Journal of Clinical Trials and Case Studies

Open access

*Commentary*

# Beyond Predictive Accuracy: Clinical Effectiveness and Actionable Explainability of AI-Based Cardiac Arrest Surveillance in General Wards

**Minsoo Kim[1], Dongjoon Yoo[2,*]**

[1]*School of Pharmacy, Sungkyunkwan University, 2066 Seobu-ro, Jangan-gu, Suwon 16419, South Korea*

[2,*]*Department of Critical Care Medicine and Emergency Medicine, Inha University Hospital, Incheon 22212, Republic of Korea*

***Correspondence:** Dongjoon Yoo**,** Department of Critical Care Medicine and Emergency Medicine, Inha University Hospital, Incheon 22212, Republic of Korea, E-mail: djinyoo@gmail.com; DOI: 10.1042/JCTCS.8.2.0040

## Abstract

The clinical management of In-Hospital Cardiac Arrest (IHCA) is undergoing a paradigm shift from reactive, score-based protocols to proactive, AI-driven surveillance. However, the majority of AI-enabled Medical Devices (SaMD) have historically relied on retrospective "*in silico*" validation, failing to demonstrate tangible improvements in patient-centered outcomes. This mini-review evaluates the Deep learning-based Cardiac Arrest Risk Score (DeepCARS®) through the lens of a recent landmark prospective trial involving over 35,000 admissions. We argue that for AI to become the "standard of care," as advocated by the American Heart Association (AHA) and JAMA, research must move beyond diagnostic metrics like AUROC toward hard endpoints such as mortality and cardiac arrest incidence. Furthermore, we critique technical explainability methods like SHAP (Shapley Additive Explanations) for their lack of clinical utility, proposing instead a model of "medical explainability" where AI alerts are functionally mapped to disease-specific interventions, such as fluid resuscitation for sepsis. The results of this trial demonstrate that integrated, actionable AI can reduce cardiac arrest incidence by 46% and mortality by 35%, providing a blueprint for the future of acute care surveillance.

**Received date:** February 14, 2026; **Accepted date:** February 26, 2026; **Published date:** March 10, 2026

**Citation:** Minsoo Kim, Dongjoon Yoo (2026) Beyond Predictive Accuracy: Clinical Effectiveness and Actionable Explainability of AI-Based Cardiac Arrest Surveillance in General Wards. J Clin Trial Case Stud, 8:2.



## Introduction: The Crisis of Ward-Based Surveillance

In Hospital Cardiac Arrest (IHCA) remains a primary cause of preventable mortality, with survival rates remaining stagnant at approximately 21% despite the widespread implementation of Rapid Response Systems (RRS) **[1]**. The "afferent limb" of RRS, the detection of deteriorating patients is currently dominated by manual Early Warning Scores (EWS) like NEWS and MEWS. These systems are limited by their linear thresholds and intermittent nature, often identifying deterioration only when it is too late for life-saving stabilization.

Artificial Intelligence (AI) and Deep Learning (DL) provide the capability to identify complex, non-linear patterns of physiological decline hours before clinical manifestation. However, a profound "implementation gap" exists between the statistical accuracy of these models and their real-world clinical utility **[2]**. This review analyzes the transition of AI from a diagnostic curiosity to a clinically effective interventional tool, focusing on the necessity of prospective validation and actionable explainability.

## The Evidence Gap: The Pitfalls of Retrospective Validation

A critical barrier to the adoption of medical AI is the reliance on retrospective validation. While retrospective studies provide a cost-effective starting point, they fail to account for the "intervention effect"; how the clinician's response to an alert alters the patient's outcome.

Recent systematic analyses published in Nature Medicine reveal a concerning lack of rigorous evidence in the AI field. An evaluation of 521 FDA-authorized AI devices found that approximately 43% lacked any published clinical validation data **[3]**. Of those that were validated, only 4% were supported by randomized controlled trials, with 28% relying solely on retrospective data **[3,4]**. This lack of "*in vivo*" testing is a latent safety risk, as algorithms can experience a performance drop when moved from curated datasets to the noisy, dynamic environment of a hospital ward **[10].** Regulatory bodies and clinical guidelines from the AHA now emphasize that prospective validation is a prerequisite for establishing the clinical effectiveness of any early warning AI **[5]**.

## The DeepCARS® Prospective Trial: A Benchmark for Clinical Effectiveness

The study "Clinical Effectiveness of an Artificial Intelligence-Based Prediction Model for Cardiac Arrest in General Ward-Admitted Patients" addresses these systemic deficits through a one-year, prospective, non-randomized interventional trial **[6]**.

### Study Design and Methodology

Conducted at a tertiary academic hospital, the trial screened 36,922 admissions, focusing on a "target cohort" of 2,906 patients whose DeepCARS® scores reached a high-risk threshold of 95 **[6]**. Patients were allocated to either an AI-SaMD–guided cohort (intervention/reassessment within 24 hours of an alert) or a usual care cohort. Unlike prior research, this study utilized Propensity Score Matching (PSM) and E-value sensitivity analyses to ensure the findings were robust to confounding, aiming to mimic the rigor of a randomized trial [6].

### Patient-Centered Outcomes

In alignment with editorial standards from JAMA and the AHA, the primary outcome was ward-based cardiac arrest, while secondary outcomes included all-cause in-hospital mortality **[2,5]**. The incidence of cardiac arrest was significantly lower in the AI-guided group compared to usual care group (1.06% *vs.* 2.07%; adjusted risk ratio, 0.54) **(Table 1, Figure 1) [6]**. Similarly, mortality rates declined significantly (adjusted RR, 0.65) [6]. These findings demonstrate that AI alerts, when acted upon, can reduce the risk of death by 35% without requiring additional institutional resources. However, improved outcomes alone may not ensure adoption, underscoring the need for explainability that supports actionable bedside decision-making, especially given the black-box nature of deep learning models.

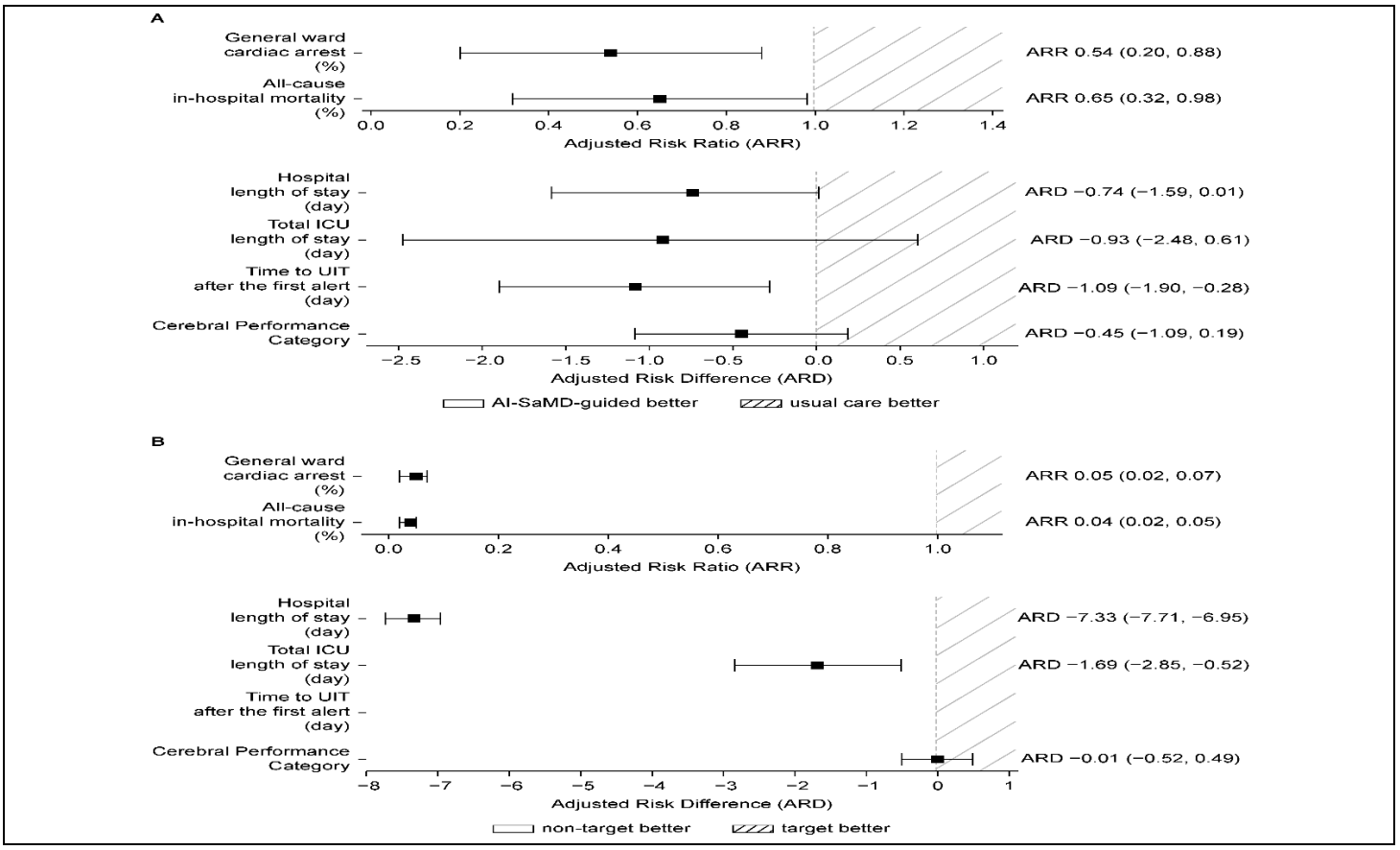


**Figure 1:** Association between AI-SaMD implementation and patient outcomes.

(A) Adjusted outcomes comparing the AI-SaMD-guided and usual care cohorts.

(B) Adjusted outcomes comparing the target and non-target cohorts.

**Table 1:** Adjusted outcomes based on AI-SaMD-guided intervention.

| Variables | AI-SaMD-Guided Cohort (n = 1409) | Usual Care Cohort (n = 1497) | Adjusted Risk Ratio or Adjusted Risk Difference (95% CI) | p-Value |
|---|---|---|---|---|
| **Primary outcome** | | | | |
| General ward cardiac arrest (n) | 15 (1.06%) | 31 (2.07%) | ARR 0.54 (0.20, 0.88) | ** |
| **Secondary outcomes** | | | | |
| All-cause in-hospital mortality (n) | 24 (1.70%) | 41 (2.74%) | ARR 0.65 (0.32, 0.98) | * |
| Hospital length of stay (days) | 9.71 (4.83, 17.71) | 10.46 (5.61, 18.15) | ARD-0.73 (-1.56, 0.11) | 0.089 |
| Total ICU length of stay (days) | 4.70 (2.64, 9.81) | 5.81 (3.93, 10.52) | ARD-0.93 (-2.48, 0.61) | 0.235 |
| Time to UIT after the first alert (days) | 0.73 (0.26, 2.48) | 1.82 (0.51, 6.95) | ARD -1.09 (-1.90, -0.28) | ** |
| Cerebral Performance Category | 4.00 ± 0.96 | 4.45 ± 0.99 | ARD -0.45 (-1.09, 0.19) | 0.168 |

**Note:** Data are presented as mean ± standard deviation, median (interquartile range), or number (percentage). * denotes p-value<0.05. ** denotes p-value <0.01.

ICU-Intensive Care Unit; UIT-Unplanned Intensive Care Unit Transfer; ARR-Adjusted Risk Ratio; ARD-Adjusted Risk Difference; CI-Confidence Interval.

## Redefining Explainability: Why SHAP is Insufficient

The "black box" nature of deep learning is often cited as a barrier to trust. To mitigate this, many systems use technical interpretability tools like SHAP. However, SHAP is increasingly criticized for its lack of "clinical actionability" **[7]**.

### Technical *vs.* Medical Explainability

SHAP provides a mathematical attribution of feature importance, but it does not represent physiological causality. A clinician cannot "treat" a high SHAP value associated with a non-modifiable factor like age. Furthermore, a recent study in npj Digital Medicine (2025) found that providing SHAP plots alone resulted in a lower "weight of advice" (clinician acceptance) compared with clinical explanations. Clinicians were found to prioritize "medical explainability" terminology familiar to frontline providers over abstract force plots **[7]**.

### Actionable insights: The case for targeted intervention

A clinically meaningful form of explainability does not focus on why a model generates a score, but on what action should follow when the score is high. In this trial, high-risk alerts were mapped to common deterioration patterns, classified by diagnostic causes recognizable to HCPs, and linked to corresponding interventions such as fluid resuscitation, antibiotic adjustment, intubation, or timely ICU transfer **(Table 2B)**. By translating a 0-100 score into a clinical phenotype, the system moved beyond "explaining the model" to "guiding the treatment," a shift that may be mandatory for real-world adoption and improved patient outcomes **[6]**.

**Table 2.** Clinical explainability analysis.

| Intervention reasons | General ward cardiac arrest (%) | | All-cause in-hospital mortality (%) | |
|---|---|---|---|---|
| | AI-SaMD-guided cohort | Usual care cohort | AI-SaMD-guided cohort | Usual care cohort |
| Respiratory | 2.53 | 3.23 | 3.8 | 7.53 |
| Sepsis | 5.88 | 7.5 | 9.8 | 10 |
| Shock w/o sepsis | 3.7 | 5.88 | 3.7 | 11.76 |
| Metabolic | 7.14 | 0 | 7.14 | 0 |
| Others | 0 | 0 | 0 | 0 |

**Table 2A:** Patient outcomes based on intervention reasons.

**Note:** Patients without documented intervention reasons were excluded from the subgroup analysis.

| Intervention reasons | Types of intervention | | | | |
|---|---|---|---|---|---|
| | Counsel for treatment plan (%) | ABGA/Image (%) | Oxygen/Airway (%) | IV/fluid (%) | UIT/DNR (%) |
| Respiratory | 97.47 | 16.46 | 26.58 | 15.19 | 10.13 |
| Sepsis | 100 | 27.45 | 9.8 | 43.14 | 35.29 |
| Shock w/o sepsis | 100 | 22.22 | 7.41 | 33.33 | 7.41 |
| Metabolic | 92.86 | 21.43 | 7.14 | 35.71 | 14.29 |
| Others | 100 | 16.67 | 16.67 | 8.33 | 0 |

**Table 2B:** Frequency of intervention types based on intervention reasons for the AI-SaMD guided cohort.

**Note:** Multiple types of interventions could be implemented for a single patient.

## Implementation Science: The Human-in-the-Loop

The effectiveness of AI is contingent upon its integration into the clinical workflow. In this non-randomized controlled trial, timing and compliance were critical drivers of success. Interventions initiated within 4 hours of the first alert were associated with significantly fewer adverse outcomes than those delayed by 20-24 hours. Furthermore, maintaining high compliance (>90%) with alerts was associated with a 2- to 4-fold lower incidence of cardiac arrest or death. This highlights that AI is not a standalone solution but a force multiplier for the existing Rapid Response Team (RRS) **[6]**.

## Considerations and Future Directions

Still, even with the promising results of this study, interpretation of this prospective nonrandomized trial should be tempered by four limits to portability across real-world settings **[2,5]**. First, single-center nonrandomized allocation leaves residual confounding; multicenter randomized confirmation is required, and the authors cite a stepped-wedge cluster RCT (KCT0010243) **[6]**. Second, even with matched inputs and endpoints, performance is not transferable across AI-SaMDs: Head-to-head evaluations show wide variability in diabetic retinopathy and tuberculosis CXR CAD, and thresholds are not interchangeable across products or subgroups **[8,9]**. Third, dataset shift can reduce discrimination, supporting local external validation and prospective "*in-vivo*" evaluation **[5,10]** Clinical effect also depends on human-AI interaction; response time and alert compliance vary, so gains must be replicated with clinician-in-the-loop implementation studies that measure outcomes, not only alert metrics **[5,6]**. Finally, the lack of formal health-economic evaluation limits policy readiness; cost-effectiveness and budget-impact analyses should accompany future deployments to guide procurement and pricing **[11]**.

## Conclusion

For AI-based surveillance to become the standard of care, three conditions must be met: Prospective validation in real-world settings, demonstration of patient-centered outcomes (mortality/morbidity), and provision of actionable medical explainability. This study represents a large-scale prospective trial with a comparative arm evaluating a regulatory-approved AI-SaMD operating as a general ward early warning system in an external environment, providing meaningful real-world evidence **[6]**. Collectively, these findings suggest that AI can shift hospital care from reactive "cardiopulmonary resuscitation" to proactive "cardiopulmonary stabilization." Future research should prioritize bedside utility and rigorous causal validation over algorithmic novelty.

## References

1) Merchant RM, Topjian AA, Panchal AR, Cheng A, Aziz K, et al. Adult basic and advanced life support, pediatric basic and advanced life support, neonatal life support, resuscitation education science, and systems of care writing groups. Part 1: Executive summary: 2020 American heart association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care. Circ. 2020, 142:S337-357.
2) Khera R, Butte AJ, Berkwits M, Hswen Y, Flanagin A, et al. AI in medicine JAMA's focus on clinical outcomes, patient-centered care, quality, and equity. JAMA. 2023, 330:818-820.
3) El Fassi SC, Abdullah A, Fang Y, Natarajan S, Masroor AB, et al. Not all AI health tools with regulatory authorization are clinically validated. Nature Med. 2024, 30:2718-2720.
4) Han R, Acosta JN, Shakeri Z, Ioannidis JP, Topol EJ, et al. Randomized controlled trials evaluating artificial intelligence in clinical practice: A scoping review. Lancet Digit Health. 2024, 6:e367-73.
5) Jain SS, Goto S, Hall JL, Khan SS, MacRae CA, et al. Pragmatic approaches to the evaluation and monitoring of artificial intelligence in health

care: A science advisory from the american heart association. Circ. 2025, 152:23.
6) Park MH, Kim M, Lee MJ, Kim AJ, Cho KJ, et al. Clinical effectiveness of an artificial intelligence-based prediction model for cardiac arrest in general ward-admitted patients: A non-randomized controlled trial. Diag. 2026, 16:335.
7) Hur S, Lee Y, Park J, Jeon YJ, Cho JH, et al. Comparison of SHAP and clinician friendly explanations reveals effects on clinical decision behavior. NPJ Digit Med. 2025, 8:578.
8) Lee AY, Yanagihara RT, Lee CS, Blazes M, Jung HC, et al. Multicenter, head-to-head, real-world validation study of seven automated artificial intelligence diabetic retinopathy screening systems. Diab care. 2021, 44:1168-1175.
9) Qin ZZ, Van der Walt M, Moyo S, Ismail F, Maribe P, et al. Computer-aided detection of tuberculosis from chest radiographs in a tuberculosis prevalence survey in South Africa: External validation and modelled impacts of commercially available artificial intelligence software. The Lancet Digit Health. 2024, 6:e605-613.
10) Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, et al. How medical Ai devices are evaluated: Limitations and recommendations from an analysis of FDA approvals. Nature Med. 2021, 27:582-584.
11) El Arab RA, Al Moosa OA. Systematic review of cost effectiveness and budget impact of artificial intelligence in healthcare. NPJ Digit Med. 2025, 8:548.