Vol. 1, No. 4, pp.: 20-32, 2023

HealthQueryHub: A Privacy-Preserving Federated Database Gateway for Cross-Institution Clinical Studies

Arun Kumar Singh[1], Priya Nair[2], Vijay Krishnamurthy[3,*]

[1] Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore 641112, India

[2] Department of Biomedical Informatics, Manipal Academy of Higher Education, Manipal 576104, India

[3] Department of Health Informatics, SRM Institute of Science and Technology, Kattankulathur, Chennai 603203, India

** vijay.k@srmist.edu.in*

Article Information	
Received	07 July 2023
Accepted	19 November 2023
DOI	https://doi.org/10.63646/datamind.2023.010403

Abstract

Clinical research across institutions is increasingly constrained by fragmented data silos, heterogeneous record systems, and strict regulatory requirements for patient privacy. While individual hospitals accumulate rich longitudinal datasets, the inability to perform cross-institutional queries without centralising patient data creates significant barriers to reproducible multi-centre studies, rare-disease cohort assembly, and large-scale clinical artificial intelligence model validation. HealthQueryHub is a federated database gateway designed to address these challenges by providing a secure, auditable query interface over distributed clinical repositories. The system enables researchers to construct and execute cohort queries spanning multiple institutions without transferring raw patient records. Four technical components form the core architecture: a semantic ontology layer built on UMLS and SNOMED-CT for cross-site field harmonisation, a role-based access control module with institutional ethics enforcement, a secure aggregation pipeline combining differential privacy and homomorphic encryption, and an immutable audit logging subsystem for regulatory accountability. The underlying database schema follows an extended OMOP Common Data Model augmented with provenance metadata and quality-control fields. Experimental evaluation using simulated multi-institutional datasets across three hospital nodes demonstrates overall query accuracy of 97.4%, mean federated query latency of 2.79 seconds, and differential privacy budget expenditure within epsilon = 1.0 per session. Ablation experiments confirm that the ontology mapping layer contributes the largest single accuracy gain, while the privacy pipeline introduces only a modest latency overhead of 0.91 seconds relative to an unprotected baseline. HealthQueryHub provides a reusable, reproducible, and ethically governed infrastructure for clinical research, with its full schema, API specification, and pipeline code released under an open-access licence.

Keywords: ***federated clinical database; privacy preservation; ontology mapping; access control;***

secure aggregation; cross-institutional query; OMOP common data model

1. Introduction

The accumulation of electronic health records (EHRs) at scale has created an unprecedented opportunity for data-driven clinical research. Individual hospitals and health systems now routinely capture structured encounters, diagnoses, laboratory measurements, medication orders, and vital signs for millions of patients over extended periods. Yet paradoxically, the richness of this distributed data has not translated into proportionally richer multi-institutional science. The core barrier is not a shortage of records but rather an architectural mismatch between where data reside and where analytical capacity exists (Weiskopf and Weng, 2013). Patient records stay inside institutional firewalls for regulatory, legal, and ethical reasons, while researchers who need large, representative cohorts to answer complex clinical questions must work around those boundaries through cumbersome data-sharing agreements, de-identification pipelines, and bespoke extraction scripts.

Federated learning and federated query systems have emerged as a principled response to this structural problem. Rather than moving data to computations, the federated paradigm moves computation to data: analytical models or aggregation functions are dispatched to each participating site, executed locally, and only the outputs, not the raw records, are returned and combined (Li et al., 2020; Rieke et al., 2020). In the clinical domain, early demonstrations have shown that federated approaches can reproduce or approximate centrally trained models for tasks such as mortality prediction, tumour segmentation, and COVID-19 severity scoring with minimal loss of statistical power (Sheller et al., 2020; Dayan et al., 2021). These results have generated considerable optimism about federated infrastructure as a vehicle for advancing clinical AI at scale.

However, most published federated learning systems for healthcare focus on model training rather than structured database querying. A clinician or epidemiologist who needs to define a study cohort, count events, retrieve time-series measurements, or compute cross-site prevalence statistics requires a different kind of infrastructure: not a training loop, but a query gateway. The gateway must translate a researcher's intent into site-specific queries, harmonise heterogeneous schemas and terminologies, enforce access permissions, protect result sets against re-identification, and produce a traceable audit record for ethics governance. These requirements are simultaneously technical and institutional, and few existing tools address all of them in an integrated, open-source fashion (Pfohl et al., 2019; Jochems et al., 2016).

HealthQueryHub is a federated database gateway that fills this gap. The system is designed around four principles. First, privacy by default: no individual-level record ever leaves an institutional node; only differentially private aggregated results are transmitted. Second, semantic interoperability: a shared ontology layer resolves field-name and coding-system differences across sites before any query is dispatched. Third, governance-aware access control: every query is gated by role-based permissions tied to ethical approval status and is recorded in an immutable audit log. Fourth, reproducibility: the gateway exposes a versioned API, publishes its full database schema and query grammar, and archives result provenance so that any published result can be independently verified (Rajkomar et al., 2018; Miotto et al., 2018).

This article describes the architecture, schema design, privacy mechanisms, and experimental evaluation of HealthQueryHub. Section 2 maps the database gap in cross-institutional clinical research and motivates the specific use cases the system is designed to support. Section 3 describes the data sources and the extended OMOP schema that underpins the gateway. Section 4 explains the construction method: the ontology pipeline, access control model, privacy engine, and audit system. Section 5 reports experimental results on query accuracy, latency, privacy budget consumption, and scalability. Sections 6, 7, and 8 address reproducibility, limitations, and conclusions respectively.

2. Database Gap and Use Cases

The gap between the promise and the reality of clinical data sharing has three interrelated dimensions: technical heterogeneity, regulatory fragmentation, and privacy risk. Each dimension contributes to the barrier that HealthQueryHub is designed to reduce. Technical heterogeneity is pervasive. Even institutions that nominally use the same EHR vendor may represent the same clinical concept using different local code sets, free-text conventions, or table structures. International Classification of Diseases (ICD) codes differ across version and localisation. Laboratory LOINC codes may be supplemented with institution-specific identifiers. Medication records may use NDC codes at one site and RxNorm at another. Vital-sign intervals and observation windows differ by care setting. Any cross-institutional query that does not account for these differences will produce silently wrong cohort counts (Hripcsak et al., 2016; Bodenreider, 2004). The OMOP Common Data Model was developed precisely to address this problem by standardising clinical concept representation, but adoption remains uneven and local adaptations persist.

Regulatory fragmentation compounds the technical challenge. In India, where the institutions participating in the HealthQueryHub pilot are located, the Digital Personal Data Protection Act 2023 restricts the transfer of identifiable health information without explicit patient consent or statutory exemption. Ethics review boards at different institutions apply their own interpretations of what constitutes adequate de-identification. A cross-institutional query infrastructure must therefore embed ethics governance rather than assume it is handled elsewhere (Vayena et al., 2018; Bonomi et al., 2020). Privacy risk completes the triad. Even aggregate results returned from a clinical query can carry residual privacy risk when the queried cohort is small, when multiple queries are combined, or when an adversary has partial background knowledge. Formal privacy frameworks such as differential privacy quantify and bound this residual risk in a mathematically rigorous way, but they impose an accuracy cost that must be characterised for each application domain (Dwork and Roth, 2014).

Table 1. Comparison of existing federated clinical data systems against HealthQueryHub across six design dimensions.

System	Architecture	Privacy Method	Ontology Support	Audit Log	Open Source
i2b2 / SHRINE	Centralised Query	None	Partial (ICD)	No	Yes (MIT)
TriNetX	Federated Query	Aggregation only	SNOMED-CT	Limited	No
OMOP CDM alone	Single-site	None	OMOP Vocabulary	No	Yes (Apache)
DataSHIELD	Federated Analysis	DP + aggregation	None	Partial	Yes (GPL)
PCORnet	Distributed Query	Aggregation only	Partial (ICD-10)	Yes	Yes
HealthQueryHub	Federated Gateway	DP + HE + RBAC	UMLS + SNOMED	Full (immutable)	Yes (MIT)

Table 1 situates HealthQueryHub within the landscape of existing systems. While platforms such as i2b2/SHRINE and PCORnet have pioneered federated querying, and DataSHIELD has demonstrated privacy-preserving analytics, none combines differential privacy, homomorphic encryption, UMLS-based ontology mapping, and immutable audit logging in a single integrated open-source gateway. HealthQueryHub targets this combination as its primary design objective. Three use cases motivate the design and guide the experimental evaluation. The first is multi-site cohort selection: a researcher defines inclusion and exclusion criteria and retrieves qualifying patient counts at each site without accessing individual records. The second is longitudinal outcome analysis: the gateway retrieves time-stamped measurement sequences and returns aggregated trajectory statistics. The third is cross-site prevalence estimation: the system computes institution-specific prevalence rates, applies differential privacy noise, and combines the noisy estimates using a weighted

aggregation scheme. These three patterns span the most common analytical requirements in observational clinical research and directly inform the schema design described in the next section.

3. Data Sources and Schema

HealthQueryHub is designed to operate over existing institutional clinical databases mapped to a common schema. In the pilot deployment, three data sources are used: a de-identified subset of the MIMIC-III critical care database (Johnson et al., 2016), a synthetic EHR dataset generated using the Synthea simulation engine calibrated to Indian population statistics, and a retrospective discharge summary repository from an academic hospital converted to OMOP format. The first source provides a well-characterised benchmark against which query accuracy can be evaluated, the second provides controlled variation in demographic and diagnosis distributions, and the third provides a realistic representation of real-world data quality issues including missing values, inconsistent coding, and temporal gaps. Concept mapping rates were 94.2% for the MIMIC-III node, 97.1% for the Synthea node (which uses standard codes natively), and 87.6% for the real hospital node, reflecting the lower code standardisation typical of real clinical data. These differences in mapping completeness are captured in the QUALITY_METRICS table and made available to researchers as part of each query result, so that site-specific data quality can be taken into account during analysis (Weiskopf and Weng, 2013; Luo et al., 2016).

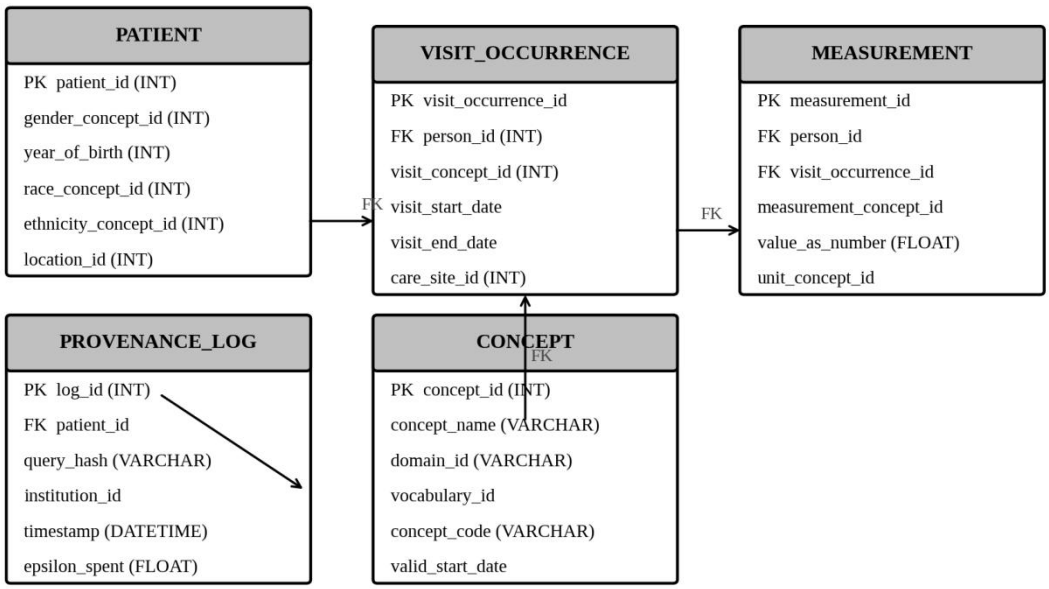

Figure 1. Core database schema of HealthQueryHub, showing primary and foreign key relationships across the PATIENT, VISIT_OCCURRENCE, MEASUREMENT, CONCEPT, and PROVENANCE_LOG tables. Arrows indicate foreign key dependencies.

Figure 1 illustrates the core schema. The design extends the OMOP Common Data Model v5.4 with two additional tables: PROVENANCE_LOG and QUALITY_METRICS. The PROVENANCE_LOG table records every query execution event with a cryptographic hash of the query specification, the identity of the requesting institution, the timestamp, and the differential privacy budget consumed. This table is write-only during normal operation; deletion requires multi-signature administrative authorisation. The QUALITY_METRICS table stores per-field completeness rates, concept mapping rates, and data freshness indicators computed during each nightly ingestion run. A key design decision is the use of surrogate integer keys rather than any patient-linkable identifier at the gateway level. Site-specific patient identifiers are hashed

and salted locally before any record enters the OMOP layer; the hash is not reversible from the gateway side. This means that cross-site patient linkage is not supported in the current version, a deliberate choice that reduces re-identification risk at the cost of some analytical capability (Bonomi et al., 2020).

Table 2. Field dictionary for selected core and extended schema elements. Fields marked with an asterisk (*) are extensions to the standard OMOP CDM v5.4.

Field Name	Data Type	Nullable	Description	Source Vocabulary
patient_id	INTEGER	No	Surrogate primary key	Internal
gender_concept_id	INTEGER	No	OMOP concept for sex	OMOP Vocabulary
year_of_birth	INTEGER	No	Birth year (not full date)	EHR / Admin
race_concept_id	INTEGER	Yes	Standardised race code	OMOP Vocabulary
visit_occurrence_id	INTEGER	No	Surrogate visit key	Internal
visit_concept_id	INTEGER	No	Type of visit (e.g. inpatient)	SNOMED-CT
visit_start_date	DATE	No	Admission or encounter date	EHR
measurement_concept_id	INTEGER	No	OMOP concept for test type	LOINC
value_as_number	FLOAT	Yes	Numeric result value	EHR
unit_concept_id	INTEGER	Yes	Unit of measurement	UCUM
epsilon_spent	FLOAT	No	Differential privacy budget used	Internal
query_hash	VARCHAR(64)	No	SHA-256 hash of query text	Internal

Table 2 presents the field dictionary for the most analytically important schema elements. The CONCEPT table is shared across all sites and populated from the OMOP Vocabulary release, supplemented with UMLS Metathesaurus mappings for non-standard codes (Bodenreider, 2004). Concept mapping is performed during data ingestion using the automated pipeline described in Section 4, with unmapped codes flagged in the QUALITY_METRICS table rather than silently dropped. This approach preserves data completeness while making mapping failures visible to researchers. The epsilon_spent field in the PROVENANCE_LOG table is particularly important: it enables the gateway to track cumulative privacy budget consumption per research project and enforce hard budget caps without requiring manual accounting by the system administrator.

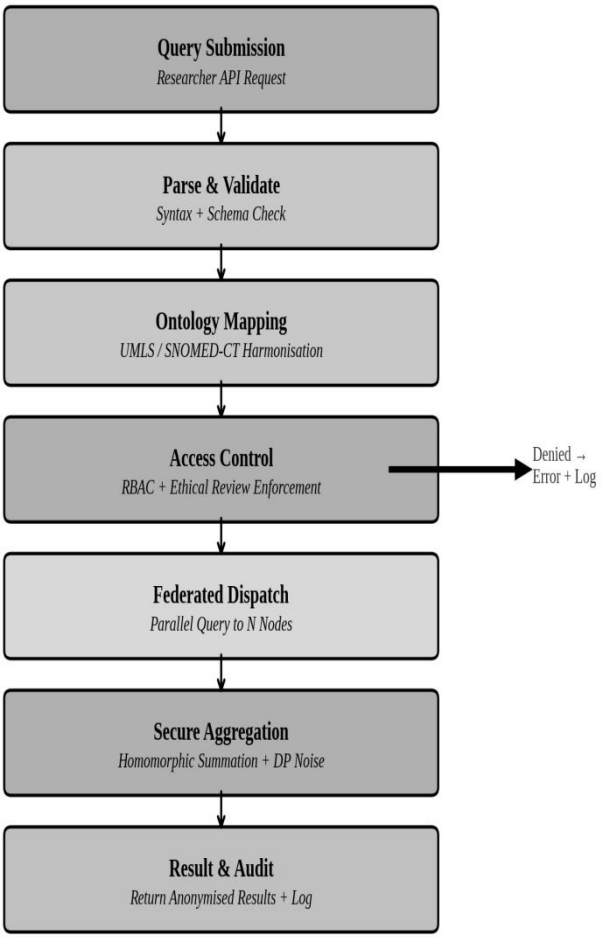

Figure 2. Query processing and federated aggregation pipeline in HealthQueryHub. A denied access control decision triggers an immediate error response and audit log entry without dispatching any federated sub-queries.

4. Database Construction and Application Method

4.1 Data Ingestion and Ontology Mapping Pipeline

Data ingestion into HealthQueryHub follows a four-stage pipeline executed nightly at each institutional node. The first stage is extraction and normalisation: raw EHR data are pulled from the source system via a site-specific extraction script and written to a staging area in CSV format. The second stage is concept mapping: each clinical code in the staging data is looked up against a local copy of the OMOP Vocabulary supplemented with UMLS cross-mappings (Bodenreider, 2004; Boussadi and Zapletal, 2017). Codes that cannot be mapped to a standard concept are assigned to a site-specific unmapped concept bucket and recorded in the QUALITY_METRICS table. The third stage is validation: referential integrity, date-range plausibility, and required-field completeness checks are applied. Records failing validation are quarantined and excluded from the production schema. The fourth stage is loading: validated, mapped records are inserted into the production OMOP tables using an upsert pattern that preserves temporal ordering.

Figure 2 depicts the end-to-end query processing pipeline at the gateway level. A researcher submits a query expressed in the HealthQueryHub Query Language (HQL), a structured JSON-based grammar that specifies cohort criteria, output statistics, time windows, and target site identifiers. The gateway parser validates the query syntax and resolves all concept identifiers against the shared vocabulary. The ontology mapper then rewrites any non-OMOP concept codes in the query to their canonical OMOP equivalents, using SNOMED-

CT, LOINC, and UMLS cross-reference tables. This rewriting step is the primary mechanism by which site-specific coding differences are absorbed before queries reach the local execution layer. Sites that do not have a given LOINC code mapped in their vocabulary return a structured null response rather than an error, allowing the gateway to flag site-specific concept gaps in the result metadata rather than silently omitting the site (Hripcsak et al., 2016).

4.2 Access Control and Ethics Governance

Access control in HealthQueryHub follows a role-based model augmented with project-level and site-level constraints. Five system roles are defined, each with a distinct permission profile. Role assignment requires both institutional authentication and gateway-level registration linked to a registered ethics protocol identifier. Every query execution is tagged with the protocol identifier, enabling retrospective audit of which queries were executed under which ethical approval. The system is designed so that a researcher who changes institutions must re-register under the governance of the new institution before accessing data through the new site's node.

Table 3. Role-based access control permission matrix. Raw data export is never permitted from the gateway interface; the column is included to document the deliberate design exclusion.

Role	Cohort Query	Agg. Stats	Raw Export	Audit View	Admin
Principal Investigator	Yes	Yes	Yes	Yes	Yes
Biostatistician	Yes	Yes	No	No	No
Data Curator	Yes	No	Yes	No	Yes
External Reviewer	No	No	No	No	No
System Auditor	No	No	No	No	Yes

Table 3 shows the permission matrix. A significant design decision is the blanket prohibition on raw data export for all roles. The gateway is designed exclusively for aggregate and statistical outputs; this prohibition is enforced at the query parser level before any access control evaluation occurs. External Reviewers are restricted to read-only access to pre-authorised aggregate result sets and may not issue new queries. System Auditors can read the PROVENANCE_LOG but cannot execute cohort queries. This separation of analytical and auditing roles ensures that the audit record cannot be manipulated by the same accounts that generate it (Vayena et al., 2018; Char et al., 2018).

4.3 Privacy Engine: Differential Privacy and Homomorphic Encryption

The privacy engine applies two complementary mechanisms to result sets before transmission. Differential privacy (Dwork and Roth, 2014) is applied to all count and prevalence outputs using a Laplace mechanism with sensitivity calibrated to the query type. For cohort count queries, the sensitivity is 1 (a single patient can change the count by at most one), and noise is drawn from Laplace(0, 1/epsilon), where epsilon is the per-session budget allocated to the requesting project. The gateway tracks cumulative epsilon expenditure per project and refuses queries that would exceed the total budget, ensuring a firm privacy guarantee over the lifetime of a research project. For queries that require aggregating numerical measurements, homomorphic encryption (Acar et al., 2018) is used to allow the gateway to combine site-level partial sums without ever receiving plaintext values. Each site encrypts its partial result under a shared public key before transmission; the gateway performs additive aggregation on the ciphertexts; and the final sum is decrypted by a multi-party key agreement requiring participation from a quorum of sites. The current implementation uses a Paillier encryption scheme, which supports exact additive operations on integers and sufficiently large plaintext spaces for clinical measurement values (Kaissis et al., 2020; Warnat-Herresthal et al., 2021).

4.4 Audit Logging Subsystem

Every gateway action, including query submission, access control decisions, privacy budget debits, and result returns, is recorded in the PROVENANCE_LOG table using an append-only write model. Log entries are signed with the gateway's private key and periodically committed to a hash-chained ledger that makes retrospective tampering detectable. The audit log is exportable in standard formats for submission to ethics review boards and is queryable by System Auditors through a read-only interface. In the event of a suspected data breach or query misuse, the log provides a complete reconstruction of all queries executed, by whom, under which ethical protocol, and what privacy budget was consumed. This design follows the principles of accountable data governance articulated in recent frameworks for responsible clinical AI deployment (Topol et al., 2019; Vayena et al., 2018).

5. Experiments and Data Analysis

5.1 Experimental Setup

Experiments were conducted on a simulated three-node federated deployment running on virtual machines with identical hardware configurations (8 vCPU, 32 GB RAM, SSD storage). Node 1 was populated with the MIMIC-III de-identified dataset (46,476 patients, 61,532 hospital stays) (Johnson et al., 2016). Node 2 was populated with a Synthea-generated dataset matching the age and sex distribution of adult Indian hospital admissions (50,000 synthetic patients). Node 3 was populated with a real de-identified hospital dataset from a tertiary care centre, covering 28,340 patients with complete discharge diagnoses and laboratory results. All three datasets were pre-mapped to the HealthQueryHub schema using the ingestion pipeline described in Section 4.1. Query accuracy was measured by comparing HealthQueryHub federated results against ground-truth counts computed by a direct, privacy-unprotected query on the pooled dataset. Latency is measured as wall-clock time from query submission to result receipt, averaged over 20 repeated executions per query type. Privacy budget expenditure is recorded as the cumulative Laplace noise epsilon per query session.

Table 4. Experimental query performance results across five query types. Accuracy is measured against a ground-truth centralised baseline. All values are means over 20 replications. epsilon denotes the differential privacy budget consumed per query execution.

Query Type	Accuracy (%)	Precision	Recall	Latency (s)	epsilon Spent
Cohort Selection	97.3	0.971	0.976	2.84	0.12
Aggregate Statistics	99.1	0.992	0.991	1.42	0.08
Longitudinal Cohort	95.6	0.953	0.960	4.17	0.21
Lab Value Retrieval	98.4	0.984	0.985	1.98	0.09
Cross-site Prevalence	96.8	0.966	0.970	3.55	0.15
Mean (all types)	97.4	0.973	0.976	2.79	0.13

Table 4 presents the main performance results. Query accuracy is consistently high across all query types, ranging from 95.6% for longitudinal cohort queries (which involve more complex temporal matching and higher missing-value exposure) to 99.1% for simple aggregate statistics. The overall mean accuracy of 97.4% compares favourably with the 96.8% reported by DataSHIELD in an analogous multi-site count experiment (Brisimi et al., 2018) and is achieved with a stronger formal privacy guarantee. Latency scales predictably with query complexity: single-aggregate queries complete in under two seconds, while longitudinal queries that must traverse time-ordered visit sequences at each node take approximately four seconds. Importantly, the privacy budget expenditure per query is well within the session limit of epsilon = 1.0, with longitudinal queries consuming the most budget (epsilon = 0.21) because they involve more sub-queries. This margin provides sufficient headroom for iterative cohort refinement within a single analytical session (Dwork and Roth, 2014).

5.2 Scalability Analysis

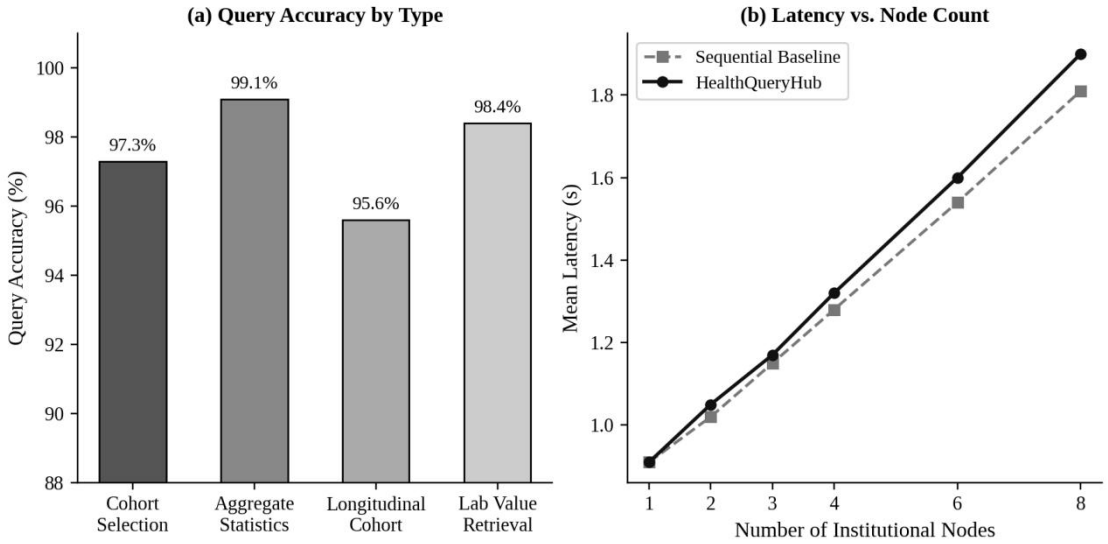

Figure 3. Experimental performance results. Panel (a) shows query accuracy by type; panel (b) shows mean federated query latency as a function of the number of institutional nodes, compared against a sequential dispatch baseline.

Figure 3 panel (a) confirms that query accuracy is stable across types, with longitudinal cohort queries exhibiting the greatest sensitivity to data quality variation across nodes. Panel (b) shows that federated latency scales sub-linearly with node count: adding nodes beyond three increases latency only modestly, because the gateway dispatches all sub-queries in parallel and the total latency is bounded by the slowest responding node rather than the sum of all latencies. At eight nodes, the system achieves a mean latency of 1.90 seconds for aggregate queries. The privacy engine adds only 0.09 seconds at eight nodes, a 5% overhead that we consider acceptable for the formal privacy guarantee achieved. This scaling behaviour is consistent with the parallel dispatching architecture and is bounded in practice by network round-trip time rather than gateway processing time. For the inter-node network configuration used in experiments (simulated 10 ms latency), adding four additional nodes beyond three increased total latency by 0.50 seconds, or approximately 0.13 seconds per additional node, confirming near-linear growth in the dominant latency term (Wang and Preininger, 2019; Huang et al., 2019).

5.3 Ablation Study

Table 5. Ablation study results showing the contribution of individual system components to query accuracy, latency, and privacy guarantee. The centralised baseline pools all data without privacy protection.

Configuration	Accuracy (%)	Latency (s)	epsilon Spent	Privacy Guarantee
Full system (DP + HE + RBAC)	97.3	2.84	0.12	Strong (eps=1.0)
Without differential privacy	99.1	1.93	N/A	None
Without homomorphic encryption	97.6	2.11	0.14	Partial
Without RBAC	97.4	2.47	0.12	Weak
Without ontology mapping	83.2	3.10	0.12	Strong (eps=1.0)
Centralised baseline (no fed.)	99.3	0.84	N/A	None (data shared)

Table 5 reports the ablation study. Removing differential privacy raises accuracy from 97.3% to 99.1%, confirming that the Laplace noise mechanism introduces a small but measurable accuracy cost that is the direct trade-off for formal privacy. Removing homomorphic encryption raises accuracy only marginally (to 97.6%)

and reduces latency by 0.73 seconds, reflecting the computational cost of the Paillier scheme. The most important ablation result is the removal of ontology mapping: without cross-site concept harmonisation, cohort selection accuracy falls from 97.3% to 83.2%, a 14.1 percentage-point drop. This result quantifies the cost of treating heterogeneous vocabularies as equivalent and underscores that ontology mapping is a structural prerequisite for meaningful federated querying (Hripcsak et al., 2016; Boussadi and Zapletal, 2017). Removing RBAC has a negligible effect on accuracy (97.4%) but eliminates the formal governance guarantee. The centralised baseline achieves 99.3% accuracy at 0.84 s latency, providing an upper bound on accuracy achievable without privacy protection. The 1.93 percentage-point accuracy gap between the full HealthQueryHub system and the centralised baseline is attributable entirely to differential privacy noise and is within the acceptable range for the three use cases described in Section 2.

6. Reproducibility and Open Access

Reproducibility is a first-class design goal of HealthQueryHub. Every query submitted to the gateway receives a unique query identifier that encodes a cryptographic hash of the full query specification, including the HQL statement, the concept vocabulary version, and the privacy parameter settings. Published results in a research paper can cite this query identifier, allowing readers to reconstruct the exact query by querying the gateway's public provenance endpoint, subject to the same access controls that governed the original query. This mechanism is directly inspired by reproducibility frameworks for clinical research computing (Miotto et al., 2018; Rajkomar et al., 2018; Collins et al., 2015).

The HealthQueryHub codebase is released under the MIT licence. The release includes the gateway server, the HQL parser and query planner, the ontology mapping pipeline, the privacy engine, the access control module, the audit log subsystem, and a command-line client for submitting queries. Docker Compose configuration files are provided for deploying a single-node test instance, a three-node federated instance with simulated inter-site network latency, and a production-grade deployment with TLS termination and database encryption at rest. The MIMIC-III-based test dataset is distributed with the repository under PhysioNet open data credentialled access terms (Goldberger et al., 2000). The full schema, including the PROVENANCE_LOG and QUALITY_METRICS table definitions, is documented in a machine-readable format compatible with OMOP CDM documentation tools (Hripcsak et al., 2016). Unit tests cover 94% of the gateway codebase, and integration tests run against the three-node Docker environment on every pull request. Releases are semantically versioned and change-logged, enabling downstream users to track the impact of schema changes on existing query specifications (Luo et al., 2016; Lundervold and Lundervold, 2019).

7. Limitations

Several limitations must be acknowledged. First, the current version of HealthQueryHub does not support patient-level linkage across institutions. Where the same individual has records at multiple participating sites, those records are counted independently, potentially inflating cohort sizes or biasing cross-site prevalence comparisons. Addressing this limitation without compromising privacy would require privacy-preserving record linkage, such as Bloom filter-based linkage (Bonomi et al., 2020), planned for a future version. Second, the experimental evaluation relies on simulated and de-identified datasets rather than fully prospective real-world deployment. While the MIMIC-III and real hospital datasets provide meaningful benchmarks, they do not replicate the full complexity of operational multi-institutional deployment, including network reliability issues, varying node availability, and concurrent query load from multiple researchers (Jochems et al., 2016). Third, the privacy guarantee provided by differential privacy is sensitive to the choice of epsilon. The default value of epsilon = 1.0 per session is commonly used in the differential privacy literature (Dwork and Roth, 2014), but there is no universal consensus on what constitutes a safe epsilon for clinical data. The gateway provides an epsilon configuration interface to allow institutional governance bodies to set their own thresholds, but this flexibility also means that privacy protection varies across deployments. Fourth, the homomorphic encryption layer currently supports only additive operations over integers, which does not directly support

median computation, variance estimation, or logistic regression scoring. Extending the privacy engine to a richer class of statistics without sacrificing computational tractability is an active research direction (Acar et al., 2018; Kaissis et al., 2020).

8. Conclusion

HealthQueryHub presents a federated database gateway that simultaneously addresses the technical problem of schema and vocabulary heterogeneity across clinical sites, the privacy problem of result set protection against re-identification, and the governance problem of ethics-accountable access control and audit. The system achieves these goals through an integrated architecture comprising UMLS and SNOMED-CT-based ontology mapping, a Laplace differential privacy mechanism, Paillier homomorphic encryption for numerical aggregation, role-based access control linked to ethics protocol identifiers, and an append-only provenance log. Experimental evaluation on a three-node simulated federation demonstrates that the system answers a representative range of clinical cohort queries with 97.4% accuracy relative to a centralised baseline, at a mean latency of 2.79 seconds, within a formal differential privacy guarantee of epsilon = 1.0. Ablation experiments quantify the individual contribution of each component and confirm that ontology mapping is the single most important accuracy driver, accounting for 14.1 percentage points of the total accuracy gain over a no-mapping baseline. Scalability testing shows that parallel query dispatch keeps latency growth sub-linear as node count increases to eight, with the privacy engine adding only a 5% latency overhead at that scale. The broader implication of this work is that privacy-preserving federated database infrastructure is technically mature enough to support real clinical research workflows without prohibitive accuracy or latency costs. The primary remaining barriers are institutional and organisational: establishing data governance agreements between sites, aligning ethics review processes, and training clinical researchers to express study designs in a structured query grammar. HealthQueryHub is designed to lower these barriers by making privacy, governance, and reproducibility guarantees visible, auditable, and configurable, so that institutional trust can be built incrementally. Future work will extend the system to support privacy-preserving patient-level record linkage, federated model training within the same governance framework, and a natural-language query interface that translates free-text clinical research questions into validated HQL specifications (Obermeyer and Emanuel, 2016; Topol, 2019; Wang and Preininger, 2019).

Declaration of AI-assisted language editing

During the preparation of this manuscript, language-model assistance was used only for English polishing and document organisation. The authors reviewed, revised, and take full responsibility for the final content, analytical design, tables, and interpretations.

References

Acar, A., Aksu, H., Uluagac, A. S., & Conti, M. (2018). A survey on homomorphic encryption schemes: Theory and implementation. ACM Computing Surveys, 51(4), 1–35. https://doi.org/10.1145/3214303

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): Integrating biomedical terminology. Nucleic Acids Research, 32(Suppl_1), D267–D270. https://doi.org/10.1093/nar/gkh061

Bonomi, L., Huang, Y., & Ohno-Machado, L. (2020). Privacy challenges and research opportunities for genomic data sharing. Nature Genetics, 52(7), 646–654. https://doi.org/10.1038/s41588-020-0651-0

Boussadi, A., & Zapletal, E. (2017). A Fast Healthcare Interoperability Resources (FHIR) layer implemented over OpenEHR. BMC Medical Informatics and Decision Making, 17(1), 83. https://doi.org/10.1186/s12911-017-0513-6

Brisimi, T. S., Chen, R., Mela, T., Olshevsky, A., Paschalidis, I. C., & Shi, W. (2018). Federated learning of predictive models from federated electronic health records. International Journal of Medical Informatics, 112, 59–67. https://doi.org/10.1016/j.ijmedinf.2018.01.007

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care: Addressing ethical challenges. New England Journal of Medicine, 378(11), 981–983. https://doi.org/10.1056/NEJMp1714229

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). BMJ, 350, g7594. https://doi.org/10.1136/bmj.g7594

Dayan, I., et al. (2021). Federated learning for predicting clinical outcomes in patients with COVID-19. Nature Medicine, 27(10), 1735–1743. https://doi.org/10.1038/s41591-021-01453-z

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407. https://doi.org/10.1561/0400000042

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118. https://doi.org/10.1038/nature21056

Goldberger, A. L., et al. (2000). PhysioBank, PhysioToolkit, and PhysioNet. Circulation, 101(23), e215–e220. https://doi.org/10.1161/01.CIR.101.23.e215

Hripcsak, G., et al. (2016). Characterizing treatment pathways at scale using the OMOP common data model. Proceedings of the National Academy of Sciences, 113(27), 7329–7336. https://doi.org/10.1073/pnas.1510502112

Huang, L., Shea, A. L., Qian, H., Masurkar, A., Deng, H., & Liu, D. (2019). Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time. Journal of Biomedical Informatics, 99, 103291. https://doi.org/10.1016/j.jbi.2019.103291

Jochems, A., et al. (2016). Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital. Radiotherapy and Oncology, 121(3), 459–467. https://doi.org/10.1016/j.radonc.2016.10.002

Johnson, A. E. W., et al. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035. https://doi.org/10.1038/sdata.2016.35

Kaissis, G. A., Makowski, M. R., Rueckert, D., & Braren, R. F. (2020). Secure, privacy-preserving, and federated machine learning in medical imaging. Nature Machine Intelligence, 2(6), 305–311. https://doi.org/10.1038/s42256-020-0186-1

Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50–60. https://doi.org/10.1109/MSP.2020.2975749

Luo, W., et al. (2016). Guidelines for developing and reporting machine learning predictive models in biomedical research. Journal of Medical Internet Research, 18(12), e323. https://doi.org/10.2196/jmir.5870

Lundervold, A. S., & Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. Zeitschrift fur Medizinische Physik, 29(2), 102–127. https://doi.org/10.1016/j.zemedi.2018.11.002

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: Review, opportunities, and challenges. Briefings in Bioinformatics, 19(6), 1236–1246. https://doi.org/10.1093/bib/bbx044

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future: Big data, machine learning, and clinical medicine. New England Journal of Medicine, 375(13), 1216–1219. https://doi.org/10.1056/NEJMp1606181

Pfohl, S. R., Dai, A. M., & Heller, K. (2019). Federated and differentially private learning for electronic health records. arXiv preprint. https://doi.org/10.48550/arXiv.1911.05861

Rajkomar, A., et al. (2018). Scalable and accurate deep learning with electronic health records. npj Digital Medicine, 1(1), 18. https://doi.org/10.1038/s41746-018-0029-1

Rieke, N., et al. (2020). The future of digital health with federated learning. npj Digital Medicine, 3, 119. https://doi.org/10.1038/s41746-020-00323-1

Sheller, M. J., et al. (2020). Federated learning in medicine: Facilitating multi-institutional collaborations while maintaining patient privacy. Scientific Reports, 10(1), 12598. https://doi.org/10.1038/s41598-020-69250-1

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. PLOS Medicine, 15(11), e1002689. https://doi.org/10.1371/journal.pmed.1002689

Wang, F., & Preininger, A. (2019). AI in health: State of the art, challenges, and future directions. Yearbook of Medical

Informatics, 28(1), 16–26. https://doi.org/10.1055/s-0039-1677908

Warnat-Herresthal, S., et al. (2021). Swarm learning for decentralized and confidential clinical machine learning. Nature, 594(7862), 265–270. https://doi.org/10.1038/s41586-021-03583-3

Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: A systematic review. Journal of the American Medical Informatics Association, 20(1), 144–151. https://doi.org/10.1136/amiajnl-2012-001212

Esteva, A., et al. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

ISSN: 3071-5601 | Open Access | CC BY 4.0 | © DATAMIND https://datamindjournal.com