



Review Article

Copyright © All rights are reserved by Avtandil Gagnidze

Artificial Intelligence in Medicine: Opportunities, Risks, and Prospects for Georgia

Avtandil Gagnidze* and Mari Manjgaladze

East European University, East-West University, Tbilisi, Georgia

*Corresponding author: Avtandil Gagnidze, East European University, Tbilisi, Georgia.

To Cite This article: Avtandil Gagnidze* and Mari Manjgaladze, *Artificial Intelligence in Medicine: Opportunities, Risks, and Prospects for Georgia*. *Am J Biomed Sci & Res*. 2026 32(5) AJBSR.MS.ID.004192, DOI: [10.34297/AJBSR.2026.32.004192](https://doi.org/10.34297/AJBSR.2026.32.004192)

Received: 📅 October 17, 2026 **Published:** 📅 October 23, 2026

Abstract

Background: Artificial Intelligence (AI) has evolved in clinical practice from an experimental technology into an everyday tool. By the end of 2025, the FDA list included 1,451 AI-enabled medical devices, approximately 76% of them in radiology. At the same time, the large language models (LLMs) most frequently used by physicians have not been authorized as medical devices.

Objective: This article assesses the opportunities offered by medical AI today, the limits of its reliability, and the conditions required for its responsible implementation in Georgia.

Methods: This narrative review draws on the author's textbook [18], the clinical studies discussed therein, a targeted literature search, and national and international institutional sources on Georgia's healthcare system.

Results: Diagnostic systems (Aidoc, IDx-DR, Paige Prostate, and DeepFAN) demonstrate high accuracy in narrowly defined tasks; in real-world practice, however, performance depends on image quality, population characteristics, and integration into workflows. The history of Clinical Decision Support Systems (CDSS)—including IBM Watson for Oncology and Epic's sepsis model—shows that implementation without independent validation is risky. In Georgia, the principal barriers include incomplete adoption of Electronic Health Records (EHRs), fragmentation of the private sector, regional inequalities, limited Georgian-language support, and an evolving regulatory framework.

Conclusion: AI does not replace physicians; it augments them. A phased approach is recommended for Georgia: AI literacy, low-risk pilots in "silent mode," local validation, an anonymized national database, and a legal framework harmonized with the European Union.

Keywords: Artificial intelligence, Clinical decision support, Large language models, Algorithmic bias, Digital health, Georgia

Introduction

Medicine is an information-intensive field: physicians process symptoms, laboratory data, images, and continually updated research literature every day. AI is particularly valuable in managing this volume of information: it accelerates image analysis, automates routine documentation, and predicts changes in patients' conditions [31,39]. However, the history of medical AI is not an uninterrupted success story. As early as the 1970s, the expert system MYCIN performed at the level of infectious-disease specialists in a blinded evaluation, yet it was never implemented in clinical practice [5,45]. Half a century later, the failure of IBM Watson for Oncology and the independent evaluation of Epic's sepsis model [44] reiterated the same lesson: technology that works in a laboratory may fail in a real

clinical setting. The central question is therefore no longer "Can AI diagnose?" but "For which task, under what conditions, and under whose responsibility?" The article's central thesis is the "human + AI" model: the best outcomes arise when AI does what it does well—speed, processing large datasets, and pattern recognition—while physicians contribute their strengths: clinical judgment, empathy, and an understanding of context and nuance [18].

This article addresses three questions:

- What types of AI systems are currently used in clinical practice, and what evidence supports them? (Sections 3–5)
- What ethical, technical, regulatory, and environmental risks accompany their use? (Section 6)

c. What conditions are required for the responsible implementation of medical AI in Georgia? (Sections 7–8)

Methodology

Type of Review

This article is a narrative review. Unlike a systematic review, it does not claim to cover all available evidence; its purpose is conceptual synthesis and identification of gaps in a rapidly evolving field characterized by substantial methodological heterogeneity [23].

Three Groups of Sources

- a. Conceptual Framework: the author’s textbook, Introduction to Artificial Intelligence for Health Professions [18] (English version [19]), and classic sources [34,39,40]. The textbook informs the conceptual framework and selection of systems, while empirical claims are based, where possible, on independent peer-reviewed sources.
- b. Clinical Evidence: a targeted search of PubMed/MEDLINE and Google Scholar. The search was conducted in April–September 2026. Screening was carried out in three stages. Title and abstract screening: the search strings retrieved approximately 600 records. After removal of duplicates and title assessment, approximately 400 records were excluded. After abstract assessment, a further approximately 150 records were excluded. Finally, about 50 sources were used. Search terms were divided into two groups and combined using AND: (A) “artificial intelligence” OR “machine learning” OR “deep learning” OR “large language model”; (B) “diagnostic accuracy” OR “clinical decision support” OR “external validation” OR “algorithmic bias” OR “ambient scribe”. The main time frame was 2018–2026; this restriction did not apply to historical and conceptual sources.

c. Institutional and Regulatory Documents: the FDA list of AI-enabled medical devices, EU regulations, and reports from the State Audit Office of Georgia, the European Observatory on Health Systems and Policies, WHO, and EUAA, selected purposively. System-selection criteria. Specific AI systems were selected according to three criteria: (1) regulatory status (FDA clearance/authorization, CE marking, or MHRA registration); (2) availability of clinical data; and (3) illustrative value, including lessons from both success and failure.

System-Selection Criteria: Specific AI systems were selected according to three criteria: (1) regulatory status (FDA clearance/ authorization, CE marking, or MHRA registration); (2) availability of clinical data; and (3) illustrative value, including lessons from both success and failure.

Limitations

Screening was not conducted by two independent reviewers, increasing the risk of selection bias. The medical AI market changes rapidly: the figures presented reflect the state of the field in 2026, and their clinical use should be checked against current manufacturer and regulator information. Manufacturer-sponsored studies are not treated as independent validation. Claims concerning Georgia rely partly on international reports because local empirical data are limited.

Conceptual Foundations

Key concepts

Physicians do not need to know how to program in order to critically assess medical AI, but they should understand how a system learns and where its capabilities end. The concepts are nested: AI ⊃ machine learning ⊃ neural networks ⊃ deep learning. Their meanings, clinical examples, and limitations are summarized in (Table 1).

Table 1: Key AI concepts in a clinical context.

Concept	Definition	Clinical Example	Main Limitation
Artificial intelligence (AI)	Systems that perform tasks associated with human intelligence; all current systems are forms of “narrow AI.”	Any of the systems listed below	The label “AI” groups together very different technologies.
Machine learning (ML)	A system learns rules from data (supervised, unsupervised, or reinforcement learning).	Sepsis-risk prediction from EHR data	May reproduce bias in historical data.
Deep learning (DL)	Multilayer neural networks; particularly effective with images.	Detecting hemorrhage on CT	A “black box”: decisions can be difficult to explain.
Natural language processing (NLP)	Understanding and processing human language.	Structuring clinical notes	Georgian language and terminology are weakly supported.
Generative AI and LLMs	Generation of new text, images, or code through statistical prediction of the next token.	AI scribes; Open Evidence	“Hallucinations”: plausible but false information.

The “black box” and Levels of Autonomy

Two concepts are particularly important for clinicians. The first is the “black-box” problem: the billions of parameters in a deep neural network cannot be inspected individually, yet physicians must be able to answer “why?”—both to justify their decisions

and to explain them to patients. Explainable AI (XAI) offers one response; for example, it can highlight a suspicious area in an image. The second concept is the level of autonomy: who makes the final decision? The overwhelming majority of medical AI systems are assistive; fully autonomous systems constitute a small and tightly bounded category (Table 2).

Table 2: Levels of autonomy in medical AI systems.

Level	System Function	Decision-Maker	Example
Triage	Flags suspicious cases and prioritizes them in the work queue	Physician, in every case	Aidoc; Nanox.AI
Second reader	Highlights suspicious areas in an image	Physician, taking the AI indication into account	Paige Prostate; mammography AI
Decision support	Assesses risk or suggests treatment options	Physician or multidisciplinary team	Epic Sepsis Model; IBM Watson
Autonomous	Produces a screening result without the physician viewing the image	System, within strictly defined limits	IDx-DR (LumineticsCore)

Can AI “understand”?

A philosophical question precedes the technical and ethical ones: does a system understand what it is doing? Alan Turing framed this as a behavioral test: if an interviewer cannot distinguish a machine from a human in a text-based exchange, debate about machine “thinking” becomes less meaningful [40]. John Searle responded with the “Chinese Room” argument: a person who does not know Chinese but selects the correct characters by following a rulebook may appear to understand the language, although they understand none of the words. Manipulating symbols (syntax) does not, by itself, produce meaning (semantics) [34]. This debate has resurfaced with contemporary LLMs. Bender and colleagues call LLMs “stochastic parrots”—systems that statistically assemble linguistic patterns without verifying their meaning or truth [4]. The debate remains open, but the clinical implication is clear even without resolving it: fluent language is not evidence of truth. This is the basis of the risk of hallucination (Section 6.4) and a reason not to treat an LLM response as equivalent to a colleague’s consultation.

Diagnostic AI Systems

Diagnostics, particularly image analysis, is the most mature area of AI: images are well-structured data, and deep learning performs strongly in pattern recognition. Of the 1,451 AI-enabled devices cleared/authorized by the FDA, 1,104 are in radiology, 141 in cardiology, and 67 in neurology [18]. How to interpret the numbers. Sensitivity indicates the proportion of patients with disease whom the system identifies; specificity indicates the proportion of healthy individuals correctly classified as healthy. Both measures should be interpreted in light of disease prevalence: for rare diseases, false alarms may outnumber true cases even when specificity is high. AUC measures discrimination alone; calibration and threshold performance must also be reported to assess clinical utility [6].

Radiology

Aidoc: analyzes CT scans in the background and moves a study to the top of the radiologist’s worklist when Intracranial

Hemorrhage (ICH) is suspected. In data submitted to the FDA, the ICH algorithm had 93.6% sensitivity and 92.3% specificity (198 cases) [11]; a later version achieved 96.2% and 94.8%, respectively [13]. In real-world clinical practice, the system was evaluated at University Hospital Basel. Of 4,450 patients undergoing emergency head CT, 3,017 were scanned after AI implementation; hemorrhage prevalence was almost identical before and after implementation (14.4% and 14.3%). Overall, AI accuracy was 93.0%, sensitivity 87.2%, specificity 93.9%, negative predictive value 97.8%, and positive predictive value only 70.5% [35]. After implementation, the time to notification of acute hemorrhage fell from 73 to 58 minutes, while the time to rule out hemorrhage in the emergency department during working hours decreased from 205 to 167 minutes. Overall report turnaround time was essentially unchanged (60 versus 63 minutes) [35]. This illustrates the principle above: real-world sensitivity (87.2%) was lower than the figure submitted to the FDA, and a positive predictive value of 70.5% means that roughly one in three or four alerts was false. The system’s practical benefit, however, lay not in increased accuracy but in accelerating workflow—the purpose of a triage system.

Google’s Mammography AI: In the first major retrospective study using data from the UK OPTIMAM database and the United States, the system reduced false-positive results by 1.2 percentage points and false-negative results by 2.7 points in UK data compared with expert radiologists. In a simulation in which AI served as the second reader, the second reader’s workload fell by 88% [28]. In a large NHS study published in 2026 (~175,000 women), AI as a second reader increased cancer detection from 7.54 to 9.33 per 1,000 women, reduced recalls by 39.3% among women attending their first screening, and detected 25% of interval cancers missed by double reading [41]. In the same study, the arbitration panel overruled 93 correct AI decisions that would have led to patient recall [22]—another example of challenges in human–AI interaction. Some of these studies are retrospective or simulation-based; replication in real-world clinical practice requires further evaluation.

Multi Disease AI: platforms move beyond the “one model—one problem” principle. a2z-Unified-Triage received FDA clearance in November 2025. For patients aged 22 years or older, it flags seven acute conditions on abdominal and pelvic CT: small-bowel obstruction, acute cholecystitis, acute pancreatitis, acute diverticulitis, hydronephrosis, free intraperitoneal air, and unruptured aortic aneurysm [14]. Annalise Enterprise CXR can identify up to 124 findings on chest radiographs [33], although this full functionality is available in Europe under CE marking, while FDA authorization in the United States covers only a subset of findings. The more findings a system searches for, the greater the total number of false alerts.

“Human + AI”: The DeepFAN paradox. DeepFAN is a model designed to distinguish benign from malignant pulmonary nodules. In a clinical study (ChiCTR2400084624; three centers; 463 pathologically confirmed nodules from 400 patients), AI achieved an AUC of 0.954 independently. The mean AUC of 12 junior radiologists (1–5 years of experience) was 0.667 without AI and increased to 0.776 with AI assistance; inter-reader agreement (κ) improved from 0.313 to 0.421 [46]. The results are noteworthy in two ways. AI substantially improved physicians’ performance, yet physicians assisted by AI performed less well than AI alone. One possible explanation is that physicians sometimes rejected correct suggestions and sometimes uncritically accepted incorrect ones; however, the study abstract does not analyze this mechanism, so it remains a hypothesis. This does not mean physicians are redundant: AI does not know the patient’s history and is not accountable for the decision. Rather, it indicates that the “human + AI” model is still evolving and that clinicians need training in when to trust a signal, when to challenge it, and how to interpret confidence measures. The study is currently available as a preprint and awaits peer review.

Pathology, Ophthalmology, and Cardiology

6.2.1.Paige Prostate: was the first AI-based system to receive FDA authorization through the De Novo pathway in September 2021 for highlighting suspicious areas on digitized prostate biopsy slides. In the pivotal study, 16 pathologists evaluated 527 slides (171 cancerous and 356 benign) twice—with and without the system. AI assistance improved cancer detection by an average of 7.3% and did not affect assessment of benign slides [12]. PathAI AISight Dx is a different type of product: a digital pathology image-management

and viewing platform that the FDA authorized in 2025 for primary diagnosis when used with specified scanners [15]. It does not itself make diagnoses; rather, it provides a digital environment in which AI algorithms can be integrated. In both cases, accuracy depends substantially on slide preparation, staining, and scanner characteristics.

IDx-DR: (LumineticsCore) was the first autonomous AI system to receive FDA authorization in 2018 for screening diabetic retinopathy without a physician viewing the image. In the pivotal study, sensitivity was 87.2% and specificity 90.7% [1]. In a prospective study at a specialized diabetes clinic in Karlsburg, Germany (875 inpatients), sensitivity for detecting referable retinopathy was 94.4% and specificity 90.5%. These figures, however, apply only to images that could be evaluated: with undilated pupils, images could not be captured for 10.5% of patients, and the system judged images from another 26.1% unsuitable for analysis. Thus, the system returned a result for only 63.4% of patients [24]. The real-world effectiveness of an autonomous system depends on equipment and trained personnel—an issue explored in the Georgian scenario (Section 7.3).

In cardiology: Cardiologs’ deep neural-network algorithm achieved 92.2% accuracy in identifying major abnormalities on emergency-department ECGs, compared with 87.2% for a conventional algorithm [37]. In a study of 1,000 Holter recordings at three university hospitals, an AI-based platform was not inferior to the conventional approach in diagnostic accuracy and reduced analysis time by more than 25% (from 6.0 to 4.4 minutes on average) [17]. AliveCor Kardia 12L is a pocket-sized 12-lead ECG device that records eight leads using five electrodes. As of January 2026, its KAI 12L algorithm was cleared for 39 cardiac determinations, including acute myocardial infarction and ischemia [2]. The 4TAVR system uses AI for surgical planning by automatically segmenting cardiac structures on CT and measuring the aortic root before Transcatheter Aortic Valve Replacement (TAVR). The FDA cleared the system in April 2025 on the basis of technical validation (929 CT scans compared with expert measurements); no clinical trial was conducted [16].

Summary

The key findings and limitations of diagnostic systems are summarized in (Table 3).

Table 3: Comparison of diagnostic AI systems.

System	Field	Autonomy	Key Finding	Main Limitation
Aidoc	Radiology (CT)	Triage	Sensitivity 87.2%; PPV 70.5%; notification time 73 → 58 min	A negative result does not rule out pathology
DeepFAN	Pulmonary nodules	Second reader	AUC: AI 0.954; physicians 0.667 → 0.776 with AI	Preprint; human-AI interaction requires training
Paige Prostate	Pathology	Second reader	FDA De Novo (2021); detection improved by 7.3%	Variation in staining and scanners
IDx-DR	Ophthalmology	Autonomous	Referable retinopathy: sensitivity 94.4%, specificity 90.5%	Result available for only 63.4% (10.5% no image; 26.1% unsuitable)
Cardiologs	Cardiology (ECG)	Second reader	Accuracy 92.2% vs 87.2%	A “normal” result does not exclude acute coronary syndrome

Overall conclusion: high accuracy on a narrowly defined task does not make a system error-free. Manufacturer-reported performance may not be reproduced with a different scanner, protocol, or population, and authorization applies only to the specified intended use.

Clinical Decision Support, LLMs, and AI Scribes

CDSS: Lessons from Failure

Clinical Decision Support Systems (CDSS) process not a single image but diverse and often incomplete patient data. Their history therefore offers more cautionary lessons than success stories. IBM Watson for Oncology was developed with Memorial Sloan Kettering Cancer Center to suggest optimal treatment options. Internal documents made public in 2018 described incorrect and unsafe recommendations; the system had been trained largely on

hypothetical cases and reflected the practice of a single center. In 2022, IBM sold Watson Health. Epic’s sepsis model was implemented in hundreds of US hospitals. An independent University of Michigan evaluation involving more than 27,000 patients found that the model missed approximately two-thirds of sepsis cases while generating many false alerts [44]. The result can be “alert fatigue,” in which staff stop paying attention to alerts. DeepMind’s acute kidney injury model predicted injury 48 hours in advance, but it was trained on data from US veterans, the majority of whom were men, and performed less well in women—an illustrative example of dataset shift.

LLM-based tools

In recent years, tools that accept clinical questions in natural language have rapidly become part of physicians’ everyday practice (Table 4).

Table 4: LLM-based clinical tools.

Tool	What it Offers	Regulatory Status	Main Risk
Open Evidence	Answers with citations; used by 88.9% of surveyed physician users of LLMs; ~15 million questions per month	Not an authorized medical device	Funded by pharmaceutical advertising; US guidelines
DoxGPT (Doximity)	Integrated into a professional network for physicians	Not an authorized medical device	Evidence of benefit comes from manufacturer research
ChatGPT for Clinicians	Clinical version based on GPT-5 (2026)	Not an authorized medical device	Lack of EHR integration and independent validation
iatroX	Based on NICE guidelines and SmPCs	Registered with the MHRA	Tailored to the UK context

The regulatory paradox and physician responsibility. As of February 2026, no product on the FDA list uses generative AI or an LLM. This means physicians often use precisely those tools that lack regulatory authorization. An LLM response is not, in itself, medical advice: if a decision based on it harms a patient, responsibility rests with the physician, and “the system was wrong” is not an adequate defense. Any clinically consequential LLM response should therefore be checked against primary sources and current guidelines. The principle of localization. The iatroX example shows that clinical AI must align with the guidelines of the country in which it is used. No comparable tool based on Georgian guidelines yet exists, although initial steps toward Georgian-language, evidence-based symptom triage have been taken [3].

AI Scribes (Ambient Scribes)

Abridge, Nuance DAX (Microsoft Dragon Copilot), Nabla, Heidi, and Tortus listen to consultations with patient consent and generate a draft structured note. In the first randomized trial (UCLA; 238 physicians, 14 specialties, approximately 72,000 visits), Nabla users’ documentation time per note decreased by 9.5% compared with the control group (from an average of 4 minutes 30 seconds to 3 minutes 49 seconds), while DAX showed no statistically significant effect (–1.7%). Both systems produced small improvements in burnout measures, although physicians reported clinically significant inaccuracies as occurring “sometimes” [27]. The technology can free up time, but independent studies find more modest effects than manufacturers promise. Risks include

hallucinated symptoms, signing notes without reading them, and storing audio as health data. The accuracy of these systems in Georgian has not been independently assessed.

Challenges and Risks

The most serious challenges of medical AI are not random glitches but arise from the nature of the technology itself. They cannot therefore be resolved simply by improving accuracy.

Transparency and Automation Bias

Physicians may find it difficult to justify recommendations from a “black-box” system, while people tend to place excessive trust in automated systems. If AI reports “normal,” clinical suspicion should not disappear: a negative AI result does not rule out pathology. A long-term risk is deskilling. Just as total reliance on GPS can weaken spatial orientation, continual reliance on AI may weaken differential-diagnostic skills [18]. AI should therefore be treated first as a “second opinion,” not as the “first and final” opinion.

Algorithmic Bias

AI learns from historical data and reproduces their inequalities, often amplifying them. The literature distinguishes selection bias (groups are underrepresented in the data), labeling bias, measurement bias, and deployment bias (a model is used in a population for which it has not been validated). A prominent clinical example is pulse oximetry. US hospital data showed that occult hypoxemia—arterial blood gases indicating oxygen

saturation below 88% while pulse oximetry showed 92–96%—was approximately three times more common among Black patients than among White patients [36]. This is measurement bias: a device calibrated primarily on lighter skin overestimates saturation in darker skin. If AI is trained on such data to predict, for example, the need for oxygen therapy or hospitalization, it may automate and scale the error: a patient who needs oxygen could be classified as “stable” by the algorithm. A second example concerns healthcare resource allocation. A commercial algorithm used previous healthcare spending as a proxy for patients’ health needs. As a result, Black patients who had historically received less care were assigned lower risk scores than White patients with comparable illness [29]. Non-medical parallels include Amazon’s hiring system and the UK A-Level grading algorithm (2020) [18]. The implication for Georgia is direct: any model trained on a foreign population requires local, stratified validation in Georgian patients.

Privacy

Three risks are especially salient in the era of generative AI: the illusion of anonymization (anonymized data may still be re-identified), images that themselves contain identifying information, and entering patient data into general-purpose chatbots. Under Georgian law, health data are a special category of personal data.

Hallucinations and Academic Integrity

LLMs can invent nonexistent articles, authors, and DOIs—a direct consequence of the “stochastic parrot” characteristics discussed in Section 3.3. According to ICMJE recommendations, an AI tool cannot be an author, and its use in a manuscript should be disclosed [25].

Regulation

Most AI medical devices receive FDA clearance through the 510(k) pathway, based on “substantial equivalence” to an existing device and often supported by retrospective data. Randomized studies demonstrating real patient benefit remain uncommon. In the European Union, medical AI is subject to two regulatory regimes: the Medical Device Regulation [9] and the AI Act [10], which treats it as high-risk and requires human oversight. LLMs, as noted above, remain in a “gray area.”

Digital Poverty and Concentration of Power

The resources needed to develop large models are concentrated in a small number of companies and economies, and the models are trained mainly on data from the “Global North.” Digital poverty has three dimensions: access (no internet or device), quality (internet is slow or expensive), and competence (the technology is available, but users lack the skills to use it) [18]. For countries with small language and medical markets, including Georgia, this can mean technological dependence and poor contextual fit.

Environmental Impact

The carbon footprint of AI is often overlooked in discussions of AI ethics. Training large models—and increasingly, using them at scale—requires substantial energy. The emissions associated with

a single training run of GPT-3 have been estimated at approximately 550 tonnes of CO₂ equivalent [30], comparable to the annual emissions of more than 100 passenger cars. The International Energy Agency estimates that one ChatGPT query consumes about 2.9 Wh of electricity on average, compared with approximately 0.3 Wh for a conventional Google search—nearly ten times less [26]. These estimates vary considerably by model and infrastructure and are declining over time.

For healthcare, which is itself a major source of emissions, this has two practical implications. First, AI should be used where its clinical benefit is justified, not “everywhere it is possible.” Second, procurement decisions should consider energy efficiency alongside accuracy and cost: a smaller model designed for a specific task may be a more sustainable choice than a large general-purpose model.

Prospects for Georgia

The main constraints on medical AI in Georgia are not a lack of algorithms but data infrastructure, institutional fragmentation, and language barriers. This section describes the context in which the systems discussed above would need to be implemented.

Healthcare System Context

Electronic health records. The EHR system began operating on a pilot basis in 2019, but the obligation to submit data was postponed several times and became mandatory only on 1 August 2023. As of 2023, no supervisory measures had been undertaken since the system was established. Of 400,681 inpatient and outpatient cases in selected state programs (diabetes, mental health, rare diseases, organ transplantation, and tuberculosis) during 2022–2023, 201,942 (50%) were not recorded in the EHR. Coverage was better for cases under the Universal Healthcare Program, 80% of which were recorded. Of 16,575 professionally active physicians (excluding family and rural physicians, who are not subject to this obligation), 24% (3,904) did not use the EHR and e-prescription systems [38]. Because AI learns from data, incomplete EHR coverage directly constrains the development and validation of local models. Challenges in digital health governance in Georgia are also discussed by Gotsadze, et al. [21]. Telemedicine. The same audit found that telemedicine equipment had been installed in 50 rural outpatient facilities since 2022, but only 3,883 (4%) of 97,381 potential beneficiaries received services during 2022–2024. Of six types of equipment purchased, the ophthalmoscope, stethoscope, and spirometer remained unused; in three of five villages inspected, outpatient facilities had no internet access [38]. These findings directly affect the prospects for remote AI diagnostics (see Section 7.3).

Private-sector dominance and financing. Approximately 80% of hospital beds are privately and commercially owned, and the private sector has shown resistance to central planning and regulation [8]. The Universal Healthcare Program has operated since 2013; public expenditure amounted to 3.1% of GDP in 2022. Estimates of out-of-pocket spending differ considerably across sources. According to the European Observatory, it declined from more than 80% in 2011 to approximately 30% in 2022 [8],

whereas Galt & Taggart estimated that it accounted for 51% of total healthcare expenditure in 2022 [20]. The discrepancy is likely methodological, and both sources should be reported. For AI, this means that data standardization requires cooperation among numerous independent private institutions. Workforce. Georgia has one of the highest physician densities in Europe (6.1 per 1,000 population in 2021), but one of the lowest numbers of nurses. In 2022, there was approximately one nurse per physician, compared with an average of two to three nurses per physician in the WHO European Region. Shortages are especially acute in mountainous and hard-to-reach areas [20,42,43]. AI may reduce

some documentation and triage workload, but it cannot replace a missing workforce. Regional inequality. Specialized services are concentrated in Tbilisi and other large cities, while access in the regions is limited [7,32]. (Table 5) shows several-fold differences between regions in hospital-bed and outpatient-contact indicators. One reason for the low use of primary care is the underdevelopment of outpatient facilities, which directs patients straight to hospitals. Another is the high level of self-medication with non-prescription medicines [20]. This gap creates the greatest potential benefit for remote AI diagnostics—and at the same time one of the greatest barriers to them (see Section 7.3).

Table 5: Regional indicators, 2021 (Source: Galt & Taggart [20], based on NCDC data).

Region	Hospital Beds Per 1,000 Population	Outpatient Contacts Per Capita
Tbilisi	8.1	9.1
Imereti	7.9	2.7
Kakheti	2.5	2
Samtskhe-Javakheti	2.8	1.4
Kvemo Kartli	2.4	1.2
Guria	1.8	2.1
Georgia (average)	5.6	4

Note*: Selected regions are shown; the source provides complete data for all 11 regions.

Opportunities

- Remote diagnostics for mountainous and remote regions: analysis of X-rays, CT scans, ECGs, and retinal images in combination with telemedicine.
- Epidemiological analytics: early identification of local trends from EHR data (e.g., increases in diabetes or cardiovascular disease) and planning of preventive programs.
- Medical education: AI-based simulation training, particularly in surgery.
- Second-opinion services: additional analysis of complex and rare cases at major clinics.
- Automated documentation: reducing physicians' administrative workload through a Georgian-language AI scribe [18].

Scenario: AI Screening for Diabetic Retinopathy in Kakheti

The scenario below is constructed by the authors. It is not based on a study conducted in Kakheti; rather, it applies the data discussed above to a practical context. The parameters used are approximate.

Why Kakheti, and why retinopathy? Hospital-bed density in Kakheti (2.5 per 1,000 population) is three times lower than in Tbilisi, while outpatient contacts (2.0 per capita) are four and a half times lower (Table 5). Diabetic retinopathy is a condition in which annual screening can prevent vision loss, while ophthalmologists are rarely available in regional primary care. An autonomous system

such as IDx-DR could theoretically address this gap by enabling screening at the primary-care level without an ophthalmologist present.

What do the data tell us? If the results of the 2026 IDx-DR study (Section 4.2) are directly applied, with both proportions calculated from the same cohort, then out of 1,000 patients attending screening, images could not be captured for approximately 105, and a further ~260 would have images unsuitable for analysis. In other words, roughly one in three patients would still need referral to an ophthalmologist in Telavi or Tbilisi. An autonomous system does not eliminate the need for ophthalmologists; it reduces it. A referral pathway must therefore be in place before screening begins.

What is needed for implementation?

- Equipment and Personnel:** a retinal camera and a trained operator (nurse or technician); image quality depends substantially on operator skill. Given the nursing shortage (Section 7.1), a dedicated workforce solution would be needed.
- Infrastructure:** reliable internet for image transfer and integration with the EHR so that results are recorded in the patient's file.
- "Silent" Pilot:** initially, AI results do not affect clinical decisions. An ophthalmologist remotely reads every image, and the two results are compared. This allows sensitivity, specificity, and the proportion of ungradable images to be measured in the local population.
- Referral and Treatment Pathway:** timely ophthalmology appointments and access to laser/anti-VEGF treatment for

patients with positive or ungradable results; without this, screening merely produces a list.

e. Legal Clarity: who is liable for a missed case-the clinic, manufacturer, or operator? This must be defined before the pilot starts.

f. Financing: include screening in the Universal Healthcare Program, since out-of-pocket payment would reduce participation in the region.

Open questions. A realistic evaluation of this scenario requires data that are not currently available: the number of people with diabetes in Kakheti, the number and distribution of ophthalmologists in the region, and the quality of internet access in rural outpatient facilities. Collecting these data could itself be the first practical step.

Discussion and Recommendations

This review leads to three main conclusions. First, AI's value is clearest in narrow, well-defined tasks (triage, screening, measurement), whereas attempts to automate complex clinical decisions (Watson, Epic) have disappointed. Second, manufacturer-reported performance and real-world outcomes are not the same: the 26.1% rate of ungradable images with IDx-DR and Aidoc's 70.5% positive predictive value are clear examples. Third, the most widely used tools (LLMs) are among the least regulated, increasing rather than reducing physicians' professional responsibility.

"Human + AI": How Collaboration Works

Emphasizing risks does not mean that AI's benefits are doubtful. Successful models of collaboration share a common logic: AI performs high-volume, repetitive work, while humans provide judgment, interpretation, and accountability [18].

a. Medicine. Aidoc reviews thousands of scans in the background and helps radiologists identify critical cases more quickly; Paige Prostate directs pathologists to areas suspicious for cancer; an AI scribe redirects physicians' attention from the computer screen to the patient.

b. Law. AI can process thousands of pages and find similar cases in hours, a task that could take a lawyer weeks; interpretation and strategy remain the lawyer's responsibility.

c. Education. AI creates personalized learning materials and provides immediate feedback; teachers focus on motivation, emotional support, and explaining difficult concepts.

d. Programming. AI writes code snippets and searches for errors; programmers make architectural decisions and review AI-generated code.

e. In all four fields, collaboration works only if people check AI outputs rather than simply approve them. The DeepFAN paradox (Section 4.1) shows that this skill does not develop automatically; it must be taught.

AI and Medical Professions

AI will affect specialties unevenly [18]. The greatest changes

are likely in fields where much of the work involves interpreting structured data: radiology (radiologists may shift toward complex cases and interventional procedures), pathology (toward rare and complex cases), and cardiology, where ECG interpretation may become routine and cardiologists may focus on complex arrhythmias and interventions. Surgery, which requires physical intervention and urgent decisions, psychiatry, where therapeutic relationships and nonverbal signals are essential, and pediatrics, where family dynamics and individualized care are central, may change less. New roles are emerging, including clinical AI specialist, medical data scientist, and medical AI ethics specialist. The professional question is less "Which specialty will disappear?" than "Which tasks will be redistributed?" A physician who uses AI may outperform a colleague who does not on certain tasks, while clinical reasoning, communication, and accountability remain at the core of the medical profession. For Georgia, this means that medical education should begin preparing physicians for this redistribution now (Recommendation 1).

Recommendations

In light of these findings, rapid technology deployment is not the right path for Georgia. Implementation should be phased, context-sensitive, and responsible.

Education:

a. Introduce a mandatory AI-literacy course in medical schools and continuing professional development programs, covering core concepts, limitations, ethics, checking AI outputs (e.g., the DOCTOR technique: Doubt, Original source, check guidelines, talk to colleagues, observe patient, Reassess), and practical human-AI interaction.

Implementation:

a. Begin with low-risk, high-benefit tasks: documentation, triage, screening, and telemedicine.

b. Implement in three stages: "silent" mode (AI operates but does not influence decisions) → visible support (the physician retains full control) → broad deployment with post-implementation monitoring [18].

c. Use an eight-criterion procurement assessment: clinical validation, data bias, transparency, integration, regulatory status, cost (including energy efficiency), staff training, and post-implementation monitoring.

Data and Infrastructure:

a. Achieve full EHR implementation and data standardization (e.g., HL7 FHIR) across private and public institutions.

b. Create an anonymized Georgian medical database with full protection of patient rights; alternatively, use federated learning, in which the model travels to the data while the data remain within the institution.

c. Develop Georgian-language medical terminology resources and an evaluation benchmark to assess the clinical accuracy of LLMs in Georgian.

- d. Establish public–private partnerships for infrastructure, especially in the regions.

Governance:

- a. Develop an ethical and legal framework for medical AI harmonized with the MDR and AI Act, addressing allocation of liability, informed consent, and documentation of AI use in the medical record.
- b. Establish a national system for recording AI-related errors

and incidents as part of patient safety.

Proposed Roadmap

The time frames in (Table 6) are a framework proposed by the authors, not an evidence-based forecast. They are not based on official government plans and require validation through institutional consensus. Progress through these stages requires not only technological but also institutional and cultural change—above all, building trust among physicians, patients, and technology developers.

Table 6: Roadmap for implementing medical AI in Georgia.

Phase	Time Frame	Main Tasks
Short term	2026–2028	AI-literacy courses and physician training; “silent” pilots in documentation, triage, and screening (e.g., Section 7.3)
Medium term	2028–2032	Full EHR implementation, data standardization, Georgian-language support, anonymized national database, legal framework
Long term	2032–2037	Personalized and predictive medicine; population health management

Conclusion

Medical AI is undergoing a transformative but still incomplete phase: it is neither hype nor a reason for disillusionment. Diagnostic systems have demonstrated convincing accuracy in narrow tasks, while the history of CDSS—from MYCIN to Watson—has taught us that technical excellence is necessary but insufficient: implementation is far more difficult than development. For Georgia, AI presents a genuine opportunity to reduce regional inequalities, ease physicians’ workload, and improve healthcare quality. Achieving these goals requires strategic vision, investment in education and data infrastructure, support for the Georgian language, and a clear legal and ethical framework. Medical AI is not a completed project but an ongoing process that requires continuous research, critical evaluation, and adaptation.

Key messages

- a. AI augments physicians rather than replacing them—but “human + AI” collaboration does not work automatically; it requires training (the DeepFAN paradox).
- b. Manufacturer-reported accuracy is not the same as real-world performance; every system should be evaluated in the local population in “silent” mode.
- c. The most frequently used tools (LLMs) are among the least regulated; responsibility remains with the physician.
- d. Bias originates in data; models trained on foreign populations require validation in Georgian patients.
- e. In Georgia, the main barriers are data and institutions, not algorithms. EHRs, Georgian-language support, and a legal framework should take priority over technology procurement.

Declarations

Ethics statement. This paper is a literature review and does not include research involving humans or animals; ethics committee

approval was not required. Conflict of interest. The first author is the author of textbooks cited in the article. No other conflict of interest is declared. Statement on the use of AI (ICMJE). A generative AI tool (Claude, Anthropic) was used to help structure and language-edit the manuscript draft. The authors checked the content, sources, and conclusions and accept full responsibility for them.

Acknowledgments

The study was conducted with support of East European University.

Conflict of Interest

None

References

1. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC (2018) Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med 1: 39.
2. AliveCor (2026) AliveCor announces FDA clearance of new cardiac determinations for Kardia 12L ECG System, bringing total to 39(13): 2026.
3. Apshinashvili I, Pkhakadze G (2026) Bridging the auditability gap in consumer medical AI: design and public deployment of a Georgian-language, evidence-based symptom triage system (SheniEkimi). Georgian Med J 1(2): 1-12.
4. Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21): 610-623.
5. Buchanan BG, Shortliffe EH (Eds.), (1984) Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project. Addison-Wesley, Reading, MA.
6. Collins GS, Reitsma JB, Altman DG, Moons KGM (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 350: g7594.
7. (2026) EUAA – European Union Agency for Asylum. Georgia: EUAA reports on paediatrics, hepatitis, diabetes, and gastrointestinal diseases

- show regional disparities in availability of care. MedCOI Topical Reports. <https://www.euaa.europa.eu>.
8. (2024) ssssEuropean Observatory on Health Systems and Policies. Health systems in action 2024: Georgia. WHO Regional Office for Europe, Copenhagen. <https://eurohealthobservatory.who.int/publications/i/health-systems-in-action-georgia-2024>.
 9. (2017) European Parliament and Council of the European Union Regulation (EU) 2017/745 of 5 April 2017 on medical devices. Off J Eur Union L 117: 1-175.
 10. (2024) European Parliament and Council of the European Union Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Off J Eur Union L series, 12 July 2024.
 11. (2018) FDA – U.S. Food and Drug Administration 510(k) premarket notification K180647: BriefCase (Aidoc Medical). FDA, Silver Spring, MD.
 12. (2021) FDA – U.S. Food and Drug Administration FDA authorize software that can help identify prostate cancer 21 September 2021. <https://www.fda.gov/news-events/press-announcements/fda-authorizes-software-can-help-identify-prostate-cancer>.
 13. (2022) FDA – U.S. Food and Drug Administration 510(k) premarket notification K221240: BriefCase (Aidoc Medical). FDA, Silver Spring, MD.
 14. (2025) FDA – U.S. Food and Drug Administration 510(k) premarket notification K252366: a2z-Unified-Triage. FDA, Silver Spring, MD.
 15. (2025) FDA – U.S. Food and Drug Administration 510(k) premarket notification K243391: AISight Dx (PathAI). FDA, Silver Spring, MD.
 16. (2025) FDA – U.S. Food and Drug Administration 510(k) premarket notification K241984: Hi-D Imaging 4TAVR (Hi-D Imaging AG). Decision date: 2 April 2025. FDA, Silver Spring, MD.
 17. Fiorina L, Maupain C, Gardella C, Manenti V, Salerno F, et al. (2022) Evaluation of an ambulatory ECG analysis platform using deep neural networks in routine clinical practice. J Am Heart Assoc 11(18): e026196.
 18. Gagnidze A (2026) Introduction to Artificial Intelligence for Health Professions [in Georgian]. East-West University, Tbilisi. ISBN 978-9941-8-9385-8.
 19. Gagnidze A (2026) Introduction to Artificial Intelligence for Health Professions. East-West University; East European University, Tbilisi. ISBN 978-9941-8-9256-1.
 20. Galt & Taggart (Bochorishvili E, Kurashvili S) (2023) Georgia's healthcare sector. Galt & Taggart, Tbilisi. https://galtandtaggart.com/files/galt/georgias-healthcare-sector-overview_eng_may-23.pdf.
 21. Gotsadze G, Akaki Zoidze, Tamar Gabunia, Brian Chin (2024) Advance governance for digital transformation in health: insights from Georgia's experience. BMJ Glob Health 9(10): e015589.
 22. (2026) Google Research Improving breast cancer screening workflows with machine learning [blog]. <https://research.google/blog/improving-breast-cancer-screening-workflows-with-machine-learning/>.
 23. Grant MJ, Booth A (2009) A typology of reviews: an analysis of 14 review types and associated methodologies. Health Info Libr J 26(2): 91-108.
 24. Hunfeld E, Tayar A, Paul S (2026) Real-world performance of the AI diagnostic system IDx-DR in the diagnosis of diabetic retinopathy and its main confounders. Sci Rep 16: 4349.
 25. ICMJE – International Committee of Medical Journal Editors (2026) Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals (updated January 2026). <https://www.icmje.org/recommendations>.
 26. IEA – International Energy Agency (2024) Electricity 2024: analysis and forecast to 2026. IEA, Paris. <https://www.iea.org/reports/electricity-2024>.
 27. Lukac PJ, William Turner, Sitaram Vangala, Aaron T Chin, Joshua Khalili, et al. (2025) Ambient AI scribes in clinical practice: a randomized trial. NEJM AI 2(12):10.
 28. McKinney SM, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, et al. (2020) International evaluation of an AI system for breast cancer screening. Nature 577(7788): 89-94.
 29. Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366(6464): 447-453.
 30. Patterson D, Joseph Gonzalez, Quoc Le, Chen Liang, Lluís Miquel Munguia, et al. (2021) Carbon emissions and large neural network training. arXiv: 2104.10350.
 31. Rajkomar A, Dean J, Kohane I (2019) Machine learning in medicine. N Engl J Med 380(14): 1347-1358.
 32. Richardson E, Berdzuli N (2017) Georgia: health system review. Health Syst Transit 19(4): 1-90.
 33. Seah JCY, Cyril HM Tang, Quinlan D Buchlak, Xavier G Holt, Jeffrey B Wardman, et al. (2021) Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. Lancet Digit Health 3(8): e496-e506.
 34. Searle JR (1980) Minds, brains, and programs. Behav Brain Sci 3(3): 417-424.
 35. Seyam M, Weikert T, Sauter A, Brehm A, Psychogios MN, et al. (2022) Utilization of artificial intelligence-based intracranial hemorrhage detection on emergent noncontrast CT images in clinical workflow. Radiol Artif Intell 4(2): e210168.
 36. Sjöding MW, Dickson RP, Iwashyna TJ, Gay SE, Valley TS, et al. (2020) Racial bias in pulse oximetry measurement. N Engl J Med 383(25): 2477-2478.
 37. Smith SW, Walsh B, Grauer K, Wang K, Rapin J, et al. (2019) A deep neural network learning algorithm outperforms a conventional algorithm for emergency department electrocardiogram interpretation. J Electrocardiol 52: 88-95.
 38. (2024) State Audit Office of Georgia. Digital Health Performance Audit [in Georgian]. State Audit Office of Georgia, Tbilisi. <https://web-api.parliament.ge/storage/files/shares/zedamxedveloba/auditis-jguifi/2025/angarishi-cifruli.pdf>.
 39. Topol E (2019) Deep medicine: how artificial intelligence can make healthcare human again. Basic Books, New York.
 40. Turing AM (1950) Computing machinery and intelligence. Mind 59(236): 433-460.
 41. Warren LM, Jenny Venton, Kenneth C Young, Mark Halling Brown, Christopher J Kelly, et al. (2026) Impact of using artificial intelligence as a second reader in breast screening including arbitration. Nat Cancer 7(3): 507-521.
 42. (2024) WHO Regional Office for Europe. Georgia moves forward to regulate nursing and midwifery [news item]. <https://www.who.int/europe/news/item/12-08-2024-georgia-moves-forward-to-regulate-nursing-and-midwifery>.
 43. (2025) WHO – World Health Organization. Joint external evaluation of IHR core capacities of Georgia: mission report. World Health Organization, Geneva. <https://www.who.int/publications/i/item/9789240102941>.
 44. Wong A, Otles E, Donnelly JP, Andrew Krumm, Jeffrey McCullough, et al. (2021) External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med 181(8):

1065-1070.

45. Yu VL, LM Fagan, SM Wraith, WJ Clancey, AC Scott, et al. (1979) Antimicrobial selection by a computer: a blinded evaluation by infectious diseases experts. JAMA 242(12): 1279-1282.

46. Zhu Z (2026) DeepFAN, a transformer-based deep learning model for human-artificial intelligence collaborative assessment of incidental pulmonary nodules in CT scans: a multi-reader, multi-case trial. arXiv: 2603.25607.