Insight Brief

Constructing Defensible AI Platforms in Healthcare

How we unlocked revenue growth potential exceeding 10x what was previously possible

Table of contents

Healthcare organizations face increasing pressure to adopt advanced technologies while navigating complex Artificial Intelligence (AI), data protection, and regulatory expectations. Based on our experience supporting both public and private institutions, we offer a defensible approach to AI platforms that delivers meaningful insights while safeguarding patient trust. As we delve deeper into digital transformation, the significance of robust and secure data and AI platforms becomes clear. Such platforms are fundamental in harnessing the full potential of health data, ensuring that innovations lead to improved health outcomes and efficiency gains while advancing standards of data integrity and protection. So, we ask ourselves:

How can healthcare organizations adopt AI without compromising trust or data protection?

What makes an AI system truly defensible, beyond compliance checklists?

Are current governance models ready to manage the risks posed by generative AI and autonomous systems?

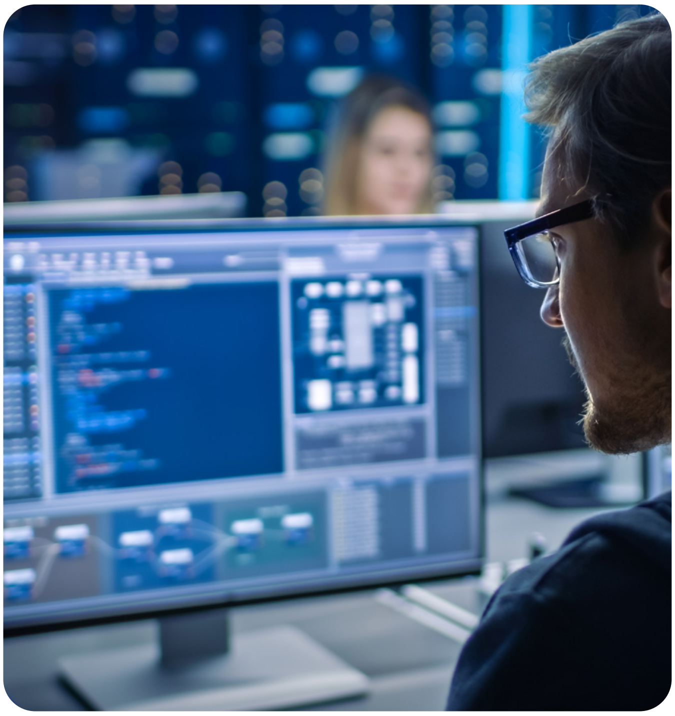
How do we ensure that technical design decisions support regulatory defensibility and patient expectations?

In 2024, we introduced *defensible AI* as data and AI systems that reliably achieve organizational objectives in a manner that meets or exceeds rigorous safety, data protection, and regulatory standards. It involves an approach where technology drives the adoption of AI safety and AI security standards. This is crucial in the healthcare sector, where data is confidential and sensitive, and the implications of mishandling data can be significant. By building platforms that are powerful while being trustworthy and transparent, we pave the way for advancements that are both innovative and secure, ensuring that patient welfare remains at the forefront of technological progress.

In this brief, we consider three core concepts to defensible AI that have emerged from our work with clients and the development of an AI platform that is achieving remarkable success. These insights are drawn from our work across healthcare sectors and are shared to help other organizations implement defensible AI solutions aligned with emerging standards.

Scalable AI management: Central to these efforts is the concept of *control management*, which involves mapping controls and requirements against best practice standards and functional needs — and continuously monitoring those controls when operating the platform. This ensures that every component of the AI system performs optimally and operates within a framework that is aligned with regulatory and patient expectations. It's also a proactive approach that ensures ongoing alignment and responsiveness to emerging challenges and technological advancement.

Secure insights: AI security is introducing new challenges to contend with, beyond considerations of the data that is accessed, and by whom. Emerging data protection technologies provide a means of ensuring only the right data, at the right time, and for the right purposes are available. But securing the data, or analytical tools used to process that data, requires new thinking to address a rapidly modernizing technology landscape. We consider the balance between data and AI security, introducing a technique we call *synthetic trends* to enhance machine learning models without compromising the security of the data used.



Effective platform design: The architecture that securely bridges health and other data to create meaningful insights plays a critical role in driving business outcomes. A key consideration is finding the right balance between central control (*a data fabric*) and decentralized autonomy (*a data mesh*). In working with sensitive data, the architecture of many platforms will involve segregated workspaces controlled by independent teams, each producing their own data products while maintaining rigorous governance independently. This structure boosts innovation and efficiency while reinforcing the security and compliance of the data and deployed AI system.

As we explore these elements further, we will delve into the specifics of control management, the innovative applications of synthetic trends in machine learning, and the strategic importance of a *secure health fabric* in sustaining the integrity and defensibility of AI platforms. Our experience has shown that each component is integral to shaping a healthcare environment where technology and patient-centric care converge seamlessly.

What does it take to manage AI at scale?

We have seen time and again that in the realm of healthcare technology, the integration of AI presents transformative potential — from enhancing diagnostic accuracy to personalizing patient care. However, the inherent complexities and the sensitive nature of health data necessitate a robust framework to ensure that these innovations are both effective and defensible, aligned with stringent regulatory standards. This is where an AI management system can become pivotal to balance governance and innovation, and it starts with control management.

Control management, at its core, involves the systematic alignment of AI functionalities with regulatory requirements and best practices through a process known as controls and requirements mapping. This process serves as the backbone for constructing AI systems that are both powerful and trustworthy. It ensures that every aspect of the AI application, from data handling to patient interaction, adheres to established guidelines and ethical norms, thus safeguarding against potential risks that could compromise patient safety or data integrity.

Controls mapping

The first step in effective control management is to define what's needed by identifying the essential controls for AI systems. These controls are informed by a comprehensive understanding of applicable regulations and best practices in the domains of data protection, privacy, and AI governance (see the sidebar controls mapping with AI). For instance, international frameworks by ISO and NIST provide a structured approach to consolidating these requirements, offering a common platform for organizations to align their AI strategies with global standards.[1]

System design

Once the necessary controls are identified, the next phase — system design — involves translating these requirements into technical specifications. This translation dictates the architecture of the AI systems, ensuring that they are built to be functional as well as secure and compliant. This might include designing data flows that respect privacy, establishing governance protocols that ensure accountability, and setting up risk assessments to continuously evaluate the credibility and reliability of AI operations.

Operating guidelines

Finally, control management extends into creating operating guidelines that ensure these controls are effectively managed and adapted over time. This includes setting up monitoring systems to oversee AI performance continuously, validating AI models periodically to ensure their accuracy, and adapting to new regulations through ongoing horizon scanning. This feedback loop ensures that AI systems remain aligned with the initial control mappings and adapt to an evolving regulatory landscape.

[1] International Organization for Standardization (ISO); U.S. National Institute of Standards and Technology (NIST)

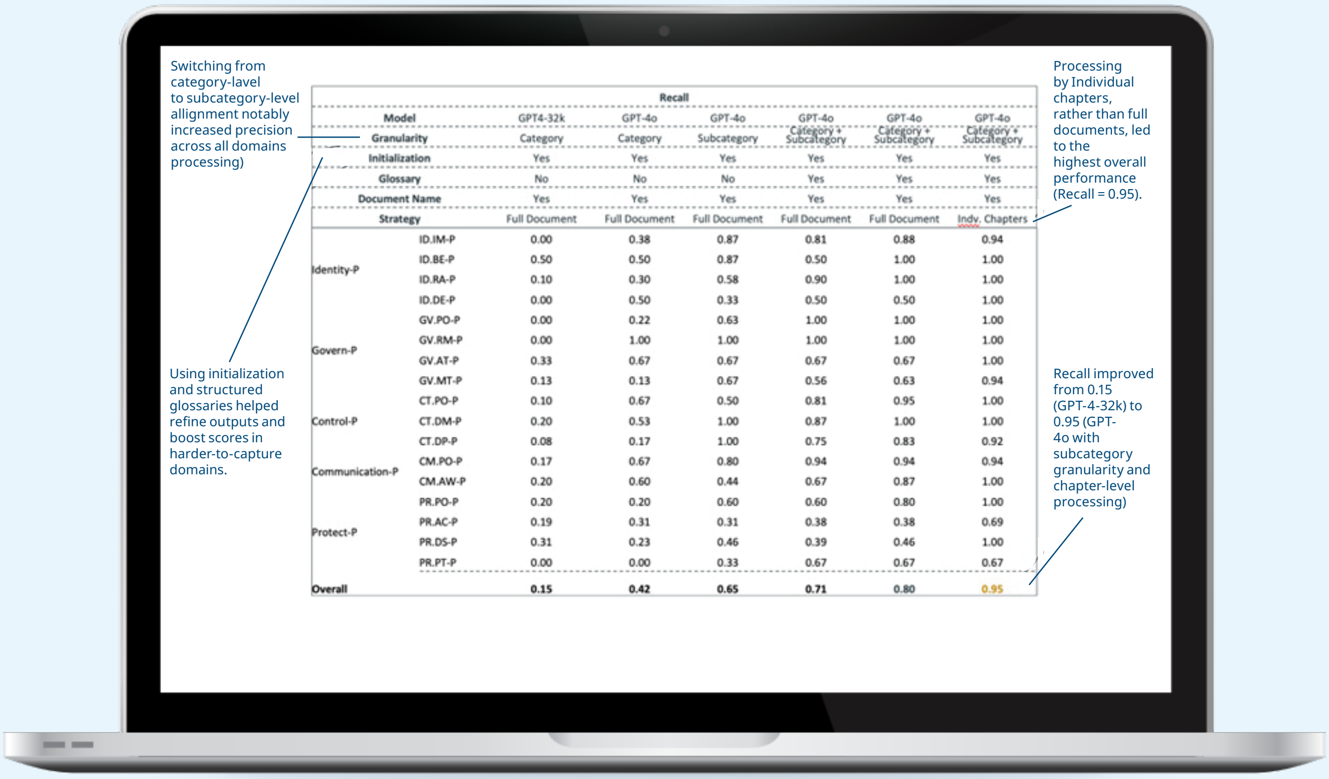
Our structured approach to AI management — underpinning the broader narrative of building defensible AI platforms — ensures that healthcare organizations can harness the power of AI responsibly. By maintaining a rigorous adherence to controls and requirements mapping, we can clarify the level and scope of oversight required within a comprehensive AI management system. This can include integrating impact assessments, performance monitoring, and governance checkpoints throughout the AI lifeycle, covering areas such as data quality, security, transparency, and explainability. A mature AI management system will also enable continuous learning and improvement so that an AI system evolves responsibly, in step with organizational goals, regulatiory expectations, and patient trust.

Controls mapping with AI

We use a Large Language Model (LLM) and Natural Language Processing (NLP) to automate controls mapping from a series of regulatory and guidance documents to well-established frameworks, such as the International Organization of Standardization (ISO) and by the U.S. National Institute of Standards and Technology (NIST). With an expert in the loop, we are able to conduct a risk-based assessment of requirements and use that to develop an implementation roadmap that align to needs with operational requirements, resulting in a clear path forward for product launch. Use cases have confirmed the framework's ability to:

- Extract and structure diverse regulatory actions into interpretable matrices
- Identify overlaps and discrepancies across frameworks through transparent scoring
- Support expert-driven validation through explainability and traceable mappings

The results below used the NIST Privacy Framework to illustrate the method's capacity to scale across domains.



Switching from category-lavel to subcategory-level allignment notably increased precision across all domains processing)

Using initialization and structured glossaries helped refine outputs and boost scores in harder-to-capture domains.

Processing by Individual chapters, rather than full documents, led to the highest overall performance (Recall = 0.95).

Recall improved from 0.15 (GPT-4-32k) to 0.95 (GPT-4o with subcategory granularity and chapter-level processing)

		Recall					
Model		GPT4-32k	GPT-4o	GPT-4o	GPT-4o	GPT-4o	GPT-4o
Granularity		Category	Category	Subcategory	Category + Subcategory	Category + Subcategory	Category + Subcategory
Initialization		Yes	Yes	Yes	Yes	Yes	Yes
Glossary		No	No	No	Yes	Yes	Yes
Document Name		Yes	Yes	Yes	Yes	Yes	Yes
Strategy		Full Document	Full Document	Full Document	Full Document	Full Document	Indv. Chapters
Identity-P	ID.IM-P	0.00	0.38	0.87	0.81	0.88	0.94
	ID.BE-P	0.50	0.50	0.87	0.50	1.00	1.00
	ID.RA-P	0.10	0.30	0.58	0.90	1.00	1.00
	ID.DE-P	0.00	0.50	0.33	0.50	0.50	1.00
Govern-P	GV.PO-P	0.00	0.22	0.63	1.00	1.00	1.00
	GV.RM-P	0.00	1.00	1.00	1.00	1.00	1.00
	GV.AT-P	0.33	0.67	0.67	0.67	0.67	1.00
	GV.MT-P	0.13	0.13	0.67	0.56	0.63	0.94
Control-P	CT.PO-P	0.10	0.67	0.50	0.81	0.95	1.00
	CT.DM-P	0.20	0.53	1.00	0.87	1.00	1.00
	CT.DP-P	0.08	0.17	1.00	0.75	0.83	0.92
Communication-P	CM.PO-P	0.17	0.67	0.80	0.94	0.94	0.94
	CM.AW-P	0.20	0.60	0.44	0.67	0.87	1.00
Protect-P	PR.PO-P	0.20	0.20	0.60	0.60	0.80	1.00
	PR.AC-P	0.19	0.31	0.31	0.38	0.38	0.69
	PR.DS-P	0.31	0.23	0.46	0.39	0.46	1.00
	PR.PT-P	0.00	0.00	0.33	0.67	0.67	0.67
Overall		**0.15**	**0.42**	**0.65**	**0.71**	0.80	0.95

In subsequent sections, we will explore other critical components such as synthetic trends and secure health fabric, further detailing how they contribute to the robustness of defensible AI platforms.

How can we secure AI models without limiting utility?

In our efforts to build defensible AI platforms, particularly in healthcare, the sophistication and security of how data is used and shared has become especially important. Robust de-identification methods, which involve removing identifying elements, can be used but the industry lacks widespread adoption of standardized practices. This absence of fixed standards provides space to explore forward-looking approaches, especially in light of emerging AI threats that will need to be addressed. [2] As AI and other emerging technologies reshape the landscape, more sophisticated strategies are needed to balance AI and data protection with responsible use.

Inspired by the manipulation of data in a latent space for signal and image processing, we conceptualized the idea of *synthetic trends* within a simplified, hidden and abstracted version of the original source data. This method significantly enhances both the utility and security of data utilized in AI models — see the case study at the end of this brief for an example of how this works in practice. This transformation process preserves the utility of the data for machine learning models while protecting sensitive information.

Abstracted data: Imagine a map that shows only the key patterns and connections, without revealing the individual details behind them. The abstracted data, an embedding space, is essentially a high-dimensional space used to convert large sets of complex data into simplified, yet richly informative, vectors or points. These vectors capture the essential patterns and relationships inherent in the data and do so in a way that abstracts and compresses the original input into a form that is non-reversible and typically uninterpretable by humans (see the sidebar reconstruction risk).

Synthetic trends: We want to maintain the underlying statistical properties and behavioral patterns necessary for effective AI modeling — without exposing or requiring access to the original, sensitive data. Synthetic trends are the distilled insights derived from the processed data within the embedding spaces. This approach aligns with the stringent AI and data protection standards and patient expectations regarding the use of their information, ensuring that data handling within AI systems is both responsible and secure.

Secure AI modeling: Machine learning models are then trained and deployed on the synthetic trends, which are highly secure by their very design, representing abstract statistical properties and patterns without disclosing their construction. By using synthetic trends, machine learning models are themselves secure from many AI threats. For example, reconstructing the underlying data from model outputs would at best reveal the synthetic trends that hide underlying features and relationships. The sensitive data that was abstracted away remains hidden and secure.

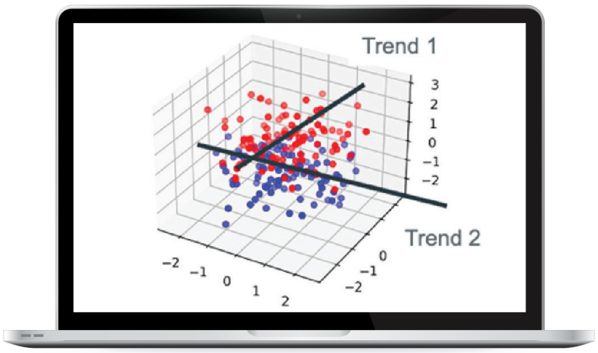
In the broader context of building defensible AI platforms, these synthetic trends allow for the robust use of data in environments where privacy and AI security must be uncompromised, such as in healthcare, where patient data is both vital and vulnerable. By leveraging abstracted data and synthetic trends, organizations can ensure that their AI systems are powerful and effective while inherently designed to protect the privacy and integrity of the data they analyze.

[2] OWASP Machine Learning Security Top Ten, 2023. Available from: https://owasp.org/www-project-machine-learning-security-top-10/

Source data

ID	AGE	SEX	ICD-10	SOURCE	RX
111	35	M	R20	Physician	Atorvastatin
222	33	M	R22	Hospital	TCA
333	37	F	R20	Hospital	
444	39	F	R21	Physician	NSAID

Abstracted



Incorporating embedding spaces and synthetic trends into AI systems requires a meticulous approach to system design and operational guidelines, as discussed in the previous section on controls management. These components need to be seamlessly integrated into the overall architecture of AI platforms, ensuring that every step — from data ingestion and processing to insight generation and application — is governed by principles of security, privacy, and compliance.

As we continue to explore the components of defensible AI platforms based on our experience, the innovative use of synthetic trends within abstracted data exemplifies how advanced technologies can be harnessed to enhance data utility while rigorously protecting data. This supports the operational needs of healthcare organizations and builds trust in AI applications, crucial for sustainable advancement in healthcare technology.

Reconstruction risk

Synthetic trends are the distilled insights derived from the processed data within abstracted embedding spaces. These embedding spaces preserve essential patterns and relationships while compressing the original input into a form that is non-reversible and typically uninterpretable by humans, such as the example below.

ID	OUTCOME	TREND 1	TREND 2	TREND 3	TREND 4	TREND 5
1	0	0.159	1.629	0.138	2.5	0.89
2	1	-1.712	-0.153	-5.8	-0.15	-0.15
3	1	1.090	0.617	11.72	0.83	0.617
4	0	1.771	-1.277	7.63	-0.28	-0.45
5	0	-1.308	-0.817	-4.6	1.14	-0.31

Because original values in the health data are transformed into an embedding space, a row-level record would need to be reconstructed before any meaningful data about an individual could be misused. We define this possibility, the chance of inferring original values from their transformed representations, as *reconstruction risk*, a precursor to data misuse.

Reconstruction risk quantifies the privacy and security exposure of the transformed embedding space. It accounts for the diminishing contribution of higher order embedding dimensions, as well as the effects of distortion and noise injection. While retaining more dimensions may preserve more detail, we deliberately limit this by applying dimensionality reduction This introduces distortion that acts as a safeguard, making it significantly harder to reconstruct original values and reducing the likelihood of reversal.

In practice, we use this method to reduce large and complex datasets into compact forms that improve computational efficiency while producing secure synthetic trends. In these cases, reconstruction risk is orders of magnitude below practical thresholds, supporting a high level of anonymization. This approach offers a practical way to assess AI and data protection through the lends of adversarial reconstruction difficulty.

Is your platform ready for AI governance?

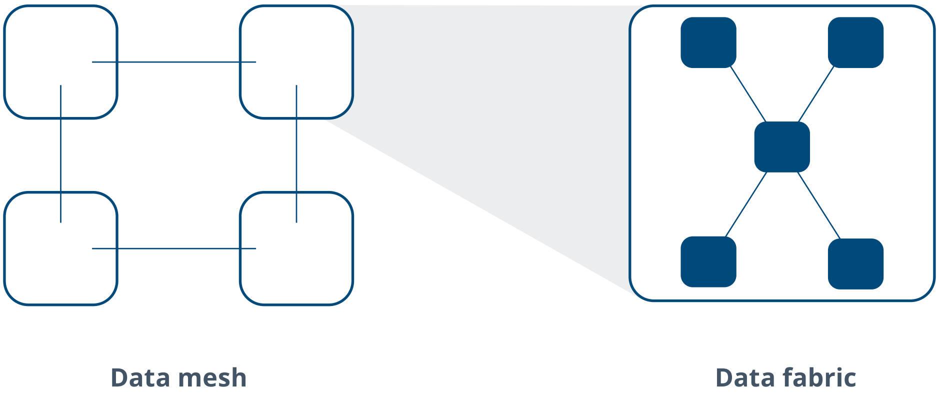
Where data-driven decision-making is critical, the architecture of data platforms in healthcare needs to strike a delicate balance between centralized control and decentralized innovation. This balance facilitates robust data governance and fosters the flexibility required for rapid technological advancements. Effective platform design is critical in navigating this landscape, especially as AI agents become integral to automating and enhancing data interactions. Together, these concepts form the pillars of modern data platform architecture in healthcare, aiming to harness the full potential of AI while ensuring the data remains secure, private, and effectively managed.

Data mesh: Decentralized data ownership and architecture can be achieved through a data mesh, allowing individual domains within an organization to control their own data within a framework that ensures interoperability and standardization. This method supports rapid innovation by empowering teams to develop and deploy data-driven solutions independently while being aligned on core principles. Oversight becomes critical to ensure continued alignment at deployment and during operational phases.

Data fabric: In the realm of healthcare AI, the creation of a secure health fabric can be pivotal for setting standards and enforcing oversight. This approach is fundamental to ensuring that data integrity and security are maintained while fostering innovation through AI applications. A secure health fabric is designed around the concept of segregated workspaces controlled by independent teams, each responsible for developing and managing their own AI and data products, such as synthetic trends and machine learning models derived from these trends.

AI agents: Representing a leap towards more dynamic and intelligent systems, AI agents are designed to operate semi-autonomously within data environments. These agents can handle tasks ranging from data monitoring and management to more complex decision-making processes. Critically, AI agents operate within predefined guardrails and logic established by human experts, enabling automation while preserving oversight. They can be particularly effective at executing established operating guidelines, with the potential to support real-time decision-making and operational governance.

The architecture of a data fabric aligned to a broader data mesh strategy, which we call a secure health fabric in this context, involves the creation of segregated workspaces, where each team operates independently under a unified technological framework. We find this setup allows for the customization of AI applications to meet specific needs without compromising the security and governance standards of the overall platform and core principles. Each workspace produces its own data products, which are developed, tested, and validated independently of others, thereby enhancing security and specialization.

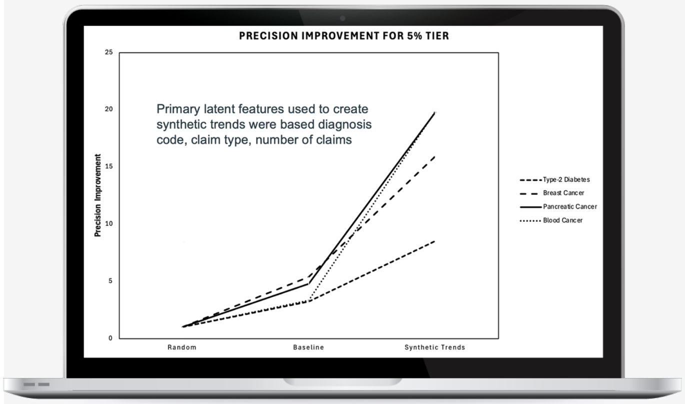


These independent workspaces in a secure health fabric are governed by a set of rigorous protocols and checks that ensure all operations are aligned with the highest standards of data protection and AI security, many of which are designed to be executed by AI agents. To ensure accountability, our architecture includes human-in-the-loop checkpoints and operational monitoring systems, reinforcing trust in how AI agents are used to manage AI -enabled workflows. This governance model is about overseeing operations and enabling them through supportive and clear, human-defined guidelines and oversight, aligned with both regulatory requirements and organizational objectives, while respecting core principles established within a broader data mesh strategy.

The secure health fabric we have developed is more than just a data management system; it's a comprehensive solution designed to support the deployment, management, and scaling of AI applications in healthcare. By ensuring that each component of the system is managed securely and efficiently — from data input and processing to model deployment and monitoring — we provide a foundation that supports current healthcare AI applications and is adaptable to future innovations and challenges. This strategic approach to AI platform operations sets a new standard for how AI can be integrated into healthcare, ensuring that it is both powerful and defensible.

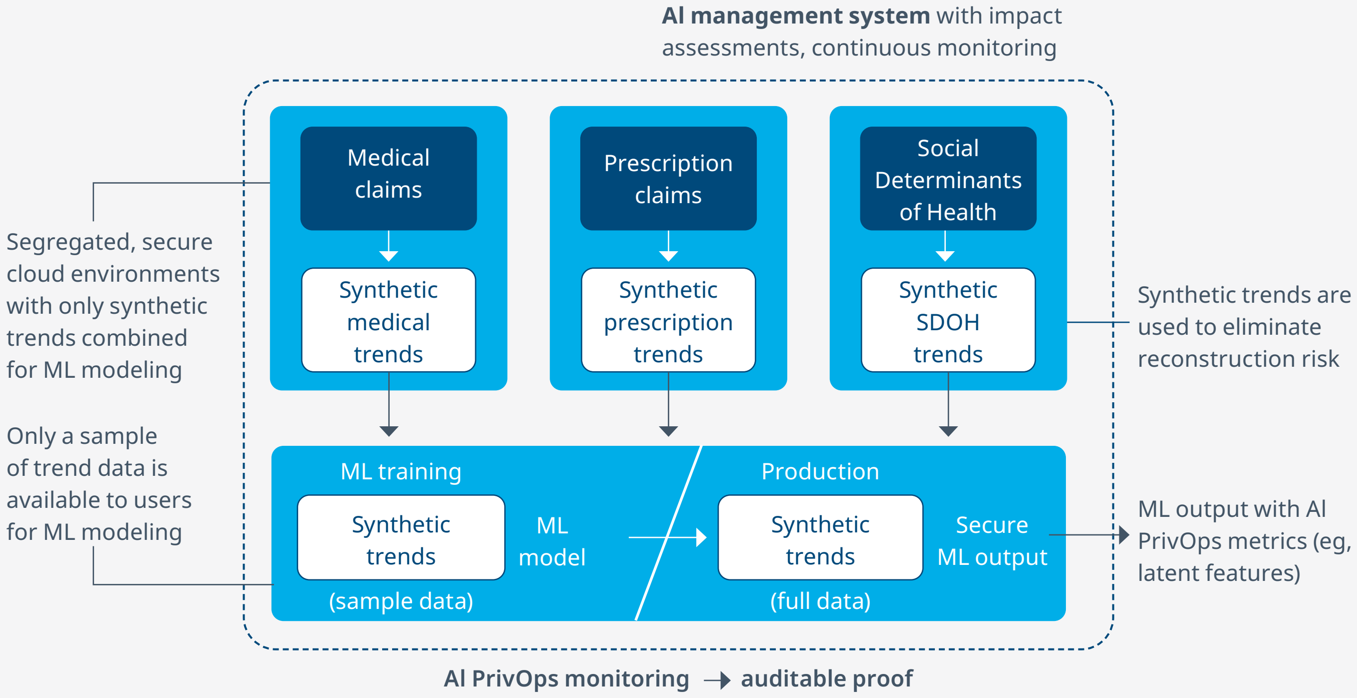
Case Study

Let's explore one example where health data was unavailable for machine learning due to AI and data protection concerns. We built a platform that introduced synthetic health trends to dramatically improve machine learning precision and unlock revenue growth potential exceeding 10x what was previously possible. Within a federated architecture, the introduction of synthetic health trends increased machine learning precision by more than 16x, without changing the baseline machine learning approach that was already in use and without feature engineering. This breakthrough enabled the business to reach previously inaccessible markets, with data that was otherwise unavailable, transforming both performance and growth trajectories.



A key feature of our secure health fabric is its reproducible architecture, which allows for plug-and-play functionality to support diverse AI use cases, such as digital health applications. This flexibility is crucial for adapting to the evolving needs of the healthcare sector, where different scenarios may call for bespoke AI solutions. Our platform differs significantly from other offerings by providing a robust, ready-to-use infrastructure that can be easily customized and scaled.

With this setup, a federated design was used to keep confidential datasets segregated while combining insights through synthetic trends. This was accomplished through a hybrid model that incorporates both horizontal and vertical strategies from federated learning, a privacy-enhancing technology encouraged by data protection authorities. This framework allows the combination of synthetic trends across different functional areas under strict security controls, ensuring that individual data points remain confidential while still providing valuable aggregated insights.


AI management system with impact assessments, continuous monitoring
Medical claims
Prescription claims
Social Determinants of Health
Synthetic medical trends
Synthetic prescription trends
Synthetic SDOH trends
Segregated, secure cloud environments with only synthetic trends combined for ML modeling
Synthetic trends are used to eliminate reconstruction risk
ML training
Production
Synthetic trends
ML model
Synthetic trends
Secure ML output
(sample data)
(full data)
Only a sample of trend data is available to users for ML modeling
ML output with AI PrivOps metrics (eg, latent features)
AI PrivOps monitoring → auditable proof

In horizontal federated learning, we align data by columns across different datasets, computing statistics such as means from each dataset and then pooling these to derive global insights. Conversely, our vertical federated learning approach aligns data by rows, matching records across datasets without ever combining them directly, instead calculating trends from each dataset and appending these to create comprehensive individual profiles.

We have presented an ambitious plan to accelerate this growth further with novel feature enhancements that will redefine what is considered best practice for machine learning and precision with sensitive information. We are expanding the platform to incorporate advanced predictive modeling techniques for longitudinal data and introducing synthetic trend reinforcement learning. Depending on the use case, modeling can occur on original attributes where permissible, and on synthetic trends to augment predictive accuracy without disclosing sensitive attributes.

Final thoughts

In the rapidly evolving landscape of healthcare, the strategic integration of defensible AI platforms represents a transformative approach — one that can drive significant improvements in patient care, operational efficiency, and innovation. To achieve these benefits, healthcare organizations will want to embrace scalable AI management frameworks, robust data security practices, and thoughtfully designed data architectures, each grounded in rigorous controls and standards.

Effective AI control management, through systematic alignment of system functionalities with regulatory requirements, ensures that healthcare innovations are impactful and trustworthy. Employing advanced methods such as synthetic trends within abstracted data spaces allows for secure, insightful analytics that respect patient privacy and comply with stringent AI and data protection expectations. Additionally, the thoughtful integration of centralized governance and decentralized innovation within data platforms — such as through a secure health fabric — offers healthcare organizations both flexibility and control, fostering a responsive yet stable AI ecosystem.

Ultimately, we believe organizations that adopt these principles will be better positioned to lead in the responsible use of AI. By embedding defensibility into the fabric of their data systems, and drawing on proven methods and expert collaboration, healthcare providers and public institutions can drive innovation while upholding the highest standards of patient care and trust.

To learn more

Our approach to transforming data into a secure representation, aligned with the stringent AI and data protection standards and patient expectations regarding the use of their information, is available with the *IQVIA Synthetic Trends Engine*. It's enabled by our secure health fabric and AI Governance and Privacy Operations (AI PrivOps) monitoring to produce auditable proof of continuous oversight and protection. The Synthetic Trends Engine can be used alone in a data cleanroom or in a federated learning approach, for a variety of health and wellness applications that require robust implementation of privacy and security measures against emerging AI threats. IQVIA Applied AI Science is a leader in developing advanced AI methods and platforms, powered by Privacy Analytics for third party assessments and privacy operations monitoring.

References

1. Yacoubian, C. 3 Guiding Principles for Defensible & Impactful AI in Healthcare. IQVIA Blog. 2025 April 01. Available at: https://www.iqvia.com/blogs/2025/04/3-guiding-principles-for-defensible-and-impactful-ai-in-healthcare

2. Safari, M. Privacy Strategy: Accelerating the Journey to Big Data Outcomes. IQVIA Insight Brief. 2024 May 15. Available at: https://www.iqvia.com/locations/united-states/library/insight-brief/privacy-strategy-accelerating-the-journey-to-big-data-outcomes

3. Arbuckle, L. A New Standard for Anonymization. IAPP Blog. 2023 March 14. Available at: https://iapp.org/news/a/a-new-standard-for-anonymization/

4. Biswal, D. An Integrated Approach to Securing AI. IQVIA Blog. 2024 Oct 1. Available at: https://www.iqvia.com/blogs/2024/09/an-integrated-approach-to-securing-ai

5. Arbuckle L, El Emam, K. Building an Anonymization Pipeline: Creating Safe Data. O'Reilly Media. 2020 April 13

6. Mian, M. Managing AI in Practice: A Structured Approach to Reliable and Defensible Systems. IQVIA Blog. 2025 Mar 31. Available at: https://www.iqvia.com/blogs/2025/03/managing-ai-in-practice

7. Reed, J. How Pharma Companies are Solving Regulatory Challenges with AI-based Technology. American Pharmaceutical Review. 2024 March 11. Available at: https://www.americanpharmaceuticalreview.com/Featured-Articles/611954-How-Pharma-Companies-Are-Solving-Regulatory-Challenges-with-AI-based-Technology/

8. Arbuckle L, Ritchie F. The Five Safes of Risk-based Anonymization. IEEE Security & Privacy. 2019 Aug 30;17(5):84-9

About the author

LUK ARBUCKLE

Global AI Practice Leader and Chief Methodologist, IQVIA Applied AI Science

Luk has written books, holds multiple patents, and led a technology team for a data protection authority. He draws from an extensive background in AI science and privacy technology with 25+ years experience. He is regularly consulted by senior executives and industry leaders across a spectrum of industry verticals for his ability to translate complex systems into simple and effective language. He also has 10 years experience in international standards development. Luk provides strategic leadership to clients and to IQVIA on the methods to get the most value from data with defensible AI.

About IQVIA Applied AI Science (AAIS)

IQVIA Applied AI Science (AAIS) draws upon its award-winning platforms and unmatched expertise to provide off-the-shelf and tailored AI solutions. Its customers are empowered to access the full potential of artificial intelligence, confident their AI tools — and the data fueling them — are aligned with regulations and fit for purpose. Novartis, NHS England, and Endeavor Health are among the many world-class brands accelerating innovation for a healthier world with AAIS.

CONTACT US

defensibleAI@iqvia.com

iqvia.com/contact