Explainability And Interoperability As Twin Pillars Of Artificial Intelligence In Healthcare

Sujata Ambekar[1], Kalpana Joshi[2]
[1,2]*Department of Computer Science, Fergusson College (Autonomous), Pune*
doi.org/10.64643/IJIRT211729

***Abstract*—Artificial intelligence (AI) now matches specialist performance on many narrow clinical tasks, yet few models reach routine care. This paper argues that two properties namely, explainability and interoperability. Rather than focusing on raw accuracy, they decide whether a model is adopted. Explainability means a clinician can see why a prediction was made whereas interoperability means the model can receive data from and return results to existing health information systems. We review current explainability techniques, covering inherently interpretable models, post hoc feature attribution and saliency mapping, with the evidence on their limits. Then we review interoperability across its technical, syntactic, semantic and organisational levels, with emphasis on the standards most relevant to AI such as HL7 FHIR, SMART on FHIR and the FAIR principles. We show that they depend on each other. An explanation is only as meaningful as the coded data behind it, and an explanation that cannot travel with the prediction through a standard interface never reaches the bedside. On this basis we propose XI-Care, a five-layer conceptual framework that treats explanation as a standards-based, exchangeable data object under a continuous governance layer. We close with emerging trends namely, large language models, federated learning, fairness auditing and the TRIPOD+AI and FUTURE-AI guidelines and with open research questions.**

***Index Terms*—explainable artificial intelligence; interoperability; HL7 FHIR; SMART on FHIR; clinical decision support; federated learning**

I. INTRODUCTION

Over the past decade, the question asked of healthcare AI has quietly changed. Early studies asked whether a model could match clinicians; the harder question now is why so few of these models are ever switched on in a hospital. Deep networks perform at specialist level on a range of imaging, risk-prediction and workflow tasks [1]. Rajkomar et al. showed that deep models trained on unprocessed EHR data from two academic hospitals could forecast in-hospital death, unplanned readmission and prolonged stays [2], and large language models (LLMs) have since carried similar gains into text-heavy work such as note summarisation and clinical question answering [3]. Routine clinical use, however, still lags well behind what these studies report.

In our reading of the literature, two barriers come up again and again. The first is opacity. The best-performing models are usually black boxes, and Rudin goes further than most critics: where the stakes are high, she argues, an opaque model should not be explained after training but replaced by one that is interpretable from the start [4]. Opacity can also conceal harm. Obermeyer et al. examined a commercial algorithm that US health systems used to target extra care and found that it systematically underestimated how sick Black patients were. The cause lay in the training label, past spending on care, which reflects access to treatment as much as illness, and the problem came to light only once researchers looked inside the model's logic [5].

The second barrier is fragmentation. Patient data are scattered across systems that use different formats and different vocabularies, and Lehne et al. make the case that no part of digital medicine, AI included, can work at scale without interoperability [6]. Put simply, a model that runs only on one hospital's exported spreadsheet remains a research prototype; it becomes a clinical tool when it can read from, and write back to, the hospital's own systems through standard interfaces.

These two topics rarely meet. Explainability is studied mainly by the machine learning community, while interoperability belongs to health informatics, and the two are usually discussed in separate venues.

This paper brings them into a single discussion and makes three contributions:

1) A structured survey of explainability techniques for clinical AI, together with what is known about how far they can be trusted (Section 2).
2) A review of the interoperability standards that an AI system must satisfy to enter clinical workflow (Section 3).
3) XI-Care, a conceptual framework in which explanations are generated, encoded and exchanged as standard data objects under continuous governance (Section 4).

The paper is a narrative review with a conceptual proposal. It reports no new experimental results; empirical validation of the framework is identified as future work.

II. EXPLAINABILITY IN CLINICAL AI

2.1. Concepts and taxonomy

Explainability is the ability of a model to give a specific audience the reasons behind its behaviour. Barredo Arrieta et al. separate transparent models, which are understandable by design, from post hoc techniques that explain an already trained opaque model [7]. Tjoa and Guan apply this distinction to medicine and note that clinical explanations carry a higher burden, because an error in the explanation can mislead a treatment decision [8]. Two further axes matter in practice: scope (a local explanation of one patient's prediction versus a global account of the model) and dependence on model type (model-agnostic versus model-specific). Figure 1 arranges the main method families along these axes.

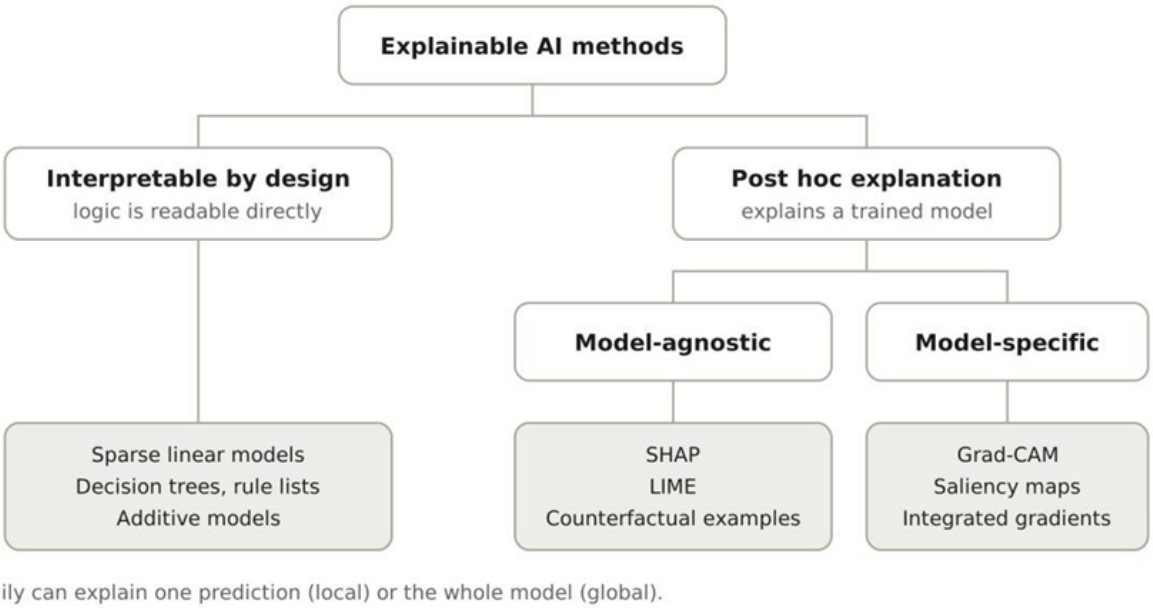

Figure 1. Taxonomy of explainability methods for clinical AI

2.2. Principal methods

Feature attribution-

SHAP assigns each input feature a contribution to a prediction using Shapley values from cooperative game theory, and unifies several earlier attribution methods under one additive framework [9]. LIME fits a simple surrogate model around a single prediction and reports the SHAP draws on Shapley values, a concept from cooperative game theory, to share out a prediction among the input features, and Lundberg and Lee showed that several earlier attribution methods are special cases of this additive scheme [9]. LIME takes a different route: it perturbs the input around one case, fits a simple local model to the black box's responses, and reports that local model's coefficients as the explanation [10]. Because neither method needs access to the model's internals, both are widely used on tabular EHR data.

Saliency mapping-

Grad-CAM takes the gradient of the target class with respect to the feature maps of the final convolutional

layer and uses it to weight those maps, giving a low-resolution heat map of the parts of the image that pushed the model towards its prediction [11]. It is now routinely applied to chest radiographs and retinal photographs, and one-dimensional variants are used on physiological signals such as the ECG.

Interpretable-by-design models-

Sparse linear models, decision trees, rule lists and generalised additive models are transparent by construction, so a clinician can inspect the whole decision logic.

Rudin argues that such models should be the starting point in high-stakes medicine, and that the supposed loss of accuracy from choosing them is frequently small or absent when the input features are carefully engineered [4].

Table 1 compares the families.

Table 1. Comparison of explainability method families used in clinical AI

Method family	Scope	Model dependence	Typical clinical data	Main limitation
Interpretable-by-design models [4]	Global and local	Intrinsic	Tabular EHR, risk scores	Harder to apply to raw images and signals
SHAP [9]	Local, aggregates to global	Agnostic	Tabular EHR, laboratory values	Costly to compute; sensitive to correlated features
LIME [10]	Local	Agnostic	Tabular data, text	Explanations vary with the sampling of neighbours
Grad-CAM [11]	Local	Specific to convolutional networks	Medical images, ECG	Coarse; shows where the model looked, not why
Counterfactual and example-based	Local	Either	Tabular data, images	May suggest clinically implausible changes

2.3. What explanations can and cannot do?

An explanation is worth only what its reader can do with it. Holzinger et al. call this quality causability, meaning how well an explanation helps a particular expert, in a particular setting, arrive at a causal understanding of the result; it can therefore only be judged for a specific user and context [12]. Amann et al. examine explainability from four angles (technical, legal, medical and the patient's) and find that the requirement differs across them. They conclude that decision-support tools which omit explainability put basic principles of medical ethics at risk [13].

Ghassemi et al. offer the sharpest warning. In their view, today's post hoc methods are not able to justify a decision about an individual patient, and relying on them can give clinicians unwarranted confidence. Trust, they argue, should rest instead on thorough internal and external validation, with explanation tools better suited to auditing models than to supporting individual decisions [14]. We find this argument persuasive, although we would not go as far as removing explanations from the bedside altogether. Our own position, which shapes the framework in Section 4, has three parts: use interpretable models whenever they perform adequately; read post hoc explanations as hypotheses that need checking rather than as evidence; and test every explanation with the clinicians who are meant to use it.

III. INTEROPERABILITY AS THE DEPLOYMENT SUBSTRATE

3.1. Four levels

Lehne et al. define interoperability as the capacity of systems not only to exchange data but also to make use of it, and they distinguish four levels [6]. At the technical level, the question is simply whether data can be transmitted from one system to another. The syntactic level concerns the structure and format of what is sent. Semantic interoperability requires that a code carries the same meaning for the receiving system as for the sender, while organisational interoperability covers the workflows, policies and legal agreements that govern the exchange. An AI model can break down at any of these levels: it may fail to connect, fail to parse the message, misread a code, or lack permission to act on its output.

3.2. Standards relevant to AI

Among syntactic standards, HL7 FHIR (Fast

Healthcare Interoperability Resources) is now the most widely adopted. FHIR breaks clinical information into small, reusable building blocks called resources (a patient, a laboratory observation, a diagnosis) that are exchanged over RESTful web APIs. SMART on FHIR builds on this with a common way to authorise and launch applications, which lets a third-party app run inside any EHR that supports the specification without bespoke integration work [15]. The consequence for AI is significant: a model wrapped as a SMART app can, in principle, be built once and deployed across many hospitals.

FHIR is seldom used on its own in research. Vorisek et al. reviewed 49 studies and found that most used it for data capture and standardisation, and that a majority paired it with other terminologies or data models, in particular LOINC, SNOMED CT, ICD-10 and OMOP. The obstacles they report include resource definitions that change between FHIR versions, legal uncertainty, and the overhead of running a FHIR server [16]. For AI, the lesson is that a FHIR pipeline does not by itself guarantee semantic consistency, which is the gap that requirement R1 of our framework addresses. The FAIR principles add a broader expectation: data and their metadata should be easy for both people and software to locate, access, combine and reuse [17].

The practical value for machine learning has been demonstrated. Rajkomar et al. mapped the full EHR records of two academic medical centres to a FHIR-based representation and trained deep models on both without site-specific data harmonisation [2].

Table 2. Interoperability levels (after [6]) and their relevance to clinical AI

Level	Question it answers	Representative standards	Why AI needs it
Technical	Can the systems connect?	Web protocols, secure transport	Real-time data feed and result return
Syntactic	Is the structure shared?	HL7 FHIR, DICOM	One input pipeline across sites
Semantic	Is the meaning shared?	SNOMED CT, LOINC, ICD-10, OMOP	Features keep their meaning across hospitals
Organisational	Are processes and rules aligned?	Consent, governance and data-sharing agreements	Lawful deployment and accountability

3.3. Interoperability and collaborative learning

In federated learning, each institution trains the shared model on its own data and sends back only model updates, so patient records never leave the hospital; this removes many of the privacy and governance obstacles that block multi-centre studies [18]. It does not, however, make interoperability unnecessary. If the participating sites encode their data in different structures or vocabularies, the global model is in effect learning from inputs that do not mean the same thing. Some degree of standardisation therefore has to be in place before collaborative training can work. Figure 2 shows the resulting path from source systems to the clinician.

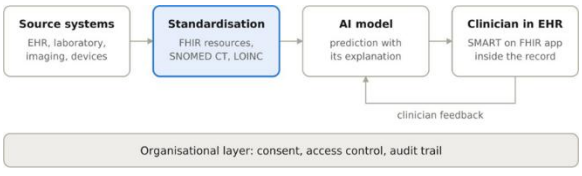

Figure 2. Interoperable data flow for clinical AI

IV. XI-CARE: A FRAMEWORK FOR EXPLAINABLE AND INTEROPERABLE CLINICAL AI

4.1. Rationale

Explainability and interoperability fail together when they are designed apart. A SHAP value attached to a local column name such as "lab_17" explains nothing to a clinician at another hospital; the same value attached to a LOINC-coded observation does. Likewise, a well-validated saliency map is wasted if the EHR interface can show only a risk score.

We therefore propose XI-Care (eXplainable, Interoperable Care), a conceptual architecture with one organising idea: the explanation is a first-class data object that is encoded in the same standards as the data it explains and travels with the prediction.

Figure 3 shows the five layers.

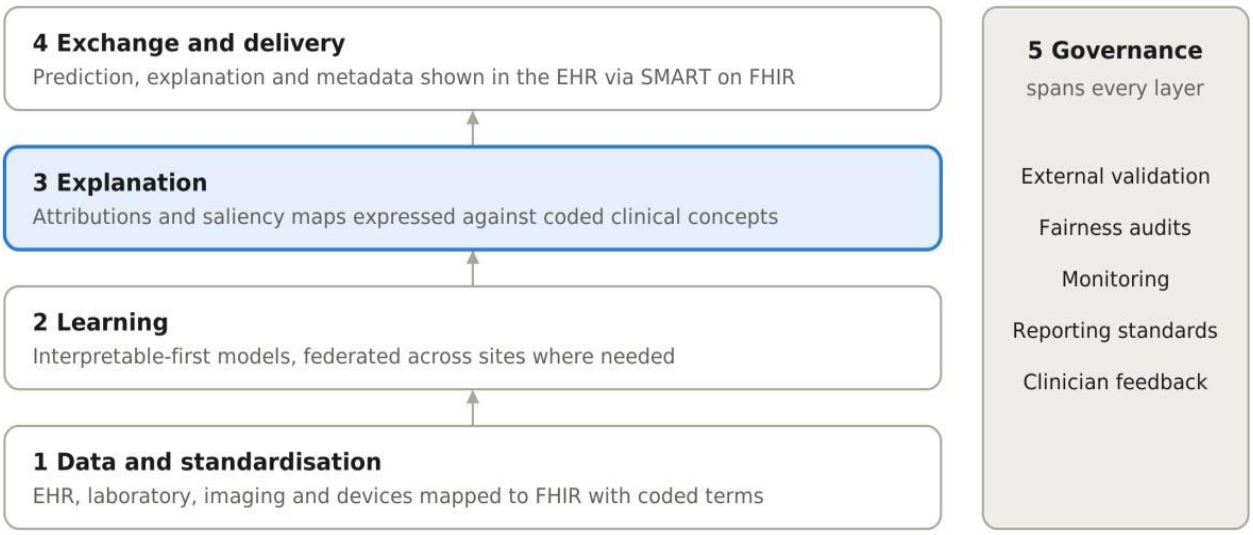

Figure 3. The proposed XI-Care framework

4.2. Layers

1. Data and standardization-

Source systems (EHR, laboratory, imaging, devices) are mapped to FHIR resources with coded terminologies. Every model feature keeps a link to its source code and resource.

2. Learning-

Models are trained on the standardised representation, interpretable-by-design where performance permits [4], and federated across sites where data cannot be pooled [18].

3. Explanation-

Each prediction is issued with an explanation suited to the data type: feature attributions for tabular inputs [9], [10], saliency maps for images and signals [11]. Attributions are expressed against coded clinical concepts, not internal variable names.

4. Exchange and delivery-

The risk score, its explanation and the model's metadata (version number, intended use, and the population on which it was validated) are bundled into a single standards-based payload, which is displayed within the clinician's own EHR by means of a SMART on FHIR app [15].

5. Governance-

A cross-cutting layer covers external validation, subgroup fairness audits, post-deployment monitoring and transparent reporting. When clinicians disagree with or comment on an explanation, that feedback is recorded and passed back to the learning layer for review.

4.3. Design requirements

Table 3. Design requirements of the XI-Care framework

ID	Requirement	Addresses
R1	Every input feature is traceable to a standard code	Semantic interoperability; meaningful attributions
R2	Every prediction carries an explanation and model metadata	Opacity; auditability
R3	Explanations are evaluated with intended users before deployment	Causability [12]; false reassurance [14]
R4	Delivery uses open, vendor-neutral interfaces	Portability across EHR systems
R5	Performance and explanations are monitored by subgroup after deployment	Hidden bias [5]; data drift

4.4. Illustrative scenario

Consider a cardiovascular risk model that combines ECG waveforms with EHR variables. Under XI-Care, laboratory results arrive as LOINC-coded FHIR Observations and diagnoses as coded Conditions (layer 1). The model returns a risk estimate (layer 2), a ranked

list of contributing coded features and a saliency trace over the ECG segment that most influenced the output (layer 3). The cardiologist sees all three in the EHR beside the patient record (layer 4), and any disagreement is recorded for audit (layer 5). The scenario is illustrative; it has not been implemented or evaluated in this paper.

V. EMERGING TRENDS AND OPEN CHALLENGES

Large language models-

LLMs are now being tried on a wide range of biomedical tasks, with results that are striking in some areas and disappointing in others [3]. They bring the two themes of this paper back in a new form. An LLM can produce a fluent, convincing rationale for its answer, but nothing guarantees that this rationale reflects how the answer was actually generated, and a plausible story is not the same thing as an explanation. On the other hand, LLMs could support interoperability by extracting coded, structured data from free-text notes, and could support explainability by turning feature attributions into plain-language summaries for patients. Neither use should be adopted without the kind of validation discussed in Section 2.3.

Federated and privacy-preserving learning-

Federated learning has begun to move out of simulation studies and into real collaborations between institutions [18]. Several problems remain open: sites differ in their patient populations and data quality, it is unclear who should explain a model that no single participant controls, and the updates that sites exchange need governance of their own.

Fairness and bias auditing-

The Obermeyer case [5] is a reminder that bias can be built into the choice of training label and still go unnoticed by standard accuracy metrics. Reviewing explanations separately for each patient subgroup, with demographic variables coded in a consistent, standard way, gives auditors a workable method for catching problems of this kind.

Reporting and consensus guidelines-

Trust is increasingly codified. TRIPOD+AI provides a 27-item checklist for reporting clinical prediction models, whether built with regression or machine learning [19]. The FUTURE-AI guideline, agreed by 117 experts from 50 countries, sets six principles (fairness, universality, traceability, usability, robustness and explainability) and 30 best practices across the AI Expectations for trustworthy clinical AI are increasingly being written down as formal guidance. TRIPOD+AI is a 27-item reporting checklist that applies to prediction models whether they are developed with regression or with machine learning [19]. FUTURE-AI, an international consensus guideline, organises its recommendations around six principles (fairness, universality, traceability, usability, robustness and explainability) and sets out 30 best practices spanning the full life cycle of a model [20]. Two of these principles map directly onto the themes of this paper: universality calls for the use of standards, and explainability is the subject of Section 2. We read this pairing as further support for treating the two as joint design goals rather than separate concerns.

Open research questions:

- How should an explanation be represented so that any compliant EHR can store, display and audit it?
- Which measures of explanation quality predict better clinical decisions, as opposed to higher user confidence [14]?
- How can explanations remain stable when the same model is deployed across sites with different coding practices?
- What is the cost, in accuracy and engineering effort, of an interpretable-first policy for multimodal data?

VI. CONCLUSION

High accuracy is a necessary condition for clinical AI, but on its own it has not been enough to secure adoption. Our review suggests that explainability and interoperability account for much of the remaining gap, and that each works best when designed alongside the other. Explanation methods are useful, but only after they have been tested with the people who will rely on them. The open standards such as FHIR, in turn, allow a model to move between institutions while keeping the meaning of its inputs intact. XI-Care brings these strands together by encoding explanations in the same standards as the clinical data they describe and delivering them through the same interfaces, within a continuous governance process. The

framework is at present conceptual. In future work we plan to build a prototype on a publicly available critical-care dataset, propose a standard representation for explanations, and carry out a clinician study to test whether standards-linked explanations lead to better clinical decisions.

REFERENCES

[1] E. J. Topol, “High-performance medicine: the convergence of human and artificial intelligence,” *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019, doi: 10.1038/s41591-018-0300-7.

[2] A. Rajkomar *et al*., “Scalable and accurate deep learning with electronic health records,” *npj Digital Medicine*, vol. 1, Art. no. 18, 2018, doi: 10.1038/s41746-018-0029-1.

[3] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, “Large language models in medicine,” *Nature Medicine*, vol. 29, no. 8, pp. 1930–1940, 2023, doi: 10.1038/s41591-023-02448-8.

[4] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019, doi: 10.1038/s42256-019-0048-x.

[5] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, “Dissecting racial bias in an algorithm used to manage the health of populations,” *Science*, vol. 366, no. 6464, pp. 447–453, 2019, doi: 10.1126/science. aax2342.

[6] M. Lehne, J. Sass, A. Essenwanger, J. Schepers, and S. Thun, “Why digital medicine depends on interoperability,” *npj Digital Medicine*, vol. 2, Art. no. 79, 2019, doi: 10.1038/s41746-019-0158-1.

[7] A. Barredo Arrieta *et al*., “Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Information Fusion*, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.

[8] E. Tjoa and C. Guan, “A survey on explainable artificial intelligence (XAI): Toward medical XAI,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2021, doi: 10.1109/TNNLS.2020.3027314.

[9] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 4765–4774. [Online]. Available: https://arxiv.org/abs/1705.07874

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should I trust you?’: Explaining the predictions of any classifier,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[12] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, “Causability and explainability of artificial intelligence in medicine,” *WIREs Data Mining and Knowledge Discovery*, vol. 9, no. 4, Art. no. e1312, 2019, doi: 10.1002/widm.1312.

[13] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, “Explainability for artificial intelligence in healthcare: A multidisciplinary perspective,” *BMC Medical Informatics and Decision Making*, vol. 20, Art. no. 310, 2020, doi: 10.1186/s12911-020-01332-6.

[14] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, “The false hope of current approaches to explainable artificial intelligence in health care,” *The Lancet Digital Health*, vol. 3, no. 11, pp. e745–e750, 2021, doi: 10.1016/S2589-7500(21)00208-9.

[15] J. C. Mandel, D. A. Kreda, K. D. Mandl, I. S. Kohane, and R. B. Ramoni, “SMART on FHIR: A standards-based, interoperable apps platform for electronic health records,” *Journal of the American Medical Informatics Association*, vol. 23, no. 5, pp. 899–908, 2016, doi: 10.1093/jamia/ocv189.

[16] C. N. Vorisek *et al*., “Fast Healthcare Interoperability Resources (FHIR) for interoperability in health research: Systematic review,” *JMIR Medical Informatics*, vol. 10, no. 7, Art. no. e35724, 2022, doi: 10.2196/35724.

[17] M. D. Wilkinson *et al*., “The FAIR Guiding Principles for scientific data management and stewardship,” *Scientific Data*, vol. 3, Art. no. 160018, 2016, doi: 10.1038/sdata.2016.18.

[18] N. Rieke *et al*., “The future of digital health with federated learning,” *npj Digital Medicine*, vol. 3, Art. no. 119, 2020, doi: 10.1038/s41746-020-00323-1.

[19] G. S. Collins *et al*., “TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods,” *BMJ*, vol. 385, Art. no. e078378, 2024, doi: 10.1136/bmj-2023-078378.

[20] K. Lekadir *et al*., “FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare,” *BMJ*, vol. 388, Art. no. e081554, 2025, doi: 10.1136/bmj-2024-081554.