CDC Group Healthcare Database

Healthcare databases manage critical patient records, appointments, billing, and insurance data securely. According to Kline: “Databases are one of the most important components of a high-functioning healthcare organization” (Kevin Kline, 2024).
The CDC Group logical database is designed to manage the clinical and administrative data of patients and invoices by integrating information from electronic medical records (EMR), picture archiving, laboratory information and communication systems, together with insurance databases, thus improving efficiency, data integration, and traceability. Microsoft SQL Server was chosen as the preferred DBMS which offers a tight integration with the Microsoft ecosystem, strong governance and built-in HIPAA compliance frameworks. An XTIVIA article supports this by stating, SQL Server integrates seamlessly with other Microsoft products, such as SharePoint and PowerBI, making it easier to work with data across platforms. (XTIVIA, 2023)

Entity-Relationship (ER)

The entity-relationship diagram visually shows the relationship between different items in a database. IBM agrees by stating, “ERDs are a high-level conceptual data model that sets the stage for more advanced database design and analysis”. (Belcic and Stryker, 2024)

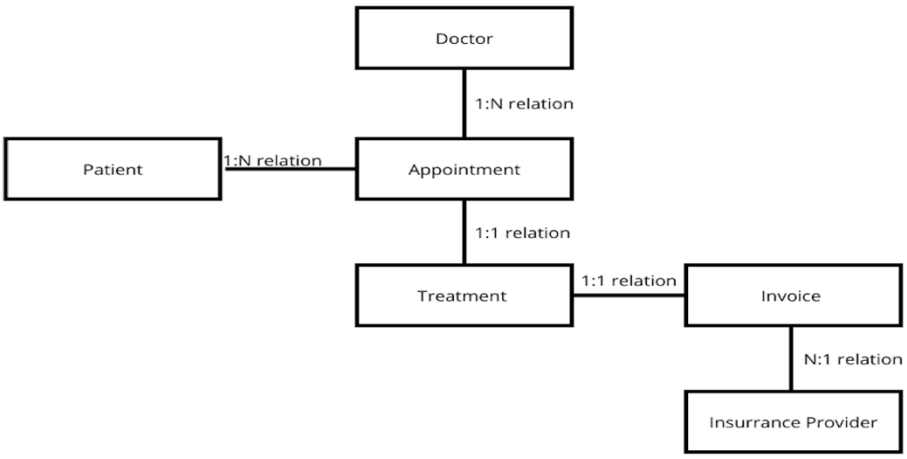
Core entities include:

- Patient: personal data and contacts.
- Doctor: personal information and their specialty.
- Specialty: medical areas of competence.
- Appointment: patient visits with the date and time.
- Treatment: clinical intervention and medical notes.
- Invoice: treatment invoice with the amount, payment type and linked insurance.
- Insurance: affiliated providers.

Each entity has a unique identifier that allows the data to be connected among the different tables. Considering the type of data in our database, the principal data types are integers because they are used as primary keys and references among the tables. In addition, there will be variable character data for textual information, date (DD-MM-YYYY), or date and time (DD-MM-YYYY HH: MM), mostly in the client and appointment entities, and decimal values for the payment amounts. It is important to underline that the primary key guarantees the records ' uniqueness, and the foreign key will be used to connect one table to another.

Relationships:

Patient-Appointment: a patient can have multiple appointments; each appointment is tied to one patient (1:N)
Doctor-Appointment: a doctor can produce multiple appointments, but each appointment is associated with one doctor (1:N)
Appointment-treatment: one appointment produces one treatment (1:1)
Treatment-invoice: one treatment generates one invoice (1:1)
Invoice-Insurance Provider: multiple invoices can refer to the same insurance provider, each invoice to one provider (N:1)



The logical model ensures consistent structuring of records imported from different systems. It adheres to ICD-10, LOINC, and DICOM standards, as recommended by Liu et al. (2023). In addition, to guarantee data quality and interoperability, the logical model employs two level standardization(openEHR and ISO13606) and ontologies (OWL and RLF). Those would support the inferences and semantic rules and address fragmented data and heterogeneous systems (De Mello et al., 2022).

CDC Group Healthcare System Data Management Pipeline Data

Data Capturing and Sources.
The implementation of the CDC Group healthcare database will be used to manage 2.2 million direct patient interactions per year in 54 medical centers and 60 blood collection points per year. The data management pipeline will use various data sources. The data sources are used to characterize medical systems of considerable size(Sujansky, 2001).

Primary Data Sources.
Demographic patient information will be drawn from the registration systems of each of the health care centers and blood collection points, from electronic intake forms at receptionist desks. Laboratory data is classified as structured data(test codes, number values, reference ranges and technician annotations). The data is achieved through processing 3.7 million tests per year generated via Laboratory Information Systems (LIS).
Secondary and External Sources.
The system is connected to insurance provider databases which contain information on approximately 8,000 private companies. This allows us to check whether data is complete and for data related to claims processing. External laboratory reference databases confirm test methodologies and quality control standards. (Naugler & Church, 2019).

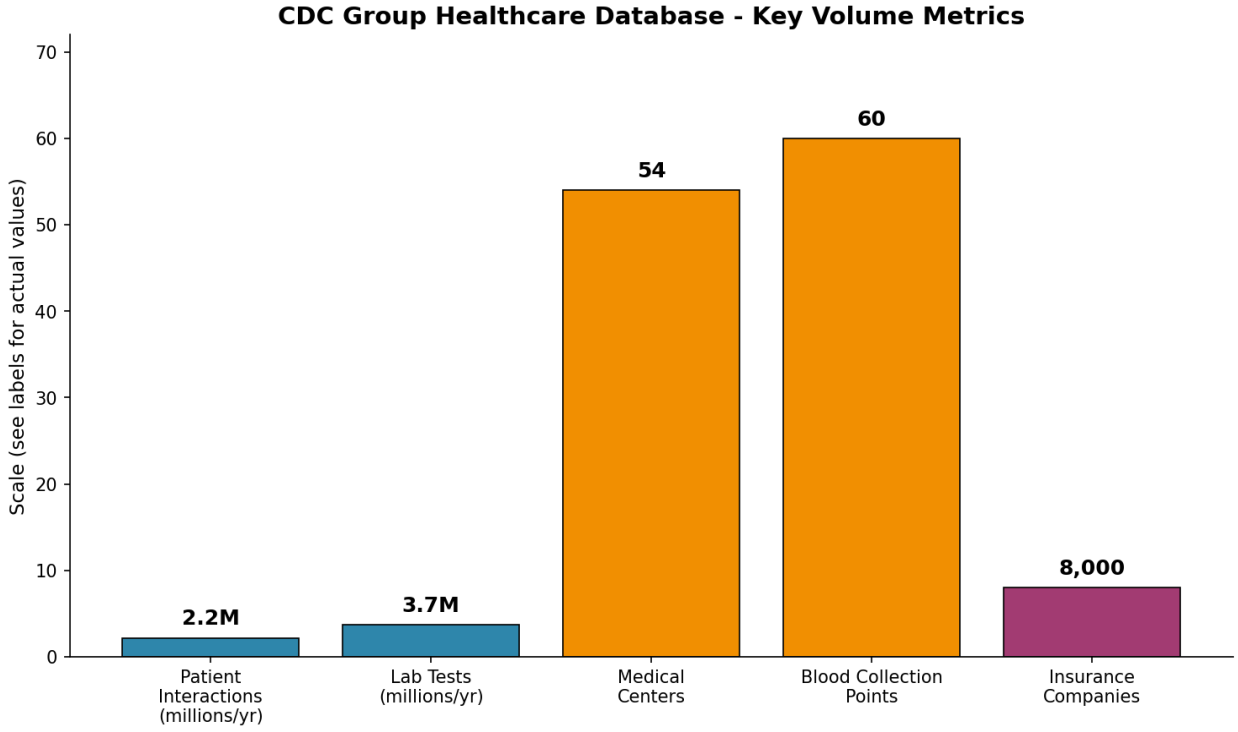
Data Cleaning Process
Considering CDC Group's scale of operations, automated quality management of data would be necessary to ensure data accuracy of millions of transactions yearly (Rahm & Do, 2000).

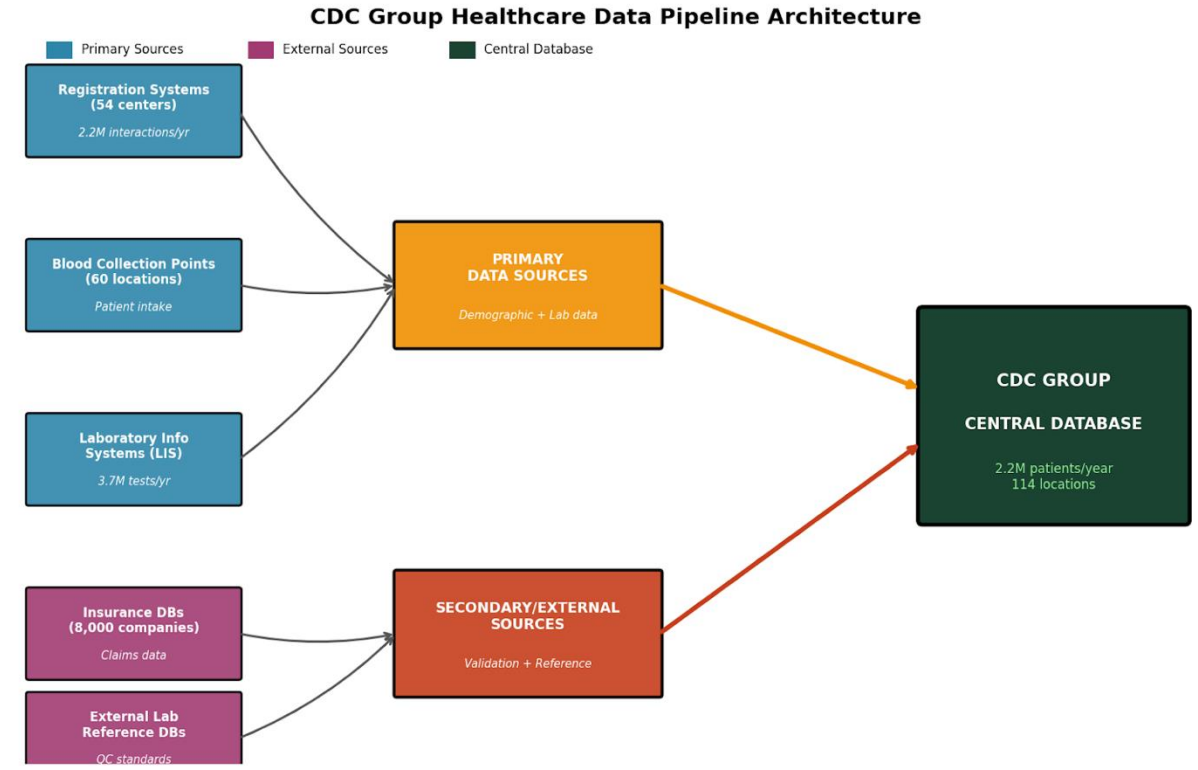
Stage 1: Real-time Validation. We used Luhn algorithms; patient identifiers are check-digitized, which avoids 15–20% of transcription errors that arise manually (Kahn et al.,

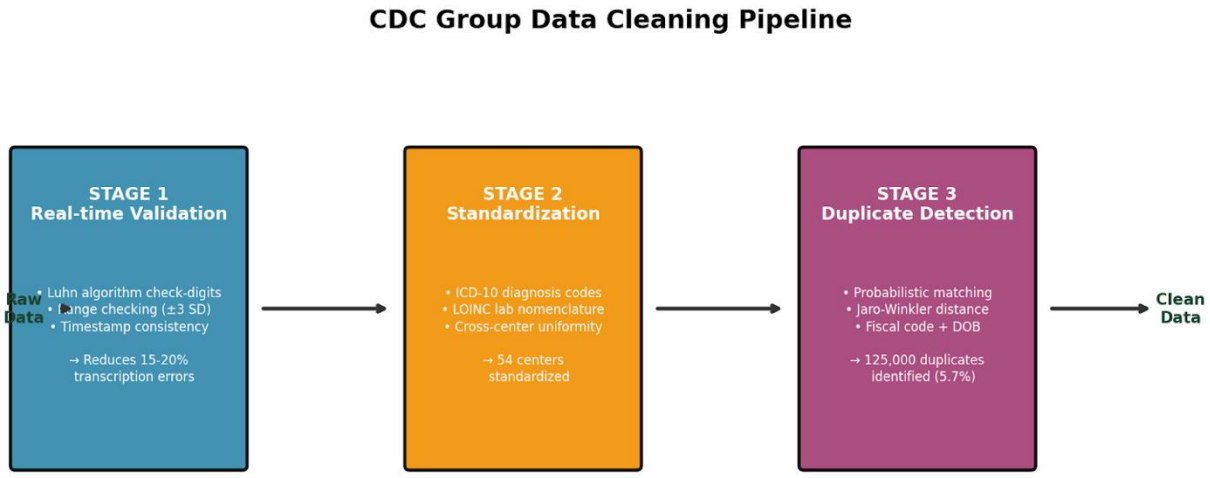
2012). For lab results, we run range checking against our relative range data points; greater than 3 standard deviations will activate automatic flags. Time markers confirm consistency, and the examination date precedes the result reporting date.

Stage 2: Standardization and Normalisation
We created scripts that standardize data and terminology across the 54-center network. Medical diagnoses are mapped to standardized ICD-10 codes and laboratory examinations coincide with LOINC (Logical Observation Identifiers Names and Codes) nomenclature.

Stage 3: Duplicate Detection and Entity Resolution
We utilize a probabilistic matching algorithm. The algorithm compares fiscal codes (codice fiscale), names with distance metrics from Jaro-Winkler's data (Jaro-Winkler distance metric), and date of birth. 125,000 duplicate records were identified by our system, (5.7% of total patient base), this demands manual finding.







Conclusion

The design of the CDC Group healthcare database provides a robust and scalable solution for managing patient information, medical records, and administrative data efficiently. The implementation of a normalized schema, enforces data integrity through constraints, and incorporating security measures such as role-based access control and encryption, the system also ensures accuracy, confidentiality, and compliance with healthcare regulations (e.g., HIPAA), building confidence among providers, regulators, and patients. “SQL Server offers a solid foundation for managing sensitive patient information, its capabilities ensure

compliance with HIPAA regulations". (Blaze.tech, 2025). Overall, this database design lays the foundation for improved data accessibility, streamlined workflows, and better decision-making, ultimately contributing to enhanced patient care and operational efficiency. To further improve the system, future enhancements should include integration with advanced analytics, migration to cloud infrastructure and incorporation of AI-driven insights for predictive care. Additionally, developing mobile and patient portal solutions together with regular security audits will strengthen patient engagement and safeguard sensitive information.

References

Liu, M., Luo, J., Li, L., Pan, X., Tan, S., Ji, W., Zhang, H., Tang, S., Liu, J., Wu, B., Chen, Z., Wu, X. and Zhou, Y. (2023). Design and development of a disease-specific clinical database system to increase the availability of hospital data in China. Health Information Science and Systems, 11(1). doi:https://doi.org/10.1007/s13755-023-00211-4.

De Mello, B.H., Rigo, S.J., da Costa, C.A., da Rosa Righi, R., Donida, B., Bez, M.R. and Schunke, L.C. (2022). Semantic interoperability in health records standards: a systematic literature review. *Health and Technology*, 12(2), pp.255–272. doi:https://doi.org/10.1007/s12553-022-00639-w.

Kahn, M. G., Raebel, M. A., Glanz, J. M., Riedlinger, K., & Steiner, J. F. (2012). A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Medical Care, 50, S21-S29.

Naugler, C., & Church, D. L. (2019). Automation and artificial intelligence in the clinical laboratory. Critical Reviews in Clinical Laboratory Sciences, 56(2), 98-110.
Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin, 23(4), 3-13.

Sujansky, W. V. (2001). Heterogeneous database integration in biomedicine. Journal of Biomedical Informatics, 34(4), 285-298.

Kevin Kline (2024). The Top Three Reasons Why Secure, Reliable Databases Are Essential for Modern Healthcare Organizations. [online] Database Trends and Applications. Available at: https://www.dbta.com/Editorial/Think-About-It/The-Top-Three-Reasons-Why-Secure-Reliable-Databases-Are-Essential-for-Modern-Healthcare-Organizations-166228.aspx.

Belcic, I. and Stryker, C. (2024). What is an Entity Relationship Diagram? | IBM. [online] www.ibm.com. Available at: https://www.ibm.com/think/topics/entity-relationship-diagram.

XTIVIA (2023). Microsoft SQL Server: Advantages & Best Practices for Technical Corporate Decision Makers. [online] Virtual-DBA Remote DBA Services & Support - Certified Database Experts. Available at: https://virtual-dba.com/blog/microsoft-sql-server-advantages-and-best-practices/.

Blaze.tech. (2025). Top HIPAA-Compliant Databases for Secure Healthcare Data Management. [online] Available at: https://www.blaze.tech/post/hipaa-compliant-database.