ERJ open research
THE BEST IN OPEN ACCESS BASIC, TRANSLATIONAL & CLINICAL RESPIRATORY RESEARCH

Early View

Mini-Review

Large Language Models in Respiratory Care: Safety and Clinical Integration

Alessandro Porcella, Antonio Fabozzi, Matteo Bonini, Paolo Palange

Please cite this article as: Porcella A, Fabozzi A, Bonini M, *et al*. Large Language Models in Respiratory Care: Safety and Clinical Integration. *ERJ Open Res* 2026; in press (https://doi.org/10.1183/23120541.01699-2025).

R1

Large Language Models in Respiratory Care: Safety and Clinical Integration

Alessandro Porcella[1], Antonio Fabozzi[1], Matteo Bonini[1], Paolo Palange[1]

[1] Department of Public Health and Infectious Diseases, Respiratory and Critical Care Division, Sapienza University of Rome, Rome, Italy

Corresponding author:

Paolo Palange, MD

Department of Public Health and Infectious Diseases,

Respiratory and Critical Care Division

Sapienza University of Rome

Piazzale Aldo Moro 5, 00185 Rome, Italy

E-mail: paolo.palange@uniroma1.it

Abstract

Background: Large language models (LLMs) are rapidly entering respiratory medicine workflows. Their clinical role remains unclear. A central concern is whether they function as autonomous decision makers or as tools that depend on clinician input.

Methods: We conducted a structured narrative review across three domains aligned with respiratory practice. Bedside clinical decision support and triage, pulmonary function test (PFT)/spirometry interpretation, and chest X-ray (CXR) reporting. We included quantitative studies comparing LLMs or specialised AI systems with clinician performance, and extracted accuracy, agreement, and safety-critical failure modes.

Results: At bedside, LLMs approached expert diagnosis in straightforward cases, but performance fell in highly-complex or multi-step management tasks. The dominant error was omission of required actions, unsafe under-triage or incomplete plans. In PFTs, guideline-anchored and spirogram-trained LLMs achieved high agreement and reduced inter-observer variability on structured inputs. In contrast, generalist LLMs showed only moderate reliability and inconsistent handling of borderline or mixed patterns. In CXR reporting, chest-specific multimodal systems outperformed generalist LLMs, with fewer hallucinations and better clinical acceptability; yet clinically relevant discrepancies persisted without radiologist review.

Conclusions: Across domains, AI does not replace respiratory expertise. When inputs are complete, structured, and guideline-concordant, outputs stabilise and add value. When inputs are ambiguous or incomplete, errors expand, often through dangerous omissions. Safe deployment therefore requires specialised training, strict input structure, and continuous clinician oversight.

Introduction

Introduction

Large language models (LLMs) are moving into clinical practice faster than local governance can adapt. They generate fluent medical text, summarise evidence, and propose actions. This creates a tempting “digital curbside consult.” Yet fluency is not safety. In high-stakes care, the key question is whether a system remains guideline-concordant and stable when case complexity or input format changes [1–2].

Respiratory medicine is a stringent test-bed for LLMs. Acute decisions require rapid triage, risk stratification, and multi-step management. Many actions are mandatory. Even small omissions can shift outcomes [3–4]. Functional respiratory diagnostics add a second layer. PFT interpretation relies on fixed thresholds, pattern hierarchies, and quality control. Errors cluster in mixed patterns and in borderline obstruction or restriction [5–6]. Imaging adds a third layer. CXR interpretation depends on subtle visual patterns and clear anatomic localisation. Generalist LLMs are not trained for that. Multimodal, chest-tuned models can reduce hallucination, but still require oversight [7–8].

Three domains map directly to respiratory workflows and expose distinct error modes.

First, bedside clinical advice and guideline adherence. Real-world and vignette studies across specialties—including hypertension, mental health, and internal medicine—show that LLMs can approach expert diagnosis in simple cases [9–14]. However, evidence indicates they frequently omit management steps and under-triage severe patients [3–4]. Reasoning-enhanced models improve concordance but do not eliminate omission-type failures.

Second, PFT and spirometry interpretation. Guideline-anchored AI systems and recent primary care trials demonstrate that AI can outperform average clinicians for pattern classification and COPD detection when curve quality is adequate [5,15–18]. Conversely, generalist LLMs (like standard GPT-4) have only moderate agreement with pulmonologists and struggle with mixed patterns unless the model is specifically spirogram-trained [19–22].

Third, chest imaging. Specialised multimodal CXR systems generally outperform generalist LLMs in report acceptability, summarization, and localisation [7–8, 23–25]. Nonetheless, systematic reviews warn that residual hallucinations and missed findings remain, especially out-of-distribution, and clinician review improves safety [26].

A primary safety pillar is out-of-distribution (OOD) failure [27]. Most LLMs are trained on common presentations. In respiratory medicine, the highest-risk cases are often precisely those that are rare, atypical, or multi-morbidity—settings where training coverage is uncertain. When faced with rare diseases (e.g., specific ILD patterns or orphan lung diseases) absent from their training sets, models may fail to declare uncertainty and instead force the presentation into a common diagnosis or produce a fluent but unsafe plan [28]. Instead, they may "hallucinate" a common diagnosis or force the case into a standard category [29]. This behavior poses a significant risk in tertiary care settings where rare diseases are prevalent. Current systems largely lack the self-awareness to flag OOD inputs, necessitating strict human oversight for atypical cases [30,31].

Across these domains the same question repeats: do current AI systems reduce risk by standardising respiratory decisions, or do they amplify risk by producing plausible but unsafe outputs? This dichotomy highlights two opposing forces. On one side, AI holds the potential to act as a "Safety Net," detecting subtle findings (e.g., small pneumothorax or rare patterns) that a weary clinician might miss due to cognitive fatigue. On the other, it introduces the risk

of "Automation Bias," where clinicians—trusting the fluent and confident tone of an LLM—may accept erroneous outputs without critical verification. This review examines whether current evidence supports AI as a reliable safety net or merely a sophisticated, yet fallible, amplifier of input quality. Crucially, LLM failure is not only 'more of the same' human error. These systems can generate novel, non-human failure modes that clinicians are not trained to anticipate—confident hallucinations, guideline-sounding but incorrect rationales, and brittle behavior to small changes in input format. This elevates automation bias from a theoretical concern to a practical safety hazard. Mitigations therefore require workflow design, not only user vigilance: structured input templates, forced cross-checks of key red-flag items, explicit uncertainty prompts, and 'no-autopilot' policies for high-risk decisions (e.g., triage level, escalation, and discharge safety-netting).

Objective

To combine and summarize numerical evaluations of AI systems in (i) bedside clinical advice and guideline adherence, (ii) PFT interpretation, and (iii) chest imaging, focusing on performance, dominant error modes, and operational constraints for safe respiratory deployment.

Methods

Review Design and Scope

This work was conducted as a structured narrative review of clinical evaluations of AI systems, following ERS and ERJ Open Research expectations for methodological transparency. The review focused on three predefined domains directly relevant to respiratory clinical workflows:

(1) bedside clinical advice and guideline-driven decision support;

(2) interpretation of pulmonary function tests (PFT) and spirometry;

(3) chest imaging, with emphasis on chest X-ray (CXR) report generation and reasoning.

These three domains were selected a priori because they represent high-volume, high-risk workflows in respiratory medicine in which large language models (LLMs) and multimodal systems are being actively tested.

Eligibility Criteria

Inclusion criteria

We included studies meeting all criteria below:

1. Population / setting: clinical, near-clinical, or clinically validated vignette settings relevant to acute care, respiratory diagnostics, or thoracic imaging.
2. Intervention: any artificial intelligence (AI) system capable of generating clinical decisions, interpretations, or reports, including general LLMs, domain-tuned LLMs, multimodal vision–language models, or rule-based AI.
3. Comparator: expert clinicians, human panels, or recognised guideline standards (ATS/ERS, GOLD, NASS, or institutional protocols).
4. Outcomes: quantitative performance metrics (accuracy, sensitivity/specificity, AUROC, or structured appropriateness ratings), plus qualitative assessment of error types as reported by the authors.
5. Design: prospective, retrospective, benchmark or observational comparisons.

Exclusion criteria

We excluded:

– studies evaluating only model architecture or training procedures without clinical endpoints;

– purely synthetic or unvalidated datasets;

– studies lacking a human or guideline gold standard;

– purely opinion, commentaries, letters, or non-quantitative viewpoints;

– radiology works not involving thoracic modalities;

– PFT studies without pattern classification or explicit diagnostic comparison.

Search Strategy

A targeted search was performed in PubMed, Scopus, and Google Scholar between January 2019 and January 2025, reflecting the emergence of modern LLMs.

Search terms combined:

- *"large language model"*, *"GPT"*, *"AI decision support"*, *"clinical reasoning"*, *"triage"*, *"pulmonary function test"*, *"spirometry interpretation"*, *"chest X-ray"*, *"vision–language model"*, *"radiology report generation"*;
- Filters: human, English, clinical, quantitative evaluation.

We added newly identified studies only if they matched the predefined domains.

Study identification and selection followed a PRISMA-style workflow. A targeted search was performed in PubMed, Scopus, and Google Scholar between January 2019 and November 2025 using predefined respiratory-relevant domains and quantitative filters. Titles/abstracts

were screened, and full texts were evaluated to confirm eligibility. In total, 68 unique records were screened; 42 were excluded based on prespecified criteria, and 26 studies met inclusion criteria and were included in the review. While this number may appear limited given the exponential growth of AI literature, it reflects our deliberate focus on "clinical safety quantification" rather than technical model architecture. A PRISMA flow diagram detailing the screening process and reasons for exclusion is provided Figure 1.

Given the mini-review scope and the heterogeneity of study designs (benchmarks, retrospective comparisons, and one randomized trial), we did not perform a formal risk-of-bias appraisal with a single tool. Instead, we explicitly distinguished evidence by setting and 'distance to practice': (i) real-world prospective/clinical deployments and randomized evaluations were interpreted as the highest-weight evidence for safety and integration; (ii) near-clinical, guideline-anchored vignette studies were considered supportive but more sensitive to prompt structure and case selection; and (iii) technical benchmark papers without a clinical gold standard were excluded a priori. This design-based weighting is reflected in the narrative synthesis and in the interpretation of safety claims.

Study Selection

Titles and abstracts were screened to determine whether the study addressed one of the three respiratory-relevant use cases. Full-text evaluation was then performed to confirm eligibility.

Priority was given to studies that:

– directly compared model output with expert decisions;

– offered granular metrics stratified by case difficulty or severity;

– characterised model errors.

No disagreements occurred because selection was performed with a priori inclusion rules.

Data Extraction

From each study, we extracted:

1. Model type (general LLMs, domain-specific LLMs, multimodal vision–language model, guideline-anchored algorithm, or DL classifier).
2. Clinical task (triage, diagnostic reasoning, therapy planning, PFT pattern recognition, COPD classification, CXR interpretation/reporting).
3. Gold standard comparator (expert panel, clinicians, official guideline).
4. Quantitative metrics (accuracy, AUROC, sensitivity/specificity appropriateness score).
5. Failure modes: omission errors, under-triage, hallucinations, guideline-drift, low-quality test sensitivity, out-of-distribution (OOD) errors.
6. Authors' statements on safety, generalisability, oversight, and deployment constraints.

No numerical recalculation was performed, consistent with narrative review methodology.

Synthesis Approach

Given heterogeneity in tasks, labels, and datasets, pooling was not performed.

Instead, we used a domain-segmented synthesis, maintaining comparability by:

– grouping bedside/triage LLMs by clinical task type,

– grouping PFT works by diagnostic pattern classification,

– grouping imaging studies by modality (all CXR).

Each domain was summarised using a structured table listing:
study identifier, task, main quantitative endpoint, and primary safety conclusion.
Narrative synthesis highlighted recurrent failure modes and operational implications.

Generative AI tools (ChatGPT, OpenAI) were used to assist with language refinement. All contents, data interpretation, and reference checking were performed and verified by the authors.

Results

Our targeted search yielded a focused set of studies that met the predefined eligibility criteria across the three domains, resulting in 26 included articles (10 bedside CDS/triage, 10 PFT/spirometry, 6 CXR imaging).

1) Bedside clinical advice / guideline adherence

In emergency cases and structured benchmarks, LLMs reached near-expert accuracy on straightforward diagnosis. Performance fell when tasks required triage, test selection, or full management plans. The dominant high-risk error was omission: missing therapies, missing follow-up, or incomplete non-pharmacologic actions. Under-triage of severe cases was repeatedly observed. Guideline-prompting and reasoning-focused models improved concordance but did not remove unsafe recommendations.

2) PFT / spirometry interpretation

Guideline-anchored AI showed high agreement with clinical PFT reports and reduced inter-rater variability. Deep-learning spirogram classifiers matched or exceeded physicians when curve quality was adequate. Generalist LLMs achieved only moderate agreement and were highly sensitive to input format and framing. Errors concentrated in mixed patterns,

restriction, and borderline small-airway disease. A spirogram-trained multimodal LLMs improved stability and COPD report performance compared with generalist systems.

3) Chest imaging (CXR)

Chest-specific multimodal systems generated more accurate CXR reports than generalist LLMs, with fewer hallucinations and better localisation. Generalist LLM still invented findings or missed subtle abnormalities, especially in atypical cases.

Discussion

Across bedside decision support, PFT and CXR, the included studies converge toward a single, consistent conclusion: AI systems do not replace clinical reasoning; they amplify it. They magnify strengths when clinicians provide structured, guideline-based inputs, and they magnify weaknesses when inputs lack completeness, clarity, or clinical relevance. This amplification principle explains the cross-domain pattern of failure modes observed in the reviewed literature.

First, bedside studies show that LLMs can reproduce expert-level diagnostic labels in straightforward cases, but this strength collapses when clinical action requires multistep reasoning. Even the highest-performing reasoning-enhanced systems match experts only when the clinical narrative is clear, complete, and explicitly anchored to guideline logic. When these conditions are absent, models do not merely fail: they amplify the incompleteness of the input, producing management plans missing therapies, safety steps, or follow-up. Authors consistently interpret these omissions as the dominant and most

clinically consequential failure mode [1–4,6,7,9]. This behaviour is not randomness; it is amplification of input gaps. If information is missing or ambiguously phrased, the model does not correct it — it expands it into a more detailed, but still incomplete, plan.

Second, the same amplification behaviour appears in the domain of PFT interpretation. Guideline-anchored AI systems perform well because their inputs are structured, numerical, and standardised. When rulesets or DL-based spirogram classifiers operate on high-quality curves, they reproduce the hierarchical diagnostic logic reliably [11–13,15,19]. But generalist LLM, when asked to interpret PFT traces or numerical reports presented as free-text, degrade sharply. They amplify the ambiguity of unstructured inputs: borderline obstruction becomes over- or under-called; restriction is missed; mixed defects are interpreted inconsistently [17,18]. In primary care settings, AI-supported spirometry improves general practitioners' diagnostic agreement and confidence compared with unaided interpretation, reinforcing its role as an assistive amplifier rather than an autonomous interpreter [14]. The common factor is not the task — it is the *structure* of the input. When the input enforces physiology and guidelines, AI performance stabilises. When the input is free-text or underspecified, the model amplifies ambiguity.

Third, chest imaging shows the amplification principle in its clearest form. Chest-specific multimodal models such as M4CXR or CXR-LLaVA improve accuracy, localisation, and faithfulness because they operate on highly structured pixel-level inputs aligned with radiological conventions [21–23,24,25]. Generalist LLMs, receiving only textual prompts, amplify linguistic priors rather than visual truth, generating hallucinated consolidations or omitting subtle findings. The failure is not stochastic: it reflects an underlying amplification of prompt bias. Radiologist-in-the-loop workflows reduce these risks precisely because they add the missing clinical context and visual correction [22][26].

Taken together, these results justify a strong, unifying interpretation: AI performance is not independent of clinician expertise; it is conditional on it. The more precise, complete, and guideline-concordant the human input, the more reliable the model output. Conversely, the more ambiguous or incomplete the input, the more severe the errors. This explains why autonomous deployment consistently fails in high-severity strata: these strata inherently contain incomplete, ambiguous, or noisy information, and the model amplifies these defects.

Implications for clinical integration and validation in respiratory medicine

This amplification principle has direct operational implications.

First, AI should not be viewed as a diagnostic engine but as a *precision multiplier*. It increases throughput and reduces variability when embedded in workflows where human domain knowledge constrains the input.

Second, oversight is not optional. It is the mechanism through which clinicians interrupt amplification of unsafe reasoning. Every domain showed that the clinician-in-the-loop configuration improves accuracy, avoids omissions, and corrects drift.

Third, local validation is mandatory because amplification interacts with local case-mix. A model trained on typical cases will amplify the distribution it knows — and mismanage the cases it has not seen.

Finally, respiratory medicine, more than other fields, relies on structured data: physiological thresholds, quality checks, standardised patterns, severity grading. This is precisely the environment where AI can be most helpful when properly constrained — and most harmful when unconstrained.

Accountability must be explicit. If an AI system ‘amplifies’ an incomplete input into an unsafe recommendation, responsibility cannot be left ambiguous between clinician, institution, and vendor. Safe integration requires auditable logs (inputs, model version, outputs), clear

escalation rules, and defined ownership of final decisions. Data security is equally limiting in real-world deployment: many generalist LLM workflows imply transmission of patient information to external servers, which may be unacceptable without enterprise agreements, strict governance, and data-minimization. Practical options include on-premise or institutionally hosted models, de-identification/redaction pipelines, and restricting LLM use to non-identifiable structured features when feasible.

Future directions should also address patient trust and the clinician–patient relationship. LLM-assisted decisions may improve consistency, but they can also erode trust if patients perceive that care is 'outsourced' to opaque systems or if AI-generated explanations are confident yet wrong. Implementation should therefore prioritize transparency (what the tool did and did not do), informed communication, and preservation of clinician accountability. Evaluations should move beyond accuracy to patient-centered outcomes—acceptability, perceived empathy, willingness to follow recommendations, and the impact of AI use on shared decision-making—particularly in chronic respiratory care where longitudinal trust is central. Morover, will require clinician 'LLM literacy' as a measurable competency: explicit training to counter automation bias, enforce structured inputs, and apply verification routines for safety-critical outputs. This shifts AI use from passive acceptance to calibrated supervision, which is also essential to preserve patient trust.

In summary, the evidence does not support the view of AI as an autonomous diagnostic entity. Instead, it supports a sharper, clinically grounded interpretation: LLMs amplify the clinician. They do not invent safety; they inherit it. They do not repair gaps; they enlarge them. In respiratory workflows — bedside, PFT, and imaging — the central driver of safety and performance is the clinician's control over input structure and oversight over outputs.

Without this, AI magnifies risk; with it, AI can standardise reasoning, surface omissions, and enhance efficiency.

Conclusion

Across bedside care, PFT interpretation, and chest imaging, current AI systems offer measurable value but fail key safety requirements for autonomy.

At the bedside, omissions and under-triage remain unacceptable.

In PFT, guideline-anchored AI is reliable; generalist LLMs are not.

In imaging, chest-specific multimodal models outperform generalist systems but still need radiologist oversight.

AI should be deployed only within supervised, task-specific, locally validated respiratory workflows.

Acknowledgements

The authors used ChatGPT Plus (version 5.1, Thinking mode; OpenAI, San Francisco, CA) to support English language refinement and editing of the manuscript. The authors reviewed and take full responsibility for the content.

Conflicts of Interest

The authors declare that they have no conflicts of interest related to this work.

Funding

No funding or financial support was received for the conduct of this study or the preparation of the manuscript.

Figure 1 Legend: PRISMA flow-diagram for selection of AI studies.

References

1. Vrdoljak J, Boban Z, Males I, et al. Evaluating large language and large reasoning models as decision support tools in emergency internal medicine. Comput Biol Med. 2025;192(Pt B):110351.
2. Qiu P, Wu C, Liu S, et al. Quantifying the reasoning abilities of large language models on real-world clinical cases (MedR-Bench). Nat Commun. 2025;16:7866. doi:10.1038/s41467-025-64769-1.
3. Colakca C, Ergin M, Ozensoy HS, et al. Emergency department triaging using ChatGPT based on emergency severity index principles: a cross-sectional study. Sci Rep. 2024;14:22106. doi:10.1038/s41598-024-73229-7.
4. Zaboli A, Brigo F, Brigiari G, et al. Chat-GPT in triage: still far from surpassing human expertise – an observational study. Am J Emerg Med. 2025;92:165–171. doi:10.1016/j.ajem.2025.03.028.
5. Topalovic M, Das N, Burgel PR, et al. Artificial intelligence outperforms pulmonologists in the interpretation of pulmonary function tests. Eur Respir J. 2019;53:1801660. doi:10.1183/13993003.01660-2018.
6. Huang YCT, et al. Development and evaluation of a computerized algorithm for the interpretation of pulmonary function tests. PLoS One. 2024;19:e0297519. doi:10.1371/journal.pone.0297519.
7. Lee RW, Lee KH, Yun JS, Kim MS, Choi HS. Comparative analysis of M4CXR, an LLM-based chest X-ray report generation model, and ChatGPT in radiological interpretation. J Clin Med. 2024;13(23):7057. doi:10.3390/jcm13237057.
8. Tanno R, Barrett DGT, Sellergren A, et al. Collaboration between clinicians and vision–language models in radiology report generation. Nat Med. 2025;31:599-608. doi:10.1038/s41591-024-03302-1.
9. Zand J, Miao J, Hommos MS, et al. Performance of large language models in analyzing common hypertension scenarios. Hypertension. 2025 Nov 3. doi:10.1161/HYPERTENSIONAHA.125.25492. Online ahead of print.

10. Perlis RH, Goldberg JF, Ostacher MJ, Schneck CD. Clinical decision support for bipolar depression using large language models. Neuropsychopharmacology. 2024;49(9):1412–1416. doi:10.1038/s41386-024-01841-2.

11. Labinsky H, Nagler LK, Krusche M, et al. Vignette-based comparative analysis of ChatGPT and specialist treatment decisions for rheumatic patients: results of the Rheum2Guide study. Rheumatol Int. 2024;44(10):2043-2053. doi:10.1007/s00296-024-05675-5.

12. Khoylyan A, Salvato J, Vazquez F, Girgis M, Tang A, Chen T. Evaluation of GPT-4 concordance with North American Spine Society guidelines for lumbar fusion surgery. North Am Spine Soc J. 2025;21:100580. doi:10.1016/j.xnsj.2024.100580.

13. Noda R, Tanabe K, Ichikawa D, Shibagaki Y. GPT-4's performance in supporting physician decision-making in nephrology multiple-choice questions. Sci Rep. 2025;15:15439. doi:10.1038/s41598-025-99774-3.

14. Hirosawa T, Harada Y, Mizuta K, et al. Evaluating ChatGPT-4's accuracy in identifying final diagnoses within differential diagnoses compared with those of physicians. JMIR Form Res. 2024;8:e59267. doi:10.2196/59267.

15. Sunjaya A, Edwards GD, Harvey J, et al. Validation of artificial intelligence spirometry diagnostic support software in primary care: a blinded diagnostic accuracy study. ERJ Open Res. 2025;11:00116-2025. doi:10.1183/23120541.00116-2025

16. Doe G, Banya W, Edwards GD, Topalovic M, El-Emir E, Evans RA, et al. AI-Assisted Spirometry Interpretation in Primary Care: A Randomized Controlled Trial. NEJM AI. 2025;2(8):e? (ahead of print). doi:10.1056/AIoa2400804

17. Saad T, Pandey R, Padarya S, et al. Application of artificial intelligence in the interpretation of pulmonary function tests. Cureus. 2025;17(4):e82056. doi:10.7759/cureus.82056.

18. Pandiarajan DSB, Mani DAP, Vadivelu DG. Enhancing pulmonary function test reporting with artificial intelligence: a retrospective observational study. TPM – Testing, Psychometrics, Methodology in Applied Psychology. 2025;32(S5):535-539.

19. Rohita S, Shanmuganathan A, Chitrakumar A, Mohan SP, Sam Selva Shruthi J. Accuracy of spirometric interpretation by artificial intelligence-based software in comparison with a qualified respiratory physician in a tertiary care center in Chengalpattu District. J Neonatal Surg. 2025;14(32S):813-820.

20. Park JH, Cheong C, Kang S, Lee I, Lee S, Yoon K, Lee H. Evaluating advanced large language models for pulmonary disease diagnosis using portable spirometer data: a comparative analysis of Gemini 1.5 Pro, GPT-4o, and Claude 3.5 Sonnet. In: Proceedings of the IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI 2024). IEEE; 2024. doi:10.1109/BHI62660.2024.10913702

21. Mei S, Long Y, Cao S, Han X, Geng S, Sun J, et al. SpiroLLM: finetuning pretrained large language models to understand spirogram time series with clinical validation in COPD reporting. arXiv. 2025;arXiv:2507.16145.

22. Yadav P, Rastogi V, Yadav A, Parashar P. Artificial Intelligence: A Promising Tool in Diagnosis of Respiratory Diseases. Intell Pharm. 2024;2(5):784–791. doi:10.1016/j.ipha.2024.05.002.

23. Lee S, Youn J, Kim H, Kim M, Yoon SH. CXR-LLaVA: a multimodal large language model for interpreting chest X-ray images. Eur Radiol. 2025;35(7). doi:10.1007/s00330-024-11339-6.

24. Serapio A, Delbrouck JB, Van Veen D, Langlotz CP, Chaudhari AS, Pauly JM, et al. An open-source fine-tuned large language model for radiology report impression summarization: multi-institutional evaluation and reader study. BMC Med Imaging. 2024;24:1435. doi:10.1186/s12880-024-01435-w.

25. Huang J, Neill L, Wittbrodt M, et al. Generative Artificial Intelligence for Chest Radiograph Interpretation in the Emergency Department. JAMA Netw Open. 2023;6(10):e2336100. doi:10.1001/jamanetworkopen.2023.36100.

26. Artsi Y, et al. Large language models in radiology reporting: systematic review of performance, safety, and clinical integration. Radiol Artif Intell. 2025;7(3):e240091. doi:10.1016/j.radain.2025.240091

27. Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, et al. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems 32 (NeurIPS 2019). Curran Associates, Inc.; 2019. p. 13991–14002.

28. Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. npj Digit Med. 2021;4(1):4. doi:10.1038/s41746-020-00367-3.

29. Savage T, Wang J, Gallo R, Boukil A, Patel V, Safavi-Naini SAA, et al. Large language model uncertainty proxies: discrimination and calibration for medical diagnosis and treatment. J Am Med Inform Assoc. 2025;32(1):139–149. doi:10.1093/jamia/ocae254.

30. Yang J, Zhou K, Li Y, Liu Z. Generalized out-of-distribution detection: a survey. arXiv [Preprint]. 2024;2110.11334. doi:10.48550/arXiv.2110.11334.

31. Kaur R, Jha S, Roy A, Park S, Sokolsky O, Lee I. Detecting OODs as datapoints with high uncertainty. arXiv [Preprint]. 2021;2108.06380. doi:10.48550/arXiv.2108.06380.

Table 1. Bedside decision support: synthetic summary

Study	Task	Main quantitative endpoint	Core safety conclusion
Vrdoljak et al., 2025 [1]	ED internal medicine decision support	Diagnostic concordance vs experts	Near-expert diagnosis in simple cases, but management omissions persist.
Qiu et al., 2025 [2]	Clinical reasoning benchmark (MedR-Bench)	Accuracy across exam/diagnosis/therapy	Highest on diagnosis, lower on tests/therapy; omission-driven risk.
Colakca et al., 2024 [3]	ED triage (ESI)	Agreement vs nurse triage	Moderate overall agreement; not adequate for autonomous triage. High-acuity recognition is better, but mis-triage risk persists without supervision
Zaboli et al., 2025 [4]	ED triage	Sens/spec by severity	Sensitivity drops in high-risk strata; autonomy unsafe.
Zand et al., 2025 [9]	Guideline vignettes (hypertension)	Appropriateness vs guideline	Unsafe recommendations persist; guardrails required.
Perlis et al., 2024 [10]	Guideline-prompted CDS	Concordance vs community clinicians	Improves choices vs community clinicians; adjunct only.
Labinsky et al., 2024 [11]	Treatment plan generation	Plan concordance vs specialists	Drift in complex/second-line therapy; supervision needed.
Khoylyan et al., 2024 [12]	Guideline concordance (surgery)	Concordance vs NASS guidance	High but imperfect agreement; adjunct only.
Noda et al., 2025 [13]	CDS in nephrology	Change in physician accuracy	Mean improvement with pockets of harm; domain validation required.
Hirosawa et al., 2024 [14]	Final diagnosis from differentials	Agreement vs physicians	Fair–good agreement; support role plausible, but real-world reliability remains uncertain and requires supervision

Table 2. PFT/spirometry: synthetic summary

Study	Task	Main quantitative endpoint	Core safety conclusion
Topalovic et al., 2019 [5]	Full PFT interpretation	Accuracy/κ vs pulmonologists	AI can exceed clinicians on pattern tasks; guideline anchoring key.
Huang et al., 2024 [6]	Rule-based PFT algorithm	Agreement vs clinical reports	High concordance; issues mainly in mildly abnormal tests.
Sunjaya et al., 2025 [15]	COPD detection from spirometry	Sens/spec/AUC vs gold standard	High accuracy for COPD; depends on test quality and oversight.
Doe et al., 2025 [16]	RCT of AI-assisted spirometry interpretation in primary care	Proportion of correctly interpreted spirometry sessions vs expert reference; subgroup accuracy for COPD and other chronic respiratory diseases.	AI decision support improves clinicians' diagnostic accuracy overall but AI alone can still misclassify complex cases; use as assistive tool, not replacement.
Saad et al., 2025 [17]	Supervised AI PFT pattern interpretation	Accuracy, sensitivity, specificity and AUC of AI models vs pulmonologists for multi-class PFT interpretation on >3,000 cases	Well-trained ML classifiers can match or exceed average pulmonologist performance for pattern recognition but still need clinical context and oversight.
Pandiarajan et al., 2025 [18]	AI-based automated reporting pipeline for full PFTs in a tertiary centre	Concordance of AI-generated reports with expert pulmonologist reports; completeness and standardisation of pattern description.	Structured AI reporting can standardise PFT interpretation and reduce variability, but final validation by a specialist remains mandatory
Rohita et al., 2025 [19]	ChatGPT-based spirometry interpretation vs respiratory physician on routine lab reports	Diagnostic accuracy and class-wise agreement vs respiratory physicians	ChatGPT achieves high agreement on straightforward patterns but shows clinically relevant errors on mixed/complex cases; cannot be used autonomously
Park et al., 2024 [20]	Head-to-head LLM on spirometry	Diagnostic accuracy, F1-score and calibration of each LLM against expert-labelled diagnoses	Variable performance across models; unstable across tasks, needs domain anchoring.
Mei et al., 2025 [21]	Spirogram→COPD report (SpiroLLM)	AUROC/report agreement	Specialised training improves consistency; OOD risk remains.

Study	Task	Main quantitative endpoint	Core safety conclusion
Yadav et al., 2024 [22]	Narrative review of AI tools for diagnosis of respiratory diseases	No primary dataset; synthesis of diagnostic performance metrics reported across AI models for respiratory conditions	AI shows promise for respiratory diagnostics but evidence is heterogeneous

Table 3. Chest imaging: synthetic summary

Study	Task	Main quantitative endpoint	Core safety conclusion
Lee et al., 2024 [7]	CXR reporting: specialised vs general LLM	Acceptability/localisation vs radiologists	Domain-tuned model outperforms generalist; fewer hallucinations.
Tanno et al., 2025 [22]	Vision-language reporting with review	Expert preference/quality	Clinician-in-the-loop yields clinically acceptable reports.
Lee et al., 2025 [23]	Multimodal CXR assistant (CXR-LLaVA)	Benchmark vs multimodal LLMs (GPT-4-vision, Gemini) and radiologist-rated report acceptability	Promising but dataset-dependent.
Serapio et al., 2024 [24]	Radiology-tuned LLM summarisation	Clinical accuracy and robustness vs base LLMs across institutions	Specialisation reduces semantic errors.
Huang J et al., 2023 [25]	AI-generated ED chest X-ray reports (clinical validation vs radiologists)	Accuracy and clinical agreement	High overall clinical accuracy; clinically significant discrepancies still occur, requiring radiologist oversight.
Artsi et al., 2025 [26]	Systematic review of LLM radiology reporting	Fidelity / hallucination rates clinical integration limits	Human oversight and guardrails remain mandatory before deployment.

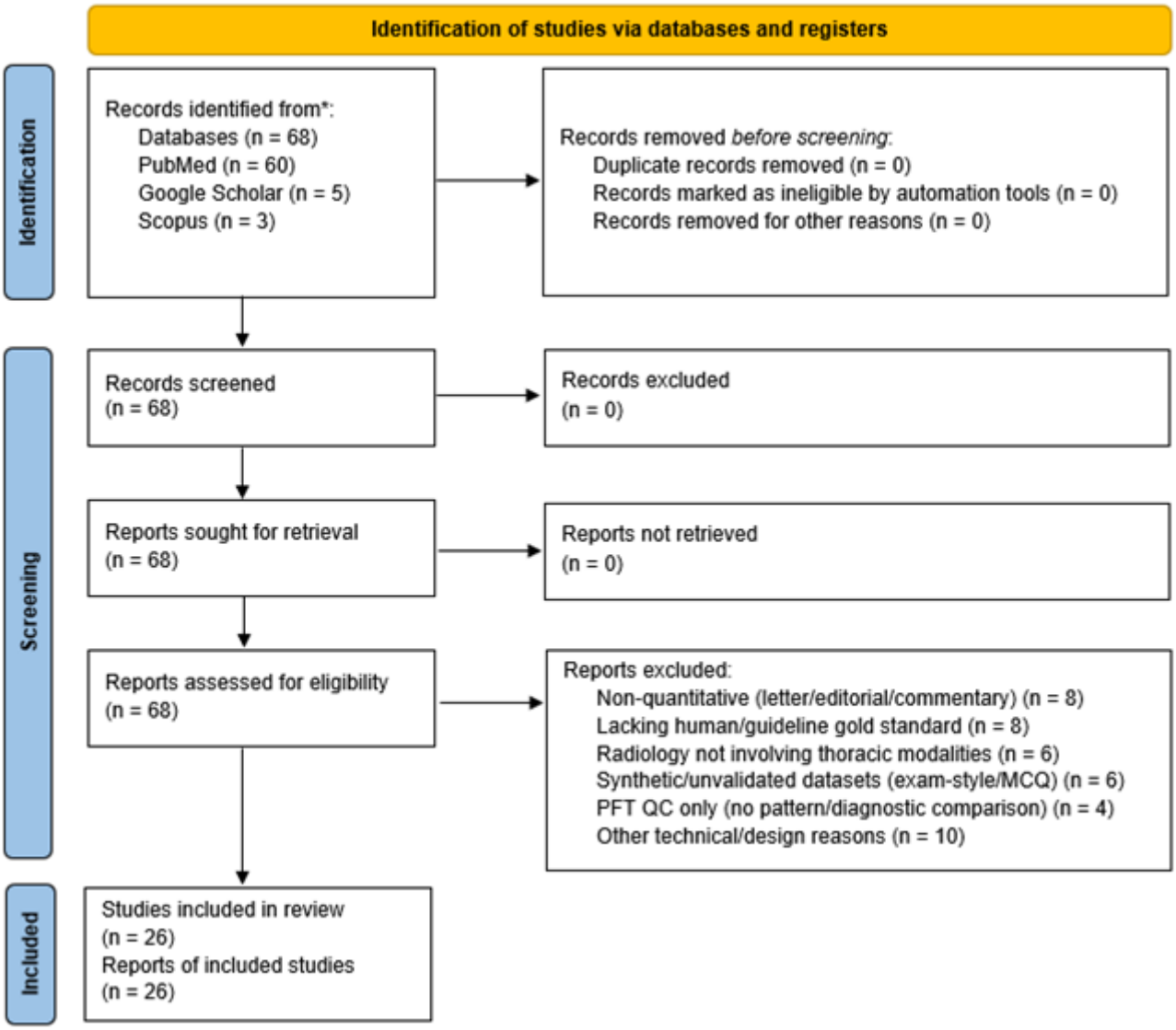

Figure 1