VINE Journal of Information and Knowledge Management Systems

When resistance becomes knowledge: governing health-care AI through federated learning and explainable AI

Giuseppe Lanfranchi
Department of Mathematics and Computer Sciences, Physical Sciences and Earth Sciences, University of Messina, Messina, Italy, and Bayes Business School, City St. George University of London, London, UK

Roberto Marino
Department of Mathematics and Computer Sciences, Physical Sciences and Earth Sciences, University of Messina, Messina, Italy

Guido Gembillo and Domenico Santoro
Department of Medicine and Surgery, University of Messina, Messina, Italy, and

Massimo Villari
Department of Mathematics and Computer Sciences, Physical Sciences and Earth Sciences, University of Messina, Messina, Italy

Received 12 January 2026
Revised 13 May 2026
Accepted 25 June 2026

Abstract

Purpose – Digital transformation in health care is frequently delayed by resistance, which is typically conceptualized as a barrier to AI adoption. This study aims to reframe resistance as a knowledge-generating signal, examining how it shapes organizational learning and knowledge governance during AI implementation in high-stakes, regulated settings.

Design/methodology/approach – The study draws on a canonical action research program conducted within the eHealth Network: Artificial Intelligence and Innovative ICT Tools Oriented toward Digital Diagnostics project. Empirical material was collected across nephrology, hepatology and diabetology through 55 interdisciplinary meetings, 17 semistructured interviews and extensive project documentation, enabling longitudinal analysis of how knowledge is produced, negotiated and institutionalized during AI implementation.

Findings – The findings show that resistance operates as diagnostic signal, revealing misalignments across conceptual, representational and evidential knowledge proximities. These tensions trigger the development of boundary infrastructures, including shared vocabularies, standardized test datasets and interdisciplinary routines, through which knowledge is created, validated and governed across professional communities. The analysis identifies two socio-technical mechanisms that mediate these knowledge tensions: federated learning, which functions as a knowledge-governance capability balancing data sovereignty with institutional learning and explainable AI, which, when institutionalized through clinical explanation rounds, supports collective sensemaking and embeds interpretability into everyday decision-making routines.

Funding: This research was conducted within the framework of the RAIdD (eHealth Network: Artificial Intelligence and Innovative ICT Tools Oriented toward Digital Diagnostics) project, funded by the Italian Ministry of Health (CUP: E63C22001410001).

VINE Journal of Information and Knowledge Management Systems
Emerald Publishing Limited
2059-5891
DOI 10.1108/VJIKMS-01-2026-0010

Originality/value – The study contributes to the knowledge management and information systems literature by reconceptualizing resistance as a generative constraint for governing knowledge in AI-enabled organizations. It advances understanding of how socio-technical mechanisms can transform resistance into an asset for organizational learning under epistemic pluralism, offering insights applicable beyond health care to other regulated and knowledge-intensive domains.

Keywords Artificial intelligence, Knowledge governance, Federated learning, Explainable AI, Boundary infrastructures, Action research, Health-care organizations

Paper type Research paper

1. Introduction

Digital transformation in health care increasingly relies on AI systems, often justified by pressures of demand, resources and quality of care (Alowais *et al.*, 2023; National Academies of Sciences, 2018). Recent studies further highlight how emerging digital infrastructures, such as e-health platforms, virtual hospitals and blockchain-based systems, are reshaping health-care delivery models and organizational practices (Biancone *et al.*, 2021; Bidoli *et al.*, 2023; Massaro, 2021). Yet AI initiatives confront organizations with a knowledge-governance problem: coordinating heterogeneous expertise, evidence standards and accountability around algorithmic systems. Empirically, health-care AI rarely unfolds as linear adoption. Instead, it is shaped by enduring frictions among professional groups, regulatory constraints, epistemic disagreements and ethical concerns that remain embedded in everyday practice (Hinings *et al.*, 2018; Vial, 2019). Information systems research has long examined resistance as a barrier to implementation (Waddell and Sohal, 1998; Scott *et al.*, 2021), while more recent work treats resistance as ambivalent and constitutive of socio-technical change (Lapointe and Rivard, 2005; Dwivedi *et al.*, 2021; Schmidt *et al.*, 2024). However, resistance is still often theorized as an object to explain, rather than as an operational condition organizations must continuously organize through when governing AI in practice (Aryal *et al.*, 2023). This gap is acute in health care, where AI intersects with professionalized work, strict regulation and entrenched epistemic commitments (Fox and Connolly, 2018; Menard and Bott, 2024), as well as emerging ethical dilemmas related to AI-supported decision-making in clinical practice (Cobianchi *et al.*, 2022). We therefore examine how resistance becomes translated into concrete knowledge arrangements that render AI workable, legitimate and accountable over time as part of broader AI-enabled shifts in organizational knowledge creation and evaluation practices (Mukherjee *et al.*, 2025). We ask:

Q. How do organizations render AI governable under persistent resistance in regulated health-care settings and how do federated learning (FL) and explainable AI (XAI) contribute to this process?

Empirically, we draw on a canonical action research (CAR) program embedded in the eHealth Network: Artificial Intelligence and Innovative ICT Tools Oriented toward Digital Diagnostics (RAidD) project across nephrology, hepatology and diabetology (Baskerville and Wood-Harper, 1996; Davison *et al.*, 2004). We conceptualize FL as a knowledge-governance architecture that redistributes data ownership and accountability under privacy constraints and XAI as a situated practice through which interpretability and professional legitimacy are negotiated in clinical work. Our contributions are: repositioning resistance as an organizing condition that informs boundary infrastructures; theorizing FL and XAI as knowledge-governance devices in high-stakes settings; and showing how CAR supports the co-construction of AI governance over time.

The remainder of the paper is structured as follows. Section 2 develops the theoretical framing by introducing the concept of knowledge proximity gaps and positioning FL and XAI as socio-technical architectures for knowledge governance. Section 3 outlines the CAR

design and empirical setting. Section 4 presents the findings across three iterative cycles. Section 5 discusses theoretical and practical implications and Section 6 concludes.

2. Theoretical framing

2.1 Limits of technology adoption models in explaining artificial intelligence adoption in health care

Technology adoption research (e.g. TOE, TAM, UTAUT) has clarified how technological characteristics, organizational readiness and user perceptions shape initial engagement with digital innovations (Tornatzky and Fleischer, 1990; Davis *et al.*, 1989; Venkatesh *et al.*, 2003; Baker, 2011). In health care, these models help locate friction points in uptake (Yang *et al.*, 2022a, 2022b; Roppelt *et al.*, 2024). However, AI differs from incremental digital technologies because it introduces statistical inference into domains governed by professional judgment and accountability (Patel and Cohen, 2022). As a result, adoption depends less on readiness or acceptance and more on sustained coordination across epistemic communities that rely on different assumptions about evidence, responsibility and control (Pieper and Gleasure, 2025). In fragmented clinical systems, these differences persist over time and shape how AI is evaluated, justified and governed in practice (Fox and Connolly, 2018; Menard and Bott, 2024; Kumar *et al.*, 2021). Adoption models thus identify conditions for starting, but offer limited leverage for explaining how AI becomes governable under persistent misalignment (Malhotra and Hinings, 2015).

2.2 Resistance, epistemic misalignment and boundary infrastructures

Early work framed resistance as a barrier to be reduced or overcome (Lawrence, 1969; Kotter and Schlesinger, 1989; Scott *et al.*, 2021). More recent perspectives conceptualize resistance as multilevel and constitutive of socio-technical change, surfacing tensions that can enable more context-attuned solutions (Lapointe and Rivard, 2005; Dwivedi *et al.*, 2021; Schmidt *et al.*, 2024; Bartunek *et al.*, 2025). In health-care AI, resistance is often institutional (privacy, regulation and fragmentation), epistemic/professional (autonomy and "black-box" skepticism) and ethical/symbolic (fairness, bias and legitimacy) (Henry *et al.*, 2022; Wang *et al.*, 2023; Grosek *et al.*, 2024).

A socio-technical lens treats AI systems as enacted configurations in which social and technical elements coevolve through situated practice (Bostrom and Heinen, 1977; Orlikowski, 1992; Venkatraman *et al.*, 2022). The frictions observed in health care are therefore durable epistemic misalignments over what counts as valid knowledge and legitimate decision-making (Tikhomirov *et al.*, 2024; Sokol *et al.*, 2025; Adams, 2025).

When such misalignment is rendered visible through recurring frictions, it operates as a signal of deeper coordination challenges in how knowledge is framed, represented and evaluated across professional communities. To sharpen the analytical focus of this study, we conceptualize these recurrent misalignments in terms of knowledge proximity gaps.

We define a knowledge proximity gap as the structured distance between actors in how they frame the focal problem, represent it through data and models and evaluate what counts as valid and actionable evidence. Unlike generic notions of socio-technical friction or broad epistemic disagreement, knowledge proximity gaps refer specifically to divergences in evaluative regimes that bear directly on decision rights, accountability and the institutionalization of knowledge practices.

We distinguish three analytically separable dimensions. Conceptual proximity concerns the extent to which actors share problem framings and clinical categories. Representational proximity refers to the alignment between how phenomena are encoded in data sets, variables and computational features. Evidential proximity captures convergence or divergence in standards of proof, plausibility and acceptable risk when translating model outputs into clinical judgment.

These dimensions are related but not reducible to one another. Actors may converge conceptually while diverging representationally or accept statistical performance while questioning evidential sufficiency for accountable clinical use.

Knowledge proximity gaps are analytically distinct from boundary objects. Whereas boundary objects describe artifacts that facilitate translation across communities, proximity gaps denote the underlying epistemic distance that renders such artifacts necessary. In this sense, proximity gaps constitute a governance problem: they signal where institutionalized routines, shared standards or infrastructural arrangements are required to stabilize cross-professional coordination without presupposing epistemic convergence. Appendix 1 details the empirical indicators used to identify each dimension consistently across the three CAR cycles.

When resistance surfaces these proximity gaps, it provides actionable signals that inform the reconfiguration of socio-technical architectures.

Our focus is how resistance, as a signal of this gap, becomes operationalized through boundary infrastructures (routines, standards, shared evaluative artifacts) that enable ongoing knowledge governance. Table 1 synthesizes the main types of resistance identified in the literature and their manifestation within health-care settings.

2.3 Federated learning and explainable artificial intelligence as socio-technical architectures for knowledge governance

FL and XAI are often framed as technical solutions, but we treat them as complementary socio-technical architectures that reorganize knowledge, authority and accountability under persistent resistance (Antunes *et al.*, 2022; Loh *et al.*, 2022). FL addresses institutional resistance by enabling collaborative learning without centralizing sensitive data, thereby redistributing decision rights and responsibilities under privacy constraints (Ali *et al.*, 2024; Li *et al.*, 2025). XAI addresses epistemic/professional resistance by supporting interrogability and justification at the point of use; however, it becomes consequential only when institutionalized in routines that connect interpretation to accountable action (Markus *et al.*, 2020; Bienefeld *et al.*, 2023; Nasir *et al.*, 2024). This framing motivates our empirical

Table 1. Forms of resistance in health-care AI adoption

Type of resistance	Description	Typical problems	References
Institutional/Organizational	Related to regulatory constraints, governance, data fragmentation and medical specializations	Interoperability challenges, regulatory compliance issues, interspecialty conflicts	Esmaeilzadeh (2024); Schmidt *et al.* (2024)
Epistemic/professional	Connected to professional identities and clinicians' established practices	Fear of loss of autonomy, skepticism toward "black-box" systems, difficulties integrating into diagnostic processes	Henry *et al.* (2022); Wang *et al.* (2023)
Ethical/symbolic	Linked to values, trust and the social perception of technologies	Concerns about transparency, fairness, algorithmic bias and the impact on inequalities	Grosek *et al.* (2024)

Source(s): Authors' own work

focus on how FL and XAI are progressively shaped and stabilized as boundary infrastructures within a CAR program.

Figure 1 contrasts a conventional black-box machine learning pipeline with an explanation-aware model in a clinical diagnosis setting. Whereas black-box systems limit clinical scrutiny despite predictive accuracy, explanation-aware models integrate variables such as SHapley Additive exPlanations (SHAP)-based feature attributions, enabling explanations to be examined and discussed within structured clinical routines (Lundberg and Lee, 2017). This shift transforms interpretability from a static model property into an operational mechanism for transparency, plausibility assessment and model governance under continuous retraining.

3. Methods

This study adopts CAR to examine AI adoption as an ongoing process of organizing and knowledge coordination under conditions of persistent resistance. Action Research is characterized by iterative cycles of diagnosis, intervention and reflection that integrate inquiry and action through direct engagement with organizational practice (Lewin, 1946; Avison *et al.*, 1999; Davison *et al.*, 2004). Unlike purely observational approaches, AR emphasizes co-construction with practitioners, allowing theory and intervention to evolve in tandem.

We employed a canonical form of AR to ensure methodological rigor through a structured logic linking theory-informed diagnosis, purposeful intervention and systematic reflection (Baskerville and Wood-Harper, 1996). CAR explicitly defines stakeholder roles, responsibilities and learning objectives across cycles, involving clinicians from multiple specialties, IT professionals and academic researchers. This approach is particularly suited to health care, where prior research highlights its effectiveness in addressing epistemic, organizational and regulatory complexity through collaborative problem-solving (Cordeiro and Soares, 2018).

Methodologically, CAR enables resistance to be treated not only as an empirical phenomenon but as a source of insight guiding the design and refinement of socio-technical arrangements. Across iterative cycles, resistance informed the evolution of governance routines, coordination mechanisms and boundary infrastructures. Regular interdisciplinary meetings and semistructured interviews supported continuous reflexive engagement, enabling the articulation of epistemic tensions, collective sensemaking and the iterative redesign of interventions as challenges emerged. This combination of structured action cycles and sustained reflexivity supports the development of practice-grounded theoretical insights into how AI-related knowledge is governed and stabilized over time in a highly

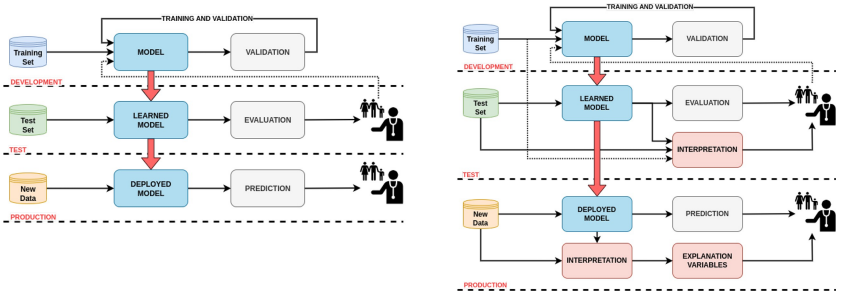

Figure 1. From black-box prediction to explanation-centered governance in clinical decision support

Source: Authors' own work

regulated health-care setting. Additional procedural and technical details are provided in Appendix 2.

3.1 Empirical context: the RAidD project

The empirical investigation is embedded in the RAidD project (eHealth Network: AI and innovative ICT tools oriented toward digital diagnostics), a large-scale, multiyear socio-technical initiative launched in December 2022 and ongoing at the time of writing (expected completion in 2027). The project aims to design and experiment with AI architectures for predictive and prognostic diagnostics while addressing organizational, epistemic and regulatory challenges across clinical domains. The study focuses on nephrology, hepatology and diabetology, three clinically complex and strategically significant domains that differ in diagnostic logics, data structures and work practices. This heterogeneity makes RAidD a suitable setting for examining knowledge coordination under conditions of epistemic plurality. The project involves an interdisciplinary configuration of actors, including clinical leaders from the three departments, research fellows with economics and statistics backgrounds and IT researchers responsible for machine learning and federated learning development.

Project activities are organized around weekly interdisciplinary meetings, which are systematically documented and serve as recurring arenas for coordination, negotiation and collective sensemaking. These meetings supported progress on clinical database construction while surfacing tensions related to data quality, model reproducibility and interpretability. In parallel, semistructured interviews with clinicians and IT researchers provided insight into situated concerns, professional judgments and perceptions of AI-related risks and opportunities. Project documentation and ethnographic field notes further complemented the empirical material.

Given its longitudinal, intervention-oriented and knowledge-intensive nature, RAidD is well suited to a CAR design (Baskerville and Wood-Harper, 1996; Davison *et al.*, 2004). The study is structured around three iterative cycles aligned with major project milestones: clinical database construction, prototype development and testing using synthetic data and validation of AI applications in real-world clinical settings.

3.2 Canonical action research cycles

Consistent with the CAR approach, the study was structured around three iterative cycles (Table 2), each following the recursive logic of diagnosis, planning, action and reflection and informed by the theoretical framework outlined in Section 2. Rather than linear stages, these cycles are treated as interconnected moments of inquiry through which resistance surfaced and shaped subsequent organizational and technical decisions.

First cycle (December 2022–September 2023) – diagnosis and problem codefinition: This phase focused on reconstructing existing practices across the three specialties and jointly defining baseline requirements for a multicenter clinical database. Resistance emerged primarily as epistemic and organizational misalignment around data definitions, inclusion criteria and variable comparability. Data were collected through interviews, document analysis and meetings, culminating in the collaborative design of an initial shared data model and the launch of data harmonization activities.

Second cycle (late 2023–2024) – development and early experimentation: This cycle centered on AI prototype development under continued constraints on real clinical data access. Synthetic data were introduced as an interim organizational solution to sustain experimentation while addressing privacy and regulatory concerns. Activities included codesign workshops, weekly monitoring meetings and interviews, with reflection focused on

Table 2. Action research cycles in the RAidD project

Cycle	Period	Main activities	Primary forms of resistance identified
1. Initial diagnosis and problem codefinition	December 2022–September 2023	Mapping of existing practices in nephrology, hepatology and diabetology; collective definition of database requirements; weekly recorded meetings; interviews with clinicians and IT researchers; analysis of clinical documentation and organizational protocols; collaborative design of the first data model; start of clinical data collection and harmonization; collective reflection sessions to consolidate a shared understanding of methodological and organizational challenges	Epistemic: misalignment in data definitions, clinical reasoning and comparability of variables across specialties; uncertainty over data ownership and responsibility
2. Development and early experimentation	Late 2023–2024	Planning and implementation of synthetic data for training and testing FL models; codesign workshops with clinicians and IT researchers; weekly monitoring meetings to evaluate reproducibility of results; individual interviews to capture perceptions and expectations; final reflection phase leading to redefinition of data collection strategies and preparation for real-world experimentation	Institutional: constraints related to data availability, privacy and regulation; skepticism toward model validity and generalizability; negotiation of acceptable performance metrics and validation criteria
3. Clinical validation	2025–2026 (ongoing)	Progressive integration of AI algorithms into clinical workflows; training sessions, interdisciplinary workshops and feedback activities; pilot tests in each department; collection of qualitative and quantitative data on impact and sustainability; joint reflection among clinicians, IT researchers and researchers to identify improvements and ensure coherence with ethical, institutional and professional requirements	Professional/ethical: concerns about interpretability, accountability, trust and the legitimacy of AI-supported clinical decisions at the point of use

Source(s): Authors own work

revising data assumptions and clarifying governance arrangements in preparation for clinical validation.

Third cycle (2025–2026) – clinical validation (ongoing): The ongoing third cycle is dedicated to validating AI prototypes in real-world clinical settings through gradual integration into clinical workflows. This phase involves training sessions, structured feedback activities, interdisciplinary workshops and pilot tests, supported by qualitative and quantitative data collection. Resistance becomes particularly salient at the point of use, where algorithmic outputs intersect with clinical judgment, accountability and ethical requirements.

Together, these cycles align RAidD project milestones with the CAR design, balancing methodological rigor and practical relevance (Avison *et al.*, 1999; Baskerville, 1999). They provide the empirical foundation for analyzing how resistance operates as a design input and how socio-technical architectures such as FL and XAI are progressively shaped through organizing work.

3.3 Data collection and analysis

Over the first two years, we assembled a diversified empirical corpus in line with methodological guidelines for CAR in information systems. Data collection and analysis were conducted iteratively and in parallel with the CAR cycles, allowing emerging insights to inform subsequent interventions and vice versa.

The empirical material includes 55 recorded and transcribed weekly project meetings (approximately 110 h), which served as the primary sites for observing how resistance, coordination challenges and epistemic disagreements emerged and were addressed during data construction, model development and validation. In addition, 17 semistructured interviews (45–60 min) were conducted with clinicians from the three specialties and members of the IT team, focusing on trust, responsibility, interpretability and professional judgment in relation to AI. The data set was complemented by approximately 20 project and organizational documents, including clinical protocols, meeting minutes, ethical assessments and technical reports.

Overall, meetings and interviews generated a corpus of approximately 450 pages of text, analyzed using a theory-informed, process-oriented coding strategy supported by NVivo (Maher *et al.*, 2018).

Analysis proceeded in three stages: open coding to identify recurring tensions and practices; axial coding to aggregate codes into higher-level categories of resistance (institutional, epistemic, professional and ethical) and corresponding organizing responses; and selective coding to trace how these categories evolved across CAR cycles and informed the institutionalization of FL and XAI practices.

Interpretive rigor was ensured through joint coding sessions, triangulation across data sources and researchers and member-checking during weekly meetings, where preliminary interpretations were discussed and refined with practitioners. Additional details on the CAR process, including the documentation of decision episodes and governance adjustments, are provided in Appendix 2.

4. Findings

4.1 First cycle – data model development (December 2022–September 2023)

The first CAR cycle generated a set of practical outputs that marked the starting point of the project while simultaneously revealing the epistemic and organizational conditions under which AI development could proceed, particularly exposing conceptual and representational knowledge proximity gaps in how clinical constructs were framed and encoded across

specialties. Central to this phase was the construction of a shared clinical data model, iteratively revised through three versions. Initial attempts exposed coordination breakdowns and persistent difficulties in aligning heterogeneous clinical logics within a single representational structure. To enable progress, the team provisionally adopted the diabetology data set as a reference point, as a temporary stabilization of knowledge that balanced feasibility, comparability and epistemic coherence while keeping the model open to future reconfiguration. Importantly, clinical reasoning styles remained distinct across specialties and the shared data model functioned as a negotiated representational compromise rather than a unified epistemic framework.

Across meetings and interviews, a pronounced knowledge proximity gap emerged between IT researchers and clinicians. IT researchers approached data as abstract features for model training, whereas clinicians interpreted data points within causal narratives and pathological processes. This divergence generated recurrent friction around variable selection, missing values and outlier interpretation, manifesting not as overt opposition but as persistent resistance that slowed decision-making and destabilized confidence in the emerging data model. An early organizational response to these tensions was the institutionalization of weekly interdisciplinary meetings, which evolved from *ad hoc* coordination events into a formal governance routine supporting collective sensemaking and alignment (Orlikowski, 2000). These meetings provided continuity, enabled the surfacing of assumptions and facilitated the negotiation of shared interpretations across professional groups. In parallel, a revised data model was approved together with standardized criteria for data cleaning and harmonization. Clinical leads also organized targeted training sessions for IT researchers, supported by shared materials, to foster a common clinical vocabulary and mutual understanding across specialties.

Taken together, the stabilization of the data model, the formalization of interdisciplinary routines and the introduction of internal training initiatives constitute the main outcomes of the first CAR cycle. More importantly, they illustrate how resistance functioned as a diagnostic signal, what we conceptualize as design telemetry, revealing misalignments in conceptual and representational knowledge and prompting the creation of boundary infrastructures that enabled continued collaboration. By the end of this cycle, the project had established a minimal but workable knowledge-governance arrangement, enabling the transition to AI prototype development.

4.2 Second cycle – development of artificial intelligence model prototypes with synthetic data and federated learning (Late 2023–2024)

During the second CAR cycle, the project shifted from problem framing to coordinated experimentation aimed at developing functional AI prototypes under persistent data and governance constraints. Proximity gaps increasingly surfaced along representational and evidential dimensions, particularly concerning the adequacy of synthetic data and the sufficiency of statistical performance metrics. The validated clinical data model was first reviewed and approved across nephrology, hepatology and diabetology through close collaboration with clinicians, who assessed variable adequacy, coding consistency and epistemic coherence rather than mere technical correctness. Given continued limitations in access to real clinical data, synthetic data were introduced as an organizational solution to sustain experimentation while complying with ethical and regulatory requirements. Data augmentation was implemented using conditional tabular generative adversarial networks (CTGAN) (Xu *et al.*, 2019), enabling the generation of clinically plausible records while preserving underlying data distributions. This approach decoupled early model development from direct reliance on sensitive patient data, thereby reducing institutional and epistemic

resistance. However, synthetic augmentation also introduced the risk of reinforcing representational biases embedded in limited local samples and the training synthetic – test real (TSTR) separation mitigated uncertainty regarding the fidelity of synthetic patterns under evolving clinical distributions.

A unified test set was constructed as a shared reference point for model evaluation, while model training followed an FL approach. Each department retained control over its local data set while contributing to the iterative training of a global model, supported by Ray.io for orchestration (Moritz *et al.*, 2018) and PyTorch for model optimization (Paszke *et al.*, 2019). From an organizational perspective, FL functioned as a boundary-spanning socio-technical architecture that redistributed data ownership, accountability and coordination across departments, addressing resistance related to data centralization and compliance (Carlile, 2002). Training proceeded through multiple federated rounds, with local models independently trained on augmented data sets and aggregated into a global model until jointly defined convergence criteria were met. Model performance was evaluated on the centralized test set using standard metrics (e.g. accuracy, area under the curve (AUC), F1-score), but results were systematically discussed within a multidisciplinary working group. These discussions acted as sensemaking practices through which statistical outputs were assessed against clinical plausibility and professional judgment. When discrepancies emerged, additional training cycles and refinements were initiated, with resistance manifesting as hesitation and reinterpretation rather than opposition. In this sense, FL redistributed accountability but did not remove the underlying evidential tension; it rendered it governable through structured coordination.

The outcome of this cycle was a set of AI models for chronic disease prediction across the three specialties. Beyond technical performance, these models reflected the negotiated and co-constructed nature of the project, illustrating how FL can support knowledge coordination and experimentation in complex health-care contexts through iterative CAR cycles. Figure 2 presents the visual representation of the FL workflow adopted in the RAidD project.

4.3 Third cycle: model explanation and clinical validation in real settings (2025–2026, ongoing)

The third CAR cycle focuses on the clinical validation of the predictive models developed in earlier phases, marking a shift from coordinated experimentation to situated use. In this phase, resistance becomes concentrated in the evidential dimension of knowledge proximity as clinicians questioned the legitimacy of statistical outputs and post-hoc explanations for accountable clinical actions. Models are evaluated on real patient data using a TSTR strategy, which allows clinical assessment while complying with regulatory and privacy constraints. This approach reflects a deliberate emphasis on methodological caution and gradual institutionalization rather than premature deployment. Interpretability plays a central role in this phase. SHAP decision plots are used to explain how individual features contribute to patient-level risk predictions, not as stand-alone diagnostic aids but as interpretive resources that enable clinicians to interrogate model plausibility and legitimacy. Yet SHAP explanations remained *post-hoc* approximations of model behavior, meaning that interpretability reduced opacity without fully resolving evidential uncertainty. Explanations thus support discussion around uncertainty, responsibility and acceptable use, rather than functioning as definitive justifications. To institutionalize this interpretive work, the project introduced clinical explanation rounds, structured interdisciplinary sessions that mirror hospital rounds while focusing on AI outputs. These rounds function as epistemic governance practices, enabling clinicians, data scientists and IT researchers to collectively interpret predictions, surface disagreements and identify epistemic concerns. Rather than

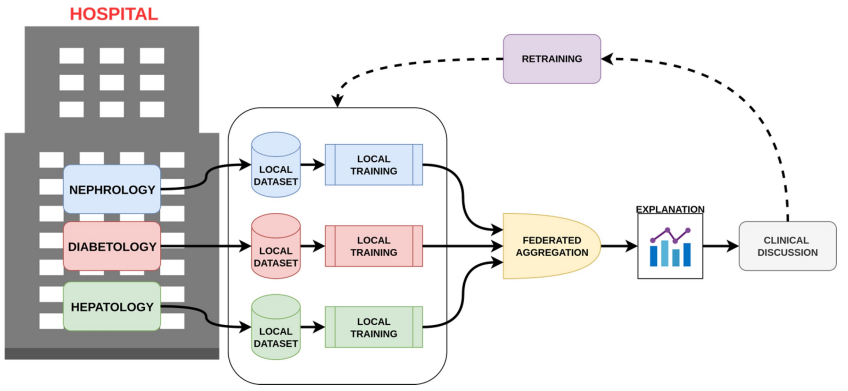

Figure 2. Federated learning architecture adopted in RAidD

Source: Authors' own work

resolving inconsistencies through technical thresholds alone, the rounds institutionalized residual ambiguity as a legitimate object of collective evaluation. At this stage, models remain in exploratory clinical validation and are explicitly not framed as decision-support tools ready for deployment, nor have they received formal ethical approval. This deliberate provisionality reflects the project's orientation toward governing AI through iterative alignment with clinical judgment, ethical standards and institutional requirements. Overall, the rounds provide a mechanism for continuous co-construction of AI systems with clinical expertise, embedding interpretability into everyday organizing practices. By linking explanation, collective sensemaking and iterative technical intervention, this phase illustrates how resistance at the point of use becomes a resource for epistemic alignment, accountability and the gradual legitimization of AI in health care. Figures 3 and 4 illustrate SHAP decision plots as interpretive artifacts used during this cycle. Table 3 summarizes how the three CAR cycles evolved through the canonical phases of diagnosis, planning, action, evaluation and reflection, highlighting the role of resistance in shaping successive organizational interventions.

5. Discussion

5.1 Theoretical contribution

Our study advances a view of AI adoption in health care as a coevolutionary, resistance-aware organizing process, rather than a linear trajectory of acceptance or implementation. Resistance is treated as an endogenous condition through which socio-technical systems are progressively shaped, with FL and XAI emerging as calibrated responses to persistent epistemic, organizational and regulatory misalignments (Brescia *et al.*, 2025; Orlikowski, 1992). This perspective complements dominant adoption models (TOE; TAM/UTAUT) by shifting attention from readiness and acceptance to the ongoing work of socio-technical realignment across heterogeneous epistemic communities. While the empirical evidence is grounded in the RAidD project, the propositions developed below should be read as theoretically informed inferences about AI governance under persistent epistemic misalignment, rather than as universally generalizable claims. Accordingly, we advance the following propositions.

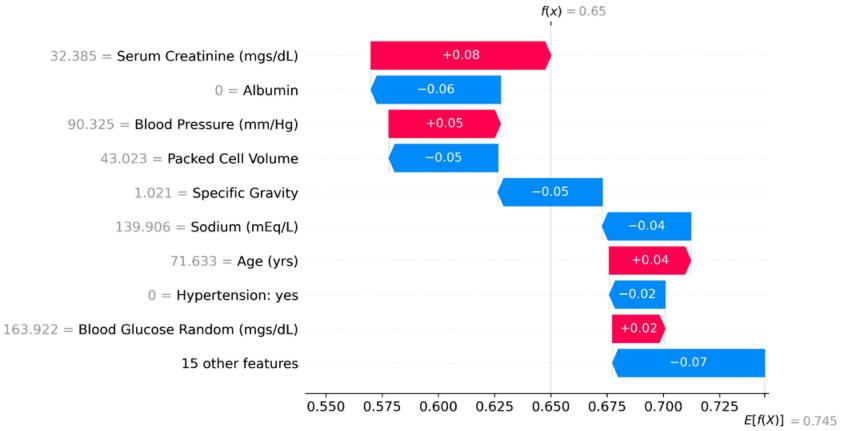

Figure 3. A SHAP decision plot was used during the third CAR cycle to support clinical interpretation of AI predictions. Each line represents the contribution of individual features to the final patient-level diagnostic prediction

Source: Authors' own work

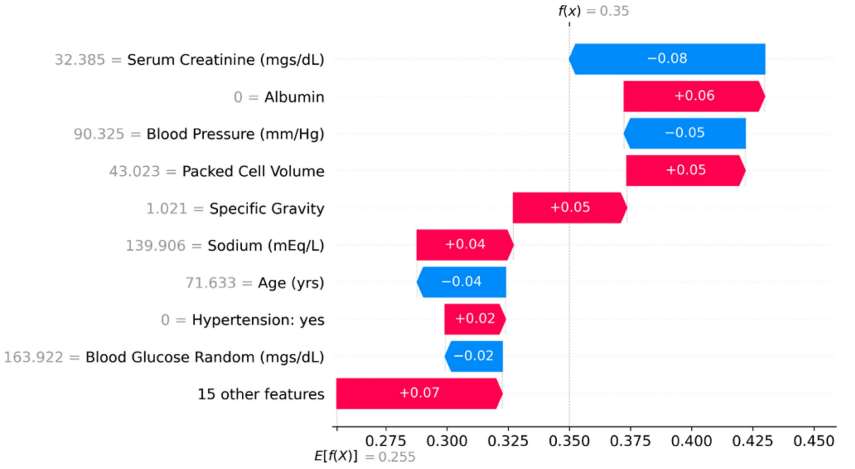

Figure 4. SHAP decision plot illustrating feature-level explanations discussed during clinical explanation rounds. Each line represents the contribution of individual features to the final patient-level diagnostic prediction

Source: Authors' own work

5.1.1 Reframing resistance: from obstacle to generative constraint. Where classical views conceptualize resistance as a problem to be contained or overcome (Lawrence, 1969; Zaltman and Lin, 1971; Kotter and Schlesinger, 1989) and much adoption research continues to frame it as a barrier to success (Scott *et al.*, 2021; Khizar *et al.*, 2025; Lyu *et al.*, 2025), our findings align with more recent work that treats resistance as a constitutive and ambivalent

Table 3. Canonical action research cycles

Cycle	Diagnosis	Planning	Action	Evaluation	Reflection
1) Data Model Development (December 2022–mid-2023) 20 weekly meetings; six interviews	Identification of missing data and communication gaps	Redesign of data model; adoption of diabetology dataset as reference	Iterative revision of model; weekly meetings; training sessions	Assessment of adequacy across departments; documentation via meetings/interviews	Recognition of knowledge proximity gap; decision to realign trajectory
2) AI Prototyping with Synthetic Data and FL (late 2023–2024) 22 weekly meetings; seven interviews	Recognition of limited data availability and cross-specialty heterogeneity	Design of CTGAN-based synthetic data strategy and FL framework	Creation of local datasets; federated training rounds; harmonization of test set	Technical evaluation (accuracy, AUC and F1) and clinical interpretation	Collective reflection through interdisciplinary discussions; retraining and feature refinements
3) Clinical validation and model explanation (2025–ongoing) 13 weekly meetings; four interviews	Identification of interpretability and validation challenges	Establishment of clinical explanation rounds with structured templates	Weekly sessions using SHAP plots; presentation of patient-level predictions	Collective review of variable contributions and clinical plausibility	Feedback loop enabling retraining, reweighting and ethical alignment

Source(s): Authors' own work

phenomenon (Lapointe and Rivard, 2005; Dwivedi *et al.*, 2021; Schmidt *et al.*, 2024). In the RAidD project, tensions around data governance, epistemic vocabularies and transparency (Wang *et al.*, 2023; Grosek *et al.*, 2024) were not eliminated but translated into durable boundary arrangements, including interdisciplinary routines, shared data standards, unified test sets and clinical explanation rounds. In this sense, resistance functioned as a design specification rather than an impediment, forcing explicit choices around data sovereignty, comparability, interpretability and accountability throughout the CAR cycles (Lewin, 1946; Davison *et al.*, 2004; Avison *et al.*, 1999; Baskerville and Wood-Harper, 1996). These specifications emerged where epistemic coordination broke down and can be understood as manifestations of a knowledge proximity gap between clinicians' causal reasoning and data scientists' probabilistic representations, spanning conceptual, representational and evidential dimensions. Accordingly, we propose:

P1. When resistance is surfaced and institutionalized through boundary infrastructures (e.g. routines, standards, shared vocabularies), it becomes a generative constraint that improves AI design quality and cross-professional alignment.

Unlike perspectives that conceptualize resistance primarily as negotiation (Lapointe and Rivard, 2005) or as an effect of socio-technical entanglement (Orlikowski, 2007), our findings emphasize resistance as a diagnostic mechanism that produces actionable design requirements, materialized in infrastructures that stabilize collaboration without requiring epistemic convergence.

5.1.2 Federated learning as a governance capability. In the RAidD project, FL was introduced as a response to persistent constraints related to data access, privacy and cross-specialty coordination. While privacy preservation is a primary motivation for FL In the RAidD project, Federated Learning was introduced as a response to persistent constraints related to data access, privacy and cross-specialty coordination. (Li *et al.*, 2025; Ali *et al.*, 2024), our findings show that FL operates more fundamentally as an organizational governance capability. FL enables organizations to balance local autonomy, rooted in departmental data sovereignty, professional accountability and domain-specific knowledge, with global integration through cross-specialty learning (Pesqueira *et al.*, 2025; Rauniyar *et al.*, 2022). In regulatory environments shaped by general data protection regulation and health insurance portability and accountability act, this balance is constitutive of trust, legitimacy and compliance (Zhang *et al.*, 2025; Madathil *et al.*, 2025). The presence of non-IID data typical of heterogeneous clinical settings rendered full data centralization both infeasible and undesirable (Efthymiadis *et al.*, 2024; Iyer, 2024). By preserving local semantic idiosyncrasies while enforcing common interfaces and aggregation rules, FL attenuated the representational dimension of the knowledge proximity gap, enabling partial alignment through shared computational representations (Appio *et al.*, 2023). Accordingly, we propose:

P2. When treated as a governance capability under strong privacy and regulatory constraints and in the presence of non-IID data heterogeneity, Federated Learning enables a more sustainable exploration–exploitation balance than centralized or strictly local approaches.

This perspective shifts the focus of FL research from technical performance alone to its role in mediating institutional and regulatory resistance by embedding decision rights, accountability and auditability directly into the learning architecture. Furthermore, this governance role also creates the conditions for controlled experimentation mechanisms, such as synthetic-data prototyping, discussed in Proposition 3.

5.1.3 Synthetic data as a conditional exploration mechanism within FL-based governance. In the RAidD project, synthetic data were introduced to sustain early experimentation under limited access to real clinical data. To address data scarcity, we adopted a TSTR strategy using CTGAN (Xu *et al.*, 2019), enabling early exploratory learning under strict regulatory constraints (Loftus *et al.*, 2022; Pati *et al.*, 2024). However, our findings highlight a specific epistemic risk: models trained on synthetic data may converge on artifacts of the data generator rather than clinically meaningful regularities. To mitigate this risk, we complemented standard performance metrics with clinician-approved real test sets and shared review routines, transforming recurrent disagreements into structured evaluative practices (Li *et al.*, 2025). The value of synthetic data thus lay not in epistemic neutrality but in controlled provisionality. Synthetic data supported rapid iteration only insofar as explicit retraining triggers and expert-led validity checks were institutionalized to detect drift and implausible patterns (Sturluson *et al.*, 2021). Accordingly, we propose:

P3. (conditional). Within FL-based governance arrangements, synthetic data supports early-stage exploration when paired with expert-led validity controls (e.g. shared real test sets, calibration and drift checks and documented review criteria); otherwise, offline gains risk masking false convergence and eroding epistemic trust.

Clinician resistance in this phase functioned as a validity checkpoint, prompting the institutionalization of guardrails that embedded synthetic data use within organizational routines. This proposition should be interpreted as conditional to Proposition 2, as synthetic data does not constitute an independent governance mechanism but operates within FL-enabled coordination structures.

5.1.4 From explainability tools to explanation-in-use. While FL primarily addresses coordination and governance across distributed data environments (Proposition 2), explainability operates at the point of use, where model outputs intersect with clinical judgment and accountability. In the RAidD project, clinical explanation rounds institutionalized explainability as a situated organizational practice rather than a static model property. During the third CAR cycle, clinical explanation rounds institutionalized explainability as a situated organizational practice rather than a static model property. SHAP decision plots (Figures 3 and 4) were collectively reviewed to connect model rationales with clinicians' experiential knowledge and disease mechanisms, enabling actionable decisions such as feature reweighting and retraining triggers (Lundberg and Lee, 2017; Yan *et al.*, 2025). Rather than adding post-hoc transparency, these rounds functioned as structured spaces that systematically linked interpretation to action, addressing clinicians' demands for accountable and trustworthy AI systems (Gastaldi *et al.*, 2018; Sadeghi *et al.*, 2024; Nasir *et al.*, 2024) and the persistent clinician–developer gap (Bienefeld *et al.*, 2023; Yang *et al.*, 2022a, 2022b). They also mitigated the limitations of conventional metrics that fail to translate into clinically meaningful reasoning without situated explanation (Tikhomirov *et al.*, 2024; Sokol *et al.*, 2025; Adams, 2025; Dlugatch *et al.*, 2024). Accordingly, we propose:

P4. XAI increases professional acceptability when explanations are embedded in recurring, multidisciplinary practices with clearly defined decision rights, systematic documentation and explicit retraining triggers that link interpretive work to concrete design actions within clinical governance routines.

This shifts XAI research from tools and interfaces toward explanation-in-use as a socio-technical governance practice through which resistance is transformed into accountability and model evolution.

The generative role of resistance observed in this study depends on the existence of structured routines capable of documenting and revisiting disagreement, rather than suppressing it. Similarly, FL and XAI function as boundary infrastructures only when decision authority and technical experimentation remain institutionally differentiated. Ultimately, while the mechanisms identified here may extend to other regulated knowledge-intensive domains, their applicability presupposes comparable professional accountability structures and governance sensitivity to evidential standards.

From a knowledge management perspective, our findings contribute by reconceptualizing how knowledge is generated, validated and stabilized in AI-enabled organizations. Rather than treating knowledge as an asset to be transferred or shared (e.g. Grant, 1996; Alavi and Leidner, 2001), we show how knowledge emerges through resistance-driven interactions across epistemically heterogeneous communities. In this process, knowledge proximity gaps identify where knowledge fails to travel, while boundary infrastructures represent organizational responses that enable coordination without requiring full epistemic convergence. Accordingly, AI governance can be understood as a form of knowledge orchestration, in which learning, validation and accountability are continuously negotiated through socio-technical arrangements.

5.2 Practical implications

For executives overseeing AI initiatives in health care, the central insight is that resistance constitutes actionable managerial information. These considerations are particularly relevant in light of ongoing transformations in health-care systems toward digitally enabled and distributed care models, such as e-health platforms and virtual hospitals (Biancone *et al.*, 2021; Bidoli *et al.*, 2023). Rather than neutralizing it, leaders should interpret resistance as design telemetry signaling where architectures, processes or evaluation criteria require revision. This requires rendering frictions visible and auditable and translating them into explicit specifications—shared data semantics, predictable interdisciplinary routines and agreed testing practices. Early effort should focus on a lean, evolvable boundary infrastructure (e.g. a core data dictionary, a shared clinical vocabulary and a common real-data test set) stabilized enough to support rapid iteration without imposing premature semantic uniformity. Three priorities follow:

(1) FL should be governed as an organizational choice, not treated as a purely technical topology. The leadership question is whether FL improves the balance between departmental data sovereignty and institutional learning under privacy and compliance constraints. Where it does, governance should codify aggregation rules, auditability, rollback procedures and convergence criteria and require evaluation against a shared, clinician-approved test set. Absent such governance, FL risks fragmenting into parallel silos.

(2) Synthetic data can accelerate learning only when its use is explicitly bounded in purpose and time. Enforcing TSTR separation, routine calibration and drift checks, and pre-authorized retraining triggers converts synthetic-data experimentation into controlled exploration rather than a proxy for real-world performance.

(3) Explainability creates value only when institutionalized as clinical governance, not when consumed as stand-alone visualization. Clinical Explanation Rounds should be formally scheduled and resourced, with SHAP plots anchoring discussion and outcomes recorded as traceable actions. Progress across stages should be gated by

alignment across three signals: technical adequacy, clinical plausibility, and observed impact in shadow use. Advancing without such alignment invites false convergence and erodes trust.

These priorities require clear roles and a minimal set of artifacts. Each specialty benefits from a Clinical Data Steward accountable for variable semantics, while an Explainability Facilitator ensures that interpretive insights translate into implemented changes. Oversight should track alignment and safety indicators alongside performance metrics, including harmonization of variables, implementation of explanation-round decisions, time from flagged issue to model change, calibration and drift on real data and site-level rollbacks. Complementing accuracy or AUC with these indicators renders AI initiatives governable in practice, enabling steady progress toward clinically plausible, technically robust and institutionally sustainable solutions.

6. Conclusions

This study examined AI adoption in health care as a coevolutionary, resistance-aware process. Across nephrology, hepatology and diabetology, we showed how institutional, organizational, epistemic and ethical forms of resistance do not simply hinder progress but actively shape it, directing architectural choices, routines and evaluation practices. Through an embedded program of inquiry, we demonstrated how FL) and XAI emerged as calibrated socio-technical responses, enabling cross-departmental learning without data centralization and transforming interpretability from a technical feature into a governance routine.

Empirically, three insights stand out. First, the knowledge proximity gap between clinical and computational reasoning was progressively translated into a boundary infrastructure—including shared data semantics, a common real-data test set and structured interdisciplinary routines—that stabilized collaboration while preserving local variation. Second, FL functioned primarily as a governance mechanism, balancing data sovereignty with cross-specialty learning under privacy and compliance constraints. Third, the combination of TSTR and clinical explanation rounds institutionalized validity and accountability: synthetic data accelerated early exploration, while SHAP-based discussions linked interpretability to concrete design actions such as feature reweighting, threshold adjustment and retraining decisions.

Theoretically, these findings extend technology-adoption research by reframing resistance as a generative design constraint and by providing process-level microfoundations for AI adoption in knowledge-intensive, regulated settings. Rather than treating resistance as a barrier to be overcome, the study shows how it operates as a diagnostic and organizing condition shaping socio-technical design and governance over time.

Managerially, the results point to a set of actionable governance choices: investing early in a lean but evolvable boundary infrastructure; treating FL as an organizational decision with explicit aggregation, audit and rollback policies; using synthetic data as a time-bounded accelerator governed by TSTR separation, calibration and drift checks and predefined retraining triggers; and institutionalizing explainability as a recurring clinical routine that produces traceable and accountable decisions. This study has limitations that open avenues for future research. Empirically, it focuses on three specialties within a single institutional context; future work could examine multi-institutional consortia to assess the scalability of federated governance arrangements. Substantively, extending the analysis to imaging or multimodal AI pipelines may surface different boundary conditions for FL, XAI and validity guardrails. Finally, longitudinal assessments of patient outcomes and workflow impacts would strengthen the link between organizational mechanisms and downstream clinical value.

Acknowledgements

The authors gratefully acknowledge the contribution of the clinicians, researchers, and technical staff involved in the project for their collaboration throughout the Canonical Action Research program.

References

Adams, J. (2025), "Ethical and epistemic implications of artificial intelligence in medicine: a stakeholder-based assessment", *AI and Society*, Vol. 40 No. 8, pp. 5935-5950, doi: 10.1007/s00146-025-02398-4.

Alavi, M. and Leidner, D.E. (2001), "Review: knowledge management and knowledge management systems: conceptual foundations and research issues", *MIS Quarterly*, Vol. 25 No. 1, pp. 107-136.

Ali, M.S., Ahsan, M.M., Tasnim, L., Afrin, S., Biswas, K., Hossain, M.M., Ahmed, M.M., Hashan, R., Islam, M.K. and Raman, S. (2024), "Federated learning in healthcare: model misconducts, security, challenges, applications, and future research directions – a systematic review", arXiv Preprint arXiv:2405, p. 13832, doi: 10.48550/ARXIV.2405.13832.

Alowais, S.A., Alghamdi, S.S., Alsuhebany, N., Alqahtani, T., Alshaya, A.I., Almohareb, S.N., … Albekairy, A.M. (2023), "Revolutionizing healthcare: the role of artificial intelligence in clinical practice", *BMC Medical Education*, Vol. 23 No. 1, p. 689.

Antunes, R.S., André da Costa, C., Küderle, A., Yari, I.A. and Eskofier, B. (2022), "Federated learning for healthcare: systematic review and architecture proposal", *ACM Transactions on Intelligent Systems and Technology (TIST)*, Vol. 13 No. 4, pp. 1-23.

Appio, F.P., La Torre, D., Lazzeri, F., Masri, H. and Schiavone, F. (2023), "Artificial intelligence: Technological advancements and methodologies", *in Impact of Artificial Intelligence in Business and Society*, Routledge, London, pp. 13-81.

Aryal, A., Truex, D. and El Amrani, R. (2023), "Lessons from enterprise systems competency centers in adopting digital transformation initiatives: an assemblage approach", *Information and Organization*, Vol. 33 No. 4, p. 100490.

Avison, D.E., Lau, F., Myers, M.D. and Nielsen, P.A. (1999), "Action research", *Communications of the ACM*, Vol. 42 No. 1, pp. 94-97.

Baker, J. (2011), "The technology–organization–environment framework", *Information Systems Theory: Explaining and Predicting Our Digital Society*, Springer, New York, NY, pp. 231-245.

Bartunek, J.M., Hansen, H., Cable, M.S. and Ianniello, A. (2025), "Generative resistance: the importance of enabling differences", *Organization Theory*, Vol. 6 No. 2, p. 26317877251351640.

Baskerville, R.L. (1999), "Investigating information systems with action research", *Communications of the Association for Information Systems*, Vol. 2 No. 3es.

Baskerville, R.L. and Wood-Harper, A.T. (1996), "A critical perspective on action research as a method for information systems research", *Journal of Information Technology*, Vol. 11 No. 3, pp. 235-246.

Biancone, P., Secinaro, S., Marseglia, R. and Calandra, D. (2021), "E-health for the future. Managerial perspectives using a multiple case study approach", *Technovation*, Vol. 120, p. 102406.

Bidoli, C., Pegoraro, V., Dal Mas, F., Bagnoli, C., Bert, F., Bonin, M., Butturini, G., Cobianchi, L., Cordiano, C., Minto, G. and Pilerci, C. (2023), "Virtual hospitals: the future of the healthcare system? An expert consensus", *Journal of Telemedicine and Telecare*, Vol. 31 No. 1, pp. 1-13.

Bienefeld, N., Boss, J.M., Lüthy, R., Brodbeck, D., Azzati, J., Blaser, M., Willms, J. and Keller, E. (2023), "Solving the explainable AI conundrum by bridging clinicians' needs and developers' goals", *Npj Digital Medicine*, Vol. 6 No. 1.

Bostrom, R.P. and Heinen, J.S. (1977), "MIS problems and failures: a socio-technical perspective. Part I: the causes", *MIS Quarterly*, Vol. 1 No. 3, pp. 17-32.

Brescia, V., Degregori, G. and Cavazza, A. (2025), "Artificial intelligence and knowledge management in healthcare: a pathway to SDGs achievement", *VINE Journal of Information and Knowledge Management Systems*.

Carlile, P.R. (2002), "A pragmatic view of knowledge and boundaries: boundary objects in new product development", *Organization Science*, Vol. 13 No. 4, pp. 442-455.

Cobianchi, L., Verde, J.M., Loftus, T.J., Piccolo, D., Dal Mas, F., Mascagni, P., Vazquez, A.G., Ansaloni, L., Marseglia, G.R., Massaro, M. and Gallix, B. (2022), "Artificial intelligence and surgery: ethical dilemmas and open issues", *Journal of the American College of Surgeons*, Vol. 235 No. 2, pp. 268-275.

Cordeiro, L. and Soares, C.B. (2018), "Action research in the healthcare field: a scoping review", *JBI Database of Systematic Reviews and Implementation Reports*, Vol. 16 No. 4, pp. 1003-1047.

Davis, F.D., Bagozzi, R.P. and Warshaw, P.R. (1989), "User acceptance of computer technology: a comparison of two theoretical models", *Management Science*, Vol. 35 No. 8, pp. 982-1003.

Davison, R.M., Martinsons, M.G. and Kock, N. (2004), "Principles of canonical action research", *Information Systems Journal*, Vol. 14 No. 1, pp. 65-86, doi: 10.1111/j.1365-2575.2004.00162.x.

Dlugatch, R., Georgieva, A. and Kerasidou, A. (2024), "AI-driven decision support systems and epistemic reliance: a qualitative study on obstetricians' and midwives' perspectives on integrating AI-driven CTG into clinical decision making", *BMC Medical Ethics*, Vol. 25 No. 1, p. 6.

Dwivedi, Y.K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V. and Williams, M.D. (2021), "Artificial intelligence (AI): multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy", *International Journal of Information Management*, Vol. 57, p. 101994.

Efthymiadis, F., Karras, A., Karras, C. and Sioutas, S. (2024), "Advanced optimization techniques for federated learning on non-IID data", *Future Internet*, Vol. 16 No. 10, pp. 370, doi: 10.3390/fi16100370.

Esmaeilzadeh, P. (2024), "Challenges and strategies for wide-scale artificial intelligence (AI) deployment in healthcare practices: a perspective for healthcare organizations", *Artificial Intelligence in Medicine*, Vol. 151, p. 102861.

Fox, G. and Connolly, R. (2018), "Mobile health technology adoption across generations: narrowing the digital divide", *Information Systems Journal*, Vol. 28 No. 6, pp. 995-1019.

Gastaldi, L., Appio, F.P., Corso, M. and Pistorio, A. (2018), "Managing the exploration-exploitation paradox in healthcare: three complementary paths to leverage on the digital transformation", *Business Process Management Journal*, Vol. 24 No. 5, pp. 1200-1234.

Grant, R.M. (1996), "Toward a knowledge-based theory of the firm", *Strategic Management Journal*, Vol. 17 No. S2, pp. 109-122.

Grosek, Š., Štivić, S., Borovečki, A., Ćurković, M., Lajovic, J., Marušić, A., Mijatović, A., Miksić, M., Mimica, S., Škrlep, E., Lah Tomulić, K. and Erčulj, V. (2024), "Ethical attitudes and perspectives of AI use in medicine between Croatian and Slovenian faculty members of school of medicine: cross-sectional study", *Plos One*, Vol. 19 No. 12, p. e0310599.

Henry, K.E., Kornfield, R., Sridharan, A., Linton, R.C., Groh, C., Wang, T., Wu, A., Mutlu, B. and Saria, S. (2022), "Human–machine teaming is key to AI adoption: clinicians' experiences with a deployed machine learning system", *Npj Digital Medicine*, Vol. 5 No. 1, p. 97.

Hinings, B., Gegenhuber, T. and Greenwood, R. (2018), "Digital innovation and transformation: an institutional perspective", *Information and Organization*, Vol. 28 No. 1, pp. 52-61.

Iyer, V.N. (2024), "A review on different techniques used to combat the non-IID and heterogeneous nature of data in FL", arXiv (Cornell University), doi: 10.48550/arxiv.2401.00809.

Khizar, H.M.U., Ashraf, A., Yuan, J. and Al-Waqfi, M. (2025), "Insights into ChatGPT adoption (or resistance) in research practices: the behavioral reasoning perspective", *Technological Forecasting and Social Change*, Vol. 215, p. 124047.

Kotter, J.P. and Schlesinger, L.A. (1989), "Choosing strategies for change", *in Readings in Strategic Management*, Macmillan Education UK, London, pp. 294-306.

Kumar, M., Singh, J.B., Chandwani, R. and Gupta, A. (2021), "Locating resistance to healthcare information technology: a bourdieusian analysis of doctors' symbolic capital conservation", *Information Systems Journal*, Vol. 32 No. 2, pp. 377-413.

Lapointe, L. and Rivard, S. (2005), "A multilevel model of resistance to information technology implementation", *MIS Quarterly*, Vol. 29 No. 3, pp. 461-491.

Lawrence, P.R. (1969), "How to deal with resistance to change", *Harvard Business Review.*

Lewin, K. (1946), "Action research and minority problems", *Journal of Social Issues*, Vol. 2 No. 4, pp. 34-46, doi: 10.1111/j.1540-4560.1946.tb02295.x.

Li, M., Xu, P., Hu, J., Tang, Z. and Yang, G. (2025), "From challenges and pitfalls to recommendations and opportunities: implementing federated learning in healthcare", *Medical Image Analysis*, Vol. 101, p. 103497, doi: 10.1016/j.media.2025.103497.

Loftus, T.J., Ruppert, M.M., Shickel, B., Ozrazgat-Baslanti, T., Balch, J.A., Efron, P.A., Upchurch, G.R., Rashidi, P., Tignanelli, C.J., Bian, J. and Bihorac, A. (2022), "Federated learning for preserving data privacy in collaborative healthcare research", *Digital Health*, Vol. 8, doi: 10.1177/20552076221134455.

Loh, H.W., Ooi, C.P., Seoni, S., Barua, P.D., Molinari, F. and Acharya, U.R. (2022), "Application of explainable artificial intelligence for healthcare: a systematic review of the last decade (2011–2022)", *Computer Methods and Programs in Biomedicine*, Vol. 226, p. 107161.

Lundberg, S.M. and Lee, S.-I. (2017), "A unified approach to interpreting model predictions", *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 30, pp. 4765-4774.

Lyu, T., Huang, K. and Chen, H. (2025), "Exploring the impact of technology readiness and innovation resistance on user adoption of autonomous delivery vehicles", *International Journal of Human–Computer Interaction*, Vol. 41 No. 12, pp. 7663-7683.

Madathil, N.T., Dankar, F.K., Gergely, M., Belkacem, A.N. and Alrabaee, S. (2025), "Revolutionizing healthcare data analytics with federated learning: a comprehensive survey of applications, systems, and future directions", *Computational and Structural Biotechnology Journal*, Vol. 28, p. 217, doi: 10.1016/j.csbj.2025.06.009.

Maher, C., Hadfield, M., Hutchings, M. and De Eyto, A. (2018), "Ensuring rigor in qualitative data analysis: a design research approach to coding combining NVivo with traditional material methods", *International Journal of Qualitative Methods*, Vol. 17 No. 1, p. 1609406918786362.

Malhotra, N. and Hinings, C.B. (2015), "Unpacking continuity and change as a process of organizational transformation", *Long Range Planning*, Vol. 48 No. 1, pp. 1-22.

Markus, A.F., Kors, J.A. and Rijnbeek, P.R. (2020), "The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies", *Journal of Biomedical Informatics*, Vol. 113, p. 103655.

Massaro, M. (2021), "Digital transformation in the healthcare sector through blockchain technology. Insights from academic research and business developments", *Technovation*, Vol. 120, p. 102386.

Menard, P. and Bott, G.J. (2024), "Artificial intelligence misuse and concern for information privacy: new construct validation and future directions", *Information Systems Journal*, Vol. 35 No. 1, pp. 322-367.

Moritz, P., Nishihara, R., Wang, S., Tumanov, A., Liaw, R., Liang, E., Elibol, M., Yang, Z., Paul, W., Jordan, M.I. and Stoica, I. (2018), "Ray: a distributed framework for emerging AI applications",

Proceedings of the 13th USENIX Symposium on Operating Systems Design and Implementation (OSDI '18), 561-577. *USENIX Association, Berkeley, CA.*

Mukherjee, S., Baral, M.M., Venkataiah, C., Wider, W. and Fauzi, M.A. (2025), "A review of the impact of generative AI in education using context, intervention, mechanism and outcome logic", *VINE Journal of Information and Knowledge Management Systems*, Vol. 56 No. 1.

Nasir, S., Khan, R.A. and Bai, S. (2024), "Ethical framework for harnessing the power of AI in healthcare and beyond", *IEEE Access*, Vol. 12, p. 31014, doi: 10.1109/access.2024.3369912.

National Academies of Sciences (2018), Medicine Division, Board on Global Health, and Committee on Improving the Quality of Health Care Globally, *Crossing the Global Quality Chasm: Improving Health Care Worldwide*, National Academies Press, Washington, DC.

Orlikowski, W.J. (1992), "The duality of technology: rethinking the concept of technology in organizations", *Organization Science*, Vol. 3 No. 3, pp. 398-427.

Orlikowski, W.J. (2000), "Using technology and constituting structures: a practice lens for studying technology in organizations", *Organization Science*, Vol. 11 No. 4, pp. 404-428.

Orlikowski, W.J. (2007), "Sociomaterial practices: exploring technology at work", *Organization Studies*, Vol. 28 No. 9, pp. 1435-1448.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A. and Chintala, S. (2019), "PyTorch: an imperative style, high-performance deep learning library", *Advances in Neural Information Processing Systems (NeurIPS 2019)*, Vol. 32, pp. 8024-8035.

Patel, V.L. and Cohen, T.A. (2022), "Clinical cognition and AI: from emulation to symbiosis", *in Intelligent Systems in Medicine and Health: The Role of AI*, Springer International Publishing, Cham, pp. 109-133.

Pati, S., Kumar, S., Varma, A., Edwards, B., Lu, C., Qu, L., Wang, J.J., Lakshminarayanan, A., Wang, S., Sheller, M., Chang, K., Singh, P., Rubin, D.L., Kalpathy-Cramer, J. and Bakas, S. (2024), "Privacy preservation for federated learning in health care", *Patterns*, Vol. 5 No. 7, p. 100974, doi: 10.1016/j.patter.2024.100974.

Pesqueira, A., Sousa, M.J. and Pereira, R. (2025), "Leveraging AI decentralization for sustainable and ESG-compliant health-care supply chains: insights from strategic leadership and dynamic capabilities", *VINE Journal of Information and Knowledge Management Systems*, pp. 1-28.

Pieper, M. and Gleasure, R. (2025), "How AI helps to compile human intelligence: an empirical study of emerging augmented intelligence for medical image scanning", *Information Systems Journal*, Vol. 35 No. 5, pp. 1399-1421. doi: 10.1111/isj.12585.

Rauniyar, A., Hagos, D.H., Jha, D., Håkegård, J.E., Bagci, U., Rawat, D.B. and Vlassov, V. (2022), Federated Learning for Medical Applications: A Taxonomy, Current Trends, Challenges, and Future Research Directions, Vol. 11 No. 5, pp. 7374-7398, doi: 10.48550/ARXIV.2208.03392.

Roppelt, J.S., Kanbach, D.K. and Kraus, S. (2024), "Artificial intelligence in healthcare institutions: a systematic literature review on influencing factors", *Technology in Society*, Vol. 76, p. 102443.

Sadeghi, Z., Alizadehsani, R., Çifçi, M.A., Kausar, S., Rehman, R., Mahanta, P., Bora, P.K., Almasri, A., Alkhawaldeh, R.S., Hussain, S., Alataş, B., Shoeibi, A., Moosaei, H., Hladík, M., Nahavandi, S. and Pardalos, P.M. (2024), "A review of explainable artificial intelligence in healthcare", *Computers and Electrical Engineering*, Vol. 118, p. 109370, doi: 10.1016/j.compeleceng.2024.109370.

Schmidt, J., Schutte, N.M., Buttigieg, S., Novillo-Ortiz, D., Sutherland, E., Anderson, M., … Van Kessel, R. (2024), "Mapping the regulatory landscape for artificial intelligence in health within the European Union", *Npj Digital Medicine*, Vol. 7 No. 1, p. 229.

Scott, I.A., Carter, S.M. and Coiera, E. (2021), "Exploring stakeholder attitudes towards AI in clinical practice", *BMJ Health and Care Informatics*, Vol. 28 No. 1, p. e100300.

Sokol, K., Fackler, J. and Vogt, J.E. (2025), "Artificial intelligence should genuinely support clinical reasoning and decision making to bridge the translational gap", *Npj Digital Medicine*, Vol. 8 No. 1, p. 345.

Sturluson, S.P., Trew, S., Muñoz-González, L., Grama, M., Passerat-Palmbach, J., Rueckert, D. and Alansary, A. (2021), "FedRAD: federated robust adaptive distillation", arXiv (Cornell University), doi: 10.48550/arxiv.2112.01405.

Tikhomirov, L., Semmler, C., McCradden, M., Searston, R., Ghassemi, M. and Oakden-Rayner, L. (2024), "Medical artificial intelligence for clinicians: the lost cognitive perspective", *The Lancet Digital Health*, Vol. 6 No. 8, pp. e589-e594.

Tornatzky, L.G. and Fleischer, M. (1990), *The Processes of Technological Innovation*, Lexington Books, Lexington, MA.

Venkatesh, V., Morris, M.G., Davis, G.B. and Davis, F.D. (2003), "User acceptance of information technology: toward a unified view", *MIS Quarterly*, Vol. 27 No. 3, pp. 425-478.

Venkatraman, S., Sundarraj, R.P. and Seethamraju, R. (2022), "Exploring health-analytics adoption in indian private healthcare organizations: an institutional-theoretic perspective", *Information and Organization*, Vol. 32 No. 3, p. 100430.

Vial, G. (2019), "Understanding digital transformation: a review and a research agenda", *Managing Digital Transformation*, pp. 13-66.

Waddell, D. and Sohal, A.S. (1998), "Resistance: a constructive tool for change management", *Management Decision*, Vol. 36 No. 8, pp. 543-548.

Wang, C., Liu, S., Yang, H., Guo, J., Wu, Y. and Liu, J. (2023), "Ethical considerations of using ChatGPT in health care", *Journal of Medical Internet Research*, Vol. 25, p. e48009.

Xu, L., Skoularidou, M., Cuesta-Infante, A. and Veeramachaneni, K. (2019), "Modeling tabular data using conditional GAN", *Advances in Neural Information Processing Systems (NeurIPS 2019)*, Vol. 32, pp. 7333-7343.

Yan, J., Husted, K. and Fath, B. (2025), "Transforming organizational knowledge creation through artificial intelligence: a systematic review of the emergent literature", *VINE Journal of Information and Knowledge Management Systems*, Vol. 56 No. 2, pp. 1-19.

Yang, G., Ye, Q. and Xia, J. (2022a), "Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: a mini-review, two showcases and beyond", *Information Fusion*, Vol. 77, pp. 29-52.

Yang, J., Luo, B., Zhao, C. and Zhang, H. (2022b), "Artificial intelligence healthcare service resources adoption by medical institutions based on TOE framework", *Digital Health*, Vol. 8, p. 20552076221126034.

Zaltman, G. and Lin, N. (1971), "On the nature of innovations", *American Behavioral Scientist*, Vol. 14 No. 5, pp. 651-673.

Zhang, F., Zhai, D., Bai, G.L., Jiang, J., Ye, Q., Ji, X. and Liu, X. (2025), "Towards fairness-aware and privacy-preserving enhanced collaborative learning for healthcare", *Nature Communications*, Vol. 16 No. 1, doi: 10.1038/s41467-025-58055-3.

Appendix 1 - Construct operationalization and experimental evidence

A1. Operationalization of knowledge proximity gaps across car cycles

To ensure consistent application of the knowledge proximity gap construct across the three canonical action research cycles, we developed a set of diagnostic criteria used during coding and iterative analysis. Rather than treating proximity gaps as generic disagreement, we identified them when divergences concerned problem framing, data representation or standards of evidential validity in ways that affected decision-making and governance arrangements. Table A1 summarizes the analytical dimensions and their empirical indicators.

Table A1. Forms of resistance in health-care AI adoption

Dimension	Diagnostic question	Empirical indicators in meetings/interviews	Typical governance response
Conceptual proximity	Are actors framing the same clinical problem using comparable categories and objectives?	Disagreement over outcome definitions; competing interpretations of diagnostic goals; divergent understandings of what the model is optimizing	Shared glossaries; cross-specialty clarification sessions; alignment workshops
Representational proximity	Are clinical phenomena encoded and structured in comparable ways across datasets and models?	Disputes over variable inclusion/exclusion; inconsistencies in coding practices; concerns over missing values or cross-specialty comparability	Standardized data dictionaries; harmonization protocols; unified test datasets
Evidential proximity	Do actors converge on what counts as sufficient and legitimate evidence for action?	Skepticism toward statistical metrics (e.g. AUC) without clinical plausibility; demands for interpretability; requests for validation pathways	Clinical explanation rounds; agreed performance thresholds; retraining triggers and documentation routines

Source(s): Authors' own work

During data analysis, these dimensions were applied iteratively in NVivo coding. Episodes were coded as proximity gaps when disagreements concerned evaluative standards rather than mere logistical coordination. Coding was refined across cycles through joint researcher discussions and member-checking during interdisciplinary meetings, ensuring that dimensions remained analytically distinct and consistently applied.

A2. Procedural experimentation: common data model stabilization through epistemic friction (canonical action research Cycle 1)

To empirically ground what we reported in Section 4.1, we show the experimental procedure aimed at assessing the stabilization of a common data model (CDM) through the systematic observation of epistemic friction. Embedded within the first canonical action research (CAR) cycle, the experiment treated data modeling as an organizational intervention rather than a preliminary technical step, with the objective of rendering resistance observable, traceable and actionable.

The experiment took as input heterogeneous clinical data sets originating from nephrology, hepatology and diabetology, each characterized by locally defined schemas and coding practices. An initial, deliberately underspecified CDM (version 0.1) was introduced to explicit misalignments across specialties. The procedural protocol consisted of weekly interdisciplinary data rounds in which clinicians and technical researchers jointly reviewed a restricted subset of variables. Each session produced a structured decision log documenting variable-level disagreements and resolutions and all modifications to the CDM were versioned and recorded to ensure traceability.

Evaluation relied on a reduced set of essential indicators capturing both procedural stabilization and epistemic alignment. First, epistemic friction was operationalized as a *Normalized Issue Rate*:

$$F_t^{norm} = \frac{|I_t|}{|V|}$$

where I_t denotes the set of unresolved issues at meeting t and V the set of CDM variables. This measure allowed us to track the evolution of semantic disputes independently of model size. Second, stabilization was assessed through time-to-consensus:

$$TTC(v) = t_s(v) - t_0(v)$$

defined as the number of meetings required for a variable v. to reach an accepted and unchanged definition, where $t_s(v)$and $t_0(v)$indicates the consensus and zero time for the feature in exam. Third, the operational applicability of the CDM was measured via the harmonization success rate, defined as:

$$H_r = \frac{1}{|S|}\sum_{s}\frac{|v \in V : viscorrectlymappedwithinS|}{|V|}$$

where *S* denotes the set of participating specialties.

Analysis of these indicators across successive CDM versions revealed a clear convergence pattern. Both the normalized issue rate and the average time-to-consensus decreased as data rounds became institutionalized, while harmonization success increased between early and stabilized CDM iterations. Importantly, disagreement was not eliminated but transformed into a structured design signal, enabling coordinated resolution without requiring full convergence of clinical reasoning styles. Figures A1–A3 illustrate the evolution of the normalized issue rate, time-to-consensus and harmonization success across the first CAR cycle.

The *outputs* of the experiment include a stabilized CDM (version 1.0), a shared data dictionary with explicit ownership and a documented audit trail linking epistemic frictions to concrete modeling decisions. Together, these results empirically substantiate the claims advanced in Section 4.1 by demonstrating how resistance during data model development can be operationalized and leveraged as a generative constraint for constructing a shared and governable representational foundation.

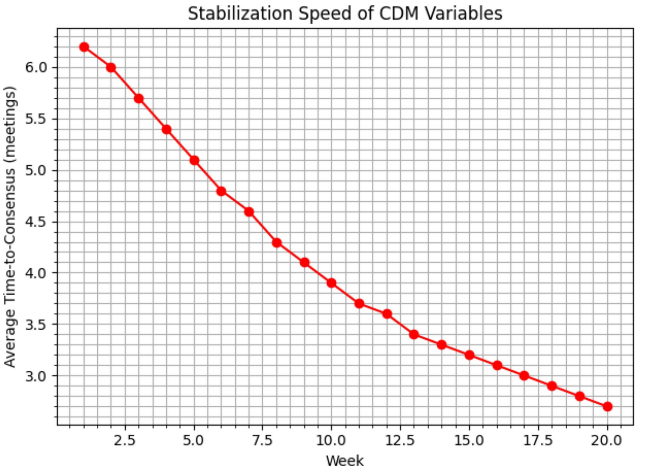

Figure A1. Epistemic friction over the first CAR cycle. Normalized issue rate showing the progressive reduction of unresolved semantic disagreements during common data model (CDM) stabilization across 20 weekly interdisciplinary meetings

Source: Authors' own work

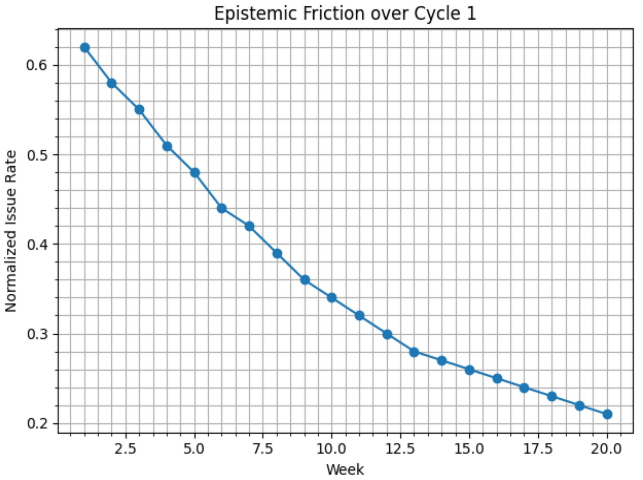

Figure A2. Stabilization speed of common data model (CDM) variables. Average time-to-consensus required to reach stable variable definitions across successive weekly meetings during the first CAR cycle

Source: Authors' own work

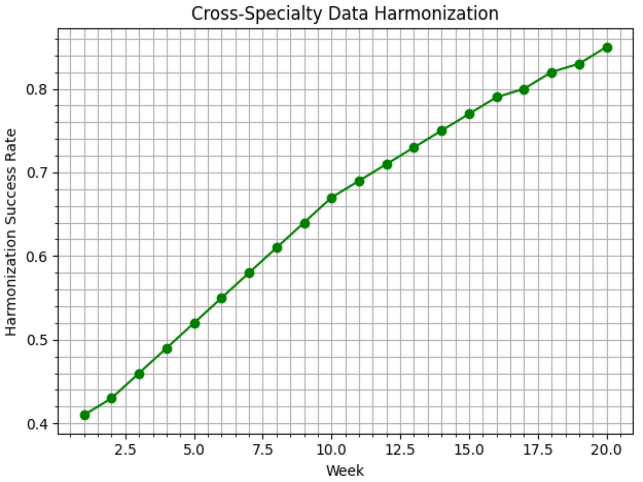

Figure A3. Cross-specialty data harmonization during the first CAR cycle. Harmonization success rate showing the progressive alignment of data structures across nephrology, hepatology and diabetology

Source: Authors' own work

A3. Procedural experimentation: federated learning with synthetic data augmentation under evolving clinical data sets (CAR Cycle 2)

To empirically support what we presented in Section 4.2, we show a set of experiments aimed at evaluating the role of FL as a governance mechanism under conditions of data heterogeneity, limited data availability and continuous data set evolution. The experimental setting involved three clinical specialties, nephrology, hepatology and diabetology, each maintaining a local data set, coherent with the shared common data model, subject to frequent changes due to patient turnover, with individuals entering and exiting the cohort over time. This dynamic context reflects realistic clinical data generation processes and precludes the assumption of a fixed or stationary training data set.

As an initial step, each specialty augmented its local data set using a conditional tabular GAN (CTGAN) to support early model development and mitigate data sparsity, particularly during the initial phases of the cycle. Synthetic data were used exclusively at the local level and never shared across sites, preserving data sovereignty while enabling local training to proceed before sufficient real-world samples accumulated. The augmented local data sets served as inputs to a federated training protocol based on the FedAvg algorithm, in which model updates were periodically aggregated to produce a shared global model without centralizing patient-level data.

Given the evolving nature of the underlying data sets, the federated model was not trained once and deployed statically but instead retrained on a biweekly schedule. Every two weeks, each site updated its local training set to reflect newly enrolled patients, discharged individuals and revised clinical records, after which a new round of federated training was initiated. This periodic retraining allowed the global model to adapt incrementally to distributional changes while maintaining continuity with prior model states. Model updates, aggregation rounds and performance metrics were logged systematically to ensure auditability and traceability across training cycles.

Evaluation focused on comparing the behavior of the federated model against locally trained baselines using a shared test set defined through the common data model introduced in Cycle 1. Performance metrics were analyzed across successive retraining rounds to assess stability under data drift, rather than peak accuracy under static conditions. The results indicate that federated aggregation enabled the construction of a common model that remained robust across specialties despite non-identically distributed and temporally evolving data, while CTGAN-based augmentation functioned as a provisional accelerator rather than a substitute for real clinical observations. Together, these experiments substantiate the claims of Section 4.2 by demonstrating how FL, combined with controlled synthetic data augmentation and periodic retraining, supports collaborative model development under realistic clinical and organizational constraints. Figures A4 and A5 summarize the similarity between synthetic and real-data distributions and the longitudinal TSTR performance of the federated model across successive retraining rounds.

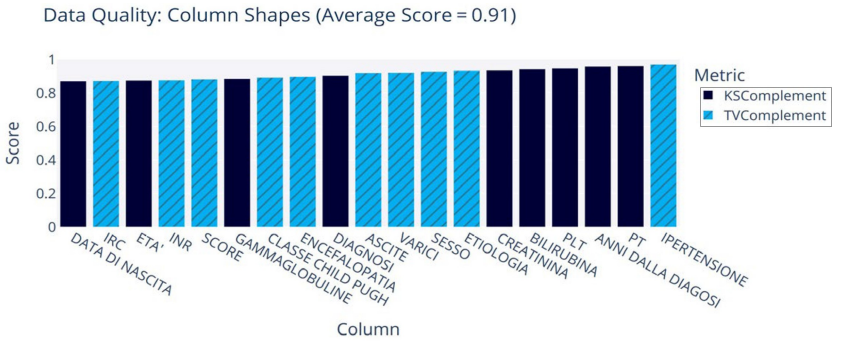

Figure A4. Similarity score of CTGAN-generated synthetic data across clinical variables. Average similarity between synthetic and real-data distributions used for local model training

Source: Authors' own work

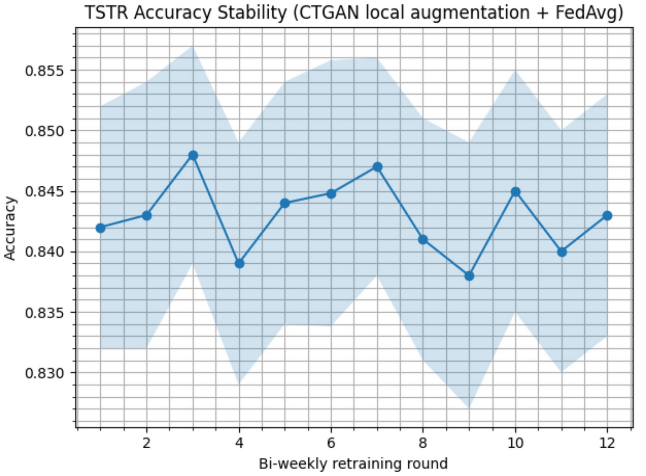

Figure A5. Biweekly TSTR accuracy of the federated model (FedAvg) trained with local CTGAN augmentation. Shaded areas indicate 95% confidence intervals computed on the real test set. The trajectory remains statistically stable across retraining rounds (permutation test on slope, $p > 0.05$)

Source: Authors' own work

A4. Procedural experimentation: TEST phase and case-based plausibility assessment with SHAP explanatory variables (CAR Cycle 3)

To empirically support what showed in Section 4.3, we annexed an experiment aimed at evaluating the consistency and clinical plausibility of model explanations under conditions of periodic retraining. Given that the federated model was updated biweekly to accommodate continuously evolving clinical data sets, the experiment assessed whether explanation artifacts, specifically SHAP-based explanations, remained stable, interpretable and actionable over time, rather than fluctuating arbitrarily across successive model versions.

A set of clinically representative *sentinel cases* was identified in collaboration with domain experts from nephrology, hepatology and diabetology. These cases reflected common and well-understood clinical profiles and were used as fixed reference points for explanation analysis across retraining rounds. For each sentinel case, SHAP explanations were computed after each federated update and evaluated along two complementary dimensions: explanation consistency and case-based clinical plausibility.

Explanation consistency was operationalized by assessing the stability of feature importance rankings across successive retraining rounds, with particular attention to a predefined *clinical core feature set* comprising serum creatinine, albumin, blood pressure, packed cell volume, specific gravity, sodium, age, hypertension status and random blood glucose. To quantify the degree to which explanations remained grounded in clinically expected variables, we introduced the Core Overlap@5 (Co.@5) indicator, measuring the proportion of the top-5 SHAP features belonging to this core set. As illustrated in Figure A6, Co.@5 remained consistently high and exhibited progressive stabilization across biweekly retraining rounds, indicating that periodic model updates did not induce arbitrary shifts in the explanatory structure.

In parallel, case-based plausibility was assessed during structured clinical explanation rounds, in which clinicians evaluated SHAP decision plots using a standardized template focusing on clinical coherence, directionality of effects and interpretability in relation to established pathophysiological reasoning. The Plausibility Acceptance Rate (PAR), defined as the proportion of explanations judged clinically coherent, is reported alongside Co.@5 in Figure A6. The close temporal alignment between the two indicators shows that explanations more strongly anchored to the clinical core were also more likely to be accepted as plausible and actionable by clinicians.

Importantly, explanations were not treated as passive artifacts. A non-trivial subset of reviewed cases triggered concrete governance actions, including targeted model review, feature inspection or documented acceptance under uncertainty. Instances of explanation instability were limited and predominantly associated with clinically ambiguous profiles, where they functioned as diagnostic signals prompting focused discussion rather than as evidence of systemic explanation failure.

Together, these results substantiate the claims of Section 4.3 by demonstrating that XAI effectiveness in this setting depends on the joint stability and clinical grounding of explanations over time. By linking explanation consistency (Co.@5) and clinician-rated plausibility (PAR) within a single temporal analysis (Figure A6), this experiment shows how SHAP-based explanations can be institutionalized as part of an ongoing governance practice rather than treated as static post-hoc outputs.

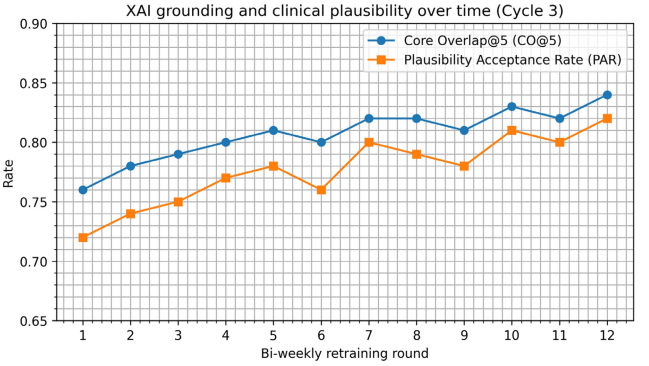

Figure A6. Temporal grounding and clinical plausibility of SHAP explanations under periodic retraining

Source: Authors' own work

Figure A6 shows the evolution of two explanation-level indicators across 12 biweekly retraining rounds of the federated model. Core Overlap@5 (CO@5) quantifies explanation grounding by measuring, for each round, the proportion of the top-5 features ranked by absolute SHAP value that belong to a predefined clinical core feature set (serum creatinine, albumin, blood pressure, packed cell volume, specific gravity, sodium, age, hypertension status and random blood glucose), aggregated across sentinel cases. PAR denotes the proportion of case-level SHAP explanations that were judged clinically coherent during structured clinical explanation rounds, based on expert assessment of feature relevance, directionality and pathophysiological plausibility. The concurrent stabilization of CO@5 and PAR indicates that explanations increasingly remain anchored to clinically expected variables and are correspondingly perceived as more plausible and actionable by clinicians. Values are aggregated across sentinel cases and reported to illustrate governance-relevant stability trends rather than individual explanation behavior.

Appendix 2 – analytical transparency and CAR process documentation

This study was conducted within a canonical action research (CAR) framework, where members of the research team participated in diagnostic discussions, facilitated interdisciplinary meetings and occasionally proposed procedural adjustments (e.g. data harmonization criteria, validation routines, retraining schedules). However, formal authority over clinical validation thresholds, data set inclusion and decisions regarding model testing and use remained with designated clinical leads. This separation of facilitation and decision rights was maintained throughout the three CAR cycles and key decisions and disagreements were systematically logged in meeting minutes and version-controlled documentation. To enhance analytical transparency, we explicitly distinguished between logistical coordination issues and evaluative disagreements. Episodes were coded as analytically relevant when tensions concerned standards of validity, accountability or acceptable abstraction rather than operational inconvenience. During axial and selective coding, competing interpretations were discussed among researchers, particularly when assessing whether observed tensions reflected temporary coordination challenges or deeper knowledge proximity gaps. These interpretations were revisited in subsequent interdisciplinary meetings to verify whether the analytical framing adequately captured practitioners' concerns. This iterative validation reduced the risk of retrospective rationalization.

Selected decision moments illustrate how disagreement informed governance arrangements rather than being resolved through implicit alignment. Selected decision moments illustrate how disagreement informed governance arrangements rather than being resolved through implicit alignment. Table A2 summarizes these episodes, specifying the nature of the disagreement, the *locus* of decision authority and the resulting governance adjustment.

For example, the provisional adoption of the diabetology data set as a reference model in *Cycle 1* followed sustained debate over cross-specialty comparability; the introduction of synthetic data under a training-synthetic/test-real (TSTR) separation in *Cycle 2* emerged from concerns about clinical plausibility and regulatory legitimacy; and the institutionalization of clinical explanation rounds in *Cycle 3* resulted from disagreements regarding whether statistical performance and post-hoc explanations were sufficient grounds for accountable clinical action. In each case, resistance generated explicit adjustments in procedures, documentation and validation routines, leaving a traceable governance outcome. Collectively, these measures enhance the transparency and replicability of the CAR process by making visible how documented tensions shaped the evolution of socio-technical arrangements across cycles.

These episodes illustrate how disagreement was documented and translated into explicit procedural adjustments rather than resolved through implicit alignment.

Table A2. Selected decision moments and governance adjustments across the three CAR cycles

CAR cycle	Decision episode	Nature of disagreement	Decision authority	Governance adjustment
Cycle 1	Adoption of reference dataset (diabetology)	Cross-specialty comparability and variable definitions	Clinical leads (after interdisciplinary debate)	Provisional shared CDM; formalized harmonization protocol
Cycle 2	Introduction of synthetic data under FL	Clinical plausibility and regulatory legitimacy	Technical proposal with formal clinical approval	TSTR separation; unified real test set; retraining schedule
Cycle 3	Institutionalization of clinical explanation rounds	Sufficiency of statistical metrics and SHAP explanations for accountable action	Clinical validation authority	Formalized explanation rounds; retraining triggers; documented review process

Source(s): Authors own work

Corresponding author
Giuseppe Lanfranchi can be contacted at: gilanfranchi@unime.it