

Volume 1 Issue 1
Article Number: JHAI20260001

Explainable AI for Electronic Health Records and Clinical Time-Series Prediction: A Systematic Review

Deepak Singh ¹ and Smita Nirkhi*²

¹Principal Solution Architect, Gainwell Technologies, Texas, United States

²Symbiosis Institute of Technology, Nagpur, Symbiosis International (Deemed University), Nagpur, Maharashtra 441248, India

Abstract

Machine learning models trained on electronic health records (EHRs) and clinical time-series data now predict sepsis onset, in-hospital mortality, and readmission risk with substantial accuracy, but their high dimensionality and temporal complexity make them especially opaque to the clinicians who must act on their output. We conducted a systematic review, following PRISMA guidelines, of explainable AI (XAI) methods applied specifically to EHR and clinical time-series prediction. From 312 records identified across five databases, 30 studies met inclusion criteria and were organized into three categories: post-hoc feature-attribution methods (chiefly SHAP and LIME) applied to gradient-boosted or deep models; ante-hoc, intrinsically interpretable architectures, including attention-based recurrent networks and additive models; and studies evaluating trust, fairness, or clinical validity of these explanations. Post-hoc attribution dominates the literature and reliably identifies established clinical risk factors, but several studies show that attention weights and feature-importance scores can track spurious correlates of clinical workflow, such as medication-administration timing, rather than the underlying physiology. Ante-hoc and constraint-based models offer a promising, less-explored alternative, and fairness-auditing studies demonstrate that interpretability tools can expose demographic disparities in model behavior. Across categories, however, evidence for a genuine causal link between explanation and improved clinical decision-making remains sparse. We argue that EHR-focused XAI research should shift emphasis from producing more post-hoc explanations toward validating their clinical utility, expanding ante-hoc model adoption, and standardizing evaluation of explanation quality and fairness.

Keywords: Explainable Artificial Intelligence (XAI); Electronic Health Records; Clinical Time-Series; Sepsis Prediction; Interpretable Machine Learning

1. Introduction

Machine learning models built on electronic health records (EHRs) now predict sepsis onset, in-hospital mortality, and readmission risk with levels of accuracy that increasingly rival or exceed traditional clinical scoring systems [1]. Early recurrent-network models for EHR sequences, such as Doctor AI, demonstrated that deep learning could forecast a patient's future diagnoses and medications directly from longitudinal visit histories [2]. Yet EHR data poses a distinct interpretability challenge relative to other clinical AI domains: it is high-dimensional, irregularly sampled, and

*Corresponding Author: Smita Nirkhi (smita811@gmail.com)

Received: 11 June 2026; Revised: 22 July 2026; Accepted: 26 August 2026; Published: 30 September 2026

© 2026 The Author(s).

This is an open access article licensed under a Creative Commons Attribution-NonCommercial 4.0 International Licence (CC BY-NC 4.0).

temporally structured, combining laboratory values, vital signs, medication orders, and free text across a variable-length patient history. This complexity makes EHR-based models especially difficult for clinicians to audit, and especially easy for them to fail silently by learning artifacts of clinical workflow rather than physiology [3].

The consequences of this opacity are not merely academic. A sepsis early-warning system that cannot indicate which vital signs or laboratory trends drove an alert offers little basis for a clinician to accept or override it under time pressure. Regulatory frameworks increasingly expect more: the EU’s GDPR implies a right to explanation for automated decisions [4], the US FDA considers explainability in its evaluation of AI-based clinical software [5], and models trained on retrospective EHR data are known to be vulnerable to learning spurious correlations from confounded or shifted populations [6].

Explainable AI (XAI) methods — most prominently SHAP [7] and LIME [8] — have been applied extensively to EHR-based models, and broader surveys have cataloged XAI techniques across healthcare generally [9]. However, EHR and clinical time-series prediction raise interpretability questions distinct from imaging: an explanation must account not only for *which* variables mattered but *when*, across a sequence of encounters, and must be robust to the fact that clinical documentation itself reflects prior clinical decisions, not only underlying disease state. This review therefore narrows its scope specifically to XAI for EHR and clinical time-series prediction, asking: which explanation methods dominate this domain; do intrinsically interpretable (“ante-hoc”) alternatives to post-hoc explanation see meaningful use; and what evidence exists that these explanations are trustworthy, fair, and clinically useful, rather than merely plausible?

The remainder of the paper is organized as follows. Section 2 describes the review methodology. Section 3 presents publication trends, a taxonomy of the reviewed methods, and a synthesis organized around three method categories: post-hoc feature attribution, ante-hoc and attention-based models, and trust/fairness evaluation. Section 4 discusses cross-cutting findings, limitations, and future directions, and Section 5 concludes.

2. Methodology

This review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [10].

2.1. Search Strategy

We searched PubMed, IEEE Xplore, ACM Digital Library, Web of Science, and Scopus using the Boolean expression (“explainable AI” OR “XAI” OR “interpretable machine learning”) AND (“electronic health records” OR “EHR” OR “clinical time-series” OR “ICU” OR “sepsis prediction” OR “mortality prediction”), adapted to each database’s syntax. The search was conducted in December 2025 and limited to English-language publications, with no restriction on publication year.

2.2. Method Taxonomy

Included studies were categorized into three groups based on how the explanation was generated and evaluated (Figure 1): **post-hoc feature-attribution methods**, which explain a separately trained black-box model, typically via SHAP, LIME, or related feature-importance techniques; **ante-hoc and attention-based models**, which are interpretable by design, including attention-weighted recurrent networks, additive models, and knowledge-distillation approaches; and **trust, fairness, and clinical validation studies**, which empirically evaluate the fairness, consistency, or clinical impact of explanations rather than proposing a new method.

2.3. Inclusion, Exclusion, and Selection Process

Studies were included if they applied, developed, or evaluated an XAI method on EHR or clinical time-series data for a clinical prediction task, and explicitly addressed model explainability rather than only predictive performance. Studies were excluded if their primary data modality was imaging, genomic, or free-text without structured time-series; if explainability was mentioned only as future work; if the study was an editorial or opinion piece without original methodology; or if it duplicated an already-retained population and model.

Selection proceeded in three stages — title/abstract screening, full-text retrieval, and eligibility assessment — each performed independently by two reviewers, with disagreements resolved by discussion.

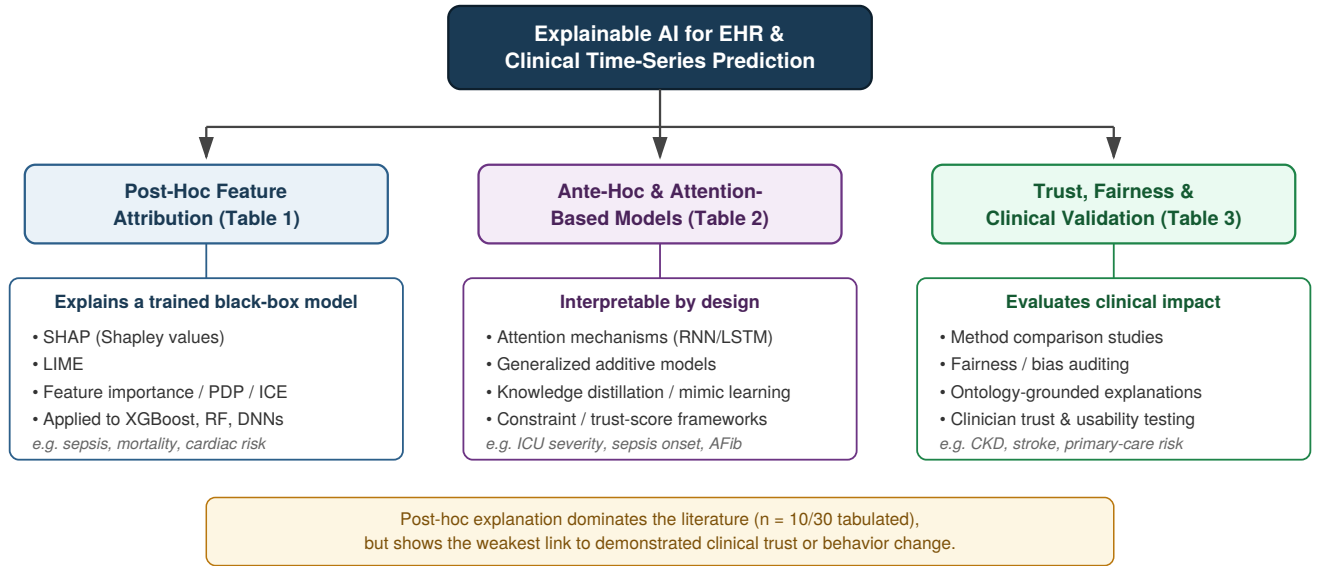


Figure 1: Taxonomy of explainable AI methods for EHR and clinical time-series prediction, organized around the three tables presented in this review.

As summarized in the PRISMA flowchart (Figure 2), 312 records were identified, 96 duplicates removed, and 216 records screened by title and abstract. Of these, 132 were excluded, leaving 84 reports sought for retrieval; 6 could not be retrieved. Of the 78 reports assessed for full-text eligibility, 48 were excluded (19 not focused on EHR or time-series data, 14 lacking a formal XAI evaluation, 9 editorials or opinion pieces, and 6 duplicate populations), leaving **30 studies** included in the review.

Key limitations of the selection process include restriction to English-language studies, inclusion of non-peer-reviewed pre-prints, and a degree of subjectivity in assigning borderline studies to the post-hoc versus ante-hoc categories, since some models combine both strategies.

3. Results

3.1. Publication Trends

Figure 3 shows the distribution of included studies by publication year. Early work appeared steadily from 2017 (2 studies) through 2019 (4 studies), reflecting the initial adoption of attention-based EHR models and early SHAP applications. Volume increased from 2021 onward (5 studies), peaking at 6 studies in 2022, before settling at 3–5 studies annually through 2023–2024; 2025 is undercounted, as the search was conducted in December of that year. This pattern suggests that EHR-focused XAI research matured somewhat earlier than imaging-focused XAI, consistent with SHAP and attention mechanisms becoming standard tooling for structured and sequential clinical data by the early 2020s.

3.2. Post-Hoc Feature-Attribution Methods

Post-hoc attribution, dominated by SHAP, is the most common approach in the reviewed literature (Table 1), typically paired with gradient-boosted tree models. Across several ICU and sepsis mortality-prediction studies, tree ensembles (XGBoost, LightGBM, random forests) consistently achieve strong predictive performance, and SHAP analyses repeatedly surface a recurring set of established risk factors — SOFA score, blood urea nitrogen, age, and vital-sign trends [11, 12]. A comparative evaluation of feature-importance, partial-dependence, and individual-conditional-expectation plots for hypertension prediction found that different techniques surface complementary, rather than identical, insights into model behavior, and cautioned against relying on a single interpretability method [13]. Applied to chronic and acute disease prediction alike — heart disease [14], heart failure mortality [15], breast cancer recurrence [16],

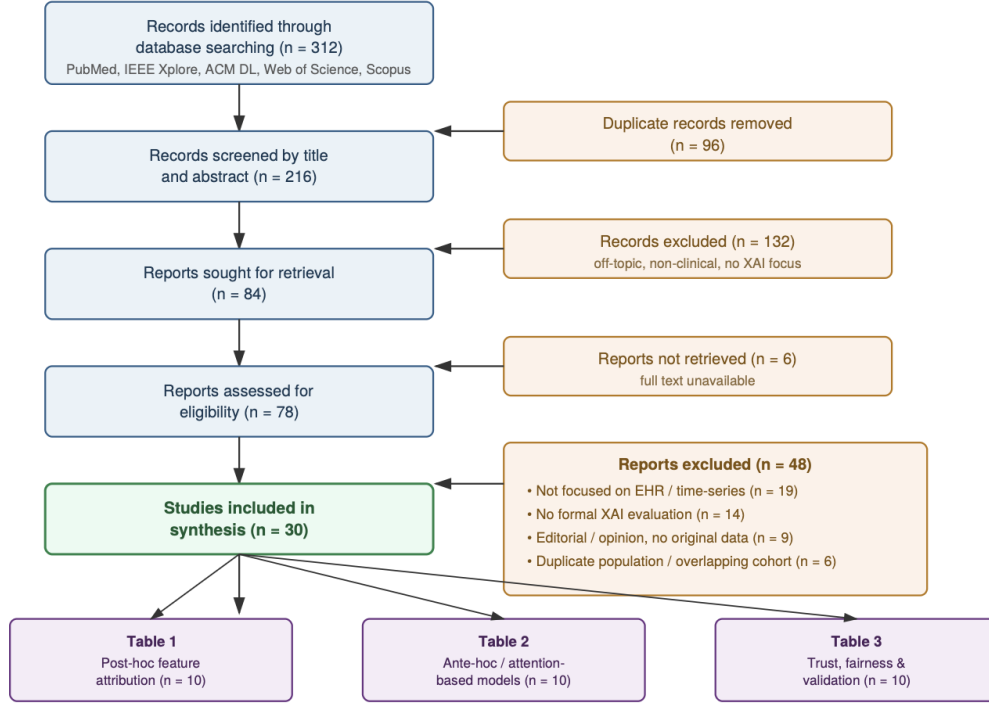


Figure 2: PRISMA flowchart of the study selection process.

COVID-19 prognosis [17], non-communicable disease risk [18], and 30-day readmission [19] — SHAP-based pipelines report strong discrimination (AUC typically 0.79–0.94) while surfacing clinically recognizable top features. Notably, an interpretability analysis of in-hospital mortality prediction went further, using SHAP not only to explain predictions but to audit fairness, finding that model reliance on demographic attributes varied unevenly across patient subgroups [20] — an important reminder that post-hoc attribution can serve an auditing function beyond simple explanation.

Table 1: Post-Hoc Feature-Attribution XAI for EHR Risk Prediction

Reference	Clinical Application	Model & XAI Method	Data / Scope	Key Finding
20	in-hospital mortality prediction	deep learning models; best-performing interpretability method	electronic health records	The best-performing interpretability method successfully identifies critical features for mortality prediction and recognizes new important featu...
19	Predicting 30-day unplanned readmission risk for elderly patients (age ≥ 50 year...)	Category boost (CatBoost), and four other machine learning approaches; Shapley additive explanations (SHAP)	Electronic health record (EHR) data including frailty,...	CatBoost model outperforms other models with a mean AUC of 0.79; prior readmissions, discharge to a rehabilitation facility, length of stay, como...
11	Prediction of sepsis mortality in ICU patients	Decision Trees, Gradient Boosting, XGBoost, LightGBM, Multilayer Perce...; SHAP (Shapley Additive Explanations)	clinical features	The Random Forest model achieved an AUROC of 0.94 with a confidence interval of ± 0.01 , outperforming other models.
12	Real-time prediction of sepsis onset in critical care	Time-phAsed machine learning model for Sepsis Prediction; Intuitive way to illustrate the impact of the input features on the ou....	Hourly-recorded clinical variables	The proposed model is accurate and interpretable for real-time prediction of sepsis onset in critical care, with utility scores of 0.430 on inter...

Continued on next page

Table 1 (continued)

Reference	Clinical Application	Model & XAI Method	Data / Scope	Key Finding
16	Prediction of breast cancer recurrence risk	AdaBoost; Shapley Additive Explanation (SHAP)	Feature importance ranking	AdaBoost algorithm had the best prediction performance for predicting breast cancer recurrence; CA125, CEA, Fbg, and tumor diameter were the most...
17	COVID-19 prognosis prediction using clinical markers	machine learning and deep learning algorithms; five explainable artificial intelligence techniques (XAI)	clinical and laboratory markers	Maximum testing accuracy of 95% was achieved; most important markers are c-reactive protein, basophils, lymphocytes, albumin, D-Dimer and neutro...
14	Predicting heart disease risk for early intervention and continuous monitoring	XGBoost; SHAP	Tabular data (symptoms and features from private and pu...	XGBoost with SF-2 feature subset achieved 97.57% accuracy, 96.61% sensitivity, 90.48% specificity, 95.00% precision, 92.68% F1 score, and 98...
15	Predicting mortality risk for patients with heart failure in intensive care unit...	Extreme Gradient Boosting (XGBoost); SHapley Additive exPlanation (SHAP)	Clinical data (first 24 hours of ICU admission)	The XGBoost model had the highest predictive performance among four models with an AUC of 0.824; the average blood urea nitrogen was recognized a...
18	Non-communicable diseases (NCDs) prediction	Deep neural network; Deep Shapley Additive Explanations (DeepSHAP)	Global and local	The proposed model outperforms various state-of-the-art models and can improve medical understanding of NCDs diagnosis by providing general insig...
13	predicting hypertension	random forest; Feature Importance, Partial Dependence Plot, Individual Conditional Ex...	cardiorespiratory fitness data	Different interpretability techniques can shed light on different insights on the model behavior where global interpretations can enable clinicia...

3.3. Ante-Hoc and Attention-Based Models

A smaller but growing body of work builds interpretability directly into the model architecture rather than explaining it after the fact (Table 2). Attention-weighted recurrent networks have been used to predict daily sepsis, myocardial infarction, and medication administration from ICU time-series, with attention maps highlighting influential clinical events [21]; notably, the same study found that attention on vancomycin-administration timing partly reflected clinician dosing behavior rather than the patient’s underlying physiology, illustrating that attention weights can encode workflow artifacts alongside genuine clinical signal. Other ante-hoc approaches include DeepSOFA, an interpretable deep learning severity score that outperformed the standard SOFA baseline for in-hospital mortality prediction [22]; hierarchical-attention recurrent networks for mortality prediction [23]; and predictive clustering models that generate explainable phenotype groupings from longitudinal EHR data [24]. Genuinely additive, rather than post-hoc, interpretability was demonstrated by a factored generalized additive model for predicting postoperative acute kidney injury and respiratory failure, which matched or exceeded black-box baselines while producing explanations with high clinical face validity [25]. Constraint-based frameworks that incorporate domain knowledge directly into the prediction pipeline, rather than explaining a black box, have also shown promise for early sepsis prediction and for improving trust calibration under model uncertainty [26, 27], and probabilistic graphical models applied to a sepsis case study demonstrated how interpretable structure can reveal a physiologically grounded relationship between organ-dysfunction scores and outcomes [28]. In cardiology, a deep learning model for atrial fibrillation detection paired its predictions with dedicated feature modules explaining the reasoning behind each classification, without relying on post-hoc attribution [29], and a tree-ensemble mortality model using intrinsic feature-interaction analysis surfaced previously overlooked risk factors

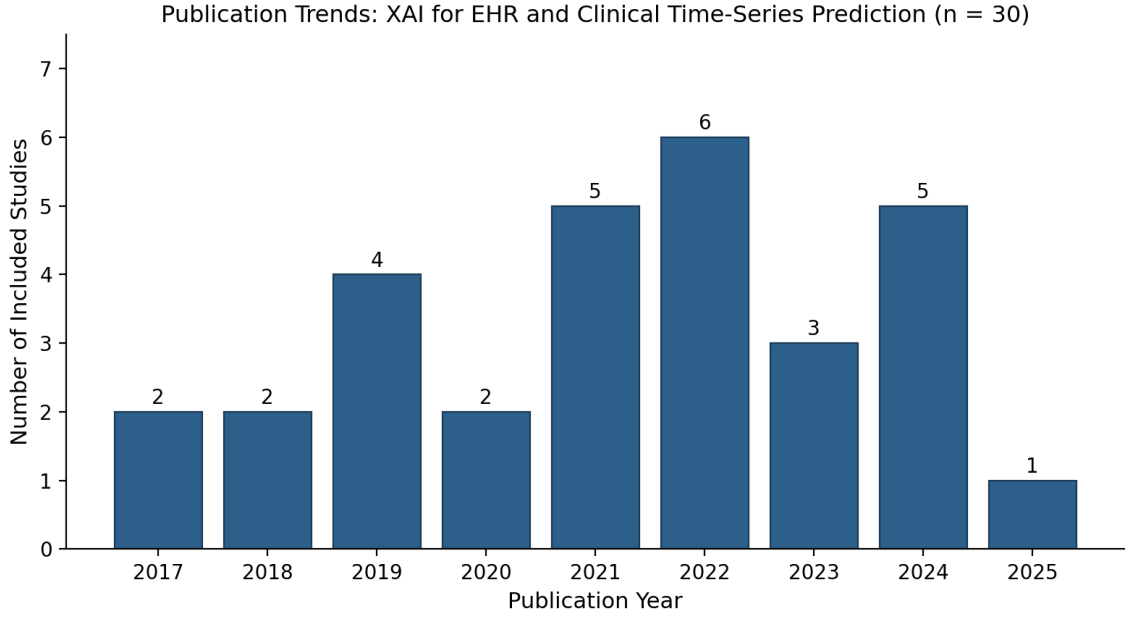


Figure 3: Publication trends for the 30 included studies, by year.

missed by linear baselines [30].

Table 2: Ante-Hoc and Attention-Based Interpretable Models for Clinical Time-Series

Reference	Clinical Application	Model & XAI Method	Data / Scope	Key Finding
21	Predicting daily sepsis, myocardial infarction (MI), and vancomycin antibiotic a...	Long short-term memory (LSTM) recurrent neural networks (RNNs) incorpo...; Attention maps	EHR data from MIMIC-III dataset	Models achieved next-day predictive AUC of 0.876 for sepsis, 0.823 for MI, and 0.833 for vancomycin administration.
22	Assessing illness severity and predicting in-hospital mortality among critically...	Deep learning (interpretable deep learning models); clinically interpretable deep learning	temporal measurements, streaming electronic health reco...	DeepSOFA yields significantly more accurate predictions of in-hospital mortality at any point during an ICU stay compared to SOFA baseline models...
23	mortality prediction	gated recurrent unit-based recurrent neural network with hierarchical...; hierarchical attention	diagnostic codes	The prediction accuracy of the model outperforms baseline models and the model demonstrates interpretability in visualizations.
24	disease risk prediction	non-parametric predictive clustering-based risk prediction model integ...; attention mechanisms and predictive clustering	longitudinal electronic health records (EHRs)	The model demonstrated effectiveness in capturing longitudinal EHR information for disease risk prediction and generated explainable evidence to...
25	Clinical decision support in the operating room for prediction of postoperative...	Factored generalized additive model (F-GAM); –	EHR / clinical time-series	F-GAM showed superior model performance in terms of AUPRC and AUROC compared to several other GAM implementations, random forests, support vector...

Continued on next page

Table 2 (continued)

Reference	Clinical Application	Model & XAI Method	Data / Scope	Key Finding
27	Early prediction of sepsis in ICU patients	Artificial Intelligence Sepsis Expert algorithm; produces a list of the most significant contributing factors	high-resolution vital signs time series and electronic...	Artificial Intelligence Sepsis Expert can accurately predict the onset of sepsis in an ICU patient 4–12 hours prior to clinical recognition, wit...
28	sepsis case study in critical care	Conditional independence models, Kernel methods; graphical models	EHR / clinical time-series	SIRS depends on both organ dysfunction measured through the Sequential Organ Failure (SOFA) score and the ICU outcome, thus concluding that SIRS...
26	early sepsis prediction	XGBoost; Trust-MAPS (trust-scores)	electronic health records	Trust-MAPS framework shows desirable behavior in handling potential errors and boosting predictive performance, achieving an AUC of 0.91 for pred...
29	detection of atrial fibrillation using electrocardiogram	deep learning model; two feature modules that describe the reason for DLM decisions	electrocardiogram (ECG)	The DLM successfully detected AF using diverse ECGs and described the reason for this decision; AUCs for detecting AF using 12-lead ECG were 0.99...
30	all-cause mortality prediction	tree ensemble model; XAI (Explainable Artificial Intelligence) technique	NHANES dataset (47,261 samples, 151 features) and UK Bi...	IMPACT models achieve higher accuracy than linear models and neural networks; identified several overlooked risk factors and interaction effects;...

3.4. Trust, Fairness, and Clinical Validation

A third group of studies steps back from proposing new models to evaluate the trustworthiness, fairness, or real-world validity of EHR- based explanations (Table 3). An ontology-grounded explanation framework for next-visit prediction found that incorporating temporal structure and medical domain knowledge improved the quality of mined explanations relative to generic post-hoc methods [31]. A comparative study of deep Taylor decomposition, Shapley values, and model coefficients for stroke outcome prediction found that, despite differing mechanisms, the methods produced broadly consistent rankings of important features [32] — an encouraging, if isolated, signal of convergent validity across explanation techniques. Disease-specific applications illustrate both the promise and the limits of current practice: an XAI framework for cardiovascular disease prediction reported only moderate accuracy despite explanation support [33]; chronic kidney disease models using SHAP and partial-dependence plots achieved high accuracy with clinically interpretable feature sets [34, 35]; and an EEG-based stroke-prediction study using LIME reported moderate (approximately 80%) accuracy, underscoring that explanation quality cannot compensate for a weaker underlying signal [36]. A combined SHAP/LIME framework spanning five distinct disease-prediction tasks reported very high accuracy and was explicitly designed to support non-specialist clinical decision-making [37], while XGBoost-based sepsis-mortality prediction using SHAP achieved strong discrimination and identified a consistent, clinically plausible set of top predictive features [38]. At the health-system level, a machine-learning risk-stratification tool deployed in a value-based primary-care setting showed early promise for identifying high-need patients, illustrating one of the few examples in this review of an XAI-supported model evaluated in an operational, rather than purely retrospective, setting [39]; a related LIME-based framework for dementia and activity-recognition classification similarly reported strong performance across multiple assessment tasks [40].

Table 3: Trust, Fairness, and Clinical Validation of EHR-Based XAI

Reference	Clinical Application	Model & XAI Method	Data / Scope	Key Finding
31	predicting the next visit of a patient based on clinical history	Doctor AI, a multilabel classifier; Doctor XAI, a model-agnostic ontology-based explainability technique	sequential, multi-labeled, ontology-linked clinical dat...	exploiting the temporal dimension in the data and the domain knowledge encoded in the medical ontology improves the quality of the mined explanat...
37	Disease prediction for Diabetes, Anaemia, Thalassemia, Heart Disease, Thrombocyt...	Decision Trees, Naive Bayes, Random Forests, XGBoost; SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model...	Feature importance for each prediction	The framework achieves 99.2% accuracy and provides understandable explanations for model outputs, enabling clinical practitioners to make inform...
32	stroke outcome prediction	logistic regression; tree boosting; multilayer perceptrons (MLPs); deep Taylor decomposition; Shapley values; model coefficients	clinical features	All models performed comparably (AUC 0.81-0.83). Interpretability analysis showed consistent results across models, rating age and stroke severity...
33	cardiovascular disease prediction	ensemble classification models, support vector machine (SVM), AdaBoost...; eXplainable Artificial Intelligence (XAI) framework	categorical, integer, and real type attributes	XAI-driven SVM, LR, and naive Bayes achieved an accuracy of 89%, which is reasonably fair compared to existing models
34	Diagnosis of chronic kidney disease	Extreme gradient boost (XGB); Shapley Additive Explanations (SHAP) and Partial Dependency Plots (PDP...	Global and local	XGB demonstrated the highest accuracy among six machine learning algorithms; a graphical user interface with explanations was developed for the f...
35	early diagnosis of chronic kidney disease	extreme gradient boosting classifier; explainability analysis (feature importance)	clinical features (hemoglobin, specific gravity, hypert...	optimal model uses 3 features (hemoglobin, specific gravity, hypertension) with accuracy 99.2% (SD 0.8) and 97.5% on unseen data; hemoglobin is...
36	stroke prediction using EEG signal	Adaptive Gradient Boosting models; Eli5 and LIME	EEG signal	The Adaptive Gradient Boosting models showed around 80% accuracy for classification of the control group and the stroke group. showing early promise in identifying high-risk patient needs and boosting clinician...
39	risk stratification in a high-need, value-based primary care setting	machine learning-based risk stratification model; -	EHR / clinical time-series	
38	Early prediction of in-hospital mortality in critically ill patients with sepsis	eXtreme Gradient Boosting (XGBoost); SHapley Additive exPlanations (SHAP)	Clinical features (57 clinical parameters collected on...	The XGBoost model achieved the best performance with an AUC of 0.884 and an accuracy of 89.5% in the validation cohort.

Continued on next page

Table 3 (continued)

Reference	Clinical Application	Model & XAI Method	Data / Scope	Key Finding
40	cognitive health assessment, health and dementia classification, activity recogn...	random forest, K-nearest neighbor, support vector machine, decision tree,...; Local Interpretable Model Agnostic (LIME)	sensor data from smart home	XAI-HAR achieves 0.96% of F-score for health and dementia classification and 0.95 and 0.97% for activity recognition of dementia and healthy in...

4. Discussion

This review synthesizes 30 studies applying XAI specifically to EHR and clinical time-series prediction. Three cross-cutting findings stand out.

Post-hoc attribution dominates, but with an unresolved validity question. SHAP is by far the most common explanation method in this literature, and it consistently identifies clinically recognizable risk factors [11, 14]. Yet a direct comparison of interpretability techniques found that different methods surface complementary rather than fully overlapping insights [13], and no study in this review directly demonstrates that SHAP-based explanations, on their own, change clinician behavior or improve patient outcomes relative to an unexplained model of equivalent accuracy. The field has largely validated technical properties of these explanations (consistency, fidelity) rather than their downstream clinical value.

Attention and feature importance can encode workflow, not just physiology. A recurring caution in this review is that attention-based and attribution-based explanations can reflect artifacts of clinical process — such as when a medication was ordered — rather than the patient’s underlying condition [21]. This is a distinct failure mode from the spurious-correlation problem documented in imaging XAI, and one that EHR-specific evaluation frameworks have only begun to address [6].

Ante-hoc models remain underused relative to their promise. Intrinsically interpretable architectures — additive models [25], constraint-based frameworks [26], and interpretable severity scores [22] — match or exceed black-box baselines in several of the reviewed studies while avoiding the post-hoc validity question entirely. Their comparatively rare appearance in this literature, relative to SHAP-based post-hoc explanation, likely reflects the convenience of applying an off-the-shelf explainer to an existing high-performing model rather than any inherent performance disadvantage of ante-hoc design.

Fairness auditing is an emerging, valuable use of XAI. Beyond explaining individual predictions, interpretability analysis has been used to expose demographic disparities in model reliance and treatment recommendations [20], and ontology-grounded explanation frameworks show that incorporating domain knowledge can improve explanation quality over generic post-hoc methods [31]. This dual-use potential — explanation and bias auditing — deserves more systematic attention in future EHR-XAI research.

Limitations. This review was restricted to English-language, EHR- and time-series-specific studies, which may exclude relevant work published as broader multimodal studies; it included pre-prints that have not undergone peer review; and category assignment for studies combining post-hoc and ante-hoc elements involved subjective judgment. Among the reviewed studies themselves, the dominant limitation is that almost all evidence is retrospective and derived from public ICU datasets (predominantly MIMIC), leaving prospective, real-world validation of EHR-XAI systems largely absent from the literature [39] being a notable exception.

Future directions. We highlight four priorities: (1) prospective and operational evaluation of EHR-XAI systems, since evidence for clinical impact remains almost entirely retrospective; (2) greater adoption of ante-hoc, intrinsically interpretable models as a genuine alternative to post-hoc explanation of black-box ensembles; (3) explanation methods and evaluation frameworks that explicitly account for clinical-workflow confounding, distinct from standard feature-importance validation; and (4) wider use of XAI as a fairness-auditing tool, building on early demonstrations in mortality prediction [20].

5. Conclusion

This systematic review examined how explainable AI is applied specifically to electronic health record and clinical time-series prediction, synthesizing 30 studies across post-hoc attribution, ante-hoc model design, and trust/fairness

evaluation. SHAP-based post-hoc explanation dominates the field and reliably surfaces clinically plausible risk factors, but the review finds limited direct evidence that these explanations improve clinical decision-making, and documents a specific and under-recognized risk: attention and attribution scores in temporal EHR models can reflect clinical workflow rather than physiology. Ante-hoc, intrinsically interpretable models offer a promising and comparatively underused alternative, and interpretability tools have shown real value as fairness-auditing instruments. Advancing this field will require less emphasis on generating additional post-hoc explanations and more on prospective validation, workflow-aware evaluation, and wider adoption of interpretable-by-design architectures for EHR and clinical time-series prediction.

Author Contributions

Deepak Singh(D.S): Conceptualization, Methodology, Investigation, Formal analysis, Software, Data curation, Visualization, Writing – original draft; Smita Nirkhi(S.N): Conceptualization, Methodology, Validation, Supervision, Writing – review editing, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The data that support the findings of this study are available from the corresponding author, S.N, upon reasonable request.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this manuscript, the authors used Paperpal and Grammarly to support language refinement, grammatical correction, and improvements in readability and presentation. Mendeley Cite was used for citation insertion, reference organization, and bibliography management. Additional AI-assisted tools were used selectively to accelerate early-stage drafting and improve the efficiency of manuscript preparation. All AI-assisted content was subsequently reviewed, edited, and verified by the authors, who remain fully responsible for the accuracy, originality, interpretation, and final content of the manuscript.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

- [1] N Tomašev, X Glorot, JW Rae, M Zielinski, H Askham, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 2019.
- [2] E Choi, MT Bahadori, A Schuetz, et al. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, 2016.
- [3] EJ Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 2019.
- [4] S Wachter, B Mittelstadt, and L Floridi. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International data privacy law*, 2017.
- [5] US Food and Drug Administration. Artificial intelligence and machine learning in software as a medical device. Technical report, US Food and Drug Administration, Silver Spring, MD, 2021.

- [6] C Zhang, S Bengio, M Hardt, B Recht, et al. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 2021.
- [7] SM Lundberg and SI Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 2017.
- [8] MT Ribeiro, S Singh, and C Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 2016.
- [9] SS Band, A Yarahmadi, CC Hsu, M Biyari, et al. Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. *Informatics in Medicine Unlocked*, 2023.
- [10] MJ Page, JE McKenzie, PM Bossuyt, et al. The prisma 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372:n71, 2021.
- [11] Jiayi Gao, Yuying Lu, Negin Ashrafi, I. Domingo, K. Alaei, and M. Pishgar. Prediction of sepsis mortality in icu patients using machine learning methods. *BMC Medical Informatics and Decision Making*, 24, 2024.
- [12] X Li, X Xu, F Xie, X Xu, Y Sun, X Liu, X Jia, et al. A time-phased machine learning model for real-time prediction of sepsis in critical care. *Critical Care Medicine*, 2020.
- [13] Radwa Elshawi, M. Al-Mallah, and Sherif Sakr. On the interpretability of machine learning-based model for predicting hypertension. *BMC Medical Informatics and Decision Making*, 19, 2019.
- [14] Hosam F. El-Sofany, B. Bouallegue, and Y. M. A. El-Latif. A proposed technique for predicting heart disease using machine learning algorithms and an explainable ai method. *Scientific Reports*, 14, 2024.
- [15] J Li, S Liu, Y Hu, L Zhu, Y Mao, and J Liu. Predicting mortality in intensive care unit patients with heart failure using an interpretable machine learning model: retrospective cohort study. *Journal of medical Internet research*, 2022.
- [16] Duo Zuo, Lexin Yang, Yu Jin, Huan Qi, Yahui Liu, and Li Ren. Machine learning-based models for the prediction of breast cancer recurrence risk. *BMC Medical Informatics and Decision Making*, 23, 2023.
- [17] Krishnaraj Chadaga, Srikanth Prabhu, Niranjana Sampathila, R. Chadaga, S. Umakanth, Devadas Bhat, and Shashi Kumar G S. Explainable artificial intelligence approaches for covid-19 prognosis prediction using clinical markers. *Scientific Reports*, 14, 2024.
- [18] Khishigsuren Davagdorj, J. Bae, Van-Huy Pham, N. Theera-Umpon, and K. Ryu. Explainable artificial intelligence based framework for non-communicable diseases prediction. *IEEE Access*, 9:123672–123688, 2021.
- [19] S. Mohanty, Deborah A Lekan, T. McCoy, M. Jenkins, and Prashanti Manda. Machine learning for predicting readmission risk among the frail: Explainable ai for healthcare. *Patterns*, 3, 2021.
- [20] Chuizheng Meng, Loc Trinh, Nan Xu, James Enouen, and Yan Liu. Interpretability and fairness evaluation of deep learning models on mimic-iv dataset. *Scientific Reports*, 12, 2022.
- [21] Deepak A Kaji, J. Zech, Jun S. Kim, Samuel K. Cho, N. Dangayach, A. Costa, and E. Oermann. An attention based deep learning model of clinical events in the intensive care unit. *PLoS ONE*, 14, 2019.
- [22] Benjamin Shickel, T. Loftus, L. Adhikari, T. Ozrazgat-Baslanti, A. Bihorac, and Parisa Rashidi. Deepsofa: A continuous acuity score for critically ill patients using clinically interpretable deep learning. *Scientific Reports*, 9, 2018.
- [23] Ying Sha and May D. Wang. Interpretable predictions of clinical outcomes with an attention-based recurrent neural network. *ACM-BCB ... : the ... ACM Conference on Bioinformatics, Computational Biology and Biomedicine. ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, 2017:233 – 240, 2017.
- [24] Shuai Niu, Qing Yin, Jing Ma, Yunya Song, Yida Xu, Liang Bai, Wei Pan, and Xian Yang. Enhancing healthcare decision support through explainable ai models for risk prediction. *Decision Support Systems*, 181:114228, 2024.
- [25] Zhicheng Cui, Bradley A. Fritz, C. King, M. Avidan, and Yixin Chen. A factored generalized additive model for clinical decision support in the operating room. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2019:343–352, 2019.
- [26] Mehak Arora, Hassan Mortagy, Nathan Dwarshuis, Jeffrey Wang, Philip Yang, A. Holder, Swati Gupta, and R. Kamaleswaran. Improving clinical decision support through interpretable machine learning and error handling in electronic health records. *Journal of the American Medical Informatics Association : JAMIA*, 2023.

- [27] S. Nemati, A. Holder, Fereshteh Razmi, Matthew D. Stanley, G. Clifford, and T. Buchman. An interpretable machine learning model for accurate prediction of sepsis in the icu. *Critical care medicine*, 46:547 – 553, 2017.
- [28] A. Vellido, V. Ribas, Carles Morales, Adolfo Ruiz Sanmartín, and Juan Carlos Ruiz Rodríguez. Machine learning in critical care: state-of-the-art and a sepsis case study. *BioMedical Engineering OnLine*, 17, 2018.
- [29] YY Jo, Y Cho, SY Lee, J Kwon, KH Kim, KH Jeon, et al. Explainable artificial intelligence to detect atrial fibrillation using electrocardiogram. *International Journal of Cardiology*, 2021.
- [30] Wei Qiu, Hugh Chen, A. Dincer, Scott M. Lundberg, M. Kaeberlein, and Su-In Lee. Interpretable machine learning prediction of all-cause mortality. *Communications Medicine*, 2, 2022.
- [31] Cecilia Panigutti, A. Perotti, and D. Pedreschi. Doctor xai: an ontology-based approach to black-box sequential data classification explanations. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020.
- [32] Esra Zihni, V. Madai, Michelle Livne, Ivana Galinovic, A. Khalil, J. Fiebach, and D. Frey. Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome. *PLoS ONE*, 15, 2019.
- [33] P. Guleria, Parvathaneni Naga Srinivasu, Shakeel Ahmed, N. Almusallam, and F. Alarfaj. Xai framework for cardiovascular disease prediction using classification techniques. *Electronics*, 2022.
- [34] Gangani Dharmarathne, Madhusa Bogahawaththa, Marion McAfee, Upaka Rathnayake, and D.P.P. Meddage. On the diagnosis of chronic kidney disease using a machine learning-based interface with explainable artificial intelligence. *Intelligent Systems with Applications*, 22:200397, 2024.
- [35] Pedro A. Moreno-Sánchez. Data-driven early diagnosis of chronic kidney disease: Development and evaluation of an explainable ai model. *IEEE Access*, 11:38359–38369, 2021.
- [36] Mohammed Saidul Islam, I. Hussain, Md Mezbaur Rahman, Se Jin Park, and Md. Azam Hossain. Explainable artificial intelligence model for stroke prediction using eeg signal. *Sensors (Basel, Switzerland)*, 22, 2022.
- [37] R Agrawal, T Gupta, S Gupta, S Chauhan, P Patel, et al. Fostering trust and interpretability: integrating explainable ai (xai) with machine learning for enhanced disease prediction and decision transparency. *Diagnostic Pathology*, 2025.
- [38] Chang Hu, Lu Li, Wei ping Huang, Tong Wu, Qiancheng Xu, Juan Liu, and Bo Hu. Interpretable machine learning for early prediction of prognosis in sepsis: A discovery and validation study. *Infectious Diseases and Therapy*, 11: 1117 – 1132, 2022.
- [39] S. Bhatt, Adam Cohon, J. Rose, Natalia Majerczyk, B. Cozzi, Drew Crenshaw, and Griffin R Myers. Interpretable machine learning models for clinical decision-making in a high-need, value-based primary care setting. *NEJM Catalyst*, 2021.
- [40] A. R. Javed, H. Khan, M. Alomari, M. U. Sarwar, Muhammad Asim, Ahmad S. Almadhor, and Muhammad Zahid Khan. Toward explainable ai-empowered cognitive health assessment. *Frontiers in Public Health*, 11, 2023.