nature health

Analysis

https://doi.org/10.1038/s44360-025-00048-z

A scoping review of silent trials for medical artificial intelligence

Received: 27 October 2025

Accepted: 15 December 2025

Published online: 16 February 2026

Check for updates

A list of authors and their affiliations appears at the end of the paper

A 'silent trial' refers to the prospective, noninterventional testing of artificial intelligence (AI) models in the intended clinical setting without affecting patient care or institutional operations. The silent evaluation phase has received less attention than in silico algorithm development or formal clinical evaluations, despite its increasing recognition as a critical phase. There are no formal guidelines for performing silent AI evaluations in healthcare settings. We conducted a scoping review to identify silent AI evaluations described in the literature and to summarize current practices for performing silent testing. We screened the PubMed, Web of Science and Scopus databases for articles fitting our criteria for silent AI evaluations, or silent trials, published from 2015 to 2025. A total of 891 articles were identified, of which 75 met the criteria for inclusion in the final review. We found wide variance in terminology, description and rationale for silent evaluations, leading to substantial heterogeneity in the reported information. Overwhelmingly, the papers reported measurements of area under the curve and similar metrics of technical performance. Far fewer studies reported verification of outputs against an in situ clinical ground truth; when reported, the approaches varied in comprehensiveness. We noted less discussion of sociotechnical components, such as stakeholder engagement and human–computer interaction elements. We conclude that there is an opportunity to bring together diverse evaluative practices (for example, from data science, human factors and other fields) if the silent evaluation phase is to be maximally effective. These gaps mirror challenges in the effective translation of AI tools from computer to bedside and identify opportunities to improve silent evaluation protocols that address key needs.

Despite the increasing deluge of papers describing the development of artificial intelligence (AI) models for healthcare applications, strikingly few of those models have proceeded to clinical use[1]. A translational gap[2] remains, partially due to the substantial difference between building a model that works in silico (that is, validation within a dataset) and creating one that is clinically useful, actionable and beneficial to patients or the healthcare system[3].

One mechanism for bridging the translational gap is conducting an evaluation following algorithmic validation, but before the clinical evaluation of the model in practice. This phase is known as a 'silent trial' (a term with many variants, including 'shadow evaluation' or 'silent testing') and is common practice among many healthcare institutions with advanced internal AI teams[4,5]. 'Silent' traditionally refers to the notion that the model's outputs are produced in parallel to (and thus separate from) the standard of care; therefore, they do not influence clinicians (Table 1).

Primarily, the silent phase of AI development is used to ascertain whether the model will maintain its performance in a live context[6]. The value of this phase is that it allows teams not only to test a model for potential utility (data pipeline stability and model drifts, among other

e-mail: lana.tikhomirov@adelaide.edu.au

https://doi.org/10.1038/s44360-025-00048-z

Table 1 | Range of definitions and nomenclature given to silent trials

Study type	Definition
Prospective clinical validation study (modern silent evaluation)	A prospective algorithmic validation involving an assessment of the model's predictions against live expert annotations to verify facts about the patient or outcome of interest. Separation is maintained between care and model evaluation.
Prospective algorithmic validation (traditional silent trial)	Running the model live while maintaining a separation between care and model evaluation; assessing model performance but not assessing against live annotations of real-world information beyond the data obtained
Prospective validation study (internal validation)	Conducting a cross-sectional assessment of a model's performance
Prospective observational study	Integrated into the clinical system; may or may not be observable to clinical users
Temporal validation	Prospective algorithmic validation with a particular focus on the model's performance over time

concerns; see the glossary in Box 1) but also to assess the financial sustainability of models in real-world evaluations without affecting care or operation[7]. During this stage, teams can make informed decisions about whether to discard a model, iteratively improve its performance or move to deployment based on local evidence[8].

The importance of local evidence is perhaps more relevant to AI tools than to historical healthcare interventions. While we would not expect the performance of a drug or device to change substantially when tested in a hospital across the street with the same patient population, this is indeed the case for AI models[6,8,9]. Even for models that have received regulatory clearance or approval based on clinical evidence, substantial differences may be apparent in local performance such that their reliability may vary across settings[10,11]. Researchers have noted the challenges of bringing AI systems to market based solely on retrospective evidence[12,13]. The silent evaluation stage may represent a low-risk bridge between retrospective and clinical evidence that may help developers decide whether a clinical trial is warranted. The regulatory science of AI involves the important consideration of which types of evidence are acceptable for determining the safety of AI as a medical device. The silent phase of translation offers a low-risk testing paradigm that reflects real-world conditions by which one might judge the performance of an algorithm. This may be a critical step before determining whether (and what type of) clinical trials should be pursued—a judgement that may be made by regulatory professionals, ethics committees or AI oversight bodies.

Given that the silent phase of AI testing offers an opportunity to evaluate performance locally using precise metrics relevant to the population and institution, yet does not affect care (thus minimizing risk to health institutions and patients alike), it is perhaps surprising that this key phase does not receive more attention. Silent trials have equivalents in other fields (for example, beta testing in software engineering, silent review in aviation, and simulations in training, which are standard practices), but, to the best of our knowledge, no reporting guidelines or authoritative publications have addressed the silent phase in medical AI. Our project group, the Collaboration for Translational AI Trials (CANAIRI), has a particular focus on building knowledge and best practices around the silent phase to facilitate local capacity-building in AI evaluations and to demonstrate accountable AI integration[14]. We conducted a scoping review and critical analysis[15] to explore the literature around the following key points: (1) How is the silent phase defined, described and justified? (2) What practices are being undertaken during this phase? (3) What are the implications of the latter in relation to the larger goal of responsibly translating AI into healthcare systems? Scoping reviews map the existing literature on a topic, identify knowledge gaps and clarify concepts. We find this method valuable because we are addressing a nascent paradigm in AI with the goal of synthesizing and reflecting on the available literature. This Analysis aims to bring clarity and consistency to the silent

BOX 1

Glossary of terms

Algorithmic bias: a systematic discrepancy in a model's performance based on a feature that would be considered unfair in relation to non-clinically relevant constructs

Automation bias: over-reliance of human decision-making on an AI model or system, leading to preventable consequences

Contextualized subgroups of interest: a group of individuals with shared relevant attributes that have known or suspected associations with disparate health outcomes related to the intended use of an AI health technology

Data drift: a usually unanticipated change in the statistical properties of a model that affects its performance

Data pipeline: the complete pathway by which information flows from its point of entry into a system to the output of that system

Data preprocessing: methods for addressing consistency and quality among data elements before training

Failure modes: systematic patterns of error in relation to a specific metric (for example, false positives)

Feature selection: the choice of model inputs

Human adaptation: a change in human behaviour in response to the presence of an AI system

Human factors: aspects pertaining to the user of technology that can affect how the technology is perceived, integrated, vetted for errors and used in a wider system

Incidental findings: the identification of an imminent and potentially harmful error in relation to a specific patient, which could prevent harm if acted on

Model downtime: the time when the model is unavailable unexpectedly due to technical issues

Scalability: whether an algorithm's use can be expanded to the entire context of its intended use

Silent: the model's outputs do not influence the act of care for patients or operational systems

Sociotechnical system: the wider system in which algorithms exist—involving human expertise; the coordination of different healthcare professionals, infrastructures and technical systems; and patient considerations

Sociotechnical: the interdependence between technology and humans

Temporal generalizability: an algorithm's applicability to new, incoming data prospectively

Verification: the process of manually or computationally assessing individual model outputs against a 'ground truth' label—whether a label captured in the health record or another clinical system—by expert evaluation (for example, reader studies), or an expert or group of experts selected to conduct a manual review

https://doi.org/10.1038/s44360-025-00048-z

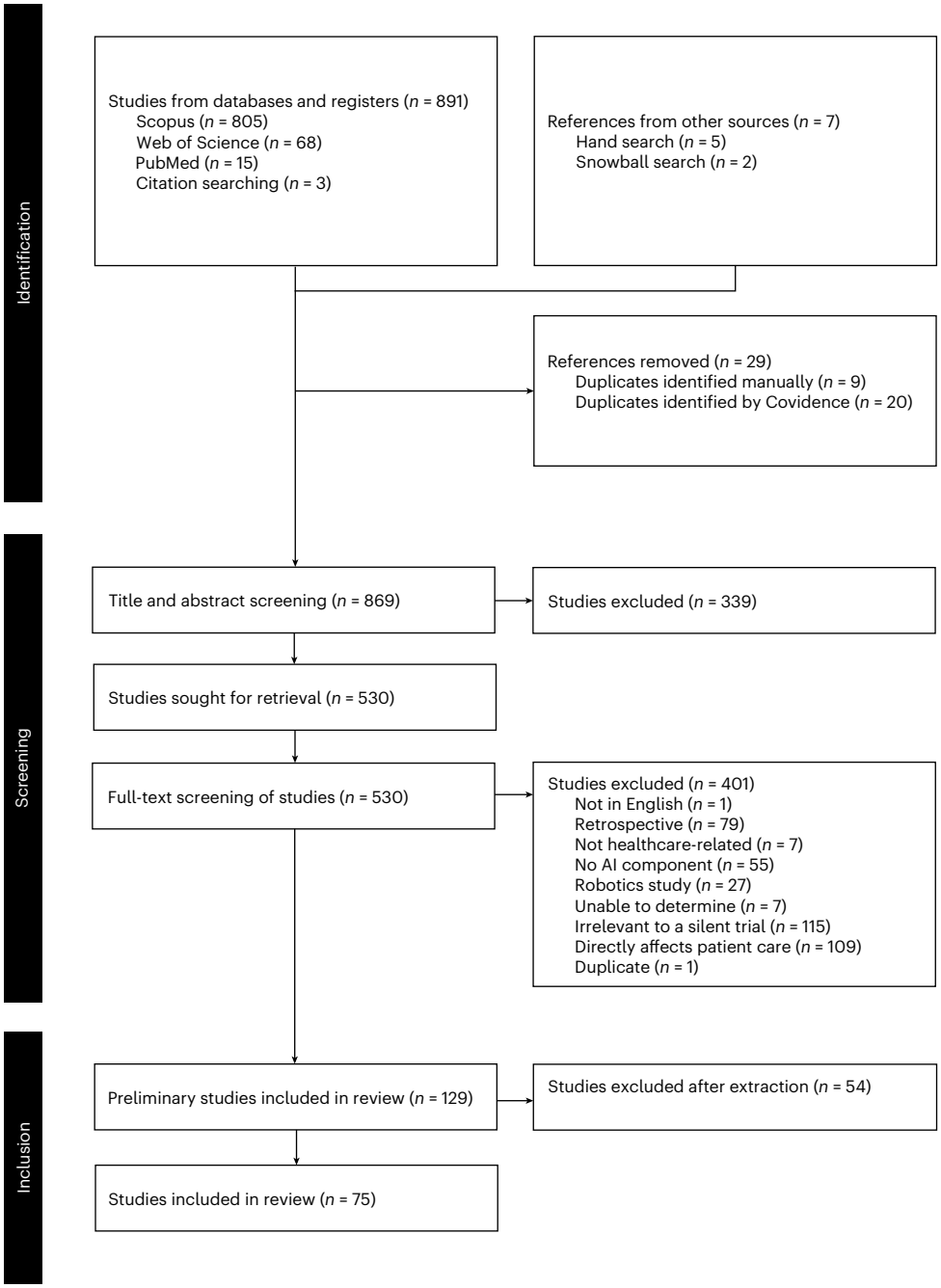

Fig. 1 | PRISMA diagram showing the identification of evidence sources from database searches and hand search methods. Following the data charting process, a further 54 papers did not meet the criteria.

phase while considering the implications of current practices for AI translation efforts.

Results

From September 2024 to October 2025, we scoped the published literature for primary research studies published in English that describe testing an AI model in a manner closely mimicking its intended use but without modifications to the standard of care, to validate the model in a 'live' context. From a total of 898 papers, we removed duplicates ($n = 29$) and screened 530 full-text articles for inclusion (Fig. 1). After excluding papers that did not describe a true live validation study, those involving substantial alterations to patient care, those with insufficient detail for us to assess the silent component of their study and those that did not involve an AI tool, we finally included 75 studies.

We then looked for papers related to the AI tools evaluated in that set of 75 studies. We identified six additional studies that provided further details about the silent evaluation. Of these, two[16,17] contained information about the original silent phase evaluation that was included in data charting, while four others[18–21] explored the later clinical, stakeholder or human factors impacts of the algorithm after the silent evaluation, during its integration into patient care. As our unit of analysis is the silent phase itself, we combined only the information retrieved about the practices undertaken during the silent phase, excluding postdeployment work. Therefore, we incorporated the information extracted from the first two papers and did not include the latter four, as they were conducted while the model was not silent (that is, live), thus falling under the exclusion criteria. The results of data charting are summarized in Table 2.

https://doi.org/10.1038/s44360-025-00048-z

Table 2 | General information about the included silent studies

Study	Aim and rationale	Model type and intended use	Model evaluation	Additional considerations	Categorization
Aakre et al. (2017)[21]	To assess an automated SOFA score calculation for patients in the ICU	Predictive machine learning	• Agreement between automated SOFA scoring and manual scoring calculation over a 1-month period • Comparison of 215 ICU inpatients' SOFA scores at 3 hospital sites, with 5,978 total scores compared • 134 random spot checks on 27 unique patients to assess the real-time accuracy of automated SOFA score calculation • Manual scoring performed independently by research team members, with a chart review for comparison	Interviewed clinicians about interface features to visualize SOFA subcomponents	Compared model outputs with clinician annotations
Afshar et al. (2023)[28]	To assess the AI tool's predictive performance and evaluative human factors	Predictive deep learning	• Algorithm performance: sensitivity and specificity • Observed 100 random encounters with adult patients • Described data flow from and to the EHR • Described scalability and computational infrastructure	• Interview guide and survey to assess user acceptability of the tool • Determined barriers and facilitators to using the tool	Framework for the design and implementation of the model
Alrajhi et al. (2022)[75]	To assess a real-time severity prediction tool for COVID-19 management	Predictive machine learning	• Algorithm performance: AUC/ROC, F1 • 185 cases for the prospective validation set • Imputed missing data; addressed class imbalances	Clinician feedback related to class imbalance issue	Algorithmic validation study
Aydın et al. (2025)[76]	To validate and compare an ML-based scoring system for paediatric appendicitis	Diagnostic machine learning	• Algorithm performance: AUC, sensitivity, specificity, PPV, NPV • Applied to 3,036 paediatric patients across 13 hospitals and 13 paediatric centres • ML-based diagnosis assessed against histopathological examination (gold standard) • Compared ML model performance against existing scoring methods	• Specified separation of care and model validation • Assessed feature interactions and ranked importance	Algorithmic validation, comparative study
Bachelot et al. (2023)[77]	To compare model performance for testicular sperm extraction	Predictive machine learning	• Algorithm performance: AUC, sensitivity, specificity • 26 patients for the prospective validation set • Described data processing	Assessed feature importance across models	Algorithmic validation study
Bedoya et al. (2020)[39]	To validate a sepsis prediction model	Diagnostic deep learning	• Algorithm performance: compared with standard EWS, compared multiple models with the standard process • 1,475 encounters over a 2-month silent trial • Model development team tracked alarm volume, resolved technical issues and identified label leakage • Calculated alarm volume	Stakeholder engagement with clinical teams used	Comparison of the model with the standard-of-care algorithm
Berg et al. (2023)[78]	To assess an AI software for classifying palpable breast masses in a low-resource setting	Predictive AI	• Algorithm performance: AUC, specificity, NPR • 758 masses in breast tissue • A single radiologist reader reviewed AI- and radiologist-assigned malignancies • Minimal training for users to mimic the conditions of intended use		Compared diagnostic performance with human readers
Brajer et al. (2020)[36]	To assess the model's ability to predict the risk of in-hospital mortality for adult patients	Predictive machine learning	• Algorithm performance: ROC, PR, AUROC • 5,273 hospitalizations, 4,525 unique adult patients in the ICU • Assessed subgroup-specific performance for sensitivity, specificity and PPV • Assessed threshold setting in different environments • Described data and model availability; updated predictions daily	• Partnered with clinical and operational leaders to design the model and evaluation • Clinical partners provided feedback into the interface • Model fact sheet iteratively designed with stakeholder input	Compared algorithmic prediction with human annotations
Butler et al. (2019)[79]	To clinically validate an AI tool for triaging brain cancer	Triage machine learning	• Algorithm performance: sensitivity, specificity • 104 patients with brain cancer • Outcome assessment was blinded to the algorithm • Some subgroup-specific analysis of under-represented cancer cases	Simulated workflow run within a research laboratory	Compared algorithmic prediction with independent clinician diagnosis

https://doi.org/10.1038/s44360-025-00048-z

Table 2 (continued) | General information about the included silent studies

Study	Aim and rationale	Model type and intended use	Model evaluation	Additional considerations	Categorization
Campanella et al. (2025)[80]	To conduct a prospective silent trial of a model for lung cancer detection	Predictive machine learning	• Algorithm performance: AUC, PPV, NPV, sensitivity, specificity • Application of an open-source foundation model with local fine-tuning • 4-month trial period • Subgrouped analysis by sample type, failure mode testing of false negatives • Assessed different thresholds against primary metrics • Described data pipeline and real-time stream	Assessed the attention areas of the model	Prospective silent trial
Chen et al. (2025)[81]	To evaluate the utility of a radiomics nomogram to predict oesophageal pathological progression	Predictive machine learning	• Algorithm performance: AUC, sensitivity, specificity, accuracy, DCA • 251 cases • Ground truth was reviewed by a pathologist and compared and combined with the model for overall clinical utility • Described the need for preprocessing due to equipment differences	DCA for utility	Clinical validation
Cheng et al. (2025)[82]	To prospectively validate a hypertension risk model	Predictive machine learning	• Algorithm performance: AUC, precision, sensitivity, specificity, calibration curves • 961,519 cases • Assessed fairness across age and sex, BMI across different risk levels, model performance, and socioeconomic factors in the high-risk group • Discussed managing data missingness and shift	Clinician-focused app to provide clinicians an opportunity to assess prediction utility and risk factor contributions	Algorithmic validation
Chiang et al. (2025)[83]	To prospectively validate an early warning haemodynamic risk model	Predictive machine learning	• Algorithm performance: AUROC, AUPRC, precision, recall, specificity, false alarm rate and missed alarm rate • 18,438 patient cases • Assessed sex and age, as well as respiratory, cardiovascular, gastrointestinal and trauma groups on AUROC and AUPRC • Model updates hourly		Algorithmic validation
Chufal et al. (2025)[84]	To prospectively and temporally validate a model predicting ineligibility for radiotherapy treatment	Predictive machine learning	• Algorithm performance: AUC • 47 patients • Compared model prediction with clinical decision on a case-by-case basis, with only the research team seeing the model predictions • Noted fairness concerns by sociodemographic groups; stated that these were addressed through consistency in the assessment method	Discussion of threshold setting based on clinical impact to patients and risk assessment	Prospective algorithmic validation with clinical verification
Coley et al. (2021)[85]	To assess an algorithm's predictive accuracy of suicide attempt within 90 days	Predictive machine learning	• Algorithm performance: sensitivity, specificity, PPV, NPV • Prospective algorithmic validation concurrent with the testing set		Temporal validation, internal algorithmic validation
Corbin et al. (2023)[86]	To conduct a silent trial of the model's prospective performance	Predictive machine learning	Algorithm performance: AUROC, ROC, calibration, net benefit, expected utility • 10,000–20,000 unique patients • Bias assessed across protected demographic classes • Mapping of data inputs to outputs across the data stream workflow		Prospective algorithmic validation
Dave et al. (2023)[87]	To evaluate the accuracy of a real-time model detecting abnormal lung parenchyma	Predictive deep learning	• Algorithm performance: AUROC, F1 • 100 patients, sample size rationale provided • Analysed by sex, race, ventilation strategy and BMI • Functionality embedded into an ultrasound machine • Assessed different classification and contiguity thresholds • Human assessment independent from predictions		Compared algorithmic prediction with human annotations

https://doi.org/10.1038/s44360-025-00048-z

Table 2 (continued) | General information about the included silent studies

Study	Aim and rationale	Model type and intended use	Model evaluation	Additional considerations	Categorization
El Moheb et al. (2025)[88]	To validate a model for automated billing coding	Administrative deep learning	• Algorithm performance: precision, recall, F1, AUPRC • 268 operative notes • Trained to predict 19 CPT codes for automated coding, compared with expert medical coders • Assessed overcoding and undercoding, as well as discrepancies against ground truth		Prospective algorithmic validation study
Escalé-Besa et al. (2023)[24]	To validate a model's diagnostic accuracy for skin diseases	Diagnostic deep learning	• Algorithm performance: accuracy, sensitivity, specificity per disease; TP, FP, TN or FN based on the top 3 most likely diagnosis • 100 patients • Failure care analysis • Clinician diagnosis and offered AI prediction	Satisfaction of GPs with AI as decision support for each case	Compared diagnostic performance with human readers
Faqar-Uz-Zaman et al. (2022)[89]	To evaluate the diagnostic accuracy of an app in the ED	Diagnostic (N/A)	• Algorithm performance: • 450 patients • Compared diagnostic accuracy for the top 4–5 diagnoses between the AI tool and the ED physician (matched between candidate diagnoses)		Compared algorithmic prediction with human annotations
Felmingham et al. (2022)[90]	To evaluate an AI tool's diagnostic accuracy for skin cancer detection	Diagnostic deep learning	• Algorithm performance: AUROC, sensitivity, specificity, FNR • 214 cases, 742 lesions • Trained on the use of a camera and software before the study • Compared diagnostic accuracy with independent diagnoses by teledermatologists • Analysis of AI errors		Compared algorithmic prediction with independent clinician diagnosis
Feng et al. (2025)[91]	To validate a diagnostic model for distinguishing thymomas from other nodules	Diagnostic machine learning	• Algorithm performance: ROC, DCA, sensitivity, specificity • 23 patients • Expert evaluation panel provided ground truth • Performance of 3 radiologists (mixed experience levels) compared with model performance using AUC • No clinical information provided to the radiologists	Described a training process for radiologists	Prospective clinical validation (silent trial)
Hanley et al. (2017)[92]	To evaluate an AI tool for predicting the need for a CT scan in patients with TBI	Triage machine learning	• Algorithm performance: AUROC, sensitivity, specificity, NPV, PPV; clinical utility • 720 patient CTs across 11 ED sites • Assessed model outputs against clinical annotations as determined by laboratory reading and imaging specialist readers according to a prespecified statistical plan • Failure mode analysis of false negatives		Compared algorithmic prediction with human annotations
Hoang et al. (2025)[93]	To evaluate SAFE-WAIT in a silent trial	Predictive machine learning	• Algorithm performance: recall, specificity, accuracy, precision, NPV, FPR, FNR, F1 score • Bias assessment conducted by sex (male, female) and age bracket (young, middle-aged, older adult)	Utility value calculation articulated in terms of clinically relevant decisions and outcomes	Silent trial (algorithmic validation)
Im et al. (2018)[94]	To validate an AI tool for diagnosing aggressive lymphomas before deployment to LMICs	Diagnostic deep learning	• Algorithm performance: specificity, sensitivity, efficiency, size measurements, staining, reproducibility • Described data quality controls • Equipment detailed • 40 patients	Computational time and system components, cost, computational infrastructure	Independent verification of AI labels against clinician assessment
Jauk et al. (2020)[19]	To evaluate a delirium prediction model in its clinical setting	Predictive machine learning	• Algorithm performance: AUROC, sensitivity, specificity, FPR, FNR, PPV, NPV • Rated against nurse assessment of the delirium risk score and the Confusion Assessment Method • Reported failure modes and exclusions • Independent assessment by nurses on 33 patients, 86 with exposure to the AI output	• Expert group of senior physicians, ward nurses, technicians, employees • Offered training for users	Compared outcomes with expert ratings

https://doi.org/10.1038/s44360-025-00048-z

Table 2 (continued) | General information about the included silent studies

Study	Aim and rationale	Model type and intended use	Model evaluation	Additional considerations	Categorization
Kim et al. (2023)[10]	To validate a commercial AI tool for detecting chest radiographic abnormalities	Diagnostic AI	• Algorithm performance: AUROC, sensitivity, specificity • Assessed pathologies on 3,047 radiographs with and without AI output across two centres • CE marking by the Ministry of Food and Drug Safety of Korea • 4 first- and third-year radiology residents as target users • Reading times and failure care analysis		Compared diagnostic accuracy with and without AI assistance
Korfiatis et al. (2023)[95]	To evaluate an AI tool detecting PDA from CT scans	Diagnostic deep learning	• Algorithm performance: AUROC, sensitivity, specificity, F1 • Simulated a screening sample of 297 consecutive abdominal CTs for validation by radiologists • Assessed failure modes using tumour-related parameters	• Reported substantial impact to clinical workflow • Used heat maps during the review process	Radiologist-verified diagnostic accuracy
Kramer et al. (2024)[96]	To validate a model predicting malnutrition in hospitalized patients	Predictive machine learning	• Algorithm performance: AUROC, sensitivity, specificity, accuracy • 159 patients • Dieticians assessed malnutrition in admitted patients, compared (masked) with real-time ML predictions		Compared algorithmic prediction with human annotations
Kwong et al. (2022)[97]	To evaluate a model predicting hydronephrosis in utero	Predictive deep learning	• Algorithm performance: AUROC, AUPRC • Assessed failure modes by age, laterality, changes in image processing and ultrasound machine • Assessed bias for sex and postal code • Looked for potential causes of drift • Recorded model downtime • 1,234 cases with prediction at the desired implementation care point and compared with later decision to proceed with surgery • Reported data stream for model evaluation related to patient data confidentiality and security	• Measured clinician engagement • Assessed usability and disruption to workflow • Used activation maps • Conducted patient and family surveys to assess receptivity	Verification of the model against the outcome label
Liu et al. (2023)[98]	To validate a model predicting postoperative pain	Predictive deep learning	• Algorithm performance: ROC, AUC, RMSE, correlation • Compared algorithmic prediction of maximum pain score with clinician preprocedure prediction in adult inpatients undergoing noncardiac surgery with general anaesthesia • Included patient race in the model but did not report performance subgrouped by race • Reported dataset drift		Compared algorithmic prediction with independent clinician rating
Liu et al. (2024)[99]	To evaluate an AI model estimating bone age	Decision support deep learning	• Algorithm performance: RMSE, MSE • Assessed performance by patient age and sex, as well as radiography vendor • 973 radiographs across 9 hospitals • 3 expert reviewers as gold standard; inter-rater reliability calculated	• Measured time to completion of reading, human versus AI • Per-bone κ values to indicate disagreements	Clinical validation study comparing AI with gold standard
Luo et al. (2019)[100]	To validate a model detecting gastrointestinal cancers	Diagnostic deep learning	• Algorithm performance: AUC, ROC, PPV, NPV, sensitivity, specificity • Reviewed false negatives plus a random subset assessed against an independent assessment by experts • 175 patients, 4,532 images collected from 5 hospitals • Noted the presence and location of tumours	Measured processing time	Algorithmic validation with verification of a random subset
Lupei et al. (2022)[101]	To evaluate the real-time performance of a COVID-19 prognostic model	Predictive machine learning	• Algorithm performance: AUC, ROC, PPV, NPV, sensitivity, specificity • 13,271 symptomatic patients with COVID-19 • Evaluated sensitivity and specificity across sex and race • Assessed label drift as a result of improved outcomes for patients	Opted out of research requests, noted in the chart and honoured by the team	Prospective algorithmic validation

https://doi.org/10.1038/s44360-025-00048-z

Table 2 (continued) | General information about the included silent studies

Study	Aim and rationale	Model type and intended use	Model evaluation	Additional considerations	Categorization
Mahajan et al. (2023)[102]	To assess a model's predictive accuracy for 30-day postoperative mortality and major adverse cardiac and cerebrovascular events	Predictive machine learning	• Algorithm performance: AUC, ROC, PPV, NPV, sensitivity, specificity • 206,353 patient cases • Compared performance with an algorithm already used in care	SHAP values applied to retrospective test only	Prospective algorithmic validation study
Major et al. (2020)[103]	To validate a model predicting short-term in-hospital mortality	Predictive machine learning	• Algorithm performance: descriptive statistics (*n* patients meeting the primary outcome) • 9-month trial with 41,728 predictions+12-week silent test in which hospitalists reviewed 104 alerts to determine whether the alert was actionable and appropriate • Assessed bias by comparing algorithmic fairness approaches	• Clinical stakeholders selected 75% PPV as the desired threshold for the model • Experimented with different thresholds, varied across sites to reflect population needs	Prospective algorithmic validation
Manz et al. (2020)[16]	To validate an algorithm predicting 180-day mortality risk in a general oncology cohort	Predictive machine learning	• Algorithm performance: AUC, AUPRC, Brier score, PPV, NPV, sensitivity, alert rate tested at different risk thresholds • 24,582 patient cases over a 2-month period • Calculated performance metrics across different groups (disease site, practice type, self-reported race, sex, insurance, stage of cancer); reported performance to be better for women or at a later stage of cancer for men • Described the model being locked; no updates made	Use of a nudging strategy described in a companion paper	Prospective algorithmic validation
Miró Catalina et al. (2024)[104]	To validate a diagnostic algorithm in radiology	Diagnostic deep learning	• Algorithm performance: TP, TN, FP, FN, sensitivity, specificity • 278 cases of 471 participants • Researchers interpreted reference radiology reports before inputting to AI to obtain a diagnosis for comparison • Error testing for certain pathologies		Compared diagnostic performance with human readers
Morse et al. (2022)[27]	To evaluate a model detecting CKD in a paediatric hospital	Evaluative machine learning	• Algorithm performance: AUROC • ML model draws data directly from the EHR in near real time • 1,270 patient admissions over ~6 months		Prospective algorithmic validation
Nemeth et al. (2023)[37]	To validate a model for detecting septic shock	Predictive machine learning	• Algorithm performance: AUC, PPV, NPV • 5,384 hospital admissions in 4,804 patients during a 6-month silent test, comparing predictions with a clinician's independent judgement • Extensive failure case analysis • Tested different time horizons • Described data flow and infrastructure for the model	• Codesign using interviews with multiple stakeholders • User acceptance testing • Alignment of model use with practice guidelines	Compared model outputs with clinician annotations
O'Brien et al. (2020)[105]	To evaluate an EWS for patient deterioration	Predictive machine learning	• Algorithm performance: PPV, sensitivity, thresholding • 4,210 encounters, 97 patients • Set up data analytics to reflect real-time streaming of live data	• Alert risk presented using red, yellow and green colour codes • Nursing consult on visualization	Algorithmic validation study
Ouyang et al. (2020)[32]	To validate a segmentation model assessing cardiac function	Predictive deep learning	• Algorithm performance: AUC, RMSE, R^2 • Measurements of cardiac function in 1,288 patients • Compared model measurements with those by human annotators, with manual blinded re-evaluation by 5 experts for cases with a large discrepancy between the model and annotations		Compared model outputs with clinician annotations
Pan et al. (2025)[106]	To validate a model predicting the utility of CT for mTBI	Predictive machine learning	• Algorithm performance: AUC, accuracy, sensitivity, specificity, PPV, NPV, F1, DCA • 86 patients • ML model compared with serum biomarkers for TBI and a statistical regression model	• SHAP values • DCA to assess clinical utility	Prospective clinical validation (silent trial)

https://doi.org/10.1038/s44360-025-00048-z

Table 2 (continued) | General information about the included silent studies

Study	Aim and rationale	Model type and intended use	Model evaluation	Additional considerations	Categorization
Pou-Prom et al. (2022)[34]	To validate an early warning system in inpatients	Predictive machine learning	• Algorithm performance: AUC, PPV, sensitivity • Determined a composite outcome label • Described the shift needed to accommodate changes due to onset of the COVID-19 pandemic • Described a detailed preprocessing plan • Evaluated the processing stream • Initially planned a 4-month trial, which was extended to 6 months • Conducted training with users	Weekly check-ins with stakeholders during the silent phase	Real-time algorithmic validation
Pyrros et al. (2023)[107]	To validate a model detecting type 2 diabetes from chest radiographs and EHR data	Predictive deep learning	• Algorithm performance: AUROC, PPV, sensitivity, specificity, F1, Youden's *J* index, PR, NPV, odds ratio, demographics • 9,943 chest radiographs • Noted the potential for health disparities; planned subgroup analysis by race/ethnicity; mentioned the need for fine-tuning due to fairness and robustness issues • Data stream and infrastructure described	Used an animated technique through an autoencoder for feature highlighting	Algorithmic validation study
Qian et al. (2025)[108]	To validate a model predicting surgical intervention need for paediatric intussusception	Predictive deep learning	• Algorithm performance: AUC, accuracy, NPV, F1, ROC • 50 patients • Reported consistent performance across different patient populations by age		Algorithmic validation
Rajakariar et al. (2020)[25]	To validate a smartwatch device for detecting atrial fibrillation	Diagnostic machine learning	• Algorithm performance: sensitivity, specificity, TP, TN, Cohen's κ for agreement • Failure case analysis for unclassified tracings assessed by 2 electrophysiologists • Described the data pipeline • 200 consecutive patients over 6 months, 439 ECGs • Cardiologist diagnosis as the reference standard		Compared device output with clinician diagnosis
Rawson et al. (2021)[109]	To validate a model detecting secondary bacterial infection during COVID-19	Predictive machine learning	• Algorithm performance: AUROC, descriptive analysis		Prospective pilot test of the algorithm
Razavian et al. (2020)[33]	To validate a model predicting outcomes for hospitalized patients with COVID-19	Predictive machine learning	• Algorithm performance: AUROC, AUPRC, PPV, thresholded sensitivity, confidence intervals • Integration through the EHR; data flow described • Described the cleaning process, feature minimization, threshold selection and time horizon • 445 patients over 474 admissions (109,913 prediction instances) • Medical students and practicing physicians assessed face validity, timing and clinical utility	• Review with medical students to assess 30 patient encounters for impact on clinical decision-making from model prediction • Interface described • Feature-level XAI	Prospective observational study (unclear of impact)
Ren et al. (2025)[110]	To evaluate a smartphone-based AI for classifying auricular deformities	Diagnostic deep learning	• Algorithm performance: AUC, ROC, sensitivity, specificity, precision, F1 score • 272 cases • Ground truth established by two independent professionals • Compared human and model performance • Scalable and low-cost diagnostic support • Guidance for proper image acquisition • Failure analysis identified discrepancies between retrospective and prospective validation sets • Described the data pipeline and inference process		Clinical validation
Schinkel et al. (2022)[111]	To validate a model predicting a positive blood culture result	Predictive machine learning	• Algorithm performance: AUROC, AUPRC, calibration, feature contributions, DCA • Described data processing in a live context • 3-month period of real-time validation		Real-time prospective algorithmic validation

https://doi.org/10.1038/s44360-025-00048-z

Table 2 (continued) | General information about the included silent studies

Study	Aim and rationale	Model type and intended use	Model evaluation	Additional considerations	Categorization
Shah et al. (2021)[112]	To validate a model predicting clinical deterioration	Predictive machine learning	• Algorithm performance: AUPRC, AUROC, PPV, NNE • Preplanned subgroup analysis by race, sex and age revealed discrepancies • 146,446 hospitalizations in 103,930 unique patients • Described data processing steps and feature importance calculations		Algorithmic validation study
Shamout et al. (2021)[113]	To validate a model predicting deterioration from COVID-19	Predictive machine learning	• Algorithm performance: AUC, PR, PPV, NPV • 375 examinations • Real-time extraction; addressed computational resources		Prospective algorithmic validation (silent trial)
Shelov et al. (2018)[38]	To validate a model predicting clinical acuity in a paediatric ICU	Machine learning decision support	• Algorithm performance: Littenberg Technology Assessment in Medicine framework • Approximately 6-month verification phase before going live • Measured the impact of the model in EHR on processing time • Validation done through a survey for project team clinicians to complete (315 forms for 182 patients) • Retrospective analysis of data quality and patients meeting the at-risk criteria • Reported on the availability of the algorithm	• Some interfaces included • Design included a multidisciplinary team comprising physicians, nurses, informaticians, respiratory therapists and improvement advisors	Prospective verification of the model against clinical judgement
Sheppard et al. (2018)[29]	To validate an algorithm for triaging patients with suspected high BP for ambulatory pressure monitoring	Triage machine learning	• Algorithm performance: sensitivity, specificity, PPV, NPV, AUROC • Compared the accuracy of the triaging strategy across subgroups (by setting, age, sex, smoking status, BMI, history of hypertension, diabetes, CKD, cardiovascular disease and BP measuring device) • 887 eligible patients with 3 same-visit BP readings • Described the rationale for excluding cases based on data missingness	Advised patients with hypertension history on the design of the project, recruitment and study literature before ethics submission	Comparison of algorithmic triaging approach against the standard
Shi et al. (2025)[114]	To evaluate a model predicting the risk of colorectal polyp recurrence	Predictive machine learning	• Algorithm performance: ROC, DCA, sensitivity, specificity • 166 patients	• DCA to assess clinical utility • Demonstrated the user interface	Prospective algorithmic validation study
Smith et al. (2024)[115]	To evaluate a model for breast cancer screening	AI decision support	• Algorithm performance: recall or no recall decision • Assessed concordant and discordant cases • 8,779 patients aged 50–70 years • Trained film readers verified the results • Assessed multiple features of patients and scan results	Regions of interest available during reviews	Compared diagnostic performance with human readers
Stamatopoulos et al. (2025)[116]	To validate a model predicting miscarriage risk	Predictive machine learning	• Algorithm performance: sensitivity, specificity, PPV, NPV • Assessor had access to ground truth and compared algorithm predictions against short-term outcomes	Inferred a lack of clinical utility due to unreliable predictions	Prospective algorithmic validation study
Stephen et al. (2023)[20]	To validate a model detecting paediatric sepsis	Predictive machine learning	• Algorithm performance: AUC, PPV • 8,608 cases (1-year period) • Thresholding for alerts to consider false alerts, alert fatigue, resources for sepsis huddle	Team of clinicians, data scientists, improvement experts and clinical informaticians; regular meetings throughout the project	Real-time algorithmic validation
Swinnerton et al. (2025)[117]	To prospectively validate a prediction tool for severe COVID-19 risk	Predictive machine learning	• Algorithm performance: AUC, calibration • 51,587 infections • Assessed subgroup performance	Feature importance	Prospective algorithmic validation study

https://doi.org/10.1038/s44360-025-00048-z

Table 2 (continued) | General information about the included silent studies

Study	Aim and rationale	Model type and intended use	Model evaluation	Additional considerations	Categorization
Tan et al. (2025)[26]	To clinically validate AI-based multispectral imaging for burn wound assessment	Classification deep learning	• Algorithm performance: sensitivity, specificity, accuracy • 40 patients, 70 burn images • Failure mode analysis affecting overdiagnosis • Bias assessment by skin pigmentation and tattoo presence • Reported on availability, feasibility and time to diagnostic result • Described the user interface • UKCA class I medical device, ISO 13485	• Reported evaluator training • Described the user interface	Prospective clinical validation study
Tariq et al. (2023)[118]	To validate a model screening for low bone density	Screening machine learning	• Algorithm performance: image label, precision, recall, *F* score, AUROC • For 2 consecutive days, curated 344 scans (with and without contrast) from patients aged ≥50 years • Some analysis of lower-performing areas	Heat maps for regions of interest	Algorithmic validation study
Titano et al. (2018)[119]	To simulate the clinical implementation of a triage algorithm for radiology	Triage deep learning	• Algorithm performance: AUC, sensitivity, specificity, accuracy, time to notify about critical findings, runtime • 180 images reviewed by a radiologist and a surgeon (50/50 split); 2 radiologists and a neurosurgeon reviewed images without access to the EMR or prior images		Prospective simulated trial with human readers
Vaid et al. (2020)[120]	To validate an outcome prediction model for COVID-19	Predictive machine learning	• Algorithm performance: AUROC, AUPRC, F1, sensitivity, specificity • 21-day trial • Assessed race as a potential contributing variable to outcome prediction	SHAP scores	Prospective algorithmic validation (silent trial)
Wall et al. (2022)[121]	To evaluate a model for supporting radiation therapy plans	Predictive machine learning	• Algorithm performance: prediction error, ROC, concordance • VQA application provides failures for features, top 5 features and 'total gain' • Reported runtime and compute power • Physicists measured 445 VMAT plans over 3 months • VQA predictions recorded alongside PSQA measurements		Prospective validation including comparison with the standard of care
Wan et al. (2025)[122]	To validate a model predicting neoadjuvant treatment response	Predictive machine learning	• Algorithm performance: AIC, ROC, PPV, NPV, DCA, calibration • 76 patients • Compared the performance of a clinical–radiomics model to that of a radiomics model, a clinical model and a radiologist's subjective assessment	DCA to assess potential clinical benefit	Clinical validation
Wang et al. (2019)[123]	To validate a model predicting new-onset lung cancer	Predictive machine learning	• Algorithm performance: AUC, ROC, PPV, sensitivity, specificity • Performance within each risk category • 836,659 patient records		Algorithmic validation study
Wang et al. (2025)[124]	To validate a model for cardiovascular disease diagnosis	Diagnostic deep learning	• Algorithmic validation: AUC, sensitivity, specificity, F1, accuracy • 62 patients • Ground truth established by 3 emergency physicians reviewing the data, compared with algorithm outputs	SHAP values	Algorithmic validation with clinical verification
Wissel et al. (2020)[125]	To validate an NLP application to assign surgical candidacy for epilepsy	Decision support machine learning	• Algorithm performance: AUC, sensitivity, specificity, PPV, NPV, NNS, number of prospective surgical candidates • Retrained the model weekly on the most recent training set based on free text notes • Verification on 100 randomly selected patient cases • Tested the inter-rater reliability of clinicians' manual classifications versus the algorithm	Interpretability analysis revealed wording associated with surgical candidacy	Algorithmic validation with verification of a random subset
Wong et al. (2021)[30]	To temporally validate a model predicting acute respiratory failure	Predictive machine learning	• Algorithm performance: AUROC, AUPRC, sensitivity, specificity, PPV, NPV • Event horizon • 122,842 encounters, 112,740 controls		Temporal validation study

https://doi.org/10.1038/s44360-025-00048-z

Table 2 (continued) | General information about the included silent studies

Study	Aim and rationale	Model type and intended use	Model evaluation	Additional considerations	Categorization
Xie et al. (2025)[126]	To validate a model diagnosing axial spondyloarthritis	Diagnostic deep learning	• Algorithmic validation: AUC, accuracy, sensitivity, specificity, F1, precision • 209 patients • Diagnostic accuracy compared with accepted clinical classification criteria for each patient	SHAP values	Algorithmic validation
Ye et al. (2019)[127]	To validate a real-time early warning system predicting high risk of inpatient mortality	Predictive machine learning	• Algorithm performance: sensitivity, specificity, PPV, ROC, *C*-statistic, hazard ratios • 11,762 patients with an assigned EWS	Top 50 important features	Algorithmic validation study
Ye et al. (2020)[128]	To validate a nomogram for predicting liver failure	Predictive machine learning	• Algorithm performance: precision, recall, accuracy, F1 • 120 patients undergoing hepatectomy		Algorithmic validation study
Yu et al. (2022)[129]	To validate a sepsis prediction model	Predictive machine learning	• Algorithm performance: F1, sensitivity, specificity, AUROC, AUPRC • 3,532 alerts; 388 met the sepsis criteria • Analysed model successes and failures • Considered scalability through compute requirements	SHAP values for a 'lite' version of the model	Algorithmic validation study
Zhang et al. (2025)[130]	To validate a model identifying atrial fibrillation after ischaemic stroke	Diagnostic deep learning	• Algorithm performance: AUC, sensitivity, specificity, PPC, NPV • 73 patients • Assessed model performance by patient age bracket • An independent researcher conducted a blinded review of predicted atrial fibrillation status and actual diagnosis after clinical workup • Described data cleaning and patient inclusion criteria		Algorithmic validation

AIC, Akaike information criterion; AUC, area under the curve; BMI, body mass index; BP, blood pressure; COVID-19, coronavirus disease 2019; CKD, chronic kidney disease; CPT, Current Procedural Terminology; CT, computed tomography; DCA, decision curve analysis; ECG, electrocardiogram; ED, emergency department; EMR, electronic medical record; EWS, early warning score; FN, false negative; FNR, false negative rate; FP, false positive; GP, general physician; ICU, intensive care unit; ISO, International Organization for Standardization; LMICs, low- to middle-income countries; ML, machine learning; MSE, mean square error; mTBI, mild traumatic brain injury; N/A, not applicable, NLP, natural language processing; NNE, number needed to evaluate; NNS, number needed to screen; NPR, negative prediction rate; NPV, negative predictive value; PDA, pancreatic ductal adenocarcinoma; PPV, positive predictive value; PR, precision–recall; PSQA, patient-specific quality assurance; RMSE, root mean square error; ROC, receiver operating characteristic; SOFA, sequential organ failure assessment; TBI, traumatic brain injury; TN, true negative; TP, true positive; UKCA, UK Conformity Assessed; VMAT, volumetric modulated arc therapy; VQA, virtual quality assurance; XAI, explainable AI.

Composition of silent evaluations

The geographical locations and institutions of the included silent evaluations were extracted. From the 75 final papers (excluding sister studies, as they share the same characteristics), we found silent evaluations performed in Australia, Austria, Canada, China, France, India, Germany, Mexico, the Netherlands, Saudi Arabia, Spain, South Korea, Taiwan, Turkey, the UK and the USA, with demographic information obtainable for 74 of the 75 papers (as shown in Fig. 2, generated using R software[22] and RStudio[23]). Most silent evaluations were conducted in the USA (48%), China (19%) and the UK (7%). A list of institutions (hospitals and research centres) where silent evaluations were performed is provided in Table 3. Nine studies reported the evaluation of a commercially available AI system. Four of the nine studies reported the approval regime[10,24–26] (for example, CE-marked, cleared device, or approved device and class rating), while the remaining papers did not provide details about the system.

Study design and purpose

Our eligibility criteria led us to papers that self-identified as silent trials, as well as to model validations under other names and forms that paralleled the silent trial methods. Importantly, only 15 studies explicitly used the term silent to describe their evaluation, highlighting that similar methodologies exhibit substantial variation in their nomenclature and conceptualization.

Definitions varied along a spectrum, ranging from technical validation of the algorithm in a live clinical environment to broad, multistage silent evaluations of the clinical setting. We note that algorithmic validation, clinical validation, temporal validation and prospective validation were often used interchangeably to describe similar methodologies but with varying scopes of evaluation (Table 2). Variation in the clinical verification of the model (human or automated annotation of ground truth for model comparison) was less predictive of the breadth and depth of clinical evaluation than the purpose of the trial itself. For instance, some papers aimed to prospectively validate the technical performance of a model (for example, "…to evaluate the ability of three metrics to monitor for a reduction in performance of a CKD model deployed at a paediatric hospital." (ref. 27)), while others purported to evaluate the potential clinical utility of the algorithm across a wider array of elements (for example, "…to assess the AI system's predictive performance in a retrospective setting and evaluate the human factors surrounding the BPA before initiating the quasi-experimental clinical study." (ref. 28)).

While we only included papers for which we could be relatively confident that there was a separation between model evaluation and clinical care, this core component of the silent phase was often not clearly articulated. When not articulated as such, we inferred separation from contextual information within the paper (for example, "Clinicians assessed patients as per usual practice."), grammatical tense (for example, "This algorithm would have identified X patients in practice.") and minor methodological cues (for example, "The research team did not intervene in the clinical management of these patients.").

The length of the evaluation phase was consistently reported, either as a specified date range or as a quantitative number of patients

https://doi.org/10.1038/s44360-025-00048-z

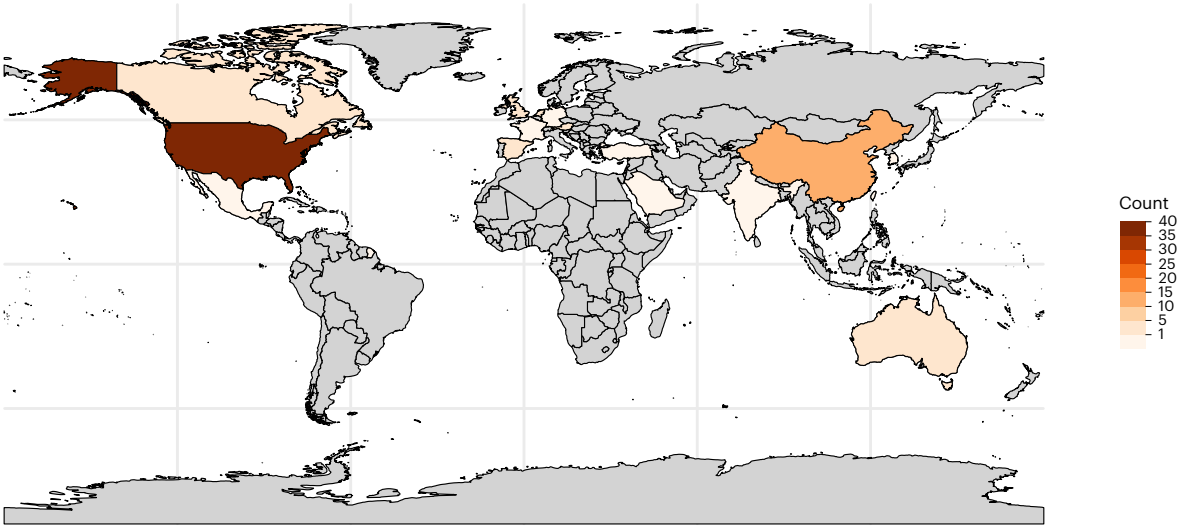

Fig. 2 | World map showing the number of silent trials identified by country. The countries of silent trials were counted once for each paper, if available (74 of 75 papers). The USA was the most represented country (36 trials), followed by China (14 trials), the UK (5 trials) and Canada (3 trials). In total, 16 countries were represented in the silent trials. Figure created using R software and RStudio (2025).

or cases; however, a justification or rationale for these choices was rarely provided. The total time period for silent evaluations ranged from 2 days to 18 months.

Model evaluation during the silent phase

Most studies described the input data and their form (for example, tabular data and images), and more than half described how the inputs were selected during the development stage. Some studies focused explicitly on technical performance-related reasons for feature selection, while others reported clinical justifications for specific variables, including the feasibility of using these variables relative to the intended use environment (and thus their relevance to evaluation during the silent phase).

Metrics of model performance included AUROC (area under the receiver operating characteristic curve), sensitivity, specificity, negative predictive value and positive predictive value, with all studies describing at least one of these. Some studies, often predominant in medical imaging, examined model performance in greater depth and included an assessment of failure modes—for example, descriptive performance on subgroups within disease categories or an exploration of a specific class of suboptimal performance, such as describing all false-negative cases.

Few studies that reported feedback to recalibrate the model included changing model thresholds to improve sensitivity or specificity, as well as updating the model based on changing demographics or features of the prospective patients. Some papers[16,29,30] reported not updating the model during the evaluation (for example, "Models were not retrained for both validations for fair assessment." (ref. 30)). Rarely did studies describe data shifts or the steps taken to address performance shifts; often, these were simply observed during the evaluation period.

A minority of studies addressed potential algorithmic biases. Typically, this meant exploring model performance among contextualized subgroups of interest (that is, algorithmic bias), which involves assessing an algorithm's performance against identified clinical (for example, specific health conditions) or demographically defined (for example, age, sex, race and ethnicity) subgroups at risk of disparate health outcomes based on the intended use of the AI tool (that is, marginalized, vulnerable or under-represented groups)[31]. Race and sex were the most common subgroups of interest; rarely was a link made to health inequities or other structural issues as a rationale for conducting this testing, and when justified, it included only a general appeal.

In addition to subgroup analyses, a subset of studies examined algorithmic bias that appeared at test time when development and evaluation settings did not match. Some reported drops in performance linked to noisy or incomplete data and inconsistencies in electronic health record (EHR) coding, while others noted reduced accuracy due to differences in data acquisition, patient populations and clinical practices. Some studies specifically linked these issues to temporal or distributional shifts between training and deployment data. A common conclusion across all studies was that a performance drop is apparent when moving from retrospective to live evaluation, showing that models often perform less reliably during silent or prospective evaluation.

A key process during the silent phase is verifying the correctness of the model's predictions in a live environment, which we have termed 'verification of model outputs'. Such verification could refer to any of the following: agreement between a model's prediction and information noted or coded in the medical record; an expert evaluator's (for example, a physician's or nurse's) assessment of the model prediction; or a case-by-case evaluation by experts independently compared with the model's outputs to determine agreement, conducted blind to the model output for comparison purposes. We categorize verification in our papers as human annotation versus automatic annotation, in which trials used either automated annotation of ground truth (obtaining algorithm performance (AUROC) by comparing with a test set of clinical information that was not transparently defined) or live human annotation (comparing the algorithm with clinical ground truth obtained through expert or novice consensus panels during the trial). When human annotation was used, only a small minority of these studies described the characteristics of evaluators, such as qualifications, role or whether they received any formal instructions for review. However, the evaluator of the algorithm—who was responsible for comparing the model with annotations and for viewing the system during the trial—was often invisible and was rarely reported. When alluded to, evaluators were used either to provide an independent assessment of the same outcome the model was predicting (for example, "Variance between performance of senior sonographers and AI measurements was compared." (ref. 32)) or to evaluate aspects of the tool itself, such as establishing clinical utility (for example, "assessed the face validity, timing, and clinical utility of predictions" (ref. 33)). In some cases, it was not clearly described whether the evaluator's role was to conduct an independent (blind) assessment of the same outcome the model was meant to predict or whether they were viewing the model output and meant to verify its accuracy.

https://doi.org/10.1038/s44360-025-00048-z

Table 3 | Demographic information of the included final 75 papers

Trial	Country	Institutions
Organ failure: Aakre et al. (2017)[21]	USA	Mayo Clinic hospitals in Rochester, MN, and Jacksonville, FL
NLP for opioid use: Afshar et al. (2023)[28]	USA	University of Wisconsin Hospital
COVID-19: Alrajhi et al. (2022)[75]	Kingdom of Saudi Arabia	King Faisal Specialist Hospital and Research Centre
Appendicitis: Aydın et al. (2025)[76]	Turkey	13 tertiary paediatric hospitals across Turkey
Sperm: Bachelot et al. (2023)[77]	France	Assistance Publique-Hopitaux de Paris, Sorbonne University, Paris
Sepsis: Bedoya et al. (2020)[39]	USA	A hospital in the Duke University Health System
Breast: Berg et al. (2023)[78]	Mexico	Hospital Valentin Gomez Farias and Hospital General de Tijuana
Mortality: Brajer et al. (2020)[36]	USA	Duke University Health System
Brain cancer: Butler et al. (2019)[79]	UK	Western General Hospital, Edinburgh
Lung cancer: Campanella et al. (2025)[80]	USA	N/A
Neoplasia: Chen et al. (2025)[81]	China	N/A
Hypertension: Cheng et al. (2025)[82]	China	4 Taklamakan Desert-adjacent regions in northwest China
Haemodynamic instability: Chiang et al. (2025)[83]	Taiwan	Taipei Veterans General Hospital
Breast cancer: Chufal et al. (2025)[84]	India	Rajiv Gandhi Cancer Institute & Research Centre
Suicide risk: Coley et al. (2021)[85]	USA	HealthPartners, Henry Ford Health System, Kaiser Permanente
DEPLOYR (also a framework): Corbin et al. (2023)[86]	USA	Stanford Health Care, Stanford, CA
Lung: Dave et al. (2023)[87]	Canada	London Health Sciences Center, London, Ontario
Breast cancer: El Moheb et al. (2025)[88]	USA	University of Virginia Medical Center
Skin lesion: Escalé-Besa et al. (2023)[24], using study protocol by Escalé-Besa et al. (2022)[131]	Spain	Primary care centres managed by Institut Catala de la Salut, Catalonia
Emergency department: Faqar-Uz-Zaman et al. (2022)[89], using study protocol by Faqar-Uz-Zaman et al. (2021)[132]	Germany	University Hospital Frankfurt
Skin cancer: Felmingham et al. (2022)[90] and results paper by Felmingham et al. (2023)[133]	Australia	Alfred Hospital and Skin Health Institute, Melbourne
Chest: Feng et al. (2025)[91]	China	Tangshan People's Hospital
Head injury: Hanley et al. (2017)[92]	USA	Allegheny General Hospital, Pittsburgh, PA; Baylor University Medical Center, Dallas, TX; Detroit Receiving Hospital, Detroit, MI; Emory University School of Medicine and Grady Memorial Hospital, Atlanta, GA; Hartford Hospital, Hartford, CT; R Adams Cowley Shock Trauma Center, Baltimore, MD; University of Rochester Medical Center, Rochester, NY; University of Texas Memorial Hermann Hospital, Houston, TX; University of Virginia Health System, Charlottesville, VA; Washington University Barnes Jewish Medical Center, St. Louis, MO; Wayne State University Sinai-Grace Hospital, Detroit, MI
Sepsis: Hoang et al. (2025)[93]	Australia	N/A
Lymphoma: Im et al. (2018)[94]	USA	Massachusetts General Hospital, Boston, MA
Delirium: Jauk et al. (2020)[19]	Austria	LKH Graz II
Chest: Kim et al. (2023)[10]	South Korea	N/A
Pancreas: Korfiatis et al. (2023)[95]	USA	N/A
Malnutrition: Kramer et al. (2024)[96]	Austria	University Hospital Graz
Hydronephrosis: Kwong et al. (2022)[97]	Canada	Hospital for Sick Children, Toronto, Ontario
Pain prediction: Liu et al. (2023)[98]	USA	Massachusetts General Hospital, Boston, MA
Bone age: Liu et al. (2024)[99]	China	Children's Hospital of Zhejiang University School of Medicine, Children's Hospital of Fudan University, The First Affiliated Hospital of Sun Yat-Sen University, Xi'an Children's Hospital Affiliated to Xi'an Jiaotong University, Tianjin Medical University General Hospital, Children's Hospital of Chongqing Medical University, Shenzhen Children's Hospital, The Second Affiliated Hospital of Nanchang University, Tongji Hospital of Tongji Medical College of Huazhong University of Science and Technology
Gastrointestinal: Luo et al. (2019)[100]	China	Sun Yat-sen University Cancer Center, North Guangdong People's Hospital, Shaoguan; Wuzhou's Red Cross Hospital, Wuzhou; Jiangxi Cancer Hospital, Nanchang; Puning People's Hospital, Puning; Jieyang People's Hospital, Jieyang
COVID-19: Lupei et al. (2022)[101]	USA	N/A

https://doi.org/10.1038/s44360-025-00048-z

Table 3 (continued) | Demographic information of the included final 75 papers

Trial	Country	Institutions
Postoperative: Mahajan et al. (2023)[102]	USA	20 hospitals in the University Pittsburgh Medical Center health network
End of life: Major et al. (2020)[103]	USA	NYU Langone Health
Morality: Manz et al. (2020)[16]	USA	University of Pennsylvania Health System (multiple practices)
Chest: Miró Catalina et al. (2024)[104], using protocol by Miró Catalina et al. (2022)[134]	Spain	Osana Primary Care Centre Catalonia
Kidney disease: Morse et al. (2022)[27]	USA	Lucile Packard Children's Hospital, Stanford University
Shock: Nemeth et al. (2023)[37]	USA	Mayo Clinic
Deterioration: O'Brien et al. (2020)[105]	USA	Duke University Hospital System
Cardiac: Ouyang et al. (2020)[32]	USA	Cedars-Sinai Medical Center
Traumatic brain injury: Pan et al. (2025)[106]	China	Wuhan Yangtze River Shipping General Hospital
Early warning system: Pou-Prom et al. (2022)[34]	Canada	St Michael's Hospital, Toronto, Ontario
Type 2 diabetes: Pyrros et al. (2023)[107]	USA	Emory Hospital and 28 geographically unique locations
Enema: Qian et al. (2025)[108]	China	Children's Hospital of Soochow University and Affiliated Changzhou Children's Hospital of Nantong University
Atrial fibrillation: Rajakariar et al. (2020)[25]	Australia	N/A
Bacterial infection: Rawson et al. (2021)[109]	UK	Three hospitals in northwest London
COVID-19: Razavian et al. (2020)[33]	USA	N/A
Paediatric: Ren et al. (2025)[110]	China	Obstetrics & Gynecology Hospital of Fudan University
Blood cultures: Schinkel et al. (2022)[111]	The Netherlands	Amsterdam UMC Location VU Medical Center
Deterioration: Shah et al. (2021)[112]	USA	Four PennMedicine hospitals
COVID-19: Shamout et al. (2021)[113]	USA	NYU Langone Health Institute
Paediatric ICU: Shelov et al. (2018)[38]	USA	Children's Hospital of Philadelphia
Blood pressure: Sheppard et al. (2018)[29], using study protocol by Sheppard et al. (2016)[135]	UK	Ten general practice surgeries and one hospital trust
Polyps: Shi et al. (2025)[114]	China	Hospital of Xuzhou Medical University
Breast: Smith et al. (2024)[115]	UK	Hospital—N/A
Pregnancy: Stamatopoulos et al. (2025)[116]	N/A	N/A
Sepsis: Stephen et al. (2023)[20]	USA	Hospital—N/A
COVID-19: Swinnerton et al. (2025)[117]	USA	EHR data from Veterans Affairs facilities nationwide
Burn: Tan et al. (2025)[26]	UK	Newcastle upon Tyne and Manchester Adult Burn Centres
Bone density: Tariq et al. (2023)[118]	USA	N/A
Neurological events: Titano et al. (2018)[119]	USA	Hospital—N/A
COVID-19: Vaid et al. (2020)[120]	USA	Five hospitals within the Mount Sinai Health System in New York City: Mount Sinai Hospital (MSH) located in East Harlem, Manhattan; Mount Sinai Morningside (MSM) located in Morningside Heights, Manhattan; Mount Sinai West (MSW) located in Midtown West, Manhattan; Mount Sinai Brooklyn (MSB) located in Midwood, Brooklyn; and Mount Sinai Queens (MSQ) located in Astoria, Queens
Radiation therapy: Wall et al. (2022)[121]	USA	Unknown local institution
Breast cancer: Wan et al. (2025)[122]	China	Renji Hospital
Lung cancer: Wang et al. (2019)[123]	USA	Hospitals in Maine
Cardiology: Wang et al. (2025)[124]	China	Zigong Fourth People's Hospital
Epilepsy: Wissel et al. (2020)[125]	USA	Cincinnati Children's Hospital Medical Center
Respiratory failure: Wong et al. (2021)[30]	USA	Emory Healthcare Network
Arthritis: Xie et al. (2025)[126]	China	N/A
Mortality: Ye et al. (2019)[127]	USA	Berkshire Health System
Liver surgery: Ye et al. (2020)[128]	China	Eastern Hepatobiliary Surgery Hospital (EHBH)
Sepsis: Yu et al. (2022)[129]	USA	Single, tertiary-care academic medical centre in St. Louis, MO
Cardiovascular: Zhang et al. (2025)[130]	China	Shanghai Sixth People's Hospital

Many studies discussed data quality issues and their management during the silent phase. While some studies described the process for removing patients with incomplete data points, conflicting data or nonstandardized data inputs, there was limited discussion on how this would be managed in a live, real-world deployment context. Some reported on elements around the data pipeline (that is, the flow of data from input to inference), including data quality issues (for example, missingness) and 'downtime' (that is, when the data flow stopped or was negatively affected, causing the model to become nonfunctional). Few studies detailed the granular elements of data flow from the point of contact through processing and analysis to generate predictions, but any such descriptions were generally comprehensive. One study describing the full processing stream for data flow noted the rationale of needing to most closely approximate the conditions of clinical integration, noting that the 'deployment server' was on the same secure private network as the clinical systems, with data pipelines monitored and continually audited by a dedicated data science team[34].

Some studies described model scalability, either as a formal assessment of the computational feasibility of the model in the clinical pipeline or as a stated assertion that the model was scalable. However, it was not always clear what scalability meant in these papers.

Sociotechnical considerations

Sociotechnical considerations concern the ways in which humans design and interact with AI tools. A minority of papers described some element of user engagement either before or during the silent phase.

Most sociotechnical evaluations analysed subjective user experience related to the prediction/interface or the overall impact of the model on workflow, either in the silent environment or presumably before the model was deployed to end users. These evaluations were often conducted in collaboration with clinicians and healthcare staff, indicating that stakeholder expertise and preferences are important. However, when these end users contributed to the usability and preferences of the model[20,28,35–39], it was often not explicitly stated that these consumers were not exposed to model predictions on live patients during the prospective testing phase to evaluate model usability.

We describe the role of human factors in the silent phase as ambiguous, much like earlier difficulties in describing model evaluators and separating the model from care. As such, the evaluation of human factors operates similarly to stakeholder engagement with end users, where feedback is used to refine the later deployment of the system, rather than to comprehensively examine the relationship between the model and the evaluator. Nevertheless, one of the papers considered cognitive factors, such as alert fatigue, in its human factors evaluation; for example, "allowed for consideration of false alerts, alert fatigue, and resources required for a sepsis huddle when designing our model. The Aware tier with high sensitivity was designed to enable situational awareness and prompt discussions about sepsis risk at the individual patient, clinical team, and unit level." (ref. 20). Further, some studies described the integration of explainability methods (for example, SHAP (SHapley Additive exPlanations), heat maps) with model outputs during the silent phase, with the aim of preparing for improved adoption following integration. However, no study assessed the potential impact of visualizations on human decision-making, such as whether the use of explainability mechanisms could prevent persuasion by incorrect AI results.

Users and stakeholders were engaged in the process of testing or designing the model most commonly through interview groups that provided feedback on the context and facilitation of the tool, often as multidisciplinary teams (for example, "This expert group was set up in order to enhance participation of health professionals, including senior physicians, ward nurses, technicians, and leading employees." (ref. 19)). The reasons behind these evaluations, if described at all, were usually to assess model accuracy, the feasibility of model integration and user acceptance. Assessments of usability and AI evaluation were conducted almost entirely before deployment. One study described an evaluator developing potential automation bias following a silent phase evaluation (referred to as the phenomenon of 'induced belief revision' (ref. 17)), which the authors note is important to address to ensure scientifically rigorous evaluation and separation of the model's testing from care[17]. In the process of assessing the model's performance against real-world information, consideration of the potential for incidental findings in the data that could have implications for patient safety was described in four papers[17,24,34,39]. None of these studies described any form of patient or consumer engagement.

Discussion

The vastness and diversity of literature reporting on silent evaluations of AI indicate that there is undoubtedly a perceived value in this paradigm for ensuring model performance in the prospective setting, linked to motivations around 'responsible AI'. The heterogeneity of the currently reported practices highlights the immense opportunity to coalesce around best practices; we hope that this work is one step in this regard. In this vein, we focus specifically on the silent phase, which is bounded by good model development on one side[40] and first-in-human studies (DECIDE-AI[41]), clinical trials (SPIRIT-AI[42], CONSORT-AI[43]) and other clinical evaluation studies on the other. Considering the silent phase not only as a means to assess the prospective performance of a model but also as a mechanism to facilitate responsible and effective downstream translation, our scoping study highlights several opportunities for enhancing practice around this critical translational stage[41].

A consistent challenge in determining whether a paper described a proper silent trial centred on the variability in the use of the term silent. Some papers used the term silent trial but then described the outputs as being visible to the care team (and thus were excluded). We adopted the multiple-reviewer method for adjudication partly because it was difficult to discern whether the model outputs were truly silent. It was common for silent evaluations to be reported in tandem with retrospective testing and/or live deployment. Due to this combination, it was similarly challenging to discern which reported aspects of the study design pertained to which of these stages. For instance, data cleaning might be described, but it was unclear whether this occurred during retrospective or prospective testing. Additionally, the number of case observations or the time period was reported as an aggregate, leaving the proportion during the silent phase unclear. In some cases, reporting on the model's performance was aggregated across the silent and live phases in a manner similar to randomized controlled trials.

We propose that, as a first step, the field should consolidate the notion of silent as a state in which the model's outputs are not visible to the treating team or clinician while the model's performance is being evaluated. This does not necessarily mean that the model itself is invisible; for example, testing user interfaces may involve exposing some staff to the system. We suggest that maintaining a silent trial requires that these staff members are not caring for the same patients for whom the model inference is being run, to prevent contamination of the trial and thus ensure an objective evaluation[17].

We further suggest that papers reporting on evaluations during this phase should clearly distinguish between model evaluation and the care environment. Understandably, resourcing can be a challenge to complete separation; in line with medical literature more broadly[44], transparency should be encouraged, with authors able to comment on the rationale for the choices they made.

An intriguing finding—and one where we feel efforts ought to be consolidated—is the gap between what is most commonly reported and what those with extensive experience deploying AI systems know to be important. Specifically, there is an overwhelmingly strong focus on model metrics (for example, AUROC and AUPRC (area under the precision–recall curve)), with far more limited discussion of workflow and systems integration, human factors, and verification of clinically relevant ground truth labels. By contrast, the NICE (National Institute

for Health and Care Excellence) standards for digital health technologies (including AI) emphasize the use of human factors and a broader set of considerations to evaluate such tools, which is far more in keeping with a healthcare environment[45].

One possible explanation is that silent suggests invisibility, and human factor evaluations require end users to engage with some aspects of the model. However, we find that most reported usability evaluations involve healthcare professionals, who we assume are the intended end users of the model. Guidelines endorsed by regulatory agencies, such as Good Machine Learning Practices[40], recommend the involvement of clinical staff in model development and evaluation, and the literature we describe here indicates some recognition of this guidance. Given that researchers are identifying emergent risks from additions like explainability[46,47], it seems important to ensure that these impacts are measured before exposing patients (and research participants) to the model's influence over their care. There is an immense opportunity to explore how human factors might be involved during the silent stage, which could reduce risk once the model reaches the integration stage in addition to improving the precision of the clinical evaluation protocol[41,48,49].

Safety-oriented metrics for model testing can include failure modes, model bias and data shift[50]—well-known limitations of AI models once they proceed to real-time deployment, during which model performance typically drops (to varying degrees)[51]. Reasons can include data quality (for example, feature set discrepancy, temporal feature leakage, operational feature constraints[52]), limitations of model generalizability, mismatch between the data available for development and the deployment environment, concept drift, and unintended changes such as data drift[6,14,53]. Importantly, failure mode testing supports the identification of systematic patterns of lower performance. In radiology, where AI tools have seen the most uptake and have undergone rigorous research on their limitations[54], failure mode reporting was much more common than for nonimaging models in our results.

Algorithmic bias is a known ethical threat in health AI, so it was somewhat surprising to see limited reporting of subgroup-specific performance testing in silent phase evaluations. It is possible that developers conducted bias testing during the development phase, with the presumption that fairness had already been addressed at that point. However, the under-reporting of subgroup-specific performance has been noted in machine learning studies[55] and randomized controlled trials of AI[56]. Assumptions behind choices regarding algorithmic fairness approaches must be verified in their real-world environments to prevent algorithmic discrimination[57–59]. This is particularly important given that some AI models may embed patterns that track patient race even when this is not explicitly coded in the algorithm[60]. Clinical use of AI tools must be informed by details of the model's performance across particular subgroups so that clinicians can properly calibrate how they weight the model's output in their clinical decision-making to avoid risk[61,62]. The silent phase is an ideal stage to test the real-time failure modes of the model and to identify mitigation strategies to prevent worsening inequities and missing clinically relevant gaps in subgroup-specific performance.

While our charting framework extends beyond the original conceptualization of silent trials[6], we note that, across the 75 studies reviewed, each element of charting was reported by some studies. We consider this to support the notion of a silent phase as offering an opportunity for more than just in situ technical validation. We suggest that, if this phase is considered a key component of AI translation, there would be considerable advantage in incorporating a more holistic set of practices. Without aligning silent phase evaluations with real-world needs, we risk implementing clinical applications incorrectly, potentially causing the optimism and momentum around AI to collapse and leading to preventable harm. The concept of translational trials, as advocated by our team[14], frames silent evaluation as a fundamental step in responsible AI translation, with methodological practices guided primarily by the intention of replicating as closely as possible the clinical conditions in which the tool will be used. This paradigm then provides maximally relevant and nuanced information about the model's performance to support more effective and precise translation.

We acknowledge that our scoping review has the limitation of being restricted to practices reported in the literature through published studies and is subject to the typical limitations of such work, including restriction to English-language papers and a subset of publication venues. It is possible that some elements we observed to be under-reported were actually undertaken by teams to facilitate translation but were not reported in the paper. We accept this limitation, although we also note that some teams did report these aspects. Therefore, we view the choice to report or not as reflective of the inherent values of the broader field. To address this limitation, our research team has planned a series of key informant interviews to investigate whether other practices were undertaken but simply not described in the paper.

Another limitation concerns the review process and the terminology. We initially focused on the term silent trial and its known variants, but it is possible that we are unaware of other terms describing analogous evaluative processes. Thus, by missing such works, this review might have failed to cover some other aspects of silent evaluations. Similarly, some silent evaluations may have been conducted by industry groups but not published in the literature, being available only through internal technical reports.

If the ultimate goal of the silent evaluation phase is to bridge the gap in the translation from bench to bedside, we need to ensure that the practices undertaken during this phase most closely approximate the needs of the translational environment. By intentionally designing silent trials to gather evidence that incorporates a sociotechnical and systems engineering[63,64] lens, there is good reason to believe that we can improve the efficacy of translation for these complex interventions[65]. What does this mean for the silent evaluation phase? We believe that by broadening the scope of practices undertaken during this translation stage, we can improve the AI implementation ecosystem in healthcare. These practices should reflect, as closely as possible, the intended implementation setting. A translational evaluation paradigm embodies this framing by explicitly positioning translation as the end goal and necessitating the collection of evidence that adequately informs this state[14]. As more attention is placed on silent evaluations, we hope to provide constructive guidance based on this work to improve the preparation, conduct and reporting of silent phase evaluations and to move towards a focus on a translational evaluation paradigm.

Methods

This scoping review follows the framework for scoping review studies outlined by Arksey and O'Malley[15]. This study complies with the methodology from the *JBI Manual for Evidence Synthesis* guidelines[66] and adheres to the PRISMA-ScR checklist (PRISMA extension for scoping reviews)[67]. This review study was preregistered with the Open Science Framework (https://osf.io/63bhx/) rather than PROSPERO, as it did not assess direct health-related outcomes. Institutional ethics approval was not required.

Information sources and search strategy

Our initial scope was to search the literature for studies reporting on a silent evaluation (including processes reported under analogous terms) of an AI tool in healthcare settings. The full search strategy was developed with a University of Adelaide librarian in collaboration with M.D.McC. and L.T. (Supplementary Table 1). The first search was conducted on 23 October 2024 and updated on 25 September 2025. Controlled vocabulary terms for nondatabase searches were derived from the database search terms.

Searches were conducted using the PubMed, Web of Science and Scopus databases. We also used reference snowballing (using reference

lists from the included papers) and hand searched the literature from these lists, including papers that fit our inclusion criteria. We chose not to include regulatory guidelines as a primary source in this review, as our focus is less on the AI product itself and more on the design and ecological validity of its local testing.

During the process, we recognized that some teams published different components of a silent phase evaluation across multiple papers (for example, one paper might describe the model evaluation while another describes the evaluation of human factors or workflows). Therefore, a complementary search strategy was added during the extraction stage, in which the reviewer (L.T.) performed an adjacent hand search for each included paper to find additional studies exploring sociotechnical evaluations of the silently tested AI system in the final set of included papers. The papers sought were primarily on human factors, stakeholder engagement, qualitative evaluation, or adjunct studies that contained trial information not discussed in the original paper. We believe that these papers provide information about the broader life cycle of translating AI into practice that may not be immediately reported in current silent phase evaluations; however, we extracted only information pertaining to the silent phase.

Eligibility criteria

We included articles that described the evaluation of an AI or machine learning model during a silent phase evaluation in a healthcare environment (for example, hospitals, clinics, outpatient settings or other environments where healthcare is provided). Due to the ambiguous nature of classifying algorithms as AI, we relied on the consensus of members with technical expertise to categorize algorithms as eligible. We define AI (or machine learning) broadly as any model that builds predictive models from input–output data[68], with training on datasets as a key process. We recognize that there may be a variety of opinions on whether some models constitute machine learning or AI; as a group, we sought to be broad in our inclusion criteria to ensure that cases in which the silent trial paradigm was used were included (encompassing many traditional machine learning approaches). We included a broad variety of machine learning and deep learning models, with more details on how papers self-classified their models available in Table 2. We excluded studies that were not related to healthcare, did not involve AI or machine learning methods, involved models unrelated to a clinical target or clinician use (for example, research-based use of machine learning in health), mentioned the silent phase but were not primary research articles, or described plans to conduct a silent evaluation (for example, protocol papers). Articles not written in English, as well as those published before 1 January 2015, were excluded, as we sought to understand current practices. Two reviewers carried out title and abstract screening, as well as full-text screening (L.T. and A.M). A third reviewer (M.D.McC.) resolved conflicts. A systematic review software (Covidence, Veritas Health Innovation[69]) was used for each stage of screening. The study selection criteria were applied to (1) title and abstract screening, (2) full-text screening with two pilot rounds and (3) full-text extraction for papers that did not meet the criteria during data charting.

While conducting the initial review of articles, we noted that the lack of consistent nomenclature and definitions made it difficult to distinguish a true silent phase from other paradigms, such as external or internal validations (see Table 1 and Box 1 for the nomenclature of testing paradigms). Through an iterative and collaborative process with extractors and the wider CANAIRI group, we identified the following elements as minimum qualifications for a silent phase evaluation: (1) the trial of the AI tool must be conducted in its intended use setting or simulate this setting as closely as possible (live), and (2) the AI tool's outputs must not be acted on by the intended users and should not be seen at the time of treatment (silent). We note that the 'live' nature of the silent phase may be limiting depending on the operational constraints of its intended context; thus, we emphasize replicating the live context as closely as possible as an important consideration. For instance, in radiology, most scans are not analysed in real time by the clinician. As such, algorithms can run on consecutive prospective patient scans, but the results can be analysed retrospectively by evaluators to mimic real-time practice as closely as possible while remaining realistic. Another important distinction of silent trials is the separation of model evaluation and care, meaning that we excluded studies in which changes were made to the patient's experience of care to suit the study's aims. For example, in diagnostic studies, model outputs may not be acted on by the treating team, but the patient may undergo study-specific procedures such as new tests or interventions[70]. As the primary objective of a silent period is to first assess the ecological validity of the model[4,6], changing the way care is delivered would contradict this goal. It should be noted that, among the various interpretations of the word 'silent', we opted for silence defined by the model prediction's lack of impact on care, not the model itself being silent in the sense of being invisible (Table 1). This distinction allowed us to include studies that engage clinical end users to test different workflow integrations, evaluate user interfaces, and conduct other preclinical testing that exposes users to an AI algorithm while maintaining at least an intended separation between model evaluation and clinical care. Very often, we needed to review the full text of the paper in extensive detail to ensure that the above two criteria were met. We used at least two, often three, team members to agree on including each of the final papers.

Our above-described criteria were iteratively refined by L.T. and M.D.McC., with input from our authorship team, until we were satisfied that the studies included in the final analysis met the described conditions. While certain aspects of the evaluation's conduct remain somewhat uncertain (see further details in the Discussion), our final list of included papers represents evaluations of AI tools that were validated live or near live in their intended implementation environment (also see Table 2 for inclusion and exclusion criteria).

Data charting process

Our data charting form was initially developed by L.T. and M.D.McC., with input from X.L., and then reviewed by the CANAIRI Steering Group. The charting process was initially drafted based on the authorship team's own experiences with running silent evaluations at their respective institutions, and we included items that were commonly reported in these protocols[71]. We triangulated these protocols with relevant reporting guidelines (for example, DECIDE-AI, TRIPOD + AI), regulatory guidance (US Food and Drug Administration, Health Canada, Therapeutic Goods Administration (Australia)) and authoritative guidance documents (for example, NICE, World Health Organization). The item categories of information for extraction are listed in Supplementary Table 1, and a glossary of terms is available in Box 1.

A key assumption we made in our charting process is that AI is a sociotechnical system[72]. Under this framing, the evaluation of an AI tool must include not only the algorithm's technical performance but also the entire system in which it operates, combined with the human element that sustains its performance. This assumption is grounded in the lived experience of many members of our CANAIRI collaboration team in developing and deploying machine learning models in healthcare settings—a perspective that is gaining increasing support within the literature[73,74]. We chose to chart information related to the evaluators, their perception of the interface, human adaptation influencing AI evaluation and the engagement of relevant stakeholders throughout the process as entry points for sociotechnical evaluation.

We completed two charting pilot rounds of six full-text papers, the first on grey literature (reports) and the second on original research from scientific journals (hand searched). Once consensus on these extractions was reached by L.T., M.D.McC. and X.L., we progressed to the official extraction. Data charting consisted of a colour-coded scheme in which items that the reviewer was unable to find were

highlighted in red, uncertain items were highlighted in orange, and charting elements found in the text were either copied directly or paraphrased by the reviewer. Data were extracted using a standardized data collection form created in Google Sheets (Alphabet). Two independent reviewers (L.T. and C.S.) charted data for 55 studies and any accompanying metadata (for example, separately published study protocols, supplementary materials) in the same repository. After the initial extraction was completed, the papers were split among seven group members (L.E., L.J.P., A.v.d.V., S.B., N.P., C.S., M. Mamdani, G.K., H.T, N.C.K, M.D.McC.) based on their areas of expertise (system, technical, sociotechnical), and the papers were accordingly categorized into these groups by L.T. Therefore, these members had separate Google Sheets with L.T.'s original charting results and were required to read the papers and compare the initial charting against their own findings, resulting in each paper undergoing a minimum of two reviews. Elements remained in red if both reviewers were unable to find them, while any conflicting responses were discussed with and resolved by M.D.McC. or X.L.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The study database, which describes our full extraction from the included studies, is publicly available at https://docs.google.com/spreadsheets/d/17CFyfViM0IMPQYnBquQ16H-fqGtYvNT9D-wCX5zZO4I/edit?usp=sharing.

References

1. Chen, J. H. & Asch, S. M. Machine learning and prediction in medicine—beyond the peak of inflated expectations. *N. Engl. J. Med.* **376**, 2507–2509 (2017).
2. Seneviratne, M. G., Shah, N. H. & Chu, L. Bridging the implementation gap of machine learning in healthcare. *BMJ Innov.* **6**, 45–47 (2020).
3. Sendak, M. P. et al. A path for translation of machine learning products into healthcare delivery. *EMJ Innov.* https://doi.org/10.33590/emjinnov/19-00172 (2020).
4. McCradden, M. D., Stephenson, E. A. & Anderson, J. A. Clinical research underlies ethical integration of healthcare artificial intelligence. *Nat. Med.* **26**, 1325–1326 (2020).
5. Sendak, M. et al. Editorial. Surfacing best practices for AI software development and integration in healthcare. *Front. Digit. Health* **5**, 1150875 (2023).
6. Wiens, J. et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.* **25**, 1337–1340 (2019).
7. Morse, K. E., Bagley, S. C. & Shah, N. H. Estimate the hidden deployment cost of predictive models to improve patient care. *Nat. Med.* **26**, 18–19 (2020).
8. McCradden, M. D. et al. A research ethics framework for the clinical translation of healthcare machine learning. *Am. J. Bioeth.* **22**, 8–22 (2022).
9. Futoma, J., Simons, M., Panch, T., Doshi-Velez, F. & Celi, L. A. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit. Health* **2**, e489–e492 (2020).
10. Kim, C. et al. Multicentre external validation of a commercial artificial intelligence software to analyse chest radiographs in health screening environments with low disease prevalence. *Eur. Radiol.* **33**, 3501–3509 (2023).
11. Wong, A. et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern. Med.* **181**, 1065–1070 (2021).
12. Harvey, H. B. & Gowda, V. How the FDA regulates AI. *Acad. Radiol.* **27**, 58–61 (2020).
13. Wu, E. et al. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat. Med.* **27**, 582–584 (2021).
14. McCradden, M. D. et al. CANAIRI: the Collaboration for Translational Artificial Intelligence Trials in healthcare. *Nat. Med.* **31**, 9–11 (2025).
15. Arksey, H. & O'Malley, L. Scoping studies: towards a methodological framework. *Int. J. Soc. Res. Methodol.* **8**, 19–32 (2005).
16. Manz, C. R. et al. Validation of a machine learning algorithm to predict 180-day mortality for outpatients with cancer. *JAMA Oncol.* **6**, 1723–1730 (2020).
17. Kwong, J. C. C. et al. When the model trains you: induced belief revision and its implications on artificial intelligence research and patient care—a case study on predicting obstructive hydronephrosis in children. *NEJM AI* **1**, AIcs2300004 (2024).
18. Sendak, M. P. et al. Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. *JMIR Med. Inform.* **8**, e15182 (2020).
19. Jauk, S. et al. Risk prediction of delirium in hospitalized patients using machine learning: an implementation and prospective evaluation study. *J. Am. Med. Inform. Assoc.* **27**, 1383–1392 (2020).
20. Stephen, R. J. et al. Sepsis prediction in hospitalized children: clinical decision support design and deployment. *Hosp. Pediatr.* **13**, 751–759 (2023).
21. Aakre, C. et al. Prospective validation of a near real-time EHR-integrated automated SOFA score calculator. *Int. J. Med. Inform.* **103**, 1–6 (2017).
22. R Core Team. R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, 2025); https://www.r-project.org/
23. Posit Team. *RStudio: Integrated Development Environment for R* (Posit Software, PBC, 2025).
24. Escalé-Besa, A. et al. Exploring the potential of artificial intelligence in improving skin lesion diagnosis in primary care. *Sci. Rep.* **13**, 4293 (2023).
25. Rajakariar, K. et al. Accuracy of a smartwatch based single-lead electrocardiogram device in detection of atrial fibrillation. *Heart* **106**, 665–670 (2020).
26. Tan, P., Nyeko-Lacek, M., Walsh, K., Sheikh, Z. & Lewis, C. J. Artificial intelligence-enhanced multispectral imaging for burn wound assessment: insights from a multi-centre UK evaluation. *Burns* **51**, 107550 (2025).
27. Morse, K. E. et al. Monitoring approaches for a pediatric chronic kidney disease machine learning model. *Appl. Clin. Inform.* **13**, 431–438 (2022).
28. Afshar, M. et al. Deployment of real-time natural language processing and deep learning clinical decision support in the electronic health record: pipeline implementation for an opioid misuse screener in hospitalized adults. *JMIR Med. Inform.* **11**, e44977 (2023).
29. Sheppard, J. P. et al. Prospective external validation of the Predicting Out-of-OFfice Blood Pressure (PROOF-BP) strategy for triaging ambulatory monitoring in the diagnosis and management of hypertension: observational cohort study. *BMJ* **361**, k2478 (2018).
30. Wong, A. I. et al. Prediction of acute respiratory failure requiring advanced respiratory support in advance of interventions and treatment: a multivariable prediction model from electronic medical record data. *Crit. Care Explor.* **3**, e0402 (2021).
31. Ganapathi, S. et al. Tackling bias in AI health datasets through the STANDING Together initiative. *Nat. Med.* **28**, 2232–2233 (2022).
32. Ouyang, D. et al. Video-based AI for beat-to-beat assessment of cardiac function. *Nature* **580**, 252–256 (2020).

33. Razavian, N. et al. A validated, real-time prediction model for favorable outcomes in hospitalized COVID-19 patients. *NPJ Digit. Med.* **3**, 130 (2020).
34. Pou-Prom, C., Murray, J., Kuzulugil, S., Mamdani, M. & Verma, A. A. From compute to care: lessons learned from deploying an early warning system into clinical practice. *Front. Digit. Health* **4**, 932123 (2022).
35. Aakre, C. A., Kitson, J. E., Li, M. & Herasevich, V. Iterative user interface design for automated sequential organ failure assessment score calculator in sepsis detection. *JMIR Hum. Factors* **4**, e14 (2017).
36. Brajer, N. et al. Prospective and external evaluation of a machine learning model to predict in-hospital mortality of adults at time of admission. *JAMA Netw. Open* **3**, e1920733 (2020).
37. Nemeth, C. et al. TCCC decision support with machine learning prediction of hemorrhage risk, shock probability. *Mil. Med.* **188**, 659–665 (2023).
38. Shelov, E. et al. Design and implementation of a pediatric ICU acuity scoring tool as clinical decision support. *Appl. Clin. Inform.* **9**, 576–587 (2018).
39. Bedoya, A. D. et al. Machine learning for early detection of sepsis: an internal and temporal validation study. *JAMIA Open* **3**, 252–260 (2020).
40. Artificial Intelligence/Machine Learning-enabled Working Group. Good Machine Learning Practice for medical device development: guiding principles. *FDA* https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles (2025).
41. DECIDE-AI Steering Group. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat. Med.* **27**, 186–187 (2021).
42. Rivera, S. C. et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Lancet Digit. Health* **2**, e549–e560 (2020).
43. Liu, X. et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Lancet Digit. Health* **2**, e537–e548 (2020).
44. Moher, D. Guidelines for reporting health care research: advancing the clarity and transparency of scientific reporting. *Can. J. Anaesth.* **56**, 96–101 (2009).
45. Evidence standards framework for digital health technologies. Section C: evidence standards tables. *NICE* https://www.nice.org.uk/corporate/ecd7/chapter/section-c-evidence-standards-tables (2018).
46. Gaube, S. et al. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digit. Med.* **4**, 31 (2021).
47. Chromik, M., Eiband, M., Buchner, F., Krüger, A. & Butz, A. I think I get your point, AI! The illusion of explanatory depth in explainable AI. In *26th International Conference on Intelligent User Interfaces* 307–317 (Association for Computing Machinery, 2021); https://doi.org/10.1145/3397481.3450644
48. Felmingham, C. M. et al. The importance of incorporating human factors in the design and implementation of artificial intelligence for skin cancer diagnosis in the real world. *Am. J. Clin. Dermatol.* **22**, 233–242 (2021).
49. Tikhomirov, L. et al. Medical artificial intelligence for clinicians: the lost cognitive perspective. *Lancet Digit. Health* **6**, e589–e594 (2024).
50. Park, Y. et al. Evaluating artificial intelligence in medicine: phases of clinical research. *JAMIA Open* **3**, 326–331 (2020).
51. Lemmon, J. et al. Evaluation of feature selection methods for preserving machine learning performance in the presence of temporal dataset shift in clinical medicine. *Methods Inf. Med.* **62**, 60–70 (2023).
52. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **17**, 195 (2019).
53. Finlayson, S. G. et al. The clinician and dataset shift in artificial intelligence. *N. Engl. J. Med.* **385**, 283–286 (2021).
54. Badgeley, M. A. et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digit. Med.* **2**, 31 (2019).
55. Bozkurt, S. et al. Reporting of demographic data and representativeness in machine learning models using electronic health records. *J. Am. Med. Inform. Assoc.* **27**, 1878–1884 (2020).
56. Plana, D. et al. Randomized clinical trials of machine learning interventions in health care: a systematic review. *JAMA Netw. Open* **5**, e2233946 (2022).
57. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
58. McCradden, M. D., Joshi, S., Mazwi, M. & Anderson, J. A. Ethical limitations of algorithmic fairness solutions in health care machine learning. *Lancet Digit. Health* **2**, e221–e223 (2020).
59. McCradden, M. et al. What's fair is... fair? Presenting JustEFAB, an ethical framework for operationalizing medical ethics and social justice in the integration of clinical machine learning: JustEFAB. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* 1505–1519 (Association for Computing Machinery, 2023); https://dl.acm.org/doi/abs/10.1145/3593013.3594096
60. Gichoya, J. W. et al. AI recognition of patient race in medical imaging: a modelling study. *Lancet Digit. Health* **4**, e406–e414 (2022).
61. Arora, A. et al. The value of standards for health datasets in artificial intelligence-based applications. *Nat. Med.* **29**, 2929–2938 (2023).
62. McCradden, M. D. et al. What makes a 'good' decision with artificial intelligence? A grounded theory study in paediatric care. *BMJ Evid. Based Med.* **30**, 183–193 (2025).
63. Assadi, A. et al. An integration engineering framework for machine learning in healthcare. *Front. Digit. Health* **4**, 932411 (2022).
64. Militello, L. G. et al. Using human factors methods to mitigate bias in artificial intelligence-based clinical decision support. *J. Am. Med. Inform. Assoc.* **32**, 398–403 (2025).
65. Campbell, N. C. et al. Designing and evaluating complex interventions to improve health care. *BMJ* **334**, 455–459 (2007).
66. Aromataris, E. et al. (eds) *JBI Manual for Evidence Synthesis* (JBI, 2024); https://synthesismanual.jbi.global
67. Tricco, A. et al. PRISMA extension for scoping reviews (PRISMAScR): checklist and explanation. *Ann. Intern. Med.* **169**, 467–473 (2018).
68. Breiman, L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat. Sci.* **16**, 199–231 (2001).
69. *Covidence Systematic Review Software* https://www.covidence.org (Veritas Health Innovation, 2025).
70. Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N. & Folk, J. C. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit. Med.* **1**, 39 (2018).
71. Tonekaboni, S. et al. How to validate machine learning models prior to deployment: silent trial protocol for evaluation of real-time models at ICU. In *Proceedings of the Conference on Health, Inference, and Learning* Vol. 174 (eds Flores, G. et al.) 169–182 (PMLR, 2022).
72. Sendak, M. et al. "The human body is a black box": supporting clinical decision-making with deep learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* 99–109 (Association for Computing Machinery, 2020); https://doi.org/10.1145/3351095.3372827

73. Balagopalan, A. et al. Machine learning for healthcare that matters: reorienting from technical novelty to equitable impact. *PLOS Digit. Health* **3**, e0000474 (2024).
74. Papoutsi, C., Wherton, J., Shaw, S., Morrison, C. & Greenhalgh, T. Putting the social back into sociotechnical: case studies of co-design in digital health. *J. Am. Med. Inform. Assoc.* **28**, 284–293 (2021).
75. Alrajhi, A. A. et al. Data-driven prediction for COVID-19 severity in hospitalized patients. *Int. J. Environ. Res. Public Health* **19**, 2958 (2022).
76. Aydın, E. et al. Diagnostic accuracy of a machine learning-derived appendicitis score in children: a multicenter validation study. *Children (Basel)* **12**, 937 (2025).
77. Bachelot, G. et al. A machine learning approach for the prediction of testicular sperm extraction in nonobstructive azoospermia: algorithm development and validation study. *J. Med. Internet Res.* **25**, e44047 (2023).
78. Berg, W. A. et al. Toward AI-supported US triage of women with palpable breast lumps in a low-resource setting. *Radiology* **307**, e223351 (2023).
79. Butler, H. J. et al. Development of high-throughput ATR-FTIR technology for rapid triage of brain cancer. *Nat. Commun.* **10**, 4501 (2019).
80. Campanella, G. et al. Real-world deployment of a fine-tuned pathology foundation model for lung cancer biomarker detection. *Nat. Med.* **31**, 3002–3010 (2025).
81. Chen, Y. et al. Endoscopic ultrasound-based radiomics for predicting pathologic upgrade in esophageal low-grade intraepithelial neoplasia. *Surg. Endosc.* **39**, 2239–2249 (2025).
82. Cheng, Y. et al. Two-year hypertension incidence risk prediction in populations in the desert regions of northwest China: prospective cohort study. *J. Med. Internet Res.* **27**, e68442 (2025).
83. Chiang, D.-H., Jiang, Z., Tian, C. & Wang, C.-Y. Development and validation of a dynamic early warning system with time-varying machine learning models for predicting hemodynamic instability in critical care: a multicohort study. *Crit. Care* **29**, 318 (2025).
84. Chufal, K. S. et al. Machine learning model for predicting DIBH non-eligibility in left-sided breast cancer radiotherapy: development, validation and clinical impact analysis. *Radiother. Oncol.* **205**, 110764 (2025).
85. Coley, R. Y., Walker, R. L., Cruz, M., Simon, G. E. & Shortreed, S. M. Clinical risk prediction models and informative cluster size: assessing the performance of a suicide risk prediction algorithm. *Biom. J.* **63**, 1375–1388 (2021).
86. Corbin, C. K. et al. DEPLOYR: a technical framework for deploying custom real-time machine learning models into the electronic medical record. *J. Am. Med. Inform. Assoc.* **30**, 1532–1542 (2023).
87. Dave, C. et al. Prospective real-time validation of a lung ultrasound deep learning model in the ICU. *Crit. Care Med.* **51**, 301–309 (2023).
88. El Moheb, M. et al. An open-architecture AI model for CPT coding in breast surgery: development, validation, and prospective testing. *Ann. Surg.* **282**, 439–448 (2025).
89. Faqar-Uz-Zaman, S. F. et al. The diagnostic efficacy of an app-based diagnostic health care application in the emergency room: eRadaR-trial. A prospective, double-blinded, observational study. *Ann. Surg.* **276**, 935–942 (2022).
90. Felmingham, C. et al. Improving Skin cancer Management with ARTificial Intelligence (SMARTI): protocol for a preintervention/postintervention trial of an artificial intelligence system used as a diagnostic aid for skin cancer management in a specialist dermatology setting. *BMJ Open* **12**, e050203 (2022).
91. Feng, W. et al. Identifying small thymomas from other asymptomatic anterior mediastinal nodules based on CT images using logistic regression. *Front. Oncol.* **15**, 1590710 (2025).
92. Hanley, D. et al. Emergency department triage of traumatic head injury using a brain electrical activity biomarker: a multisite prospective observational validation trial. *Acad. Emerg. Med.* **24**, 617–627 (2017).
93. Hoang, M. T. et al. Evaluating the utility of a clinical sepsis AI tool in emergency waiting rooms: a preliminary silent trial. *Stud. Health Technol. Inform.* **329**, 307–311 (2025).
94. Im, H. et al. Design and clinical validation of a point-of-care device for the diagnosis of lymphoma via contrast-enhanced microholography and machine learning. *Nat. Biomed. Eng.* **2**, 666–674 (2018).
95. Korfiatis, P. et al. Automated artificial intelligence model trained on a large data set can detect pancreas cancer on diagnostic computed tomography scans as well as visually occult preinvasive cancer on prediagnostic computed tomography scans. *Gastroenterology* **165**, 1533–1546 (2023).
96. Kramer, D. et al. Machine learning-based prediction of malnutrition in surgical in-patients: a validation pilot study. *Stud. Health Technol. Inform.* **313**, 156–157 (2024).
97. Kwong, J. C. C. et al. The silent trial—the bridge between bench-to-bedside clinical AI applications. *Front. Digit. Health* **4**, 929508 (2022).
98. Liu, R. et al. Development and prospective validation of postoperative pain prediction from preoperative EHR data using attention-based set embeddings. *NPJ Digit. Med.* **6**, 209 (2023).
99. Liu, Y. et al. Validation of an established TW3 artificial intelligence bone age assessment system: a prospective, multicenter, confirmatory study. *Quant. Imaging Med. Surg.* **14**, 144–159 (2024).
100. Luo, H. et al. Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: a multicentre, case–control, diagnostic study. *Lancet Oncol.* **20**, 1645–1654 (2019).
101. Lupei, M. I. et al. A 12-hospital prospective evaluation of a clinical decision support prognostic algorithm based on logistic regression as a form of machine learning to facilitate decision making for patients with suspected COVID-19. *PLoS ONE* **17**, e0262193 (2022).
102. Mahajan, A. et al. Development and validation of a machine learning model to identify patients before surgery at high risk for postoperative adverse events. *JAMA Netw. Open* **6**, e2322285 (2023).
103. Major, V. J. & Aphinyanaphongs, Y. Development, implementation, and prospective validation of a model to predict 60-day end-of-life in hospitalized adults upon admission at three sites. *BMC Med. Inform. Decis. Mak.* **20**, 214 (2020).
104. Miró Catalina, Q. et al. Real-world testing of an artificial intelligence algorithm for the analysis of chest X-rays in primary care settings. *Sci. Rep.* **14**, 5199 (2024).
105. O'Brien, C. et al. Development, implementation, and evaluation of an in-hospital optimized early warning score for patient deterioration. *MDM Policy Pract.* **5**, 2381468319899663 (2020).
106. Pan, Y. et al. An interpretable machine learning model based on optimal feature selection for identifying CT abnormalities in patients with mild traumatic brain injury. *EClinicalMedicine* **82**, 103192 (2025).
107. Pyrros, A. et al. Opportunistic detection of type 2 diabetes using deep learning from frontal chest radiographs. *Nat. Commun.* **14**, 4039 (2023).
108. Qian, Y.-F., Zhou, J.-J., Shi, S.-L. & Guo, W.-L. Predictive model integrating deep learning and clinical features based on ultrasound imaging data for surgical intervention in intussusception in children younger than 8 months. *BMJ Open* **15**, e097575 (2025).
109. Rawson, T. M. et al. Supervised machine learning to support the diagnosis of bacterial infection in the context of COVID-19. *JAC Antimicrob. Resist.* **3**, dlab002 (2021).

110. Ren, L.-J. et al. Artificial intelligence assisted identification of newborn auricular deformities via smartphone application. *EClinicalMedicine* **81**, 103124 (2025).
111. Schinkel, M. et al. Diagnostic stewardship for blood cultures in the emergency department: a multicenter validation and prospective evaluation of a machine learning prediction tool. *EBioMedicine* **82**, 104176 (2022).
112. Shah, P. K. et al. A simulated prospective evaluation of a deep learning model for real-time prediction of clinical deterioration among ward patients. *Crit. Care Med.* **49**, 1312–1321 (2021).
113. Shamout, F. E. et al. An artificial intelligence system for predicting the deterioration of COVID-19 patients in the emergency department. *NPJ Digit. Med.* **4**, 80 (2021).
114. Shi, Y.-H. et al. Construction and validation of machine learning-based predictive model for colorectal polyp recurrence one year after endoscopic mucosal resection. *World J. Gastroenterol.* **31**, 102387 (2025).
115. Smith, S. J., Bradley, S. A., Walker-Stabeler, K. & Siafakas, M. A prospective analysis of screen-detected cancers recalled and not recalled by artificial intelligence. *J. Breast Imaging* **6**, 378–387 (2024).
116. Stamatopoulos, N. et al. Temporal and external validation of the algorithm predicting first trimester outcome of a viable pregnancy. *Aust. N. Z. J. Obstet. Gynaecol.* **65**, 128–134 (2025).
117. Swinnerton, K. et al. Leveraging near-real-time patient and population data to incorporate fluctuating risk of severe COVID-19: development and prospective validation of a personalised risk prediction tool. *EClinicalMedicine* **81**, 103114 (2025).
118. Tariq, A., Patel, B. N., Sensakovic, W. F., Fahrenholtz, S. J. & Banerjee, I. Opportunistic screening for low bone density using abdominopelvic computed tomography scans. *Med. Phys.* **50**, 4296–4307 (2023).
119. Titano, J. J. et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nat. Med.* **24**, 1337–1341 (2018).
120. Vaid, A. et al. Machine learning to predict mortality and critical events in a cohort of patients with COVID-19 in New York City: model development and validation. *J. Med. Internet Res.* **22**, e24018 (2020).
121. Wall, P. D. H., Hirata, E., Morin, O., Valdes, G. & Witztum, A. Prospective clinical validation of virtual patient-specific quality assurance of volumetric modulated arc therapy radiation therapy plans. *Int. J. Radiat. Oncol. Biol. Phys.* **113**, 1091–1102 (2022).
122. Wan, C.-F. et al. Radiomics of multimodal ultrasound for early prediction of pathologic complete response to neoadjuvant chemotherapy in breast cancer. *Acad. Radiol.* **32**, 1861–1873 (2025).
123. Wang, X. et al. Prediction of the 1-year risk of incident lung cancer: prospective study using electronic health records from the state of Maine. *J. Med. Internet Res.* **21**, e13260 (2019).
124. Wang, L., Wu, H., Wu, C., Shu, L. & Zhou, D. A deep-learning system integrating electrocardiograms and laboratory indicators for diagnosing acute aortic dissection and acute myocardial infarction. *Int. J. Cardiol.* **423**, 133008 (2025).
125. Wissel, B. D. et al. Prospective validation of a machine learning model that uses provider notes to identify candidates for resective epilepsy surgery. *Epilepsia* **61**, 39–48 (2020).
126. Xie, Z. et al. Enhanced diagnosis of axial spondyloarthritis using machine learning with sacroiliac joint MRI: a multicenter study. *Insights Imaging* **16**, 91 (2025).
127. Ye, C. et al. A real-time early warning system for monitoring inpatient mortality risk: prospective study using electronic medical record data. *J. Med. Internet Res.* **21**, e13719 (2019).
128. Ye, J.-Z. et al. Nomogram for prediction of the International Study Group of Liver Surgery (ISGLS) grade B/C posthepatectomy liver failure in HBV-related hepatocellular carcinoma patients: an external validation and prospective application study. *BMC Cancer* **20**, 1036 (2020).
129. Yu, S. C. et al. Sepsis prediction for the general ward setting. *Front. Digit. Health* **4**, 848599 (2022).
130. Zhang, Z. et al. Development of an MRI based artificial intelligence model for the identification of underlying atrial fibrillation after ischemic stroke: a multicenter proof-of-concept analysis. *EClinicalMedicine* **81**, 103118 (2025).
131. Escalé-Besa, A. et al. Using artificial intelligence as a diagnostic decision support tool in skin disease: protocol for an observational prospective cohort study. *JMIR Res. Protoc.* **11**, e37531 (2022).
132. Faqar-Uz-Zaman, S. F. et al. Study protocol for a prospective, double-blinded, observational study investigating the diagnostic accuracy of an app-based diagnostic health care application in an emergency room setting: the eRadaR trial. *BMJ Open* **11**, e041396 (2021).
133. Felmingham, C. et al. Improving skin cancer management with ARTificial intelligence: a pre–post intervention trial of an artificial intelligence system used as a diagnostic aid for skin cancer management in a real-world specialist dermatology setting. *J. Am. Acad. Dermatol.* **88**, 1138–1142 (2023).
134. Miró Catalina, Q., Fuster-Casanovas, A., Solé-Casals, J. & Vidal-Alaball, J. Developing an artificial intelligence model for reading chest x-rays: protocol for a prospective validation study. *JMIR Res. Protoc.* **11**, e39536 (2022).
135. Sheppard, J. P., Martin, U., Gill, P., Stevens, R. & McManus, R. J. Prospective Register Of patients undergoing repeated OFfice and Ambulatory Blood Pressure Monitoring (PROOF-ABPM): protocol for an observational cohort study. *BMJ Open* **6**, e012607 (2016).

Acknowledgements

This scoping review is part of a larger study exploring and expanding considerations for silent phase evaluations of healthcare AI. The Project CANAIRI Steering Group has been involved in the design of this overall work, and we would like to thank L. Oakden-Rayner, M. Mamdani, J. Louise and L. A. Smith. No funding source supported this scoping study. M.D.McC. gratefully acknowledges salary support from The Hospital Research Foundation Group. Additional support for CANAIRI-related activities has been generously provided by the Australian Institute for Machine Learning through the Centre for Augmented Reasoning.

Author contributions

Study design and conceptualization: M.D.McC., X.L., L.T. Methodology: M.D.McC., X.L., L.T., C.S., A.v.d.V., M.P.S., K.V., J.A.A., S.J., A.J.L., A.M., I.S., I.A., S.S., L.E., L.-A.F. Data extraction and analysis: L.T., A.M., H.T., N.C.K., G.K., S.B., N.P., A.v.d.V., L.E., L.J.P., C.S. Data interpretation and synthesis: all authors. Writing—first draft: L.T. Writing—review and editing: all authors. Project supervision: M.D.McC.

Funding

Open access funding provided by Adelaide University.

Competing interests

S.R.P. is an employee of Google and may own stock as part of a standard compensation package. X.L. is an employee of Microsoft. M.P.S. is a co-inventor of software licensed from Duke University to Cohere Med, Inc., KelaHealth, Fullsteam Health and Clinetic. M.P.S. owns equity in Clinetic. M.D.McC. discloses financial support related to independent ethics consultation activities for Google Health (USA), Cephalgo and iheed. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44360-025-00048-z.

Correspondence and requests for materials should be addressed to Lana Tikhomirov.

Peer review information *Nature Health* thanks Stephen Gilbert and Pearse Keane for their contribution to the peer review of this work. Primary Handling Editor: Lorenzo Righetto, in collaboration with the *Nature Health* team.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2026

Lana Tikhomirov[1,2] ✉, Carolyn Semmler[1,2], Noah Prizant[3], Srijan Bhasin[3,4], Georgia Kenyon[1], Anton van der Vegt[5], Lauren Erdman[6,7], Nikhil Cherian Kurian[1], Humphrey Thompson[1,8], Lyle J. Palmer[1,8], Abdullahi Mohamud[9], Judy Wawira Gichoya[10], Seyi Soremekun[11], Mark P. Sendak[3], James A. Anderson[9,12,13], Stephen R. Pfohl[14], Ian Stedman[15], Daniel Ehrmann[16,17], Karin Verspoor[18,19], Jethro C. C. Kwong[20,21], Lesley-Anne Farmer[22], Alex John London[23], Ismail Akrout[24], Shalmali Joshi[25], Elena Dicus[26], Xiaoxuan Liu[27] & Melissa D. McCradden[1,9,28]

[1]Australian Institute for Machine Learning, Adelaide University, Adelaide, South Australia, Australia. [2]School of Psychology, Adelaide University, Adelaide, South Australia, Australia. [3]Duke Institute for Health Innovation (DIHI), Duke University School of Medicine, Durham, NC, USA. [4]Duke University School of Medicine, Durham, NC, USA. [5]Queensland Digital Health Centre, The University of Queensland, Brisbane, Queensland, Australia. [6]Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [7]University of Cincinnati School of Medicine, Cincinnati, OH, USA. [8]School of Public Health, Adelaide University, Adelaide, South Australia, Australia. [9]SickKids Research Institute, Toronto, Ontario, Canada. [10]Department of Radiology and Imaging Sciences, Emory School of Medicine, Atlanta, GA, USA. [11]Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, UK. [12]Bioethics Department, The Hospital for Sick Children, Toronto, Ontario, Canada. [13]Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada. [14]Google Research, Mountain View, CA, USA. [15]School of Public Policy and Administration, York University, Toronto, Ontario, Canada. [16]Division of Cardiology, Department of Pediatrics, University of Michigan Medical School, Ann Arbor, MI, USA. [17]The Weil Institute for Critical Care Research and Innovation, University of Michigan Medical School, Ann Arbor, MI, USA. [18]School of Computing Technologies, RMIT University, Melbourne, Victoria, Australia. [19]School of Computing and Information Systems, The University of Melbourne, Melbourne, Victoria, Australia. [20]Division of Urology, Department of Surgery, University of Toronto, Toronto, Ontario, Canada. [21]Temerty Centre for AI Research and Education in Medicine, University of Toronto, Toronto, Ontario, Canada. [22]Tethyan Consulting, Canberra, Australian Capital Territory, Australia. [23]Carnegie Mellon University, Pittsburgh, PA, USA. [24]The Hospital for Sick Children, Toronto, Ontario, Canada. [25]Columbia University, New York, NY, USA. [26]Central Adelaide Local Health Network, Adelaide, South Australia, Australia. [27]Department of Applied Health Sciences, University of Birmingham, Birmingham, UK. [28]Women's and Children's Health Network, Adelaide, South Australia, Australia. ✉e-mail: lana.tikhomirov@adelaide.edu.au

nature portfolio

Corresponding author(s): Lana Tikhomirov

Last updated by author(s): Aug 12, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	*Provide a description of all commercial, open source and custom code used to collect the data in this study, specifying the version used OR state that no software was used.*
Data analysis	R Software (2025) and RStudio (2025) were used to generate Figure 3.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Database containing extracted studies is available at: https://docs.google.com/spreadsheets/d/17CFyfViM0IMPQYnBquQ16H-fqGtYvNT9D-wCX5zZO4I/edit?usp=sharing

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Available in PRISMA diagram (Figure 1)
Data exclusions	Exclusion criteria was pre-established and is available in the methods section
Replication	As a scoping study, we support replicability by transparently reporting our methods undertaken.
Randomization	N/A
Blinding	We ensured independent review of each included study

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A