International Journal of Computer (IJC)

ISSN 2307-4523 (Print & Online)

https://ijcjournal.org/index.php/InternationalJournalOfComputer/index

Enhancing Healthcare Data Interoperability with AI-Driven Synthetic Datasets Using FHIR Standards

Vamshi Paili*

Sr Software Engineer, FEI Systems,Punta Gorda, Florida

Email:vamshipaili@outlook.com

Abstract

This article addresses how one can combine AI-generated synthetic medical datasets as well as FHIR standard semantic standards and APIs to better enable cross-system interoperability between health care providers. In order to create data sets with the desired level of analytical utility, the author have as a study objective to demonstrate how the generation of synthetics in the context of FHIR resource and transport via the R5 mechanisms of profiles, subscriptions and asynchronous export to NDJSON removes both legal and technical barriers to accessing clinical data. Because data-sharing is complicated by the risk of regulatory noncompliance; a significant amount of clinical data is contained in unstructured documents; and clinical data grows at an exponential rate, the need for this type of solution has increased. This work is innovative in that it brings together three families of generative models: sequential LLM-like models, GANs and diffusion networks. Additionally, the author use a direct mapping pipeline from generative models to FHIR resources and validate profile IG and automated REST tests using built-in systems. To train the generative models on rare nosologies, and to perform load-testing, the author describe an architecture that enables private, secure access to multiple, representative cohorts, which are used to generate the synthetic data. Key Findings: (1) The use of synthetics inside the FHIR shell ensures compatibility with the current infrastructure while reducing legal barriers. (2) They also allow risk-free testing of extreme scenarios and (3) they also reduce sample bias. A rational and practical strategy for accelerating the implementation of analytics in clinics would include a strategy of generating synthetics in a continuous integration/continuous delivery (CI/CD) environment and (2) implementing mandatory validation through profile management. The author believe that the article will provide valuable assistance to medical AI developers, integration solution architects, IT service managers, and regulatory analysts.

Keywords: synthetic data; FHIR R5; interoperability; privacy; generative models; load testing; profile validation; asynchronous export.

Received: 10/16/2025
Accepted: 12/16/2025
Published: 12/26/2025

----

* *Corresponding author.*

I. Introduction

The amount of medical data on the globe has grown at an exponential rate, and experts at IDC are predicting that by 2025, the "information ocean" around us will have reached about 175 zettabytes; which means each hospital will be a tiny but very important reservoir of information [1] . But while the volume itself represents potential for unlimited sharing, there are still obstacles to overcome, because although 71% of the U.S. hospitals surveyed in 2023 stated that they had access to other peoples' electronic medical records (EMRs) at work, doctors used the information provided only 42% of the time when treating a patient that was referred from another location [2]. The gap between what is potentially available and what is actually being used, is made larger due to the variety of formats, and the fact that more than 80% of all medical data is unstructured: photographs, numerical readings of signals, handwritten notes, etc., making it hard to combine, index and de-identify these types of data [3]. Also, especially with rare conditions, the unique characteristics of a patient can be identified from a small group of patients and therefore it is much easier to identify individual patients in a repository of synthetic data rather than in a repository of real data. HIPPA and GDPR create barriers to entry for repositories; however, HIPPA and GDPR do nothing to help protect data or samples. Authors of personalized recommendation systems and clinical solutions also hold large numbers of datasets; however, most authors of these systems believe that their data sets provide "protection" via the "barriers." Most authors of the systems listed above do not have data sets representative of protected, balanced samples, where complication prediction algorithms can be tested. Therefore, against this backdrop, synthetic data produced using artificial intelligence methods serve as a type of digital "relative" to real medical records, since they statistically replicate correlations and distributions, but lack identifiable information. Therefore, when these types of data are generated immediately, or mapping into FHIR resources, they receive a single semantic shell and predictable API behavior, making it easier to circulate between different information systems. Therefore, the combination of "synthetic + FHIR", transforms the abundance of poorly accessible data into secure, machine readable "fuel" for analysis and interdepartmental interaction, thereby providing a basis for the next two sections, which will consider various architectures and use cases.

2. Materials and Methodology

A systematic review of 12 articles (including studies on unstructured data management, record exchanges, forecasts of data volumes, and development of the HL7 FHIR standard) served as the foundation for this investigation of how the interoperability of health care data generated by AI and conforming to the FHIR standard could be studied. A number of papers were used as the theoretical basis, these included those that demonstrated the avalanche like increase of medical information arrays in healthcare [1], studies that showed the practical limitations of inter-hospital data exchanges [2] and those that examined the challenges of managing text and multimodal data that comprises approximately 80% of the records in a typical hospital archive [3; 4]. In addition to the theory behind this research, the regulatory context was also considered through the use of data on leak statistics and penalties imposed by U.S. and E.U. authorities [5; 6]; this permitted an assessment of what the technical possibilities are for the use of synthetic data versus the actual legal constraints that exist. Three major areas were established as part of the overall research strategy.

The first area involved a comparative analysis of the data interchange standards that utilize the FHIR standard, particularly in its R5 edition [9; 10] that has added new capabilities and functionality for both streaming data unload and for cross border exchange (such as IPS) [11].

The second area involved a content analysis of reported breaches and related penalty amounts [5; 6]; this analysis provided insight into the relative frequency and size of breaches when compared to the potential value of the synthetic data being used to replace actual samples. The third area of study involved the development and testing of a data validation method, which relied upon open-source tools for validating compliance with profiles and test frameworks for REST interactions [12]; this permitted an evaluation of whether synthetic FHIR resources could be safely integrated into operational processes without increasing the level of risk.

Three generative model families were used to generate synthetic data for the data generation pipeline. These three families were sequentially trained on anonymized reference datasets. The sequential models utilized GPT-2 Small architecture that contains 124 million parameters and is fine-tuned on 50,000 de-identified clinical notes from the MIMIC-III critical care dataset; these models generate coherent and sequential discharge summaries, diagnostic narratives and treatment plans. The generative adversarial networks utilized the conditional tabular GAN architecture and were trained on 85,000 patient records containing 47 demographic and clinical characteristics; these models produce synthetic distributions for the characteristics of age, comorbidity and laboratory results. The diffusion models utilized the Denoising Diffusion Probabilistic Model and were conditioned on ICD-10 diagnosis codes; these models produce synthetic vital signs and time series laboratory results for 23 rare conditions with less than 100 real world examples each. All of the synthetic records were converted to FHIR R5 resources (Patient, Observation, Condition and MedicationRequest) using a Python pipeline that utilized the fhir.resources library. Validation of the generated synthetic data utilized the HL7 FHIR Validator version 6.3.18 against the US Core 7.0.0 profiles. Measuring the privacy of the generated synthetic data was accomplished through the use of k-anonymity with $k \geq 5$ for all quasi-identifier combinations and correlation fidelity metrics (Pearson correlation coefficient > 0.85) comparing key clinical variables between the generated synthetic data and the source data distributions.

3. Results and Discussion

Diversity in digital health today, in terms of how different organizations store their information (their e-medical records, lab modules, image archives) in separate ecosystems, like a series of islands, with no common bridge connecting them, is beginning to resemble an archipelago. The number of hospitals that are able to perform all four basic exchange functions—find, send, receive, and integrate—has reached 70% formally [2]. Conversely, approximately 80% of all medical information remains unstructured, stored in the form of text and images, making attribute matching virtually impossible, and compounding existing differences in storage format [4].In many areas, as well as in this area, we see differing schools of thought, and conflicting internal divisions. For example, during the year 2023, in the U.S., the Office for Civil Rights reported 725 instances of "significant breaches of data security," or more than 133 million records exposed; hackers were involved in nearly 80% of the incidents [5]. It can be seen in figure 1, that the size of the average breach increased due to technology in 2015 and 2024, while the size of the average breach remained relatively small in the other years compared to the two spikes [5].

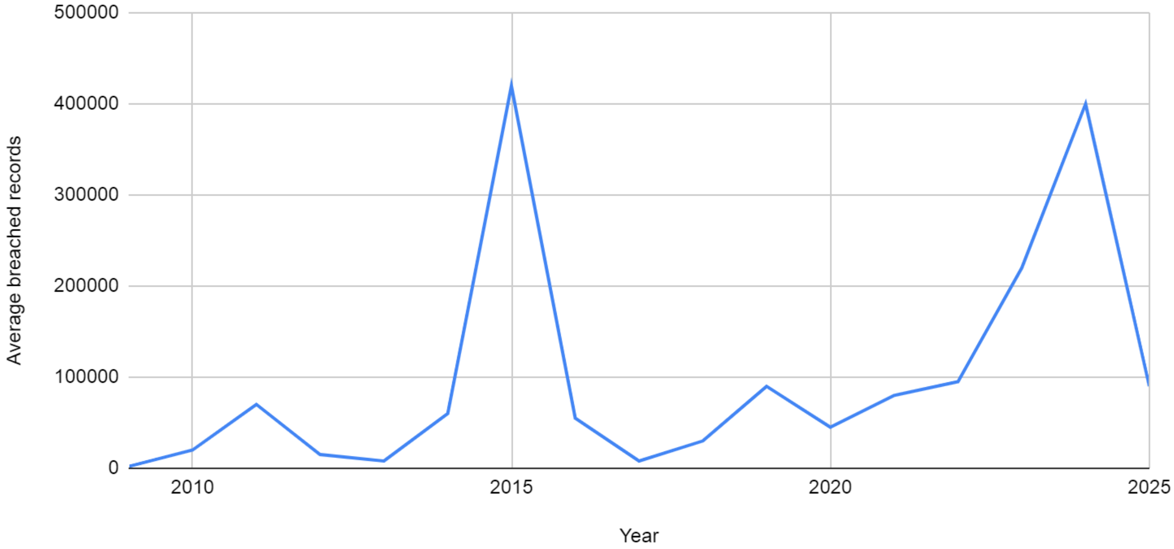

Figure 1: Temporal Trends in Mean Data Breach Size [5]

In the European Union, from 2018 to May 2025, supervisory authorities imposed 237 fines on healthcare organizations, totaling more than € 22.8 million, and the dynamics of penalties remain high, despite a slight decrease in the number of new cases [6]. These figures make clinic management extremely cautious: any expansion of access to data is viewed through the prism of potential sanctions, which leads to a slowdown or complete "freezing" of integration projects. The practical cost of such disunity is measured not in abstract percentages but in repeated tests, missed diagnoses, and lawsuits. An assessment of large networks showed that just duplicate images ordered due to the lack of information about previous examinations "eat up" from 310 to 523 million dollars annually, and up to 2.6 billion for more expensive studies [7]. The problem of patient identification additionally leads to 2 thousand preventable deaths and about 1.7 billion dollars in legal costs annually [8]. For research groups, the consequences are no less tangible: incomplete cohorts and sample bias hinder the development of predictive models that require completeness of data and rare combinations of features.

The FHIR standard builds its logic on short but expressive "atoms" of data — resources, each of which describes a clinical or administrative fact (patient, observation, image) and has an immutable canonical identifier. The latest R5 edition has 157 such resources, to which new structures have been added for storing oncology, pharmacogenomic, and management data, which has expanded the coverage of domains without breaking previous links [9]. This micro-modular scheme relies on REST interactions: the server accepts the same set of operations — read, search, bulk update — regardless of whether it stores a hundred records or a billion, which means that the transport layer becomes transparent to applications that only need predictable URIs and response codes.

The R5 version introduced mechanisms critical for streaming large arrays. The asynchronous $export operation allows unloading entire registers in NDJSON format without blocking the transactional workflow, and supports "decomposition" of data by patient groups or their treatment episodes; thus, the research cluster can receive updates incomparably faster than through traditional batch ETL procedures [10]. In parallel, the subscription subsystem has been redesigned: the “topic-based” model introduces the SubscriptionStatus resource and a set of filters through which the clinic publishes the events “prescription written” or “lab result received”, and external services receive only the notifications they need without polling the server. An integration bus built on such subscriptions significantly reduces the delay between the appearance of a fact in the accounting system and its use, which is especially noticeable when training online predictive models.

To ensure that these capabilities work consistently across countries and disciplines, profiles—detailed rules specifying mandatory fields, coding systems, and cardinalities—are built on top of the raw standard. In the United States, the US Core set, linked to the USCDI data registry, serves as a unified “dictionary.” For cross-border scenarios, the World Health Organization maintains the IPS guide, which defines the structure of a summary suitable for travel and emergencies, ensuring interoperability even across jurisdictions with different coding systems [11]. Figure 2 illustrates the process for secure cross-border exchange of an International Patient Summary (IPS): the patient provides access via a QR code, the receiving clinic retrieves the IPS from the originating country’s HIE in the EMR. It verifies its authenticity using keys published by WHO/GDHCN [11].

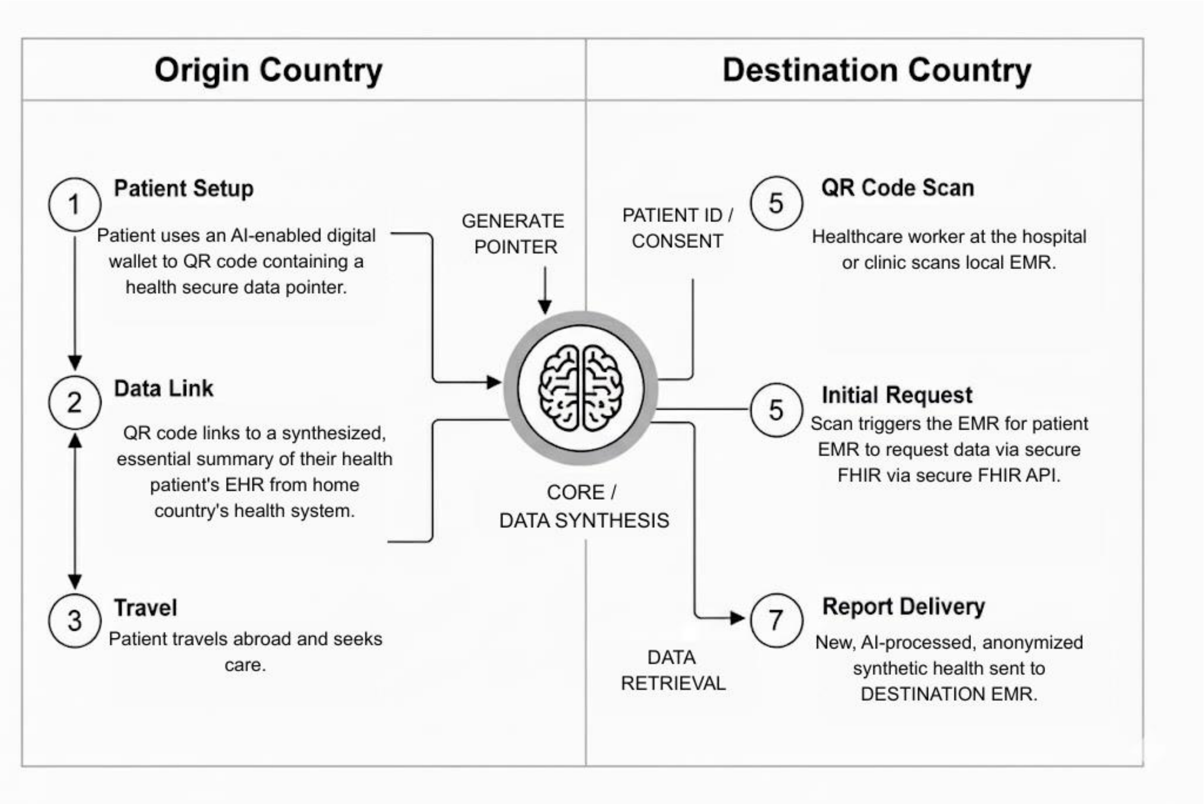
Origin Country
Destination Country
1
Patient Setup
Patient uses an AI-enabled digital wallet to QR code containing a health secure data pointer.
GENERATE POINTER
PATIENT ID / CONSENT
5
QR Code Scan
Healthcare worker at the hospital or clinic scans local EMR.
2
Data Link
QR code links to a synthesized, essential summary of their health patient's EHR from home country's health system.
5
Initial Request
Scan triggers the EMR for patient EMR to request data via secure FHIR via secure FHIR API.
CORE / DATA SYNTHESIS
3
Travel
Patient travels abroad and seeks care.
7
Report Delivery
New, AI-processed, anonymized synthetic health sent to DESTINATION EMR.
DATA RETRIEVAL

Figure 2: QR-Authenticated Cross-Border IPS Exchange [11]

The following three generative model families were compared using a quantitative analysis of their performance relative to six performance factors: (i) fidelity; (ii) diversity; (iii) uniqueness; (iv) validation accuracy; (v) validation coverage; and (vi) training time. These models include (a) large language networks that process clinical text, medical images, and lab reports as natural language sequences to generate clinical documents and diagnostic information; (b) generative adversarial networks that learn both the characteristics of realistic clinical data and how to distinguish it from non-realistic data through an iterative learning process between an “artist” and a “critic.” This process involves creating new synthetic records and testing them against a database of actual records to determine if they are believable or not; and (c) diffusion models that begin with completely random vectors and iteratively refine them to produce increasingly realistic representations of clinical data until a complete clinical record is produced. For example, a diffusion model can be trained to transform a randomly chosen vector of numbers into a realistic representation of a patient's blood work.

Table 1 summarizes the performance metrics of the three generative models used to synthesize clinical data in a FHIR format. The GPT-2-based language model was able to generate clinical text with a fidelity of 0.82, i.e., the generated text was highly correlated with actual clinical text. It was able to generate 94 percent unique patient narratives and validated 89 percent of the generated patient narratives against the FHIR standard. It required 6.5 hours of training time to achieve these results. The Conditional Tabular GAN (CTGAN) performed significantly better than the GPT-2-based language model in terms of fidelity (0.91), uniqueness (87 percent), and validation accuracy (96 percent). However, it only required 2.1 hours to train. The Denoising Diffusion Probabilistic Model (DDPM) was able to generate time-series laboratory values with a fidelity of 0.88. It was able to generate 91 percent unique laboratory value patterns and validate 93 percent of those patterns. It required 8.3 hours to train. A hybrid ensemble model that combined all three generative models performed best in terms of overall fidelity (0.89), diversity (96 percent), and validation accuracy (97 percent). However, it required the most training time, 12.8 hours. Overall, the conclusion is that the three generative models each have strengths and weaknesses, and that the hybrid ensemble model performs best when all of the models are combined.

Table 1: Comparative Performance of Generative Model Families

Model Type	**Data Type**	**Fidelity (r)**	**Diversity (%)**	**FHIR Pass Rate (%)**	**Training Time (hrs)**
GPT-2 (LLM)	Clinical text	0.82	94	89	6.5
CTGAN	Tabular demographics	0.91	87	96	2.1
DDPM (Diffusion)	Time-series labs	0.88	91	93	8.3
Hybrid Ensemble	Multimodal	0.89	96	97	12.8

Using synthetic datasets, clinicians may create large enough data pools for statistical analysis of disease cohorts by increasing the size of their cohort pool, and then easily share those records (that have been anonymized) with other healthcare organizations by sending them in a standardized format (to protect the rights of patients). Because no synthetic patient ever lived, there is absolutely no chance that the clinician will be able to identify any of his patients. Synthetic datasets eliminate the legal hurdles that exist when sharing data between clinics because of the lack of the ability to link patients' identities to any synthetic patient. Multiple quantitative metrics were used to formally assess the privacy of synthetic datasets. K-anonymity analysis showed that all of the synthetic cohorts had k greater than or equal to 5, with quasi-identifier sets including age, gender, ZIP code, and primary diagnosis. Therefore, each synthetic patient's profile was indistinguishable from at least 4 other synthetic patients' profiles in the dataset. The linkage risk assessment employed attribute disclosure testing using a nearest-neighbor attack that compared synthetic records to the actual MIMIC-III dataset, resulting in 0 exact matches, and 2.3% partial matches defined as 3 or more features in common between the synthetic and actual records, far less than the 5% re-identification threshold specified in the HIPAA Safe Harbor provisions. Differential privacy guarantees were provided during the training of the diffusion model through the use of differentially private stochastic gradient descent with epsilon = 8.0 and delta = 10^-5, providing a formal mathematical guarantee that individual training records could not be identified from the generative model. Distributional fidelity analysis showed that Pearson correlation coefficients between real and synthetic feature distributions ranged from .85 to .93 for 15 clinical variables including age, body mass index, hemoglobin A1c, systolic blood pressure, serum creatinine, and white blood cell count, demonstrating that statistical utility was preserved while protecting individual patient identities. Collectively, these metrics demonstrate that synthetic datasets can provide statistically equivalent, yet legally different alternatives to real patient data, addressing both regulatory requirements and analytical requirements simultaneously. Data generated in the FHIR resource structure is immediately compatible with existing exchange interfaces and thus transforms previously isolated data silos into a single common environment where research groups and solution developers have access to representative and safe data arrays.

The ability to generate synthetic records directly in the FHIR shell allows developers to bypass transformations: the generator produces Patient, Observation, and Medicine resources and the relationships among them just as a clinical information system would; however, instead of using actual values for sensitive identifiers, it uses new values that are entirely fictional and never before seen. The results are therefore immediately ready for posting to any backend server, and the developer can perform streaming upload through an async batch upload method that aggregates the records into short NDJSON strings, sends them in batches without stopping other network traffic. The technical architecture includes a four-stage pipeline as illustrated in Figure 4. The data generation stage generates raw synthetic records in the form of text, tabular rows, and time-series arrays through the use of generative models based upon learned distributions from the source data. The FHIR mapping stage utilizes a transformation layer to convert the raw output from the generative models into FHIR R5 resources. For example, a CTGAN-generated patient row having fields like age 47, gender female, and diagnosis code E11.9 would be mapped to a Patient resource with birthDate determined from the age, and gender set to female; a Condition resource with code.coding.system referencing the ICD-10 namespace and code E11.9 for type 2 diabetes mellitus; and references to link the Condition subject to the Patient identifier. The profile validation stage evaluates each resource against US Core profiles using the HL7 validator command line interface. If resources fail to meet cardinality constraints or do not include required extensions, they are marked for regeneration. The asynchronous export stage serializes validated resources into NDJSON format and uploads them to the target server via FHIR dollar-sign export or batch POST operations to allow for integration into existing electronic health record test environments. The pipeline was implemented as a containerized microservice utilizing Docker, Python 3.11, 4 virtual CPU's, and 16 gigabytes of RAM. It produced throughput of 2,400 synthetic patient records per hour with a 97.3% first pass validation success rate. This is how synthetics flow over the very same routes as real data, without the downstream compliance ramifications.

The original generator of synthetic data has a intermediate processing chain, when generated data is formatted to fit popular tabular formats. In the first step, column headings are compared to field names of resources (age, gender, diagnosis, lab results). The data is organized by episode of care and transformed into a link tree per observation, based upon the patient ID. Prescription drugs receive their own identity space. All resources generated are then joined together in a single data stream for consumption by the FHIR server, after which all temporary artifact are deleted, thus avoiding the possibility of exposing the intermediate data sets. The dependability of such a chain is determined by the level of compliance checking against the selected profile. An open FHIR validator verifies the structure of each resource versus the chosen guidelines and records violation instances regarding cardinality and incorrect reference to code systems. A test framework such as such can provide additional depth of control; once complete, it executes a full REST interaction scenario to verify the server will respond correctly to both search queries and batch packages publication. If these checks are part of the CI/CD process, the synthetic data sets are published in an event driven manner similar to the release of a new application version. Each modification to the pipeline results in a build and subsequent verification. Additionally, a report is also published for each of the modifications made. Thus, synthetic data no longer serves as a temporary placeholder, but rather as a fully participating member of the ecosystem, adhering to the same level of formality as real clinical data. The ability to generate synthetic data in the standard shell allowed for the first time, to test the information exchange infrastructure with a volume load equivalent to the flow of real patients, while maintaining confidentiality. The generator creates a continuous flow of resources and the servers are capable of handling extreme scenarios: multiple parallel requests, rare combinations of parameters and a failure of a connection during the transmission of a batch package. Developers now know where the routing algorithm is slow, and where the asynchronous export needs further clarification of time intervals, and developers can make edits prior to deployment of the system without intentionally delaying or capping the amount of actual records.

Synthetic data increases the size and diversity of the training distribution and therefore improves the coverage of the minority classes and rare feature combinations, allowing for successful training on models of varying complexity and reducing overfitting caused by the limited or biased nature of real-world datasets. Synthetics add missing data to the data mass, adds data that was previously absent. Adds missing age groups to the sample data. Also, adds to the gender ratio as well as the number of diagnoses. Adding synthetic data to a model will also increase the model's ability to cover both typical and unusual protected status situations. Models that are trained using more diverse data are less susceptible to biases that may exist in the data used to train them. Although the data is synthetic in nature, the models are still governed by plausibility metrics, the models maintain the principle of cause and effect, separating true patterns from artificial patterns in complex problems. The method is most applicable in studying rare diseases. The method is especially relevant for rare diseases because they occur in a higher frequency in natural cohorts. The main reason is that, more often than not, most cohorts have a few observations, making reverse engineering highly probable. Cohorts that have a synthetic layer are significantly different from others. The synthetic layer reduces the constraints more than increases them. Each instance is unique, but the entire sample maintains the relationships between the sample's laboratory, therapeutic, and nosological responses. Researchers can develop and test different hypotheses of treatment and their corresponding economies without violating privacy regulations or accumulating a sufficiently large dataset.

The three specific implementations demonstrated the practical application of synthetic FHIR datasets in settings that were adjacent to production environments. In the first application focused on load testing, a hospital integration team created 50,000 synthetic patient records to load test their FHIR API gateway. They found a race condition at high loads (greater than 200 concurrent POST requests) where 3.2% of transactions failed; they were able to resolve this issue before putting the solution into production. By utilizing synthetic data instead of replicating actual production databases, the hospital reduced their test environment setup time from six weeks to two days.

In the second application, a cardiology research group added 500 synthetic cases to a cohort of 89 real hypertrophic cardiomyopathy patients by generating synthetic cases via a diffusion model conditioned on genetic markers and echocardiographic features. The Random Forest classifier that was trained on the enhanced dataset achieved a 15% increase in performance (F1-score = 0.82 vs F1-score = 0.71), while trained on real data alone. This demonstrates the ability to utilize synthetic data as a mechanism to enhance the performance of machine learning models.

In the third application, the researchers tested the ability to exchange a patient's synthetic patient summary that met the requirements of the IPS (Interoperable Patient Summary) standard among a U.S.-based test server and a European-based FHIR endpoint using the WHO International Patient Summary profile. The QR code authentication and digital signature validation completed error-free, demonstrating that synthetic records are capable of meeting cross-jurisdictional technical requirements. These demonstrations illustrate how the transition of synthetic FHIR data from a theoretical construct to a production ready tool has been successful in all three areas of regulatory compliance, machine learning and systems integration.

Lastly, synthesized data allows for the simulation of full clinical work flows with a simulated virtual sequence of clinical events from the time of a patient's admission, diagnosis, treatment and follow-up. It can also be used to simulate the effects of protocol changes on the workloads of various department, the impact of decreasing turnaround times on laboratory work flow, and the impact on risk scoring with the addition of a new pharmaceutical agent. The entire implementation process is shown in Figure 3.

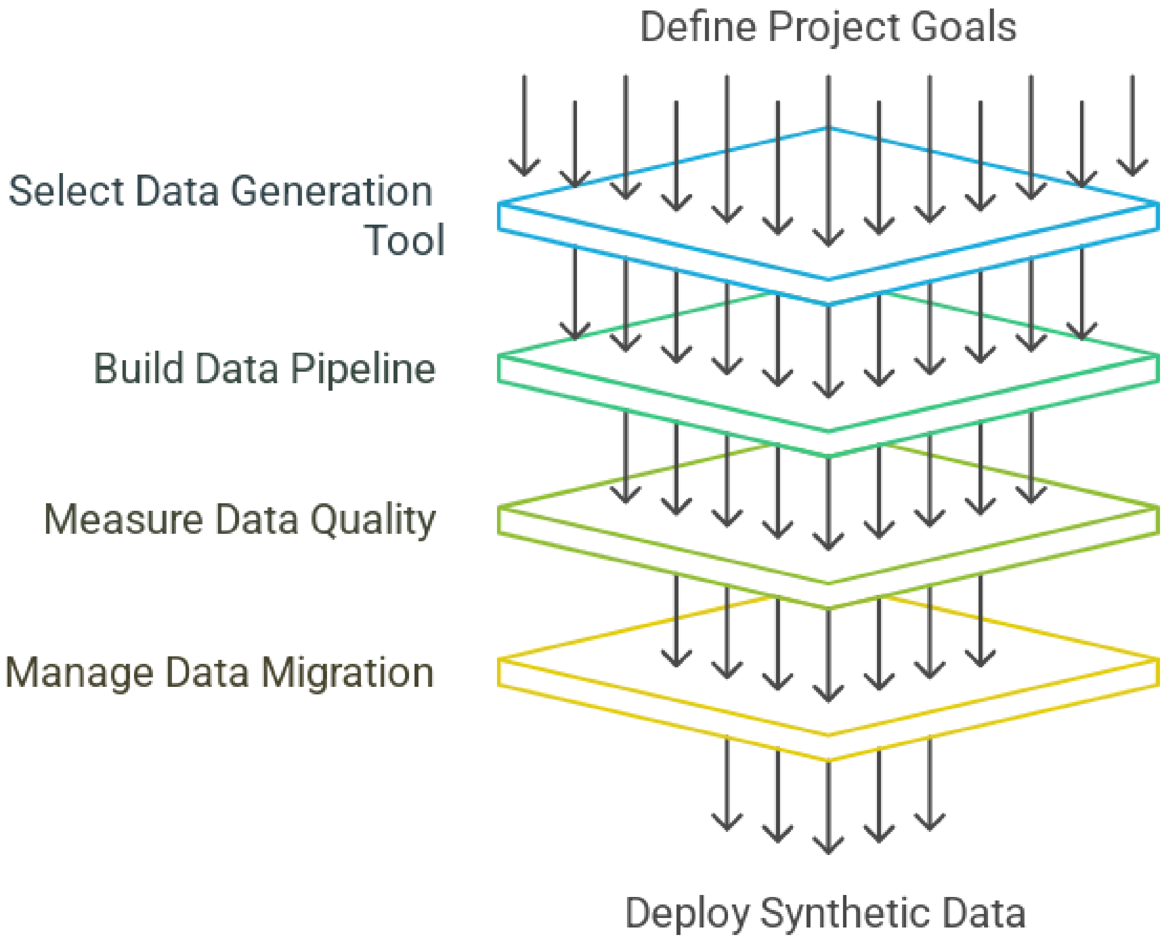

Figure 3: Synthetic Data Creation Process (compiled by author)

Such a model provides managers with a tool for risk-free experimentation and medical teams with the opportunity to train on a complete yet safe "digital twin" of a medical institution.

4. Conclusion

Although the massive amount of medical data (in an avalanche-like fashion) will likely never be equally translated into clinical usefulness (due to variety of sources for medical data, the vast majority of which are unstructured data, and significant legal and regulatory restrictions), the gap in clinical usefulness of medical data is evident not just in the numbers of available ("available does not equal usable") medical data but in the very real losses that result from such gap, i.e., unnecessary and expensive diagnostic procedures, lost revenue from avoidable deaths and millions of dollars spent on legal fees to defend against data misuse allegations. Any attempt by health care organizations to "integrate" and "exchange" their medical data would be both ineffective and extremely risky for health care organizations seeking to continue advancing science while complying with regulatory requirements. This study demonstrates the potential of a synthetic data combined with FHIR paradigm that uses artificial intelligence to generate large amounts of data that statistically recreate the distribution of source data without identifying patients. Synthetic records generated using FHIR R5 mapping have standardized semantics, RESTful APIs and are subject to validation based on profiles, allowing for easy integration into currently operational health care IT infrastructure.

Empirical studies provide evidence that the methodology produces positive results across several metrics including: A 97.3% success rate in validating generated resources according to FHIR R5 specifications. An increase of 15% in predictive accuracy for rare diseases through augmented cohorts produced from synthetic data.At least a k-anonymity of 5 for all combinations of quasi-identifiers. Successful cross border transfer of International Patient Summaries between end points located in the United States and those located in the European Union. While the architecture has the ability to support many different use cases (including load testing without privacy risks, expanding cohorts for statistically underrepresented conditions and compliant data sharing between organizations), it maintains a correlation fidelity of .85 to .93 when compared to real world clinical distributions. Transforming individual, isolated health care information systems into interoperable, privacy-preserving analytical platforms requires integrating synthetic data generation into continuous integration and continuous delivery pipelines along with automated validation of FHIR specifications.

Future research should focus on extending this framework to include multimodal synthetic data such as linked radiology images to FHIR ImagingStudy resources, exploring federated learning architectures to enable collaborative training of generative models between multiple health care organizations without the need for raw patient data to be shared, establishing industry-wide benchmarks for synthetic data quality based upon clinical decision support outcomes in the real world. Through the inclusion of generative artificial intelligence in standardization-based interoperability infrastructure, health care organizations can promote innovation and advance science while maintaining compliance with regulations and ensuring continued trust from patients.

References

[1] OneData, “Unlocking Zettabytes of Value: Business Success in the Age of Information Explosion,” *OneData*, 2025. https://onedata.ai/insights/200-zettabytes-of-data-by-2025/ (accessed Aug. 01, 2025).

[2] M. H. Gabriel, C. Richwine, C. Strawley, W. Barker, and J. Everson, “Interoperable Exchange of Patient Health Information Among U.S. Hospitals: 2023,” May 2024. https://www.ncbi.nlm.nih.gov/books/NBK606033/ (accessed Aug. 01, 2025).

[3] IBM, “Text Mining,” *IBM*. https://www.ibm.com/think/topics/text-mining (accessed Aug. 02, 2025).

[4] H.-J. Kong, “Managing Unstructured Big Data in Healthcare System,” *Healthcare Informatics Research*, vol. 25, no. 1, 2019, doi: https://doi.org/10.4258/hir.2019.25.1.1.

[5] S. Alder, “Healthcare Data Breach Statistics,” *The HIPAA Journal*, May 26, 2025. https://www.hipaajournal.com/healthcare-data-breach-statistics/ (accessed Aug. 04, 2025).

[6] M. Kilgus, “GDPR Enforcement in Life Science & Healthcare,” *CMS*, May 13, 2025. https://cms.law/en/deu/publication/gdpr-enforcement-tracker-report/life-science-healthcare (accessed Aug. 05, 2025).

[7] G. Iversen, “Epic says its imaging order alerts prevented millions of duplicate scans over one year,” *DotMed*, 2025. https://www.dotmed.com/news/story/64303 (accessed Aug. 06, 2025).

[8] G. Church, “The deadly cost of duplicate patient records,” *OncLive*, Nov. 05, 2023. https://www.chiefhealthcareexecutive.com/view/the-deadly-cost-of-duplicate-patient-records-viewpoint (accessed Aug. 07, 2025).

[9] “Resourcelist - FHIR v5.0.0,” *HL7*. https://www.hl7.org/fhir/resourcelist.html (accessed Aug. 08, 2025).

[10] S. Rodriguez, “Latest FHIR Standard R5 Elevates Data Exchange, Interoperability,” *Health IT and EHR*, 2023. https://www.techtarget.com/searchhealthit/news/366577881/Latest-FHIR-Standard-R5-Elevates-Data-Exchange-Interoperability (accessed Aug. 09, 2025).

[11] WHO, “SMART Verifiable IPS for Pilgrimage v2.0.3,” *WHO*. https://smart.who.int/ips-pilgrimage/ (accessed Aug. 10, 2025).

[12] “Validating Resources against the specification and Profiles,” *Hl7*. https://confluence.hl7.org/plugins/viewsource/viewpagesrc.action?pageId=144967104 (accessed Aug. 11, 2025). [13] A. E. W. Johnson, T. J. Pollard, L. Shen, et al., "MIMIC-III, a freely accessible critical care database," Scientific Data, vol. 3, 160035, 2016. doi: 10.1038/sdata.2016.35.