International Journal of Elite Research and Innovation

Vol. 01 Issue 02, July, 2026

Explainable Machine Learning in Healthcare Current Developments, Clinical Applications, and Future Perspectives

David A. Roser

Department of Gastroenterology, Augsburg University Hospital, Augsburg, Germany

Article History

Received Date: 13/06/2026
Revised Date: 14/07/2026
Accepted Date: 21/07/2026
Published Date: 28/07/2026

Abstract
Explainable AI is becoming more relevant in making decisions in the medical domain based on complex AI predictions that can be explained. This study summarizes the theoretical background, main techniques, ongoing research, clinical uses, assessment needs, and prospects of Explainable Machine Learning (XAI) in healthcare. There are several major approaches such as feature-importance analysis, model explanation given by SHAP and LIME, saliency mapping, counterfactual explanations, rule-based explanations, decision-tree models, attention mechanisms, example-based reasoning, and natural-language explanations or multimodal explanations. They are used in applications such as disease diagnosis, disease risk prediction, medical imaging, personalised treatment, drug discovery, critical-care support, remote monitoring and population-health surveillance. The study emphasizes that the quality of the explanation is dependent not only on visual clarity, but also on fidelity, stability, robustness, clinical relevance, actionability and consistency across settings and populations. Human-centred evaluation is also crucial, as explanations can enhance understanding while also fostering automation bias or reliance. To implement this in clinical practice, external validation, fairness assessment, workflow integration, transparent reporting, reproducibility, data governance, and continuous monitoring are essential. Facing these are a lack of standardized evaluation measures, biased and unrepresentative datasets, privacy/security concerns, regulatory uncertainty, and misleading post-hoc explanations. Future efforts should focus on communicating with stakeholders, future clinical trials, communicating uncertainty, and collaboration across disciplines. Finally, trustworthy explainable systems should empower patient involvement, enhance clinical decision-making, and facilitate safe, equitable and clinically relevant human-AI interactions.

Keywords:

- Explainable artificial intelligence
- machine learning
- healthcare
- clinical decision support
- model interpretability

1. Introduction

In healthcare, machine learning plays a crucial role in modern-day medicine by helping computer systems recognize complex patterns in clinical data and making predictions that aid in diagnosis, prognosis, treatment planning and monitoring of patients. Applications are now being made in medical imaging and electronic health records, disease risk assessment, rehabilitation, precision medicine and clinical decision-support systems. Recent advances prove that machine-learning models can handle the large amount of data and data heterogeneity quickly compared to traditional statistical models, enhancing the efficiency of healthcare. The use of AI has thus become a key research focus to break down these computational tools into clinically relevant and comprehensible information (Aravindkumar et al., 2026), known as explainable AI. However, this complexity of predictive algorithms has led to significant issues in how the output of the algorithm is produced and understood.

This growing trend of deep learning and other sophisticated algorithms has driven a greater need for explainable machine learning (XAI) in clinical settings. Explainability is defined as the ability to communicate the factors, relationships or processes that underlie the predictions of an AI system that can be understood by relevant stakeholders. To make computational models more interpretable and to enable meaningful interaction between humans and AI-driven systems, explainable AI (XAI) techniques have been developed that make otherwise opaque models interpretable (Adadi & Berrada, 2018). In the health sector, there are differences in how often and to what extent different parties (including clinicians, patients, researchers, developers, regulators, and policy makers) need to be explained. Hence, explainability needs to be considered as a multi-disciplinary requirement that is technically valid and clinically relevant, ethically acceptable, and communicates effectively (the Precise4Q consortium et al., 2020).

Black-box models can make very accurate predictions, but not enough information provided about the variables or relationship(s) that lead to predictions. This opacity makes it hard to test the clinical relevance of the model and to know whether it learns appropriate medical patterns or whether it takes a shortcut. For instance, a deep learning method for hip fracture prediction was discovered to have utilized non-clinically meaningful information on the health and patient, in addition to only clinically important image features (Badgeley et al., 2019). Similarly, saliency maps used to indicate relevant areas of the image can also be visually convincing even if the model parameters or training labels change significantly, questioning the reliability of saliency maps (Adebayo et al., 2018). Medical imaging studies also suggest that not all saliency techniques will be consistent in localizing abnormalities and/or provide accurate insights into the logic of the model (Arun et al., 2021).

The transparency of models is critical to understanding if the model output is clinically plausible, scientifically valid and consistent with accepted medical knowledge. But explanations should not be assumed to be proof of the reliability of an algorithm. Unsatisfactory explanations or explanations that are not followed through can lead to false confidence and result in unwarranted recommendations by clinicians (Antoniadi et al., 2021; Babic et al., 2021). Trust should thus be justified based on the validation, fidelity to the explanation, robustness and clinical benefits of the explanation, and not by its visual appeal. Accountability is also critical, given that the impact of health care decisions

is felt on the patient's safety, professional responsibility, informed consent, and health equity. Therefore, technical, legal, ethical and human-centred factors need to be included in the explainability frameworks (Barredo Arrieta, 2020).

This study is intended to be a critical examination of the recent developments in explainable machine learning (XAI) and to review its important clinical applications. It evaluates some of the top explanation techniques, how they are implemented in medical data and medical imaging, and how the quality of the explanation is measured. The study also examines issues with clinical validation, model reliability, user trust from users, workflow, standardisation, and regulatory accountability. The weaknesses and prospects of explainable clinical decision support systems are presented.

2. Conceptual Foundations of Explainable Machine Learning

2.1 Definition and Principles of Explainable Artificial Intelligence

Explainable AI is defined as techniques and systems that can be used to explain how predictions, classifications or recommendations are made by machine learning models. Central the principles are transparency, comprehensibility, fidelity, consistency, and clinical relevance. In healthcare, the three elements of explainability are important because a technically correct explanation have little use if it cannot be understood or applied by a healthcare professional (Cutillo et al., 2020).

2.2 Interpretability versus Explainability

Interpretability is a property of a model, as is explainability, but interpretability is more about the "how" of a model's working, while explainability is more about the "how" of a model's working for "how" a specific user. These terms are often confused, but interpretability is sometimes considered to be a quality of the model rather than an analytical technique, and explainability can be obtained by other methods. The effectiveness should be evaluated by aspects beyond just visual representation, such as model fidelity, comprehensiveness, stability and usefulness (Carvalho et al., 2019).

2.3 Intrinsically Interpretable and Post-Hoc Explainability Methods

Intrinsically interpretable models such as decision trees, rule-based systems and sparse regression models make predictions via a structure that can be looked at directly. On the other hand, post-hoc approaches are used after training to explain complex black-box models using feature attribution, surrogate modelling, saliency mapping or counterfactual analysis. Post hoc explanations may however mask the use of clinically irrelevant shortcuts. For example, it has been demonstrated that COVID-19 models based on radiography show artefacts of the data sets instead of true pathological features (DeGrave et al., 2021).

2.4 Local and Global Explanations

Local explanations make sense of why a model made a certain prediction for a given patient, while global explanations explain the overall behaviour of the model in the population. Local approaches are particularly important in personalised clinical decision-making as they focus on influential variables for individual cases. There are also differences in local interpretability methods that can lead to different explanations for the same prediction of the health care system, potentially resulting in inconsistencies and clinical uncertainty (ElShawi et al., 2021).

2.5 Model-Specific and Model-Agnostic Approaches

Model-specific methods rely on specific algorithms and use their internal structure to produce explanations. The relationships between the inputs and outputs can be used to apply model-agnostic approaches to different predictive models. While the model-agnostic methods are flexible for deployment, they do not necessarily reflect the thinking of the original model. Furthermore, evaluation of explanation systems needs to be done under real-world operational conditions and not just based on experimental datasets (Bhatt et al., 2020). The ability to be manipulated adversely also highlights how predictive models and explanations can be manipulated on purpose (Finlayson et al., 2019). The conceptual aspects of XAI can be classified based on model transparency, the scope of the explanation, methodological dependency, and stakeholder needs, as shown in Table 1.

Table 1. Conceptual Dimensions of Explainable Machine Learning in Healthcare

Dimension	Categories	Healthcare relevance	References
Model transparency	Interpretable; black-box	Indicates whether clinicians can directly understand model decisions.	(Lipton, 2018)
Explanation type	Intrinsic; post-hoc	Distinguishes transparent models from externally generated explanations.	(Guidotti et al., 2019)
Explanation scope	Local; global	Supports patient-specific and population-level interpretation.	(Lundberg et al., 2020)
Method dependence	Model-specific; model-agnostic	Guides method selection according to algorithm type.	(Vilone & Longo, 2021)
Intended user	Clinician; patient; developer; regulator	Ensures explanations meet different stakeholder requirements.	(the Precise4Q consortium et al., 2020)

2.6 Stakeholders and Their Explanation Requirements

There are different requirements for explanation from clinicians, patients, developers, administrators and regulators. Clinicians want information that is actionable and medically plausible, patients want information that is accessible and enables informed participation. Technical evidence is needed to help developers debug, while documentation is necessary to show safety and accountability to regulators. Technical evidence is needed for developers to help them debug, and documentation is necessary for regulators to know it is safe and accountable. Design and implementation of explainable healthcare systems (Char et al., 2018) thus need to be informed by ethical issues related to autonomy, fairness, privacy, and responsibility. Clinical deployment studies also demonstrate that explanations need to be compatible with current work routines and consider that there are many factors influencing both health care workers and patients, that can impact the actual use of the clinical computer application (Beede et al., 2020).

2.7 Role of Explainability in Clinical Decision-Making

Explainability can support the verification of the prediction, detection of errors, comparison of algorithmic recommendations with medical knowledge, and communicating decisions to patients. However, explanations also lead to automation bias as it makes it easier for the user to accept seemingly plausible outputs without careful consideration. There are cognitive forcing strategies that can help minimize over-dependence on AI systems by first having users independently evaluate recommendations before viewing them (Buçinca et al., 2021). Explanation should thus complement and not supplant clinical decision making. Figure 1 shows the interaction between clinical data, predictive models, explanation approaches and interpretation of these models by stakeholders, for transparent and accountable healthcare decisions.

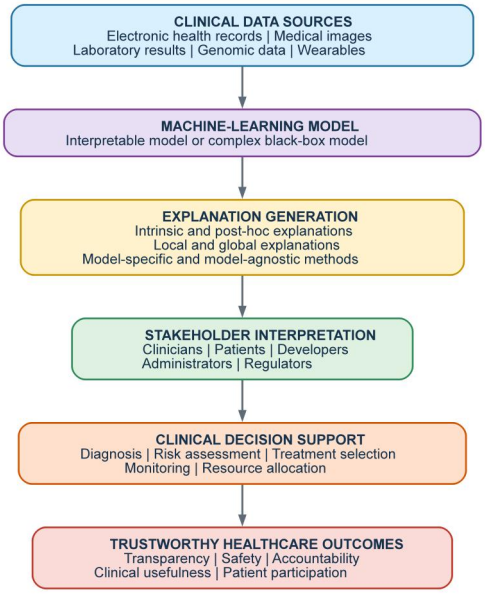

Figure 1. Conceptual Framework of Explainable Machine Learning in Healthcare

3. Major Explainability Methods Used in Healthcare

3.1 Feature-Importance and Feature-Attribution Methods

Both feature-importance and feature-attribution techniques determine the variables that play the most significant role in a model's prediction. In health care, these methods identify laboratory values, symptoms, demographic factors, medications, and/or physiologic measurements that are linked to diagnostic or prognostic outcomes. However, feature importance may reflect some biases of the electronic health record, such as unequal representation, data missingness, and variations in documentation practices (Gianfrancesco et al., 2018). Therefore, attribution scores must be used in conjunction with clinical knowledge and data-quality assessments. This selection process is summarized as a structured pathway in Figure 2, which guides users to the right explainability method based on clinical data type, explainability purpose, target stakeholder and desired output.

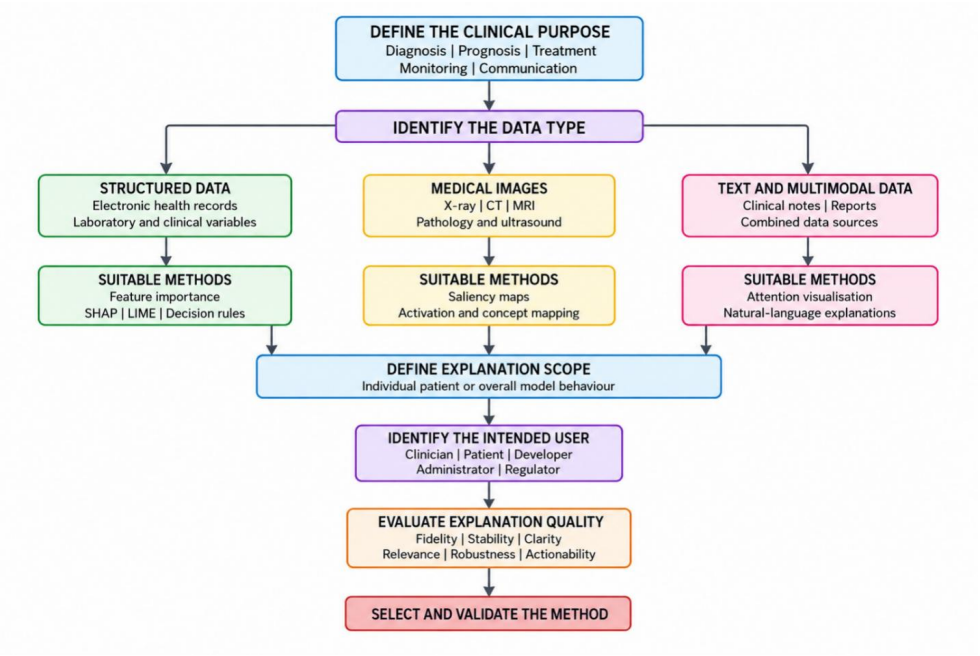

Figure 2. Explainability-Method Selection Pathway for Healthcare Applications

3.2 SHAP and LIME-Based Explanations

For generating local explanations, SHAP and LIME are commonly used techniques that are model agnostic in nature. SHAP uses cooperative game theory to assign a contribution value to each feature, while LIME uses a simpler, surrogate model to approximate a complex model around an individual prediction. These approaches can be used for structured clinical data, imaging, risk-prediction models and so on. However, post hoc explanations can over-simplify complex decision-making processes and give the wrong idea of the reasoning behind the model (Ghassemi et al., 2021). Their output should therefore be subjected to fidelity checks, checks for stability and checks for clinical relevance.

3.3 Saliency Maps and Visual Explanation Techniques

In radiology, pathology, dermatology and ophthalmology, saliency maps, class-activation maps, heat maps and gradient-based visualisation methods are popular. They indicate parts of the image that they have determined to be influential to a model's classification. Although they look intuitive, the explainability concept is not consistently applied in computer-aided radiological diagnosis, and many studies are limited in their clinical evaluation of visual explanations (Groen et al., 2022). Furthermore, some methods of visual explanations in neural networks are quite fragile; minor changes in the input result in significant differences in how the neural network is interpreted without altering the prediction (Ghorbani et al., 2019).

3.4 Counterfactual Explanations

Counterfactual explanations explain the minimal clinically important changes needed to generate an alternate prediction. For instance, if a patient's estimated cardiovascular risk would be reduced by a certain level if blood pressure or cholesterol were lowered to a certain level, it would be mentioned. These are action-focused explanations that can be of benefit for shared decision-making. But alternatives generated should be medically possible and not recommend a change of an immutably set or clinically inappropriate variable. The counterfactual approaches are a subset of the black-box model interpretation techniques (Guidotti et al., 2019).

3.5 Rule-Based and Decision-Tree Explanations

Decision Trees and Rule-based models: These models give predictions in terms of comprehensible conditional structures. They enable clinicians to understand how specific combinations of variables lead to a recommendation, by virtue of their transparent pathways. They facilitate auditability and are particularly useful if clinical accountability is favoured. With the rise of explainable AI, there has been attention on systems that can predict and reason in an understandable way, as opposed to only providing explanations for opaque systems (Gunning et al., 2019).

3.6 Attention-Based Explanation Methods

Attention mechanisms assign different weights to elements of the input such as clinical notes, image regions or time-series observations. Attention visualisations used to show what information was used to make a prediction, but attention weights cannot always be seen as faithful explanations. They do not have a consistent meaning and are dependent on the model architecture, training, and clinical context. However, there is a need for better validation of attention-based outputs for use in clinical decision-making in the current state of medical XAI research (González-Alday et al., 2023).

3.7 Example-Based and Prototype-Based Explanations

Example-based methods rely on retrieval of similar past patients, representative prototypes, and/or contrasting cases to explain the predictions. These explanations are in line with the way clinicians think in terms of cases and support users to compare an algorithmic recommendation with recognised disease patterns. However, bad choice of examples may support bias or foster inappropriateness. In clinical decision-support research, clinical decision-makers sometimes ignore correct recommendations when they are explained to them (Gaube et al., 2021).

3.8 Natural-Language and Multimodal Explanations

Natural-language explanations convert the model output to textual rationales, summaries or patient-specific suggestions. Multimodal systems feature a synthesis of text and images, numerical data and temporal records, and visual evidence. While these techniques may help to make the content more accessible, they should not create explanations that are convincing but not supported. Neurological evidence indicates the need to carefully evaluate the explanation design's impact on clinician reliance and decision performance, suggesting the necessity of careful human-centred evaluation (Gombolay et al., 2024). As summarized in Table 2, the different principal explainability techniques provide different types of outputs, have varying clinical applications, have different advantages, and have different methodological limitations.

Table 2. Comparison of Major Explainability Methods Used in Healthcare

Method	Typical output	Main application	Limitation	References
Feature importance	Ranked variables	Risk prediction and prognosis	Does not establish causality.	(Gianfrancesco et al., 2018)
SHAP	Feature-contribution values	Patient-level risk assessment	Computationally intensive.	(Lundberg et al., 2018, 2020)
LIME	Local surrogate explanation	Clinical classification	May produce unstable results.	(Ghassemi et al., 2021)
Saliency maps	Highlighted image regions	Radiology and pathology	May be visually persuasive but unreliable.	(Arun et al., 2021)
Counterfactuals	Alternative clinical scenario	Treatment and risk reduction	Recommendations may be infeasible.	(Holzinger et al., 2020)
Decision rules	Conditional pathways	Diagnosis and triage	Complex models may require many rules.	(Carvalho et al., 2019)
Attention methods	Weighted input elements	Clinical text and imaging	Attention may not reflect true reasoning.	(Mesinovic et al., 2025)
Natural-language explanations	Textual rationale	Clinical communication	May generate unsupported explanations.	(Langer et al., 2021)

4. Current Developments in Explainable Healthcare Machine Learning

4.1 Explainable Deep Learning for Medical Data

In the clinical domain, explainable deep-learning systems are being created that integrate analysis of complex clinical data and give interpretable evidence of their predictions. These systems integrate a combination of deep neural networks and feature attribution, visualisation, rule-extraction and uncertainty-estimation techniques. In medicine, explainability is of paramount importance as clinicians need to know that predictions are made based on proper physiological patterns and not spurious correlations. Hence, medical Artificial Intelligence should be capable of giving explanations which should be helpful for clinical scrutiny and responsible decision-making (Kundu, 2021).

4.2 Explainability in Medical Imaging and Computer Vision

Medical imaging continues to be a field with a great amount of activity for explainable machine learning. To visualise the anatomical regions which impact diagnostic predictions, saliency maps, class-activation mapping, concept-based explanations and region-level visualisations are provided. But clinically useful explanations need to be accurate, stable and comprehensible by radiologists, not just good-looking. New guidelines have been introduced for evaluation, which include localisation performance, explanation consistency, clinical relevance, and correspondence with expert annotations (Jin et al., 2023). There are also new methods that focus on explanation based on human reasoning and radiological interpretation (Lee, 2026).

4.3 Explainable Models for Electronic Health Records

Electronic health records include longitudinal data that include diagnoses, medication prescriptions, lab results, procedures, and physiological measurements. Explainable models can be used to understand which variables and temporal patterns are important for predicting outcomes. For instance, an interpretable model that was built to forecast acute critical illness based on EHR data has been able to deliver explanations along with prediction scores on a per-patient basis, which has the potential to support explainability for early clinical action (Lauritsen et al., 2020).

4.4 Explainable Natural-Language Processing in Clinical Practice

Explainable NLP is used in clinical notes, discharge summaries, pathology reports and medical literature. Current systems are able to identify influential words, clinical concepts, and create textual rationales for classifications. However, explanations produced need to be evaluated for information and clinical logic. It is essential to consider

workflow compatibility, data quality, prospective evaluation, and measurable clinical benefit to achieve successful implementation (Kelly et al., 2019).

4.5 Explainable Federated and Privacy-Preserving Learning

Federated and privacy-preserving learning can be used to train models collaboratively without any direct exchange of sensitive patient data between healthcare institutions. Incorporating explainability now is being adopted to look for patterns and pinpoint inconsistent model behaviors in these systems on an institutional level. Access to the information needed for detailed explanations may, however, be limited due to privacy mechanisms. Explanation formats must thus provide transparency, confidentiality, usability and adherence to regulations to stakeholders (Langer et al., 2021).

4.6 Causal Explainability and Counterfactual Reasoning

Causal explainability aims to go beyond the identification of statistical association, to find potentially causal relationships that might have a meaningful impact on clinical outcomes. A similar evaluation of how changes in patient characteristics might affect a prediction is called "counterfactual reasoning. The idea of causability is about the effectiveness of explanations in helping a human user to understand cause-and-effect relationships, which relates the computational explanation to human reasoning (Holzinger et al., 2019).

4.7 Explainability in Foundation and Generative AI Models

Clinical summarisation, question answering, diagnostic support, and multimodal analysis are among the uses of these models that are gaining in popularity: foundation and generative models. They are large and complex, introducing new risks of fabricated rationales, unbacked recommendations and reduced traceability. Thus, one can see a shift in the field of explainability research towards evidence-supported output, source attribution, uncertainty communication, and means of detecting clinical information in the source of generated responses (Buçinca et al., 2021).

4.8 Human-Centred and Interactive Explainability Systems

Human-centred explainability enables clinicians to engage with explanations, consider counterfactuals and ask for information specific to certain decisions. Such explanations should be judged from a technical and humanistic point of view. System Causability Scale is a structured scale to assess the clarity, usefulness, completeness, and cognitive effectiveness of explanations produced by humans and machines (Holzinger et al., 2020). Figure 3 depicts the relationship between the machine-generated explanations and professional clinical judgement which was iterative.

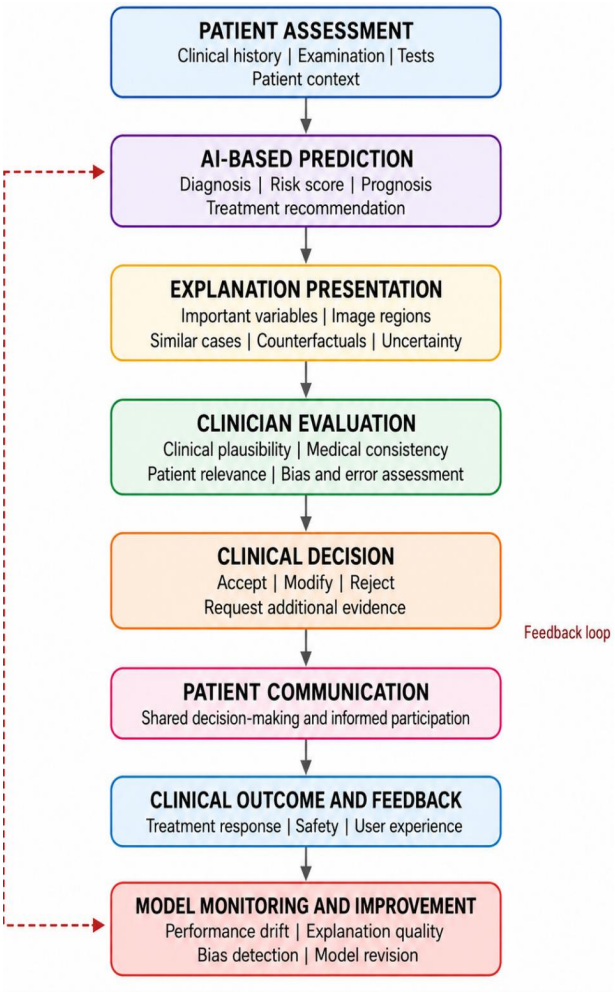

Figure 3. Human-Centred Explainable Clinical Decision-Support Cycle

5. Clinical Applications of Explainable Machine Learning

5.1 Disease Diagnosis and Early Detection

Explainable machine learning models are becoming more prevalent for early detection of diseases from clinical, laboratory, imaging and genomic information. Explanations facilitate clinicians to understand the contribution of various symptoms, biomarkers or patient characteristics to a diagnostic prediction. In the healthcare sector, XAI is very useful in the diagnosis of neurological diseases, cardiovascular diseases, oncological diseases, and metabolic diseases, as well as in making the output of the automated systems more understandable for the human user (Loh et al., 2022). But care needs to be taken when evaluating the quality of explanations as apparent interpretability does not necessarily imply that the model's reasoning is rendered faithfully (Lipton, 2018).

5.2 Clinical Risk Prediction and Prognostic Assessment

Clinical risk-prediction models are used to predict the risk of complications, disease progression, readmission, or mortality. Explainability helps the clinician to distinguish patient-specific risk factors from population-level factors. Tree-based explainability approaches can combine local explanations, explaining patient-specific predictions, with global explanations, providing insights into the general model behaviour, which can be used to support both personalised explanation and institutional risk management (Lundberg et al., 2020). Reliable explanations thus enhance clinical confidence if they are stable, clinically relevant, and if they are validated accordingly (Markus et al., 2021).

5.3 Medical Imaging and Radiological Interpretation

In radiology, explainable models are used to assist in the interpretation of X-rays, computed tomography, magnetic resonance imaging, mammography and ultrasound. The use of visual explanation techniques can help to reveal the image regions corresponding to abnormalities and enable radiologists to compare the model attention with well-known anatomical evidence. However, even if systems are very accurate, ethical and clinical issues can arise when practitioners are unable to understand or question the recommendations (London, 2019).

5.4 Personalised Medicine and Treatment Recommendation

Explainable machine learning can be applied to personalised medicine to uncover the effect of patient identity, comorbidities, genetic makeup and prior treatment responses on treatment recommendations. Patient-specific explanations may be useful for the clinician to discuss different treatment options and to highlight the expected benefits and/or the risks. But explanations for models should not imply that fairness is simply a technical issue; structural, social and institutional factors that cannot be addressed through algorithmic changes can also contribute to health disparities (McCradden et al., 2020).

5.5 Drug Discovery and Therapeutic Response Prediction

To identify the important biological aspects of molecules, to predict the interactions between drug and target, to estimate toxicity, and to predict treatment response, explainable models are used in drug discovery. Explanations aid researchers in understanding why certain compounds are targeted and why any relationships that have been identified are scientifically plausible. Emerging language and foundation models have the potential to further facilitate literature synthesis and therapeutic hypothesis generation, though explanations should be vetted as generated rationales persuasive yet lacking data (Mesinovic et al., 2025).

5.6 Intensive-Care and Emergency Decision Support

Patients deteriorating in intensive care and EMS settings need to be rapidly assessed. Explainable systems can forecast a sepsis event, respiratory failure, cardiac events and more, as well as pinpoint the factors that increase the risk. An explainable ML system created for surgical care showed that patient-specific predictions can help prevent intraoperative hypoxaemia and provide clinicians with information about the risk factors that influence their prediction (Lundberg et al., 2018).

5.7 Remote Patient Monitoring and Digital Health

Wearable devices and mobile-health platforms gather data on activity, sleep, glucose levels, heart rate and other parameters continuously. These data can be translated into human-readable alerts, behavioural suggestions and risk summaries using explainable models. While these explanations enhance patient engagement, they need to be brief, appropriate and tailored to the context and patient's health literacy level (the Precise4Q consortium et al., 2020).

5.8 Population Health and Public-Health Surveillance

Explainable machine learning can, at a population level, detect patterns of disease, vulnerable populations, trends in health care usage and potential public health threats. Explanations enable policy makers to look at how demographic, environmental and socio-economic factors affect predictions. However, for population-level applications, careful consideration must be taken to avoid decisions regarding resource allocation or surveillance being made based on biased data (Gianfrancesco et al., 2018). Explainable machine learning can enable a wide variety of healthcare tasks, such as individual diagnosis to population-level surveillance, as shown in Table 3.

Table 3. Clinical Applications and Functions of Explainable Machine Learning

Application	Explanation function	Expected benefit	References
Disease diagnosis	Identifies influential symptoms, biomarkers, or features.	Earlier and more reliable detection.	(Kundu, 2021)
Risk prediction	Explains patient-specific risk factors.	Improved prioritisation and prevention.	(Markus et al., 2021)
Medical imaging	Highlights relevant anatomical regions.	Supports radiological verification.	(Groen et al., 2022)
Personalised treatment	Explains treatment recommendations.	More individualised clinical decisions.	(London, 2019)
Drug discovery	Identifies influential molecular features.	Better candidate selection and interpretation.	(Loh et al., 2022)

Critical care	Reveals factors linked to deterioration.	Faster clinical intervention.	(Lauritsen et al., 2020)
Remote monitoring	Converts sensor data into understandable alerts.	Supports early intervention and self-management.	(Cutillo et al., 2020)
Public-health surveillance	Explains population-level risk patterns.	Improved planning and resource allocation.	(Parikh et al., 2019)

6. Evaluation, Validation, and Clinical Implementation

6.1 Technical Evaluation of Explanation Quality

Technical analysis finds out whether an explanation is a precise account of the behaviour of the machine-learning model being explained. Popular dimensions of assessment are completeness, comprehensibility, sparsity, localisation accuracy, computational efficiency and consistency with known domain knowledge. Both algorithmic performance and the target users should be evaluated since what is needed to explain something is based on the purpose of presenting information (Mohseni et al., 2021). Systematic comparison of deep-learning interpretation approaches is also required to assess the ability of the approaches to detect meaningful input features, as opposed to visually convincing but irrelevant patterns (Montavon et al., 2018). A clinically deployable explainable-AI system must meet requirements of technical, human-centred, organisational, and governance, as outlined in Table 4.

Table 4. Evaluation and Clinical Implementation Requirements for Explainable AI

Dimension	Main assessment	Clinical importance	References
Fidelity	Agreement between explanation and model behaviour	Prevents misleading explanations.	(Mohseni et al., 2021)
Stability	Similar explanations for similar patients	Ensures consistent interpretation.	(Vilone & Longo, 2021)
Robustness	Resistance to noise and manipulation	Improves system safety.	(Tabassi, 2023)
Clinical relevance	Alignment with medical knowledge	Supports meaningful decisions.	(Petch et al., 2022)
Actionability	Ability to guide clinical action	Improves practical usefulness.	(Miller, 2019)
Human factors	Effects on trust and decision performance	Reduces inappropriate reliance.	(Nicolson et al., 2025)
External validity	Performance across populations and settings	Establishes generalisability.	(Kelly et al., 2019)
Fairness	Consistency across demographic groups	Reduces health inequities.	(Obermeyer et al., 2019)
Workflow compatibility	Integration with clinical systems	Reduces burden and alert fatigue.	(Beede et al., 2020)
Reproducibility	Transparent reporting of methods and data	Supports independent validation.	(Tabassi, 2023)

6.2 Fidelity, Stability, Consistency, and Robustness

Fidelity is the degree to which an explanation faithfully reflects the prediction process of an original model. Stability is the extent to which similar clinical cases produce similar explanations, and consistency looks at consistency across repeated analyses or other related methods. Robustness assesses how well explanations are stable in the face of small changes in the inputs. These properties are necessary since unstable explanations can still deceive clinicians even in the case where predictive accuracy does not vary. Artificial-intelligence risk management frameworks thus suggest ongoing testing, documentation, monitoring, and governance over the lifecycle of the system (Tabassi, 2023).

6.3 Clinical Relevance and Actionability of Explanations

The explanations provided clinically need to take into consideration medically relevant factors and justify adequate action, which may be further examination, alteration of treatment, or closer attention. The explanations must be related to clinical logic and not just show statistical relationships. Explainable models can enhance the interpretation of risks in cardiology, but they are of limited value when the outputs of explanations are not related to actionable clinical choices (Petch et al., 2022).

6.4 Human-Centred Evaluation with Healthcare Professionals

Human-centre assessment engages clinicians in the process of understanding the explanations to be understandable, useful, timely, and consistent with professional reasoning. According to social-science research, efficient explanations are selective, contrastive and tailored to the knowledge and anticipations of their recipients (Miller, 2019). Assessment must thus encompass interviews, usability studies, simulated scenarios, and prospective clinical trials as opposed to basing on computational measures only.

6.5 Effects on Clinical Trust and Decision Performance

Explainability enhances trust, and increased trust does not always yield better decisions. Clinicians vary in experience, confidence and vulnerability to automation, which implies that the same justification can lead to the right level of reliance in one user and excessive reliance in another. Recent findings indicate a high degree of clinician variability in trust, reliance, and performance in the interaction with explainable systems (Nicolson et al., 2025).

6.6 External Validation and Real-World Generalisability

External validation determines whether a model and its explanations can be trusted across diverse hospitals, populations, devices and geographic locations. Models that are trained on small datasets recreate latent institutional or demographic biases. An example of a population-health algorithm that was widely used was inadequate in estimating the healthcare needs of Black patients since expenditure was the proxy of illness severity (Obermeyer et al., 2019). Assessment of bias should be thus included in the validation processes before implementation (Parikh et al., 2019).

6.7 Integration with Clinical Workflows and Health Information Systems

Elucidable systems should be able to bring harmony with electronic health records and current decision processes. The descriptions should be brief, provided at the right point of care, and should not add a heavy load of documentation and alert fatigue. Real-world electronic health record studies indicate the necessity of enhanced prospective assessment and implementation of explainable models in a workflow (Payrovnaziri et al., 2020).

6.8 Reporting Standards and Reproducibility Requirements

Datasets, preprocessing, model architecture, explanation techniques, evaluation metrics, subgroup analysis, and validation processes should be documented by way of transparent reporting. Adequate information should also be given by researchers to reproduce predictions as well as explanations. Comparability and methodological weaknesses Standardised reporting helps facilitate regulatory review and responsible clinical implementation (London, 2019). Figure 4 shows the stepwise lifecycle to transition an explainable machine-learning model into safe and sustainable clinical use.

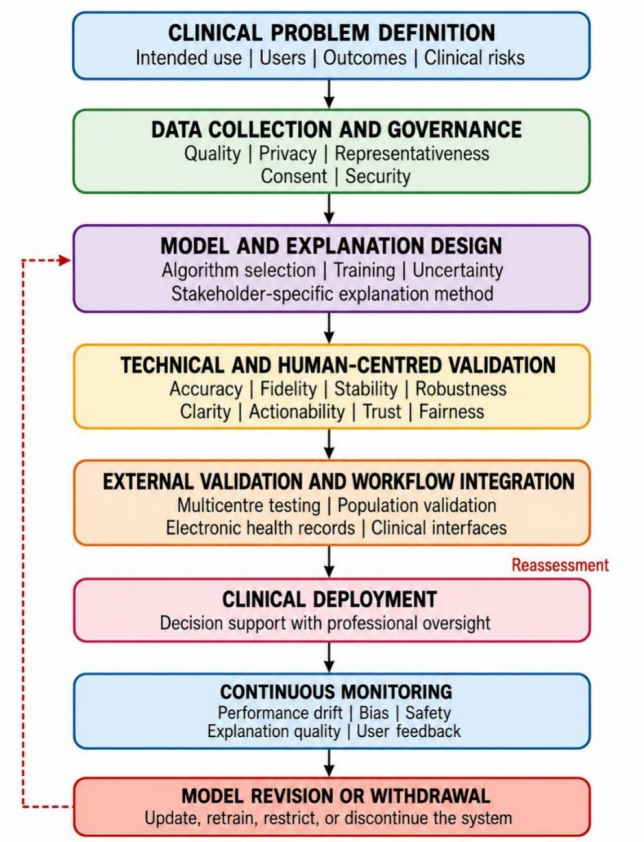

Figure 4. Clinical Validation and Implementation Lifecycle for Explainable AI

7. Challenges, Research Gaps, and Future Perspectives

7.1 Standardization and Evaluation Gaps

One of the challenges of explainable machine learning is the lack of a widely accepted formal measurement of the quality of an explanation. It is difficult to compare results across existing studies because various assessors used different scales for the assessment of fidelity, comprehensiveness, stability, completeness, and usefulness. Furthermore, explanations that seem straightforward cannot always increase human understanding or accuracy of decision making. Experimental results have shown that there is a potential difference between perceived interpretability and task performance, which requires accurate evaluation frameworks that are user-centred and objective (Poursabzi-Sangdeh et al., 2021). Future research is needed to provide clinically validated benchmarks to evaluate technical faithfulness and usefulness (Vilone & Longo, 2021).

7.2 Bias, Fairness, Privacy, and Data Governance

Explanations do not necessarily prevent a system from showing discriminatory patterns, but they do not remove the bias when the datasets are unrepresentative, there is unequal access to healthcare, or there are inappropriate proxy variables. Performance should thus be evaluated for demographic, socioeconomic, and clinical subgroups in the context of fairness evaluation. But privacy and data-governance issues are also significant, as a detailed explanation reveals sensitive features of the patient or information about the training data. Strong governance frameworks are needed to establish who owns the data, who has access, who is responsible, who monitors and who is held accountable (Reddy et al., 2020; Vayena et al., 2018).

7.3 Regulatory, Ethical, and Legal Considerations

The concept of clinical explainability is intimately related to informed consent, patient autonomy, professional responsibility and liability. Healthcare professionals need to be able to challenge algorithmic recommendations as well as recognize the restrictions of automated outputs. In addition, there needs to be transparency around data usage, model uncertainty, possible harm and who is responsible when the model is used in the clinic. It should therefore be required that regulations include documented evidence of validation, regular testing and monitoring, and a procedure for handling model failures.

7.4 Accuracy, Interpretability, and Unfaithful Explanations

Complex models might be good predictors but are not easily understood. On the contrary, simpler models can be more transparent and yield lower accuracy in some tasks. This trade-off must be weighed clinically risk-wise instead of looking at it technically. In addition, post-hoc explanations also have the risk of providing plausible, but unfaithful explanations that do not actually reflect model reasoning. These explanations can lead to false confidence and further compound automation bias.

7.5 Stakeholder-Specific and Clinically Deployable Explainable AI

Future Explainable Systems should be able to provide individual explanations for clinicians, patients, regulators and developers. Clinicians need concrete evidence, while patients need easily digestible information that is health-literate. Future clinical trials, multi-disciplinary cooperation, external validation, uncertainty communication and real-world workflow integration will be the key factors in progress. Trustworthy explainable AI should ultimately enhance critical human judgement rather than replace professional and patient participation.

8. Conclusion

Explainable machine learning plays a key role in responsible AI adoption in healthcare. It does not just enhance model transparency; it also aids clinical verification, holds accountability, communicates with patients, and aids in detecting bias, data restrictions, and unsafe patterns of prediction. The applications currently under development show significant promise in the areas of diagnosis, prognosis, medical imaging, personalised treatment, critical-care monitoring, digital health, drug discovery and population-health management. But the value of an explanation as a clinical tool is related to its fidelity, stability, comprehensibility, relevance, and capacity to facilitate appropriate action. Several existing explainability techniques complement both each other and have their own limitations, such as feature-attribution, saliency maps, counterfactual explanations, interpretable rules, example-based methods, and natural-language explanations. Explanations that are visually convincing or technically detailed, but do not accurately describe the model behaviour, can still be misleading. Thus, it is not sufficient to consider explainability as an alternative to model validation, external testing, fairness evaluation, uncertainty modelling, and ongoing clinical monitoring. Standardised evaluation metrics, explanation designs tailored to individual stakeholders, more robust reporting systems, robust data-governance systems and future studies in actual clinical settings are required to support future progress. Multidisciplinary collaboration is going to be vital between clinicians, the patient, data scientists, regulators, ethicists and system developers. Finally, reliable explainable machine learning should be used to complement the professional decision-making process, to improve equity in healthcare delivery, and to facilitate safe, transparent, and clinically relevant collaboration between humans and AI.

References

1. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138–52160.
2. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. Advances in Neural Information Processing Systems, 31. https://proceedings.neurips.cc/paper/8160-sanity-checks-for-saliency-maps
3. Antoniadi, A. M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B. A., & Mooney, C. (2021). Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review. Applied Sciences, 11(11), 5088.
4. Aravindkumar, A., Ramadoss, M., Fakhruddin Ahmed, S. A., Sampath, V., & Lakshminarayanan, K. (2026). Explainable AI in healthcare: A systematic review of XAI use cases in imaging, diagnostics, and rehabilitation. Frontiers in Artificial Intelligence, 9, 1749527.
5. Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J., Gidwani, M., Adebayo, J., Li, M. D., & Kalpathy-Cramer, J. (2021). Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging. Radiology: Artificial Intelligence, 3(6), e200267. https://doi.org/10.1148/ryai.2021200267
6. Babic, B., Gerke, S., Evgeniou, T., & Cohen, I. G. (2021). Beware explanations from AI in health care. Science, 373(6552), 284–286. https://doi.org/10.1126/science.abg1834
7. Badgeley, M. A., Zech, J. R., Oakden-Rayner, L., Glicksberg, B. S., Liu, M., Gale, W., McConnell, M. V., Percha, B., Snyder, T. M., & Dudley, J. T. (2019). Deep learning predicts hip fracture using confounding patient and healthcare variables. NPJ Digital Medicine, 2(1), 31.
8. Barredo Arrieta, A. (2020). D az-Rodr guez. N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F.
9. Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., & Vardoulakis, L. M. (2020). A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic

Retinopathy. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–12. https://doi.org/10.1145/3313831.3376718
10. Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M. F., & Eckersley, P. (2020). Explainable machine learning in deployment. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 648–657. https://doi.org/10.1145/3351095.3375624
11. Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), 1–21. https://doi.org/10.1145/3449287
12. Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. Electronics, 8(8), 832.
13. Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—Addressing ethical challenges. The New England Journal of Medicine, 378(11), 981.
14. Cutillo, C. M., Sharma, K. R., Foschini, L., Kundu, S., Mackintosh, M., Mandl, K. D., & Noel 1, M. in H. W. W. G. B. T. 1 C. E. 1 C. C. 1 G. K. 1 G. V. 1 J. R. 8 S. B. 9 S. (2020). Machine intelligence in healthcare—Perspectives on trustworthiness, explainability, usability, and transparency. NPJ Digital Medicine, 3(1), 47.
15. DeGrave, A. J., Janizek, J. D., & Lee, S.-I. (2021). AI for radiographic COVID-19 detection selects shortcuts over signal. Nature Machine Intelligence, 3(7), 610–619.
16. ElShawi, R., Sherif, Y., Al-Mallah, M., & Sakr, S. (2021). Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. Computational Intelligence, 37(4), 1633–1650. https://doi.org/10.1111/coin.12410
17. 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
18. Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., Coughlin, J. F., Guttag, J. V., Colak, E., & Ghassemi, M. (2021). Do as AI say: Susceptibility in deployment of clinical decision-aids. NPJ Digital Medicine, 4(1), 31.
19. Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11), e745–e750.
20. Ghorbani, A., Abid, A., & Zou, J. (2019). Interpretation of neural networks is fragile. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 3681–3688. https://aaai.org/ojs/index.php/AAAI/article/view/4252
21. Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. JAMA Internal Medicine, 178(11), 1544–1547.
22. Gombolay, G. Y., Silva, A., Schrum, M., Gopalan, N., Hallman-Cooper, J., Dutt, M., & Gombolay, M. (2024). Effects of explainable artificial intelligence in neurology decision support. Annals of Clinical and Translational Neurology, 11(5), 1224–1235. https://doi.org/10.1002/acn3.52036
23. González-Alday, R., García-Cuesta, E., Kulikowski, C. A., & Maojo, V. (2023). A scoping review on the progress, applicability, and future of explainable artificial intelligence in medicine. Applied Sciences, 13(19), 10778.
24. Groen, A. M., Kraan, R., Amirkhan, S. F., Daams, J. G., & Maas, M. (2022). A systematic review on the use of explainability in deep learning systems for computer aided diagnosis in radiology: Limited use of explainable AI? European Journal of Radiology, 157, 110592.
25. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys, 51(5), 1–42. https://doi.org/10.1145/3236009
26. Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. Science Robotics, 4(37), eaay7120. https://doi.org/10.1126/scirobotics.aay7120
27. Holzinger, A., Carrington, A., & Müller, H. (2020). Measuring the Quality of Explanations: The System Causability Scale (SCS): Comparing Human and Machine Explanations. KI - Künstliche Intelligenz, 34(2), 193–198. https://doi.org/10.1007/s13218-020-00636-z
28. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. WIREs Data Mining and Knowledge Discovery, 9(4), e1312. https://doi.org/10.1002/widm.1312
29. Jin, W., Li, X., Fatehi, M., & Hamarneh, G. (2023). Guidelines and evaluation of clinical explainable AI in medical image analysis. Medical Image Analysis, 84, 102684.
30. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17(1), 195. https://doi.org/10.1186/s12916-019-1426-2
31. Kundu, S. (2021). AI in medicine must be explainable. Nature Medicine, 27(8), 1328–1328.
32. Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artificial Intelligence, 296, 103473.
33. Lauritsen, S. M., Kristensen, M., Olsen, M. V., Larsen, M. S., Lauritsen, K. M., Jørgensen, M. J., Lange, J., & Thiesson, B. (2020). Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nature Communications, 11(1), 3852.
34. Lee, S. (2026). Explainable Artificial Intelligence in Medical Imaging: Explanations That Make Sense to Humans. American Journal of Roentgenology, AJR.26.35232. https://doi.org/10.2214/AJR.26.35232
35. Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is

both important and slippery. Queue, 16(3), 31–57. https://doi.org/10.1145/3236386.3241340
36. Loh, H. W., Ooi, C. P., Seoni, S., Barua, P. D., Molinari, F., & Acharya, U. R. (2022). Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). Computer Methods and Programs in Biomedicine, 226, 107161.
37. London, A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. Hastings Center Report, 49(1), 15–21. https://doi.org/10.1002/hast.973
38. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67.
39. Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.-W., Newman, S.-F., & Kim, J. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nature Biomedical Engineering, 2(10), 749–760.
40. Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. Journal of Biomedical Informatics, 113, 103655.
41. McCradden, M. D., Joshi, S., Mazwi, M., & Anderson, J. A. (2020). Ethical limitations of algorithmic fairness solutions in health care machine learning. The Lancet Digital Health, 2(5), e221–e223.
42. Mesinovic, M., Watkinson, P., & Zhu, T. (2025). Explainability in the age of large language models for healthcare. Communications Engineering, 4(1), 128.
43. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38.
44. Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. ACM Transactions on Interactive Intelligent Systems, 11(3–4), 1–45. https://doi.org/10.1145/3387166
45. Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. Digital Signal Processing, 73, 1–15.
46. Nicolson, A., Bradburn, E., Gal, Y., Papageorghiou, A. T., & Noble, J. A. (2025). The human factor in explainable artificial intelligence: Clinician variability in trust, reliance, and performance. Npj Digital Medicine, 8(1), 658.
47. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342
48. Parikh, R. B., Teeple, S., & Navathe, A. S. (2019). Addressing bias in artificial intelligence in health care. Jama, 322(24), 2377–2378.
49. Payrovnaziri, S. N., Chen, Z., Rengifo-Moreno, P., Miller, T., Bian, J., Chen, J. H., Liu, X., & He, Z. (2020). Explainable artificial intelligence models using real-world electronic health record data: A systematic scoping review. Journal of the American Medical Informatics Association, 27(7), 1173–1185.
50. Petch, J., Di, S., & Nelson, W. (2022). Opening the black box: The promise and limitations of explainable machine learning in cardiology. Canadian Journal of Cardiology, 38(2), 204–213.
51. Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and Measuring Model Interpretability. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–52. https://doi.org/10.1145/3411764.3445315
52. Reddy, S., Allan, S., Coghlan, S., & Cooper, P. (2020). A governance model for the application of AI in health care. Journal of the American Medical Informatics Association, 27(3), 491–497.
53. Tabassi, E. (2023). Artificial intelligence risk management framework (AI RMF 1.0). https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10?ref=decisionarts.io
54. the Precise4Q consortium, Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20(1), 310. https://doi.org/10.1186/s12911-020-01332-6
55. Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. PLoS Medicine, 15(11), e1002689.
56. Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. Information Fusion, 76, 89–106.