RESEARCH ARTICLE OPEN ACCESS

AI in Healthcare Frameworks that Prioritize Fairness, Security, and Accountability

Adepeju Ayotunde Adeyinka[1], Yejide Lamina[2], Yewande Iyimide Adeyeye[3], Andrew Yaw Minkah[4], Obah Edom Tawo[5]

[1]Department of Data Science and Engineering, North Carolina A&T State University, North Carolina, USA.
[2]Lagos State University, College of Medicine, Lagos, Nigeria
[3]Dumfries and Galloway royal infirmary, Dumfries, Scotland
[4]Department of Healthcare Administration, University of the Potomac
[5]Southern Alberta Institute of Technology, Alberta, Canada (SAIT).

Abstract:

The integration of artificial intelligence (AI) into healthcare systems offers transformative potential for improving diagnostic accuracy, clinical decision-making, and operational efficiency. However, the deployment of AI in high-stakes medical environments raises significant concerns regarding fairness, security, and accountability. This review explores the foundational principles, emerging technologies, and regulatory frameworks shaping the development of responsible AI in healthcare. It examines sources of algorithmic bias, privacy vulnerabilities, and explainability challenges while highlighting real-world case studies and mitigation strategies such as debiasing techniques, federated learning, and model governance. The paper also discusses the role of decentralized identity, regulatory technology (RegTech), and blockchain in enhancing data control and auditability. Emphasizing the need for cross-jurisdictional alignment and adaptive compliance systems, the review synthesizes key insights and offers strategic recommendations for building equitable, secure, and ethically grounded AI infrastructures. By aligning technical innovation with ethical imperatives, this work contributes to the evolving discourse on trustworthy AI in modern healthcare ecosystems.

Keywords: Artificial Intelligence, Healthcare, Fairness, Security, Accountability

1. Introduction

1.1 Background and Motivation for AI Adoption in Healthcare

The integration of artificial intelligence (AI) into healthcare systems marks a transformative evolution driven by the need to enhance efficiency, accuracy, and personalization in medical practice. AI technologies such as machine learning, natural language processing, and computer vision are being rapidly adopted to address persistent challenges in healthcare, including diagnostic errors, workflow inefficiencies, and variability in clinical decision-making (Jiang et al., 2017). As healthcare systems confront mounting pressures from aging populations and chronic disease burdens, AI provides the capacity to process vast datasets at speeds and scales beyond human capability, thereby supporting predictive modeling, early detection, and tailored treatment interventions (Topol, 2019).

Figure 1 shows a healthcare professional discussing the primary barriers to successful AI integration in clinical and administrative settings. These include data governance issues, alignment of AI with business strategy, talent shortages, and implementation gaps. The figure highlights that 82% of enterprises fail to scale AI to achieve meaningful business outcomes.

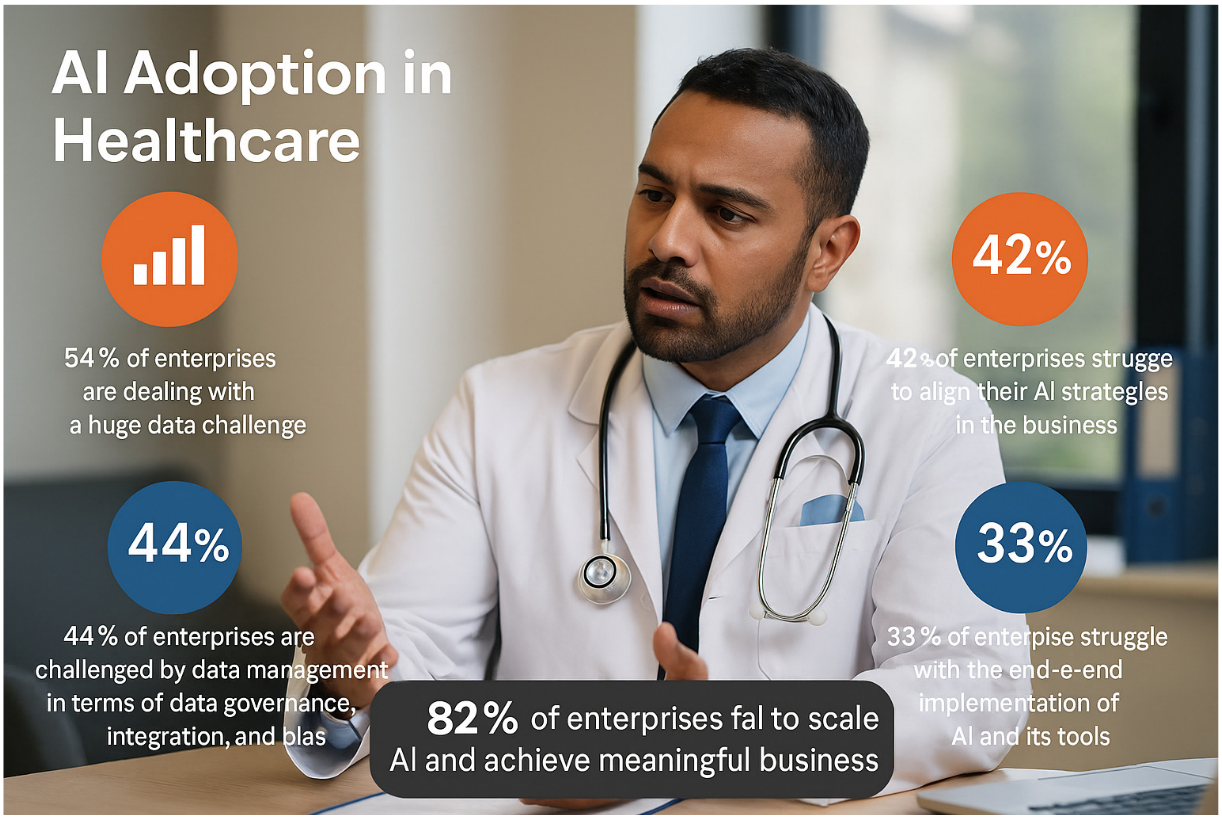
AI Adoption in Healthcare
54 % of enterprises are dealing with a huge data challenge
44%
44 % of enterprises are challenged by data management in terms of data governance, integration, and blas
42%
42%of enterprises strugge to align their AI strategles in the business
33%
33 % of enterpise struggle with the end-e-end implementation of AI and its tools
82% of enterprises fal to scale AI and achieve meaningful business

Figure 1: Key Challenges in AI Adoption Across Healthcare Enterprises

The motivation for AI deployment is also closely tied to the digitization of health records and the availability of large, diverse datasets generated through electronic health records (EHRs), wearable sensors, and genomic sequencing (Dilsizian & Siegel, 2014). These data-rich environments are fertile ground for AI systems designed to uncover latent patterns and insights that inform clinical practice and public health strategies (Obermeyer & Emanuel, 2016). Furthermore, AI is viewed as a pivotal tool for operational efficiency, enabling hospitals and health organizations to optimize resource allocation, reduce costs, and improve patient engagement through automated systems and intelligent triage (Davenport & Kalakota, 2019). The accelerating convergence of computational power, medical data, and clinical need underscores the global motivation to adopt AI as a cornerstone of next-generation healthcare delivery.

1.2 The Critical Need for Fairness, Security, and Accountability in AI Healthcare Systems

The deployment of AI in healthcare demands heightened attention to fairness, security, and accountability due to the sensitive nature of medical data and the profound impact AI-driven decisions can have on patient outcomes. Bias in AI algorithms—especially racial, gender, and socioeconomic disparities—can perpetuate existing healthcare inequities if not systematically addressed. For instance, a widely used population health algorithm was found to underestimate the healthcare needs of Black patients, raising significant ethical concerns (Obermeyer et al., 2019).

Ensuring fairness requires proactive debiasing strategies and transparent reporting of algorithmic performance across diverse populations (Raji & Buolamwini, 2019). However, fairness alone is insufficient without robust safeguards to protect patient data privacy. The exponential growth of medical big data has increased the risk of unauthorized access and breaches, necessitating advanced encryption, secure multi-party computation, and differential privacy techniques (Price & Cohen, 2019).

Accountability, often neglected in technical development, is essential for ethical AI deployment. The "right to explanation" as mandated by the EU's GDPR illustrates the growing demand for explainable and interpretable AI systems in healthcare (Goodman & Flaxman, 2017). Moreover, translating high-level AI ethics principles into actionable governance frameworks remains a key challenge for practitioners and regulators alike (Morley et al., 2020). Without integrated accountability mechanisms, AI systems risk becoming opaque decision-makers with little recourse for error, undermining trust and safety in clinical environments.

1.3 Research Questions and Scope of the Review

As artificial intelligence becomes increasingly integrated into healthcare systems, critical questions arise regarding how to ensure that these technologies are fair, secure, and accountable. This review aims to investigate how current AI frameworks in healthcare address the challenges of algorithmic bias, data privacy, and explainability. Specifically, the following research questions guide this review: (1) What technical and ethical approaches are being used to mitigate bias in AI healthcare systems? (2) How are issues of data security and patient privacy managed in AI-enabled medical environments? (3) What mechanisms are in place to ensure transparency and accountability in clinical AI decision-making?

The scope of this review encompasses peer-reviewed academic literature, case studies, and regulatory frameworks that focus on AI systems used in diagnostics, treatment recommendations, patient monitoring, and administrative operations within healthcare. It emphasizes both the technical design of AI algorithms and the policy mechanisms that govern their deployment (Vokinger et al., 2021). Recognizing the tension between innovation and safety, this review takes a multidisciplinary approach that includes perspectives from data science, medicine, ethics, and law (Wiens et al., 2019). By doing so, it aims to synthesize best practices while identifying critical gaps in the development of trustworthy AI systems in healthcare (Panch et al., 2019).

1.4 Methodology for Literature Selection and Analysis

This review adopts a scoping review methodology to map the existing literature on AI frameworks in healthcare that emphasize fairness, security, and accountability. Following the foundational framework by Arksey and O'Malley (2005), the review includes five core stages: identifying the research question, identifying relevant studies, selecting studies based on inclusion/exclusion criteria, charting the data, and collating and summarizing results. To enhance rigor, the methodological enhancements proposed by Levac et al. (2010) were incorporated, including stakeholder consultation and iterative team review.

Studies were selected through comprehensive searches across databases such as PubMed, IEEE Xplore, Scopus, and Google Scholar. Only peer-reviewed journal articles, conference proceedings, and white papers published up to 2022 were included. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) was employed to ensure transparent reporting of study identification, screening, eligibility, and inclusion processes (Tricco et al., 2018).

The data extraction and thematic synthesis process followed the JBI Manual for Evidence Synthesis (Peters et al., 2020), allowing the mapping of current practices, identification of knowledge gaps, and classification of methodologies addressing fairness, security, and accountability in AI-driven healthcare systems. This approach ensures comprehensive coverage and analytic clarity, supporting a multidimensional understanding of the research landscape.

1.5 Structure of the Paper

This review paper is organized into five core sections to provide a coherent and structured exploration of AI frameworks in healthcare that prioritize fairness, security, and accountability. Following the introduction, Section 2 presents the foundational landscape of artificial intelligence in healthcare, including a technical overview of common AI applications and the limitations that necessitate ethical governance.

Section 3 delves into fairness-centric frameworks, examining the origins of algorithmic bias, the socio-demographic implications of inequitable model performance, and debiasing strategies that have been implemented in real-world clinical settings. This section also introduces fairness metrics and equity assessment tools relevant to AI in health contexts.

Section 4 focuses on the critical pillars of security and accountability, addressing data protection protocols, privacy-preserving techniques like federated learning, and transparency-enhancing practices such as model explainability and auditability. Legal and regulatory considerations governing responsible AI use in healthcare are also discussed.

The paper concludes with Section 5, which synthesizes the findings, outlines current best practices, and proposes future research directions. It offers a forward-looking perspective on how integrated frameworks can systematically embed fairness, security, and accountability in AI systems to ensure safe, ethical, and inclusive healthcare delivery. This structure supports a multidisciplinary analysis aimed at informing policy, guiding practitioners, and shaping ethical AI development in health technology ecosystems.

2. Foundations of AI in Healthcare

2.1 Overview of AI Applications Across Clinical and Administrative Domains

Artificial intelligence has emerged as a transformative force across clinical and administrative domains of healthcare, facilitating more efficient, accurate, and data-driven medical practice. In clinical settings, AI technologies have demonstrated exceptional performance in image-based diagnostics, such as dermatological screening, radiological interpretation, and retinal analysis for diabetic retinopathy (Esteva et al., 2019). Natural language processing (NLP) enables the extraction of meaningful clinical information from unstructured electronic health records (EHRs), improving disease surveillance and personalized treatment pathways.

Figure 2 shows a conceptual visualization of the four primary domains where artificial intelligence is applied in healthcare. These include health services management, predictive medicine, patient data and diagnostics, and clinical decision-making. The arrangement emphasizes the interconnected nature of these domains with AI at the core, signifying its central role in modern healthcare transformation.

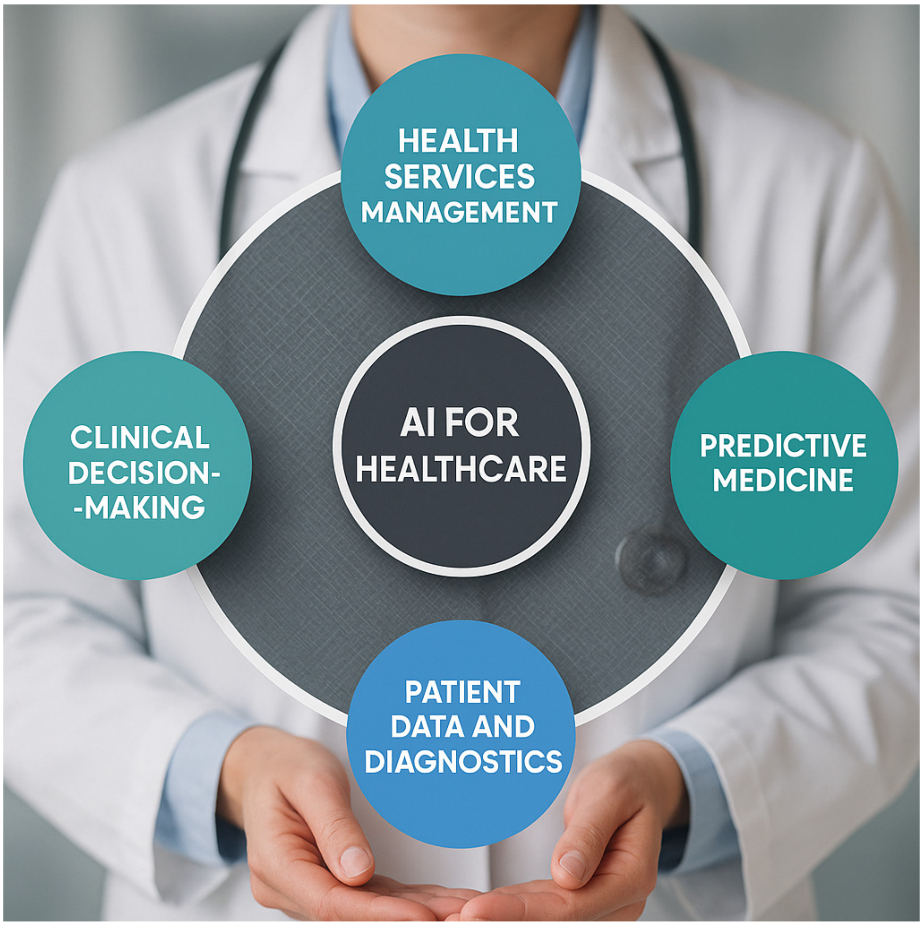
HEALTH SERVICES MANAGEMENT
CLINICAL DECISION- -MAKING
AI FOR HEALTHCARE
PREDICTIVE MEDICINE
PATIENT DATA AND DIAGNOSTICS

Figure 2: Core Domains of AI Integration in Healthcare

Moreover, AI models support real-time patient monitoring through integration with wearable devices, enabling early detection of critical conditions such as sepsis, atrial fibrillation, or respiratory failure. These models can synthesize data streams from physiological sensors, clinical histories, and laboratory values to inform predictive analytics and risk stratification tools (Rajkomar et al., 2019). In administrative domains, AI streamlines hospital operations through intelligent resource allocation, automated medical billing, claims processing, and scheduling optimization, reducing overhead and minimizing human error.

Recent systematic reviews show that AI systems, particularly deep learning models, can match or outperform clinicians in diagnostic tasks under specific conditions, although generalizability remains an ongoing challenge (Shen et al., 2021). This wide applicability underscores AI's dual role as a clinical decision support tool and an administrative optimizer, with ongoing research focused on harmonizing both functions to achieve holistic improvements in healthcare delivery.

2.2 Types of AI Models in Healthcare (ML, NLP, Computer Vision, Robotics)

The deployment of artificial intelligence (AI) in healthcare encompasses a broad spectrum of model types, each designed to address unique tasks in diagnostics, treatment planning, administrative support, and patient engagement. Machine learning (ML), particularly supervised learning algorithms like logistic regression, decision trees, support vector machines, and neural networks, dominates the predictive modeling landscape. These models are trained on historical healthcare data to predict disease onset, patient deterioration, and treatment response. For example, ML models have been effectively used to predict hospital readmission and sepsis risk with high sensitivity (Rajkomar et al., 2019).

Figure 3 shows six practical implementations of artificial intelligence in modern healthcare settings, represented through real-life imagery. These include machine learning, diagnosis and treatment apps, natural language processing, robotic process automation, rule-based expert systems, and physical robots. The figure highlights how AI tools are integrated into clinical workflows to enhance diagnostics, treatment precision, and operational efficiency.

Figure 3: Real-World Applications of AI Technologies in Healthcare

Natural language processing (NLP) is another vital component of healthcare AI. It enables systems to analyze unstructured clinical notes, extract relevant medical entities, and translate physician-patient conversations into structured datasets. NLP models such as BERT and its domain-specific variants like ClinicalBERT have been increasingly utilized to enhance clinical documentation and facilitate clinical decision support systems (Shickel et al., 2018).

Computer vision techniques, powered by convolutional neural networks (CNNs), are central to AI applications in medical imaging. These models enable high-accuracy detection and classification of anomalies in X-rays, MRIs, CT scans, and histopathological images, often surpassing human expert performance in controlled environments (Esteva et al., 2019). Finally, robotics integrated with AI facilitates minimally invasive surgeries, autonomous medical assistance, and patient rehabilitation through intelligent navigation and real-time sensory feedback (Yang et al., 2017). Collectively, these AI models provide scalable and adaptive solutions across diverse healthcare settings, continually evolving with advances in algorithmic design and clinical data integration.

2.3 Role of Big Data and EHRs in Enabling AI-Driven Solutions

Big data and electronic health records (EHRs) serve as foundational components for the development and implementation of artificial intelligence (AI) solutions in healthcare. The sheer volume, variety, and velocity of healthcare data generated from clinical encounters, medical imaging, laboratory results, genomics, and wearable sensors provide fertile ground for training machine learning models that support early disease detection, personalized treatment, and operational optimization (Raghupathi & Raghupathi, 2014). This wealth of information enables AI systems to identify complex patterns and correlations that are often imperceptible to human clinicians.

Figure 4 shows how artificial intelligence enhances various domains of public health through real-world applications. These include disease forecasting, risk prediction, health diagnosis, spatial modeling, public health surveillance, and pandemic/epidemic modeling. Each component is visually represented and linked to a central hub, emphasizing AI's integrative role in strengthening public health infrastructure.

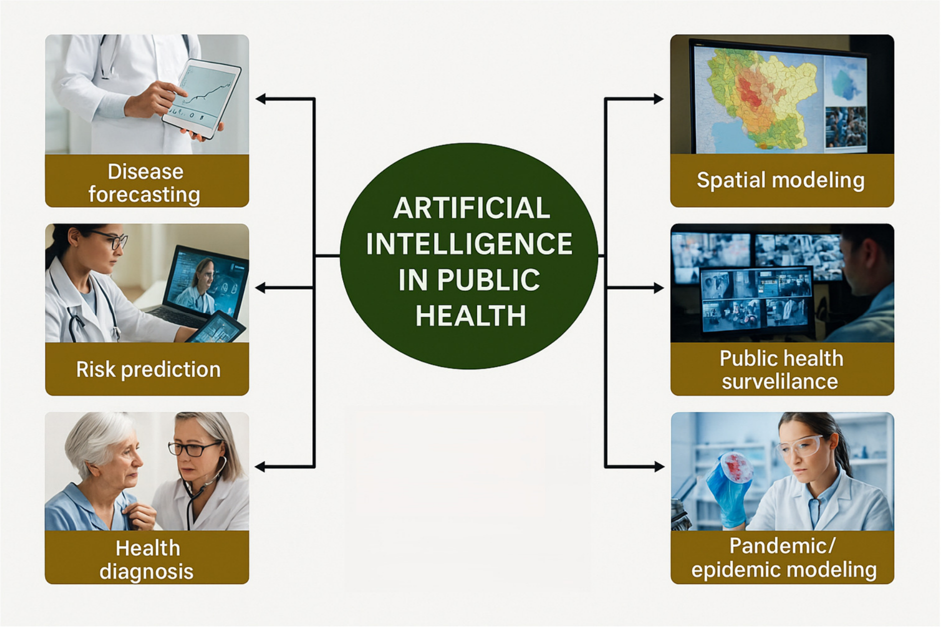
ARTIFICIAL INTELLIGENCE IN PUBLIC HEALTH
Disease forecasting
Risk prediction
Health diagnosis
Spatial modeling
Public health survelilance
Pandemic/ epidemic modeling

Figure 4: Core Applications of Artificial Intelligence in Public Health Practice

EHRs, in particular, are central to data-driven healthcare AI because they contain longitudinal and multimodal patient data, such as structured data (e.g., lab values, prescriptions) and unstructured text (e.g., physician notes). These records not only support retrospective analyses but also facilitate real-time clinical decision support when integrated with AI algorithms (Jensen et al., 2012). Moreover, the integration of EHRs with big data analytics has enabled the creation of predictive models for patient readmission, adverse event prevention, and population health management.

However, challenges persist due to data silos, inconsistencies in data standards, and privacy concerns. Nevertheless, initiatives promoting interoperability and standardized data representation—such as HL7 FHIR—are advancing the usability of EHR data for AI applications (Mehta & Pandit, 2018). As EHR systems evolve, their role in enabling scalable, explainable, and actionable AI-driven solutions is expected to become even more critical in shaping precision medicine and efficient healthcare delivery.

2.4 Limitations of Traditional AI Models: Bias, Opacity, and Vulnerability

Despite their growing influence in healthcare, traditional artificial intelligence (AI) models often suffer from critical limitations that undermine their reliability and trustworthiness. One of the most pressing issues is algorithmic bias, which occurs when AI systems trained on non-representative or historically biased datasets produce discriminatory outcomes. This is particularly problematic in healthcare, where underrepresentation of minority populations in training data can result in inaccurate diagnostics or inappropriate treatment recommendations (Obermeyer et al., 2019).

Another major limitation is model opacity—often described as the "black box" nature of complex AI systems like deep learning models. These systems can generate highly accurate predictions, but they frequently lack transparency, making it difficult for clinicians and stakeholders to understand the rationale behind medical decisions (Doshi-Velez & Kim, 2017). This lack of interpretability raises concerns about clinical accountability, especially in high-stakes scenarios where explainability is crucial for informed consent and medical liability.

Additionally, traditional AI systems are increasingly exposed to security vulnerabilities, including adversarial attacks, where subtle data manipulations can cause the model to make incorrect predictions. These attacks pose serious risks in clinical contexts, particularly in medical imaging and diagnostic support systems (Finlayson et al., 2019). The absence of robust defense mechanisms makes conventional AI tools susceptible to manipulation and exploitation, challenging their safe deployment in real-world healthcare settings.

Table 1 highlights the critical limitations of traditional artificial intelligence (AI) systems in healthcare. It outlines three core challenges—algorithmic bias, model opacity, and security vulnerability—alongside their implications. These issues undermine the reliability, fairness, and safety of AI-powered clinical tools.

Table 1: Key Limitations of Traditional AI Models in Healthcare

Limitation	Description	Implications	Examples / Sources
Algorithmic Bias	Results from training AI on non-representative or biased datasets.	Can lead to discriminatory diagnostics or treatments, especially for minority populations.	Obermeyer et al. (2019)
Model Opacity	Complex models (e.g., deep learning) function as "black boxes" lacking transparency.	Hinders clinical trust, informed consent, and accountability in medical decision-making.	Doshi-Velez & Kim (2017)
Security Vulnerability	Susceptibility to adversarial attacks that subtly alter input data to cause incorrect outputs.	Threatens the reliability and safety of diagnostic tools, especially in imaging and critical care applications.	Finlayson et al. (2019)

As these limitations become more apparent, there is a growing emphasis on developing fair, interpretable, and secure AI models tailored to the sensitive and complex nature of healthcare environments.

2.5 Ethical and Regulatory Landscape (HIPAA, GDPR, FDA AI/ML Regulations)

As artificial intelligence (AI) becomes deeply embedded in healthcare systems, the need for comprehensive ethical and regulatory oversight has become increasingly critical. In the United States, the Health Insurance Portability and Accountability Act (HIPAA) sets the foundation for safeguarding patient data privacy and security. However, HIPAA was not originally designed to address the complexities introduced by machine learning and predictive analytics, thus creating gaps in the regulation of data used for AI development and deployment (Price & Cohen, 2019).

In Europe, the General Data Protection Regulation (GDPR) provides a more expansive framework for data protection and individual rights. Notably, GDPR mandates transparency in algorithmic decision-making, granting individuals the “right to explanation” when automated systems are used to influence significant outcomes, such as clinical diagnoses or treatment recommendations (Goodman & Flaxman, 2017). These requirements impose strict obligations on developers and healthcare institutions to ensure the interpretability, fairness, and accountability of AI systems.

In addition, regulatory bodies such as the U.S. Food and Drug Administration (FDA) have introduced adaptive frameworks to evaluate and approve AI/ML-based Software as a Medical Device (SaMD). The FDA’s proposed "total product lifecycle" approach emphasizes continuous monitoring, transparency, and real-world performance assessment to ensure that AI systems remain safe and effective post-deployment (He et al., 2019). These evolving ethical and regulatory paradigms underscore the urgency of embedding compliance, accountability, and public trust into the design and implementation of healthcare AI technologies.

3. Fairness in AI Healthcare Frameworks

3.1 Sources of Bias in Healthcare Datasets and Model Development

Bias in artificial intelligence (AI) systems used in healthcare is often introduced during the data collection and model development phases, leading to skewed predictions and unequal healthcare outcomes. One major source of bias stems from non-representative training datasets, which fail to adequately reflect the diversity of the patient population. Historical data may overrepresent certain demographic groups (e.g., middle-aged white males) while underrepresenting minorities, women, or rural populations, resulting in models that generalize poorly across different groups (Chen et al., 2019).

Figure 5 shows the real-world implications of AI bias in healthcare through five illustrated blocks. These include inaccurate risk prediction, missed diagnoses, incorrect treatment recommendations, healthcare disparities, and systemic discrimination. Each element emphasizes how bias in AI models can compromise patient safety, equity, and clinical decision-making.

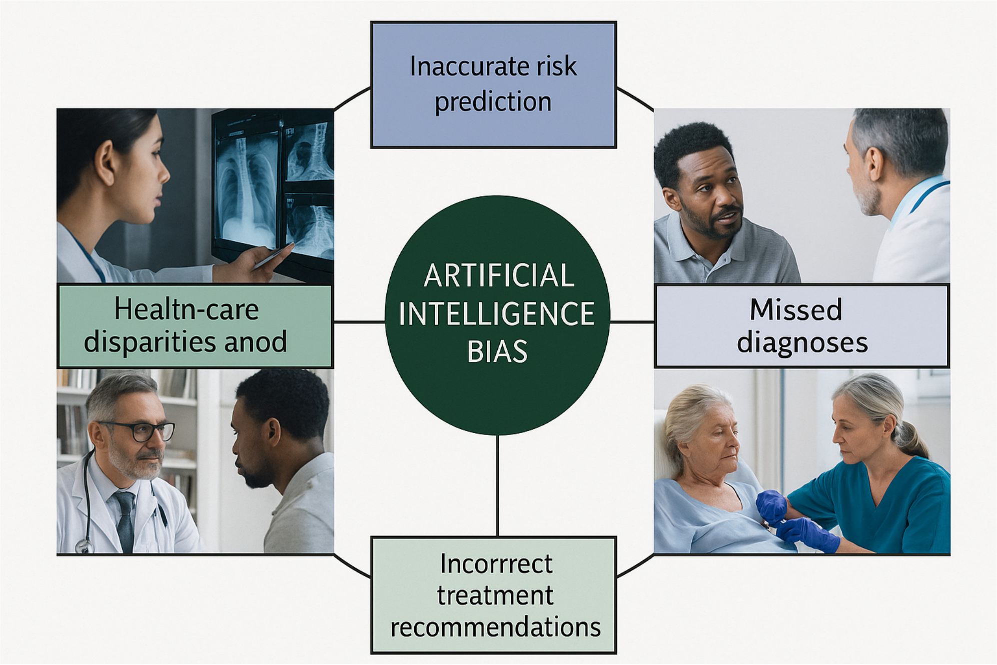
Inaccurate risk prediction
ARTIFICIAL INTELLIGENCE BIAS
Healtn-care disparities anod
Missed diagnoses
Incorrrect treatment recommendations

Figure 5: Consequences of Artificial Intelligence Bias in Healthcare

Another source of bias is labeling inconsistency or subjectivity, particularly in datasets that rely on clinician annotations. Variability in clinical judgment, documentation practices, or cultural interpretations of symptoms can propagate errors that disproportionately affect marginalized communities (Rajkomar et al., 2018).

Moreover, proxy variables—such as healthcare expenditure or visit frequency—often used in risk prediction models, can embed systemic inequalities rather than reflect actual health needs. For example, a well-known case demonstrated how a commercial algorithm significantly underestimated the health needs of Black patients because it used healthcare cost as a proxy for health status (Obermeyer et al., 2019).

Additionally, bias may be amplified during feature selection or model tuning if developers unknowingly select variables correlated with race, income, or gender. Without rigorous fairness auditing and transparency, these embedded biases can persist undetected, undermining the equity and trustworthiness of AI-enabled healthcare systems.

3.2 Debiasing Techniques: Re-sampling, Adversarial Training, and Fairness Constraints

To address bias in AI healthcare systems, researchers and practitioners have developed various debiasing techniques aimed at improving model fairness without sacrificing performance. One of the foundational approaches is re-sampling, which involves either oversampling underrepresented groups or undersampling overrepresented ones to create a more balanced training dataset. This method is particularly effective when dealing with class imbalance in diagnostic prediction tasks, ensuring that minority groups receive equitable attention during training (Zou & Schiebinger, 2018).

Another powerful method is adversarial training, which introduces an adversary during model training to penalize the learning of patterns correlated with protected attributes like race or gender. The goal is to make predictions accurate while ensuring they are not biased against specific subgroups. This technique has been successfully applied in healthcare applications such as mortality prediction and disease classification, reducing disparate impacts across demographic lines (Beutel et al., 2019).

Fairness constraints—also known as fairness-aware optimization—can be explicitly integrated into the objective function of machine learning algorithms. These constraints enforce parity in key performance metrics (e.g., false positive rates or predictive parity) across protected groups. Algorithms like equalized odds and demographic parity are commonly implemented in this regard. Studies have shown that imposing such constraints during model development can significantly reduce algorithmic discrimination in tasks like risk scoring and triage without degrading overall accuracy (Agarwal et al., 2018).

Table 2 presents key debiasing strategies used to mitigate algorithmic bias in AI healthcare systems. It describes re-sampling, adversarial training, and fairness constraints, focusing on their methods and outcomes. These techniques aim to ensure equitable performance across diverse demographic groups.

Table 2: Debiasing Techniques in AI Healthcare Models

Debiasing Technique	Description	Outcome	Examples / Sources

Re-sampling	Adjusts dataset distribution by oversampling underrepresented groups or undersampling dominant groups.	Improves class balance and ensures minority populations are adequately represented in training.	Zou & Schiebinger (2018)
Adversarial Training	Introduces a counter-model during training to penalize learning patterns tied to protected attributes.	Enhances fairness while preserving predictive performance in sensitive applications.	Beutel et al. (2019)
Fairness Constraints	Incorporates fairness-aware rules into the model's objective function, such as equalized odds.	Reduces discriminatory outcomes across groups while maintaining accuracy.	Agarwal et al. (2018)

These debiasing strategies, when properly validated and contextually applied, contribute to the development of AI systems that promote equitable healthcare outcomes across diverse populations.

3.3 Equity in Predictive Modeling: Race, Gender, and Socio-Economic Dimensions

Ensuring equity in predictive modeling is a critical challenge in the application of AI to healthcare, particularly when considering dimensions such as race, gender, and socio-economic status. These factors can significantly influence both the input data and the interpretability of model outcomes, often resulting in systemic disparities if not explicitly accounted for during model development. Racial disparities have been widely documented in predictive algorithms, where models trained on data reflecting historical biases may systematically underdiagnose or deprioritize care for racial minorities (Obermeyer et al., 2019).

Gender bias also presents a significant concern, especially in clinical applications like cardiovascular diagnostics where women are historically underrepresented in trial data. This imbalance can lead to poor generalization of AI systems and inaccurate risk predictions for female patients. Incorporating sex-specific variables and stratified model evaluation helps mitigate this form of bias, ensuring that predictive outputs are both clinically relevant and demographically sensitive (Larrazabal et al., 2020).

Table 3 outlines the equity challenges in healthcare predictive modeling related to race, gender, and socio-economic status. It highlights how systemic disparities in data representation and healthcare access can affect model fairness and utility. Incorporating stratified evaluation and fairness-aware learning is vital for developing inclusive AI tools.

Table 3: Equity Considerations in AI-Based Predictive Modeling for Healthcare

Equity Dimension	Description	Challenge	Examples / Sources
Race	Historical bias in healthcare data can lead to underdiagnosis and deprioritization of racial minorities.	Models may reflect and reinforce systemic inequities unless racially diverse data and evaluations are used.	Obermeyer et al. (2019)
Gender	Women are underrepresented in clinical datasets, especially for conditions like cardiovascular disease.	Risk prediction models may underperform for female patients without sex-specific variables or validation.	Larrazabal et al. (2020)
Socio-Economic Status (SES)	SES influences health via access to care and living conditions; often captured using proxy variables like insurance.	Models risk deprioritizing low-income groups unless explicitly tested across SES strata.	Chen et al. (2019)

Similarly, socio-economic status (SES) influences health outcomes through access to care, living conditions, and health behaviors. Predictive models that use proxies like insurance status or healthcare utilization often encode SES-related disparities. If left unaddressed, such models may perpetuate inequities by allocating fewer resources or less aggressive interventions to low-income populations. Approaches such as fairness-aware learning and stratified validation across SES groups are essential to developing more inclusive and ethically sound models (Chen et al., 2019).

3.4 Case Studies of Fair AI Deployment in Clinical Decision Support Systems

Real-world case studies provide valuable insights into how fairness can be operationalized in clinical decision support systems (CDSS) using artificial intelligence (AI). One notable example is the deployment of a machine learning model at Mount Sinai Health System for predicting severe COVID-19 outcomes. The developers incorporated fairness constraints to ensure that the model did not disproportionately assign high-risk scores to certain racial or socio-economic groups. By including race-stratified validation and adjusting for social determinants of health, the model demonstrated improved equity in triage decisions without sacrificing predictive accuracy (Wang et al., 2021).

Another case is IBM Watson for Oncology, which faced early criticism for biased recommendations due to training on data predominantly sourced from a single cancer center. This limitation led to treatment suggestions that were not generalizable across diverse populations. As a result, subsequent deployments adopted localized retraining and human-AI collaborative protocols to align outputs with the clinical context of underrepresented populations, illustrating the importance of inclusive data and clinician oversight in fair CDSS implementation (Strickland, 2019).

Table 4 showcases real-world case studies that illustrate fairness-conscious implementation of AI in clinical decision support systems (CDSS). These examples emphasize strategies such as race-stratified validation, localized retraining, and transparency reporting. Each case underscores the critical role of equity, inclusive data, and oversight in ethical AI deployment in healthcare.

Table 4: Case Studies of Fair AI Deployment in Clinical Decision Support Systems

Case Study	Fairness Strategy	Outcome	Examples / Sources
Mount Sinai COVID-19 Risk Model	Race-stratified validation and adjustment for social determinants of health.	Enhanced equity in risk predictions without compromising model performance.	Wang et al. (2021)
IBM Watson for Oncology	Localized retraining and human-AI collaborative protocols after biased recommendations.	Improved generalizability and alignment with diverse clinical contexts.	Strickland (2019)
Model Cards by Google	Standardized documentation of model fairness, performance across demographics, and intended uses.	Increased transparency and better-informed deployment decisions in clinical settings.	Mitchell et al. (2019)

Furthermore, the "Model Cards" approach proposed by Google researchers has been applied in healthcare settings to increase transparency and fairness. These cards document performance metrics across demographic groups, data sources, and intended use cases, enabling stakeholders to better assess fairness and suitability before deployment (Mitchell et al., 2019). Collectively, these case studies highlight the importance of deliberate fairness-aware design, localized data representation, and transparent reporting as foundational principles for ethical AI in clinical environments.

3.5 Metrics and Benchmarks for Evaluating Fairness in Healthcare AI

The evaluation of fairness in healthcare AI systems requires the use of well-defined metrics and benchmarks that quantify disparities in performance across different population subgroups. Traditional model performance metrics—such as accuracy or area under the ROC curve—are insufficient for fairness analysis because they may mask systematic errors affecting marginalized groups. Instead, fairness-specific metrics such as demographic parity, equal opportunity, and equalized odds are employed to assess whether models produce equitable outcomes (Hardt et al., 2016). These metrics evaluate whether predictions are independent of protected attributes like race, gender, or age, or whether true and false positive rates are balanced across subgroups.

In healthcare contexts, these fairness metrics are increasingly integrated into model validation pipelines. For example, equalized odds ensures that the sensitivity and specificity of a model are consistent across demographic groups—a critical requirement for diagnostic tools where unequal error rates can have life-threatening consequences (Rajkomar et al., 2018). Moreover, healthcare institutions have begun to adopt subgroup-specific performance audits to benchmark algorithmic behavior across clinical sites, populations, and use cases, making fairness evaluation a routine component of AI governance (Chen et al., 2019).

Furthermore, open-source toolkits such as IBM's AI Fairness 360 and Google's What-If Tool provide standardized benchmarks and visualizations that facilitate bias detection and comparative analysis. These platforms are vital for ensuring that healthcare AI systems adhere to ethical principles, mitigate discrimination, and align with equity-focused regulatory expectations.

4. Security and Accountability in AI-Driven Healthcare

4.1 Security Challenges: Data Breaches, Adversarial Attacks, and Model Inversion

As artificial intelligence (AI) becomes increasingly integrated into healthcare infrastructure, the threat landscape for cybersecurity expands, presenting several critical challenges. Data breaches remain one of the most prevalent security issues, primarily due to the large volumes of sensitive patient information stored in electronic health records (EHRs) and AI-driven health information systems. These breaches can occur through unauthorized access, ransomware attacks, or misconfigured databases, often resulting in financial loss, patient harm, and erosion of public trust (Ben-Assuli, 2015).

Figure 6 shows four major types of adversarial attacks on artificial intelligence systems using realistic character-based scenarios. These include evasion attacks that mislead AI through subtle alterations, model inversion attacks that extract sensitive data, poisoning attacks that corrupt training datasets, and exploit attacks that manipulate AI biases for malicious outputs. The central cone and concentric rings symbolize AI's vulnerability landscape across different threat vectors.

Evasion Attacks
Subtle modifications deceive AI systems
Poisoning Attacks
Malicious code corruts AI training processes
Model Inversion Attacks
AI models are reverse-engineered to extract data
Exploit Attacks
AI biases are induced to produce harmful outputs

Figure 6: Adversarial Attacks on AI Systems

Another emerging concern is adversarial attacks, wherein malicious actors introduce subtle, often imperceptible, perturbations to input data—such as medical images or physiological signals—that lead AI models to make incorrect predictions. These attacks exploit the sensitivity of deep learning models and pose serious risks in diagnostic applications like radiology or dermatology, where even minor misclassifications can result in delayed or inappropriate treatment (Finlayson et al., 2019). The healthcare sector's limited expertise in AI-specific threat mitigation further compounds the vulnerability of such systems.

Model inversion attacks also represent a growing threat. In these attacks, adversaries exploit access to a trained model's outputs to reconstruct sensitive training data, including identifiable patient attributes. This threat undermines the confidentiality guarantees of machine learning applications and challenges compliance with privacy regulations such as HIPAA and GDPR (Fredrikson et al., 2015). These security challenges demand robust defense mechanisms—including encryption, secure multi-party computation, and adversarial training—to ensure the resilience of AI systems in clinical environments.

4.2 Privacy-Preserving AI Methods: Federated Learning, Differential Privacy, and Secure Computation

Privacy concerns are central to the ethical deployment of artificial intelligence (AI) in healthcare, where patient data is highly sensitive and protected by strict regulatory frameworks such as HIPAA and GDPR. Traditional centralized AI training methods often require pooling patient data into a single repository, which increases the risk of exposure in the event of a breach. To mitigate this, researchers have developed privacy-preserving AI techniques such as federated learning, differential privacy, and secure computation.

Federated learning (FL) enables AI models to be trained across decentralized devices or institutions without sharing raw data. Instead, only model updates (e.g., gradients or weights) are transmitted, which enhances data security and allows for collaborative learning across healthcare institutions while maintaining local data privacy (Yang et al., 2019). This approach is particularly valuable in medical imaging and cross-institutional studies, where data sharing is legally and ethically constrained.

Differential privacy (DP) introduces mathematically bounded noise into datasets or query results to obscure individual-level information, providing provable privacy guarantees even when data is accessed or analyzed. In healthcare AI, DP can be used to protect against re-identification attacks during model inference or training, particularly in sensitive applications such as genomics or mental health prediction (Dwork & Roth, 2014).

Table 5 summarizes key privacy-preserving AI techniques—federated learning, differential privacy, and secure computation—used in healthcare. These methods address regulatory and ethical concerns by minimizing data exposure during model training and inference. They are crucial for enabling safe, collaborative, and compliant AI innovation in sensitive clinical environments.

Table 5: Privacy-Preserving AI Methods in Healthcare

Method	Description	Use Case in Healthcare	Examples / Sources
Federated Learning (FL)	Trains models across decentralized data sources without transferring raw data; only model updates are shared.	Enables privacy-preserving collaboration across hospitals, especially in imaging and multi-institutional research.	Yang et al. (2019)
Differential Privacy (DP)	Adds mathematical noise to data or outputs to obscure individual-level information.	Protects patient identity during data analysis in areas like genomics and mental health.	Dwork & Roth (2014)
Secure Computation	Includes SMPC and homomorphic encryption to perform computations on encrypted data.	Allows collaborative AI analysis without revealing underlying clinical data.	Kaissis et al. (2020)

Secure multi-party computation (SMPC) and homomorphic encryption are also gaining traction in clinical research. These methods allow multiple parties to compute functions over encrypted data without revealing the underlying information, thereby preserving privacy while enabling complex AI workflows (Kaissis et al., 2020). Together, these technologies form a multilayered defense against privacy risks in AI-driven healthcare systems.

4.3 Explainability and Transparency for Accountable AI Decisions

In healthcare, the adoption of artificial intelligence (AI) hinges not only on model performance but also on explainability and transparency, which are essential for clinical accountability, regulatory compliance, and patient trust. Black-box models such as deep neural networks often deliver high accuracy but lack interpretability, posing challenges when medical professionals must justify treatment decisions or understand the rationale behind algorithmic outputs (Doshi-Velez & Kim, 2017).

To address these concerns, a range of explainable AI (XAI) techniques have been developed. Methods like SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) provide post hoc explanations by attributing model predictions to specific input features. These tools are especially valuable in clinical scenarios such as diagnostic imaging and risk prediction, where identifying contributing factors (e.g., symptoms, lab results) enhances clinician trust and supports evidence-based decision-making (Lundberg & Lee, 2017).

Beyond technical tools, transparency mechanisms like "model cards" and "datasheets for datasets" have emerged to document model design, training data characteristics, intended use cases, and performance across demographic subgroups. These frameworks promote transparency throughout the AI lifecycle and help prevent misuse or overgeneralization of models in unintended contexts (Mitchell et al., 2019). By embedding explainability and transparency into AI systems, developers and healthcare institutions can foster a culture of accountability that aligns with ethical standards and supports equitable patient care.

4.4 Model Governance, Audit Trails, and Documentation Frameworks

Robust model governance, comprehensive audit trails, and standardized documentation frameworks are essential for ensuring accountability and trustworthiness in AI-enabled healthcare systems. As AI applications influence high-stakes medical decisions, governing their lifecycle—from development to deployment and post-market surveillance—has become a top priority for healthcare organizations and regulatory bodies. Effective model governance includes defining roles, setting performance benchmarks, enforcing compliance, and managing risks related to model drift, retraining, and clinical integration (Sendak et al., 2020).

Audit trails play a vital role in establishing accountability by capturing detailed logs of model predictions, user interactions, and clinical outcomes. These records support retrospective analyses, help identify sources of error or bias, and provide documentation in cases of malpractice investigations or regulatory reviews (Kraska et al., 2019). Auditability ensures that every AI-generated decision can be traced back to its underlying inputs and algorithmic logic, enabling clinicians and institutions to respond to adverse outcomes with transparency.

In addition, structured documentation frameworks such as model cards and datasheets for datasets improve transparency by standardizing the reporting of AI models' development context, training data characteristics, evaluation metrics, and ethical considerations. These tools encourage responsible model deployment and promote alignment with institutional, legal, and ethical standards (Mitchell et al., 2019). Together, governance protocols, auditable systems, and structured documentation form the foundation for operationalizing responsible AI in healthcare settings.

4.5 Regulatory and Legal Mechanisms for AI Accountability in Healthcare

The increasing integration of artificial intelligence (AI) into healthcare has prompted the development of regulatory and legal frameworks designed to ensure algorithmic accountability, protect patient rights, and promote responsible innovation. At the core of these mechanisms is the requirement for transparency, fairness, and traceability, particularly in systems that influence diagnosis, treatment, and resource allocation.

In the United States, the Food and Drug Administration (FDA) has introduced a regulatory framework for Software as a Medical Device (SaMD), including AI/ML-based tools. The FDA's proposed "Total Product Lifecycle" (TPLC) approach emphasizes premarket evaluation, post-market monitoring, and adaptive learning control to manage risks while encouraging iterative model improvement (U.S. FDA, 2021). This framework promotes accountability by requiring developers to submit detailed algorithm change protocols and performance monitoring plans.

In the European Union, the General Data Protection Regulation (GDPR) mandates strict data governance and introduces a legal basis for algorithmic transparency through the "right to explanation." This ensures that patients can inquire into the rationale behind AI-driven decisions, especially when such decisions significantly impact their health outcomes (Goodman & Flaxman, 2017).

Additionally, emerging legislation such as the EU Artificial Intelligence Act classifies AI applications in healthcare as "high-risk," requiring conformity assessments, human oversight, and clear documentation of risk management strategies (Stix & Maas, 2021). Collectively, these mechanisms serve to standardize accountability practices, reinforce ethical AI deployment, and build public trust in healthcare technologies.

5. Future Directions and Integrated Frameworks

5.1 Emerging Trends: Decentralized Identity, CBDCs, and RegTech Synergies

Emerging technological trends are reshaping the landscape of accountable AI in healthcare, particularly through decentralized identity systems, central bank digital currencies (CBDCs), and the integration of regulatory technology (RegTech). These innovations are increasingly being aligned with healthcare AI frameworks to enhance transparency, improve patient control over data, and strengthen compliance infrastructures.

Decentralized identity (DID) models leverage blockchain and distributed ledger technologies to empower patients with ownership of their digital health identities. This framework mitigates reliance on centralized health data repositories, reducing the risk of identity theft and enabling secure, consent-driven data sharing in AI-driven health platforms. DID systems promote privacy by design and support granular access control, a critical feature for ensuring ethical AI implementation (Dahiya & Mathew, 2021).

While CBDCs are traditionally associated with financial systems, their programmable nature and integration with identity verification mechanisms offer potential synergies in public health administration, such as conditional healthcare subsidies or incentives. In AI-powered healthcare systems, CBDC-enabled smart contracts can facilitate secure, auditable transactions tied to patient services, reducing fraud and streamlining reimbursements (Auer et al., 2021).

Simultaneously, RegTech is gaining traction in healthcare by automating regulatory compliance through AI-enabled monitoring, real-time reporting, and anomaly detection. These tools support auditability and enhance the traceability of AI decisions, reinforcing legal accountability in clinical environments (Zetzsche et al., 2020). Together, these converging technologies provide a foundation for more secure, fair, and accountable AI adoption in the evolving digital healthcare ecosystem.

5.2 Strategic Guidelines for Designing Blockchain-Ready Compliance Systems

As artificial intelligence (AI) becomes embedded in healthcare systems, blockchain-ready compliance architectures are gaining traction for their potential to enhance transparency, immutability, and auditability in data management and regulatory oversight. To ensure that AI systems operate within ethical and legal boundaries, strategic design principles must integrate blockchain technologies with compliance workflows, especially in contexts involving sensitive patient data and automated decision-making.

One foundational guideline is the integration of smart contracts to automate regulatory compliance and access control. Smart contracts allow healthcare institutions to encode compliance rules into programmable logic, enabling real-time monitoring and enforcement of data usage policies without relying on intermediaries. This approach ensures adherence to consent management protocols and audit logging, thereby improving trust in AI decisions (Zhang et al., 2020).

Second, healthcare institutions should adopt interoperable and modular blockchain architectures that support integration with legacy systems, AI pipelines, and external regulatory databases. By using permissioned blockchain frameworks like Hyperledger Fabric, organizations can selectively share information with regulators, auditors, and patients while maintaining data confidentiality and control (Kuo et al., 2019).

Lastly, the incorporation of privacy-preserving techniques such as zero-knowledge proofs, secure multiparty computation, and differential privacy is essential for ensuring that blockchain-enabled systems comply with privacy laws like HIPAA and GDPR. These tools help reconcile the tension between data immutability and the right to be forgotten, making compliance systems robust, scalable, and legally sound (Angraal et al., 2017). Together, these strategic guidelines pave the way for the deployment of blockchain-aligned infrastructures that promote fairness and accountability in healthcare AI ecosystems.

5.3 Policy Recommendations for Multi-Jurisdictional Alignment

In the context of globalized healthcare delivery and AI adoption, ensuring multi-jurisdictional regulatory alignment is essential for promoting interoperability, safeguarding patient rights, and fostering ethical innovation. AI systems deployed across borders must comply with diverse legal frameworks—such as the General Data Protection Regulation (GDPR) in the European Union, the Health Insurance Portability and Accountability Act (HIPAA) in the United States, and emerging frameworks in Asia and Africa—each with unique provisions regarding data privacy, consent, and accountability.

One key recommendation is the harmonization of data protection standards across jurisdictions. International bodies and multilateral agreements should support baseline privacy and transparency criteria, similar to the OECD Privacy Guidelines, to reduce compliance fragmentation and legal uncertainty for AI developers (Taddeo

& Floridi, 2018). This includes mutual recognition of certification frameworks and third-party audits that validate AI systems against internationally accepted ethical principles.

Second, countries should establish cross-border regulatory sandboxes to pilot and evaluate AI systems under controlled conditions, encouraging safe innovation while informing transnational policy development. These sandboxes foster regulatory cooperation and data sharing between jurisdictions without compromising sovereignty or security (Zwitter et al., 2020).

Third, interoperable legal and technical standards—such as the use of standardized APIs, metadata schemas, and governance protocols—should be mandated to facilitate data portability and real-time compliance monitoring. Global health authorities, such as the World Health Organization (WHO), can play a coordinating role in defining such norms to ensure equitable access and algorithmic accountability in AI-powered healthcare systems worldwide (Leslie, 2020).

5.4 Research Directions in Privacy-Aware and Adaptive Compliance Models

The growing integration of artificial intelligence (AI) in healthcare has intensified the need for privacy-aware and adaptive compliance models that can dynamically respond to evolving legal, ethical, and technological landscapes. Traditional static compliance systems are increasingly inadequate for managing the complex data flows and algorithmic behaviors associated with modern AI applications. As such, future research must explore the development of systems capable of continuous learning, contextual adaptation, and real-time regulatory alignment.

One promising direction is the advancement of context-aware compliance frameworks, where AI systems can interpret the legal and ethical requirements relevant to specific jurisdictions, patient populations, or clinical scenarios. This approach relies on embedding regulatory logic into AI workflows, enabling dynamic rule-checking and policy enforcement based on the use context (Veale & Edwards, 2018). Such systems must incorporate updatable knowledge bases that reflect legislative changes and institutional policies.

Another critical research focus is on adaptive risk modeling, where AI compliance systems continuously monitor algorithmic behavior for drift, bias, or emergent vulnerabilities. This entails integrating AI auditing mechanisms that provide real-time alerts and generate compliance reports for stakeholders and regulators, facilitating proactive risk management (Wachter et al., 2017).

Additionally, privacy-preserving machine learning methods such as federated analytics, encrypted computation, and hybrid differential privacy models are being explored to ensure data protection while maintaining AI utility. These models must be tested in longitudinal studies to evaluate their resilience in large-scale, distributed healthcare environments (Rieke et al., 2020). These research pathways aim to establish a new generation of AI systems that are not only technically robust but also ethically aligned and regulatory-ready.

5.5 Summary of Key Findings and Recommendations

This review has highlighted the growing need for AI frameworks in healthcare that uphold the principles of fairness, security, and accountability. Across all stages of AI system development and deployment—from data acquisition to real-time inference—critical vulnerabilities and inequities persist unless intentional design, governance, and evaluation mechanisms are implemented. A synthesis of the findings reveals that ensuring equitable access and outcomes requires addressing structural biases in datasets, embedding explainability in model logic, and deploying secure, privacy-preserving infrastructures.

A key recommendation is to incorporate fairness-aware algorithms and perform subgroup-specific performance audits during development. These practices reduce the risk of perpetuating health disparities and ensure the equitable application of AI across diverse patient populations (Rajkomar et al., 2018). Furthermore, deploying privacy-enhancing technologies such as federated learning, differential privacy, and homomorphic encryption ensures compliance with regulations like GDPR and HIPAA while safeguarding patient autonomy and trust (Dwork & Roth, 2014; Rieke et al., 2020).

Finally, AI systems must be embedded within adaptive, transparent compliance frameworks that facilitate real-time auditability, stakeholder accountability, and cross-border regulatory alignment. Incorporating tools such as model cards, algorithmic impact assessments, and blockchain-enabled audit trails ensures traceability and ethical deployment. By institutionalizing these practices, healthcare systems can build resilient AI infrastructures that are responsive to evolving ethical, legal, and technological demands.

5.6 Final Thought

As artificial intelligence continues to transform healthcare, the urgency to build systems that are fair, secure, and accountable cannot be overstated. The integration of AI into clinical environments offers remarkable potential to enhance diagnostics, personalize treatment, and improve operational efficiency. However, these advancements must not come at the expense of ethical responsibility, patient trust, or regulatory compliance. Future AI systems must be designed with human-centered values at their core, ensuring that they serve not only technological innovation but also the broader goal of health equity and justice. By fostering interdisciplinary collaboration and embedding transparent governance mechanisms, the healthcare sector can confidently harness AI's benefits while safeguarding against its risks.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. *Proceedings of the 35th International Conference on Machine Learning*, 60–69. https://proceedings.mlr.press/v80/agarwal18a.html

Angraal, S., Krumholz, H. M., & Schulz, W. L. (2017). Blockchain technology: Applications in health care. *Circulation: Cardiovascular Quality and Outcomes, 10*(9), e003800. https://doi.org/10.1161/CIRCOUTCOMES.117.003800

Arksey, H., & O'Malley, L. (2005). Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology, 8*(1), 19–32. https://doi.org/10.1080/1364557032000119616

Auer, R., Cornelli, G., & Frost, J. (2021). Rise of the central bank digital currencies: Drivers, approaches and technologies. *BIS Working Papers, No. 880.* https://www.bis.org/publ/work880.pdf

Beutel, A., Chen, J., Zhao, Z., & Chi, E. H. (2019). Putting fairness principles into practice: Challenges, metrics, and improvements. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 453–459. https://doi.org/10.1145/3306618.3314281

Ben-Assuli, O. (2015). Electronic health records, adoption, quality of care, legal and privacy issues and their implementation in emergency departments. *Health Policy, 119*(3), 287–297. https://doi.org/10.1016/j.healthpol.2014.11.014

Chen, I. Y., Joshi, S., Ghassemi, M., & Liu, Y. (2019). Treating health disparities with artificial intelligence. *Nature Medicine, 25*(10), 1567–1569. https://doi.org/10.1038/s41591-019-0589-4

Dahiya, M., & Mathew, S. (2021). Decentralized digital identity systems: Opportunities and challenges. *Journal of Information Security Research, 12*(2), 45–53. https://doi.org/10.6025/jisr/2021/12/2/45-53

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal, 6*(2), 94–98. https://doi.org/10.7861/futurehosp.6-2-94

Dilsizian, S. E., & Siegel, E. L. (2014). Artificial intelligence in medicine and cardiac imaging. *Current Cardiology Reports, 16*(1), 441. https://doi.org/10.1007/s11886-013-0441-8

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*. https://doi.org/10.48550/arXiv.1702.08608

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science, 9*(3–4), 211–407. https://doi.org/10.1561/0400000042

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine, 25*(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. *Science, 363*(6433), 1287–1289. https://doi.org/10.1126/science.aaw4399

Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks. *Proceedings of the 22nd ACM SIGSAC Conference*, 1322–1333. https://doi.org/10.1145/2810103.2813677

Goodman, B., & Flaxman, S. (2017). EU regulations on algorithmic decision-making. *AI Magazine, 38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 3315–3323. https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

He, J., Baxter, S. L., Xu, J., Zhou, X., & Zhang, K. (2019). AI technologies in medicine. *Nature Medicine, 25*(1), 30–36. https://doi.org/10.1038/s41591-018-0307-0

Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining EHRs. *Nature Reviews Genetics, 13*(6), 395–405. https://doi.org/10.1038/nrg3208

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., ... & Wang, Y. (2017). AI in healthcare. *Stroke and Vascular Neurology, 2*(4), 230–243. https://doi.org/10.1136/svn-2017-000101

Kaissis, G., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure and federated learning in medical imaging. *Nature Machine Intelligence, 2*, 305–311. https://doi.org/10.1038/s42256-020-0186-1

Kraska, T., Beutel, A., Chi, E. H., Dean, J., & Polyzotis, N. (2019). ML lifecycle management. *ACM SIGMOD Conference*, 2155–2160. https://doi.org/10.1145/3299869.3324959

Kuo, T. T., Kim, H. E., & Ohno-Machado, L. (2019). Blockchain for healthcare. *JAMIA, 26*(6), 1211–1220. https://doi.org/10.1093/jamia/ocz065

Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., & Ferrante, E. (2020). Gender imbalance in imaging datasets. *PNAS, 117*(23), 12592–12594. https://doi.org/10.1073/pnas.1919012117

Leslie, D. (2020). Understanding AI ethics and safety. *The Alan Turing Institute*. https://doi.org/10.5281/zenodo.3240529

Levac, D., Colquhoun, H., & O'Brien, K. K. (2010). Advancing scoping review methodology. *Implementation Science, 5*, 69. https://doi.org/10.1186/1748-5908-5-69

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *NeurIPS*, 4765–4774. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Mehta, N., & Pandit, A. (2018). Big data analytics and healthcare. *IJMI, 114*, 57–65. https://doi.org/10.1016/j.ijmedinf.2018.03.013

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. *FAT*, 220–229. https://doi.org/10.1145/3287560.3287596

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). AI ethics tools and practices. *Science and Engineering Ethics, 26*, 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future — big data, ML, and clinical medicine. *NEJM, 375*(13), 1216–1219. https://doi.org/10.1056/NEJMp1606181

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in algorithms. *Science, 366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Panch, T., Mattie, H., & Celi, L. A. (2019). The "inconvenient truth" about AI in healthcare. *NPJ Digital Medicine, 2*, 77. https://doi.org/10.1038/s41746-019-0155-4

Peters, M. D., Godfrey, C. M., McInerney, P., Munn, Z., Tricco, A. C., & Khalil, H. (2020). Chapter 11: Scoping reviews (2020 version). *JBI Manual for Evidence Synthesis*. https://doi.org/10.46658/JBIMES-20-12

Price, W. N., & Cohen, I. G. (2019). Privacy in medical big data. *Nature Medicine, 25*(1), 37–43. https://doi.org/10.1038/s41591-018-0272-7

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare. *Health Information Science and Systems, 2*(1), 3. https://doi.org/10.1186/2047-2501-2-3

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *NEJM, 380*(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in ML. *Annals of Internal Medicine, 169*(12), 866–872. https://doi.org/10.7326/M18-1990

Raji, I. D., & Buolamwini, J. (2019). Actionable auditing of biased AI. *AIES Conference*, 429–435. https://doi.org/10.1145/3306618.3314244

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. *NPJ Digital Medicine, 3*, 119. https://doi.org/10.1038/s41746-020-00323-1

Sendak, M. P., Ratliff, W., Sarro, D., Alderton, E., Futoma, J., Gao, M., ... & Balu, S. (2020). Sepsis deep learning in clinical care. *JMIR Medical Informatics, 8*(7), e15182. https://doi.org/10.2196/15182

Shen, J., Zhang, C. J. P., Jiang, B., Chen, J., Song, J., Liu, Z., & He, Z. (2021). AI vs. clinicians in diagnosis.

JMIR Medical Informatics, 9(3), e24098. https://doi.org/10.2196/24098

Stix, C., & Maas, M. M. (2021). Algorithmic Accountability Act in the EU. *Philosophy & Technology, 34*, 949–971. https://doi.org/10.1007/s13347-021-00441-8

Strickland, E. (2019). IBM Watson in healthcare. *IEEE Spectrum, 56*(4), 24–31. https://doi.org/10.1109/MSPEC.2019.8674535

Taddeo, M., & Floridi, L. (2018). AI as a force for good. *Science, 361*(6404), 751–752. https://doi.org/10.1126/science.aat5991

Topol, E. J. (2019). High-performance medicine. *Nature Medicine, 25*, 44–56. https://doi.org/10.1038/s41591-018-0300-7

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., ... & Straus, S. E. (2018). PRISMA extension for scoping reviews. *Annals of Internal Medicine, 169*(7), 467–473. https://doi.org/10.7326/M18-0850

U.S. Food and Drug Administration (FDA). (2021). *AI/ML-Based Software as a Medical Device Action Plan.* https://www.fda.gov/media/145022/download

Veale, M., & Edwards, L. (2018). GDPR and automated decision-making. *Computer Law & Security Review, 34*(2), 398–404. https://doi.org/10.1016/j.clsr.2017.12.002

Vokinger, K. N., Feuerriegel, S., & Kesselheim, A. S. (2021). Mitigating bias in ML. *Communications Medicine, 1*, 25. https://doi.org/10.1038/s43856-021-00028-8

Wang, F., Casalino, L. P., Khullar, D., & Shah, N. H. (2021). Ensuring fairness in ML. *Annals of Internal Medicine, 174*(4), 580–585. https://doi.org/10.7326/M20-6076

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., ... & Goldenberg, A. (2019). Do no harm: Responsible ML. *Nature Medicine, 25*(9), 1337–1340. https://doi.org/10.1038/s41591-019-0548-6

Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning. *ACM Transactions on Intelligent Systems and Technology, 10*(2), 1–19. https://doi.org/10.1145/3298981

Zhang, P., White, J., Schmidt, D. C., Lenz, G., & Rosenbloom, S. T. (2020). FHIRChain for clinical data sharing. *CIN: Computers, Informatics, Nursing, 38*(7), 340–348. https://doi.org/10.1097/CIN.0000000000000586

Zetzsche, D. A., Buckley, R. P., Arner, D. W., & Barberis, J. N. (2020). Regulating LIBRA. *UNSW Law Research Series, 44*. https://doi.org/10.2139/ssrn.3414401

Zou, J. Y., & Schiebinger, L. (2018). AI can be sexist and racist—make it fair. *Nature, 559*(7714), 324–326. https://doi.org/10.1038/d41586-018-05707-8

Zwitter, A., Gstrein, O. J., & Yap, E. (2020). Digital identity and blockchain. *Frontiers in Blockchain, 3*, 26. https://doi.org/10.3389/fbloc.2020.00026