Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A systematic review of clinical translation of large language models using electronic health records

Taiyu Zhu

taiyu.zhu@kcl.ac.uk

King's College London

Mandy Miresmaeili

King's College London

Christopher Tomlinson

King's College London

Linglong Qian

King's College London

James Teo

King's College Hospital NHS Foundation Trust

Zina Ibrahim

King's College London

Richard Dobson

King's College London

Article

Keywords:

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11651581/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. J.T. and R.D. are directors of CogStack Ltd. The other authors declare no competing interests.

A systematic review of clinical translation of large language models using electronic health records

Taiyu Zhu[1,*], Mandy Miresmaeili[1], Christopher Tomlinson[1,2], Linglong Qian[1], James Teo[1,3,4], Zina Ibrahim[1], Richard Dobson[1,2]

[1]Department of Biostatistics & Health Informatics, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, United Kingdom

[2]Institute of Health Informatics, University College London, London, United Kingdom

[3]King's College Hospital NHS Foundation Trust, London, United Kingdom

[4]Guy's and St Thomas' NHS Foundation Trust, London, United Kingdom

[*]Correspondence: taiyu.zhu@kcl.ac.uk

Abstract

We mapped clinical translation of large language models (LLMs) evaluated using electronic health record data. PubMed, Embase, Scopus and IEEE Xplore were searched for peer-reviewed empirical reports published between 1 January 2022 and 1 June 2026. LLM agent-assisted screening and profiling, human validation and targeted full-text review produced a descriptive map of 1,870 reports. Primary classifications came from full texts for 181 reports and titles and abstracts for the remainder. Most evaluations were retrospective or offline (92.3%); 56.2% used a single dataset or setting. Integrated, operational, interventional or outcome-impact evaluation occurred in 4.6%; 3.1% combined these evaluations with validation across multiple datasets or settings. Evidence connecting broader validation with clinical use remains limited. Translation into practice requires transportability, accountable human oversight, prospective evaluation in real-world workflows and clinically meaningful outcomes.

Generative large language models (LLMs) have progressed from question-answering benchmarks to evaluation in clinical settings, including documentation, information extraction, summarisation, patient communication, decision support, and operational tasks. The text-rich content of electronic health records (EHRs), including clinical notes, discharge summaries, radiology and pathology reports, patient portal messages, and administrative records, represents a prominent application domain. Recent studies have extended LLM evaluation from complex medical question answering and differential diagnosis to tasks using real-world clinical text and EHR-derived cases[1–3]. Evaluations in EHR-based settings have also identified limitations in autonomous clinical decision-making and workflow robustness[4].

EHR data, however, are generated within institutions and health systems that differ in documentation conventions, patient populations, coding practices, and data quality, all of which also change over time. Strong performance on a single retrospective extract or public benchmark does not establish transportability to another hospital, specialty, or workflow, nor does it

demonstrate effects on clinician behaviour, documentation quality, safety, or patient outcomes. These concerns are familiar from previous EHR-based modelling[5]; generative LLMs can also produce unsupported statements and omissions in clinical reasoning[4].

Previous reviews have summarised medical and clinical LLM research across broader healthcare settings[6–9]. Du and colleagues catalogued clinical specialties, tasks, models and evaluation metrics in 196 generative LLM–EHR studies identified through 9 November 2024[10]. For EHR-based applications, an unresolved question is how often evaluation across datasets or settings coincides with prospective, user-facing, operational or interventional assessment. Their joint distribution can distinguish growth in technical evaluation from the development of evidence relevant to clinical implementation.

The scale and pace of this literature also make consistent evidence synthesis difficult. Searches can become outdated quickly, while repeated screening and profiling of thousands of records require substantial reviewer time. Recent systematic and scoping reviews have used LLMs to support screening and data extraction[11,12]. We developed CALREP (Clinical Agent for Literature Review and Evidence Profiling) to apply prespecified criteria at scale while retaining human adjudication and validation. The workflow links record-level screening decisions, evidence profiles and review flags to an interactive atlas designed for periodic updating.

We systematically reviewed LLM studies using real-world EHR-derived data published through 1 June 2026 across four databases. We mapped reports across two distinct dimensions: dataset-validation breadth and clinical translation stage (Table 1). Targeted full-text review examined reports initially classified as prospective, user-facing, operational, interventional or unclear, together with two reports without abstracts. Additional profiles captured models, datasets, clinical tasks, human involvement, governance, safety assessment, outcomes and trial evidence. Our primary aim was to establish how often broader validation and clinically advanced evaluation co-occur. The resulting study-linked evidence map also characterises publication growth and identifies priorities for further evaluation.

Results

Search, screening and human validation

The searches retrieved 11,693 records: 4,702 from Scopus, 3,423 from Embase/Ovid, 2,407 from PubMed and 1,161 from IEEE Xplore (Fig. 1). After deduplication and publication-version reconciliation, 7,507 records remained for title-and-abstract screening.

We compared inclusion decisions from ten model configurations with a human consensus reference for 376 screening records (5.01% of those screened). GPT-5.6 Luna at high reasoning had the highest screening Cohen's κ (0.887) and tied for the highest agreement (95.7%). We selected this configuration for screening the deduplicated library (Supplementary Table 1).

Using Luna at high reasoning, 1,909 records (25.4%) met the screening criteria and entered evidence profiling; 5,598 (74.6%) were excluded. Under the hierarchical rule, the primary exclusion reason was C1 (ineligible publication type or status) for 2,634 records (47.1%), C2 (no central evaluation of an eligible generative LLM) for 1,249 (22.3%), C3 (no real-world EHR-

derived data) for 1,704 (30.4%) and C4 (no eligible EHR-relevant task) for 11 (0.2%).

We then compared the ten configurations with a human consensus reference for 96 of the 1,909 initially eligible reports (5.03%) on dataset breadth and clinical translation. Luna at high reasoning had the highest agreement and κ for dataset breadth (90.6%; $\kappa = 0.849$); its agreement on translation stage was 96.9% ($\kappa = 0.830$). We retained it for evidence profiling (Supplementary Table 2).

To characterise reports closest to clinical use more precisely, we sought 223 full texts for review. These comprised all 90 reports classified as prospective, real-time, silent or standalone user-facing (Stage B), all 94 classified as integrated, operational, interventional or outcome-impact (Stage C), all 37 with unclear translation, and two title-only records classified as retrospective or offline (Stage A). Three full texts remained under repository embargo. Human reviewers read all 220 available full texts. Of these, 181 remained eligible and 39 were excluded (22 for no real-world EHR-derived data, 14 for ineligible publication type or status and three for no central evaluation of an eligible generative LLM). Among the 181 eligible reports, dataset-validation labels changed for 83 and translation-stage labels for 64 (Supplementary Fig. 1).

Substituting these primary labels and removing the 39 ineligible reports produced an adjusted profile of 1,870 reports (Fig. 2; Supplementary Table 3). The three reports without an available full text retained their title-and-abstract classifications. The adjusted map therefore comprised 181 full-text and 1,689 title-and-abstract primary classifications. Fifty-eight reports (3.1%) combined broader validation across multiple datasets or settings (Level 2 or 3) with Stage C evaluation, here termed the frontier; all had undergone full-text review. Among the frontier reports, 33 described operational deployment and 20 integration into routine workflow; five were randomised clinical trials, six were single-arm pilot implementations and four were other interventional studies, while 42 stated no trial design and one was unclear. Study-level datasets, evaluation timing, workflow status, trial design, human roles and outcomes are reported in Supplementary Table 4.

Publication growth and evidence maturity

In the adjusted profile, eligible report counts rose from seven in 2022 and 60 in 2023 to 426 in 2024, 882 in 2025 and 495 by 1 June 2026 (Fig. 3; Supplementary Table 5). Overall, 1,803 reports (96.4%) were published from 2024 onwards; the 2026 count already exceeded that for all of 2024. The Level 1 proportion decreased from 58.2% in 2024 to 55.1% in 2025 and was 56.8% in the partial 2026 interval. Over the same period, the proportion classified as Level 2 increased from 16.0% to 29.5%, while the Level 3 proportion increased from 1.9% to 3.6%. The proportion combining Level 2 or 3 validation with Stage B or C evaluation increased from 1.9% in 2024 to 5.3% in both 2025 and 2026. Database coverage was non-exclusive: 1,589 reports (85.0%) appeared in Scopus, 1,207 (64.5%) in PubMed, 889 (47.5%) in Embase/Ovid and 170 (9.1%) in IEEE Xplore.

Datasets and data sources

Dataset breadth was classified as one EHR-derived dataset, cohort, institution or health system (Level 1) in 1,051 reports (56.2%); multiple datasets, cohorts, sites, institutions, health systems or a distinct temporal validation cohort (Level 2) in 450 (24.1%); both public or shareable and local or private EHR data (Level 3) in 53 (2.8%); and unclear in 316 (16.9%). Thus, 503 reports (26.9%) supported evaluation across multiple EHR datasets, cohorts, sites, institutions or data-source types. Broader validation did not necessarily establish independent external validation.

The secondary provenance extraction classified 569 reports (30.4%) as using local or institutional EHR data only, 367 (19.6%) as using public or shareable data only and 53 (2.8%) as combining public and local data; provenance was unclear in 881 (47.1%; Fig. 4). Among the 989 reports with assignable provenance, 569 (57.5%) used only local or institutional data. Frequently named public resources included MIMIC-IV (124 reports), MIMIC-III (95), MIMIC-CXR (79) and i2b2 or n2c2 corpora (25); these categories were non-exclusive (Supplementary Table 6).

Clinical translation and workflow integration

Clinical translation was classified as retrospective or offline (Stage A) in 1,726 reports (92.3%); prospective, real-time, silent or standalone user-facing (Stage B) in 58 (3.1%); and integrated, operational, interventional or outcome-impact (Stage C) in 86 (4.6%). All reports in the adjusted profile had an assigned translation stage. Stage C is heterogeneous and should not be interpreted as a deployment count alone.

Joint analysis quantified the distribution across validation breadth and clinical translation (Fig. 2). In total, 990 reports (52.9%) were both Level 1 and Stage A. Of the 144 Stage B or C reports, 81 (56.2%; 4.3% of all reports) were Level 2 or 3. The frontier comprised 53 Level 2/Stage C reports and five Level 3/Stage C reports; the individual reports and their study characteristics are presented in Supplementary Table 4. Thus, 58 of 86 Stage C reports (67.4%) had broader validation, although these reports represented 3.1% of the complete map (Fig. 5). Conversely, 45 of 53 Level 3 reports (84.9%) remained retrospective or offline, and 27 of 86 Stage C reports (31.4%) remained Level 1. Among the 181 eligible full texts, eight reports (4.4%) were classified as randomised clinical trials, 15 (8.3%) as single-arm pilot implementations and eight (4.4%) as other interventional designs; 148 (81.8%) had no trial or interventional design stated, and two (1.1%) were unclear (Fig. 5; Supplementary Table 7). Trial designation and translation stage were not interchangeable because translation stage also incorporated prospective, workflow-integrated, operational and outcome-impact evaluation.

Outcomes, human involvement, governance and safety

The title-and-abstract descriptive fields of the 1,870 retained reports indicated that evaluation remained centred on technical performance. Model accuracy or task performance was reported in 1,747 reports (93.4%), whereas workflow or time outcomes appeared in 163 (8.7%), clinician experience or burden in 52 (2.8%), patient outcomes or experience in 46 (2.5%) and clinician acceptance, editing or override rates in 21 (1.1%; Fig. 5; Supplementary Table 8).

Human involvement included clinician adjudication or error review in 297 reports (15.9%), clinician-created reference labels in 292 (15.6%), direct clinician comparison in 184 (9.8%) and a human-in-the-loop workflow in 80 (4.3%). The abstracts identified hallucination or fabrication assessment in 122 reports (6.5%), omission assessment in 73 (3.9%) and a harm or severity taxonomy in 27 (1.4%). Governance and safety information was not reported in 1,814 (97.0%) and 1,529 (81.8%) title-and-abstract profiles, respectively.

Model families, use modes and EHR tasks

Model characteristics are summarised in Supplementary Table 9. A closed or proprietary model was the only reported access type in 485 reports (25.9%), an open-weight model in 305 (16.3%) and both types in 152 (8.1%); access was unclear or outside these categories in 928 (49.6%). After post hoc normalisation, GPT-4 was the most frequent named model category (355 reports; 19.0%), followed by Llama (344; 18.4%), GPT-4o (209; 11.2%), GPT-3.5 (138; 7.4%) and version-unspecified GPT or ChatGPT (108; 5.8%). GPT-2 or GPT-Neo and GPT-OSS were retained as separate categories.

Prompting was reported in 940 reports (50.3%), information extraction in 797 (42.6%), fine-tuning in 382 (20.4%), embeddings or classification in 313 (16.7%), summarisation in 263 (14.1%), retrieval-augmented generation in 194 (10.4%), multimodal language modelling in 157 (8.4%) and agentic workflows in 70 (3.7%). A report could receive more than one label. In broad, non-exclusive task groupings, 727 reports (38.9%) addressed information extraction or structuring; 439 (23.5%) diagnostic reasoning, treatment planning or decision support; 418 (22.4%) summarisation, documentation or report generation; and 396 (21.2%) clinical classification, coding or staging (Supplementary Table 10).

Preprint sensitivity and interactive atlas

Re-screening 281 database-indexed preprint records while waiving publication status alone identified 164 additional eligible reports. For a like-for-like title-and-abstract comparison, adding these reports to the 1,909-report primary profile produced 2,073 reports. The frontier changed from 25 of 1,909 (1.3%) to 29 of 2,073 (1.4%), and the proportion reaching Stage B or C remained 9.6%. Adding the same preprints to the adjusted map yielded 62 frontier reports among 2,034 (3.0%), compared with 58 of 1,870 (3.1%; Supplementary Table 11). The preprints were profiled from titles and abstracts.

Across its peer-reviewed and optional preprint layers, the interactive atlas links 2,034 reports to 1,175 entities through 31,544 report–entity connections. Its aggregate pathway view shows how reported dataset provenance connects to validation breadth and clinical translation (Supplementary Fig. 2), while the network view contains 113 characteristics and 494 co-reporting relationships. The atlas retains report-level links to support inspection of the underlying evidence (`https://max-llm.github.io/CALREP_atlas/`). The workflow is designed for repeat updates: refreshed database exports can pass through the same deduplication, screening, profiling and atlas-construction steps, while ambiguous record links and clinically advanced reports remain routed to human review.

Discussion

This review identified rapid growth in LLM–EHR research alongside a much smaller body of clinically advanced evidence. Of the retained reports, 96.4% were published from 2024 onwards; by the June 2026 cutoff, the number identified for 2026 already exceeded that for all of 2024. After targeted full-text adjustment, Stage A reports outnumbered Stage C reports by approximately 20 to one, and 990 reports were both Level 1 and Stage A. Although 58 of 86 Stage C reports (67.4%) had broader validation, these reports represented only 3.1% of the adjusted map. The principal evidence gap is therefore the scarcity of studies combining broader validation with evaluation in clinical use. The preprint sensitivity analysis did not materially alter this interpretation. The evidence gap was therefore robust to inclusion of preprint records captured by the database searches.

Several indicators of evidence maturity increased over time. The Level 1 proportion was 58.2% in 2024, 55.1% in 2025 and 56.8% in the partial 2026 interval. Its higher value in 2022 (71.4%) was based on only seven reports, and the yearly pattern was not monotonic. From 2024 to the 1 June 2026 cutoff, the proportion classified as Level 2 increased from 16.0% to 29.5%, while the Level 3 proportion increased from 1.9% to 3.6%. The proportion combining Level 2 or 3 validation with Stage B or C evaluation rose from 1.9% in 2024 to 5.3% in 2025 and was also 5.3% in the partial 2026 interval. Together, these shifts indicate a gradual redistribution towards broader and more clinically situated evaluation, although most reports remained Level 1 or retrospective and offline. Across the complete map, 27 of 86 Stage C reports were Level 1, whereas 45 of 53 Level 3 reports remained Stage A.

The distinction matters because EHR documentation, coding behaviour, patient populations, language and workflow vary across settings. Retrospective evaluation is essential for development, but cannot establish prospective reliability, effects on clinician behaviour, or benefits and harms in care. Conversely, a locally integrated system can reveal workflow effects without establishing performance across settings. The dimensions expose two complementary questions: whether an LLM is evaluated across settings, and whether it is evaluated in prospective or operational use.

The Stage C corpus included several distinct evaluation designs. EyeFM combined multicountry validation with a randomised clinical evaluation of an ophthalmic copilot; PsychFound paired multicentre EHR development with a prospective two-arm study in psychiatric practice; and an EHR-embedded decision-support system in Kenya was evaluated across 16 clinics with explicit analysis of harmful and beneficial changes to clinician documentation[13–15]. A cardiology trial evaluated LLM-assisted clinicians using real patient records, but retained retrospective cases and a standalone study workflow[16]. These studies address different questions about transportability, clinician–model interaction, operational safety and patient consequences. Subsequent reports published after the publication cutoff and therefore excluded from the review counts further illustrate this distinction. An autonomous agent navigated a sandboxed EHR in simulations based on real patient cases, whereas a prospective emergency-department evaluation found declining adoption over four weeks and no difference in length of stay[17,18]. A pragmatic cluster-randomised trial of the Kenyan system found no significant reduction in 14-day treat-

ment failure despite no identified safety signal[19]. Together, these results reinforce the need to distinguish technical capability, successful deployment and demonstrated patient benefit.

The provenance analysis further characterises limitations in transportability. Among reports with assignable provenance, 57.5% used only local or institutional data. Such data are clinically valuable and can represent local workflows, but without evaluation on independently sourced data they provide limited evidence that performance will transfer across documentation systems, populations, languages and institutions. Frequently reused public resources can themselves concentrate evidence within a small number of health systems. The field therefore requires deliberate combinations of local clinical relevance, public reproducibility and independent external validation.

Task performance dominated the descriptive evidence map, while reported workflow, clinician and patient outcomes were less frequent. The governance and safety findings concern what was recoverable from abstracts, where operational details are often omitted. Broader benchmark initiatives such as MedHELM cover clinical decision support, note generation, patient communication, medical research and administrative tasks, but this breadth remains distinct from prospective workflow evaluation[20]. Adjacent randomised evidence outside the EHR scope found that strong standalone model performance did not improve diagnostic or disposition decisions when members of the public used the same systems, illustrating why human–model interaction must be evaluated directly[21]. Progress towards clinical implementation will require evaluated pathways that connect external validity and accountable human use with operational reliability, predefined responses to failure and clinical outcomes.

Model choice also remains concentrated. GPT-4 and Llama were the most frequently reported model categories; model-access status was unclear or outside the structured categories in nearly half of reports. Prompting and information extraction were reported more often than retrieval-augmented generation or agentic workflows. As models and access routes change rapidly, reporting of model version, access, adaptation and hosting is necessary for reproducibility and for distinguishing model capability from implementation design.

Model access also has consequences for clinical implementation. Proprietary application programming interfaces (APIs) can provide capable models without the hardware and maintenance demands of local hosting, but use with identifiable EHR data can depend on approved contractual arrangements, data residency, private connectivity, audit controls and continued vendor availability. Changes in model behaviour, pricing, policy or regional service availability can constrain reproducibility and portability between health systems. Reported secure implementations have required institutional infrastructure and governance[22]. Locally deployed open-weight models can retain data within institutional boundaries and permit version control, adaptation and local monitoring; however, they transfer the burden of cybersecurity, computing, validation, maintenance and clinical assurance to the health system[23]. A post-cutoff study paired fully on-premise deployment with decision-time reliability signals, although evaluation remained retrospective and benchmark-based rather than operational clinical use[24]. Future evaluations should report deployment location, data flows and retention, model versioning, contractual or licensing constraints, and the resources required for sustained operation.

The two-dimensional framework deliberately avoids collapsing heterogeneous evidence into a single hierarchy. A multi-dataset retrospective benchmark and a locally integrated intervention answer different questions; neither should automatically outrank the other. The adjusted profile contained 503 Level 2 or 3 reports and 144 Stage B or C reports, providing focused corpora for deeper methodological appraisal. Their intersection comprised 81 reports; 58 combined broader validation specifically with Stage C evaluation.

The evidence map links the joint distribution of validation breadth and clinical translation to the reports, models, datasets, outcomes and safeguards from which it was derived. Targeted full-text review showed that fuller descriptions of datasets and workflows changed which reports met the frontier definition. Title-and-abstract classification provides an initial evidence layer at scale, an approach also used in an LLM-assisted systematic review of clinical medicine[11]. CALREP retains the decisions, rationales and human consensus references supporting this layer. Refreshed exports can pass through the same workflow, with ambiguous or consequential decisions directed to human review. Linking these decisions to the atlas supports targeted appraisal and repeat updates as the literature grows.

These findings inform research funding, procurement and health-system governance. Evaluations should address validation breadth and position in the clinical pathway, with prospectively specified measures of safety, workload, patient experience, equity and clinical outcomes. A staged evaluation pathway could progress from external retrospective validation to silent prospective evaluation, controlled workflow integration and monitoring after changes to models, data or workflows. Silent trials can test intended-use conditions; governance should reflect local infrastructure; dynamic red-teaming can reveal failures missed by static benchmarks; and patient-centred evaluation should address trust, acceptability and safety[25–28]. Procurement arrangements should specify model versions, material changes, audit access, local monitoring and data portability. Public funding can support multi-institutional evaluation in under-represented health systems and publication of negative or null implementation findings.

To make this evidence interpretable, future reports should state the number and type of datasets, institutions, health systems and temporal validation samples, distinguish public from local data, and identify model version, access, adaptation, data flows and monitoring arrangements. TRIPOD-LLM and CONSORT-AI provide relevant reporting frameworks[29,30].

This analysis has several limitations. Screening and initial profiling used titles, abstracts and metadata, consistent with a previous large LLM-assisted review[11]. Human consensus references covered 376 screening records and 96 profiles, and only the two primary dimensions underwent profile-agreement analysis. These comparisons informed model selection and should not be interpreted as independent external validation of the workflow. The agreement analyses address title-and-abstract screening and profiling; descriptive fields were not independently double-extracted. We did not pool effect estimates or undertake formal study-level risk-of-bias, reporting-bias or certainty-of-evidence assessments; the map characterises evaluation design and reporting rather than clinical efficacy. Finally, English-language restrictions and exclusion of synthetic-only studies, medical-knowledge benchmarks and abstract-only reports limit scope. The preprint analysis covered database-indexed records rather than a dedicated search of preprint servers.

The LLM–EHR evidence base has expanded rapidly, but studies combining broader validation with evaluation in clinical use remain scarce. Addressing this gap requires studies that establish transportability across health systems, evaluate models under accountable human oversight and measure outcomes that are meaningful to clinicians and patients.

Methods

Search strategy and eligibility criteria

Reporting of this review was guided by the PRISMA 2020 checklist and its abstract checklist, which accompany the manuscript. We searched PubMed, Embase via Ovid, Scopus and IEEE Xplore on 2 October 2026 for English-language records published between 1 January 2022 and 1 June 2026. The search combined three concept blocks covering generative LLMs, EHR-derived data sources, and evaluation or clinical task terms. Synonyms within blocks were joined with OR and the blocks with AND. Model terms covered generative and foundation models, instruction tuning, named model families and retrieval-augmented generation. EHR terms covered records, notes, reports, portal messages and named datasets; task terms covered development, validation, extraction, generation, prediction, documentation, decision support and workflow use. The bare acronym RAG was excluded after pilot searches showed insufficient specificity. Verbatim executable queries are provided in Supplementary Note 1.

PubMed combined Title/Abstract terms with the Electronic Health Records MeSH term and applied date and English-language filters. Review and meta-analysis publication types and records with survey in the title were excluded. Embase matched title and abstract fields and applied the available year and language limits. Scopus used TITLE-ABS-KEY with publication-year and English-language restrictions. IEEE Xplore used All Metadata; non-English records were removed during screening where a language filter was unavailable. Records with dates after 1 June 2026 were removed during export or harmonisation.

Eligible reports centrally evaluated a generative LLM, used real-world EHR-derived data and addressed an EHR-relevant clinical, research, operational or workflow task. Peer-reviewed full-length journal articles, empirical letters, implementation or quality-improvement reports with original results, and archival conference papers were eligible. EHR-derived data included structured variables, clinical notes and reports, patient-portal messages, routine-care documentation, and named EHR or health-system datasets. Use during training, fine-tuning, retrieval, development, validation or testing could satisfy the data criterion. Studies combining synthetic and real EHR data were eligible; synthetic-only or vignette-only studies were not. We excluded preprints from the primary analysis, abstract-only publications, reviews, editorials, protocols without results, non-EHR tasks and records limited to encoder-only or conventional machine learning models. The review was registered on the Open Science Framework (`https://doi.org/10.17605/OSF.IO/TCM8Q`). The protocol is available in the registered OSF record. No amendments were made to the registered methods.

Record management and publication reconciliation

Database exports were normalised to a common schema. Automatic deduplication used DOI, PubMed identifier and exact normalised titles, with checks for conflicting identifiers. Near-duplicate title matches at similarity of at least 0.81 and metadata-conflict groups were exported for human adjudication. Of 26 candidate groups, 23 were confirmed as duplicates and three were retained separately (Supplementary Fig. 3).

Preprint-indexed records were reconciled with peer-reviewed versions published by the cutoff using bibliographic identifiers, title matching, publication dates and manual review of ambiguous links. The manual review covered 32 links: 31 were accepted and one rejected; 30 accepted links were applied and one was already resolved. The final library contained 7,507 records after removal of 4,186 duplicate or version records; publication-status reconciliation was recorded for 50 retained records. Stable identifiers and source-database links were retained throughout.

Agent-assisted screening and profiling with human validation

We implemented CALREP in Python using LangGraph for resumable orchestration. A deterministic deduplication stage was followed by title-and-abstract screening and structured evidence profiling. Human decisions resolved ambiguous duplicate and publication-version links. Each model assessment was restricted to supplied record information and did not browse independently. Prompts, model configuration, record identifiers and outputs were retained as part of the review record.

Two reviewers with complementary clinical and LLM-methods expertise contributed independent annotations and adjudication. Independent annotations were collected in separate Label Studio projects using identical task files. The reviewers remained blinded to one another's labels until both annotation exports were complete. The screening reference comprised 376 randomly selected records (5.01% of the deduplicated library), with 92 included and 284 excluded by the final human reference. Reference decisions comprised 330 independent agreements and 46 resolved through re-review and adjudication. Individual reviewer labels were available for 375 records; comparisons used the consensus reference. Neither the screening nor the profiling reference records were used to refine the prompts (Supplementary Note 2).

Five candidate models were each assessed at low and high reasoning settings, giving ten configurations. The API identifiers were `gpt-5.6-luna`, `gpt-5.6-sol`, `gpt-5.6-terra`, `claude-sonnet-5` and `claude-opus-5`. The accompanying code was tested with Python 3.11.15, LangGraph 1.2.5, OpenAI Python SDK 2.38.0 and Anthropic Python SDK 0.104.1. We compared their inclusion decisions with the 376-record human reference using the same screening prompt. Performance measures were agreement, Cohen's κ, sensitivity and specificity (Supplementary Table 1). GPT-5.6 Luna at high reasoning was selected for production screening on the basis of its screening agreement and κ.

The screening instructions advanced records when all four criteria were directly supported or reasonably indicated, excluded records when a criterion clearly failed, and retained genuinely uncertain records for review. The ordered criteria were eligible publication type and status (C1),

central evaluation of an eligible generative LLM (C2), real-world EHR-derived data (C3), and an eligible EHR-relevant task (C4). Exclusions used the first applicable failed criterion in this hierarchy. Screening the 7,507 deduplicated records yielded 1,909 reports for evidence profiling.

From these 1,909 reports, 96 (5.03%) were randomly selected for the profiling reference; 22 required adjudication of one or both primary dimensions. We compared the same ten configurations with this human reference using the same profiling prompt, measuring agreement and κ for dataset-validation breadth and clinical translation stage (Supplementary Table 2). The comparison supported retaining Luna at high reasoning for production profiling.

During production profiling, each eligible report received two primary labels (Table 1). Dataset validation was categorised as one EHR-derived dataset, cohort, institution or health system (Level 1); multiple datasets, cohorts, sites, institutions, health systems or a distinct temporal validation cohort (Level 2); explicit evaluation on both public or shareable and local or private EHR data (Level 3); or unclear. Clinical translation was categorised as retrospective or offline (Stage A); prospective, real-time, silent or standalone user-facing evaluation (Stage B); integrated, operational, interventional or outcome-impact evaluation (Stage C); or unclear. Clinician labels or ratings of retrospectively generated outputs alone did not raise Stage A to Stage B. Stage C did not require technical integration into the EHR interface when actual routine-workflow use or an intervention was evaluated. These dimensions draw on the clinical AI evaluation continuum reflected in TRIPOD-LLM, DECIDE-AI and CONSORT-AI[29–31].

Additional fields captured models, model access, LLM use modes, datasets, clinical specialty and task, synthetic-data use, human involvement, governance, safety and outcomes. The secondary provenance extraction used titles, abstracts, named datasets and the existing profile rationale. It classified data as public/shareable only, local/institutional only, mixed public and local, or unclear. Recognised public datasets remained public when their originating hospital was named; unnamed clinical data were not assumed to be local.

Ninety-five per cent confidence intervals (CIs) for agreement and κ used 20,000 percentile bootstrap resamples of records; diagnostic proportions used Wilson intervals. No hypothesis tests were performed. The profile comparison assessed the two primary dimensions.

Targeted full-text review and adjusted evidence map

After title-and-abstract profiling, we sought full texts for 223 reports: 221 assigned Stage B, Stage C or unclear translation after title-and-abstract profiling, and two Stage A records without abstracts. Three full texts were unavailable because of repository embargoes. Human reviewers read all 220 available full texts. Full-text classifications followed the same prespecified criteria for dataset breadth and clinical translation used in the agent's profiling prompt (Supplementary Note 5).

For each report, the review recorded page-referenced evidence, dataset or site descriptions, evaluation timing, workflow status, trial design, human roles and outcomes. We replaced the two primary labels for 181 eligible reports and removed 39 ineligible reports. All other reports, including the three targeted full texts unavailable owing to repository embargo, retained their title-and-abstract labels.

The full-text review did not re-extract the complete descriptive schema. Model, task, human-involvement, governance, safety and outcome summaries therefore use the title-and-abstract fields for the 1,870 retained reports. Full-text trial categories and concise study characteristics of the frontier reports are reported separately. The frontier was defined as Level 2 or 3 combined with Stage C (Supplementary Note 3).

Preprint sensitivity analysis

We independently re-screened all 281 database-indexed preprint candidates after waiving publication status alone and retaining the other eligibility criteria. Publication-version reconciliation preceded this analysis; the 164 eligible preprints did not overlap the primary profile. All were profiled from titles and abstracts using GPT-5.6 Luna at high reasoning. We compared the original 1,909-report title-and-abstract profile with a combined 2,073-report profile, and separately added the preprint layer to the adjusted map (2,034 reports). No additional full-text review of preprints was included (Supplementary Note 4).

Data synthesis and atlas construction

We summarised classifications descriptively, with reports as the unit of analysis. Principal adjusted-profile counts used 1,870 reports; initial title-and-abstract counts used 1,909; full-text characteristics used the 181 eligible reviewed reports. Source-database coverage and most descriptive fields were non-exclusive. Outcome categories were non-exclusive, and a report could contribute to more than one category. Model names were harmonised with fixed string rules. Free-text clinical tasks were grouped post hoc with fixed case-insensitive keyword rules; original descriptions were retained

The synthesis mapped evaluation design and reporting descriptively. Primary dimensions were cross-tabulated and yearly proportions calculated to distinguish publication growth from evidence maturity. The 2026 interval ended on 1 June and was not treated as a complete year. A study–entity graph and searchable atlas were constructed from the final profiles. The linked Markdown structure was adapted from Karpathy's LLM Wiki pattern[32]; the evidence schema, classification rules and human validation procedures were developed for this review. Edges represent study-level co-reporting and do not establish causal relationships. The prompts are reproduced verbatim as supporting resources in Supplementary Note 5.

Data availability

The final peer-reviewed evidence profiles, separately labelled preprint profiles and study-linked knowledge graph are available at `https://github.com/MAX-LLM/CALREP_atlas`. The interactive atlas is available at `https://max-llm.github.io/CALREP_atlas/`.

Code availability

The screening, profiling, deduplication and validation code, together with the prompts used for the review, are available at `https://github.com/MAX-LLM/CALREP`.

Acknowledgments

This work was supported by the School of Mental Health and Psychological Sciences Early Career Researchers Grant at King's College London. The funder had no role in study design, data collection, analysis, interpretation, manuscript preparation or the decision to submit for publication.

Author contributions

T.Z. conceived and designed the study, developed the review workflow and software, analysed the data, prepared the figures and drafted the manuscript. T.Z. and M.M. manually reviewed and annotated the studies and evaluated the screening and profiling models. C.T. and J.T. provided clinical input and contributed to interpretation of the findings. L.Q. contributed to prompt development, and Z.I. led prompt refinement. R.D. supervised the study and provided critical feedback on the manuscript. All authors reviewed and approved the final manuscript.

Competing interests

J.T. and R.D. are directors of CogStack Ltd. The other authors declare no competing interests.

References

[1] Singhal, K. et al. Toward expert-level medical question answering with large language models. *Nat. Med.*, 31:943–950, 2025. `https://doi.org/10.1038/s41591-024-03423-7`.

[2] McDuff, D. et al. Towards accurate differential diagnosis with large language models. *Nature*, 642:451–457, 2025. `https://doi.org/10.1038/s41586-025-08869-4`.

[3] Wu, J. et al. BRIDGE: benchmarking large language models for understanding real-world clinical practice texts. *Nat. Biomed. Eng.*, 2026. `https://doi.org/10.1038/s41551-026-01719-2`; Published online 17 June 2026.

[4] Hager, P. et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Med.*, 30:2613–2622, 2024. `https://doi.org/10.1038/s41591-024-03097-1`.

[5] Rajkomar, A. et al. Scalable and accurate deep learning with electronic health records. *npj Digit. Med.*, 1:18, 2018. `https://doi.org/10.1038/s41746-018-0029-1`.

[6] Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.*, 29:1930–1940, 2023. `https://doi.org/10.1038/s41591-023-02448-8`.

[7] Preiksaitis, C. et al. The role of large language models in transforming emergency medicine: Scoping review. *JMIR Med. Inform.*, 12:e53787, 2024. https://doi.org/10.2196/53787.

[8] Chen, D., Alnassar, S. A., Avison, K. E., Huang, R. S. & Raman, S. Large language model applications for health information extraction in oncology: Scoping review. *JMIR Cancer*, 11:e65984, 2025. https://doi.org/10.2196/65984.

[9] Bedi, S. et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA*, 333:319–328, 2025. https://doi.org/10.1001/jama.2024.21700.

[10] Du, X. et al. Testing and evaluation of generative large language models in electronic health record applications: a systematic review. *J. Am. Med. Inform. Assoc.*, 33:743–753, 2026. https://doi.org/10.1093/jamia/ocaf233.

[11] Chen, S. F. et al. LLM-assisted systematic review of large language models in clinical medicine. *Nat. Med.*, 32:1152–1159, 2026. https://doi.org/10.1038/s41591-026-04229-5.

[12] Sounack, T. et al. AI-assisted scoping review of code sharing in clinical prediction model research. *Nat. Med.*, 2026. doi: 10.1038/s41591-026-04691-1.

[13] Wu, Y. et al. An eyecare foundation model for clinical assistance: a randomized controlled trial. *Nat. Med.*, 31:3404–3413, 2025. https://doi.org/10.1038/s41591-025-03900-7.

[14] Wang, R. et al. A domain-adapted large language model to support clinicians in psychiatric clinical practice. *Nat. Mach. Intell.*, 8:690–707, 2026. https://doi.org/10.1038/s42256-026-01224-w.

[15] Agweyu, A. et al. Safety of a large language model-based clinical decision support system in African primary healthcare. *Nat. Health*, 1:607–618, 2026. https://doi.org/10.1038/s44360-026-00082-5.

[16] O'Sullivan, J. W. et al. A large language model for complex cardiology care. *Nat. Med.*, 32:616–623, 2026. https://doi.org/10.1038/s41591-025-04190-9.

[17] Ferber, D. et al. Towards autonomous medical artificial intelligence agents. *Nature*, 655: 1282–1291, 2026. https://doi.org/10.1038/s41586-026-10675-5.

[18] Leibovitch, L. et al. Prospective evaluation of a large language model clinical decision support system in the emergency department. *Nat. Med.*, 32:3741–3747, 2026. https://doi.org/10.1038/s41591-026-04601-5.

[19] Agweyu, A. et al. Generative AI-enabled clinical decision support system in primary care: a pragmatic, cluster-randomized trial. *Nat. Med.*, 32:3032–3039, 2026. https://doi.org/10.1038/s41591-026-04503-6.

[20] Bedi, S. et al. Holistic evaluation of large language models for medical tasks with MedHELM. *Nat. Med.*, 32:943–951, 2026. https://doi.org/10.1038/s41591-025-04151-2.

[21] Bean, A. M. et al. Reliability of LLMs as medical assistants for the general public: a randomized preregistered study. *Nat. Med.*, 32:609–615, 2026. https://doi.org/10.1038/s41591-025-04074-y.

[22] Ng, M. Y., Helzer, J., Pfeffer, M. A., Seto, T. & Hernandez-Boussard, T. Development of secure infrastructure for advancing generative artificial intelligence research in healthcare at an academic medical center. *J. Am. Med. Inform. Assoc.*, 32:586–588, 2025. https://doi.org/10.1093/jamia/ocaf005.

[23] Dennstädt, F., Hastings, J., Putora, P. M., Schmerder, M. & Cihoric, N. Implementing large language models in healthcare while balancing control, collaboration, costs and security. *npj Digit. Med.*, 8:143, 2025. https://doi.org/10.1038/s41746-025-01476-7.

[24] Zhang, L. et al. On-premise medical AI agents for reliable clinical decision-making. *Nat. Med.*, 32:3705–3716, 2026. https://doi.org/10.1038/s41591-026-04609-x.

[25] Tikhomirov, L. et al. A scoping review of silent trials for medical artificial intelligence. *Nat. Health*, 1:532–554, 2026. https://doi.org/10.1038/s44360-025-00048-z.

[26] Ong, J. C. L. et al. Large language models in global health. *Nat. Health*, 1:35–47, 2026. https://doi.org/10.1038/s44360-025-00024-7.

[27] Pan, J. et al. Addressing benchmarking gaps in large language models for health and medicine with dynamic red-teaming. *Nat. Health*, 2026. https://doi.org/10.1038/s44360-026-00152-8.

[28] Lai, S. et al. Patient factors in medical artificial intelligence: a systematic review. *Nat. Health*, 2026. https://doi.org/10.1038/s44360-026-00190-2.

[29] Gallifant, J. et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat. Med.*, 31:60–69, 2025. https://doi.org/10.1038/s41591-024-03425-5.

[30] Liu, X. et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat. Med.*, 26:1364–1374, 2020. https://doi.org/10.1038/s41591-020-1034-x.

[31] Vasey, B. et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat. Med.*, 28:924–933, 2022. https://doi.org/10.1038/s41591-022-01772-9.

[32] Andrej Karpathy. LLM Wiki: A pattern for building personal knowledge bases using LLMs. GitHub Gist, 2026. https://gist.github.com/karpathy/442a6bf555914893e9891c11519de94f (accessed 3 October 2026).

Tables

Table 1: Evidence-profile dimensions used to classify eligible reports.

Dimension	Categories
Dataset validation level	Level 1: one EHR dataset, cohort, institution or health system; Level 2: multiple datasets, cohorts, sites, institutions, health systems or a distinct temporal validation cohort; Level 3: both public/shareable and local/private EHR data; unclear
Clinical translation stage	Stage A: retrospective/offline; Stage B: prospective, silent, real-time or standalone user-facing; Stage C: integrated, operational, interventional or outcome-impact; unclear
Model characteristics	Model and system names; access profile; prompting, retrieval augmentation, fine-tuning, extraction, summarisation, classification, agentic, multimodal or other use mode
Dataset and clinical characteristics	Named dataset or setting; synthetic-data use; clinical specialty; clinical task
Human and translational evidence	Trial evidence; human comparison or review; oversight and governance; safety evaluation; outcomes measured
Classification rationale	Brief evidence-based rationale for the two primary labels

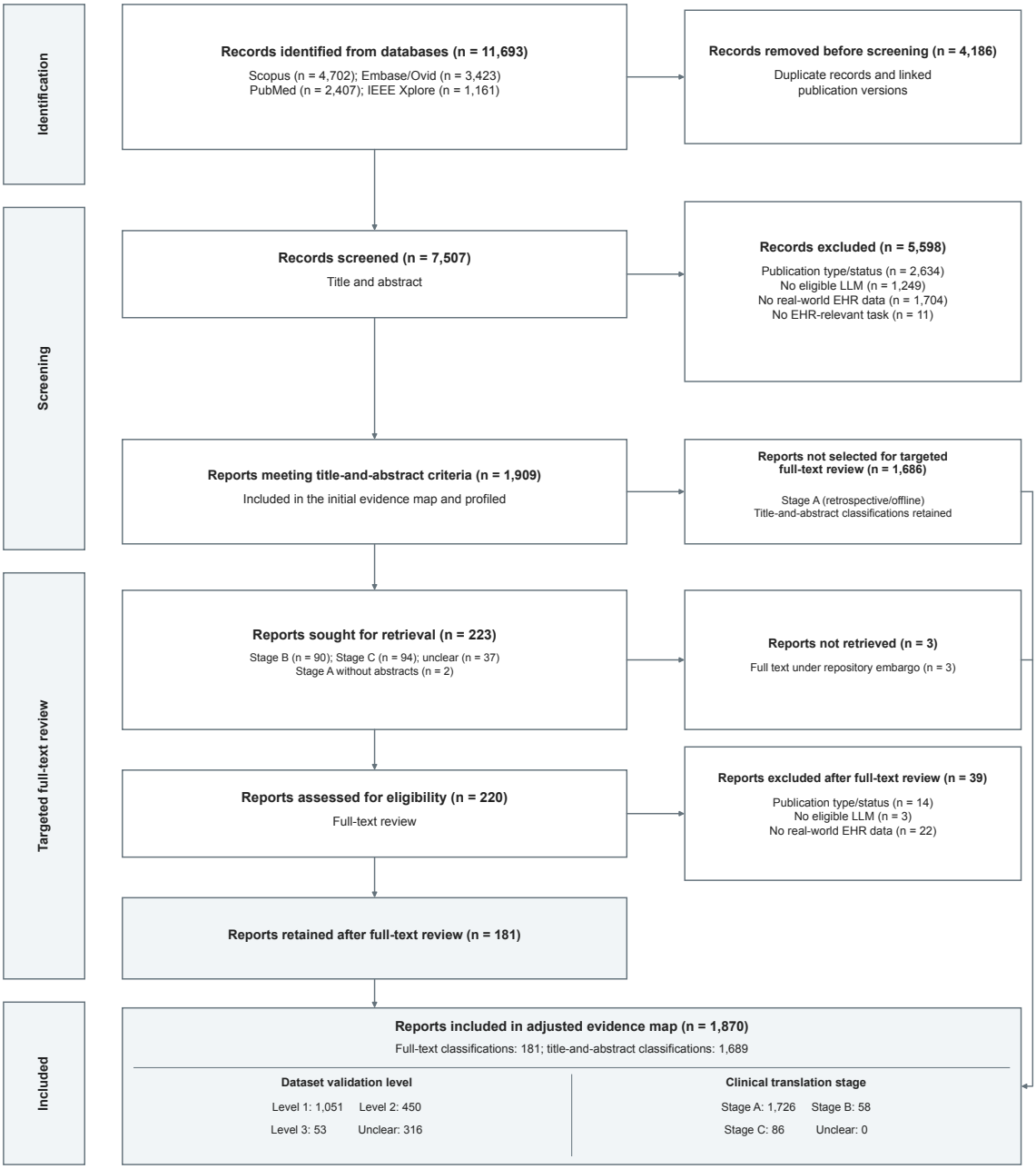


Fig. 1: PRISMA 2020 flow diagram adapted for targeted full-text review. The four database searches yielded 11,693 records. After deduplication and publication-version reconciliation, 7,507 records were screened and 1,909 met the title-and-abstract criteria. Of these, 223 were selected for targeted full-text review: 221 reports classified as prospective, real-time, silent or standalone user-facing (Stage B), integrated, operational, interventional or outcome-impact (Stage C), or unclear, and two records without abstracts classified as retrospective or offline (Stage A). Full-text review showed that one was an invited commentary and the other a letter to the editor; both were excluded as non-empirical publication types. The remaining 1,686 Stage A reports retained their title-and-abstract classifications. Three reports were not retrieved because their full texts remained under repository embargo. Thirty-nine of the 220 assessed reports were excluded. The adjusted evidence map therefore included 1,870 reports: 181 with full-text classifications and 1,689 with title-and-abstract classifications.

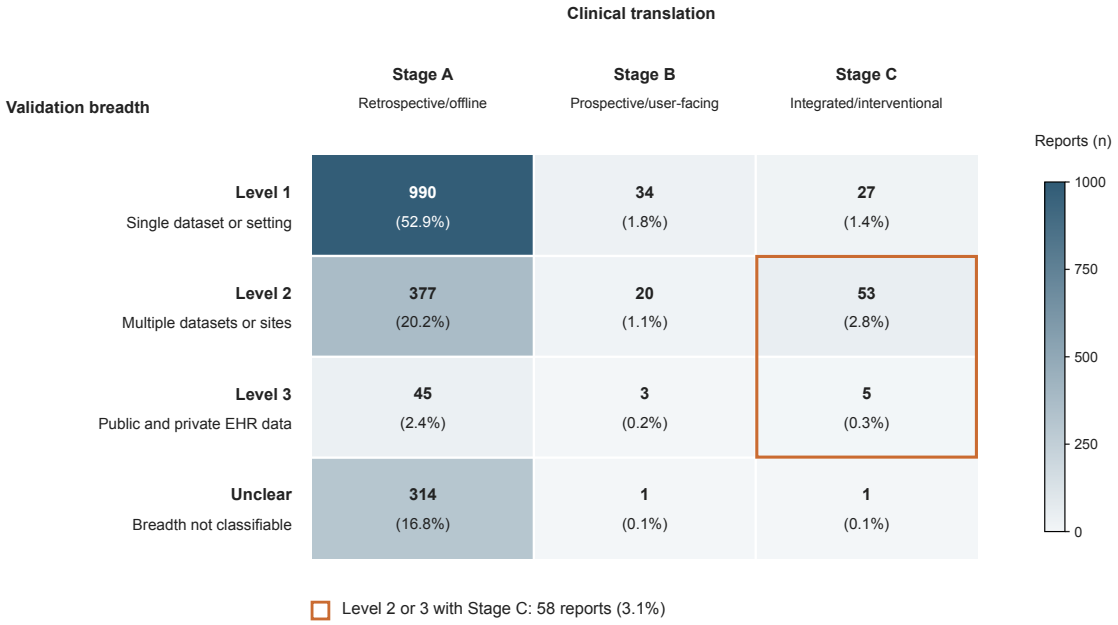


Fig. 2: Joint distribution of validation breadth and clinical translation in the adjusted evidence map. Counts and percentages use 1,870 reports as the denominator. The orange outline identifies the 58 reports (3.1%) combining broader validation across multiple datasets, sites or temporal cohorts, or public and local data (Level 2 or 3), with integrated, operational, interventional or outcome-impact evaluation (Stage C); all had full-text primary classifications. Colour intensity represents report counts on a linear scale. Broader validation does not necessarily establish independent external validation.

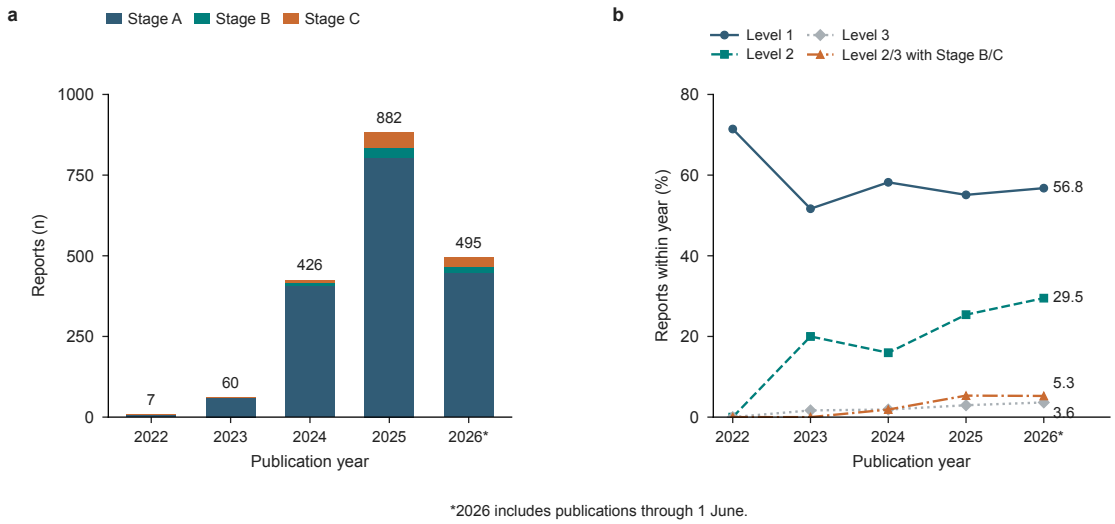


Fig. 3: Publication growth and evidence maturity. **a**, Reports in the adjusted map by publication year and translation stage: retrospective or offline (Stage A); prospective, real-time, silent or standalone user-facing (Stage B); and integrated, operational, interventional or outcome-impact (Stage C). **b**, Within-year proportions at validation Levels 1, 2 and 3, representing one dataset or setting, multiple datasets or settings, and mixed public and local data, respectively; the fourth series combines Level 2 or 3 validation with Stage B or C evaluation. Yearly denominators are 7, 60, 426, 882 and 495 reports for 2022–2026, respectively. The 2026 interval ends on 1 June. Lines connect descriptive yearly proportions; no temporal hypothesis test was performed.

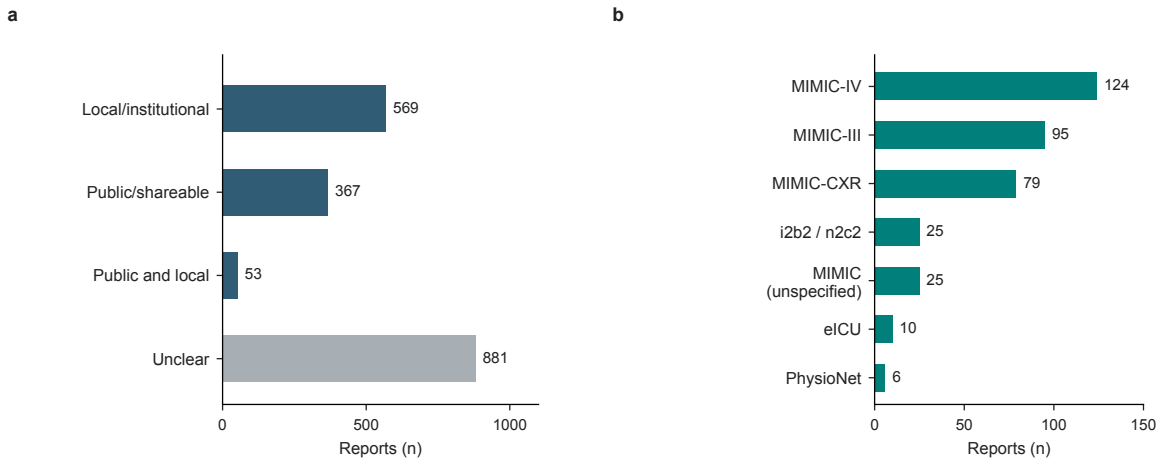


Fig. 4: EHR dataset provenance and named public resources. **a**, Secondary provenance classifications among 1,870 retained reports. **b**, Named public EHR resources extracted with fixed string patterns; categories are non-exclusive. Provenance uses titles, abstracts, named datasets and profile rationales. Named resources derive from title-and-abstract profiling. Grey denotes unclear provenance, with insufficient reporting for assignment.

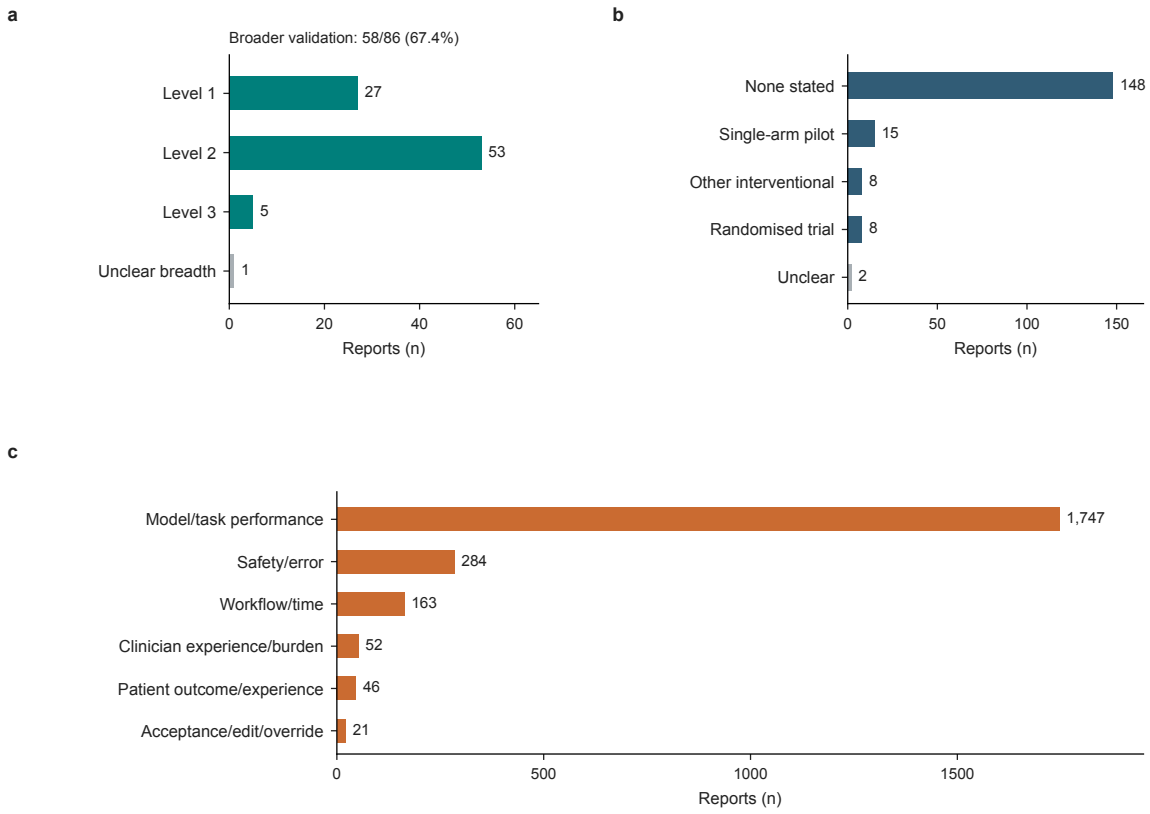


Fig. 5: Evidence supporting clinical translation. **a**, Validation breadth among the 86 reports with integrated, operational, interventional or outcome-impact evaluation (Stage C); 58 (67.4%) had broader validation across multiple datasets or settings, or public and local data (Level 2 or 3). **b**, Trial and implementation-study categories among the 181 eligible full texts. **c**, Selected reported outcomes among 1,870 retained reports. Outcome categories are non-exclusive and derive from title-and-abstract profiles; trial categories derive from the full-text review. Stage C is not equivalent to evidence of clinical benefit.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- llmehrsystematicreviewsupplement.pdf