6

Medicine

Overview

AI in medicine advanced on multiple fronts in 2025, but strong model performance has not consistently translated into real-world clinical impact. In molecular biology, where understanding protein behavior is fundamental to drug development, AI-driven protein research continued to grow, and smaller, more specialized models matched or outperformed larger general-purpose systems on protein structure prediction, genomics, and drug discovery. On clinical reasoning tasks, leading AI models now score higher than most physicians on structured clinical evaluations, yet nearly half of clinical AI studies still rely on simulated scenarios rather than real patient data.
The tools gaining traction in practice are those that support clinicians' existing workflows, such as ambient AI scribes and sepsis prediction systems. Authorizations from the U.S. Food and Drug Administration (FDA) for AI-enabled medical devices increased, but clinical evidence continues to lag behind. Patients are also encountering AI-generated health information directly in search results, often before they speak to a clinician, with even less oversight or vetting than the tools moving through formal regulatory channels. AI's impact on medicine is clear, but realizing it at scale will require clinical evidence, governance, and ethical frameworks.

Contents

Chapter Highlights

1. **In molecular biology, smaller models are outperforming larger ones.** MSAPairformer, a 111-million-parameter protein language model, outperformed previous leading methods on the benchmark, ProteinGym; and GPN-Star, a 200-million-parameter genomics model, outperformed a model with 40 billion parameters.

2. **Virtual cell models emerged as a new frontier in 2025, with major releases including Evo 2 from the Arc Institute, STATE, and DeepMind's AlphaGenome.** These models aim to predict cellular responses to drugs and genetic perturbations without running wet-lab experiments, though current systems still require experimental validation.

3. **Like other areas of AI, biological model development is increasingly bottlenecked on data rather than architecture.** With cofolding models now representing all structure types in the Protein Data Bank, 2025 saw a turn toward distilled datasets of AI-predicted structures and training on combined experimental data sources, expanding training sets from hundreds of thousands of entries to tens of millions.

4. **AI tools that automatically generate clinical notes from patient visits saw broad adoption in 2025.** Across multiple hospital systems, physicians reported they were spending up to 83% less time writing notes, experiencing significant reductions in burnout, with one hospital system reporting a 112% return on investment.

5. **The FDA authorized 258 AI medical devices in 2025, most through pathways that do not require new clinical trials.** The vast majority entered the market via device-modification pathways that rely on existing safety and efficacy evidence rather than new randomized trials, with only 2.4% of devices with clinical studies supported by randomized trial data.

6. **A multi-agent AI system scored 85.5% on complex published case studies, versus 20% for unaided physicians.** Microsoft's AI Diagnostic Orchestrator, paired with OpenAI's o3, was tested on challenging cases drawn from the medical literature against physicians working without their usual tools. Multi-agent frameworks more broadly have shown diagnostic accuracy gains of 7% to over 60% over single-agent baselines.

7. **AI-generated summaries now appear at the top of 84% to 92% of health-related Google searches.** Symptom and common health questions trigger an AI Overview 92% of the time, followed by treatment and condition queries. These summaries are now a routine feature of health information searches, shaping the initial interpretation of users' questions.

8. **Ethics discussion in medical AI publications more than doubled in 2025, but the conversation is narrow.** Governance dominates the discourse, while algorithm accountability, biosecurity, and global health equity remain underexplored.

9. **Research interest in medical digital twins is growing fast, and where rigorous trials exist, early results are promising.** In a randomized trial of 150 diabetes patients, 71% achieved healthy blood sugar levels over one year while safely reducing their medications.

6.1 The Central Dogma

AI models for molecular biology span the pathway from gene sequence to protein structure to therapeutic design. This section tracks advances in protein language models, structure prediction, protein design, virtual cell models, and multimodal foundation models for biomedical discovery. The analysis draws on PubMed publication counts, benchmark evaluations including ProteinGym and FoldBench, and model release data from 2024 and 2025. A recurring pattern across these areas is the tension between scale and specialization. In several areas, smaller or more targeted models matched or outperformed larger general-purpose systems.

Research Volume

AI-driven protein research grew approximately 71% between 2024 and 2025 (Figure 6.1.1). Total publications across four categories—function prediction, protein structure prediction, protein-drug interactions, and synthetic protein design—rose from 2,259 in 2024 to 3,855 in 2025. Protein-drug interactions represented the largest share of output in both years, accounting for 49.9% of papers in 2024 and rising to 54.4% in 2025. Protein structure prediction constituted the second-largest category in 2024 at 28.7%, though its relative share declined to 23.9% in 2025. Function prediction and synthetic protein design each remained comparatively stable, with function prediction increasing from 9.7% to 10.4% and synthetic protein design decreasing slightly from 11.7% to 11.3%. The shift in relative share toward protein-drug interactions, even as absolute counts grew across all categories, may reflect maturing structure prediction methods and growing interest in therapeutics applications. Publications focused specifically on AI for drug discovery have followed a similar upward trajectory (Figure 6.1.2).

Number of AI-driven protein research publications, 2024 vs. 2025

Source: RAISE Health, 2026 | Chart: 2026 AI Index report

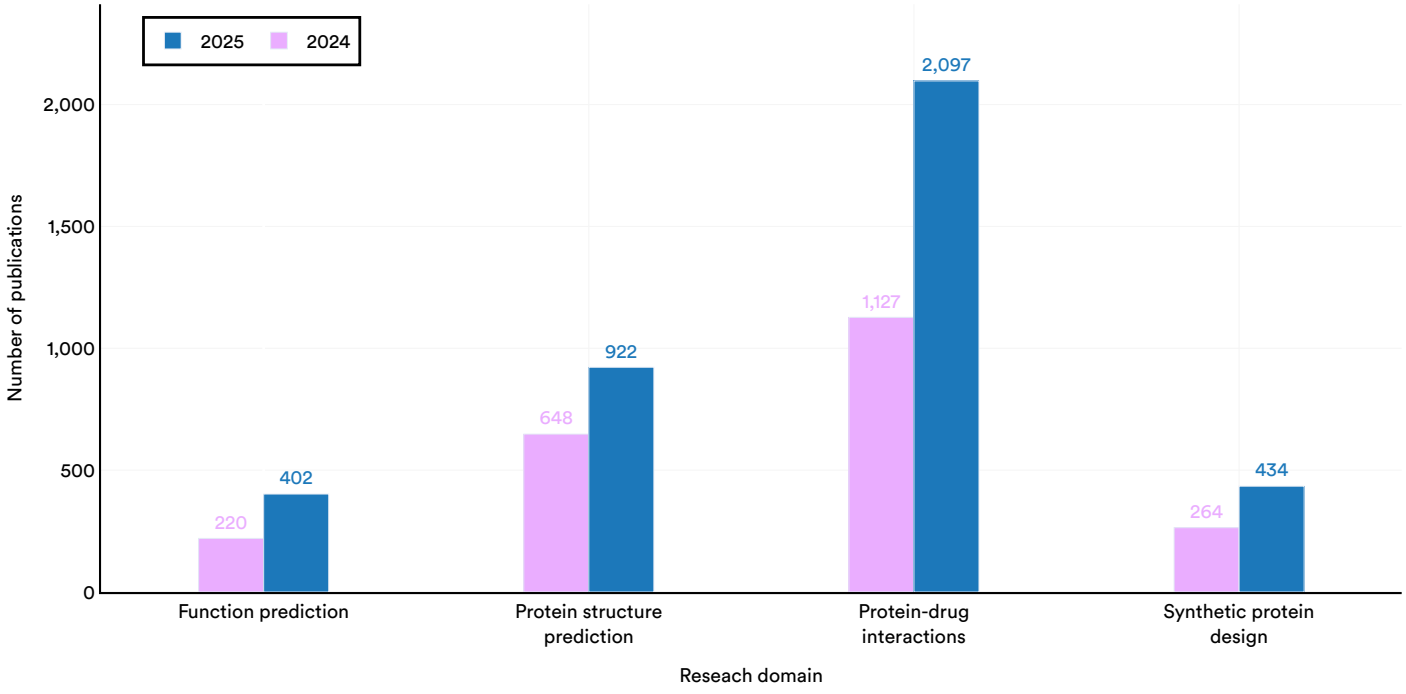


Figure 6.1.1

Number of publications on AI for drug discovery, 2018–25
Source: RAISE Health, 2026 | Chart: 2026 AI Index report

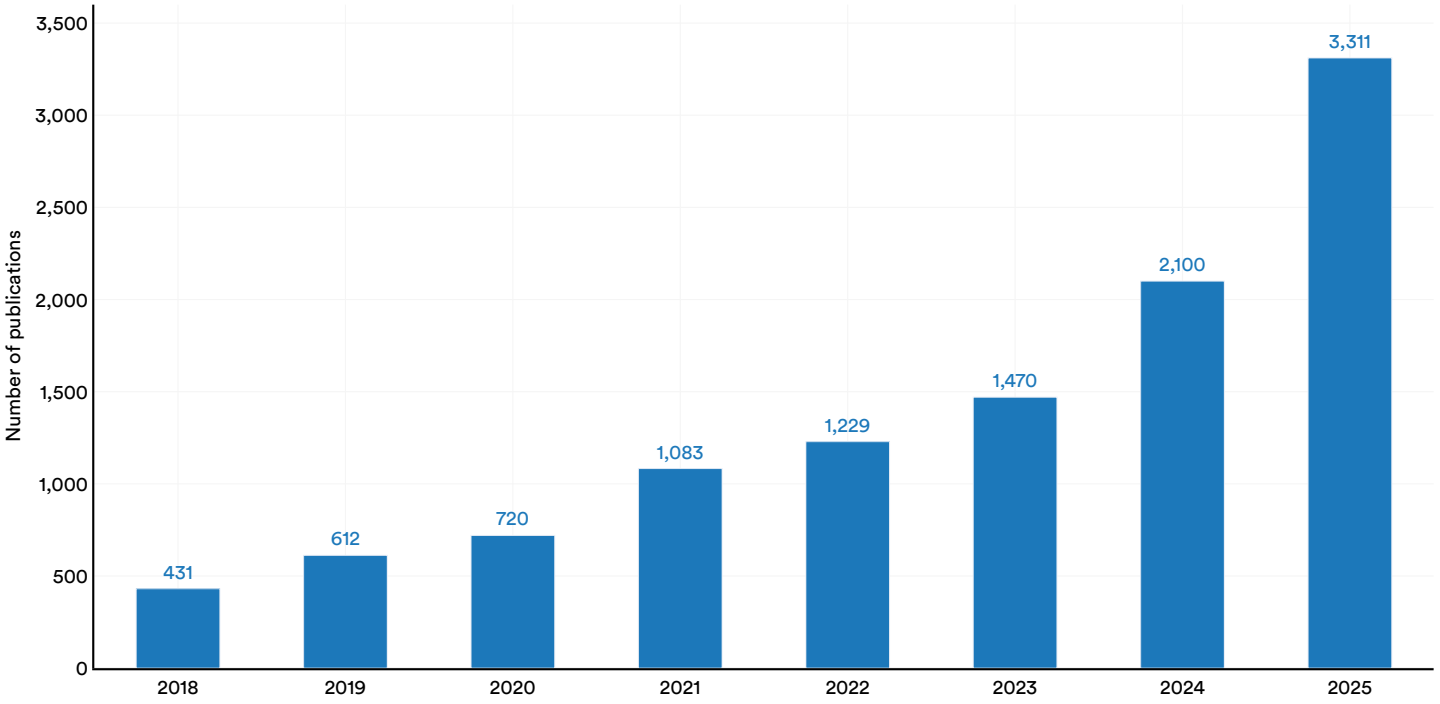


Figure 6.1.2

Public Datasets

Molecular and Cellular Biology

Demand for training data has continued to grow as AI models have gained further adoption in biology. Rapidly collecting new biological data is typically time-consuming and expensive. In 2025, biological AI models increasingly trained on multiple datasets with different types of experimental measurements. Several cofolding methods (where two or more molecules are modeled simultaneously), for example, began training on both structural data from the Protein Data Bank (PDB) and experimental small-molecule binding affinity measurements from repositories such as PubChem, ChEMBL, and BindingDB.

The scale of publicly available biological datasets has grown by several orders of magnitude since the PDB's founding in 1971, though direct size comparisons across databases should be interpreted with caution. The unit of measurement differs by source: An "entry" may represent an experimentally solved protein structure, a bioactivity measurement, a gene sequence, or a single-cell observation.

Other models are trained on synthetically generated data. AlphaFold 3 and its open-source replications, including Boltz-2 and OpenFold3, all train on "self-distillation" datasets of predicted protein structures from AlphaFold 2 that have been filtered for quality. Meta FAIR released Open Molecules 2025 (OMol25), a dataset of over 100 million quantum mechanics calculations of molecules.

New experimental datasets also debuted in 2025. Tahoe-100M, the largest publicly released, single-cell sequencing dataset, contains measurements from over 50 cancer cell types exposed to more than 1,100 drugs. BaseData features over 9.8 billion genes obtained through metagenomic mining (Figure 6.1.3).

Size of public datasets for molecular and cellular biology
Source: RAISE Health, 2026 | Chart: 2026 AI Index report

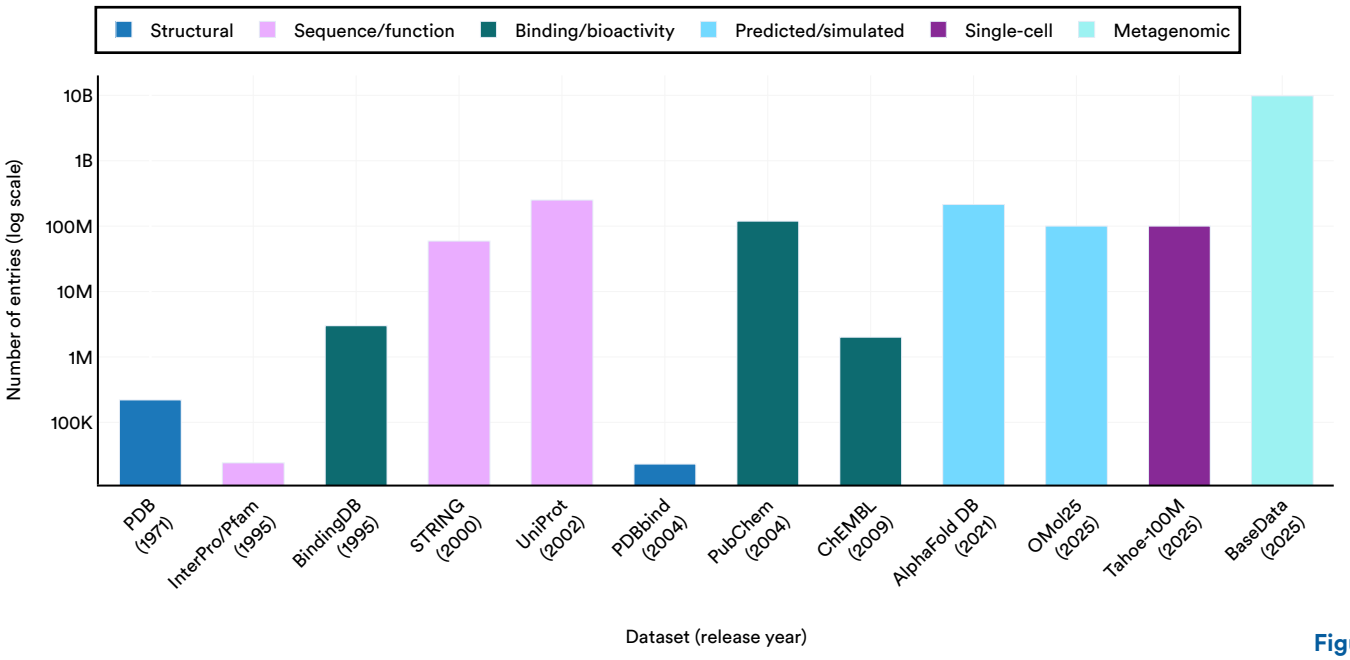


Figure 6.1.3

Data for Biomedical Vision-Language Models

Training biomedical vision-language models requires large repositories of images and captions that can be transformed into a continual pretraining dataset. In the general domain, data scaling is often considered a mature or saturated research direction, but this does not appear to hold in the biomedical setting. Newer datasets extend beyond a single specialty and incorporate a broader range of modalities and biomedical domains (Figure 6.1.4).

Size of select biomedical datasets at release, 2020–25
Source: RAISE Health, 2026 | Chart: 2026 AI Index report

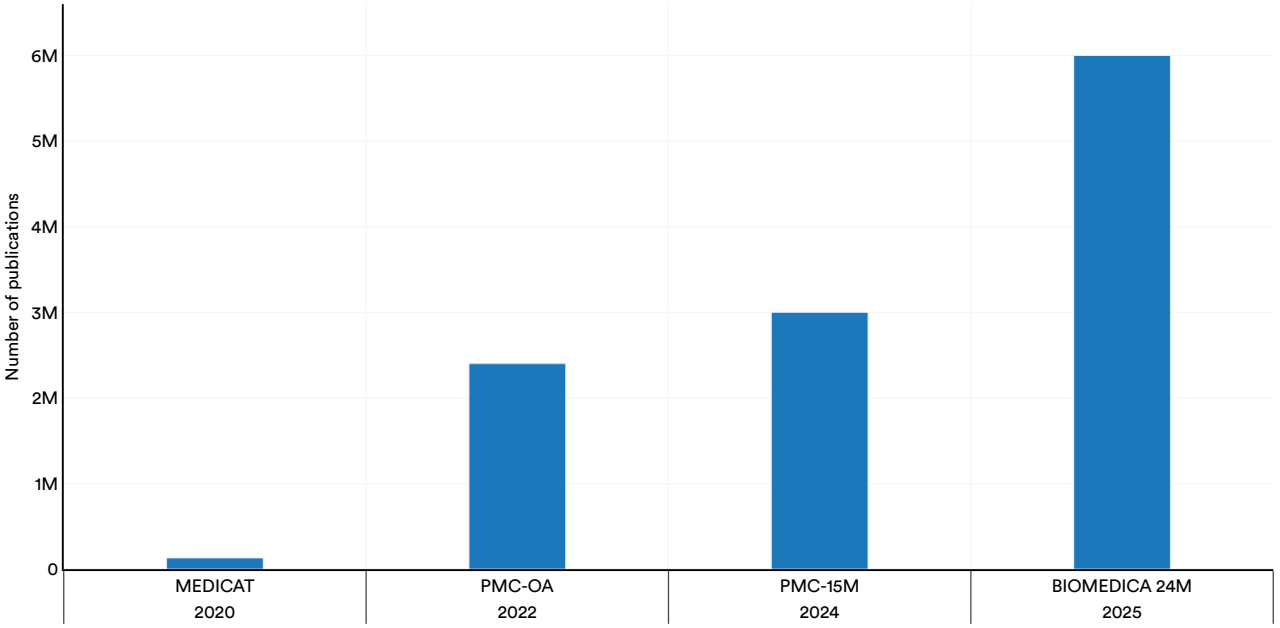


Figure 6.1.4

Sequence-Based Models: Protein Language Models

The trend in protein language models (PLMs) shifted in 2025 from scaling model size to improving model efficiency and specialization. In 2024, efforts culminated in the 98-billion-parameter ESM3. In 2025, the focus turned to smaller architectures trained on curated data or augmented with retrieval methods (Figure 6.1.5).

ProteinGym is a comprehensive benchmark suite for protein fitness prediction and design, comprising over 250 standardized deep mutational scanning assays with millions of mutated sequences and curated clinical datasets with expert-annotated mutation effects. MSAPairformer, a 111 million-parameter model trained on multiple sequence alignments and physical constraints, surpassed previous state-of-the-art methods at a fraction of the training and parameter budget (Figure 6.1.6). The Profluent E1 series similarly set new performance standards by combining a smaller model with a retrieval augmented generation (RAG) approach. While certain tasks still benefit from larger models, others—such as predicting cellular localization—appear to vary more with training method and data than with parameter count.

Size of protein structure prediction models, 2021–25

Source: RAISE Health, 2026 | Chart: 2026 AI Index report

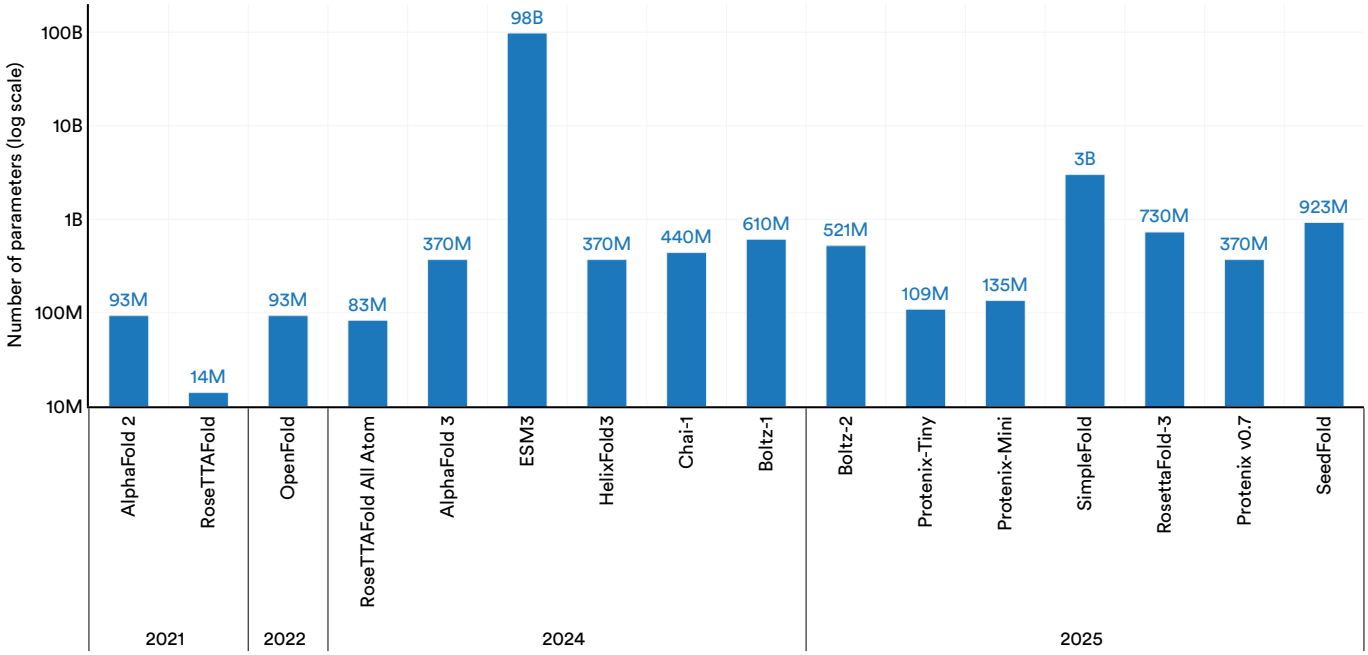


Figure 6.1.5

Performance of protein language models on ProteinGym, 2021–25

Source: RAISE Health, 2026 | Chart: 2026 AI Index report

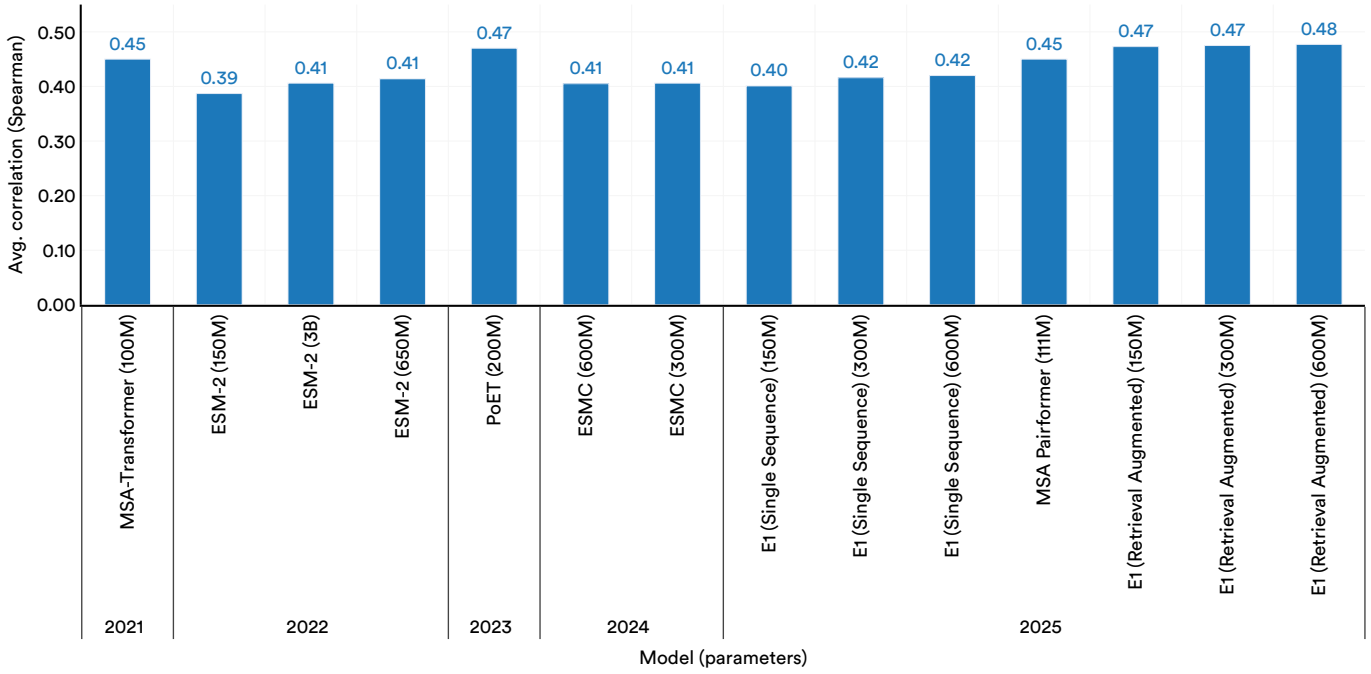


Figure 6.1.6

Beyond benchmark performance, PLMs have also become more task-specific. The ESM-C series demonstrated that smaller models geared toward a single task, such as representation learning, could be successful without the full feature set of the ESM3 family. A fine-tuned ProGen model (6 billion parameters) was used to design a novel CRISPR-Cas protein, OpenCRISPR-1, which demonstrated improved specificity relative to standard SpCas9.

Structure Prediction and Cofolding Models

Multiple open-source structure prediction models were released in 2025, inspired by the architecture of AlphaFold 3. These models tackle the task of "cofolding"—predicting the three-dimensional structures formed by combinations of proteins, nucleic acids, drugs, and other biomolecules. While AlphaFold 3 retains a performance advantage on certain tasks, most cofolding models have demonstrated similar performance across protein structure and biomolecular complex modeling tasks. Some, including the Boltz series and OpenFold3, are released under commercially permissive licenses.

Because cofolding models can now represent all structure types available in the PDB, further performance gains will likely require new data sources or deeper extraction of signal from existing data. One strategy is the use of "distilled" datasets, in which AI-predicted protein structures supplement experimentally determined ones, scaling training datasets from hundreds of thousands of entries to tens of millions. Boltz-2 illustrates a complementary approach—training on both structural information from the PDB and experimental binding affinity measurements—to combine specialization and scale (Figure 6.1.7). The four leading cofolding models share a common foundation of experimental and distilled structural data, but they diverge in their use of additional sources such as molecular dynamics simulations, binding affinity measurements, and RNA structure.

Training data sources for cofolding models, 2025

Model	Structural data (experimental)	Structural data (distilled)	Molecular dynamics	Binding affinity	RNA structure
AlphaFold 3	✓	✓		✓	
Boltz-2	✓	✓	✓	✓	
SimpleFold	✓	✓	✓		
OpenFold3	✓	✓		✓	✓

Figure 6.1.6[1]

Similar to trends in other areas of AI, bigger models have not necessarily translated to better performance in protein structure prediction. Following the release of AlphaFold 3, subsequent models have converged on a similar parameter scale rather than continuing to grow (Figure 6.1.8). FoldBench is a benchmark that tests whether a model can correctly predict how a small molecule, such as a drug candidate, physically binds to a target protein. AlphaFold 3's performance on FoldBench has yet to be significantly surpassed even though several larger models have been released since (Figure 6.1.9). These results suggest that data, rather than model size, is an important bottleneck in protein structure prediction.

1 Rows represent individual cofolding models. Columns indicate whether each model incorporated a given data type during training, including experimentally determined and distilled protein structures, molecular dynamics simulations, binding affinity measurements, and RNA structural data. A check mark indicates that the data type was used.

Size of protein structure prediction models, 2021–25

Source: RAISE Health, 2026 | Chart: 2026 AI Index report

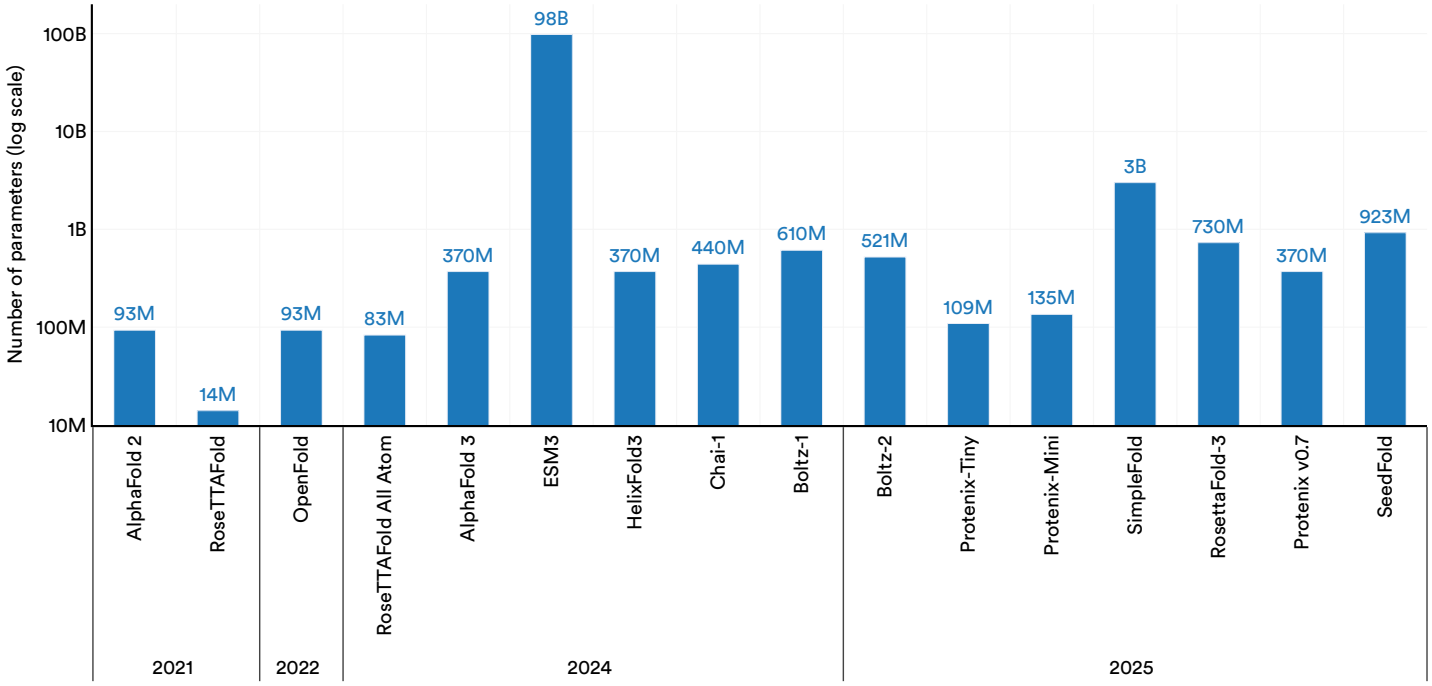


Figure 6.1.8

FoldBench: protein cofolding performance, 2024–25

Source: RAISE Health, 2026 | Chart: 2026 AI Index report

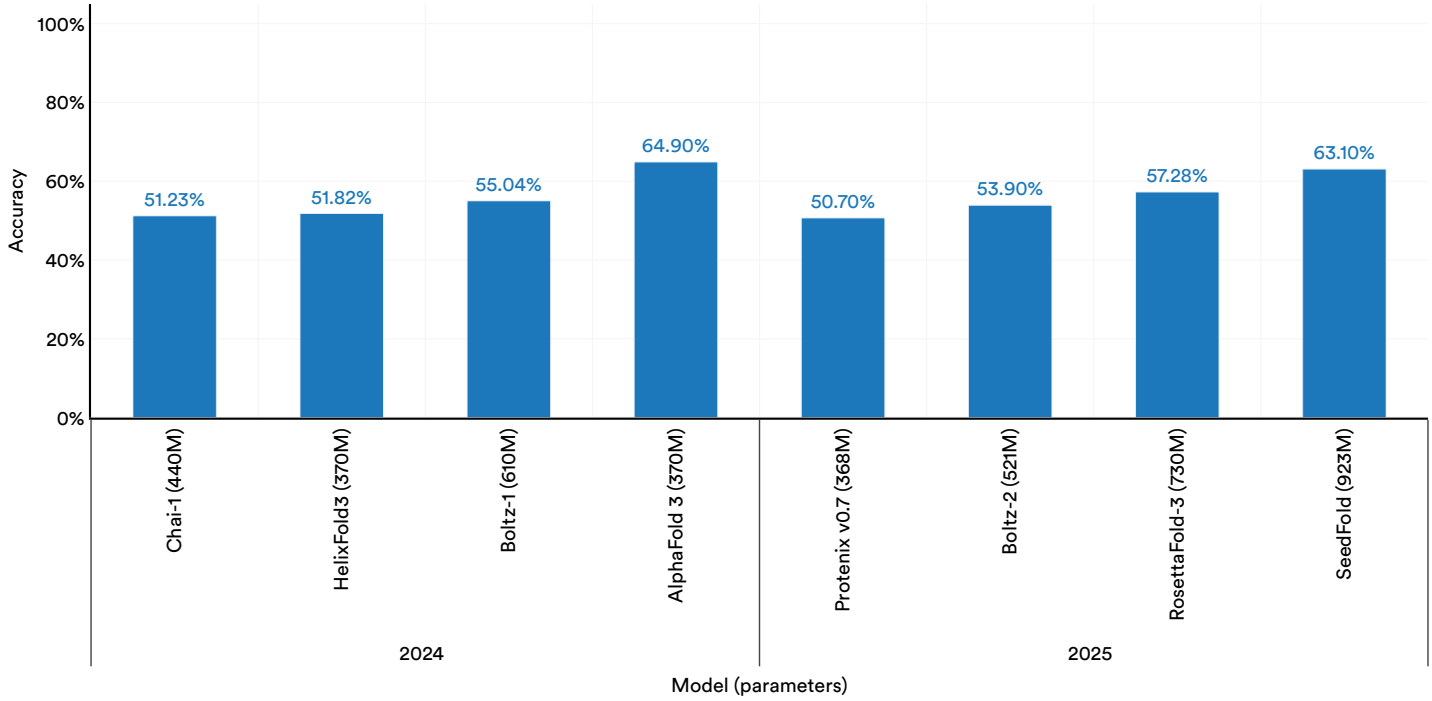


Figure 6.1.9

Protein Design and Generative Models for Therapeutics

Advances in cofolding have enabled a new generation of generative models for protein design, including methods for designing antibodies, nanobodies, and peptides. Approaches range from workflows built around existing structure prediction methods (e.g., BindCraft, Germinal) to models directly trained for generating binders (e.g., RFDiffusion, BoltzGen).

A protein design challenge hosted by Adatpyv Bio in 2025 provided a controlled comparison. Multiple methods were tested on the task of designing a binder targeting Nipah virus. Of the thousand-plus designs tested, 99 proteins were confirmed to bind, and none neutralized the targeted protein. The specialized method Mosaic, which combines multiple tools with expert tuning, outperformed general-purpose approaches (Figure 6.1.10).

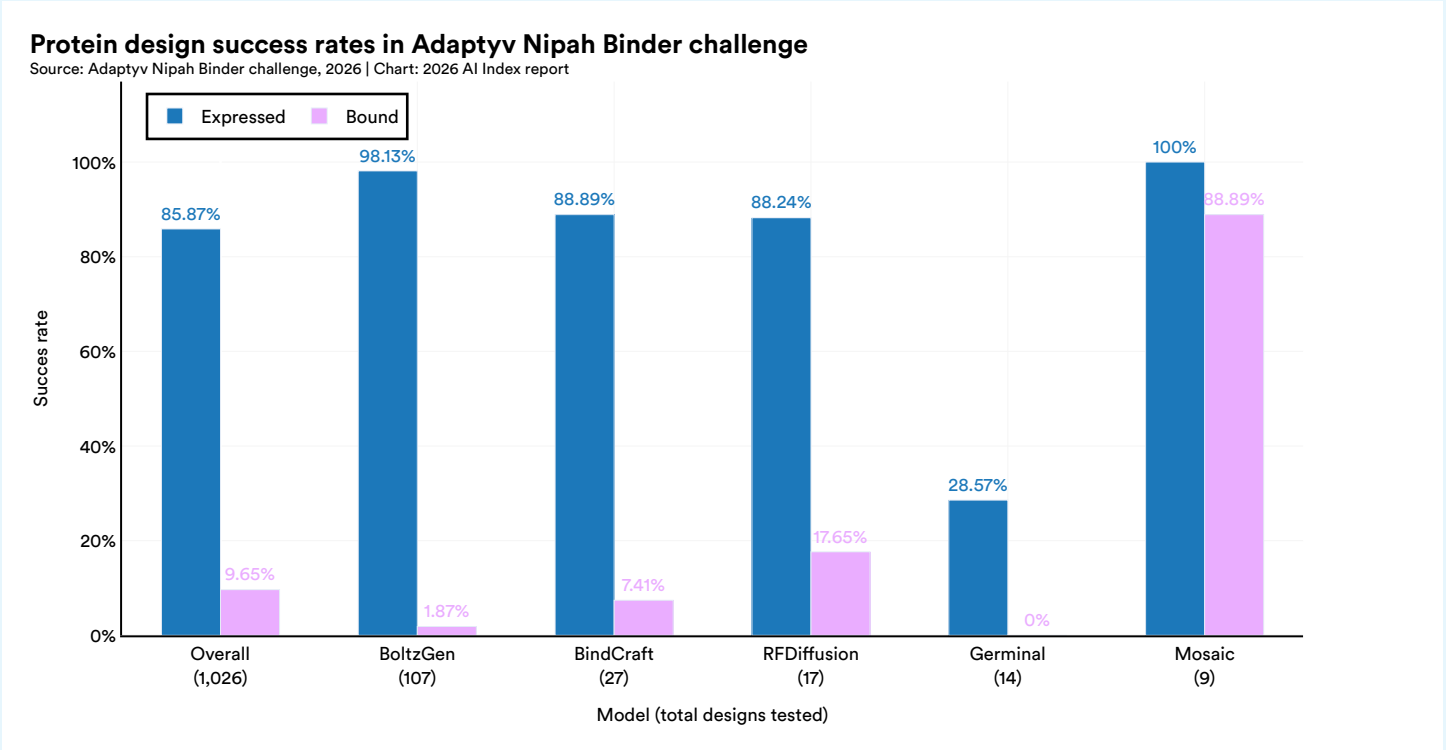


Figure 6.1.10

Virtual Cell Models and Genomic Foundation Models

Research on "virtual cell" models—AI systems that model cellular states and responses to stimuli—increased substantially in 2025, as reflected in growing PubMed publication counts (Figure 6.1.11). Notable releases included Evo 2, a DNA language model from the Arc Institute; STATE, a perturbation-response model; and AlphaGenome, a multimodal model from DeepMind.

However, current virtual cell and genomic foundation models still lag behind smaller, task-specific models on several benchmarks. GPN-Star, a 200-million-parameter model focused on functional and regulatory genomics, outperformed Evo 2 (40B parameters) on multiple variant effect prediction tasks (Figure 6.1.12). These results suggest that scale alone is not yet sufficient, and that training method and data curation remain important determinants of performance.

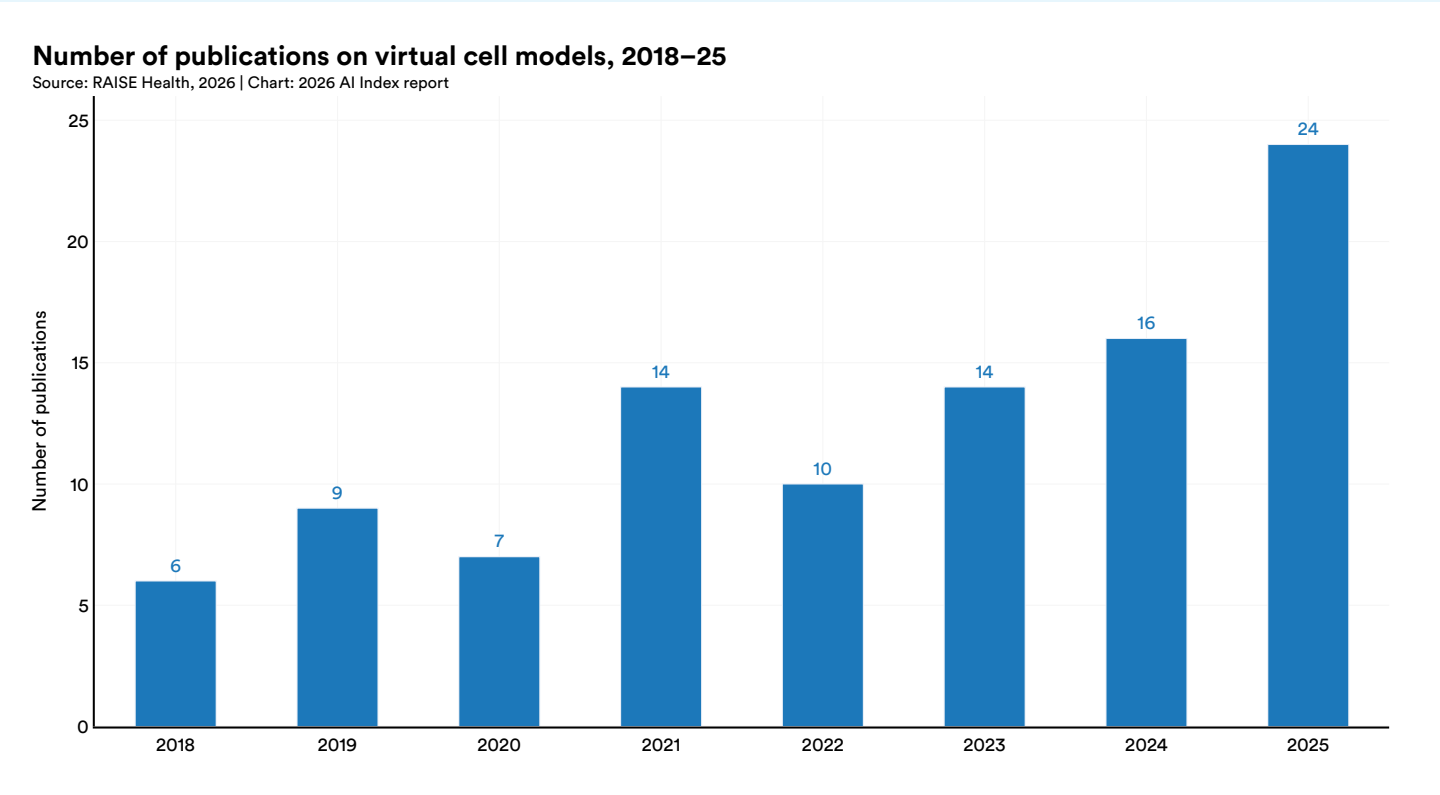


Figure 6.1.11

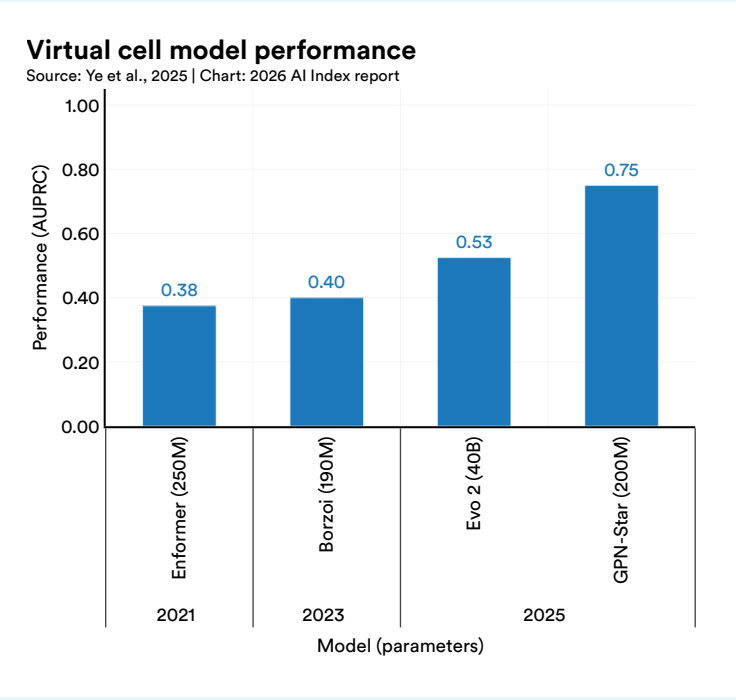


Figure 6.1.12

Multimodal Foundation Models for Biomedical Discovery

Scientific publications on multimodal foundation models for biomedical discovery have been growing rapidly since 2021 (Figure 6.1.13). While several subfields within multimodal biomedical AI gained traction in 2025, two areas have been especially impactful: Vision-language models pair medical or biological images with text, while vision-omics models integrate imaging with genomic or transcriptomic data.

Number of publications on multimodal biomedical AI, 2021–25

Source: RAISE Health, 2026 | Chart: 2026 AI Index report

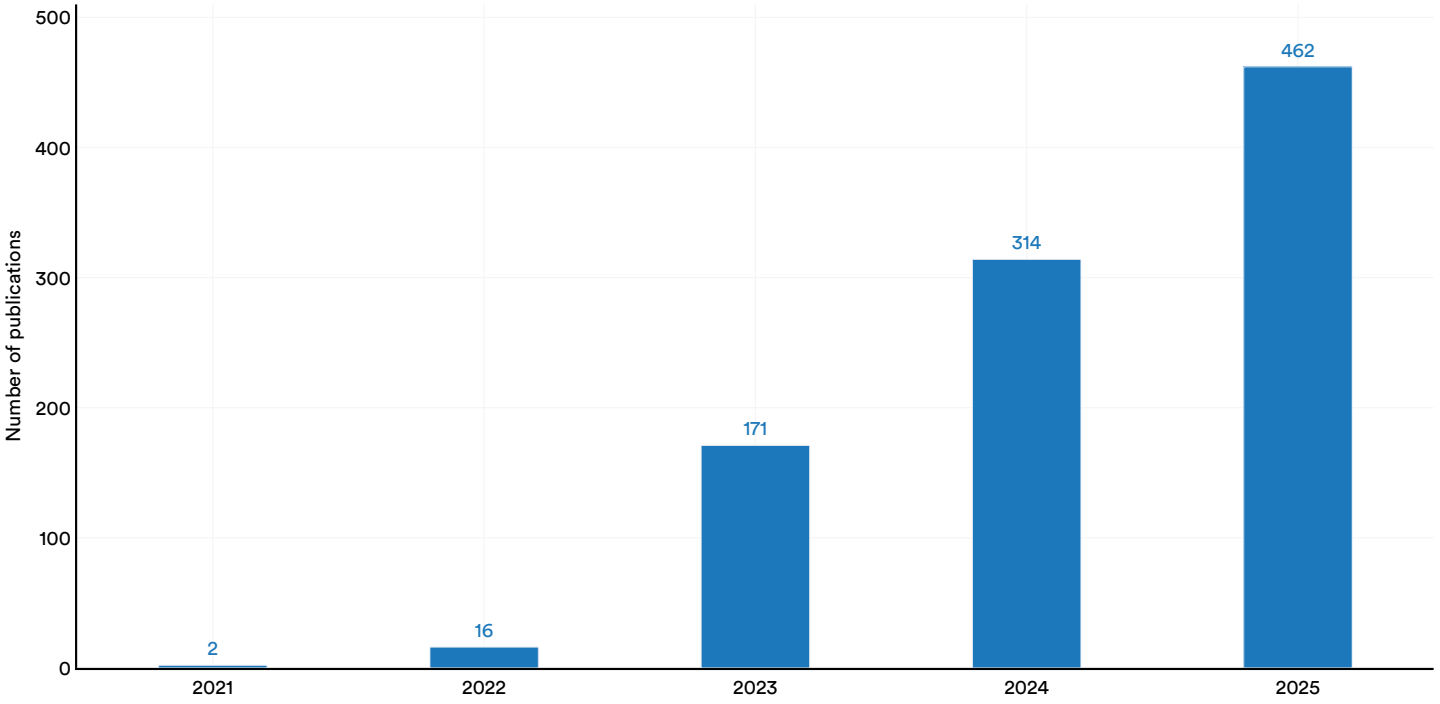


Figure 6.1.13

HIGHLIGHT:

Automated and Agentic Biomedical Discovery

In 2025, efforts to automate scientific discovery focused on integrating digital reasoning with physical laboratory validation. Robin, an automated discovery framework, linked literature-based hypothesis generation with experimental data analysis, identifying the ROCK inhibitor ripasudil as a novel candidate for dry age-related macular degeneration. STELLA, an autonomous bioinformatics agent, expanded its own technical capabilities by discovering and integrating new software tools rather than relying on manually curated toolsets. Biomni, a general-purpose biomedical AI agent developed at Stanford University, mapped a unified biomedical action space across 25 subfields, integrating 150 specialized tools, 105 software packages, and 59 databases.

Collaborative multiagent frameworks also emerged. Agent Laboratory, developed by AMD and Johns Hopkins, assigns distinct roles to PhD, Postdoc, and ML Engineer agents within a simulated lab structure. The Virtual Lab uses an LLM Principal Investigator to orchestrate specialized scientist agents, producing 92 novel nanobody binder designs for SARS-CoV-2. These systems are part of an early-stage trend toward multiagent coordination in biomedical research, though their outputs still require experimental validation.

6.2 Clinical Applications

The molecular and cellular AI advances described in section 6.1 provide the upstream models on which clinical tools increasingly depend. This section tracks how AI is being applied in clinical settings, from medical imaging and diagnostic reasoning to workflow integration, regulatory authorization, and enterprise-scale deployment. The analysis draws on prospective trial counts, FDA device authorization data, benchmark evaluations, and published apportionment outcomes from health systems. Across these areas, strong benchmark results have yet to translate reliably to measurable clinical outcomes.

Imaging

Data Scale and Availability

Training data for medical imaging AI remains roughly 100 times smaller in raw sample count than for nonmedical AI (Figure 6.2.1). MAIRA-2, a radiology-focused model trained on approximately 1.4 million chest radiographs, compared with DINOv3, a general-purpose vision transformer trained on 1.7 billion unlabeled natural images. On the multimodal side, RadFM trained on approximately 16 million mixed 2D and 3D medical scans paired with clinical text, while OpenCLIP trained on LAION-5B, comprising approximately 5.85 billion image–text pairs. Data scarcity is especially pronounced for three-dimensional modalities such as CT and MRI, and fragmentation across institutions further limits the development of large-scale medical foundation models.

Training data volume in medical and nonmedical AI: imaging-only vs. multimodal models

Source: RAISE Health, 2026 | Chart: 2026 AI Index report

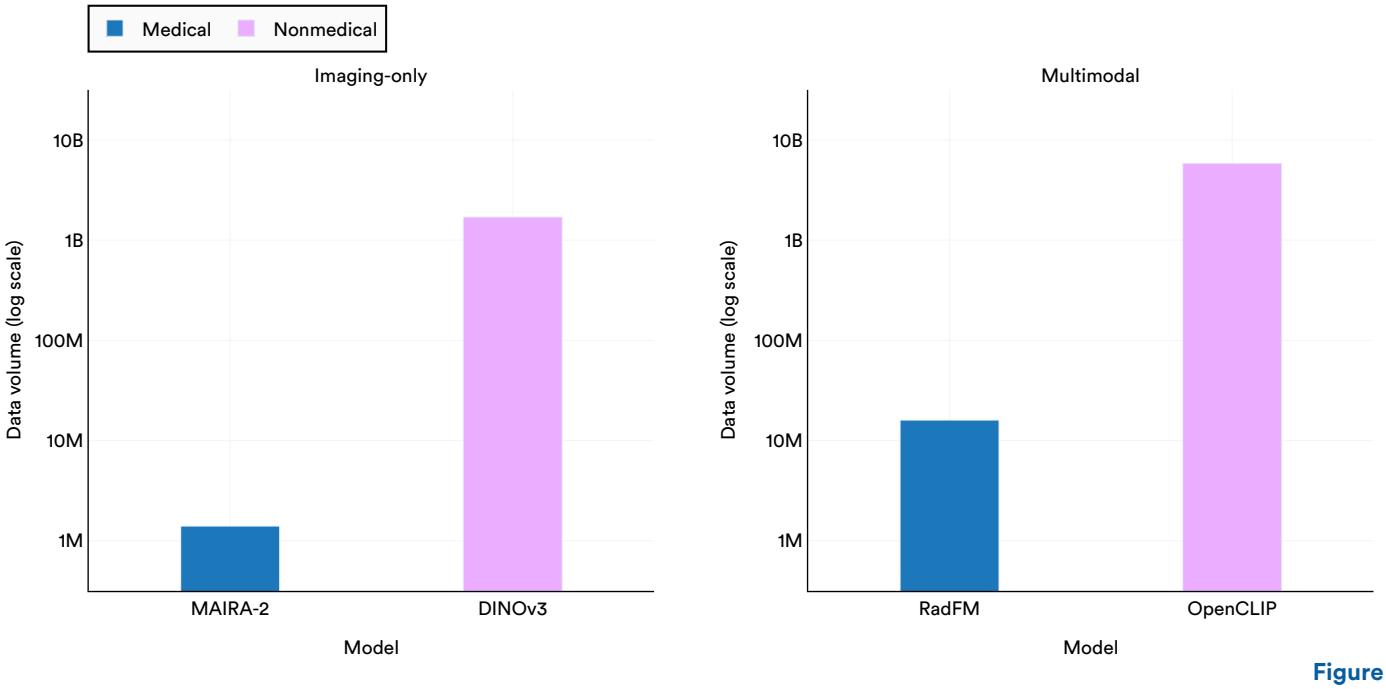


Figure 6.2.1

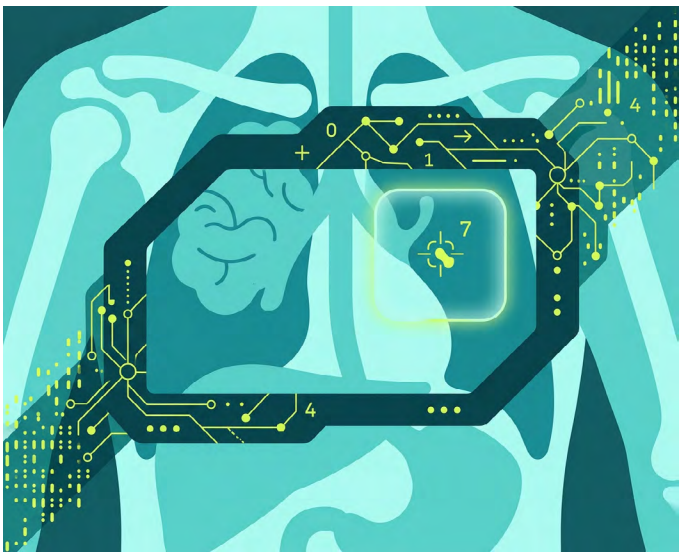
Modeling Approaches

Vision language models (VLMs) for medical imaging have expanded beyond radiology into pathology, dermatology, ophthalmology, and cardiology (Figure 6.2.2). Across six clinical disciplines, the number of research models and FDA-cleared commercial products grew, with pathology seeing the greatest concentration of new research releases. The Merlin model demonstrated that a highly capable CT foundation model could be trained on a single 40GB GPU by leveraging both radiology reports and ICD codes during training, making advanced medical AI accessible even in resource-constrained settings.

Medical imaging AI lacks the standardized cross-model benchmarks common in general-domain machine learning. Models in different specialties are typically evaluated on different datasets, making direct performance comparisons across disciplines difficult. Recent MICCAI 2025 challenges (CHIMERA, UNICORN) represent early efforts to address this gap. Human-centered evaluation, where clinicians manually review model outputs, has become more prevalent in publications and provides stronger evidence of clinical utility than lexical metrics.

Discipline	Notable releases	Analogous FDA-cleared models
Cardiology	EchoJEPA (2026) PanEcho (2025) EchoFM (2025) EchoPrime (2025)	Bunkerhill ECG-EF Heartflow Plaque Analysis
Oncology	MUSK (2025)	Allix5, Clairity (2025) Transpara (2.1.0) (2024)
Ophthalmology	EyeCLIP (2025) Meta-EyeFM (2025) RETFound-Green (2025)	CLARUS (700), Carl Zeiss (2025)
Pathology	Virchow2G (2026) KRONOS (2025) VORTEX (2025) Threads(2025) mSTAR (2025) PRISM2 (2025) MPath (2025) H-Optimus-0 (2024)	ArteraAI Prostate (2025) Ibex Prostate Detect (2025)
Radiology	MedGemma 1.5 (2026) COLIPRI (2026) TTE 3D CT (2025) 3DINO-ViT (2025) CT-FM (2025) RadFM (2025) Merlin (2025)	BriefCase Triage: CARE Multi-triage CT Body (2026) a2z-Unified-Triage (2025) Bunkerhill BMD (2025) Bunkerhill AAQ (2025) Brainomix 360 Triage Stroke (2025) Ezra Flash (2025)

Figure 6.2.2



Prospective Clinical Trials

The number of prospective trials validating medical imaging AI models grew by 28.5% year over year, from 417 in 2024 to 536 in 2025 (Figure 6.2.3). This growth signals the field is moving beyond studies that evaluate AI on past patient data toward live clinical trials, the kind of evidence required before hospitals will adopt these tools. Recent trials include MASAI, a randomized screening accuracy study of AI-assisted mammography, and NOTIFY-1 and NOTIFY-EXTEND, which tested whether flagging early signs of heart disease that AI spotted on routine CT scans led doctors to prescribe more preventive cholesterol medication.

Number of papers reporting prospective trials of clinical imaging ML/AI models, 2010–25

Source: RAISE Health, 2026 | Chart: 2026 AI Index report

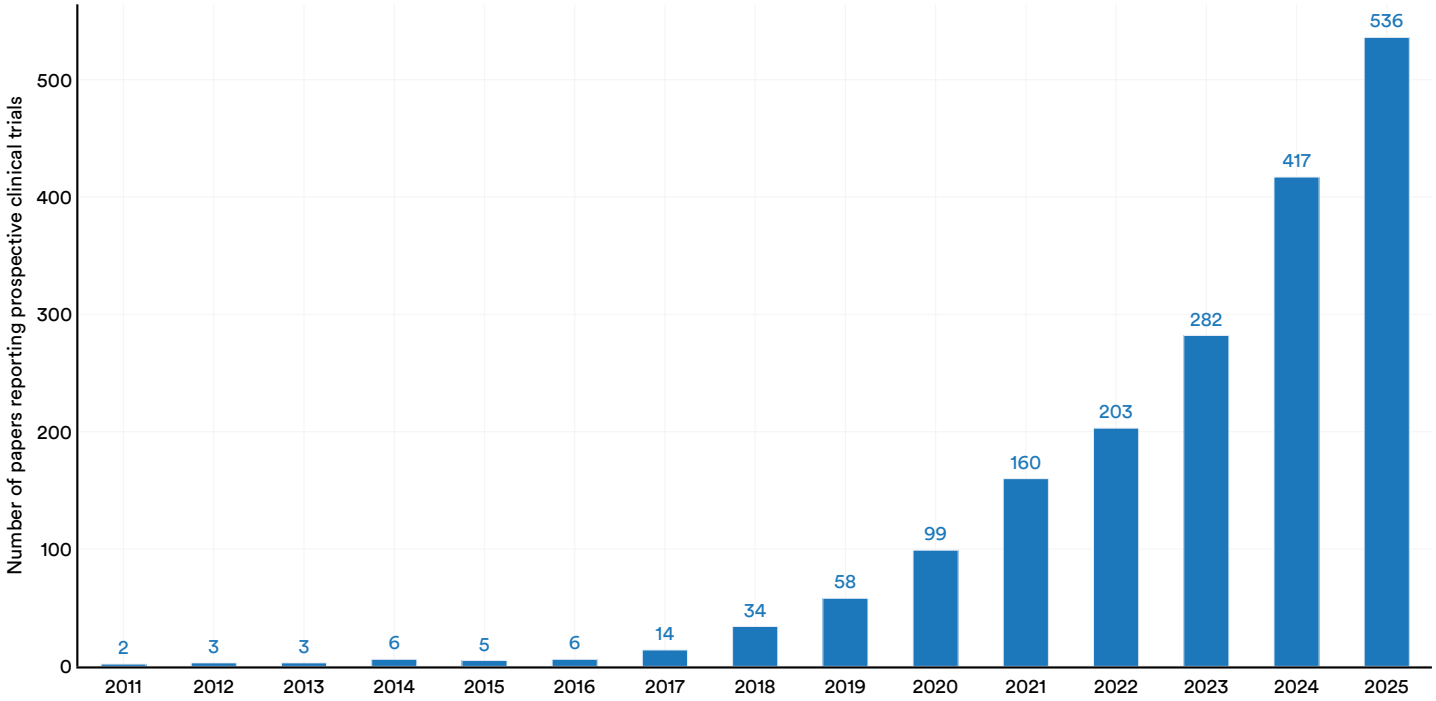


Figure 6.2.3

HIGHLIGHT:

LLM Clinical Reasoning Performance

In a multi-experiment evaluation, OpenAI's o1-preview reasoning model was tested on diagnostic reasoning tasks, management reasoning vignettes, probabilistic reasoning scenarios, and real emergency department (ED) cases with blinded expert scoring (Brodeur et al., 2025). On New England Journal of Medicine (NEJM) clinicopathological conferences (n=143), the model included the correct diagnosis in its differential 78% of the time, with 52% top-1 accuracy. On NEJM Healer cases (80 responses), it achieved a perfect revised-IDEA score in 78 of 80, compared with 47 of 80 for GPT-4, 28 of 80 for attending physicians, and 16 of 80 for residents. On management reasoning, o1 preview's median score was 86%, versus 42% for GPT-4 only, 41% for physicians with access to GPT-4, and 34% for physicians with conventional resources (Figure 6.2.4). In 76 real ED cases, o1 produced diagnoses rated "exact/very close" in 67%–83% of cases across three diagnostic stages, surpassing two attending physicians at each stage.

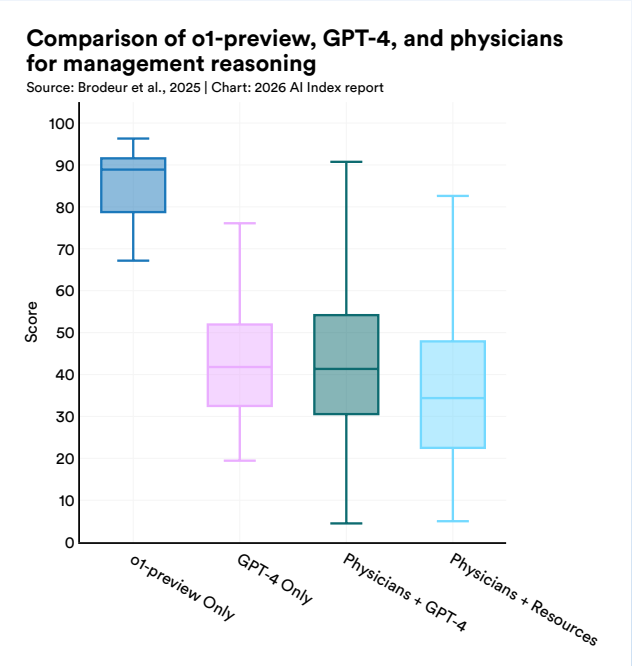


Figure 6.2.4[2]

These results suggest that current LLMs have surpassed most existing clinical reasoning benchmarks, but they reflect isolated cognitive evaluations rather than real-world clinical integration. Whether AI-assisted reasoning translates to improved patient outcomes remains an open question requiring prospective trials.

HIGHLIGHT:

AI Agents in Clinical Medicine

Autonomous and semiautonomous AI agents have emerged as a major development in clinical AI in 2025–26. Unlike conventional AI models that generate predictions or classifications in isolation, these systems reason across multiple steps, access external tools and data sources, and coordinate with other AI agents or human clinicians to complete complex clinical tasks.

Multiagent frameworks, in which multiple AI agents take on specialized roles—such as diagnostician or pharmacist—and collaborate through structured reasoning protocols, have shown early promise on benchmark evaluations. Diagnostic accuracy gains over single-agent baselines ranged from 7% to over 60%, depending on the complexity of the clinical task (Gorenshtein et al., 2025; Zheng et al., 2025; Liu et al., 2025). Microsoft's AI Diagnostic Orchestrator (MAI-DxO), paired with OpenAI's o3 reasoning model,

2 Box plot of normalized management reasoning points by LLMs and physicians on Gray Matters management cases. Five cases were included. Three o1-preview responses were generated for each case. The prior study collected five GPT-4 responses to each case, 176 responses from physicians with access to GPT-4, and 199 responses from physicians with access to conventional resources.

HIGHLIGHT:

achieved 85.5% accuracy on diagnostically challenging cases from the New England Journal of Medicine, compared with approximately 20% among 21 practicing physicians with five to 20 years of clinical experience working under comparable conditions.

A new set of benchmarks specifically designed to evaluate these agentic systems has started to appear. A 2025 scoping review identified 43 studies evaluating agentic AI in healthcare, 36 of which (84%) were published in 2025. On MedAgentBench (Jiang et al., 2025), which evaluates LLM agents in a virtual electronic health record (EHR) environment across 300 clinically derived tasks, the best performing model achieved a task success rate of 69.7%. These results suggest that, despite access to advanced capabilities such as tool use and iterative reasoning, the evidence base for reliable autonomous clinical AI agents remains early-stage.

Deployment, Implementation, and Deimplementation

This section tracks the regulatory, institutional, and evidentiary dimensions of clinical AI deployment, from FDA device authorizations to enterprise-scale outcomes.

FDA-Authorized AI/ML-Enabled Devices

In the United States, FDA 510(k) is the most common regulatory pathway for AI medical devices, requiring manufacturers to demonstrate that a new device is substantially equivalent to one already on the market rather than conducting new clinical trials. The number of 510(k)-cleared AI/ML-related devices reached 246 in 2025, continuing a steep upward trajectory that began with 16 devices in 2016 (Figure 6.2.5).

Because most cleared radiology AI solutions are offered commercially rather than as open-source tools, healthcare systems are typically required to complete financial clearance and cost-effectiveness justification before implementation. Comp2Comp, a notable exception, is an open-source python package for CT imaging analysis with two modules (bone mineral density and abdominal aortic quantification) that secured FDA clearance in 2025.

FDA 510(k)-cleared AI/ML-enabled imaging-related medical devices, 2011–25

Source: FDA, 2025 | Chart: 2026 AI Index report

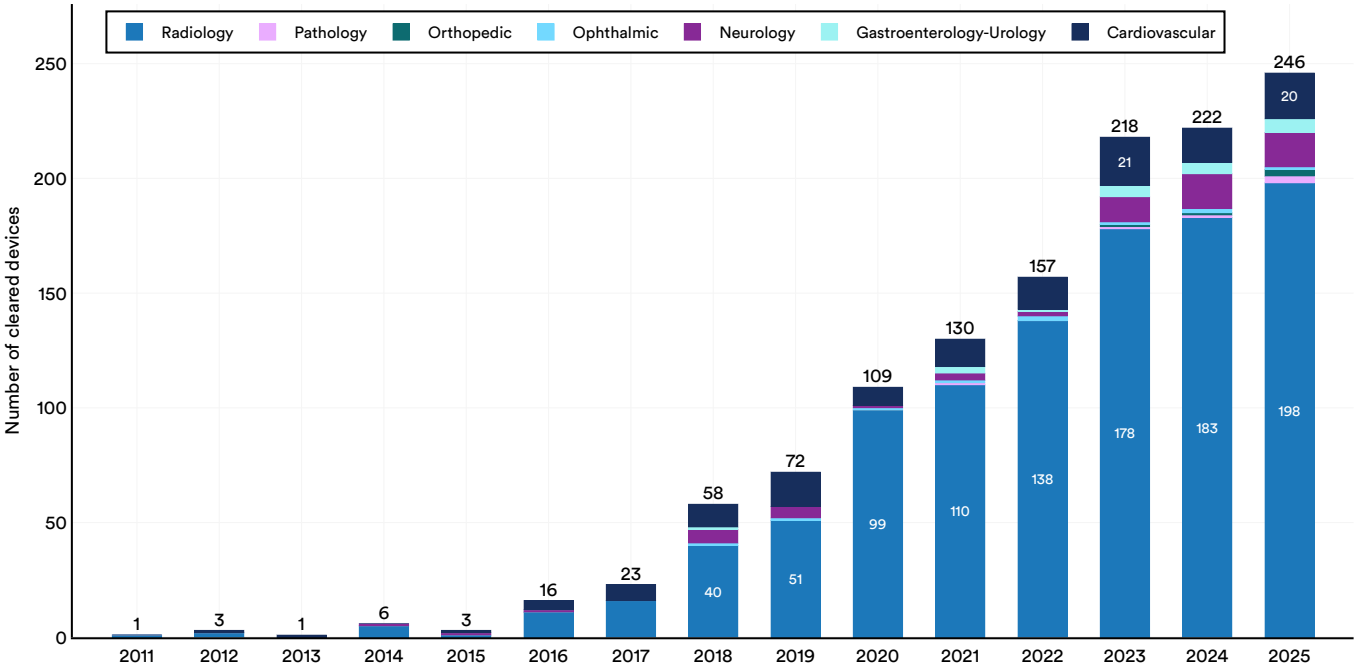


Figure 6.2.5

By December 2025, the FDA had authorized a total of 1,357 AI/ML-enabled medical devices from 693 different companies across 17 clinical specialties (Figure 6.2.6). Annual authorizations reached 258 through September 2025, already surpassing all prior full-year totals. The cumulative total crossed the 1,000-device milestone in 2024. Ninety-eight new companies entered the space in 2025, continuing a trend of broadening market participation (103 new entrants in 2023, 109 in 2024).

Number of AI medical devices approved by the FDA, 1995–2025

Source: FDA, 2025 | Chart: 2026 AI Index report

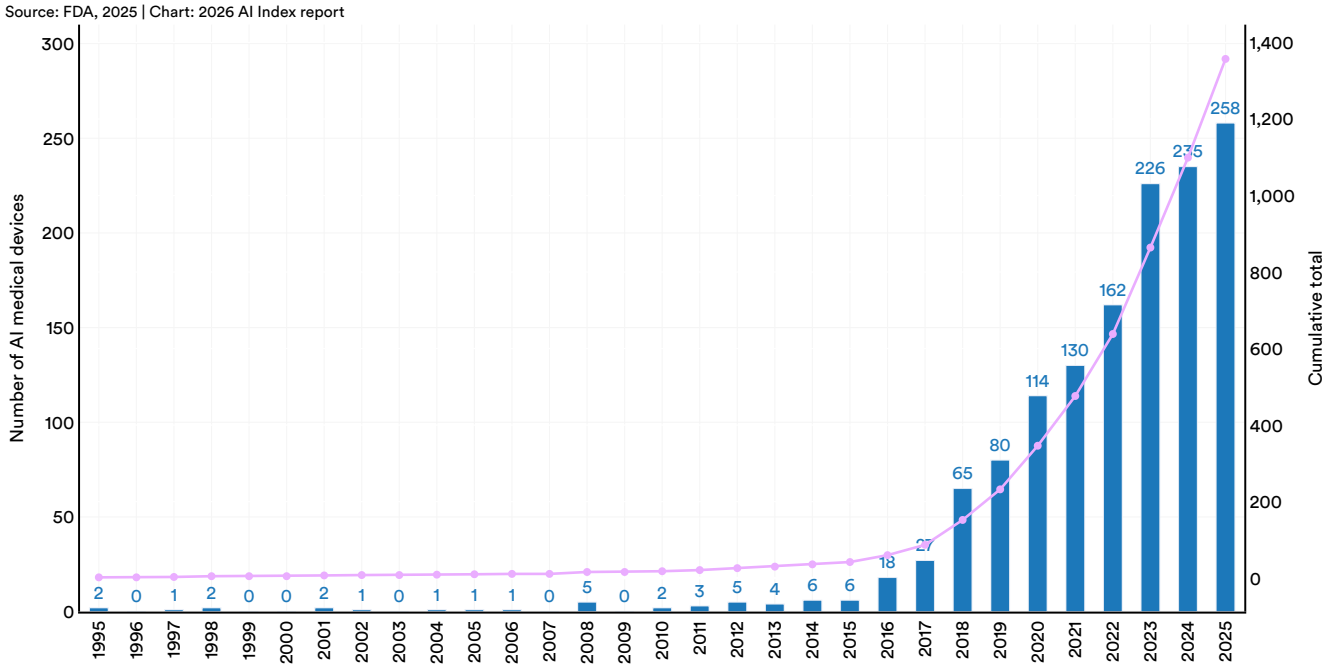


Figure 6.2.6

Devices by Clinical Specialty

Radiology accounts for the largest share of authorized AI/ML devices at 1,039 of 1,357 (76.6%), followed by cardiovascular (130 devices, 9.6%) and neurology (61 devices, 4.5%) (Figure 6.2.7). Non-radiology authorizations have increased from 7 in 2016 to 60 in 2025 (Figure 6.2.8). Cardiology, neurology, anesthesiology, and gastroenterology-urology have all seen acceleration since 2020, suggesting that AI is beginning to spread from imaging-centric applications to broader clinical domains.

Number of AI/ML medical devices approved by the FDA by top companies, 2016–25

Source: FDA, 2025 | Chart: 2026 AI Index report

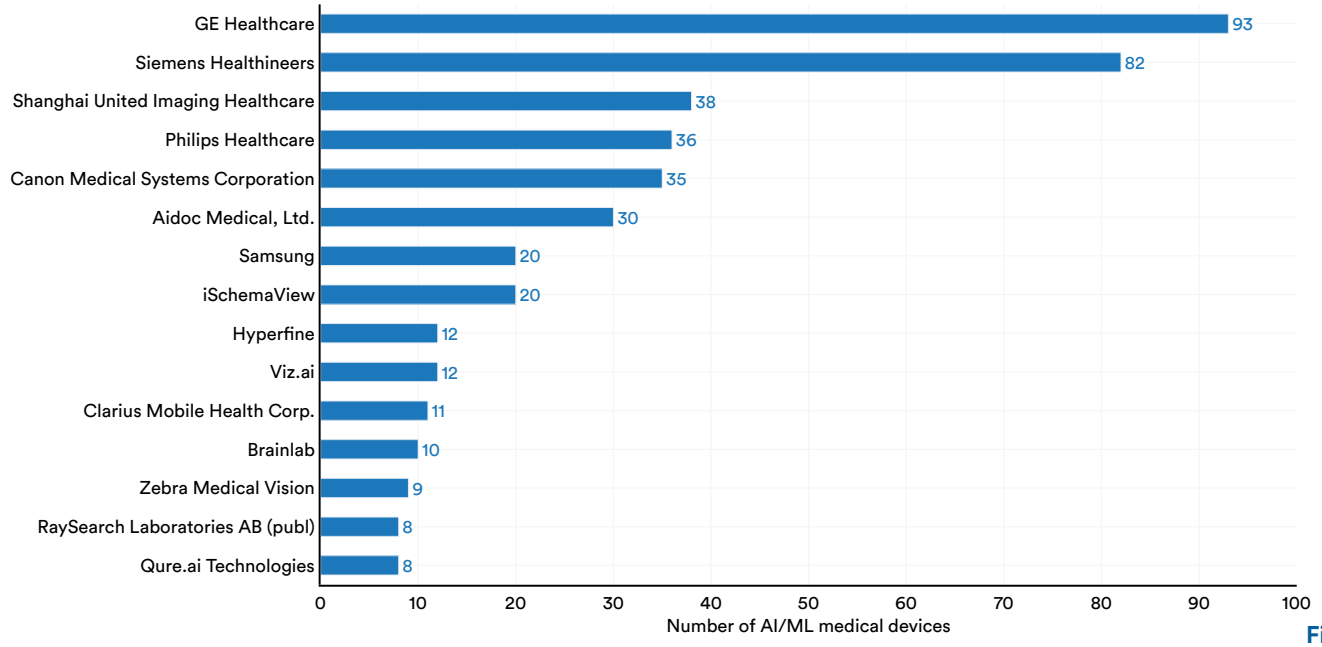


Figure 6.2.7

Number of AI/ML medical devices approved by the FDA by specialty, 2016–25

Source: FDA, 2025 | Chart: 2026 AI Index report

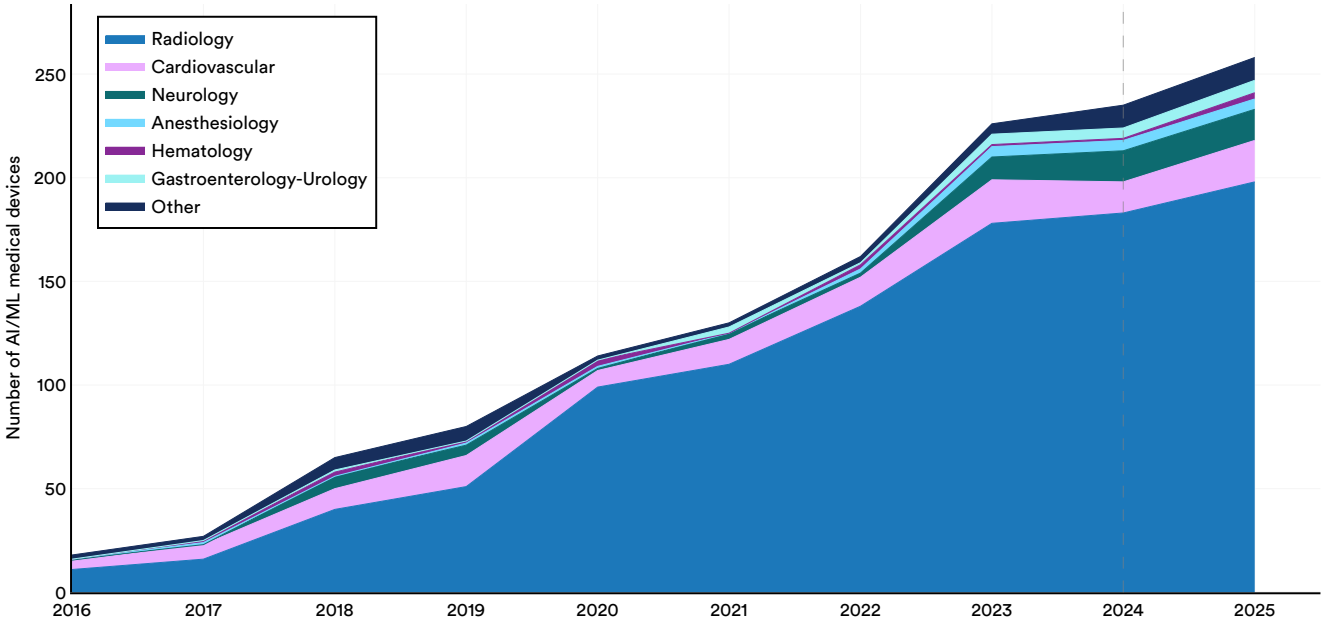


Figure 6.2.8

Industry Landscape

FDA clearance does not equal clinical adoption. Financial, operational, and institutional barriers often stand between regulatory authorization and real-world deployment. The authorized device market is concentrated at the top but fragmented overall. GE Healthcare leads with 93 devices, followed by Siemens Healthineers (82), Shanghai United Imaging Healthcare (38), Philips Healthcare (36), and Canon Medical Systems (35), and Aidoc Medical (30) (Figure 6.2.9). Of the 626 companies with at least one authorized device, the large majority hold only one or two, reflecting a broad ecosystem of specialized entrants alongside established manufacturers.

Number of AI/ML medical devices approved by the FDA by top companies, 2016–25

Source: FDA, 2025 | Chart: 2026 AI Index report

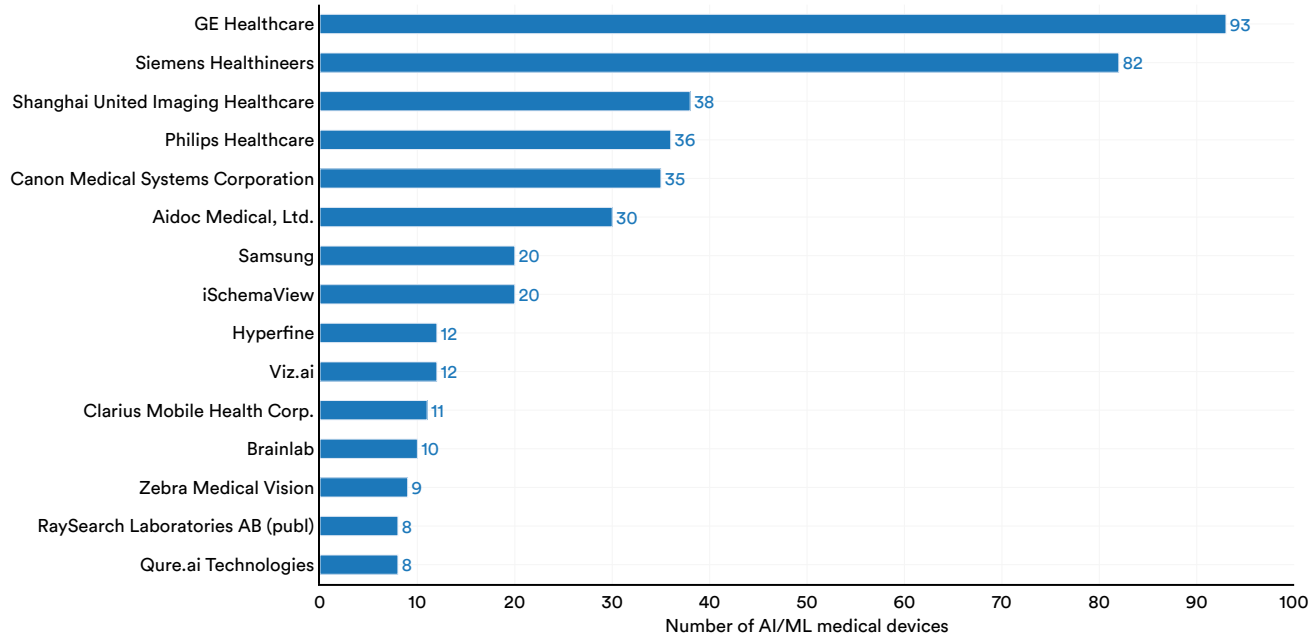


Figure 6.2.9

In January 2025, the FDA issued draft guidance on AI-enabled device software functions applying a Total Product Life Cycle approach. Predetermined Change Control Plans , a mechanism that permits iterative updates after initial market authorization, were used in approximately 10% of 2025 clearances. Despite this growth, a peer-reviewed analysis of all 1,016 authorizations through December 2024 (Singh et al., 2025) found that only 2.4% of devices with clinical studies were supported by randomized controlled trial data, with nearly all devices entering via the 510(k) pathway.

Enterprise-Scale Deployments in 2025

Clinical AI moved from pilot-stage initiatives to enterprise-scale deployments in 2025, with health systems reporting measurable outcomes across clinical and operational domains. The most published evidence was from ambient AI documentation, AI-powered sepsis prediction, and generative AI integration into clinical workflows.

Ambient AI Documentation

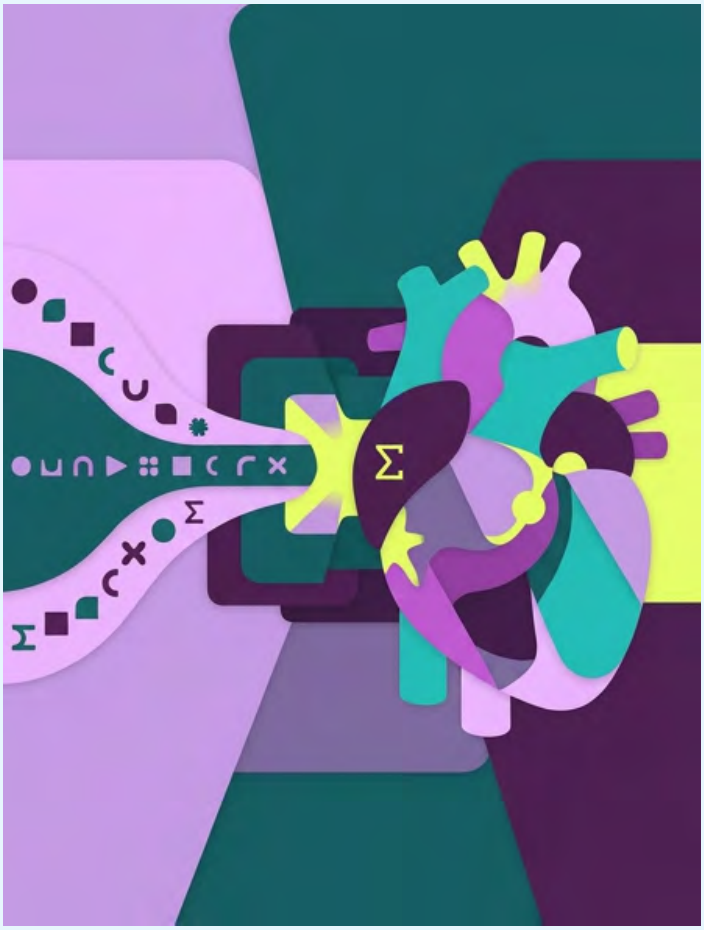
Ambient AI scribes, tools that automatically generate clinical documentation from patient–clinician conversations, saw the broadest adoption of any clinical AI category in 2025. Abridge, one of the leading

platforms, expanded from approximately 100 to over 150 health systems, including Kaiser Permanente's deployment across 40 hospitals and more than 600 medical offices. Adoption reached 63% among hospitals using Epic's electronic health record system.

Outcomes were consistent across multiple institutions. Sharp HealthCare reported an 83% reduction in note-writing effort and a 3.5%–6% increase in work relative value units—a standard measure of physician clinical productivity—per encounter. The University of Chicago Medicine reported a 47% reduction in cognitive load and a 58% increase in undivided patient attention. MaineHealth reported a 23% reduction in time spent on clinical notes, with the tool used in 70.3% of encounters. At Northwestern Medicine, physicians using the tool in more than half of encounters saw 11.3 additional patients per month and a 24% reduction in documentation time, with a reported 112% return on investment. At Stanford Health Care, a prospective study of 48 physicians published in JAMIA (February 2025) found statistically significant reductions in task load and burnout, with physicians reporting a median time savings of 20 minutes per half day of clinic.

AI-Powered Sepsis Prediction

Two sepsis prediction systems reported mortality reductions in large-scale deployments in 2025. The Targeted Real-time Early Warning System, developed at Johns Hopkins and commercialized by Bayesian Health, was deployed across 13 Cleveland Clinic hospitals. Reported outcomes included an 18.7% relative reduction in sepsis mortality, a 1.85-hour reduction in median time to first antibiotic order, the correct identification of 82% of sepsis cases, an 89% clinician adoption rate, and a 10% reduction in intensive care unit utilization. COMPOSER, a deep learning model at UC San Diego Health monitoring over 150 variables per patient, reported a 17% reduction in sepsis mortality (1.9% absolute) across 6,217 admissions, a 5% increase in sepsis bundle compliance, and an estimated 50 lives saved annually.



Generative AI in Clinical Workflows

Health systems began embedding LLM-powered tools directly into electronic health records. ChatEHR, a system generating plain-language summaries of patient records, logged 23,000 sessions across 1,075 trained users within three months of broad rollout. 60% of usage occurred through automated prompts and 40% through interactive interfaces. Separately, an AI tool generating plain-language explanations of laboratory, imaging, and pathology results was evaluated in a study published in *JAMA Network Open* (August 2025). Of 93 survey respondents, 85% considered the tool user-friendly, 72% found it beneficial for laboratory results, and 63% for imaging results. OpenEvidence, a real-time evidence retrieval platform, reported adoption by 40% of U.S. physicians.

Evidence Gaps and Governance

The inaugural State of Clinical AI Report (January 2026), published by the Stanford-Harvard ARISE Network, reviewed over 500 clinical AI studies and found that nearly half used exam-style questions rather than real patient data. Only 5% used real clinical data. The report concluded that AI performs most effectively when supporting rather than replacing clinician judgment.

Separately, the NOHARM benchmark found that leading LLMs produced 11.8 to 14.6 severely harmful recommendations per 100 clinical cases, with 76.6% being errors of omission (e.g., failing to recommend a critical test). These findings apply to general-purpose LLMs evaluated on open-ended clinical reasoning tasks, not to the narrower, task-specific tools driving current adoption. Ambient scribes and sepsis alerts, for example, operate within constrained workflows with clinician oversight.

Governance frameworks have also advanced. Stanford Health Care's FURM framework now governs all new AI tool adoptions at that institution, and the GUIDE-AI Lab is working to make the framework available to other health systems.

HIGHLIGHT:

Digital Twins in Medicine

A medical digital twin is a dynamic, data-linked computational representation of an individual patient that updates over time and supports forecasting, simulation, and treatment optimization. Research activity in this area has grown rapidly, with publication counts rising from near 0 in 2015 to 372 in 2025 (Figure 6.2.10). Patent filings in healthcare digital twins (CPC class G16H) tell a similar story, with filings increasing from 30 in 2016 to 4,926 in 2025 (Figure 6.2.11).

However, conceptual clarity has not kept pace with publication growth. A 2025 scoping review in npj Digital Medicine assessed 149 human digital twin studies published between 2017 and 2024 and found that only 12.1% (18 studies) satisfied the National Academies of Sciences, Engineering, and Medicine (NASEM) definition of a digital twin. That definition requires three elements: personalization, dynamic updating, and predictive capability (Sadée et al., 2025). Only 19% of systems were tested in real healthcare environments.

Clinical trials incorporating digital twin elements accelerated in 2025, particularly in oncology and diabetes. A pilot trial in prostate cancer using adaptive therapy concluded in 2025 with significantly increased survival (Zhang et al., 2022). New trials extended the approach to breast cancer (Mayo Clinic phase II) and ovarian cancer (ACTOv phase II RCT, n=80). For diabetes, a randomized controlled trial (n=150) of Twin Health's Whole Body Digital Twin platform found that 71% of participants achieved an HbA1c below 6.5% within twelve months, while safely reducing their intake of blood sugar–lowering medications.

HIGHLIGHT:

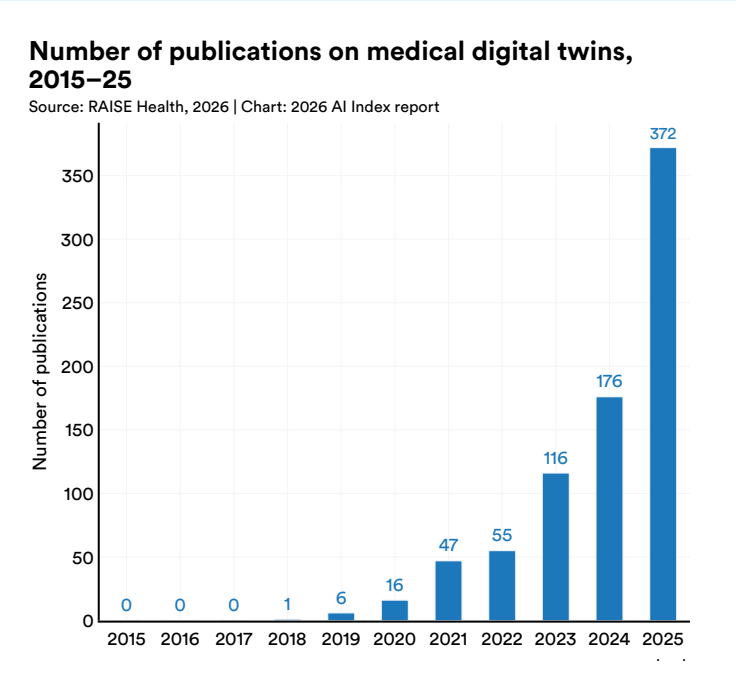


Figure 6.2.11[3]

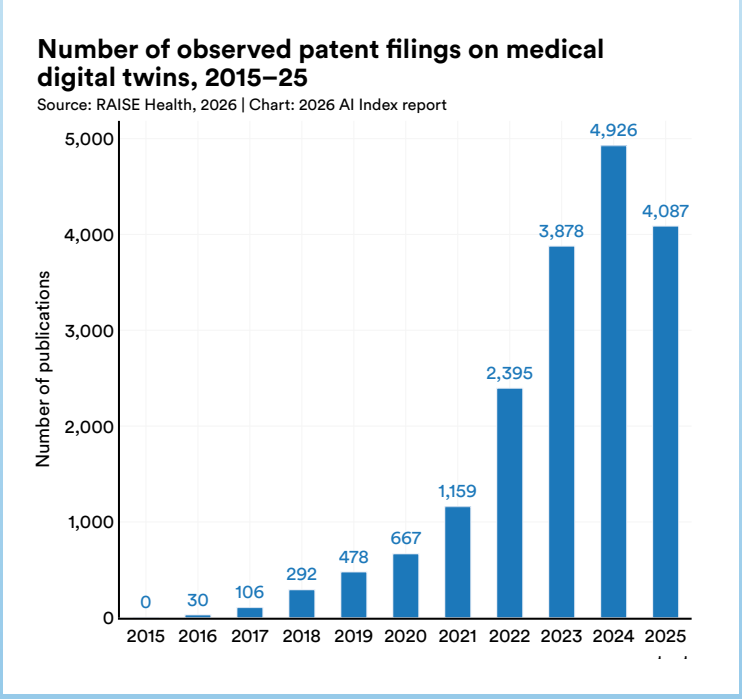


Figure 6.2.12

3 The bar in 2025 appears lower than in 2024 because not all patents filed in 2025 have been published or become publicly available yet.

6.3 Patient Engagement

As patients interact more with AI tools—both through clinical workflows and consumer-facing platforms, efforts have been made to understand how they perceive these technologies. This section examines AI-generated health search results, patient attitudes toward AI in healthcare, and the emerging evidence base for patient-facing AI tools.

AI Overviews for Health-Related Searches

AI-generated summary responses, referred to by Google as "AI Overviews," now appear at the top of most health-related search results. On average, 84%–92% of health-related queries triggered an AI Overview across five primary query types (Figure 6.3.1). Symptom and common health questions were the most likely to trigger an overview (92%), followed by treatment-related queries (90%) and condition-based queries (84%–88%). AI-generated summaries are a routine feature of health information searches, shaping the initial interpretation of questions posed by most users.

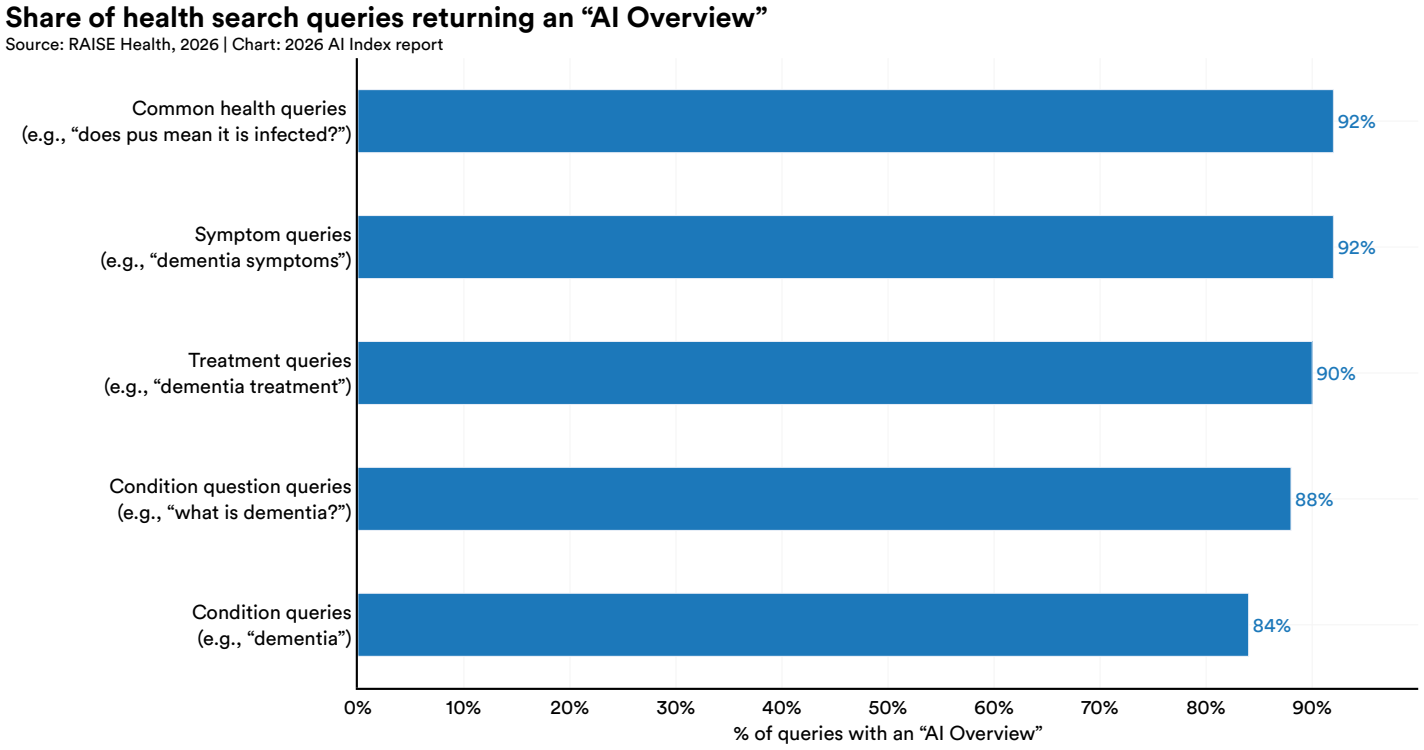


Figure 6.3.1

Patient Perspectives on AI in Healthcare

Publication volume on the patient perspective of AI in healthcare grew tenfold between 2020 and 2025 (Figure 6.3.2). Conditional acceptance emerged as a prevalent perspective across the literature. Patients tended to endorse AI in assistive roles rather than autonomous decision-making, particularly in high-stakes clinical contexts (Fee et al., 2025; Allen et al., 2025; Hmido et al., 2025). Demographic disparities in acceptance—patterned by age, gender, education, and race—were documented across multiple studies (Labinsky et al., 2025; Ogu et al., 2025; Li et al., 2025).

Preservation of the human relationship emerged as a consistent theme, with patients identifying the potential loss of empathic care as a primary concern (Carl et al., 2025; Davis et al., 2025). Trust in AI appeared to be clinician-mediated rather than technology-evaluated. Provider endorsement functioned as a key determinant of patient acceptance (Berger et al., 2025; Machado et al., 2025; Nong et al., 2025). Transparency and disclosure of AI use were similarly prioritized across populations, and emerging disclosure frameworks offer practical guidance for clinical settings (Figure 6.3.3).

Number of publications on patient perceptions of AI in healthcare, 2020–25

Source: RAISE Health, 2026 | Chart: 2026 AI Index report

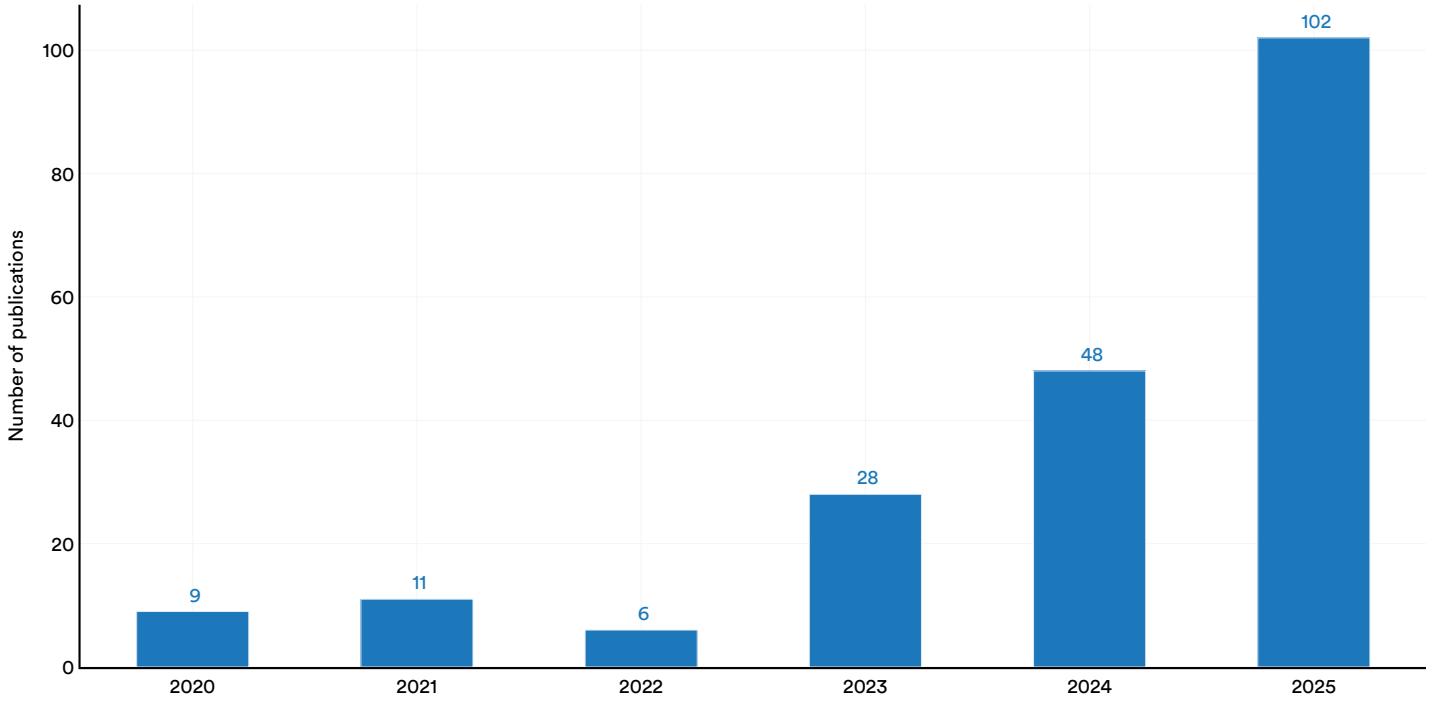


Figure 6.3.2

Artificial intelligence use cases and recommendations for patient notification
Source: Mello et al., 2025

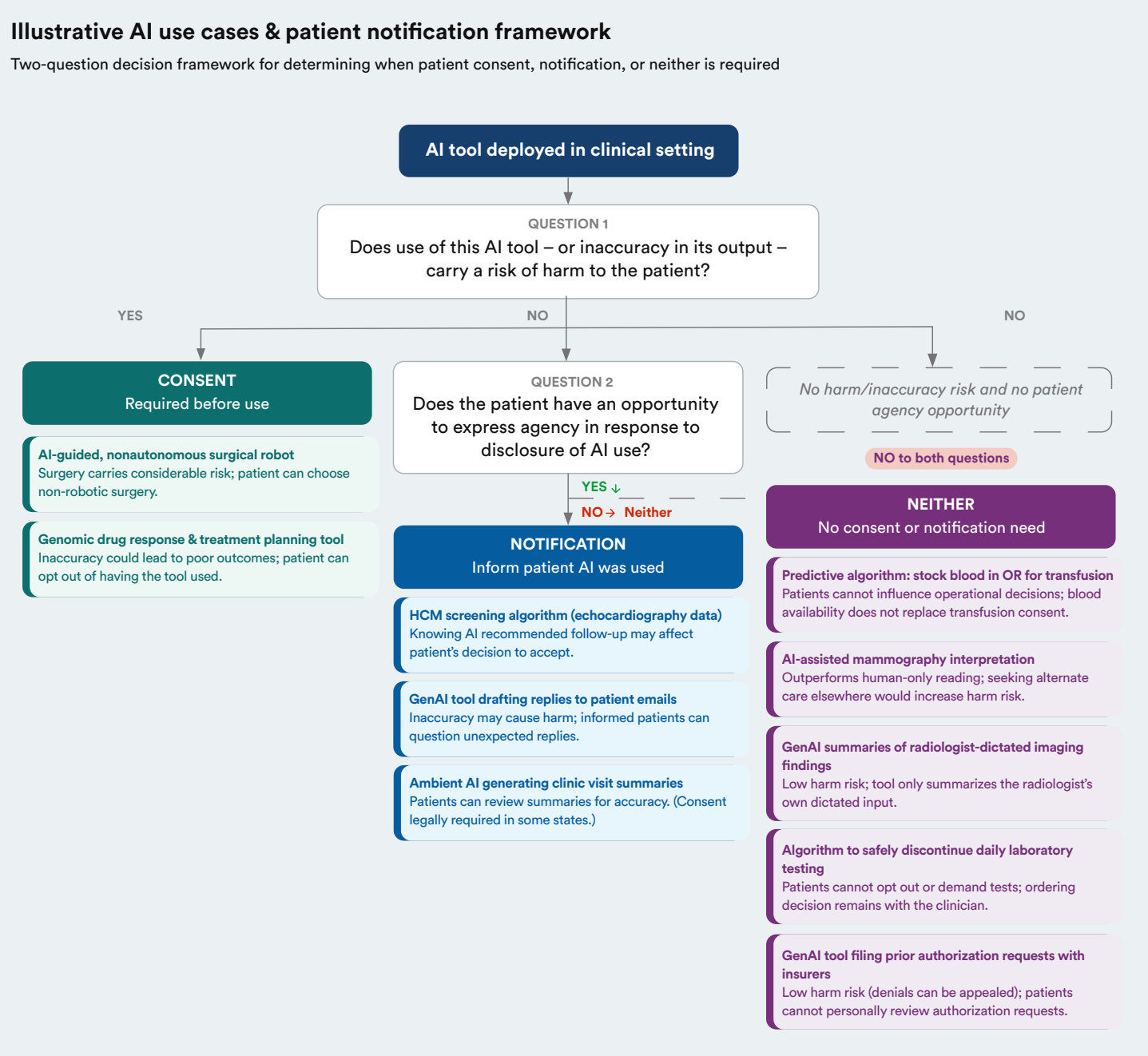


Figure 6.3.3

Internal medicine, radiology, and oncology were the most frequently represented specialties in this literature (Figure 6.3.4). The United States, United Kingdom, and Germany account for the greatest number of publications, while studies from sub-Saharan Africa, Latin America, and Southeast Asia remain underrepresented (Figure 6.3.5). Studies that include children and adolescents as participants, rather than drawing solely on parent or caregiver perspectives, remain rare.

Medical specialties represented in publications exploring patient perceptions of AI in healthcare, 2020–25

Source: RAISE Health, 2026 | Chart: 2026 AI Index report

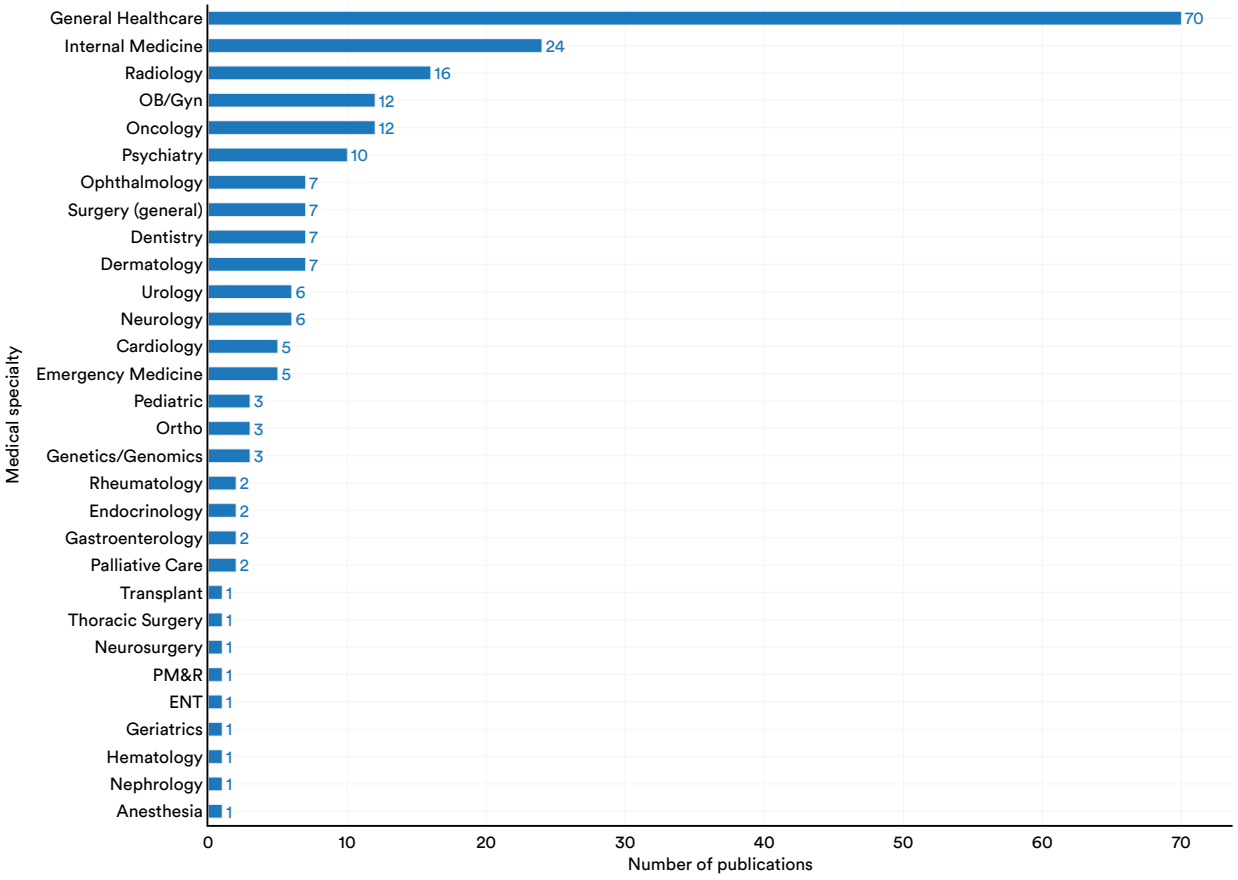


Figure 6.3.4 [4]

Geographic distribution of publications on patient perceptions of AI in healthcare by country, 2020–25

Source: RAISE Health, 2026 | Chart: 2026 AI Index report

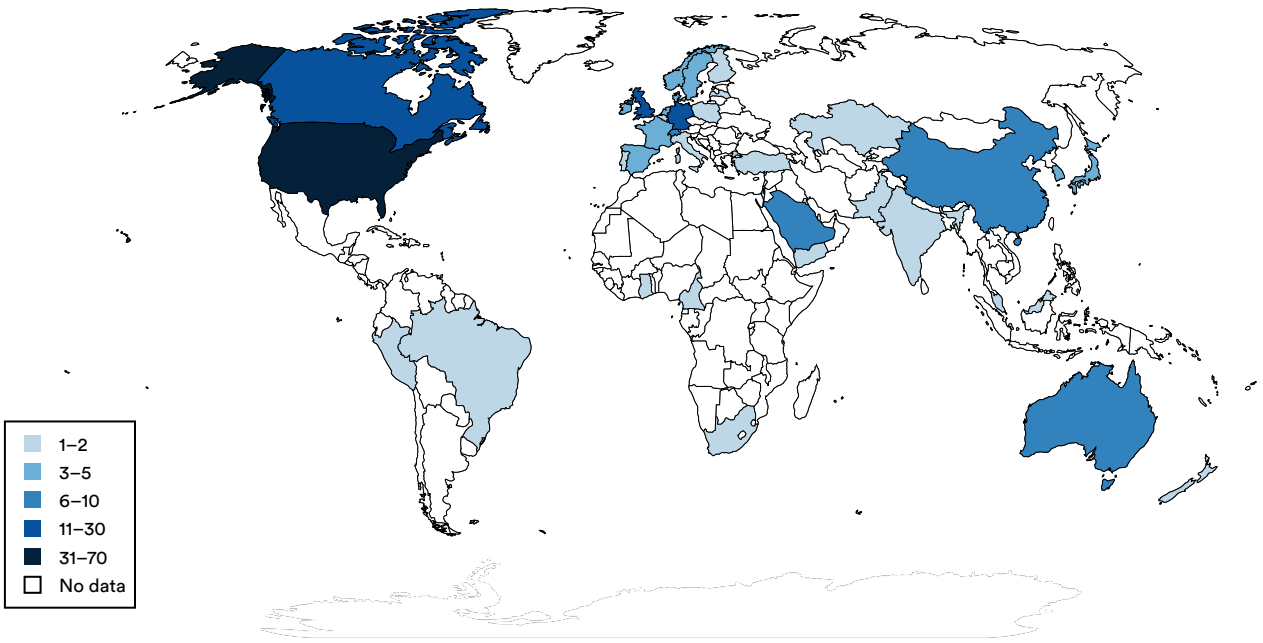


Figure 6.3.5[5]

4 Publications may be tagged with multiple specialties.

5 Publications may be tagged with multiple countries; countries of origin are included. In total, 39 countries are represented (N = 204).

6.4 Ethical Considerations

This section tracks the volume and focus of ethical disclosure in medical AI publications, drawing on a bibliometric analysis of PubMed Central from January 2021 to December 2025. Publications were identified using search terms for medical AI and ethics, and then categorized by emphasis on data sharing, algorithm sharing, biosecurity, and global health. Ethics topics were either grouped under algorithmic, governance, or societal concerns.

Volume and Concentration

Of the total number of medical AI publications in 2025, 43.4% discussed ethics topics—up from 37.1% in 2024 (Figure 6.4.1). The absolute number of such publications more than doubled between the two years. Among the specific topics discussed, the growth was concentrated on governance, outpacing algorithmic and societal concerns (Figure 6.4.2). In 2025, the number of governance-related publications reached 1,228, compared with 896 for algorithmic concerns and 874 for societal concerns.

Despite the attention paid to biosecurity in policy discussions, the subject is relatively unexplored in medical AI publications. In 2025, only 14 of these publications discussed biosecurity, with even fewer directly addressing the ethical implications of misuse or dual use (Figure 6.4.3).

Number of medical AI and ethics publications, 2021–25

Source: RAISE Health, 2026 | Chart: 2026 AI Index report

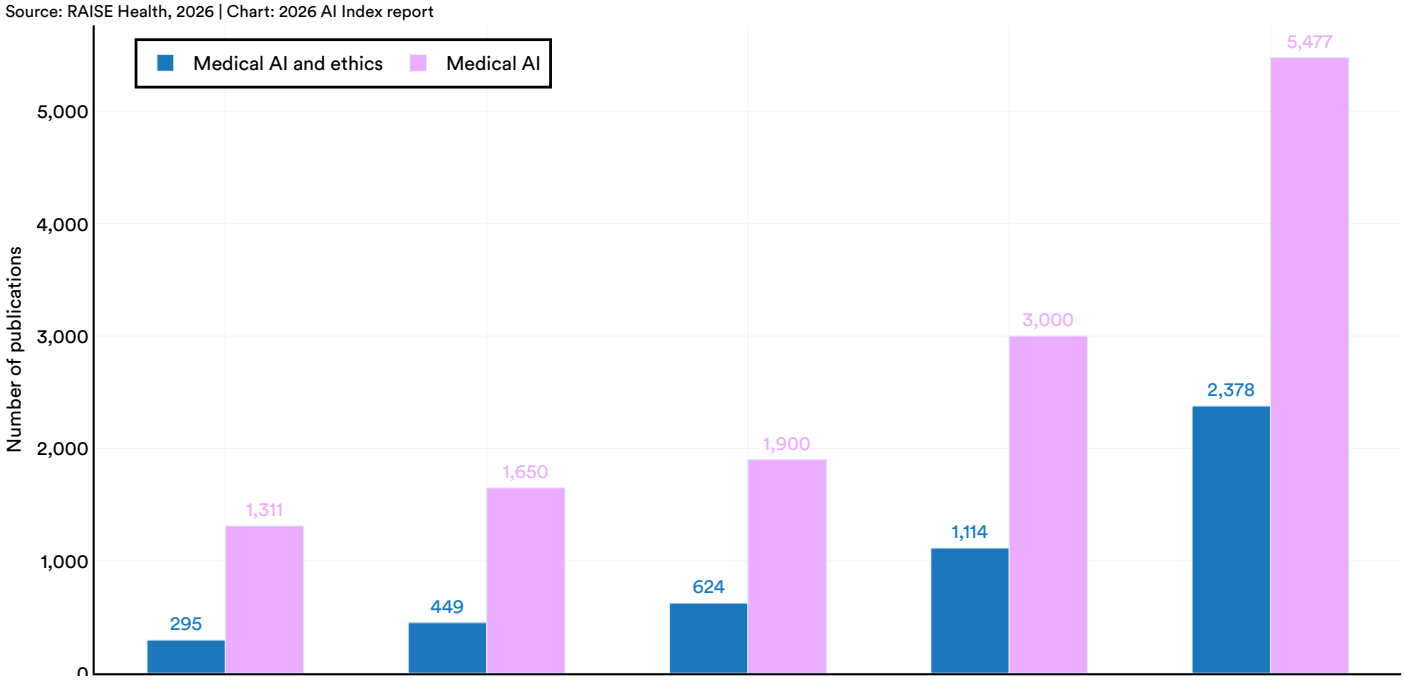


Figure 6.4.1

Number of medical AI publications by ethics topics, 2021–25

Source: RAISE Health, 2026 | Chart: 2026 AI Index report

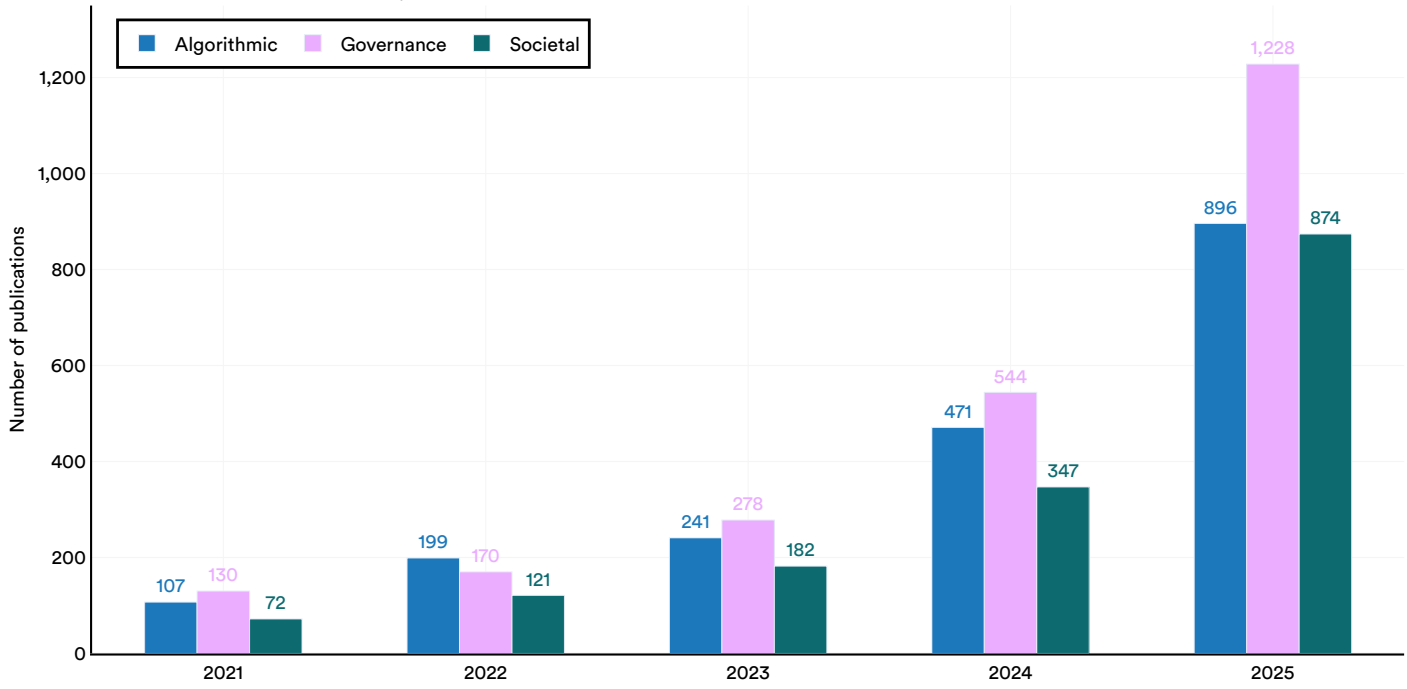


Figure 6.4.2

Number of medical AI and biosecurity publications, 2021–25

Source: RAISE Health, 2026 | Chart: 2026 AI Index report

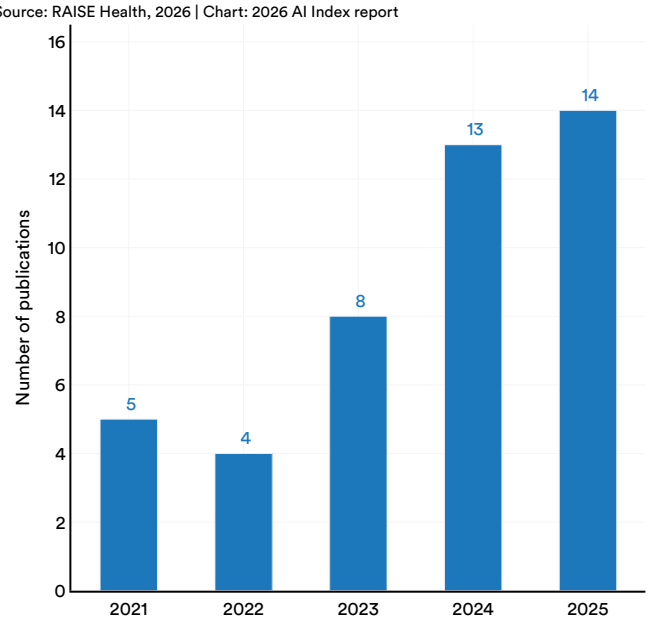


Figure 6.4.3

Global Health: A Different Ethical Focus

Global health is an exception to the governance-dominated pattern. Among publications addressing global health in 2025, 51.8% (100 of 193) also mentioned ethics topics (Figure 6.4.4). Europe led with 38 publications, followed by East Asia (31) and North America (28), while sub-Saharan Africa, Latin America, and Oceania each produced fewer than five (Figure 6.4.5). In a departure from every other subcategory examined, societal concerns—including equity, justice, and accessibility—ranked highest in the global health context, surpassing both governance and algorithmic concerns (Figure 6.4.6). Researchers studying AI for global health are raising different questions from their peers working in the broader field.

Number of medical AI, global health, and ethics publications, 2021–25

Source: RAISE Health, 2026 | Chart: 2026 AI Index report

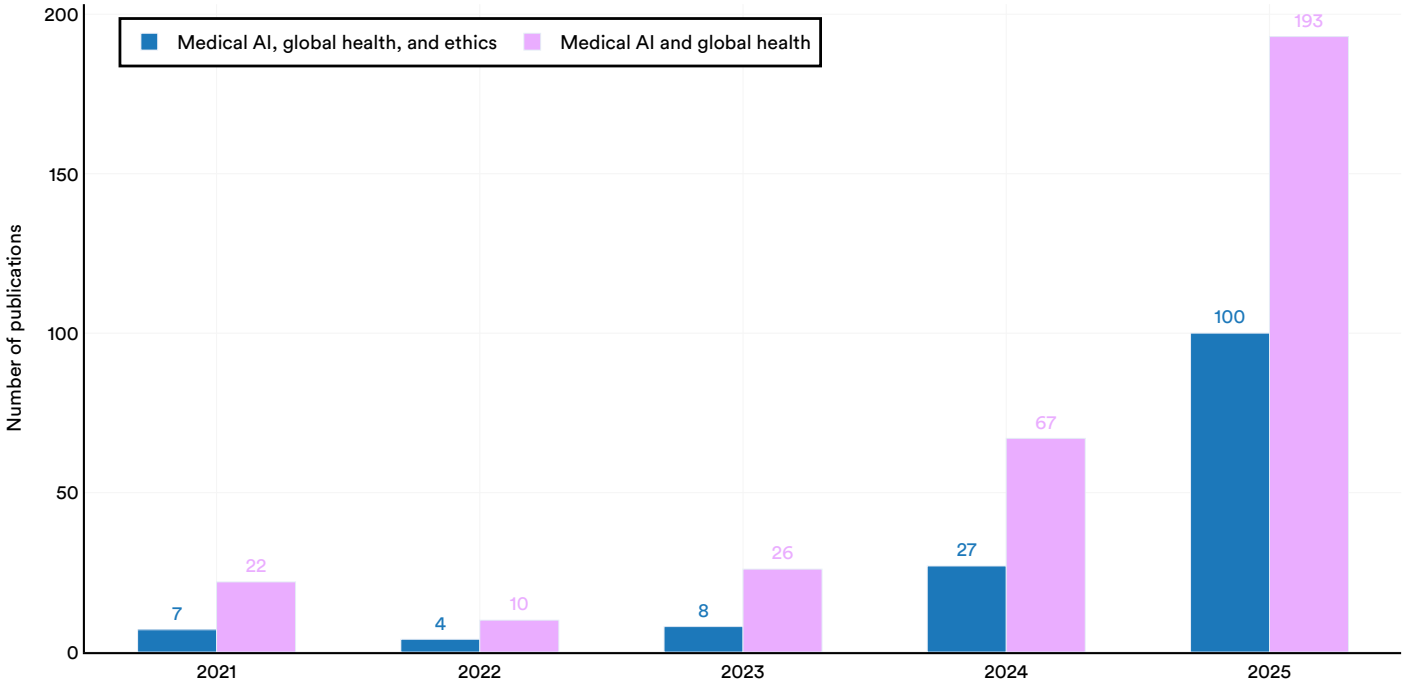


Figure 6.4.2

Number of medical AI, global health, and ethics publications, 2021–25

Source: RAISE Health, 2026 | Chart: 2026 AI Index report

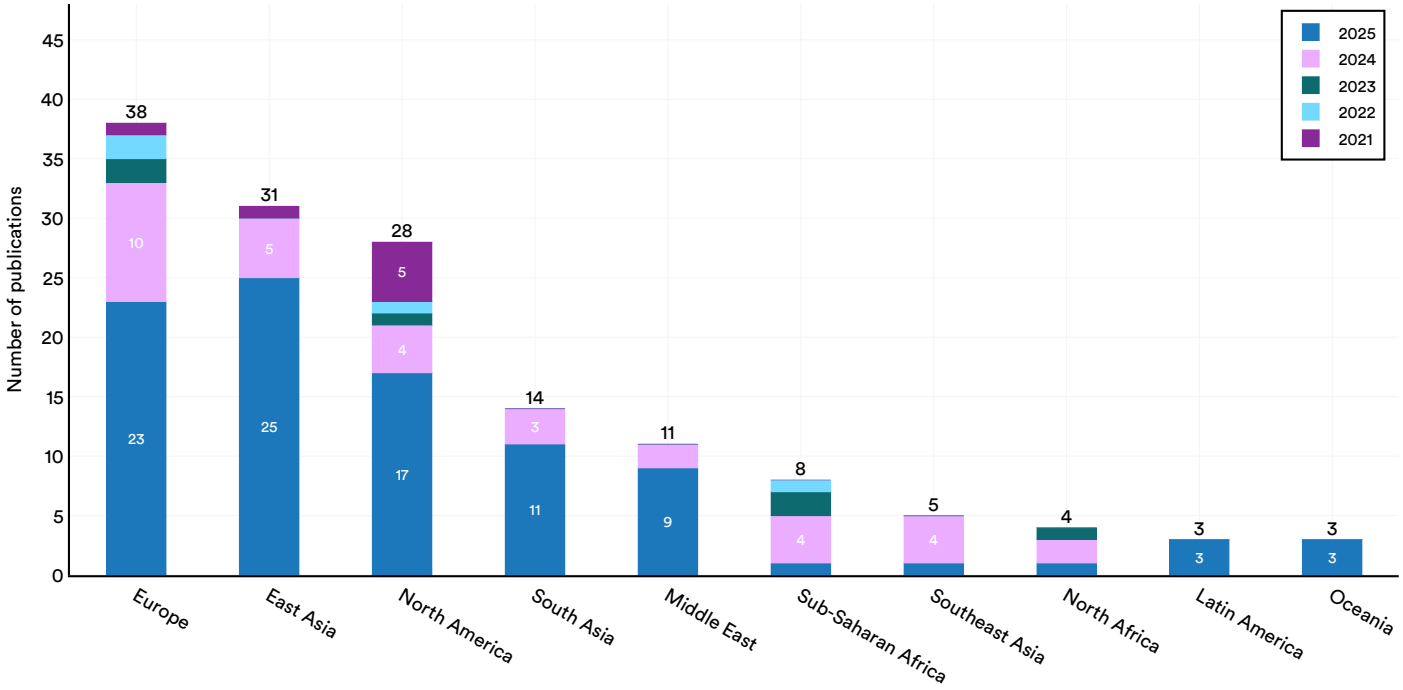


Figure 6.4.5

Number of medical AI and global health publications by ethics topics, 2021–25

Source: RAISE Health, 2026 | Chart: 2026 AI Index report

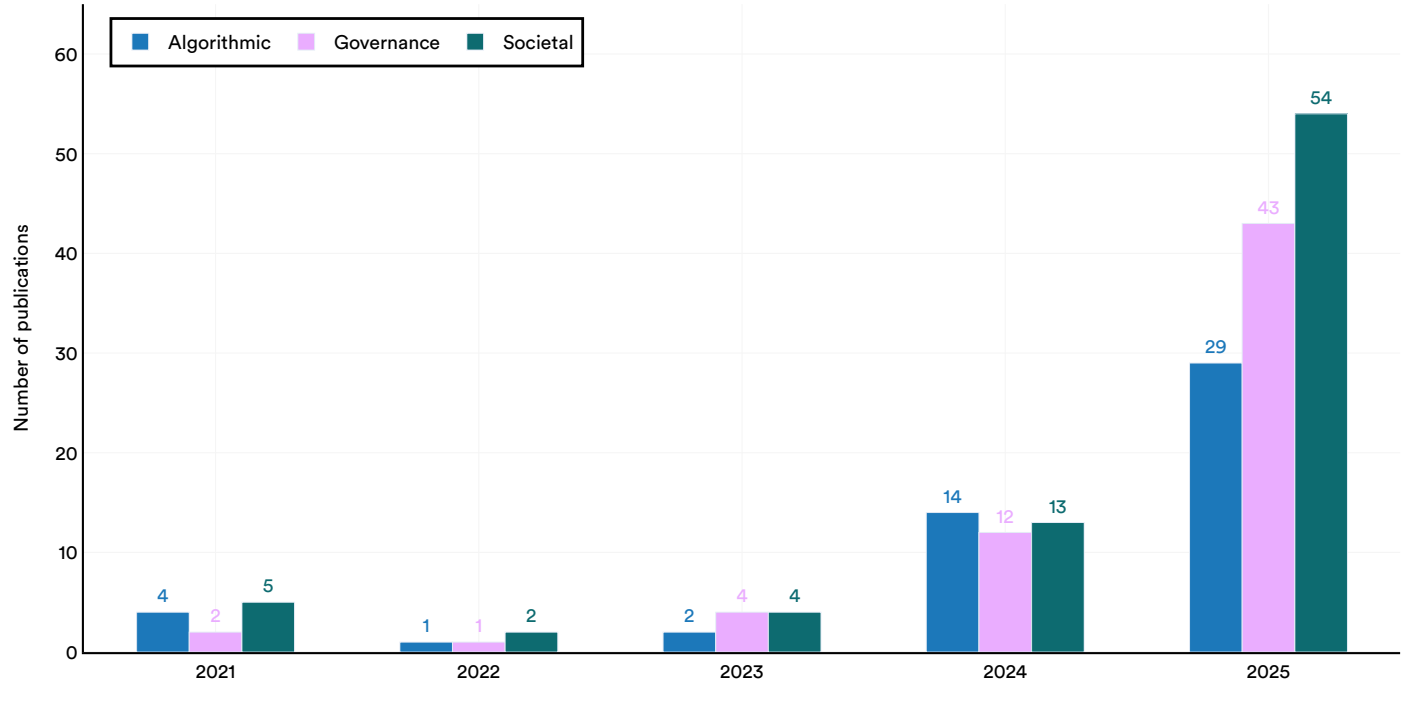


Figure 6.4.6