The team has been tasked with designing and building a secure, scalable, and compliant database solution for a global private organisation conducting clinical trials.  Key considerations include high-level encryption to maintain data security, auditability, and flexibility to handle mixed data, allowing for the analysis of sensitive clinical and personal data across multiple regions while meeting strict regulatory requirements such as GDPR.  Data should be retrievable for analytic work and should meet regulatory requirements regarding standardised coding of information.

Real time data streaming would not be required, and therefore batch processing and cleaning is proposed, however frequent access to updated clean data by multiple users (high availability) on a role-based need is essential.  Key stakeholders for data base utilisation include the trial site personal, trial sponsors, statisticians, medical monitors, data managers, as well as regulatory bodies.

This report outlines the logical design, data management pipeline, database model, management system selection and hosting options, to provide a compliant, secure and user-friendly solution to meet client needs.

The data management process follows a structured pipeline to ensure accuracy and compliance.

1. Data capture process: Data will be entered via electronic data capture systems in a standardised structured manner, a non-exhaustive example of entities is shown in figure 1.   Additional structured data from wearable and electronic devices, and semi structured/unstructured data such as clinical notes and medical images will also be captured.
2. Validation: During data entry, mandatory fields must be completed. Drop-down menus will be used to ensure consistency for clinical terms. Numeric fields will include range checks (eg: age), while dates and identifiers will follow strict format rules. These measures will prevent erroneous entries and reduce cleaning workload.
3. Cleaning: must be undertaken in an auditable manner; all changes to data will be time-stamped and traceable.  Following data input, automated logic queries will identify impossible values. Units, values, and medical conditions will be standardised. Machine learning approaches would be adopted to spot pattern anomalies and flag these to appropriate users.  Missing values would not be imputed unless explicitly described in the study documentation (for regulatory reasons), however if pre-defined, then data will be imputed as appropriate based on last known value, data average or through machine learning methods (Weissler et al, 2021).
4. Storage: PostgreSQL will be used as detailed below.
5. Retrieval: Validated and cleaned data will be accessed through role-based permission for analysis, monitoring, or auditing. This ensures users only see data appropriate to their responsibilities. Given the sensitivity of clinical trial data, all data will be encrypted at rest (storage) and in transit. Audit trails will log all access, edits, and queries with time stamps, user IDs, and reasons for change

To support relational integrity between different entities, primary and foreign keys will be used. Each entity will have a unique identifier (primary key) to help distinguish between records, and foreign keys will be used to help link related entities, for instance, Trial Participant will be linked to Medical History by Participant ID.

Since this database will collect and store sensitive information, strong encryption, and adherence to GDPR or equivalent guidelines is critical.  Data will be stored in an anonymised manner wherever possible using deidentified patient ID codes and all tables will be normalised to Third Normal Form (3NF) to decrease redundancy thereby minimising duplication of sensitive information and limiting the potential for data conflicts or inconsistencies.

**Figure 1.  ERD highlighting example data items, their relationships, and data types utilised.**

The ERD shows the main entities:Trial Participant;  subject information (ID, demographic)

Clinical Trial, Trial Drug, Trial Centers;  trial-level details, site location, and trial drug details

Visits; participant trial visit details and tests

Medication;  stores medication name and standardised code (MedDRA)

Adverse Event; captures on trial safety data

Medical History ; stores participants pre-existing medical conditions

Medical Images; stores list and details of available medical images and links to external cloud storage site where image held.

Unique identifiers will be stored as BIGINT to accommodate the very large number of patients, visits, and trial events expected in global studies. TEXT is used for clinical notes (such as adverse event description), VARCHAR is applied where controlled string lengths are required, NUMERIC is used for dose amounts, ENUM for categorical variables such as gender, and DATE/TIMESTAMPTZ to ensure precise tracking of visits, adverse events, and administration of medications.  Medical Images are securely stored in the cloud; the database contains a table linking image metadata with the external URL.

A diagram of a medical organization

AI-generated content may be incorrect.

**DBMS Selection**

PostgreSQL is recommended as it offers reliability, flexible storage for mixed data, and allows for auditing. It provides strict relational storage for questionnaire answers and trial records alongside JSON support for device data, and can link to secure external storage (for example a managed cloud file store, similar to a secure image archive hospitals use) for unstructured data, such as medical images, which will be stored under strict permissions (Hong, 2024)

While not a core requirement, PostgreSQL can also support machine learning which will aid in future enhancements like predictive modelling and may be utilised to spot potential pattern anomalies as previously mentioned.

By contrast, Microsoft SQL server offers similar security and tooling but at higher licensing cost and tighter vendor dependency. NoSQL solutions provide speed and flexibility but lacks automatic enforcement of relationships and auditability, making it less suitable for regulated, clinical data.

Therefore, PostgreSQL represents the most feasible and compliant choice.

**Hosting Solutions**

Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP) all offer managed PostgreSQL services, encryption, and scalability.

AWS offers Amazon RDS for PostgreSQL, with automated backups, high availability through multi-region deployment, and compliance with GDPR (Introduction to amazon web services, 2025). For unstructured data such as medical images, AWS S3 provides secure, scalable storage.

Azure provides similar functionality with strong Microsoft ecosystem integration, where Azure Blob Storage can be used to store unstructured data.  However, there may be increased dependency on Microsoft, and this solution was considered more costly. GCP provides comparable compliance and strong analytical tools, though it is less widely adopted within the clinical research sector.

Given the client’s need for global access, regulatory compliance, sector adoption, support for mixed data, and bearing in mind feasibility, AWS is the recommended hosting solution.

**Conclusion**:

The proposed design balances flexibility, security, scalability, and compliance to meet the client’s needs. Implementing this system as a normalised, relational model in PostgreSQL, hosted on AWS with role-based access will support global clinical trials whilst safeguarding patient data. Having a structured data pipeline, strong encryption, and high auditability will ensure regulatory compliance and data integrity.

**~~Possible database build proposal outline (some of this could be included in an appendix as the brief allows for that given the word count is 1000).~~**

~~~~ **~~Project scope and objectives~~** ~~- Define the system purpose (authoritative clinical-trial store for enrolment, visits, labs, dosing, device data and imaging), success criteria (data integrity, auditability, secure access) and out-of-scope items.~~

~~~~ **~~Key stakeholders~~** ~~- Sponsor, trial managers, data managers, site staff, clinicians, IT/security, and the analytics team; identify single points of contact for decisions and sign-off.~~

~~~~ **~~Needs analysis~~** ~~- Capture functional needs (CRF ingestion, coding, query management, audit trail, reporting), non-functional needs (security, availability, scalability) and data sources (EDC, labs, wearables, imaging).~~

~~~~ **~~Functional specification~~** ~~- List required capabilities: enrolment/visit/dosing/lab/event tables, questionnaire/device ingestion, image metadata and retrieval, role-based views, audit log and de-identification controls.~~

~~~~ **~~Technical specification~~** ~~- DBMS choice (Managed PostgreSQL), object storage for large files, JSONB for flexible payloads, RLS for scoped access, audit logging and an ETL staging pipeline.~~

~~~~ **~~Data governance and compliance~~** ~~- Consent tracking, pseudonymisation, retention policy, SAR workflows and vocabulary/version control for MedDRA/LOINC.~~

~~~~ **~~Project plan and milestones~~** ~~- Define deliverables and decision gates (requirements sign-off, schema design, ingestion prototype, validation and user acceptance, handover); specify responsibilities and approval owners - timelines to be agreed with the sponsor.~~

~~~~ **~~Budget and resourcing~~** ~~- Identify roles required (DBA, data engineer, security lead, data manager, project lead), expected licensing and cloud hosting categories and contingency for regulatory validation.~~

~~~~ **~~Deliverables~~** ~~- ERD, DDL, data dictionary, JSON schema files, sample ingest scripts, test dataset, validation report and an operations guide.~~

Potential References

<https://www.postgresql.org/about/> (PostgreSQL website)

<https://nebius.com/blog/posts/postgresql-in-context-of-ml> (overview of PostgreSQL, machine learning capabilities and challenges)

Weissler, E.  2021 <https://pmc.ncbi.nlm.nih.gov/articles/PMC8365941/>  /machine learning in clinical trials