Journal Entry One

DAT-375

Brandon Hobbs

March 6, 2023

I am being asked to help a hospital with an internal study. They want me to, “track the most common injuries to the age group ‘Infant/Child’ within the last 10 years in comparison to the most common injuries of the previous 10 years for the same age group at this hospital.” I think it might also be a good idea to compare this to a nationwide statistic.

To start this analysis, I would want to break this down into three sections: Problem Identification, Data Requirements, and Extraction Selections.

**Problem Identification**

In this section I will attempt to clarify the problem and identify the data I need.

To start, a definition of *Infant/Child* must be established. The CDC (2021) defines an infant as a human from birth to one years old. A child is similarly defined as a human less than 18 years of age. These child years are broken down by the CDC into many subclassifications, e.g., middle childhood as 6-8 years.

An injury will be defined as any condition that prompted the child to be brought to the attention of the hospital staff. Therefore, any patient registered that is below 18 years of age shall be considered in this analysis.

The injury will be taken from the official discharge paperwork and not the intake. If more than one injury is recorded, then they will be added to the dataset individually and independently.

Lastly, the current year of analysis will begin January 1 at midnight central time. The prior years’ will be January 1, midnight 2022 until December 31, 2022, at 23:59. Any patient entered into the hospital’s system prior to any of these dates will be excluded.

**Data Requirements**

In this section I will attempt to discuss the data I would isolate and extract to produce results.

As the hospital is interested in injuries this needs to be the central theme of the data. The first step will be finding a reliable source of data. As this hospital uses ICD-11 (WHO) when coding for billing this will be the data used for the source of injury.

The date, as per the definition in the previous section, will be the discharge time.

Each patient record will have any PII (personal identifiable information) stripped, e.g., name, social security number, phone number, etc., before adding the data to the central database for this analysis. After removing the PII each record fitting the age cutoff will be moved into the database marked *raw.*

This raw data will be left for traceability and backup purposes.

The raw data will then be run through an ETL pipeline where each ICD-11 code will be extracted, along with other characteristic data, namely: age at discharge, date of discharge, and sex. The patient ID will be substituted with a system generated hash code. This hash code will be reversible if data reconciliation is needed.

The ETL pipeline will place each record into a Google BigQuery table labeled *transformed*.

The data will be visualized and analyzed through Tableau to take advantage of its simple UX. The analysis of this data will start simple:

* Histogram of ICD-11 code, sorted in descending order, per year
* Histogram of ICD-11 code by sex, per year
* Histogram of ICD-11 code by age brackets (as per CDC), per year
  + 0-1 years
  + 1-2 years
  + 2-3 years
  + 3-5 years
  + 6-8 years
  + 9-11 years
  + 12-14 years
  + 15-17 year

**Extraction Selections**

In this section I will attempt to justify my thought process and choices.

As this is a comparative analysis, year on year, it is imperative to have a standard that existed across the period in question. If not, then the data would need to be transformed to a new standard which could introduce error. By using the ICD-11 codes, instead of doctor’s notes or other sources, e.g., patient complaint, the data will be mastered. That is, there is no need for lemmatization or stemming and allows the analysis to be more quantitative than qualitative as there will be no judgement needed for “injury type”. Moreover, as ICD-11 came out in 2022, and the hospital quickly updated its software, there will be no need to transform codes between ICD-11 and the older ICD-10.

Stripping the PII data makes the data less sensitive and less secretive and eases security without causing a loss in value. The value lies in making connections between sex, age bracket, and injury. For example, do 8-year-old boys have the same injuries as 8-year-old girls and has this changed year on year? Dividing the data up and looking at the injuries per bracket is also valuable, i.e., do breaks and strains happen at the same frequency across all brackets?

Lastly, by preserving the patient ID, via reversible hash code, in the transformed dataset data reconciliation is possible.

**References**

CDC. *Child development: infants (0-1 years).* (2021, November 29). <https://www.cdc.gov/ncbddd/childdevelopment/positiveparenting/infants.html>

World Health Organization. *ICD-11*. (2022, April 26). https://icd.who.int/en