THE EUROPEAN SOCIETY OF MEDICINE
Medical Research Archives, Volume 13 Issue 12

RESEARCH ARTICLE

A scoping review of AI/ML algorithm updating practices for model continuity and patient safety using a simplified checklist

Ahmed Umar Otokiti[1], MD ,MPH, MBA; Huan-ju Shih[2], PhD(c); Makuochukwu Maryann Ozoude[3], MD; Ilse Siguachi[4], MS; Leyla B. Warsame[5], MD; Karmen S. Williams[6], DrPH, MBA; Seyi John Akinloye[7], BDS.

[1]Digital Health 360 Degrees at Digital Health Solutions, LLC, White Plains, NY, United States.
[2]College of Public Health, George Mason University, Fairfax, VA, United States.
[3]Zaporozhye State Medical University, Zaporizhzhia, Ukraine.
[4]Department of Epidemiology and Biostatistics, Graduate School of Public Health and Management, City University of New York, NY, United States.
[5]House Calls Physician, NY, United States.
[6]Department of Health Policy and Management, Graduate School of Public Health and Management, City University of New York, NY, United States.
[7] School of Dentistry, University of Alabama at Birmingham, Birmingham, USA

OPEN ACCESS

PUBLISHED
31 December 2025

CITATION
Otokiti, A.U., et al., 2025. A scoping review of AI/ML algorithm updating practices for model continuity and patient safety using a simplified checklist. Medical Research Archives, [online] 13(12). https://doi.org/10.18103/mra.v13i12.7083

DOI
https://doi.org/10.18103/mra.v13i12.7083

ISSN
2375-1924

ABSTRACT

Objective: To evaluate the extent to which clinical artificial intelligence (AI) and machine learning (ML) models prioritize updating, transparency, and demographic reporting in the published literature.

Patients and Methods: This study conducted a systematic review of clinical AI/ML models using PRISMA guidelines from March 2020 until December 2021. A new checklist and scoring system were introduced to assess model quality, with additional evaluation of demographic reporting, particularly by ethnicity and race. A comprehensive search was performed across six major databases, including Ovid Embase, MEDLINE, and Cochrane Library. Across various study designs, eligible studies included human-based predictive or prognostic AI/ML models using supervised learning and at least two predictors. Studies not meeting these criteria were excluded.

Results: Out of 390 AI/ML studies reviewed, only 9% mentioned plans or methods for future model updates. The vast majority (98%) of models were still in the research phase, and only 2% had reached production. Additionally, only 12% adhered to best practices in model development, and 84% failed to report demographic composition by race or ethnicity.

Conclusion: These findings highlight key limitations in the current clinical AI landscape—especially a lack of transparency, limited readiness for deployment, and minimal consideration for inclusivity or generalizability. Greater focus on model updating, adherence to development standards, and demographic transparency is essential to improve the safety, reliability, and equity of clinical AI/ML models.

Keywords: Artificial intelligence, machine learning, model updating

Introduction

There is a surge in healthcare artificial intelligence (AI) and machine learning (ML) algorithms due to their potential to improve clinical efficiency and overall quality of care[1]. These algorithms can automate insights directly from data without using standard computer programming and analyze large data sets with high dimensionality to yield insights and predictions on complex associations without prior assumptions from traditional statistical methods, differentiating AI and ML models from other statistical models[2–4]. Supervised and unsupervised learning are the two general methods of gaining insights from data in AI: supervised learning involves making predictions based on a set of prespecified input, references, and output variables, whereas unsupervised learning is used to draw inferences from data sets consisting of input data without labeled responses[5].

Just like biological systems undergo programmed cell death, a well-known characteristic of AI and ML algorithms is that their performance degrades over time due to the occurrence of model calibration (calibration drift), which refers to a shift in the accuracy of risk estimates in terms of the agreement between the predicted risks of events and their actual observed frequencies[6]. Calibration drift arises due to deploying a model in a dynamic environment, with the resulting difference between the population or setting in which the model was trained and that in which it was implemented[7].

Patient-level algorithm predictions must prioritize consistency and accuracy due to the risk of patient harm; therefore, an appropriate model-updating process is essential across the model's lifetime[8]. The best practice is to update a clinical model rather than abandon the model, build another, or repeat the selection of predictors, which leads to a loss of the previous scientific information captured[9,10].

The existence of multiple models for the same clinical scenario without model-updating methods declared *ab initio* leaves clinicians uncertain of which model is appropriate to use[9]. For example, there are more than 80 models for the prognosis of stroke[11], more than 20 models predicting intensive care unit stay after cardiac surgery[12], more than 100 published algorithms for prognosis after neurotrauma[13], and over 50 models to predict outcomes after breast cancer[14].

Subtle population demographic changes, in addition to changes in healthcare access and the heterogeneity of health insurance coverage (health disparity), can also deteriorate a model's future output[1,15]. Changes in best practice clinical guidelines and variations in practice preferences across different healthcare providers can also be a source of data shift, resulting in sub-optimal model output[1–8,10,16–20]. Health centers can update or change aspects of their information systems, database and data archiving systems, and digital health tools such as imaging software and EHRs. In addition, there is constant change in clinical nomenclature and disease coding, which can also affect the output[21]. The healthcare regulatory landscape is constantly evolving as well[22,23]. The enactment of the Affordable Care Act was associated with many sweeping reforms to healthcare delivery and redefining value in healthcare delivery[24], and, as such, a model built to produce outputs based on previous standards of care will likely be suboptimal.

Most AI tools are developed based on the nuances of specific local healthcare workflows and the data they generate; for example, consider an algorithm developed to predict sepsis based on a patient's lactate level. The algorithm will learn to correlate the physician's lactate orders with a high possibility of sepsis; however, model quality would be reduced if a policy change required more frequent ordering of lactate tests. Model validation will show reduced performance in this situation, as the learned pattern does not generalize across sites and circumstances[15,23,(p31)]. In addition, there is systemic bias in the geographic distribution of patient cohorts, as algorithms trained on US data were disproportionately trained on patients from just 3 states (New York, California, and Massachusetts)[25].

Label and causality leakage phenomena occur when the model's prediction target is directly or indirectly present in the training data set[26]. An example is a model developed to predict hospital mortality in patients admitted to the intensive care unit. An AI model trained naively on all data will learn to correlate extubating and turning off the ventilator with the death of a patient and ultimately produce a near-perfect predictive performance yet with absolutely no clinical utility[26]. Causality leakage in the clinical model can occur in a situation whereby a clinician orders a test based on a high index of suspicion of a clinical outcome that the algorithm is meant to predict; the algorithm then uses the test to generate an alert that results in an action[27].

The need to prioritize model updating is equally significant in the application of Large Language Models (LLMs). Generative Artificial Intelligence (Gen A.I)[28,29] types of LLM, such as Generative Pre-trained Transformers (GPT), are gradually being integrated into various aspects of healthcare operations and clinical care[29]. Applications include medical text summarization, translation, clinical decision support, clinical documentation, patient education, adverse effect detection, and clinical research data management[30] Gen A.I. models differ from traditional rule-based systems as they operate on much higher data dimensionality and volume. For instance, GPT-4 was trained on data with one trillion parameters (OpenAI)[28,29]. Another defining characteristic is the use of techniques like Reinforcement Learning from Human Feedback (RLHF), which incorporates few-shot learning and chain-of-thought reasoning[30]. However, LLMs also present challenges such as a lack of explainability, hallucinations, bias propagation, overdependence bias, and the potential weakening of clinicians' critical thinking abilities[31]. The complexity of Gen A.I. models, coupled with their lack of complete explainability—leading to issues such as hallucination and bias propagation—underscores the importance of model updating, especially for Gen A.I. and LLMs[32].

Agentic AI systems are models that can act independently and autonomously to achieve predefined objectives[33]. Although many modern AI implementations combine both capabilities of Gen A.I and agentic A.I, the autonomous state of agentic A.I models sets them apart from Gen A.I[34]. Considering that these higher level models are built on multiple layers of neural networks and volumes of data, the need for their model updating becomes even more important to ensure safety and continuity. There are several methods that address the data shift required to update models[1,9,21,35,36]. The least complex method involves adjusting the model intercept to a different prevalence or incidence rate according to the new population assuming risk factors still confer the same level of risk. Another option is to adjust the population prevalence rate and add a single adjustment to all risk factors in the model; one or more risk factor relationships may also need to be adjusted, given the changes in relationships over time. A more complex method involves adjusting both the prevalence and the coefficients and adding new risk factors into the model. The last option involves refitting the entire model based on a new data sample, either alone or in combination with the addition of new potential risk factors; this essentially remodels the problem from scratch based on a new sample[36].

Real-time calibration drift detection and updating is the most computationally intensive approach; however, real-time detection provides users with the peace of mind that their models are accurate at the time of use without requiring manual steps[36]. A similar approach is incremental updating, in which models are updated based on new instances as they become available[35,36]. Fixed and batch updating at specified intervals is another option, with models evaluated and updated at specific intervals[35,36].

Generally speaking, Gen A.I and Agentic AI models can be updated by the following methods; long context, fine tuning and Retrieval-augmented generation (RAG)[37–39], RLHF (Reinforcement Learning from Human Feedback)[40,41];RLAIF (Reinforcement Learning from AI Feedback), experience-based learning, iterative optimization, self-reflection[42,43].

Our main study objective is to evaluate clinical model updating in peer-reviewed AI and ML models and assess model updating practices used in direct patient-provider clinical decision-making. Phases of model development pertaining to applicability and reproducibility (model updating, impact assessment, and implementation) have received less attention in the scientific literature[34]. Clinical model-updating processes seek to prevent model deterioration with adverse consequences of model inaccuracy. The lack of model updating in clinical settings can impact the generalizability and reproducibility of clinical models[44]. The model-updating processes of clinical algorithms should be determined proactively from the time of initial model development[23,44] to ensure patient safety and quality of care. Additionally, identifying possible algorithmic risk in the form of pre-deployment risk assessment should be an integral part of determining the level of aggressive ongoing auditing and reassessment required during model implementation.

Methods

INCLUSION AND EXCLUSION CRITERIA

Our original protocol was developed based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) protocols[45]. All studies that were published from March 2020 until December 2021 were reviewed (**Textbox 1**). We chose to include papers from March 2020 to December 2021 because we were more interested in the latest model updating practices amongst healthcare AI models that have contemporary implications on patient safety. Bearing in mind, the rapid acceleration of AI implementation in healthcare settings, this duration corresponded to the previous recent years as at the time of our review and literature search. Studies were included without geographic or regional preferences. More detailed justifications for our inclusion and exclusion criteria items can be found in our earlier published protocol of this systematic review[46]. Failure to meet eligibility criteria resulted in exclusion from the review (**Textbox 2**).

OUTCOMES

The primary outcome of this scoping review is the percentage of published algorithms that prioritize model-updating methods (model updating is considered prioritized if it is part of the algorithm protocol). We identified the type of relationship between studies that prioritize model updating and the following model characteristics: geographic region, quality of studies and by setting of model development. In addition, we assessed how frequently EHR records were used for model development[47–49].

As a secondary end point, we captured the amount (percentages) of models reporting the demographic breakdown of their data (ethnic background and gender); this is of particular importance owing to potential societal harm and resulting AI and ML algorithm setbacks due to the use of nonrepresentative data[50].

SEARCH STRATEGY

A comprehensive literature search was conducted using the following databases: Ovid Embase, Ovid MEDLINE, Ovid PsycINFO, Web of Science Core Collection, Scopus, and the Cochrane Library. Searches were originally limited to articles published from January 1, 2018, to December 31, 2021; however, the study team decided only to include articles from March 2020-December 2021. We were more interested in the latest model updating practices amongst healthcare AI models. Bearing in mind, the accelerating wave in AI evolution, this duration corresponded to the previous recent years as at the time of our review and literature search.

The search strategy for each database was developed by a medical librarian (SW) in concert with the rest of the team. Each search strategy used a combination of keywords and subject headings related to ML, predictive algorithms, medical diseases and disorders, and study design (**Appendix A**).

MODEL ANALYSIS

To evaluate the reporting quality, we adapted a verified tool available for model quality assessment[51]. The CHARMS is an 11-item checklist, with each

item created to assess the model on the domains of risk of bias and applicability (**Appendix B**). The checklist is a comprehensive guide created from a combination of eight other published guides that include both criteria to ascertain applicability and reproducibility with implications for patient safety, as well as technical validity of a model's results, some of which are beyond the scope of our review.

CREATION OF A CLINICAL MODEL QUALITY ASSESSMENT TOOL (MODIFIED CHARMS CHECKLIST)

We created our quality assessment checklist by extracting criteria that are more specific to applicability and reproducibility analysis that could potentially impact patient safety and quality of care at the clinical model deployment level, resulting in our six-item checklist for study quality assessment (**Textbox 3**).

Our goal was to focus on established factors and best practices that indicate a study's applicability and low risk of bias to ensure generalizability beyond the model's technical output as follows[52–54]:

- *Applicability:* the extent to which the study fits within the inclusion and exclusion criteria
- *Risk of bias*: the extent to which any flaws in the study lead to overly optimistic estimates of predictive performance measures
- *Generalizability*: the degree to which the study results are relevant to the larger population
- *Reproducibility:* the ability to duplicate the study using the same methods as in the original study

The target users of our modified checklist are clinicians interested in the objective assessment (based on the dimensions of applicability and reproducibility) of published AI/ML models who may not be technically savvy about the inner workings of data science or machine learning algorithms.

The original CHARMS checklist comes with some technical variables that are more advanced than the comprehension of such clinicians and some variables that are not directly related to applicability and reproducibility, hence our rationale for the creation of the modified checklist.

A total of five items out of our six-item checklist were adapted from the CHARMS checklist (CHARMS checklist has 11 items); our last criterion, the model development checklist standard, was obtained from the literature review of best practices for model development.

Our determination of criteria that are more specific to applicability and reproducibility at the level of care was based on the recommendation of the CHARMS checklist as published by Moons et'al[51]. In the published CHARMS checklist, apart from identifying the 11 domains of model quality assessment, they also highlighted specific items within the domains that impacts applicability, generalizability and overall model quality.

Since our goal is to assess model applicability and generalizability at the clinical level and not to evaluate the technical validity of healthcare models, we took the initiative to exclude domains that focus on technical validity and with less impact on applicability and generalizability as per CHARMS checklist recommendations. We then unanimously excluded these six domains listed below from the CHARMS checklist leaving us with five domains to build our modified CHARMS checklist.

1. Participants
2. Candidate predictors
3. Sample size
4. Model development
5. Model performance
6. Results.

Our 6th checklist item (Model development and reporting standards) was included based on our findings from literature search showing that adhering to a model development and reporting standard can ensure study reproducibility and applicability[55–57].

Most clinical models do not declare any model development and reporting standards they may have adopted. Despite the availability of these guidelines,

there is poor overall reporting of adopted model development standards in many published AI models[54–56,58] .

We reviewed the following reporting standards as part of our literature review for this study; SPIRIT-AI (Standard Protocol Items: Recommendation for Interventional Trials-Artificial Intelligence), CONSORT-AI (Consolidated Standards of Reporting Trials-Artificial Intelligence), TRIPOD (Transparent Reporting of a Multivariable Prediction Model of Individual Prognosis Or Diagnosis), REMARK (Reporting Recommendations for Tumour Marker Prognostic Studies), and GRIPS (Genetic Risk Prediction Studies)[59–67].

RATIONALE FOR CHECKLIST ITEMS

Checklist Items Adapted from the CHARMS Checklist:

- Study Design/Data Source for Model Development: The data used to develop the algorithm may be sourced from retrospective and prospective cohorts including RCTs and cross-sectional studies. In addition, there is a proliferation of sourcing model data from registries, databases, and EHRs. Although RCTs are considered the gold standard, they also have shortcomings similar to all other methods. Although RCTs are designed to reduce biased outcomes, their findings can lead to impaired generalizability of outcomes in real-life clinical scenarios owing to the rigid eligibility criteria of study participants. Data sources for model development are critical for the predictive accuracy, applicability, and reproducibility of any algorithm[10,51,64]
- Outcomes: the lack of well-defined study outcomes increases risk of bias and adversely affects model reproducibility in real-life clinical scenarios[51]. For example, 40% of cancer prognostic model studies were found to have poorly defined outcomes[65]. For our quality assessment, a well-defined outcome is considered to occur when the definition and measurement of the outcome events or target disease clearly correspond to the outcome definition of the study objective[51] .
- Model testing and evaluation methods: model validation is the process of quantifying model performance in other individuals beyond the training and testing data set used to develop the model[66]. Whenever the predictive performance of a model is estimated using the same data set that was used to develop the model, it is referred to as "apparent performance"[51]. Regardless of which modeling technique is used, apparent performance tends to be biased, as it can overestimate performance relative to the performance of other individuals. It is very important that all models be evaluated in an independent data set (external validation) before deployment[50]. Externally validated models (either temporal or geographic validation) provided the best insights into the usefulness of the model for other individuals, centers or settings, and regions. Several reviews have shown that external validation studies are generally uncommon[5,13,67,68], as most studies are only internally validated by a random split sample of the data into development and validation samples[5]. (Table 1)
- Model updating method recommendation: in the event that an existing model shows poor performance when evaluated in other settings (geographic or temporal), it is best practice to adjust, update, or recalibrate the original model to increase performance[51], as there are well-established methods to achieve successful model updating. It is also best practice that the potential techniques for updating a model on external deployment can be identified before deployment[1,23].
- Model interpretation and generalizability concerns: best practice guidelines for reporting medical studies recommend discussing strengths, weaknesses, and future challenges with regard to the generalizability of the studies[60,69,70]. For models, these studies should

therefore provide insight into the model's applicability, usefulness, and intended users[51]. This discussion also serves as a basis for comparison with other studies. Therefore, our quality checklist will include a score (1 star) for a study that mentions the strengths and weaknesses of their model in the Discussion section.

Table 1. Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies.

Domains and key items		General	Applicability	Risk of bias
Source of data				
	Source of data (cohort, case-control, randomized trial participants, or registry data)		✓	✓
Participants				
	Participant eligibility and recruitment methods (consecutive participants, location, number of centers, setting, and inclusion and exclusion criteria)	✓	✓	
	Participant descriptions	✓	✓	
	Details of treatment received, if relevant		✓	✓
	Study dates	✓	✓	
Outcome to be predicted				
	Definition and methods for outcome measurements		✓	✓
	Determine if the same outcome definition and method for measurement was used in all patients			✓
	Type of outcome (single or combined end points)	✓	✓	
	Determine if the outcome was assessed without knowledge of candidate predictors (blinded)			✓
	Determine if candidate predictors were part of the outcome (in panel or consensus diagnosis)			✓
	Time of outcome occurrence or summary of duration of follow-up		✓	

Candidate Predictors (or Index Test)				
	Number and type of predictors (demographics, patient history, physical examination, additional testing, and disease characteristics)	✓		
	Definition and method for measuring candidate predictors		✓	✓
	Timing of predictor measurement (patient presentation, diagnosis, and treatment initiation)		✓	
	Determine if predictors were assessed blinded for outcome and for each other (if relevant)			✓
	Handling predictors in the modeling (continuous, linear, and nonlinear transformation or categorized)			✓
Sample size				
	Number of participants and number of outcomes or events	✓		
	Number of outcomes or events in relation to the number of candidate predictors (events per variable)			✓
Missing data				
	Number of participants with any missing values (including predictors and outcomes)	✓		✓
	Number of participants with missing data for each predictor			✓
	Handling of missing data (complete case analysis, imputation, or other methods)			✓
Model development				
	Modeling methods (logistics, survival, neural networks, or machine learning techniques)	✓		
	Modeling assumptions satisfied			✓
	Method for selecting predictors for inclusion in multivariable modeling (all candidate predictors and preselection based on unadjusted association with the outcome)			✓

	Methods for selecting predictors during multivariable modeling (full model approach backward or forward selection) and criteria used (*P* value and Akaike Information Criterion)			✓
	Shrinkage of predictor weights or regression coefficients (no shrinkage, uniform shrinkage, and penalized estimation)		✓	✓
Model performance				
	Calibration (calibration plots, calibration slope, and Hosmer-Lemeshow test) and discrimination (*C*-statistic, *D*-statistic, and log-rank) measures with CIs		✓	
	Classification measures (sensitivity, specificity, predictive values, and net reclassification improvement) and whether a priori cut points were used			✓
Model evaluation				
	Method used for testing model performance: development data set only (random split of data, resampling methods, bootstrap or cross-validation, or none) or separate external validation (temporal, geographic, different settings, and different investigators)			✓
	In case of poor validation, whether the model was adjusted or updated (intercept recalibrated, predictor effects adjusted, or new predictors added)		✓	✓
Results				
	Final and other multivariable models (basic, extended, and simplified) presented, including predictor weights or regression coefficients, intercept, baseline survival, and model performance measures (with SEs or CIs)	✓	✓	

	Any alternative presentation of the final prediction models (sum score, nomogram, score chart, and predictions for a specific risk subgroup with performance)	✓	✓	
	Comparison of the distribution of predictors (including missing data) for development and validation data sets			✓
Interpretation and discussion				
	Interpretation of presented models (confirmatory, model useful for practice vs exploratory, and more research needed)	✓	✓	
	Comparison with other studies, discussion of generalizability, strengths, and limitations	✓	✓	

Other CHARMS Checklist Items

The remaining 6 items in CHARMS were excluded from our assessment tool because they were already considered during the initial screening stage of our review process (participant characteristics and predictors). We also excluded items that focused on technical assessment, as that is beyond the scope of our study objective of real-life clinical applicability (technical process of model development, model performance, results, and sample size). Although the checklist still needs to be validated, our adapted checklist captures the essence of our review.

Checklist Items Based on a Literature Review of Best Practices of Clinical Model Studies: Model Development Reporting Standards

The best practice standards for reporting primary prognostic and predictive model studies exist in the literature[56] and include SPIRIT-AI (Standard Protocol Items: Recommendation for Interventional Trials-Artificial Intelligence), CONSORT-AI (Consolidated Standards of Reporting Trials-Artificial Intelligence), TRIPOD (Transparent Reporting of a Multivariable Prediction Model of Individual Prognosis Or Diagnosis), REMARK (Reporting Recommendations for Tumour Marker Prognostic Studies), and GRIPS (Genetic Risk Prediction Studies)[59–61]. Adhering to these guidelines may ensure study reproducibility and could improve future real-life applications[55–57]. Despite the availability of these guidelines, there is poor overall quality of reporting in many published AI models[54–56,58]. Therefore, we have included declaring a reporting standard as part of our checklist (reporting standard scores will receive 1 star).

For each checklist item fulfilled by the study reviewed, studies will be scored with 1 or 2 stars as described above, with a possible maximum score of 10 stars for each study.

Quantitative assessment of the quality of the reviewed studies using our modified CHARMS checklist

Based on our literature search, it was evident that there is no universally accepted standard definition for assessing the quality of studies and evaluating the risk of bias in research papers related to our study. To address this, we established a baseline for quality assessment in this study:

Research Study Design and Handling of Missing Data: For studies conducted using randomized controlled trials, we assigned 2 points. For other research sources and designs, including cohorts, registries, and convenience sampling, 1 point each was allocated.

Articles demonstrating effective handling of missing data were assigned 1 point. This criterion is not to assess the level or degree of missing data in a study enough to affect its technical validity. Rather, we are interested in whether the studies have declared how they handled missing data. As such, this criterion has a binary answer; either they declared it or did not declare it.

Clearly Defined Primary Outcome: Studies explicitly defining their primary outcome were given 1 point.

Model Testing and Evaluation Methods: Because of the higher impact of external validation on model applicability in real-life clinical scenarios, we prioritize these models in our checklist. Research papers that incorporated separate external validation methods spanning geographical, temporal, and population variations received 2 points. Studies relying solely on the same development data for validation, including random splits like 80/20 or 70/30 and reassembly techniques (e.g., bootstrap and cross-validation), were allocated 1 point.

Model Updating Information: The primary outcome of our review is the proactive determination of possible model-updating methods. As such, we will prioritize any study that proactively suggests a model-updating method as part of its study method by scoring it as 2 points.

Declaration of Model limitations and strength: If a paper included an evaluation of the model's strengths, weaknesses, and risk of bias, it was given 1 point. If this information was absent, no points were awarded.

Adherence to Model Development and Reporting Standards: If the study conformed to recognized best practice standards for model development and reporting, and if it cited relevant standards such as CONSORT-AI, SPIRIT-AI, DECIDE-AI, NEUR-UPDA ML, TRIPOD-ML, PROBAST-ML, and STROBE, it received 1 point.

A study was considered to meet the baseline if it scored at least 5 points (**Textbox 4**), with a basic acceptable score being 5, in the following categories: Study design (1 point); Handling of missing data (1 point); Well-defined primary outcome (1 point); Adequate model testing and evaluation methods (1 point); Model update (N/A); Model interpretation and limitation concerns (1 point); Model reporting and development standard (N/A). Subsequently, we also created a new variable named "quality baseline," which was categorized as either "yes" for papers meeting the quality baseline (i.e., scoring 5 points or more) or "no" for those falling short of this baseline. This quality assessment framework provided a structured approach for evaluating the research papers in our study.

STUDY SELECTION AND DATA EXTRACTION

All search results were imported into Covidence software for deduplication and screening[61]. Covidence facilitates a blind review process, and results from multiple databases can be imported, deduplicated, and screened for eligibility. Following the title and abstract screening phase, the full text of all included abstracts were gathered and imported into the Covidence software. Covidence created a PRISMA flowchart and facilitated data extraction and quality appraisal phases[62].

Two reviewers (team members A.O and H.S) used the Covidence software to screen the title and abstract of each article and the full text of all included abstracts. Two independent reviewers resolved disparities whenever there was a lack of agreement in the papers selected.

Data extraction and quality assurance were conducted by all team members simultaneously using the Covidence software. For any particular data point to be accepted, at least two reviewers must agree with the data extracted. It was also set up to resolve conflict between two reviewers by allowing for a third

reviewer to serve as a tiebreaker in any particular study with a contested data point.

STATISTICAL ANALYSIS

The screening process was documented and presented using the PRISMA flow diagram (**Figure 1**). Prior to title and abstract screening, the review team met to screen a random sample of 50 records to validate the inclusion and exclusion criteria.

For each checklist item fulfilled by the study reviewed, studies were scored with 1 or 2 points as described above in **Textbox 3**, with a possible maximum score of 10 stars for each model reviewed.

Our preliminary search on geographic clusters of reported AI adoption and model implementation revealed that AI and ML adoption is mostly clustered in the United States, Canada, the United Kingdom, Australia, the European Union, China, Taiwan, and Israel[49,60]. We added the following categories to our geographical regions based on the clusters of models from our preliminary results: Japan, Korea, India, Pakistan, South America, Other Asia, and Others (other countries not specified in the predefined categories).

After extracting data from the studies, we conducted a narrative synthesis. Data were summarized using descriptive statistics, figures, and tables for visualization. Categorical data were presented as percentages. The distribution of continuous data such as sample size and the number of predictors were described using means and SDs for normally distributed data using median and 25th and 75th percentiles for nonnormally distributed data. The results were characterized by study design, outcomes, service delivery type, ML techniques, and model-updating properties.

ETHICS APPROVAL

On August 13, 2021, our systematic review protocol was registered with the International PROSPERO (Prospective Register of Systematic Reviews) CRD42021245470[71]. Our protocol was developed based on the PRISMA-P (Preferred Reporting Items for Systematic Reviews and Meta-Analysis Protocols) 2015 statement[45]. Our study does not require an ethics committee review because our research does not directly involve human subject data, and it was conducted on publicly available data from published articles.

DEVIATION FROM OUR REGISTERED PROTOCOL

The main deviation from our registered protocol was the duration of studies included in the systematic review. Our initial protocol was to include the previous ten years of A.I research. We included only the last two years to capture the latest AI study characteristics in the review as explained in our method section above.

Results

The search resulted in 390 articles for exaction. Most aims of these publications were predictive in nature (300, 75.8%) and carried out in academic centers (261, 66%), with mainly neural network algorithms (288, 72.7%) (**Table 2;**). Cardiology (67, 16.9%), Neurology (55, 13.9%), Respiratory (47, 11.9%), ID (38, 9.6%), and GU (33, 8.33) models were most prevalent (**Table 2; Figure 3**). Most models were also in the research phase (388, 98%); only (8, 2%) were in the production phase. Geographically, most models in our sample were from China, Taiwan, EU, US, Japan, and Korea (**Figure 2**). Only 16% of the models were built using accessible open data registries.

Based on our endpoint and quality assessment tool components (Modified CHARMS checklist) (**Table 3**), only 6% of the total studies were randomized controlled trials (RCTs). Furthermore, a mere 9% of studies attempted to ensure the model would be updated in the future; over half of the studies (53%) neglected to declare their approach to handling missing data. Furthermore, only 32% of these studies accurately defined their primary outcomes, and a mere 27% diligently tested their models using external validation methods; only 12% reported following a best practice standard.

[1]**Table 2.** Study characteristics

Characteristics		**Results**
Total	396	**Numbers (%)**
Publication Setting		
	Academic	261 (65.91)
	Non-academic medical center	112 (28.28)
	Vendor/Industry	12 (3.03)
	Governmental	11 (2.78)
Study Aim		
	Predictive	300 (75.76)
	Prognostic	96 (24.24)
Disease-biological system of study		
	Neurology	55 (13.89)
	Endocrinology	29 (7.32)
	ENT	6 (1.52)
	Cardiovascular	67 (16.92)
	Respiratory	47 (11.87)
	Gastroenterology	27 (6.82)
	Genitourinary	33 (8.33)
	Orthopedics/MSK	31 (7.83)
	Infectious disease	38 (9.60)
	Dermatology	2 (0.51)
	Multi-systemic	28 (7.07)
	Rheumatology	7 (1.77)
	Hematology Oncology	8 (2.02)
	Other: Anesthesiology	1 (0.25)
	Other: ICU	2 (0.51)
	Other: Ophthalmology	7 (1.77)
	Other: Opioid	2 (0.51)
	Other: Patient Priorities Care	2 (0.51)
	Other: Surgery	1 (0.25)
	Other: Traumatology	1 (0.25)
	Other: cognitive function	1 (0.25)

Type of algorithm		
	Traditional machine learning	9 (2.27)
	Deep neural network	288 (72.72)
	Both traditional machine learning/deep neural networks	99 (25)
Geographical region of publication		
	US.	70 (17.68)
	Canada	8 (2.02)
	China/Taiwan	126 (31.82)
	UK	12 (3.03)
	Australia/New Zealand	5 (1.26)
	EU	86 (21.72)
	Israel	5 (1.26)
	India/Pakistan	8 (2.02)
	Japan/Korea	43 (10.86)
	South America	13 (3.03)
	Other Asia	14 (3.54)
	Other unclassified	5 (1.26)
Stage of model implementation		
	Production/post research	8 (2.02)
	Research	388 (97.98)
Data sources registry		
	Closed registry/proprietary	330 (83.33)
	Open registry/open source	65 (16.41)
	Missing	1 (0.25)
Oncology study		
	Yes	59 (14.90)
	No	337 (85.10)
COVID-19 Study		
	Yes	30 (7.58)
	No	366 (92.42)

Table 3. Primary end point with other items in the quality assessment checklist (Modified CHARMS checklist)

Characteristics		**Results**
Total	396	**Numbers (%)**
Study design and missing data		
	RCT	24 (6.06)
	Others	371 (93.69)
	Missing	1 (0.25)
Handling of missing data		
	Yes	181 (45.71)
	No	210 (53.03)
	Missing	5 (1.26)
Primary outcome is well defined		
	Yes	127 (32.07)
	No	268 (67.68)
	Missing	1 (0.25)
Model testing and evaluation methods		
	Yes	107 (27.02)
	No	289 (72.98)
Model updating method (primary end point)		
	Yes	38 (9.60)
	No	358 (90.40)
Model limitation and applicability concerns		
	Yes	302 (76.26)
	No	94 (23.74)
Model reporting and development standard		
	Yes	48 (12.12)
	No	348 (87.88)

Based on a two-way T-tests, the average quality score of studies that recommended model updating was higher than those that did not recommend any model updating method t(9394) = 2.5. p<0.001. There was, however, no significant relationship between the setting of model development (academic vs. non-academic) or geographical region and quality of study scores. A multiple/mixed regression analysis controlling for site of model development and nature of the model revealed a positive relationship between studies that suggested model updating and higher quality scores.

In our sample, 31% of worldwide models did not disclose their gender composition while 23% of US focused models did not disclose gender (**Table 4**). 84% of worldwide had no ethnicity composition reported while 44% of the US focused models did not show the ethnic composition of their training data.

Table 4. Secondary end points

Characteristics		**Results; Numbers(%)**	
Break down by ethnicity		**Worldwide**	**USA only**
	Yes	61 (15.40)	39 (55.71)
	No	335 (84.60)	31 (44.29)
Break down by gender			
	Yes	272 (68.69)	54 (77.14)
	No	124 (31.31)	16 (22.86)

Discussion

PRINCIPAL FINDINGS AND COMPARISON TO PRIOR WORK

Our study objective was to evaluate prioritization of clinical model updating in peer-reviewed/published AI and ML models that can be used in direct patient-provider clinical decision-making. We also tested the relationship between the quality of published AI clinical models and prioritization of the model updating process. Our secondary outcomes included AI/ML model geographic distribution and inclusion of demographic data.

In recent years, there has been a growing interest in the development and implementation of clinical AI/M) models in healthcare to improve patient outcomes and assist healthcare providers in decision-making. However, our scoping review reveals that most studies were primarily predictive rather than prognostic in nature, suggesting that the focus has been on predicting and identifying certain conditions or outcomes (i.e., sepsis, readmission, deterioration risk, or CDI) rather than assessing long-term prognosis or patient trajectories. Predictive modeling can be complex and opaque thereby enhancing the distrust in AI systems used in clinical practice. Furthermore, our analysis revealed that neural network algorithms were the most commonly used given their efficacy in handling complex medical data but most neural network algorithms lack clarity due to their opaque nature (no explainability).

Additionally, our analysis found that the majority of these studies were conducted in academic centers where research is prioritized, infrastructure/resources are available, and there is an established culture of pedagogy; this is further illustrated by our findings that most of the studies were in the research phase and not deployed in a clinical setting. Therefore, while there is significant interest and potential in clinical AI/ML models, there are still challenges and barriers to their widespread adoption in real-world clinical settings.

Generalizability is defined as the ability of a model to perform well on datasets that have different characteristics from training data[72]. Most AI/ML models developed in academic centers are trained on homogenous patient populations that do not reflect subpopulations found in non-academic institutions and would generalize poorly in other

settings. External validation can be used to mitigate poor generalizability by evaluating model performance on datasets not used to develop the model[73]. In our review, only 27% of the studies externally validated their models.

Certain subspecialties, such as cardiology, neurology, respiratory medicine, infectious diseases, and genitourinary medicine, were the most represented in terms of the number of AI/ML models developed. These findings highlight the potential impact of AI/ML models in these subspecialties and the need for further research in other areas of healthcare. Our analysis revealed that a significant number of studies on clinical AI/ML models were conducted in China, Taiwan, the European Union, the United States, Japan, and Korea, indicating a global interest and involvement in the development of these models but also an increasing gap between developed countries with economic resources/infrastructure and low resource countries. With most AI tools being developed in these regions, the training data excludes population characteristics found in these low resource regions, thereby exacerbating unintended bias and potentiating poor outcomes. Furthermore, our study found that only a small percentage (16%) of the AI/ML models utilized open data registries, highlighting the need for improved data sharing and accessibility in the field of clinical AI/ML. Proprietary/closed registry data negate the ability to independently validate models, thus going against the cardinal principle of replicability in research.

Our study showed that a low percentage (6%) of the total studies were randomized controlled trials while 32% of studies accurately identified their primary outcomes, suggesting a need for more rigorous study designs to evaluate the effectiveness of clinical AI/ML models. Only 12% of the studies reported following the best practice standard for model development, while 53% of studies did not explain how they dealt with missing data. Handling of missing data can have a large impact on the AI/ML models outcomes especially when used in healthcare[73]. In terms of model updating, our findings indicate that a small proportion of studies (9%) made attempts to ensure that AI/ML models would be updated in the future through recommendations or guidelines. There was a statistically significant positive relationship between our model quality scores and models declaring a model updating method. Quality scores were not affected by location of study (academic/non-academic) or by geographic location. When controlling location of study and nature of model (predictive/prognostic), studies that recommended model updating had higher quality scores.

In terms of gender and ethnic composition, our study found that, worldwide, 31% did not disclose gender and 84% did not disclose ethnicity. In the US, 23% of the studies did not disclose gender, and 44% did not provide ethnicity composition. However, due to the homogeneity of populations outside of the US, the lack of disclosure for ethnicity is not as surprising. Nevertheless, we believe reporting gender and ethnicity composition in US models is still sub-par considering the diverse nature of the US population.

STRENGHTS AND LIMITATIONS

There is no standardized, accepted simple checklist for clinical model assessment. Our results suggest a positive relationship between studies that recommended model updating and higher quality of study scores based on our modified CHARMS checklist tool. Based on this identified relationship our modified checklist may be validated further to serve as a quick validation tool for clinical models amongst clinicians who are not invested in the rigors of data science but rather more interested in the utility, safety, and applicability of a potential algorithm.

With healthcare AI rapidly advancing toward agentic systems—AI that operates with high levels of independence and autonomy—we are entering dangerous territory without adequate safeguards. Clinicians urgently need accessible, standardized assessment tools to evaluate these systems before deployment. Without such checklists, we risk deploying AI agents that make critical clinical decisions without proper validation, potentially leading to

preventable patient harm and medical errors. The time to act now, before autonomous AI systems are integrated into clinical workflows without the safety mechanisms necessary to protect patients.

Study Limitations

Interpretation of our review should bear some limitations in mind. First, AI and ML implementations in health care are novel and lack standardization across different regions and clinical specialty domains. Although we established our literature search strategy (**Appendix A**), this lack of standards can impact on the scope and sensitivity of our search and render the reproducibility of our review challenging. While the terms "AI" and "ML" are included in our search, terms used to describe models and modeling are not standardized, and therefore, it is possible that our strategy did not capture possible emerging or lesser-known terms. In addition, our search included only English-language publications, so we cannot generalize our findings to publications in other languages. Book chapters, theses, short papers, editorials, non–peer-reviewed reports, or conference abstracts were also not included. Another factor to consider in interpreting our results is that the studies we reviewed were published during the global COVID-19 pandemic. The impact of the pandemic on nature and type of AI and ML studies published during this time is unknown.

An important limitation to bear in mind is that we excluded 80 studies that were proprietary and did not disclose their AI/ML methodologies. Based on this, we are unable to verify if those studies considered model updating in their model development due to their opaque reporting. We cannot verify the extent to which their models would have skewed our result if they had actually reported their methodologies.

Conclusion

In conclusion, model updating is an essential part of the maintenance of a model to ensure optimal output during implementation. Contemporary models heavily ignore this important process and can adversely affect patient safety. There is a need to report the breakdown of gender and ethnicity data used to build models. Without this disclosure, there will likely be a worsening of gender and racial disparity in the implementation of models similar to what it obtains in biomedical therapeutics and device development.

There are no consensus-accepted standards for evaluating and screening proposed clinical models at the bedside. Subject to further validation, our modified CHARMS checklist may serve as a quick screening tool for clinicians who are not savvy data scientists. Future considerations include validating the modified CHARMS checklist to confirm its applicability for different healthcare models. There is need for more models assessing long-term prognosis or patient trajectories rather than the present excess of models that predict patient immediate conditions like sepsis. Additionally, we hope that moving forward more clinical models will utilize open data registries as training data for ease of independent verification and enhanced applicability of the model.

Acknowledgements

We would like to acknowledge Martin, Lily librarian at the Levy Library at the Icahn school of Medicine in Mount Sinai, New York for her assistance with our search and setting up Covidence software. We would also like to acknowledge the following investigators who assisted in our preliminary data abstraction.

- Rasheedat A. sadiq-onilenla MD. MBA. MPH. Msc
- Maxwell Edomwande MBBS, MBA, CDIP, LSSBB (Umass Amherst)
- Osazuwa Ighodaro, MBBS, Ekpoma
- Gloria Yeesuf, MB, BS, MPH, MBA
- Omotayo Olusola, MBBS, M.A, MPH
- Soji Akin Ojo, MD - Pharmaceutical Product Development (PPD), Thermo Fisher Scientific, Wilmington, NC, United States.

We also like to acknowledge the following subject matter experts who were the source of inspiration for conducting this review;

- Robert Freeman, RN; Division of Data-Driven and Digital Medicine, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, New York
- Matthew Levin, MD, PhD; Department of Anesthesiology, Perioperative and Pain Medicine, Icahn School of Medicine at Mount Sinai, New York, New York

Data Availability

The data that support the findings of this study are available from the corresponding author [AO] upon reasonable request.

References:

1. Matheny ME, Whicher D, Thadaney Israni S. Artificial Intelligence in Health Care: A Report From the National Academy of Medicine. *JAMA*. 2020; 323(6):509-510. doi:10.1001/jama.2019.21579

2. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is Machine Learning? A Primer for the Epidemiologist. *Am J Epidemiol*. 2019;188(12):2222-2239. doi:10.1093/aje/kwz189

3. Navarro CLA, Damen JAAG, Takada T, et al. Protocol for a systematic review on the methodological and reporting quality of prediction model studies using machine learning techniques. *BMJ Open*. 2020;10(11):e038832. doi:10.1136/bmjopen-2020-038832

4. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. *BMC Med Res Methodol*. 2019;19:64. doi:10.1186/s12874-019-0681-4

5. Brnabic A, Hess LM. Systematic literature review of machine learning methods used in the analysis of real-world data for patient-provider decision making. *BMC Med Inform Decis Mak*. 2021;21(1):54. doi:10.1186/s12911-021-01403-2

6. Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230. doi:10.1186/s12916-019-1466-7

7. Koola JD, Ho SB, Cao A, et al. Predicting 30 Day Hospital Readmission Risk in a National Cohort of Patients with Cirrhosis. *Dig Dis Sci*. 2020;65(4): 1003-1031. doi:10.1007/s10620-019-05826-w

8. Moons KGM, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart*. 2012;98 (9):691-698. doi:10.1136/heartjnl-2011-301247

9. Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol*. 2008;61 (1):76-86. doi:10.1016/j.jclinepi.2007.04.018

10. Steyerberg EW, Borsboom GJJM, van Houwelingen HC, Eijkemans MJC, Habbema JDF. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med*. 2004;23(16):2567-2586. doi:10.1002/sim.1844

11. Counsell C, Dennis M. Systematic Review of Prognostic Models in Patients with Acute Stroke. *Cerebrovasc Dis*. 2001;12(3):159-170. doi:10.1159/000047699

12. Prediction Models for Prolonged Intensive Care Unit Stay After Cardiac Surgery | Circulation. Accessed August 31, 2025. https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.109.926808?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed

13. Perel P, Edwards P, Wentz R, Roberts I. Systematic review of prognostic models in traumatic brain injury. *BMC Med Inform Decis Mak*. 2006;6:38. doi:10.1186/1472-6947-6-38

14. Phung MT, Tin Tin S, Elwood JM. Prognostic models for breast cancer: a systematic review. *BMC Cancer*. 2019;19:230. doi:10.1186/s12885-019-5442-6

15. Saria S, Subbaswamy A. Tutorial: Safe and Reliable Machine Learning. *arXiv*. Preprint posted online April 15, 2019. doi:10.48550/arXiv.1904.07204

16. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol*. 2015;68(3):279-289. doi:10.1016/j.jclinepi.2014.06.018

17. Kappen TH, Vergouwe Y, van Klei WA, van Wolfswinkel L, Kalkman CJ, Moons KGM. Adaptation of Clinical Prediction Models for Application in Local Settings. *Med Decis Making*. 2012;32(3):E1-E10. doi:10.1177/0272989X12439755

18. Schulam P, Saria S. Can You Trust This Prediction? Auditing Pointwise Reliability After Learning. In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. PMLR; 2019:1022-1031. Accessed August 31, 2025. https://proceedings.mlr.press/v89/schulam19a.html

19. Shah ND, Steyerberg EW, Kent DM. Big Data and Predictive Analytics: Recalibrating Expectations.

JAMA. 2018;320(1):27-28. doi:10.1001/jama.2018.5602

20. Toll DB, Janssen KJM, Vergouwe Y, Moons KGM. Validation, updating and impact of clinical prediction rules: a review. *J Clin Epidemiol*. 2008;61(11):1085-1094. doi:10.1016/j.jclinepi.2008.04.008

21. Davis SE, Greevy RA Jr, Fonnesbeck C, Lasko TA, Walsh CG, Matheny ME. A nonparametric updating method to correct clinical prediction model drift. *J Am Med Inform Assoc*. 2019;26(12):1448-1457. doi:10.1093/jamia/ocz127

22. Lipsitz LA. Understanding Health Care as a Complex System: The Foundation for Unintended Consequences. *JAMA*. 2012;308(3):243-244. doi:10.1001/jama.2012.7551

23. Schulam P, Saria S. Reliable decision support using counterfactual models. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Curran Associates Inc.; 2017:1696-1706.

24. Hall MA, Lord R. Obamacare: what the Affordable Care Act means for patients and physicians. *BMJ*. 2014;349:g5376. doi:10.1136/bmj.g5376

25. Kaushal A, Altman R, Langlotz C. Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. *JAMA*. 2020;324(12):1212-1213. doi:10.1001/jama.2020.12067

26. Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. A Review of Challenges and Opportunities in Machine Learning for Health. *AMIA Summits Transl Sci Proc*. 2020;2020:191-200.

27. Constructing the world: Active causal learning in cognition | Bramley Computational Cognitive Science Lab. Accessed August 31, 2025. https://www.bramleylab.ppls.ed.ac.uk/publication/2017-01-01_bramley2017phdthesis/

28. How ChatGPT and our foundation models are developed. OpenAI Help Center. Accessed August 31, 2025. https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed

29. The Internet May Be Too Small for the AI Boom, Researchers Say - The Wall Street Journal Google Your News Update - WSJ Podcasts. The Wall Street Journal. Accessed August 31, 2025. https://www.wsj.com/podcasts/google-news-update/the-internet-may-be-too-small-for-the-ai-boom-researchers-say/a424f137-a5a4-46b7-b746-c7fd3d0a483d

30. Yu P, Xu H, Hu X, Deng C. Leveraging Generative AI and Large Language Models: A Comprehensive Roadmap for Healthcare Integration. *Healthcare*. 2023;11(20):2776. doi:10.3390/healthcare11202776

31. Busch F, Hoffmann L, Rueger C, et al. Current applications and challenges in large language models for patient care: a systematic review. *Commun Med*. 2025;5:26. doi:10.1038/s43856-024-00717-2

32. Kwong JCC, Wang SCY, Nickel GC, Cacciamani GE, Kvedar JC. The long but necessary road to responsible use of large language models in healthcare research. *Npj Digit Med*. 2024;7(1):177. doi:10.1038/s41746-024-01180-y

33. Towards Urban Planing AI Agent in the Age of Agentic AI. Accessed August 31, 2025. https://arxiv.org/html/2507.14730

34. White J. Building Living Software Systems with Generative & Agentic AI. *arXiv*. Preprint posted online August 3, 2024. doi:10.48550/arXiv.2408.01768

35. Guajardo JA, Weber R, Miranda J. A model updating strategy for predicting time series with seasonal patterns. *Appl Soft Comput*. 2010;10(1):276-283. doi:10.1016/j.asoc.2009.07.005

36. Davis SE, Greevy RA, Lasko TA, Walsh CG, Matheny ME. Detection of calibration drift in clinical prediction models to inform model updating. *J Biomed Inform*. 2020;112:103611. doi:10.1016/j.jbi.2020.103611

37. Singh A, Pandey N, Shirgaonkar A, Manoj P, Aski V. A Study of Optimizations for Fine-tuning Large Language Models. *arXiv*. Preprint posted online June 6, 2024. doi:10.48550/arXiv.2406.02290

38. Gao Y, Xiong Y, Gao X, et al. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv*. Preprint posted online March 27, 2024. doi:10.48550/arXiv.2312.10997

39. Lewis P, Perez E, Piktus A, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: *Advances in Neural Information Processing Systems*. Vol 33. Curran Associates, Inc.; 2020:9459-9474. Accessed August 31, 2025. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

40. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. *arXiv*. Preprint posted online March 4, 2022. doi:10.48550/arXiv.2203.02155

41. González Barman K, Lohse S, de Regt HW. Reinforcement Learning from Human Feedback in LLMs: Whose Culture, Whose Values, Whose Perspectives? *Philos Technol*. 2025;38(2):35. doi:10.1007/s13347-025-00861-0

42. van Stein N, Vermetten D, V. Kononova A, Bäck T. Explainable Benchmarking for Iterative Optimization Heuristics. *ACM Trans Evol Learn Optim*. 2025;5(2): 13:1-13:30. doi:10.1145/3716638

43. Mitrevski A, Plöger PG, Lakemeyer G. Representation and Experience-Based Learning of Explainable Models for Robot Action Execution. In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2020:5641-5647. doi:10.1109/IROS45743.2020.9341470

44. de Hond AAH, Leeuwenberg AM, Hooft L, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *Npj Digit Med*. 2022;5(1):2. doi:10.1038/s41746-021-00549-7

45. Shamseer L, Moher D, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ*. 2015;349:g7647. doi:10.1136/bmj.g7647

46. Otokiti AU, Ozoude MM, Williams KS, et al. The Need to Prioritize Model-Updating Processes in Clinical Artificial Intelligence (AI) Models: Protocol for a Scoping Review. *JMIR Res Protoc*. 2023;12(1): e37685. doi:10.2196/37685

47. Bell SK, Delbanco T, Elmore JG, et al. Frequency and Types of Patient-Reported Errors in Electronic Health Record Ambulatory Care Notes. *JAMA Netw Open*. 2020;3(6):e205867. doi:10.1001/jamanetworkopen.2020.5867

48. Diaz-Garelli JF, Strowd R, Wells BJ, Ahmed T, Merrill R, Topaloglu U. Lost in Translation: Diagnosis Records Show More Inaccuracies After Biopsy in Oncology Care EHRs. *AMIA Summits Transl Sci Proc*. 2019;2019:325-334.

49. Tse J, You W. How Accurate is the Electronic Health Record? – A Pilot Study Evaluating Information Accuracy in a Primary Care Setting. In: *Health Informatics: The Transformative Power of Innovation*. IOS Press; 2011:158-164. doi:10.3233/978-1-60750-791-8-158

50. Zou J, Schiebinger L. Ensuring that biomedical AI benefits diverse populations. *eBioMedicine*. 2021;67. doi:10.1016/j.ebiom.2021.103358

51. Moons KGM, de Groot JAH, Bouwmeester W, et al. Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. *PLoS Med*. 2014;11(10): e1001744. doi:10.1371/journal.pmed.1001744

52. Kim AA, Rachid Zaim S, Subbian V. Assessing reproducibility and veracity across machine learning techniques in biomedicine: A case study using TCGA data. *Int J Med Inf*. 2020;141:104148. doi:10.1016/j.ijmedinf.2020.104148

53. Li J, Liu L, Le TD, Liu J. Accurate data-driven prediction does not mean high reproducibility. *Nat Mach Intell*. 2020;2(1):13-15. doi:10.1038/s42256-019-0140-2

54. Stevens LM, Mortazavi BJ, Deo RC, Curtis L, Kao DP. Recommendations for Reporting Machine Learning Analyses in Clinical Research. *Circ Cardiovasc Qual Outcomes*. 2020;13(10):e006556. doi:10.1161/CIRCOUTCOMES.120.006556

55. Bouwmeester W, Zuithoff NPA, Mallett S, et al. Reporting and Methods in Clinical Prediction Research: A Systematic Review. *PLoS Med*. 2012;9(5):e1001221. doi:10.1371/journal.pmed.1001221

56. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD):

the TRIPOD Statement. *BMC Med*. 2015;13(1):1. doi:10.1186/s12916-014-0241-z

57. Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. *BMJ*. 2020;370:m3210. doi:10.1136/bmj.m3210

58. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health*. 2019;1(6):e271-e297. doi:10.1016/S2589-7500(19)30123-2

59. Collins GS, Dhiman P, Navarro CLA, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. 2021;11(7):e048008. doi:10.1136/bmjopen-2020-048008

60. Altman DG, McShane LM, Sauerbrei W, Taube SE. Reporting recommendations for tumor marker prognostic studies (REMARK): explanation and elaboration. *BMC Med*. 2012;10:51. doi:10.1186/1741-7015-10-51

61. Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Lancet Digit Health*. 2020;2(10):e549-e560. doi:10.1016/S2589-7500(20)30219-3

62. Janssens ACJW, Ioannidis JPA, Bedrosian S, et al. Strengthening the reporting of genetic risk prediction studies (GRIPS): explanation and elaboration. *Eur J Epidemiol*. 2011;26(4):313-337. doi:10.1007/s10654-011-9551-z

63. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ*. 2010;340:c869. doi:10.1136/bmj.c869

64. Steyerberg EW, Moons KGM, Windt DA van der, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. *PLOS Med*. 2013;10(2):e1001381. doi:10.1371/journal.pmed.1001381

65. Mallett S, Royston P, Dutton S, Waters R, Altman DG. Reporting methods in studies developing prognostic models in cancer: a review. *BMC Med*. 2010;8(1):20. doi:10.1186/1741-7015-8-20

66. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med*. 2000;19(4):453-473. doi:10.1002/(sici)1097-0258(20000229)19:4<453::aid-sim350>3.0.co;2-5

67. Dieren S van, Beulens JWJ, Kengne AP, et al. Prediction models for the risk of cardiovascular disease in patients with type 2 diabetes: a systematic review. *Heart*. 2012;98(5):360-369. doi:10.1136/heartjnl-2011-300734

68. Reilly BM, Evans AT. Translating Clinical Research into Clinical Practice: Impact of Using Prediction Rules To Make Decisions. *Ann Intern Med*. 2006;144(3):201-209. doi:10.7326/0003-4819-144-3-200602070-00009

69. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD Statement for Reporting Studies of Diagnostic Accuracy: Explanation and Elaboration. *Ann Intern Med*. 2003;138(1):W1-12. doi:10.7326/0003-4819-138-1-200301070-00012-w1

70. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *PLOS Med*. 2009;6(7):e1000100. doi:10.1371/journal.pmed.1000100

71. Artificial Intelligence in Healthcare. O'Reilly Online Learning. Accessed August 31, 2025. https://www.oreilly.com/library/view/artificial-intelligence-in/9780128184394/

72. Lu C, Ahmed SR, Singh P, Kalpathy-Cramer J. Estimating Test Performance for AI Medical Devices under Distribution Shift with Conformal Prediction. *arXiv*. Preprint posted online July 12, 2022. doi:10.48550/arXiv.2207.05796

73. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol*. 2014;14(1):40. doi:10.1186/1471-2288-14-40

