination of the current state of external validation and uncertainty quantification in medical AI. Second, we offer practical recommendations for researchers, emphasizing the need for diverse and representative datasets, multidimensional evaluation metrics, and iterative validation processes. Third, we debunk prevalent misconceptions, such as the notion that external validation is prohibitively expensive or limited to large datasets, showcasing alternative approaches that balance rigor and feasibility.

By addressing these issues, our work sets a new standard for responsible research in medical AI. It underscores the necessity of aligning model development with real-world clinical needs, thereby fostering greater trust and utility among healthcare professionals and patients. We believe that adopting these practices will not only improve the reliability of AI systems but also pave the way for their broader adoption in clinical workflows, ultimately enhancing patient outcomes and safety. At the same time, we argue that external validation and uncertainty quantification are not merely methodological refinements but constitute epistemological and ethical imperatives. They contribute to the epistemic sustainability of clinical knowledge and are essential for maintaining trust in AI-enabled decision-making.

Future studies should build upon these insights by exploring innovative methods for uncertainty quantification and validating AI systems in increasingly complex and dynamic healthcare environments. By raising the bar for evaluation standards, we aim to contribute to a more robust, transparent, and impactful body of research in medical informatics.

Summary Table

What was already known on the topic

- External validation involves testing ML models with data from different settings to ensure generalizability and robustness.
- Uncertainty quantification in AI models helps in understanding the reliability of predictions, crucial for clinical decision-making.
- The reproducibility crisis in medical AI highlighted the need for external validation and uncertainty management in developing reliable ML models.
- Few studies in medical informatics perform external validation or incorporate uncertainty quantification, leading to concerns about the reliability of reported results.

What this study added to our knowledge

- This review provides a detailed analysis of the current state of external validation and uncertainty quantification in medical AI, highlighting significant gaps and areas for improvement.
- It debunks common myths about external validation and uncertainty quantification, clarifying misconceptions and emphasizing their importance in developing robust AI models.
- The study underscores the importance of diverse and representative datasets for external validation to improve the generalizability of ML models across different clinical settings.
- The review offers practical recommendations for researchers and practitioners to ensure transparency, reproducibility, and practical applicability of medical AI systems.

CRediT authorship contribution statement

Luca Marconi: Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Data curation. **Federico Cabitza:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

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 article.

Acknowledgements

The authors would like to express their sincere gratitude to Bastien Rance, Maxim Topaz, and Li Zhou, associated editors of the International Journal of Medical Informatics, for their constructive feedback, insightful comments and suggestions on an early draft, which significantly enhanced the quality and rigor of the work.

Appendix A. Justification of recommendations

To ensure that our recommendations are grounded in the findings of our review and relevant literature, we present the detailed justifications in the tables below. Each recommendation is supported by key findings or established research, as cited in the main text.

Table A.1
Comprehensive Justification of Recommendations - Part 1. This table presents the rationale behind each recommendation proposed in this study. Each justification is grounded in empirical findings from the literature or insights drawn from the review.

Recommendation No.	Justification
1	Our review highlights that fewer than 4% of studies use external validation datasets, with most relying solely on internal validation. This is a critical shortfall that undermines the reliability of ML models in medical informatics. Studies such as Peng et al. (2020) and Banda et al. (2019) demonstrate that testing models on datasets from different institutions, geographic locations, and demographic groups ensures generalizability and external validity [34,35]. Without diverse external datasets, ML models risk limited generalizability and potential overfitting to the development environment, as emphasized by Dou et al. (2021) [37]. Shashikumar et al. (2021) highlight the necessity of validating models across multiple settings and time periods to maintain performance reliability [10].
2	External validation should not focus solely on a single performance metric like accuracy but should cover multiple dimensions. Findings from Dou et al. (2021) emphasize the need for multidimensional evaluation strategies [37]. Shashikumar et al. (2021) highlighted that external validation using multiple metrics is essential to comprehensively assess the reliability of models in high-stakes applications such as sepsis prediction [10]. Their findings highlighted that relying on a single metric would not capture critical dimensions of model performance, including the ability to provide early and actionable predictions, reduce false alarms, and maintain high specificity across different datasets and temporal conditions [10]. Buddenkotte et al. (2023) further demonstrate that calibrated ensemble models enhance reliability by aggregating predictions from multiple models trained with different loss functions [50].
3	External validation is often perceived as expensive and logistically challenging, due to the need for large datasets [76,77]. However, studies suggest that meaningful insights can be derived from smaller sample sizes if selected and analyzed correctly. Collins et al. (2015) indicate that a minimum effective sample size of around 100 events can be sufficient for external validation [54]. Riley et al. (2021) provide frameworks to optimize sample sizes for meaningful validation results [55]. Decentralized training methods, such as federated learning, can further facilitate external validation without compromising privacy [56].

Table A.2
Comprehensive Justification of Recommendations - Part 2. This table continues the rationale behind each recommendation. Each justification is grounded in empirical findings from the literature or insights drawn from the review.

Recommendation No.	Justification
4	Uncertainty quantification in medical AI should be an integral part of validation to assess reliability in diagnoses and treatments [22]. Lue et al. (2022) illustrate how conformal prediction can effectively quantify uncertainty in medical imaging [57]. Amir et al. (2024) show that conformal prediction improves model trustworthiness by providing actionable insights through mean coverage errors [58].
5	Enhancing model robustness requires selecting predictors directly relevant to the target outcome. Steyerberg et al. (2018) and Sauerbrei et al. (2007) emphasize that predictor selection based on clinical relevance and external knowledge ensures validity and generalizability [59,60]. This aligns with data minimization principles, ensuring models remain actionable in clinical settings [61,62].
6	Integrating uncertainty quantification and external validation is essential for building robust models. Lambert et al. (2022) highlight that uncertainty quantification identifies unreliable predictions during training, reducing risks and preventing errors [21]. Qian et al. (2023) stress the importance of iterative external validation to confirm generalizability and address biases [63]. Birkenbihl et al. (2020) emphasize systematic validation in refining model development [64].
7	Aligning validation practices with regulatory standards and assessing clinical impact are critical for AI integration into healthcare. Meskó and Topol (2023) advocate for robust regulatory oversight to mitigate privacy risks [65]. Nagendran (2020) highlights the need for more prospective and real-world studies to assess AI's clinical impact [66]. Parikh et al. (2019) stress the necessity of meaningful clinical endpoints to demonstrate the real-world applicability of AI models [78]. Frameworks such as CLARITY AI provide structured approaches to integrate dimensions like ethical governance, data handling, and usability, ensuring AI systems remain scientifically robust, ethically sound, and practically relevant [67].

Appendix B. Additional information

This appendix provides additional details on key methodological aspects discussed in the main text. It is structured into three sections: (1) validation strategies for machine learning, which outlines different approaches to ensure model robustness and generalizability; (2) predictors and biases in medical AI, which examines the impact of biases on model performance and reproducibility; and (3) false myths in medical AI, which addresses common misconceptions regarding external validation and uncertainty quantification. These supplementary materials serve to enhance methodological transparency and support the rigorous evaluation of medical AI models.

B.1. Validation strategies for machine learning

Machine learning (ML) model validation is a crucial step in assessing robustness and generalizability, particularly in medical AI applications. Table B.1 outlines the four primary validation strategies: hyperparameter selection, internal validation with hold-out test data, external validation with test data from other facilities, and external validation with data coming from the same facility at different times.

While hyperparameter selection and internal validation are necessary steps in model development, they are not sufficient to guarantee the model's generalizability and reliability in practical, real-world settings. External validation is the gold standard for confirming a model's robustness and applicability across different contexts. Without external validation, there is a significant risk that a model may fail to generalize to new, unseen environments, thereby undermining its clinical effectiveness and safety. Therefore, external validation is indispensable for certifying an ML model's reliability and robustness in diverse clinical settings, ensuring it meets the high standards required for health technology assessment and medical device certification. Temporal external validation evaluates stability over time rather than across locations, providing a more comprehensive understanding of model reliability in dynamic healthcare environments.

Understanding the importance of internal and external validation requires recognizing their specific meanings and implementations within the ML and health informatics communities, as outlined in Table B.2.

By comparing these perspectives, we observe that while the principles of internal and external validation are similar, health informatics places greater emphasis on clinical utility, regulatory requirements, and real-world applicability. This distinction reflects the high stakes and critical importance of deploying ML models in healthcare settings.

B.2. Predictors and bias in medical AI

Developing robust and reliable ML models for medical AI requires careful selection of relevant predictors and consideration of associated biases. The biases inherent in the predictors used for making predictions have significant implications for data uncertainty, robustness assessment, and reproducibility. Robustness depends not only on data quantity but also on the selection of domain-specific predictors that are closely related to the target outcomes. For example, incorporating unrelated factors such as religion or education level into a predictive model for cancer could compromise its robustness. Therefore, enhancing robustness involves selecting predictors that are directly relevant to the medical condition being predicted, ensuring that the model remains reliable and valid across different settings. Table B.3 outlines the key types of biases and their characteristics, Tables B.4 and B.5 detail the impacts of these biases on robustness assessment and reproducibility, respectively, and Table B.6 presents strategies for effectively mitigating these biases.

Biases in predictors pose significant challenges in developing reliable and robust ML models for medical AI. Addressing these biases through careful predictor selection, comprehensive bias mitigation strategies, and continuous monitoring is essential. By implementing these practices, researchers and practitioners can enhance the robustness, fairness, and reproducibility of AI systems in healthcare, leading to more trustworthy and effective clinical applications.

Table B.1
Validation Strategies in Machine Learning.

Strategy	Concepts
Hyperparameter Selection	Involves tuning parameters that govern the training process of the model, typically using techniques like cross-validation. It prevents overfitting by ensuring the model is not excessively tailored to specific patterns within the training set.
Internal Validation	Involves splitting the dataset into separate training and hold-out test sets. The hold-out set is used exclusively for evaluating the model's performance after training, providing an unbiased assessment of the model's generalizability.
External Validation	Testing the model on data from entirely different facilities. This demonstrates the model's robustness and generalizability across different populations, clinical settings, and data collection processes.
Temporal External Validation	Testing models on data collected from the same facility but at different times, assessing robustness to temporal changes in data distribution.

Table B.2
Differences Between Machine Learning and Health Informatics Communities.

Aspect	Machine Learning Community	Health Informatics Community
Internal Validation	Focuses on model evaluation, hyperparameter tuning, and overfitting prevention within a single dataset.	Emphasizes clinical utility assessment, data integrity, and preliminary efficacy within a specific clinical dataset.
External Validation	Aims at testing generalizability, robustness, and bias identification using independent datasets.	Focuses on inter-institutional generalization, regulatory compliance, clinical impact, and reproducibility across diverse clinical environments.

Table B.3
Types of Biases and Their Characteristics.

Type of Bias	Characteristics
Selection Bias	Occurs when the training data is not representative of the target population, leading to poor generalization to broader populations.
Measurement Bias	Arises from inaccuracies in data collection processes, such as inconsistencies in how different hospitals record patient data, leading to biased predictions.
Confounding Bias	Occurs when an extraneous variable influences both the predictor and the outcome, resulting in spurious associations.
Sampling Bias	Involves the over- or under-representation of certain subpopulations within the training dataset, skewing the model's predictions towards overrepresented groups.
Algorithmic Bias	Reflects biases embedded in the algorithms, which can arise from model choices, parameter settings, or optimization criteria, potentially leading to unfair outcomes.

Table B.4
Impact of Biases on Robustness Assessment.

Aspect	Impact on Robustness Assessment
Generalizability	Models trained on biased datasets may fail to generalize to new, unseen data. For example, a model trained in one region may not perform well in another with different demographics or conditions.
Fairness	Biases can lead to unfair treatment of certain groups. In healthcare, this can result in disparities in the quality of care received by different populations. For example, an ML model biased towards younger patients may underpredict the risk of diseases in older adults.
Performance Metrics	The evaluation metrics themselves can be biased. A model might show high accuracy in the training environment but fail when applied to a more diverse test set.

Table B.5
Impact of Biases on Reproducibility.

Aspect	Impact on Reproducibility
Data Heterogeneity	Differences in data collection methods, patient demographics, and healthcare practices across institutions can lead to variability in model performance. A model trained on one dataset may not reproduce the same results on another due to these differences.
Model Interpretability	Biases can obscure the underlying reasons for a model's predictions, making it harder to interpret and replicate the results in different settings.
Validation Protocols	BBiases can influence the validation protocols used during model development. For instance, if the validation set is not representative of the broader population, the reported performance metrics may not be reproducible in other settings.

Table B.6
Mitigation Strategies for Biases.

Mitigation Strategy	Details
Diverse and Representative Data	Ensuring that training data covers various demographic groups and clinical conditions to mitigate biases.
Bias Detection and Correction	Implementing techniques such as re-weighting samples, fairness-aware algorithms, and subgroup analyses to detect and correct biases in data and models.
Transparent Reporting	Clearly reporting the characteristics of training and validation datasets, along with performance metrics for different subgroups, to improve understanding.
Continuous Monitoring	Monitoring model performance post-deployment to identify and address emerging biases over time, ensuring ongoing reliability and fairness.

B.3. False myths

Common misconceptions regarding external validation and uncertainty quantification can hinder the adoption of best practices in medical AI. Tables B.7 and B.8 debunk these myths, providing accurate clarifications.

By addressing these misconceptions, we can improve the understanding and implementation of external validation and uncertainty quantification, resulting in more reliable, generalizable, and trustworthy ML models in health informatics.

References

[1] F.R. Kolbinger, G.P. Veldhuizen, J. Zhu, D. Truhn, J.N. Kather, Reporting guidelines in medical artificial intelligence: a systematic review and meta-analysis, Commun. Med. 4 (1) (2024) 71, https://doi.org/10.1038/s43856-024-00492-0.
[2] A. Qayyum, J. Qadir, M. Bilal, A. Al-Fuqaha, Secure and robust machine learning for healthcare: a survey, IEEE Rev. Biomed. Eng. 14 (2021) 156–180, https://doi.org/10.1109/RBME.2020.3013489.
[3] A.L. Beam, A.K. Manrai, M. Ghassemi, Challenges to the reproducibility of machine learning models in health care, JAMA 323 (4) (2020) 305–306, https://doi.org/10.1001/jama.2019.20866.
[4] O. Ciobanu-Caraus, A. Aicher, J.M. Kernbach, L. Regli, C. Serra, V.E. Staartjes, A critical moment in machine learning in medicine: on reproducible and interpretable learning, Acta Neurochir. 166 (2024) 14, https://doi.org/10.1007/s00701-024-05892-8.
[5] E. Hüllermeier, W. Waegeman, Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods, Mach. Learn. 110 (2021) 457–506, https://doi.org/10.1007/s10994-021-05946-3.

Table B.7
False myths about external validation.

Misconception	Clarification
External validation requires large amounts of data, making it expensive	The necessary data size depends on the acceptable level of uncertainty. Even a small sample size, such as 100 cases, can provide meaningful insights. Some external validation is always better than none, as it offers critical information about the model's generalizability.
External validation must be performed in multiple settings	While validating in various settings can enhance robustness, it is not an absolute requirement. Validation in a single independent setting can still provide valuable insights into the model's performance and generalizability, as long as the setting is sufficiently different from the original development environment.
External validation requires prospective data	Retrospective data can also be used for external validation, as long as it is independent of the training data. This approach still allows for a valid assessment of the model's generalizability.
External validation focuses solely on accuracy	External validation evaluates multiple aspects, including evaluating the model's compliance with principles of data minimization and relevance, ensuring that the model is appropriate and lawful for its intended use.
Strong internal validation is sufficient	Internal validation alone does not guarantee generalizability to new, unseen data from different settings. External validation is essential for identifying potential biases and ensuring reliable real-world performance.
External validation is only necessary for final models	Early and iterative external validation can identify weaknesses and guide improvements throughout the development process, resulting in a more robust final model.
External validation does not need to consider data heterogeneity	External data should reflect the variability and diversity present in real-world applications to ensure the model's robustness across different subpopulations and clinical practices.

Table B.8
False myths about uncertainty quantification.

Misconception	Clarification
Uncertainty quantification is difficult to understand and confusing for users.	While some techniques are complex, methods like confidence intervals or conformal prediction sets can be intuitive once explained. Educating users about these concepts enhances understanding and supports better decision-making.
Uncertainty quantification is only useful when data is highly variable	Uncertainty quantification is valuable regardless of data variability, as it helps assess prediction reliability and sufficiency of data.
Uncertainty quantification is redundant if model accuracy is high.	Even highly accurate models benefit from uncertainty quantification by identifying cases where predictions may be uncertain.
Uncertainty quantification slows down decision-making	While it adds a step, uncertainty quantification ultimately supports more informed decisions, reducing the risk of costly errors.
Uncertainty quantification is limited to predictive models	Uncertainty quantification applies beyond predictive models, including classification, clustering, and regression.
Uncertainty quantification is only for advanced practitioners.	Some methods require expertise, but many are accessible with a basic understanding of statistics and ML.

[6] B. Kompa, J. Snoek, A.L. Beam, Second opinion needed: communicating uncertainty in medical machine learning, npj Digit. Med. 4 (1) (2021) 4, https://doi.org/10.1038/s41746-020-00367-3, https://www.nature.com/articles/s41746-020-00367-3.
[7] 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.
[8] H. Olsson, K. Kartasalo, N. Mulliqi, M. Capuccini, P. Ruusuvuori, H. Samaratunga, B. Delahunt, C. Lindskog, E.A.M. Janssen, A. Blilie, I.P.I.E. Panel, L. Egevad, O. Spjuth, M. Eklund, Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction, Nat. Commun. 13 (1) (2022) 7761, https://doi.org/10.1038/s41467-022-34945-8.
[9] A. Campagner, L. Agnello, A. Carobene, et al., Complete blood count and mdw-based machine learning algorithms for sepsis detection: a multicentric development and external validation study, J. Med. Internet Res. (2024), https://doi.org/10.2196/55492.
[10] S.P. Shashikumar, G. Wardi, A. Malhotra, S. Nemati, npj Digit. Med. 4 (1) (2021), https://doi.org/10.1038/s41746-021-00504-6, cited by: 38; All Open Access, Gold Open Access, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85115012762&doi=10.1038%2fs41746-021-00504-6&partnerID=40&md5=c891aecaf03516cf063a676e9d02a942.
[11] A. Stupple, D. Singerman, L.A. Celi, The reproducibility crisis in the age of digital medicine, npj Digit. Med. 2 (1) (2019) 2.
[12] E. Coiera, E. Ammenwerth, A. Georgiou, F. Magrabi, Does health informatics have a replication crisis?, J. Am. Med. Inform. Assoc. 25 (8) (2018) 963–968.
[13] M. Hutson, Artificial intelligence faces reproducibility crisis, Science 359 (6377) (2018) 725–726, https://doi.org/10.1126/science.359.6377.725.
[14] A.L. Beam, A.K. Manrai, M. Ghassemi, Challenges to the reproducibility of machine learning models in health care, JAMA 323 (4) (2020) 305–306.
[15] 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.
[16] J.P. Ioannidis, Why most published research findings are false, PLoS Med. 2 (8) (2005) e124.
[17] P. Tugwell, J.A. Knottnerus, Clinical prediction models are not being validated, J. Clin. Epidemiol. 68 (1) (2015) 1–2.
[18] G.C. Siontis, J.P. Ioannidis, Response to letter by forike et al.: more rigorous, not less, external validation is needed, J. Clin. Epidemiol. 69 (2016) 250–251.
[19] G.C. Siontis, I. Tzoulaki, P.J. Castaldi, J.P. Ioannidis, External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination, J. Clin. Epidemiol. 68 (1) (2015) 25–34.
[20] E. Begoli, T. Bhattacharya, D. Kusnezov, The need for uncertainty quantification in machine-assisted medical decision making, Nat. Mach. Intell. 1 (1) (2019) 20–23.
[21] B. Lambert, F. Forbes, S. Doyle, H. Dehaene, M. Dojat, Trustworthy clinical ai solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis, Artif. Intell. Med. 150 (2024) 102830, https://doi.org/10.1016/j.artmed.2024.102830, https://www.sciencedirect.com/science/article/pii/S0933365724000721.
[22] S. Seoni, V. Jahmunah, M. Salvi, P.D. Barua, F. Molinari, U.R. Acharya, Application of uncertainty quantification to artificial intelligence in healthcare: a review of last decade (2013–2023), Comput. Biol. Med. 165 (2023) 107441, https://doi.org/10.1016/j.compbiomed.2023.107441, https://www.sciencedirect.com/science/article/pii/S001048252300906X.
[23] E.W. Steyerberg, F.E. Harrell, Prediction models need appropriate internal, internal–external, and external validation, J. Clin. Epidemiol. 69 (2016) 245–247.
[24] T.P. Debray, Y. Vergouwe, H. Koffijberg, D. Nieboer, E.W. Steyerberg, K.G. Moons, A new framework to enhance the interpretation of external validation studies of clinical prediction models, J. Clin. Epidemiol. 68 (3) (2015) 279–289.
[25] F. Cabitza, J.-D. Zeitoun, The proof of the pudding: in praise of a culture of real-world validation for medical artificial intelligence, Ann. Transl. Med. 7 (8) (2019).

[26] D.Y. Kang, P.N. DeYoung, J. Tantiongloc, T.P. Coleman, R.L. Owens, Statistical uncertainty quantification to augment clinical decision support: a first implementation in sleep medicine, npj Digit. Med. 4 (1) (2021) 142.
[27] A. Peluso, I. Danciu, H.-J. Yoon, J.M. Yusof, T. Bhattacharya, A. Spannaus, N. Schaefferkoetter, E.B. Durbin, X.-C. Wu, A. Stroup, J. Doherty, S. Schwartz, C. Wiggins, L. Coyle, L. Penberthy, G.D. Tourassi, S. Gao, Deep learning uncertainty quantification for clinical text classification, J. Biomed. Inform. 149 (2024) 104576, https://doi.org/10.1016/j.jbi.2023.104576, https://www.sciencedirect.com/science/article/pii/S1532046423002976.
[28] A.C. Yu, B. Mohajer, J. Eng, External validation of deep learning algorithms for radiologic diagnosis: a systematic review, Radiol. Artif. Intell. 4 (3) (2022) e210064.
[29] G. Gulati, J. Upshaw, B.S. Wessler, R.J. Brazil, J. Nelson, D. van Klaveren, C.M. Lundquist, J.G. Park, H. McGinnes, E.W. Steyerberg, et al., Generalizability of cardiovascular disease clinical prediction models: 158 independent external validations of 104 unique models, Circ. Cardiovasc. Qual. Outcomes 15 (4) (2022) e008487.
[30] A. Wong, E. Otles, J.P. Donnelly, A. Krumm, J. McCullough, O. DeTroyer-Cooley, J. Pestrue, M. Phillips, J. Konye, C. Penoza, et al., External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients, JAMA Intern. Med. 181 (8) (2021) 1065–1070.
[31] J.S. Chen, A.S. Coyner, S. Ostmo, K. Sonmez, S. Bajimaya, E. Pradhan, N. Valikodath, E.D. Cole, T. Al-Khaled, R.P. Chan, et al., Deep learning for the diagnosis of stage in retinopathy of prematurity: accuracy and generalizability across populations and cameras, Ophthalmol. Retina 5 (10) (2021) 1027–1035.
[32] M. Xue, S. Jia, L. Chen, H. Huang, L. Yu, W. Zhu, Comput. Methods Programs Biomed. 230 (2023), https://doi.org/10.1016/j.cmpb.2023.107356, cited by: 4, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85146652766&doi=10.1016%2fj.cmpb.2023.107356&partnerID=40&md5=d69ace6a984484e1b443c42a5b602ef5.
[33] T.K. Yoo, I.H. Ryu, G. Lee, Y. Kim, J.K. Kim, I.S. Lee, J.S. Kim, T.H. Rim, npj Digit. Med. 2 (1) (2019), https://doi.org/10.1038/s41746-019-0135-8, cited by: 53; All Open Access, Gold Open Access, Green Open Access, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85089606293&doi=10.1038%2fs41746-019-0135-8&partnerID=40&md5=7065539720582bcbb6d3f06c717a298a.
[34] Y. Peng, T.D. Keenan, Q. Chen, E. Agrón, A. Allot, W.T. Wong, E.Y. Chew, Z. Lu, npj Digit. Med. 3 (1) (2020), https://doi.org/10.1038/s41746-020-00317-z, cited by: 36; All Open Access, Gold Open Access https://www.scopus.com/inward/record.uri?eid=2-s2.0-85089982425&doi=10.1038%2fs41746-020-00317-z&partnerID=40&md5=404ddb299514d0ca2a1f88cddba05408.
[35] J.M. Banda, A. Sarraju, F. Abbasi, J. Parizo, M. Pariani, H. Ison, E. Briskin, H. Wand, S. Dubois, K. Jung, S.A. Myers, D.J. Rader, J.B. Leader, M.F. Murray, K.D. Myers, K. Wilemon, N.H. Shah, J.W. Knowles, npj Digit. Med. 2 (1) (2019), https://doi.org/10.1038/s41746-019-0101-5, cited by: 66; All Open Access, Gold Open Access https://www.scopus.com/inward/record.uri?eid=2-s2.0-85089606094&doi=10.1038%2fs41746-019-0101-5&partnerID=40&md5=008efa3b187514531e01af7d7a89584d.
[36] D. Eng, C. Chute, N. Khandwala, P. Rajpurkar, J. Long, S. Shleifer, M.H. Khalaf, A.T. Sandhu, F. Rodriguez, D.J. Maron, S. Seyyedi, D. Marin, I. Golub, M. Budoff, F. Kitamura, M.S. Takahashi, R.W. Filice, R. Shah, J. Mongan, K. Kallianos, C.P. Langlotz, M.P. Lungren, A.Y. Ng, B.N. Patel, npj Digit. Med. 4 (1) (2021), https://doi.org/10.1038/s41746-021-00460-1, cited by: 70; All Open Access, Gold Open Access, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85107441829&doi=10.1038%2fs41746-021-00460-1&partnerID=40&md5=4c9ff88122e7d006b1bf16379795f26f.
[37] Q. Dou, T.Y. So, M. Jiang, Q. Liu, V. Vardhanabhuti, G. Kaissis, Z. Li, W. Si, H.H.C. Lee, K. Yu, Z. Feng, L. Dong, E. Burian, F. Jungmann, R. Braren, M. Makowski, B. Kainz, D. Rueckert, B. Glocker, S.C.H. Yu, P.A. Heng, npj Digit. Med. 4 (1) (2021), https://doi.org/10.1038/s41746-021-00431-6, cited by: 146; All Open Access, Gold Open Access, Green Open Access, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85103580182&doi=10.1038%2fs41746-021-00431-6&partnerID=40&md5=4d6804a7e9d4a890ea2671aae1d41d28.
[38] E. Yang, C.K. Kim, Y. Guan, B.-B. Koo, J.-H. Kim, Comput. Methods Programs Biomed. 215 (2022), https://doi.org/10.1016/j.cmpb.2022.106616, cited by: 15, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85122529377&doi=10.1016%2fj.cmpb.2022.106616&partnerID=40&md5=497fe1f709074dc666acecf9a24a9db7.
[39] Y.-H.J. Chen, C.-S. Lin, C. Lin, D.-J. Tsai, W.-H. Fang, C.-C. Lee, C.-H. Wang, S.-J. Chen, J. Med. Syst. 47 (1) (2023), https://doi.org/10.1007/s10916-023-01980-x, cited by: 2, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85166194545&doi=10.1007%2fs10916-023-01980-x&partnerID=40&md5=cc0674620dfb7423966035b39ea1c555.
[40] A. Garifullin, L. Lensu, H. Uusitalo, Comput. Biol. Med. 136 (2021), https://doi.org/10.1016/j.compbiomed.2021.104725, cited by: 33; All Open Access, Hybrid Gold Open Access, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85112340185&doi=10.1016%2fj.compbiomed.2021.104725&partnerID=40&md5=2f76a77da3c6380ab9ecf5d6ea0293fb.
[41] M. Abdar, M. Samami, S. Dehghani Mahmoodabad, T. Doan, B. Mazoure, R. Hashemifesharaki, L. Liu, A. Khosravi, U.R. Acharya, V. Makarenkov, S. Nahavandi, Comput. Biol. Med. 135 (2021), https://doi.org/10.1016/j.compbiomed.2021.104418, cited by: 136; All Open Access, Green Open Access, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85106889973&doi=10.1016%2fj.compbiomed.2021.104418&partnerID=40&md5=fe807bd2520050dcdbe20fde224f5ccb.
[42] H.M.D. Kabir, S. Khanam, F. Khozeimeh, A. Khosravi, S.K. Mondal, S. Nahavandi, U.R. Acharya, Comput. Biol. Med. 143 (2022), https://doi.org/10.1016/j.compbiomed.2022.105246, cited by: 11, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85124209734&doi=10.1016%2fj.compbiomed.2022.105246&partnerID=40&md5=1cfb8090b7bbf485f55e15acef323062.
[43] A. Goncalves, P. Ray, B. Soper, D. Widemann, M. Nygård, J.F. Nygård, A.P. Sales, J. Biomed. Inform. X 4 (2019), https://doi.org/10.1016/j.yjbinx.2019.100059, cited by: 5; All Open Access, Hybrid Gold Open Access, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85074886072&doi=10.1016%2fj.yjbinx.2019.100059&partnerID=40&md5=a43e2d6d1cea301aecb2b5913aba68ab.
[44] T. Zhou, S. Zhu, Comput. Biol. Med. 163 (2023), https://doi.org/10.1016/j.compbiomed.2023.107142, cited by: 4, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85162082101&doi=10.1016%2fj.compbiomed.2023.107142&partnerID=40&md5=4e767da745f6cbc37d62c01a7febacab.
[45] A. Pepe, J. Egger, M. Codari, M.J. Willemink, C. Gsaxner, J. Li, P.M. Roth, D. Schmalstieg, G. Mistelbauer, D. Fleischmann, Comput. Biol. Med. 165 (2023), https://doi.org/10.1016/j.compbiomed.2023.107365, cited by: 1; All Open Access, Green Open Access, Hybrid Gold Open Access, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85169039340&doi=10.1016%2fj.compbiomed.2023.107365&partnerID=40&md5=6db28c9b2c8c7805efd02e3bb4aef6d5.
[46] M. Salvador, F. Regazzoni, L. Dede', A. Quarteroni, Comput. Methods Programs Biomed. 231 (2023), https://doi.org/10.1016/j.cmpb.2023.107402, cited by: 9; All Open Access, Green Open Access, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85147608291&doi=10.1016%2fj.cmpb.2023.107402&partnerID=40&md5=fa97034608e50dc6779d7aa17aadf733.
[47] S. Toledo-Cortés, D.H. Useche, H. Müller, F.A. González, Comput. Biol. Med. 145 (2022), https://doi.org/10.1016/j.compbiomed.2022.105472, cited by: 14; All Open Access, Green Open Access, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85128165660&doi=10.1016%2fj.compbiomed.2022.105472&partnerID=40&md5=c5d6006248ef0fc3951892e3e53a231f.
[48] V. Jahmunah, E. Ng, R.-S. Tan, S.L. Oh, U.R. Acharya, Comput. Methods Programs Biomed. 229 (2023), https://doi.org/10.1016/j.cmpb.2022.107308, cited by: 30; All Open Access, Bronze Open Access, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85144366929&doi=10.1016%2fj.cmpb.2022.107308&partnerID=40&md5=d585b06c5b259f22b987a456bfc9480b.
[49] L. Wang, X. Ye, L. Ju, W. He, D. Zhang, X. Wang, Y. Huang, W. Feng, K. Song, Z. Ge, Comput. Biol. Med. 158 (2023), https://doi.org/10.1016/j.compbiomed.2023.106714, cited by: 4; All Open Access, Hybrid Gold Open Access, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85151283034&doi=10.1016%2fj.compbiomed.2023.106714&partnerID=40&md5=137ed1ad42cd9aafa07a7a35e51680eb.
[50] T. Buddenkotte, L. Escudero Sanchez, M. Crispin-Ortuzar, R. Woitek, C. McCague, J.D. Brenton, O. Öktem, E. Sala, L. Rundo, Comput. Biol. Med. 163 (2023), https://doi.org/10.1016/j.compbiomed.2023.107096, cited by: 11; All Open Access, Hybrid Gold Open Access, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85161555981&doi=10.1016%2fj.compbiomed.2023.107096&partnerID=40&md5=eabe85a8e260dcc217b67c642ac98636.
[51] J. Verhaeghe, T. De Corte, C.M. Sauer, T. Hendriks, O.W. Thijssens, F. Ongenae, P. Elbers, J. De Waele, S. Van Hoecke, Int. J. Med. Inform. 175 (2023), https://doi.org/10.1016/j.ijmedinf.2023.105086, cited by: 5; All Open Access, Hybrid Gold Open Access, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85156145834&doi=10.1016%2fj.ijmedinf.2023.105086&partnerID=40&md5=fa7b589275331518dc99c9cbd93de4ac.
[52] J. Yan, S. Cai, X. Cai, G. Zhu, W. Zhou, R. Guo, H. Yan, Y. Wang, Comput. Methods Programs Biomed. 240 (2023), https://doi.org/10.1016/j.cmpb.2023.107674, cited by: 2; All Open Access, Hybrid Gold Open Access, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85162173014&doi=10.1016%2fj.cmpb.2023.107674&partnerID=40&md5=7cdd6f8e7bfd8b0c3ceed23405eaa8d4.
[53] F. Pennestrì, F. Cabitza, N. Picerno, et al., Sharing reliable information worldwide: healthcare strategies based on artificial intelligence need external validation, Position paper, BMC Med. Inform. Decis. Mak. 25 (2025) 56, https://doi.org/10.1186/s12911-025-02883-2.
[54] G.S. Collins, E.O. Ogundimu, D.G. Altman, Sample size considerations for the external validation of a multivariable prognostic model: a resampling study, Stat. Med. 35 (2) (2016) 214–226, https://doi.org/10.1002/sim.6787, arXiv: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.6787, https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6787.
[55] R.D. Riley, T.P.A. Debray, G.S. Collins, L. Archer, J. Ensor, M. van Smeden, K.I.E. Snell, Minimum sample size for external validation of a clinical prediction model with a binary outcome, Stat. Med. 40 (19) (2021) 4230–4251, https://doi.org/10.1002/sim.9025, arXiv: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.9025, https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9025.
[56] Y. Kawamura, A. Vafaei Sadr, V. Abedi, R. Zand, Many models, little adoption—what accounts for low uptake of machine learning models for atrial fibrillation prediction and detection?, J. Clin. Med. 13 (5) (2024), https://doi.org/10.3390/jcm13051313, https://www.mdpi.com/2077-0383/13/5/1313.

[57] C. Lu, A.N. Angelopoulos, S. Pomerantz, Improving trustworthiness of ai disease severity rating in medical imaging with ordinal conformal prediction sets, in: L. Wang, Q. Dou, P.T. Fletcher, S. Speidel, S. Li (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2022, Springer Nature, Switzerland, Cham, 2022, pp. 545–554.
[58] A.M. Vahdani, S. Faghani, Deep conformal supervision: leveraging intermediate features for robust uncertainty quantification, J. Digit. Imag. (2024), https://doi.org/10.1007/s10278-024-01286-5.
[59] E.W. Steyerberg, H. Uno, J.P. Ioannidis, B. van Calster, C. Ukaegbu, T. Dhingra, S. Syngal, F. Kastrinos, Poor performance of clinical prediction models: the harm of commonly applied methods, J. Clin. Epidemiol. 98 (2018) 133–143, https://doi.org/10.1016/j.jclinepi.2017.11.013, https://www.sciencedirect.com/science/article/pii/S0895435617307886.
[60] W. Sauerbrei, P. Royston, H. Binder, Selection of important variables and determination of functional form for continuous predictors in multivariable model building, Stat. Med. 26 (30) (2007) 5512–5528, https://doi.org/10.1002/sim.3148, arXiv: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.3148, https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3148.
[61] D. Gonçalves-Ferreira, M. Sousa, G.M. Bacelar-Silva, S. Frade, L.F. Antunes, T. Beale, R. Cruz-Correia, et al., Openehr and general data protection regulation: evaluation of principles and requirements, JMIR Med. Inform. 7 (1) (2019) e9845.
[62] K. Stöger, D. Schneeberger, P. Kieseberg, A. Holzinger, Legal aspects of data cleansing in medical ai, Comput. Law Secur. Rev. 42 (2021) 105587, https://doi.org/10.1016/j.clsr.2021.105587, https://www.sciencedirect.com/science/article/pii/S0267364921000601.
[63] X. Qian, S. Xian, S. Yifei, G. Wei, H. Liu, X. Xiaoming, C. Chu, Y. Yilong, Y. Shuang, M. Kai, C. Mei, Q. Yi, External validation of a deep learning detection system for glaucomatous optic neuropathy: a real-world multicentre study, Eye 37 (2023) 3813–3818, https://doi.org/10.1038/s41433-023-02622-9.
[64] C. Birkenbihl, M.A. Emon, H. Vrooman, S. Westwood, S. Lovestone, M. Hofmann-Apitius, H. Fröhlich, Differences in cohort study data affect external validation of artificial intelligence models for predictive diagnostics of dementia: lessons for translation into clinical practice, EPMA J. 11 (2020) 367–376, https://doi.org/10.1007/s13167-020-00216-z.
[65] B. Meskó, E.J. Topol, The imperative for regulatory oversight of large language models (or generative ai) in healthcare, npj Digit. Med. 6 (1) (2023) 120, https://doi.org/10.1038/s41746-023-00873-0.
[66] M. Nagendran, Y. Chen, C.A. Lovejoy, A.C. Gordon, M. Komorowski, H. Harvey, E.J. Topol, J.P.A. 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, https://doi.org/10.1136/bmj.m689.
[67] L. Marconi, E. Pirovano, F. Cabitza, Clarity ai: a comprehensive checklist integrating established frameworks for enhanced research quality in medical ai studies, in: F. Calimeri, M. Dragoni, F. Stella (Eds.), Artificial Intelligence for Healthcare 2024: Proceedings of the 3rd AIxIA Workshop on Artificial Intelligence for Healthcare (HC@AIxIA 2024), CEUR Workshop Proceedings, 2024, pp. 1–14, http://ceur-ws.org/Vol-3880/paper1.pdf.
[68] D. Shin, Debiasing AI: Rethinking the Intersection of Innovation and Sustainability, 1st edition, Routledge, 2025.
[69] D. Shin, Artificial misinformation, Exploring human-algorithm interaction online, 2024.
[70] R.U. Shah, A.P. Bress, A.J. Vickers, Do prediction models do more harm than good?, Circ. Cardiovasc. Qual. Outcomes 15 (4) (2022) e008667.
[71] J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché Buc, E. Fox, H. Larochelle, Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program), J. Mach. Learn. Res. 22 (1) (2021) 7459–7478.
[72] N. Barnes, Publish your computer code: it is good enough, Nature 467 (7317) (2010) 753.
[73] E. Ammenwerth, M. Rigby, Evidence-based health informatics and the scientific development of the field, in: Evidence-Based Health Informatics: Promoting Safety and Efficiency Through Scientific Methods and Ethical Policy 222, 2016, p. 14.
[74] C.L. Ramspek, K.J. Jager, F.W. Dekker, C. Zoccali, M. van Diepen, External validation of prognostic models: what, why, how, when and where?, Clin. Kidney J. 14 (1) (2020) 49–58, https://doi.org/10.1093/ckj/sfaa188, arXiv: https://academic.oup.com/ckj/article-pdf/14/1/49/36184810/sfaa188.pdf.
[75] L. Huang, S. Ruan, Y. Xing, M. Feng, A review of uncertainty quantification in medical image analysis: probabilistic and non-probabilistic methods, Med. Image Anal. 97 (2024) 103223, https://doi.org/10.1016/j.media.2024.103223, https://www.sciencedirect.com/science/article/pii/S1361841524001488.
[76] R.D. Riley, J. Ensor, K.I.E. Snell, T.P.A. Debray, D.G. Altman, K.G. Moons, G.S. Collins, External validation of clinical prediction models using big datasets from e-health records or ipd meta-analysis: opportunities and challenges, BMJ 353 (2016), https://doi.org/10.1136/BMJ.I3140.
[77] D.G. Altman, M. Trivella, F. Pezzella, A.L. Harris, U. Pastorino, Systematic review of multiple studies of prognosis: the feasibility of obtaining individual patient data, in: Advances in Statistical Methods for the Health Sciences: Applications to Cancer and AIDS Studies, Genome Sequence Analysis, and Survival Analysis, 2007, pp. 3–18.
[78] R.B. Parikh, Z. Obermeyer, A.S. Navathe, Regulation of predictive analytics in medicine, Science 363 (6429) (2019) 810–812, https://doi.org/10.1126/science.aaw0029, arXiv: https://www.science.org/doi/pdf/10.1126/science.aaw0029, https://www.science.org/doi/abs/10.1126/science.aaw0029.