Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/



Research Paper Open Access

Explainable Artificial Intelligence in Personalized Patient Engagement and Digital Healthcare Decision-Making: A Comprehensive Study

Aasim Ghaffar[1], Vu H P Ho[2], Naisthika Kumar[3], Shahid Ali[4], Stefano Mancini[5]

[1]Department of Physics, Astronomy and Mathematics, University of Hertfordshire.
[2]Faculty of Artificial Intelligence, Universiti Teknologi Malaysia (UTM)
[3]Department of Cardio-thoracic Surgery, Royal Papworth Hospital, UK.
[4]Master in Data Science, Department of Computer Science, Avila University Kansus, Missouri.
[5]Department of Internal Medicine, IRCCS -AUSL of Reggio Emilia, Civil Hospital of Guastalla (RE).

Corresponding Author: Aasim Ghaffar , Email:aasimghaffar442925@gmail.com

Abstract
Artificial intelligence (AI) is increasingly used in digital healthcare to support patient engagement, personalized services and clinical decision-making. The opacity of many models reduces clinical trust, acceptance of patients and responsible deployment. This review focuses on how explainable artificial intelligence (XAI) can enhance the transparency, accountability and user trust for personalizing patient engagement and digital healthcare decision-making, and how interpretable machine learning (iML) can support these efforts. We discuss the possibility of integrating demographic, clinical, behavioral and digital interaction data with explanation methods in the context of personalized recommendations and patient-cented interventions. We then assess the evidence for XAI in clinical decision support, risk prediction, treatment personalization, patient communication and digital health platforms. The principles of interpretability, fairness, privacy, reliability, and human oversight are considered key pillars of trustworthy AI-driven care. Evidence suggests that explanations can play a part in facilitating collaboration among clinicians, patients, and AI systems, by providing reasons for predictions, but that explanations alone are insufficient to ensure appropriate levels of trust and safer decisions. In conclusion, we summarize key technical and ethical issues and identify the areas of AI research most likely to result in clinically responsible, patient-focused and transparent AI frameworks.

Keywords: Explainable artificial intelligence, Patient engagement, Digital health, Clinical decision support, Personalized healthcare, Trustworthy AI, Interpretability.

1. Introduction

AI is being adopted in hospitals and digital health applications all over the world. Medical images can be analyzed by machine learning models, electronic health records can be used for health risk prediction, and mobile apps and chatbots can support patients using machine learning (Haug & Drazen, 2023; Rajpurkar et al., 2022). Meanwhile, people's health and activities are recorded by wearable devices, patient portals and mobile apps. These technologies facilitate health care providers to gain insight into the needs of the individual and provide more individualized care. An important objective of precision medicine and personalized digital healthcare is this (Allen, 2024; Johnson et al., 2021).

But the customization of medicine relies on the AI models that are not always easy to understand. The doctors and patients may not understand why the system gives a particular recommendation; deep learning or other sophisticated AI techniques can make accurate predictions. This confusion can lead to lower levels of trust, increased risk of error and challenges when an issue arises regarding responsibility (Hulsen, 2023; Markus et al., 2021). To solve this issue, Explainable Artificial Intelligence (XAI) has recently attracted attention. XAI can be used to explain the reasoning behind an AI system's decision, the factors that influence its predictions, and how changing conditions might impact the outcomes. It is not only regulatory authorities that are monitoring explainability. The U.S. Food and Drug Administration has emphasized the need to interpret outputs from AI in its clinical decision support (CDS) draft guidance (Khera et al., 2023). In a comparable way, the EU's AI Act classifies some regulated medical products as

falling under the high-risk category, where they are subject to further obligations (Aboy et al., 2024; Busch et al., 2024).

While these advantages are present, there is ongoing debate regarding the effectiveness of AI explanations. There is a debate about whether existing explanation methods actually give a true understanding of the explanation of a particular decision that an AI system makes for a specific patient. Carefully testing and validating AI models to ensure safety for patients may be more important, they suggest (Ghassemi et al., 2021). Explanations did not mitigate the negative impacts of AI models that yielded biased outcomes in a study of 457 clinicians (Jabbour et al., 2023). The methods of XAI and their usage within medical care systems have been talked about in past reviews (Ali, Akhlaq, et al., 2023; Band et al., 2023; Loh et al., 2022). Other research has been directed at clinicians' belief in AI (Rosenbacke et al., 2024) and use of AI for clinical decision support (Salimparsa et al., 2025). Less attention has been given, however, to patient engagement, personalized care and clinical decision making, all at once. This is crucial because AI explanations should cater to the needs of various audiences, including patients, doctors, and other healthcare professionals.

This review seeks to fill this void by highlighting the key concepts and methodologies of eXplainable AI in Healthcare. It delves into the potential benefits of AI explanations for enhancing patient engagement and facilitating personalized digital health services. It also explores the functions of XAI in clinical decision-making, health risk prediction, and tailored therapy. Additionally, the review covers key topics including trust, fairness, privacy, and regulations that need to be taken into account for the safe and responsible use of AI in healthcare. Lastly, it outlines avenues for future research. The conceptual framework in figure 1 is used to cluster the main topics discussed in this review.

2. Review Approach

This study is narrative review, which is primarily based on research articles published from 2021 to 2025. The search terms used were "explainable artificial intelligence", "interpretability", "patient engagement", "digital health", "clinical decision support", "trust", "fairness", "privacy", and "regulation" in open scholarly sources, such as PubMed Central and publishers' websites. The papers we focused on were primarily systematic reviews, research studies involving doctors or patients, and papers on healthcare rules. Studies were accepted that directly addressed ways in which AI systems explain or how these explanations are assessed, or when they are applied in a healthcare setting. We did not combine the results statistically because the selected studies had different research methods, outcomes and areas of healthcare.

3. Conceptual Foundations of Explainable AI in Healthcare

3.1 Interpretability, explainability and explanation

The terms interpretability and explainability are often used with similar meanings in research papers. They're a bit different, however. An interpretable AI model is one that can be understood directly, for instance, a simple statistical model or a scoring system. An explainable AI system on the other hand, offers explanations of its predictions and decisions, typically by employing another technique after the model is trained (Markus et al., 2021).

In healthcare, the distinction holds significance as AI-driven decisions can directly influence the treatment and safety of the patients. Rudin et al. (2022) recommended the use of simple and understandable AI models whenever possible for crucial medical decisions. This is because explanations produced after making a prediction using a complex model could not represent the decision made by the model. This topic was also explored in the review by Markus et al. (2021) about “AI explainability in healthcare.” There was also scarce proof showing that these explanations are actually helpful in the practical world of health care, they said. Other reviews have also provided detail on the various explainable AI techniques and their key strengths and weaknesses (Ali, Abuhmed, et al., 2023; Dwivedi et al., 2023).

3.2 Families of explanation methods and their evaluation

Table 1 shows key method types for the explanations of AI predictions. Feature attribution methods, such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations), are commonly used with clinical data. These methods aim to uncover the most influential factors linked to patient health that impact an AI prediction (Band et al., 2023; Chaddad et al., 2023). If medical images are used, saliency maps can be used to enable the user to understand which parts of an image have an influence on the decision made by the model (van der Velden et al., 2022).

Yet another approach is called counterfactual explanation: it presents patients with information that differs in some minor way, and highlights how the change could result in a different prediction. This method may be useful for patients to comprehend what adjustments need to be made to enhance their health outcomes (Verma et al., 2024). Example-based methods: predictions are made by comparing them with other patient cases or other examples identified. Moreover, natural language explanations employ common language to

explain AI decisions, while some more recent work relies on large language models to produce such explanations (Thirunavukarasu et al., 2023).

But explanations are only helpful if they are accurate, clear, and reliable. It is still not easy to assess the quality. A systematic review revealed that many studies evaluated explanations based on poorly formulated evidence, e.g., personal observations, but not by strong evaluation procedures. Important aspects like accuracy, completeness and simplicity were highlighted in the review (Nauta et al., 2023). The same issues have been encountered in healthcare studies. Jung et al. (2023) conducted a literature review of 882 papers on explainable AI with clinical data, but found only six of them that satisfied their criteria for assessing the effectiveness of explanations. The discovery points to a major missing link – while there are numerous approaches to explain the actions of AI, very few have been rigorously evaluated to understand how effective they are for doctors and patients.

Table 1: **Main Families of Explanation Methods Used in Healthcare**

Approach	Typical techniques	Typical healthcare use	Main limitation	Sources
Intrinsically interpretable models	Sparse regression, rule lists, scoring systems	Risk scores, triage and screening tools	May need careful feature design; can be less suited to raw images or free text	Rudin et al., 2022; Markus et al., 2021
Feature attribution (post hoc)	SHAP, LIME	Risk and survival models on tabular records	Can be unstable and may mislead for individual patients	Band et al., 2023; Ghassemi et al., 2021; Moncada-Torres et al., 2021
Visual saliency	Grad-CAM, layer-wise relevance propagation	Medical image analysis	Shows where the model looked, not why it decided	van der Velden et al., 2022; Ghassemi et al., 2021
Counterfactual explanations	Minimal-change what-if scenarios	Patient-facing guidance; questions about what could change	Suggested changes may be infeasible or causally invalid	Verma et al., 2024
Example-based explanations	Similar-case retrieval, prototypes	Support for clinical reasoning by analogy	Depends on case selection and raises privacy concerns	Dwivedi et al., 2023; Chaddad et al., 2023
Natural-language rationales	Template or language-model text	Patient communication; summaries in decision support	Fluent text may not reflect the model's actual reasoning	Thirunavukarasu et al., 2023; Meskó & Topol, 2023

4. XAI for Personalized Patient Engagement and Digital Health Platforms

4.1 From heterogeneous data to individual explanations

There are various categories of information associated with personalized healthcare, such as demographic data, medical history, lifestyle, and engagement with digital health services. Figure 1 explains how this information is used by AI models to make predictions. These predictions are then explained by an explanation layer, that will help doctors and patients to understand those predictions. Fairness, privacy, reliability, human oversight, and regulations are other key elements that contribute to the responsible use of these systems.

There are new AI techniques that can integrate various health data types, like medical records, images, genetic information, and information captured by wearables. This allows AI systems to develop a more complete understanding of an individual's health (Acosta et al., 2022; Kline et al., 2022). A new approach being explored is the application of digital twins which are digital representations of individuals to simulate health conditions and explore potential outcomes (Katsoulakis et al., 2024).

But, when multiple data sources are brought together, AI models become more complex and the reason behind decisions more difficult to explain. A complex model, which may contain medical record, imagery, or genetic data, may not be as intuitive as a simple model that contains only one type of data. Allen (2024) conducted a review of 27 studies on digital health and precision medicine explainable machine learning. The review resulted in a conclusion that explanation methods still have to be improved and tested in order to have a reliable application for personalized healthcare.

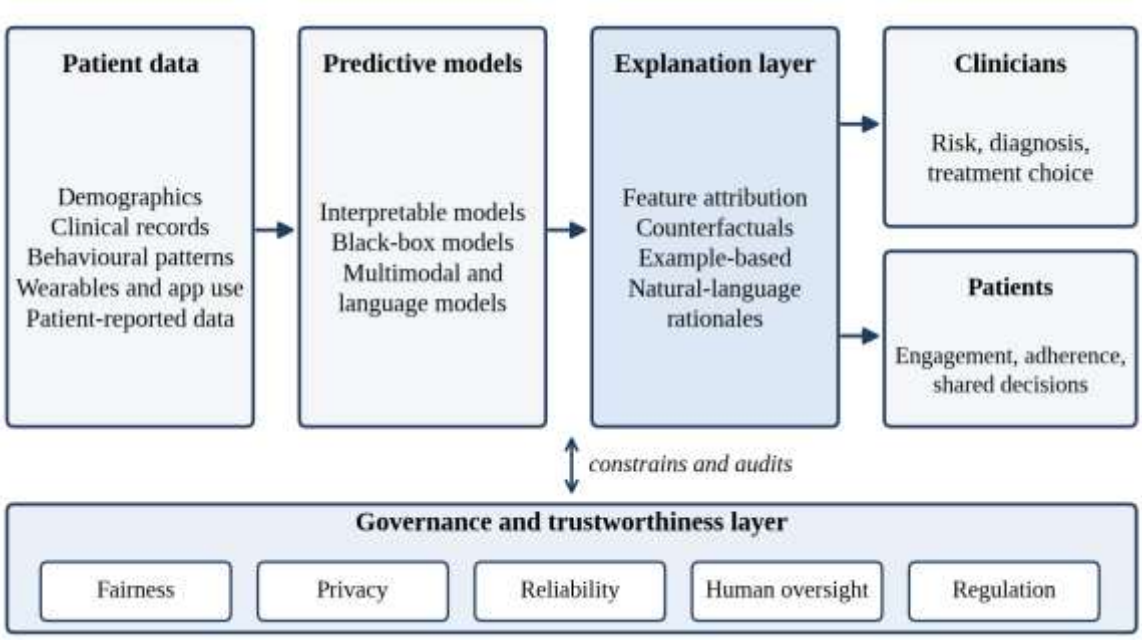


Figure 1: **Conceptual Framework for Explainable AI in Personalized Patient Engagement and Digital Healthcare Decision-Making**

4.2 Patient-facing explanation and communication

Engagement tools are the first and most obvious place where patients interact with AI. There is systematic evidence of the use of AI chatbots to support health behavior change, yet there are differences in the quality of the studies and in outcome measures (Aggarwal et al., 2023). The expectations have been raised even more with the advent of large language models. Large language models have raised expectations even higher. The licensed professionals' responses to questions posed in a public social media forum were compared to the chatbot's responses, and in most evaluations, the licensed professionals preferred the chatbot's answer over the physician's answer (Ayers et al., 2023). Fluent text is not the same as transparent reasoning, however. A generated rationale may be convincing but do not reflect the way the system arrives to its output; the use of these tools in care still raises suggestions about the need for regulation (Meskó & Topol, 2023; Thirunavukarasu et al., 2023). Generalist foundation models can start to create a disconnect between what it is capable of doing and what can be substantively checked. Generalist foundation models could start to create a disconnect between what they are capable of doing and what can be substantively checked.

Patients aren't just the less technical audience. Qualitative studies reveal that patients have specific worries about AI in care, such as safety, privacy and how this impacts their connection with the clinician (Richardson et al., 2021). An explanation that is adequate for a specialist (e.g., a list of the features of a model) will not answer these questions. Bienefeld et al. (2023) discovered that there are different mental models of the purpose of an explanation among both clinicians and developers and suggested personalized explanations as one design variation. Engagement tools might be appropriate for counterfactual statements around modifiable behaviors, such as how estimated risk would change based on different activities (Verma et al., 2024). That assumption should be tested in patients and the causal validity of any proposed change should be demonstrated before it is given as a suggestion for advice.

5. XAI in Clinical Decision Support, Risk Prediction and Treatment Personalization

Explainability has been researched most in the context of clinical decision support. Several studies have found that technical solutions have been created without considering the input from the clinicians who will use them (Antoniadi et al., 2021; Salimparsa et al., 2025). The three-stage process suggested by Salamparsa et al. (2025) starts by selecting methods that are user-centred, followed by co-design of the interface, and wraps up with iterative evaluation. Co-design studies demonstrate the need for this. Both the goals for XAI, sources of truth, and attitudes towards exploring new options varied between the two groups, with the 112 developers and clinicians developing a decision support tool for intensive care. While clinicians emphasized clinical plausibility and the patient in front of them, developers emphasized the interpretability of the model and the data (Bienefeld et al., 2023).

Does explanation raise trust? There is some evidence that clinician trust in XAI is higher than in standard AI, especially if explanations are clear, concise, and relevant to clinical practice, but complex or

contradictory explanations might lead to erosion of trust (Rosenbacke et al., 2024). The same review warns that trust in itself is not considered a good thing as AI tips can be incorrect.

The gain from explanation in risk prediction is sometimes analytic, too. Moncada-Torres et al. (2021) demonstrated that explainable machine learning model outperformed Cox regression for breast cancer survival and that, by feature-attribution, it yielded insight into the influence of patient characteristics on the prediction. The example showcases a typical scenario where an attribution approach is applied to ensure that a model is using clinically reasonable variables, and where it provides information to support the patient-level estimate.

Personalizing treatment is a greater challenge. Patient-specific factors must provide justification for a different recommendation, and explanations should include the factors that underlie an individualized recommendation, when applicable, to the clinician and (when appropriate) the patient in shared decision-making (Johnson et al, 2021; Kline et al, 2022). There are few systems that have been studied in the context of routine care. Jung et al. (2023) pointed out that there is lack of integration of XAI techniques in real-time patient care. There were a number of studies (93 in total) conducted, and a synthesis of these highlighted the importance of the tools such as LIME and SHAP in delivering feature importance (Ali, Akhlaq, et al., 2023), leaving feature attribution to have a significant practical load, and consequently, the known weaknesses are important. However, as Ghassemi et al. (2021) note, there are potential pitfalls of such explanations to the individual patient that are difficult to detect.

6. Trust, Human Oversight and Human–AI Collaboration

Though AI explanations are intended to assist people in making better decisions, their true value lies in how people use them in real-world situations. In an experiment, Jabbour et al. (2023) provided an explanation using AI to 457 clinicians who were evaluating patients with acute respiratory failure with and without AI explanation. The initial diagnostic accuracy was 73% for the clinicians. They had a 4.4% higher rate of accuracy when the AI predictions were provided with explanations as well. The accuracy dropped, however, when the AI system gave biased predictions that included explanations, by 11.3%. This demonstrates that explanations are not always enough to prevent errors and can sometimes result in wrong decisions.

Khera et al. (2023) pointed out that medical practitioners might overlook their own expertise and medical judgment and trust the AI recommendations. Hence, AI systems should be thoroughly evaluated in actual clinical settings before they become popular. The same results have been seen in mammography, where the AI suggestions proved to be inaccurate and consequently hamstrung the accuracy of the image interpretation. It was found that less experienced readers were more affected than experienced professionals (Dratsch et al., 2023). The results highlight the need for human oversight, consideration, and assessment in the application of AI in the medical field.

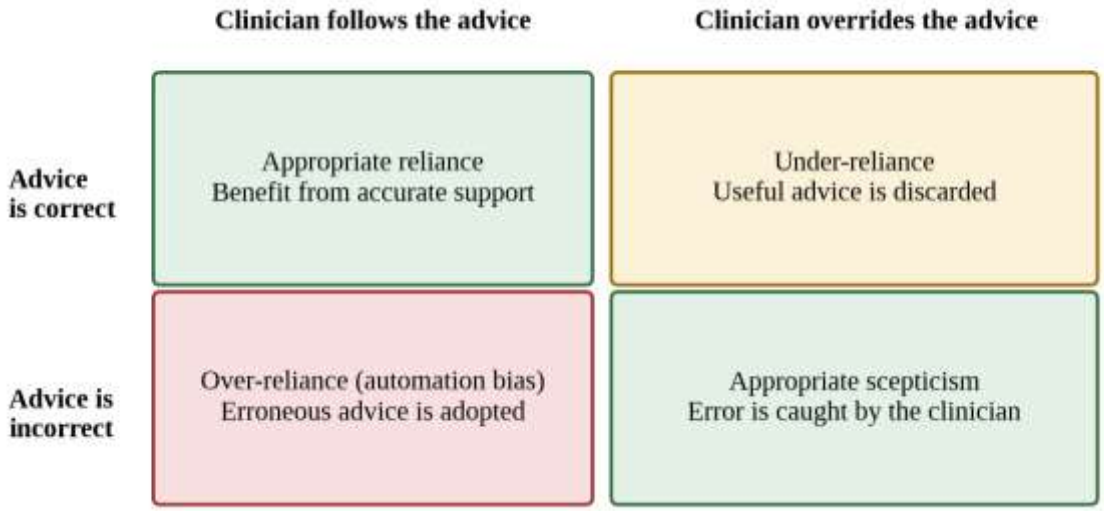


Figure 2: Trust Calibration in Clinician–AI Decision-Making

The primary objective of AI in healthcare is to enable medical professionals to use AI recommendations only when they are accurate, as illustrated in Figure 2. The goal is not to tell doctors to take the advice of every AI prediction, but to tell them when they can rely on it. Doctors might be able to identify good recommendations when they are presented in clear and clinically useful explanations (Rosenbacke et al., 2024). But explanations do not necessarily ensure an AI model is accurate or unbiased. Ghassemi et al (2021) posited that using the proper datasets and in the appropriate clinical environment to test AI models

is a more direct route for increased safety. They also cautioned against requiring as a prerequisite the explainability for all AI solutions. This debate demonstrates that there remains disagreement among researchers as to the level of explanation required in the context of healthcare (Reddy, 2022). The results of these findings provide valuable insights into the design of AI systems. Interfaces should show the uncertainty of predictions, provide the ability for physicians to challenge AI suggestions and provide information in a way that facilitates clinical decision-making. Another factor is training - as different doctors have different levels of experience, they might respond differently to AI recommendations (Dratsch et al., 2023).).

7. Fairness, Privacy and Regulation

a. Fairness: Unfair AI models can result in unfair treatment and poor health outcomes for patients. Bias can be introduced in a model via the data it is trained on, or via population differences in the data, or via label inconsistencies in the data (Chen et al., 2023). Various approaches have been proposed to detect and mitigate these issues, such as fairness metrics, bias mitigation techniques, and AI model auditing tools (Xu et al., 2022). When evaluating AI models, one way that Ex explainable AI can assist is in determining if there are factors associated with protected characteristics like race or sex that are used in making predictions. In the effort to decrease bias, Chen et al. (2023) have identified explainability, federated learning, and disentanglement as potential solutions. But, explanations are not always obvious when doctors do not know what the problem is (Jabbour et al., 2023).

b. Personalization: Access to sensitive patient data is crucial for personalized healthcare, raising significant privacy concerns. AI explanations can also provide detailed information regarding the model or the data that the model was trained on. One potential solution is federated learning, which lets AI models learn from data that is located at various places without regularly moving the original data to a central server. The authors of the systematic review of 612 articles reported that the majority of them were still at proof of concept, and just 5.2% of the studies had real-world applications (Teo et al., 2024). The benefits and drawbacks of federated learning and other privacy-preserving techniques, such as impact on accuracy, computational needs, and system complexity, have been discussed in other reviews (Khalid et al., 2023; Nguyen et al., 2022).

c. Regulation: An all-in-one law for AI regulation (Busch et al., 2024). In the medical device industry, this law is complemented by the medical device regulations and data protection laws. Some regulated medical products AI systems are considered “high-risk” and have further legal requirements (Aboy et al., 2024). It is vital to identify the kind of explanation required, the level of detail, and if the explanations are appropriate for whom, among others, doctors and patients. But, explainability is not sufficient to make an AI system trustworthy. Other factors are also critical, such as reliability, fairness, privacy, and responsibility (Ali, Abuhmed, et al., 2023). The primary challenges are highlighted and solutions are proposed in Table 2.

Table 2: **Principal Challenges for Trustworthy XAI in Digital Healthcare and Practical Responses**

Challenge	Why it matters	Practical response	Sources
Evaluating explanations	Few studies test whether explanations help users	Task-grounded, human-centered evaluation with shared properties	Nauta et al., 2023; Jung et al., 2023
Over-reliance on AI	Biased or wrong advice can lower accuracy	Trust calibration, uncertainty display, local validation, training	Jabbour et al., 2023; Dratsch et al., 2023; Khera et al., 2023
Stakeholder mismatch	Clinicians, developers and patients want different things	Co-design and personalized explanations	Bienefeld et al., 2023; Salimparsa et al., 2025
Algorithmic bias	Disparities arise in data, labels and deployment	Subgroup audits, mitigation methods, explanation-assisted checks	Chen et al., 2023; Xu et al., 2022
Privacy	Personalization needs sensitive data	Federated and other privacy-preserving learning; review of what explanations disclose	Teo et al., 2024; Khalid et al., 2023; Nguyen et al., 2022

Regulatory compliance	High-risk obligations apply to medical AI	Documented transparency, oversight and risk management	Aboy et al., 2024; Busch et al., 2024
Language-model outputs	Fluent rationales can be unfaithful	Regulatory oversight, verification of evidence, human review	Meskó & Topol, 2023; Thirunavukarasu et al., 2023

8. Research Gaps, Implications and Limitations

8.1 Research priorities

The results of this review identify some key points for future research. This involves difficulties in evaluation, healthcare management, legal issues, and human relations and communication (Bharati et al., 2024). First, the explainable AI (XAI) techniques must be validated in real health care environments. Very few have been applied in clinical practice (Jung et al., 2023). The same issues are present in federated learning, where numerous studies have yet to transition into clinical practice (Teo et al., 2024). Further studies should be directed to enhancing the outcomes of the patient/clinical decision making in addition to measuring user satisfaction. Second, more effective ways of assessing explanations are required. To evaluate their accuracy and clarity, impact on task performance, appropriate trust in AI, and the capability to discover and fix errors (Nauta et al., 2023; Jung et al., 2023). Thirdly, explanations need to be tailored to the needs of patients, taking into account their cultural background, language and health knowledge. Researchers need to enhance counterfactual and causal explanations to make sure that their recommendations are medically acceptable and dependable (Bienefeld et al., 2023; Verma et al., 2024). Four, the rise of the multimodal AI and large foundation models poses new challenges. They should be the subject of further research to explain their outputs accurately, and to identify evidence for recommendations (Meskó & Topol, 2023, Moor et al., 2023). Lastly, explainability should be a part of managing the AI system, such as model validation, ongoing surveillance, and regulatory documentation. Longo et al. (2024) highlighted the need for interdisciplinary work to tackle the unsolved issues in this area. Such a strategy is relevant in healthcare, where effective explanations are required to take into account the needs of patients, healthcare professionals, technical experts, and legal authorities.

8.2 Implications for practice

For developers and health systems, the evidence points to a short set of practices. Choose the simplest model that meets the clinical need, and prefer an interpretable model where performance is comparable (Rudin et al., 2022). Design explanation formats with the intended users, treating clinicians and patients as separate audiences, and test them in the setting of use (Salimparsa et al., 2025). Validate models locally and audit them across patient subgroups before relying on explanations as a safeguard (Chen et al., 2023; Ghassemi et al., 2021). Make it easy to record and review the cases in which clinicians disagree with the AI. Finally, document the explanation methods used and their known limits as part of regulatory and governance files (Aboy et al., 2024).

8.3 Limitations of this review

This review has limits. It is narrative rather than systematic, so the selection of sources reflects the authors' judgment, and no formal risk-of-bias appraisal was applied. Much of the evidence discussed is retrospective or vignette-based, and some fields, notably imaging and critical care, are better represented than others. Research on patient-facing explanation is sparse, and findings from vignette studies may not carry over to routine practice. The rapid growth of language-model applications also means that parts of this literature will date quickly. The conclusions should therefore be read as a synthesis of current evidence and not as a quantitative estimate of effect.

9. Conclusion

Explanable AI (XAI) is a technique that aids in interpreting the rationale behind a particular prediction or recommendation made by an AI system. In the context of personalized digital healthcare, where the effectiveness of AI is both about its predictive power and its ability to be trusted, this is crucial. The reviewed research in this paper demonstrates advances in explanation methods, engaging healthcare practitioners in system design and enhanced regulatory focus. Several issues remain however. There are numerous methods of explanation and the results of these methods may be different for each individual patient. Other studies have demonstrated that explanations are not always sufficient to stop doctors from implementing AI recommendations that have a bias. Furthermore, patients have not been as well studied as healthcare workers in previous research. Explaining digital health care is essential for transparency, but is not the only guarantee of safety and reliable decision making. It should be used in conjunction with correct model validations, fairness evaluations, privacy safeguards, human intervention, and medical specialists and patients. The integration of these methods and their effects on clinical decisions and patient

outcomes need to be further explored and analyzed in future studies, to ensure AI-powered healthcare becomes more personalized, reliable, and trusted.

References

1. Aboy, M., Minssen, T., & Vayena, E. (2024). Navigating the EU AI Act: Implications for regulated digital medical products. npj Digital Medicine, 7, Article 237. https://doi.org/10.1038/s41746-024-01232-3
2. Acosta, J. N., Falcone, G. J., Rajpurkar, P., & Topol, E. J. (2022). Multimodal biomedical AI. Nature Medicine, 28(9), 1773–1784. https://doi.org/10.1038/s41591-022-01981-2
3. Aggarwal, A., Tam, C. C., Wu, D., Li, X., & Qiao, S. (2023). Artificial intelligence–based chatbots for promoting health behavioral changes: Systematic review. Journal of Medical Internet Research, 25, Article e40789. https://doi.org/10.2196/40789
4. Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. Information Fusion, 99, Article 101805. https://doi.org/10.1016/j.inffus.2023.101805
5. Ali, S., Akhlaq, F., Imran, A. S., Kastrati, Z., Daudpota, S. M., & Moosa, M. (2023). The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review. Computers in Biology and Medicine, 166, Article 107555. https://doi.org/10.1016/j.compbiomed.2023.107555
6. Allen, B. (2024). The promise of explainable AI in digital health for precision medicine: A systematic review. Journal of Personalized Medicine, 14(3), Article 277. https://doi.org/10.3390/jpm14030277
7. 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), Article 5088. https://doi.org/10.3390/app11115088
8. Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Internal Medicine, 183(6), 589–596. https://doi.org/10.1001/jamainternmed.2023.1838
9. Band, S. S., Yarahmadi, A., Hsu, C.-C., Biyari, M., Sookhak, M., Ameri, R., Dehzangi, I., Chronopoulos, A. T., & Liang, H.-W. (2023). Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. Informatics in Medicine Unlocked, 40, Article 101286. https://doi.org/10.1016/j.imu.2023.101286
10. Bharati, S., Mondal, M. R. H., & Podder, P. (2024). A review on explainable artificial intelligence for healthcare: Why, how, and when? IEEE Transactions on Artificial Intelligence, 5(4), 1429–1442. https://doi.org/10.1109/TAI.2023.3266418
11. Bienefeld, N., Boss, J. M., Lüthy, R., Brodbeck, D., Azzati, J., Blaser, M., Willms, J., & Keller, E. (2023). Solving the explainable AI conundrum by bridging clinicians' needs and developers' goals. npj Digital Medicine, 6, Article 94. https://doi.org/10.1038/s41746-023-00837-4
12. Busch, F., Kather, J. N., Johner, C., Moser, M., Truhn, D., Adams, L. C., & Bressem, K. K. (2024). Navigating the European Union Artificial Intelligence Act for healthcare. npj Digital Medicine, 7, Article 210. https://doi.org/10.1038/s41746-024-01213-6
13. Chaddad, A., Peng, J., Xu, J., & Bouridane, A. (2023). Survey of explainable AI techniques in healthcare. Sensors, 23(2), Article 634. https://doi.org/10.3390/s23020634
14. Chen, R. J., Wang, J. J., Williamson, D. F. K., Chen, T. Y., Lipkova, J., Lu, M. Y., Sahai, S., & Mahmood, F. (2023). Algorithmic fairness in artificial intelligence for medicine and healthcare. Nature Biomedical Engineering, 7(6), 719–742. https://doi.org/10.1038/s41551-023-01056-8
15. Dratsch, T., Chen, X., Rezazade Mehrizi, M., Kloeckner, R., Mähringer-Kunz, A., Püsken, M., Baeßler, B., Sauer, S., Maintz, D., & Pinto dos Santos, D. (2023). Automation bias in mammography: The impact of artificial intelligence BI-RADS suggestions on reader performance. Radiology, 307(4), Article e222176. https://doi.org/10.1148/radiol.222176
16. Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. ACM Computing Surveys, 55(9), Article 194. https://doi.org/10.1145/3561048
17. 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. https://doi.org/10.1016/S2589-7500(21)00208-9
18. Haug, C. J., & Drazen, J. M. (2023). Artificial intelligence and machine learning in clinical medicine, 2023. New England Journal of Medicine, 388(13), 1201–1208. https://doi.org/10.1056/NEJMra2302038

19. Hulsen, T. (2023). Explainable artificial intelligence (XAI): Concepts and challenges in healthcare. AI, 4(3), 652–666. https://doi.org/10.3390/ai4030034
20. Jabbour, S., Fouhey, D., Shepard, S., Valley, T. S., Kazerooni, E. A., Banovic, N., Wiens, J., & Sjoding, M. W. (2023). Measuring the impact of AI in the diagnosis of hospitalized patients: A randomized clinical vignette survey study. JAMA, 330(23), 2275–2284. https://doi.org/10.1001/jama.2023.22295
21. Johnson, K. B., Wei, W.-Q., Weeraratne, D., Frisse, M. E., Misulis, K., Rhee, K., Zhao, J., & Snowdon, J. L. (2021). Precision medicine, AI, and the future of personalized health care. Clinical and Translational Science, 14(1), 86–93. https://doi.org/10.1111/cts.12884
22. Jung, J., Lee, H., Jung, H., & Kim, H. (2023). Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. Heliyon, 9(5), Article e16110. https://doi.org/10.1016/j.heliyon.2023.e16110
23. Katsoulakis, E., Wang, Q., Wu, H., Shahriyari, L., Fletcher, R., Liu, J., Achenie, L., Liu, H., Jackson, P., Xiao, Y., Syeda-Mahmood, T., Tuli, R., & Deng, J. (2024). Digital twins for health: A scoping review. npj Digital Medicine, 7, Article 77. https://doi.org/10.1038/s41746-024-01073-0
24. Khalid, N., Qayyum, A., Bilal, M., Al-Fuqaha, A., & Qadir, J. (2023). Privacy-preserving artificial intelligence in healthcare: Techniques and applications. Computers in Biology and Medicine, 158, Article 106848. https://doi.org/10.1016/j.compbiomed.2023.106848
25. Khera, R., Simon, M. A., & Ross, J. S. (2023). Automation bias and assistive AI: Risk of harm from AI-driven clinical decision support. JAMA, 330(23), 2255–2257. https://doi.org/10.1001/jama.2023.22557
26. Kline, A., Wang, H., Li, Y., Dennis, S., Hutch, M., Xu, Z., Wang, F., Cheng, F., & Luo, Y. (2022). Multimodal machine learning in precision health: A scoping review. npj Digital Medicine, 5, Article 171. https://doi.org/10.1038/s41746-022-00712-8
27. 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, Article 107161. https://doi.org/10.1016/j.cmpb.2022.107161
28. Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Páez, A., Samek, W., Schneider, J., Speith, T., & Stumpf, S. (2024). Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. Information Fusion, 106, Article 102301. https://doi.org/10.1016/j.inffus.2024.102301
29. 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, Article 103655. https://doi.org/10.1016/j.jbi.2020.103655
30. Meskó, B., & Topol, E. J. (2023). The imperative for regulatory oversight of large language models (or generative AI) in healthcare. npj Digital Medicine, 6, Article 120. https://doi.org/10.1038/s41746-023-00873-0
31. Moncada-Torres, A., van Maaren, M. C., Hendriks, M. P., Siesling, S., & Geleijnse, G. (2021). Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. Scientific Reports, 11, Article 6968. https://doi.org/10.1038/s41598-021-86327-7
32. Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. Nature, 616(7956), 259–265. https://doi.org/10.1038/s41586-023-05881-4
33. Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., & Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. ACM Computing Surveys, 55(13s), Article 295. https://doi.org/10.1145/3583558
34. Nguyen, D. C., Pham, Q.-V., Pathirana, P. N., Ding, M., Seneviratne, A., Lin, Z., Dobre, O., & Hwang, W.-J. (2022). Federated learning for smart healthcare: A survey. ACM Computing Surveys, 55(3), Article 60. https://doi.org/10.1145/3501296
35. Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31–38. https://doi.org/10.1038/s41591-021-01614-0
36. Reddy, S. (2022). Explainability and artificial intelligence in medicine. The Lancet Digital Health, 4(4), e214–e215. https://doi.org/10.1016/S2589-7500(22)00029-2
37. Richardson, J. P., Smith, C., Curtis, S., Watson, S., Zhu, X., Barry, B., & Sharp, R. R. (2021). Patient apprehensions about the use of artificial intelligence in healthcare. npj Digital Medicine, 4, Article 140. https://doi.org/10.1038/s41746-021-00509-1

38. Rosenbacke, R., Melhus, Å., McKee, M., & Stuckler, D. (2024). How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: Systematic review. JMIR AI, 3, Article e53207. https://doi.org/10.2196/53207
39. Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistics Surveys, 16, 1–85. https://doi.org/10.1214/21-SS133
40. Salimparsa, M., Sedig, K., Lizotte, D. J., Abdullah, S. S., Chalabianloo, N., & Muanda, F. T. (2025). Explainable AI for clinical decision support systems: Literature review, key gaps, and research synthesis. Informatics, 12(4), Article 119. https://doi.org/10.3390/informatics12040119
41. Teo, Z. L., Jin, L., Liu, N., Li, S., Miao, D., Zhang, X., Ng, W. Y., Tan, T. F., Lee, D. M., Chua, K. J., Heng, J., Liu, Y., Goh, R. S. M., & Ting, D. S. W. (2024). Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture. Cell Reports Medicine, 5(2), Article 101419. https://doi.org/10.1016/j.xcrm.2024.101419
42. Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. Nature Medicine, 29(8), 1930–1940. https://doi.org/10.1038/s41591-023-02448-8
43. van der Velden, B. H. M., Kuijf, H. J., Gilhuijs, K. G. A., & Viergever, M. A. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Medical Image Analysis, 79, Article 102470. https://doi.org/10.1016/j.media.2022.102470
44. Verma, S., Boonsanong, V., Hoang, M., Hines, K., Dickerson, J., & Shah, C. (2024). Counterfactual explanations and algorithmic recourses for machine learning: A review. ACM Computing Surveys, 56(12), Article 312. https://doi.org/10.1145/3677119
45. Xu, J., Xiao, Y., Wang, W. H., Ning, Y., Shenkman, E. A., Bian, J., & Wang, F. (2022). Algorithmic fairness in computational medicine. eBioMedicine, 84, Article 104250. https://doi.org/10.1016/j.ebiom.2022.104250