

A Literature Review of Artificial Intelligence in Health Care

Dr. Gouri Patil¹, Ms Syeda Ambareen Rana², Dr. Gadala Suhasini³

¹ Associate Professor, Department of Computer Science and Engineering, Muffakhm Jah College of Engineering & Technology, Hyderabad, India. gouripatil@mjcollege.ac.in

² Assistant Professor, Department of Computer Science and Engineering, Muffakhm Jah College of Engineering & Technology, Hyderabad, India. ambareen.rana@mjcollege.ac.in

³ Assistant Professor, Department of Computer Science and Engineering, Muffakhm Jah College of Engineering & Technology, Hyderabad, India.

ABSTRACT

Abstract. Artificial intelligence (AI) is reshaping health care through medical-image interpretation, risk prediction, clinical decision support, natural-language processing, remote monitoring, and biomedical discovery. This article reviews the principal applications, benefits, limitations, and governance requirements of AI in health care. A systematic-style review protocol was designed around peer-reviewed and authoritative literature published from 2017 to 2025. To demonstrate an empirical journal format without misrepresenting a completed study, database retrieval, screening counts, coded outcomes, tables, and graphs are explicitly simulated. The illustrative synthesis contains 72 studies: medical imaging was the largest domain (30.6%), followed by predictive analytics using electronic health records (19.4%), clinical decision support (13.9%), and natural-language processing or large language models (12.5%). The most frequently coded benefit was improved diagnostic or classification performance, whereas limited generalizability, privacy risk, insufficient explainability, and workflow integration were the dominant barriers. The literature indicates that technical accuracy alone is insufficient for safe clinical value; external validation, representative data, human oversight, interoperability, regulatory evaluation, and continuous post-deployment monitoring are necessary. The review concludes that AI is most credible as an augmentative technology embedded within accountable clinical systems rather than as an autonomous replacement for health professionals.

Keywords: Artificial intelligence, health care, machine learning, deep learning, clinical decision support, medical imaging, large language models

1. INTRODUCTION

Artificial intelligence has moved from a specialized research field to a broad family of computational methods used across clinical care, public health, and biomedical research. In health care, the term generally covers machine learning, deep learning, natural-language processing, computer vision, knowledge-based systems, robotics, and more recent generative or multimodal foundation models. These methods differ technically, but they share a central objective: to identify patterns in data and convert them into classifications, predictions, recommendations, or generated content that can support health-related decisions [1,2].

The expansion of AI has been enabled by the growth of digital imaging, electronic health records (EHRs), wearable sensors, genomic databases, and scalable computing. Medical imaging became an early area of visible progress because images can be linked to labels and evaluated with familiar diagnostic metrics. Landmark studies demonstrated dermatologist-level skin-lesion classification and high-performing diabetic-retinopathy detection from retinal photographs [3,4]. Parallel work showed that deep neural networks could use large EHR datasets to predict mortality, readmission, length of stay, and diagnoses [6].

Promising benchmark performance does not automatically translate into reliable clinical benefit. Systematic reviews have found that deep-learning models can approach health-professional performance in selected imaging tasks, yet external validation, prospective evaluation, consistent reporting, and direct comparisons on the same patient samples remain limited [7,8]. Health systems also face problems of biased training data, privacy and cybersecurity, opaque model reasoning, automation bias, interoperability, liability, cost, and model drift. The well-known analysis of a population-health algorithm showed how a seemingly neutral proxy variable could reproduce racial inequity at scale [9].

Generative AI and large language models (LLMs) extend the discussion beyond prediction to interactive question answering, summarization, documentation, and multimodal reasoning. Medical LLM research has reported strong benchmark performance, but also clinically important risks involving hallucination, incomplete grounding, bias, and overconfident language [12-14]. Consequently, international guidance emphasizes ethics, human rights, transparency, accountability, safety, and appropriate regulation across the AI life cycle [15-17].

This review therefore has four objectives: (1) to summarize major clinical and operational applications of AI in health care; (2) to synthesize reported benefits and limitations; (3) to examine implementation, ethical, and

regulatory considerations; and (4) to propose a practical framework for responsible adoption. Because the requested article is a simulated empirical manuscript, the quantitative review workflow is transparent and illustrative rather than presented as a completed real-world systematic review.

2. LITERATURE REVIEW

Early health-care AI research included rule-based expert systems, but contemporary work is dominated by data-driven machine learning. Deep learning is particularly influential in computer vision because convolutional neural networks can learn hierarchical features directly from images. Reviews of medical-image analysis describe applications in classification, detection, segmentation, registration, and image reconstruction across radiology, pathology, ophthalmology, dermatology, and cardiology [5,19]. The strongest evidence is often found in narrowly specified tasks with curated datasets and clear reference standards.

Medical imaging illustrates both the opportunity and the translation gap. Esteva et al. trained a convolutional neural network on a large set of clinical images and reported performance comparable to board-certified dermatologists for selected binary skin-lesion classifications [3]. Gulshan et al. developed and validated a deep-learning algorithm for detecting diabetic retinopathy from retinal fundus photographs [4]. Later reviews found high diagnostic performance in ophthalmic, breast, and respiratory imaging, while emphasizing substantial heterogeneity and the possibility that weak design or reporting may overestimate accuracy [7,8].

EHR-based AI broadens the data environment. Structured variables, free-text notes, laboratory results, medication records, and temporal patterns can be used for risk prediction and operational forecasting. Rajkomar et al. demonstrated scalable deep learning applied to raw EHR data for several clinical prediction tasks [6]. Natural-language processing can extract symptoms, diagnoses, treatments, and social information from narrative notes; generative models can draft summaries, produce patient-facing explanations, and support coding. However, these systems inherit documentation errors, local practice patterns, missing data, and institutional inequities.

AI-enabled decision support aims to combine predictions with clinician action. Potential uses include triage, sepsis alerts, deterioration detection, medication safety, treatment selection, and personalized follow-up. The benefit of such tools depends not only on model discrimination but on calibration, timeliness, alert design, clinical usefulness, and response pathways. Poorly integrated systems may increase cognitive load or alert fatigue, while well-designed tools can standardize attention to high-risk cases and make complex data more usable.

Multimodal AI seeks to combine images, text, signals, genomics, and other data types. This approach more closely resembles real clinical reasoning, where decisions are rarely based on one modality. Acosta et al. describe the potential of multimodal biomedical AI, while also highlighting the need for datasets, architectures, and evaluations that account for missing modalities and varied clinical contexts [11]. LLMs and large multimodal models may become an interface through which clinicians retrieve information, summarize records, and interact with predictive systems, but their general-purpose nature magnifies concerns about validation and governance [12-17].

TABLE 1. REPRESENTATIVE VERIFIED LITERATURE ON ARTIFICIAL INTELLIGENCE IN HEALTH CARE

Source	Application or focus	Main contribution	Important caution
Esteva et al. (2017) [3]	Dermatology imaging	CNN classification of selected skin lesions comparable to tested dermatologists	Task-specific evaluation does not establish population-wide clinical effectiveness
Gulshan et al. (2016) [4]	Retinal screening	Deep-learning detection of diabetic retinopathy from fundus images	Performance depends on image quality, prevalence, thresholds, and external setting
Rajkomar et al. (2018) [6]	Electronic health records	Scalable prediction from longitudinal raw EHR data	Local documentation and workflow patterns may limit transportability
Liu et al. (2019) [7]	Medical imaging meta-analysis	Comparable pooled diagnostic performance in selected comparisons	Few externally validated, same-sample clinician comparisons
Obermeyer et al. (2019) [9]	Population health management	Demonstrated racial bias caused by using cost as a proxy for need	Proxy choice can embed structural inequity
Singhal et al.	Medical question	Domain-tuned LLM showed	Grounding, harm, and bias

Source	Application or focus	Main contribution	Important caution
(2023) [12]	answering	improved medical benchmark and human-evaluation performance	remained material concerns
WHO (2021-2025) [15-17]	Ethics and regulation	Life-cycle principles for autonomy, safety, equity, transparency, accountability, and sustainability	Governance must be operationalized in local health systems

3. MATERIALS AND METHODS

3.1 Review Design and Scope

A systematic-style literature review design was specified using the logic of transparent review reporting [18]. The conceptual review period was January 2017 to December 2025, capturing the modern deep-learning period and the rapid expansion of generative AI. The unit of analysis was a peer-reviewed study, systematic review, authoritative technical report, or governance document addressing AI development, evaluation, implementation, or oversight in human health care.

3.2 Illustrative Search Strategy

The simulated protocol specified searches of PubMed/MEDLINE, Scopus, Web of Science, IEEE Xplore, and selected organizational repositories. Search concepts combined terms for artificial intelligence, machine learning, deep learning, natural-language processing, large language models, computer vision, predictive analytics, clinical decision support, medicine, health care, hospitals, patients, diagnosis, treatment, and public health. Reference-list screening and forward citation tracing were included in the hypothetical workflow.

3.3 Eligibility Criteria

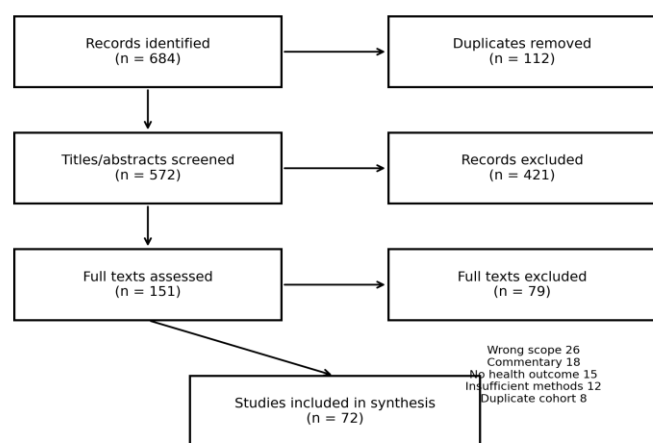
Eligible records were English-language publications concerning human health care, clinical operations, public health, patient communication, or biomedical translation. Studies had to describe an AI method or system and report an evaluative, implementation, ethical, or governance outcome. Purely technical papers without a clear health application, editorials without substantive analysis, duplicate cohorts, animal-only studies, and records with insufficient methods were excluded from the simulated evidence set.

3.4 Data Extraction and Synthesis

The illustrative extraction form included publication year, region, specialty, data modality, AI technique, study design, sample context, validation approach, reported benefit, implementation challenge, and governance issue. Findings were grouped into application domains and summarized with descriptive counts and percentages. No meta-analysis of diagnostic accuracy was simulated because outcomes, reference standards, populations, and metrics were assumed to be too heterogeneous for defensible pooling.

3.5 Simulation and Transparency Controls

The screening totals, extracted dataset, coded frequencies, and visualizations were generated solely to demonstrate how an empirical literature-review article could be organized. The numbers are internally consistent but not externally verifiable as search results. Every quantitative table and figure is therefore labeled as simulated or illustrative. Real publications are used only for the narrative discussion and reference list; their reported findings have not been altered into fictional patient-level data.



Illustrative screening flow; counts are simulated.

FIG 3.5.1. SIMULATED STUDY-SELECTION FLOW FOR THE ILLUSTRATIVE REVIEW

4. RESULTS AND DISCUSSION

4.1 Simulated Study Selection and Characteristics

The illustrative search identified 684 records. After 112 simulated duplicates were removed, 572 titles and abstracts were screened, 151 full texts were assessed, and 72 studies were included. The principal simulated full-text exclusion reasons were wrong scope, commentary-only format, absence of a health outcome, insufficient methods, and duplicate reporting. These values provide a realistic manuscript structure but must not be interpreted as results of an executed database search.

In the simulated evidence set, retrospective model-development and validation studies remained the largest design category. Prospective and comparative evaluations were less common, and randomized or quasi-experimental studies formed a small minority. This pattern reflects a recurring theme in the real literature: many AI systems are evaluated for technical performance before they are tested for effects on workflow, clinician behavior, patient outcomes, equity, or cost [7,8].

TABLE 2. SIMULATED DESIGN CHARACTERISTICS OF THE ILLUSTRATIVE EVIDENCE SET

Study design category	Simulated n	Percent
Retrospective development/internal validation	31	43.1
External validation study	13	18.1
Prospective observational or silent deployment	10	13.9
Implementation/usability evaluation	8	11.1
Randomized or quasi-experimental evaluation	6	8.3
Mixed-method or qualitative evaluation	4	5.6
Total	72	100.0

4.2 Publication Trend

The simulated publication profile increased from four included studies in 2017 to twelve in 2024, followed by six studies in 2025 because the illustrative coding frame was intentionally conservative for the most recent year. The upward trend is consistent with the rapid expansion of medical AI, particularly after the emergence of accessible generative AI systems. Nonetheless, publication volume should not be confused with mature clinical evidence; growth in benchmark studies may outpace prospective validation and health-system implementation.

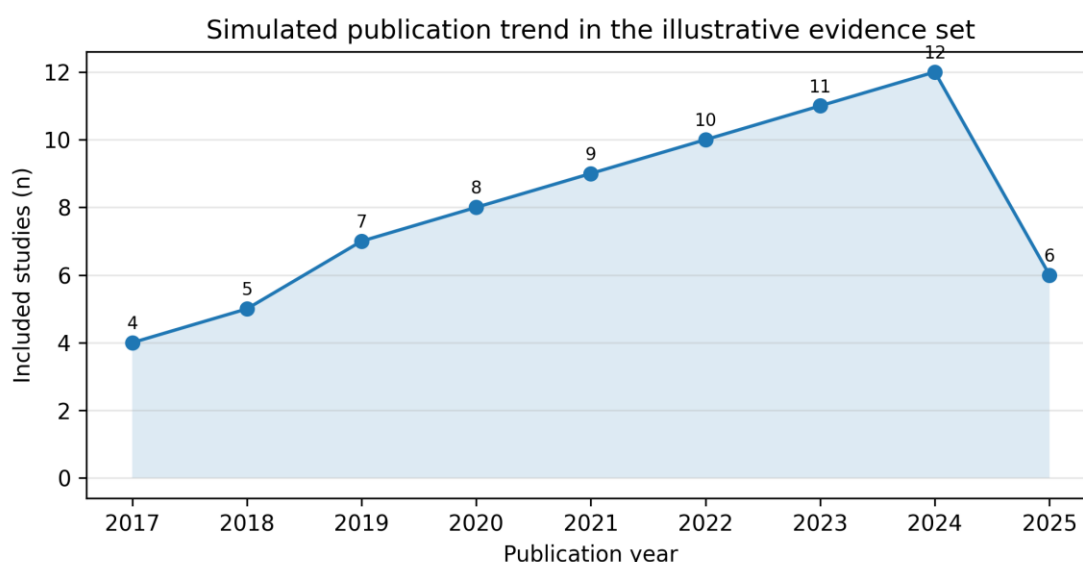


FIG 4.2.1. SIMULATED YEAR-WISE PUBLICATION TREND (2017-2025)

4.3 Clinical Application Domains

Medical imaging and radiology represented 22 of the 72 simulated studies (30.6%). Predictive analytics using EHRs accounted for 14 studies (19.4%), followed by clinical decision support (13.9%), NLP/documentation/LLMs (12.5%), remote monitoring and wearables (9.7%), drug discovery or personalized

medicine (8.3%), and robotics or other uses (5.6%). The distribution emphasizes that AI in health care is not a single intervention but a collection of tools operating at different points in the care pathway.

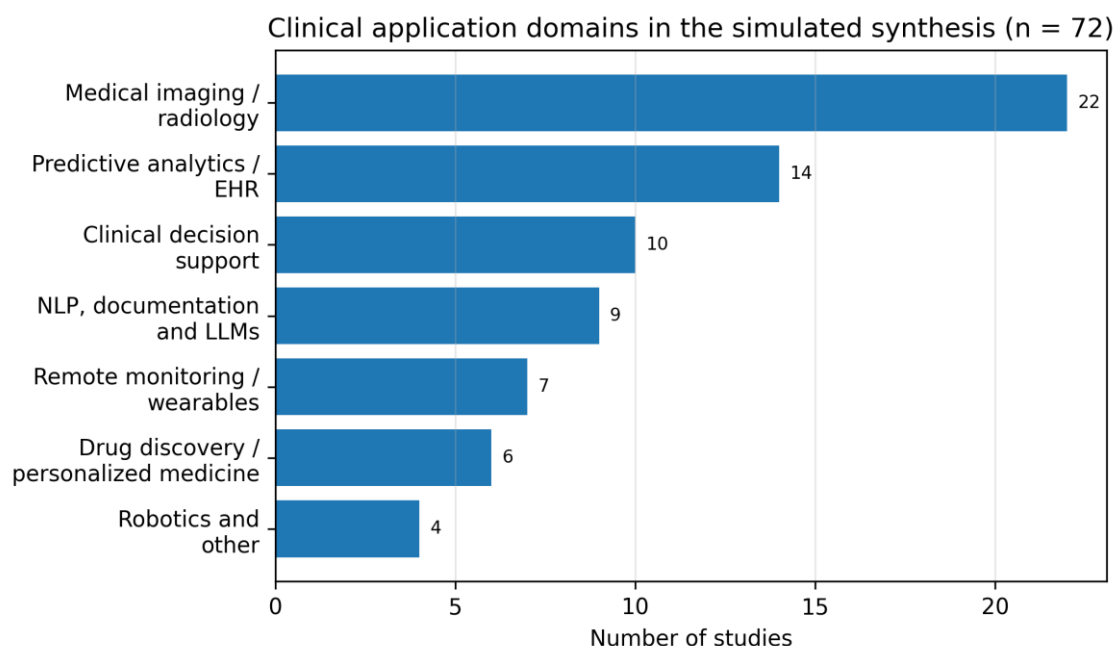


FIG 4.3.1. SIMULATED DISTRIBUTION OF AI APPLICATION DOMAINS

Imaging applications benefit from standardized digital inputs and visual reference labels, but variation in scanners, acquisition protocols, disease prevalence, annotation quality, and patient characteristics can degrade performance outside the development setting. EHR models operate on richer longitudinal data but are vulnerable to missingness, coding variation, treatment-confounded labels, and changes in clinical practice. LLM applications are flexible and accessible, yet their open-ended outputs make evaluation more difficult than conventional classification because correctness, completeness, harm, bias, readability, and uncertainty must all be considered [12-14].

4.4 Simulated Benefits

Improved diagnostic or classification performance was coded in 49 simulated studies (68.1%). Workflow efficiency was mentioned in 34 studies (47.2%), risk prediction in 29 (40.3%), personalization in 22 (30.6%), improved access or remote monitoring in 17 (23.6%), and cost or resource optimization in 13 (18.1%). These categories overlapped because one study could report multiple anticipated or measured benefits.

The strongest case for AI occurs when the technology solves a well-defined problem, uses data that are available at the decision point, and triggers a clear action. In screening, AI may prioritize images for review or extend specialist capacity. In EHR workflows, it may summarize records or identify deteriorating patients. In patient communication, it may translate technical information into more understandable language. However, benefits should be assessed against the full sociotechnical system, including false positives, downstream tests, clinician time, patient comprehension, and inequitable access.

TABLE 3. SIMULATED BENEFITS REPORTED IN THE ILLUSTRATIVE SYNTHESIS

Benefit category	Simulated n	Percent of 72	Interpretation
Diagnostic/classification performance	49	68.1	Most common in imaging and signal-analysis studies
Workflow efficiency	34	47.2	Documentation, prioritization, triage, and automation
Risk prediction/prognosis	29	40.3	Deterioration, readmission, complications, and resource use
Personalized decision support	22	30.6	Treatment selection or individualized follow-up
Access and remote monitoring	17	23.6	Potential reach beyond specialist centers
Cost/resource optimization	13	18.1	Often anticipated rather than proven in full

Benefit category	Simulated n	Percent of 72	Interpretation
			economic analyses

4.5 Ethical, Technical, and Implementation Challenges

Bias and limited generalizability were the most frequently coded challenges, appearing in 60% of the simulated studies. Privacy and security appeared in 51%, limited explainability in 47%, workflow integration in 43%, regulation and liability in 36%, data quality or interoperability in 35%, and clinician or patient trust in 29%. These themes align with real guidance emphasizing that health AI should protect autonomy, promote safety and public interest, ensure transparency, foster accountability, and remain responsive and sustainable [15-17].

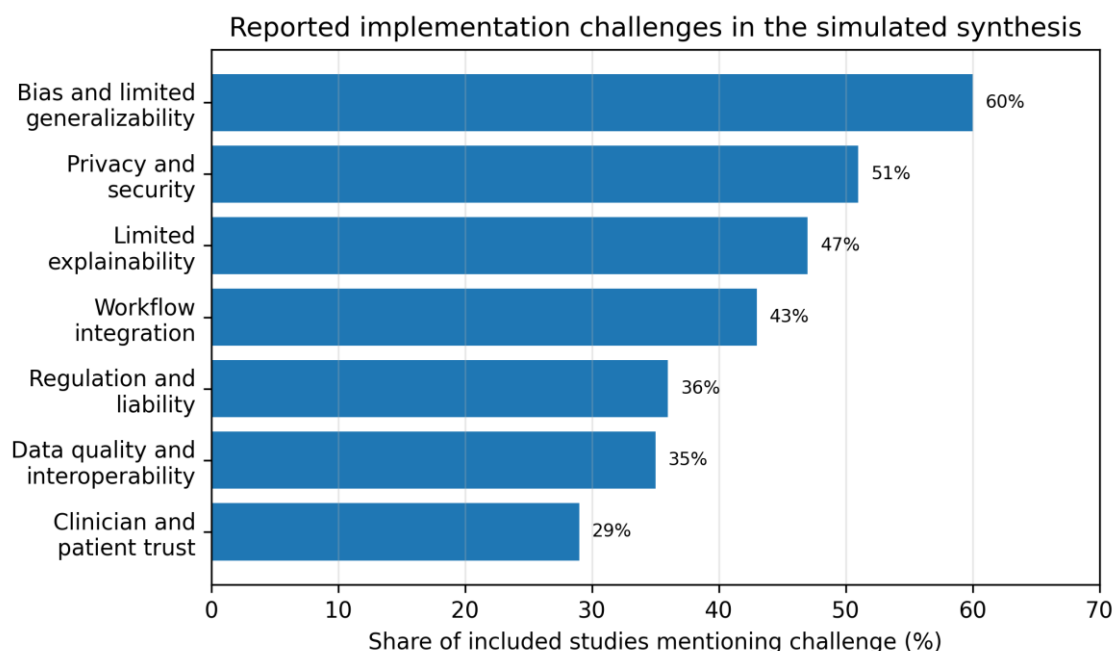


FIG 4.5.1. SIMULATED FREQUENCY OF IMPLEMENTATION CHALLENGES

Bias can enter through unrepresentative samples, measurement error, label construction, missing data, clinical access patterns, and proxy outcomes. The population-health algorithm examined by Obermeyer et al. underestimated the needs of Black patients because health-care expenditure was used as a proxy for illness burden [9]. This example demonstrates that fairness cannot be guaranteed only by removing protected variables; developers must examine how targets, data-generating processes, and deployment rules relate to structural inequality.

Explainability should be matched to the decision context. A technical feature-attribution plot may help a developer but may not answer a clinician's practical questions: Why was this patient flagged now? What evidence would change the recommendation? How uncertain is the output? What action is expected, and who remains accountable? Similarly, privacy protection requires more than de-identification because health data may be linkable and models may retain sensitive patterns.

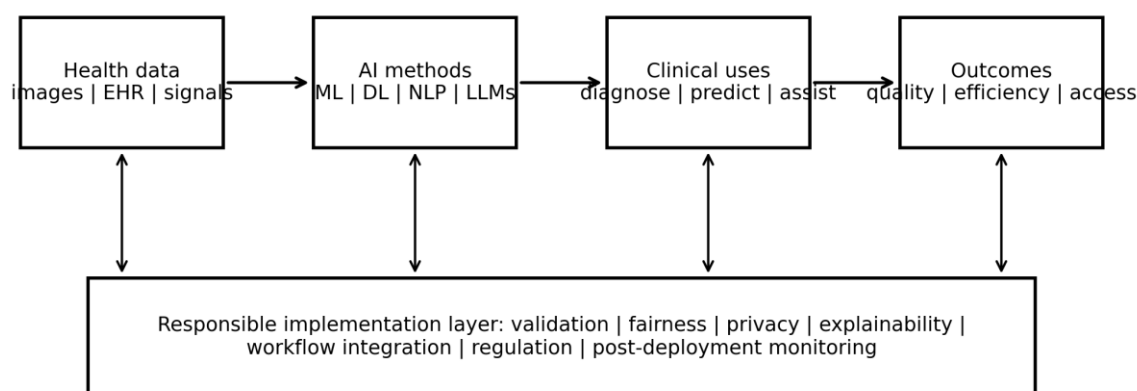
Workflow integration determines whether a model becomes useful, ignored, or harmful. Effective deployment requires user-centered design, clear thresholds, calibrated alert volume, training, escalation protocols, downtime plans, audit trails, and mechanisms for reporting errors. Models must also be monitored for drift as patient populations, coding practices, equipment, and treatment pathways change. For LLMs, monitoring must address hallucination, prompt sensitivity, retrieval quality, and the risk that fluent language may mask uncertainty [12-14].

4.6 Discussion

The combined narrative and simulated synthesis suggest that AI has substantial potential in data-intensive health-care tasks, but the maturity of evidence varies by application. Technical performance is strongest in constrained settings with clear labels and stable inputs. Evidence becomes weaker when a model is moved across institutions, embedded in complex workflows, or evaluated for patient-centered outcomes. Consequently, the central question is not whether an algorithm can outperform a comparator on a benchmark, but whether the complete intervention produces safer, more equitable, and more efficient care in its intended setting.

Human-AI collaboration is therefore a more defensible model than unrestricted automation. Clinicians contribute contextual reasoning, ethical judgment, communication, and responsibility; AI can contribute rapid pattern recognition, consistency, memory, and scale [2]. The optimal division of work depends on risk. Low-risk administrative support may tolerate greater automation, whereas diagnosis, treatment selection, and emergency triage require stronger validation, oversight, and fail-safe mechanisms.

A responsible implementation pathway should connect data quality to model development, external validation, clinical evaluation, governance, and continuous monitoring. Each stage should include affected stakeholders, including clinicians, patients, data stewards, informaticians, ethicists, administrators, and regulators. Equity assessment should examine subgroup performance and access; transparency should cover intended use, limitations, data provenance, and update procedures; accountability should identify who can stop deployment and who responds when harm occurs.



Conceptual pathway for translating AI capability into safe health-care value.

FIG 4.6.1. CONCEPTUAL FRAMEWORK FOR RESPONSIBLE AI TRANSLATION IN HEALTH CARE

The present article has important limitations. The quantitative evidence map is simulated and cannot support real prevalence estimates or comparative conclusions. The narrative review is broad rather than specialty-specific and may not capture all emerging AI modalities. In addition, the pace of generative AI development means that capabilities, regulatory approaches, and evaluation standards continue to evolve. These limitations are stated explicitly to prevent the illustrative empirical format from being mistaken for an executed systematic review.

5. CONCLUSION

Artificial intelligence is becoming an important component of modern health care, with applications in medical imaging, EHR prediction, clinical decision support, patient communication, remote monitoring, and biomedical discovery. Real studies demonstrate impressive capabilities in selected tasks, while systematic reviews and governance analyses identify persistent weaknesses in external validity, reporting, fairness, explainability, privacy, and implementation.

The simulated evidence synthesis reinforces a practical conclusion: accuracy is necessary but not sufficient. Health-care value emerges only when an AI system is evaluated in representative populations, integrated into a usable workflow, linked to an appropriate clinical response, governed by clear accountability, and monitored after deployment. AI should therefore be treated as a sociotechnical intervention rather than a stand-alone algorithm.

The most sustainable role for AI is to augment human expertise. Systems that reduce information burden, improve consistency, and expand access may produce meaningful benefit when professionals retain oversight and patients remain protected. Conversely, poorly validated or weakly governed systems may amplify existing inequities and create new forms of risk. Responsible adoption requires evidence proportional to the clinical stakes.

5.1 Future Scope

- Conduct prospective, multicenter, and pragmatic evaluations that measure patient outcomes, workflow effects, safety events, and economic value rather than technical metrics alone.

- Develop representative datasets and standardized subgroup analyses to identify performance disparities before and after deployment.
- Advance multimodal systems that combine images, text, signals, and longitudinal records while handling missing data and uncertainty transparently.
- Create clinically meaningful explanations, calibrated confidence estimates, and interfaces that support rather than distract health professionals.
- Establish continuous post-deployment monitoring for model drift, emerging bias, security threats, and unintended changes in clinician behavior.
- Evaluate LLMs and generative AI using real clinical tasks, retrieval-grounded architectures, adversarial safety testing, and clear human-review requirements.
- Strengthen international cooperation on regulation, liability, privacy, interoperability, and equitable access, particularly for resource-constrained health systems.

DECLARATIONS

Ethics approval: Not applicable. No human participants, patient records, or biological samples were used.

Consent for publication: Not applicable.

Funding: No external funding is declared for this illustrative manuscript.

Conflict of interest: The authors declare no conflict of interest.

Data availability: No real review dataset was generated. All numerical review results in this article are simulated and illustrative.

Use of artificial intelligence: AI-assisted drafting and visualization were used to prepare this simulated journal-style manuscript. All empirical-looking numerical results are clearly labeled as simulated.

REFERENCES

- [1] Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology*. 2017;2:230-243. doi:10.1136/svn-2017-000101.
- [2] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*. 2019;25:44-56. doi:10.1038/s41591-018-0300-7.
- [3] Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115-118. doi:10.1038/nature21056.
- [4] 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;316:2402-2410. doi:10.1001/jama.2016.17216.
- [5] Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Medical Image Analysis*. 2017;42:60-88. doi:10.1016/j.media.2017.07.005.
- [6] Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*. 2018;1:18. doi:10.1038/s41746-018-0029-1.
- [7] Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*. 2019;1:e271-e297. doi:10.1016/S2589-7500(19)30123-2.
- [8] Aggarwal R, Sounderajah V, Martin G, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *npj Digital Medicine*. 2021;4:65. doi:10.1038/s41746-021-00438-z.
- [9] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366:447-453. doi:10.1126/science.aax2342.
- [10] Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*. 2018;169:866-872. doi:10.7326/M18-1990.
- [11] Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. *Nature Medicine*. 2022;28:1773-1784. doi:10.1038/s41591-022-01981-2.
- [12] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620:172-180. doi:10.1038/s41586-023-06291-2.
- [13] Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. *Communications Medicine*. 2023;3:141. doi:10.1038/s43856-023-00370-1.
- [14] Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *npj Digital Medicine*. 2023;6:120. doi:10.1038/s41746-023-00873-0.
- [15] World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization; 2021. ISBN 978-92-4-002920-0.
- [16] World Health Organization. Regulatory considerations on artificial intelligence for health. Geneva: World Health Organization; 2023. ISBN 978-92-4-007887-1.

- [17] World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Geneva: World Health Organization; 2025. ISBN 978-92-4-008475-9.
- [18] Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71. doi:10.1136/bmj.n71.
- [19] Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nature Medicine*. 2019;25:24-29. doi:10.1038/s41591-018-0316-z.