# Assess Data Warehousing Needs

Gladys Murage

School of Technology, North Central University

## TIM-7020 v5: DATABASES AND BUSINESS INTELLIGENCE(8329676021)

Dr JENNY DU

January 25 , 2025

**Summary of the Starting Schema**

The starting schema chosen for my base design compromises of 6 core components. The schema is highly relational, with patient\_id serving as the central key linking diagnoses, medications and genomic data, visits and healthcare facility interactions, and insurance, billing claims and outcomes. The strengths of the starting schema are first its comprehensive design which covers patient demographics, medical history and financial information. Secondly are its well-defined relationships with clear connections between entities like patients, visits and outcomes. Finally, is the support for advanced analytics by including genomic and socioeconomic data for predictive and comparative studies. The summary of the starting schema is as follows:

1. **Patient information**

This consists of the table patients and captures demographics and identifying details of patients such as patient\_id (primary key), name, gender, race and birthdate. It also contains details and other personal information. The patient table serves as the central entity linked to various tables for patient-related data.

1. **Medical Records**

Medical records consist of 3 key tables: diagnoses, medications and covid\_19\_records.

1. Diagnoses: this entity stores patient diagnoses, including diagnoses codes, dates and descriptions . It also links to patients via patient\_id.
2. Medications: this entity tracks prescriptions or medications administered to patients. It includes medication\_name, dosage and prescription\_date.
3. Covid\_19\_records: this entity captures COVID-19 related information such as infection dates and severity. It links to patient via patient\_id.
4. **Healthcare Access and Visits**

Contains 2 tables: healthcare\_facilities and visits.

1. Healthcare\_facilities: this entity contains data about hospitals, clinics or healthcare providers. It includes facility\_id, location and specialization details.
2. Visits: this entity logs patients visits to healthcare facilities. It includes visit\_id, visit\_date and links to facilities and patients.
3. **Outcomes and Risk Factors**

Contains 2 tables: outcomes and socio-economic factors.

1. Outcomes: this entity tracks health outcomes such as recovery, hospitalization or mortality. It includes outcome\_type and related dates.
2. Socioeconomic\_factors: this entity stores socioeconomic details such as income level and education. It is linked to patients to study the impact of social determinants of health.
3. **Genomic Data**

This contains one entity table, the genomic\_data, which stores patient genetic information and risk factors. This is useful for studying genetic predispositions to diseases for the patient.

1. **Billing and Insurance**

It contains 2 key entities: insurance\_information and billing\_claims.

1. Insurance\_information: this entity tracks the patient’s insurance details such as coverage type and expiration date. It links back to patients through patient\_id.
2. Billing\_claims: this entity records billing and claims for medical services. It includes claim\_id, amount, and approval\_status.

This schema forms a strong base to integrate new data sources and to enhance analytic capabilities for deeper insights into patient health, healthcare outcomes and resource utilization. This schema provides a robust foundation for adding external data sources: firstly, is the public health statistics to allow benchmarking of patient data against population level health trends. Secondly, are the social determinants of health(SDOH) in order to study the influence of environment, community and social factors. Third is clinical research data to link patient diagnoses and treatments with clinical trial outcomes. Finally, it is wearable and lifestyle data to analyze real-time health metrics like activity levels and sleep patterns.

**Data Sources to Integrate to Bolster Research**

The purpose of the research is to find out if COVID-19 disproportionately causes new onset hypertension among people of different races and ethnicities, and to identify whether biological and socioeconomic factors contribute to these disparities. Firstly, it would therefore benefit this research and bolster it by adding public health data sets such as those from the Centers for Disease Control (CDC), World Health Organization (WHO), National Institutes of Health (NIH) and state health departments. This would allow analyzing the trends in COVID-19 prevalence of hypertension and allow benchmarking of my patient data against population health statistics. The integration strategy would be to create a new table called public\_health\_statistics to store aggregated data, it would include columns like disease\_name, region, age\_group, year, prevalence\_rate.

Secondly, is the integration of clinical research studies from sites such as clinicltrials.gov and research publications. This allows the linkage of patient conditions to clinical trials or emerging treatments, allowing prediction of outcomes based on evidence-based practices. The integration strategy would entail adding a research\_links table to associate diagnoses or outcomes with trial data.

Third, for social determinants of health (SDOH) is to integrate US Census Bureau data along with any available neighborhood demographics or socioeconomic indicators. The potential cases are to understand how social factors such as income, education, or environment affect patient outcomes. This helps in developing targeted interventions for high-risk groups. The integration strategy is to enhance the socioeconomic\_factors table with more granular data such as zip\_code, housing\_status, transportation\_access but to also create an entity called census\_data.

Refer to Figure 1 and 2 below for the new 3 entities embedded into the already existing ERD with foreign keys: census\_data, public\_health\_statistics, and research\_links

**Figure 1**

*An entity relationship diagram (schema) of mydatabase in MySQL Workbench with 26 tables*

A screenshot of a computer

Description automatically generated

**Analytics-Driven Use Cases**

With the integrated data, I am now able to perform more advanced analytics by increasing the breadth and depth as follows:

1. **Predictive Analytics:** There is now capability to use machine learning to predict which patients are at risk of developing hypertension post COVID-19 based on genetic, socioeconomic and lifestyle data.
2. **Geospatial Analysis:** It is now possible to identify hotspots for hypertension by mapping public health, and patient data by regions and even zip codes.
3. **Cost Analysis:** There is now the ability to evaluate the economic impact of hypertension on different patient groups.
4. **Outcome Optimization:** There is capability to identify the most effective treatments by linking the outcomes of disease to clinical trial data.

**Figure 2**

*A reverse engineered ERD diagram of the schema with an additional 3 tables for a total of 26*

A computer screen shot of a computer screen

Description automatically generated

**Data Storage and Management**

According to Alderden et al. (2023), good stewardship of research data begins with a comprehensive data management plan to enable the organization and all users to access information. According to another research, around the world massive amounts of data are being collected and analyzed and this stresses the importance for building excellent database management for better patient diagnosis, treatment, improving public health systems, assisting government agencies in designing and implementing public health policies and in instilling confidence if future generations who desire a better public health system (Yogesh & Karthikeyan, 2022).

Secondly, It is important to anticipate that the needs of the organization and complexity will only increase with time and this according to Tarigan and Sembiring (2023) is also true for health services which compromise health institutions for the community and do certainly involve a lot of data in their operations. Starting with the registration process, medical/paramedical examinations, hospitalization, purchasing drugs, and laboratories. These data require a container that can be used to manage the data needed in the transaction. That way later on these data can be easily integrated and processed into a useful output. Database is one of the media that can be used to accommodate all supporting data; by using a database and the features of a database management system (DBMS) can help companies carry out their operations more effectively and efficiently.

**Conclusion**

For the current schema provided in this assignment with 26 tables, and the versatile and diverse requirements of an academic research hospital, MySQL is the best choice for a relational database management system (RDBMS). The schema suggests a highly structured data set with clear relationships between entities such as parents, healthcare facilities, and outcomes. MySQL RDBMS is well suited firstly for complex relationships because MySQL supports foreign keys, which enables enforceability of data integrity across related tables. Secondly it ensures atomicity, consistency, isolation and durability (ACID) compliance, thereby providing data consistency and reliability especially for critical and sensitive healthcare data. Third, MySQL provides query power, allowing complex queries for analytics and reporting. Fourthly, MySQL can handle large data sets with appropriate index and partitioning.

In addition, there are alternative solutions to consider starting with cloud-based RDBMS such as Amazon RDS or Google Cloud SQL. Cloud-based RDBMS becomes necessary as the organization needs grow in scalability. They require minimal maintenance. The advantages are that there are automated backups, scaling and high availability. The con is that these cloud solutions have higher costs than on premise MySQL if the dataset grows significantly fast. The schema lacks unstructured data, if this were the case, combination with NoSQL systems would have been necessary.

The second consideration in addition to MySQL is data warehouses for example snowflake and BigQuery. The schema calls for analytics using large volumes of historical data such as the census data, clinical research and public medical data sets. Data warehouses become necessary to run such analytics so as to establish predictions between hypertension and COVID-19. The drawbacks are once again, the higher cost, and the fact that they are less suitable for frequent updates typical in operational systems.

Alderden et al. (2023) developed a relational database with a customized dashboard and query instrument to manage large, diverse and complex data collected for a large clinical trial. This relational database serves as an exemplary for healthcare research data management, which highlights the effectiveness of relational databases in organizing and providing access to complex healthcare data. This peer reviewed research further supports the reason as to why an RDBMS, specifically MySQL is the recommendation for operational data management for mydatabase. The reason being that for this academic research hospital there is a well-structured schema with the dataset already modeled for relational databases. Secondly, there is cost effectiveness being that MySQL is free and open source. Third, MySQL supports scalability through partitioning and clustering for large datasets. Finally, there is consideration for integration. It is easy to integrate with analytic tools such as tableau, python, or R for advanced analysis .

**References**

Alderden, J. G., Sharkey, P.D., Kennerly, S.M., Ghosh, S., Barrett, R.S., Horn, S.D., Ghosh, S., & Yap, T.L. (2023). Developing a relational database for best practice data management: The turn everyone and move for ulcer prevention database. *Computers, Informatics, Nursing*. *41*(2):59-65. <https://doi.org/10.1097/CIN.0000000000001011>

Tarigan, W., & Sembiring, R. S. B., (2023). Relational database for health care. *Jurnal Penelitian Pendidikan*. *9*(9). 7354-7360. <https://doi.org/10.29303/jppipa.v9i9.4856>

Yogesh, M. J., & Karthikeyan, J. (2022) Health informatics: Engaging modern healthcare units: A brief overview. *Front Public Health*.*10*: 1-13. <https://doi.org/10.3389/fpubh.2022.854688>