**Introduction**

The design of a scalable and secure database system to support a global clinical trial organization is summarized and evaluated in this report. The core of the solution was PostgreSQL, hosted by Amazon Web Services (AWS) to facilitate data integrity, audit trails, and adherence to European Medicines Agency (EMA, 2023) and General Data Protection Regulation (GDPR, 2016) standards. The database services are both semi-structured as well as structured data with a standardized pipeline to validation, cleaning, and encrypted storage. This report assesses how well they meet technical and regulatory requirements and offers suggestions for improving performance, auditability, and future scalability.

A close-up of a sign

AI-generated content may be incorrect. The Data Management Pipeline (Figure 1), which specifies the data capture workflow, validation, cleaning, storage, and retrieval, displays the main design elements.

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)

**Summary of Work Conducted**

The project aimed to develop a scalable, secure, and compliant database for managing global clinical trial data. Reliable capture, traceability, and regulatory compliance were to be attained without compromising usability or performance. Because of its flexibility, open environment, and capacity to manage both relational and semi-structured data, PostgreSQL was used to run the system. To minimize redundancy and maintain referential integrity between important entities like participants, clinical trials, visits, and adverse events, the tables were normalized to Third Normal Form (3NF). Although full normalization increases consistency, it may result in additional joints and slightly slower queries; therefore, indexing and query optimization are required to maintain speed.

The logical architecture that underpins these procedures is shown in the Entity–Relationship Diagram (Figure 2). The relationship between entities is defined by primary and foreign keys, and datatypes like BIGINT and TIMESTAMPTZ provide both precision and scalability. The need to maintain balanced relational design is validated by the fact that, while PostgreSQL can handle semi-structured content using JSONB, excessive use of this can bloat indexing and impair performance.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).

A predetermined pipeline of capture, validation, cleaning, storage, and retrieval was used for data management. Standardized formats were used for data entering the database from clinical notes, linked devices, and electronic data capture systems. To guarantee consistency, validation employed range checking, controlled vocabularies, and mandatory fields; however, excessively stringent validation may make the site less usable for site staff. Cleaning was done in an audit-worthy manner, labeling each modification with a user ID and time-stamped. Although success depends on well-trained models that could result in false positives, machine learning-based anomaly detection was proposed as a means to identify suspicious values (Weissler et al., 2021).

The general pipeline, which links the technical flow to regulatory control, is depicted in Figure 1 (GDPR, 2016; EMA, 2023). Throughout, auditability and encryption were incorporated. In accordance with Article 32 of the GDPR on data security, everything is encrypted while it is in transit and at rest. Strong encryption, however, reflects the trade-off between performance and compliance, adding computational overhead and potentially increasing storage costs.

Secure references were used to integrate PostgreSQL with AWS S3 cloud storage for unstructured data, such as medical images. To ensure data availability and resilience, Amazon RDS for PostgreSQL offers multi-region replication and automated backups (Amazon Web Services, 2025). Periodic vendor-risk assessments are crucial because, although this strategy promotes worldwide accessibility, it also adds reliance on AWS infrastructure and the related expenses.   
Overall, the implemented solution demonstrates how both technological and regulatory considerations were integrated into a single architecture. Strong data governance and good compliance standard alignment are guaranteed by the solution design, which also appropriately acknowledges real-world limitations like vendor dependency, performance tuning, and usability constraints. These factors serve as the basis for the assessment in the section that follows.

**Critical Evaluation of Concepts**

Database design struck a balance between system scalability and efficiency and regulatory compliance. Because of its open-source nature, which offered high levels of flexibility, rich relational modeling support, and the ability to store semi-structured data in JSONB, PostgreSQL served as the main database management system (DBMS). While PostgreSQL and Microsoft SQL Server support comparable auditability and integrity controls, PostgreSQL has the advantage of being less expensive and having no licensing requirements. Despite providing enterprise-level support and deeper integration within the Microsoft environment, SQL Server has higher vendor lock-in and maintenance costs (Microsoft, 2025).   
Although NoSQL databases, such as MongoDB, can handle unstructured data more quickly, they lack strict relation enforcement and audit trail functionality, which makes them less suitable for use in regulated clinical settings (Sethi & Panda, 2024).

Additionally, PostgreSQL offers dependability through granular permission systems, role-based access control, and ACID compliance. It needs to be carefully adjusted, though, to effectively handle very big data sets. High transaction rates that may call for partitioning or indexing schemes, as well as extensive use of JSONB fields, can negatively affect query performance. To improve reading performance with high time-series data, like streaming patient monitoring data, PostgreSQL table partitioning with declarative syntax and BRIN indexes can be used in future instances. Notwithstanding these limitations, it is adaptable enough to be used in clinical research, where it is necessary to securely and consistently link unstructured notes and structured patient data (Sarkar & Roychowdhury, 2019).

The chosen architecture makes use of S3 object storage and AWS RDS hosting with PostgreSQL. The GDPR's preference to guarantee data protection as a design-default is in line with the use of encryption, automated backup, and multi-region replication, all of which are provided by AWS (European Union, 2016). Reliance on a single cloud provider, however, raises governance concerns and may result in future cost increases. The former risk could be mitigated with scalability and disaster recovery capability maintained through a hybrid or multi-cloud approach (EMA, 2023).

The compliance was seen as a key design principle, not something to think about later. GDPR Article 32 stresses the importance of keeping data safe while it is being processed by using encryption and access controls that are built into the system. The EMA (2023) guideline also talks about how important it is for electronic data capture to be able to be audited, versioned, and traced. The system, on the other hand, has full audit trails, user identification, and logging of timestamps. It makes it easier to keep track of these things, but it uses up storage space faster and needs to be checked more often to keep the audit logs from getting too big. Pseudonymization and anonymization of participant identity maintain identity while complicating data linkage across studies during dataset merging.

Weissler et al. (2021) say that the data management pipeline was made to make sure that data quality is high by using standardization and machine learning to check it. Its structured flow from capture to retrieval helps keep trials readable and consistent. But in sensitive research situations, automation alone can't take the place of human review. False anomaly flags can cause unnecessary revalidation cycles, so it's better to use hybrid workflows that mix automated and manual review.

The system demonstrates a proactive application of the "privacy by design" concept in terms of governance. Each party can only view or alter information that is relevant to them thanks to role-based access controls, and the risk of exposure is reduced by encryption both in transit and at rest. However, such controls should be complemented with effective authentication policies as well as frequent key rotation to be effective.

The team had looked into on-premises hosting and hybrid models of the NoSQL type to weigh various approaches. Since the costs of keeping regulatory certification are high and out of proportion to international trials, they had to be disqualified. Thus, PostgreSQL on AWS produces a balanced result in terms of cost, performance, and compliance. Data lake architectures could be introduced in the future to handle the growing variety of data, particularly from wearable and imaging modalities.

**Legal and Regulatory Compliance**

The regulatory compliance remained a foremost factor during the entire database design procedure. The system was developed in accordance with the European Medicines Agency's (EMA, 2023) guidelines for computerized systems in clinical trials and the General Data Protection Regulation (GDPR, 2016). The guidelines set forth principles that directly affect the system's structure and controls, including data integrity, auditability, accountability, and privacy by design.

GDPR Articles 5 and 25 encourage privacy and data minimization by design. Role-based permission restrictions, retaining only necessary identifiers, and using pseudonymization when direct identifiers could not be avoided were ways to exercise these values. While version control and detailed audit trails provide accounting under Article 30, encryption, both in transit and at rest, satisfies Article 32 on processing security. In addition to improving compliance through audit trails and encryption, both require regular performance monitoring and add computational overhead.

These requirements are further expanded in the EMA guideline of 2023 to cover audit trails, traceable data management, and system validation. The PostgreSQL database records each transaction with time stamps and user identity to be fully traceable. The database follows specified procedures for data entry, cleaning, and review to be validation compliant. In accordance with Good Clinical Practice (GCP), all modifications are recorded to prevent unauthorized alteration or deletion. However, such thorough auditing creates the potential for the log to grow quickly, adding to the maintenance load and potentially increasing the cost of cloud storage.

With its certified centers that can support GDPR readiness, ISO 27001 compliance, and the possibility of HIPAA compliance, AWS hosting makes compliance easier (Amazon Web Services, 2025). However, since the data controller bears legal responsibility for compliance, configuration mistakes or inadequate internal governance may still result in breaches. To guarantee ongoing compliance, regular checks of encryption keys, access guidelines, and planned automatic backups are still necessary.

The foundation of clinical research integrity is data ethics, in addition to form compliance. Working with bias in algorithms, data provenance transparency, and equitable data sharing are all covered by ethical governance. Maintaining explainability becomes essential to preserving trial participant and regulatory trust as automation and machine learning make working with anomaly detection easier. By combining technical and ethical audits, accountability could be consolidated to make sure that compliance is in line with moral obligations rather than just regulatory ones.

Overall, the database design exhibits both procedural and technical compliance. The database acknowledges that maintaining legal compliance is a continuous process rather than a one-time occurrence and incorporates the security, validation, and traceability features necessary for regulated clinical research.

**Conclusions and Recommendations**

With regulatory standards guiding design from the start, the developed database solution demonstrates that compliance, scalability, and data integrity can all be achieved under a single architecture. With its strong auditability, encryption, and organized data management, the PostgreSQL setup on AWS satisfies EMA and GDPR regulations while maintaining operational agility. A functional model for multi-site clinical research is produced by combining relational storage with cloud scalability.

However, some restrictions still exist. Encryption and audit trails augment computational overhead, with the potential to impair performance with high volumes of transactions. Although extensible, PostgreSQL's JSONB and unstructured data management becomes resource-intensive in the absence of any indexing and partitioning techniques. Reliance on AWS services results in vendor lock-in and the potential for cost creep, which means that contingency planning and regular vendor re-evaluation are necessary.

Future improvements ought to focus on automating monitoring and quality control tasks. Predictive analytics could help with proactive data validation, anomaly detection packages could reduce manual labor for data cleaning, and hybrid storage architectures, integrating PostgreSQL with a lightweight NoSQL or data lake platform, could provide wearable, sensor, and image data with more scalability.

Long-term scientific value would also be added by future versions of the system that adhered to the FAIR data principles: Findable, Accessible, Interoperable, and Reusable. In order for authorized researchers to locate, analyze, and reuse clinical data from one study among others, FAIR compliance promotes open metadata practices and machine-readable formats. Incorporating FAIR principles would improve interoperability and situate the database in the larger framework of open, cooperative biomedical research.

In short, the design achieves its goals of safe, legal, and effective data management to support international clinical trials. Maintaining long-term usability and regulatory compliance will be fueled by ongoing governance, performance optimization, and storage diversification.

**References**

Amazon Web Services (2025) *Amazon RDS for PostgreSQL – Features.* Available at: <https://aws.amazon.com/rds/postgresql/> (Accessed: 13 October 2025)

European Medicines Agency (2023). Guideline on computerized 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>

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> (Accessed: 13 October 2025)

Microsoft (2025) *Azure SQL Database – Documentation.* Available at: <https://learn.microsoft.com/en-us/azure/azure-sql/> (Accessed: 13 October 2025)

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

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://www.scirp.org/journal/paperinformation?paperid=135134).

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.