Cassandra Palmer

BIOI 4870

Final Report

EHR Database Modeling

# Abstract

EHR Database Modeling accesses the open-source health record datasets provided by MIMIC-III, MIMIC-IV, and eICU to generate a micro-scale version of an EHR database. The querying capabilities of this database is implemented in the user-facing webpage, which accepts input choices of gender, ethnicity, and selection of diagnoses or procedures to output. The output of the query is a descending table of the diagnosis or procedure descriptions, as well as the frequency of the descriptions within patients of the selected gender and ethnicity. The results of the EHR Database Model in this project support the need for common data normalization, standardization, and storage processes for continuation of health data creation, updating, sharing, and analyzing.

# Research Question

This project analyzes a total of 100 health records from the MIMIC-III demo database, 100 health records from the MIMIC-IV demo database, and 2,174 health records from the eICU demo database in order to create a model Electronic Health Record (EHR) relational database. Each of the records are for distinct patients and hospital stays and creates a simplified representation of how health data is collected, secured, accessed, and available for reuse.

# Background

Electronic Health Records (EHRs) are maintained by healthcare systems to collect, organize, and extract clinical data regarding patient medical history (Alzu'bi et al., 2021). As healthcare treatments are being supported by an increase in data, EHRs are becoming more important for organizing said data and increasing the patient and healthcare worker satisfaction (Aguirre et. al, 2019). EHR implementation and adoption varies between healthcare systems (Upadhyay & Opoku-Agyeman, 2021), therefore it is important to create a seamless database to benefit EHR interoperability and Health Information Exchanges (HIEs). HIE organizations are typically seen implemented in hospitals and outpatient providers in order to exchange health records, results, provider notes, medications, and more. The exchange of health data has become increasingly important for continuation of care and transparency with medical histories (Martin et.al, 2018).

EHRs contain individual patient identification and care notes (Ehrenstein et. al, 2019), thus a publicly available database of EHRs should utilize proper HIPAA privacy and security standards and a simple way to extract useful information. EHR user interfaces can be limited and confusing between varied systems, which can cause errors when inputting, extracting, or querying data (Bowman 2013). For example, the EHR data contained in the MIMIC-III database includes admission dates of patients, but these dates are inconsistently shifted from the real date of patient admission. Therefore, these dates should not be used to analyze hospital stay trends in regard to the specific day of the week or time of year (Wang & Wu, 2019). The datasets collected from MIMIC-III, MIMIC-IV, and eICU that were used in this model database had differing date formats, naming conventions, patient identification styles, and database schemas. The different schemas of each database show the complexity and wide range of possibilities when collecting and storing deidentified health data, which also contributes to the complexity and wide range of possibilities of accessing deidentified health data for HIEs.

There are existing software and services available for health data deidentification, health data analysis, and machine learning regarding EHR databases (Bulgarelli et. al, 2020). These tools are valuable for speeding up the cycle of health data collection, storage, and implementation. Common data models also increase the turnaround of health data because they provide a common ontology to classify data from different sources in the same manner (Bulgarelli et. al, 2020). MIMIC-III, MIMIC-IV, and eICU utilize common data model and data normalization techniques such as the Observational Medical Outcomes Partnership (OMOP) analytic capabilities, LOINC terminology standardization, and the National Library of Medicine’s Unified Medical Language System (UMLS) International Statistical Classification of Diseases and Health Related Problems (ICD) (Reimer & Milinovich, 2020).

Other databases have been created to identify, relate, and organize data from various sources as seen in the Transport Data Repository constructed by Reimer and Milinovich (2020). Reimer and Milinovich concluded that the lower mapping results of their relational database is due to non-standardized data generation and storage from a third-party source (2020). The results of Transport Data Repository mapping suggest that medical record data collection, storage, access, and reuse is impacted by inconsistent data models, standardization, and normalization. The model EHR database of this project, although extremely simplified, emphasizes the constraints of accessing and reusing deidentified health data from various resources.

# Database Diagram

Text

Description automatically generated

# Methods

## Code

Repository Link: https://github.com/sassiecassie/EHR-Database-Modeling.git

Sources for database development:

* eICU-CRD Demo: Version 2.0.1
  + Total uncompressed file size: 130.6 MB
  + Can be accessed via SQLite, google cloud, or data dump from website https://physionet.org/content/eicu-crd-demo/2.0.1/ as a .zip file
* MIMIC-III Database: Version 1.4
  + Total uncompressed file size: 103.0 MB
  + Can be accessed by google cloud or data dump from website https://physionet.org/content/mimiciii-demo/1.4/ as a .zip file
* MIMIC-IV Demo Data: Version 0.9
  + Total uncompressed file size: 73.0 MB
  + Can accessed using google cloud or data dump from website https://physionet.org/content/mimic-iv-demo-omop/0.9/ as a .zip file

## Data Provenance

Sources for data:

Note: All final csv files were parsed to account for null values, quotations around values, internal commas, and other factors that affected csv file delimitation of values

* eICU-CRD Demo: Version 2.0.1
  + Data dump of files diagnosis.csv.gz, patient.csv.gz, and treatment.csv.gz from https://physionet.org/content/eicu-crd-demo/2.0.1/#files-panel
    - diagnosis.csv.gz: 352.1 KB
      * Select the columns containing the patient\_unit\_stay\_id, description, and icd9code
      * Final version size:
    - patient.csv.gz: 133.6 KB
      * Select the columns containing the patient\_unit\_stay\_id, patient\_health\_system\_stay\_id, gender, and ethnicity
      * Final version size:
    - treatment.csv.gz: 450.1 KB
      * Select the columns containing the patient\_unit\_stay\_id and treatment\_description
      * Final version size:
* MIMIC-III Database: Version 1.4
  + Data dump of files ADMISSIONS.csv, DIAGNOSES\_ICD.csv, D\_ICD\_DIAGNOSES.csv, D\_ICD\_PROCEDURES.csv, PATIENTS.csv, and PROCEDURES\_ICD.csv from https://physionet.org/content/mimiciii-demo/1.4/
    - ADMISSIONS.csv: 26.2 KB
      * Select the columns containing the subject\_id, hosp\_stay\_id, ethnicity, and diagnosis
      * Final version size:
    - DIAGNOSES\_ICD.csv: 47.8 KB
      * Select the columns containing the subject\_id, hosp\_stay\_id, and icd9code
      * Final version size:
    - D\_ICD\_DIAGNOSES.csv: 1.3 MB
      * Select the columns containing the icd9code and long description
      * Final version size:
    - D\_ICD\_PROCEDURES.csv: 283.1 KB
      * Select the columns containing the icd9code and long description
      * Final version size:
    - PATIENTS.csv: 8.4 KB
      * Select the columns containing the subject\_id and gender
      * Final version size:
    - PROCEDURES\_ICD.csv: 13.2 KB
      * Select the columns containing the subject\_id, hosp\_stay\_id, and icd9code
      * Final version size:
* MIMIC-IV Demo Data: Version 0.9
  + Data dump of files person.csv, procedure\_occurrence.csv, and visit\_detail.csv from https://physionet.org/content/mimic-iv-demo-omop/0.9/#files-panel
  + Data dump of gcpt\_proc\_itemid.csv from https://github.com/OHDSI/MIMIC/blob/master/custom\_mapping\_csv/gcpt\_proc\_itemid.csv
    - person.csv: 8.2 KB
      * Select the columns containing the person\_id, gender, and ethnicity
      * Final version size:
    - procedure\_occurrence.csv: 2.3 MB
      * Select the columns containing the person\_id and procedure\_concept\_id
      * Final version size:
    - visit\_detail.csv: 3.6 MB
      * Select the columns containing the visit\_detail\_id and person\_id
      * Final version size:
    - gcpt\_proc\_itemid.csv: 16.9 KB
    - Select the columns containing the concept\_name and target\_concept\_id
    - Final version size:

## Webpage

http://odin.unomaha.edu/~cassandrapalmer/EHRdatabasemodeling.php

Sample Query 1:

Gender: Male

Ethnicity: White

Same diagnoses selected

A picture containing text

Description automatically generatedOutput:

Sample Query 2:

Gender: Female

Ethnicity: Black

Same procedures selected

Graphical user interface, text, application, email

Description automatically generated

# Results

The results of my project demonstration show the variety of procedure and diagnosis description format as well as the similarity between the descriptions. Some of the procedure and diagnosis descriptions available from the different sources were NULL, therefore they were not included in the results. This demonstration meets my original expectations of being able to sort and select patient health records. I had originally anticipated being able to combine many columns and queries to produce more specific results. The lack of computing power, algorithm mastery, and detailed open-source data contributes to the extremely simplified health data querying power. These results however do indicate the current complexities and future advancements in EHR databases, specifically in the structuring, naming, scaling, and mapping of health record fields.

# Discussion/Conclusions

The EHR Database Model presented in this project shows the difficulties researchers, health care providers, and patients may encounter when looking for specific health data. The incomplete and inconsistent fields between the MIMIC-III, MIMIC-IV, and eICU databases is understandable given the large variety of EHR systems used by different healthcare facilities and networks, which alludes to the difficulty in implementing interoperable HIEs and current dissatisfaction of healthcare providers in regards to EHR use (Brian et. al, 2017).

Analysis of the EHR Database Model in this project demonstrates a detail-oriented representation of health data. As previously stated, the source databases were non-synonymous when accessing records and mappings of values. In order to sort and standardize the data, it was necessary to create tables from each source in the Odin server, then sort through and trim the tables based off of similar attributes. Following this, the trimmed datasets were inserted into different tables in order to create relations, then finally inserted into “master” tables. Although it could have been possible to query each of the many relations for results, this would have required additional queries and data analysis.

While completing this project, I ran into many challenges while trimming the datasets for the valuable tuples. There is an immense amount of data for hundreds of patients and their respective hospital stays, therefore I decided to select unique patient IDs and the lowest hospital stay ID for the patient ID to represent the patient’s first recorded hospital stay. I also ran into challenges while inserting CSV attributes into MySQL due to thousands of NULL values and internal commas inside of string values. In order to combat this, I found it necessary to use VIM/REGEX search and replace processes to change internal commas into semicolons and include NULL inside of repeating comma patterns. Although my initial project scope was extremely broad and difficult to achieve, I feel accomplished with my increased awareness of health databases and managing my own database. I also feel accomplished with being able to get my database to produce meaningful results, as I had issues achieving such results when querying the source databases. EHR databases and datasets have immense records that when comparable, can produce important results for analyzing health trends within demographic, time, and system constraints.

# References

Aguirre, R. R., Suarez, O., Fuentes, M., & Sanchez-Gonzalez, M. A. (2019). Electronic Health Record Implementation: A Review of Resources and Tools. *Cureus*, *11*(9), e5649. https://doi.org/10.7759/cureus.5649

Alzu'bi, A. A., Watzlaf, V., & Sheridan, P. (2021). Electronic Health Record (EHR) Abstraction. *Perspectives in health information management*, *18*(Spring), 1g.

Bowman S. (2013). Impact of electronic health record systems on information integrity: quality and safety implications. *Perspectives in health information management*, *10*(Fall), 1c.

Brian G., Arndt et al., (2017) “Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations,” *The Annals of Family Medicine* 15, no. 5 419–26, doi: 10.1370/afm.2121.

Bulgarelli L., Núñez-Reiz A., Deliberato R.O. (2020) Building Electronic Health Record Databases for Research. In: Celi L., Majumder M., Ordóñez P., Osorio J., Paik K., Somai M. (eds) Leveraging Data Science for Global Health. Springer, Cham. https://doi.org/10.1007/978-3-030-47994-7\_4

Ehrenstein V, Kharrazi H, Lehmann H, et al. (2019) Obtaining Data From Electronic Health Records. In: Gliklich RE, Leavy MB, Dreyer NA, editors. Tools and Technologies for Registry Interoperability, Registries for Evaluating Patient Outcomes: A User’s Guide, 3rd Edition, Addendum 2 [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2019 Oct. Chapter 4. Available from: https://www.ncbi.nlm.nih.gov/books/NBK551878/

Johnson, A., Pollard, T., & Mark, R. (2019). MIMIC-III Clinical Database Demo (version 1.4). *PhysioNet*. <https://doi.org/10.13026/C2HM2Q>.

Johnson, A., Pollard, T., Badawi, O., & Raffa, J. (2021). eICU Collaborative Research Database Demo (version 2.0.1). *PhysioNet*. <https://doi.org/10.13026/4mxk-na84>.

Kallfelz, M., Tsvetkova, A., Pollard, T., Kwong, M., Lipori, G., Huser, V., Osborn, J., Hao, S., & Williams, A. (2021). MIMIC-IV demo data in the OMOP Common Data Model (version 0.9). *PhysioNet*. <https://doi.org/10.13026/p1f5-7x35>.

Martin, T. J., Ranney, M. L., Dorroh, J., Asselin, N., & Sarkar, I. N. (2018). Health Information Exchange in Emergency Medical Services. *Applied clinical informatics*, *9*(4), 884–891. https://doi.org/10.1055/s-0038-1676041

Reimer, A. P., & Milinovich, A. (2020). Using UMLS for electronic health data standardization and database design. *Journal of the American Medical Informatics Association : JAMIA*, *27*(10), 1520–1528. https://doi.org/10.1093/jamia/ocaa176

Upadhyay, S., & Opoku-Agyeman, W. (2021). Factors that Determine Comprehensive Categorical Classification of EHR Implementation Levels. *Health services insights*, *14*, 11786329211024788. https://doi.org/10.1177/11786329211024788

Wang, L., & Wu, X. (2019). Caution should be taken when using electronic health record database. *Critical care (London, England)*, *23*(1), 345. https://doi.org/10.1186/s13054-019-2612-5