Proceedings of ICBioMed 2026 Symposium: Biomedical Engineering and Public Health: Real-Time Wearables for Disease Detection
DOI: 10.54254/2753-8818/2026.ES37007

Applications, Safety Challenges and Future Directions of AI Agents in Cardiovascular Prediction: A Review of Technological Evolution, System Architecture and Clinical Translation

Yijian Zheng

School of Public Health, Wenzhou Medical University, Wenzhou, China
zhengyijian04430@qq.com

Abstract. Cardiovascular disease remains a major contributor to mortality and long-term morbidity worldwide. For this reason, risk prediction, early detection, and decision support have become central tasks in digital health research. Cardiovascular prediction has developed from statistical scores to machine learning and deep learning models using clinical data, electrocardiograms, imaging, electronic health records, and multimodal inputs. AI agents add a system layer to this trajectory: they connect large language models, retrieval-augmented generation, tool use, code execution, multi-agent collaboration, and evidence tracing. Taking AgentMD, HeartAgent, and CardioAgent as representative systems, this review examines the architecture, clinical roles, safety risks, and future development of AI agents in cardiovascular prediction. The central argument is that AI agents are valuable not merely because they may improve prediction metrics, but because they can link risk calculators, deep learning models, medical knowledge bases, and clinical workflows into prediction systems that are explainable and auditable. At the same time, their deployment introduces prompt injection, hallucination, tool misuse, privacy leakage, adversarial attacks, multi-agent error propagation, and accountability risks. Future research should give more attention to lightweight reasoning, privacy-preserving implementation, prospective validation, and human-agent evaluation.

Keywords: AI agents, cardiovascular prediction, risk stratification, large language models, multi-agent systems, medical safety, clinical decision support

1. Introduction

1.1. Research background

The clinical burden of cardiovascular diseases is not only reflected in mortality, but also in the risks of long-term follow-up and readmission. According to 2025 AHA statistics, heart disease and stroke remain the leading causes of death and account for the majority of medical expenditures [1]. In clinical practice, cardiovascular risk prediction is used for tasks such as identifying high-risk

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.ES37007

individuals, assisting emergency triage, supporting treatment decision-making and conducting risk assessment.

Traditional cardiovascular prediction research includes population cohort studies and statistical models. These models provide clinicians with simple and easily deployable risk stratification tools, but most of them rely on limited variables, fixed time windows and preset functional forms, making it difficult to incorporate electronic information such as health records, electrocardiogram waveforms, imaging data and patient narratives.

Machine learning and deep learning have expanded this capability. Machine learning can capture non-linear relationships and variable interactions in routine clinical data [2]. Weng et al. [3] demonstrated that cardiovascular risk prediction using methods such as random forests, gradient boosting and neural networks can achieve better discrimination performance than traditional models. Deep learning further integrates electrocardiograms and imaging into the prediction process. These advances indicate that cardiovascular prediction has shifted from "scoring with a small number of variables" to "multimodal pattern recognition" [4].

1.2. Limitations of existing cardiovascular prediction and the motivation for AI agent intervention

However, existing models still have four categories of limitations: (1) They are disconnected from workflows, only outputting probabilities without providing action guidance; (2) A single model is difficult to integrate multi-source information such as structured data, text, electrocardiograms and imaging; (3) Their black-box nature undermines clinical trust; (4) Static prediction lacks active retrieval, tool invocation and interaction capabilities.

AI agents are being introduced to address these issues. Centered on LLMs as the reasoning core, they can perceive context, plan tasks, invoke tools, retrieve knowledge, execute actions and receive feedback. AgentMD, HeartAgent and CardioAgent demonstrate multi-agent differential diagnosis and integrated service ecosystems, marking the transformation of cardiovascular AI from "single-model prediction" to "agent-based decision support".

1.3. Scope and organization of the review

To ensure systematicity, this paper uses "AI agent CVD prediction" as keywords, and mainly retrieves relevant literature published in core journals at home and abroad in the past decade. It should be noted that this paper focuses on the application progress of AI agents in cardiovascular prediction models, while only briefly mentioning the basic theories and early classic models in cardiovascular prediction without in-depth discussion. This paper first sorts out the evolution context of cardiovascular prediction models; secondly, analyzes the role of AI agents in current cardiovascular prediction; then discusses the controversies and deficiencies in current research, such as security threats and defense strategies; subsequently, prospects possible future breakthrough directions based on existing defects. Finally, the full text is summarized.

2. Technological evolution of cardiovascular prediction models

Cardiovascular prediction models have evolved from traditional prediction risk models to deep learning. As shown in Figure 1, in the traditional statistical stage, risk factors from cohort studies were converted into generalizable and calculable scores. For example, Framingham integrated

DOI: 10.54254/2753-8818/2026.ES37007

multiple variables to estimate overall coronary heart disease risk [5], while SCORE assessed patient risk with fatal events as the endpoint [6].

Weng and Ambale-Venkatesh et al. applied machine learning to cardiovascular event prediction in the MESA study, demonstrating the improved prediction performance of machine learning in routine and imaging data [3, 7]. Hyland et al. used machine learning for early prediction of circulatory failure, providing a pre-intervention window for high-risk patients [8].

In the deep learning era, ECG has become the core input: Hannun et al. proved that deep neural networks can achieve expert-level arrhythmia detection in ambulatory electrocardiograms [9]; Attia et al. used AI-enabled ECG to screen for left ventricular systolic dysfunction, indicating that ECG contains structural and functional information that is difficult to directly identify through traditional manual interpretation [10].

In addition to ECG, deep learning has also expanded to retina, imaging and EHR. Poplin et al. predicted cardiovascular risk factors from fundus photographs, demonstrating that microvascular images can provide clues to systemic cardiovascular risk [11]. Rajkomar et al. proposed a scalable deep learning framework for EHR, demonstrating end-to-end learning capabilities in multiple clinical prediction tasks [12]. The advantages of deep learning include reduced manual feature engineering, direct processing of complex inputs, and potential in multimodal fusion.

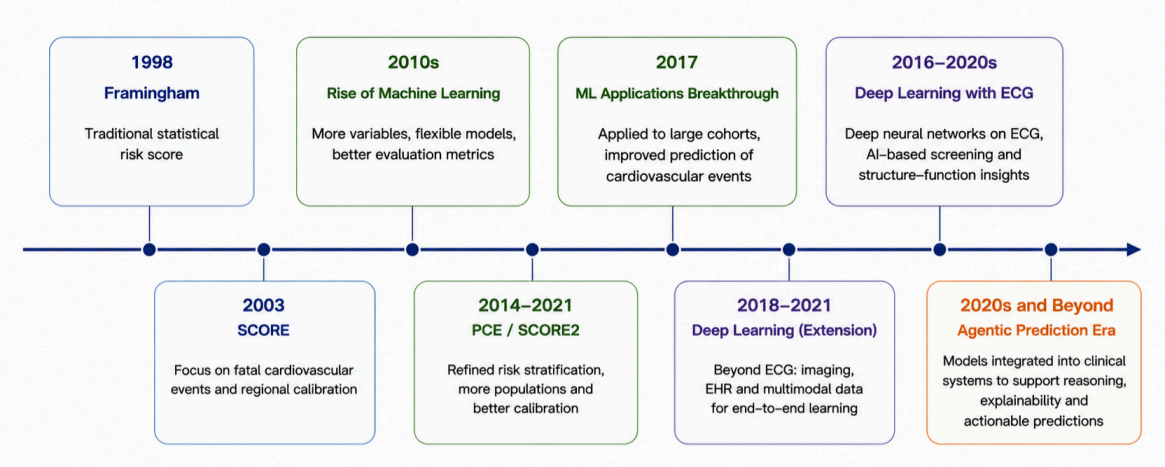

Figure 1. Temporal evolution of cardiovascular prediction models from Framingham to deep learning

The evolution from statistics to deep learning does not automatically solve workflow problems. AI agents are a systematic response to this task. Agent-based prediction does not negate existing models, but embeds them into clinical reasoning and collaboration systems—the next stage of cardiovascular prediction is to build systems that understand when, why and how to use models [13, 14].

3. Applications of AI agents in cardiovascular prediction

3.1. Core concepts and architecture of AI agents

The agent architecture distributes complex tasks to multiple subsystems with different roles. As shown in Figure 2, a coordinating agent is responsible for parsing intentions, splitting tasks and integrating outputs; specialist agents are responsible for specific disease, imaging, electrocardiogram or medication tasks; review agents are responsible for self-inspection, rebuttal and quality control;

DOI: 10.54254/2753-8818/2026.ES37007

citation or evidence agents are responsible for retrieving and verifying supporting materials. Medical multi-agent research has shown potential in zero-shot reasoning and medical decision-making. For example, MDAgents explores adaptive collaborative medical decision-making [15]. Multi-agent design is particularly suitable for cardiovascular prediction, as cardiovascular cases often involve emergency care, imaging, electrocardiography, laboratory tests, long-term medical history and non-cardiac differential diagnosis.

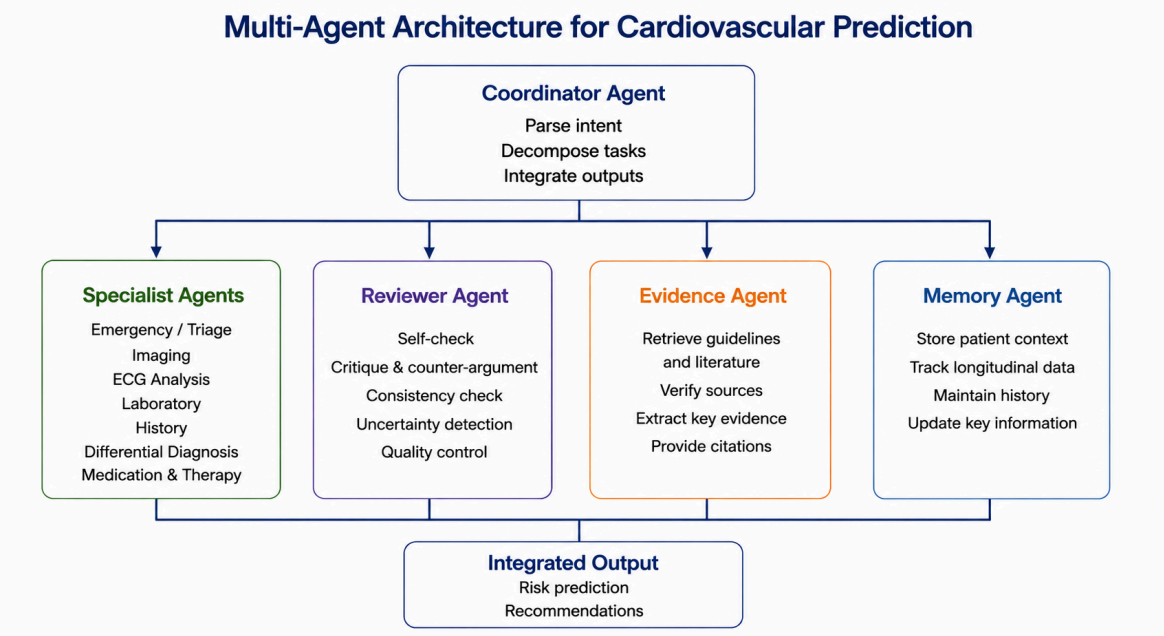

Figure 2. Schematic diagram of a general cardiovascular AI agent architecture: a coordinating agent as the center connecting multiple types of subsystem agents

3.2. Key roles of AI agents in cardiovascular prediction

3.2.1. Multi-source data fusion and dynamic profile construction

One of the core capabilities of AI agents is to integrate heterogeneous medical data into dynamically updated patient profiles. Kaushik constructed a unified multimodal learning framework that integrates genomic variation, cardiac MRI, electrocardiogram waveforms, wearable streaming data and structured EHR data to predict individualized CVD risk [16]. Nargesi developed an LLM-driven AI agent that can retrieve information from multimodal data sources such as clinical notes, electrocardiograms and echocardiograms to classify cardiovascular diseases [17].

3.2.2. In-depth mining of complex risk patterns

Traditional cardiovascular prediction models mainly assess risk based on predefined variables and static algorithms. Although deep learning has improved the ability to capture non-linear relationships, most models remain limited to specific tasks and process only a single form of data.

AI agents further expand this capability by actively integrating heterogeneous clinical information and performing dynamic reasoning across multiple data sources. In the field of cardiovascular disease prediction, AI agents can simultaneously analyze electronic health records, laboratory test results, electrocardiograms (ECG), cardiac imaging data, wearable sensor data,

DOI: 10.54254/2753-8818/2026.ES37007

medication information and longitudinal follow-up data. Deep learning can discover hidden cardiovascular disease patterns, while AI agents analyze these patterns against the background of patients' clinical history, multimodal evidence and medical knowledge, thereby generating interpretable and practically applicable prediction results. For example, AI ECG can identify the risk of low ejection fraction, and clinicians can verify whether the result is consistent with symptoms, previous history of heart failure, ultrasound findings and medication adjustments [10, 18].

3.2.3. Autonomous reasoning and clinical recommendation generation

The third role of AI agents is to convert prediction results into clinically readable recommendations. Large language models have been proven to encode certain medical knowledge, and conversational diagnostic AI also shows potential for interaction with patients and doctors. The ZODIAC framework proposed by Zhou is an important progress in this direction—it is an LLM-driven multi-agent collaboration framework that processes patients' multimodal data through the collaboration of multiple specialized agents [19]. Li systematically evaluated the performance of six multimodal large language models in ejection fraction prediction, view classification, aortic stenosis severity assessment and cardiovascular disease classification, and found that domain-specific pre-training and model adaptation are crucial for cardiovascular decision support [20].

3.3. Overall level and limitations of current applications

Currently, AI agents in cardiovascular prediction are still in the stage from proof of concept to early clinical validation. Their advantages are mainly reflected in system integration, tool invocation, explanation generation and evidence tracking, rather than single model metrics. The long-term development of medical AI has expanded from single-point algorithms to the co-evolution of data, knowledge, models and organizational processes [21]. Studies such as GeneGPT and MedCPT demonstrate that domain tools and biomedical retrieval models can improve the ability of large language models to access specialized knowledge [22, 23]. Research on clinical calculation tools augmented with ChatGPT also shows that doctor-defined tools can improve performance in medical calculation tasks [24].

Limitations cannot be ignored. First, LLMs exhibit identifiable gender bias and entrenched racial bias in cardiovascular risk attribution [25, 26]. Second, the quality of explanations still requires manual verification. The hallucination problem is particularly dangerous in medical texts [27]. Third, most cardiovascular AI models perform excellently in retrospective studies but are difficult to generalize in prospective clinical settings [28, 29]. Fourth, the chain of responsibility in multi-agent systems is unclear. If the coordinating agent integrates contradictory conclusions from multiple sub-agents, clinical practice needs to know which step went wrong and how to correct it.

4. Safety challenges and countermeasures of AI agents in cardiovascular prediction

4.1. Classification of security threats

AI agents face a wider range of security risks than ordinary prediction models, up to seven categories, as shown in Figure 3. The first category is input attacks. Adversarial example research shows that neural networks can be misled by perturbations imperceptible to humans [30]. In electrocardiogram, imaging or wearable device data, noise, device differences or malicious tampering may all alter model outputs. The second category is prompt injection and instruction

DOI: 10.54254/2753-8818/2026.ES37007

hijacking. Indirect prompt injection can inject malicious instructions into LLM-integrated applications through web pages, documents or retrieved content, causing agents to ignore original goals or leak information [31].

The third category is tool abuse and privilege escalation. If permission control is not strict, the model may execute wrong commands, or even trigger action recommendations without doctor authorization. The fourth category is hallucination and evidence mismatch. When evidence is insufficient, agents may forcibly match irrelevant evidence, leading to wrong decisions [27]. The fifth category is privacy attacks. Membership inference, model inversion and training data extraction studies show that machine learning systems may leak training data or patient attributes. When agents use clinical notes, imaging, electrocardiograms and long-term follow-up data, privacy risks increase significantly [32, 33].

The sixth category is multi-agent error propagation. As agents reinforce each other's decisions, errors may spread across the network. Since downstream decisions depend on upstream signals, errors in subsequent AI models may amplify errors in earlier systems [34]. The seventh category is supply chain risks. Contamination in any link—knowledge bases, model weights, prompt templates, retrieval indexes, third-party APIs—may be amplified and spread to the entire application [35, 36].

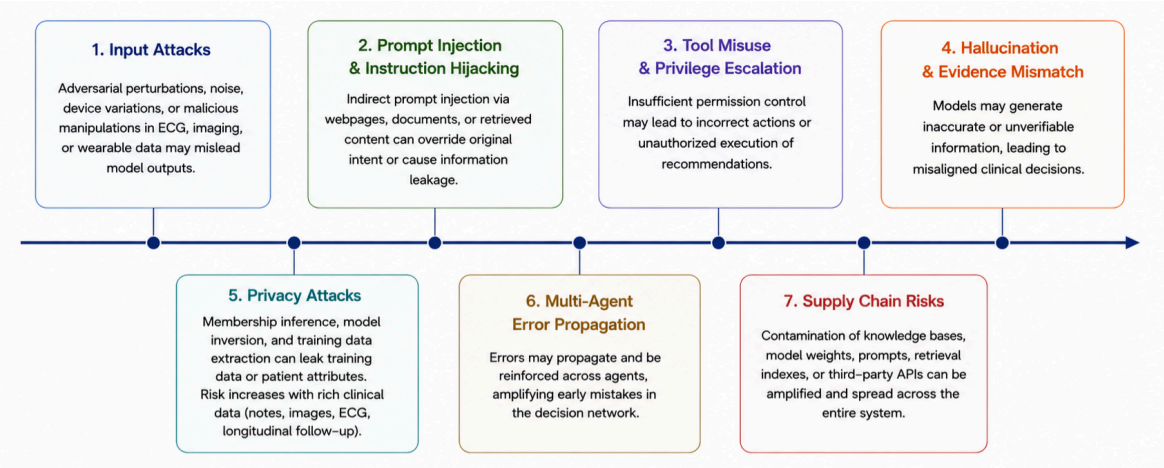

Figure 3. Distribution of seven types of security threats across input, model/prompt, tool, knowledge, collaboration and output processes

4.2. Representative defense technologies and strategies

Defending against AI agent risks needs to be carried out in layers. The input layer should include data validation, anomaly detection and source labeling. The TAZAI framework extends the zero-trust principle to AI agents, covering identity authentication, context-aware policies and real-time behavior monitoring [37]. The model layer can adopt adversarial training, calibration, external validation and subgroup fairness analysis; interpretability methods such as LIME and SHAP can help identify the main basis of model predictions [38, 39].

The tool layer requires the principle of least privilege. Agents should only call tools related to the current task, and high-risk actions must be confirmed by humans. Code interpreters should be sandboxed, clinical database queries should be audited, and the ability to write to EHR or generate medical orders should be disabled by default [40]. The RAG layer should distinguish levels of trusted sources, and label guidelines, textbooks, peer-reviewed papers, preprints and web information. Citation verification should check whether retrieved fragments actually support the

DOI: 10.54254/2753-8818/2026.ES37007

generated conclusions [41]. The MedCite framework adopts a hierarchical two-stage retrieval strategy in medical question answering, prioritizing relevant documents [42]; the Biomed-R2 framework achieves patient-aligned retrieval and iterative evidence verification through two-stage RAG [43]. These solutions all embody the idea of combining lexical recall with semantic re-ranking.

The privacy layer can adopt technologies such as federated learning and differential privacy. Federated learning allows multiple institutions to train models without sharing original patient data [44], and has been widely discussed in digital health and medical imaging [45-47]. Differential privacy can reduce the risk of individual training samples being inferred [48], but may sacrifice model accuracy. For cardiovascular agents, privacy-enhancing technologies should not only protect the training phase, but also cover clinician feedback.

The governance layer requires documentation and reporting specifications. Work should follow established reporting and evaluation guidelines, including frameworks such as TRIPOD+AI, CONSORT-AI and DECIDE-AI. Although these frameworks are not specifically designed for AI agents, they provide an important basis for transparent reporting of agent-based clinical systems. Future extended versions are expected to incorporate agent-specific features, including autonomous planning capabilities, tool use capabilities, memory functions, multi-agent collaboration mechanisms and human supervision mechanisms.

4.3. Special safety and compliance requirements in cardiovascular prediction scenarios

The particularity of cardiovascular prediction scenarios lies in the time sensitivity and high consequences of many tasks. In atrial fibrillation stroke risk assessment, wrong recommendations may affect anticoagulation decisions [49]. AI agents bring new risks beyond traditional regulation. Compliance should meet the following requirements: clear scope of application, product use should be defined in the declaration, adequate human supervision, traceable evidence chain, auditable logs, and continuous post-deployment monitoring. Current frameworks such as the FDA, the EU AI Act and China's *Opinions on the Standardized Application of Agents* are evolving from static approval to full-lifecycle dynamic governance, requiring developers to establish a compliance system throughout design, operation and retirement.

5. Future development directions

5.1. Security gaps in long-term operation, multi-agent systems and the physical world

For cardiovascular agents oriented towards long-term management, it is necessary to simultaneously address system drift caused by memory saturation and error accumulation [50]. The system needs to track patient status, medication adjustments and risk changes, as well as limit error accumulation during long-term operation. VeriGuard provides security guarantees for LLM agents through offline formal verification and online runtime monitoring [51]. However, multi-agent collaboration still requires clearer security protocols, otherwise the system may adopt harmful recommendations.

In terms of the physical world, wearable device errors, network delays and data tampering all affect prediction [52]. Cybersecurity, device quality control and clinical risk management need to be designed under the same framework.

DOI: 10.54254/2753-8818/2026.ES37007

5.2. Lightweight interpretable architectures

Large model capabilities have an important impact on agent performance, but high-performance models are often accompanied by high costs, difficulties in closed-source auditing and privacy limitations. One of the future directions is to build lightweight interpretable architectures: smaller specialized models are responsible for specific tasks. Studies show that Meissa, as a lightweight medical language model, replicates advanced agent capabilities through trajectory distillation and can run offline, reducing latency and enhancing privacy compared to cutting-edge models [53]. In the cardiovascular field, lightweight heart sound classification models based on knowledge distillation can be deployed on wearable devices to realize real-time monitoring and feedback for patients with cardiovascular diseases [54].

5.3. Clinical validation pathways and human-agent collaboration evaluation

Clinical validation of AI agents should not stop at retrospective accuracy. More reasonable pathways include technical validation, clinical expert evaluation, workflow simulation studies, and prospective clinical trials or real-world studies, as shown in Figure 4. Evaluation indicators need to cover tool selection, calculation accuracy, retrieval support rate, citation authenticity, explanation usability, missed dangerous diagnoses, workflow efficiency, patient outcomes and adverse events [55-58].

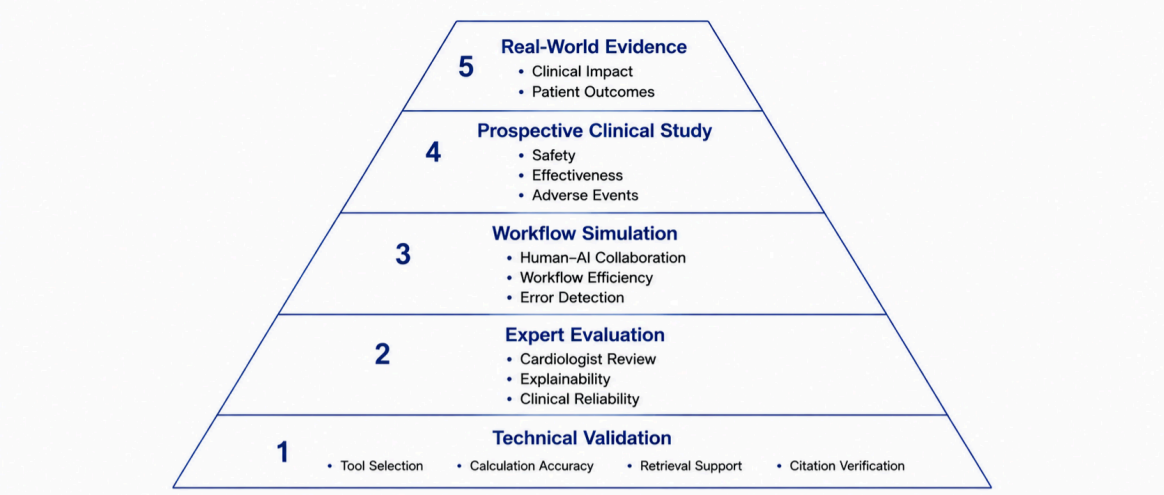

Figure 4. Clinical validation evidence pyramid for AI agents from technical validation to prospective clinical studies

Human-agent collaboration evaluation is particularly critical. Results from physician collaboration with HeartAgent suggest that the optimal positioning of agents may be to augment rather than replace physicians [59]. Future research should evaluate different collaboration modes, and evaluation indicators should also include whether physicians over-rely on the system, whether they can identify wrong recommendations, whether patients understand risks, and whether the system changes doctor-patient communication. The value of AI agents in cardiovascular prediction can only be fully demonstrated when technical indicators, clinical behavior and patient outcomes improve simultaneously.

DOI: 10.54254/2753-8818/2026.ES37007

5.4. Regulatory approval, standardization and payment responsibility frameworks

In addition to model performance and privacy technologies, the translation of cardiovascular AI agents also depends on regulatory approval, standardization and payment responsibility arrangements. At the regulatory level, evaluation should not be based solely on a single algorithm, but should combine the characteristics of "software as a medical device" and clinical decision support systems to clarify the scope of application, risk level, modifiable boundaries and human supervision requirements. For systems that can dynamically retrieve, invoke tools or generate recommendations, it is also necessary to specify re-verification trigger conditions after updates to knowledge bases, prompt templates, tool versions and model iterations.

In terms of standardization, the AHA Scientific Advisory recommends dividing AI evaluation into three stages: pre-deployment, implementation and post-deployment, and emphasizes local validation, bias assessment and post-deployment monitoring [60]. Subsequent efforts are needed to unify task classification, input data quality standards, tool invocation logs, RAG evidence chains, failure mode reports and monitoring indicators, so that different institutions can compare performance and trace responsibilities.

In terms of payment and insurance, the Belgian Society of Cardiology has proposed AI reimbursement strategies in the cardiovascular field [61]. However, agent systems still need to answer three questions: whether their assistance constitutes necessary diagnostic and treatment support, which scenarios can be included in reimbursement, and how to divide responsibilities between system recommendations and the physician's final decision.

Approval frameworks, technical standards, responsibility attribution and payment mechanisms need to mature together. Only when these conditions are gradually met can cardiovascular AI agents move from research prototypes to a sustainable clinical service system.

6. Conclusion

The evolution of cardiovascular prediction models reflects the transformation of medical AI from tools to systems: early stages saw statistical scoring with a small number of variables, followed by machine learning for high-dimensional clinical data, and then deep learning using ECG, imaging, EHR and multimodal inputs. AI agents represent a new system layer. They can organize risk models, clinical calculators, deep learning modules, medical knowledge bases, RAG, tool invocation and human-computer interaction, so that prediction no longer stays at "giving a score", but further extends to situational understanding, tool selection, evidence explanation and collaborative decision-making.

However, agentification is not inherently equivalent to safety or reliability. Cardiovascular prediction has the characteristics of high risk, high time sensitivity and strong compliance. Prompt injection, hallucination, privacy leakage, tool misuse or multi-agent error propagation may all affect clinical decision-making. The future focus is to connect the auditability of traditional models, the pattern recognition capabilities of machine learning and deep learning, the tool-based reasoning capabilities of agents, and clinical governance rules, rather than making a simple trade-off between traditional models and agent systems. Only when privacy protection, human supervision, prospective validation and post-deployment monitoring are sufficient can AI agents become a reliable component in cardiovascular prediction clinical practice.

DOI: 10.54254/2753-8818/2026.ES37007

References

[1] Martin, S. S., et al. (2025). 2025 heart disease and stroke statistics: A report of US and global data from the American Heart Association. *Circulation, 151*(8), e41–e660. https://doi.org/10.1161/cir.0000000000001303

[2] Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H. F., & van der Schaar, M. (2019). Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PLOS ONE, 14*(5), e0213653. https://doi.org/10.1371/journal.pone.0213653

[3] Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One, 12*(4), e0174944. https://doi.org/10.1371/journal.pone.0174944

[4] Kim, D. S., Patel, S. N., & Yang, E. (2026). More is better: Harnessing artificial intelligence and big data for enhanced cardiovascular risk prediction. *European Journal of Preventive Cardiology, 33*(4), 461–463. https://doi.org/10.1093/eurjpc/zwaf645

[5] Wilson, P. W. F., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., & Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation, 97*(18), 1837–1847. https://doi.org/10.1161/01.CIR.97.18.1837

[6] Conroy, R. M., et al. (2003). Estimation of ten-year risk of fatal cardiovascular disease in Europe: The SCORE project. *European Heart Journal, 24*(11), 987–1003. https://doi.org/10.1016/s0195-668x(03)00114-3

[7] Ambale-Venkatesh, B., et al. (2017). Cardiovascular event prediction by machine learning: The multi-ethnic study of atherosclerosis. *Circulation Research, 121*(9), 1092–1101. https://doi.org/10.1161/circresaha.117.311312

[8] Hyland, S. L., et al. (2020). Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine, 26*(3), 364–373. https://doi.org/10.1038/s41591-020-0789-4

[9] Hannun, A. Y., et al. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine, 25*(1), 65–69. https://doi.org/10.1038/s41591-018-0268-3

[10] Attia, Z. I., et al. (2019). Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. *Nature Medicine, 25*(1), 70–74. https://doi.org/10.1038/s41591-018-0240-2

[11] Poplin, R., et al. (2018). Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering, 2*(3), 158–164. https://doi.org/10.1038/s41551-018-0195-0

[12] Rajkomar, A., et al. (2018). Scalable and accurate deep learning with electronic health records. *npj Digital Medicine, 1*(1), 18. https://doi.org/10.1038/s41746-018-0029-1

[13] M. Heidari et al. (2025). Echo-CoPilot: A multiple-perspective agentic framework for reliable echocardiography interpretation. arXiv:2512.09944. doi: 10.48550/arXiv.2512.09944.

[14] Dunn, P., Conard, S., Ali, A., Henley, C., & Kolankarai, R. (2026). From prediction to action: The role of neuro symbolic agentic AI in cardiovascular care. *International Journal of Cardiology Innovations, 2*, 100002. https://doi.org/10.1016/j.ijci.2026.100002

[15] Y. Kim et al. (2024). MDAgents: An adaptive collaboration of LLMs for medical decision-making. arXiv:2404.15155. doi: 10.48550/arXiv.2404.15155.

[16] R. Kaushik & E. Kaushik. Causal and federated multimodal learning for cardiovascular risk prediction under heterogeneous populations. arXiv:2601.06140. doi: 10.48550/arXiv.2601.06140.

[17] Nargesi, A., Kitonyo, J., Maddah, M., & Ellinor, P. (2025). A large language model agent for multimodal evaluation of cardiovascular disease. *European Heart Journal, 46*(Supplement_1). doi: 10.1093/eurheartj/ehaf784.4504

[18] Yao, X., et al. (2021). Artificial intelligence–enabled electrocardiograms for identification of patients with low ejection fraction: A pragmatic, randomized clinical trial. *Nature Medicine, 27*(5), 815-819. doi: 10.1038/s41591-021-01335-4

[19] Zhou, Y., et al. (n.d.). Zodiac: A cardiologist-level LLM framework for multi-agent diagnostics. arXiv:2410.02026. https://doi.org/10.48550/arXiv.2410.02026

[20] Li, J., et al. (2025). Exploring multimodal large language models on transthoracic echocardiogram (TTE) tasks for cardiovascular decision support. *Journal of Biomedical Informatics, 171*, 104930. https://doi.org/10.1016/j.jbi.2025.104930

[21] Jiang, F., et al. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology, 2*(4), 230–243. https://doi.org/10.1136/svn-2017-000101

[22] Jin, Q., Yang, Y., Chen, Q., & Lu, Z. (2024). GeneGPT: Augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics, 40*(2). https://doi.org/10.1093/bioinformatics/btae075

DOI: 10.54254/2753-8818/2026.ES37007

[23] Jin, Q., et al. (2023). MedCPT: Contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics, 39*(11). https://doi.org/10.1093/bioinformatics/btad651
[24] Goodell, A. J., Chu, S. N., Rouholiman, D., & Chu, L. F. (2025). Large language model agents can use tools to perform clinical calculations. *npj Digital Medicine, 8*(1), 163. https://doi.org/10.1038/s41746-025-01475-8
[25] Chan, J. T. N., & Kwek, R. K. (2025). Uncovering bias and variability in how large language models attribute cardiovascular risk. *Frontiers in Digital Health, 7*. https://doi.org/10.3389/fdgth.2025.1710594.
[26] Cau, R., Pisu, F., Suri, J. S., & Saba, L. (2025). Addressing hidden risks: Systematic review of artificial intelligence biases across racial and ethnic groups in cardiovascular diseases. *European Journal of Radiology, 183*, 111867. https://doi.org/10.1016/j.ejrad.2024.111867.
[27] Ji, Z., et al. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), Article 248. https://doi.org/10.1145/3571730.
[28] Delaporte, A., Adham, A., Coeckelenbergh, S., & Joosten, A. (2026). Artificial intelligence guided hemodynamic management: Discrepancies between retrospective findings and prospective evidence. *Minerva Anestesiologica, 92*(3), 236–237. https://doi.org/10.23736/s0375-9393.25.19469-8.
[29] Venturiello, D., Vignaroli, W., & Nasso, G. (2026). AI-enhanced electrocardiography as a digital biomarker platform in cardiovascular medicine: Clinical applications, validation gaps and future implementation pathways. *Open Heart, 13*(2). https://doi.org/10.1136/openhrt-2026-004211.
[30] Goodfellow, I. J., Shlens, J., & Szegedy, C. Explaining and harnessing adversarial examples. arXiv:1412.6572. https://doi.org/10.48550/arXiv.1412.6572.
[31] Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security,* Copenhagen, Denmark. https://doi.org/10.1145/3605764.3623985
[32] Carlini, N., et al. (2020). Extracting training data from large language models. arXiv:2012.07805. https://doi.org/10.48550/arXiv.2012.07805
[33] Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2016). Membership inference attacks against machine learning models. arXiv:1610.05820. https://doi.org/10.48550/arXiv.1610.05820
[34] Athni, T. S. (2026). Emerging risks of AI-to-AI interactions in health care: Lessons from Moltbook. *Journal of Medical Internet Research, 28*, e96199. https://doi.org/10.2196/96199
[35] Fogel, A., Hofman, O., Cohen, E., & Vainshtein, R. (2026). Inference-time backdoors via chat templates: From LLM supply chains to agentic system compromise. arXiv:2602.04653. https://doi.org/10.48550/arXiv.2602.04653
[36] Dehghantanha, A., & Homayoun, S. (2026). SoK: The attack surface of agentic AI -- Tools, and autonomy. arXiv:2603.22928. https://doi.org/10.48550/arXiv.2603.22928
[37] Qi, J., et al. (n.d.). Towards trustworthy agentic AI: A comprehensive survey of safety, robustness, privacy, and system security. arXiv:2605.23989. https://doi.org/10.48550/arXiv.2605.23989
[38] Ribeiro, A. H., et al. (2020). Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications, 11*(1), 1760. https://doi.org/10.1038/s41467-020-15432-4
[39] Lundberg, S., & Lee, S.-I. (n.d.). A unified approach to interpreting model predictions. arXiv:1705.07874. https://doi.org/10.48550/arXiv.1705.07874
[40] Kearns, L. (n.d.). Why trust your agent? Empirical security gains from TRiSM-guided agentic workflows in healthcare. arXiv:2606.28666. https://doi.org/10.48550/arXiv.2606.28666
[41] Qian, H., et al. (n.d.). VeriCite: Towards reliable citations in retrieval-augmented generation via rigorous verification. arXiv:2510.11394. https://doi.org/10.48550/arXiv.2510.11394
[42] Wang, X., et al. (n.d.). MedCite: Can language models generate verifiable text for medicine? arXiv:2506.06605. https://doi.org/10.48550/arXiv.2506.06605
[43] Guan, H., Li, C., Zhang, W., Lian, Y., Guo, J., & Lu, W. (2026). Biomed-R2: Joint diversity retrieval and evidence reasoning for biomedical question answering. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 16597–16601). https://doi.org/10.1109/ICASSP55912.2026.11463381
[44] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. Communication-efficient learning of deep networks from decentralized data. arXiv:1602.05629. https://doi.org/10.48550/arXiv.1602.05629
[45] Rieke, N., et al. (2020). The future of digital health with federated learning. *npj Digital Medicine, 3*(1), 119. https://doi.org/10.1038/s41746-020-00323-1
[46] Sheller, M. J., et al. (2020). Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports, 10*(1), 12598. https://doi.org/10.1038/s41598-020-69250-1
[47] Kaissis, G. A., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence, 2*(6), 305–311. https://doi.org/10.1038/s42256-

DOI: 10.54254/2753-8818/2026.ES37007

020-0186-1
[48] Berini, A. D. E., et al. (2026). Security and privacy in LLMs: A comprehensive survey of threats and mitigation strategies. *Information Fusion, 132*, 104241. https://doi.org/10.1016/j.inffus.2026.104241
[49] Islam, M. M., & Okere, A. N. (2026). Advancing stroke prevention in atrial fibrillation: A systematic review of machine learning–based risk prediction models. *International Journal of Medical Informatics, 217*, 106504. https://doi.org/10.1016/j.ijmedinf.2026.106504.
[50] Bousetouane, F. AI agents need memory control over more context. arXiv:2601.11653. https://doi.org/10.48550/arXiv.2601.11653.
[51] Miculicich, L., et al. VeriGuard: Enhancing LLM agent safety via verified code generation. arXiv:2510.05156. https://doi.org/10.48550/arXiv.2510.05156.
[52] Srinivasa Rao, G., & Muneeswari, G. 18 - IoT based cardiovascular health monitoring: Ensuring security with machine learning and AI techniques. In S. K. H. Islam, D. Samanta, A. K. Pal, & U. Iqbal (Eds.), *IoT Security* (pp. 305–331). Academic Press.
[53] Chen, Y., Bai, X., Pan, Y., Zhou, Z., & Yuille, A. Meissa: Multi-modal medical agentic intelligence. arXiv:2603.09018. https://doi.org/10.48550/arXiv.2603.09018.
[54] Song, Z., et al. (2023). Cutting weights of deep learning models for heart sound classification: Introducing a knowledge distillation approach. *Annu Int Conf IEEE Eng Med Biol Soc, 2023*, 1–4. https://doi.org/10.1109/embc40787.2023.10340704.
[55] Collins, G. S., et al. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ, 385*, e078378. https://doi.org/10.1136/bmj-2023-078378
[56] Liu, X., et al. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nature Medicine, 26*(9), 1364–1374. https://doi.org/10.1038/s41591-020-1034-x
[57] Cruz Rivera, S., et al. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. *Nature Medicine, 26*(9), 1351–1363. https://doi.org/10.1038/s41591-020-1037-7
[58] Vasey, B., et al. (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature Medicine, 28*(5), 924–933. https://doi.org/10.1038/s41591-022-01772-9
[59] Zhou, S., et al. HeartAgent: An autonomous agent system for explainable differential diagnosis in cardiology. arXiv:2603.10764. https://doi.org/10.48550/arXiv.2603.10764
[60] Casado Arroyo, R., Jensen, M. T., Dilaveris, P., Locati, E. T., & De Lucia, R. (2024). The ESC Working Group on e-Cardiology. *European Heart Journal, 45*(32), 2908–2910. https://doi.org/10.1093/eurheartj/ehae332
[61] P. Lancellotti, M. De Pauw, M. Claeys, and o. t. B. S. o. Cardiology,(2020), The Belgian Society of Cardiology in Perspective, *European Heart Journal, 41*(31), 2932-2933, doi: 10.1093/eurheartj/ehaa320.