

Computer vision in clinical medicine: an LLM-powered large-scale systematic review

Eric Oermann

`Eric.Oermann@nyulangone.org`

NYU Langone Health <https://orcid.org/0000-0002-1876-5963>

John Markert

NYU Langone Health

Nicolas Goff

Department of Neurosurgery, NYU Langone Health <https://orcid.org/0000-0002-6005-1008>

Anton Alyakin

Washington University of St. Louis

Jin Vivian Lee

NYU Langone Health <https://orcid.org/0000-0002-8961-0125>

Krithik Vishwanath

NYU Langone Health

Eunice Yang

Columbia University Vagelos College of Physicians and Surgeons

Anjali Kapoor

NYU Langone Health

Joanne Choi

Augusta University/University of Georgia Medical Partnership

Robert Steele

NYU Langone Health

Annelene Schulz

NYU Langone Health

Sully Chen

Duke University School of Medicine <https://orcid.org/0000-0001-7719-469X>

Jinseok Lee

Kyung Hee University College <https://orcid.org/0000-0002-8580-490X>

Analysis

Keywords:

Posted Date: June 5th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9861346/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: **Yes** there is potential Competing Interest. E.K.O. reports consulting with Sofinnova Partners and Google, income from Merck & Co. and Mirati Therapeutics, and equity in Artisight. The other authors declare no competing interests.

Abstract

Since AlexNet, computer vision (CV) has increasingly been investigated and applied in clinical settings. Now, almost 15 years since that result, what is the state of computer vision in healthcare? We sought to answer this question using an LLM-based framework for a critical analysis of the current literature regarding CV models in healthcare. Our LLM-assisted analysis identified 64,553 articles (January 2012–March 2026) subdivided by level of evidence. 0.2% of articles were randomized control trials, and an additional 2.8% were prospective; roughly 85% of articles (n=54,811) were designed purely for benchmarking on static datasets. Within this subset, 77% (n=39,434; 61% of total included) did not report external validation. Other reporting measures were also sparse including reporting of sample size, primary statistical metrics, and human comparisons. Across all studies, investigation of failure modes, model explainability, and human-AI interaction analyses were not included in >90% of studies. Despite increased interest in CV in healthcare in recent years, there is still an endemic gap in the quality of published evidence, highlighting a need for more prospective studies and improved frameworks for deploying and trialing these technologies to enable widespread clinical adoption.

Introduction

AlexNet's success¹ on the ImageNet 2012 challenge² powered by the transition to GPU-accelerated training led to an explosion of deep learning literature—first in general computer vision³ and then in applications to clinical medicine^{4–8}. More recently, the advent of both generalist and specialist vision foundation models accelerated this trend by enabling strong performance across tasks with minimal task-specific training.^{9–12} While some computer vision models have demonstrated robust performance in the clinical setting and are augmenting real clinical workflows^{13,14}, a translational gap related to the disciplinary disconnect between computer scientists and clinicians remains⁸.

For computer scientists, success is typically defined by performance on clean, standardized benchmarks and leaderboards^{15–17}, which fail to capture workflow integration, demographic generalizability¹⁸, the temporal nature of the disease in question¹⁹, and the interpretability required for clinical decision-making. Clinicians, by contrast, operate within well-defined hierarchies of evidence – from case series to randomized controlled trials – in which clinical validity and reproducibility determine the strength of a study^{20,21}. The majority of computer vision studies in medicine, having been conducted on benchmark or retrospective data and evaluated on held-out test sets, occupy an ambiguous position within this hierarchy.^{22,23}

Studies are often conducted on publicly available image datasets or rely on curated data not representative of real-world scenarios and lack external or temporal validation. Furthermore, they often demonstrate confounded results, yet claim state-of-the-art performance and the ability to outperform clinicians.^{18,22–24} Academic incentives, emphasizing novelty and de-emphasizing confirmatory studies and negative-results publishing²⁵, as well as infrastructural, institutional, and regulatory obstacles

associated with deployment trials further perpetuate this issue²⁶. In recent years, more studies have been recognizing these potential pitfalls and are pivoting towards real, prospective data as well as focusing on interpretability, providing more reliable, replicable, and trustworthy comparisons between different computer vision models¹² and physicians^{27,28}. The emphasis has also been shifting from performing tasks that replace physician judgment, such as diagnosis, to tasks that augment clinician-in-the-loop judgement, such as triage²⁸ and screenings²⁹.

Amidst this landscape, an understanding of the real-world clinical impact of computer vision models is needed. Certain criteria have been proposed to assess the strength of a computer vision study in clinical medicine, including the CLAIM³⁰, as well as AI more broadly, including TRIPOD-AI³¹, CONSORT-AI³², and STARD-AI³³. These standards offer reasonable criteria for studies to follow as a baseline, but no consistent, unified framework exists for grading a study's clinical readiness. Additionally, due to the sheer volume and velocity of computer vision literature, most systematic reviews are bound along a single axis, such as a single subspecialty, pathology, task^{23,34}, or study design³⁵, and do not characterize the broader landscape of computer vision models in medicine. No prior review, to our knowledge, has evaluated computer vision studies against a unified clinical-readiness framework across subspecialties, tasks, and study designs, particularly given the recent shift toward generalist and foundation vision models. A current, cross-cutting assessment on a unified readiness scale is therefore warranted.

To fill these knowledge gaps, we adopted the methodology of Chen et al.³⁶, which demonstrated the reliability of a large-scale LLM-assisted review, to conduct a critical analysis of the real-world performance, clinical reliability, and applicability of contemporary computer vision models. To meet this goal, we propose a tiering evaluation rubric that assesses the level of evidence, or “readiness” of a model for the clinic.

In this systematic review, we obtained computer vision-related medical studies from three different databases and utilized a frontier LLM to 1) screen based on our pre-defined inclusion/exclusion criteria, 2) tier each included study based on level of evidence criteria for the applicability and clinical readiness of a study, and 3) extract relevant metadata of the included studies to conduct post-hoc analysis. Both screening and tiering steps were statistically validated by comparison with an independent, blinded review of a sample of studies conducted by two groups of human evaluators. This study aimed 1) to identify the quality and area(s) of focus of current clinical computer vision studies and 2) to better understand the publication landscape.

Results

Collecting and Filtering Clinical Computer Vision Studies

A search for articles published between 1 January 2012 and 8 March 2026 was conducted according to a composite query searching for strings such as ‘computer vision’ and ‘convolutional neural network’ and the intersection of terms like ‘artificial intelligence’ and ‘imaging’ (see **Supplementary Information** for full

query). The search returned 83,424, 56,215, and 86,527 studies using PubMed, Embase, and Scopus, respectively. Deduplication yielded a total of 160,436 unique studies, 64,553 of which were found to meet inclusion criteria by GPT-5.4 (reasoning effort set to high). We performed a human validation study of the filtering performance by randomly sampling 500 of the candidate works and assigning them to two independent reviewers for inclusion and exclusion. The two humans achieved an inter-rater agreement, as measured by Cohen's Kappa, of 0.577 (Bootstrap 95% CI: 0.504-0.648). Tiebreaking was conducted via mutual agreement. When the human consensus was compared with the LLM, the inter-rater agreement was found to have a Cohen's kappa of 0.596 (Bootstrap 95% CI: 0.531-0.660). The paired bootstrap CI on the difference (human kappa - LLM kappa = -0.019) was -0.112 to 0.075.

Tiering Validation

Studies were then categorized into four tiers based on level of evidence as in Chen et al.³⁶ though we selected different criteria based on validation type (see Methods for full criteria). 54,811 were found to be Tier III (internal validation only/benchmark studies), 7,799 were Tier II (external validation), 1,808 Tier I (prospective studies), and 135 S Tier (randomized control trial) (Fig 1). We performed a human validation with 400 studies. Prior to tiering, 1 study from our human validation sample was found to not fit the inclusion/exclusion criteria despite passing through LLM-screening; this study was excluded from tiering. A human inter-rater agreement was found to have a Cohen's kappa of 0.587 (Bootstrap 95% CI: 0.445-0.712). Tiebreaking was conducted via mutual agreement. When the human consensus was compared with the LLM, the inter-rater agreement was found to have a Cohen's kappa of 0.552 (Bootstrap 95% CI: 0.402-0.683). The paired bootstrap CI on the difference (human kappa - LLM kappa = 0.0352) was -0.146 to 0.214.

Metadata Extraction

Metadata extraction was not validated by humans due to the number of parameters extracted in this study; yet other studies support its use for data extraction as a viable, more efficient alternative to human metadata extraction³⁷. Most categories were not mutually exclusive; a single study could be assigned to multiple specialties, modalities, or clinical tasks. Percentages therefore reflect the proportion of studies within each category relative to the total study count and should not be summed or compared pairwise.

Publishing and Computer Science Trends

For trend modeling, we excluded studies from the year 2026 due to indexing lag; we included studies from 2026 for all other results. Between our defined search start and endpoints, we found an exponential increase in the number of clinical computer vision AI studies published across our searched databases, with publication rates growing by approximately 2.5% per month, doubling every ~28 months (Fig. 2a, $r^2=0.89$).

We categorized journals into categories based on subject matter covered (Fig 2b): interdisciplinary dominated the landscape at 29,184 included studies (39.0%), followed by medical/clinical at 22,832 studies (35.4%), and computer/engineering at 11,404 studies (17.7%). When comparing whether studies were published (Fig. 2c), the vast majority were written as full articles (83.0%; n=53,570).

The predominant model family used was an unspecified CNN (27.8%; n=19,658), followed by classical ML methods (15.8%; n=11,143), unspecified deep learning methods (11.3%; n=7,672), U-Net (10.5%, n=7,207), ResNet (8.1%; n=5,575) and ViT/Transformers (6.8%; n=4,696) as shown in Fig 2d. Unspecified CNNs became the most adopted model after surpassing classical ML, the previous dominant method in computer vision, around 2018 and have continued to dominate ever since. Interestingly, since the application of the transformer to computer vision in 2020, transformers/ViT have exponentially increased in use, and in 2025 was the third-most used model family. We observe similar rates of adoption among studies that utilize U-Nets and ResNets. SVM was the most-commonly used model family until 2021, when ResNet models surpassed it and have remained the most-used model family since (Fig. 2f).

Supervised learning was found to be the top training method (n=51,114; 68.7%), with the second training strategy transfer learning encompassing significantly fewer studies (n=6,908; 9.3%) as shown in Fig. 2g. For technical tasks (Fig. 2h), classification was overwhelmingly most common (n=38,787; 51.3%), followed by segmentation (n=21,280; 28.1%); these were also found to be the most dominant technical tasks year by year (Fig. 2i).

Interestingly, regarding study metrics, we do not observe an increase in AUC (Extended Data Fig. 1a) or accuracy (Extended Data Fig. 1b); we do observe an increase in the sample size range used over time (Extended Data Fig. 1c).

Medical Trends

The brain and central nervous system were the most studied anatomical region at 11,771 studies (18.2%), followed by lungs and chest at 9,364 studies (14.5%), and breast at 4,824 studies (7.5%). The most common clinical task was diagnosis, with 36,069 attributed studies (55.6%, Fig. 3b), followed by more advanced tasks such as treatment planning (n=8,368; 12.9%), screening (n=5,979; 9.2%), prognosis (n=4,574; 7.1%), and monitoring (n=3,657; 5.6%). This same ranking of tasks was present when examining year-by-year in Fig. 3c. Interestingly, there is an increasing trend over this period in computer vision models' use for treatment planning, indicating an increased interest in developing models for more complex tasks.

Interventional/diagnostic radiology was the most-studied specialty (Fig. 3d) at 39,695 studies (38.0%), as it deals with imaging extensively. Oncology (n=19,955; 19.1%) surpassed pathology (n=7,505; 7.2%), which could be, in part, attributable to study specialty overlap between oncology and radiology. These were followed by neurology (n=6,046; 5.8%), ophthalmology (n=4,651; 4.4%), and cardiology (n=4,130; 3.9%). These trends held throughout the entirety of the period investigated (Fig. 3e). Notably, we

observed an early adoption of computer vision in radiology in 2012 not seen in other specialties until around 2016-17, when oncology studies began to take off.

The top two most utilized modalities in clinical computer vision were MRI at 15,757 studies (21.0%; Fig. 3f), followed closely by CT at 14,450 studies (19.3%); X-ray (n=7,728; 10.3%), and histopathology at 6,494 studies (8.7%) also were commonly used. Imaging modality trends mostly held over the years (Fig. 3g), apart from CT briefly surpassing MRI from 2021-23.

Foundation Models

The greatest percentage of imaging modalities that used foundation models were in multimodal (≥ 3 modalities) (7.7%; n=3,135) and EHR/Tabular (7.0%; n=2,404) (Extended Data Fig. 2a). The greatest absolute numbers were found in MRI (n=15,757) and CT (n=14,450). The specialties with the greatest percentage of foundation model usage were found to be pathology (3.3%; n=7,505), dermatology (2.8%; n=2,140), and ophthalmology (2.5%; n=4,651) as shown in Extended Data Fig. 2b.

Interventional/diagnostic radiology dominated in terms of absolute totals (n=39,695). The earliest foundation model studies were found in 2021; since then, SAM and GPT were the most prominently featured models year by year (Extended Data Fig. 2c).

Tiering Trends

Fig. 4a depicts the tier distribution of the 64,553 studies categorized by GPT-5.4 high reasoning, which included 54,811 studies (84.9%) rated as Tier III (internal validation only/benchmark studies), 7,799 (12.1%) Tier II, 1,808 (2.8%) Tier I, and 135 (0.2%) S Tier (randomized control trial) studies. Trends over time are shown in Extended Data Fig. 3.

When comparing imaging modality and study tier, endoscopy had the highest proportion of Tier S studies at 2.2% (n=65), while all other imaging types surveyed remained below 1% (Fig. 4b). Studies incorporating EHR/tabular data alongside a CV model had the lowest percentage of Tier III studies at 79.7% (n=1,917). A similar pattern emerged across specialties: other than gastroenterology at 2.2% (n=78), all specialties surveyed have <1% of their studies in Tier S (Fig. 4c). Additionally, gastroenterology also led in Tier I (prospective) studies at 6% (n=224) and had the lowest proportion of Tier III studies at 77% (n=3,325), followed by pulmonology at 79% (n=2,066).

Medical/clinical journals published the fewest proportion of Tier III studies (74%) and the most Tier I and Tier II studies at 6% and 20%, respectively (Fig. 4d). Interdisciplinary journals had a moderate degree of Tier II studies at 10%, but majority Tier III (89%), while most studies in computer/engineering journals fell into Tier III (95%). While full papers generally had some mix of tiers (83% Tier III, 14% Tier II, 3% Tier I) as shown in Fig. 4e, conference abstracts were overwhelmingly rated as Tier III at 97%, possibly due to shorter word limits/space constraints.

We also further categorized journals based on impact factor as reported on Pubshub Journals and Congresses. 1,719 studies (2.7%) were found in journals with an impact factor <2, 21,822 studies

(33.8%) were found in journals with an impact factor between 2 and 4.9, 5,616 studies (8.7%) were found in journals with an impact factor between 5 and 9.9, 2,888 studies (4.5%) were found in journals with an impact factor between 10 and 19.9, 375 studies (0.6%) were found in journals with an impact factor above 20, and 32,133 studies (49.8%) were either found in a journal without an impact factor or were not found within the Pubshub Journals and Congresses database (Extended Data Fig. 4a). After accounting for the number of journals in each category, the normalized counts were: 1.8 for fewer than 2 journals, 9.6 for 2–4.9, 12.1 for 5–9.9, 17.7 for 10–19.9, and 4.4 for 20 or more journals (Extended Data Fig. 4b).

Journal impact factor distributions differed significantly across all study tiers (Fig. 4f). For Tier S studies, impact factor was not a strong predictor of publication venue: the largest shares appeared in journals with an impact factor of 2–4.9 (24%) or no listed impact factor (23%), followed by journals with impact factors ≥ 20 (19%), 10–19.9 (17%), and 5–9.9 (14%) (Extended Data Fig. 4c). Lower tiers leaned more heavily on lower-impact journals: unlisted and 2–4.9 impact factor journals together accounted for 73% of Tier I, 71% of Tier II, and 85% of Tier III studies.

Pearson residual analysis (Extended Data Fig. 4d) revealed Tier S and Tier II were notably over-represented in journals with impact factor ≥ 20 ($r=+22.6$ and $+13.1$), and Tier II also showed over-representation in the 10–19.9 range ($r=+17.2$). Tier I showed comparatively modest over-representation in high-impact journals ($r=+9.2$ in both the ≥ 20 and 10–19.9 bins). Accuracy and AUC distributions by impact factor bin are shown in Extended Data Fig. 5.

Study Quality Metrics

Notably, we found over 90% of included studies lacked a discussion of failure modes (Fig. 5a), explainability (Fig. 5b), or used a human in the loop (Fig. 5c) in their abstracts. Only 7% of studies had failure modes discussed ($n=4,458$; Fig. 5a) or explainability ($n=4,748$; Fig. 5b); just 8% had a human in the loop ($n=4,922$; Fig. 5c).

Sample sizes differed significantly across most tier pairings (Tier S vs. I and III; Tier I vs. II and III; Tier II vs. III; all $p<0.001$), with an additional difference between Tiers I and III ($p<0.01$; Fig. 5d). A similar pattern emerged for sample size reporting rates, where every comparison involving Tier II or III reached $p<0.001$ (Fig. 5e). For the proportion of studies reporting a primary metric (Fig. 5f) and a human evaluator (Fig. 5g), Tier III consistently differed from all other tiers, while Tier I also differed significantly from Tier II for human evaluator reporting (all significant comparisons $p<0.01$ or $p<0.001$). Citation counts (log-transformed; Fig. 5h) were highest for Tier S (median: 15.5, IQR: 2.8–53), with Tiers I and II comparable (both median: 8, IQR: ~ 2 –23/24) and Tier III the lowest (median: 7, IQR: 1–22). Tier S differed significantly from all other tiers, and Tier I from Tier III ($p<0.01$ to $p<0.001$). The same downward trend held across impact factor bins (Fig. 5i), where every pairwise comparison was significant ($p<0.001$): journals with impact factor ≥ 20 had the highest citations (median: 80, IQR: 26–203), followed by IF 10–19.9 (median: 23, IQR: 6–60), IF 5–9.9 (median: 18, IQR: 5–47), IF 2–4.9 (median: 10, IQR: 3–25), and IF <2 (median: 4, IQR: 1–13).

Interestingly, despite differences in levels of evidence between tiers, we observed significantly higher reported accuracies between Tier III and Tier I as well as Tier III and Tier II (Extended Data Fig. 6a) and AUCs between Tier III and Tier II (Extended Data Fig. 6b), indicating higher metric reporting in studies with less rigorous designs.

Tier III Sub-Analyses

Given that roughly 85% (n=54,811) of studies fell within the benchmarking category Tier III, we sought to further describe this distribution, by dividing between other forms of validation that do not influence level of evidence, including internal validation with a held-out test set, cross-validation only, another type of validation (human comparison only, etc.), or none reported. The vast majority (76.9%; n=39,434; 61% of all studies) did not report any validation, 14.2% (n=7,298) used internal, held-out validation, 4.7% (n=2,385) used cross-validation only, and 4.2% (n=2,130) used other forms of evaluation (Fig. 6a). Trends within Tier III can be found in Extended Data Fig. 7.

Sample size was found to be significantly higher in internal validation studies relative to all others within Tier III (Fig. 6b). No reported evaluation had a significantly higher sample size in comparison to cross-validation-only studies. Significant differences were observed regarding sample size reporting for interval validation and cross-validation relative to the other evaluation category and no reported evaluation (Fig. 6c). Significant differences were also found between other evaluations and no evaluations; these same significant differences were also found for reporting primary metrics (Fig. 6d). Validation group distributions among imaging modalities (Fig. 6e) and specialties (Fig. 6f) are also shown; no reported evaluation is the dominant percentage of studies across all imaging modalities and specialties. Extended Data Figure 8 demonstrates the reporting incompleteness of other factors within the Tier III that lacked validation reporting, including sample size and primary metrics as well as specialty, imaging modality, and journals with an impact factor.

Discussion

Despite an exponential increase in 2.5% new studies per month since 2012, the methodology of the field of computer vision has not kept up to this pace; over 80% of the studies published since then lack external validation, and only a handful are prospective (2.8%, n=1,808) or randomized control trials (0.2%, n=135). This disproportionate split reflects a widening gap between publication volume and evidentiary rigor in computer vision. Simpler computer vision tasks, such as classification and segmentation, as well as simpler clinical tasks, such as diagnosis, are increasing more rapidly in absolute terms than more complex tasks like triage, treatment planning, and screening. Across specialties, the lowest proportion of Tier III studies was observed in gastroenterology (77%), which nonetheless represents an overwhelming majority. Tier III studies lacking external validation similarly dominate across all journal categories and domains.

Beyond validation deficits, the reporting landscape compounds these concerns. Studies overwhelmingly do not report failure modes or explainability in their abstracts, or fail to include humans in the loop, as

would be required in the clinic in some form or fashion. The lack of a sample size being reported as well as a primary metric indicates a need for more robust criteria for abstract publishing across the board. Each of these omissions has direct clinical impacts: a model whose failure modes are unknown cannot be safely deployed, and a model without human oversight has no mechanism for error correction in high-stakes settings. Although the highest sample size median was observed in external validation studies (Tier II), this does not offset the decrease in methodological rigor present in both prospective (Tier I) and randomized control trials (Tier S).

To contextualize how little has changed, consider one of the earliest studies investigating artificial intelligence in medicine in 2020, written in the CNN era prior to the popularization of the transformer, was Nagendran, et. al.²³, which found a mere 10 randomized control trials and 9 prospective studies, with a suboptimal adherence to reporting standards across the 81 studies surveyed. Our findings suggest that the subsequent wave of publications driven by a broader AI hype cycle has not only failed to correct these deficiencies, but has in fact worsened them proportionally, with randomized control trials and prospective studies now representing a smaller percentage of total output today than they did then.

While abstract-based analyses may underestimate full-text rigor, the failure to report key methodological details *in the abstract* is itself a meaningful finding. Extensive existing guidelines are present in the literature, including guidelines specific for abstracts, such as the TRIPOD criteria^{31,38}, in which the 2015 version instructs authors to “provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions” in the abstract, yet our findings suggest that these are not being enforced effectively among the vast majority of studies, indicating that the landscape has not only not changed significantly since Nagendran in 2020²³, the proportion of high-evidence studies has declined relative to total output, as a fewer percentage of randomized control trials and prospective trials exist in the literature (only 0.2% and 2.8%, respectively). The lack of an emphasis on error reporting in the studies surveyed also indicates a lack of an assessment of shortcut learning, which due to the high-stakes nature of clinical medicine, means that these models are still far from being ready for clinical deployment across most specialties and applications.³⁹ Additionally, while human-in-the-loop is not part of many abstracts, this information is still important to clinical deployment and should be emphasized further. Our findings regarding citation count by degree of evidence indicate that there are many studies being published despite their lack of generalizable findings, a pattern consistent with Ioannidis's broader argument that incentive structures in academic publishing systematically favor positive, novel findings over rigorous ones, and one that warrants serious attention from journal editors and reviewers alike.⁴⁰

This study was limited by several factors. First, we were limited by our ability to classify based on abstracts, due to the infeasibility of 1) extracting full manuscripts with many existing behind paywalls and 2) the expense of running the number of studies screened through GPT-5.4 high reasoning using a full corpus. The main way in which this limitation affected our study was with regard to level of evidence comparisons, as when training split was not reported, all of these studies defaulted to tier III; the

significant differences we observed across other reported metrics, and between studies with known evidence tiers, however, suggest our core conclusions are robust to this constraint. Second, we were limited in that we only utilized GPT-5.4 rather than conducting a survey of frontier models for performance, an approach we acknowledge carries its own methodological irony given this paper's critique of under-validated AI. We consider it a defensible choice given the expense and scope of the corpus, but a comprehensive evaluation of LLM performance on systematic review tasks remains an important priority for the field. Additionally, we did not conduct a risk of bias assessment in this study due to both the qualitative nature of it as well as the lack of an ability to provide a rigorous assessment without full-text studies; future studies investigating computer vision in clinical medicine should incorporate this into their analyses.

The evidence presented here supports two clear conclusions. We hope that this LLM-powered systematic review serves as both an assessment of where the field of clinical computer vision has come and a framework for how we should move forward. The field must move away from the proliferation of internally validated studies and toward genuine external validation across diverse clinical settings. We call for more Tier I and Tier II studies as a necessary precondition for an eventual increase in Tier S randomized control trials that use computer vision in a way that directly impacts patient outcomes. Until that evidentiary foundation exists, deployment of computer vision models in pivotal clinical moments is premature, and the current literature does not yet provide sufficient justification to proceed otherwise.

Methods

This study was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses PRISMA guidelines (2020)⁴¹.

Search Strategy and Study Selection

Studies between January 1st, 2012 and March 8th, 2026 were acquired from Pubmed, Embase, and Scopus according to the search queries in Supplementary Information. In summary, we utilized a structure that looked for the intersection of imaging terms, artificial intelligence terms, and clinical terms. We excluded all reviews, meta-analyses, comments, editorials, letters, errata, case reports, and animal studies that did not include human participants. Studies were deduplicated by exact titles and DOIs. Abstracts less than 50 words were automatically excluded due to lack of sufficient information for accurate classification. Studies were included if 1) they utilized a computer vision model in the context of healthcare or medicine and 2) were used in clinical evaluation. Studies were excluded if they 1) used non-computer vision AI models (i.e. tabular models, LSTMs, LLMs in isolation), 2) were used for non-clinical tasks (i.e. isolated laboratory techniques) or not used in a clinical manner. EEG and EKG models as well as audio models were included if they projected input signal into an image prior to being applied (e.g. spectrograms).

Automated Screening

Given the quantity of included studies, it was impractical to screen using human screeners. Therefore, we implemented a similar approach to Chen et. al,³⁶ using an LLM-based screening strategy for inclusion and exclusion, tiering, and extraction of relevant metadata using GPT-5.4 (high reasoning mode), the latest frontier model available at the time this study was conducted. Due to the expensive nature of the task and a previous indication of reliability for performing systematic review-style tasks, we relied solely on GPT-5.4 rather than performing a survey of existing frontier models. GPT-5.4 low reasoning was used for metadata classification. Prompts for all steps are available in our GitHub repository.

For all studies after deduplication and removal of short abstracts (<50 words), each title and abstract pair was submitted with a prompt detailing our above search strategy (see GitHub repo), with the LLM instructed to include or exclude as well as the reasoning behind the decision.

Human Screening Validation

We performed a human validation study of the filtering performance by randomly sampling 500 of the candidate works and assigning them to two independent reviewers for inclusion and exclusion. Tiebreaking was conducted via mutual agreement. Inter-rater agreement was calculated via Cohen's kappa as well as 95% confidence intervals via bootstrapping. Human and LLM concordance was also calculated via Cohen's kappa as well as 95% confidence intervals via bootstrapping.

Automated Tiering

We established a tiering system for level of evidence based on the degree of validation of each study. Tier S was defined as randomized control trials in a real-world setting using medical AI computer vision algorithms, with examples including IDx-DR (diabetic retinopathy screening, FDA authorized with RCT data)¹³, AI-ECG for detecting low ejection fraction with demonstrated clinical utility⁴², and mammography AI with RCT showing improved cancer detection rates²⁹. A study cannot reach Tier S level if it lacks randomization, doesn't measure clinical outcomes, or lacks inclusion within a clinical workflow.

Tier I studies were defined as prospective (or pseudoprospective) studies of medical AI computer vision algorithms, with examples including chest X-ray pneumonia detection tested prospectively in ED, skin lesion classification validated on consecutively enrolled patients, or fundus imaging models tested prospectively across multiple clinics. Studies in this tier cannot allow cherry picking of cases or have a lack of temporal validation and would not include randomized control trials.

Tier II studies were defined as external validation conducted but on retrospectively collected data only, where the validation cohort is independent from the training dataset. These studies must have some documentation of dataset characteristics and training splits. Examples of studies within this category include pathology slide analysis validated on external histology databases or retinal OCT models tested on data from different imaging devices. Studies in this tier cannot have any inclusion within a live clinical

workflow, any clinician-model interaction, and cannot have an external validation set that is a future, prospective dataset.

Tier III studies were defined as performance evaluation on a public benchmark with no validation; this evaluation may compare other studies that have utilized the same benchmark and generally consists of proof-of-concept studies. Tier III studies demonstrate algorithms at the development and optimization stage, the furthest away from impacting patient care. Some examples of studies within this category include models tested on public datasets only (ChestX-ray⁸⁴³, Kaggle, etc) and studies in this category cannot include any claims of clinical readiness nor evidence of performance across different populations. All studies that lacked validation specification in the abstract and/or title were included within this tier.

Human Tiering Validation

For tiering validation, we extracted 400 studies to be reviewed by two reviewer groups. Prior to review, one study was excluded for not meeting our inclusion-exclusion criteria. We again assigned these 399 studies to two independent reviewers for tiering. Tiebreaking was conducted via mutual agreement. Inter-rater agreement was calculated via Cohen's kappa as well as 95% confidence intervals via bootstrapping. Human and LLM concordance was also calculated via Cohen's kappa as well as 95% confidence intervals via bootstrapping.

Metadata Extraction

Model families, training strategies, technical tasks, anatomical region, clinical applications, specialty, imaging modality, failure modes, explainability, use of humans in the loop, sample sizes, primary metrics, human evaluators were extracted using prompts found in our GitHub repository. GPT-5.4 low reasoning was used to sort these into groups (i.e. raw imaging types such as SPECT and CTA were both mapped to CT, a more general category). OpenAlex⁴⁴ was used for obtaining citation metrics based on DOI.

Statistical analysis

Among tiers, we recorded the distribution of reported sample sizes, whether a sample size was even reported, whether statistical metrics of any type were reported, and whether the use of a human evaluator for model comparison was reported in the abstract. We performed Mann-Whitney *U* tests for the actual sample size and journal impact factor (if available) across LLM Tiers and a Chi-Square test for reporting sample size, statistical metrics, and the use of a human evaluator in study abstracts, both utilizing FDR (Benjamini-Hochberg) corrections. For citation counts, a Mann-Whitney *U* test with FDR correction (Benjamini-Hochberg) was applied across study design tiers and journal impact factor bins, with citation data available for 63,638 included studies (98.6%). For our sub-analyses in Tier III, we performed analogous statistical tests among degree of validation instead of tiering for sample size, journal impact factor (if available), whether a study reported sample size, and whether a study reported statistical metrics.

Declarations

Competing interests: E.K.O. reports consulting with Sofinnova Partners and Google, income from Merck & Co. and Mirati Therapeutics, and equity in Artisight. S.F.C. reports equity in OpenAI. The other authors declare no competing interests.

Acknowledgments. This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) with a grant funded by the Ministry of Science and ICT (MSIT) of the Republic of Korea in connection with the Global AI Frontier Lab International Collaborative Research. (No. RS-2024-00469482 & RS-2024-00509257).

References

1. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).
2. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. in *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (2009). doi:10.1109/CVPR.2009.5206848.
3. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
4. Gulshan, V. *et al.* Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **316**, 2402–2410 (2016).
5. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
6. Rajpurkar, P. *et al.* CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. Preprint at <https://doi.org/10.48550/arXiv.1711.05225> (2017).
7. De Fauw, J. *et al.* Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342–1350 (2018).
8. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. *Nat. Med.* **28**, 31–38 (2022).
9. Wang, X. *et al.* Transformer-based unsupervised contrastive learning for histopathological image classification. *Med. Image Anal.* **81**, 102559 (2022).
10. Xu, H. *et al.* A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**, 181–188 (2024).
11. Lyu, Y. *et al.* Learning neuroimaging models from health system-scale data. *Nat. Biomed. Eng.* <https://doi.org/10.1038/s41551-025-01608-0> (2026) doi:10.1038/s41551-025-01608-0.
12. Alyakin, A. *et al.* CNS-Obsidian: A Neurosurgical Vision-Language Model Built From Scientific Publications. *Neurosurgery* 10.1227/neu.0000000000004070 (2026) doi:10.1227/neu.0000000000004070.
13. 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).

14. Lång, K. *et al.* Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *Lancet Oncol.* **24**, 936–944 (2023).
15. Menze, B. H. *et al.* The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans. Med. Imaging* **34**, 1993–2024 (2015).
16. Irvin, J. *et al.* CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. Preprint at <https://doi.org/10.48550/arXiv.1901.07031> (2019).
17. Antonelli, M. *et al.* The Medical Segmentation Decathlon. *Nat. Commun.* **13**, 4128 (2022).
18. Zech, J. R. *et al.* Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med.* **15**, e1002683 (2018).
19. Link, K. E. *et al.* Longitudinal deep neural networks for assessing metastatic brain cancer on a large open benchmark. *Nat. Commun.* **15**, 8170 (2024).
20. The hierarchy of research evidence - from well conducted meta-analysis down to small case series | Health Knowledge. <https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1a-epidemiology/hierarchy-research-evidence>.
21. OCEBM Levels of Evidence. <https://www.cebm.ox.ac.uk/resources/levels-of-evidence/ocebm-levels-of-evidence>.
22. Liu, X. *et al.* A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit. Health* **1**, e271–e297 (2019).
23. Nagendran, M. *et al.* Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* **368**, m689 (2020).
24. Roberts, M. *et al.* Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* **3**, 199–217 (2021).
25. Varoquaux, G. & Cheplygina, V. Machine learning for medical imaging: methodological failures and recommendations for the future. *Npj Digit. Med.* **5**, 48 (2022).
26. Wiens, J. *et al.* Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.* **25**, 1337–1340 (2019).
27. Wolf, R. M. *et al.* Autonomous artificial intelligence increases screening and follow-up for diabetic retinopathy in youth: the ACCESS randomized control trial. *Nat. Commun.* **15**, 421 (2024).
28. Kondepudi, A. *et al.* Health system learning achieves generalist neuroimaging models. *Res. Sq.* rs.3.rs-8166797 (2025) doi:10.21203/rs.3.rs-8166797/v1.
29. Gommers, J. *et al.* Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a

- randomised, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial. *Lancet* **407**, 505–514 (2026).
30. Tejani, A. S. et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. *Radiol. Artif. Intell.* **6**, e240300 (2024).
 31. Collins, G. S. et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378 (2024).
 32. Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J. & Denniston, A. K. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat. Med.* **26**, 1364–1374 (2020).
 33. Sounderajah, V. et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nat. Med.* **31**, 3283–3289 (2025).
 34. Aggarwal, R. et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *Npj Digit. Med.* **4**, 65 (2021).
 35. Plana, D. et al. Randomized Clinical Trials of Machine Learning Interventions in Health Care. *JAMA Netw. Open* **5**, e2233946 (2022).
 36. Chen, S. F. et al. LLM-assisted systematic review of large language models in clinical medicine. *Nat. Med.* **32**, 1152–1159 (2026).
 37. Gartlehner, G. et al. Artificial Intelligence–Assisted Data Extraction With a Large Language Model: A Study Within Reviews. *Ann. Intern. Med.* **178**, 1763–1771 (2025).
 38. Moons, K. G. M. et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann. Intern. Med.* **162**, W1-73 (2015).
 39. Geirhos, R. et al. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2**, 665–673 (2020).
 40. Ioannidis, J. P. A. Why Most Published Research Findings Are False. *PLoS Med.* **2**, e124 (2005).
 41. Page, M. J. et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* **372**, n71 (2021).
 42. Yao, X. et al. Artificial intelligence–enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. *Nat. Med.* **27**, 815–819 (2021).
 43. Wang, X. et al. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 3462–3471 (2017). doi:10.1109/CVPR.2017.369.
 44. Priem, J., Piwowar, H. & Orr, R. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. Preprint at <https://doi.org/10.48550/arXiv.2205.01833> (2022).

Figures

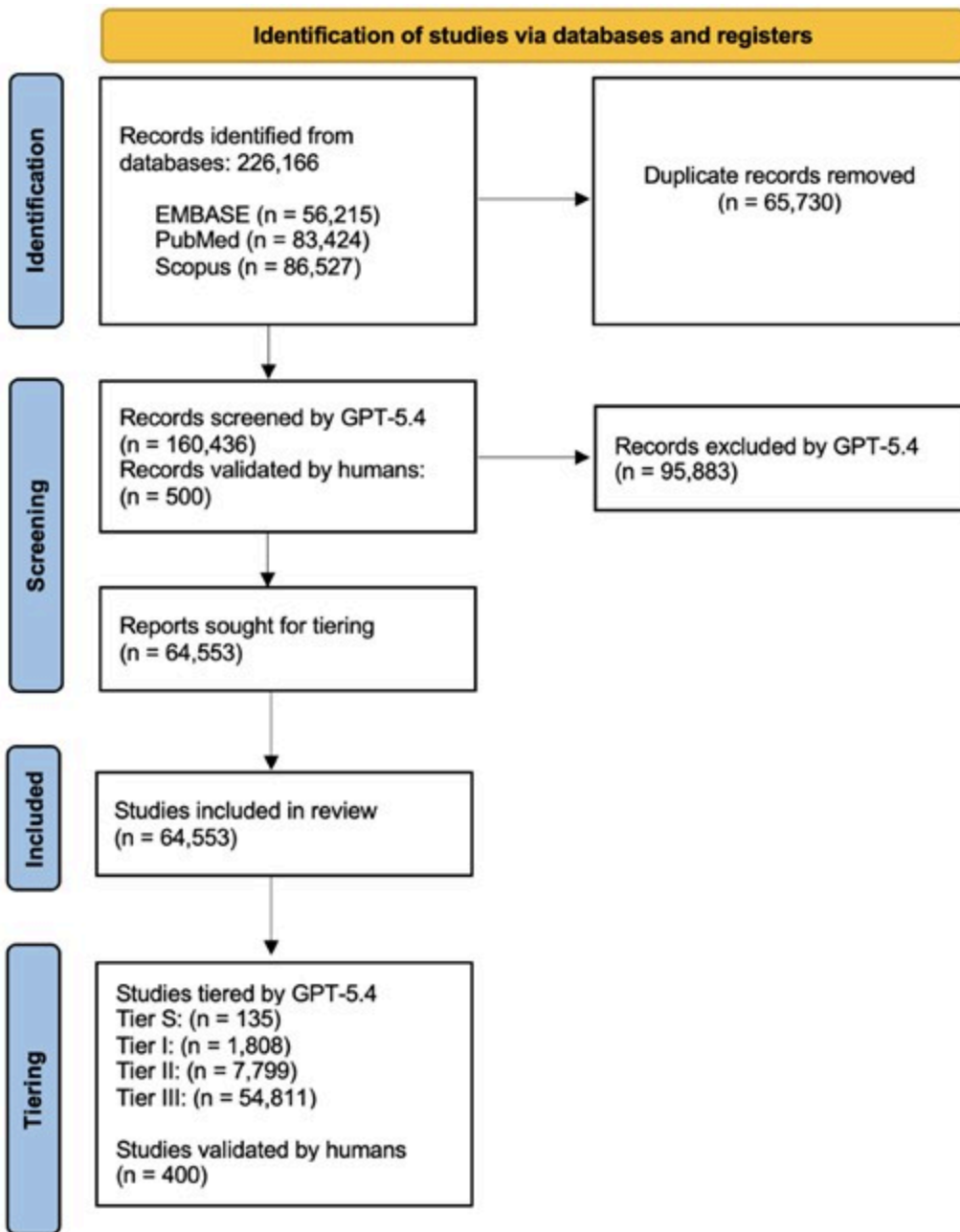


Figure 1

After scraping EMBASE, PubMed, and Scopus, 226,166 were found. Deduplication of DOI and Title resulted in 160,436 studies for screening by GPT-5.4, of which 64,553 passed this criteria. 500 studies were validated by human reviewers for inclusion-exclusion. Then, GPT-5.4 tiered these studies for level of evidence, with 400 validated by humans.

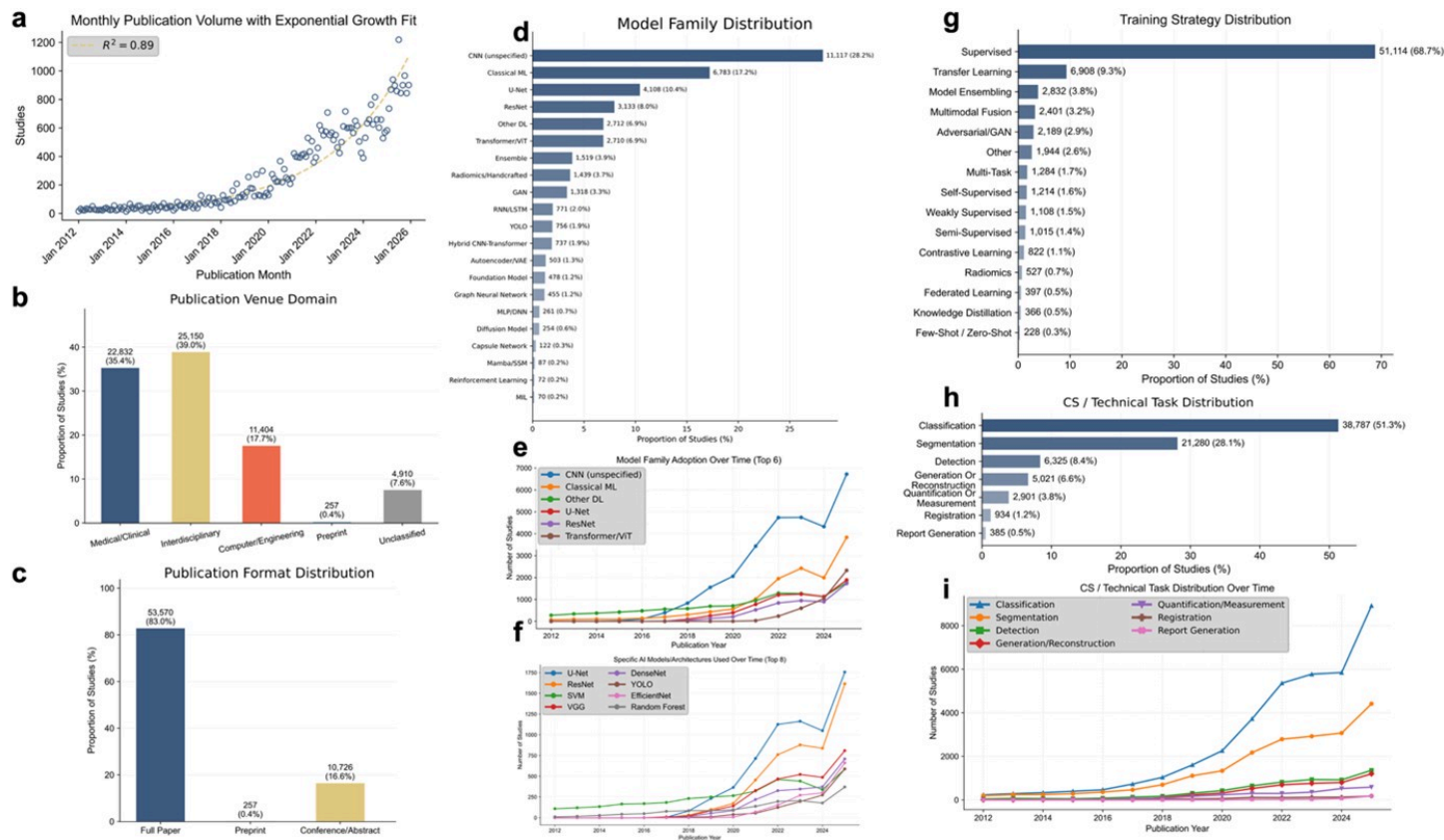


Figure 2

a) Monthly exponential distribution of publications, with a 2.5% increase per month, from January 2012 to December 2025 ($r^2=0.89$). **b)** Subject domain of publications. **c)** Publication format (Full Paper, Preprint, and Conference). **d)** Model family distributions overall and **e)** trends over time. **f)** Specific model family adoption over time. **g)** Training strategies used among studies. **h)** Computer vision tasks among studies and **i)** trends over time.

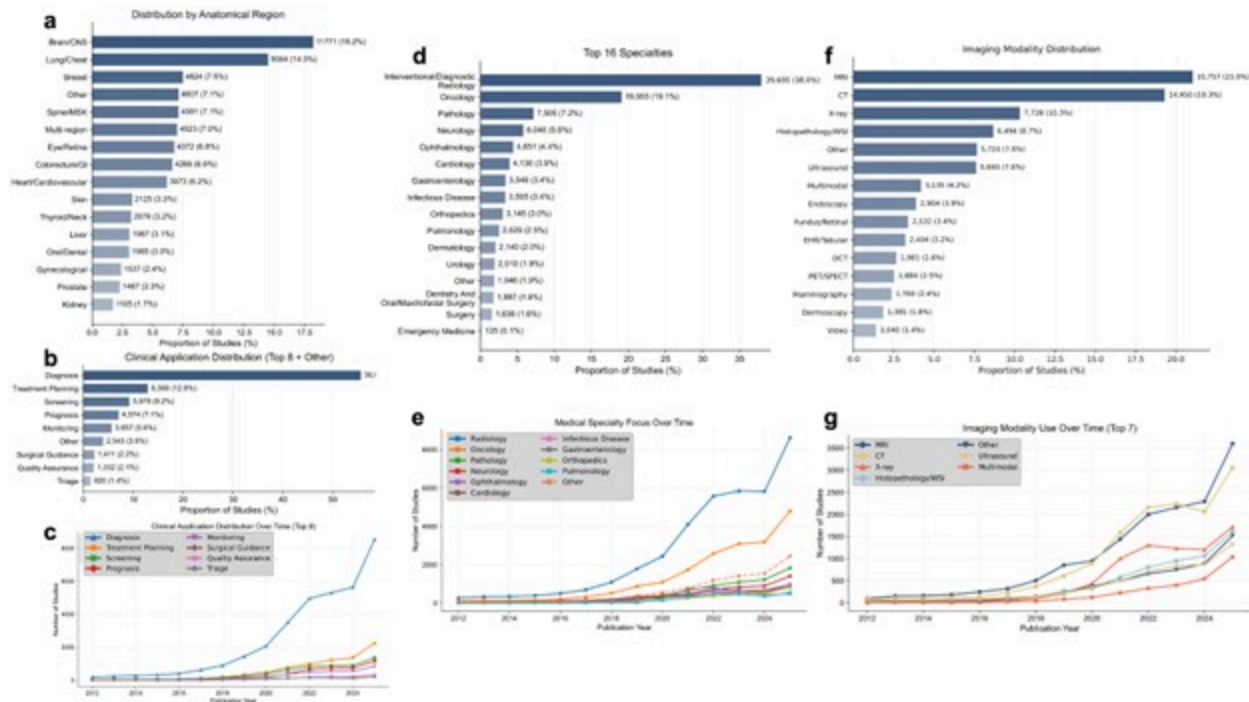


Figure 3

a) Distribution of studies by anatomical region. **b)** Distribution of clinical applications used by computer vision models and **c)** trends over time. **d)** Top specialties utilizing computer vision models and **e)** trends over time. **f)** Top imaging modalities used in studies and **g)** trends over time.

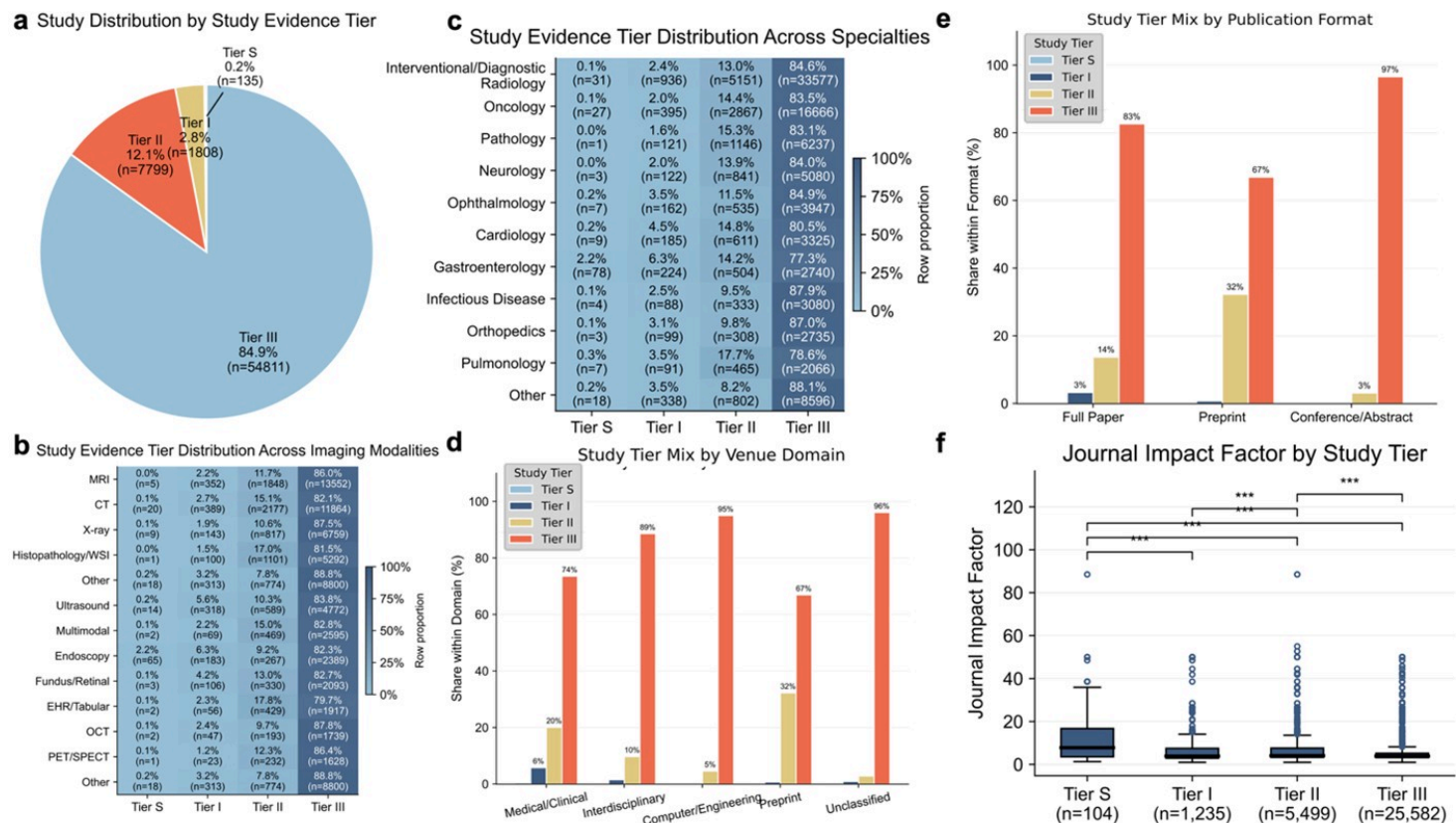


Figure 4

a) Overall study distribution metrics by level of evidence and **b)** level of evidence CDF by year. Distribution of level of evidence of studies **c)** among specialties, **d)** among journal domains, and **e)** among publication formats. **f)** Distribution of impact factor by study level of evidence with Mann-Whitney U test results with FDR (Benjamini-Hochberg) corrections. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

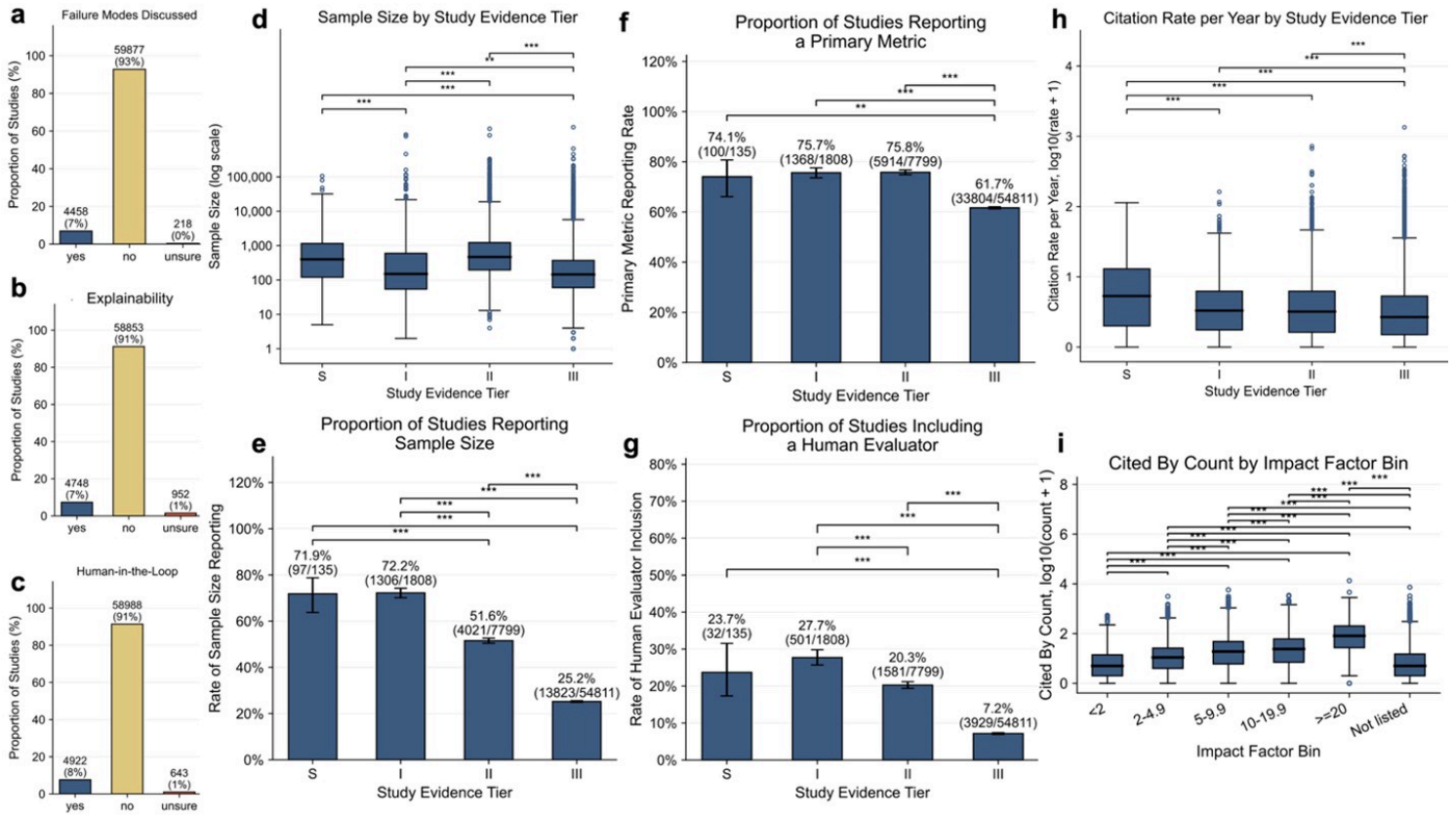


Figure 5

Summaries of GPT-5.4 replies for studies with **a)** failure modes discussed, **b)** explainability and **c)** human-in-the-loop. Mann-Whitney U test for **d)** sample sizes and Chi-Square test results for **e)** reporting sample size, **f)** statistical metrics, and **g)** the use of a human evaluator in study abstracts, and **h)** the number of citations, all utilizing FDR (Benjamini-Hochberg) corrections among study evidence tiers. **i)** Comparisons of citation metrics by journal impact factor bins. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

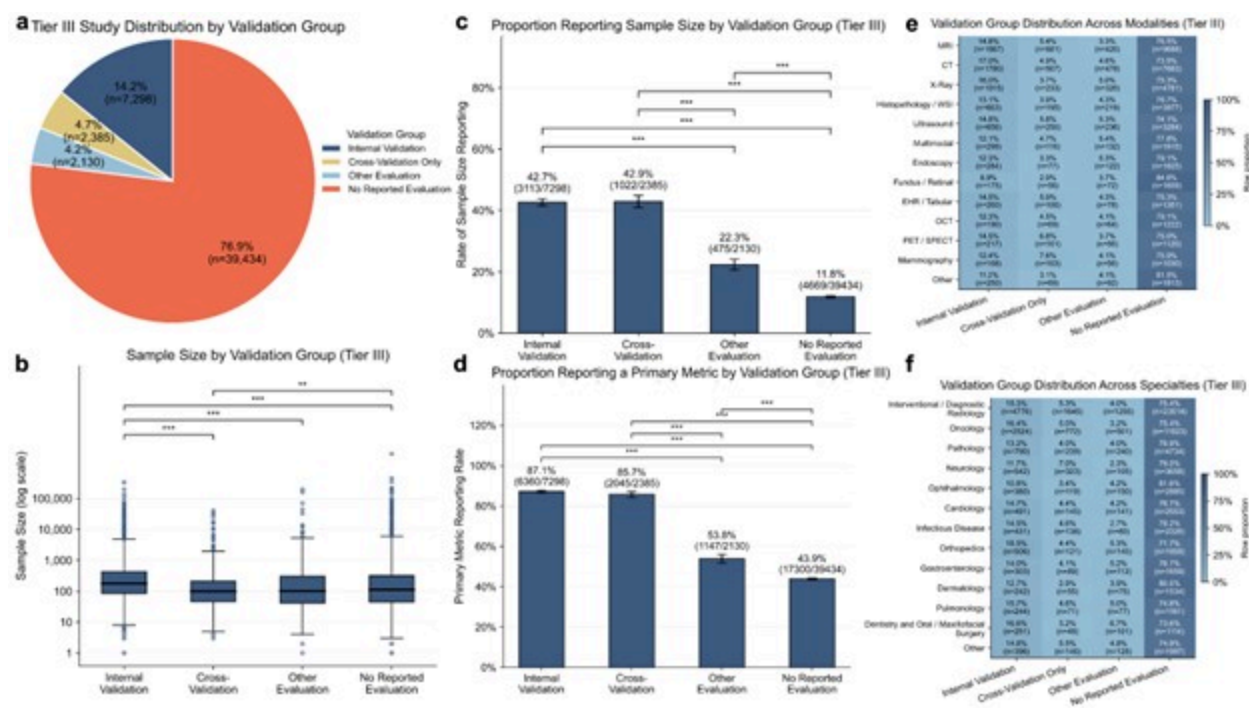


Figure 6

a) Studies within Tier III by degree of validation. Mann-Whitney U test for b) sample sizes and Chi-Square test results for c) reporting sample size and d) statistical metrics. Distribution of studies by validation group among e) imaging modalities and f) specialties. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Supplementary Files

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

- [NMEDAN152853SupplementaryInformation.docx](#)
- [NMEDAN152853ExtendedData.docx](#)