FORM 2

THE PATENTS ACT, 1970 (39 OF 1970) & THE PATENTS RULES, 2003

COMPLETE SPECIFICATION

(SEE SECTION 10 AND RULE 13)

A SYSTEM AND A METHOD FOR PRIVACY-PRESERVING SYNTHETIC HEALTHCARE DATA GENERATION

APPLICANT NAME**: GAURAV RAJ CHATTARKI**

NATIONALITY**: INDIAN**

ADDRESS**: NO 101, 6TH CROSS, 10TH MAIN,**
BIKASIPURA, BENGALURU,
KARNATAKA- 560111, INDIA

The following specification particularly describes the disclosure and the manner in which is to be performed.

EARLIEST PRIORITY DATE

This Application claims priority from a provisional patent application filed in India having Patent Application No. **202641026065**, filed on **March 05, 2026**, and titled "A SYSTEM AND A METHOD FOR PRIVACY-PRESERVING SYNTHETIC HEALTHCARE DATA GENERATION".

FIELD OF INVENTION

[0001] The present invention relates to the field of data generation. More particularly, the present invention relates to a system and a method for privacy-preserving synthetic healthcare data generation.

BACKGROUND

[0002] Current synthetic data solutions suffer from three structural flaws that prevent adoption in regulated healthcare environments. One significant limitation is the lack of regulatory provenance.

[0003] Existing tools generate data that appears realistic but provide no auditable evidence of privacy preservation. There is no mechanism to prove to regulators such as Food and Drug Administration (FDA), Institutional Review Board (IRB) that a specific synthetic dataset satisfies strict budgets under epsilon differential privacy while maintaining clinical utility. These systems lack an immutable audit log linking privacy consumption to utility outcomes.

[0004] Another limitation is the destruction of clinical correlations. Standard differential privacy mechanisms add noise uniformly. In small datasets such as rare diseases with fewer than 200 patients, this noise may destroy critical clinical relationships, for example the correlation between HbA1c levels and diabetes. This renders the data useless for clinical research or trial design.

[0005] A further limitation is the multi-institutional trust failure. Collaborative research currently relies on slow legal contracts taking 12-18 months to finalize because institutions cannot trust a central party with raw data. There is no existing framework that allows multiple institutions to pool privacy

budgets and validate data utility without exposing raw patient records to a central server or to one another.

[0006] A still further limitation is the absence of cross-border data collaboration frameworks. Healthcare data localization requirements, including the Digital Personal Data Protection Act 2023 in India and data residency provisions under the General Data Protection Regulation in the European Union, prevent raw patient data from crossing jurisdictional boundaries. There is no existing mechanism that enables multi-jurisdictional healthcare research while satisfying data sovereignty requirements of each participating jurisdiction.

[0007] Hence, there is a need for an improved system and method for privacy-preserving synthetic healthcare data generation to address the aforementioned issue(s).

OBJECTIVES OF THE INVENTION

[0008] A primary objective of the invention is to provide a privacy-preserving synthetic healthcare data generation system capable of transforming healthcare data into a synthetic healthcare dataset while maintaining privacy protection of patient information and preserving usability of the synthetic healthcare dataset for analytical applications.

[0009] Another objective of the invention is to preserve clinically meaningful relationships associated with healthcare data during synthetic data generation by preventing distortion of correlations among clinical attributes while satisfying differential privacy requirements.

[0010] Another objective of the invention is to provide an auditable governance framework that tracks privacy budget consumption, maintains immutable processing records, and enables traceability of privacy-preserving operations associated with generation and downstream use of a synthetic healthcare dataset.

[0011] Another objective of the invention is to enable secure multi-institution collaboration by enforcing privacy budget controls, supporting

distributed trust mechanisms, and facilitating compliant sharing of synthetic healthcare datasets across multiple participating institutions without exposing underlying healthcare data.

[0012] Yet another objective of the invention is to enable regulatory compliance and cross-jurisdictional interoperability by generating compliance validation records, enforcing jurisdiction-specific processing restrictions, and producing synthetic healthcare datasets in interoperable formats suitable for healthcare research, clinical analysis, and regulatory submissions.

SUMMARY

[0013] In accordance with an embodiment of the present disclosure, a system for privacy-preserving synthetic healthcare data generation is disclosed. The system includes a processor. The system includes a memory coupled to the processor, wherein the memory comprises instructions that, when executed by the processor, cause the processor to: receive a healthcare data corresponding to one or more patients from one or more healthcare data sources and a user operating a user device via user interface; retrieve the healthcare data provided according to a healthcare interoperability standard; parse a plurality of healthcare resource graphs associated with the healthcare data to identify one or more privacy-sensitive attributes associated with the healthcare data; transform the healthcare data into a synthetic healthcare dataset by applying calibrated noise to the healthcare data based on one or more privacy parameters; preserve one or more clinical relationships associated with the healthcare data while satisfying differential privacy requirements; learn one or more feature distributions corresponding to the healthcare data by applying differentially private noise to one or more marginal distributions associated with the healthcare data; generate one or more probabilistic constraints corresponding to one or more clinical attribute pairs associated with the healthcare data; discard one or more synthetic samples that violate the one or more probabilistic constraints; treat a privacy budget as a finite computational resource and deduct one or more privacy units from the privacy budget for each statistical query executed against the healthcare data; generate an append-only audit log comprising one or more hashed transaction records linking consumption of the privacy budget with one or more utility scores

corresponding to the synthetic healthcare dataset; and generate the synthetic healthcare dataset formatted according to the healthcare interoperability standard and generate a compliance certificate corresponding to the synthetic healthcare dataset.

[0014] In accordance with an embodiment of the present disclosure, a method for privacy-preserving synthetic healthcare data generation is disclosed. The method includes receiving a healthcare data corresponding to one or more patients from one or more healthcare data sources and a user operating a user device via a user interface. The method includes retrieving the healthcare data provided according to a healthcare interoperability standard. The method includes parsing a plurality of healthcare resource graphs associated with the healthcare data to identify one or more privacy-sensitive attributes associated with the healthcare data. The method includes transforming the healthcare data into a synthetic healthcare dataset by applying calibrated noise to the healthcare data based on one or more privacy parameters. The method includes preserving one or more clinical relationships associated with the healthcare data while satisfying differential privacy requirements. The method includes learning one or more feature distributions corresponding to the healthcare data by applying differentially private noise to one or more marginal distributions associated with the healthcare data. The method includes generating one or more probabilistic constraints corresponding to one or more clinical attribute pairs associated with the healthcare data. The method includes discarding one or more synthetic samples that violate the one or more probabilistic constraints. The method includes treating a privacy budget as a finite computational resource and deducting one or more privacy units from the privacy budget for each statistical query executed against the healthcare data. The method includes generating an append-only audit log comprising one or more hashed transaction records linking consumption of the privacy budget with one or more utility scores corresponding to the synthetic healthcare dataset. The method includes generating the synthetic healthcare dataset formatted according to the healthcare interoperability standard and generating a compliance certificate corresponding to the synthetic healthcare dataset.

To further clarify the advantages and features of the present disclosure, a more particular description of the disclosure will follow by reference to specific embodiments thereof, which are illustrated in the appended figures. It is to be appreciated that these figures depict only typical embodiments of the disclosure and are therefore not to be considered limiting in scope. The disclosure will be described and explained with additional specificity and detail with the appended figures.

BRIEF DESCRIPTION OF THE DRAWINGS

[0015] The disclosure will be described and explained with additional specificity and detail with the accompanying figures in which:

[0016] FIG. 1 illustrates a network environment of a system for privacy-preserving synthetic healthcare data generation in accordance with an embodiment of the present disclosure;

[0017] FIG. 2 illustrates a schematic diagram of a user device of FIG. 1, in accordance with an example implementation of the present subject matter;

[0018] FIG. 3 illustrates a schematic diagram of a system for privacy-preserving synthetic healthcare data generation of FIG. 1, in accordance with an embodiment of the present disclosure;

[0019] FIG. 4 (a) is a flow chart representing the steps involved in a method for privacy-preserving synthetic healthcare data generation, in accordance with an embodiment of the present disclosure; and

[0020] FIG. 4 (b) illustrates continued steps of the method of FIG. 4 (a) in accordance with an embodiment of the present disclosure.

[0021] Further, those skilled in the art will appreciate that elements in the figures are illustrated for simplicity and may not have necessarily been drawn to scale. Furthermore, in terms of the construction of the device, one or more components of the device may have been represented in the figures by conventional symbols, and the figures may show only those specific details that are pertinent to understanding the embodiments of the present disclosure so as not

to obscure the figures with details that will be readily apparent to those skilled in the art having the benefit of the description herein.

DETAILED DESCRIPTION

[0022] For the purpose of promoting an understanding of the principles of the disclosure, reference will now be made to the embodiment illustrated in the figures and specific language will be used to describe them. It will nevertheless be understood that no limitation of the scope of the disclosure is thereby intended. Such alterations and further modifications in the illustrated system, and such further applications of the principles of the disclosure as would normally occur to those skilled in the art are to be construed as being within the scope of the present disclosure.

[0023] The terms “comprises”, “comprising”, or any other variations thereof, are intended to cover a non-exclusive inclusion, such that a process or method that comprises a list of steps does not include only those steps but may include other steps not expressly listed or inherent to such a process or method. Similarly, one or more devices or subsystems or elements or structures or components preceded by "comprises... a" does not, without more constraints, preclude the existence of other devices, sub-systems, elements, structures, components, additional devices, additional sub-systems, additional elements, additional structures or additional components. Appearances of the phrase "in an embodiment", "in another embodiment" and similar language throughout this specification may, but not necessarily do, all refer to the same embodiment.

[0024] Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by those skilled in the art to which this disclosure belongs. The system, methods, and examples provided herein are only illustrative and not intended to be limiting.

[0025] In the following specification and the claims, reference will be made to a number of terms, which shall be defined to have the following meanings. The singular forms “a”, “an”, and “the” include plural references unless the context clearly dictates otherwise.

[0026] FIG. 1 illustrates a network environment of a system for privacy-preserving synthetic healthcare data generation in accordance with an embodiment of the present disclosure.

[0027] Referring to FIG. 1, a user device (104) corresponding to a user (108) may be communicatively coupled to a system (102). The user (108) may access the system (102) over a network (106). Examples of the user device (104) includes, but is not limited to, a mobile phone, desktop computer, portable digital assistant (PDA), smart phone, tablet, ultra-book, netbook, laptop, multi-processor system, microprocessor-based or programmable consumer electronic system, or any other communication device that a user may use. It will be appreciated that the system (102) may be presented to the user on the user device (104) as a web application accessed through a browser, through a software application on the user device, or, particularly for smartphones, through a mobile application installed at the smartphone. It will be appreciated that, within the context of the disclosure herein, web application refers to a utility implemented on a networked computing system accessible by user device over the Internet (e.g. through browsers) wherein the bulk of the processing takes place at the networked computing system, mobile applications refer to applications installed on smartphones that may communicate with a networked computing system, and a “software” application refers generally to applications other than web browsers installed on other types of user device that may communicate with a networked computing system over the network (106).

[0028] The network (106) may be a single communication network or a combination of multiple communication networks and may use a variety of different communication protocols. The network (106) may be a wireless network, a wired network, or a combination thereof. Examples of such individual personalized networks include, but are not limited to, Global System for Mobile Communication (GSM) network, Universal Mobile Telecommunications System (UMTS) network, Personal Communications Service (PCS) network, Time Division Multiple Access TDMA) network, Code Division Multiple Access (CDMA) network, Next Generation Network (NON), Public Switched Telephone Network (PSTN). Depending on the technology, the personalized network (106)

may include various network entities, such as gateways and routers; however, such details have been omitted for the sake of brevity of the present description.

[0029] The system (102) may have a homepage that is presented to the user (108) accessing a top-level web address for web applications presented to the user (108) in a browser or a welcome screen for software and mobile applications. The homepage may include links to a user log-in interface or general information about the system (102) and the option to register as user (108). It will be appreciated that the presentation of a homepage may not be necessary, for example, if a user bypasses it by directly inputting a web address corresponding to a user log-in page, or if a separate mobile application is designed for users.

[0030] A new or unregistered user can access the user log-in interface, fill out the log-in information corresponding to the user's account, and indicate that the user wishes to sign in. It will be appreciated that any conventional registration and log-in techniques for web applications, software application, and mobile applications may be used, whichever is appropriate for the user. While registering the user may be prompted to provide username and corresponding user credentials, not limited to, password, geographical location, and contact information and upon receipt of the foregoing information, a corresponding user-profile may be created and stored on a respective database of the system (102).

[0031] In an embodiment, a healthcare data sources (110) are configured to store, manage, and provide access to healthcare data corresponding to one or more patients for processing by the system 102. The healthcare data sources (110) may include internal healthcare repositories, external healthcare systems, or distributed clinical databases that maintain structured, semi-structured, or unstructured healthcare records. The healthcare data sources (110) provide the healthcare data to the ingestion module (308) through secure communication channels for downstream privacy-preserving processing. The healthcare data stored within the healthcare data sources (110) may include patient demographic records, diagnostic records, laboratory observations, medication records, encounter records, procedure records, imaging records, and treatment histories. In one embodiment, the healthcare data sources (110) may include, but are not limited to, an electronic health record system, a hospital database, a pharmaceutical database,

a research institution database, a clinical trial database, a laboratory information system, and a healthcare data exchange platform. It must be noted that the healthcare data sources (110) enable secure and scalable acquisition of healthcare data while preserving traceability of source-origin information associated with the healthcare data processed by the system (102).

[0032] In accordance with an embodiment of the present disclosure, a system for privacy-preserving synthetic healthcare data generation is disclosed. The system includes a processor. The system includes a memory coupled to the processor, wherein the memory comprises instructions that, when executed by the processor, cause the processor to: receive a healthcare data corresponding to one or more patients from one or more healthcare data sources and a user operating a user device via user interface; retrieve the healthcare data provided according to a healthcare interoperability standard; parse a plurality of healthcare resource graphs associated with the healthcare data to identify one or more privacy-sensitive attributes associated with the healthcare data; transform the healthcare data into a synthetic healthcare dataset by applying calibrated noise to the healthcare data based on one or more privacy parameters; preserve one or more clinical relationships associated with the healthcare data while satisfying differential privacy requirements; learn one or more feature distributions corresponding to the healthcare data by applying differentially private noise to one or more marginal distributions associated with the healthcare data; generate one or more probabilistic constraints corresponding to one or more clinical attribute pairs associated with the healthcare data; discard one or more synthetic samples that violate the one or more probabilistic constraints; treat a privacy budget as a finite computational resource and deduct one or more privacy units from the privacy budget for each statistical query executed against the healthcare data; generate an append-only audit log comprising one or more hashed transaction records linking consumption of the privacy budget with one or more utility scores corresponding to the synthetic healthcare dataset;; and generate the synthetic healthcare dataset formatted according to the healthcare interoperability standard and generate a compliance certificate corresponding to the synthetic healthcare dataset.

[0033] It may be noted that the foregoing system is an exemplary system and may be implemented as computer executable instructions in any computing or processing environment, including in digital electronic circuitry or in computer hardware, firmware, device driver, or software. As such, the system is not limited to any specific hardware or software configuration.

[0034] FIG. 2 illustrates a schematic diagram of a user device, in accordance with an example implementation of the present subject matter. Referring to FIG. 2, the user device (104) may comprise a processor(s) (202), a memory(s) (204) coupled to and accessible by the processor(s) (202), and an interface (210) coupled to the memory(s) (204). The user device (104) disclosed herein may be same as the user device (104) described in FIG. 1. The functions of various elements shown in the figs., including any functional blocks labelled as "processor(s)", may be provided through the use of dedicated hardware as well as hardware capable of executing instructions. When provided by a processor, the functions may be provided by a single dedicated processor, by a single shared processor, or by a plurality of individual processors, some of which may be shared. Moreover, explicit use of the term "processor" would not be construed to refer exclusively to hardware capable of executing instructions, and may implicitly comprise, without limitation, digital signal processor (DSP) hardware, network processor, application specific integrated circuit (ASIC), field programmable gate array (FPGA). Other hardware, standard and/or custom, may also be coupled to the processor(s) (202). The user device (104) may further include a display (206) in addition to other components such as, but not limited to, keyboard, sensors, logic circuits etc. Further, the user device (104) may include data (208) which may include data (208) that may be stored, utilized or generated during the operation of the user device (104).

[0035] The memory(s) (204) may be a computer-readable medium, examples of which comprise volatile memory (e.g., RAM), and/or non-volatile memory (e.g., Erasable Programmable read-only memory, i.e. EPROM, flash memory, etc.). The memory(s) (204) may be an external memory, or internal memory, such as a flash drive, a compact disk drive, an external hard disk drive, or the like. The user device (104) may further include an interface (210) that may

allow the connection or coupling of the user device (104) with one or more other devices, through a wired (e.g., Local Area Network, i.e., LAN) connection or through a wireless connection (e.g., Bluetooth®, Wi-Fi), for example, for connecting to the system shown in FIG. 1. The interface may also enable intercommunication between different logical as well as hardware components of the user device (104).

[0036] FIG. 3 illustrates a schematic diagram of a system for privacy-preserving synthetic healthcare data generation of FIG. 1, in accordance with an embodiment of the present disclosure. Referring to FIG. 3, the system (102) includes a processor(s) (302), a memory(s) (304) coupled to and accessible by the processor(s) (302), and database (346) coupled to the memory(s) (304).

[0037] The system (102) disclosed herein is the same as the system (102) described in FIG. 1. The functions of various elements shown in the figs., including any functional blocks labelled as "processor(s)", may be provided through the use of dedicated hardware as well as hardware capable of executing instructions. When provided by a processor, the functions may be provided by a single dedicated processor, by a single shared processor, or by a plurality of individual processors, some of which may be shared. Moreover, explicit use of the term "processor" would not be construed to refer exclusively to hardware capable of executing instructions, and may implicitly comprise, without limitation, digital signal processor (DSP) hardware, network processor, application specific integrated circuit (ASIC), field programmable gate array (FPGA). Other hardware, standard and/or custom, may also be coupled to the processor(s) (302). The system (102) may further include other components such as, but not limited to, keyboard, sensors, logic circuits, input/output interfaces etc. Further, the system (102) may include data which may include data that may be stored, utilized or generated during the operation of the computer implemented system (102).

[0038] The memory(s) (304) may be a computer-readable medium, examples of which comprise volatile memory (e.g., RAM), and/or non-volatile memory (e.g., Erasable Programmable read-only memory, i.e. EPROM, flash memory, etc.). The memory(s) (304) may be an external memory, or internal memory, such as a flash drive, a compact disk drive, an external hard disk drive,

or the like. The system (102) may further include the user interface (348) that may allow the connection or coupling of the system (102) with one or more other devices, through a wired (e.g., Local Area Network, i.e., LAN) connection or through a wireless connection (e.g., Bluetooth®, Wi-Fi)., for example, for connecting to the user device (104) as shown in FIG. 1. The user interface (348) may also enable intercommunication between different logical as well as hardware components of the system (102).

[0039] The system (102) may be provided with a database (346) to store healthcare data, append-only audit log records, synthetic data provenance records, cross-jurisdictional transfer records, cumulative privacy loss ledger records, consent registry records. In an example implementation of the system (102) including one or more servers, the databases may databases local to the server or may be remote to the server. It may be noted that the data in the databases may be stored as a table or may be pre-stored as a mapping with the other. This application is not limited thereto.

[0040] The system (102) may include module(s). The module(s) may include an ingestion module (308), a differential privacy module (310), a marginal distribution learning module (312), a correlation constraint module (314), a rejection sampling module (316), a clinical audit trail module (318), an audit streaming module (320), an alert module (322), a regulatory detection module (324), an adaptive budget optimization module (326), a jurisdiction processing module (328), a privacy budget composition tracking module (330), a consent lifecycle management module (332), and an output generation module (336). In one example, the module(s) may be implemented as a combination of hardware and firmware. In an example described herein, such combinations of hardware and firmware may be implemented in several different ways. For example, the firmware for module(s) may be processor (302) executable instructions stored on a non-transitory machine-readable storage medium and the hardware for the module(s) may include a processing resource (for example, implemented as either single processor or combination of multiple processors), to execute such instructions. Further, the hardware for the module(s) may include communication apparatuses, control circuitries involving electrical and electronics components,

sensors, and interface devices, which may be in communication with each other for multi-directional communication therebetween.

[0041] Further, the system (102) includes data. The data may include data that is either stored or generated as a result of functions implemented by the system. It may be further noted that information stored and available in data may be utilized by the engine(s) for performing various functions by the system. In an example, data may include a healthcare data (350), a synthetic healthcare dataset (352), a privacy-sensitive data (354), a privacy budget data (356), utility score data (358), and a provenance record data (360). It may be noted that such examples of the various functions are only indicative. The present approaches may be applicable to other examples without deviating from the scope of the present subject matter.

[0042] In the present examples, the non-transitory machine-readable storage medium may store instructions that, when executed by the processing resource, implement the functionalities of modules(s). In such examples, the system (102) may include the machine-readable storage medium storing the instructions and the processing resource to execute the instructions. In other examples of the present subject matter, the machine-readable storage medium may be located at a different location but accessible to the system (102) and the processor(s) (302).

[0043] In operation, the ingestion module (308) is configured to receive healthcare data corresponding to one or more patients from one or more healthcare data sources (110, FIG. 1) and a user (108, FIG. 1) operating a user device (104, FIG. 1) via user interface (348). The ingestion module (308) establishes one or more communication interfaces with the one or more healthcare data sources and performs controlled data acquisition based on predefined access policies and data retrieval protocols. The healthcare data comprises structured and semi-structured clinical information associated with the one or more patients, including demographic attributes, diagnostic records, laboratory observations, medication records, and encounter details. The ingestion module (308) ensures that the received healthcare data is aligned with a predefined schema corresponding to a healthcare interoperability standard and maintains integrity of the healthcare data during transmission and reception. The ingestion module (308) further associates

metadata with the healthcare data, wherein the metadata includes source identifiers, timestamps, and data lineage indicators to support downstream processing by subsequent modules. The ingestion module (308) performs validation of the healthcare data to ensure completeness, consistency, and compliance with format requirements prior to forwarding the healthcare data for further processing. An example of the ingestion module (308) includes, but is not limited to, an interface layer configured to receive healthcare data from an electronic health record system, a hospital database, a clinical laboratory information system, or a remote healthcare data repository through secure communication protocols.

[0044] An example of the healthcare data is as follows:

Patient Resource:

Patient identifier, demographic information, insurance details

Observation Resource:

Blood glucose values, blood pressure readings, laboratory test results

Condition Resource:

Diabetes diagnosis, hypertension diagnosis

Medication Resource:

Insulin prescription, antihypertensive medication records

Encounter Resource:

Hospital admission records, outpatient consultation records

Procedure Resource:

Surgical procedure details, diagnostic procedure records

[0045] In an embodiment, the ingestion module (308) is further configured to retrieve the healthcare data provided according to a healthcare interoperability

standard. The ingestion module (308) identifies the healthcare interoperability standard associated with the healthcare data and retrieves the healthcare data in a format compliant with the healthcare interoperability standard to ensure structural consistency and semantic interoperability. The healthcare interoperability standard defines data schemas, resource definitions, and exchange protocols that govern representation of the healthcare data across the one or more healthcare data sources. The ingestion module (308) processes the healthcare data in accordance with the healthcare interoperability standard to preserve hierarchical relationships, resource linkages, and attribute mappings inherent in the healthcare data. The ingestion module (308) further validates compliance of the healthcare data with the healthcare interoperability standard by verifying conformance of resource structures, field types, and data encoding formats prior to downstream processing. An example of the healthcare interoperability standard includes, but is not limited to, HL7 Fast Healthcare Interoperability Resources defining patient resources, observation resources, condition resources, and encounter resources for structured representation of the healthcare data.

[0046] In an embodiment, the ingestion module (308) is further configured to parse a plurality of healthcare resource graphs associated with the healthcare data to identify one or more privacy-sensitive attributes associated with the healthcare data. The ingestion module (308) decomposes the healthcare data into interconnected healthcare resource graphs representing relationships among clinical entities, wherein each healthcare resource graph comprises nodes corresponding to healthcare resources and edges representing dependencies and associations among the healthcare resources. The ingestion module (308) traverses the plurality of healthcare resource graphs to extract attribute-level information and applies a sensitivity identification process to classify the one or more privacy-sensitive attributes based on predefined sensitivity criteria. The one or more privacy-sensitive attributes include identifiers, quasi-identifiers, and clinically sensitive attributes that may contribute to re-identification risk or disclosure of confidential clinical information. The ingestion module (308) further tags the one or more privacy-sensitive attributes with sensitivity labels to enable differentiated processing by subsequent modules within the system (102). An example of the plurality of healthcare resource graphs includes, but is not limited

to, patient resources linked with observation resources, condition resources, medication resources, and encounter resources, wherein the ingestion module (308) identifies attributes such as patient identifiers, demographic attributes, diagnostic codes, and laboratory values as the one or more privacy-sensitive attributes based on the sensitivity criteria.

[0047] In an embodiment, the healthcare interoperability standard comprises HL7 Fast Healthcare Interoperability Resources, wherein the ingestion module (308) is configured to retrieve and process the healthcare data structured according to resource-based specifications defined by the healthcare interoperability standard. The healthcare interoperability standard defines a plurality of standardized resource types, data elements, and exchange protocols that enable consistent representation of the healthcare data across the one or more healthcare data sources. The ingestion module (308) interprets the healthcare data in accordance with the healthcare interoperability standard by mapping the healthcare data to corresponding resource definitions and maintaining linkage among related resources within the plurality of healthcare resource graphs. The ingestion module (308) further ensures that the healthcare data adheres to syntactic and semantic constraints defined by the healthcare interoperability standard, thereby enabling interoperability and consistency across subsequent modules within the system (102). An example of the healthcare interoperability standard includes, but is not limited to, patient resources representing demographic information, observation resources representing clinical measurements, condition resources representing diagnoses, and encounter resources representing clinical interactions defined under HL7 Fast Healthcare Interoperability Resources.

[0048] In an embodiment, the plurality of healthcare resource graphs comprises at least one of patient resources, observation resources, condition resources, medication resources, encounter resources, or procedure resources, wherein the ingestion module (308) is configured to organize the healthcare data into structured graph representations corresponding to the plurality of healthcare resource graphs. Each of the patient resources represents demographic and identification attributes of the one or more patients, each of the observation

resources represents clinical measurements and laboratory results, each of the condition resources represents diagnostic information, each of the medication resources represents prescribed or administered medications, each of the encounter resources represents clinical interactions, and each of the procedure resources represents medical procedures associated with the one or more patients. The ingestion module (308) establishes relationships among the plurality of healthcare resource graphs by linking associated resources through reference identifiers and dependency mappings, thereby forming an interconnected graph structure of the healthcare data. The ingestion module (308) further utilizes the plurality of healthcare resource graphs to enable contextual interpretation of the healthcare data and to support identification of the one or more privacy-sensitive attributes across the interconnected resources. An example of the plurality of healthcare resource graphs includes, but is not limited to, a patient resource linked with an observation resource capturing laboratory values, a condition resource indicating a diagnosis, a medication resource corresponding to prescribed treatment, an encounter resource associated with a clinical visit, and a procedure resource corresponding to a performed medical procedure.

[0049] In an embodiment, the one or more healthcare data sources comprise at least one of an electronic health record system, a hospital database, a pharmaceutical database, a research institution database, or a clinical trial database, wherein the ingestion module (308) is configured to establish communication with the one or more healthcare data sources to retrieve the healthcare data corresponding to the one or more patients. The electronic health record system stores longitudinal clinical records associated with the one or more patients, the hospital database stores operational and clinical data generated within a healthcare facility, the pharmaceutical database stores drug-related information and medication records, the research institution database stores datasets generated from clinical research activities, and the clinical trial database stores structured data collected during clinical trial studies. The ingestion module (308) interfaces with the one or more healthcare data sources using secure data exchange mechanisms and retrieves the healthcare data while preserving structure and integrity of the healthcare data. The ingestion module (308) further harmonizes the healthcare data obtained from the one or more healthcare data sources to

ensure compatibility with the healthcare interoperability standard prior to downstream processing. An example of the one or more healthcare data sources includes, but is not limited to, an electronic health record system storing patient demographic data and clinical history, a hospital database storing admission and discharge records, a pharmaceutical database storing medication dispensing information, a research institution database storing observational study datasets, and a clinical trial database storing trial outcome data.

[0050] In an embodiment, the differential privacy module (310) operatively coupled to the ingestion module (308) is configured to transform the healthcare data into the synthetic healthcare dataset by applying calibrated noise to the healthcare data based on the one or more privacy parameters. The differential privacy module (310) receives the healthcare data from the ingestion module (308) and applies a noise injection mechanism that perturbs the healthcare data in a controlled manner while preserving statistical characteristics of the healthcare data. The one or more privacy parameters define a magnitude and distribution of the calibrated noise to be applied, wherein the calibrated noise is determined in accordance with a differential privacy framework to limit disclosure risk associated with the healthcare data. The differential privacy module (310) ensures that transformation of the healthcare data into the synthetic healthcare dataset satisfies formal privacy guarantees while maintaining utility of the synthetic healthcare dataset for downstream analytical tasks. The differential privacy module (310) further performs attribute-level or record-level perturbation of the healthcare data based on the one or more privacy parameters to balance privacy preservation and data fidelity. An example of the differential privacy module (310) includes, but is not limited to, a noise generation component configured to apply Laplace noise or Gaussian noise to numerical attributes of the healthcare data and to apply randomized response mechanisms to categorical attributes of the healthcare data based on the one or more privacy parameters.

[0051] In an embodiment, the differential privacy module (310) is further configured to preserve one or more clinical relationships associated with the healthcare data while satisfying differential privacy requirements. The differential privacy module (310) applies the calibrated noise to the healthcare data in a

manner that maintains dependency structures and correlation patterns among clinical attributes present in the healthcare data. The differential privacy module (310) enforces preservation of the one or more clinical relationships by incorporating constraint-aware transformations that ensure consistency between related clinical attributes during generation of the synthetic healthcare dataset. The differential privacy module (310) balances application of the calibrated noise with preservation of the one or more clinical relationships by adapting the one or more privacy parameters in accordance with sensitivity of the healthcare data and importance of the one or more clinical relationships. The differential privacy module (310) further ensures that preservation of the one or more clinical relationships does not violate the differential privacy requirements by limiting information leakage through controlled noise application. An example of the one or more clinical relationships includes, but is not limited to, an association between a diagnosis attribute representing diabetes and a corresponding observation attribute representing elevated glucose levels, an association between a condition attribute representing hypertension and an observation attribute representing high blood pressure, and an association between a medication attribute and a corresponding treatment indication defined within the healthcare data.

[0052] In an embodiment, the differential privacy module (310) is configured to generate the synthetic healthcare dataset for a rare disease cohort comprising fewer than 200 patients. The differential privacy module (310) identifies the rare disease cohort based on a cohort size threshold and applies adaptive privacy parameter selection tailored to the limited size of the healthcare data corresponding to the rare disease cohort. The differential privacy module (310) adjusts the calibrated noise distribution and magnitude to preserve statistical relevance of the healthcare data while maintaining differential privacy guarantees despite the reduced sample size. The differential privacy module (310) further prioritizes preservation of the one or more clinical relationships associated with the healthcare data by allocating a proportion of the privacy budget toward maintaining correlations among clinical attributes that are critical for rare disease analysis. The differential privacy module (310) ensures that the synthetic healthcare dataset generated for the rare disease cohort retains representative

patterns of the healthcare data without exposing identifiable information corresponding to the one or more patients. An example of the rare disease cohort comprising fewer than 200 patients includes, but is not limited to, a cohort defined by a condition attribute representing a specific genetic disorder, a cohort defined by a diagnosis attribute representing a rare oncology subtype, or a cohort defined by a clinical phenotype captured within the healthcare data corresponding to the one or more patients.

[0053] In an embodiment, the marginal distribution learning module (312) operatively coupled to the differential privacy module (310) is configured to learn one or more feature distributions corresponding to the healthcare data by applying differentially private noise to one or more marginal distributions associated with the healthcare data. The marginal distribution learning module (312) receives the healthcare data from the differential privacy module (310) and computes the one or more marginal distributions for individual attributes present in the healthcare data. The one or more marginal distributions represent statistical distributions of independent features including numerical attributes, categorical attributes, and temporal attributes associated with the healthcare data. The marginal distribution learning module (312) applies differentially private noise to the one or more marginal distributions in accordance with the one or more privacy parameters to prevent disclosure of information corresponding to the one or more patients while preserving statistical characteristics of individual features. The marginal distribution learning module (312) further generates learned feature distributions that are utilized during subsequent generation of the synthetic healthcare dataset. An example of the one or more feature distributions includes, but is not limited to, an age distribution corresponding to patient demographics, a glucose level distribution corresponding to laboratory observations, a diagnosis frequency distribution corresponding to condition attributes, and a medication frequency distribution corresponding to prescription records within the healthcare data.

[0054] In an embodiment, the correlation constraint module (314) operatively coupled to the differential privacy module (310) is configured to generate one or more probabilistic constraints corresponding to one or more clinical attribute pairs associated with the healthcare data. The correlation

constraint module (314) receives the healthcare data and the learned feature distributions from the marginal distribution learning module (312) and identifies statistically significant dependencies among clinical attributes present in the healthcare data. The correlation constraint module (314) analyses co-occurrence patterns, conditional probabilities, and dependency relationships among the one or more clinical attribute pairs to generate the one or more probabilistic constraints that govern preservation of clinically meaningful associations during generation of the synthetic healthcare dataset. The one or more probabilistic constraints define acceptable probability ranges, dependency thresholds, and relationship rules associated with the one or more clinical attribute pairs. The correlation constraint module (314) further stores the one or more probabilistic constraints for use during synthetic sample generation to ensure preservation of clinical consistency. An example of the one or more clinical attribute pairs includes, but is not limited to, a diagnosis attribute representing diabetes and an observation attribute representing elevated glucose levels, a condition attribute representing hypertension and an observation attribute representing elevated blood pressure values, and a medication attribute corresponding to insulin therapy associated with a diagnosis attribute representing diabetes.

[0055] In an embodiment, the rejection sampling module (316) operatively coupled to the differential privacy module (310) is configured to discard one or more synthetic samples that violate the one or more probabilistic constraints. The rejection sampling module (316) receives candidate synthetic samples generated based on the learned feature distributions and evaluates each synthetic sample against the one or more probabilistic constraints generated by the correlation constraint module (314). The rejection sampling module (316) determines whether each synthetic sample satisfies dependency thresholds, conditional probability requirements, and relationship rules associated with the one or more clinical attribute pairs. Upon determining that a synthetic sample violates the one or more probabilistic constraints, the rejection sampling module (316) discards the synthetic sample and initiates generation of an alternate synthetic sample that conforms to the one or more probabilistic constraints. The rejection sampling module (316) further retains synthetic samples that satisfy the one or more probabilistic constraints for inclusion in the synthetic healthcare dataset. An

example of the one or more synthetic samples that violate the one or more probabilistic constraints includes, but is not limited to, a synthetic sample indicating a diagnosis attribute representing diabetes without a corresponding observation attribute representing elevated glucose levels, a synthetic sample indicating hypertension without an associated elevated blood pressure observation, or a synthetic sample indicating prescription of a specialized medication without a corresponding diagnosis attribute.

[0056] In an embodiment, the clinical audit trail module (318) operatively coupled to the ingestion module (308) and the differential privacy module (310) is configured to maintain one or more records corresponding to each data transformation operation. The clinical audit trail module (318) continuously monitors processing activities performed on the healthcare data and records each data transformation operation executed during generation of the synthetic healthcare dataset. Each of the one or more records captures transformation metadata associated with the healthcare data, including an operation identifier, a transformation type, a processing timestamp, a module identifier, and a transformation outcome associated with the corresponding data transformation operation. The clinical audit trail module (318) further associates each of the one or more records with a corresponding stage of processing performed by the ingestion module (308), the differential privacy module (310), the marginal distribution learning module (312), the correlation constraint module (314), and the rejection sampling module (316). The clinical audit trail module (318) stores the one or more records in a traceable sequence to enable reconstruction of processing history associated with the synthetic healthcare dataset. An example of the one or more records corresponding to each data transformation operation includes, but is not limited to, a record indicating retrieval of the healthcare data by the ingestion module (308), a record indicating application of calibrated noise by the differential privacy module (310), a record indicating generation of the one or more probabilistic constraints by the correlation constraint module (314), and a record indicating rejection of invalid synthetic samples by the rejection sampling module (316).

[0057] In an embodiment, the clinical audit trail module (318) is further configured to treat a privacy budget as a finite computational resource and deduct one or more privacy units from the privacy budget for each statistical query executed against the healthcare data. The clinical audit trail module (318) initializes the privacy budget for the healthcare data prior to execution of privacy-consuming operations and maintains a cumulative accounting ledger associated with the privacy budget throughout generation of the synthetic healthcare dataset. Each statistical query executed against the healthcare data consumes a quantifiable portion of the privacy budget based on sensitivity of the statistical query, complexity of the statistical query, and the one or more privacy parameters applied during execution of the statistical query. The clinical audit trail module (318) deducts the one or more privacy units from the privacy budget after execution of each statistical query and updates remaining privacy capacity associated with the healthcare data. The clinical audit trail module (318) further monitors cumulative consumption of the privacy budget to ensure that privacy guarantees associated with the healthcare data are preserved throughout processing. An example of the statistical query executed against the healthcare data includes, but is not limited to, a query for calculating an age distribution associated with the healthcare data, a query for identifying diagnosis frequency patterns associated with condition attributes, a query for calculating laboratory observation distributions, and a query for evaluating dependency relationships among clinical attributes.

[0058] In an embodiment, the clinical audit trail module (318) is further configured to generate an append-only audit log comprising one or more hashed transaction records linking consumption of the privacy budget with one or more utility scores corresponding to the synthetic healthcare dataset. The clinical audit trail module (318) creates a transaction record for each privacy-consuming operation performed during generation of the synthetic healthcare dataset and applies a cryptographic hashing operation to each transaction record to generate the one or more hashed transaction records. Each of the one or more hashed transaction records stores information associated with privacy budget consumption, including an amount of consumed privacy units, an associated statistical query identifier, processing timestamps, and corresponding utility measurements generated after execution of the privacy-consuming operation. The

one or more utility scores represent performance indicators associated with the synthetic healthcare dataset and quantify usability of the synthetic healthcare dataset after application of privacy-preserving transformations. The clinical audit trail module (318) appends each of the one or more hashed transaction records sequentially to the append-only audit log to preserve chronological integrity of privacy consumption records. An example of the one or more utility scores corresponding to the synthetic healthcare dataset includes, but is not limited to, a statistical fidelity score measuring similarity between the healthcare data and the synthetic healthcare dataset, a correlation preservation score measuring retention of clinical relationships, and a clinical plausibility score measuring consistency of generated synthetic samples.

[0059] In an embodiment, each audit trail entry of the append-only audit log comprises an epsilon parameter, a delta parameter, and a noise calibration record corresponding to generation of the synthetic healthcare dataset. The clinical audit trail module (318) records the epsilon parameter representing a quantifiable privacy loss threshold associated with a privacy-preserving operation and records the delta parameter representing a probability threshold associated with permissible privacy leakage during execution of the privacy-preserving operation. The noise calibration record stores configuration details associated with application of calibrated noise during transformation of the healthcare data into the synthetic healthcare dataset. The noise calibration record includes parameters associated with noise distribution type, noise magnitude, sensitivity values, allocation of the privacy budget, and module identifiers corresponding to execution of privacy-preserving operations. The clinical audit trail module (318) associates the epsilon parameter, the delta parameter, and the noise calibration record with a corresponding audit trail entry to enable traceability of privacy-preserving configurations used during generation of the synthetic healthcare dataset. An example of the noise calibration record includes, but is not limited to, a record identifying application of Laplace noise to numerical attributes of the healthcare data, a record identifying application of Gaussian noise to observation values, and a record identifying allocation of the privacy budget across feature distribution learning and correlation preservation operations.

[0060] In an embodiment, the clinical audit trail module (318) is further configured to allocate a larger proportion of the privacy budget to preservation of the one or more clinical relationships when a patient cohort size is below a predefined threshold. The clinical audit trail module (318) continuously monitors the patient cohort size associated with the healthcare data and compares the patient cohort size with the predefined threshold to identify limited-size cohorts that require adaptive privacy allocation. Upon determining that the patient cohort size is below the predefined threshold, the clinical audit trail module (318) dynamically reallocates the privacy budget by assigning a greater proportion of privacy units toward operations responsible for preserving the one or more clinical relationships associated with the healthcare data. The clinical audit trail module (318) coordinates with the differential privacy module (310), the marginal distribution learning module (312), and the correlation constraint module (314) to reduce excessive noise distortion in clinically dependent attributes while maintaining required privacy guarantees. The clinical audit trail module (318) further records the adaptive allocation decision within the append-only audit log for subsequent traceability. An example of the predefined threshold includes, but is not limited to, a cohort threshold corresponding to fewer than 200 patients associated with a rare disease cohort, a cohort threshold corresponding to a specialized oncology population, or a cohort threshold corresponding to a limited paediatric patient group represented within the healthcare data.

[0061] In an embodiment, the clinical audit trail module (318) is further configured to reject one or more synthetic healthcare dataset generation requests when cumulative consumption of the privacy budget exceeds a global privacy threshold across a plurality of participating institutions. The clinical audit trail module (318) monitors privacy budget consumption associated with the plurality of participating institutions contributing the healthcare data and maintains a cumulative privacy consumption record corresponding to all privacy-preserving operations executed across the plurality of participating institutions. The global privacy threshold defines a maximum allowable level of cumulative privacy consumption for the healthcare data shared among the plurality of participating institutions. Upon receiving a synthetic healthcare dataset generation request, the clinical audit trail module (318) evaluates remaining privacy capacity by

comparing cumulative consumption of the privacy budget with the global privacy threshold. When cumulative consumption of the privacy budget exceeds the global privacy threshold, the clinical audit trail module (318) automatically rejects the synthetic healthcare dataset generation request and records the rejection event within the append-only audit log. An example of the plurality of participating institutions includes, but is not limited to, multiple hospitals contributing the healthcare data for collaborative clinical research, pharmaceutical organizations participating in multi-site drug development programs, and research institutions contributing datasets for federated healthcare analytics.

[0062] In an embodiment, the clinical audit trail module (318) is further configured to maintain a synthetic data provenance record corresponding to the synthetic healthcare dataset. The clinical audit trail module (318) generates and stores the synthetic data provenance record to establish lifecycle traceability associated with creation, processing, distribution, and downstream usage of the synthetic healthcare dataset. The synthetic data provenance record maintains metadata associated with each stage of generation of the synthetic healthcare dataset, including source references, transformation history, privacy configuration details, utility measurements, access records, and distribution events. The clinical audit trail module (318) continuously updates the synthetic data provenance record as additional processing events occur within the system (102) to ensure that the synthetic healthcare dataset remains traceable throughout its operational lifecycle. The synthetic data provenance record further enables association of the synthetic healthcare dataset with corresponding privacy-preserving operations executed by the differential privacy module (310), the marginal distribution learning module (312), the correlation constraint module (314), and the rejection sampling module (316). An example of the synthetic data provenance record includes, but is not limited to, a provenance record containing dataset creation timestamps, transformation identifiers, privacy budget allocation records, utility evaluation records, access logs, and distribution records associated with the synthetic healthcare dataset.

[0063] In an embodiment, the synthetic data provenance record comprises a generation provenance identifier corresponding to a source audit entry that records

the one or more privacy parameters, a privacy mechanism type, consumption of the privacy budget, and the one or more utility scores. The clinical audit trail module (318) generates the generation provenance identifier as a unique reference associated with a specific generation event of the synthetic healthcare dataset. The generation provenance identifier links the synthetic healthcare dataset to the source audit entry stored within the append-only audit log, thereby enabling traceability of privacy-preserving operations performed during generation of the synthetic healthcare dataset. The source audit entry records the one or more privacy parameters applied during transformation of the healthcare data, the privacy mechanism type utilized by the differential privacy module (310), cumulative or operation-level consumption of the privacy budget, and the one or more utility scores generated to evaluate usability of the synthetic healthcare dataset. The clinical audit trail module (318) uses the generation provenance identifier to retrieve generation-specific records for validation, auditing, and reproducibility of the synthetic healthcare dataset. An example of the privacy mechanism type includes, but is not limited to, a Laplace mechanism applied to numerical healthcare attributes, a Gaussian mechanism applied to observational healthcare attributes, or a randomized response mechanism applied to categorical healthcare attributes.

[0064] In an embodiment, the synthetic data provenance record further comprises source attribution information corresponding to the plurality of participating institutions contributing the healthcare data. The clinical audit trail module (318) records the source attribution information to identify each participating institution that contributes a portion of the healthcare data used for generation of the synthetic healthcare dataset. The source attribution information includes institutional identifiers, contribution timestamps, contribution volumes, authorization credentials, and dataset references associated with the plurality of participating institutions. The clinical audit trail module (318) links the source attribution information with the synthetic healthcare dataset to maintain traceability of institutional contributions throughout the lifecycle of the synthetic healthcare dataset. The source attribution information further enables verification of contribution authenticity and supports governance of multi-institutional data collaboration workflows. An example of the plurality of participating institutions

includes, but is not limited to, hospitals contributing patient treatment records, pharmaceutical organizations contributing drug trial datasets, academic medical centres contributing research datasets, and diagnostic laboratories contributing clinical observation datasets associated with the healthcare data.

[0065] In an embodiment, the synthetic data provenance record further comprises a downstream use registry configured to maintain one or more records corresponding to one or more consuming entities, one or more use categories, one or more timestamps, and one or more purpose descriptions associated with downstream use of the synthetic healthcare dataset. The clinical audit trail module (318) maintains the downstream use registry to track every authorized downstream interaction involving the synthetic healthcare dataset after generation of the synthetic healthcare dataset. The one or more records stored in the downstream use registry identify the one or more consuming entities accessing the synthetic healthcare dataset, classify intended usage through the one or more use categories, record the one or more timestamps associated with each access event, and document the one or more purpose descriptions corresponding to downstream utilization of the synthetic healthcare dataset. The clinical audit trail module (318) updates the downstream use registry whenever the synthetic healthcare dataset is transferred, accessed, processed, or utilized by an authorized consuming entity. The downstream use registry further enables traceability of downstream lifecycle events associated with the synthetic healthcare dataset. An example of the one or more consuming entities includes, but is not limited to, research institutions performing clinical studies, pharmaceutical organizations conducting drug development analysis, healthcare providers performing operational analytics, and artificial intelligence developers utilizing the synthetic healthcare dataset for machine learning model training.

[0066] In an embodiment, the synthetic data provenance record further comprises one or more regulatory submission records corresponding to the synthetic healthcare dataset. The clinical audit trail module (318) stores the one or more regulatory submission records to document formal submissions involving the synthetic healthcare dataset to regulatory authorities, compliance agencies, or approval bodies. Each of the one or more regulatory submission records includes

submission identifiers, submission timestamps, regulatory authority identifiers, submission status information, supporting compliance documentation references, and references to corresponding versions of the synthetic healthcare dataset. The clinical audit trail module (318) links the one or more regulatory submission records with the synthetic data provenance record to maintain traceability between regulatory interactions and generation history of the synthetic healthcare dataset. The one or more regulatory submission records further enable validation of whether the synthetic healthcare dataset has been utilized in regulatory workflows and whether associated privacy and utility requirements were satisfied during submission. An example of the one or more regulatory submission records includes, but is not limited to, a submission record associated with a clinical trial filing to FDA, a submission record associated with institutional review documentation submitted to an Institutional Review Board, and a submission record associated with privacy compliance reporting submitted under HIPAA requirements.

[0067] In an embodiment, the append-only audit log is maintained using at least one of a permissioned blockchain ledger, a multi-party computation trust layer, a trusted execution environment trust layer, or a conflict-free replicated data type trust layer. The clinical audit trail module (318) stores the append-only audit log within a distributed trust infrastructure to ensure integrity, tamper resistance, and verifiable synchronization of audit records associated with the synthetic healthcare dataset. The permissioned blockchain ledger maintains sequential audit entries through cryptographically linked blocks accessible only to authorized participating institutions. The multi-party computation trust layer enables distributed validation of audit records without exposing underlying healthcare data to any single participating institution. The trusted execution environment trust layer isolates execution of audit validation processes within a protected hardware environment to prevent unauthorized access or manipulation of audit records. The conflict-free replicated data type trust layer enables synchronized replication of the append-only audit log across distributed systems while ensuring consistency of audit entries during concurrent updates. The clinical audit trail module (318) selects one or more trust mechanisms based on deployment architecture, institutional participation requirements, and security policies

associated with the healthcare data. An example of the permissioned blockchain ledger includes, but is not limited to, a consortium blockchain deployed across hospitals, an example of the multi-party computation trust layer includes distributed validation nodes operated by research institutions, an example of the trusted execution environment trust layer includes secure enclave execution environments, and an example of the conflict-free replicated data type trust layer includes distributed ledger replicas maintained across geographically distributed healthcare networks.

[0068] In an embodiment, each audit trail entry of the append-only audit log comprises a cryptographic hash generated using contents of a preceding audit trail entry and contents of a current audit trail entry to form a cryptographically linked audit trail hash chain configured to provide tamper detection for the append-only audit log. The cryptographic hash is generated by combining audit information associated with the current audit trail entry and a hash value associated with the preceding audit trail entry. The cryptographically linked audit trail hash chain establishes a sequential dependency among audit trail entries such that each audit trail entry is mathematically linked to a preceding audit trail entry. The append-only audit log therefore maintains an immutable chronological sequence of audit activities associated with generation of the synthetic healthcare dataset. An example of the cryptographic hash includes, but is not limited to, a SHA-256 hash, a SHA-3 hash, a Blake2 hash, and a cryptographic digest generated using a secure hashing algorithm.

[0069] In an embodiment, the clinical audit trail module (318) further comprises a bilateral trust mode for deployments involving two participating institutions or three participating institutions. The bilateral trust mode enables coordinated governance of the healthcare data and the synthetic healthcare dataset in deployment environments where a limited number of participating institutions collaboratively contribute the healthcare data. The clinical audit trail module (318) activates the bilateral trust mode when the deployment architecture includes either the two participating institutions or the three participating institutions and applies trust validation rules specifically configured for limited-participant collaboration environments. The bilateral trust mode manages validation of privacy budget

consumption, synchronization of audit records, authorization of synthetic healthcare dataset generation requests, and verification of institutional participation rights associated with the healthcare data. The clinical audit trail module (318) further ensures that each participating institution maintains visibility into audit activities associated with shared processing operations while preventing unilateral modification of records associated with the healthcare data. An example of the two participating institutions includes, but is not limited to, a hospital and a pharmaceutical organization collaboratively generating the synthetic healthcare dataset for drug development analysis, and an example of the three participating institutions includes a hospital, a research institution, and a diagnostic laboratory collaboratively generating the synthetic healthcare dataset for clinical research.

[0070] In an embodiment, for deployments involving three participating institutions, the privacy budget transactions utilize a two-of-three signature threshold consensus mechanism. The clinical audit trail module (318) implements the two-of-three signature threshold consensus mechanism to validate privacy budget transactions associated with the healthcare data when the healthcare data is contributed by the three participating institutions. Under the two-of-three signature threshold consensus mechanism, approval of a privacy budget transaction requires authorization from at least two participating institutions before execution of the privacy budget transaction. The clinical audit trail module (318) transmits transaction requests to the three participating institutions, receives digital authorization signatures from the participating institutions, verifies authenticity of the digital authorization signatures, and executes the privacy budget transaction only after the required threshold is satisfied. The privacy budget transactions include allocation of the privacy budget, approval of synthetic healthcare dataset generation requests, modification of privacy configuration parameters, and validation of cumulative privacy budget consumption associated with the healthcare data. An example of the three participating institutions includes, but is not limited to, a hospital contributing clinical treatment records, a pharmaceutical organization contributing trial datasets, and a research institution contributing observational healthcare data, wherein approval from any two participating institutions authorizes execution of the privacy budget transaction.

[0071] In an embodiment, the audit streaming module (320) operatively coupled to the clinical audit trail module (318) is configured to publish each audit trail entry of the append-only audit log to one or more external monitoring systems as each audit trail entry is generated. The audit streaming module (320) continuously monitors creation of each audit trail entry within the append-only audit log and automatically transmits each audit trail entry to the one or more external monitoring systems in real time upon generation of each audit trail entry. The audit streaming module (320) converts each audit trail entry into a transmission-compatible format and securely transfers each audit trail entry through one or more communication channels to enable continuous monitoring of privacy-preserving operations associated with the healthcare data and the synthetic healthcare dataset. The one or more external monitoring systems receive each audit trail entry and perform operational monitoring, compliance monitoring, anomaly detection, and audit verification activities associated with the append-only audit log. The audit streaming module (320) further ensures synchronization between the append-only audit log maintained by the clinical audit trail module (318) and the one or more external monitoring systems. An example of the one or more external monitoring systems includes, but is not limited to, a regulatory compliance monitoring platform, a security information and event management platform, an institutional governance dashboard, a healthcare network monitoring server, and an external audit verification platform.

[0072] In an embodiment, the audit streaming module (320) is further configured to transmit a cryptographic hash corresponding to each audit trail entry of the append-only audit log for independent integrity verification. The audit streaming module (320) generates or retrieves the cryptographic hash corresponding to each audit trail entry after creation of each audit trail entry within the append-only audit log and transmits the cryptographic hash to the one or more external monitoring systems through a secure communication channel. The cryptographic hash uniquely represents contents of each audit trail entry and enables receiving systems to independently verify whether each audit trail entry has been altered, deleted, or replaced after generation of each audit trail entry. The audit streaming module (320) enables the one or more external monitoring systems to compare the cryptographic hash received during transmission with a

locally generated cryptographic hash derived from a corresponding audit trail entry for validation of data integrity. The audit streaming module (320) further maintains transmission records corresponding to the cryptographic hash for audit traceability. An example of the cryptographic hash includes, but is not limited to, a SHA-256 hash generated for each audit trail entry, a SHA-3 hash generated for distributed validation environments, and a cryptographic digest transmitted to an external compliance verification platform for integrity validation.

[0073] In an embodiment, the alert module (322) operatively coupled to the clinical audit trail module (318) is configured to generate one or more notifications when consumption of the privacy budget, the one or more utility scores, or one or more access patterns exceed one or more predefined thresholds. The alert module (322) continuously monitors operational metrics associated with the clinical audit trail module (318), including cumulative consumption of the privacy budget, performance measurements represented by the one or more utility scores, and access behaviour associated with the synthetic healthcare dataset. The alert module (322) compares real-time operational metrics with the one or more predefined thresholds to identify abnormal conditions, policy violations, or operational risks associated with generation and usage of the synthetic healthcare dataset. Upon detecting that consumption of the privacy budget exceeds a predefined threshold, that the one or more utility scores fall below an acceptable threshold, or that the one or more access patterns indicate abnormal activity, the alert module (322) generates the one or more notifications and transmits the one or more notifications to authorized entities for corrective action. The alert module (322) further records generation of the one or more notifications within the append-only audit log for traceability. An example of the one or more notifications includes, but is not limited to, a privacy budget exhaustion alert transmitted to system administrators, a low utility score alert transmitted to data governance teams, and an unauthorized access pattern alert transmitted to institutional security personnel.

[0074] In an embodiment, the regulatory detection module (324) operatively coupled to the clinical audit trail module (318) is configured to determine one or more applicable privacy regulations based on at least one of an institutional

location, a data type, an intended use declaration, or a jurisdictional requirement. The regulatory detection module (324) analyses regulatory attributes associated with the healthcare data and identifies legal and compliance obligations applicable to generation, storage, transfer, and downstream use of the synthetic healthcare dataset. The institutional location determines geographic applicability of privacy laws, the data type determines sensitivity classification and regulatory handling requirements, the intended use declaration identifies purpose-specific compliance obligations, and the jurisdictional requirement defines territorial restrictions applicable to processing activities. The regulatory detection module (324) compares the regulatory attributes with predefined regulatory rules and automatically determines the one or more applicable privacy regulations that govern processing of the healthcare data and the synthetic healthcare dataset. The regulatory detection module (324) further transmits the identified regulatory requirements to the clinical audit trail module (318) for documentation and compliance enforcement. An example of the one or more applicable privacy regulations includes, but is not limited to, Digital Personal Data Protection Act, 2023 requirements applicable to healthcare data originating from India, HIPAA requirements applicable to healthcare data originating from the United States, and General Data Protection Regulation requirements applicable to healthcare data.

[0075] In an embodiment, the adaptive budget optimization module (326) operatively coupled to the clinical audit trail module (318) is configured to analyse historical audit trail entries to build a statistical model of institutional query patterns including query frequency, consumption of the privacy budget, and distribution of the one or more utility scores. The adaptive budget optimization module (326) retrieves historical audit trail entries from the clinical audit trail module (318) and analyses prior privacy-consuming activities associated with the healthcare data and the synthetic healthcare dataset. The adaptive budget optimization module (326) identifies recurring institutional query behaviours by evaluating query frequency associated with participating institutions, historical consumption of the privacy budget for different query types, and distribution trends associated with the one or more utility scores generated after execution of prior queries. The statistical model of institutional query patterns is generated to predict future privacy budget requirements and optimize allocation of the privacy

budget for subsequent synthetic healthcare dataset generation requests. The adaptive budget optimization module (326) continuously updates the statistical model based on newly generated audit trail entries to improve predictive accuracy over time. An example of the institutional query patterns includes, but is not limited to, repeated clinical research queries executed by research institutions, recurring analytical queries executed by pharmaceutical organizations, frequent statistical reporting queries executed by hospitals, and query sequences associated with specific privacy budget consumption behaviours.

[0076] In an embodiment, the adaptive budget optimization module (326) is further configured to generate predictive privacy budget allocation recommendations for the plurality of participating institutions. The adaptive budget optimization module (326) utilizes the statistical model of institutional query patterns to forecast future privacy budget requirements associated with each participating institution contributing the healthcare data. The adaptive budget optimization module (326) evaluates historical consumption of the privacy budget, anticipated query demand, institutional usage behaviour, and projected utility requirements to generate predictive privacy budget allocation recommendations that optimize distribution of the privacy budget across the plurality of participating institutions. The predictive privacy budget allocation recommendations specify recommended privacy budget allocations, reservation limits, expected consumption windows, and priority levels associated with future synthetic healthcare dataset generation requests. The adaptive budget optimization module (326) further transmits the predictive privacy budget allocation recommendations to the clinical audit trail module (318) for implementation and tracking. An example of the predictive privacy budget allocation recommendations includes, but is not limited to, allocating a larger proportion of the privacy budget to a research institution performing repeated clinical trial analysis, reserving a defined portion of the privacy budget for a hospital conducting population health analytics, and limiting privacy budget allocation for low-priority exploratory queries initiated by a participating institution.

[0077] In an embodiment, the adaptive budget optimization module (326) is further configured to select a privacy composition mechanism from sequential

composition, advanced composition, and Rényi divergence-based composition based on a remaining privacy budget and one or more query characteristics. The adaptive budget optimization module (326) evaluates the remaining privacy budget associated with the healthcare data and analyses the one or more query characteristics prior to execution of a new privacy-consuming operation. The one or more query characteristics include query sensitivity, query frequency, expected privacy loss, computational complexity, and intended analytical purpose associated with a synthetic healthcare dataset generation request. Based on the evaluation, the adaptive budget optimization module (326) dynamically selects the privacy composition mechanism that optimizes preservation of the remaining privacy budget while maintaining required utility of the synthetic healthcare dataset. Sequential composition is selected when independent privacy-consuming operations require cumulative privacy accounting, advanced composition is selected when controlled relaxation of privacy loss bounds improves operational efficiency, and Rényi divergence-based composition is selected when tighter privacy accounting is required across repeated queries. The adaptive budget optimization module (326) further records selection of the privacy composition mechanism within the clinical audit trail module (318) for audit traceability. An example of the one or more query characteristics includes, but is not limited to, repeated low-sensitivity analytical queries processed using sequential composition, high-frequency research queries processed using advanced composition, and complex multi-institution analytical queries processed using Rényi divergence-based composition.

[0078] In an embodiment, the adaptive budget optimization module (326) is further configured to compute a Pareto frontier of privacy budget values and the one or more utility scores corresponding to a synthetic healthcare dataset generation request and automatically release unredeemed privacy budget reservations exceeding a predefined timeout period. The adaptive budget optimization module (326) evaluates multiple combinations of privacy budget values and corresponding utility outcomes associated with the synthetic healthcare dataset generation request to identify an optimal balance between privacy preservation and utility performance. The Pareto frontier represents a set of optimal privacy budget values where improvement in the one or more utility

scores results in increased privacy consumption and reduction in privacy consumption results in reduced utility performance. The adaptive budget optimization module (326) selects one or more recommended privacy budget values from the Pareto frontier based on institutional priorities, regulatory constraints, and operational objectives associated with generation of the synthetic healthcare dataset. The adaptive budget optimization module (326) further monitors reserved portions of the privacy budget allocated for pending synthetic healthcare dataset generation requests and automatically releases unredeemed privacy budget reservations when the predefined timeout period expires without execution of the corresponding request. An example of the synthetic healthcare dataset generation request includes, but is not limited to, a clinical research dataset request requiring high statistical fidelity, a pharmaceutical analytics request requiring preservation of clinical relationships, and a low-priority exploratory dataset request requiring limited privacy budget allocation.

[0079] In an embodiment, the jurisdictional processing module (328) operatively coupled to the ingestion module (308) and the clinical audit trail module (318) is configured to tag the healthcare data with a jurisdiction identifier. The jurisdictional processing module (328) analyses geographic, institutional, and regulatory metadata associated with the healthcare data to determine an originating jurisdiction corresponding to the healthcare data. The jurisdiction identifier is assigned to the healthcare data based on geographic location of the one or more healthcare data sources, location of the plurality of participating institutions, applicable regulatory territory, or origin of the one or more patients associated with the healthcare data. The jurisdictional processing module (328) embeds the jurisdiction identifier within metadata associated with the healthcare data to ensure traceability of jurisdiction-specific compliance requirements during subsequent processing of the healthcare data and generation of the synthetic healthcare dataset. The jurisdictional processing module (328) further transmits the jurisdiction identifier to the clinical audit trail module (318) for recording and compliance monitoring. An example of the jurisdiction identifier includes, but is not limited to, a jurisdiction identifier corresponding to India for healthcare data subject to Digital Personal Data Protection Act, 2023 requirements, a jurisdiction identifier corresponding to the United States for healthcare data subject to HIPAA

requirements, and a jurisdiction identifier corresponding to the European Union for healthcare data subject to General Data Protection Regulation requirements.

[0080] In an embodiment, the jurisdictional processing module (328) is further configured to prevent transmission of the healthcare data outside a jurisdiction of origin. The jurisdictional processing module (328) continuously monitors movement of the healthcare data across processing environments and compares each transmission request with the jurisdiction identifier associated with the healthcare data. The jurisdictional processing module (328) determines whether a requested transmission destination is located outside the jurisdiction of origin and automatically blocks the requested transmission when the requested transmission violates jurisdictional restrictions applicable to the healthcare data. The jurisdictional processing module (328) enforces geographic containment policies, regulatory transfer restrictions, and institutional data residency requirements to ensure that the healthcare data remains within the jurisdiction of origin unless authorized processing conditions are satisfied. The jurisdictional processing module (328) further records each blocked transmission event and each authorized transmission event within the clinical audit trail module (318) for traceability. An example of the jurisdiction of origin includes, but is not limited to, healthcare data originating from India restricted under Digital Personal Data Protection Act, 2023 requirements, healthcare data originating from the United States subject to HIPAA restrictions, and healthcare data originating from the European Union subject to General Data Protection Regulation cross-border transfer restrictions.

[0081] In an embodiment, the jurisdictional processing module (328) is further configured to permit transfer of the synthetic healthcare dataset across jurisdictions based on generation of a compliance certificate. The records associated with cross-jurisdictional transfer of the synthetic healthcare dataset are stored in a database (346). The jurisdictional processing module (328) evaluates whether the synthetic healthcare dataset satisfies privacy, regulatory, and jurisdictional transfer requirements prior to authorizing cross-jurisdictional transfer of the synthetic healthcare dataset. Upon determining that the synthetic healthcare dataset complies with applicable privacy regulations and no longer

contains identifiable healthcare data, the jurisdictional processing module (328) permits transfer of the synthetic healthcare dataset across jurisdictions after generation of the compliance certificate by the output generation module (336). The compliance certificate validates that the synthetic healthcare dataset satisfies privacy thresholds, regulatory requirements, and jurisdiction-specific transfer conditions applicable to the healthcare data. The jurisdictional processing module (328) stores records associated with cross-jurisdictional transfer of the synthetic healthcare dataset, including destination identifiers, transfer timestamps, authorization records, and compliance references for future traceability. An example of cross-jurisdictional transfer of the synthetic healthcare dataset includes, but is not limited to, transfer of the synthetic healthcare dataset from India to the United States for pharmaceutical analytics, transfer of the synthetic healthcare dataset from India to the United States for clinical research, transfer of the synthetic healthcare dataset from the United States to the European Union for machine learning model development, and transfer of the synthetic healthcare dataset from the European Union to India for healthcare analytics.

[0082] In an embodiment, the jurisdictional processing module (328) is deployed within a network boundary corresponding to a healthcare data holder, including an on-premise infrastructure environment, a private cloud environment, a virtual private cloud environment, an institutional data centre environment, or an edge computing environment. The processor (302) executes within the network boundary such that healthcare data remains stored and processed within the network boundary throughout ingestion, privacy-preserving transformation, privacy budget accounting, and synthetic healthcare dataset generation. The jurisdictional processing module (328) prevents transmission of the healthcare data outside the network boundary and permits external transmission of only the synthetic healthcare dataset, the compliance certificate, and one or more cryptographic hashes corresponding to audit trail entries. An example of the network boundary includes, but is not limited to, a hospital data centre, a research institution private cloud environment, a pharmaceutical organization virtual private cloud environment, and a government healthcare infrastructure environment.

[0083] In an embodiment, the privacy budget composition tracking module (330) operatively coupled to the clinical audit trail module (318) is configured to maintain a cumulative privacy loss ledger corresponding to the healthcare data using sequential composition, advanced composition, or Rényi divergence-based composition. The privacy budget composition tracking module (330) continuously tracks cumulative privacy loss resulting from multiple privacy-consuming operations performed on the healthcare data during generation of the synthetic healthcare dataset. The privacy budget composition tracking module (330) calculates cumulative privacy consumption by applying sequential composition when privacy losses from independent operations are aggregated, applying advanced composition when bounded relaxation of cumulative privacy loss is required, or applying Rényi divergence-based composition when tighter privacy accounting is required for repeated analytical operations. The cumulative privacy loss ledger stores privacy consumption records associated with each privacy-preserving operation, cumulative privacy loss values, applied composition methodologies, and remaining privacy budget values corresponding to the healthcare data. The privacy budget composition tracking module (330) further synchronizes the cumulative privacy loss ledger with the clinical audit trail module (318) to ensure continuous audit traceability. An example of the privacy-consuming operations includes, but is not limited to, marginal distribution analysis performed by the marginal distribution learning module (312), correlation preservation operations performed by the correlation constraint module (314), and repeated synthetic healthcare dataset generation requests executed by the differential privacy module (310).

[0084] In an embodiment, data associated with the privacy composition mechanism applied to one or more privacy-consuming operations is stored by the privacy budget composition tracking module (330) within the cumulative privacy loss ledger and the clinical audit trail module (318). The privacy budget composition tracking module (330) records operational details corresponding to each privacy composition mechanism selected for execution of the one or more privacy-consuming operations performed on the healthcare data. The stored data includes identifiers of the privacy composition mechanism, timestamps of execution, privacy loss values associated with each operation, remaining privacy

budget values, query identifiers, and references to corresponding modules executing the one or more privacy-consuming operations. The privacy budget composition tracking module (330) links the stored data with corresponding entries in the cumulative privacy loss ledger to maintain traceability of privacy accounting decisions throughout generation of the synthetic healthcare dataset. The privacy budget composition tracking module (330) further enables retrieval of the stored data for regulatory validation, audit review, and operational optimization. An example of the one or more privacy-consuming operations includes, but is not limited to, a statistical query processed using sequential composition, a repeated analytical query processed using advanced composition, and a high-frequency multi-institution query processed using Rényi divergence-based composition.

[0085] In an embodiment, the consent lifecycle management module (332) operatively coupled to the ingestion module (308) is configured to maintain a consent registry corresponding to the one or more patients. The consent lifecycle management module (332) stores and manages consent-related information associated with the one or more patients whose healthcare data is received by the ingestion module (308). The consent registry maintains records identifying consent status, consent scope, consent duration, consent modification history, and authorized usage categories associated with the healthcare data corresponding to the one or more patients. The consent lifecycle management module (332) continuously updates the consent registry based on newly received consent records, modifications to existing consent permissions, and revocation events associated with the one or more patients. The consent lifecycle management module (332) further enables validation of consent permissions prior to allowing processing of the healthcare data for generation of the synthetic healthcare dataset. An example of the consent registry includes, but is not limited to, a registry storing patient authorization for clinical research usage, a registry storing patient authorization for pharmaceutical analytics usage, and a registry storing patient restrictions prohibiting use of healthcare data for commercial machine learning applications.

[0086] In an embodiment, the consent lifecycle management module (332) is further configured to exclude the healthcare data corresponding to a patient from subsequent generation of the synthetic healthcare dataset upon receiving a consent revocation event and generate a consent revocation certificate. The consent lifecycle management module (332) continuously monitors the consent registry for updates indicating withdrawal of previously granted consent by the patient. Upon receiving the consent revocation event, the consent lifecycle management module (332) identifies the healthcare data corresponding to the patient and prevents further utilization of the healthcare data in subsequent synthetic healthcare dataset generation processes. The consent lifecycle management module (332) communicates the exclusion instruction to the ingestion module (308), the differential privacy module (310), and the clinical audit trail module (318) to ensure that future processing operations do not utilize the healthcare data corresponding to the patient. The consent lifecycle management module (332) further generates the consent revocation certificate documenting the consent revocation event, exclusion actions performed, timestamps associated with the consent revocation event, and identifiers corresponding to affected datasets. An example of the consent revocation event includes, but is not limited to, withdrawal of patient authorization for clinical research usage, revocation of consent for pharmaceutical analytics usage, and revocation of consent for artificial intelligence model development usage.

[0087] In an embodiment, the output generation module (336) operatively coupled to the differential privacy module (310) and the clinical audit trail module (318) is configured to generate the synthetic healthcare dataset formatted according to the healthcare interoperability standard and generate the compliance certificate corresponding to the synthetic healthcare dataset. The output generation module (336) receives privacy-preserved synthetic records from the differential privacy module (310) and retrieves audit records, provenance records, and compliance-related metadata from the clinical audit trail module (318). The output generation module (336) structures the synthetic healthcare dataset in accordance with the healthcare interoperability standard to ensure compatibility with downstream healthcare systems, research platforms, and analytical environments. The output generation module (336) preserves resource relationships, attribute

mappings, and formatting requirements associated with the healthcare interoperability standard while generating the synthetic healthcare dataset. The output generation module (336) further generates the compliance certificate to document privacy compliance status, utility validation outcomes, provenance references, and regulatory verification information associated with the synthetic healthcare dataset. An example of the synthetic healthcare dataset formatted according to the healthcare interoperability standard includes, but is not limited to, a synthetic patient resource dataset, a synthetic observation resource dataset, a synthetic condition resource dataset, and a synthetic medication resource dataset structured in accordance with HL7 Fast Healthcare Interoperability Resources.

[0088] In an embodiment, the compliance certificate comprises privacy validation data, statistical fidelity metrics, correlation preservation metrics, and clinical plausibility metrics corresponding to the synthetic healthcare dataset. The output generation module (336) compiles the compliance certificate by aggregating validation information generated during processing of the healthcare data and generation of the synthetic healthcare dataset. The privacy validation data confirms that privacy-preserving operations executed by the differential privacy module (310) satisfy defined privacy requirements and records privacy parameters associated with generation of the synthetic healthcare dataset. The statistical fidelity metrics evaluate similarity between the healthcare data and the synthetic healthcare dataset at an attribute distribution level. The correlation preservation metrics evaluate whether clinically significant relationships associated with the healthcare data are retained within the synthetic healthcare dataset. The clinical plausibility metrics evaluate whether synthetic records generated within the synthetic healthcare dataset satisfy medically consistent patterns and clinically valid relationships. The output generation module (336) further links the compliance certificate with the synthetic data provenance record maintained by the clinical audit trail module (318). An example of the statistical fidelity metrics includes, but is not limited to, distribution similarity scores and attribute variance comparisons, an example of the correlation preservation metrics includes diagnosis-to-observation correlation retention scores, and an example of the clinical plausibility metrics includes rule-based validation scores confirming medically consistent synthetic records.

[0089] In an embodiment, the synthetic healthcare dataset is generated in at least one of JavaScript Object Notation format, extensible markup language format, or a format compatible with the healthcare interoperability standard. The output generation module (336) determines an output format requirement associated with downstream systems, regulatory workflows, or analytical applications prior to generation of the synthetic healthcare dataset. The output generation module (336) serializes synthetic records into JavaScript Object Notation format when lightweight structured data exchange is required, serializes synthetic records into extensible markup language format when hierarchical document representation is required, or generates the synthetic healthcare dataset in a format compatible with the healthcare interoperability standard when interoperability with healthcare systems is required. The output generation module (336) preserves structural integrity, attribute mappings, and resource relationships during conversion of the synthetic healthcare dataset into the selected output format. The output generation module (336) further validates the generated output format to ensure compatibility with intended receiving systems. An example of the format compatible with the healthcare interoperability standard includes, but is not limited to, HL7 Fast Healthcare Interoperability Resources JavaScript Object Notation bundles, HL7 Fast Healthcare Interoperability Resources extensible markup language resources, and structured interoperability packages configured for clinical research platforms.

[0090] It must be noted that the system (102) addresses a technical problem associated with preventing transmission of raw healthcare data across a network boundary or a jurisdictional boundary while preserving analytical utility of healthcare data for distributed processing across a plurality of participating institutions. Conventional distributed healthcare data processing architectures require transmission or centralized aggregation of healthcare data, thereby increasing risks associated with unauthorized disclosure, re-identification, regulatory non-compliance, and network-based exposure of sensitive patient information. Further, conventional audit mechanisms generally lack independently verifiable integrity protection capable of detecting unauthorized modification of audit information generated during privacy-preserving processing.

[0091] It must be noted that the system (102) provides a technical effect by confining storage and processing of the healthcare data within a network boundary corresponding to a healthcare data holder while permitting transmission of only the synthetic healthcare dataset, the compliance certificate, and cryptographic hashes corresponding to audit trail entries outside the network boundary. The append-only audit log utilizes a cryptographically linked audit trail hash chain to provide independently verifiable and tamper-evident integrity verification of audit information. Accordingly, the system (102) improves operation of a networked computing environment by reducing transmission of raw healthcare data across network boundaries, preserving jurisdictional processing containment, enabling secure multi-institution processing, and providing cryptographic integrity verification of distributed audit information while preserving analytical utility of the synthetic healthcare dataset.

[0092] In a non-limiting exemplary implementation, the system (102) is deployed across a multi-institution healthcare research network comprising a hospital, a pharmaceutical organization, and a research institution seeking to collaboratively develop predictive treatment models for diabetes without exposing raw healthcare data of patients. The system (102) is deployed on a cloud-based healthcare analytics infrastructure, wherein the processor (302) is hosted on secure servers accessible to authorized participating institutions through encrypted communication channels. The hospital provides patient diagnostic records, laboratory observations, medication histories, and encounter records through the ingestion module (308). The pharmaceutical organization provides clinical trial datasets related to diabetes drug outcomes, and the research institution provides longitudinal observational datasets associated with treatment efficacy.

[0093] The ingestion module (308) receives healthcare data from electronic health record systems, hospital databases, pharmaceutical databases, and research institution databases. The ingestion module (308) retrieves the healthcare data structured according to HL7 Fast Healthcare Interoperability Resources and parses healthcare resource graphs comprising patient resources, observation resources, condition resources, medication resources, and encounter resources. During parsing, the ingestion module (308) identifies privacy-sensitive attributes

including patient identifiers, diagnosis records, genomic markers, and treatment histories.

[0094] The differential privacy module (310) transforms the healthcare data into a synthetic healthcare dataset by applying calibrated noise based on privacy parameters. For example, blood glucose records and medication usage records are transformed while preserving relationships between diabetes diagnoses and glucose levels. The marginal distribution learning module (312) learns age distributions, diagnosis frequency distributions, and laboratory value distributions. The correlation constraint module (314) generates probabilistic constraints ensuring that synthetic diabetic patients continue to exhibit clinically valid glucose patterns. The rejection sampling module (316) discards synthetic samples that violate clinical constraints, such as synthetic patients receiving insulin prescriptions without diabetes diagnoses.

[0095] The clinical audit trail module (318) records every transformation event performed on the healthcare data. Each statistical query deducts privacy units from a privacy budget. The clinical audit trail module (318) generates an append-only audit log containing hashed transaction records linking privacy consumption with utility scores. When the pharmaceutical organization submits repeated dataset generation requests that exceed a global privacy threshold, the clinical audit trail module (318) automatically rejects additional requests. The clinical audit trail module (318) further maintains a synthetic data provenance record documenting source attribution information for the hospital, the pharmaceutical organization, and the research institution.

[0096] The audit streaming module (320) publishes each audit entry to external compliance systems operated by the participating institutions and transmits cryptographic hashes for independent verification. The alert module (322) generates alerts when privacy budget consumption exceeds predefined thresholds. The automatic regulatory detection module (324) determines whether HIPAA, Digital Personal Data Protection Act, 2023, or General Data Protection Regulation requirements apply based on institutional location and intended use declarations.

[0097] The adaptive budget optimization module (326) analyses historical research queries and allocates privacy budgets to each participating institution. The adaptive budget optimization module (326) selects Rényi divergence-based composition for repeated analytical queries and computes a Pareto frontier balancing privacy preservation and dataset utility. The jurisdictional processing module (328) tags healthcare data originating from India, the United States, and the with corresponding jurisdiction identifiers. Raw healthcare data remains within the jurisdiction of origin, while the synthetic healthcare dataset is permitted to move across jurisdictions after validation of a compliance certificate.

[0098] The privacy budget composition tracking module (330) maintains cumulative privacy loss records for repeated research operations. The consent lifecycle management module (332) manages patient consent records and removes healthcare data when a patient revokes participation.

[0099] Finally, the output generation module (336) generates the synthetic healthcare dataset in HL7 Fast Healthcare Interoperability Resources JavaScript Object Notation format and generates a compliance certificate. The pharmaceutical organization utilizes the synthetic healthcare dataset to train drug response models, the research institution performs population-level diabetes analytics, and the hospital evaluates treatment effectiveness without exposing raw healthcare data. This deployment enables real-world collaborative healthcare innovation while preserving privacy, maintaining compliance, and reducing delays associated with traditional data-sharing agreements.

[0100] FIG. 4 (a) is a flow chart representing the steps involved in a method for privacy-preserving synthetic healthcare data generation, in accordance with an embodiment of the present disclosure; and FIG. 4 (b) illustrates continued steps of the method of FIG. 4 (a) in accordance with an embodiment of the present disclosure.

[0101] The method (400) includes receiving a healthcare data corresponding to one or more patients from one or more healthcare data sources and a user operating a user device via user interface in step 404. The healthcare data includes clinical and administrative information associated with the one or more patients,

including demographic records, diagnostic records, laboratory observations, medication records, encounter records, and procedure records. The one or more healthcare data sources provide access to the healthcare data through secure communication channels for downstream privacy-preserving processing. The method further validates integrity and completeness of the healthcare data prior to subsequent operations.

[0102] The method (400) includes retrieving the healthcare data provided according to a healthcare interoperability standard in step 406. An example of the healthcare interoperability standard includes, but is not limited to, HL7 Fast Healthcare Interoperability Resources comprising patient resources, observation resources, condition resources, medication resources, encounter resources, and procedure resources.

[0103] The method (400) includes parsing a plurality of healthcare resource graphs associated with the healthcare data to identify one or more privacy-sensitive attributes associated with the healthcare data in step 408. An example of the plurality of healthcare resource graphs includes, but is not limited to, patient resources linked with observation resources, condition resources, medication resources, encounter resources, and procedure resources.

[0104] The method (400) includes transforming the healthcare data into a synthetic healthcare dataset by applying calibrated noise to the healthcare data based on one or more privacy parameters in step 410. The method applies the calibrated noise to clinical attributes contained within the healthcare data to reduce disclosure risks associated with the one or more patients while preserving statistical characteristics required for downstream analysis. The one or more privacy parameters define noise magnitude, privacy thresholds, and privacy loss constraints associated with generation of the synthetic healthcare dataset. The method further evaluates transformed outputs to ensure compliance with differential privacy requirements. An example of the calibrated noise includes, but is not limited to, Laplace noise applied to numerical healthcare attributes, Gaussian noise applied to laboratory observations, and randomized response mechanisms applied to categorical healthcare attributes.

[0105] The method (400) includes preserving one or more clinical relationships associated with the healthcare data while satisfying differential privacy requirements in step 412. An example of the one or more clinical relationships includes, but is not limited to, an association between a diabetes diagnosis and elevated glucose levels, an association between hypertension and elevated blood pressure values, and an association between prescribed medication and a corresponding medical condition.

[0106] The method (400) includes learning one or more feature distributions corresponding to the healthcare data by applying differentially private noise to one or more marginal distributions associated with the healthcare data in step 414. An example of the one or more feature distributions includes, but is not limited to, age distributions, diagnosis frequency distributions, laboratory value distributions, and medication frequency distributions.

[0107] The method (400) includes generating one or more probabilistic constraints corresponding to one or more clinical attribute pairs associated with the healthcare data in step 416. An example of the one or more clinical attribute pairs includes, but is not limited to, a diabetes diagnosis and elevated glucose levels, hypertension and elevated blood pressure readings, and prescribed medication associated with a corresponding diagnosis.

[0108] The method (400) includes discarding one or more synthetic samples that violate the one or more probabilistic constraints in step 418. The method evaluates generated synthetic samples against the one or more probabilistic constraints to determine whether each synthetic sample satisfies predefined dependency rules, conditional probability requirements, and clinical consistency requirements associated with the healthcare data. Upon determining that a synthetic sample violates the one or more probabilistic constraints, the method removes the synthetic sample from the synthetic healthcare dataset generation process and initiates generation of an alternate synthetic sample. The method further retains synthetic samples satisfying the one or more probabilistic constraints.

[0109] The method (400) includes treating a privacy budget as a finite computational resource and deduct one or more privacy units from the privacy budget for each statistical query executed against the healthcare data in step 422. The method initializes the privacy budget prior to execution of privacy-preserving operations and tracks consumption of the privacy budget during each statistical query performed on the healthcare data. The one or more privacy units are deducted based on query sensitivity, query complexity, and privacy requirements associated with the statistical query. The method further updates remaining privacy capacity after execution of each statistical query. An example of the statistical query includes, but is not limited to, age distribution analysis, diagnosis frequency analysis, laboratory value analysis, and clinical correlation analysis.

[0110] The method (400) includes generating an append-only audit log comprising one or more hashed transaction records linking consumption of the privacy budget with one or more utility scores corresponding to the synthetic healthcare dataset audit trail entry in step 424. An example of the one or more utility scores includes, but is not limited to, statistical fidelity scores, correlation preservation scores, and clinical plausibility scores.

[0111] In an embodiment, each audit trail entry of the append-only audit log comprises a cryptographic hash generated using contents of a preceding audit trail entry and contents of a current audit trail entry to form a cryptographically linked audit trail hash chain configured to provide tamper detection for the append-only audit log.

[0112] The method (400) further includes maintaining the append-only audit log using at least one of a permissioned blockchain ledger, a multi-party computation trust layer, a trusted execution environment trust layer, or a conflict-free replicated data type trust layer, and transmitting a cryptographic hash corresponding to each audit trail entry for independent integrity verification. An example of the trust layer includes, but is not limited to, a permissioned blockchain ledger, a multi-party computation trust layer, a trusted execution environment trust layer, and a conflict-free replicated data type trust layer.

[0113] The method (400) further includes implementing a bilateral trust mode for deployments involving two or three participating institutions, wherein, for deployments involving two participating institutions, privacy budget transactions require bilateral cryptographic signatures from the two participating institutions and tamper detection is performed through cross-verification of replicated audit trail hash chains, and wherein, for deployments involving three participating institutions, the privacy budget transactions utilize a two-of-three signature threshold consensus mechanism. An example of the participating institutions includes, but is not limited to, hospitals, pharmaceutical organizations, research institutions, and clinical trial centres.

[0114] The method (400) further includes tagging the healthcare data with a jurisdiction identifier, preventing transmission of the healthcare data outside a jurisdiction of origin, and permitting transfer of the synthetic healthcare dataset across jurisdictions based on generation of the compliance certificate, wherein the method (400) is performed within a network boundary corresponding to a healthcare data holder such that the healthcare data remains stored and processed within the network boundary throughout ingestion, transformation, privacy budget accounting, and generation of the synthetic healthcare dataset.

[0115] The method (400) includes generating the synthetic healthcare dataset formatted according to the healthcare interoperability standard and generate the compliance certificate corresponding to the synthetic healthcare dataset in step 464. The method receives privacy-preserved synthetic records and structures the synthetic healthcare dataset in accordance with formatting requirements, schema definitions, and interoperability protocols associated with the healthcare interoperability standard. The method preserves resource relationships and attribute mappings while generating the synthetic healthcare dataset for compatibility with downstream healthcare systems and analytical platforms. The method further generates the compliance certificate containing validation information associated with privacy preservation and dataset quality. An example of the healthcare interoperability standard includes, but is not limited to, HL7 Fast Healthcare Interoperability Resources.

[0116] In an embodiment, the synthetic healthcare dataset is generated in at least one of JavaScript Object Notation format, extensible markup language format, or a format compatible with the healthcare interoperability standard. An example of the format compatible with the healthcare interoperability standard includes, but is not limited to, HL7 Fast Healthcare Interoperability Resources bundles, resource collections, and interoperable healthcare message structures.

[0117] Thus, various embodiments of the system and method for privacy-preserving synthetic healthcare data generation provides several benefits. The invention provides significant advantages by enabling generation of a synthetic healthcare dataset while preserving privacy of healthcare data corresponding to one or more patients through controlled privacy-preserving mechanisms. The invention preserves clinically meaningful relationships within the healthcare data, thereby improving realism and analytical usability of the synthetic healthcare dataset for research, pharmaceutical analysis, clinical studies, and artificial intelligence applications. The invention further enables immutable audit traceability through maintenance of audit records, privacy budget tracking, and synthetic data provenance records, thereby improving regulatory transparency and compliance readiness. The invention supports secure collaboration among a plurality of participating institutions without exposing raw healthcare data and prevents unauthorized cross-jurisdictional transmission of regulated healthcare data. Additionally, the invention improves consent management, adaptive privacy allocation, interoperability, and operational efficiency while reducing risks of re-identification, privacy leakage, and regulatory non-compliance.

[0118] The techniques described in this disclosure may be implemented, at least in part, in hardware, software, firmware, or any combination thereof. For example, various aspects of the described techniques may be implemented within one or more processors, including one or more microprocessors, digital signal processors (DSPs), application-specific integrated circuits (ASICs), field-programmable gate arrays (FPGAs), or any other equivalent integrated or discrete logic circuitry, as well as any combinations of such components. The term "processor" or "processing subsystem" may generally refer to any of the foregoing logic circuitry, alone or in combination with other logic circuitry, or any other

equivalent circuitry. A control unit including hardware may also perform one or more of the techniques of this disclosure.

[0119] Such hardware, software, and firmware may be implemented within the same device or within separate devices to support the various techniques described in this disclosure. In addition, any of the described units, modules, or components may be implemented together or separately as discrete but interoperable logic devices. Depiction of different features as modules or units is intended to highlight different functional aspects and does not necessarily imply that such modules or units must be realized by separate hardware, firmware, or software components. Rather, functionality associated with one or more modules or units may be performed by separate hardware, firmware, or software components, or integrated within common or separate hardware, firmware, or software components.

[0120] It will be understood by those skilled in the art that the foregoing general description and the following detailed description are exemplary and explanatory of the disclosure and are not intended to be restrictive thereof.

[0121] While specific language has been used to describe the disclosure, any limitations arising on account of the same are not intended. As would be apparent to a person skilled in the art, various working modifications may be made to the method in order to implement the inventive concept as taught herein.

[0122] The figures and the foregoing description give examples of embodiments. Those skilled in the art will appreciate that one or more of the described elements may well be combined into a single functional element. Alternatively, certain elements may be split into multiple functional elements. Elements from one embodiment may be added to another embodiment. For example, the order of processes described herein may be changed and are not limited to the manner described herein. Moreover, the actions of any flow diagram need not be implemented in the order shown; nor do all of the acts need to be necessarily performed. Also, those acts that are not dependent on other acts may be performed in parallel with the other acts. The scope of embodiments is by no means limited by these specific examples.

I CLAIM:

1. A system for privacy-preserving synthetic healthcare data generation, comprising:

a processor;

a memory coupled to the processor, wherein the memory comprises instructions that, when executed by the processor, cause the processor to:

receive a healthcare data corresponding to one or more patients from one or more healthcare data sources and a user operating a user device via a user interface;

retrieve the healthcare data provided according to a healthcare interoperability standard;

parse a plurality of healthcare resource graphs associated with the healthcare data to identify one or more privacy-sensitive attributes associated with the healthcare data;

transform the healthcare data into a synthetic healthcare dataset by applying calibrated noise to the healthcare data based on one or more privacy parameters;

preserve one or more clinical relationships associated with the healthcare data while satisfying differential privacy requirements;

learn one or more feature distributions corresponding to the healthcare data by applying differentially private noise to one or more marginal distributions associated with the healthcare data;

generate one or more probabilistic constraints corresponding to one or more clinical attribute pairs associated with the healthcare data;

discard one or more synthetic samples that violate the one or more probabilistic constraints;

treat a privacy budget as a finite computational resource and deduct one or more privacy units from the privacy budget for each statistical query executed against the healthcare data;

generate an append-only audit log comprising one or more hashed transaction records linking consumption of the privacy budget with one or more utility scores corresponding to the synthetic healthcare dataset; and

generate the synthetic healthcare dataset formatted according to the healthcare interoperability standard and generate a compliance certificate corresponding to the synthetic healthcare dataset.

2. The system as claimed in claim 1, wherein the healthcare interoperability standard comprises HL7 Fast Healthcare Interoperability Resources.

3. The system as claimed in claim 1, wherein the plurality of healthcare resource graphs comprises at least one of patient resources, observation resources, condition resources, medication resources, encounter resources, or procedure resources.

4. The system as claimed in claim 1, to cause the processor to:

maintain one or more records corresponding to each data transformation operation;

implement a bilateral trust mode for deployments involving two or three participating institutions,

wherein, for deployments involving two participating institutions, privacy budget transactions require bilateral cryptographic signatures from the two participating institutions and tamper detection is performed through cross-verification of replicated audit trail hash chains, and

wherein, for deployments involving three participating institutions, the privacy budget transactions utilize a two-of-three signature threshold consensus mechanism; and

reject one or more synthetic healthcare dataset generation requests when cumulative consumption of the privacy budget exceeds a global privacy threshold across the plurality of participating institutions.

5. The system as claimed in claim 1, to cause the processor to maintain a synthetic data provenance record corresponding to the synthetic healthcare dataset, wherein each audit trail entry of the append-only audit log comprises an epsilon parameter, a delta parameter, and a noise calibration record corresponding to generation of the synthetic healthcare dataset.

6. The system as claimed in claim 1, to cause the processor to:

publish each audit trail entry of the append-only audit log to one or more external monitoring systems as each audit trail entry is generated;

transmit a cryptographic hash corresponding to each audit trail entry for independent integrity verification, wherein the append-only audit log is maintained using at least one of a permissioned blockchain ledger, a multi-party computation trust layer, a trusted execution environment trust layer, or a conflict-free replicated data type trust layer; and

maintain the append-only audit log using at least one of a permissioned blockchain ledger, a multi-party computation trust layer, a trusted execution environment trust layer, or a conflict-free replicated data type trust layer.

7. The system as claimed in claim 6, wherein each audit trail entry of the append-only audit log comprises a cryptographic hash generated using contents of a preceding audit trail entry and contents of a current audit trail entry to form a cryptographically linked audit trail hash chain configured to provide tamper detection for the append-only audit log.

8. The system as claimed in claim 1,

wherein the synthetic data provenance record comprises a generation provenance identifier corresponding to a source audit entry that records the one or more privacy parameters, a privacy mechanism type, consumption of the privacy budget, and the one or more utility scores,

wherein the synthetic data provenance record further comprises a downstream use registry configured to maintain one or more records corresponding to one or more consuming entities, one or more use categories, one

or more timestamps, and one or more purpose descriptions associated with downstream use of the synthetic healthcare dataset,

wherein the synthetic data provenance record further comprises one or more regulatory submission records corresponding to the synthetic healthcare dataset, and

wherein the compliance certificate comprises privacy validation data, statistical fidelity metrics, correlation preservation metrics, and clinical plausibility metrics corresponding to the synthetic healthcare dataset.

9. The system as claimed in claim 5, to cause the processor to determine one or more applicable privacy regulations based on at least one of an institutional location, a data type, an intended use declaration, or a jurisdictional requirement, wherein the compliance certificate comprises privacy validation data, statistical fidelity metrics, correlation preservation metrics, and clinical plausibility metrics corresponding to the synthetic healthcare dataset.

10. The system as claimed in claim 1, to cause the processor to generate the synthetic healthcare dataset for a rare disease cohort comprising fewer than 200 patients.

11. The system as claimed in claim 1, to cause the processor to allocate a larger proportion of the privacy budget to preservation of the one or more clinical relationships when a patient cohort size is below a predefined threshold.

12. The system as claimed in claim 1, to cause the processor to:

analyse historical audit trail entries to build a statistical model of institutional query patterns including query frequency, consumption of the privacy budget, and distribution of the one or more utility scores,

generate predictive privacy budget allocation recommendations for the plurality of participating institutions;

select a privacy composition mechanism from sequential composition, advanced composition, and Rényi divergence-based composition based on a remaining privacy budget and one or more query characteristics; and

compute a Pareto frontier of privacy budget values and the one or more utility scores corresponding to a synthetic healthcare dataset generation request and automatically release unredeemed privacy budget reservations exceeding a predefined timeout period.

13. The system as claimed in claim 1, to cause the processor to:

generate one or more notifications when consumption of the privacy budget, the one or more utility scores, or one or more access patterns exceed one or more predefined thresholds;

tag the healthcare data with a jurisdiction identifier;

prevent transmission of the healthcare data outside a jurisdiction of origin;

permit transfer of the synthetic healthcare dataset across jurisdictions based on generation of the compliance certificate;

maintain a cumulative privacy loss ledger corresponding to the healthcare data using sequential composition, advanced composition, or Rényi divergence-based composition;

store data associated with a privacy composition mechanism applied to one or more privacy-consuming operations; and

maintain a consent registry corresponding to the one or more patients;

exclude the healthcare data corresponding to a patient from subsequent generation of the synthetic healthcare dataset upon receiving a consent revocation event and generate a consent revocation certificate,

wherein the one or more healthcare data sources comprise at least one of an electronic health record system, a hospital database, a pharmaceutical database, a research institution database, or a clinical trial database.

14. The system as claimed in claim 1, wherein the synthetic healthcare dataset is generated in at least one of JavaScript Object Notation format, extensible markup language format, or a format compatible with the healthcare interoperability standard.

15. The system as claimed in claim 1, to cause the processor to:

prevent transmission of the healthcare data outside a jurisdiction of origin, permit transfer of the synthetic healthcare dataset across jurisdictions based on generation of the compliance certificate;

maintain a cumulative privacy loss ledger corresponding to the healthcare data using sequential composition, advanced composition, or Rényi divergence-based composition;

store data associated with a privacy composition mechanism applied to one or more privacy-consuming operations; and

exclude the healthcare data corresponding to a patient from subsequent generation of the synthetic healthcare dataset upon receiving a consent revocation event and generate a consent revocation certificate, wherein the one or more healthcare data sources comprise at least one of an electronic health record system, a hospital database, a pharmaceutical database, a research institution database, or a clinical trial database.

16. A method for privacy-preserving synthetic healthcare data generation, comprising:

receiving a healthcare data corresponding to one or more patients from one or more healthcare data sources and a user operating a user device via a user interface;

retrieving the healthcare data provided according to a healthcare interoperability standard;

parsing a plurality of healthcare resource graphs associated with the healthcare data to identify one or more privacy-sensitive attributes associated with the healthcare data;

transforming the healthcare data into a synthetic healthcare dataset by applying calibrated noise to the healthcare data based on one or more privacy parameters;

preserving one or more clinical relationships associated with the healthcare data while satisfying differential privacy requirements;

learning one or more feature distributions corresponding to the healthcare data by applying differentially private noise to one or more marginal distributions associated with the healthcare data;

generating one or more probabilistic constraints corresponding to one or more clinical attribute pairs associated with the healthcare data;

discarding one or more synthetic samples that violate the one or more probabilistic constraints;

treating a privacy budget as a finite computational resource and deducting one or more privacy units from the privacy budget for each statistical query executed against the healthcare data;

generating an append-only audit log comprising one or more hashed transaction records linking consumption of the privacy budget with one or more utility scores corresponding to the synthetic healthcare dataset; and

generating the synthetic healthcare dataset formatted according to the healthcare interoperability standard and generating a compliance certificate corresponding to the synthetic healthcare dataset.

17. The method as claimed in claim 16, wherein each audit trail entry of the append-only audit log comprises a cryptographic hash generated using contents of a preceding audit trail entry and contents of a current audit trail entry to form a cryptographically linked audit trail hash chain configured to provide tamper detection for the append-only audit log.

18. The method as claimed in claim 16, further comprising maintaining the append-only audit log using at least one of a permissioned blockchain ledger, a multi-party computation trust layer, a trusted execution environment trust layer, or a conflict-free replicated data type trust layer, and transmitting a cryptographic hash corresponding to each audit trail entry for independent integrity verification.

19. The method as claimed in claim 16, further comprising implementing a bilateral trust mode for deployments involving two or three participating

institutions, wherein, for deployments involving two participating institutions, privacy budget transactions require bilateral cryptographic signatures from the two participating institutions and tamper detection is performed through cross-verification of replicated audit trail hash chains, and wherein, for deployments involving three participating institutions, the privacy budget transactions utilize a two-of-three signature threshold consensus mechanism.

20. The method as claimed in claim 16, further comprising tagging the healthcare data with a jurisdiction identifier, preventing transmission of the healthcare data outside a jurisdiction of origin, and permitting transfer of the synthetic healthcare dataset across jurisdictions based on generation of the compliance certificate, wherein the method is performed within a network boundary corresponding to a healthcare data holder such that the healthcare data remains stored and processed within the network boundary throughout ingestion, transformation, privacy budget accounting, and generation of the synthetic healthcare dataset.

21. The method as claimed in claim 16, wherein the synthetic healthcare dataset is generated in at least one of JavaScript Object Notation format, extensible markup language format, or a format compatible with the healthcare interoperability standard.

Dated this **06th day of August 2026**

Signature

Manish Kumar
Patent Agent (IN/PA 5059)
Agent for the Applicant

A SYSTEM AND A METHOD FOR PRIVACY-PRESERVING SYNTHETIC HEALTHCARE DATA GENERATION

ABSTRACT

A system and a method for privacy-preserving synthetic healthcare data generation is disclosed. System includes processor and memory coupled to processor, wherein memory comprises instructions that, when executed by processor, cause processor to receive healthcare data corresponding to patients from healthcare data sources; retrieve healthcare data according to healthcare interoperability standard; parse healthcare resource graphs to identify privacy-sensitive attributes; transform healthcare data into synthetic healthcare dataset using calibrated noise based on privacy parameters; preserve clinical relationships; learn feature distributions; generate probabilistic constraints; discard synthetic samples violating probabilistic constraints; maintain records for data transformation operations; manage privacy budget consumption; generate append-only audit log linking privacy budget consumption with utility scores; maintain synthetic data provenance record; publish audit trail entries; generate notifications; determine applicable privacy regulations; generate predictive privacy budget allocation recommendations; tag healthcare data with jurisdiction identifier; maintain consent registry; and generate synthetic healthcare dataset and compliance certificate.

FIG. 3