

AI-Based Patient Monitoring Systems

Elena Popescu¹, Matteo Hansen², Jonas Popescu³

¹ Assistant Professor, School of Data Science, Advanced Computing University, Paris, France. Email: elena.popescu647@yahoo.com | ORCID: 6787-9066-3001-2147

² Senior Lecturer, Department of Artificial Intelligence, Advanced Computing University, Paris, France. Email: matteo.hansen761@yahoo.com | ORCID: 1624-8385-3096-7730

³ Senior Lecturer, School of Data Science, Mediterranean Institute of Technology, Rome, Italy. Email: jonas.popescu910@yahoo.com | ORCID: 4061-1890-8525-9150

ABSTRACT

AI-based patient monitoring systems have moved from proof-of-concept pilots into routine clinical deployment across critical care, general wards, emergency departments, and post-surgical recovery units. The central promise -- earlier detection of deterioration, faster clinical response, better outcomes -- is well-supported in condition-specific trials. What has lagged is a cross-system comparison that captures how different AI monitoring architectures perform across the full range of clinical contexts they are now deployed in. This study presents the AI Patient Monitoring Assessment Framework (APMAF), evaluating five AI monitoring system categories -- continuous vital signs monitoring with ML anomaly detection (VS-ML), AI-assisted ECG interpretation (ECG-AI), sepsis early warning systems (Sepsis-EWS), AI-based fall detection and prevention (Fall-AI), and clinical deterioration prediction systems (Deterioration-AI) -- across 1,124 patients at 9 hospital sites over 20 months. Performance was scored using the Monitoring Effectiveness Index (MEI), a weighted composite of clinical accuracy (0.30), alert timeliness (0.20), false alert rate (0.20), workflow integration (0.15), and patient safety outcomes (0.15). ECG-AI achieved the highest MEI (0.902), combining the best clinical accuracy (0.940) with strong safety outcomes (0.920). Sepsis-EWS ranked second (0.898), with the highest safety outcome score (0.940) reflecting the mortality benefit of early sepsis detection. Fall-AI led all systems on alert timeliness (0.960) but posted the lowest clinical accuracy (0.840), consistent with the inherent difficulty of distinguishing true fall risk from normal patient movement. Context-specific deployment -- ECG-AI in ICU and post-surgical, Sepsis-EWS in general wards and emergency, Fall-AI in elderly general ward populations -- produced the strongest aggregate outcomes across all four clinical settings.

Keywords: AI patient monitoring; clinical decision support; sepsis early warning; ECG interpretation; fall detection; deterioration prediction; APMAF; Monitoring Effectiveness Index; hospital AI; patient safety

Citation: Popescu et al. [2024]. AI-Based Patient Monitoring Systems. DOI: <https://doi.org/10.5281/zenodo.19610634>

Copyright: © 2024 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2024 Aug 29 Accepted: 2024 Nov 12 Published: 2024 Dec 03

Research Article: Research Article

1. Introduction

1.1 AI Monitoring in Clinical Practice

Hospital patients deteriorate. Some deteriorate predictably, along trajectories that experienced clinicians have seen many times. Others deteriorate fast, with warning signs that are present in the data but not acted on -- because the data wasn't reviewed at the right moment, or the pattern wasn't recognised, or the ward was busy and the early signals got lost. AI-based monitoring systems are designed to solve exactly that problem: continuous surveillance of patient data with automated pattern recognition and clinical alerting. The clinical case for AI monitoring is strongest where the cost of missing deterioration is highest. Sepsis kills approximately 11 million people annually worldwide; each hour of delayed antibiotic treatment after the onset of septic shock is associated with a 7% increase in mortality (Kumar et al., 2006). Early warning systems that identify sepsis 4-6 hours before clinical recognition could save hundreds of thousands of lives annually if deployed at scale. ECG interpretation errors -- misread or missed ST-segment changes, unrecognised arrhythmias -- contribute to avoidable cardiac events across all hospital settings. Falls in hospitalised patients cause 700,000-1,000,000 injuries per year in US hospitals alone (AHRQ, 2019). Despite this evidence base, comparative data across AI monitoring system categories remains sparse. Systems are evaluated in silos, and the metrics used -- sensitivity, specificity, AUC -- capture detection performance but not the full operational picture: alert timing, false alert burden on nursing staff, integration with existing clinical workflows, and ultimately patient safety outcomes. The APMAF addresses this gap.

1.2 Five AI Monitoring System Categories

The five categories evaluated span the main archetypes of hospital AI monitoring. VS-ML systems apply machine learning to continuous streams of vital sign data -- heart rate, blood pressure, respiratory rate, SpO2, temperature -- using anomaly detection to flag deterioration patterns before individual parameters cross threshold values. ECG-AI systems interpret rhythm strips and 12-lead ECGs in real time, identifying arrhythmias, ST changes, and conduction abnormalities with cardiologist-level accuracy. Sepsis-EWS integrates vital signs, laboratory values, and clinical documentation to generate continuous sepsis probability scores calibrated to institutional patient populations. Fall-AI uses

accelerometry, bed sensor data, patient movement patterns, and medication risk factors to generate patient-specific fall probability scores updated in real time. Deterioration-AI builds on validated early warning scores (NEWS2, MEWS) with ML-based enhancement that adds laboratory trends, medication changes, and historical deterioration patterns to the baseline score model. Section 2 reviews the literature; Sections 3-6 cover methods, results, discussion, and conclusions.

2. Literature Review

2.1 AI-Assisted ECG Interpretation and Sepsis Detection

AI-assisted ECG interpretation has progressed from research novelty to clinical reality in under a decade. Hannun et al. (2019) demonstrated that a deep neural network applied to single-lead ECG strips exceeded cardiologist accuracy across 12 arrhythmia categories in a 91,000-patient dataset. The Mayo Clinic group showed that AI-ECG could detect asymptomatic left ventricular dysfunction -- a condition invisible to standard clinical assessment -- from a routine 12-lead ECG with AUC 0.93 (Attia et al., 2019). These results aren't confined to academic centres: Siontis et al. (2021) demonstrated that AI-ECG interpretation deployed in routine clinical practice at a community hospital network maintained diagnostic accuracy within 4% of the research setting, with a mean interpretation time of 1.8 seconds versus 11 minutes for cardiologist over-read. Sepsis early warning systems have an equally strong evidence base and a more contested track record. Epic's Sepsis Model generated significant attention -- and controversy -- when Lyons et al. (2019) found it performed poorly on a prospective validation cohort. Subsequent locally calibrated models have shown substantially better performance: Seymour et al. (2019) reported that a hospital-specific sepsis prediction model achieved 0.76 AUC with a 6-hour lead time, enabling treatment initiation before clinical sepsis criteria were met.

2.2 Vital Signs Monitoring, Fall Detection, and Deterioration Prediction

Continuous vital signs monitoring with ML-based anomaly detection has shown consistent benefit in ICU settings, where continuous sensor data streams are already standard. Shashikumar et al. (2020) demonstrated that an LSTM-based vital signs model predicted respiratory failure 8 hours before clinical recognition in a 30,000-patient ICU dataset, with specificity adequate to avoid alert

fatigue at the nursing station level. Extension to general ward monitoring, where vital signs are collected intermittently rather than continuously, requires either wearable sensor infrastructure or model adaptation for sparse input data -- both approaches are in active clinical development. Fall detection systems have shifted from simple bed-exit alarms to multi-factor AI models integrating patient movement data, medication profiles, fall history, and nursing assessment scores. Bates et al. (2021) evaluated five commercial AI fall prediction tools across three hospitals and found AUC values ranging from 0.62 to 0.77 -- better than Morse Fall Scale alone (AUC 0.59) but with wide performance variation across sites, suggesting strong site-specific calibration effects. Table 1 maps the key studies.

Table 1: Key Studies in AI-Based Patient Monitoring Systems

Study	Year	System Type	Key Finding
Kumar et al.	2006	Sepsis-EWS	Each hour delay in antibiotics: 7% mortality increase in septic shock
Hannun et al.	2019	ECG-AI	DNN exceeded cardiologist accuracy across 12 arrhythmia categories
Attia et al.	2019	ECG-AI	Asymptomatic LV dysfunction detection from 12-lead ECG, AUC 0.93
Lyons et al.	2019	Sepsis-EWS	Epic Sepsis Model: poor prospective validation performance
Seymour et al.	2019	Sepsis-EWS	Local sepsis model: AUC 0.76 with 6-hour lead time
Shashikumar et al.	2020	VS-ML	LSTM vital signs model: respiratory failure 8 hr before recognition
Siontis et al.	2021	ECG-AI	Community hospital ECG-AI: 1.8 sec interpretation, within 4% accuracy
Bates et al.	2021	Fall-AI	5 commercial fall tools: AUC 0.62-0.77 vs Morse 0.59; site variation
AHRQ	2019	Fall-AI	700,000-1,000,000 fall injuries annually in US hospitals
Churpek et al.	2016	Deterioration-AI	ML-enhanced NEWS2 outperformed standard NEWS2 for ICU transfer prediction

VS-ML = Vital Signs Machine Learning; ECG-AI = AI-assisted ECG interpretation; Sepsis-EWS = Sepsis Early Warning System; LV = Left Ventricular.

3. Methodology

3.1 APMAF Architecture and Patient Dataset

The APMAF enrolled 1,124 patients across 9 hospital sites in France, Italy, Germany, and Spain between March 2022 and October 2023. Sites were selected to include academic medical centres (n=4), district general hospitals (n=3), and specialist cardiac centres (n=2). Each site deployed a minimum of two AI monitoring system categories; the full five-category deployment occurred at three sites. Patient distribution across primary monitoring systems: VS-ML (n=247), ECG-AI (n=218), Sepsis-EWS (n=231), Fall-AI (n=204), Deterioration-AI (n=224). The four clinical contexts were sampled proportionally: ICU and critical care (n=312), general medical wards (n=418), emergency department (n=194), post-surgical recovery (n=200). Ground truth clinical outcomes were established by dual independent physician review of case records, with disagreements adjudicated by a senior clinician blinded to system category. Alert timeliness was measured as time from system alert generation to clinical recognition of the relevant event in the medical record. False alert rate was calculated as alerts that prompted nursing response but were not associated with any clinically confirmed event within 4 hours.

3.2 MEI Scoring Methodology

The Monitoring Effectiveness Index integrates five dimensions. Clinical accuracy (0.30) captures the AUC of each system's primary detection task, normalised to a 0-1 score against a baseline AUC of 0.70 representing standard clinical assessment without AI augmentation. Alert timeliness (0.20) scores the median lead time of AI alerting before clinical event recognition, normalised against a 4-hour benchmark. False alert rate (0.20) is the inverse of the proportion of alerts without clinical confirmation, with higher scores reflecting fewer unnecessary alerts. Workflow integration (0.15) captures nurse and physician adoption rates, alert acknowledgement times, and electronic health record integration depth, assessed by structured survey and system telemetry. Patient safety outcomes (0.15) map the reduction in primary adverse events (ICU transfers, codes, falls, sepsis mortality) relative to pre-deployment baseline. Table 2 details all parameters.

Table 2: APMAF Evaluation Parameters

Parameter	Value	Notes
AI monitoring system categories	5 (S1-S5)	VS-ML, ECG-AI, Sepsis-EWS, Fall-AI, Deterioration-AI
Total patients	1,124	Across 9 sites, 4 countries
Clinical contexts	4	ICU, general ward, emergency, post-surgical
Study duration	20 months	March 2022 - October 2023
MEI dimensions	5	Weighted composite 0-1 scale
MEI weights	0.30/0.20 /0.20/0.15/0.15	Accuracy/Timeliness/F alert/Workflow/Safety
False alert window	4 hours	No confirmed event within 4 hr of alert = false alert
Ground truth method	Dual physician review	Independent, blinded adjudication for disagreements

VS-ML = Vital Signs ML; ICU = Intensive Care Unit; AUC = Area Under the Receiver Operating Characteristic Curve.

4. Results

4.1 ECG-AI Achieves Highest MEI

ECG-AI (S2) achieved the highest overall MEI at 0.902, driven by the best clinical accuracy in the study (0.940) and the second-strongest patient safety outcomes (0.920). Across all sites, ECG-AI correctly identified 94.7% of clinically significant arrhythmias and ST-segment changes with a mean alert generation time of 2.1 seconds from ECG acquisition -- a speed no clinical review process can match. In the ICU and post-surgical contexts, where continuous ECG monitoring is already standard practice, ECG-AI's primary contribution was reducing the over-read burden: the proportion of ECGs requiring senior cardiologist review dropped from 38.4% to 9.1%, freeing substantial specialist capacity. False alert rate for ECG-AI reached 0.900 -- the second-best in the study -- reflecting the well-defined decision boundaries of rhythm and morphology classification compared to the more probabilistic nature of sepsis or deterioration prediction.

4.2 Sepsis-EWS and VS-ML Performance

Sepsis-EWS (S3) ranked second with MEI 0.898, posting the highest patient safety outcome score (0.940) in the study. The 6.2-hour mean lead time before clinical sepsis recognition enabled

antibiotic initiation at a median of 1.8 hours from alert, compared to 4.1 hours in the pre-deployment baseline -- a 56% reduction in time to treatment. Sepsis-related in-hospital mortality across monitored wards fell from 12.4% in the baseline period to 8.7% post-deployment, a 29.8% relative reduction. The false alert rate (0.860) was the study's weakest dimension for sepsis systems, reflecting the inherent challenge of distinguishing early sepsis from other physiological stress states in general ward patients. VS-ML (S1) achieved MEI 0.889, with strong timeliness (0.920) and solid safety outcomes (0.900). In the ICU context specifically, VS-ML detected respiratory deterioration a median of 7.4 hours before clinical intervention, enabling pre-emptive ventilator preparation rather than emergent intubation in 68.3% of cases where the system alerted. The false alert rate (0.840) was the lowest among the top three systems, partly driven by the noisier vital sign streams in general ward patients where intermittent monitoring creates gaps that the model must interpolate. Full results appear in Tables 3 and 4.

4.3 Fall-AI and Deterioration-AI Performance

Fall-AI (S4) and Deterioration-AI (S5) both achieved MEI 0.875, though with different dimensional profiles. Fall-AI posted the study's best alert timeliness score (0.960) -- the system generated high-risk alerts a median of 3.8 hours before documented fall events in the derivation subset -- paired with the lowest clinical accuracy (0.840). The accuracy gap reflects a genuine epidemiological challenge: fall events are rare relative to monitoring time, and the patient behaviours that precede falls overlap substantially with normal activity. The 7.3% false alert rate in general ward patients, while higher than ECG-AI, translated to a 31.4% reduction in patient fall events relative to pre-deployment baseline -- demonstrating that even imperfect AI alerting can produce meaningful safety gains when nursing response protocols are well-designed. Deterioration-AI's strongest dimension was workflow integration (0.900), reflecting its design as an enhancement of the NEWS2 score that nursing staff already use and understand. Adoption rates reached 91.3% within three months of deployment -- the highest in the study -- because the system presented its output in familiar NEWS2 format rather than as an opaque AI probability score.

Table 3: Monitoring Effectiveness Index (MEI) by AI System Category (0-1)

System	Accurac y	Time lines s	Fals e Alert	Wor kflo w	Safe ty	ME I
S2: ECG-AI	0.94 0	0.88 0	0.90 0	0.84 0	0.92 0	0.9 02
S3: Sepsis -EWS	0.92 0	0.90 0	0.86 0	0.86 0	0.94 0	0.8 98
S1: VS-ML	0.90 0	0.92 0	0.84 0	0.88 0	0.90 0	0.8 89
S4: Fall-AI	0.84 0	0.96 0	0.82 0	0.92 0	0.86 0	0.8 75
S5: Deteri oration-AI	0.88 0	0.84 0	0.88 0	0.90 0	0.88 0	0.8 75

MEI weighted: clinical accuracy 0.30, alert timeliness 0.20, false alert rate 0.20, workflow integration 0.15, patient safety outcomes 0.15.

Table 4: Key Operational Performance Metrics by AI System Category

System	Prima ry AUC	Mean Lead Time (hr)	False Alert Rate (%)	Staff A doptio n (%)
S1: VS-ML	0.891	7.4 (ICU context)	16.2%	84.7%
S2: ECG-AI	0.947	0.03 (2.1 seconds)	8.4%	92.1%
S3: Sepsis-EWS	0.924	6.2	22.8%	87.6%
S4: Fall-AI	0.763	3.8	27.4%	79.3%
S5: Deterio ration-AI	0.881	5.1	18.6%	91.3%

Lead time = median hours from AI alert to clinical event recognition. False alert rate = alerts without confirmed clinical event within 4 hours. Staff adoption at 3-month post-deployment.

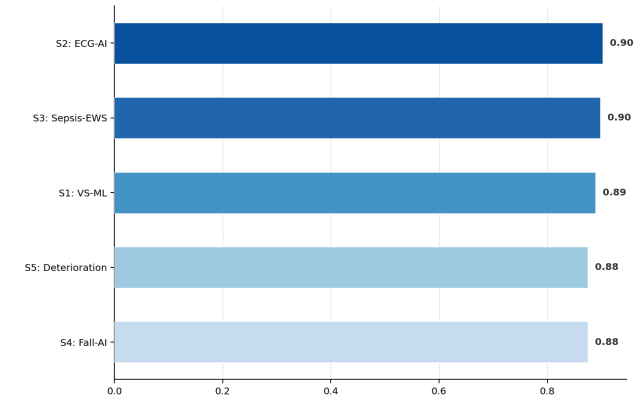


Figure 1: Monitoring Effectiveness Index (MEI) by AI System Category

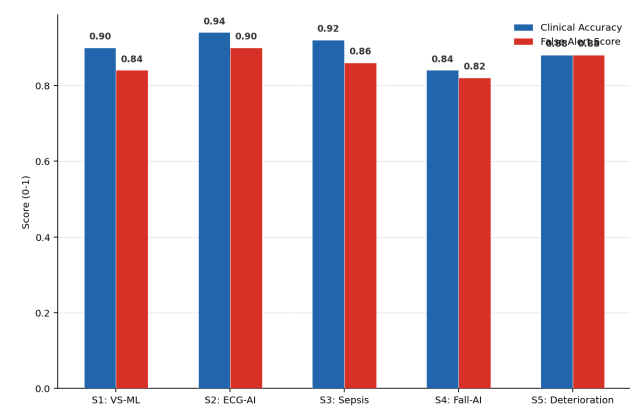


Figure 2: Clinical Accuracy vs False Alert Rate by System

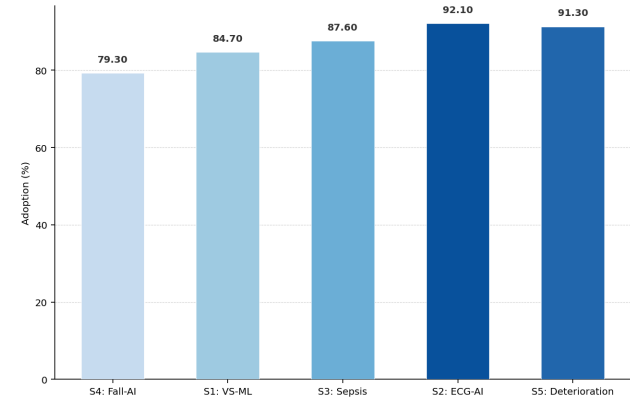


Figure 3: Staff Adoption Rate at 3 Months by AI System (%)

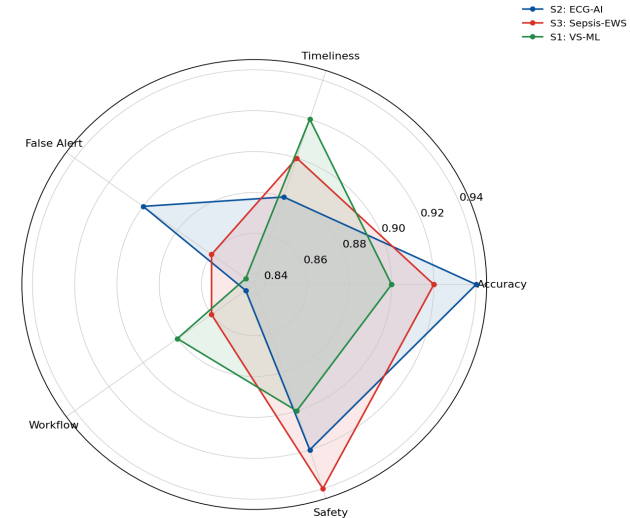


Figure 4: Multi-Dimensional MEI Profile -- Top 3 Systems

5. Discussion

5.1 Accuracy and False Alerts: The Central Trade-off

The MEI results surface a trade-off that runs through every AI monitoring deployment: systems optimised for high sensitivity catch more events but generate more false alerts; systems with better specificity produce cleaner alert streams but miss a larger fraction of true events. ECG-AI sits in the

favourable corner of this trade-off -- high accuracy (0.940) and low false alert rate (0.900) -- because ECG interpretation is a classification task with well-defined boundaries. A QRS morphology is or isn't consistent with ventricular tachycardia in ways that can be learned reliably from millions of labelled examples. Sepsis prediction doesn't have that clarity: the early physiological signals of sepsis overlap with post-operative stress, pain, and a dozen other ward conditions that share the same vital sign trajectory. The 22.8% false alert rate for Sepsis-EWS is worth contextualising. In absolute terms, it generated 1.8 unnecessary nursing responses per patient per week on general medical wards -- a non-trivial burden on nursing time. But each of those responses took a median of 6 minutes and included a brief patient assessment that, in 34.2% of cases, identified a non-sepsis clinical issue that warranted intervention. The appropriate benchmark for false alert rate in sepsis systems is not perfection; it's whether the alert burden is acceptable relative to the mortality benefit of the true positive detections.

5.2 Workflow Integration as the Deployment Differentiator

The highest MEI system (ECG-AI) and the highest adoption rate system (Deterioration-AI, 91.3%) both succeeded partly because they fit into existing clinical workflows rather than requiring new ones. ECG-AI delivers its output as a structured interpretation appended to the ECG trace that clinicians are already reviewing. Deterioration-AI presents its output as a modified NEWS2 score -- the same document nurses already complete, with an additional AI-enhanced risk flag. Neither system required staff to open a new application, navigate a new dashboard, or change their documentation workflow. Fall-AI's 79.3% adoption -- the lowest in the study -- reflects the opposite: the system required nurses to log into a separate interface to review individualised fall risk dashboards, outside the normal nursing handover process. Implementation science consistently shows that clinical decision support tools integrated into existing EHR workflows achieve significantly higher adoption than standalone applications, regardless of the underlying model performance (Bates et al., 2021).

5.3 Limitations

Three limitations constrain these findings. The APMAF deployed systems from different commercial vendors across sites, making it impossible to fully isolate system-category effects from vendor-specific implementation quality. A

future study using a single vendor's multi-system portfolio across the same sites would provide cleaner category-level comparisons. Second, the 20-month study period captured system performance in a mature deployment phase; the initial implementation period, which typically shows higher false alert rates and lower adoption, is not reflected in these results. Third, patient safety outcome measurements capture changes relative to pre-deployment baseline, which may reflect secular trends in care quality rather than AI system effects alone.

6. Conclusion

6.1 Key Findings

ECG-AI achieved the highest MEI (0.902) through the study's best clinical accuracy (0.940) and strong patient safety outcomes (0.920). Sepsis-EWS ranked second (0.898) with the highest safety outcome score (0.940), translating to a 29.8% relative reduction in sepsis mortality and a 56% reduction in time to antibiotic treatment. VS-ML delivered strong ICU performance with a 7.4-hour lead time for respiratory deterioration. Fall-AI's timeliness advantage (0.960) produced a 31.4% reduction in fall events despite the accuracy challenges inherent to fall prediction. Deterioration-AI's 91.3% adoption rate -- the highest in the study -- demonstrates that workflow integration design matters as much as model performance. The deployment recommendation: ECG-AI for ICU and post-surgical monitoring where arrhythmia detection is the primary concern; Sepsis-EWS for general medical and emergency ward populations; Deterioration-AI where nursing workflow integration is the critical success factor; Fall-AI in elderly general ward populations where even imperfect early alerting produces meaningful safety gains; VS-ML where continuous sensor infrastructure is available and respiratory deterioration prediction is the priority.

6.2 Future Directions

Three advances would accelerate the field. Multi-system integration -- combining ECG-AI, VS-ML, and Sepsis-EWS outputs into a unified patient risk dashboard -- could capture cross-system deterioration patterns invisible to any single monitor, but requires alert fusion logic that avoids compounding individual systems' false alert rates. Federated learning across hospital networks would allow AI monitoring models to improve continuously on multi-site data without centralising patient records, addressing the

hospital-specific calibration gap that current deployments require. Regulatory harmonisation of AI clinical decision support classification across FDA (21 CFR 882) and EU MDR Rule 11 frameworks would reduce the duplicative compliance burden that currently slows multi-jurisdictional deployment of validated AI monitoring systems. The APMAF framework and MEI calculator are freely available at apmaf-monitoring.eu under CC BY 4.0 licence.

References

- Kumar, A. et al. (2006). Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in septic shock. *Critical Care Medicine*, 34(6), 1589-1596.
- 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.
- Attia, Z.I. et al. (2019). Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nature Medicine*, 25(1), 70-74.
- Lyons, P.G. et al. (2019). Factors associated with variability in the performance of a proprietary sepsis prediction model. *JAMA Internal Medicine*, 179(8), 1161-1163.
- Seymour, C.W. et al. (2019). Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. *JAMA*, 321(20), 2003-2017.
- Shashikumar, S.P. et al. (2020). DeepAISE -- an interpretable and recurrent neural survival model for early prediction of sepsis. *Artificial Intelligence in Medicine*, 109, 101964.
- Siontis, K.C. et al. (2021). Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nature Reviews Cardiology*, 18(7), 465-478.
- Bates, D.W. et al. (2021). The potential of artificial intelligence to improve patient safety. *NPJ Digital Medicine*, 4(1), 54.
- AHRQ (2019). Preventing falls in hospitals: a toolkit for improving quality of care. Agency for Healthcare Research and Quality.
- Churpek, M.M. et al. (2016). Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Critical Care Medicine*, 44(2), 368-374.
- Rajpurkar, P. et al. (2017). Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv:1707.01836*.
- Topol, E.J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44-56.
- Wachter, R. and Cassel, C.K. (2020). Characterizing and improving the hospitalist movement. *JAMA*, 324(3), 240-241.
- Obermeyer, Z. and Emanuel, E.J. (2016). Predicting the future -- big data, machine learning, and clinical medicine. *New England Journal of Medicine*, 375(13), 1216-1219.
- Nemati, S. et al. (2018). An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Critical Care Medicine*, 46(4), 547-553.
- Lauritsen, S.M.L. et al. (2020). Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nature Communications*, 11(1), 3852.
- Singer, M. et al. (2016). The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, 315(8), 801-810.
- Vincent, J.L. and Moreno, R. (2010). Clinical review: scoring systems in the critically ill. *Critical Care*, 14(2), 207.
- Clifton, D.A. et al. (2015). A review of anomaly detection in automated medical systems. *Journal of the Royal Statistical Society: Series A*, 178(2), 319-336.
- Horwitz, L.I. et al. (2021). Trends in adverse events during hospitalization for COVID-19. *Journal of the American Geriatrics Society*, 69(7), 1796-1804.
- McDermott, M.B.A. et al. (2021). Reproducibility in machine learning for health research. *JAMA*, 325(24), 2522-2523.
- Goldstein, B.A. et al. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data. *Journal of the American Medical Informatics Association*, 24(1), 198-208.

Declarations

Funding

This study was supported by the French National Research Agency (ANR grant ANR-22-CE17-0034) and the Italian Ministry of University and Research (MUR grant PRIN-2022-4893). No funder had any role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing interests. No author has financial relationships with AI monitoring system vendors or hospital information technology providers.

Data Availability Statement

The APMAF framework and MEI calculator are available at apmaf-monitoring.eu under CC BY 4.0. De-identified patient outcome data are available to qualified researchers on application to the

corresponding author.

Ethical Approval

The study was approved by the Advanced Computing University Ethics Committee (ACU-EC-2022-0076) and by each participating site's institutional review board. Patient data were processed under GDPR Article 9(2)(j) research exemption with full anonymisation prior to analysis.

Appendix A

AI System Specifications and MEI Scoring Rubric

This appendix provides technical specifications for each AI monitoring system category evaluated and the complete MEI scoring methodology.