**Introduction**

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 analysis of sensitive clinical and personal data across regions while meeting GDPR requirements (European Union, 2016).  Data should be retrievable for analytics and meet coding standards.

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 with role-based access is essential (Sarkar & Roychowdhury, 2019).  Key stakeholders for data base utilisation include the trial site personnel, trial sponsors, statisticians, medical monitors, data managers, as well as regulatory bodies (EMA, 2023).

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 2. 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. To guarantee uniformity for clinical terms, drop-down menus will be utilized, dates and identifiers must follow format guidelines, while numeric fields will incorporate range checks (e.g. age). These measures reduce errors and cleaning workload. Rigid validation guidelines, however, can occasionally limit site staff usability, contributing to a trade-off between operational effectiveness and data quality.
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, data will be imputed using last known value, data average or machine learning methods (Weissler et al, 2021). While these approaches improve accuracy, careful governance controls are required as audit trails and anomaly detection can increase system overhead and reviewer workload.
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 access, edits, and queries with timestamps, user IDs, and reasons for change.

A close-up of a sign

AI-generated content may be incorrect.These sequential stages are illustrated in Figure 1, which shows the proposed data management pipeline for the clinical trial database.

Figure 1 Data Management Pipeline – Clinical Trial Database   
(Developed by team based on course materials from Williams, 2025; regulatory compliance elements referenced from European Union, 2016; EMA, 2023)

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.

**ERD – Clinical Trial Database**

Figure 2 illustrates the core entities and relationships within the database: Participants, Clinical Trials, Visits, Medications, Medical History, and Medical Images. This model appropriately displays relational integrity through primary and foreign keys, uses scalable datatypes such as BIGINT for identifiers and TIMESTAMPTZ for time-sensitive records, thus meeting the regulatory requirements for precision and auditability (Sarkar & Roychowdhury, 2019; EMA, 2023).

A diagram of a medical organization

AI-generated content may be incorrect.

Figure 2 Entity–Relationship Diagram (ERD) – Clinical Trial Database   
(Developed by the team; compliance considerations referenced from Sarkar & Roychowdhury, 2019; EMA, 2023).

**DBMS Selection**

PostgreSQL is recommended for reliable, flexible storage of mixed data and auditing features, providing strict relational storage for questionnaire answers, trial records, JSON support for device data, and secure linkage to external storage (e.g. managed cloud file store) for unstructured files, such as medical images (Sarkar & Roychowdhury, 2019). Despite not being a requirement, PostgreSQL can also support machine learning, which could further enhance predictive modelling and be trained in anomaly detection, as previously mentioned. Nonetheless, PostgreSQL may require advanced tuning to handle very large volumes of unstructured data, and JSONB queries can be less efficient than native NoSQL solutions.

Microsoft SQL Server does offer similar security and tooling at higher cost and tighter vendor dependency, with NoSQL providing speed and flexibility, but lacking automatic enforcement of relationships and auditability (Sethi & Panda, 2024).

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 (AWS, 2025; EMA, 2023). 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 (Microsoft, 2025).  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 (Google Cloud, 2025).

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

**Conclusion**

The proposed design balances flexibility, security, scalability, and compliance to meet the client’s needs. Global clinical trials will be facilitated whereas patient data is safeguarded through implementing a normalized system with a relational model in PostgreSQL, hosted on AWS with role-based access. Having a structured data pipeline, strong encryption, and high auditability will ensure regulatory compliance and data integrity. Nevertheless, performance tuning for unstructured data and vendor dependency on AWS remain challenges, demonstrating that strict governance and hybrid strategies are required for mitigation purposes.

**References**

European Union (2016) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Official Journal of the European Union, L119, 1–88. Available at: [https://eur-lex.europa.eu/eli/reg/2016/679/oj](https://eur-lex.europa.eu/eli/reg/2016/679/oj%20) (Accessed: 01 September 2025).

Sarkar, T. & Roychowdhury, S. (2019) Data wrangling with Python : creating actionable data from raw sources. 1st edition. Birmingham ; Packt Publishing Ltd.

European Medicines Agency (2023). *Guideline on computerised systems and electronic data in clinical trials (EMA/INS/GCP/112288/2023).* EMA, Amsterdam. Available at: <https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/guideline-computerised-systems-and-electronic-data-clinical-trials_en.pdf>

Weissler, E. H., Naumann, T., Andersson, T., Ranganath, R., Elemento, O., Luo, Y., Freitag, D. F., Benoit, J., Hughes, M. C., Khan, F., Slater, P., Shameer, K., Roe, M., Hutchison, E., Kollins, S. H., Broedl, U., Meng, Z., Wong, J. L., Curtis, L., Huang, E., … Ghassemi, M. (2021). The role of machine learning in clinical research: transforming the future of evidence generation. Trials, 22(1), 537. <https://doi.org/10.1186/s13063-021-05489-x>

Williams, G. (2025) Deciphering Big Data [module materials]. University of Essex Online, July–October.

Microsoft (2025) Azure Database for PostgreSQL – Documentation. Available at: <https://learn.microsoft.com/en-us/azure/postgresql/> (Accessed: 01 September 2025).

Google Cloud (2025) Cloud SQL for PostgreSQL. Available at: [https://cloud.google.com/sql/postgresql](https://cloud.google.com/sql/postgresql%20) (Accessed: 01 September 2025).

Sethi, S. and Panda, S. (2024) SQL or NoSQL—Practical Aspect and Rational behind Choosing Data Stores. *Journal of Computer and Communications*, **12**, 1-20. doi: [10.4236/jcc.2024.128001](https://doi.org/10.4236/jcc.2024.128001).

Amazon Web Services (2025). Amazon RDS for PostgreSQL – Features. Available at: <https://aws.amazon.com/rds/postgresql/> (Accessed: 01 September 2025)

**Group Members:**

* + **Andreea Nicoara** (Student ID 12699266)
  + **Dean Carey** (Student ID 12698710)
  + **Monique Mizz**
  + **Sonya Jackson** (Student ID 12698873)