Cureus
Part of SPRINGER NATURE

Open Access Review Article

Artificial Intelligence (AI) in Healthcare

Divyanshi Baijal [1], Himanshi Khattar [1], Piyush Kumar [1]

1. Radiation Oncology, Shri Ram Murti Smarak Institute of Medical Sciences, Bareilly, IND

Corresponding author: Piyush Kumar, piykumagr@gmail.com

Review began 08/19/2026
Review ended 11/12/2026
Published 11/17/2026

© **Copyright** 2026
Baijal et al. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

DOI: 10.7759/cureus.119829

Abstract

Artificial intelligence (AI) is steadily transforming healthcare, not through a single breakthrough but through a growing range of applications that support clinicians across diagnosis, treatment planning, workflow management, and drug development. However, the degree to which these technologies are ready for routine clinical use varies considerably between specialties. This review examines evidence from 26 peer-reviewed studies, systematic reviews, and policy papers covering ophthalmology, breast and prostate pathology, lung cancer, gastrointestinal endoscopy, gynecologic oncology, radiotherapy, chronic disease management, clinical documentation, consumer health applications, drug discovery, and the regulatory and educational frameworks surrounding AI in medicine. The strongest evidence consistently comes from imaging-based applications, where deep learning systems have achieved diagnostic performance comparable to, and in some cases exceeding, that of experienced specialists while improving efficiency and reducing workload. In contrast, AI models for treatment planning, toxicity prediction, and cancer prognosis remain promising but are often supported by retrospective, single-center studies with limited external validation. Large language models are beginning to ease the burden of clinical documentation, although concerns regarding bias, privacy, transparency, and hallucinations remain important barriers to widespread adoption. Consumer-facing AI tools and generative AI for drug discovery show considerable potential but require more rigorous validation before routine implementation. Across all domains, a common theme emerges: the greatest challenges are no longer solely technical but relate to generalizability, regulatory oversight, clinician training, and responsible integration into healthcare systems. Current evidence suggests that AI is most effective as a tool that augments, rather than replaces, clinical expertise. Continued progress will depend on robust prospective validation, transparent reporting, and thoughtful collaboration between clinicians, researchers, regulators, and healthcare institutions.

Categories: Healthcare Technology
Keywords: artificial intelligence, clinical decision support, deep learning, digital pathology, healthcare, large language models, machine learning, medical imaging, oncology, precision medicine

Introduction And Background

A field-by-field reckoning with clinical AI

Artificial intelligence (AI) has entered medicine not through one grand entrance, but through dozens of side doors such as a camera attachment in an eye clinic, a second opinion on a mammogram, a language model condensing a hospital chart, and a chatbot triaging a worried parent at 2 a.m. This review draws on 26 peer-reviewed studies, systematic reviews, and policy papers spanning more than a dozen specialties to ask a specific question: by the standards clinicians actually use - diagnostic accuracy, external validation, reproducibility - how far has each of these applications really progressed?

The answer varies sharply by domain, and the variation itself is instructive. In imaging-heavy fields, such as reading a retinal photograph, grading a pathology slide, and flagging a polyp mid-colonoscopy, several AI systems have demonstrated high diagnostic performance, with some reporting post area-under-the-curve (AUC) scores above 0.90, benchmarked against consensus panels of specialists rather than single readers, and in a growing number of cases have been tested across multiple institutions, scanners, and patient populations. Even here progress has been uneven: anterior-segment conditions such as cataracts lagged behind fundus-based screening for years, held back by the absence of standardized imaging protocols, until slit-lamp video capture gave AI a comparably rich data source to work with [1]. In other domains, radiotherapy toxicity prediction, cancer recurrence modeling, and gynecologic oncology, the underlying signal looks genuinely promising, but the supporting evidence still leans heavily on small, single-center cohorts with little prospective or external validation. In some of the most consequential parts of medicine - regulatory oversight, clinician training, organizational leadership - the open questions aren't really about model performance at all, but about the human and institutional infrastructure needed to deploy these tools safely.

What follows proceeds specialty by specialty: ophthalmology, breast and prostate imaging, lung cancer, gastrointestinal endoscopy, gynecologic and cervical oncology, radiotherapy planning, oncologic risk prediction, clinical documentation, population health, consumer-facing health apps, drug discovery, and the regulatory and educational scaffolding around all of it. Each section reports what the underlying studies

actually measured, at what scale, and against what benchmark - the details that determine whether a given result is close to clinical readiness or still several validation steps away.

Literature search and selection

This article is a narrative review exploring the evolving role of AI in healthcare, with a particular focus on its applications across different medical specialties, including oncology. The aim was to bring together relevant published evidence to understand the current clinical applications of AI, assess the progress made so far, and highlight the challenges that may influence its integration into routine clinical practice.

Relevant literature was identified through searches of PubMed/MEDLINE, Scopus, Web of Science, and Google Scholar. The literature search was conducted over a one-month period, from 1 July 2026 to 1 August 2026, to identify relevant publications on the applications of AI in healthcare and oncology.

The search included combinations of keywords such as artificial intelligence, machine learning, deep learning, neural networks, generative AI, clinical decision support, medical imaging, cancer diagnosis, radiotherapy planning, toxicity prediction, oncology, and drug discovery. Additional specialty-specific terms were used to identify studies addressing applications in ophthalmology, breast and prostate cancer, lung cancer, gastrointestinal endoscopy, gynecological oncology and other areas of clinical practice.

The review considered original research articles, systematic reviews, meta-analyses and relevant policy papers that discussed the development, evaluation or clinical application of AI in healthcare. Particular attention was given to studies reporting diagnostic performance, clinical outcomes, treatment planning applications, predictive models and the practical challenges of implementing AI-based technologies. Publications with limited relevance to the scope of the review, duplicate reports, or insufficient clinical or methodological information were not considered.

The selection of literature was based on its relevance to the objectives of the review and its contribution to understanding the role of AI across different areas of medicine. Titles and abstracts were reviewed initially, followed by assessment of the relevant full-text articles. Reference lists of important publications and review articles were also examined to identify additional studies that might otherwise have been missed.

A total of 26 publications were included in the final review. These were selected to provide a representative overview of the major developments in clinical AI, covering both established applications and emerging areas of research. The intention was not to include every publication available on the subject, but rather to highlight studies that offered meaningful clinical evidence, demonstrated important technological developments, or addressed key challenges associated with the adoption of AI in healthcare.

Since this article is a narrative review and not a systematic review, the literature search was not conducted according to a formal systematic-review protocol. No formal risk-of-bias assessment or meta-analysis was performed. The selection of publications involved an element of author judgement, and the possibility of selection bias therefore cannot be ruled out. The findings should be interpreted as a narrative synthesis of relevant published evidence rather than an exhaustive assessment of all available literature.

Review

Eye care: cataracts, diabetic eyes, and glaucoma

Vision loss from cataracts, diabetic retinopathy, and glaucoma is largely preventable when caught early, but early detection depends on access to an ophthalmologist, which much of the world simply doesn't have. That access gap is where AI has made its earliest and most rigorously validated case in all of medicine: a photograph of the eye, acquired using a hospital-grade fundus camera or a smartphone-mounted imaging attachment, can be analyzed using a trained neural network to assist in the early detection and prediction of ophthalmic conditions such as diabetic retinopathy and glaucoma. AI-based image analysis can also support cataract screening using appropriate anterior-segment imaging.

One study trained an EfficientNetV2-based model on 38,320 diagnosable frames drawn from 1,812 cataract-affected eyes, filmed with a smartphone-attachable slit-lamp device, and graded nuclear cataract severity against the WHO's four-tier system; it achieved AUCs between 0.923 and 0.967 across the four grades, with overall accuracy, sensitivity, and specificity of 0.94, 0.96, and 0.93 - closely tracking conventional slit-lamp assessment [1]. A separate, far larger effort led by Google trained an Inception-v3 network on 128,175 retinal photographs, each graded three to seven times by a panel of 54 ophthalmologists, then validated it on 9,963 US images and 1,748 French images graded independently by seven to eight board-certified specialists; the algorithm reached AUCs of 0.991 and 0.990 on the two datasets respectively, with sensitivity as high as 97.5% at the operating point tuned for screening. Notably, performance plateaued once training data passed roughly 60,000 images, meaning further gains came mainly from better-quality grading rather than sheer data volume [2]. A third study, trained on 39,745 fundus photographs from Chinese clinics and validated on a held-out set of 8,000 images, detected glaucomatous optic neuropathy with an AUC of 0.986, 95.6% sensitivity, and 92.0% specificity, a false-positive rate of roughly 8% in this validation set. A detailed error

analysis found that most false negatives occurred when glaucoma coexisted with high myopia, a combination that confounds human graders too, and that 95.4% of the 480 false-positive cases in fact showed a genuine, if different, optic disc abnormality, most commonly physiologic cupping [3].

The pattern across all three studies is consistent: AUCs at or above 0.92, built and validated largely within a single country, health system, or ethnic group. The open question is no longer whether these systems can read an eye scan accurately - the evidence says they can - but whether that accuracy transfers across the full range of ethnic backgrounds, lighting conditions, and camera hardware found in clinics worldwide, a question none of the three studies was designed to fully answer.

Breast cancer: from screening to the pathology slide

Breast cancer screening generates an enormous volume of imaging that needs a trained eye, alongside a persistent, well-documented problem of both missed cancers and false alarms that send healthy women for unnecessary biopsies. In 2020, a DeepMind and Google Health team, working with Imperial College London and several NHS trusts, published a widely covered Nature study using an ensemble of three convolutional neural networks - separate lesion, breast, and full-case models - to read screening mammograms, reigniting debate over whether software could ease this burden without lowering the bar on accuracy [4].

That system, trained and tested on 121,455 UK and 22,225 US screening cases, achieved an AUC of 0.889 in the UK test set and 0.811 in the US set, and at selected operating points outperformed the average individual radiologist while matching the accuracy of the UK's standard double-reading protocol - though the paper drew sharp public criticism from other scientists over incomplete code and hyperparameter disclosure, a reproducibility dispute the authors partly defended on patient-safety grounds [4]. A broader review of AI across the full breast cancer care pathway found comparable strength elsewhere in the pipeline: pooling 16 studies, standalone AI matched or exceeded individual radiologists on sensitivity, specificity, and AUC when reading digital mammography, and reduced tomosynthesis reading time by up to 50%; a Swedish randomized trial found AI-supported single reading achieved cancer-detection rates comparable to full double reading while meaningfully cutting radiologist workload; one ultrasound-reading system cut false-positive readings by 37.3% and unnecessary biopsy requests by 27.8% while preserving sensitivity, outperforming an average of ten board-certified breast radiologists; and pathology-slide algorithms reached AUCs of 0.98-0.99 for detecting invasive carcinoma and ductal carcinoma in situ, with one open-challenge model reaching an AUC of 0.994 for lymph node metastasis detection, exceeding time-constrained pathologist performance. Digital scoring tools also showed strong correlation with manual grading of estrogen- and progesterone-receptor and Ki-67 status, and early models are now predicting BRCA mutation status and gene-expression-based recurrence risk directly from routine H&E-stained slides [5].

Taken together, this is one of the most thoroughly tested corners of clinical AI - several hundred related devices had already secured US FDA clearance at the time of the review. But most of the strongest results still rest on retrospective rather than prospective data, and the review's authors raise an explicit equity concern: because training data still skews toward majority populations, models built this way risk widening, rather than narrowing, disparities in breast cancer outcomes for underrepresented groups [5].

Prostate cancer and the pathologist's microscope

Grading a prostate biopsy - assigning it a Gleason score that directly shapes treatment decisions - is notoriously subjective, with even experienced pathologists disagreeing with one another on a meaningful share of cases. It is a natural target for AI, and one where researchers have pushed the technology to genuinely large scale, training systems not merely to match a single expert's opinion but to hold up against an entire panel of them.

One team built a fully automated grading system, an extended U-Net architecture, trained using a semi-automatic labelling technique that leveraged pathology-report text rather than the pixel-level manual annotation that usually limits how much data computational pathology can learn from. Tested against a three-pathologist consensus on 535 biopsies, it achieved a quadratic Cohen's kappa of 0.918; in a separate observer study, it outperformed 10 of 15 independent pathologists (kappa 0.854 versus a panel median of 0.819), and it held up reasonably well - kappa of 0.707-0.723 - on an external Swiss dataset scanned and stained differently. Tellingly, most of its errors clustered at the Grade Group 2/3 and 4/5 boundaries, the exact distinctions where human graders also show the lowest agreement [6]. A far larger effort trained a multiple-instance-learning framework on 44,732 whole-slide images from 15,187 patients across prostate, basal cell carcinoma, and breast lymph node datasets - roughly 88 times the pixel volume of the entire ImageNet corpus - reaching AUCs of 0.986, 0.986, and 0.965, respectively (0.991 for prostate once a recurrent neural network aggregation step was added), and at a clinical-grade threshold of 100% sensitivity, could safely exclude 65-75% of slides from manual pathologist review. Accuracy dropped by roughly 3 percentage points when tested on a different scanner, and by roughly 6 points on external consultation slides from other institutions - a smaller degradation, notably, than a comparably trained model saw when tested outside its original well-curated benchmark dataset (AUC falling from 0.930 to 0.727), evidence that large, messy, real-world training data may generalize better than small, pristine research datasets [7].

These are genuinely large, methodologically rigorous studies, and the direction of travel - letting pathologists concentrate on the hardest cases while software clears the routine ones - looks close to practical reality. The remaining hurdle is less about raw accuracy and more about consistency across institutions; it is also worth noting that several authors on the larger study disclosed financial ties to a commercial computational-pathology company, a conflict-of-interest profile worth weighing alongside the results [7].

Lung cancer: catching it on a scan, reading it under the microscope

Lung cancer kills more people worldwide than any other cancer, and low-dose CT screening - while proven to save lives - walks a difficult line between catching disease early and sending too many patients for unnecessary follow-up. AI has been tested on both ends of the lung cancer pathway: reading the screening scan itself, and, more surprisingly, reading the tumor's underlying biology directly off a pathology slide.

Predicting a tumor's genetics from its microscopic appearance, without a genetic test, is a genuine leap beyond simple diagnosis - but it remains early-stage: some of the rarer mutations were tested on only a handful of positive slides, raising overfitting risk, and one clinically important mutation type, ALK rearrangement, could not be predicted by the model at all [8]. It is a preview of where pathology AI is headed more than a tool ready for routine use today.

One widely cited Nature Medicine study built an end-to-end 3D deep-learning model that, when a patient's prior scan was available for comparison, outperformed six practicing radiologists at estimating cancer risk from a chest CT, reducing both false positives and false negatives relative to radiologist-only assessment. A separate study trained an Inception-v3 network on 1,634 whole-slide images from The Cancer Genome Atlas to classify lung tissue as normal, adenocarcinoma, or squamous cell carcinoma, reaching an AUC of roughly 0.97 for distinguishing the two cancer subtypes and 0.99 for tumor-versus-normal discrimination - comparable to three independent thoracic pathologists - and held up with AUCs of 0.83-0.98 across three independent validation cohorts using frozen sections, paraffin-embedded sections, and biopsies. In a methodologically novel second step, the same pipeline was retrained to predict the mutational status of ten frequently mutated genes directly from tissue appearance alone, with no molecular sequencing input; six genes, including STK11, EGFR, and TP53, were predictable with AUCs of 0.733-0.856, and an external validation on a 63-slide EGFR cohort retained an AUC of 0.75 [9].

Digestive health: finding what the naked eye misses

Colonoscopy is remarkably effective at preventing colon cancer, but only if every precancerous polyp gets spotted - and human endoscopists reliably miss a meaningful share of them, particularly small or flat lesions. This is one of the clearest cases in medicine of a task suited to a computer that never tires or loses concentration partway through a long procedure, and the supporting evidence here is unusually strong, including randomized-trial data rather than retrospective comparisons alone.

A comprehensive review of the endoscopy literature reported strong figures across the full care pathway: AI detected Helicobacter pylori infection with meta-analytic sensitivity and specificity of 87% and 86%, spotted early gastric cancer with 92.2% sensitivity, and - pooling 21 randomized trials involving 18,232 patients - cut the adenoma miss rate by roughly 55%; cholangioscopy- and endoscopic-ultrasound-based models reached 90-99% accuracy for malignant biliary strictures [10]. A focused pooled analysis of five of those randomized trials (4,354 patients, four from China and one from Italy) found that AI-assisted colonoscopy detected adenomas in 36.6% of patients versus 25.2% with standard colonoscopy alone (risk ratio 1.44), a benefit the GRADE-rated moderate-quality analysis found held independent of a polyp's size, location, or morphology - evidence the software may genuinely be attending to different visual cues than human endoscopists, rather than simply automating what they already do well [11]. A separate real-time detection system, trained on 8,641 hand-labeled images and validated with VGG19, ResNet50, and comparable architectures, achieved 96.4% accuracy and an AUC of 0.991, with a localization Dice score of 0.83 versus 0.55 for a prior comparable method, and processed video at 98 frames per second - roughly four times the speed needed for genuine real-time use. In a prospective video test, three expert colonoscopists using the system's overlay detected 45 polyps versus 36 without it, and the system independently flagged all 28 polyps the original colonoscopist had actually removed, the same study reported frame-level false-positive rates of roughly 5-7% in its other video-based validation experiments, though a per-procedure false-positive count specific to the three-colonoscopist overlay test was not reported [12].

This is about as strong a case as clinical AI has made for itself outside of ophthalmology: large, randomized evidence, a benefit that doesn't simply replicate existing human performance, and technology demonstrably fast enough for live use. The main caveat is geographic concentration - four of the five randomized trials came from single centers in China, with comparatively low control-arm detection rates, so how the benefit translates to other health systems and baseline accuracy levels remains to be confirmed [11].

Women's health: cervical and gynecologic cancer

Cervical cancer is among the most preventable cancers in medicine, but prevention depends on sustained,

well-resourced screening infrastructure - cytology, HPV testing, colposcopy - that strains health systems even in wealthy countries and is often simply unavailable elsewhere. AI has been tested across nearly the entire gynecologic-oncology pathway, from the initial HPV test through diagnosis, staging, and radiotherapy planning.

A large review citing more than 30 individual studies found a machine-learning decision system could identify women at high risk of CIN2/3-or-worse precancerous changes with 94.32% specificity from extended HPV genotyping, and traced cervical cytology screening from early commercial systems (PAPNET in 1992, ThinPrep in 2004) through modern deep-learning classifiers now reaching accuracy figures from roughly 93% up past 99% depending on dataset; in one prospective study of 703,103 women, AI-assisted cytology increased detection sensitivity for precancerous cells by 5.8 percentage points over manual reading alone [13,14]. A colposcopic AI-assisted diagnostic system agreed with final pathology results 82.2% of the time, versus 65.9% for conventional colposcopy, and pelvic MRI models reached dice coefficients of 0.8-0.98 for tumor segmentation, an AUC up to 0.921 for predicting parametrial invasion, and an AUC of 0.933 for the best-performing hybrid lymph node metastasis model [14]. A separate meta-analysis pooling 12 studies and 793 patients found MRI-based radiomic features predicted lymph node metastasis with 80% pooled sensitivity, 76% specificity, and an AUC of 0.83 - meaningfully outperforming traditional morphology-based imaging (historically just 38-56% sensitivity) - though the same analysis found average methodological quality (Radiomics Quality Score) of only 33.8% of the maximum possible, with no study prospectively designed and just two of the 12 externally validated [15].

The breadth of what's being explored here - screening, diagnosis, staging, even radiotherapy contouring - is genuinely impressive, and individual results, like the 5.8-point sensitivity gain across more than 700,000 women, are backed by real scale. But much of the supporting evidence remains retrospective and single-center, and the field's own quality-scoring exercise suggests the underlying research base should be read as encouraging rather than conclusive, and strictly as a diagnostic aid rather than a replacement for clinician judgment.

Radiotherapy: planning treatment and predicting side effects

Radiotherapy is a balancing act: enough radiation to control a tumor, without so much that it damages the healthy tissue nearby. Predicting which patients will suffer significant treatment-related toxicity - and planning treatment to avoid it - has long relied on clinical experience and dose-volume rules of thumb, which is exactly the kind of pattern-recognition problem machine learning is suited to.

A comprehensive review of 53 studies applying classical machine-learning classifiers - logistic regression, random forest, support vector machines, and artificial neural networks - to radiotherapy toxicity prediction across eight cancer sites (13 head-and-neck studies, 15 lung, 16 prostate, and smaller numbers elsewhere) found reported AUCs ranging widely, roughly 0.6 to 0.9, with a consistent and concerning pattern: the best-performing models, above 0.85, were almost universally derived from small-to-medium cohorts of fewer than 150 patients - a classic signature of a model that may be memorizing its training data rather than learning something generalizable. Genetic factors known to influence radio-sensitivity were incorporated in only two of the 53 studies, despite well-established evidence of their relevance, and the review's authors recommend wider adoption of the TRIPOD reporting checklist to improve comparability of future toxicity-prediction research [16].

The signal here is real but early-stage. Before any of these models could reasonably guide a radiation oncologist's dosing decisions, the field needs larger, multi-institution studies that hold up on patients the model has never seen - the same external-validation gap that recurs across nearly every domain in this review.

Cancer risk, recurrence, and survival: the older toolkit

Before today's deep-learning boom, an earlier generation of machine-learning methods - artificial neural networks, decision trees, support vector machines, and Bayesian networks - was already being applied to a foundational question in oncology: given a patient's clinical, imaging, or genomic data, can software predict who is at risk of cancer, whose cancer will recur, and who will survive it?

A wide-ranging 2015 survey of that earlier literature, organizing dozens of studies into tables by predictive task, found some eye-catching figures: an artificial-neural-network breast cancer risk model built on 48,774 mammographic findings reached an AUC of 0.965 under ten-fold cross-validation; a support-vector-machine model using a 50-gene expression signature predicted breast cancer survival with 97% accuracy in a cohort of 295 patients; an SVM-based recurrence model outperformed both neural networks and Cox proportional-hazards regression at 84.6% accuracy; and one semi-supervised model, tested on a SEER-derived database of 162,500 patients, outperformed both SVM and neural-network approaches at scale. But the same survey was candid about a recurring weak spot: one study claiming 100% accuracy for predicting oral cancer recurrence had been tested on just 86 patients, and the majority of the recurrence- and survival-prediction studies reviewed included fewer than 300 patients in total, with little to no external, multi-institution validation and inconsistent reporting of how predictive features were even selected [17].

This older body of work is best read as the direct methodological ancestor of the deep-learning imaging studies profiled elsewhere in this review - and its central lesson, that small, single-source samples can produce impressively high accuracy scores that don't necessarily generalize, is one the field has had to relearn at every subsequent stage of the technology's development.

Easing the paperwork: AI and clinical documentation

Ask almost any physician what's driving burnout, and chart review comes up fast: the average doctor spends roughly 1.5 hours a day working through patient records, and 75% of burnout cases cite electronic health record workflow as a primary driver. Large language models - the technology behind modern AI chatbots - are now being tested as a way to summarize and simplify that mountain of documentation.

Early case studies are encouraging: NLP summarization tools such as BARTcnn and GPT-3's DaVinci model condensed lengthy neuroradiology reports to under 20% of their original length, and a Nature Medicine evaluation found that GPT-4, adapted with in-context learning, outperformed physicians on some summarization and hallucination-avoidance metrics. A separate soft-prompting framework, SPeC, has focused specifically on improving the consistency of language-model outputs from one note to the next, since inconsistency remains one of the biggest practical obstacles to trusting an AI-written summary in a clinical setting [18].

The promise is real, but so are the risks: median outpatient note length has grown 60% since 2009, with note redundancy exceeding 50%, straining the context window of even large language models; training data has repeatedly been shown to underrepresent certain demographic groups - pulse-oximeter measurement bias by skin tone is one well-documented example cited alongside this issue - and informed-consent gaps in large-scale secondary data use remain unresolved, echoing the 2017 controversy over DeepMind's NHS Streams data-sharing arrangement. Researchers have also demonstrated that some language models are vulnerable to membership-inference attacks reaching up to 80% precision, meaning they can, in principle, be manipulated into leaking fragments of the patient data they were trained on. None of this is disqualifying, but it means AI-assisted documentation is being adopted alongside - not instead of - careful, ongoing attention to privacy and fairness [18].

Managing chronic disease at the population scale

Diabetes, hypertension, and chronic obstructive pulmonary disease (COPD) together account for an enormous share of global healthcare spending, and a growing body of research looks at AI not as a diagnostic tool for one patient, but as infrastructure for managing entire populations of them. Rather than testing a single algorithm, a recent bibliometric analysis mapped the research landscape itself, screening 758 publications and retaining 341 for review, spanning 775 institutions across 55 countries from the field's inception through August 2024 [19].

The picture that emerged was lopsided: the United States produced 50.7% of all publications and carried the highest total citation score, followed distantly by China at 11.1%, with the research clustering into four dominant themes - diagnosis, care delivery, telemedicine, and underlying technology - and diabetes, hypertension, and COPD as the most-studied conditions. "Machine learning" was the single most prominent emergent keyword among the most recent papers analyzed [19].

Two findings stand out as cautionary notes. First, even as publication volume has risen near-exponentially since 2013, the field's total local citation score - a proxy for field-internal academic impact - has actually been declining in recent years, suggesting growth in quantity has outpaced growth in methodological rigor or influence. Second, contributions from Africa and South America were minimal in both output and international collaboration, a global equity gap worth remembering whenever a chronic-disease AI tool built almost entirely on US and Chinese data is proposed for use in a different health system [19].

AI in your pocket: symptom checkers and cycle-tracking apps

Long before hospitals adopted AI, patients were already using consumer versions of it - symptom-checker websites and apps, and menstrual cycle-tracking apps used by hundreds of millions of people worldwide. These tools sit outside clinical settings entirely, and the research suggests they have been held to a noticeably lower validation standard than the imaging systems reviewed above.

One widely cited BMJ audit tested 23 publicly available symptom checkers against 45 standardized patient vignettes - 770 diagnosis evaluations and 532 triage evaluations in total - and found that the correct diagnosis was listed first in only 34% of evaluations (58% within the top 20 possible differentials), with appropriate triage advice given in 57% of cases; performance was strongly urgency-dependent, with triage accuracy reaching 80% for genuinely emergent presentations but falling to just 33% for cases where self-care was actually appropriate, meaning the tools systematically over-refer low-acuity symptoms to formal care [20]. A separate review of menstrual cycle-tracking applications, screening 150 articles and retaining 49 for full review across four research domains, found calendar-based ovulation-day prediction accuracy ranging from 17% to 89% depending on method; the FDA-cleared Natural Cycles app, which incorporates basal body

temperature, achieved a typical-use Pearl Index of 7.0 contraceptive failures per 100 person-years versus 24 for other fertility-awareness methods generally - but in one dataset of 1.4 million eligible cycles, ovulation could not be assigned at all in 665,603 of them, three-quarters of which were missing temperature readings for at least half the cycle. No app assessed against a standardized quality checklist scored above 13 of 15 possible points. This scale of missing data is clinically relevant: Natural Cycles' own published typical-use Pearl Index (6.9-7.0) sits far above its perfect-use Pearl Index (~1.0), and this gap is understood to be driven substantially by exactly this kind of incomplete real-world logging rather than by algorithmic error alone. When temperature data is missing, the app's fallback logic is inherently more conservative and closer to calendar-based estimation, which is less reliable at pinpointing ovulation, meaning the aggregate typical-use Pearl Index likely averages over a wide range of true individual risk, from consistent loggers approaching perfect-use performance to sparse loggers whose real risk may approach that of calendar-based methods generally. None of the reviewed literature reports failure rates stratified by data completeness, which represents a meaningful gap for assessing the app's real-world safety profile for lower-adherence users specifically [21].

It is worth noting the symptom-checker research predates today's far more capable large-language-model chatbots and may not reflect what those newer tools can do. But both studies point to the same underlying issue: consumer health AI has been validated far less rigorously than its clinical imaging counterparts, and both literatures show real gaps in whose data - and which users - the underlying research actually reflects, with cycle-tracking studies in particular over-representing white, European, and higher-education users.

Designing new drugs from scratch

Traditional drug discovery is slow and staggeringly expensive, typically 10 to 20 years and up to $2.6 billion to bring a single new medicine to patients. Generative AI, the same broad class of technology behind image and text generation, has been tested as a way to compress that timeline by designing entirely novel drug candidates rather than simply screening existing compound libraries.

In one landmark 2019 demonstration, researchers combined a generative autoencoder, a tensor-train decomposition step, and a reinforcement-learning loop rewarded via a self-organizing map, to design novel small-molecule inhibitors of DDR1, a kinase implicated in fibrosis. Starting from nothing, the full pipeline - generating candidate structures, synthesizing them in a lab, and testing them in cells - produced six real compounds in just 21 days. Four of the six showed measurable biological activity against the target; two, with IC50 values of 10.3 and 5.8 nanomolar, performed well enough in cell-based assays to warrant further study; and the strongest candidate went on to show favorable pharmacokinetics in mouse testing [22].

This is one of the earliest published examples of an AI-designed drug candidate making it all the way to live biological testing, and the speed is genuinely striking when measured against a decade-plus conventional baseline. But it remains a single successful proof of concept on one drug target, with a handful of candidate molecules and no head-to-head comparison to how a conventional discovery effort would have performed on the identical problem - and three of the paper's authors disclosed financial stakes in the generative-AI technology used, a detail worth weighing alongside an otherwise compelling result [22].

Rules, training, and leadership: the human infrastructure behind the technology

None of the tools above will do much good if regulators can't approve them responsibly, if clinicians aren't trained to question their output, or if hospital culture resists using them at all. A cluster of research outside the clinical-trial literature looks at exactly this - the surrounding scaffolding that determines whether AI actually reaches patients.

US regulators have proposed a Predetermined Change Control Plan framework specifically for AI/ML-based software that continues learning after deployment, requiring manufacturers to pre-specify - through a defined SaMD Pre-Specifications (SPS) and Algorithm Change Protocol (ACP) - exactly how a system is permitted to change over time and how those changes will be monitored, rather than requiring a fresh review for every update [23]. Medical educators have separately proposed that tomorrow's clinicians need three specific new competencies: functioning as an Evaluator (judging when an AI tool's use is appropriate), an Interpreter (understanding where its errors are likely to originate), and a Communicator (explaining its output to a patient in plain language) - citing Microsoft's 2016 Tay chatbot, which began generating offensive content within 24 hours of public release, as a cautionary illustration of unsupervised AI behavior [24]. A pilot of adaptive e-learning software in medical education has also been reported, although the evidence base for such interventions remains limited and largely uncontrolled [25]. In a large, methodologically separate survey of 2,813 physicians at a major US academic medical center (72.2% response rate), a study of general leadership dynamics that does not itself examine AI or digital health tools, but whose findings this review applies by analogy, researchers found that supervisor leadership quality, scored on a validated 12-to-60-point scale, was strongly associated with both burnout and job satisfaction: each one-point increase in a leader's score was associated with a 3.3% reduction in the odds of burnout among their physicians and a 9.0% increase in satisfaction, with leadership scores alone statistically explaining roughly 11% of the variance in burnout and 47% of the variance in satisfaction [26].

Put together, these findings make a point that's easy to miss amid diagnostic-accuracy headlines: even where an algorithm performs beautifully in a research paper, getting it safely into everyday use is fundamentally a matter of regulation, training, and local leadership - a people problem, in other words, at least as much as it is a technical one.

Conclusions

Step back from the individual studies, and a clear hierarchy of evidentiary maturity emerges. Imaging-based AI, reading eyes, mammograms, pathology slides, and colonoscopy video, is the most mature part of the field, regularly matching or edging out expert human performance and, increasingly, evaluation at multi-institution scale. Radiotherapy planning, gynecologic and cervical cancer applications, and oncologic risk prediction are moving in the same direction but remain a step behind, held back by smaller patient cohorts and limited prospective or external validation. Consumer-facing tools such as symptom checkers and cycle trackers lag furthest in methodological rigor, while workflow-focused AI, including chart summarization and population-health analytics, remains useful but is entangled with unresolved questions of bias, consent, and data quality. Generative AI for drug design remains at an early proof-of-concept stage, with compelling individual demonstrations but no established track record of reproducible success.

The clearest takeaway is that the barriers standing in the way are, in most domains, no longer primarily about whether the underlying algorithms work - they often perform at levels competitive with trained specialists. The harder task is demonstrating that this performance holds outside the original hospital or dataset, while building the regulatory and educational scaffolding needed for responsible deployment and strengthening the human systems - leadership, training, and institutional trust - required to use these tools effectively. On the evidence assembled here, medicine's algorithmic future looks genuinely promising, but it is not finished yet.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Divyanshi Baijal, Himanshi Khattar, Piyush Kumar

Acquisition, analysis, or interpretation of data: Divyanshi Baijal, Himanshi Khattar, Piyush Kumar

Drafting of the manuscript: Divyanshi Baijal, Himanshi Khattar, Piyush Kumar

Critical review of the manuscript for important intellectual content: Divyanshi Baijal, Himanshi Khattar, Piyush Kumar

Supervision: Piyush Kumar

Disclosures

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

Claude AI and ChatGPT were used for data compilation.

References

1. Shimizu E, Tanji M, Nakayama S, et al.: AI-based diagnosis of nuclear cataract from slit-lamp videos . Sci Rep. 2023, 13:22046. 10.1038/s41598-023-49563-7
2. Gulshan V, Peng L, Coram M, et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016, 316:2402-10. 10.1001/jama.2016.17216
3. Li Z, He Y, Keel S, Meng W, Chang RT, He M: Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. Ophthalmology. 2018, 125:1199-206. 10.1016/j.ophtha.2018.01.023
4. McKinney SM, Sieniek M, Godbole V, et al.: International evaluation of an AI system for breast cancer screening. Nature. 2020, 577:89-94. 10.1038/s41586-019-1799-6
5. Ahn JS, Shin S, Yang SA, et al.: Artificial intelligence in breast cancer diagnosis and personalized medicine . J Breast Cancer. 2023, 26:405-35. 10.4048/jbc.2023.26.e45

6. Bulten W, Pinckaers H, van Boven H, et al.: Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. Lancet Oncol. 2020, 21:233-41.
7. Campanella G, Hanna MG, Geneslaw L, et al.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med. 2019, 25:1301-9. 10.1038/s41591-019-0508-1
8. Coudray N, Ocampo PS, Sakellaropoulos T, et al.: Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med. 2018, 24:1559-67. 10.1038/s41591-018-0177-5
9. Ardila D, Kiraly AP, Bharadwaj S, et al.: End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med. 2019, 25:954-61. 10.1038/s41591-019-0447-x
10. Ali H, Muzammil MA, Dahiya DS, et al.: Artificial intelligence in gastrointestinal endoscopy: a comprehensive review. Ann Gastroenterol. 2024, 37:133-41. 10.20524/aog.2024.0861
11. Hassan C, Spadaccini M, Iannone A, et al.: Performance of artificial intelligence in colonoscopy for adenoma and polyp detection: a systematic review and meta-analysis. Gastrointest Endosc. 2021, 93:77-85.e6. 10.1016/j.gie.2020.06.059
12. Urban G, Tripathi P, Alkayali T, Mittal M, Jalali F, Karnes W, Baldi P: Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. Gastroenterology. 2018, 155:1069-1078.e8. 10.1053/j.gastro.2018.06.037
13. Khattar H, Goel R, Kumar P: Artificial intelligence in gynaecological malignancies: perspectives of a clinical oncologist. Cureus. 2023, 15:e45660. 10.7759/cureus.45660
14. Hou X, Shen G, Zhou L, Li Y, Wang T, Ma X: Artificial intelligence in cervical cancer screening and diagnosis. Front Oncol. 2022, 12:851367. 10.3389/fonc.2022.851367
15. Li L, Zhang J, Zhe X, Tang M, Zhang X, Lei X, Zhang L: A meta-analysis of MRI-based radiomic features for predicting lymph node metastasis in patients with cervical cancer. Eur J Radiol. 2022, 151:110243. 10.1016/j.ejrad.2022.110243
16. Isaksson LJ, Pepa M, Zaffaroni M, et al.: Machine learning-based models for prediction of toxicity outcomes in radiotherapy. Front Oncol. 2020, 10:790. 10.3389/fonc.2020.00790
17. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI: Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J. 2015, 13:8-17. 10.1016/j.csbj.2014.11.005
18. Lee C, Vogt KA, Kumar S: Prospects for AI clinical summarization to reduce the burden of patient chart review. Front Digit Health. 2024, 6:1475092. 10.3389/fdgth.2024.1475092
19. Pan M, Li R, Wei J, et al.: Application of artificial intelligence in the health management of chronic disease: bibliometric analysis. Front Med (Lausanne). 2024, 11:1506641. 10.3389/fmed.2024.1506641
20. Semigran HL, Linder JA, Gidengil C, Mehrotra A: Evaluation of symptom checkers for self diagnosis and triage: audit study. BMJ. 2015, 351:h3480. 10.1136/bmj.h3480
21. Schantz JS, Fernandez CS, Anne Marie ZJ: Menstrual cycle tracking applications and the potential for epidemiological research: a comprehensive review of the literature. Curr Epidemiol Rep. 2021, 8:9-19. 10.1007/s40471-020-00260-3
22. Zhavoronkov A, Ivanenkov YA, Aliper A, et al.: Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat Biotechnol. 2019, 37:1038-40. 10.1038/s41587-019-0224-x
23. US Food and Drug Administration. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD): discussion paper and request for feedback. (2019). Accessed: July 7, 2026: https://apo.org.au/node/228371.
24. Lomis K, Jeffries P, Palatta A, Sage M, Sheikh J, Sheperis C, Whelan A: Artificial intelligence for health professions educators. NAM Perspect. 2021, 2021: 10.31478/202109a
25. Sharma N, Doherty I, Dong C: Adaptive learning in medical education: the final piece of technology enhanced learning?. Ulster Med J. 2017, 86:198-200.
26. Shanafelt TD, Gorringe G, Menaker R, et al.: Impact of organizational leadership on physician burnout and satisfaction. Mayo Clin Proc. 2015, 90:432-40. 10.1016/j.mayocp.2015.01.012