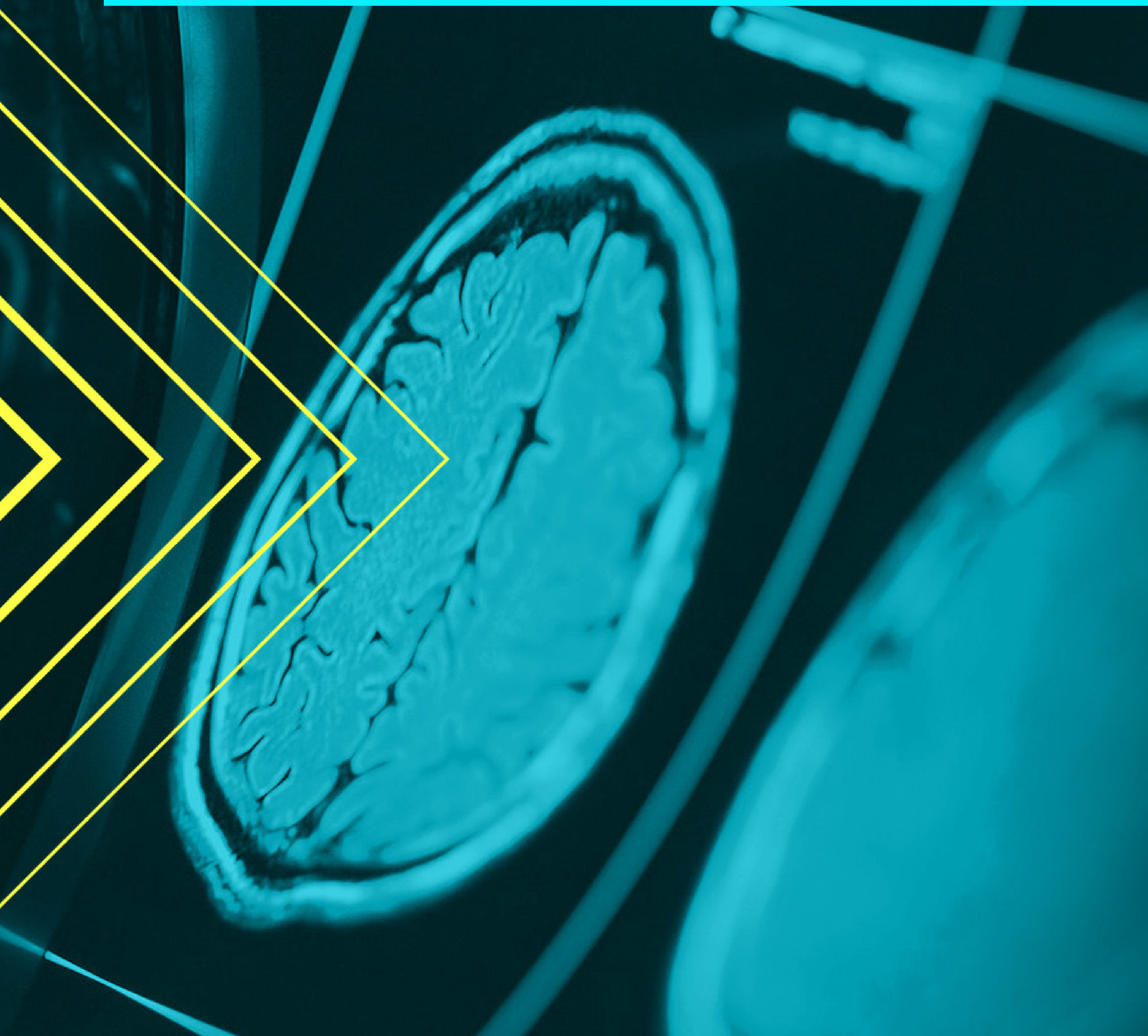
Together Towards Tomorrow: Collaborating to Overcome Data Challenges and Accelerate AI Innovation

Thiago Ramos dos Santos[1], Franziska Lobig[2] and Steffen Vogler[2]

[1] Zühlke Engineering GmbH, [2] Bayer AG



1

Executive Summary

Artificial Intelligence (AI) has rapidly advanced on the foundation of open-source frameworks and vibrant collaborative communities, revolutionizing many industries. In healthcare, and particularly in medical imaging, there is significant potential for AI-driven transformation. However, while AI technologies are plentiful and accessible, obtaining data can be challenging. Medical imaging data, essential for building robust, generalizable AI models, remains locked behind stringent privacy regulations, fragmented systems, and complex legal barriers. This creates a paradox: the tools for progress are widely available, yet the critical ingredient, diverse and representative health data, is often out of reach.

Bayer's AI Innovation Platform seeks to resolve this disconnect by fostering a secure, collaborative environment where the promise of open-source AI can be united with responsible and privacy-conscious access to medical data. By recognizing best practices in data stewardship, transparency, and security, the platform empowers innovators to develop and validate healthcare AI applications. Through bridging the gap between technological capability and data-driven insight, Bayer's initiative aspires to unlock the true potential of AI in medical imaging, ultimately striving to improve patient outcomes and advance the future of healthcare.

2

Introduction and Current Landscape

Artificial Intelligence (AI) technology has advanced in recent years, propelled by the openness and agility of open-source frameworks and global communities. Tools and libraries such as TensorFlow[1], PyTorch[2], NumPy[3], and others have made cutting-edge technology accessible to researchers and innovators everywhere, lowering barriers to entry and accelerating progress at a pace never imagined.

In stark contrast, access to medical imaging data, the lifeblood for any radiology AI application, remains tightly restricted. Health data is highly protected by privacy regulations, isolated within fragmented systems, and enmeshed in layers of legal complexity. While the technological building blocks are abundant and refined, the essential resource to fuel meaningful advances is often just out of reach. This entrenched division between the freedom of technological innovation and the constraints of data access stands as a significant challenge to realizing the full transformative power of AI in medicine.

Unless this fundamental challenge is addressed, the true promise of AI in healthcare may remain unfulfilled. We risk confining breakthroughs to the lab, never allowing them to reach the patients and clinicians who need them most. Today, there are approximately 1,000 AI algorithms cleared by the FDA[4], most of them are image-based, reflecting the remarkable pace of technological progress. However, many of these solutions focus on relatively simple use cases, such as expediting the detection of acute conditions like stroke, or making screening processes more efficient in terms of timing and number of cancer cases detected[5]. The transformative power of AI in healthcare remains not fully realized.

Without bridging the gap between technology and data, many compelling possibilities for the future will remain out of reach. For example, what if we could anticipate Alzheimer's, heart disease, and diabetes a decade before symptoms emerge by analyzing blood protein patterns with machine learning[6]? What if we could forecast patient survival and response to immunotherapy across cancers using just a handful of clinical and genomic features[7]? What if we could detect pancreatic cancer on scans, sometimes more than a year before traditional diagnosis, and reveal hidden disease that eludes the human eye[8]? What if we could diagnose rare conditions earlier and more precisely in routine care[9], and replace invasive liver biopsies with fast, accurate imaging for better patient outcomes[10]?

Many such breakthroughs, and many more yet unimagined, may remain confined within labs unless we find ways to responsibly and securely unlock health data for innovation. This will help AI to fulfill its potential to transform patient care and the future of medicine.

To the best of our knowledge, solutions available today either focus on: (1) Providing infrastructure for storing and processing imaging data; (2) providing algorithms for data organization, annotation, etc.; and (3) algorithms and models for image analysis. The gap between technology and data remains wide open. Research conducted by Bayer, shows that the lack of standards in data selection, curation, and access is the biggest pain point in the development of medical algorithms (Figure 1). Aware of this, ARPA-H (US Advanced Research Projects Agency for Health) released back in 2024 the now finished INDEX program[11], which aimed "to create a platform that seamlessly links data providers, data users, and service providers with high quality images that will enable robust, trustworthy AI tool development for pathology and radiology."

RESPONDENTS RATED FRUSTRATION LEVEL OF PAIN POINTS IN AI DEVELOPMENT AND VALIDATION PRIOR TO DEPLOYMENT

73% **Multidisciplinary knowledge missing**
In-depth regulatory, medical, clinical, commercial, reimbursement, data science, and legal information

73% **Data challenge**
Challenges exist with data selection, curation, access, affordability, and annotation

66% **Inefficiencies**
Optimizing paper work, contracting, payment processes and cloud cost transparency

64% **Fragmented development environments**
Providing secure collaborative areas to allow working on sensitive data securely, setting up and maintaining development environments over time, and meeting regulatory standards

59% **AI based validation of clinical trial sites**
Clinical trials are costly and time-consuming. In addition, obtaining access to sites for validating studios and meeting regulatory standards, can be challenging.

3

Challenges

In the rapidly evolving landscape of healthcare AI, access to diverse, high-quality data and robust development infrastructure remains a persistent challenge. Through an examination of access barriers, innovative platform features, and practical approaches to privacy and interoperability, we chart a path toward more scalable, inclusive, and impactful medical AI innovation.

Access to Data

Access to medical imaging data is stratified into three informal levels, which we discuss below.

First-Level Access

Direct access to data from clinics and hospitals. This is typically limited to academic institutions with internal collaborations. However, it comes with additional steps such as Institutional Review Board (IRB) approvals, data anonymization, and legal reviews. As a result, academics often avoid using real-world data in favor of synthetic or public datasets to meet publication deadlines.

Second-Level Access

Aggregated datasets from commercial providers like SegMed or RadImageNet. These are curated and standardized, making them attractive to industry players. However, the cost of access can be prohibitive for smaller startups and academic labs. Beyond cost, there are additional barriers: researchers may lack awareness of the range of available data providers, and comparing terms, quality, and coverage between different offerings can be confusing or opaque. These obstacles, combined, can contribute to significant delays in AI development and deployment.

An important consequence of this reliance on public datasets is a lack of diversity and variability in the data used for research. Many researchers gravitate toward the same well-known public datasets[12], leading to several challenges:

GENERALIZABILITY
Models trained on narrow, homogeneous datasets may perform poorly when applied to data from different populations, scanners, or clinical settings.

BECHMARK SATURATION
As leading datasets become overused, incremental improvements reported in the literature may represent tuning to idiosyncrasies of the dataset rather than true advances in clinical utility.

BIAS AND REPRODUCIBILITY
Repeated use of similar datasets can perpetuate inherent biases and limit the reproducibility of results across diverse environments.

Third-Level Access

Platforms like Bayer's AI Innovation Platform (AIIP)[13], aggregate both public datasets and commercial aggregators. AIIP offers a unified interface, compliance-ready infrastructure, and negotiation support, making it viable for both academia and industry.

AFFORDABILITY

Even when data is available, the cost of acquiring, curating, and ensuring compliance remains high, compounded by the fact that only a handful of major data aggregators control most of the market. Because these few providers set their own prices and there is no standard way of offering data, pricing lacks transparency and true competition is minimal. As a result, the natural balancing effect of a competitive market is absent. Academic labs, in particular, lack the resources to build large, high-quality datasets, while the time required for data cleaning, harmonization, and annotation rarely translates to academic recognition. This leads to a situation where:

- Labs prioritize fast publication cycles (e.g., ICML, MICCAI, NeurIPS).
- Data curation is seen as a sunk cost with little academic return.
- Only large institutions can afford to build robust datasets, creating a significant barrier to entry.

LEGISLATION

Regulations such as GDPR, HIPAA, and MDR play a vital role in safeguarding patient privacy and protecting societal interests. While these frameworks are crucial for ensuring ethical data usage, they can also add layers of complexity to research and innovation:

- Data sharing agreements often require careful legal review.
- Cross-border data transfers face stringent requirements.
- Anonymization standards can differ across regions.
- Innovation, especially for smaller players who lack dedicated legal teams.

These challenges may slow down innovation or make it harder for smaller organizations to participate, but they ultimately serve to uphold the trust, safety, and rights of individuals in the healthcare ecosystem.

4

Innovations and Solutions

AI researchers and developers operate within a landscape of competing pressures, including the need for speed, data quality, risk management, and the pursuit of impactful results. Many promising innovations struggle to move beyond initial prototypes, a phenomenon often described as the “last mile” problem in translational AI research[14]: the challenge of turning laboratory breakthroughs into real-world solutions.

To address these hurdles, we propose the concept of a “solutionsphere”: a collaborative ecosystem where shared infrastructure, frameworks for security and privacy, and standardized pipelines can accelerate data work and AI development. The solutionsphere is designed to foster an environment in which data and tools are accessible, reliable, and compliant with security and privacy regulations, enabling participants to contribute, access, and benefit from a centralized platform. In essence, the solutionsphere aims to bridge the gap between research and deployment, supporting scalable and inclusive innovation in fields like healthcare AI.

Bayer’s AI Innovation Platform represents a foundational step in this direction. By centralizing data offerings and development tools in a single, streamlined platform, AIIP empowers researchers, startups, and institutions to collaborate and innovate more effectively, all within a secure and compliant setting. This approach lays the groundwork for a new era of accessible and impactful AI-driven solutions.

Bayer’s AIIP is designed to break down barriers to data access by establishing a trusted and secure ecosystem tailored for AI development in radiology. The platform enables startups, hospitals, and researchers to build innovative AI applications with an infrastructure and streamline data access to accelerate meaningful advancements in medical AI. AIIP aims to foster an environment where pioneering ideas can thrive and be translated into impactful solutions.

VISUALIZATION OF A SOLUTIONSPHERE

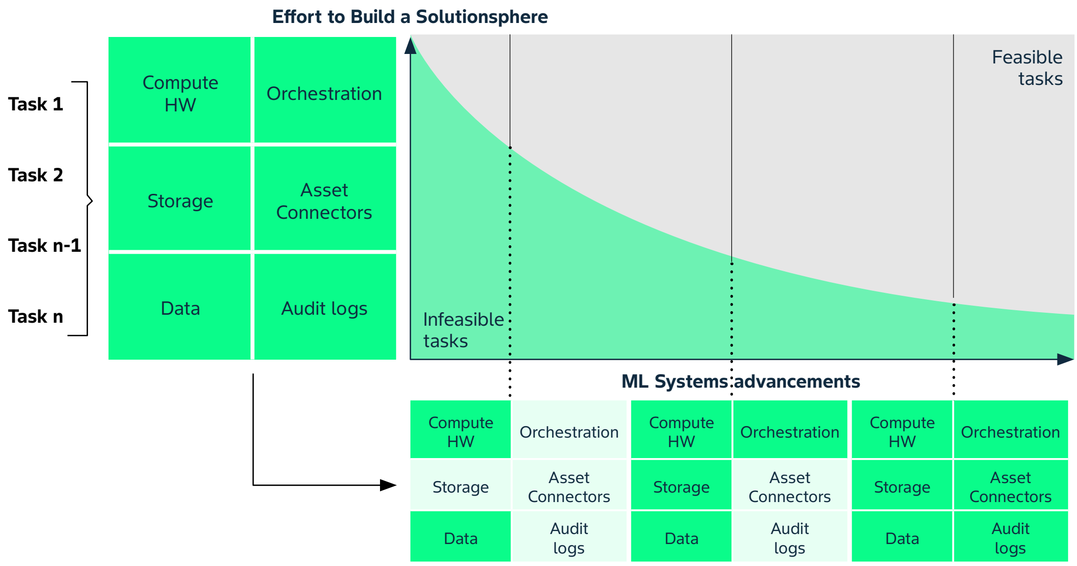


Fig 2: Visualization of a solutionsphere concept, showing that smooth integration of tools and technical capabilities makes it increasingly easier to work effectively on challenging tasks for healthcare AI applications[15].

Connecting data, tools and security for image-based AI models

The Data Marketplace

The data marketplace within AIIP automatically connects developers with data providers based on specific search criteria. It displays statistics about available datasets, enabling informed decision-making. Developers can negotiate terms and order data directly from well-known data providers through the platform. The data is delivered securely to their workspaces. This streamlined process benefits both data providers and developers by facilitating secure and efficient data exchange.

The Development Environment

AIIP’s development environment is designed for reliability and regulatory awareness, supporting the full lifecycle of AI model creation, from data ingestion to validation and deployment. It integrates smoothly with open-source tools like TensorBoard and MONAI Label, offering developers a familiar and robust toolkit within a secure and supportive infrastructure. This environment empowers teams to innovate confidently while maintaining best practices in medical software development.

Privacy and Compliance by Design

AIIP is built with privacy and security at its core. Data never leaves the secure environment, and access is tightly controlled. Comprehensive audit trails, encryption, and access controls ensure that all activities are transparent.

5

Conclusion

The integration of open-source AI tools with proprietary health data represents a pivotal challenge in the journey toward transformative healthcare solutions. Without secure, scalable frameworks, the promise of artificial intelligence, its potential to improve diagnostics, personalize treatment, and accelerate research, may remain largely untapped. Bayer's AI Innovation Platform (AIIP) rises to this challenge by offering a comprehensive and robust infrastructure designed to bridge the gap between innovation and real-world clinical practice.

Through its secure data marketplace, AIIP enables healthcare organizations to share and access real-world medical imaging data to help satisfy regulatory requirements, while maintaining privacy and security standards. These safeguards help to ensure that sensitive patient information remains protected, fostering trust among participants and paving the way for collaborative advancements. The qualified development environment provides researchers, clinicians, and developers with state-of-the-art tools, pre-trained models, and APIs to accelerate experimentation and deployment of novel AI applications.

Collaborations with technology providers e.g. Google Cloud, equip the platform with scalable computational resources and advanced generative AI capabilities, ensuring that innovation is not limited by infrastructure constraints.

The platform's community-centric approach encourages active engagement from developers and clinicians.

Collaborative challenges and a vibrant marketplace for AI solutions create an environment where new ideas can flourish and find rapid adoption.

By serving as a dynamic blueprint for the integration of AI into healthcare, AIIP not only streamlines the development and deployment of medical AI but also establishes new ways for data accessibility, workflow efficiency, and patient outcomes. We warmly invite developers, researchers, healthcare providers, and industry partners to join the AIIP community. Together, we can drive the next era of medical innovation, one where secure, intelligent, and collaborative solutions help shape the future of patient care.

References

1. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray and B. Steiner, “TensorFlow: a system for large-scale machine learning,” in USENIX conference on Operating Systems Design and Implementation, 2016.
2. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani and Chila, “PyTorch: an imperative style, high-performance deep learning library,” in International Conference on Neural Information Processing Systems, 2019.
3. C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk and Br, “Array programming with NumPy,” Nature, vol. 585, p. 357–362, 2020.
4. D. Fornell, “FDA has now cleared more than 1,000 AI models, including many in cardiology,” 10 January 2025. [Online]. Available: https://cardiovascularbusiness.com/topics/artificial-intelligence/fda-has-cleared-more-1000-ai-algorithms-many-cardiology. [Accessed 4 August 2025].
5. F. Lobig, J. Graham, A. Damania, B. Sattin, J. Reis and P. Bharadwaj, “Enhancing patient outcomes: the role of clinical utility in guiding healthcare providers in curating radiology AI applications,” Frontiers in Digital Health, vol. 6, 2024.
6. D. A. Gadd, R. F. Hillary, Z. Kuncheva, T. Mangelis, Y. Cheng, M. Dissanayake, R. Admanit, J. Gagnon, T. Lin, K. L. Ferber and H. Runz, “Blood protein assessment of leading incident diseases and mortality in the UK Biobank,” Nature Aging, vol. 4, p. 939–948, 2024.
7. T.-G. Chang, Y. Cao, H. J. Sfreddo, S. R. Dhruba, S.-H. Lee, C. Valero, S.-K. Yoo, D. Chowell, L. G. T. Morris and E. Ruppin, “LORIS robustly predicts patient outcomes with immune checkpoint blockade therapy using common clinical, pathologic and genomic features,” Nat Cancer, vol. 5, no. 8, pp. 1158-1175, 2024.
8. P. Korfiatis, G. Suman, N. G. Patnam, K. H. Trivedi, A. Karbhari, S. Mukherjee, C. Cook, J. R. Klug, A. Patra, H. Khasawneh, N. Rajamohan, J. G. Fletcher, M. J. Truty and S. Majumder, “Automated Artificial Intelligence Model Trained on a Large Data Set Can Detect Pancreas Cancer on Diagnostic Computed Tomography Scans As Well As Visually Occult Preinvasive Cancer on Prediagnostic Computed Tomography Scans,” Gastroenterology , vol. 165, no. 6, pp. 1533-1546, 2023.
9. D. P. Germain, D. Gruson, M. Maces and N. Garçon, “Applying artificial intelligence to rare diseases: a literature review highlighting lessons from Fabry disease,” Orphanet Journal of Rare Diseases, vol. 20, no. 186, 2025.
10. M. Bajre, M. Moawad, E. Shumbayawonda, J. E. Carolan, J. Hart, E. Culver and M. Heneghan, “LiverMultiScan as an alternative to liver biopsy to monitor autoimmune hepatitis in the National Health Service in England: an economic evaluation,” BMJ Open, vol. 12, no. 9, 2022.
11. Advanced US Research Projects Agency for Health (ARPA-H), “INDEX - Imaging Data Exchange,” 2025. [Online]. Available: https://arpa-h.gov/explore-funding/programs/index. [Accessed 4 August 2025].
12. M. Graziani, M. Molnar, I. E. Morales, J. Cadow-Gossweiler and T. Laino, Making Sense of Data in the Wild: Data Analysis Automation at Scale, arXiv, 2025.
13. Bayer AG, “AI Innovation Platform,” [Online]. Available: https://app.innovationplatform.ai. [Accessed 15 August 2025].
14. E. Coiera, “The Last Mile: Where Artificial Intelligence Meets Reality,” Journal of Medical Internet Research, vol. 21, no. 11, 2019.
15. M. Akhtar, O. Benjelloun, C. Conforti, P. Gijsbers, J. Giner-Miguelez, N. Jain, M. Kuchnik, Q. Lhoest, P. Marcenac, M. Maskey, P. Mattson, L. Oala, P. Ruyssen, R. Shinde and E. Simperl, “Croissant: A Metadata Format for ML-Ready Datasets,” in Proceedings of the Eighth Workshop on Data Management for End-to-End Machine Learning, 2024.

PP-DIGI-US-0040-1 October 2025