Deloitte. | databricks

REALISTIC SYNTHETIC CLAIMS
DATA AND PROVIDER NOTE GENERATION ON DATABRICKS

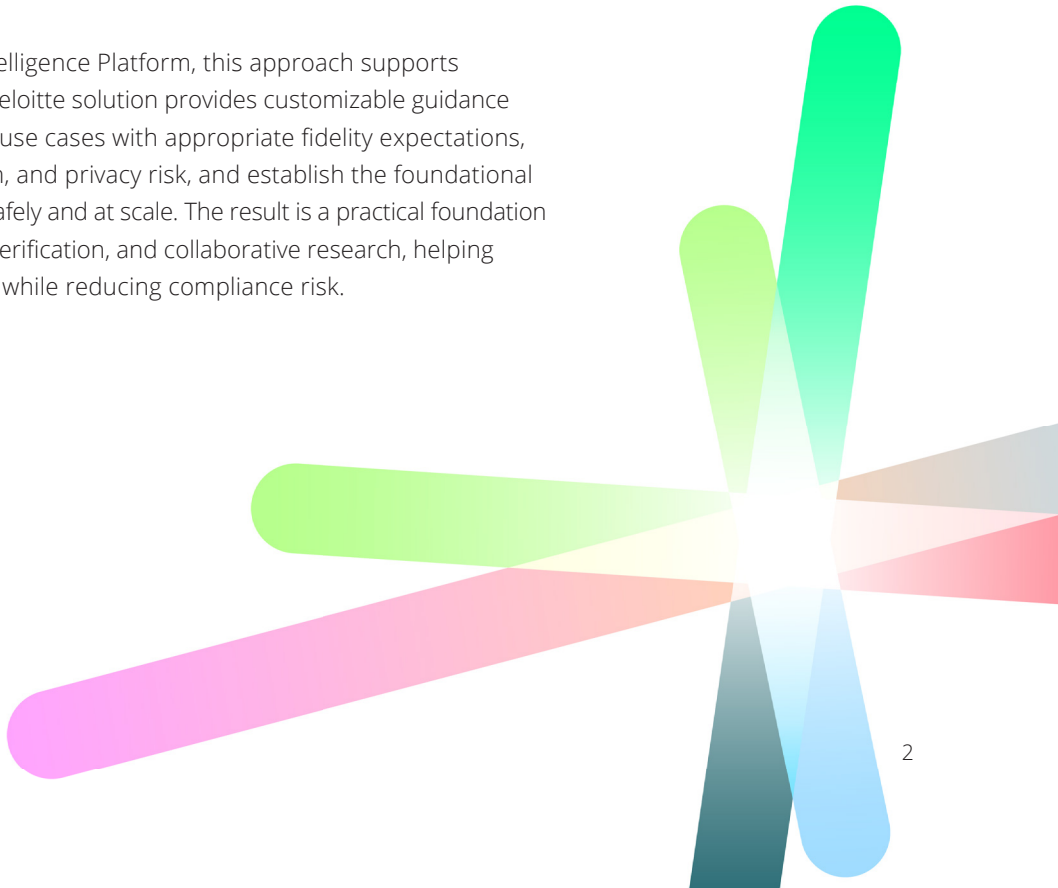
EXECUTIVE ***SUMMARY***

Access to large-scale, realistic health care data is often slowed by legal, privacy and compliance constraints. Synthetic data can help organizations innovate faster by replicating the structure and utility of real patient records while reducing exposure to sensitive information.

Health care data presents a dual challenge: preserving complex relationships in structured data—such as demographics, diagnoses and procedural codes—and generating medically plausible free-text clinical narratives, such as provider notes.

Our approach addresses both needs through a dual-model solution that combines Generative Adversarial Networks (GANs) for structured data and Large Language Models (LLMs) for text generation and automated quality review. GANs learn clinically meaningful representations from embeddings of concise descriptions derived from ICD diagnosis codes, capturing relationships in the vector space to maintain structural integrity and realistic correlations. LLMs generate contextually appropriate clinical notes and serve as artificial intelligence (AI) automated "judge" models to evaluate note quality (e.g., medical plausibility, relevance and coherence).

Integrated with the Databricks Data Intelligence Platform, this approach supports repeatable generation workflows. The Deloitte solution provides customizable guidance that helps teams align their healthcare use cases with appropriate fidelity expectations, apply practical checks for utility, realism, and privacy risk, and establish the foundational practices needed to use synthetic data safely and at scale. The result is a practical foundation for safe model development, workflow verification, and collaborative research, helping teams accelerate health care initiatives while reducing compliance risk.



BACKGROUND AND ***INTRODUCTION***

Rich, detailed patient data is essential for health care innovation: it enables predictive modeling, software validation and training. However, real patient records come with real-world constraints. Privacy requirements, legal agreements and sensitivity concerns can create months-long delays—and even then, teams may still lack comprehensive datasets for development, analytics or operational improvement.

To bridge this gap, we generate synthetic data that both resembles real-world patient encounters and is designed for safer use and sharing. Synthetic data is generated in two primary forms: tabular data—such as demographics, diagnoses, procedures, claim charge amounts and billing fields—and narrative data, such as provider notes and post-visit summaries. When deployed on the Databricks Data Intelligence Platform, these assets can be governed, versioned and shared through controlled workflows.

For synthetic data to be useful, it must maintain realistic relationships; symptoms should align with diagnoses and narrative, clinical notes should logically reflect the patient journey—from first presentation through long-term care and resolution.

Medical data complexity makes this difficult. Decades of clinical nuance, hidden correlations and massive, high-cardinality code sets mean simple anonymization is not sufficient for many use cases. We address this by combining GAN-based structured data synthesis with fine-tuned LLMs for clinical narratives, supported by multi-layered privacy and quality validation.

This approach is designed to retain realistic relationships while reducing the likelihood of tracing records back to individuals in a source dataset. The result: datasets that are medically accurate for intended development and testing use cases, privacy-preserving through layered controls and release gates, and safe for innovation and analysis.

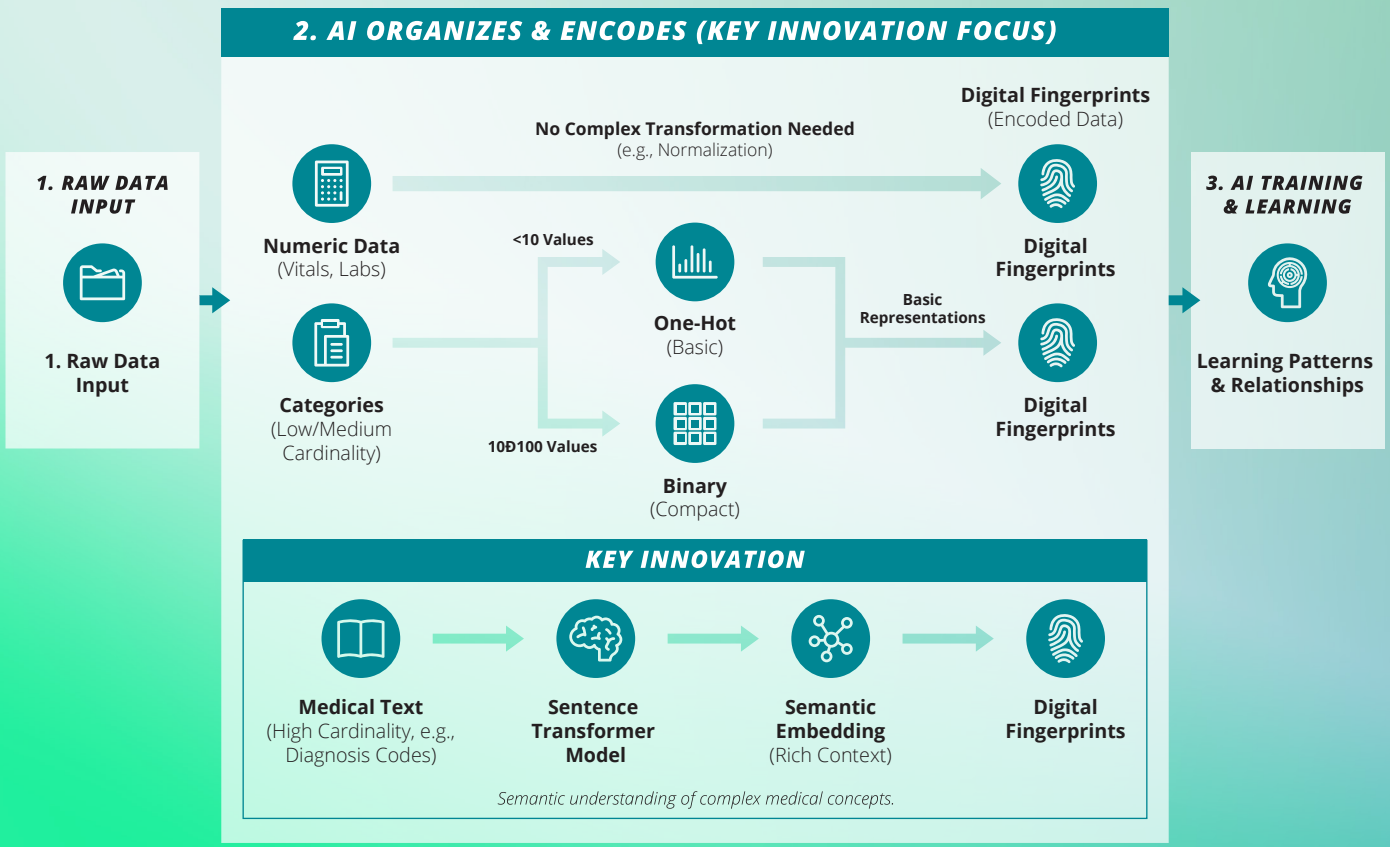
THE VALUE OF MULTIPLE **GENERATIVE MODEL APPROACHES**

Generating high-fidelity tabular data

Health care data is complex, with rare clinical events and high-cardinality code systems like ICD-10. Many standard synthetic data techniques, including traditional GANs or variational autoencoders (VAEs), struggle to maintain fidelity across these domains. Our approach integrates several techniques to improve realism and utility:

- **Advanced data encoding:** Autoencoders help pre-compress the feature space. Pre-training a combined (numerical + categorical features) autoencoder ensures behavioral integrity during compression.
- **Transformer-based models:** For code-heavy features (like diagnosis and procedure codes), pre-trained sentence transformer embeddings capture clinical similarities better than one-hot or binary encoding.
- **Customized GAN training with safeguards:** GAN training is tuned for health care feature spaces, with a training regime tailored to reduce the chance of reproducing exact patient source records while still maintaining the overall distribution.

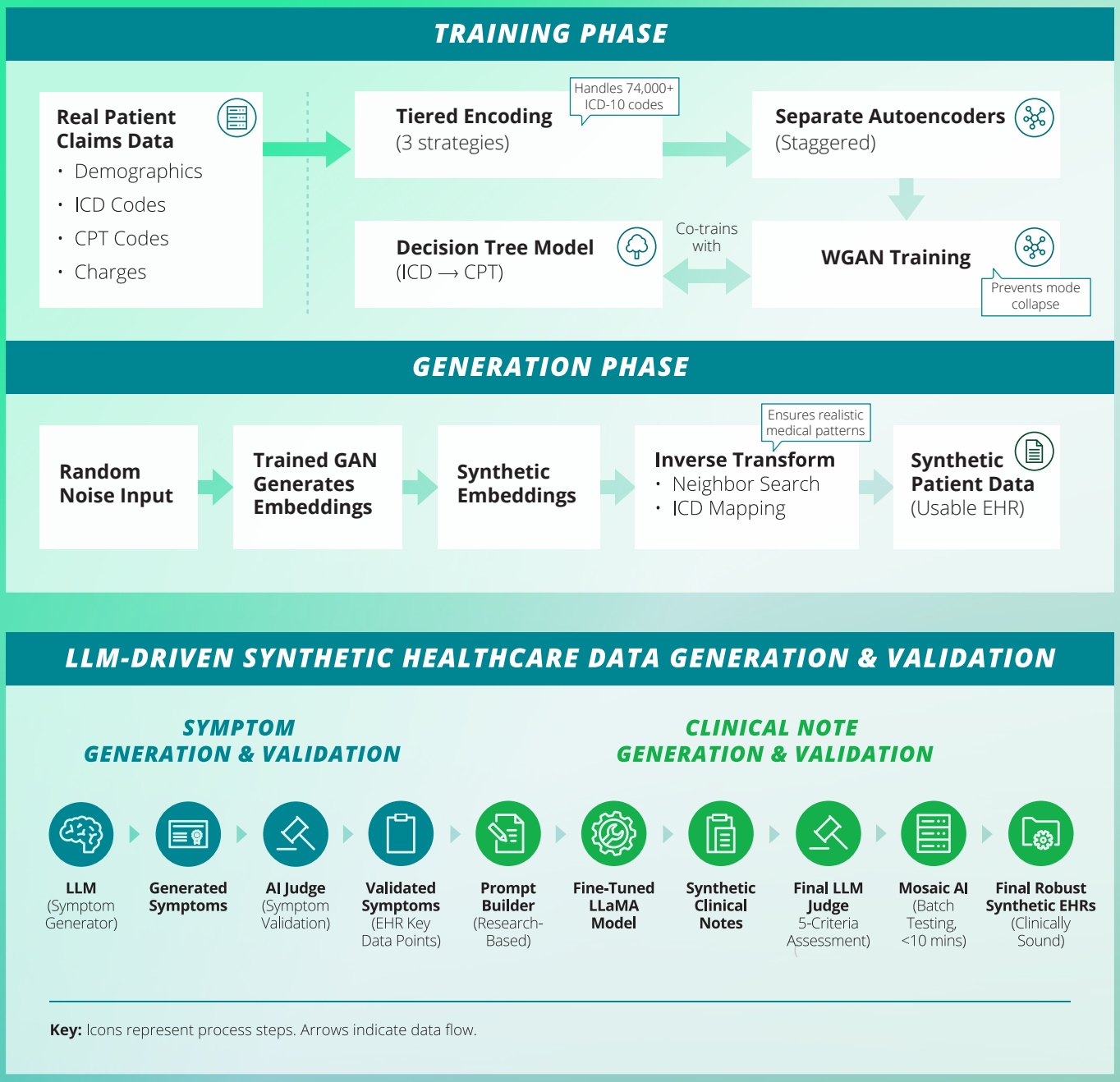
Our training and validation approach is designed to ensure that synthetic claims outputs are consistent, reviewable and ready for governed reuse on Databricks.



Generating realistic clinical narratives with LLMs

Capturing the nuance of free-text clinical notes is important for training modern health care software and validating downstream workflows.

- **Symptom generation:** An LLM generates clinically plausible symptoms for each synthetic record; outputs are reviewed by an automated AI judge model for relevance.
- **Clinical note synthesis:** Validated symptoms become prompts for a fine-tuned LLM, such as Llama-family models, to produce detailed provider notes.
- **Quality review:** Synthetic notes undergo LLM-based evaluation, assessing medical plausibility, realism and coherence across a "virtual patient" journey. Automated quality review hinges on LLMs finetuned on medical data.



BUSINESS ***VALUE DRIVERS***

Realistic data on demand

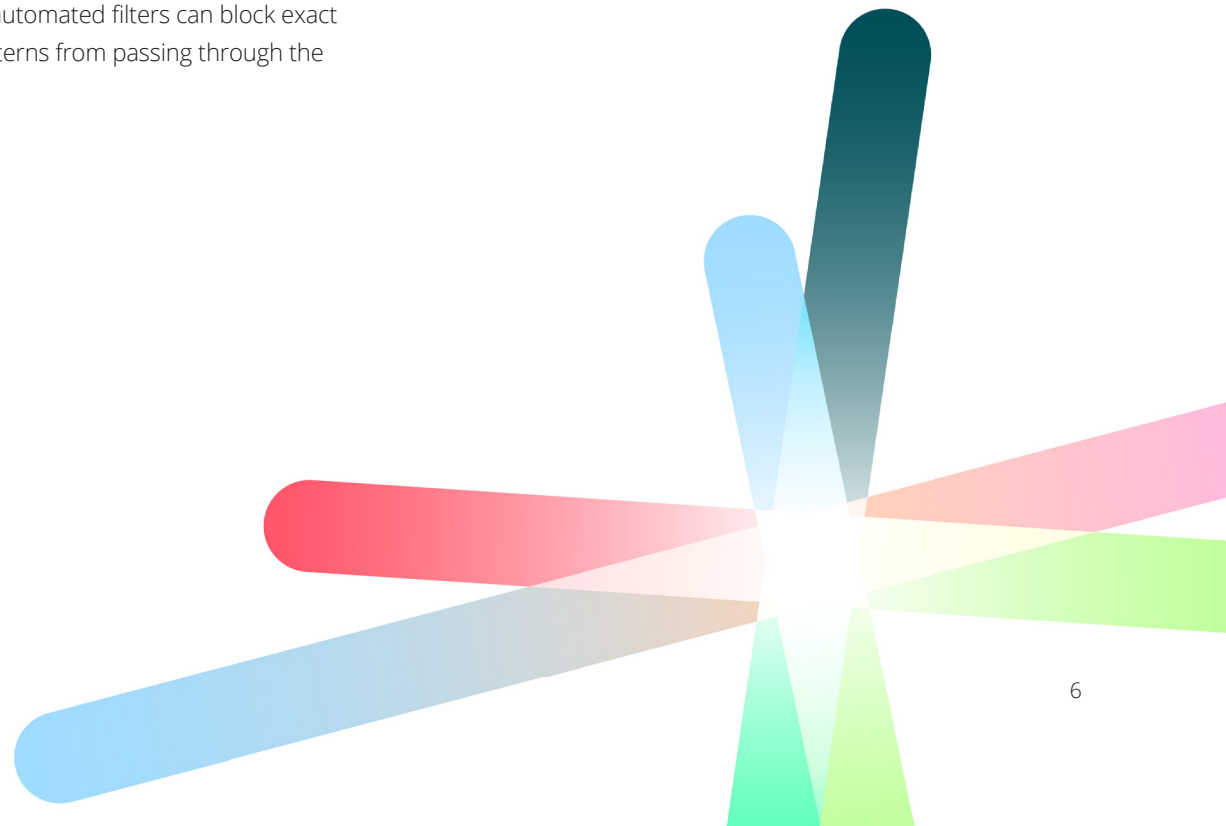
Synthetic data can accelerate access to robust datasets for analytics, modeling and application development. It can also help researchers, clinicians and developers reduce delays associated with legal review and the residual risk of using de-identified data for certain use cases. Teams can:

- **Accelerate model development:** Prototype and validate new AI/machine learning (AI/ML) models on high-fidelity data without accessing protected health information (PHI).
- **Verify clinical workflows:** Simulate patient journeys to stress-test electronic health record (EHR) integration, note parsing and clinical decision support (CDS) logic before deployment.
- **Enable collaborative research:** Share privacy-safer datasets across institutions to benchmark algorithms and conduct large-scale studies.

Reduced privacy and compliance risk (designed for safer sharing)

Our synthetic datasets are engineered to reduce exposure to direct and indirect patient identifiers and to help minimize re-identification risk through layered controls and release criteria. For example, personal details such as names and precise addresses can be systematically replaced with generalized values, and automated filters can block exact matches and unique patterns from passing through the synthetic data pipeline.

Audit-oriented artifacts, such as detailed logs and lineage information, can accompany each dataset to support traceability and regulatory readiness, particularly when data must be shared internally, with external partners or across delivery locations.



Reduced licensing friction: Speed, savings and flexibility

Real claims and clinical datasets may come with licensing restrictions, redistribution limits and strict usage constraints. Synthetic data can reduce these hurdles by accelerating access to representative data for development and testing—supporting operational agility while reducing dependence on lengthy contracting cycles.

Deep dive: Technical approach mapped to business needs

Producing synthetic health care data is both a technical and business challenge; success requires balancing realism, privacy and utility. The following illustrates how the approach maps to executive concerns:

- **Privacy protection:** Controls such as exact-match filtering, generalization (e.g., region-based), and embedding-based checks help reduce the likelihood that a synthetic record can be linked back to an individual.
- **Validation and benchmarking:** To assess statistical fidelity and note quality, each release is tested across multiple quality metrics, including root mean squared error (RMSE), earth mover's distance (EMD), nearest neighbor distance ratio (NNDR), and others. Where appropriate, results may be benchmarked against commercial tools, such as Datacebo, using a clearly defined evaluation protocol.
- **Scalable, transparent workflows:** Orchestrating end-to-end steps in Databricks supports tracking, lineage and repeatability—helpful for demonstrating controls and consistency to stakeholders.

Deloitte helps make this measurable and repeatable by codifying the validation harness, defining release thresholds tied to the intended use case, and integrating these checks into the Databricks workflow to consistently execute and evidence controls.

INFRASTRUCTURE & ORCHESTRATION ***WITH DATABRICKS***

Our approach relies on the Databricks Data Intelligence Platform to unify data engineering, model training and Generative AI (GenAI) deployment in a secure, governed environment—important for sensitive health care data. Primary architecture components include:

UNIFIED GOVERNANCE (UNITY CATALOG):

Databricks Unity Catalog provides centralized governance for data assets, models and embeddings, including:

- **Lineage and traceability:** Capture of data/model lineage to connect synthetic outputs to training data versions and model versions.
- **Secure access:** Centralized permissions across tabular and unstructured clinical data, enabling access to synthetic outputs while restricting source data.

SEMANTIC RETRIEVAL (AGENT BRICKS VECTOR SEARCH):

Vector search supports medical code integrity by:

- **Embedding management:** Transforming high-dimensional ICD codes into vectors stored in a managed index.
- **Inverse transformation:** Mapping synthetic vectors back to clinically valid codes to improve medical plausibility.

Our approach translates clinical validity requirements into retrieval/validation tests, including code-set constraints, outlier handling and plausibility checks.

EXPERIMENT TRACKING (MLFLOW):

Given the experimental nature of synthetic data generation, MLflow supports:

- **Model comparison:** Manage and compare GAN lifecycles, tracking hyperparameters and generative approaches.
- **Model registry:** Registering approved models to support reproducible dataset generation.

SCALABLE INFERENCE (AGENT BRICKS MODEL SERVING AND DATABRICKS AI FUNCTIONS FOR SQL):

Models can be deployed through optimized endpoints for generative Llama models (clinical notes) and "judge" LLMs.

- **Batch generation with AI functions:** High-throughput inference for population-scale synthesis by calling LLMs from SQL queries.

We also design guardrails and run-time checks—prompt templates, output validation and failure handling—to support safe, consistent note generation.

FUTURE ROADMAP AND PRODUCTION SCALING:

We plan to further automate and scale with:

- **Databricks Asset Bundles (DABs):** Infrastructure-as-code for pipelines and model configurations, supporting reproducible deployment and continuous integration/ continuous delivery (CI/CD).
- **Lakehouse monitoring:** Automated monitoring to detect data drift and track statistical validity over time.
- **Databricks LangChain integration:** Foundation for next-generation applications , such as retrieval-augmented generation (RAG)-enhanced use cases and multi-stage quality validation pipelines.

RESULT:

Deployment is designed to be rapid, reproducible and well-governed—minimizing risk while maximizing velocity.

CONCLUSION: UNLOCKING HEALTH CARE INNOVATION WITH ***HIGH-FIDELITY SYNTHETIC EHRS***

Our Databricks-powered synthetic EHR pipeline helps organizations unlock the value of health care data while reducing the associated privacy, security and access risks. By combining GenAI, rigorous validation and scalable infrastructure, we provide a dual-model approach that bridges the gap between data utility and patient privacy.

Primary synthetic EHR value: The primary value of this solution is expanded access to high-fidelity data for three core objectives:

- **Safe model development:** Build, train and validate clinical AI models on realistic datasets without exposing PHI and with governed access, versioning and release approvals.
- **Workflow verification:** Use simulated provider notes to test EHR integration, note parsing, and Clinical Decision Support (CDS) workflows before working with live patient records using repeatable validation gates tied to the intended workflow.
- **Collaborative research:** Decouple data utility from patient identity to enable broader benchmarking and joint algorithm development with documented permitted-use guidance and audit-ready evidence of controls.

Future potential: Extended use cases: Beyond data generation, high-fidelity synthetic data can also help teams explore next-generation clinical intelligence such as:

- **RAG-powered diagnostic support:** Use vector embeddings of synthetic notes to retrieve relevant historical cases for decision support in safe sandboxes before validating on approved real-world data.
- **Interactive case triage:** Apply statistical mining on synthetic EHRs to support real-time recommendations, symptom checks and triage guidance with clear guardrails on where synthetic data is, and is not, appropriate.

This solution offers more than just data—it offers confidence and repeatability. Leaders can better demonstrate the effectiveness of privacy safeguards and governance tools, while teams move faster to build, test and improve health care products and workflows.

Deloitte. | databricks

AUTHORS

Vashishtha J Bhatt
Engineering As a Service Senior Consultant
Deloitte Consulting LLP
vjbhatt@deloitte.com

Hemanth Yarlaggada
AI & Data Senior Consultant
Deloitte Consulting LLP
hemyarlagadda@deloitte.com

Andrew Wright
Lead AI and Data Science Engineer II
Deloitte Consulting LLP
andrwright@deloitte.com

David Heagy
AI & Data Consultant
Deloitte Consulting LLP
dheagy@deloitte.com