Beyond the black box: interpretability, accountability, and responsible clinical integration of AI-driven heart rate variability models-a narrative

frontiers in Computer Science

TYPE Review
PUBLISHED 13 May 2026
DOI 10.3389/fcomp.2026.1847389

OPEN ACCESS

EDITED BY
Nicholas Gerald Murray,
University of Nevada, United States

REVIEWED BY
Francisco Epelde,
Parc Taulí Foundation, Spain
Tambi Medabala,
Netaji Subhas National Institute of Sports, India

*CORRESPONDENCE
Oana Geman
✉ geman@chalmers.se

RECEIVED 04 April 2026
REVISED 19 April 2026
ACCEPTED 28 April 2026
PUBLISHED 13 May 2026

CITATION
Burlacu A, Olariu M, Geman O, Iftene A, Bogdan-Goroftei R-E and Brinza C (2026) Beyond the black box: interpretability, accountability, and responsible clinical integration of AI-driven heart rate variability models—a narrative review.
Front. Comput. Sci. 8:1847389.
doi: 10.3389/fcomp.2026.1847389

Beyond the black box: interpretability, accountability, and responsible clinical integration of AI-driven heart rate variability models—a narrative review

Alexandru Burlacu[1,2], Maria Olariu[3], Oana Geman[4]*, Adrian Iftene[3], Roxana-Elena Bogdan-Goroftei[5] and Crischentian Brinza[1]

[1]Faculty of Medicine, University of Medicine and Pharmacy "Grigore T Popa", Iasi, Romania, [2]Institute of Cardiovascular Diseases "Prof. Dr. George I.M. Georgescu", Iasi, Romania, [3]Department of Computer Science, Faculty of Informatics, Alexandru Ioan Cuza University, Iasi, Romania, [4]Division of Data Science and Artificial Intelligence, Computer Science and Engineering, Chalmers University of Technology, Gothenburg University, Gothenburg, Sweden, [5]Faculty of Medicine, "Dunarea de Jos" University, Galati, Romania

Background: Heart rate variability (HRV) is a widely used digital biomarker reflecting autonomic regulation and has been associated with diverse cardiovascular, critical care, and stress-related outcomes. In parallel, AI and machine-learning methods have expanded rapidly in HRV-based prediction, often achieving strong predictive performance. However, clinical translation remains constrained by limited interpretability, unclear accountability, and challenges in workflow integration, particularly for black-box models used in high-stakes settings.
Methods: This narrative, concept-driven review examined conceptual, methodological, clinical, and governance dimensions of interpretability in AI-driven HRV prediction. A structured literature search was performed in PubMed/MEDLINE, Scopus, and Embase databases.
Results: The analysis showed that interpretability is not a binary property but varies by model design and deployment context. Post-hoc explainability methods may increase transparency, yet they can also be unstable, incomplete, or misleading, with potential to increase automation bias. Clinical adoption is further limited by signal-quality variability (ECG vs. wearable PPG), insufficient external validation, workflow misalignment, and unclear medico-legal responsibility. A four-step pragmatic implementation framework is proposed: data governance and signal integrity; robust model development and validation; workflow-compatible clinical integration with human oversight; and continuous post-deployment monitoring and governance.
Conclusion: HRV-AI systems should be treated as socio-technical interventions. Responsible adoption requires proportional transparency, explicit accountability structures, and lifecycle governance beyond predictive accuracy alone.

KEYWORDS
artificial intelligence, clinical decision support systems, explainable AI, heart rate variability, model governance

1 Introduction

Heart rate variability (HRV) has emerged as an important digital biomarker reflecting the dynamic interplay between the sympathetic and parasympathetic branches of the autonomic nervous system (Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology, 1996; Shaffer and Ginsberg, 2017). Unlike static physiological measures, HRV captures complex, time-dependent regulatory processes that integrate cardiovascular, metabolic, inflammatory, and neurohumoral signals (Thayer et al., 2010). Over the past two decades, HRV metrics derived from short-term and long-term ECG recordings have been associated with a wide range of clinical outcomes, including cardiovascular morbidity and mortality, autonomic dysfunction, critical illness evolution, and stress-related conditions (Tsuji et al., 1996; Buccelletti et al., 2009; Ahmad et al., 2009; Chalmers et al., 2014). This physiological richness has positioned HRV as an attractive target for advanced computational modeling aimed at prediction rather than mere description (Faust et al., 2018).

In parallel with the growing clinical interest in HRV, artificial intelligence (AI) and machine-learning (ML) techniques have rapidly expanded within HRV analytics (Acharya et al., 2017; Clifford et al., 2017). Contemporary studies increasingly employ complex models, such as ensemble methods, deep neural networks, and hybrid architectures, to predict outcomes ranging from arrhythmia detection and acute deterioration to long-term prognosis and behavioral or psychological states. These approaches frequently demonstrate impressive predictive performance, often outperforming traditional statistical models and established clinical scores. As a result, HRV-based AI systems are progressively framed as decision-support tools with potential applicability across preventive, acute, and critical care settings (Hannun et al., 2019; Kwon et al., 2018).

Despite these advances, the clinical translation of AI-driven HRV prediction is constrained by a fundamental tension between predictive performance and interpretability. Many of the most accurate models operate as so-called "*black boxes*," offering limited insight into how specific HRV features contribute to predictions or how outputs relate to known physiological mechanisms. While opacity may be acceptable in purely technical contexts, clinical decision-making relies on transparency, causal plausibility, and the ability to integrate algorithmic outputs into structured medical reasoning. Consequently, high performance alone is insufficient to guarantee clinical trust, adoption, or responsible use (Rudin, 2019; Castelvecchi, 2016; Wiens et al., 2019).

The lack of interpretability in HRV-based black-box models raises practical challenges for clinicians, institutions, and healthcare systems. Opaque predictions complicate bedside decision-making, hinder meaningful communication with patients, and generate uncertainty regarding medico-legal responsibility when algorithmic recommendations influence care. These concerns are amplified in contexts where HRV-based predictions could influence high-stakes interventions or resource allocation (Rudin, 2019; Wiens et al., 2019; Kelly et al., 2019; Price et al., 2019).

Beyond clinical practicality, black-box HRV models introduce ethical and governance dilemmas related to accountability, automation bias, and asymmetries between algorithm developers and end-user clinicians. HRV reflects sensitive physiological dynamics related not only to cardiovascular function but also to stress and neuroautonomic regulation. As HRV analytics increasingly extend to wearable monitoring, occupational health, and high-risk operational environments, opaque AI systems may generate consequences that extend beyond traditional clinical care. In such settings, proportional transparency and clearly defined responsibility frameworks become essential for maintaining trust and enabling responsible deployment (Thayer et al., 2010; Price et al., 2019; London, 2019).

Against this background, the present narrative review critically examines the role of interpretability in AI-driven HRV prediction, moving beyond a simplistic dichotomy between black-box and interpretable models (Figure 1). Particular attention is given to the

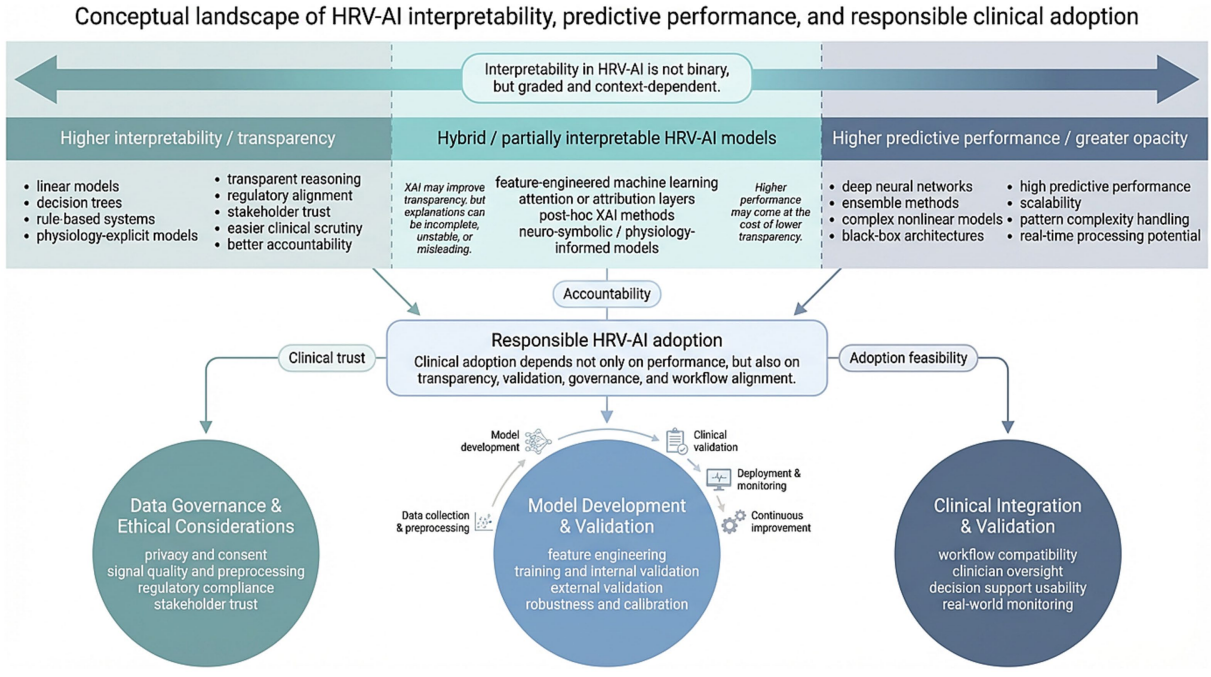

FIGURE 1
Conceptual landscape of HRV-AI interpretability.

emerging field of explainable artificial intelligence (XAI), which seeks to enhance transparency through post-hoc or intrinsically interpretable techniques without necessarily sacrificing predictive performance.

We aim to clarify the conceptual and practical limits of explainability, distinguishing between technical explanations and clinically meaningful understanding. Furthermore, this review explores when XAI approaches may genuinely support trust, accountability, and clinical integration, and when they risk providing only superficial reassurance. Building on this analysis, we propose a pragmatic implementation framework designed to guide the responsible clinical adoption of HRV-based AI systems across diverse healthcare contexts.

Although this review engages with broader literature on medical AI, explainability, human factors, and governance, its primary focus is HRV-based AI systems. HRV poses distinctive translational challenges because model behavior may be shaped not only by algorithmic architecture, but also by signal modality, preprocessing decisions, recording conditions, physiological context, and the interpretive ambiguity of certain feature families. Accordingly, examples from adjacent domains are used here to contextualize implementation issues, but they should not be interpreted as substitutes for HRV-specific evidence.

2 Methodological approach and scope

This article was conducted as a narrative, concept-driven critical review. The objective was not to perform a systematic review or quantitative meta-analysis, but to examine conceptual, methodological, clinical, and governance dimensions of interpretability in AI-driven heart rate variability (HRV) prediction. The methodological approach was designed to ensure transparency of the literature identification process, and analytical rigor, while acknowledging the heterogeneity inherent to HRV analytics.

2.1 Search strategy and time frame

A structured literature search was performed in PubMed/MEDLINE, Scopus, and Embase databases. The search was conducted between 1 November 2025 and 31 December 2025. Search terms included combinations of the following keywords: "heart rate variability," "HRV," "machine learning," "artificial intelligence," "deep learning," "neural network," "random forest," "gradient boosting," "support vector machine," "explainable AI," "interpretable machine learning," and "black-box model." Boolean operators were used to combine physiological and methodological terms. No language restrictions were applied during the initial search. Reference lists of relevant articles were manually screened to identify additional publications of conceptual or methodological relevance.

The database search yielded 812 records, with an additional 14 records identified through reference screening. After removal of duplicates, 536 unique records were screened by title and abstract. The final narrative synthesis was informed by 53 publications considered directly relevant to one or more of the following domains: (1) HRV-based AI prediction studies; (2) interpretability or explainability methods relevant to healthcare AI; (3) clinical implementation, human factors, or workflow integration issues; and (4) ethical, legal, or governance analyses relevant to high-stakes clinical AI.

2.2 Study selection and prioritization

Two reviewers independently screened titles and abstracts for relevance. Articles were considered for full-text evaluation if they met at least one of the following criteria: (1) original research applying AI or machine-learning approaches to HRV-based prediction tasks; (2) methodological studies addressing model interpretability, explainability, or transparency in healthcare AI; (3) ethical, legal, or governance analyses concerning opaque AI systems in clinical decision support; or (4) foundational physiological standards relevant to HRV interpretation. Because this article was designed as a narrative, concept-driven synthesis rather than a systematic review, study selection was not intended to support exhaustive quantitative comparison. Instead, publications were prioritized when they were conceptually informative, methodologically representative, or clinically relevant to the central question of how interpretability affects the responsible deployment of HRV-based AI systems. Disagreements regarding inclusion were resolved through discussion and consensus. Formal risk-of-bias tools were not applied, however, methodological rigor, validation strategy, reporting transparency, and relevance to the review objectives were considered during study selection and interpretation.

2.3 Data extraction and analytical synthesis

For included studies, data were extracted independently by the two reviewers using a predefined framework. When available, extracted elements included HRV feature domains (time-domain, frequency-domain, and nonlinear indices), signal source and acquisition context (e.g., ECG versus wearable PPG), modeling architecture, prediction task, validation strategy, performance metrics (e.g., AUROC, accuracy, sensitivity, specificity), and information relevant to model transparency, explainability, or black-box behavior.

Due to substantial heterogeneity in signal acquisition protocols, preprocessing pipelines, feature engineering strategies, outcome definitions, validation approaches, and performance reporting, quantitative pooling was not considered methodologically appropriate. Instead, studies were synthesized narratively and comparatively, with emphasis on four complementary dimensions: (2) degree of model transparency and architectural complexity; (1) clinical context and risk profile of deployment; (3) ethical and governance implications, including accountability and automation bias; and (4) translational feasibility and integration into clinical reasoning.

The final body of literature included in the narrative synthesis comprised studies spanning several recurring HRV-AI application domains, including rhythm-related classification, deterioration and risk prediction, autonomic or stress-related state assessment, wearable monitoring, and broader translational or governance-oriented analyses relevant to HRV deployment. Across this body of literature, the most frequent signal modalities were ECG-derived HRV and wearable PPG-derived variability signals, while the most commonly encountered model families included conventional machine-learning algorithms, ensemble methods, and deep-learning approaches. Interpretability strategies, when reported, ranged from intrinsically interpretable models to post-hoc explainability methods, although the depth of reporting, validation practices, and deployment relevance varied substantially across studies.

3 Conceptual definitions

The debate surrounding artificial intelligence in heart rate variability (HRV) prediction is frequently framed in terms of "black-box" versus "interpretable" models. However, these terms are often used imprecisely, obscuring important conceptual distinctions. In the context of HRV analytics, a black-box model refers to a computational system whose internal decision-making processes are not directly accessible or understandable to human users, particularly clinicians (Rudin, 2019; Topol, 2019). Such models—often based on deep neural networks, complex ensemble methods, or high-dimensional nonlinear transformations—can generate highly accurate predictions while providing limited direct insight into how specific HRV features or physiological patterns contribute to the output (Topol, 2019; Tjoa and Guan, 2021).

By contrast, interpretable approaches typically refer to models whose structure allows users to trace how inputs are transformed into predictions. Examples include linear regression models with clearly defined coefficients, decision trees with explicit branching logic, or rule-based systems grounded in predefined thresholds (Rudin, 2019; McKelvey et al., 2018). In HRV research, such models may align more directly with established physiological constructs—for instance, distinguishing between time-domain, frequency-domain, or non-linear indices in a manner that preserves conceptual continuity with autonomic regulation theory (Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology, 1996; Shaffer and Ginsberg, 2017). Importantly, interpretability in this sense does not necessarily imply simplicity, but rather structural transparency and retraceability (Rudin, 2019).

While the distinction between black-box and interpretable models is analytically useful, it may risk oversimplifying a more nuanced reality. Many contemporary systems incorporate layers of varying transparency, combining interpretable components with opaque transformations, or applying post-hoc explanation techniques to otherwise complex architectures (Tjoa and Guan, 2021; Ribeiro et al., 2016). Consequently, it may be more appropriate to consider interpretability not as a strictly binary attribute, but as a property that varies in degree depending on model design, feature engineering, and contextual deployment (Ribeiro et al., 2016). Rather than asking whether a model is interpretable or not, a more clinically relevant question may be how interpretable it is, for whom, and under which conditions of use (Floridi et al., 2018).

Within HRV analytics, this gradation becomes particularly salient. A model that provides global feature importance scores may offer a different level of insight than one that delivers patient-specific explanations, temporal attribution maps, or counterfactual scenarios. Similarly, a system that reveals statistical associations between HRV metrics and outcomes may still obscure mechanistic pathways linking autonomic physiology to clinical events (Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology, 1996; Tjoa and Guan, 2021). The practical meaning of interpretability therefore depends not only on mathematical transparency, but also on the alignment between model outputs and established physiological knowledge. In this sense, interpretability may be viewed as relational: it emerges at the interface between algorithmic structure and clinical reasoning (London, 2019; Floridi et al., 2018).

A further conceptual clarification concerns the distinction between technical explainability and clinically meaningful interpretability. Technical explainability refers to methods that describe aspects of a model's internal functioning—such as feature attribution scores, gradient-based saliency maps, or local surrogate approximations (Rudin, 2019; Tjoa and Guan, 2021; Ribeiro et al., 2016). These tools can enhance transparency at the computational level. Clinically meaningful interpretability, however, requires more than insight into algorithmic mechanics. It implies that the explanation can be integrated into pathophysiological understanding, support counterfactual reasoning, and inform actionable medical decisions (London, 2019; Floridi et al., 2018). An explanation that identifies a specific HRV feature as influential is technically informative; it becomes clinically meaningful only if the clinician can relate that feature to plausible autonomic mechanisms, patient context, and therapeutic implications (Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology, 1996; Shaffer and Ginsberg, 2017).

Clarifying these conceptual boundaries is not a purely semantic exercise. In high-stakes domains such as HRV-based prediction of deterioration, arrhythmia risk, or stress-related dysfunction, the level and type of interpretability required may vary according to clinical risk, regulatory context, and ethical burden (London, 2019; Floridi et al., 2018). A screening tool deployed for low-risk stratification may tolerate lower degrees of interpretability than a system influencing invasive interventions or resource allocation. Establishing shared definitions therefore provides the analytical foundation for evaluating when black-box approaches may be acceptable, when enhanced transparency is warranted, and how proportional interpretability can be implemented in clinical AI systems (London, 2019; Stix, 2021).

Across the reviewed HRV-AI literature, several broad application domains were recurrent, including rhythm-related classification tasks, prediction of deterioration or adverse clinical events, autonomic or stress-related state assessment, and wearable or remote-monitoring applications. These use cases differed not only in clinical objective, but also in signal source, recording duration, feature engineering strategy, model family, validation depth, and the degree of interpretability expected for deployment. This heterogeneity is important because the practical meaning of transparency in HRV-AI depends strongly on whether the model is intended for low-risk monitoring, intermediate-risk triage support, or high-stakes clinical decision support.

4 Clinical resistance and governance challenges in black-box HRV models

The discussion in this section is intended to move from signal- and model-level vulnerabilities, to clinician-facing interpretability and workflow challenges, and finally to broader issues of accountability, regulation, and deployment governance. Although these dimensions are closely interconnected, distinguishing them analytically helps clarify how technical fragility in HRV-AI systems may translate into clinical usability problems and, ultimately, into governance and oversight concerns.

While the conceptual distinction between interpretable and black-box models provides a useful analytical framework, the implications of algorithmic opacity become most visible in real-world clinical environments. In healthcare settings, opaque AI systems may influence clinical reasoning, accountability structures, and patient trust. Consequently, the adoption of black-box HRV prediction models raises not only technical questions but also cognitive, ethical, and governance challenges that must be addressed before widespread clinical deployment.

4.1 Accountability gaps in opaque AI systems

In the clinical setting, AI-based clinical decision support systems (AI-CDSS) are increasingly used to assist clinical decision making. While AI is capable of enhancing diagnostics, planning treatment, and monitoring patients, the implementation of such technologies requires addressing the challenges regarding data privacy, algorithmic bias, model interpretability, regulatory and human clinical oversight (O'Sullivan et al., 2022).

Data usage policies (GDPR) and AI systems development legislation (AI Act), provide guidelines regarding how the data should be used, as well as what can be included and how to develop an AI system. A mandatory requirement of the AI Act for systems that are used in the healthcare context is that human critical thinking must be maintained. This means that the systems must be designed so that they can be effectively overseen and the limitations and capabilities properly understood by clinicians. They should be able to remain aware of possible tendency to over rely on the produced output.

An additional recommendation of the AI Act states that the human operators must be able to correctly understand and interpret the system's output using for example interpretation tools and methods available, in order to decide whether to disregard or override the output. The Montreal Declaration for the Responsible Development of Artificial Intelligence from 2018 states that "only human beings can be held responsible for decisions stemming from recommendations made by AIS (Artificial Intelligence Systems), and the actions that proceed therefrom" (paragraph 9.1). As a result of these requirements, the necessary quality of the proposed medical AI solutions has increased, due to the need of implementing transparency, retraceability and comprehensibility. This also means that the focus has shifted from accuracy towards answering the question of why (O'Sullivan et al., 2022).

Accuracy remains an important factor in assessing the usability of a proposed solution, and as a result, some discussions follow the trade-off between explainability and accuracy of models. From an outcome standpoint, the employment of opaque AI models is justified. However, from a responsibility assignment point-of-view, the "black box" models raise the question of accountability in the decision-making process (Freyer et al., 2024).

Clinical Decision Support Systems (CDSS) that use AI have been debated by many authors throughout the years. Adams argues in favor that alongside the traditional for principles of bioethics, principle of beneficence, non-maleficence, respect for autonomy and justice, a fifth principle of explicability must be added (Adams, 2023). Floridi et al. note that explicability allows for the four core principles to be achieved in the context of AI use (Floridi et al., 2018; Adams, 2023).

Other discussions state that neither developers, nor healthcare providers can be held responsible for the consequences resulting from the use of "black-box" AI-DSS, deriving into *a responsibility gap*. According to Kempt, HCP are rationally obligated and in consequence are responsible for taking decisions considering the support provided by the AI-DSS. This also translates into discarding incorrect assessments. In an earlier work of the same author, disagreements between the HCP and the AI model should be solved by bringing a second HCP (Freyer et al., 2024).

In the context of contemporary European regulatory and governance frameworks, transparency, traceability, data governance, and human oversight are increasingly emphasized for higher-risk clinical AI applications. However, the precise operational obligations and conformity requirements depend on the intended purpose and regulatory pathway of the specific system (O'Sullivan et al., 2022). These attributes are also found in a study that highlights the patient concerns regarding oversight and the need for clear accountability structures involving regulatory bodies and care teams (Stroud et al., 2025). It is shown that patients prefer knowing what oversight procedures are employed when using AI tools for their care. The patient's perspective provided by this article highlights the need for accountability measures, as well as transparent solutions, especially when their data will likely be used for improvement of the systems (Stroud et al., 2025).

4.2 Automation bias and over-reliance on algorithmic outputs

Beyond accountability concerns, another major challenge associated with opaque AI systems is automation bias. Post-hoc methods could provide a false sense of security, which in turn increases the risk of automation bias (Freyer et al., 2024). This risk becomes a center point of discussion, especially when considering the cognitive load that the introduction of an AI system weighs down on. Over reliance is identified as a potential consequence of opaque models, due to the lack of interpretability while often delivering high accuracy. This poses a threat especially among less experienced clinicians who might lack the expertise for critical analysis of the AI output (Freyer et al., 2024; Tun et al., 2025).

In broadening the term of over-reliance, an article examines "inappropriate compliance," meaning accepting bad advice, and "inappropriate reliance," meaning rejecting good advice. Highly experienced specialists performed worse when the AI's logic was perceived as very clear, while the performance of the participants with the least neurology experience improved with the use of XAI. A solution proposed by this article is personalized, user-centered design of AI systems, due to the fact that a clinician's professional background impacts how they interpret and rely on automated advice (Gombolay et al., 2024).

While analyzing the cognitive risks of AI support, an article proposes frictional AI, namely pro-hoc explanations, to mitigate over reliance and automation bias. These explanations consist of providing similar cases with alternative outcomes, which encourages critical reflection. The frictional behavior comes from inserting a "cognitive forcing function" that encourages users to think critically, rather than automatically accepting the AI's first suggestion. Less experienced physicians found the pro-hoc support as beneficial, even if their diagnostic accuracy has not radically improved (Cabitza et al., 2024).

A study argues that the use of lower-level Class Activation Maps (CAMs) yields higher diagnostic accuracy among experienced physicians. In this case the features extracted from earlier layers of the convolutional neural network and traditional coloring schemes were more effective at increasing diagnostic accuracy, in the task of detecting thoracolumbar fractures from the vertebral X-rays, then higher level semantic designs. The findings of this article highlight the impact of explanation design on user reliance, and the importance of human centered solutions (Famiglini et al., 2024).

4.3 Ethical asymmetry between algorithm developers and clinicians

The ethical landscape of AI in healthcare reveals a structural asymmetry of responsibility between algorithm developers and

clinicians. While software developers face technical and quality management standards, namely ISO 14971 and IEC 62304, the regulatory framework ultimately states that there must be a "human-in-control" (O'Sullivan et al., 2022). The asymmetry is that developers design systems towards the best performance, but if the clinicians follow along suggestion, the developer is protected by the clause of "human-in-control," while the clinician might face the malpractice claim (O'Sullivan et al., 2022).

They must also take into account the unpredictability of developing a continuous learning system, however clinicians bear ultimate legal and moral weight of the final decision (O'Sullivan et al., 2022). The use of machine learning methods, specifically deep learning, where algorithms can train themselves or self-learn must take into account that the model will be learning from local clinical information. The problem will arise when the same model is used in another environment that might be dissimilar to the initial one. Another issue is that the bias, from either insufficient data or self-learning, might be leading towards unreliable future predictions and poor generalizability (Busnatu et al., 2022).

4.4 Relevance to high-risk and dual-use deployment contexts

Under the EU AI Act, AI used for medical diagnosis, treatment planning, or monitoring physiological processes, such as an AI-guided ECG analysis for heart failure, is classified as high risk, since it also falls under the classes IIa, IIb or III of the Medical Device Regulation (MDR) (European Parliament, 2024; European Parliament, 2017). This entails that the systems must undergo a "third party conformity assessment" by a regulatory body, that evaluates the robustness of the data quality, transparency, and human-oversight mechanisms.

Research that can benefit and harm humanity or dual-use research is a topic discussed in medical ethics. The surge in the use of artificial intelligence for tools and research has asked for a call to action, especially since AI is becoming less expensive to use and, in some cases, requires less training (European Parliament, 2017; Bobier et al., 2025). Additionally, while the need for traceability and reproducibility imposes for research to be shared on repositories for "open science," the risk of misuse must also be taken into account. This leads to a debate between scientific openness and public security, requiring a look into restricting the publication of datasets or algorithm features, such as weights (Bobier et al., 2025).

The AI Act also states that a robust risk management system is crucial, when developing AI for healthcare. It calls for a response protocol in the case of errors and biases which should be continuously monitored (European Parliament, 2024).

4.5 Signal-level and data-quality risks specific to HRV analytics

The monitoring and analysis of Heart Rate Variability (HRV) can be typically done using precise electrocardiograms (ECGs) in a clinical setting, or photoplethysmography (PPG) sensors from wearables. However, PPG-based data collection is highly susceptible to noise, variations in temperature, and poor peripheral perfusion, all of which can compromise the integrity of the data (Tasmurzayev et al., 2025). Other measurement inaccuracies include the reliance on specific HRV metrics, such as the LF/HF ratio for sympathovagal balance measurement, which is highly debated. If context is not accounted for these metrics can be influenced by respiratory rate, posture and non-stationary dynamics, which can mislead users or clinicians (Burlacu et al., 2026). The context should also include gender and age, as younger men present with a lower complexity of heartbeat generation than women of the same age (Voss et al., 2015). These details are highly relevant in the case of training predictive AI models, where non-diverse datasets can produce misdiagnoses, and discriminatory outcomes (Radanliev, 2025).

In the context of wearables protecting privacy and confidentiality is crucial as these devices capture sensitive information regarding the individual's health (Radanliev, 2025). While General Data Protection Regulation (GDPR) imposes severe penalties to organizations that handle personal data inadequately, it is important to note that anonymizing data does not always guarantee privacy (Radanliev, 2025). The collection of raw physiological data is hard to anonymize completely, as it can sometimes be reverse engineered to identify individuals, which represents another reason for urgent updates to existing privacy laws, cybersecurity and data stewardship frameworks (Radanliev, 2025).

One of the risks of using wearables sensors' data and AI systems is the black-box nature of the deep learning models used for interpreting HRV data, whose decisions are difficult to interpret and understand. Another risk is the over-reliance on these algorithms which can influence people's decisions based on the guidance from AI, thus potentially reducing human agency and decision-making (Radanliev, 2025).

These signal-level and preprocessing-related issues are not only technical concerns, but also shape how clinicians interpret, trust, and operationalize HRV-AI outputs in practice.

4.6 Barriers to real-world clinical adoption

The real-world clinical adoption of AI systems remains limited by the opacity of complex ML models and the uncertainty they create for clinicians and healthcare institutions (Tun et al., 2025; Busnatu et al., 2022). The lack of insight makes it difficult for clinicians to interpret the pathophysiological rationale behind a prediction (Hu et al., 2019). Furthermore, the integration of AI tools that misalign with the existing workflow often causes cognitive overload and alert fatigue, thus rendering the decision support ineffective (Wong et al., 2025).

The lack of interpretability directly impacts the medico-legal accountability and responsibility concerns (O'Sullivan et al., 2022). Clinicians have expressed their concern regarding medical liability, especially concerning who is legally at fault if the AI system makes an incorrect recommendation that results in patient harm. Regulatory frameworks, such as the EU AI Act, that classify AI in healthcare as "high-risk" and impose strict human oversight, insist that human clinicians remain in control and bear the ultimate responsibility for the medical decisions. However, the legal and ethical "responsibility gap" that comes from whether or not to hold clinicians accountable for black-box outputs they cannot understand the reasoning behind, is a deterrent for the adoption of such systems (O'Sullivan et al., 2022; Freyer et al., 2024).

These transparency and accountability barriers do not only impact the clinicians' trust, but also patient communication and broader institutional acceptance (Tun et al., 2025). As physicians struggle to adequately explain AI-driven diagnoses or treatment plans to their patients, the shared decision-making and informed consent is hindered, thus compromising the patient-clinician relationship (Freyer et

al., 2024; Stroud et al., 2025). Additionally, institutional acceptance is also inhibited by the disruption of established clinical roles, the dehumanization of patient care and fear of job displacement (Tun et al., 2025).

Ultimately, these interconnected barriers have profound implications for healthcare investment and resource allocation (Agard et al., 2026). Rather than focusing on algorithmic performance, healthcare investments must prioritize interdisciplinary collaboration, human-centered explainable AI (XAI) design, and robust infrastructure to ensure these tools are safely and effectively integrated into real-world workflows (Agard et al., 2026; Raszke et al., 2025).

Taken together, these clinical, cognitive, ethical, and regulatory challenges illustrate why the black-box nature of many high-performance ML models remains a major barrier to their responsible adoption in healthcare. While predictive accuracy remains essential, the limitations described above highlight the need for approaches capable of reconciling model performance with transparency, interpretability, and human oversight. In recent years, a growing body of research has therefore focused on methods designed to bridge this gap, including explainable artificial intelligence (XAI), hybrid modeling strategies, and physiology-informed ML approaches. This tension has stimulated growing interest in methodological approaches capable of improving model transparency while preserving predictive performance.

When these interpretability and reliability limitations are transferred into real clinical environments, they become workflow, communication, and human-factor challenges rather than purely computational ones.

5 Approaches to reconcile predictive performance and interpretability

A critical appraisal of black-box HRV models should not be interpreted as a categorical rejection of model complexity. In physiological time-series analysis, higher-capacity models may capture nonlinear, multiscale, or temporally distributed patterns that are difficult to represent with simpler architectures. In some settings, this may generate genuine clinical value. The central question is therefore not whether complexity should be eliminated, but under what conditions the performance gains of more complex models are sufficiently robust, clinically relevant, and governable to justify reduced transparency.

In response to the limitations associated with opaque ML models, a growing body of research has explored methods aimed at improving transparency while preserving predictive performance. These approaches aim to reduce the tension between predictive performance and interpretability that characterizes many contemporary ML applications in healthcare. Rather than replacing complex models with fully transparent ones, current research increasingly explores intermediate strategies capable of improving transparency while preserving predictive power.

5.1 Post-hoc explainability methods and their limitations

Post-hoc explainability refers to the methods applied to models that are not interpretable by design, with the purpose of enhancing the understanding of the mechanisms behind the output provided (Barredo Arrieta et al., 2020). Some examples of the methods used can be categorized based on the presentation of the explanation, inspired by the ways humans explain, namely text, visual, local, by example, by simplification and feature relevance techniques. Text explanations enhance model explainability by generating symbols that map the model's rationale. Visual techniques visualize model behavior using dimensionality reduction, facilitating understanding for non-experts. Local explanations focus on simpler solution subspaces for clearer insights. Explanations by example extract relevant data to elucidate inner model relationships. Simplification techniques build new models that retain performance while reducing complexity. Lastly, feature relevance methods assess and compare variable sensitivity to outputs, indirectly revealing model functionality (Barredo Arrieta et al., 2020).

The categorization of explainability methods can focus on the usability besides for the model proposed (Linardatos et al., 2021). The distinction between model-agnostic methods and model-specific methods is whether or not they can be used on a single or group of models, or they can be applied to any model. The generation of local explanations can be done using LIME, SHAP, and Grad-CAM, due to their ability highlight the features that contribute most to the model's output. LIME, or Local Interpretable Model-agnostic Explanations, use local surrogate models that explain model predictions by generating fake instances around a target instance, applying the complex model to these instances, and training an interpretable model to capture the original model's behavior in that area. SHAP (Shapley Additive Explanations) offers a unified framework for interpretation based on Shapley values, which measure the contribution of each feature to model predictions. Comparisons between LIME and SHAP demonstrate their effectiveness in explaining a deep learning model's decision in medical datasets (Tursunalieva et al., 2024).

Both LIME and SHAP, despite their advantages, have significant limitations affecting their reliability (Hooshyar, 2024). These methods' approximating models are indirectly related to the original model, often failing to provide accurate explanations for the model's output. LIME's assumptions of local linearity and feature independence limit its use with complex data, leading to unstable explanations. SHAP, while offering detailed contribution analysis, faces challenges in computational efficiency and precision due to similar restrictive assumptions (Hooshyar, 2024). Research from Harvard, MIT, Drexel, and Carnegie Mellon University found that these explanation methods frequently produce conflicting outputs for models like neural networks, causing machine learning practitioners to rely on potentially misleading explanations in decision-making scenarios (Krishna et al., 2022).

For more complex algorithms, such as Artificial Neural Networks (ANNs), explanations should focus more on the depth of the hidden layers than the surface-level ones that LIME and SHAP provide. By combining predictive power with the transparency of symbolic reasoning, neural-symbolic AI has the potential to make ANNs easier to understand (Hooshyar, 2024).

The critics of methods such as LIME and SHAP argue that they are vulnerable to adversarial attacks, meaning that manipulated data can mislead model outcomes without human detection (Hooshyar, 2024). This undermines the reliability of explanations in XAI, as minor changes, especially in image classification, can lead to inaccurate explanations or failure to accurately identify relevant features (Garouani et al., 2024).

Emerging evidence indicates that explanations can negatively impact clinical decision making by creating undue confidence in AI

outputs. The phenomena of post-hoc explanations deluding the user into understanding without providing actual insight into the behavior of the model is coined as "explainability trap" (Singh et al., 2025). A study by Ghassemi et al. revealed that physicians exposed to saliency maps with AI recommendations were more likely to accept incorrect suggestions than those who only received predictions, thus potentially undermining clinical judgment (Ghassemi et al., 2021).

A study argues that the limitations of explanations of an AI can be classified as follows: unfaithful, where the explanation conflicts with model data; irrelevant, when it does not meet user need; unstable, producing different outputs to the same data and model; incoherent; revealing, exposing sensitive information; unnecessary, providing redundant details; cherry-picked, by selectively presenting information while ignoring contradictions (Martens et al., 2025).

Conversely, a fully interpretable model is not automatically preferable if its predictive performance is inferior in a high-consequence clinical task. A proportionate view is therefore needed: the degree and form of interpretability required should depend on deployment risk, intended clinical use, availability of human oversight, and the consequences of model error. In low-risk or supportive applications, limited opacity may be acceptable if performance is robust and outputs are embedded within appropriate oversight structures. In higher-risk settings, however, stronger transparency, clearer uncertainty communication, and more explicit accountability mechanisms become increasingly important.

5.2 Cognitive risks of explainability: the explainability trap

While the explanations improve decision accuracy when the output is correct, a paradox emerges from the accuracy decrease when the AI is incorrect. A study involving 257 medical students and 3,855 diagnostic decisions validates this transparency paradox, arguing that the convincing reasons provided by the AI systems made it hard to differentiate between correct and incorrect advice. The persuasiveness leads to potential over-reliance, discernment failure and false confidence (Khanna et al., 2025).

Post-hoc explanations may seem to provide transparency, however they pose a risk of misleading users. In an effort to satisfy human demands, developers might create algorithms that present convincing yet false narratives about a model's true reasoning. Relying on superficial explanations threaten to reproduce a pathological human behavior of using socially acceptable justifications to mask biases. These explanations offer an illusion of understanding while hiding the model's actual, potentially flawed decision-making process (Singh et al., 2025; Lipton, 2018).

5.3 Pro-hoc explainability and frictional AI

A form of explanations that consists in offering alternative explanations, by using examples of similar cases for each category of the possible outcomes is termed pro-hoc explanations. Instead of giving direct recommendation or post-hoc explanations, pro-hoc is part of frictional AI, aimed at balancing decision effectiveness with cognitive risks, such as over-reliance or automation-bias mitigation. In an article employing this type of explanations, a group of physicians was presented with an X-ray image, got their opinions noted, were then provided with three similar cases, and finally had their final decisions collected. This analysis resulted in the orthopedists detecting vertebral fractures with an improved accuracy, and a reduction in specific errors. Another important aspect revealed in this study was that the physician's confidence in their final decisions has increased (Cabitza et al., 2024).

By encouraging the user to verify the AI's logic against their own clinical knowledge, the system prevents "inappropriate compliance." Their aim is to be "cognitively non-invasive" and "non-substitutive," thus ensuring that the physicians remain the primary decision-makers. Several typologies have been identified for integrating friction to manage user cognitive load: Judicial Protocols contrast various arguments to avoid reliance on single AI classifications; Cautious Protocols utilize uncertainty quantification to address ambiguous cases; and Decentralized Protocols delay information until users complete specific tasks, minimizing automation bias. This friction serves as a tool for responsible knowledge work, requiring careful calibration to avoid frustration or passive reliance (Cabitza et al., 2024).

The principles of Frictional AI may lead towards design ethos that may seem contrary to the industry's trend towards efficiency and specificity in AI systems. Frictional AI emphasizes the value of human expertise while embracing ambiguity, and the alignment to the clinicians' work practices can enhance the usability of this technology in real-world scenarios, such as in radiology (Anichini et al., 2024).

5.4 Hybrid modeling strategies combining interpretable and black-box components

Hybrid strategies are revolutionizing the understanding of complex dynamics through "gray-box" identification. Systems biology models often only know some of the details of how things work, so they need to find new rules by looking at small amounts of noisy clinical data (Ahmadi Daryakenari et al., 2024). Symbolic regression methods are mixed with Physics-Informed Neural Networks (PINNs) and the Extreme Theory of Functional Connections (X-TFC) in the AI-Aristotle framework to solve this problem. Initially, deep neural networks approximate the missing, unobserved physical dynamics, functioning as the black-box component. Subsequently, symbolic regression algorithms turn these learned neural representations into compact analytical expressions. This framework was tested on biomedical challenges, such as single-dose pharmacokinetics and ultradian glucose-insulin endocrine models, successfully recovering hidden biological interactions, therefore translating opaque neural weights into transparent mathematical equations (Ahmadi Daryakenari et al., 2024).

Another domain-specific framework is OIKAN (Optimized Interpretable Kolmogorov-Arnold Network), which enhances predictive tasks by merging the neural network's expressiveness with symbolic transparency (Yedilkhan et al., 2025). Unlike traditional multilayer perceptrons, which use dense weight matrices, OIKAN utilizes a neural network which learns intermediate representations which are then processed using a two-stage sparse symbolic regression pipeline with ElasticNet regularization. This process identifies key variables and applies non-linear transformations to create interpretable mathematical formulas. By integrating deep learning's pattern recognition with symbolic logic, OIKAN produces clinically relevant models that are both accurate and mathematically verifiable (Yedilkhan et al., 2025).

In clinical text analysis, integrating LLMs with rule-based expert systems represents a major step towards decision-making transparency (Prenosil et al., 2025). While models like GPT-4 adequately

handle unstructured data, they often provide unclear reasoning. To solve this issue, researchers developed a semantic-neuro-symbolic framework that integrates GPT-4 with a prolog-based expert system Plato-3, to extract structured clinical data from free-text found in PET/CT reports for prostate cancer patients. The proposed system is able to transform unstructured diagnostic reports into structured, traceable outputs and reduce risks associated with data privacy and inaccuracies (Prenosil et al., 2025).

5.5 Physiology-informed and constraint-based AI approaches

Integrating biological laws into machine learning model architecture represents a shift from purely data-driven methods. Researchers are using "partial-knowledge" models that combine the mechanistic models' predictability with deep learning's data processing abilities. An important technique for partial knowledge is "in-processing," where medical principles are embedded into the algorithm. This approach uses regularization terms to penalize deviations from physiological norms, promoting biologically consistent and interpretable outputs. The result is models that are both accurate and grounded in real-world scenarios (de Rooij et al., 2025).

An approach of constraint-based AI is the development of Universal Differential Equation (UDE) systems that are enhanced by physiology-informed regularization (de Rooij et al., 2025). UDEs combine mechanistic differential equations with neural networks to enable learning of unknown biological interactions from sparse biomedical data. To mitigate potential overfitting, physiology-informed regularization imposes penalties on biologically implausible predictions, thus ensuring accurate outcomes. For example, constraints like non-negativity and area-under-the-curve, in glucose appearance from a meal, direct the neural network towards stable and physiologically valid results. This approach not only stabilizes training with limited data, but also minimizes unrealistic behavior, producing models that are more transparent (de Rooij et al., 2025).

6 A pragmatic implementation framework for HRV-AI systems

The translation of AI-based HRV prediction models from experimental research into clinical practice requires more than algorithmic performance alone. Although numerous studies have demonstrated the predictive potential of ML models applied to physiological signals, real-world clinical implementation remains limited due to technical, organizational, and governance barriers (Wiens et al., 2019; Kelly et al., 2019; Golden et al., 2024). The integration of AI systems into healthcare environments must therefore address multiple dimensions simultaneously, including data quality, model reliability, workflow compatibility, and long-term monitoring (Wiens et al., 2019; Kelly et al., 2019).

In clinical environments, predictive models frequently operate as components of clinical decision support systems (CDSS), where they interact directly with healthcare professionals and influence diagnostic or therapeutic reasoning. However, the successful deployment of AI-enabled CDSS depends on careful alignment with existing clinical workflows and decision-making processes. Systems that are poorly integrated into clinical routines may increase cognitive burden, disrupt clinical judgement or contribute to alert fatigue, ultimately altering clinician trust and adoption (Poly et al., 2020; Abdelrahman Mohamed et al., 2025). Previous studies evaluating AI-assisted CDSS have emphasized that usability, transparency, and workflow compatibility are essential determinants of successful implementation in healthcare settings (Golden et al., 2024; Wang et al., 2021).

Clinical environments are dynamic, characterized by evolving patient populations, changing diagnostic protocols, and variable data acquisition conditions. These factors can alter the relationship between model inputs and clinical outcomes over time, leading to performance degradation if models are not continuously evaluated and recalibrated. Consequently, responsible deployment of HRV-based AI systems requires a structured implementation strategy that addresses the entire lifecycle of data, models, and clinical interaction (Botha et al., 2024).

To address these challenges, we propose a pragmatic four-step implementation framework designed to guide the responsible translation of HRV-based AI systems from experimental models to real-world clinical environments (Figure 2). The framework emphasizes data governance, robust model development, integration into clinical workflows, and continuous monitoring of algorithmic performance.

6.1 Step 1—data governance and signal integrity

The first and most fundamental stage of HRV-AI implementation concerns the quality, reliability, and governance of physiological data. HRV analysis relies on accurate detection of inter-beat intervals, making predictive models particularly sensitive to signal noise, acquisition artifacts, and preprocessing variability.

Electrocardiography (ECG) remains the gold standard for HRV measurement in clinical environments, providing high temporal resolution and precise detection of R-R intervals. In contrast, photoplethysmography (PPG) sensors embedded in wearable devices offer scalable monitoring solutions but are more susceptible to motion artifacts, variations in peripheral perfusion, and environmental noise. These limitations may introduce significant variability in HRV features and can influence model performance when algorithms trained on ECG-derived data are applied to wearable-derived signals (Tasmurzayev et al., 2025).

Standardization of signal preprocessing is therefore essential. Recommended procedures typically include artifact detection and correction, filtering of ectopic beats, interpolation of missing intervals, and harmonized feature extraction across time-domain, frequency-domain, and nonlinear HRV metrics. Lack of standardization across datasets has been identified as a major barrier to reproducibility and generalizability in physiological signal analytics (Shaffer and Ginsberg, 2017).

Beyond signal quality, responsible data governance must also address issues related to privacy, consent, and regulatory compliance. Physiological signals collected through wearable devices may reveal sensitive information regarding both physical and psychological states. As digital health infrastructures expand, robust governance frameworks aligned with data protection regulations become essential for the ethical management of patient-generated health data (Radanliev, 2025).

6.2 Step 2—model development and validation

The second step involves the development and validation of predictive models capable of extracting clinically meaningful information

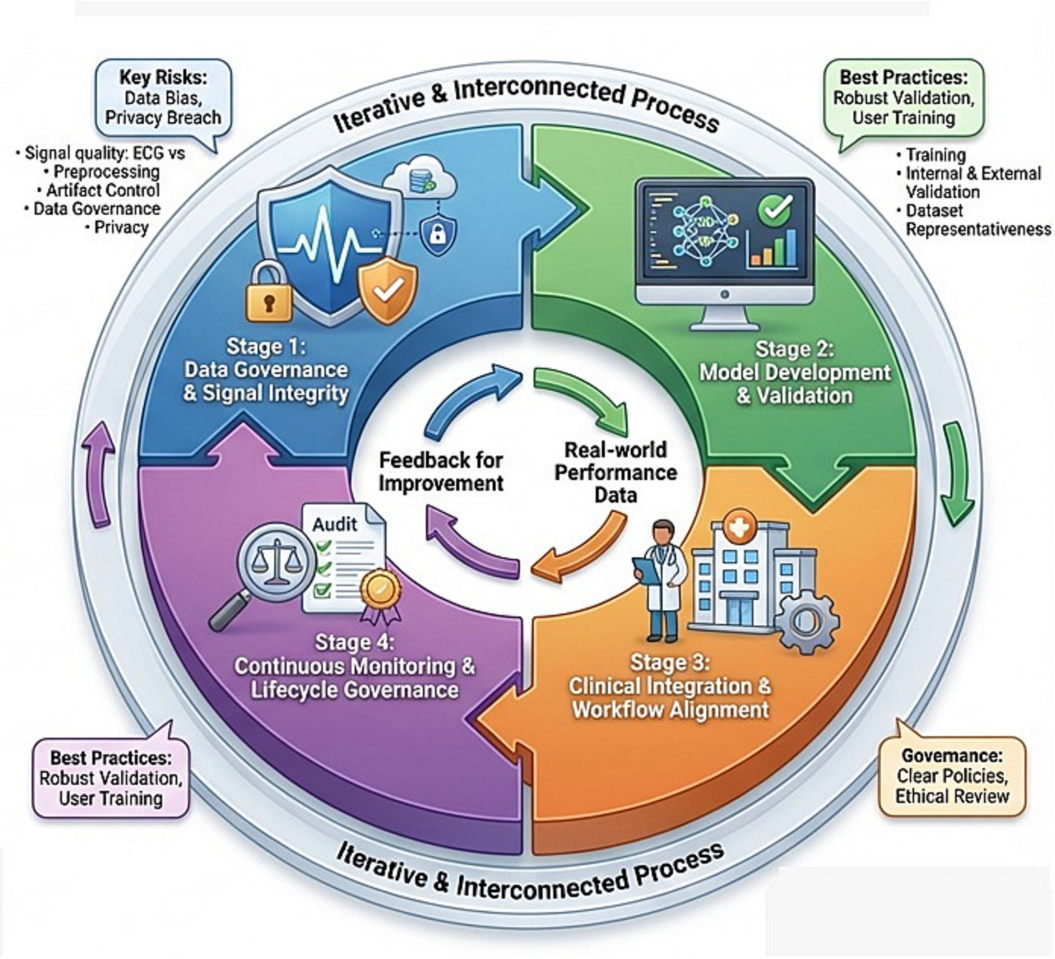

FIGURE 2
Lifecycle framework for responsible implementation of HRV-AI systems.

from HRV signals. Machine learning approaches commonly applied to HRV prediction include support vector machines, random forests, gradient boosting methods, and deep neural networks capable of modeling complex nonlinear physiological relationships (Faust et al., 2018).

While advanced machine-learning architectures may achieve superior predictive performance, their adoption in clinical practice requires rigorous validation strategies. Internal validation methods such as cross-validation provide initial estimates of model performance but may overestimate generalizability when applied to homogeneous datasets. External validation using independent datasets from different institutions or populations is therefore critical to assess model robustness and reduce the risk of dataset bias (Botha et al., 2024).

Another key consideration is dataset representativeness. HRV patterns may vary according to age, sex, comorbidities, medication use, and physiological context. Models trained on narrowly defined populations may perform poorly when applied to broader clinical settings. Ensuring demographic and clinical diversity in training datasets is therefore essential for developing generalizable HRV-based predictive models (Shaffer and Ginsberg, 2017).

In HRV applications, this challenge is further amplified by the physiological sensitivity of many features to respiration, posture, circadian timing, autonomic state, medication effects, and preprocessing decisions. Accordingly, model generalizability in HRV-AI depends not only on sample size and external validation, but also on the reproducibility of the signal-processing pipeline and the physiological comparability of the populations in which the model is trained and deployed.

6.3 Step 3—clinical integration and workflow alignment

Even highly accurate predictive models may fail to generate clinical value if they are not effectively integrated into existing clinical workflows. The third step of the framework therefore focuses on the operational integration of HRV-AI systems into clinical decision support environments.

Clinical decision support systems augmented with AI have demonstrated potential to enhance diagnostic accuracy and assist clinicians in complex decision-making scenarios. However, poorly designed CDSS interfaces may disrupt clinical workflows and contribute to cognitive overload. Alert fatigue represents a well-recognized challenge in digital clinical systems, where excessive or poorly contextualized alerts may lead clinicians to ignore or override recommendations, thereby reducing the effectiveness of decision support tools (Poly et al., 2020).

Studies examining AI-driven clinical systems have emphasized the importance of designing interfaces that complement clinicians'

cognitive processes rather than replacing them. Human-centered design approaches encourage the development of AI systems that function as collaborative tools supporting clinician judgement rather than autonomous decision-makers. Within this paradigm, AI systems provide risk estimates or pattern recognition insights while clinicians retain full authority over diagnostic and therapeutic decisions (Golden et al., 2024).

Transparency in model outputs is equally important for clinical acceptance. Interfaces that provide contextual explanations, uncertainty estimates, or visual summaries of physiological trends may improve clinician understanding and trust in AI-supported predictions.

6.4 Step 4—continuous monitoring and lifecycle governance

The final step addresses the long-term governance of AI systems after deployment. Unlike traditional clinical tools, machine-learning models operate within evolving environments where changes in patient populations, clinical practices, or data acquisition technologies may alter the relationship between inputs and outcomes.

This phenomenon could gradually degrade predictive performance if algorithms are not periodically evaluated and recalibrated. Continuous monitoring frameworks therefore play an important role in maintaining the reliability of deployed AI systems (Botha et al., 2024).

Model governance strategies typically include periodic performance auditing, recalibration of prediction thresholds, and evaluation of model behavior across demographic or clinical subgroups. These procedures help identify unintended biases or performance deterioration that may emerge over time. Importantly, governance structures should also define clear accountability pathways among clinicians, healthcare institutions, and algorithm developers. Transparent oversight mechanisms contribute to maintaining clinical trust while ensuring that AI systems remain aligned with patient safety standards (Botha et al., 2024).

Taken together, these four stages emphasize that the safe deployment of HRV-based AI systems requires attention not only to model performance, but also to data integrity, workflow integration, and long-term governance. Rather than viewing AI models as isolated predictive tools, this framework highlights their role as components of broader socio-technical systems within healthcare environments. Importantly, these stages should not be interpreted as strictly sequential but as interacting components of an iterative governance process throughout the lifecycle of clinical AI systems.

7 Limitations

This review has several limitations. First, it was designed as a narrative, concept-driven review rather than a systematic review. Accordingly, although a structured literature search was performed, study selection was not intended to be fully exhaustive and no formal risk-of-bias assessment was conducted. Second, the included literature was heterogeneous with respect to source type, signal source, preprocessing methods, HRV feature definitions, prediction tasks, validation strategies, and study design, which precluded quantitative synthesis. Third, part of the argument developed in this review draws on broader healthcare AI, explainability, ethics, and governance literature rather than exclusively HRV-specific empirical studies, reflecting the still limited number of deployment-oriented investigations in this exact niche. Finally, the four-step implementation framework proposed here is conceptual and pragmatic in nature and should be regarded as a translational guide rather than an empirically validated implementation model. Future work should test, refine, and validate this framework in real-world HRV-AI deployment settings.

8 Conclusion

AI applied to heart rate variability (HRV) analysis offers promising opportunities for improving risk stratification and physiological monitoring across multiple clinical domains. However, the transition from experimental models to real-world clinical use remains constrained by important challenges related to interpretability, accountability, and integration into existing healthcare workflows.

The relationship between algorithmic accuracy and interpretability should therefore not be approached as a strict dichotomy. Highly accurate models often rely on complex architectures that are difficult to interpret, while overly simplified models may fail to capture the physiological complexity reflected in HRV signals. The central challenge is not the elimination of model opacity, but the development of governance and implementation strategies that allow these systems to be used responsibly in clinical environments.

Explainability techniques can contribute to improving transparency, yet they also present methodological limitations and may sometimes generate misleading impressions of understanding. Consequently, explanations should be interpreted cautiously and considered supportive tools rather than definitive representations of algorithmic reasoning.

In this context, the pragmatic implementation framework proposed in this article emphasizes four essential dimensions for the responsible adoption of HRV-based AI systems: ensuring data quality and signal integrity, developing and validating robust predictive models, integrating these tools into clinical workflows with appropriate human oversight, and maintaining continuous monitoring of algorithm performance after deployment.

Ultimately, HRV-AI systems should be understood not merely as predictive algorithms but as components of complex socio-technical healthcare systems. Their clinical value will depend not only on predictive accuracy, but also on transparent governance structures, careful implementation strategies, and the ability of clinicians to critically interpret algorithmic outputs within the broader context of patient care.

Author contributions

AB: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization,

Writing – original draft, Writing – review & editing. MO: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. OG: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. AI: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. R-EB-G: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. CB: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that Generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Abdelrahman Mohamed, H., Sayed, A., and Samah, B. (2025). Recent advances in artificial intelligence for healthcare: a comprehensive analysis of image processing and next-generation applications. *Inform. Technol. Bull. Comput. Inform.* 7, 1–17. doi: 10.21608/fcihib.2025.353639.1136

Acharya, U., Fujita, H., Oh, S. L., Hagiwara, Y., Tan, J. H., and Abdul Rahim, M. A. (2017). Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network. *Inf. Sci.* 405:12. doi: 10.1016/j.ins.2017.04.012

Adams, J. (2023). Defending explicability as a principle for the ethics of artificial intelligence in medicine. *Med. Health Care Philos.* 26, 615–623. doi: 10.1007/s11019-023-10175-7

Agard, G., Hraiech, S., and Gauss, T. (2026). From promise to practice: a roadmap for artificial intelligence in critical care. *J. Crit. Care* 91:155263. doi: 10.1016/j.jcrc.2025.155263

Ahmad, S., Ramsay, T., Huebsch, L., Flanagan, S., McDiarmid, S., Batkin, I., et al. (2009). Continuous multi-parameter heart rate variability analysis heralds onset of sepsis in adults. *PLoS One* 4:e6642. doi: 10.1371/journal.pone.0006642

Ahmadi Daryakenari, N., De Florio, M., Shukla, K., and Karniadakis, G. E. (2024). AI-Aristotle: a physics-informed framework for systems biology gray-box identification. *PLoS Comput. Biol.* 20:e1011916. doi: 10.1371/journal.pcbi.1011916

Anichini, G., Natali, C., and Cabitza, F. (2024). Invisible to machines: designing AI that supports vision work in radiology. *Comput. Support. Coop. Work* 33, 993–1036. doi: 10.1007/s10606-024-09491-0

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* 58, 82–115. doi: 10.1016/j.inffus.2019.12.012

Bobier, C., Hurst, D. J., and Obeid, J. (2025). Artificial intelligence, pharmaceutical development and dual-use research of concern: a call to action. *J. Med. Ethics*:jme-2025-110750. doi: 10.1136/jme-2025-110750

Botha, N. N., Segbedzi, C. E., Dumahasi, V. K., Maneen, S., Kodom, R. V., Tsedze, I. S., et al. (2024). Artificial intelligence in healthcare: a scoping review of perceived threats to patient rights and safety. *Arch. Public Health* 82:414. doi: 10.1186/s13690-024-01414-1

Buccelletti, E., Gilardi, E., Scaini, E., Galiuto, L., Persiani, R., Biondi, A., et al. (2009). Heart rate variability and myocardial infarction: systematic literature review and meta-analysis. *Eur. Rev. Med. Pharmacol. Sci.* 13, 299–307.

Burlacu, A., Brinza, C., Geman, O., Karppa, M., and Hemanth, D. J. (2026). Heart rate variability as a dual-use digital biomarker: integrating clinical, AI, and operational perspectives on human performance and resilience. *BMC Cardiovasc. Disord.* 26:87. doi: 10.1186/s12872-026-05543-z

Busnatu, Ş., Niculescu, A. G., Bolocan, A., Petrescu, G. E. D., Păduraru, D. N., Năstasă, I., et al. (2022). Clinical applications of artificial intelligence-an updated overview. *J. Clin. Med.* 11:265. doi: 10.3390/jcm11082265

Cabitza, F., Natali, C., Famiglini, L., Campagner, A., Caccavella, V., and Gallazzi, E. (2024). Never tell me the odds: investigating pro-hoc explanations in medical decision making. *Artif. Intell. Med.* 150:102819. doi: 10.1016/j.artmed.2024.102819

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature* 538, 20–23. doi: 10.1038/538020a

Chalmers, J. A., Quintana, D. S., Abbott, M. J., and Kemp, A. H. (2014). Anxiety disorders are associated with reduced heart rate variability: a meta-analysis. *Front. Psych.* 5:80. doi: 10.3389/fpsyt.2014.00080

Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology. (1996). Heart rate variability: standards of measurement, physiological interpretation and clinical use. Task force of the European Society of Cardiology and the north American Society of Pacing and Electrophysiology. *Circulation* 93, 1043–1065.

Clifford, G. D., Liu, C., Moody, B., Lehman, L. H., Silva, I., Li, Q., et al. (2017). AF classification from a short single Lead ECG recording: the PhysioNet/computing in cardiology challenge 2017. *Comput. Cardiol.* 44:22489. doi: 10.22489/CinC.2017.065-469

de Rooij, M., Erdős, B., van Riel, N. A. W., and O'Donovan, S. D. (2025). Physiology-informed regularisation enables training of universal differential equation systems for biological applications. *PLoS Comput. Biol.* 21:e1012198. doi: 10.1371/journal.pcbi.1012198

European Parliament (2017) Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (Text with EEA relevance. Available online at: https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng (Accessed March 5, 2026).

European Parliament (2024) Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). Available online at: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng (Accessed March 5, 2026).

Famiglini, L., Campagner, A., Barandas, M., La Maida, G. A., Gallazzi, E., and Cabitza, F. (2024). Evidence-based XAI: an empirical approach to design more effective and explainable decision support systems. *Comput. Biol. Med.* 170:108042. doi: 10.1016/j.compbiomed.2024.108042

Faust, O., Hagiwara, Y., Tan, J. H., Oh, S. L., and Acharya, U. (2018). Deep learning for healthcare applications based on physiological signals: a review. *Comput. Methods Prog. Biomed.* 161, 1–13. doi: 10.1016/j.cmpb.2018.04.005

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., et al. (2018). AI4People-an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Mind. Mach.* 28, 689–707. doi: 10.1007/s11023-018-9482-5

Freyer, N., Groß, D., and Lipprandt, M. (2024). The ethical requirement of explainability for AI-DSS in healthcare: a systematic review of reasons. *BMC Med. Ethics* 25:104. doi: 10.1186/s12910-024-01103-2

Garouani, M., Mothe, J., Barhrhouj, A., and Aligon, J. Investigating the Duality of Interpretability and Explainability in Machine Learning. *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)* (2024) 861–867 p.

Ghassemi, M., Oakden-Rayner, L., and Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digital Health* 3, e745–e750. doi: 10.1016/S2589-7500(21)00208-9

Golden, G., Popescu, C., Israel, S., Perlman, K., Armstrong, C., Fratila, R., et al. (2024). Applying artificial intelligence to clinical decision support in mental health: what have we learned? *Health Policy Technol.* 13:100844. doi: 10.1016/j.hlpt.2024.100844

Gombolay, G. Y., Silva, A., Schrum, M., Gopalan, N., Hallman-Cooper, J., Dutt, M., et al. (2024). Effects of explainable artificial intelligence in neurology decision support. *Ann. Clinic. Transl. Neurol.* 11, 1224–1235. doi: 10.1002/acn3.52036

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., et al. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.* 25, 65–69. doi: 10.1038/s41591-018-0268-3

Hooshyar, D. (2024). Problems with SHAP and LIME in interpretable AI for education: a comparative study of post-hoc explanations and neural-symbolic rule extraction. *IEEE Access* 12, 137472–137490. doi: 10.1109/ACCESS.2024.3463948

Hu, S. Y., Santus, E., Forsyth, A. W., Malhotra, D., Haimson, J., Chatterjee, N. A., et al. (2019). Can machine learning improve patient selection for cardiac resynchronization therapy? *PLoS One* 14:e0222397. doi: 10.1371/journal.pone.0222397

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., and King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* 17:195. doi: 10.1186/s12916-019-1426-2

Khanna, M., Wang, Z., Wei, L., and Xue, L. When medical AI explanations help and when they harm. arXiv [Preprint] (2025). doi: 10.48550/arXiv.2512.08424

Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., et al. The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective (2022).

Kwon, J. M., Lee, Y., Lee, Y., Lee, S., and Park, J. (2018). An algorithm based on deep learning for predicting in-hospital cardiac arrest. *J. Am. Heart Assoc.* 7:678. doi: 10.1161/JAHA.118.008678

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2021). Explainable AI: a review of machine learning interpretability methods. *Entropy* 23:18. doi: 10.3390/e23010018

Lipton, Z. (2018). The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 31–57. doi: 10.1145/3236386.3241340

London, A. J. (2019). Artificial intelligence and black-box medical decisions: accuracy versus Explainability. *Hast. Cent. Rep.* 49, 15–21. doi: 10.1002/hast.973

Martens, D., Shmueli, G., Evgeniou, T., Bauer, K., Janiesch, C., Feuerriegel, S., et al. Beware of "explanations" of AI. arXiv [Preprint] (2025) doi: 10.48550/arXiv.2504.06791

McKelvey, T., Ahmad, M., Teredesai, A., and Eckert, C. Interpretable Machine Learning in Healthcare. (2018) New York, NY: IEEE.

O'Sullivan, S., Janssen, M., Holzinger, A., Nevejans, N., Eminaga, O., Meyer, C. P., et al. (2022). Explainable artificial intelligence (XAI): closing the gap between image analysis and navigation in complex invasive diagnostic procedures. *World. J. Urol.* 40, 1125–1134. doi: 10.1007/s00345-022-03930-7

Poly, T. N., Islam, M. M., Muhtar, M. S., Yang, H. C., Nguyen, P. A. A., and Li, Y. J. (2020). Machine learning approach to reduce alert fatigue using a disease medication-related clinical decision support system: model development and validation. *JMIR Med. Inform.* 8:e19489. doi: 10.2196/19489

Prenosil, G. A., Weitzel, T. K., Bello, S. C., Mingels, C., Manzini, G., Meier, L. P., et al. (2025). Neuro-symbolic AI for auditable cognitive information extraction from medical reports. *Commun. Med.* 5:491. doi: 10.1038/s43856-025-01194-x

Price, W. N., Gerke, S., and Cohen, I. G. (2019). Potential liability for physicians using artificial intelligence. *JAMA* 322, 1765–1766. doi: 10.1001/jama.2019.15064

Radanliev, P. (2025). Privacy, ethics, transparency, and accountability in AI systems for wearable devices. *Front. Digital Health* 7:1431246. doi: 10.3389/fdgth.2025.1431246

Raszke, P., Giebel, G. D., Abels, C., Wasem, J., Adamzik, M., Nowak, H., et al. (2025). User-oriented requirements for artificial intelligence-based clinical decision support Systems in Sepsis: protocol for a multimethod research project. *JMIR Res. Protoc.* 14:e62704. doi: 10.2196/62704

Ribeiro, M. T., Singh, S., and Guestrin, C. "why should I trust you?": explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; San Francisco: Association for Computing Machinery; (2016). p. 1135–44.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intellig.* 1, 206–215. doi: 10.1038/s42256-019-0048-x

Shaffer, F., and Ginsberg, J. P. (2017). An overview of heart rate variability metrics and norms. *Front. Public Health* 5:258. doi: 10.3389/fpubh.2017.00258

Singh, Y., Hathaway, Q. A., Keishing, V., Salehi, S., Wei, Y., Horvat, N., et al. (2025). Beyond post hoc explanations: a comprehensive framework for accountable AI in medical imaging through transparency, interpretability, and explainability. *Bioengineering* 12:879. doi: 10.3390/bioengineering12080879

Stix, C. (2021). Actionable principles for artificial intelligence policy: three pathways. *Sci. Eng. Ethics* 27:15. doi: 10.1007/s11948-020-00277-3

Stroud, A. M., Minteer, S. A., Zhu, X., Ridgeway, J. L., Miller, J. E., and Barry, B. A. (2025). Patient information needs for transparent and trustworthy cardiovascular artificial intelligence: a qualitative study. *PLoS Digit. Health* 4:e0000826. doi: 10.1371/journal.pdig.0000826

Tasmurzayev, N., Amangeldy, B., Imankulov, T., Imanbek, B., Postolache, O. A., and Konysbekova, A. (2025). A wearable IoT-based measurement system for real-time cardiovascular risk prediction using heart rate variability. *Eng* 6:259. doi: 10.3390/eng6100259

Thayer, J. F., Yamamoto, S. S., and Brosschot, J. F. (2010). The relationship of autonomic imbalance, heart rate variability and cardiovascular disease risk factors. *Int. J. Cardiol.* 141, 122–131. doi: 10.1016/j.ijcard.2009.09.543

Tjoa, E., and Guan, C. (2021). A survey on explainable artificial intelligence (XAI): toward medical XAI. *IEEE Trans. Neural Networks Learn. Syst.* 32, 4793–4813. doi: 10.1109/TNNLS.2020.3027314

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* 25, 44–56. doi: 10.1038/s41591-018-0300-7

Tsuji, H., Larson, M. G., Venditti, F. J., Manders, E. S., Evans, J. C., Feldman, C. L., et al. (1996). Impact of reduced heart rate variability on risk for cardiac events. The Framingham heart study. *Circulation* 94, 2850–2855. doi: 10.1161/01.cir.94.11.2850

Tun, H. M., Rahman, H. A., Naing, L., and Malik, O. A. (2025). Trust in Artificial Intelligence-Based Clinical Decision Support Systems among Health Care Workers: systematic review. *J. Med. Internet Res.* 27:e69678. doi: 10.2196/69678

Tursunalieva, A., Alexander, D. L. J., Dunne, R., Li, J., Riera, L., and Zhao, Y. (2024). Making sense of machine learning: a review of interpretation techniques and their applications. *Appl. Sci.* 14:496. doi: 10.3390/app14020496

Voss, A., Schroeder, R., Heitmann, A., Peters, A., and Perz, S. (2015). Short-term heart rate variability--influence of gender and age in healthy subjects. *PLoS One* 10:e0118308. doi: 10.1371/journal.pone.0118308

Wang, D., Wang, L., Zhang, Z., Wang, D., Zhu, H., Gao, Y., et al. (2021). "Brilliant AI Doctor," in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment, (New York, NY: Association for Computing Machinery).

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., et al. (2019). Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.* 25, 1337–1340. doi: 10.1038/s41591-019-0548-6

Wong, K. K. L., Han, Y., Cai, Y., Ouyang, W., Du, H., and Liu, C. (2025). From Trust in Automation to trust in AI in healthcare: a 30-year longitudinal review and an interdisciplinary framework. *Bioengineering* 12:1070. doi: 10.3390/bioengineering12101070

Yedilkhan, D., Zhalgasbayev, A., Saleshova, S., and Khaimuldin, N. (2025). OIKAN: a hybrid AI framework combining symbolic inference and deep learning for interpretable information retrieval models. *Algorithms* 18:639. doi: 10.3390/a18100639