Geographies of Trust: AI, Biomedicine, and the Next Era of Federated and Visiting Data Models

Authors: Patricia Buendia, Seonyoung Kim, Natalie Meyers, Francis P. Crawley, Gavin Farrell, Ronit Purian, EOSC-Future/RDA Artificial Intelligence Data Visitation Working Group

Published: 24 November 2025

Version: 1.1c DOI: 10.5281/zenodo.16929066

INTRODUCTION

Biomedical research increasingly relies on vast, sensitive datasets—from genomic sequences to clinical records. Yet the ethical, legal, and logistical challenges of moving or copying this data to centralized repositories are immense. In response, a diverse ecosystem of platforms has emerged, each offering a unique approach to **data protection**, **distributed analysis**, and **privacy-preserving computation**.

This report categorizes these platforms into three primary models: **Data Visitation (DV)**, **Centralized DV**, and **Federated**. There is also a Hybrid Centralized/Federated model to consider. This report examines how each model addresses the challenge of enabling research while maintaining data security.

Beyond current implementations, the report also explores the **evolving landscape of AI technologies**—particularly those involving distributed machine learning, synthetic data generation, and privacy-preserving computation—and assesses their potential to reshape biomedical and health research data access and analysis frameworks. By engaging with emerging trends, the report offers a forward-looking perspective on how these models must adapt to ensure ethical, secure, and scalable research in increasingly complex data ecosystems.

COMPARATIVE MODELS OF DATA PROTECTION

Three Models (and a Hybrid Model)

This section provides a visual comparison or decision guide to help choose the most appropriate technology for the sharing of access-restricted data.

- **Data Visitation (DV):** A decentralized model where tools visit locally governed data without requiring a central repository or uniform architecture.
- **Federated:** A distributed approach where data remains local but requires a uniform architecture and initial setup to enable tool-based access across sites.
- **Centralized Data Visiting:** A centralized model where data is pooled into a central repository, and researchers access it directly, necessitating uniform architecture and setup.

Table 1 outlines the key distinguishing features of the three models, while Table 2 categorizes their additional performance-related attributes into low, medium, and high tiers. A detailed description of each model, accompanied by examples, follows in the next section. A comprehensive comparative table of platforms is being maintained and curated here: Table Comparing Federated Data Platforms and Data Visiting Technologies and is discussed in Platform Comparison Table.

Table 1: Main Features of Data-In-Place Analytics Platforms

Feature	DV	Federated	Centralized DV
Data stays in place with local governance	✅	✅	❌
Researchers visit the data	❌	❌	✅
Uniform architecture required	❌	✅	✅
Central repository	❌	❌	✅
Tool/model visits data	✅	✅	❌
Initial data setup is required before DV	❌	✅	✅

Table 2: Categorization of Performance Features

Feature	DV	Federated	Centralized DV
Performance/latency	Medium (depends on network) [1,2]	Low to Medium (depends on architecture) [1,2]	High (local processing) [3]
Security/privacy risks	Low (data stays local) [4,5]	Medium (some exposure through federation) [4.5]	High (data centralized) [6]
Compliance with jurisdictional laws	High (data never leaves jurisdiction) [7,8]	Medium to High (depends on node compliance) [7,8]	Low to Medium (central repository may cross borders) [7,8]

Infrastructure cost distribution	Distributed (each site covers its own) [9,10]	Shared/Distributed [9,10]	Centralized (host bears the cost) [9,10]
Interoperability	Medium (requires compatible tools) [11,12,13]	High (uniform architecture) [11,12,13]	High (centralized control) [11,12,13]
Data freshness	High (real-time possible) [14,15]	Medium to High (depends on sync) [14, 15]	High (data centralized and updated centrally [14,15]

Centralized DV

Definition: Data and tools are maintained in a central location. Researchers access the data by visiting the secure environment where analysis tools are provided.

Key Features:

- Data remains in a secure enclave.
- Researchers access tools like Jupyter, RStudio, or custom dashboards.
- Strong governance and audit trails.

Examples:

- **All of Us Research Hub:** A centralized, secure cloud-based research platform hosting genomic, EHR, survey, and physical measurement data from diverse U.S. participants.
- **UK Biobank**: Offers centralized access to genomic and phenotypic data through secure portals. As of 2024, UK Biobank no longer allows researchers to download individual-level data. The data is visited and not moved.
- **NCATS N3C by Palantir**: COVID-19 data enclave with harmonized datasets and built-in analytics.
- **HDRUK TREs**: Trusted environments for UK health data, with strict access controls.
- **FAIR Square**: Centralized FAIRness assessment tools and metadata curation.

Strengths:

- High control and security within a single environment.
- Rich toolsets available in one place.
- High performance and low latency due to local processing.
- Strong interoperability through centralized architecture.
- Clear, standardized onboarding process for new data sources once pipeline is established.

Limitations:

- Requires data movement to the central location.
- May face cross-border data sharing and localization restrictions.
- Higher security/privacy risks if the central environment is compromised.
- Centralized infrastructure cost burden on the host organization.
- Scalability is limited by the capacity of the central infrastructure.
- Initial data setup and harmonization required before integration.

Data Visitation

Definition: Data can reside anywhere and be of any type. The technology or ML model visits the data, often without requiring it to be part of a formal network.

Key Features:

- Flexible architecture.
- Ideal for heterogeneous or sensitive datasets.
- Often used in early-stage pilots or ad hoc collaborations.

Examples:

- **FAIR Data Train**: Enables metadata-driven access and visitation across diverse datasets.
- **FAIRlyz**: Uses LLMs and AI tools to curate and analyze data remotely.
- **OSSDIP**: Focuses on secure access protocols and metadata registries.
- **Overture**: Metadata orchestration tools that support decentralized data discovery.

Strengths:

- Minimal data movement – data remains in its original location
- Adaptable to various data types, formats, and governance models.
- High data freshness due to real-time or near real-time access.
- Low secruity/privacy risk since data does not leave its host environment.
- Distributed infrastructure cost across participating sites.
- No mandatory initial setup—can access data without prior harmonization.

Limitations:

- May lack standardization or network cohesion.
- Tool interoperability can be challenging across heterogeneous sites.
- Variable performance depending on network quality.
- Interoperability depends on compatible tools at each site.
- Compliance management may be inconsistent across sites.
- Scalability for many new sources can be challenging without prior harmonization.

Federated

Definition: Data is standardized and stored in nodes within a network. Each node follows uniform architecture and SOPs. Technology or ML models visit the data, not the other way around.

Key Features:

- Strong standardization (e.g., OMOP, FHIR).
- Enables federated learning and analytics.
- Often used in large-scale consortia.

Examples:

- **DataSHIELD**: R-based statistical analysis across distributed nodes.
- **TriNetX/i2b2**: Real-time cohort discovery across hospital networks.
- **Swarm Learning Networks**: Federated AI model training across institutions.
- **ATLAS by OHDSI**: OMOP-based analytics across global nodes.
- **EUCAIM**: Imaging data federation with AI model training.
- **Genomic Data Infrastructure (GDI)**: Federated genomic analysis using GA4GH standards.
- **ELIXIR on Cloud**: Cross-site federation for life science data.
- **Federated EGA:** discovery and access of sensitive human omics and associated data consented for secondary use.

Strengths:

- High scalability - easily add new standardized nodes once standards are in place.
- A consistent onboarding process for new nodes ensures quality and compliance.
- Strong standardization (e.g., OMOP, FHIR) enables interoperability.
- Data sovereignty preserved at local sites.
- Enables federated learning and large-scale analytics.
- High potential for consistent compliance across nodes.
- Ideal for multi-institutional research.

Limitations:

- Requires significant upfront investment in harmonization.
- Governance complexity across multiple organizations and nodes.
- Performance may be impacted by inter-node latency.
- Medicum security/privacy risk due to networked access.
- Infrastructure costs are shared but require sustained investment.
- Initial data setup and alignment to network standards required before participation.

Hybrid Centralized/Federated

Definition: Data storage is centralized, but governance and access protocols are federated. It's a hybrid model—not true federation in architecture, but federated in policy.

Key Features:

- Centralized infrastructure with federated access logic.
- Often used in national or multi-agency initiatives.

Examples:

- **NCPI AnVIL**: Centralized genomic data with federated governance across NIH institutes.
- **Lifebit** and **DARE UK**: Federated TREs with compute environments

Strengths:

- Balances control with collaboration.
- Easier to manage infrastructure.

Limitations:

- Still involves centralized storage.
- May not meet strict data localization requirements.

PLATFORM COMPARISON TABLE

This dynamic comparative table catalogs platforms that support **DV**, **Federated**, **Hybrid Centralized/Federated**, and **Centralized DV** models in biomedical and health research. It is actively maintained to reflect the latest developments and platform capabilities. The latest version can be accessed here: Table Comparing Federated Data Platforms and Data Visiting Technologies.

These platforms represent a diverse and evolving ecosystem designed to enable secure, privacy-preserving analysis of sensitive data across institutions, without unnecessary duplication or centralization.

Each entry includes key attributes such as:

- Data visitation type and architecture
- Software availability and tools
- Data models and standards used (e.g., OMOP, FHIR, GA4GH)
- Stakeholders served (e.g., researchers, hospitals, pharma)
- Outputs generated (e.g., statistical models, cohort insights)
- Whether tools/models are moved to the data or expect standardized inputs

This table is intended to support strategic planning, platform selection, and policy development for organizations seeking to collaborate across data boundaries while maintaining compliance and trust.

NEW AI TECHNOLOGIES

There's a wave of emerging technologies reshaping how federated platforms and data visitation systems operate. These innovations go beyond GenAI and LLMs, though many intersect with them. Below is a curated list of cutting-edge technologies that are either being actively explored or show strong potential to impact these domains.

Emerging Technologies Impacting Data Visitation & Federated Platforms

1. Agentic AI

- Combines LLMs with autonomous planning and execution capabilities.
- Enables "virtual coworkers" that can visit data sources, perform multi-step tasks, and return results.
- Useful for federated workflows where human-in-the-loop is limited but agents are human-supervised.

Key Resources

1. Agentic Artificial Intelligence to Support Autonomous Medical Operations
   NASA's Human Research Program (HRP) is leveraging the power and efficiency of artificial intelligence (AI) tools to reduce human system risk in space medicine operations.[16]
2. Agentic AI in Biomedical Research: A New Era of Intelligent Collaboration
   ICA.ai offers AI solution for biomedical research and public health. [17]
3. Sample Agents for Healthcare and Life Sciences on AWS
   One-click deployment of infrastructure and Interactive interface for human-agent chat with pre-built agents.[18]

2. Secure Multiparty Computation (SMPC)

- SMPC allows multiple parties to jointly compute a function over their inputs while keeping those inputs private.
- Ideal for federated analytics in finance, genomics, and healthcare.
- Often paired with federated learning or data visitation models.

Key Resources

1. Secure Multi-Party Computation – SpringerLink
   Use cases discussed: secure machine learning and privacy-preserving network monitoring. [19]
2. SMPC for Privacy-Preserving Data Analysis – IJCRT
   Explores SMPC's role in healthcare and cross-institutional research. [20]

3. Causal AI

- Goes beyond correlation-based models to infer cause-effect relationships.
- Valuable in federated health research where understanding causality is key but data sharing is restricted.

Key Resources

1. Federated causal discovery with missing data in a multicentric study on endometrial cancer
   A novel federated causal discovery algorithm capable of pooling information from multiple sources with heterogeneous missing data to learn a graph representing cause-effect relationships. [21]
2. Causal discovery from observational and interventional data
   Distributed and federated causal discovery approaches for decentralized scenarios in healthcare that have privacy and regulatory data access constraints. [22]

4. Graph Data Science (GDS)

- Applies graph theory to model relationships between distributed datasets.
- Enhances federated discovery and linkage of data across institutions.

Key Resources

1. Federated biomedical knowledge graph-based question-answering system
   Biomedical discovery through an informatics platform that enables exploration and reasoning over an open-source, federated KG-based ecosystem. [23]
2. How Federated Knowledge Graphs Support AI Automation
   AI automation relies not only on data, but also on understanding. A federated knowledge graph provides that understanding in several ways. [24]

6. Application-Specific Semiconductors

- Custom chips designed for federated AI workloads (e.g., privacy-preserving training).
- Improves efficiency and scalability of federated platforms.

Key Resources

1. NVIDIA Custom GPU Architectures for Federated Learning
   NVIDIA FLARE is built on NVIDIA Clara Train, which runs on NVIDIA GPUs designed for high-throughput AI workloads enabling distributed training across hospitals and research centers without sharing patient data. [25]
2. Azure TEE is part of Azure Confidential Computing, which enables privacy-preserving AI and federated learning
   A Trusted Execution Environment is a segregated area of memory and CPU that's protected from the rest of the CPU by using encryption. [26]

7. Synthetic Data Generation or Twin Datasets

- Creates realistic but artificial datasets for training and validation.
- Can be used in federated settings to simulate data environments without exposing real data.

Key Resources

1. "Digital twin" datasets from complex EHR and wearable-device records
   Synthetic clinical data generation maximizes biomedical resource utilization and minimizes participant re-identification risks. [27]
2. Immune digital twins for complex human pathologies
   A bioinformatics ecosystem is proposed for data analysis, integration and modelling in Immune digital twins implementations. [28]

8. Neuro-Symbolic AI

- Combines symbolic AI reasoning with neural networks.
- Useful for federated systems that require explainability and rule-based governance.

Key Resources

1. Neurosymbolic AI can be leveraged in medical diagnostics
   Applied to Mental health diagnosis and conversational assistance. [29]
2. Neurosymbolic AI for reasoning on biomedical knowledge graphs (KG)
   KG completion (KGC) can help researchers make predictions to inform tasks like drug repositioning with hybrid approaches based on neurosymbolic artificial intelligence becoming more popular. [30]

9. Reinforcement Learning in Federated Settings

- Enables adaptive learning across distributed nodes.
- Can optimize resource allocation, model updates, and privacy trade-offs.

Key Resources

1. Federated Learning in Smart Healthcare
   A Comprehensive Review on Privacy, Security, and Predictive Analytics with IoT Integration. [31]
2. Federated learning applications for biomedical data
   Collaboratively training machine learning models without sharing raw data via 'federated learning' enables investigators to train a model locally on their own data, and share the parameters of the model with others to generate a central model. [32]

10. Privacy-Preserving MLOps

- Integrates privacy tools (e.g., differential privacy, homomorphic encryption) into model deployment pipelines.
- Ensures end-to-end compliance in federated AI workflows

Key Resources

1. Privacy-Preserving Machine Learning for Healthcare
   A guide for the development of private and efficient ML models in healthcare. [33]
2. Data Protection: Privacy-Preserving Data Collection With Validation
   Proposes a protocol for privacy-preserving data collection using autoencoders and enclave-based validation—aligns with data visitation principles. [34]

CONCLUSION

The platforms explored in this report have laid the groundwork for secure, scalable, and privacy-preserving biomedical research. But what comes next will be transformative. As Generative AI (GenAI), Large Language Models (LLMs), and autonomous agents evolve, they will redefine how researchers interact with data—especially in federated and data visitation contexts.

These technologies won't just enhance analysis—they'll reshape the architecture of research itself:

- LLM-powered agents will autonomously navigate metadata catalogs, generate cohort definitions, and orchestrate multi-step analyses across distributed nodes.
- GenAI tools will assist in data curation, annotation, and even synthetic data generation—reducing the burden on human experts while preserving privacy.
- AI-driven governance frameworks will dynamically enforce access policies, monitor compliance, and adapt to evolving ethical standards.
- Federated MLOps pipelines will enable seamless deployment of models across institutions, with built-in privacy-preserving mechanisms.

To stay ahead, platforms must embrace AI not just as a tool—but as a strategic layer woven into their infrastructure. This means:

- Investing in interoperable AI interfaces that can plug into diverse data environments.
- Building explainable and auditable AI systems to maintain trust and transparency.
- Leveraging agentic AI to reduce manual overhead and scale collaborative research.

The future of biomedical research will be distributed, intelligent, and deeply collaborative. Platforms that integrate AI thoughtfully—balancing innovation with ethics—will lead the way in unlocking insights from sensitive data without ever compromising its integrity.

As global health challenges grow more complex, these platforms will be essential in enabling **secure, scalable, and ethical research** across borders and institutions.

Acknowledgments

Funded by the European Union

This DV4RDA project has received funding through RDA TIGER from the European Union's Horizon Europe framework programme under grant agreement No. 101094406. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or institutions represented here. Neither the European Union nor the institutions can be held responsible for them.

References

1. ISO/IEC JTC 1. Data Virtualization Reference Architecture. ISO/IEC 19592, 2021.
2. Soudan, B., Abbas, S., Kubba, A. et al. Scalability and performance evaluation of federated learning frameworks: a comparative analysis. Int. J. Mach. Learn. & Cyber. 16,3329–3343 (2025). https://doi.org/10.1007/s13042-024-02453-4
3. Jagadish, H.V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J.M., Ramakrishnan, R., & Shahabi, C. “Big Data and Its Technical Challenges.” *Communications of the ACM*, 57(7), 2014.
4. Shokri, R. & Shmatikov, V. “Privacy-Preserving Deep Learning.” *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, 2015.
5. Lee, C., Lin, H., & Lin, T. “Federated Learning for Healthcare: A Literature Review.” *ACM Computing Surveys*, 55(1), 2022.
6. National Cancer Institute. “NCI Cancer Research Data Commons – Security & Governance resources.” https://datacommons.cancer.gov
7. Sui, D. & DeLyser, D. “Cross-Border Data Flow and Data Sovereignty.” *Annals of GIS*, 2021.
8. Global Alliance for Genomics and Health (GA4GH). “Framework for Responsible Sharing of Genomic and Health-Related Data.” 2020. https://www.ga4gh.org
9. McKinsey Digital. “Revisiting Data Architecture for Next-Gen Data Products.” 2022.
10. Gadepally, V., Mattson, T., Reuther, A., Samsi, S., & Kepner, J. “Massachusetts Open Cloud: Achieving Infrastructure-as-a-Service Economy of Scale in a Shared Cloud.” *IEEE Cloud Computing*, 2016.
11. OHDSI. “OMOP Common Data Model.” https://ohdsi.org/data-standardization/the-common-data-model/
12. HL7. “FHIR Overview.” https://hl7.org/fhir/overview.html
13. GA4GH. “Standards.” https://www.ga4gh.org
14. Atlan. “Benefits of a Data Repository.” https://atlan.com/benefits-of-a-data-repository/
15. Rudderstack. “What is Data Federation?” https://www.rudderstack.com/blog/data-federation/
16. Utilization of Artificial Intelligence-Based Tools to Support Autonomous Medical Operations. Aerospace Medical Association.June 1, 2025. https://ntrs.nasa.gov/citations/20240012663

17. Agentic AI in Biomedical Research: A New Era of Intelligent Collaboration. Apr 24, 2025. https://www.ica.ai/insights/agentic-ai-in-biomedical-research-a-new-era-of-intelligent-collaboration/

18. Sample Agents for Healthcare and Life Sciences on AWS. https://aws-samples.github.io/amazon-bedrock-agents-healthcare-lifesciences/

19. Merino, LH., Cabrero-Holgueras, J. (2023). Secure Multi-Party Computation. In: Mulder, V., Mermoud, A., Lenders, V., Tellenbach, B. (eds) Trends in Data Protection and Encryption Technologies . Springer, Cham. https://doi.org/10.1007/978-3-031-33386-6_17

20. Andrew, James & Nowygrod, Roman. (2025). Secure Multi-Party Computation (SMPC) for Privacy- Preserving Financial Analytics in the Cloud.

21. Zanga A, Bernasconi A, Lucas PJF, et al. Federated causal discovery with missing data in a multicentric study on endometrial cancer. J Biomed Inform. Published online July 22, 2025. doi:10.1016/j.jbi.2025.104877

22. Adam Li, Amin Jaber, Elias Bareinboim. Causal discovery from observational and interventional data across multiple environments. https://proceedings.neurips.cc/paper_files/paper/2023

23. Fecho K, Bizon C, Issabekova T, et al. An approach for collaborative development of a federated biomedical knowledge graph-based question-answering system: Question-of-the-Month challenges. Journal of Clinical and Translational Science. 2023;7(1):e214. doi:10.1017/cts.2023.619

24. Actian Corporation. July 23, 2025. https://www.actian.com/blog/data-intelligence/why-federated-knowledge-graphs-are-the-missing-link-in-your-ai-strategy/

25. Federated Learning With FLARE.November 29, 2021. https://blogs.nvidia.com/blog/federated-learning-ai-nvidia-flare/

26. Trusted Execution Environment (TEE). 05/07/2025. https://learn.microsoft.com/en-us/azure/confidential-computing/trusted-execution-environment

27. Marino, S., Cassidy, R., Nanni, J. et al. Medical data sharing and synthetic clinical data generation – maximizing biomedical resource utilization and minimizing participant re-identification risks. npj Digit. Med. 8, 526 (2025). https://doi.org/10.1038/s41746-025-01935-1

28. Niarakis, A., Laubenbacher, R., An, G. et al. Immune digital twins for complex human pathologies: applications, limitations, and challenges. npj Syst Biol Appl 10, 141 (2024). https://doi.org/10.1038/s41540-024-00450-5

29. A. Sheth, K. Roy and M. Gaur, "Neurosymbolic Artificial Intelligence (Why, What, and How)" in IEEE Intelligent Systems, vol. 38, no. 03, pp. 56-62, May-June 2023, doi: 10.1109/MIS.2023.3268724.

30. DeLong, L. N., Mir, R. F., Ji, Z., Smith, F. N. C., & Fleuriot, J. D. (2023). Neurosymbolic AI for reasoning on biomedical knowledge graphs. ArXiv. https://doi.org/10.48550/arXiv.2307.08411

31. Abbas SR, Abbas Z, Zahir A, Lee SW. Federated Learning in Smart Healthcare: A Comprehensive Review on Privacy, Security, and Predictive Analytics with IoT Integration. Healthcare. 2024; 12(24):2587. https://doi.org/10.3390/healthcare12242587

32. Crowson MG, Moukheiber D, Arévalo AR, Lam BD, Mantena S, Rana A, et al. (2022) A systematic review of federated learning applications for biomedical data. PLOS Digit Health 1(5): e0000033. https://doi.org/10.1371/journal.pdig.0000033

33. Guerra-Manzanares, A., Lopez, L.J.L., Maniatakos, M., Shamout, F.E. (2023). Privacy-Preserving Machine Learning for Healthcare: Open Challenges and Future Perspectives. Lecture Notes in Computer Science, vol 13932. Springer, Cham. https://doi.org/10.1007/978-3-031-39539-0_3

34. J. Hou et al., "Data Protection: Privacy-Preserving Data Collection With Validation" in IEEE Transactions on Dependable and Secure Computing, vol. 21, no. 04, pp. 3422-3438, July-Aug. 2024, doi: 10.1109/TDSC.2023.3326299