Review

Explainable Artificial Intelligence in Precision Medicine: Methods, Applications, and Challenges

Umakant Singh * and Punit Kumar Chaubey

Posted Date: 29 May 2026

doi: 10.20944/preprints202605.2039.v1

Keywords: explainable AI; interpretability; precision medicine; clinical decision support; SHAP; LIME; counterfactual explanations and model transparency

 doi:10.20944/preprints202605.2039.v1

Review

Explainable Artificial Intelligence in Precision Medicine: Methods, Applications and Challenges

Umakant Singh [1,*] and Punit Kumar Chaubey [2]

[1] Computer Science & Engineering, United University, Prayagraj, Uttar Pradesh, 211012, India

[2] Computer Science & Engineering, Bansal Institute of Engineering and Technology, Lucknow, Uttar Pradesh, 226021, India

* Correspondence: umakant@uniteduniversity.edu.in or umakantsinghsrm@gmail.com

Abstract

Precision medicine focuses on customizing diagnostic, prevention and treatment approaches by accounting for the individual characteristics of each patient. This personalization draws on diverse sources of information including clinical records, genomic data, medical imaging, lifestyle patterns and environmental factors. As the volume and complexity of such multimodal healthcare data continue to expand, machine learning (ML) and deep learning (DL) techniques have become crucial for identifying complex patterns, estimating disease risk, and supporting personalized treatment decisions. Despite their efficiency, many of these models function as opaque systems, generating forecasts without clearly indicating the reasoning behind them. This lack of transparency can undermine clinician confidence, hinder adoption in clinical practice, and raise ethical as well as regulatory concerns, particularly in healthcare contexts where decisions must be explainable and defensible. Explainable Artificial Intelligence (XAI) addresses these challenges by providing methods that make model behaviour more transparent and interpretable. Techniques such as SHAP, LIME, saliency and attention-based visualizations, counterfactual analysis, and rule-based explanations enable clinicians to inspect the rationale behind predictions, evaluate alignment with established medical knowledge, and identify potential sources of bias within data or algorithms. From a patient perspective, explain-ability improves communication, supports informed consent, and strengthens trust in AI-supported care. Regulatory authorities also depend on transparent and interpretable systems to ensure accountability, traceability and compliance with clinical safety requirements. This paper offers a comprehensive examination of explainable AI in the context of precision medicine. It introduces fundamental XAI concepts, organizes key methodological approaches, and reviews applications spanning genomics, medical imaging, and electronic health record (EHR) analytics. The chapter also discusses methods for assessing explanation quality, highlights the role of human-centred design, and addresses critical ethical and legal considerations. It concludes by outlining ongoing challenges and future research directions aimed at developing reliable, interpretable AI systems that can be effectively integrated into advanced personalized healthcare.

Keywords: explainable AI; interpretability; precision medicine; clinical decision support; SHAP; LIME; counterfactual explanations and model transparency

1. Introduction

Precision medicine marks a shift in healthcare from generalized treatment protocols to approaches designed around the individual patient. Rather than relying solely on population averages, it draws upon detailed information such as genetic profiles, diagnostic images, laboratory findings, wearable health data, lifestyle habits, environmental influences, and long-term medical records. While this wealth of information enables more personalized care, it also creates analytical challenges that conventional statistical techniques are often unable to address due to the complexity and non-linear nature of the data.

 doi:10.20944/preprints202605.2039.v1

Machine learning has therefore become an essential tool in modern precision medicine. Established methods such as support vector machines and ensemble learning algorithms continue to perform well for structured clinical datasets, while newer deep learning models have shown strong capabilities in image analysis, longitudinal health monitoring, and data integration. These approaches have been applied successfully to tasks including disease prediction, treatment planning, and outcome forecasting. However, the growing reliance on complex models has highlighted an important limitation: many of these systems offer little insight into how their predictions are generated.

In medical decision-making, transparency is not optional. Clinicians and patients must be able to understand the rationale behind AI-assisted recommendations to ensure safety, fairness, and trust. Explainable Artificial Intelligence seeks to address this need by providing clear, human-understandable explanations of model behaviour. By making AI systems more transparent and accountable, explainability supports their ethical and effective use in precision medicine.

1.1. Background: Interpretability vs. Explainability

Interpretability describes how easily a human can understand the reasoning of a model by directly examining its structure, without relying on external explanation tools. Models such as linear and logistic regression, simple decision trees, generalized additive models (GAMs), and sparse rule-based systems are commonly regarded as interpretable because their logic is transparent and their parameters have clear meanings. This clarity allows clinicians to follow how specific input variables contribute to a prediction and to judge whether the model's conclusions align with established medical knowledge and clinical experience.

Explainability, on the other hand, focuses on making the predictions of complex models understandable when their internal workings are not inherently transparent. Modern high-capacity models—including deep neural networks, ensemble learning methods, and other sophisticated algorithms—often rely on internal representations that are difficult for humans to interpret directly. To bridge this gap, post-hoc explanation techniques such as SHAP, LIME, integrated gradients, saliency maps, prototype-based explanations, and counterfactual analyses are widely applied. These approaches do not modify the original model; instead, they offer approximate or descriptive insights that help users understand why a particular decision or prediction was produced.

In practical healthcare settings, the distinction between interpretability and explainability is not always clear-cut, and both concepts play important roles. Interpretable models are often preferred for high-stakes clinical decisions because their decision logic is visible and easier to validate, reducing the risk of hidden errors or misleading conclusions. However, many biomedical domains—such as genomics, digital pathology, and other forms of high-dimensional medical data—require the expressive power of complex models to achieve acceptable predictive performance. In such cases, post-hoc explanation methods become essential for gaining insight into model behaviour.

Recent research has also emphasized the need for caution, as explanations that appear intuitive may not faithfully reflect how a model actually operates, potentially creating a false sense of trust. Therefore, the choice between inherently interpretable models and post-hoc explainability techniques should be guided by a careful balance among predictive performance, transparency, clinical reliability, and patient safety.

2. Literature Review

Artificial intelligence (AI) and machine learning (ML) have become influential forces in contemporary healthcare, enabling advanced methods for disease diagnosis, outcome prediction, clinical decision support, and personalized treatment planning. Early academic discussions, particularly the work of Evans [1], drew attention to the regulatory, safety, and governance issues surrounding clinical decision support systems, stressing that technological progress must be accompanied by appropriate oversight and accountability. As the field matured, more focused reviews began to document the growing presence of ML-based tools in everyday clinical practice.

 doi:10.20944/preprints202605.2039.v1

For example, Adadi [2] examined applications in gastroenterology, while Shin [3] reviewed developments across digital health, highlighting the gradual but steady integration of AI across multiple medical specialties.

In subsequent studies, increasing emphasis has been placed on transparency and clinician trust. Research by Ahirwar [4] and Bharati [21] suggests that clinical acceptance of AI systems depends not only on predictive accuracy but also on the ability of these systems to provide understandable and interpretable reasoning. Similar concerns are echoed in studies of medical imaging by Coppola [5] and Hong [22], which acknowledge the strong diagnostic capabilities of AI while also identifying unresolved issues related to robustness, generalizability, and ethical responsibility when models are deployed in real-world clinical settings.

More recent literature has expanded the discussion to emerging concepts such as digital twins. Reviews by Wickramasinghe [6] and Iqbal [10] describe how virtual representations of patients can be used to support precision treatment planning, particularly in areas such as oncology and the long-term management of chronic diseases. At the same time, the rapid growth of mobile health applications, remote patient monitoring systems, and population-level screening programs—examined by Bhatt [7] and Vorisek [26]—illustrates the widening role of AI in preventive care and community health. Despite these advances, the literature consistently notes challenges related to data quality, sustained patient engagement, and the risk of demographic bias, underscoring the need for careful design and evaluation of AI-driven healthcare solutions.

Precision medicine has emerged as a key area in which explainable and multimodal artificial intelligence approaches offer substantial benefits. Several comprehensive reviews, including those by Kline [13], Laccourreye [14], and Gerussi [9], describe how the integration of heterogeneous data sources—such as genomic profiles, medical imaging, and electronic health records—can support more personalized diagnostic assessments and treatment decisions. Complementing these broad perspectives, domain-specific studies have examined the impact of AI within individual medical specialties. For instance, Roy [15] focuses on diagnostic imaging, Shazly [16] explores applications in obstetrics and gynecology, and Wellnhofer [17] reviews advances in cardiovascular medicine, collectively demonstrating the wide-ranging clinical relevance of AI-driven methods.

Alongside these technical developments, growing attention has been directed toward the ethical, social, and fairness-related implications of AI adoption in healthcare. Baumgartner [20] addresses concerns surrounding equitable access to AI-enabled care and the fair distribution of healthcare resources, while Albahri [19] provides a detailed analysis of bias, trust, and transparency challenges associated with explainable AI systems. From a broader perspective, studies by Khanna [12] and Narayan [25] emphasize that successful healthcare AI solutions must not only be accurate but also scalable, economically viable, and sustainable in real-world settings.

More recent survey articles by Zafar [27] and Kuwaiti [24] highlight rapid advancements in deep learning and intelligent systems for genomics, biomedical imaging, and precision therapeutics. Collectively, this body of literature illustrates the expanding role of AI across the entire healthcare continuum—from early detection and diagnosis to treatment planning and long-term disease management—while consistently underscoring the importance of transparency, ethical responsibility, and trustworthy deployment.

Table 1. Literature Review Summary.

Ref.	Focus Area	Brief Contribution	Identified Research Gap
[1]	Regulation of clinical decision-support software	Examines the regulatory landscape for AI-driven clinical decision tools and highlights policy challenges.	Does not assess how existing regulations affect AI deployment and performance in real clinical settings.
[2]	AI applications in gastroenterology	Summarizes machine-learning uses and	Provides no standardized datasets or benchmarking

 doi:10.20944/preprints202605.2039.v1

		challenges within gastrointestinal medicine.	framework for evaluating model performance.
[3]	Digital health developments	Outlines national progress and priorities in digital health technologies.	Lacks empirical evaluation of model accuracy, usability, and patient-centered impacts.
[4]	Interpretable machine learning	Reviews transparent ML models and their relevance to healthcare.	Limited exploration of how clinicians interpret or trust the explanations generated by XAI systems.
[5]	Adoption of AI in medical imaging	Critically discusses AI's potential and challenges in imaging workflows.	Does not include quantitative assessments of proposed imaging AI approaches.
[6]	Digital twins in oncology	Proposes using digital twin models to support personalized cancer treatment.	Clinical implementation, validation, and scalability aspects remain untested.
[7]	AI in mobile health (mHealth)	Provides a broad overview of AI-enhanced mobile health applications.	Interoperability issues and data privacy concerns are not deeply analyzed.
[8]	Dementia prediction using ML	Utilizes interpretable ML models to anticipate dementia progression.	Study relies on small datasets; uncertainty remains regarding applicability to diverse populations.
[9]	AI in autoimmune liver disease	Reviews precision-medicine applications of AI for liver disorders.	Does not include comparative evaluation of ML models or performance metrics.
[10]	Ethical concerns in digital twins	Discusses privacy, responsibility, and ethical implications of digital twin technology.	Offers limited practical guidance or technical safeguards to address ethical issues.
[11]	Ethics in AI-driven digital health	Surveys key ethical risks associated with AI use in digital health systems.	No actionable framework is presented for implementing ethical standards in clinical environments.
[12]	Economics of healthcare AI	Evaluates cost-effectiveness of diagnostic and therapeutic AI tools.	Conclusions are theoretical; lacks real-world economic analyses or case studies.
[13]	Multimodal ML for precision health	Reviews methods for integrating diverse biomedical data sources.	Data fusion, missing modalities, and harmonization challenges remain insufficiently addressed.
[14]	Explainable ML for microbiome studies	Applies XAI approaches to longitudinal microbiome and multi-omic datasets.	Findings are domain-specific and not validated across broader biomedical applications.
[15]	ML and Healthcare 4.0	Surveys supervised machine learning developments in diagnostic medicine.	Does not investigate model stability or reliability in real-time clinical use.
[16]	ML in obstetrics and gynecology	Provides an overview of ML applications in maternal health.	Highlights need for curated datasets and standardized clinical AI workflows.
[17]	AI in cardiovascular imaging	Reviews regulatory, clinical, and operational	Lacks evaluation of real-time implementation and clinician–AI interaction.

 doi:10.20944/preprints202605.2039.v1

		aspects of cardiovascular imaging AI.	
[18]	Explainable ML for EHR-based prediction	Presents XAI approaches for cardiovascular risk prediction using EHR data.	Not validated across institutions or heterogeneous patient populations.
[19]	Trustworthy and explainable AI	Systematic overview of trust, transparency, and safety principles in XAI.	No unified framework for assessing trustworthiness.
[20]	Fairness in biomedical AI	Explores equitable adoption of AI tools and fairness considerations.	Lacks healthcare-specific fairness metrics and evaluation strategies.
[21]	Explainability in healthcare AI	Reviews XAI approaches and challenges in clinical contexts.	Does not assess performance risks in high-stakes decisions.
[22]	AI in radiology	Identifies pitfalls in supervised ML imaging systems.	Recommendations remain theoretical without validation.
[23]	AI adoption in pathology	Discusses conceptual frameworks for AI uptake in pathology.	No empirically tested strategy for resource-limited settings.
[24]	AI innovations in healthcare	Reviews recent AI advancements and trends.	Too general; lacks domain-specific performance insights.
[25]	AI strategy for diabetes control	Proposes an AI roadmap for diabetes management in India.	No computational validation of the proposed framework.
[26]	Bias in healthcare AI	Surveys bias sources and user perceptions.	Does not propose technical mitigation solutions.
[27]	Deep learning in genomics	Reviews DL techniques for genomics and biomedical AI.	Scalability and large-dataset integration remain unresolved.

3. Why Explainability Matters in Precision Medicine

Explainability is a core requirement for the safe, ethical, and effective use of artificial intelligence in precision medicine. In contrast to consumer-facing applications, where mistakes may be inconvenient but largely harmless, errors in medical AI can have serious consequences for patients, healthcare providers, and regulatory compliance. As a result, it is not sufficient for an AI system to produce accurate predictions alone; clinicians must also be able to understand the reasoning behind those predictions. This need for insight into model behaviour underscores the critical importance of explainability in healthcare, as it supports informed clinical judgment, accountability, and responsible decision-making.

Clinical Trust and Adoption

Clinical practice is grounded in careful evaluation of evidence, adherence to established guidelines, and professional judgment developed through experience. When an AI system offers a prediction without revealing how that conclusion was reached, it can clash with the reasoning processes clinicians depend on. Explainable models enable healthcare professionals to assess whether an AI system's logic aligns with medical knowledge, identify results that appear inconsistent with clinical understanding, and determine when a recommendation merits confidence or further scrutiny. In the absence of such transparency, even models with strong predictive performance may face distrust or be used in ways that compromise patient care.

Accountability and Auditability

 doi:10.20944/preprints202605.2039.v1

Medical decision-making is inherently tied to ethical duties, legal accountability, and regulatory compliance. When an outcome influenced by an AI system leads to patient harm or an adverse event, it becomes essential to understand the role the system played in that decision. Explainable models make this possible by exposing the reasoning process behind an output and providing a clear record of how and why a specific recommendation was produced. Such transparency supports systematic auditing and aligns with emerging regulatory expectations, including guidance from the FDA, provisions of the EU AI Act, and international clinical safety standards that emphasize traceability and justification in AI-supported care. Moreover, well-articulated explanations help distinguish whether an error arose from limitations in the data, shortcomings of the model, or inappropriate use by end users, thereby supporting accountability and corrective action.

Identifying Bias and Failure Modes

Healthcare datasets frequently reflect underlying imbalances or subtle correlations that can unintentionally influence how AI systems learn and make decisions. When left unchecked, these biases may lead to uneven diagnostic performance or unequal treatment recommendations across different patient populations. Explainability techniques help uncover which factors most strongly influence a model's predictions, allowing both developers and clinicians to identify cases where the system relies on spurious signals, institutional practices, or population-specific artifacts rather than clinically meaningful information. Detecting such issues at an early stage is essential for enhancing fairness, improving generalization across diverse groups, and avoiding the introduction of harmful disparities in precision medicine.

Patient Communication and Shared Decision-Making

As precision medicine moves toward more individualized care, patients are playing a growing role in decisions about their own treatment. When AI systems contribute to choices such as therapy selection, risk evaluation, or genetic analysis, patients have a legitimate expectation to understand how those recommendations were formed. Explainable AI enables clinicians to convey complex model outputs in clear, accessible terms, supporting informed consent and building confidence in the care process. By making AI-driven insights more transparent, patients are better equipped to ask informed questions, participate actively in discussions, and make decisions that reflect their personal values and health priorities.

Scientific Discovery and Biomedical Insight

Explainability is not limited to supporting clinical decisions; it also plays an important role in advancing scientific research. By revealing which features, interactions, or biological markers most strongly influence model outputs, explainable AI techniques can point researchers toward new hypotheses or uncover disease mechanisms that were previously overlooked. This capability is especially valuable in fields such as genomics, proteomics, radiomics, and systems biology, where models must analyze extremely large and complex datasets. Insights gained from interpretable results help separate biologically meaningful patterns from background noise, aiding the identification of potential therapeutic targets and contributing to deeper understanding of underlying biological processes.

4. Taxonomy of Xai Methods

Explainable Artificial Intelligence (XAI) encompasses a broad set of techniques designed to make the decisions and behaviour of machine-learning systems accessible to human understanding. While certain models are inherently transparent and allow their reasoning to be examined directly, others rely on supplementary methods to clarify how complex and opaque architectures arrive at their outputs. Classifying these approaches into a structured taxonomy helps clarify the role each method plays in supporting interpretability, particularly within the context of precision medicine.

 doi:10.20944/preprints202605.2039.v1

4.1. Intrinsically Interpretable Models

Models in this group are designed with transparency in mind, allowing their internal logic to be inspected without additional explanatory tools. The rules they follow, the parameters they use, or the examples they rely on are openly accessible, which makes validation and critical review more straightforward. Although these models may not always match the performance of deep neural networks on highly complex or high-dimensional problems, their clear and understandable structure makes them especially well-suited to clinical settings, where trust, consistency, and interpretability are essential.

- Linear Models (supported by careful feature design and sparsity)

Linear and logistic regression models, including their sparse and regularized forms, remain widely used in medical analytics. Their strength comes from the clear and direct way in which each input variable relates to the predicted outcome. When paired with carefully designed feature engineering—such as clinically informed transformations, interaction terms, or regularization strategies—these models can deliver transparent risk estimates and clearly indicate which factors have the greatest influence on the final prediction.

- Generalized Additive Models (GAMs)

Generalized Additive Models maintain the transparent structure of linear models while extending their capability to model smooth, nonlinear relationships for each individual feature. The influence of each variable can usually be visualized clearly, enabling clinicians to easily interpret how specific inputs affect the final prediction. Recent extensions, such as Explainable Boosting Machines, are designed to preserve this interpretability while delivering improved predictive performance.

- Decision Trees and Rule-Based Approaches

Decision trees express their reasoning as a series of clear if–then conditions, allowing users to trace the exact sequence of steps that leads to a particular prediction. Related approaches, such as rule-based models including RuleFit and Bayesian Rule Lists, convey decisions through compact collections of rules that are easy for humans to read and understand. One drawback of decision trees is their susceptibility to minor changes in the training data, which can lead to noticeably different tree structures when the model is retrained.

- Prototype and Case-Based Models

Case-based reasoning approaches justify their predictions by referencing past examples that closely resemble the current case. Traditional methods such as k-nearest neighbours follow this idea, as do newer models based on learned prototypes. This style of explanation aligns naturally with clinical practice, where physicians frequently draw on similarities to prior patient cases when making diagnostic or treatment decisions.

4.2. Post-Hoc Explainers (Model-Agnostic and Model-Specific)

Post-hoc explanation methods are designed to shed light on how an already trained model—especially one that functions as a black box—produces its predictions. These techniques can be applied either to gain an overall understanding of a model's behaviour or to clarify the reasoning behind a single, case-specific output. As such, they are essential for interpreting complex models commonly employed in precision medicine.

Feature Attribution Approaches

- **SHAP (Shapley Additive Explanations)**

SHAP is based on Shapley values from cooperative game theory and assigns a quantitative contribution to each input feature in a model's prediction. The resulting explanations are additive and internally consistent, allowing them to be used for both individual case interpretation and broader, population-level analysis. Because of these properties, SHAP is widely adopted in

 doi:10.20944/preprints202605.2039.v1

applications such as clinical risk assessment, genomic analysis, and other forms of structured biomedical data.

- **LIME (Local Interpretable Model-Agnostic Explanations)**

LIME focuses on explaining single predictions by modeling the behaviour of a complex system in the local neighborhood of a specific input. It fits a simple, human-readable surrogate model—commonly a sparse linear approximation—to identify the features that most strongly influence that particular outcome. This localized perspective helps clinicians understand why predictions may differ from one patient to another.

- **Integrated Gradients and Saliency Techniques**

Gradient-based explanation techniques are commonly applied to deep learning models, especially in imaging domains such as radiology and pathology. These methods emphasize parts of the input—such as image regions or sequence elements—that contribute most to the final prediction. The resulting visualizations can offer intuitive insight into how convolutional and other differentiable models recognize patterns in complex biomedical data.

Global Interpreters and Visualization Methods

- **Partial Dependence Plots (PDPs) and Accumulated Local Effects (ALE)**

Partial Dependence Plots show how variations in a single feature influence a model's predictions when averaged across the dataset. However, because PDPs assume feature independence, they can be misleading when inputs are correlated—a common situation in clinical data such as laboratory values or physiological measurements. Accumulated Local Effects plots address this limitation by accounting for feature dependencies, providing a more trustworthy representation of feature influence in real-world healthcare datasets.

- **Surrogate Models**

Another way to interpret complex systems is to approximate their behaviour using a simpler, transparent model, such as a small decision tree or a shallow neural network. Although these surrogate models cannot fully reproduce the original model's complexity, they can reveal overarching decision patterns that help clinicians and researchers grasp the general logic guiding a black-box system.

Causal and Rule-Driven

- **Causal Graphs and Rule Generation Techniques**

Causal techniques use structured causal models to describe relationships in a way that aligns with clinical reasoning, offering insights that extend beyond simple statistical correlations. Rule-based methods, in contrast, translate model behaviour into explicit logical rules that domain experts can examine and validate. When applied thoughtfully, both approaches can produce explanations that more closely reflect underlying biological processes rather than merely highlighting associations.

4.3. Human-Centered Explanations

Human-centered explanation strategies recognize that interpretability is not purely a technical issue, but also one of communication, cognition, and practical usability. These methods prioritize presenting explanations in formats that clinicians can readily understand and apply in real clinical environments, placing emphasis on clarity and interaction rather than mathematical detail.

- **Natural Language Explanations**

Natural language approaches translate a model's reasoning into clear, narrative descriptions that resemble everyday clinical communication. For example, instead of presenting abstract numerical scores, a system might state that a patient is considered at elevated cardiovascular risk due

 doi:10.20944/preprints202605.2039.v1

to high LDL cholesterol levels and a documented family history of heart disease. Such summaries enable clinicians to quickly understand the basis of a prediction.

- **Visual Explanations**

Visual explanation tools are particularly effective for AI systems that process medical images. Techniques such as heatmaps, attention maps, or saliency overlays can identify regions of a CT scan, MRI, or pathology slide that most influenced the model's output. For EHR-based models, visual timelines or trend charts can show how changes in laboratory results, vital signs, or medication use contributed to the predicted outcome, helping bridge numerical outputs and clinical interpretation.

- **Interactive Explanations**

Interactive explanation interfaces allow clinicians to actively explore a model's behaviour. Features such as adjustable inputs, scenario simulations, and what-if analysis dashboards enable users to modify variables and observe how predictions change in response. This interactive exploration supports a deeper understanding of model sensitivity, highlights key risk drivers, and helps clinicians assess whether the model behaves in a clinically reasonable manner across different patient scenarios.

5. Applying Xai in Precision Medicine Areas

5.1. Genomics and Molecular Diagnostics

AI plays a central role in genomic and molecular research, supporting tasks such as detecting disease-causing genetic variants, categorizing tumor types, and identifying biomarkers linked to drug sensitivity. These applications involve highly complex, high-dimensional datasets with intricate biological relationships and significant variation across populations. Explainable AI techniques help address these challenges by identifying the genes or variants that have the greatest impact on model predictions, summarizing results at the pathway level to produce biologically meaningful interpretations, and applying counterfactual analysis to examine how changes at the molecular level could influence predicted outcomes.

5.2. Medical Imaging

In radiology and pathology, AI systems are widely used for functions such as identifying abnormalities, segmenting anatomical structures, staging disease, and assisting with treatment planning. Explainability in these models is often achieved through visual approaches, including Grad-CAM, integrated gradients, and other saliency-based methods that highlight image regions most relevant to a prediction. More advanced techniques, such as Testing with Concept Activation Vectors (TCAV), connect internal network representations to clinically interpretable concepts like shape, texture, or tissue patterns. Multimodal explanation strategies further enhance understanding by combining image-based visualizations with structured clinical information. However, saliency-based explanations must be interpreted cautiously, as they can be sensitive to noise and may be misleading if not supported by quantitative validation.

5.3. Electronic Health Records (EHR) and Clinical Risk Prediction

Models built on electronic health records are commonly used for clinical risk prediction, including estimating hospital readmission risk, early detection of sepsis, and medication recommendation. Explainability in this context often focuses on temporal dynamics, illustrating how patient data change over time and identifying the periods that most strongly influence predictions. Rule-based approaches and extracted decision logic can further enhance transparency. Counterfactual explanations are particularly useful, as they allow clinicians to explore how alternative treatments or changes in patient condition might affect predicted risks. Given the fast-paced nature of clinical environments, concise explanations that emphasize key drivers, convey uncertainty, and offer actionable insights are generally preferred.

 doi:10.20944/preprints202605.2039.v1

5.4. Drug Discovery and Treatment Recommendation

AI is increasingly used in drug discovery to predict interactions between compounds and biological targets, as well as to estimate patient-specific responses to therapies. Explainability approaches in this area often make use of attention mechanisms that highlight molecular substructures or biological features most relevant to a prediction. Model-agnostic explanation tools also assist researchers in identifying signals associated with treatment effectiveness. In addition, causal inference methods are applied to estimate treatment effects, supporting more informed and personalized therapy recommendations.

6. Evaluating Dimensions

Assessing explainable AI systems requires consideration of multiple dimensions that determine how effectively explanations support clinical reasoning and safe medical decision-making. These dimensions help ensure that an XAI approach not only makes model behaviour more transparent but also delivers real, practical benefits within healthcare settings.

- **Fidelity / Faithfulness**

Fidelity measures how closely an explanation represents the model's actual decision process. High-fidelity explanations accurately reflect the internal logic used to generate predictions. Methods such as input perturbation or feature ablation are commonly employed to verify whether an explanation truly corresponds to the model's reasoning.

- **Stability / Robustness**

An explanation should remain consistent when inputs change only marginally. If patients with nearly identical profiles receive very different explanations, the explanation method may lack reliability. Robustness is particularly important in clinical applications, where small data variations are unavoidable.

- **Usefulness / Actionability**

Explanations should meaningfully assist clinicians by improving understanding, supporting quicker decisions, or increasing confidence in model outputs. Evaluations often rely on user studies, workflow analyses, or decision-support performance measures to determine whether explanations contribute positively to clinical practice.

- **Simplicity / Parsimony**

Overly complex explanations can confuse rather than clarify. Clear, concise, and focused explanations are more likely to be accepted and trusted by clinicians. Parsimonious explanations emphasize the most influential factors without overwhelming users with unnecessary detail.

- **Clinical Relevance / Correctness**

Explanations must be medically sound and consistent with established clinical knowledge, validated biomarkers, or plausible physiological processes. Expert review and retrospective case analyses are commonly used to assess whether explanations align with real-world biomedical understanding.

- **Computational Efficiency**

For AI systems to function effectively in routine care, explanations must be produced within reasonable time limits. Applications that operate in real-time or near real-time require efficient XAI techniques that integrate smoothly into clinical workflows without causing delays.

6.1. Common Approaches to Evaluating XAI

- **Synthetic Benchmarks**

 doi:10.20944/preprints202605.2039.v1

These evaluations rely on artificially constructed datasets in which the true importance of each feature is known beforehand. This allows researchers to assess how accurately an explanation method is able to recover the underlying decision logic of the model.

- **Human-in-the-Loop Assessments**

In this approach, clinicians or domain experts examine the generated explanations and assess them in terms of clarity, usefulness, and clinical credibility. The emphasis here is on practical relevance and interpretability in real clinical settings, rather than solely on technical correctness.

- **Comparison Against Clinical Biomarkers or Covariates**

Explanations are compared with well-known medical risk factors, biomarkers, or clinically validated covariates to evaluate whether the model highlights features that are consistent with existing biological and clinical understanding.

7. Ethical, Legal, and Regulatory Considerations

- **Bias and Fairness**

Explainable AI can be used to identify whether a model depends on variables that may disadvantage particular demographic or socioeconomic groups. However, if explanations are interpreted superficially, they may hide deeper structural biases rather than expose them. Promoting fairness therefore requires thorough examination of training data, ongoing performance monitoring, and the adoption of modelling strategies specifically designed to reduce unequal outcomes across patient populations.

- **Accountability and Liability**

Within clinical environments, it is essential to clearly establish who is responsible for decisions, especially when AI systems influence diagnoses or treatment choices. While explainable outputs can shed light on how a model arrived at a recommendation, they do not remove the ethical or legal responsibilities of healthcare professionals or institutions. Robust governance and accountability frameworks are necessary to address errors and adverse outcomes.

- **Privacy**

Certain explanation approaches—particularly those that rely on referencing similar patient cases or examples—may inadvertently expose sensitive personal information. As a result, XAI systems must incorporate strong safeguards for data protection and comply with de-identification and privacy standards to ensure that patient confidentiality is preserved.

- **Regulation**

Regulatory authorities, including the FDA and their international counterparts, are increasingly emphasizing transparency in clinical AI applications. Although explainability can assist in regulatory assessment, especially for risk evaluation and model validation, regulators also require clear evidence of safety, reliability, and clinical usefulness. Explanations alone are insufficient without rigorous testing and validation.

- **Patient Autonomy and Informed Consent**

Patients have a right to receive explanations that are clear, accessible, and appropriate to their level of medical understanding. Transparent communication about the role of AI in clinical care helps build trust, supports informed consent, and encourages shared decision-making. By making AI-driven recommendations easier to understand, explainability enables patients to participate more actively in planning and managing their treatment.

8. Challenges and Limitations

- **Misleading Explanations**

 doi:10.20944/preprints202605.2039.v1

Post-hoc explanation methods may generate outputs that seem intuitive and convincing, yet do not accurately represent the model's actual decision-making process. These seemingly reasonable explanations can create unwarranted trust and may mislead clinicians.

- **Lack of Standardized Evaluation Benchmarks**

Currently, there is no widely accepted framework for judging the quality of explanations in clinical contexts. The lack of common benchmarks, datasets, and evaluation metrics makes it difficult to compare different XAI approaches or define best practices.

- **Scalability Constraints**

Producing detailed, individualized explanations—particularly in large healthcare environments—can be computationally demanding. Maintaining explanation quality while ensuring efficiency and scalability continues to be a significant challenge.

- **Human-Centered Design Difficulties**

Clinicians vary in their experience levels, cognitive styles, and informational needs. In addition, personal biases can shape how explanations are interpreted. Designing explanation systems that are effective and usable across diverse user groups remains difficult.

- **Confusion between Correlation and Causation**

Many explanation techniques emphasize statistical associations rather than true causal relationships. Without careful interpretation, users may incorrectly infer causality from correlational patterns. Incorporating causal reasoning into explainable AI methods is still an evolving research direction.

9. Future Directions

Causal Explainable AI

Future efforts will focus on merging explainability with methods for identifying causes and predicting the impact of interventions. This will produce explanations that go beyond patterns in data to reveal how altering specific factors could change a patient's prognosis.

Customized Explanations

Explanations must adapt to the person using them. Physicians with different backgrounds and patients with varying understanding of health issues need options like simplified overviews, in-depth technical breakdowns, alternative visuals, or easier-to-follow wording. Tailoring in this way boosts comprehension, confidence, and practical value.

Common Standards for Assessment

Real progress requires agreed-upon ways to measure how good explanations are. This means developing shared clinical datasets that include expert-marked ideal explanations, along with studies involving multiple hospitals to see how these tools hold up in everyday medical practice.

Interactive Tools Handling Multiple Data Types

The next wave of systems will pull together different kinds of information such as scans, genetic data, and ongoing health records—into one cohesive, hands-on platform. Clinicians and others will be able to investigate predictions by running "what-if" tests, zooming into details, and trying out various scenarios.

Regulatory Standards for Explainability

 doi:10.20944/preprints202605.2039.v1

As oversight agencies insist on greater openness, dedicated frameworks are needed to test how dependable and steady explanations remain over time. Such rules will influence approval decisions, long-term safety checks, and continued supervision of AI applications in patient care.

10. Conclusion

Explainable AI (XAI) is essential for the responsible and effective use of machine learning in precision medicine. While many explanation methods exist, their real value comes from rigorous testing, close alignment with how clinicians think, and full compliance with ethical and regulatory guidelines. For these explanations to truly help patients and advance research, they need to be more than just technically accurate—they must be clear, consistent, and useful to the doctors who depend on them. XAI brings together knowledge from machine learning, clinical practice, human-computer interaction, and biomedical ethics. This multidisciplinary approach shows that the quality of explanations isn't just a technical issue; it's a broader human and organizational challenge influenced by daily clinical routines, doctors' experience, hospital policies, and what patients expect. Looking ahead, advances in XAI will depend on creating tools that fit naturally into routine clinical work, provide different levels of detail for various users, and adapt to the real-world messiness of healthcare settings. Long-term safety and reliability will require continuous monitoring after rollout, including checks for model degradation, regular input from clinicians, and proactive identification of new risks. Ethical values—fairness, accountability, equity, and patient privacy—must shape every stage of development and deployment. As regulators demand greater transparency and traceable decision-making, explainability will be key to meeting legal standards and building trustworthy oversight of AI in healthcare. Ultimately, XAI is not an add-on but a fundamental requirement for dependable precision medicine. By helping clinicians understand and trust AI recommendations, it improves diagnostic precision, enables truly personalized treatment plans, and fosters a healthcare system that is safer, more open, and better attuned to patients' needs.

Funding: The authors received no financial support for the research, authorship and publication of this article.

Conflicts of Interest / Competing Interests: The authors declare that they have no conflict of interest.

References

1. Evans BJ 2018 The challenge of regulating clinical decision support software after 21st Century Cures. American Journal of Law and Medicine
2. Adadi A and Berrada M 2019 Gastroenterology meets machine learning: Status quo and quo vadis. Advances in Bioinformatics
3. Shin SY and Lee JH 2019 Current status and future direction of digital health in Korea. Korean Journal of Physiology and Pharmacology
4. Ahirwar R and Mondal PR 2020 Interpretable machine learning in health care: Survey and discussions. International Journal of Innovative Research in Technology and Management
5. Coppola F and Faggioni L 2021 Human all too human? An all-around appraisal of the artificial intelligence revolution in medical imaging. Frontiers in Psychology
6. Wickramasinghe N and Bandara W 2021 A vision for leveraging the concept of digital twins to support the provision of personalized cancer care. IEEE Internet Computing
7. Bhatt C and Kumar I 2022 Emerging artificial intelligence-empowered mHealth: Scoping review. JMIR mHealth and uHealth
8. Chun S and Shin JH 2022 Prediction of conversion to dementia using interpretable machine learning in patients with amnestic mild cognitive impairment. Frontiers in Aging Neuroscience
9. Gerussi A and Cristoferi L 2022 Artificial intelligence for precision medicine in autoimmune liver disease. Frontiers in Immunology
10. Iqbal M and Pappas Y 2022 The use and ethics of digital twins in medicine. Journal of Law Medicine and Ethics
11. Ishengoma D and Mtaho A 2022 Artificial intelligence in digital health: Issues and dimensions of ethical concerns. Innovacion y Software

 doi:10.20944/preprints202605.2039.v1

12. Khanna S and Srivastava R 2022 Economics of artificial intelligence in healthcare: Diagnosis vs treatment. Healthcare
13. Kline A and Luo Y 2022 Multimodal machine learning in precision health: A scoping review. npj Digital Medicine
14. Laccourreye O and Garcia D 2022 Explainable machine learning for longitudinal multi-omic microbiome. Mathematics
15. Roy S and Banerjee S 2022 Demystifying supervised learning in Healthcare 4.0: A new reality of transforming diagnostic medicine. Diagnostics
16. Shazly SA and Grobman WA 2022 Introduction to machine learning in obstetrics and gynecology. Obstetrics and Gynecology
17. Wellnhofer E and Pinto DS 2022 Real-world and regulatory perspectives of artificial intelligence in cardiovascular imaging. Frontiers in Cardiovascular Medicine
18. Wesolowski R and Szymanski P 2022 An explainable artificial intelligence approach for predicting cardiovascular outcomes using electronic health records. PLOS Digital Health
19. Albahri AS and Albahri OS 2023 A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality bias risk and data fusion. Information Fusion
20. Baumgartner CF and Koch LM 2023 Fair and equitable AI in biomedical research and healthcare: Social science perspectives. Artificial Intelligence in Medicine
21. Bharati S and Podder P 2023 A review on explainable artificial intelligence for healthcare: Why how and when? IEEE Transactions on Artificial Intelligence
22. Hong W and Lee S 2023 Overcoming the challenges in the development and implementation of artificial intelligence in radiology: A comprehensive review of solutions beyond supervised learning. Korean Journal of Radiology
23. King B and Patel R 2023 What works where and how for uptake and impact of artificial intelligence in pathology: Review of theories for a realist evaluation. Journal of Medical Internet Research
24. Kuwaiti M and Nazer K 2023 A review of the role of artificial intelligence in healthcare. Journal of Personalized Medicine
25. Narayan KMV and Ali MK 2023 A strategic research framework for defeating diabetes in India: A 21st-century agenda. Journal of the Indian Institute of Science
26. Vorisek K and Lehne M 2023 Artificial intelligence bias in health care: Web-based survey. Journal of Medical Internet Research
27. Zafar A and Khan NM 2023 Reviewing methods of deep learning for intelligent healthcare systems in genomics and biomedicine. Biomedical Signal Processing and Control.