Received: 16 September 2024 | Revised: 9 May 2026 | Accepted: 1 June 2026 | Published online: 16 July 2026

Journal of Data Science and Intelligent Systems
2026, Vol. 00(00) 1–15
DOI: 10.47852/bonviewJDSIS62024354

REVIEW

Generative AI in Healthcare: Applications and Challenges

Wael Rahhal[1,*]

[1] *Department of Data Science, International University of Applied Sciences, Germany*

Abstract: This scoping review synthesizes recent literature on the applications and challenges of generative artificial intelligence (AI) in healthcare, with the aim of providing a structured overview that is useful to clinicians, data scientists, and health-system decision-makers. Using a Preferred Reporting Items for Systematic Reviews and Meta-Analyses-guided search across PubMed, IEEE Xplore, ACM Digital Library, ScienceDirect, Scopus, and Google Scholar, we identified and thematically analyzed 86 studies. The review organizes applications into seven themes—medical image processing, medical imaging and radiology interpretation, clinical documentation, drug design, clinical decision and patient–provider support, synthetic data generation, and medical education and training—and maps the model families most commonly used in each theme. On the basis of this synthesis, generative AI demonstrably accelerates diagnostic workflows, augments scarce clinical datasets, personalizes communication, and supports discovery pipelines. However, the review also identifies four interlocking challenge domains that constrain clinical translation: data and security (privacy, bias, contextualization), technology and infrastructure (scalability, interpretability, vendor dependence), ethics and governance (regulation, trust, certification), and human resources (workforce skills, change management). The paper concludes with a translational path and research priorities aimed at closing these gaps. The review is intended to serve as a reference for researchers, practitioners, and policymakers working on the responsible deployment of generative AI in clinical settings.

Keywords: generative AI, applications, challenges, healthcare, survey

1. Introduction: generative AI

Generative artificial intelligence (AI) is a class of machine learning that learns the underlying statistical structure of training data in order to synthesize new, previously unseen samples that are consistent with that distribution [1, 2]. Because these models can produce images, text, molecules, signals, and structured records that closely resemble real medical data, they have the potential to address a fundamental bottleneck in healthcare artificial intelligence: the scarcity, fragmentation, and privacy sensitivity of high-quality clinical data. A further defining feature of generative models is their ability to perform unsupervised or self-supervised learning, allowing them to acquire useful representations without relying on explicit labels [2, 3]; this is particularly valuable in medical contexts where expert annotation is expensive and slow. In contrast to discriminative models, which concentrate on classification, regression, or decision-making over existing data, generative models focus on synthesizing new entities by learning the joint distribution of the inputs [2, 3].

In the landscape of healthcare AI applications, several key generative models have emerged with distinct capabilities. Generative Adversarial Networks (GANs), built on a competitive architecture between generator and discriminator networks, have found extensive applications in medical image synthesis, image-to-image translation, anomaly detection, and data augmentation tasks in healthcare settings [2].

Variational Autoencoders (VAEs) introduce structured constraints in their encoding process, enabling continuous representations that prove valuable in healthcare through anomaly detection, image denoising, and latent space exploration of medical data. Their ability to generate new data while maintaining structured representations makes them particularly useful in medical imaging applications [2].

Autoregressive models, which predict sequential outputs based on previous data, have demonstrated effectiveness in medical image synthesis, speech and text modeling, and anomaly detection in healthcare contexts. Their sequential prediction capabilities make them particularly suitable for time-series medical data analysis [2].

Energy-based models (EBMs) employ an energy function approach to distinguish data distributions, finding applications in medical image synthesis and restoration, pattern recognition in clinical data, and unsupervised/semi-supervised learning tasks in healthcare settings [2].

Diffusion models have shown exceptional promise in healthcare through high-fidelity image generation, molecular structure generation, and audio synthesis. Their gradual denoising approach has proven particularly effective in generating high-quality medical imaging data [2].

Transformer models, leveraging self-attention mechanisms, have revolutionized clinical text analysis, medical report

***Corresponding author:** Wael Rahhal, Department of Data Science, International University of Applied Sciences, Germany. Email: vael.rahhal@iu-study.org

© The Author(s) 2026. Published by BON VIEW PUBLISHING PTE. LTD. This is an open access article under the CC BY License (https://creativecommons.org/licenses/by/4.0/).

generation, healthcare dialog systems, text summarization, and biomedical text mining. Their ability to handle sequential data has made them invaluable in processing medical documentation and literature [2].

Finally, flow-based models, utilizing the change of variables formula for distribution representation, have found applications in medical image synthesis, anomaly detection, image denoising, and data augmentation. Their unique ability to generate samples while maintaining efficient inference capabilities makes them particularly valuable in medical imaging applications [2]. Table 1 summarizes these principal model families, their core principles, and representative healthcare applications.

Although generative AI has advanced rapidly in the past three years, its translation into routine clinical practice remains uneven. Hospitals face pressing, interrelated problems: radiologists and pathologists are overloaded, clinical documentation consumes a large share of clinician time, drug-discovery pipelines are slow and costly, patient-level data required to train robust models cannot be shared freely because of privacy legislation, and the workforce lacks both the AI literacy and the governance structures needed to deploy these systems safely. Generative models directly target each of these problems by generating synthetic images and text, automating documentation, augmenting limited datasets, and supporting decision-making. At the same time, they introduce new risks that are specific to healthcare, including hallucinated clinical content, bias amplification, opaque decision-making, unclear regulatory status, and dependence on a small number of large model providers.

The problem this review addresses is therefore twofold. First, the literature on generative AI in medicine is scattered across sub-disciplines (imaging, pathology, genomics, clinical natural language processing (NLP), drug design, education, patient communication), and individual studies tend to focus narrowly on a single model or task without situating their contribution within an overall map of applications and limitations. Second, existing reviews often emphasize either the technical capabilities or the ethical concerns, rather than providing an integrated view that is useful to clinicians, data scientists, and health-system decision-makers simultaneously.

Accordingly, the present review has three objectives: (i) to identify and thematically organize the main applications of generative AI currently reported in the healthcare literature; (ii) to synthesize the principal technical, ethical, regulatory, and organizational challenges that constrain its translation into clinical practice; and (iii) to outline a translational path and a set of research directions that can help bridge the gap between model development and safe, equitable deployment.

The originality of this work lies in its integrative, multimodality perspective. In contrast to earlier surveys that concentrate on a single model family (e.g., GANs or large language models [LLMs]) or a single application area (e.g., medical imaging), this review synthesizes evidence across seven thematic application areas and four thematic challenge domains and then links them to a concrete translational pipeline. The contribution of the paper is therefore threefold: it provides a structured taxonomy of generative AI use cases in healthcare that practitioners can map onto their own workflows; it articulates the corresponding risk surface in a form that supports governance and procurement decisions; and it identifies evidence gaps that future empirical research should prioritize. The remainder of the paper is organized as follows. Section 2 states the research question. Section 3 describes the scoping review methodology and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)-guided paper selection process. Section 4 presents the thematic applications. Section 5 summarizes the translational path. Section 6 analyzes the challenges. Sections 7–9 present the conclusion, potential biases, and recommendations for future research.

2. Research Question

Our primary research question is: "What key themes emerge from the recent literature discussing the applications and challenges of generative AI in healthcare?" The aim is to

Table 1
Principal families of generative AI models and their reported healthcare applications

Model family	Key principle	Representative healthcare applications
Generative Adversarial Networks (GANs)	Adversarial game between a generator and a discriminator	Medical image synthesis, image-to-image translation, anomaly detection, data augmentation
Variational Autoencoders (VAEs)	Probabilistic encoder–decoder with a structured latent space	Anomaly detection, image denoising, latent space exploration of medical data
Autoregressive models	Sequential prediction conditioned on prior tokens or pixels	Medical image synthesis, clinical text and speech modeling, time-series anomaly detection
Energy-based models (EBMs)	Learn an energy function that characterizes the data distribution	Medical image restoration, pattern recognition in clinical data, semi-supervised learning
Diffusion models	Iterative denoising from a simple prior to the data distribution	High-fidelity image generation, molecular structure generation, audio synthesis
Transformer models	Self-attention over sequences	Clinical text analysis, radiology report generation, dialog systems, biomedical text mining
Flow-based models	Invertible transformations with tractable likelihoods	Medical image synthesis, anomaly detection, image denoising, data augmentation

identify themes within recent literature related to the potential applications and challenges of generative AI in healthcare, with the goal of guiding future research and informing clinical translation.

3. Methodology

To comprehensively explore the applications and challenges of generative AI in healthcare and to answer the research question, we conducted a scoping literature review guided by the PRISMA 2020 statement. The PRISMA guidance was adapted to the objectives of a scoping (rather than a strictly systematic) review, because the goal was to map thematic breadth rather than to pool quantitative effect sizes. The search strategy was implemented using Boolean logic across the following databases: PubMed, IEEE Xplore, ACM Digital Library, ScienceDirect, Scopus, and Google Scholar. The search string combined four concept groups: ("generative artificial intelligence" OR "generative AI" OR "diffusion model" OR "variational autoencoder" OR "generative adversarial network") AND ("large language model" OR "LLM" OR "transformer") AND ("healthcare" OR "health care" OR "medicine" OR "clinical") AND ("applications" OR "challenges" OR "opportunities" OR "limitations").

Inclusion criteria were (i) peer-reviewed journal articles, conference papers, or recognized preprints describing an application of generative AI in a healthcare or biomedical setting; (ii) publications in English; (iii) publications between January 2017 and March 2026 to capture the relevant technical lineage (from the introduction of the Transformer architecture onward), while prioritizing the 2021–2026 window for clinical applications; and (iv) papers that reported either a concrete application, an empirical evaluation, or a substantive discussion of challenges. Exclusion criteria were (i) editorials, news items, and nontechnical commentary without original analysis, (ii) duplicate records, (iii) publications not related to human or biomedical data, and (iv) studies in which generative modeling was not the primary methodological contribution.

The selection process followed the four PRISMA phases of identification, screening, eligibility, and inclusion. During identification, 1248 records were retrieved across the databases listed above. After removing 266 duplicates, 982 unique records were screened by title and abstract; 703 were excluded because they did not meet the inclusion criteria at that level (e.g., non-healthcare applications or non-generative models). The remaining 279 full-text articles were assessed for eligibility; 193 were excluded because they lacked a clear generative AI contribution, were duplicate methodological reports, or focused exclusively on non-clinical benchmarking. Eighty-six studies were finally included in the qualitative synthesis. The complete selection process is summarized in Table 2, which presents the PRISMA-guided flow of records from identification to inclusion.

Data extraction was carried out using a standardized template capturing study identifier, year of publication, generative-model family (GAN, VAE, diffusion, autoregressive, energy-based, transformer, flow-based), clinical domain (imaging, documentation, drug design, education, patient communication, synthetic data, decision support), dataset and evaluation metrics, and reported limitations. Thematic synthesis was then used to group studies along the two high-level axes of applications (Section 4) and challenges (Section 6). To strengthen the trustworthiness of the synthesis, extracted data and thematic assignments were reviewed iteratively, and discrepancies were resolved by re-reading the original source before final categorization.

Table 2
PRISMA 2020 flow summary of the study selection process. Identification: 1248 records from PubMed, IEEE Xplore, ACM Digital Library, ScienceDirect, Scopus, and Google Scholar; 266 duplicates removed. Screening: 982 records screened by title/abstract; 703 excluded. Eligibility: 279 full-text articles assessed; 193 excluded. Included: 86 studies in qualitative synthesis

Phase	Description	*n*
Identification	Records identified from PubMed, IEEE Xplore, ACM Digital Library, ScienceDirect, Scopus, and Google Scholar	1248
Identification	Duplicate records removed before screening	266
Screening	Records screened by title and abstract	982
Screening	Records excluded (not healthcare/not generative AI/non-English/editorial)	703
Eligibility	Full-text articles assessed for eligibility	279
Eligibility	Full-text articles excluded (no clear generative AI contribution; duplicate methodological reports; purely benchmarking)	193
Included	Studies included in the qualitative synthesis	86

The eligibility rules supporting this PRISMA-guided selection process are summarized in Table 3; the table distinguishes the accepted study types, language and timeframe limits, topic focus, methodological reporting requirements, and ethics-related exclusions.

4. Applications of Generative AI in Healthcare

4.1. Image processing

4.1.1. Image reconstruction

High-fidelity medical image reconstruction is essential for accurate clinical decision-making, while simultaneously minimizing patient exposure and imaging cost [4]. Two of the most widely used modalities are:

1) Computed tomography (CT), which reconstructs cross-sectional anatomical representations from multiple X-ray projections.
2) Magneticresonance imaging (MRI), which leverages strong magnetic fields and radiofrequency pulses to obtain high soft-tissue contrast.

Table 3
Inclusion and exclusion criteria applied during study selection

Criterion	Inclusion	Exclusion
Publication type	Peer-reviewed journal articles, conference papers, recognized preprints	Editorials, news items, nontechnical commentary
Language	English	Non-English publications without available translation
Timeframe	January 2017–March 2026 (emphasis on 2021–2026)	Pre-2017 work except for seminal methodological references
Topic focus	Application or evaluation of generative AI in a healthcare or biomedical setting	Non-healthcare domains; studies in which generative modeling is not the primary contribution
Methodological reporting	Clear description of dataset, model, and evaluation	Studies with insufficient methodological detail or unreproducible results
Ethics	Human studies approved by an ethics committee or using publicly released datasets	Studies lacking ethical approval where required

Both modalities are, however, constrained by trade-offs between acquisition time, radiation dose (in CT), and patient comfort, which has motivated research into accelerated acquisition strategies [5].

Accelerated imaging protocols reduce the quantity of raw data acquired, which shortens examination duration and dose but typically degrades image quality. The principal manifestations of this degradation are a reduced signal-to-noise ratio (SNR) and a reduced contrast-to-noise ratio (CNR), resulting in images that are noisier and less diagnostically reliable.

Specifically, these quality limitations are characterized by:

1) SNR, which quantifies the clarity of diagnostic information relative to background noise.
2) CNR, which quantifies the separability of adjacent tissues within the reconstructed image.

Consequently, advanced reconstruction algorithms are required to recover diagnostic quality from limited measurements [6]. Diffusion models have recently emerged as a promising class of deep generative approaches that can restore image fidelity in under-sampled or low-dose regimes.

In the field of medical image reconstruction, two notable studies have made significant advances. Research by Özbey et al. [7] introduced AdaDiff, a novel approach for MRI reconstruction that demonstrated superior outcomes compared to existing methods. Similarly, study by Xie and Li [8] presented MC-DDPM (Monte Carlo Denoising Diffusion Probabilistic Models) as a promising technique specifically designed for reconstructing under-sampled medical images, offering new possibilities in medical imaging processing where complete data sampling is not feasible.

These developments represent important steps forward in improving the quality and reliability of medical image reconstruction, particularly in scenarios where traditional imaging methods face limitations or constraints.

4.1.2. Image-to-image translation

In recent years, generative models have been applied to the problem of converting one medical imaging modality into another—a task commonly referred to as image-to-image translation. The foundational work of Isola et al. [9] on pix2pix and Zhu et al. [10] on CycleGAN established the feasibility of learning cross-modality mappings from paired and unpaired data, respectively, and is particularly relevant to clinical practice, where different modalities convey complementary diagnostic information but are not always simultaneously available. Diffusion models have since extended this line of research by enabling the synthesis of missing modalities from those already acquired. A representative application is MRI-to-CT translation [11]: CT is rapid and widely available but provides limited soft-tissue contrast, whereas MRI offers superior soft-tissue detail at the cost of acquisition time and expense. Acquiring both sequentially is resource intensive and introduces registration errors. Using a Denoising Diffusion Probabilistic Model (DDPM), Graf et al. [11] generated anatomically consistent CT volumes from MRI inputs with higher fidelity than earlier convolutional neural network (CNN)- and GAN-based methods, thereby reducing the need for additional scans. In parallel, Özbey et al. [7] proposed SynDiff, an adversarial diffusion model for medical image translation that outperformed competing approaches in terms of accuracy, inference speed, and perceptual image quality across multiple datasets, demonstrating robustness to variations in acquisition protocol.

4.1.3. Image generation

Diffusion models have rapidly become a dominant approach for generating medical images across diverse modalities and dimensionalities [12]. Reported use cases span synthetic 2D and 3D imagery, including X-rays, MRI, and CT [13–16]. One notable example is SHAPR [17], an autoencoder-based framework that infers three-dimensional cell shape from a single 2D microscopy image. Because clinical microscopy is predominantly 2D, yet many biological questions are inherently three-dimensional, SHAPR fills a practical gap: it improved cell-type classification for erythrocytes, transferred successfully to stem cells, and captured nuclear geometry, all while reducing acquisition time and storage requirements. More recent work has pushed these ideas into higher-fidelity regimes. Müller-Franzes et al. [18] introduced Medfusion, a latent diffusion model that generated ophthalmologic, chest-CT, and 3D brain MRI images of higher perceptual quality than comparable GAN baselines, with improved training stability and interpretability. Pan et al. [19] proposed a transformer-based DDPM capable of synthesizing 2D chest radiographs, cardiac MRI, and abdominal CT, achieving favorable Inception Score and Fréchet Inception Distance metrics. Collectively, these studies suggest that diffusion-based generators are well positioned to support downstream tasks such as dataset augmentation, rare-disease modeling, and medical training simulation.

4.1.4. Image classification

Medical image classification is a central task in clinical AI because it enables the automated detection and stratification of disease patterns from large volumes of imaging data, thereby accelerating decision-making and reducing inter-observer variability [12, 20]. A persistent difficulty in this domain is class imbalance, particularly in pathology workflows in which rare cell types or lesions are clinically important but statistically underrepresented. Oh and Jeong [21] addressed this issue with DiffMix, a diffusion-based data-synthesis framework that improves nuclei segmentation and classification on imbalanced pathology datasets, outperforming prior methods such as HoVer-Net and SONNET and yielding more equitable performance across cell classes. Yang et al. [22] subsequently introduced DiffMIC, a dual-guidance diffusion network that combines denoising dynamics with class- and image-level conditioning. DiffMIC achieved consistent gains on several benchmark datasets and was particularly effective in handling noisy or imbalanced inputs, attributable to its ability to attend to salient discriminative features while preserving structural coherence. More recent transformer-based diffusion pipelines have also been explored to classify arrhythmias from electrocardiographic images using deformable attention [23] and to screen for microvascular changes in systemic sclerosis from capillaroscopy images [24], further illustrating the breadth of generative-classification approaches across imaging modalities.

4.1.5. Image registration

Deformable image registration aligns two scans of the same anatomical region that have been acquired at different times, in different positions, or with different scanners and is a precondition for longitudinal comparison, image-guided intervention, and atlas construction [12]. Kim et al. [25] proposed DiffuseMorph, a diffusion-based registration framework that jointly learns the appearance transformation and the spatial deformation field, using the diffusion process to regularize intensity variation while a deformation module resolves the geometric mismatch. Because the two components are trained end-to-end from paired images, the model generalizes across modality pairs and produces smooth, anatomically plausible warps. This makes DiffuseMorph directly relevant to clinical tasks such as surgical planning, longitudinal treatment monitoring, and multimodal image fusion.

4.1.6. Image segmentation

Image segmentation underpins many downstream medical imaging tasks because it isolates anatomical or pathological structures within a scan [12]. Supervised deep-learning segmentation models typically require large corpora of pixel-accurate annotations, which are difficult to obtain in healthcare because expert labeling is costly, slow, and demands specialist knowledge [26]. Diffusion models have been explored as a means of reducing this annotation burden by generating realistic labeled examples. Amit et al. [26] demonstrated that DDPMs can synthesize annotated medical images and thereby accelerate downstream segmentation. Zhang et al. [27] subsequently introduced DiffuSeg, a domain-driven diffusion framework that segments vascular structures under limited supervision, supporting the detection of aneurysms and occlusions while minimizing manual effort. Taken together, these studies indicate that diffusion-based synthesis can make segmentation workflows more annotation-efficient without compromising accuracy.

4.2. Medical imaging

4.2.1. Medical imaging and radiology interpretation

Transformer-based architectures have been adopted across a wide spectrum of radiology and medical imaging tasks, ranging from report summarization and disease classification to error detection, multimodal fusion, and automated report generation [12]. In the area of radiology summarization, Balouch and Hussain [28] reported that Biobart-V2 achieved state-of-the-art ROUGE scores while preserving clinically salient content. For classification from chest radiographs, Bhattacharya et al. [29] proposed RadioTransformer, which models radiologists' visual search through a cascaded global–focal attention mechanism and a dedicated visual-attention loss, improving diagnostic accuracy over baselines. Chaudhari et al. [30] fine-tuned a domain-specific Radiology BERT to detect speech-recognition errors in dictated reports, demonstrating the value of contextual adaptation in clinical NLP. Jacenkow et al. [31] combined a BERT text encoder with CNN-based image features for multimodal disease classification, outperforming single-modality baselines. Li et al. [32] introduced an instance-dependent prompt-tuning framework on top of BERT to classify actionable findings in tinnitus reports. For structured information extraction, Azizi et al. [33] developed a transformer named entity recognition (NER) pipeline that produced coherent, structured outputs superior to prior methods. Moezzi et al. [34] introduced TrMRG for medical report generation, capturing long-range dependencies between image features and descriptive text; Mohsan et al. [35] combined a vision transformer encoder with a language-model decoder for radiology report generation and reported competitive results on chest-radiograph benchmarks; while Nimalsiri et al. [36] proposed MERGIS, which couples segmentation with a transformer encoder–decoder to produce comprehensive radiology reports. Taken together, these studies illustrate how transformer-based models now permeate the radiology workflow and contribute to the accuracy, efficiency, and interpretability of downstream clinical documentation.

4.2.2. Diagnostic assistance

Transformer-based models have also demonstrated strong performance across a range of diagnostic tasks [12]. Azizi et al. [33] combined a transformer encoder with CNN features to identify neurological signs from clinical narrative, reporting gains in both sensitivity and specificity over rule-based baselines. Chen et al. [37] introduced FIT-Net, a feature interaction transformer designed for the diagnosis of pathologic myopia from optical coherence tomography, which delivered more stable and accurate predictions than comparable CNN models. Chen et al. [38] applied multi-view vision transformers to mammogram analysis, capturing long-range correlations across breast views and handling input variability consistently. In neuroimaging, Alp et al. [39] combined vision transformers with sequential modeling for Alzheimer's disease classification from 3D brain MRI, exceeding 99% accuracy on the ADNI dataset and outperforming CNN and recurrent neural network (RNN) baselines. Hosain et al. [40] used vision transformers with transfer learning for wireless capsule endoscopy, improving over DenseNet201 on gastrointestinal abnormality detection. Hu et al. [41] proposed a dual transformer hashing framework that aligns histopathology whole-slide images with their diagnostic reports. Mo et al. [42] developed HoVer-Transformer for ROI-free breast ultrasound diagnosis, matching expert radiologists while eliminating the need for manual region annotation. Finally, Zhou et al. [43] introduced a unified multimodal transformer for COVID-19 that integrated imaging, clinical, and laboratory inputs and outperformed single-modality baselines in outcome prediction. These results collectively indicate that transformer encoders, particularly their multimodal variants, have become central tools for decision support across radiology, pathology, ophthalmology, and cardiology.

4.3. Clinical documentation

4.3.1. Clinical documentation and information extraction

Clinical documents are a rich but structurally challenging source of information: they combine semi-structured layout, domain-specific vocabulary, and institution-specific conventions. Xu et al. [44] developed a layout-aware pipeline that parses PDF records, classifies each line into a functional role (body, margin, footer), and then retains only clinically relevant content, feeding the resulting segments into downstream NLP. LayoutLMv2, also reported in Reference [44], jointly models text, layout, and image features and consequently improves extraction accuracy from scanned clinical PDFs. For German clinical notes, Lentzen et al. [45] pre-trained BioGottBERT, which outperformed the general-domain GottBERT on named entity recognition and underscored the benefits of in-domain pre-training. Li et al. [46] proposed ClinicalLongformer and Clinical-BigBird to handle very long clinical texts, with consistent gains on question answering, entity recognition, and document classification. Moon et al. [47] demonstrated that BERT and Sentence-BERT can support entity extraction and clustering of diagnoses, tests, and treatments across heterogeneous clinical corpora. Oh et al. [48] compared XLNet, RoBERTa, and BERT for de-identifying protected health information, finding that XLNet's bidirectional-autoregressive formulation offered the best performance. Searle et al. [49] applied BART to oncology-focused hospital-course summarization, producing problem-list-oriented summaries suitable for clinical use. Sivarajkumar and Wang [50] proposed HealthPrompt, a prompt-based framework that outperformed supervised baselines in low-resource clinical NLP. Solarte-Pabón et al. [51] showed that specialized Spanish breast-cancer corpora yielded substantial gains over generic pre-trained models. Finally, Wei et al. [52] introduced ClinicalLayoutLM for classifying scanned clinical documents across sixteen categories, leveraging both text and layout to outperform optical character recognition (OCR)-only pipelines.

4.3.2. Medical coding and billing

Medical coding and billing assign standardized codes, most commonly ICD-9 or ICD-10, to the diagnoses and procedures documented in clinical notes. Transformer-based models have been applied to make this process more accurate, efficient, and explainable. Liu et al. [53] proposed the Hierarchical Label-Wise Attention Transformer, evaluated on MIMIC-III, which uses per-label attention to improve prediction on the 50 most frequent ICD-9 codes. López-García et al. [54] combined multilingual transformers (XLM-RoBERTa, mBERT, BETO) with a two-step pipeline that first recognizes medical entities and then normalizes them to standard codes, which outperformed joint learning of the two steps and underscored the value of task separation in clinical coding. Coutinho and Martins [55] designed a transformer model tailored to Portuguese death-certificate ICD-10 assignment that incorporated hierarchical attention and document structure, with strong generalization to adjacent languages and settings.

Falissard et al. [56] presented a deep-learning pipeline for automatic ICD-10 coding of unstructured French medical text, combining NLP with multi-label classification and reporting strong accuracy and F1 scores. Collectively, these studies show that hierarchical and multilingual transformer architectures can bring both performance and transparency to automated medical coding.

4.4. Drug design

4.4.1. Protein structure prediction

Protein structure prediction is central to molecular biology and therapeutic development, and transformer-based protein language models have driven much of the recent progress [12]. Behjati et al. [57] introduced ProtAlbert, whose attention maps recover structural motifs and binding sites from sequence alone. Xiang et al. [58] proposed Functional Annotation of Proteins using Multimodal models (FAPM), a multimodal model that couples protein sequences with an LLM to generate natural-language functional annotations, supporting more explainable predictions for under-characterized proteins. Boadu et al. [59] combined sequences and structures through transformers and equivariant graph neural networks in TransFun, outperforming single-modality baselines. Geffen et al. [60] distilled ProtBert into DistilProtBert, preserving classification performance while reducing computation, which matters for large-scale drug-discovery pipelines. Castro et al. [61] introduced ReLSO to optimize protein sequences under functional and structural constraints. Ferruz et al. [62] proposed ProtGPT2 to generate de novo sequences with native-like biochemical properties, with potential applications in protein engineering. Brandes et al. [63] released ProteinBERT, which reported strong results on structure prediction, function annotation, and protein–protein interaction prediction by capturing both local and global sequence patterns. Oliveira et al. [64] introduced TEMPROT and TEMPROT+, which improved the extraction and analysis of protein sequence embeddings over earlier methods.

4.4.2. Molecular representation and drug design

Transformer-based approaches also support molecular design and property prediction [12]. Bagal et al. [65] proposed MolGPT, a generative model that produces molecules matching user-specified scaffolds and property targets, which is useful for lead optimization. Li et al. [66] introduced KPGT, a knowledge-guided pre-trained graph transformer for molecular property prediction, which outperformed self-supervised baselines on standard benchmarks. These methods illustrate how transformer inductive biases can be adapted from text to graph-structured molecular data, supporting drug-discovery workflows.

4.5. Clinical support

4.5.1. Clinical decision support

Transformer-based models are increasingly used in clinical decision support because they can combine unstructured text, structured data, and temporal signals within a single architecture [12]. When these capabilities are embedded in a clinical workflow, generative AI can surface patient-specific recommendations that help clinicians reach better-informed decisions and potentially reduce avoidable errors [67]. LLMs further extend this capability by digesting broad bodies of medical literature and longitudinal records to produce evidence-based suggestions for diagnosis, treatment selection, and prognosis [68]. Recent empirical work illustrates the range of applications. Hu and Wang [69] benchmarked BERT, BioBERT, RoBERTa, and DistilBERT on glaucoma-progression prediction, concluding that domain-adapted transformers are well suited to ophthalmology decision support. Huang [70] proposed the Surgical Action (SA) model, an encoder–decoder architecture that recognizes and anticipates surgical actions; although forecasting performance was modest, real-time action classification proved valuable for intra-operative decision support. Wang et al. [71] developed a hybrid rule-based/transformer system for extracting family-history information from clinical text, with recall comparable to established baselines. These studies suggest that transformer-based decision support is maturing from single-task demonstrators to workflow-embedded tools, although rigorous prospective evaluation remains limited.

4.5.2. Patient–provider communication

LLMs are becoming a practical tool for mediating communication between patients and healthcare providers, and there is emerging evidence that they improve patient engagement and satisfaction [68, 72]. ChatGPT, based on the GPT family, is the most widely studied representative: it is pre-trained on broad text corpora and can be fine-tuned or prompted for healthcare-specific tasks. Moshirfar et al. [73] evaluated ChatGPT in ophthalmology and reported that GPT-4 outperformed GPT-3.5 and human experts across most StatPearls categories, with the exception of "Lens and Cataract." Singhal et al. [74] introduced Med-PaLM, a domain-adapted LLM that achieved promising performance on medical question answering. Luo et al. [75] released BioGPT, a biomedical pre-trained generator that outperformed prior models across six biomedical NLP tasks, including relation extraction and biomedical question answering. These results suggest that, with appropriate domain adaptation and guardrails, LLMs can plausibly support clinician–patient communication, patient education, and triage tasks, although further work is required to measure their clinical utility in real-world deployments.

Collectively, the studies above demonstrate the breadth and maturity of diffusion-based and transformer-based generative models across medical tasks—from image synthesis to clinical decision support—and motivate the thematic synthesis presented in Table 4.

4.6. Data generation

4.6.1. Synthetic data generation and data augmentation

Synthetic data produced by generative AI provides a practical way to balance data access with patient privacy. Trained on de-identified corpora, generative models can synthesize realistic yet non-identifying patient records that support research, benchmarking, and education [2, 76, 77]. GANs have been used to create electronic health record (EHR) data by learning the underlying feature distributions, which mitigates some privacy concerns while maintaining downstream analytic utility. Synthetic data also alleviates class imbalance and rare-event under-representation, which are persistent obstacles for machine learning models trained on clinical data. Because researchers can sample synthetic cohorts along specific axes—for example, age or disease severity—the approach supports hypothesis exploration and stress-testing of clinical models [2, 76, 77]. Responsible use, however, requires careful validation against real cohorts and explicit disclosure of the synthetic origin of the data.

4.7. Education and training

4.7.1. Medical education and training

In medical education and training, generative AI can produce a diverse range of virtual patient cases spanning different conditions, demographics, and clinical scenarios, providing a scalable learning environment for students and practicing clinicians [2, 78, 79].

4.7.2. Patient education

Generative AI can also personalize patient education [80]. Models can tailor content to a patient's diagnosis, language, literacy level, and concerns—for example, producing individualized guidance on blood-glucose management, diet, physical activity, and medication adherence for a patient with type-2 diabetes. Interactive dialog interfaces allow patients to ask follow-up questions and receive grounded, conversational responses [2], which can support understanding and self-management.

Table 4
Thematic applications of generative AI in healthcare identified in this review

Application theme	Representative tasks	Primary model families
Image processing	Reconstruction, modality translation, generation, classification, registration, segmentation	Diffusion, GANs, Transformers
Medical imaging and radiology	Report generation, disease classification, error detection, multimodal fusion	Transformers (BERT, GPT, ViT), multimodal models
Clinical documentation	Entity recognition, summarization, coding and billing, layout-aware document analysis	BERT variants, BART, Longformer
Drug design	Protein structure and function prediction, molecular generation, property prediction	Protein language models, graph transformers, autoregressive models
Clinical support	Decision support, surgical-action recognition, family-history extraction, triage Patient–provider communication	BioBERT, transformer encoders, LLMs GPT family, Med-PaLM, BioGPT
Synthetic data generation	EHR synthesis, data augmentation, privacy-preserving datasets	GANs, VAEs, diffusion
Education and training	Virtual patient cases, personalized educational content	LLMs, multimodal generative models

5. Generative AI Translational Path

The integration of generative AI into healthcare workflows, known as the translational path, has the potential to reshape care delivery but is intricate and requires careful planning, execution, and governance [80, 81]. Reddy [2] proposed a staged framework for this translational path, reproduced in Figure 1 [2]. The framework articulates the sequence of activities that moves a generative model from algorithmic development to responsible clinical deployment, including problem framing, data curation, model development and validation, implementation science, governance, post-deployment monitoring, and continuous learning.

6. Challenges in the Application of Generative AI in Healthcare

The application of generative AI in healthcare presents many challenges as follows.

6.1. Data and security

6.1.1. Data privacy, security, and robustness

Integrating generative AI into healthcare elevates a longstanding tension between analytic utility and patient privacy. Because these models are trained on large volumes of sensitive data, any breach has disproportionate downstream consequences for individuals and for public trust in the health system [82–84]. LLMs exacerbate the risk profile because they can memorize rare training examples and regurgitate them under carefully crafted prompts [68]. Foundational safeguards—end-to-end encryption, fine-grained access control, authenticated communication channels, and auditable logging—should be complemented by techniques developed specifically for machine learning, including differential privacy, federated learning, and de-identification prior to training [85]. Regulatory frameworks are also required to prevent misuse and to clarify accountability in the event of model-mediated harm [86]. Finally, deployed systems must be robust against adversarial inputs and data-poisoning attacks, because even a modest perturbation to clinical imagery or documentation can produce clinically consequential errors [82–84].

6.1.2. Geopolitical influences on data governance

Healthcare data is increasingly international in scope, and its governance involves governments, multilateral organizations, and private data holders whose interests are not always aligned. Deploying generative AI across borders therefore requires navigating a patchwork of regulatory frameworks and geopolitical considerations, which can materially constrain both data access and model portability [82].

6.1.3. Contextualization problem

Generative AI systems must incorporate relevant clinical context—patient history, local protocols, medication formularies, and prevailing guidelines—to produce outputs that are both safe and useful. Inadequate contextualization leads to plausible-sounding but clinically inappropriate recommendations [83]. Techniques such as retrieval-augmented generation and fine-tuning on institution-specific corpora partially address this limitation but introduce their own data governance and maintenance obligations.

6.1.4. Data quality and bias

The quality, scope, and representativeness of the training data dictate the accuracy and fairness of generative outputs. When these data under-represent particular populations, comorbidities, or care settings, the resulting models can amplify existing health inequities rather than alleviate them [83]. Systematic auditing of training cohorts, demographic performance reporting, and bias-aware evaluation benchmarks are therefore essential before clinical deployment.

6.2. Technology and infrastructure

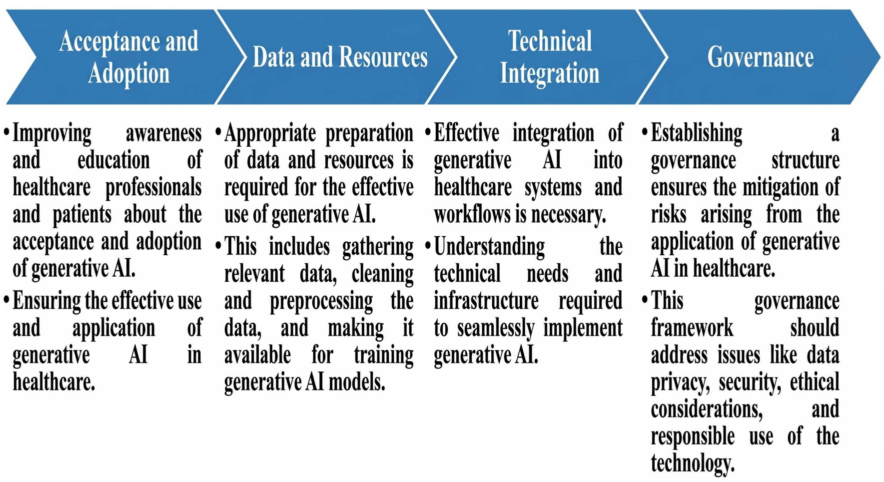
6.2.1. Dependency on specific technology platforms

Contemporary generative AI systems—most visibly GPT-4—are developed and operated by a small number of large technology companies. This creates the familiar risks of vendor lock-in, limited interoperability, and outsized influence over clinical workflows, and it motivates interest in open-weight alternatives, federated deployment, and contractual safeguards for health-system customers [82].

6.2.2. Maintenance of generative AI systems

Sustaining generative AI systems in healthcare is a substantial engineering challenge. Because both clinical practice and underlying data distributions evolve, models require periodic retraining, revalidation, and fine-tuning to remain safe and effective. This lifecycle demands dedicated MLOps infrastructure,

Figure 1
Generative AI translational path in healthcare



monitoring, and budget, which can be prohibitive for smaller healthcare organizations [82].

6.2.3. Scalability and feasibility of deployment

Scaling generative AI across a healthcare enterprise is not primarily a modeling problem; it is an integration problem. Production systems must interoperate with legacy electronic health records, clinical information systems, and regulated messaging standards such as HL7 and FHIR, all while respecting clinical latency budgets and audit requirements [82, 83]. LLMs add their own constraints: inference cost, hardware availability, and the organizational capacity to maintain prompt templates and retrieval pipelines over time [68, 83]. Practical deployments therefore benefit from modular architectures in which generative components are isolated behind service boundaries that allow targeted monitoring, versioning, and rollback.

6.2.4. Hardware requirements

Large language and diffusion models consume substantial computational resources, with floating-point operations and GPU-hour requirements that are orders of magnitude greater than those of earlier deep-learning architectures [86]. These demands translate into procurement, energy, and sustainability obligations that health systems must weigh against the expected clinical benefit.

6.2.5. Complexity and interpretability

Enterprise generative AI stacks are often complex and difficult to interpret, which constrains the explanations that clinicians and auditors can offer patients. Interpretability is not optional in medicine: it underpins informed consent, accountability, and post-deployment learning. Transparent model cards, saliency methods, and structured explanations are therefore prerequisites for clinical adoption [82, 86].

6.2.6. Data availability

Healthcare organizations routinely struggle to assemble the large, well-labeled datasets required to train robust generative models, particularly in sub-specialties where the case mix is narrow and annotation expertise is concentrated in a handful of experts [82].

6.3. Ethics and governance

6.3.1. Regulatory and ethical compliance

LLMs in clinical settings must satisfy stringent regulatory requirements, including obligations relating to data protection, patient consent, and transparency about the use of AI in decision-making. Existing frameworks such as the EU AI Act, the US FDA's guidance on software as a medical device, and national data protection laws impose overlapping but nonidentical constraints that model providers must navigate carefully [68].

6.3.2. Trust and reliability

Trust in clinical AI depends on the reliability and interpretability of its outputs, particularly in high-stakes scenarios where errors are consequential. The "black-box" character of many generative models therefore remains an obstacle to adoption and motivates ongoing work on calibrated uncertainty estimates, counterfactual explanations, and prospective audit trails [83, 84].

6.3.3. Clinical evaluation, regulation, and certification

The rapid release cadence of contemporary LLMs and generative AI systems creates a structural tension with clinical evaluation and regulation. By the time a model has been evaluated through a formal clinical pathway, newer versions are often already in use, which complicates the definition of the "locked" artifact that regulators are accustomed to certifying [84]. Regulatory agencies are adapting by articulating mechanisms for continuously learning systems, predetermined change-control plans, and post-market surveillance, but practical implementation remains in flux.

6.4. Human resources

6.4.1. Talent and skill gaps

Deploying enterprise-grade generative AI requires a combination of clinical, data-science, MLOps, and governance expertise that is scarce in most healthcare organizations. Attracting and retaining this talent is a significant constraint and one that is unlikely to be relieved without sustained investment in cross-disciplinary training and career paths [82, 83].

6.4.2. Change management

Adopting generative AI disrupts established clinical workflows and role boundaries. Successful implementation therefore depends on structured change-management practices, including stakeholder engagement, iterative pilots, training and support, and clear escalation paths when model outputs are ambiguous or incorrect [82].

Table 5 consolidates these four challenge domains and links each set of risks to illustrative mitigation strategies, showing how the detailed discussion above translates into practical governance priorities.

Table 5
Principal challenges to the deployment of generative AI in healthcare, organized by thematic domain

Domain	Key challenges	Illustrative mitigations
Data and security	Data privacy, security breaches, adversarial robustness, geopolitical data governance, contextualization, data quality and bias	Encryption, differential privacy, federated learning, bias audits, representative datasets
Technology and infrastructure	Vendor lock-in, model maintenance, scalability, hardware and GPU demand, interpretability, data availability	Open-weight alternatives, MLOps pipelines, model compression, explainability tooling

(*Continued*)

Table 5
(*Continued*)

Domain	Key challenges	Illustrative mitigations
Ethics and governance	Regulatory and ethical compliance, trust and reliability, clinical evaluation and certification	Transparent reporting (TRIPOD-AI), regulatory sandboxes, continuous post-market surveillance
Human resources	Talent and skill gaps, change management, and clinician adoption	Cross-disciplinary training, clinician-in-the-loop workflows, organizational change management

7. Conclusion

Generative AI shows significant potential in healthcare, with the most mature evidence base currently concentrated in medical imaging and disease diagnostics. The thematic distribution of applications identified in this review is summarized in Figure 2, while the four domains of challenges discussed in Section 6 are consolidated in Figure 3. Across these themes, the technology faces persistent obstacles relating to data privacy, security, infrastructure demands, the interpretability of complex architectures, and the regulatory pathways required for clinical certification. The present emphasis on CT and MRI modalities may underserve ultrasound-based workflows, where generative AI could materially improve diagnostic consistency and access in resource-constrained settings; expanding research to include a wider range of imaging modalities is therefore a priority [12]. Future work should compare generative AI models with alternative machine learning families under identical clinical evaluation protocols, report performance on heterogeneous patient populations to mitigate demographic bias, and couple technical benchmarking with prospective clinical validation and post-deployment monitoring. Addressing these gaps will be essential to translate the demonstrated technical capabilities of generative AI into measurable, equitable improvements in patient outcomes.

The evidence reviewed here also suggests that the most significant near-term gains are likely to come from hybrid pipelines that couple generative components with human expert oversight, rather than from end-to-end autonomous systems. Concretely, this calls for clinician-in-the-loop designs, transparent uncertainty reporting, institutional governance structures that can procure and audit foundation models, and shared benchmarks that measure both technical fidelity and downstream clinical impact. Without these complementary investments, the gap between impressive in silico performance and durable clinical value will persist.

8. Potential Biases

The present literature review is subject to several potential biases that merit explicit acknowledgment:

Publication bias. The review is likely to overrepresent studies reporting positive or statistically significant findings, because negative or inconclusive work is less frequently published.

Selection bias. Inclusion may have been influenced by the author's research interests, the coverage of the databases consulted, and the dominance of English-language publications.

Interpretation bias. The author's background and prior assumptions may have shaped the thematic synthesis and the relative weight assigned to different studies.

9. Recommendations for Future Research

To mitigate these biases and to consolidate the evidence base, future work in this area should:

Figure 2
Thematic distribution of generative AI applications identified in the reviewed literature

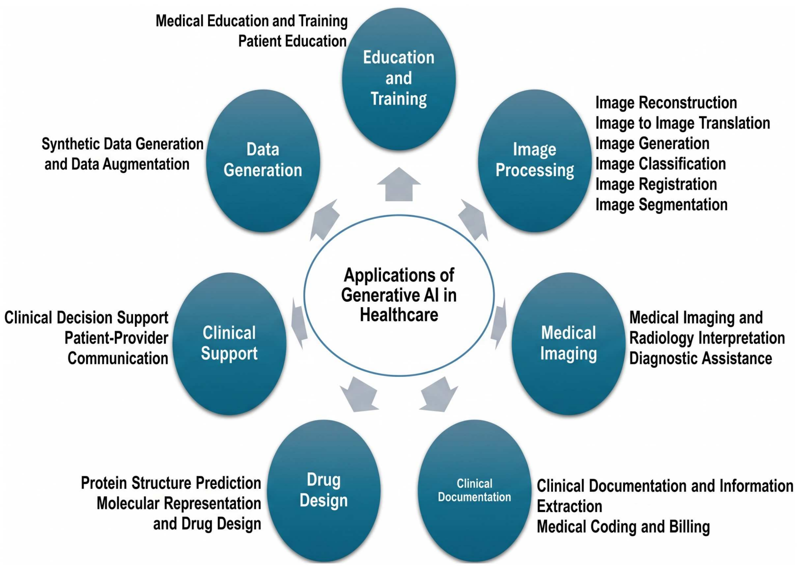
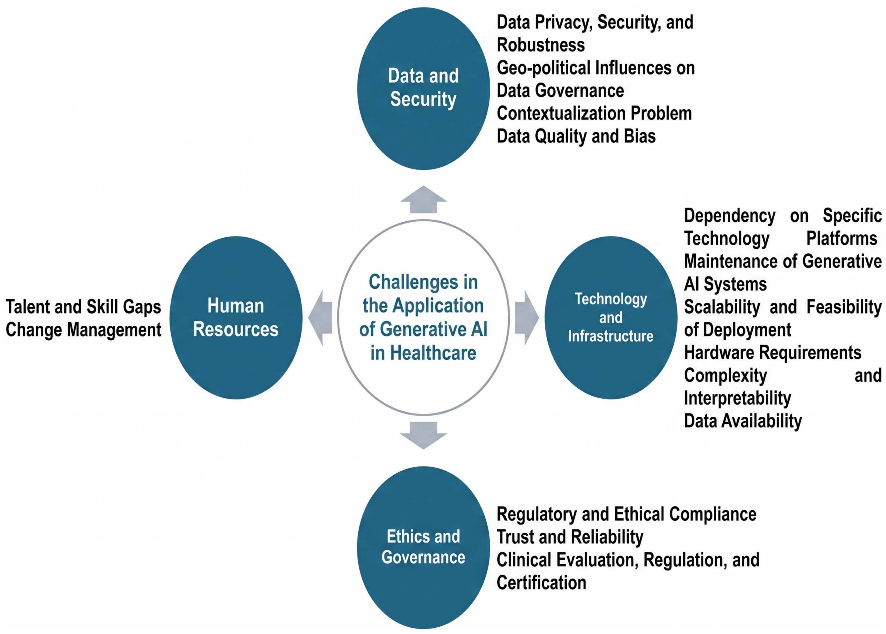


Figure 3
Overview of the four thematic challenge domains for generative AI in healthcare



Adopt a broader and more systematic literature-search strategy, including gray literature and preprints, and apply structured tools to assess study quality and risk of bias.

Encourage the publication of both positive and negative findings to yield a more balanced picture of the field; involve reviewers from diverse clinical, technical, and geographic backgrounds; and explicitly report the limitations of the review process and their implications for the conclusions drawn.

Acknowledgment

The author is grateful to JDSIS for their support in the publication of this research review.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by the author.

Conflicts of Interest

The author declares that he has no conflicts of interest to this work.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Author Contribution Statement

Wael Rahhal: Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

References

[1] Epstein, Z., Hertzmann, A., Akten, M., Farid, H., Fjeld, J., Frank, M. R., . . . , & Smith, A. (2023). Art and the science of generative AI. *Science*, *380*(6650), 1110–1111. https://doi.org/10.1126/science.adh4451

[2] Reddy, S. (2024). Generative AI in healthcare: An implementation science informed translational path on application, integration and governance. *Implementation Science*, *19*(1), 27. https://doi.org/10.1186/s13012-024-01357-9

[3] Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). *Generative AI at work*. In NBER Working Papers (*31161*). National Bureau of Economic Research. https://doi.org/10.3386/w31161

[4] Levac, B., Jalal, A., & Tamir, J. I. (2023). Accelerated motion correction for MRI using score-based generative models. In *2023 IEEE 20th International Symposium on Biomedical Imaging*, 1–5. https://doi.org/10.1109/ISBI53787.2023.10230457

[5] Korkmaz, Y., Yurt, M., Dar, S. U. H., Özbey, M., & Cukur, T. (2021). Deep MRI reconstruction with generative vision transformers. In *Machine Learning for Medical Image Reconstruction: 4th International Workshop*, 54–64. https://doi.org/10.1007/978-3-030-88552-6_6

[6] Gothwal, R., Tiwari, S., & Shivani, S. (2022). Computational medical image reconstruction techniques: A comprehensive review. *Archives of Computational Methods in Engineering*, *29*(7), 5635–5662. https://doi.org/10.1007/s11831-022-09785-w

[7] Özbey, M., Dalmaz, O., Dar, S. U. H., Bedel, H. A., Özturk, Ş., Güngör, A., & Çukur, T. (2023). Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, *42*(12), 3524–3539. https://doi.org/10.1109/TMI.2023.3290149

[8] Xie, Y., & Li, Q. (2022). Measurement-conditioned denoising diffusion probabilistic model for under-sampled medical image reconstruction. In *Medical Image Computing and Computer*

Assisted Intervention: 25th International Conference, 655–664. https://doi.org/10.1007/978-3-031-16446-0_62

[9] Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 5967–5976. https://doi.org/10.1109/CVPR.2017.632

[10] Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision*, 2242–2251. https://doi.org/10.1109/ICCV.2017.244

[11] Graf, R., Schmitt, J., Schlaeger, S., Möller, H. K., Sideri-Lampretsa, V., Sekuboyina, A., . . . , & Kirschke, J. S. (2023). Denoising diffusion-based MRI to CT image translation enables automated spinal segmentation. *European Radiology Experimental*, *7*(1), 70. https://doi.org/10.1186/s41747-023-00385-2

[12] Templin, T., Perez, M. W., Sylvia, S., Leek, J., & Sinnott-Armstrong, N. (2024). Addressing 6 challenges in generative AI for digital health: A scoping review. *PLOS Digital Health*, *3*(5), e0000503. https://doi.org/10.1371/journal.pdig.0000503

[13] Dorjsembe, Z., Odonchimed, S., & Xiao, F. (2022). Three-dimensional medical image synthesis with denoising diffusion probabilistic models. In *Medical Imaging with Deep Learning*, 1–3.

[14] Kim, B., & Ye, J. C. (2022). Diffusion deformable model for 4D temporal medical image generation. In *Medical Image Computing and Computer Assisted Intervention: 25th International Conference*, 539–548. https://doi.org/10.1007/978-3-031-16431-6_51

[15] Moghadam, P. A., van Dalen, S., Martin, K. C., Lennerz, J., Yip, S., Farahani, H., & Bashashati, A. (2023). A morphology focused diffusion probabilistic model for synthesis of histopathology images. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision*, 1999-2008 https://doi.org/10.1109/WACV56688.2023.00204

[16] Pinaya, W. H. L., Tudosiu, P.-D., Dafflon, J., da Costa, P. F., Fernandez, V., Nachev, P., . . . , & Cardoso, M. J. (2022). Brain imaging generation with latent diffusion models. In *Deep Generative Models: Second MICCAI Workshop*, 117–126. https://doi.org/10.1007/978-3-031-18576-2_12

[17] Waibel, D. J. E., Kiermeyer, N., Atwell, S., Sadafi, A., Meier, M., & Marr, C. (2022). SHAPR predicts 3D cell shapes from 2D microscopic images. *iScience*, *25*(11), 105298. https://doi.org/10.1016/j.isci.2022.105298

[18] Müller-Franzes, G., Niehues, J. M., Khader, F., Arasteh, S. T., Haarburger, C., Kuhl, C., . . . , & Truhn, D. (2023). A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Scientific Reports*, *13*(1), 12098. https://doi.org/10.1038/s41598-023-39278-0

[19] Pan, S., Wang, T., Qiu, R. L. J., Axente, M., Chang, C.-W., Peng, J., . . . , & Yang, X. (2023). 2D medical image synthesis using transformer-based denoising diffusion probabilistic model. *Physics in Medicine & Biology*, *68*(10), 105004. https://doi.org/10.1088/1361-6560/acca5c

[20] Shehab, M., Abualigah, L., Shambour, Q., Abu-Hashem, M. A., Shambour, M. K. Y., Alsalibi, A. I., & Gandomi, A. H. (2022). Machine learning in medical applications: A review of state-of-the-art methods. *Computers in Biology and Medicine*, *145*, 105458. https://doi.org/10.1016/j.compbiomed.2022.105458

[21] Oh, H.-J., & Jeong, W.-K. (2023). DiffMix: Diffusion model-based data synthesis for nuclei segmentation and classification in imbalanced pathology image datasets. In *Medical Image Computing and Computer Assisted Intervention: 26th International Conference*, 337–345. https://doi.org/10.1007/978-3-031-43898-1_33

[22] Yang, Y., Fu, H., Aviles-Rivero, A. I., Schönlieb, C.-B., & Zhu, L. (2023). DiffMIC: Dual-guidance diffusion network for medical image classification. In *Medical Image Computing and Computer Assisted Intervention: 26th International Conference*, 95–105. https://doi.org/10.1007/978-3-031-43987-2_10

[23] Dong, Y., Zhang, M., Qiu, L., Wang, L., & Yu, Y. (2023). An arrhythmia classification model based on vision transformer with deformable attention. *Micromachines*, *14*(6), 1155. https://doi.org/10.3390/mi14061155

[24] Garaiman, A., Nooralahzadeh, F., Mihai, C., Perez Gonzalez, N., Gkikopoulos, N., Becker, M. O., . . . , & Maurer, B. (2023). Vision transformer assisting rheumatologists in screening for capillaroscopy changes in systemic sclerosis: An artificial intelligence model. *Rheumatology*, *62*(7), 2492–2500. https://doi.org/10.1093/rheumatology/keac541

[25] Kim, B., Han, I., & Ye, J. C. (2022). DiffuseMorph: Unsupervised deformable image registration using diffusion model. In *Computer Vision: 17th European Conference*, 347–364. https://doi.org/10.1007/978-3-031-19821-2_20

[26] Azad, R., Heidari, M., Shariatnia, M., Aghdam, E. K., Karimijafarbigloo, S., Adeli, E., & Merhof, D. (2022). TransDeepLab: Convolution-free transformer-based Deeplab v3+ for medical image segmentation. In *Predictive Intelligence in Medicine: 5th International Workshop*, 91–102. https://doi.org/10.1007/978-3-031-16919-9_9

[27] Zhang, L., Wu, F., Bronik, K., & Papiez, B. W. (2025). DiffuSeg: Domain-driven diffusion for medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, *29*(5), 3619–3631. https://doi.org/10.1109/JBHI.2025.3526806

[28] Balouch, B. A. K., & Hussain, F. (2023). A transformer based approach for abstractive text summarization of radiology reports. In *5th International Conference on Applied Engineering and Natural Sciences*, *1*(1), 476–486. https://doi.org/10.59287/icaens.1042

[29] Bhattacharya, M., Jain, S., & Prasanna, P. (2022). RadioTransformer: A cascaded global-focal transformer for visual attention–guided disease classification. In *Computer Vision: 17th European Conference*, 679–698. https://doi.org/10.1007/978-3-031-19803-8_40

[30] Chaudhari, G. R., Liu, T., Chen, T. L., Joseph, G. B., Vella, M., Lee, Y. J., . . . , & Sohn, J. H. (2022). Application of a domain-specific BERT for detection of speech recognition errors in radiology reports. *Radiology: Artificial Intelligence*, *4*(4), e210185. https://doi.org/10.1148/ryai.210185

[31] Jacenkow, G., O'Neil, A. Q., & Tsaftaris, S. A. (2022). Indication as prior knowledge for multimodal disease classification in chest radiographs with transformers. In *2022 IEEE 19th International Symposium on Biomedical Imaging*, 1–5. https://doi.org/10.1109/ISBI52829.2022.9761567

[32] Li, J., Lin, Y., Zhao, P., Liu, W., Cai, L., Sun, J., . . . , & Wang, Z. (2022). Automatic text classification of actionable radiology reports of tinnitus patients using bidirectional encoder representations from transformer

(BERT) and in-domain pre-training (IDPT). *BMC Medical Informatics and Decision Making*, *22*(1), 200. https://doi.org/10.1186/s12911-022-01946-y

[33] Azizi, S., Hier, D. B., & Wunsch, D. C. (2022). Enhanced neurologic concept recognition using a named entity recognition model based on transformers. *Frontiers in Digital Health*, *4*, 1065581. https://doi.org/10.3389/fdgth.2022.1065581

[34] Moezzi, S. A. R., Ghaedi, A., Rahmanian, M., Mousavi, S. Z., & Sami, A. (2023). Application of deep learning in generating structured radiology reports: A transformer-based technique. *Journal of Digital Imaging*, *36*(1), 80–90. https://doi.org/10.1007/s10278-022-00692-x

[35] Mohsan, M. M., Akram, M. U., Rasool, G., Alghamdi, N. S., Baqai, M. A. A., & Abbas, M. (2023). Vision transformer and language model based radiology report generation. *IEEE Access*, *11*, 1814–1824. https://doi.org/10.1109/ACCESS.2022.3232719

[36] Nimalsiri, W., Hennayake, M., Rathnayake, K., Ambegoda, T. D., & Meedeniya, D. (2023). Automated radiology report generation using transformers. In *2023 3rd International Conference on Advanced Research in Computing*, 90–95. https://doi.org/10.1109/ICARC57651.2023.10145699

[37] Chen, S., Wu, Z., Li, M., Zhu, Y., Xie, H., Yang, P., . . . , & Lei, B. (2023). FIT-Net: Feature interaction transformer network for pathologic myopia diagnosis. *IEEE Transactions on Medical Imaging*, *42*(9), 2524–2538. https://doi.org/10.1109/TMI.2023.3260990

[38] Chen, X., Zhang, K., Abdoli, N., Gilley, P. W., Wang, X., Liu, H., . . . , & Qiu, Y. (2022). Transformers improve breast cancer diagnosis from unregistered multi-view mammograms. *Diagnostics*, *12*(7), 1549. https://doi.org/10.3390/diagnostics12071549

[39] Alp, S., Akan, T., Bhuiyan, M. S., Disbrow, E. A., Conrad, S. A., Vanchiere, J. A., . . . , & Bhuiyan, M. A. N. (2024). Joint transformer architecture in brain 3D MRI classification: Its application in Alzheimer's disease classification. *Scientific Reports*, *14*(1), 8996. https://doi.org/10.1038/s41598-024-59578-3

[40] Hosain, A. K. M. S., Islam, M., Mehedi, M. H. K., Kabir, I. E., & Khan, Z. T. (2022). Gastrointestinal disorder detection with a transformer based approach. In *2022 IEEE 13th Annual Information Technology, Electronics and Mobile Communication Conference*, 0280–0285. https://doi.org/10.1109/IEMCON56893.2022.9946531

[41] Hu, D., Xie, F., Jiang, Z., Zheng, Y., & Shi, J. (2022). Histopathology cross-modal retrieval based on dual-transformer network. In *2022 IEEE 22nd International Conference on Bioinformatics and Bioengineering*, 97–102. https://doi.org/10.1109/BIBE55377.2022.00028

[42] Mo, Y., Han, C., Liu, Y., Liu, M., Shi, Z., Lin, J., . . . , & Liang, C. (2023). HoVer-trans: Anatomy-aware hover-transformer for ROI-free breast cancer diagnosis in ultrasound images. *IEEE Transactions on Medical Imaging*, *42*(6), 1696–1706. https://doi.org/10.1109/TMI.2023.3236011

[43] Zhou, H.-Y., Yu, Y., Wang, C., Zhang, S., Gao, Y., Pan, J., . . . , & Li, W. (2023). A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nature Biomedical Engineering*, *7*(6), 743–755. https://doi.org/10.1038/s41551-023-01045-x

[44] Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., . . . , & Zhou, L. (2021). LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, *1*, 2579–2591. https://doi.org/10.18653/v1/2021.acl-long.201

[45] Lentzen, M., Madan, S., Lage-Rupprecht, V., Kühnel, L., Fluck, J., Jacobs, M., . . . , & Fröhlich, H. (2022). Critical assessment of transformer-based AI models for German clinical notes. *JAMIA Open*, *5*(4), ooac087. https://doi.org/10.1093/jamiaopen/ooac087

[46] Li, Y., Wehbe, R. M., Ahmad, F. S., Wang, H., & Luo, Y. (2022). *Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences*. arXiv. https://doi.org/10.48550/ARXIV.2201.11838

[47] Moon, S., He, H., & Liu, H. (2022). Sublanguage characteristics of clinical documents. In *2022 IEEE International Conference on Bioinformatics and Biomedicine*, 3280–3286. https://doi.org/10.1109/BIBM55620.2022.9995620

[48] Oh, S. H., Kang, M., & Lee, Y. (2022). Protected health information recognition by fine tuning a pre-training transformer model. *Healthcare Informatics Research*, *28*(1), 16–24. https://doi.org/10.4258/hir.2022.28.1.16

[49] Searle, T., Ibrahim, Z., Teo, J., & Dobson, R. J. B. (2023). Discharge summary hospital course summarisation of in patient Electronic Health Record text with clinical concept guided deep pre-trained Transformer models. *Journal of Biomedical Informatics*, *141*, 104358. https://doi.org/10.1016/j.jbi.2023.104358

[50] Sivarajkumar, S., & Wang, Y. (2023). HealthPrompt: A zero-shot learning paradigm for clinical natural language processing. In *Proceedings of the American Medical Informatics Association Annual Symposium*, 972–981.

[51] Solarte-Pabón, O., Montenegro, O., García-Barragán, A., Torrente, M., Provencio, M., Menasalvas, E., & Robles, V. (2023). Transformers for extracting breast cancer information from Spanish clinical narratives. *Artificial Intelligence in Medicine*, *143*, 102625. https://doi.org/10.1016/j.artmed.2023.102625

[52] Wei, Q., Zuo, X., Anjum, O., Hu, Y., Denlinger, R., Bernstam, E. V., . . . , & Xu, H. (2022). ClinicalLayoutLM: A pre-trained multi-modal model for understanding scanned document in electronic health records. In *2022 IEEE International Conference on Big Data*, 2821–2827. https://doi.org/10.1109/BigData55660.2022.10020569

[53] Liu, L., Perez-Concha, O., Nguyen, A., Bennett, V., & Jorm, L. (2022). Hierarchical label-wise attention transformer model for explainable ICD coding. *Journal of Biomedical Informatics*, *133*, 104161. https://doi.org/10.1016/j.jbi.2022.104161

[54] López-García, G., Jerez, J. M., Ribelles, N., Alba, E., & Veredas, F. J. (2023). Explainable clinical coding with in-domain adapted transformers. *Journal of Biomedical Informatics*, *139*, 104323. https://doi.org/10.1016/j.jbi.2023.104323

[55] Coutinho, I., & Martins, B. (2022). Transformer-based models for ICD-10 coding of death certificates with Portuguese text. *Journal of Biomedical Informatics*, *136*, 104232. https://doi.org/10.1016/j.jbi.2022.104232

[56] Falissard, L., Morgand, C., Ghosn, W., Imbaud, C., Bounebache, K., & Rey, G. (2022). Neural translation and automated recognition of ICD-10 medical entities from natural language: Model development and performance assessment. *JMIR Medical Informatics*, *10*(4), e26353. https://doi.org/10.2196/26353

[57] Behjati, A., Zare-Mirakabad, F., Arab, S. S., & Nowzari-Dalini, A. (2022). Protein sequence profile prediction using ProtAlbert transformer. *Computational Biology and Chemistry*, *99*, 107717. https://doi.org/10.1016/j.compbiolchem.2022.107717

[58] Xiang, W., Xiong, Z., Chen, H., Xiong, J., Zhang, W., Fu, Z., . . . , & Shi, Q. (2024). FAPM: Functional annotation of proteins using multimodal models beyond structural modeling. *Bioinformatics*, *40*(12), btae680. https://doi.org/10.1093/bioinformatics/btae680

[59] Boadu, F., Cao, H., & Cheng, J. (2023). Combining protein sequences and structures with transformers and equivariant graph neural networks to predict protein function. *Bioinformatics*, *39*(Suppl_1), i318–i325. https://doi.org/10.1093/bioinformatics/btad208

[60] Geffen, Y., Ofran, Y., & Unger, R. (2022). DistilProtBert: A distilled protein language model used to distinguish between real proteins and their randomly shuffled counterparts. *Bioinformatics*, *38*(Supplement_2), ii95–ii98. https://doi.org/10.1093/bioinformatics/btac474

[61] Castro, E., Godavarthi, A., Rubinfien, J., Givechian, K., Bhaskar, D., & Krishnaswamy, S. (2022). Transformer-based protein generation with regularized latent space optimization. *Nature Machine Intelligence*, *4*(10), 840–851. https://doi.org/10.1038/s42256-022-00532-1

[62] Ferruz, N., Schmidt, S., & Höcker, B. (2022). ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, *13*(1), 4348. https://doi.org/10.1038/s41467-022-32007-7

[63] Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., & Linial, M. (2022). ProteinBERT: A universal deep-learning model of protein sequence and function. *Bioinformatics*, *38*(8), 2102–2110. https://doi.org/10.1093/bioinformatics/btac020

[64] Oliveira, G. B., Pedrini, H., & Dias, Z. (2023). TEMPROT: Protein function annotation using transformers embeddings and homology search. *BMC Bioinformatics*, *24*(1), 242. https://doi.org/10.1186/s12859-023-05375-0

[65] Bagal, V., Aggarwal, R., Vinod, P. K., & Priyakumar, U. D. (2022). MolGPT: Molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling*, *62*(9), 2064–2076. https://doi.org/10.1021/acs.jcim.1c00600

[66] Li, H., Zhao, D., & Zeng, J. (2022). KPGT: Knowledge-guided pre-training of graph transformer for molecular property prediction. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 857–867. https://doi.org/10.1145/3534678.3539426

[67] Chen, Y., & Esmaeilzadeh, P. (2024). Generative AI in medical practice: In-depth exploration of privacy and security challenges. *Journal of Medical Internet Research*, *26*, e53008. https://doi.org/10.2196/53008

[68] Ustymenko, S., & Phadke, A. (2024). Promise and challenges of generative AI in healthcare information systems. In *Proceedings of the 2024 ACM Southeast Conference*, 223–228. https://doi.org/10.1145/3603287.3651196

[69] Hu, W., & Wang, S. Y. (2022). Predicting glaucoma progression requiring surgery using clinical free-text notes and transfer learning with transformers. *Translational Vision Science & Technology*, *11*(3), 37. https://doi.org/10.1167/tvst.11.3.37

[70] Huang, G. (2022). Surgical action recognition and prediction with transformers. In *2022 IEEE 2nd International Conference on Software Engineering and Artificial Intelligence*, 36–40. https://doi.org/10.1109/SEAI55746.2022.9832094

[71] Wang, L., He, H., Wen, A., Moon, S., Fu, S., Peterson, K. J., . . . , & Liu, H. (2023). Acquisition of a lexicon for family history information: Bidirectional encoder representations from transformers–assisted sublanguage analysis. *JMIR Medical Informatics*, *11*, e48072. https://doi.org/10.2196/48072

[72] Waters, M. R., Aneja, S., & Hong, J. C. (2023). Unlocking the power of ChatGPT, artificial intelligence, and large language models: Practical suggestions for radiation oncologists. *Practical Radiation Oncology*, *13*(6), E484–E490. https://doi.org/10.1016/j.prro.2023.06.011

[73] Moshirfar, M., Altaf, A., Stoakes, I. M., Tuttle, J. J., & Hoopes, P. C. (2023). Artificial intelligence in ophthalmology: A comparative analysis of GPT-3.5, GPT-4, and human expertise in answering StatPearls questions. *Cureus*, *15*(6), e40822. https://doi.org/10.7759/cureus.40822

[74] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., . . . , & Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, *620*(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2

[75] Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T.-Y. (2022). BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, *23*(6), bbac409. https://doi.org/10.1093/bib/bbac409

[76] Lan, L., You, L., Zhang, Z., Fan, Z., Zhao, W., Zeng, N., . . . , & Zhou, X. (2020). Generative adversarial networks and its applications in biomedical informatics. *Frontiers in Public Health*, *8*, 164. https://doi.org/10.3389/fpubh.2020.00164

[77] Carrasco Limeros, S., Majchrowska, S., Zoubi, M. K., Rosén, A., Suvilehto, J., Sjöblom, L., & Kjellberg, M. (2023). Assessing GAN-based generative modeling on skin lesions images. In *Digital Interaction and Machine Intelligence: Proceedings of MIDI'2022 – 10th Machine Intelligence and Digital Interaction – Conference*, 93–102. https://doi.org/10.1007/978-3-031-37649-8_10

[78] Eysenbach, G. (2023). The role of ChatGPT, generative language models, and artificial intelligence in medical education: A conversation with ChatGPT and a call for papers. *JMIR Medical Education*, *9*, e46885. https://doi.org/10.2196/46885

[79] Khan, R. A., Jawaid, M., Khan, A. R., & Sajjad, M. (2023). ChatGPT - Reshaping medical education and clinical management. *Pakistan Journal of Medical Sciences*, *39*(2), 605–607. https://doi.org/10.12669/pjms.39.2.7653

[80] Haupt, C. E., & Marks, M. (2023). AI-generated medical advice-GPT and beyond. *JAMA*, *329*(16), 1349–1350. https://doi.org/10.1001/jama.2023.5321

[81] Uprety, D., Zhu, D., & West, H. (2023). ChatGPT—A promising generative AI tool and its implications for cancer care. *Cancer*, *129*(15), 2284–2289. https://doi.org/10.1002/cncr.34827

[82] Jain, S., Subzwari, S. W. A., & Subzwari, S. A. A. (2024). Generative AI for healthcare engineering and technology challenges. In *Transfer, Diffusion and Adoption of Next-Generation Digital Technologies: IFIP WG 8.6 International Working Conference on Transfer and Diffusion of IT*, 68–80. https://doi.org/10.1007/978-3-031-50188-3_7

[83] Sai, S., Gaur, A., Sai, R., Chamola, V., Guizani, M., & Rodrigues, J. J. P. C. (2024). Generative AI for transformative healthcare: A comprehensive study of emerging models,

applications, case studies, and limitations. *IEEE Access*, *12*, 31078–31106. https://doi.org/10.1109/ACCESS.2024.3367715

[84] Zhang, P., & Kamel Boulos, M. N. (2023). Generative AI in medicine and healthcare: Promises, opportunities and challenges. *Future Internet*, *15*(9), 286. https://doi.org/10.3390/fi15090286

[85] Kuzlu, M., Xiao, Z., Sarp, S., Catak, F. O., Gurler, N., & Guler, O. (2023). The rise of generative artificial intelligence in healthcare. In *2023 12th Mediterranean Conference on Embedded Computing*, 1–4. https://doi.org/10.1109/MECO58584.2023.10155107

[86] Korngiebel, D. M., & Mooney, S. D. (2021). Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. *npj Digital Medicine*, *4*(1), 93. https://doi.org/10.1038/s41746-021-00464-x

How to Cite: Rahhal, W. (2026). Generative AI in Healthcare: Applications and Challenges. *Journal of Data Science and Intelligent Systems*. https://doi.org/10.47852/bonviewJDSIS62024354