Review | Received 7 June 2026; Revised 3 October 2026; Accepted 26 October 2026; Published 6 November 2026
https://doi.org/10.55092/aias20260007

Generative artificial intelligence in medicine: a systematic review of models, applications, challenges, and future directions

Yifei Li, Yipeng Liao, Liqun Lin* and Tiesong Zhao

Fujian Key Laboratory for Intelligent Processing and Wireless Transmission of Media Information, College of Physics and Information Engineering, Fuzhou University, Fuzhou, China

* Correspondence author; E-mail: lin_liqun@fzu.edu.cn.

Highlights:

- Four generative models are compared systematically for medical applications.
- Applications span clinical support, synthetic imaging, drug discovery, literature, and education.
- Major issues include hallucination, privacy concerns, bias, black box nature, and regulatory gap.
- Further development needs trustworthy AI, ethical governance, and collaboration.

Abstract: Generative artificial intelligence (AI), which uses the probability distribution of real-world data and generates new content by learning it, is becoming a powerful tool for use in medicine. In this paper, we will systematically review the progress made in research, major applications, and challenges of the use of generative AI in the medical field. Firstly, we give a brief introduction to the probabilistic background of generative models and then describe four kinds of models, including generative adversarial network, diffusion model, variational autoencoder (VAE), and large language model. Technical features and their potential usage in the medical field will be discussed. Secondly, we systematically investigate their applications in five fields, which include clinical decision-making, synthetic medical imaging, drug discovery, processing of medical literature, and medical education. From our review, it is clear that generative AI has a lot of capabilities when it comes to increasing efficiency and accuracy in health care delivery, drug discovery, and medical education. Nonetheless, there are several challenges associated with this technology, including factual errors, hallucinations, data privacy concerns, algorithmic bias, and difficulties in attributing responsibility. In light of these limitations, it is vital for generative AI technology to play the role of an assistant to medical professionals as opposed to taking their roles. The future will require the creation of reliable technologies like uncertainty quantification, explainable AI, and retrieval-augmented generation (RAG).

Keywords: generative AI; medical imaging; clinical decision support; drug discovery; medical education

Copyright©2026 by the authors. Published by ELSP. This work is licensed under Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium provided the original work is properly cited.

1. Introduction

1.1. Challenges facing the medical field

Due to changes in the demographic structure of the world and rapid progress in technology, contemporary medicine is facing challenges which are unprecedented in their nature. They are not only related to the changing spectrum of diseases but also to the structural contradictions inherent to the health care system [1]. All this has led to the revolution in the model of health care and development of new technologies. First of all, the problem of data explosion in medicine has been highlighted [2]. It is crucial to derive useful information from the data mass to help make proper diagnosis and treatment.

Moreover, the disproportion between the supply and demand of healthcare resources on the global scale is growing more and more obvious. In the first place, the process of population aging aggravates, and the rate of chronic diseases is growing, thus significantly raising the need for healthcare services in society [3]. At the same time, the long and expensive training of highly skilled medical professionals makes it difficult to satisfy the increased need for healthcare services in certain regions. The most complicated problem arises in remote and underdeveloped areas, where the shortage of well-trained healthcare staff becomes critical both by quantity and by their allocation. Moreover, the labor resources for the provision of primary healthcare are mostly concentrated in economically advanced territories, which makes the work of primary healthcare facilities even less effective [4,5].

The last challenge that arises due to heterogeneity concerns the disease pathology itself. Heterogeneity of pathogenesis of many diseases, including cancer, is extremely high [6]. It means that one disease can show completely different features regarding clinical course and treatment depending on the patient. Thus, there is a need among professionals for precision and personalized medicine that would require decisions to be made by taking into account a full and detailed understanding of the patient [7]. But then again, to process such complicated information, one needs a high degree of competency, which at present is missing.

1.2. Comparison with existing surveys

Recent studies on generative AI in medicine have also included several surveys; however, the scope of the majority of these surveys has mainly been limited to either one family of models or one field of applications. Most surveys have been devoted exclusively to large language models (LLMs), describing their development from early pre-trained language models, evaluating their performance for various clinical tasks, and discussing problems of fairness, accountability, transparency, and ethics [8,9]. Several surveys have concentrated on only one type of architecture used for generative models, for example, Generative adversarial networks (GANs) used for analysis of medical images [10]. In addition, some reviews have been based on certain applications of generative models in clinical settings, such as medical education [11], precision and personalized medicine [12], and de novo drug design [13,14]. Ibrahim *et al.* [15] covered several classes of models and different types of data but were limited to one type of application, namely data generation and evaluation of generative models.

This study, on the other hand, uses an all-encompassing approach, making it different from the existing studies. We will not confine ourselves to one family of models or one application domain, but instead discuss four major families of generative models and juxtapose their probability basis, technical aspects, and applicability to medicine. Following from that, we consider their application in five key

fields of medicine and incorporate issues and future directions of each of them in one coherent framework. Table 1 provides a comparison between this study and several existing reviews.

Table 1. Comparison of this review with representative existing surveys on generative AI in medicine.

Survey	Model Classes Covered	Application Domains	Primary Perspective
He *et al.* [8]	LLMs	Healthcare applications	Development roadmap from PLMs to LLMs; data, applications, accountability and ethics
Zhou *et al.* [9]	LLMs	Clinical diagnostics, medical education, *etc.*	Development principles and deployment of medical LLMs; performance benchmarking
Heng *et al.* [10]	GANs	Medical image analysis	GAN architectures for image synthesis, segmentation, classification, and reconstruction
Preiksaitis & Rose [11]	Generative AI	Medical education	Scoping review of opportunities, challenges, and future directions in education
Ghebrehiwet *et al.* [12]	Deep generative models	Precision and personalized medicine	Synthetic data generation and DGMs for personalized medicine
Ibrahim *et al.* [15]	GANs, VAEs, diffusion, LLMs	Synthetic data across multiple modalities	Synthetic data generation techniques, applications, and evaluation
Cheng *et al.* [13]	Deep generative models	Drug discovery	De novo molecular design
Tang *et al.* [14]	Generative AI	Drug discovery	Small-molecule and protein generation
This review	GANs, diffusion, VAEs, LLMs	Clinical decision support, synthetic imaging, drug discovery, literature, education	Unify probabilistic foundations of four model classes with cross-domain applications, challenges, and governance

1.3. Contributions

The contribution of the paper can be outlined in the following manner:

(1) A detailed, up-to-date review is carried out that specifically combines the probabilistic background, technical peculiarities, and medical application potential of four leading generative models (GANs, diffusion models, VAEs, and LLMs). Previous reviews in the field have not included any comparisons of various models.

(2) The applications of generative AI in medicine have been analyzed in five main areas, with representative papers reviewed, and their clinical potential assessed. This is supposed to give an organized roadmap for the specialists in the field.

(3) Major obstacles have been outlined, and a strategic framework proposed for technical and governance purposes.

1.4. Outline of paper

The structure of the rest of this paper is as follows. Figure 1 presents the general framework for the applications of generative AI in medicine. In Section 2, the basic probabilistic concepts in generative AI and their systematic discussion are provided for four fundamental model classes. The applications of generative AI in clinical decision support systems, including intelligent diagnostic reasoning, personal treatment planning, and clinical documentation automation are presented in Section 3. Synthetic medical imaging, involving data augmentation, cross-modal transformation, and few-shot generation, is presented in Section 4. Applications of generative AI in drug discovery, involving molecular generation, clinical

trials simulation, and patient recruitment, are presented in Section 5. The seventh section analyses the influence of the technology on medical education and training, covering personalized learning, virtual simulation of the clinical practice, and intelligent assessment technologies. The eighth section focuses on the critical issues faced by the discipline, which include factual errors, "hallucinations," ethical issues connected with privacy and bias problems, and lack of academic and regulatory standards. Finally, the ninth section presents future perspectives, analyzing the technological trends and governance techniques necessary for developing a reliable and human-centered intelligent healthcare ecosystem. The tenth section concludes the research and restates the role of generative AI as an assistance technology.

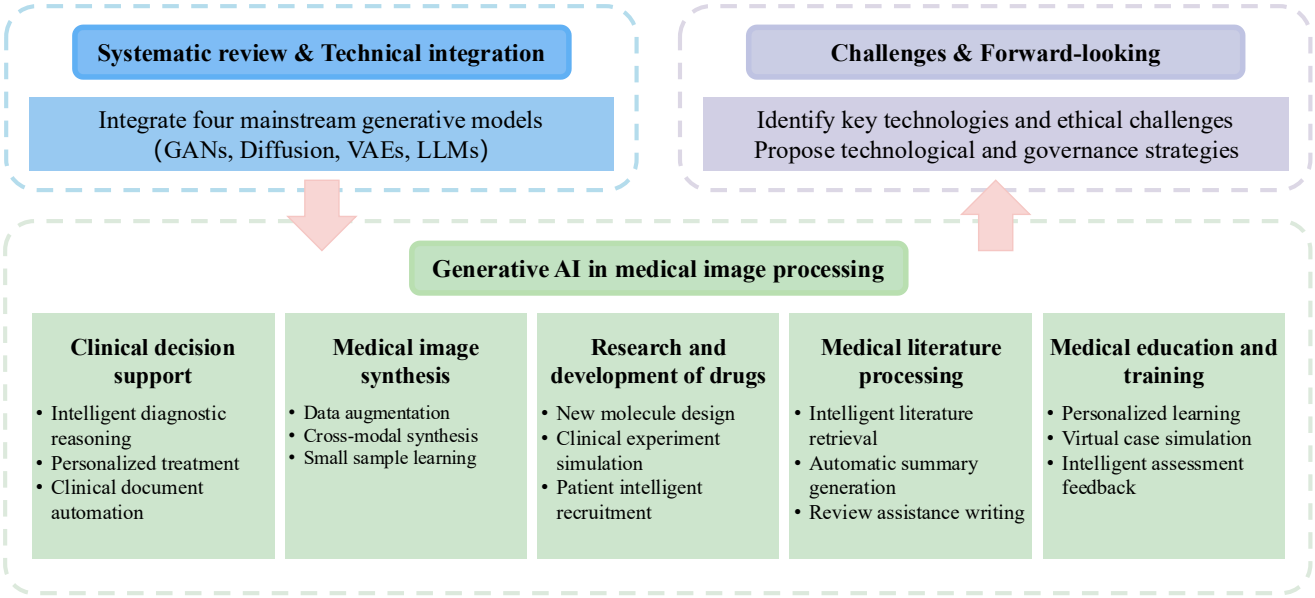


Figure 1. The main application framework of generative AI.

2. Generative models in medicine

2.1. The probabilistic nature of generative AI

Early and contemporary applications of AI in medicine have mostly centered on classification algorithms. The crux of these methods lies in learning decision boundaries, that is, assigning labels or categories to the input data. Simply put, these models try to model the conditional probability distribution, where is the probability that the output y belongs to a particular category given the input x. While discriminative AI models like support vector machines, random forests, and early deep learning techniques have been hugely successful in certain applications such as detecting lung nodules, thereby enhancing efficiency and accuracy of diagnosis, their area of application is comparatively narrow; they can be seen as "analyzers" or "classifiers" and are not suited to deal with situations in clinics that demand creativity, reasoning, and content creation.

Over the last few years, fueled by advancements in computing power and algorithms, generative AI has come into being. In terms of probabilistic modeling, the main aim of generative AI is to learn the intrinsic probability distribution of the data and sample from it for creating new and varied content.

Suppose that observed data come from an unknown and highly complex true probability distribution. The task of a generative model is to approximate this true distribution through training using a limited

number of observed samples, thereby constructing a parameterized model distribution, where θ represents the model parameters. Once the model successfully learns, new data instances can be sampled from this distribution. These new samples should be statistically "identically distributed" to the real data, thus exhibiting features similar to, yet innovative compared to, the real data.

To achieve more controllable generation, the concept of conditional generative models was introduced. Their goal is to learn the conditional probability distribution of the data given certain conditional information c The model is trained by maximizing the conditional likelihood, enabling it to generate new data x that meets the specified requirements, with diversity, based on the given condition c.

In recent years, the technological drivers of generative AI have primarily been a series of deep generative models capable of effectively fitting high-dimensional, complex probability distributions, such as GAN [16,17], Diffusion Models [18,19], Variational Autoencoders [20], and LLMs that have demonstrated remarkable capabilities on large-scale corpora [21–23]. Below is an overview of the deep generative models that have become popular due to their use of the principle of probability distribution in different ways.

2.2. Main generative models

The fast progress of generative AI is due to the creation and continuous evolution of a set of core models. Even if their realization methods differ from each other, the key aim of all of them is the study of the data distribution and the generation of new content. The next section describes several generative models with great application potential in medicine.

GANs comes from game theory. They are implicit generative models. Rather than specifying a particular probability density function, they learn to convert a simple prior into the complicated data distribution. It is achieved by adversarial training between Generator (G) and Discriminator (D). Generator's target is to make the distribution of its outputs as similar to the data distribution as possible, whereas the objective of the discriminator is to differentiate between samples generated by and those sampled from. At the end of the day, this game gives the possibility to generate highly realistic data. In the field of medicine, GANs are highly applicable for generation of medical images and data augmentation when there is a lack of annotated data. Their main features are high generation speed and image quality, although training can be rather unstable and generation diversity insufficient.

Diffusion models represent a family of likelihood-based models, whose ideas come from non-equilibrium thermodynamics. In this case, the diffusion process is chosen in advance, and the model performs a fixed forward diffusion process which transforms data to pure noise. Then, it learns a parameterized reverse process, meaning learning how to transform noise back to the initial data through gradual "denoising". In other words, from the probabilistic point of view, the model learns how to revert a gradually noising Markov chain and, consequently, learn the true data distribution. With the right choice of architecture and training scheme, the training objective of diffusion models can be simplified to predicting the noise at each step of the chain. The major advantage of such an approach is the possibility of generating high-resolution complex medical images with a controllable process. The only downsides are the low sampling speed and high computational cost.

Likewise, the VAE belongs to the likelihood family of generative models. With a neural network called the encoder, input data x is projected to the latent space to learn the posterior distribution of the latent variables. Following that, another neural network is used to reconstruct the input from the

latent variables, learning the conditional distribution. The model is optimized by the Evidence Lower Bound (ELBO) that is a trade-off between the accuracy of reconstruction and the regularization of the latent space distribution. Once training is completed, we can sample z from the prior distribution and generate new data using the decoder. Generally, the VAE-generated samples have high diversity, and the latent space is continuous, making it possible to perform interpolation and attribute editing on the latent space. In medicine, VAE can be applied for multimodal image or molecular structure generation, although the quality of the images generated is not as high as GANs or diffusion models.

In most cases, LLMs use the Transformer framework, which is a kind of autoregressive model for discrete sequences. From the point of view of probability theory, the essence of LLMs is learning the joint probability distribution of words. Due to the chain rule, it is possible to write the joint probability in the form of a product of conditional probabilities. The model makes predictions of the next word xt using the conditional probability of it depending on all previous words. With training on large text collections, LLMs are able to capture language patterns and knowledge, and, thus, generate reasonable texts on a certain prompt. In the field of medicine, LLMs can be used for clinical text generation, assisting diagnostic reasoning, medical literature summarization, patient education material creation, and other NLP tasks. Their advantages include language understanding capability and multi-task adaptability, yet the drawbacks include the "hallucination" problem when LLMs generate plausible, but factually wrong text and, therefore, threaten the factual correctness of the output. Table 2 presents a comparative summary of the common uses, strengths, and weaknesses of these four generative model types in the medical domain.

Table 2. Typical applications, advantages and disadvantages of generative models in medicine.

Generative Model	Suitable Task Types	Typical Applications in Medicine	Advantages	Disadvantages
GAN	Image generation and enhancement	Synthetic medical images, data augmentation, pathological feature generation	High image quality, fast generation speed, suitable for small sample learning	Unstable training, prone to mode collapse, limited generation diversity
Diffusion Models	High-fidelity image generation and cross-modal conversion	High-resolution medical image generation, cross-modal conversion	Extremely high image quality, controllable generation process, suitable for complex structure generation	High computational cost, slow generation speed, high data requirements for training
VAE	Molecular generation and latent space interpolation	Molecular generation, multimodal data generation, image reconstruction	Good generation diversity, continuous and interpretable latent space, stable training	Image clarity often inferior to GANs or diffusion models
LLMs	Natural language processing and text generation	Clinical diagnostic reasoning, automatic document generation, patient education text, literature summaries	Strong natural language understanding, strong multi-task adaptability, good interactivity	"Hallucination" phenomenon, dependence on training data quality, poor interpretability

3. Application of generative AI in clinical decision support

AI technology is becoming an integral component of clinical decision-making support systems, not as an independent tool, but rather as a composite system of support encompassing the cognitive, therapeutic, and documentation spheres of medicine. As illustrated in Figure 2, its clinical applications

can be organized into three complementary pillars: intelligent diagnostic reasoning and differential generation, personalized treatment plan generation, and clinical documentation automation.

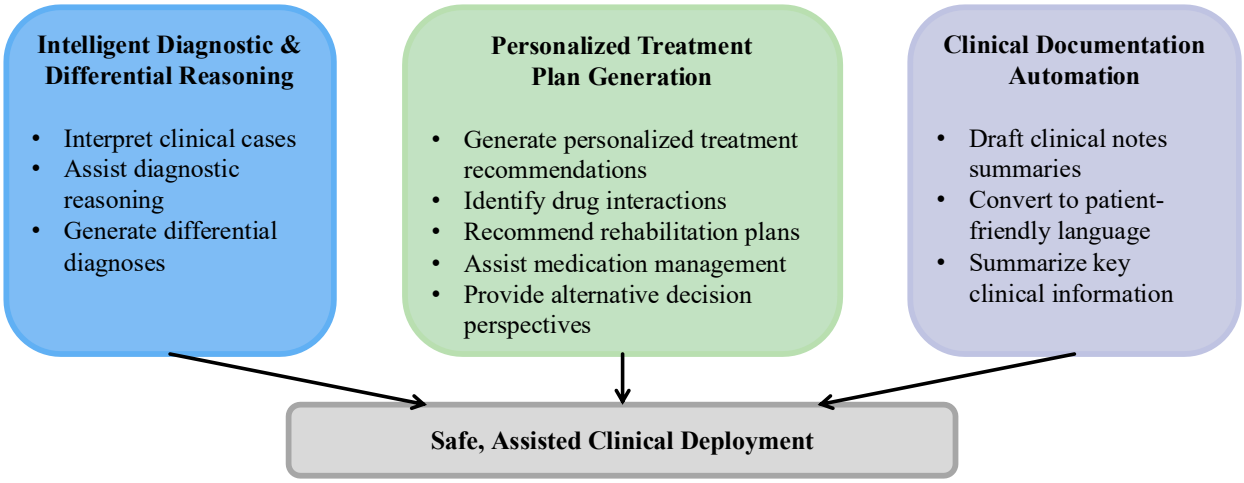


Figure 2. The application framework of generative AI in clinical decision support.

3.1. Intelligent diagnostic reasoning and differential generation

Several research studies have assessed the performance of the AI models with respect to their diagnostic capability based on clinical cases and real-world clinical data. For instance, Pagano *et al.* reported that the performance of the GPT-4 model is very impressive with respect to its ability to diagnose various diseases with 100% accuracy and its ability to provide treatment recommendations with 83% accuracy when compared with the expertise of physicians with respect to diagnosing and treating the patients with osteoarthritis of knee and hip joint [24]. The performance of ChatGPT model has been tested by Rao *et al.* based on 36 clinical snippets and it provided an accuracy of 71.7% with 76.9% accuracy for the last diagnosis-related snippet [25]. In order to mitigate the knowledge-related limitations of the generic model, Li *et al.* proposed the ChatDoctor model [26]. Moreover, Hoppe *et al.* reported that, apart from performing better than GPT-3.5 in terms of diagnostic precision in emergencies of internal medicine, the performance of GPT-4s was better than that of the emergency medicine residents, and GPT-4s could swiftly generate differential diagnoses that could guide the physicians [27]. In addition to this, it has been shown through several studies that there exists sequence dependence in the perceptual decision making involved in the analysis of medical imaging, which means that perceptual decisions in relation to the analysis of medical images can be affected by the previous images seen by the individuals. For instance, Ren *et al.* reported that when individuals look at authentic GAN-generated mammograms, then their decision making is systematically affected by the images they have previously observed [28].

The above-discussed studies show the enormous potential of the use of generative AI in diagnostic reasoning and differential diagnosis; especially when there is sufficient knowledge involved, the performance of generative AI in diagnosis can be comparable to or even exceed that of residents. But most of the existing studies are retrospective and conducted on standardized patients or on certain diseases, without any prospective and machine-controlled study in a clinical setting. The performance of the model is very much dependent on the nature and completeness of the data provided, and its generalizability and validity are questionable when used for rare diseases or complex complications. Table 3 provides the performance of selected generative AI models used in the discussed diagnostic studies.

Table 3. A performance comparison of generative AI in diagnostic tasks.

Study	Model	Task/Scenario	Sample Size	Main Results	Control Group Performance
Pagano *et al.* [24]	GPT-4	Knee/Hip Osteoarthritis Diagnosis & Treatment	100 retrospective cases	100% diagnostic concordance; 83% treatment concordance	Comparison with orthopedic specialists, complete diagnostic match
Rao *et al.* [25]	ChatGPT (GPT-3.5)	Full-process clinical decision-making	36 standardized cases	Overall accuracy 71.7%; final diagnosis 76.9%; differential diagnosis 60.3%	No direct comparison with physicians; accuracy improved with more information
Li *et al.* [26]	ChatDoctor (LLaMA fine-tuned)	Understanding and responding to real patient dialogues	10,000 test dialogues	F1 score 0.8446, significantly better than ChatGPT's 0.8406 ($P < 0.001$)	Superior to general ChatGPT, especially skilled at understanding patient descriptions
Hoppe *et al.* [27]	GPT-4 *vs* GPT-3.5	Internal medicine emergency diagnosis	100 real emergency patients	GPT-4 score 1.76/2.0, superior to GPT-3.5 (1.51) and residents (1.59)	GPT-4 significantly outperformed residents ($P = 0.01$)

Despite impressive accuracy statistics, these studies should be regarded as proof-of-concept experiments rather than clinically verified facts. The majority of studies are based on retrospective data [24,27] or controlled and standardized trials with predetermined outcomes [25], which remove all uncertainties associated with actual medical practice. In addition, there is still no generative AI-based diagnosis technology that was successfully tested during a prospective randomized control trial conducted in a real-world medical environment, nor have any been officially approved by the regulators as an independent tool. Thus, while current research results show a great potential for further investigations, they cannot be confused with practical readiness for clinical application because of close connection between the quality of data provided and clinical relevance of a case [25].

3.2. Personalized treatment plan generation

AI that generates customized treatments based on the specific clinical information of each patient is expected to assist physicians in creating more precise treatment plans. In orthopedics, for instance, Li *et al.* claimed that generative AI could precisely adjust the rehabilitation plan based on the condition of individual patients to ensure training safety and achieve optimal results. Moreover, with the use of wearable technology and virtual reality, real-time monitoring and interactive guidance can be obtained in the process of rehabilitation, thus increasing the quality of rehabilitation at home [29]. In precision oncology, the technology is able to analyze multidimensional patient-specific information such as cancer type, TNM staging, histological type, molecular markers, and expression status to create highly customized recommendations for treatment plans [30]. In terms of medication management, Roosan *et al.* discovered that ChatGPT could accurately diagnose drug interactions, propose treatment plans, and formulate management plans for 100% accuracy of both simple and complex cases [31]. A study by Liu *et al.* also showed that although AI-generated clinical decision support recommendations are not widely accepted directly, they could give new perspectives and help experts rethink treatment plans [32].

Nevertheless, the body of evidence that supports the above-mentioned applications is still mostly preliminary. The studies use retrospective data, standardized small sample sizes, or preselected small

cohorts with minimal prospective or randomized validation in real life [30,31]. Despite the perfect accuracy found by Roosan *et al.* in carefully selected cases, the controlled environment hindered the applicability to the complexity of real-life practice [31]. Likewise, a meta-analysis of Watson for Oncology showed the high consistency in cases of early stages of cancer but significant heterogeneity and lack of consistency in cases of advanced stages of cancer or region-specific cases [30]. Thus, the key point differentiating feasibility from validation of an algorithm needs to be highlighted: current generative AI tools are supposed to serve as research aids.

While generative AI presents a lot of opportunities for individualized treatment planning, its clinical application poses several distinctive challenges. First of all, one can mention the fact that generative AI utilizes population-based data for the creation of standardized treatment plans, while decision-making in practice involves dynamic balancing considering the complexity of the specific case. What is more, treatment planning is an ongoing process requiring continuous learning from the patient's response to therapy. At present, generative AI cannot be considered a sufficiently flexible tool that can facilitate such an approach. Thus, generative AI should be seen as a tool for clinical decision-making, not as a decision-making tool itself.

3.3. Automation of clinical documentation and intelligent summarization

Generative AI is capable of transforming medical texts for specialists to patient-friendly language and automatically creating clinical notes from consultations between doctors and patients or from structured data. In such way, generative AI helps to decrease the documentation load on healthcare professionals.

Regarding patient-oriented conversion, LLMs may help translate complicated medical jargon and sentence structures used in discharge summaries into more readable and comprehensible text. As a case in point, one study revealed that the Flesch-Kincaid readability score for discharge summaries was decreased from 11.0 to 6.2 after conversion by the LLMs, while the Patient Education Materials Assessment Tool (PEMAT) comprehensibility score increased from 13% to 81%. Nevertheless, the safety challenges associated with potential information omissions and errors remain an issue during the conversion process, which is why the physicians must review the generated text [33]. During the process of generating summary of clinical conversations, ChatGPT can extract important information from the doctor-patient dialogue and create well-structured clinical documents. The results regarding the coherence, consistency, and fluency of such texts are close to human performance and even outperform those of some other dedicated summarization systems [34]. Therefore, it provides a feasible solution for creating medical records automatically, especially those containing descriptive data such as chief complaints and current medical history [35]. Additionally, for clinical encounters not conducted in English, open-source-based summarization tools have exhibited some level of effectiveness when it comes to producing preliminary discharge summaries from structured datasets. Still, challenges are there with regard to temporal reasoning, terminology, and completeness of the information. For instance, summaries produced by LLaMA3 had 2.84 errors on average per summary and only 60% were categorized as "complete" [36].

Nevertheless, methodological limitations impede the reclassification of these technologies into clinically validated interventions due to their current state. The overwhelming majority of available research utilizes small sample size and retrospective data from a single center with very little statistical power and generalization [33,34,36]. Moreover, measures such as Recall-Oriented Understudy for

Gisting Evaluation (ROUGE) reward lexical match instead of clinical accuracy, which results in a problematic situation when an algorithm can receive a high score while generating factual hallucinations or omitting important information [34,37]. Importantly, there is no validation of any of these algorithms using prospective randomized trial and regulatory benchmarking since the lack of multicenter randomized trials connecting output with clinical outcomes makes them promising research technologies but not clinically applicable solutions. Thus, until proper evidence is available, their application should be limited to documentation support tools only.

Generally speaking, generative AI is not only able to assist doctors in accomplishing documentation work easily but can also enhance the efficiency of medical information transfer and patient interaction through smart summarization, offering a structured knowledge base for clinical decision support [37]. Nonetheless, the existing systems depend on physicians in terms of their correctness, comprehensiveness, and standardization of terminology and perform inconsistently when processing complicated scenarios, temporal reasoning, and multilingual medical documents.

4. Application of generative AI in synthetic medical imaging

Generative AI is increasingly being employed in the field of medical imaging, and synthetic medical imaging is one of the major topics of research in this realm. Synthetic medical imaging intends to develop generative models to synthesize realistic medical images for issues like data paucity, data privacy, and multimodal data acquisition difficulties. Using the data distribution knowledge, it is possible to generate large amounts of synthetic images that resemble actual images. Figure 3 illustrates how this field is segmented into three different but complementary domains.

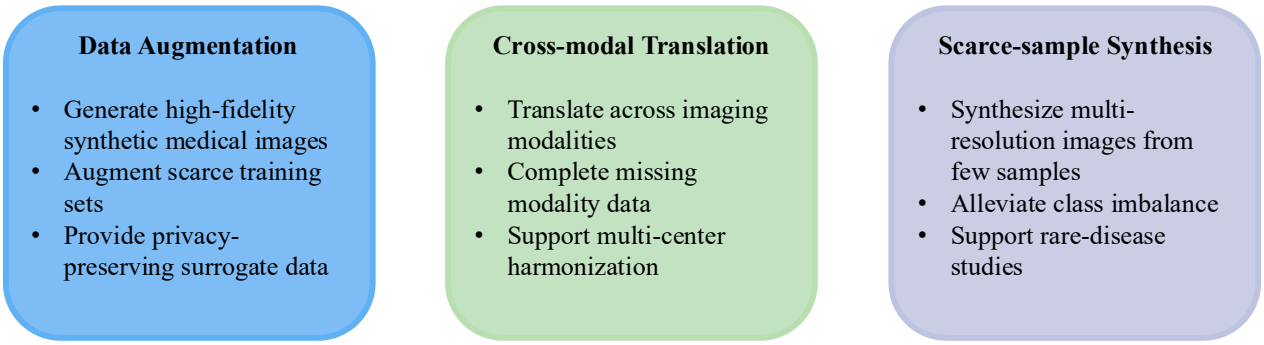


Figure 3. The application framework of generative AI in synthetic medical imaging.

Currently, GANs and diffusion models are the prevailing approaches to image generation, and they work very effectively when generating medical images. For instance, Iglesias *et al.* devised a diffusion model-based technique that is able to generate good quality synthetic medical images. The effectiveness of this system was measured with the help of a convolutional neural network, and the potential of using synthetic images in data augmentation was illustrated [38]. When it comes to the generation of 3D magnetic resonance imaging (MRI) images in AD research, scientists used cGAN, which can generate continuous 2D slices along the axial view in order to produce realistic and coherent 3D MR volumes [39]. Given that labeled datasets may be scarce, it was possible to use an auxiliary classifier GAN (ACGAN) to generate multi-resolution images with diagnostic attributes based on a small number of thyroid

ultrasound images. The visual realism and structural plausibility of the generated images were proven using the Fréchet Inception Distance metric and clinicians' ratings [40].

Apart from being used for the purpose of image synthesis, there is also an application of Generative AI in Cross-modal fusion generation. In this context, Biswas *et al.* have designed a CycleGAN model which can generate Verhoeffs-van Gieson (EVG) stained images out of the hyperspectral Hematoxylin and Eosin (H&E) stained images, thus accomplishing both staining conversion and modality conversion at once [41]. This highlights how Generative AI has potential in reducing the necessity for costly and time-consuming physical staining procedures in pathologic diagnosis through digital image synthesis. Pan Yongsheng *et al.* have discussed the advancements made from conventional approaches to deep learning.

This is used for image substitution, data completion, and multi-center data harmonization to support diagnostic tasks [42]. Sangeetha *et al.* proposed a Multimodal Fusion Deep Neural Network (MFDNN) that maps computed tomography (CT) image features, the Cancer Genome Atlas (TCGA) gene expression vectors, and electronic health records (EHR) clinical variables into a unified latent space via deep generative methods, subsequently generating fused representations for classification. Experiments showed a 92.5% accuracy in binary lung cancer classification tasks, suggesting that generative AI can achieve "virtual patient" representation generation and diagnosis at a multimodal level [43]. Hu *et al.* utilized a CT-US multimodal AI fusion system to combine preoperative CT and intraoperative ultrasound in real-time into an "enhanced new modality image" to assist puncture procedures. Applied during percutaneous nephrolithotomy (PCNL) for non-hydronephrotic renal stones, generating CT-enhanced ultrasound images increased the single-puncture success rate from 63.27% to 91.49% and shortened operative time by approximately 12 minutes [44].

Apart from these technological limitations, most studies are also limited by methodological limitations associated with retrospective and single-center designs, where diagnostic equivalence is rarely determined prospectively [42]. While perceptual metrics such as Fréchet Inception Distance (FID) determine visual similarity, they do not ensure retention of pathology that is critical for clinical decision making. Only a small number of studies, including one conducted by Hu *et al.* [44], have determined clinical endpoints while other studies limit themselves to algorithm benchmarking. Consequently, present uses of synthetic imaging can only be considered as potential areas for research and cannot be used clinically until diagnostic fidelity benchmarks and prospective regulations have been developed.

There is no doubt that generative AI shows immense promise for medical image synthesis in areas such as data augmentation, cross-modality, and small sample synthesis. Despite this, generative models suffer from issues such as instability during training and mode collapse. More precisely, in the synthesis of medical images, which are high-dimensional and multi-modal in nature, the realism and pathology correctness of the synthesized images need to be considered.

5. Application of generative AI in drug discovery

5.1. Molecular generation

There is a revolutionary change brought about by the generative AI in the domain of drug discovery. The technology, by virtue of learning and mimicking the distribution of molecules in the chemical space, can design new drugs that have biological activity. This significantly increases the speed at which the drugs are discovered. In the case of antibiotics discovery, the generative model called SyntheMol

introduced by Swanson *et al.* can design structurally unique small molecule antibiotics that can easily be synthesized from 30 billion molecules. The study has created 58 molecules for experimental validation, and six of them have shown an inhibitory effect against bacterial pathogens such as Acinetobacter baumannii [45]. This is the evidence of exploration of unknown chemical regions by generative AI. In cancer drug design, Joo *et al.* used (Conditional Variational Autoencoder) CVAE to combine the molecular fingerprints with GI50 (growth inhibition at 50%) values to generate new molecular fingerprints with anticancer properties. Then, they selected potential drug candidates from publicly available databases using the technique of similarity search [46]. AI technologies will allow increasing the efficiency and effectiveness of drug development through the analysis of a huge amount of data with improved accuracy [47].

Although there is quite a lot of promising proof-of-concept studies in favor of this approach, there is a significant translational gap between molecule generation via algorithms and clinical treatment. Most of the studies mentioned above, as well as others, are limited to preclinical validation via computational and biochemical means, and only a few reach the point of animal modeling and human trials [45–47]. However, the reason for this is that generative algorithms are prone to focusing on biological activity without considering such crucial elements as synthetic accessibility, pharmacokinetics, and adverse side effects that constitute the majority of reasons for high attrition rates in drug discovery and development [48]. Moreover, the datasets used in training by these algorithms are derived from biased and incomplete chemical libraries that raise concerns about the applicability of the molecules generated by algorithms [47]. Thus, despite being a powerful tool for exploration of the chemical space and identification of lead compounds, generative AI cannot be viewed as a platform for developing therapies yet.

Despite the substantial progress that generative AI has made in the field of drug discovery, its practical application in this field remains hampered by certain issues including those of data quality, model interpretability, and molecular synthesis complexity.

5.2. Clinical trial simulation and virtual validation

Generative AI is able to work without using lots of actual participants. It makes highly realistic virtual patients that mimic the effects of the drug in the body in order to perform initial assessment and optimization of drugs safety, efficiency, and dosage. Thus, it helps to reduce the risk of trial failures and wasting of resources.

Models like GANs may create a great amount of artificial patient data, which is consistent with real-world populations, using limited information about real-world patients. This helps build a virtual cohort covering a wide range of clinical parameter combinations. For example, the Artificial Clinic Intelligence platform uses GANs to construct a “patient virtual space.” Combined with virtual intervention permeability scores, it systematically evaluates the degree of response to a trial drug in patient subgroups with different clinical characteristics, enabling the identification of patient subpopulations most likely to benefit before the trial begins [49]. This generative AI-based virtual trial simulation not only provides prospective patient stratification rationale pre-enrollment but also dynamically predicts drug efficacy and toxicity in different populations during the trial design phase, reducing the risk of trial failure. CURATE.AI is an input mechanism-agnostic, small data-powered personalized drug delivery solution. CURATE.AI takes only patient’s own tumor marker data and dosage of the drug to create dose-response surfaces, which are specific to each patient, and personalize the administration of

chemotherapeutic treatments based on the efficacy of drug delivery. PRECISE CURATE.AI was successfully used to personalize the dose of capecitabine treatment in a pilot clinical trial for patients with metastatic solid tumors, showing the feasibility of personalized dosing in clinical studies [50].

However, some limitations have dampened the hopes of near-term clinical implementation. The ACI system is currently still a theoretical working draft and its GAN algorithm suffers from issues of training instability and model collapse due to its inability to handle the high dimensionality of small sample size omics data, with its synthesized results being expressly stated to be used for simulation purposes only [49]. Likewise, PRECISE CURATE.AI was conducted as a pilot study to investigate its feasibility and did not undergo randomization and survival end points, hence its claims on efficacy is exploratory at best [50]. In essence, current AI systems utilized for clinical trial simulation and dose optimization need to be understood as tools for research purposes for cohort enhancement and protocol design only until randomized prospective multicenter trials are done.

Once trials have been optimally designed, it is essential to identify patients that can be included in the trials. There are also clear benefits associated with the use of generative AI in this respect.

5.3. Precise recruitment of patients for clinical trials

Generative AI is slowly revolutionizing the conventional recruitment model through manual selection with its efficient and precise process of data handling and matching, helping overcome problems like inefficiency and restricted reach.

Generative AI is capable of using hospital big data systems to consolidate information from various sources like EHRs, LISs, and medical imaging to create standardized full-lifecycle patient databases. With the help of natural language processing and medical knowledge graph techniques, it is capable of automatically parsing trial eligibility and exclusion criteria and matching them with patients' features. This leads to marked enhancement of recruitment efficiency and precision. It was demonstrated that such an intelligent recruitment system managed to enroll more than 300 cases, with a conversion rate of 20.39% in the proactive reminders to physicians. For studies with rare biomarker targets, the system was able to match 60 patients, 15 of which were able to pass the screening process—a conversion rate of 25%. Consistency with the system's inclusion and exclusion criteria was as high as 60% [51]. In oncology clinical trials, AI-driven recruitment systems use machine learning algorithms to automatically mine EHRs in real-time mode and accurately identify eligible subjects, which significantly reduces the recruitment process duration and increases the conversion rates of the recruitment process. Their ability to integrate data from different organizations removes geographical restrictions and expands the recruitment scope [52].

Generative AI has proved useful even in recruiting patients for trials aimed at certain diseases. For instance, in a study of chronic subdural hematoma, the platform with an AI-enabled imaging analysis tool can automatically detect hematomas from non-contrast computed tomography scans and calculate some metrics such as the volume and midline shift. Subsequently, information about potentially eligible patients would be immediately pushed to the trial team. After the introduction of this platform, the monthly recruitment rate increased from 0.99 to 1.35, representing a 36% increase in the enrollment rate, with a positive predictive value of 81.4% [53]. With regard to cognitive disorders such as Alzheimer's disease, generative AI could also be integrated with language models to create intelligent pre-screening chatbots capable of analyzing symptoms or voice characteristics of patients to determine the risk level.

The system will also integrate with electronic health records and use it to automatically match patient data with trial eligibility criteria [54].

Although there are efficiency benefits that can be gained by using AI for subject recruiting, its implementation also faces some difficulties regarding data integrity, ethical norms, and integration of the system. Efficiency of the recruitment process is greatly influenced by the standardization of the hospital's database. In cases where inclusion/exclusion criteria demand complex judgment or rare tests, generative AI plays a minor part [51]. Besides, when advocating for these kinds of systems at other centers or even regions, problems like data interoperability, patients' privacy, and fairness of algorithms have yet to be solved.

The use of generative AI has slowly started to seep into all aspects of drug discovery process, creating a technological continuum ranging from target identification to post-marketing surveillance. In the stage of target identification and validation, generative AI is used for predicting possible disease targets through omics data and literature knowledge to facilitate the target prioritization process [45,55]. In the lead compound discovery stage, generative models are able to come up with new molecular structures of desired biological function, which helps generate novel molecules in the chemical space exploration [45–47]. In the preclinical research stage, generative AI is used to improve the ADMET characteristics of the compounds and predict their *in vivo* efficacy and toxicity through virtual clinical trials [48–50]. In the clinical trial stage, generative AI helps in the creation of virtual patient groups, optimal dosing strategies, and participant recruitment [49–54]. At last, in the post-market surveillance stage, generative AI analyzes real-world data for pharmacovigilance and indication extension. Figure 4 illustrates an end-to-end intelligent workflow process starting from the early discovery stage to post-market surveillance. An end-to-end intelligent workflow process will help in reducing the overall time taken for developing drugs, which is normally 12–15 years long and takes around $1 billion [56].

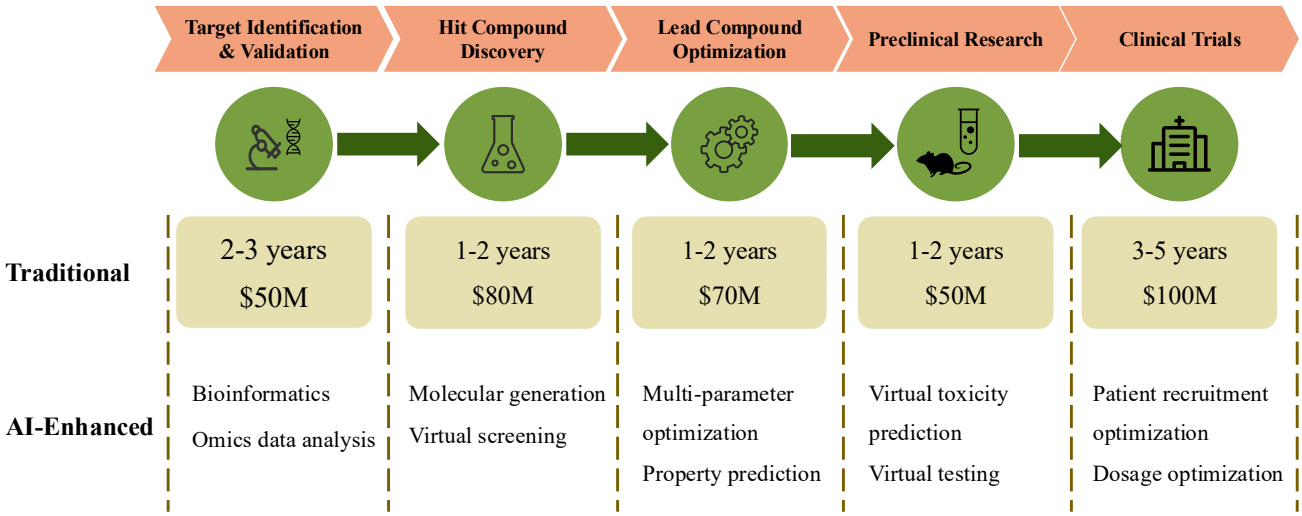


Figure 4. The framework of the process for drug discovery driven by generative AI.

Nevertheless, the existing scientific literature that supports the idea of generative AI being employed for recruitment is methodologically immature. Currently, only retrospective pilot single-center studies have been carried out with small sample size and without any control group [51,53]. Also, the outcomes that are considered in these studies reflect only workflow efficiency and not the effectiveness measured using some clinical criteria like randomization and trial completion. In the same way, although AI

matching systems have been suggested for recruitment in case of cancer [52] and Alzheimer's disease [54], they are still more theoretical and no randomized prospective trials that would compare AI-based recruitment to the usual recruitment process exist.

6. Application of generative AI in medical literature processing and knowledge synthesis

The emergence of generative AI has led to significant changes in the way that medical literature search, management, and synthesis is done. Through its ability to comprehend and generate natural language, generative AI will be able to assist in finding relevant literature within large volumes and also in summarizing and reviewing these pieces of literature.

Intelligent retrieval and filtering allow generative AI to understand complex user intent and carry out literature retrieval with high precision. The conventional Boolean searches rely on the use of particular syntax, while specialized tools that make use of natural language processing are capable of understanding queries formulated by the users using natural language and performing multi-stage literature retrieval, which includes quality assessment and filtering for relevance. In a pilot study in the field of vascular surgery, the use of AI in searches resulted in a much higher efficiency of the process being carried out in an hour and achieving higher precision of 48.6%, as compared to a manual search, which took 10 hours to perform, had 16.9% precision, and yielded no irrelevant results [57]. This shows the potential of generative AI in carrying out targeted literature retrieval. Nevertheless, the use of general large language models for real-time literature retrieval for systematic reviews results in poor performance. In a study, it was shown that ChatGPT, built on GPT-3.5, returned the lowest number of directly relevant articles (0.5%) for disease treatment topics, as well as numerous "hallucinated" citations [58]. On the other hand, although Bing AI proved more accurate in terms of its search results, the total amount of articles obtained was still not enough for conducting an independent systematic review [58]. In the future, the development of retrieval AI will shift towards specialization according to different scenarios [59].

Regarding content creation and summarization, generative AI can automatically write a literature review or an abstract based on search results or certain instructions, which will serve as a preliminary synthesis of knowledge for researchers. Studies have proven that, by iterative rapid engineering, GPT-4 can rapidly create draft literature reviews, incorporating a variety of factors in relation to various complex topics, showcasing a good breadth of knowledge [60]. During an assessment of musculoskeletal radiology, GPT-4 created narrative reviews of all relevant topics to a given topic and achieved the average rating, which was higher than the rating of a specialized AI tool (literatures.com) [61]. Thus, it can be used as a convenient source of a quick overview of the subject. However, there are serious limitations of the current technology. AI-based reviews lack the knowledge depth and context comparison and might miss some relevant literature or include incorrect data [60,61]. For instance, even when evaluating musculoskeletal radiology, GPT-4 did not manage to find all the relevant articles [61]. As a result, any AI-generated abstracts or reviews should be verified and polished manually; it serves more as an assistant research associate than an author [60].

Though these proof-of-concept studies are promising, current assessments suffer from the methodological restrictions of retroactive, topic-specific designs without any standardized criteria or clinical endpoints. Gwon *et al.* demonstrated that ChatGPT was able to find 0.5% of directly relevant trials, 16.4% of the outputs had fictitious bibliographic information, and Bing AI could not retrieve enough articles to conduct the independent evidence synthesis [58]. Mostafapour *et al.* also revealed that

GPT-4 starts hallucinating about non-relevant parameters once it surpasses the knowledge saturation point [60], and Jenko *et al.* confirmed that GPT-4 cannot recognize seminal articles in musculoskeletal radiology [61]. It is important to mention that none of the literature tools based on generative AI has been prospectively validated by several centers yet, nor any of them has been approved as a clinical decision-making aid. The crucial difference between algorithmic and clinical validation should be emphasized here: current AI tools are not meant to be used as autonomous tools of evidence-based medicine but as research tools for hypothesis generation and framing.

To conclude, the use of generative AI in the realm of literature can be applied in optimizing search capabilities and helping in writing peer reviews. The key advantages of the technology include speed and capacity to work with large amounts of information, which allow it to serve as a research assistant capable of quickly summarizing, finding connections and producing texts. However, it cannot yet serve as an independent tool for performing comprehensive searches and producing reliable reviews due to limitations concerning its tendency to produce "hallucinations" and inability to conduct a contextual analysis. Consequently, it should serve merely as an assistance tool for researchers in their exploratory and writing stages.

7. Application of generative AI in medical education and training

Generation AI is transforming medical education on several fronts. As depicted in Figure 5 below, the use of generation AI in medical education is diverse but falls into three main categories that include personalization, virtual patients, and assessment and feedback systems. The following sections elaborate on these categories.

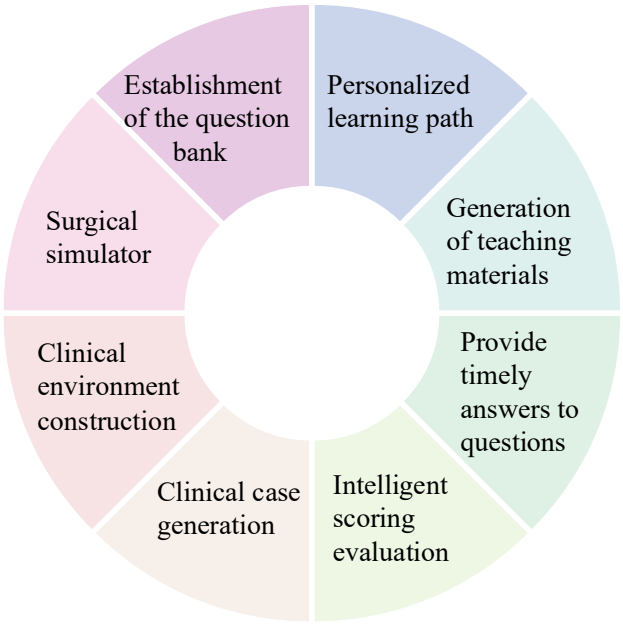


Figure 5. Applications of generative AI in medical education.

7.1. Personalized learning and teaching support

Generative AI helps to create a personalized learning journey for the students. For instance, in the course of Biostatistics, generative AI can help evaluate the prior knowledge of the students using placement tests, personalize the learning modules, generate curiosity among students about learning, and develop

practical skills [62]. With the help of generative AI technologies that assist literature organization, concept development, and academic writing, nursing students can benefit from self-directed learning, which is efficient both in terms of gaining new knowledge and writing skills [63]. Also, generative AI technologies are able to provide immediate feedback along with explanations. As proved by Kung *et al.*, ChatGPT has a high reasoning and clinical insight when providing answers to medical questions with internal consistency of questions and answers being 94.6% [64]. Further investigation through qualitative research by Naamati Schneider revealed that, during their interaction with ChatGPT, 88.9% of the students indicated that they had gained one important insight, which was defined by "novelty, non-obviousness, and validity" that enabled them to expand their minds [65]. Students should also be cognizant of the problems related to the accuracy of information, as it was revealed through the research that, although the use of generative AI is very practical, there is a risk of error diagnosis in the field of medicine of more than 50% [63,65].

Generative AI can help the instructor save much of their effort and optimize the process of designing the course. The teacher will be able to use the generative AI to build the learning materials, such as test questions, case studies, and knowledge graphs. As a result, the systematic instructional materials will be created [62,66,67]. For instance, in pharmacology, Sridharan *et al.* found out that the automatic generation of different types of questions (for example, multiple-choice questions and short-answer questions) is possible with the help of AI. Moreover, the accuracy of generating type A multiple-choice questions reaches up to 60–90%. At the same time, this approach is able to develop OSPE cases that are "technically accurate and appropriate for the learners' level" [66]. In nursing education, ChatGPT helps teachers analyze student performance data, providing personalized teaching suggestions, thus enhancing instructional effectiveness [67]. Moreover, generative AI enables instructors to develop intelligent databases of cases and questions. For instance, integrating the real world medical data in biostatistics can add practicality and up-to-date content in teaching [62]. The emerging role of instructors also demands updating their teaching techniques where they should act as guides rather than transmitters of knowledge along with collaboration with generative AI for improving instructional quality [62,65]. Specifically, instructors need systematic training on developing AI-integrated courses through deep integration of AI technologies in learning objectives and cases, optimization of instructional design in order to enhance its quality and clear articulation of learning objectives in order to increase student's engagement [68]. In another study conducted by Pang *et al.* and using the method of structural equation modeling on 523 medical students, it was shown that teaching quality and external environment have an indirect impact on learning outcomes through enhancing learning motivation (mediation effect of 0.343 and 0.287, respectively) [69]. It means that instructors are able to effectively motivate their students in AI integrated courses. Educators, on the other hand, themselves need systematic "AI literacy" training in order to increase their engagement and pedagogical technical knowledge in AI courses [70].

Despite the existing literature highlighting the potential of using generative AI for teaching in medicine, the current state of knowledge about such applications is methodologically limited. In most cases, evaluations use retrospective design or one-site pilot studies involving assessments of self-reported results of learning, with very limited number of randomized controlled trials testing long-term competency acquisition [62,63,68]. Additionally, all evaluations have been done at Level 1 and 2 of Kirkpatrick's framework and did not include assessments of changes to clinical practice and performance resulting from

learning with AI [68]. Thus, such interventions can be regarded only as interesting areas of research and need further validation by experts and against the competency framework [64,66].

The strengths of generative AI are quite evident when it comes to personalized assistance and the creation of resources for theoretical learning. However, at the same time, its shortcomings become clear as well. First of all, the veracity of the generated information still needs human verification, and excessive use can negatively affect the development of students' critical thinking skills. Secondly, the systematic inclusion of AI into the curricula of different medical programs is yet another obstacle.

7.2. Virtual cases and clinical simulation

Using generative AI, the creation of highly realistic virtual patients and disease scenarios makes it possible to emulate complex clinical scenarios that happen in practice. This way, students get an opportunity to practice their skills of diagnosis, treatment, and communication in a risk-free environment. For instance, when teaching pathophysiology, ChatGPT can be utilized to create various clinical scenarios involving patients of different genders, genetic composition, and social status [71]. ChatGPT provides dynamic experiences for medical students by allowing them to communicate with simulated patients via free-text dialogue. Based on the information about the patient provided to them, students can ask for examinations, treatment or any other intervention, which will change the disease scenario instantly [72]. Wu Fangqin *et al.* developed a generative scenario-based virtual teaching system that dynamically generates cases and disease progressions, allowing students to practice repeatedly without risk. This system has been used in the learning of 328 students from 7 classes at their own institution and 2854 students from other schools, effectively enhancing the sense of presence and engagement [73]. Nevertheless, it is necessary to consider possible errors that may appear in the generated content, and this requires the checking of the teacher to make sure about the clinical logical correctness [72]. One more proof of the interactivity of generative AI is the possibility to model doctor-patient conversations, such as the use of natural language dialogues to practice the explanation of diagnoses or conversation with the family, lowering students' psychological pressure in real situations [71]. Generative AI creates a flexible and highly customizable platform for clinical simulations for medical education purposes. The area of robotic surgical training witnessed the development of the HoloLens 2-based augmented reality pancreatic surgery simulator by Pisla *et al.* The key element of this system is the implementation of the EfficientNetV2-B0 CNN model for prediction of real-time "instrument-pancreas" interaction forces without using physical sensors. It resulted in the average mean absolute error of 0.0244 N and latency of 66 ms. It serves as the proof-of-concept of vision-based haptic feedback for minimally invasive surgery training [74].

Unfortunately, the existing body of evidence in this field is mainly descriptive and methodologically limited. It consists of the majority of studies with single center designs assessing subjective satisfaction or knowledge acquisition outcomes [71,73] and lacking comparison to established simulation methods or Objective Structured Clinical Examination scores. Moreover, the impressive technical metrics reported, such as the force prediction accuracy in the AR surgical trainer [74], constitute engineering proof-of-concept rather than clinically validated interventions, as none have been prospectively linked to patient-level outcomes, received regulatory clearance, or been integrated into competency-based curricula. Hence, the use of such platforms today should be seen as a kind of experimental tool that still requires extensive validation to be classified as proven means of training.

The generative AI technology presents a scalable solution for the simulation of clinical skills, which is quite useful in case of resource deficiency. Nevertheless, there is still a problem in the logic of the simulation cases created using the current generation technology, as they do not correspond to the complexity of the reality of clinical cases and require a strict review by teachers in terms of medical correctness and logic of the development of diseases.

7.3. Intelligent assessment and feedback systems

In the aspect of assessment, generative AI can automatically generate various types of questions, such as multiple-choice and case analysis, which makes the evaluation content more diverse and closer to reality [75]. For example, through the application of generative AI, fixed-type questions can be changed into comprehensive ones, which are associated with clinical practice, so that students' overall ability of handling clinical cases and their professional skills can be evaluated [76]. Moreover, intelligent scoring is supported by generative AI especially for subjective questions, including short answers and case analyses. With the help of natural language processing technology, it is able to do multi-dimensional evaluation based on the pre-set scoring criteria, which improves the fairness and consistency of the evaluation [75]. Meanwhile, it can also monitor and document students' learning process in real time, such as the behavior of online learning, problem-solving, and collaboration skills, and construct the indicators of comprehensive evaluation through process data, which enables the shift of evaluation mode from "exam score only" to "process and results together" [76]. Although information accuracy is still a possible concern, RAG can increase the reliability of assessment [77].

Regarding the issue of feedback, generative AI enables timely, personalized learning feedback, which will greatly promote the improvement of teaching and learning. The use of AI systems can help gather information about students' learning, such as their assignment scores, exam grades, and class performance in order to conduct real-time assessment and analysis of individual learning performance, which makes teachers be aware of students' learning condition and customize their guidance [76]. With respect to students, generative AI is capable of providing clear explanation of answers and interactive feedback for difficult concepts learning and application of the acquired knowledge. For instance, question-answering systems using the RAG technology can give accurate and trustworthy medical knowledge resources along with reference materials, which allows students to have in-depth learning and explore on their own [77]. Besides, with the help of generative AI, one could provide students with personalized recommendations concerning the learning path according to their progress and shortcomings, promoting self-directed learning [78]. Yet, it should be noted that too much dependence on the generative AI could affect critical thinking development, therefore feedback should focus on the rational evaluation of the material generated by AI [78].

Despite this clear functional proof of concept, key methodological issues prohibit moving these assessment generators from being a proof of concept to clinically validated tools for use in education. Current literature on this subject depends nearly entirely on retrospective, single-center data sets that dramatically reduce the ability to generate valid statistics and generalize [75,76]. Commonly used NLP metrics like ROUGE evaluate lexical similarity and not clinical or educational similarity, posing the dangerous possibility that highly scored outputs are filled with factually inaccurate information or dangerous omissions [77]. Most significantly, all of these programs have not been tested with prospective randomized trials in a medical educational environment. Without these multi-center trials,

which could demonstrate correlation between AI-generated assessments and hard educational outcomes, like licensing exam performance or knowledge retention, these models must be considered research tools only. They must be kept in this realm until further evidence is available.

The incorporation of AI into the evaluation process through automation could help save time for teaching. However, the criteria for the evaluation process in the form of AI is based on some sort of bias that is embedded in the training dataset. It is also not clear how the process-oriented evaluation data should be interpreted or applied in practice, and intelligent feedback systems are still weak in many areas like clinical reasoning and humanistic traits.

7.4. Crosslinguistic performance of generative AI in medical licensing examinations

Although generative AI systems like GPT-4 have exhibited almost expert performance levels in English-based medical licensing examinations, the proficiency of such AI systems in languages other than English needs further investigation since medicine requires an understanding of the local language, culture, laws, and epidemiology. The evaluation of LLMs in diverse languages is crucial before deploying them in global medical education and practice.

For the purposes of this comparative analysis, we examined studies which tested GPT-4 and GPT-4o on national medical licensing examinations conducted in seven different languages: English, Chinese (NMLE of the Mainland and SPTEMD of Taiwan), Japanese, Korean, and Spanish. Multiple-choice examination questions were used in all tests while image-based questions were not included if required.

Table 4 demonstrates that GPT-4 always surpassed the passing threshold in all tested languages, with accuracy from 66.2% (Korean traditional medicine) to 95.7% (Taiwanese SPTEMD). Nonetheless, a notable performance difference based on the language is apparent, with GPT-4 providing 95.7% for Taiwanese SPTEMD, while it gives 66.2% for Korean K-NLEKMD and 72.3% for mainland Chinese NMLE. GPT-4o helped to improve the performance of GPT-4 in non-English languages, bringing the accuracy of mainland Chinese NMLE to 88.2% and Taiwanese SPTEMD to 96.7%.

Table 4. Performance of GPT-4 and GPT-4o across different languages.

Language	Examination	GPT-4 Accuracy	GPT-4o Accuracy	Passing Threshold	Human Average Accuracy
English (USA) [79]	ABNS Self-Assessment Exam 1	83.4%	Not reported	69%	73.7%
Chinese (Mainland) [80]	NMLE 2021	72.3%	88.2%	60%	Not reported
Chinese (Taiwan) [81]	SPTEMD Stage 1	95.7%	96.7%	60%	Not reported
Japanese [82]	JMLE	79.9%	Not reported	Essential knowledge: 80%; Others: 74.6%	84.9%
Japanese [83]	NMLE	81.5%	Not reported	72%	Not reported
Korean [84]	K-NLEKMD	66.2%	Not reported	60%	76.7%
Korean [85]	KPLE	86.5%	Not reported	60%	72.0%
Spanish (Peru) [86]	ENAM	86.0%	Not reported	52.8%	55%

There are two key aspects responsible for the observed differences. First, cultural and legal specificity—a model may prescribe something that would be illegal or unethical in the target country, such as euthanasia in Japan or misinterpretation of the specific Chinese medical policy [87]. Second, localization—the models may lack terminology and regional medical data [84].

These results imply something important in relation to the use of generative AI in teaching medicine. Though GPT-4 and GPT-4o may be used effectively as tutors for exam preparation in several languages, teachers should inform their students about possible errors on culturally sensitive issues and those limited by law. In this case, retrieval-augmented generation (RAG) will help solve the problem to some extent.

8. Challenges of generative AI in medicine

The challenges will be structured into four layers that interrelate with each other, namely the data layer, the model layer, the application layer, and the system governance layer.

8.1. Data level risks

Medical training data is sensitive, biased over time, and structurally incomplete. The decades-long tradition of male-centric and Eurocentric medical research have yielded datasets in which the anatomy of white males is taken as a standard template against which all others are exceptions [88,89]. A popular commercial risk prediction model gives African American patients scores that are 46% higher compared to white patients having the same illness, thus undervaluing their health care needs [90]. Meanwhile, the error rate of commercial gender recognition technology is 34.7% in case of women with dark skin and 0.8% or less for men with light skin color [91]. All these facts demonstrate how societal biases have been translated to digital in the form of data and further validated by technological systems [89].

Privacy threats amplify the risks of representational problems. Generative models have the ability to memorize and reproduce the real-life training examples, thus allowing for the reverse engineering of patient identity [92]. Despite the suggestion of the use of synthetic data as a solution to address the problem of privacy, benchmarks concerning quality of synthetic data and privacy and representational unbiasedness issues are not yet fully developed; insufficiently validated synthetic data may solve some ethical concerns but raise others [92].

Bias and lack of adequate data only increase the level of bias and lack of integrity in data. There is always some form of bias on the part of humans performing the annotations. When marginalized communities have been under-represented, their disease phenotypes will not be learned by models. The absence of diagnosis creates a lack of necessity for data collection, hence creating a self-perpetuating cycle [90]. Solving these problems would entail employing a participatory design approach where the affected community is at the heart of data governance [89,90].

8.2. Model level pathologies

Since the probability-based approach used by generative AI contradicts the medical need for accuracy, three related problems occur: hallucination, “black-box,” and amplification of bias.

Big language models produce output that is very sure but actually wrong or false [93]. In an instance reported by Eysenbach, ChatGPT made up a fake research paper and gave it a fake DOI when asked to summarize the literature [93]. This pathology has been systematically validated by Guleria *et al.*: in

response to being asked to write a research paper, ChatGPT provided numerous references that seemed valid on their face, but all were made up [94]. These invented citations were not identified by GPTZero, while the plagiarism check revealed 0% similarity, hence, making all conventional checks obsolete [94]. Doyal *et al.* argued that information disseminated by LLMs is presented in an authoritative way, and thus makes clinical reasoning mistakes extremely deceptive [95]. There are three types of medical hallucinations: factual hallucinations, source hallucinations, and reasoning hallucinations, each demanding unique approaches to their mitigation.

The lack of transparency increases this danger. The black-box approach to deep generative models makes it impossible to explain decision-making processes, violating the principle of transparency [92,95]. If the model does not diagnose heart diseases among women correctly, the reasoning process cannot be used for human verification [88].

Not only does bias persist through the process, it becomes exaggerated through the technical process. In terms of feature representation, word embeddings create stereotypic relationships within the vectors spaces [89]. Architecturally, objective functions, which are concerned solely with predictive power, are blind to the fairness constraints that exist for the various sub-groups [89]. Evaluation-wise, benchmark tests lack population stratification and permit the optimization to majority groups while concealing performance failure to minorities [89,92].

Solutions to this problem would involve uncertainty quantification for the identification of unreliable results, retrieval-based generation from trusted medical information sources, and critically, cross-domain bias audits that consider the combined effects of race, gender, age, and economic background rather than simple statistical parities [88,92].

8.3. Application-level risks

Deployment of the technology within the clinical workflow and academia brings about particular socio-technological challenges. Misdiagnosis by using AI systems would place the responsibility not only on data providers but also on AI model developers, IT department of the hospitals, and clinicians [92,95]. There is no tiered system of accountability in governance at present, when “human-in-the-loop” is just an illusion, as it is impossible for the clinicians to validate the output of the AI [95].

In scientific writing, the rapid development of the AI technology exceeds the existing regulatory mechanisms. In their study, Guleria *et al.* proved that the content generated by AI could easily escape detection by plagiarism detector software and AI detection tools and thus become a high-fidelity forgery pipeline [94]. The listing of AI as one of the co-authors, which has precedents in scientific literature, goes against the principle that the authors should be responsible for the integrity of the article [94,95]. While the practice of AI authorship is strictly prohibited by all reputable publishers, it is hard to enforce due to the limitations of current detection technologies [94].

AI-assisted generation comes with the danger of suppressing creative thinking and reasoning. In the case where the patients tend to always take AI-provided differential diagnoses or medical students employ chatbots without cross-checking them, there will be cognitive offloading, which means that the specialized knowledge that pertains to medicine will be lost [95]. According to Ning *et al.*, ethical disclosure should be made part of peer review by using standard checklists for this purpose [92]. In the clinical context, generative AI should officially be categorized as a decision support system and not as

an autonomous system, and the usage of AI should compulsorily be recorded in all clinical records and research papers [94,95].

8.4. Systemic governance gaps

These layers make up a chain reaction where risk is reproduced because the historically underrepresented communities are reinforced at each layer: underrepresentation in the datasets is reflected in the model development process, which results in biases that cause dangerous clinical implications, loss of trust, and ultimately further exclusion of these communities in the future data collection. To overcome this vicious circle of AI risks, it is necessary to address them structurally rather than technologically.

Lau's analysis of data feminist critique of the EU AI Act shows that the narrative of regulation understands bias as a technical problem that can be corrected [88]. Iloanusi and Chun maintain that algorithmic fairness is impossible as long as the power imbalances within healthcare organizations persist regardless of de-biasing of the models [90]. Bias audits conducted in the form of compliance checklists and without community involvement can lead to legitimization of inequalities in the name of procedure justice [88].

The governance of generative AI should be considered throughout the whole lifecycle of the technology. At the data level, this means co-designing with the impacted communities and data sovereignty [88,90]. At the model level, this means using cross-equity constraints, uncertainty quantification, and causal inference modules. From the standpoint of application, this calls for a layered accountability framework, rules for AI transparency, and technology for detecting AI [92,95]. According to Liu, this type of governance mechanism represents an ecosystem consisting of government regulation, corporate accountability, civic engagement, and international coordination [89].

The credibility of generative AI is unlikely to increase due to the development of more models and data. It is necessary to change one's mindset from technological solutionism, which presupposes that technological solutions to all problems are needed, to a different approach involving governance and treating data epistemology, model architecture, clinical process, and regulation design as an integrated whole. It is only by putting the perspectives of historically marginalized communities in medicine at the forefront that generative AI will not cause the injustice of large-scale automation.

9. Future outlook

9.1. Trustworthy architectures

Future success calls for the inclusion of clinical credibility in the model architecture rather than applying post-hoc solutions. The challenge of automatic detection of clinically significant hallucinations is a big one. While existing benchmarks concentrate on the open-domain tasks, the development of clinically-relevant hallucinations implies the creation of a specialized classification system that should differentiate between factual hallucinations, invented sources and reasoning mistakes. Eysenbach found out that ChatGPT invented DOIs (Digital Object Identifiers) to publish fake papers [93] and Guleria *et al.* discovered that GPTZero was unable to detect the existence of AI-generated references, while plagiarism software found only 0% similarity [94]. It is a very risky loophole because the current technology for detecting these flaws can detect all fabricated references, while the riskier mistakes, such as wrong

anchoring in differential diagnosis, or wrong sequence of events in the course of the disease, remain unnoticed and pose a threat to the health of the patients. The future development calls for the creation of the clinically-relevant hallucination benchmark by specialists with annotation at the level of reasoning and severity ratings related to patient safety outcome

Of equal importance is the construction of RAG in non-stationary evidence contexts. While RAG models associate their output with medical knowledge base resources [77], clinical guidelines are continually changing; for instance, guidelines related to oncology or infectious diseases can change within a few months' time. Existing designs do not incorporate any temporal reasoning component; hence, it is impossible for them to discount outdated trials while keeping the consistent anatomical knowledge base information intact and to handle conflicting guidelines apart from selecting the one that has the most citations [58].

In addition, uncertainty quantification needs to be operationalized as opposed to just being another statistical measure. Though the employment of UQ to signal uncertain predictions has been suggested [92], there is no agreement on how cognitive uncertainty can be mapped to threshold clinical decisions. There is a need for future research to develop uncertainty quantification processes that would automatically trigger referral to experts once the uncertainty passes validation thresholds for certain rare conditions and/or complications.

9.2. Equitable data ecosystems

It is impossible to ensure clinical feasibility without removing deeply ingrained structural biases from the healthcare data ecosystem. One of the most promising yet understudied methods involves conducting cross-bias audit for synthesized medical images. GANs and diffusion models become increasingly popular for augmenting data sets; nevertheless, when trained on biased historical corpora, they tend to produce synthesized populations that are underrepresented and anatomically unrealistic, thus creating an illusory sense of diversity. According to the data-feminist critique by Lau, traditional bias audits approach discrimination as an issue that could be fixed with a technical fix, whereas it is a structural product of discriminatory knowledge production process [88]. At the same time, Iloanusi and Chun show that algorithms tend to exclude racial and ethnic minorities more due to their underrepresentation in training datasets [90]. In addition, Liu highlights the fact that gender biases are reinforced through data collecting, algorithm creation, and deployment processes [89]. It is essential to conduct further benchmarking to verify whether the synthesized images of dark-skinned women or elderly people have pathological and anatomically valid features.

Federated learning has been extensively employed as a way to eliminate data siloing while preserving privacy; however, as highlighted by this study, most of the primary care facilities are located in economically advanced areas [4,5]. Such an inequality can result in the exacerbation of the federated bias problem, where gradient contributions from densely populated urban hospitals outweigh the global model due to statistical representation, thus unintentionally encoding the properties of the majority population group. Further studies should focus on federated learning approaches that guarantee fairness, such as using demographic parity as an aggregation constraint.

In addition, the cross-lingual performance disparities seen in Table 4 are indicative of broader epistemological differences, as shown in [79–86]. According to the findings presented by Kasai *et al.*, the problems faced by GPT-4 during Japanese medical tests have a lot more to do with legal barriers

specific to that culture than anything else, especially when it comes to issues like euthanasia [87]. Any evaluation framework developed in the future should therefore transcend translation criteria and begin to consider the ability of models to adjust treatments according to epidemiology, drugs available, and legalities in their respective cultures.

9.3. Human AI governance

Patient-centered care entails putting into practice oversight concepts in a way that will translate into effective workflows informed by evidence. In this paper, the concept of "the human in the loop" is critiqued, noting that responsibility is shared by those providing the data, those creating the machine learning algorithms, and those who implement their use on the frontlines, with providers being unable to find time and technical expertise to validate AI products [93,96]. Human factors work needs to be conducted in future in order to understand how to effectively intervene in high-throughput workflows, for instance whether the discharge summaries created by LLMs need complete validation from physicians or stratification and selective reviews will do the job [33,36]. The MetricTree model by Drukker *et al.* offers a framework for choosing evaluation metrics that will be suitable for particular tasks [96]; this sort of framework is necessary to determine appropriate levels of supervision based on clinical risk. Similarly, there needs to be quantification of differences in automation bias in different settings of emergency, outpatient, and surgery.

Generative AI can help students learn through personalized pathways and simulation techniques in their medical training, but the long-term cognitive impacts have yet to be studied empirically. While Naamati-Schneider observed that 88.9% of students felt they acquired new insights from engaging with ChatGPT [65], Doyal *et al.* suggest that constantly offloading one's cognitive function by relying on AI-generated, customized information might hinder the development of the critical thinking ability that defines professional behavior in the field of medicine [95]. Further empirical studies would use longitudinal cohort designs comparing students trained with AI-generated cases and those in conventional programs by taking into account licensing exam results and diagnostic reliability in unique situations.

Finally, there must be a transition from pre-market, static frameworks to dynamic post-deployment governance frameworks. As highlighted by this paper, "the rapid pace of development in generative AI technology has exceeded that of its regulation" [94]. Ning *et al.* also point out that static checklists will not be able to stop ethical infringements due to continuous adjustments of models [92]. In future, the regulatory science should aim at developing pharmacovigilance-like frameworks to oversee post-deployment hallucination rate, bias changes, and any adversarial attacks involving poisoning the system. It is essential that such frameworks be coupled with an accountability framework that delineates responsibilities among model developers, IT hospital staff, and clinicians in cases of negative AI outcomes.

10. Conclusion

In summary, there have been numerous advancements in medicine as a result of generative AI in terms of decision support systems, synthetic imaging, drug discovery, literature analysis, and medical education. However, such an innovation has some limitations like hallucinations, incorrect information, data security issues, algorithmic biases, lack of transparency, and inadequate regulations. These factors hinder the large-scale use of generative AI, which can currently serve only an assistive role and cannot

fully replace clinicians. It is expected that further development of AI in a responsible manner will be determined by several factors.

Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors used generative AI tools only to improve language and readability. Specifically, the authors used GPT-4 for language polishing only. The authors take full responsibility for the content of the manuscript.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant No. 62571132 and No. 62571131).

Authors' contribution

Conceptualization, Y.L. and L.L.; methodology, Y.L. and T.Z.; investigation, Y.L. and Y.L.; writing—original draft preparation, Y.L.; writing—review and editing, Y.L., Y.L., L.L. and T.Z.; visualization, T.Z.; supervision, L.L.; project administration, L.L.; funding acquisition, L.L. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

The authors have no relevant financial or non-financial interests to disclose.

References

[1] Wang Z, Shi L, Lu L. Modern medical model: difficulties, challenges and opportunities. *Chin. Med. Ethics* 2019, 32(7):837–839.

[2] Yan Y, Qin X, Fan J, Wang L. A review of big data research in medicine & healthcare. *e-Science Technol. Appl.* 2014, 5(6):3–16.

[3] Xia C, Li J. Healthy aging or suffering aging?—prevalence of chronic disease in elderly population in the view of healthy China strategy. *Explor. Free Views* 2018, (10):115–121.

[4] Xia Y, Zhang H, Duan X, Zhu W. Study on the equity in the allocation of primary medical and health resources in China in the last decade. *Health Econ. Res.* 2023, 40(6):1–6.

[5] Qin A, Qin W, Hu F, Wang M, Yang H, *et al.* Does unequal economic development contribute to the inequitable distribution of healthcare resources? Evidence from China spanning 2001–2020. *Glob. Health* 2024, 20(1):20.

[6] Miller KD, Nogueira L, Devasia T, Mariotto AB, Yabroff KR, *et al.* Cancer treatment and survivorship statistics, 2022. *CA Cancer J. Clin.* 2022, 72(5):409–436.

[7] Marques L, Costa B, Pereira M, Silva A, Santos J, *et al.* Advancing precision medicine: a review of innovative in silico approaches for drug development, clinical pharmacology and personalized healthcare. *Pharmaceutics* 2024, 16(3):332.

[8] He K, Mao R, Lin Q, Ruan Y, Lan X, *et al.* A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Inf. Fusion* 2025, 118:102963.

[9] Zhou H, Liu F, Gu B, Zou X, Huang J, *et al.* A survey of large language models in medicine: progress, application, and challenge. *arXiv* 2023, arXiv:2311.05112.

[10] Heng Y, Ma Y, Khan FG, Khan A, Ali F, *et al.* Survey: application and analysis of generative adversarial networks in medical images. *Artif. Intell. Rev.* 2025, 58(2):39.

[11] Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. *JMIR Med. Educ.* 2023, 9(1):e48785.

[12] Ghebrehiwet I, Zaki N, Damseh R, Mohamad MS. Revolutionizing personalized medicine with generative AI: a systematic review. *Artif. Intell. Rev.* 2024, 57(5):128.

[13] Cheng Y, Gong Y, Liu Y, Song B, Zou Q. Molecular design in drug discovery: a comprehensive review of deep generative models. *Brief. Bioinform.* 2021, 22(6):bbab344.

[14] Tang X, Dai H, Knight E, Wu F, Li Y, *et al.* A survey of generative AI for de novo drug design: new frontiers in molecule and protein generation. *Brief. Bioinform.* 2024, 25(4):bbae338.

[15] Ibrahim M, Al Khalil Y, Amirrajab S, Sun C, Breeuwer M, *et al.* Generative AI for synthetic data across multiple medical modalities: a systematic review of recent developments and challenges. *Comput. Biol. Med.* 2025, 189:109834.

[16] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, *et al.* Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* 2014, 27:2672–2680.

[17] Nayak AA, Venugopala PS, Ashwini B. A systematic review on generative adversarial network (GAN): challenges and future directions. *Arch. Comput. Methods Eng.* 2024, 31(8):4739–4772.

[18] Sohl-Dickstein J, Weiss E, Maheswaranathan N, Ganguli S. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, July 6–11, 2015, pp. 2256–2265.

[19] Yan Z, Zhou Z, Li X. Survey on generative diffusion model. *Comput. Sci.* 2024, 51(1):273–283.

[20] Kingma DP, Welling M. Auto-encoding variational bayes. *arXiv* 2013, arXiv:1312.6114.

[21] Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. 2018. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (accessed on 28 March 2026).

[22] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, *et al.* Attention is all you need. *Adv. Neural Inf. Process. Syst.* 2017, 30:5998–6008.

[23] Yao Y, Duan J, Xu K, Cai Y, Sun Z, *et al.* A survey on large language model (LLM) security and privacy: the good, the bad, and the ugly. *High-Confid. Comput.* 2024, 4(2):100211.

[24] Pagano S, Holzapfel S, Kappenschneider T, Meyer M, Maderbacher G, *et al.* Arthrosis diagnosis and treatment recommendations in clinical practice: an exploratory investigation with the generative AI model GPT-4. *J. Orthop. Traumatol.* 2023, 24(1):61.

[25] Rao A, Pang M, Kim J, Kamineni M, Lie W, *et al.* Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. *J. Med. Internet Res.* 2023, 25:e48659.

[26] Li Y, Li Z, Zhang K, Dan R, Jiang S, *et al.* ChatDoctor: a medical chat model fine-tuned on a large language model Meta-AI (LLaMA) using medical domain knowledge. *Cureus* 2023, 15(6):e40895.

[27] Hoppe JM, Auer MK, Strüven A, Massberg S, Stremmel C. ChatGPT with GPT-4 outperforms emergency department physicians in diagnostic accuracy: retrospective analysis. *J. Med. Internet Res.* 2024, 26:e56110.

[28] Ren Z, Canas-Bajo T, Ghirardo C, Manassi M, Yu SX, *et al.* Serial dependence in perception across naturalistic generative adversarial network-generated mammogram. *J. Med. Imaging* 2023, 10(4):045501.

[29] Li P, Chen X, Zhang L, Chen J, Ning N. Application progress and thinking of generative artificial intelligence in orthopedics clinical practice. *West China Med. J.* 2025, 40(9):1371–1375.

[30] Zhou J, Zeng Z, Li L. A meta-analysis of Watson for Oncology in clinical application. *Sci. Rep.* 2021, 11(1):5792.

[31] Roosan D, Padua P, Khan R, Khan H, Verzosa C, *et al.* Effectiveness of ChatGPT in clinical pharmacy and the role of artificial intelligence in medication therapy management. *J. Am. Pharm. Assoc.* 2024, 64(2):422–428.e8.

[32] Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, *et al.* Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *J. Am. Med. Inform. Assoc.* 2023, 30(7):1237–1245.

[33] Zaretsky J, Kim JM, Baskharoun S, Zhao Y, Austrian J, *et al.* Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. *JAMA Netw. Open* 2024, 7(3):e240357.

[34] Fraile Navarro D, Coiera E, Hambly TW, Triplett Z, Asif N, *et al.* Expert evaluation of large language models for clinical dialogue summarization. *Sci. Rep.* 2025, 15(1):1195.

[35] Gao X, Diao X, Xu F, Li H, Wu X, *et al.* Exploration of the application of generative artificial intelligence to the challenge of medical record writing. *Chin. Hosp. Manage.* 2025, 45(5):76–79.

[36] Ganzinger M, Kunz N, Fuchs P, Lyu CK, Loos M, *et al.* Automated generation of discharge summaries: leveraging large language models with clinical data. *Sci. Rep.* 2025, 15(1):16466.

[37] Searle T, Ibrahim Z, Teo J, Dobson RJ. Discharge summary hospital course summarisation of in patient electronic health record text with clinical concept guided deep pre-trained transformer models. *J. Biomed. Inform.* 2023, 141:104358.

[38] Iglesias JA, Monterrubio JM, Sesmero MP, Sanchis A. Generation and evaluation of medical images based on diffusion models. In *Proceedings of the 2024 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, Madrid, Spain, May 23–24, 2024, pp. 1–8.

[39] Jung E, Luna M, Park SH. Conditional GAN with 3D discriminator for MRI generation of Alzheimer's disease progression. *Pattern Recogn.* 2023, 133:109061.

[40] Atri H, Shadi M, Sargolzaei M. Generating synthetic medical images with limited data using auxiliary classifier generative adversarial network: a study on thyroid ultrasound images. *J. Ultrasound* 2024, 27(1):105–121.

[41] Biswas T, Suzuki H, Ishikawa M, Kobayashi N, Obi T. Generative adversarial network based digital stain conversion for generating RGB EVG stained image from hyperspectral H&E stained image. *J. Biomed. Opt.* 2023, 28(5):056501.

[42] Pan Y, Ma H, Xia Y, Zhang Y. Application of content generation in medical images. *J. Image Graph.* 2025, 30(6):1985–2000.

[43] Sangeetha S, Mathivanan SK, Karthikeyan P, Rajadurai H, Shivahare BD, *et al.* An enhanced multimodal fusion deep learning neural network for lung cancer classification. *Syst. Soft Comput.* 2024, 6:200068.

[44] Hu Q, Xu H, Liu M, Liu Y, Zhang Q, *et al.* Application of multimodal AI fusion imaging in percutaneous nephrolithotomy for non-hydronephrotic renal calculi. *J. Clin. Radiol.* 2026, 45(1):117–121.

[45] Swanson K, Liu G, Catacutan DB, Arnold A, Zou J, *et al.* Generative AI for designing and validating easily synthesizable and structurally novel antibiotics. *Nat. Mach. Intell.* 2024, 6(3):338–353.

[46] Joo S, Kim MS, Yang J, Park J. Generative model for proposing drug candidates satisfying anticancer properties using a conditional variational autoencoder. *ACS Omega* 2020, 5(30):18642–18650.

[47] Ruksakulpiwat S, Kumar A, Ajibade A. Using ChatGPT in medical research: current status and future directions. *J. Multidiscip. Healthc.* 2023, 16:1513–1520.

[48] Blanco-Gonzalez A, Cabezon A, Seco-Gonzalez A, Conde-Torres D, Antelo-Riveiro P, *et al.* The role of AI in drug discovery: challenges, opportunities, and strategies. *Pharmaceuticals* 2023, 16(6):891.

[49] Ung CY, Correia C, Zhang Z, Caya C, Zhu S, *et al.* Artificial Clinic Intelligence (ACI): a generative AI-powered modeling platform to optimize patient cohort enrichment and clinical trial optimization. *Cancers* 2025, 17(21):3543.

[50] Tan BKJ, Teo CB, Tadeo X, Peng S, Soh HPL, *et al.* Personalised, rational, efficacy-driven cancer drug dosing via an artificial intelligence system (PRECISE): a protocol for the PRECISE CURATE.AI pilot clinical trial. *Front. Digit. Health* 2021, 3:635524.

[51] Feng C, He Z, Ge J, Cao Y, Li C. Intelligent recruitment practice of clinical trial subjects guided by big data and AI. *China Mod. Dr.* 2024, 62(17):84–87.

[52] Nashwan AJ, Hani SB. Transforming cancer clinical trials: the integral role of artificial intelligence in electronic health records for efficient patient recruitment. *Contemp. Clin. Trials Commun.* 2023, 36:101223.

[53] Hassan AE, Ravi S, Desai S, Saei HM, McKennon E, *et al.* An artificial intelligence (AI)-based approach to clinical trial recruitment: the impact of Viz RECRUIT on enrollment in the EMBOLISE trial. *Interv. Neuroradiol.* 2025, 31(6):739–744.

[54] Yigamawano FK, Odom AR, Xue C, Pandey HK, Kolachalama VB. AI-augmented frameworks for enhancing Alzheimer's disease clinical trials: a memory clinic perspective. *J. Prev. Alzheimer's Dis.* 2026, 13(1):100396.

[55] Hughes JP, Rees S, Kalindjian SB, Philpott KL. Principles of early drug discovery. *Br. J. Pharmacol.* 2011, 162(6):1239–1249.

[56] Deore AB, Dhumane JR, Wagh R, Sonawane R. The stages of drug discovery and development process. *Asian J. Pharm. Res. Dev.* 2019, 7(6):62–67.

[57] Roumengas R, Di Lorenzo G, Salhi A, de Buyer P, Chaudhuri A, *et al.* Natural language processing for literature search in vascular surgery: a pilot study testing an artificial intelligence based application. *EJVES Vasc. Forum* 2023, 60:48–52.

[58] Gwon YN, Kim JH, Chung HS, Jung EJ, Chun J, *et al.* The use of generative AI for scientific literature searches for systematic reviews: ChatGPT and Microsoft Bing AI performance evaluation. *JMIR Med. Inform.* 2024, 12:e51187.

[59] Jin Q, Leaman R, Lu Z. PubMed and beyond: biomedical literature search in the age of artificial intelligence. *eBioMedicine* 2024, 100:104988.

[60] Mostafapour M, Fortier JH, Pacheco K, Murray H, Garber G. Evaluating literature reviews conducted by humans *versus* ChatGPT: comparative study. *JMIR AI* 2024, 3:e56537.

[61] Jenko N, Ariyaratne S, Jeys L, Evans S, Iyengar KP, *et al.* An evaluation of AI generated literature reviews in musculoskeletal radiology. *Surgeon* 2024, 22(3):194–197.

[62] Wang L, Cao B, Wang J, Wang C, An J, *et al.* Reform and exploration of biostatistics curriculum driven by generative AI. *Basic Med. Educ.* 2025, 27(10):989–994.

[63] Liu T, Yang F, Yang L, Yang X, Han J, *et al.* A descriptive qualitative study on the experience of using artificial intelligence generated content tools among nursing postgraduate students. *Chin. J. Nurs. Educ.* 2024, 21(9):1046–1051.

[64] Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, *et al.* Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit. Health* 2023, 2(2):e0000198.

[65] Naamati-Schneider L. Enhancing AI competence in health management: students' experiences with ChatGPT as a learning tool. *BMC Med. Educ.* 2024, 24(1):598.

[66] Sridharan K, Sequeira RP. Artificial intelligence and medical education: application in classroom instruction and student assessment using a pharmacology & therapeutics case study. *BMC Med. Educ.* 2024, 24(1):431.

[67] Zhang C. Research on the application of ChatGPT in nursing education. *Ind. Sci. Trib.* 2025, 24(16):207–209.

[68] Tolentino R, Baradaran A, Gore G, Pluye P, Abbasgholizadeh-Rahimi S. Curriculum frameworks and educational programs in AI for medical students, residents, and practicing physicians: scoping review. *JMIR Med. Educ.* 2024, 10:e54793.

[69] Pang X, Zou J, Zhang X, Li Y, Zhang H, *et al.* The impact of artificial intelligence-assisted teaching on medical students' learning outcomes: an integrated model based on the ARCS model and constructivist theory. *BMC Med. Educ.* 2025, 25(1):1309.

[70]Runge I, Hebibi F, Lazarides R. Acceptance of pre-service teachers towards artificial intelligence (AI): the role of AI-related teacher training courses and AI-TPACK within the technology acceptance model. *Educ. Sci.* 2025, 15(2):167.

[71] Zhuang S, Yuan M, Lin X, Chen J, Lin Y, *et al.* Discussion on the application of generative artificial intelligence in case/problem based teaching of pathophysiology. *China Mod. Med.* 2024, 31(34):114–118.

[72] Scherr R, Halaseh FF, Spina A, Andalib S, Rivera R. ChatGPT interactive medical simulations for early clinical education: case study. *JMIR Med. Educ.* 2023, 9:e49877.

[73] Wu F, Wu Y, Wang Y, Wang Q, Wang H, *et al.* Construction and application of a virtual teaching system with generative scenarios for clinical thinking training. *Chin. J. Nurs. Educ.* 2025, 22(9):1029–1035.

[74] Pisla D, Hajjar NA, Rus G, Gherman B, Ciocan A, *et al.* Development of an augmented reality surgical trainer for minimally invasive pancreatic surgery. *Appl. Sci.* 2025, 15(7):3532.

[75] Zhou Y, Sun J, Liu S, Yang Y, Yuan S, *et al.* The applications and challenges of generative artificial intelligence in theoretical and case analysis assessment for resident physician education. *Med. J. Peking Union Med. Coll. Hosp.* 2025, 16(5):1352–1356.

[76] Lu L, Ding F, Xiao S. The influence of generative artificial intelligence technology on the evaluation of medical education and suggestions for countermeasures. *J. Hubei Open Vocat. Coll.* 2025, 38(10):158–160.

[77] Ding N, Song Y, Shan Z, Dong X, Yu M. Exploration of the construction of a medical teaching assistant intelligent question-answering system based on retrieval-augmented generation technology. *China Med. Educ. Technol.* 2025, 39(1):1–5.

[78] Song Y, Li Z, Ding N. Exploration and analysis of the impact of generative artificial intelligence on medical education. *China Med. Educ. Technol.* 2024, 38(3):281–286.

[79] Ali R, Tang OY, Connolly ID, Zadnik Sullivan PL, Shin JH, *et al.* Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. *Neurosurgery* 2023, 93(6):1353–1365.

[80] Luo D, Liu M, Yu R, Liu Y, Jiang W, *et al.* Evaluating the performance of GPT-3.5, GPT-4, and GPT-4o in the Chinese national medical licensing examination. *Sci. Rep.* 2025, 15(1):14119.

[81] Wu YC, Wu YC, Chang YC, Yu CY, Wu CL, *et al.* Advancing medical AI: GPT-4 and GPT-4o surpass GPT-3.5 in Taiwanese medical licensing exams. *PLoS One* 2025, 20(6):e0324841.

[82] Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. *JMIR Med. Educ.* 2023, 9:e48002.

[83] Yanagita Y, Yokokawa D, Uchida S, Tawara J, Ikusaka M. Accuracy of ChatGPT on medical questions in the national medical licensing examination in Japan: evaluation study. *JMIR Form. Res.* 2023, 7:e48023.

[84] Jang D, Yun T, Lee CY, Kwon YK, Kim CE. GPT-4 can pass the Korean national licensing examination for Korean medicine doctors. *PLoS Digit. Health* 2023, 2(12):e0000416.

[85] Jin HK, Kim E. Performance of GPT-3.5 and GPT-4 on the Korean pharmacist licensing examination: comparison study. *JMIR Med. Educ.* 2024, 10:e57451.

[86] Flores-Cohaila JA, García-Vicente A, Vizcarra-Jiménez SF, De la Cruz-Galán JP, Gutiérrez-Arratia JD, *et al.* Performance of ChatGPT on the peruvian national licensing medical examination: cross-sectional study. *JMIR Med. Educ.* 2023, 9:e48039.

[87] Kasai J, Kasai Y, Sakaguchi K, Yamada Y, Radev DR. Evaluating GPT-4 and ChatGPT on Japanese medical licensing examinations. *arXiv* 2023, arXiv:2303.18027.

[88] Lau PL. Rewriting the narrative of AI bias: a data feminist critique of algorithmic inequalities in healthcare. *Law Technol. Hum.* 2025, 7(2):8–26.

[89] Liu S. Gender bias in artificial intelligence: manifestations, generative mechanisms and governance strategies. *J. China Women's Univ.* 2025, 37(6):78–85.

[90] Iloanusi NJ, Chun SA. AI impact on health equity for marginalized, racial, and ethnic minorities. In *Proceedings of the 25th Annual International Conference on Digital Government Research*, Taipei, China, June 11–14, 2024. pp. 841–848.

[91] Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification. *Proc. Mach. Learn. Res.* 2018, 81:77–91.

[92] Ning Y, Teixayavong S, Shang Y, Savulescu J, Nagaraj V, *et al.* Generative artificial intelligence and ethical considerations in health care: a scoping review and ethics checklist. *Lancet Digit. Health* 2024, 6(11):e848–e856.

[93] Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. *JMIR Med. Educ.* 2023, 9(1):e46885.

[94] Guleria A, Krishan K, Sharma V, Kanchan T. ChatGPT: ethical concerns and challenges in academics and research. *J. Infect. Dev. Ctries.* 2023, 17(9):1292–1299.

[95] Doyal AS, Sender D, Nanda M, Serrano RA. ChatGPT and artificial intelligence in medical writing: concerns and ethical considerations. *Cureus* 2023, 15(8):e43292.

[96] Drukker K, Sahiner B, Hu T, Kim GH, Whitney HM, *et al.* MIDRC-MetricTree: a decision tree-based tool for recommending performance metrics in artificial intelligence-assisted medical image analysis. *J. Med. Imaging* 2024, 11(2):024504.