Data Quality Assessment

Logan Correa

Background

Data quality is a critical factor when it comes to decision-making based on datasets. In order to ensure adequate data quality, it is necessary to perform a data quality assessment before working with raw data. The research paper: *Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research*, by Weiskopf and Weng (2013) defines several dimensions of a data quality assessment. These components are completeness, correctness, concordance, plausibility, and currency.

Completeness represents the extent to which data is present, correctness represents the extent to which data is accurate, concordance represents the agreement between elements within the data, plausibility represents whether an element makes sense in the context of the data, and currency represents how relevant the data within the dataset is.

Methods

In order to adequately assess data dimension quality within the TNX_20200129 Diagnosis and Medication domains, several data quality assessment methods were selected from Weiskopf and Weng's (2013) paper:

1. Data element agreement: Two or more elements within an EHR are compared to see if they report the same or compatible information.
2. Element presence: A determination is made as to whether or not desired or expected data elements are present.
3. Validity check: Data in the EHR are assessed using various techniques that determine if values 'make sense'.
4. Log review: Information on the actual data entry practices (eg, dates, times, edits) is examined.

For this data quality assessment, completeness will be evaluated using data element agreement, and element presence testing; correctness will be evaluated using data element agreement, validity check, and log review testing; concordance will be evaluated using data element agreement testing; plausibility will be evaluated using data element agreement, and validity check testing; and currency will be evaluated using log review testing. The tests and graphical representations of the results were created and performed using the Python programming language. The code for these tests was uploaded to a GitHub repository to be available for secondary review.

Several assessment methods (gold standard, data source agreement, and distribution comparison) were excluded from this data quality assessment due to the nature of data we are working with and a lack of source/external datasets to compare to.

Results

Data Element Agreement

In order to test data element agreement, the test codes in the Diagnosis and Medication domains were cross-checked against the test codes within the Terminology domain. This revealed that 100% of codes within the Diagnosis and Medication domains were properly mapped to codes within the Terminology domain.

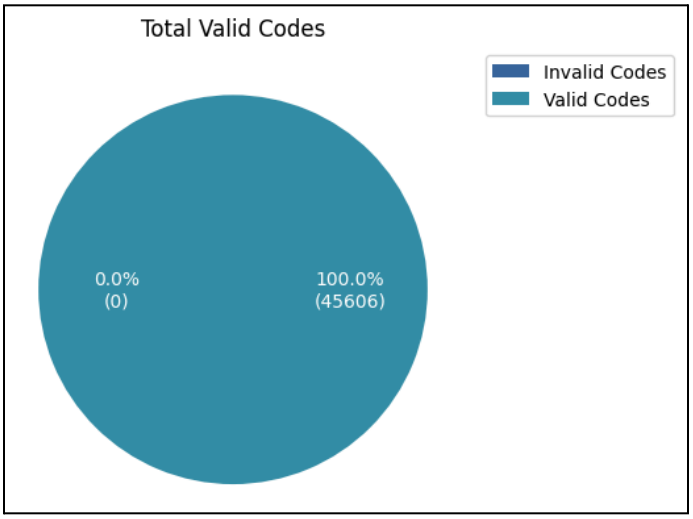

Figure 1. Valid and invalid code percentages for diagnosis and medication domains. Assessment of data revealed that all codes matched to the terminology domain.

Element presence

In order to test element presence within this dataset, null and duplicate values were identified within the Diagnosis and Medication domain. Assessment of the data revealed that there were no null values present within these domains with 100% of the entries properly filled.

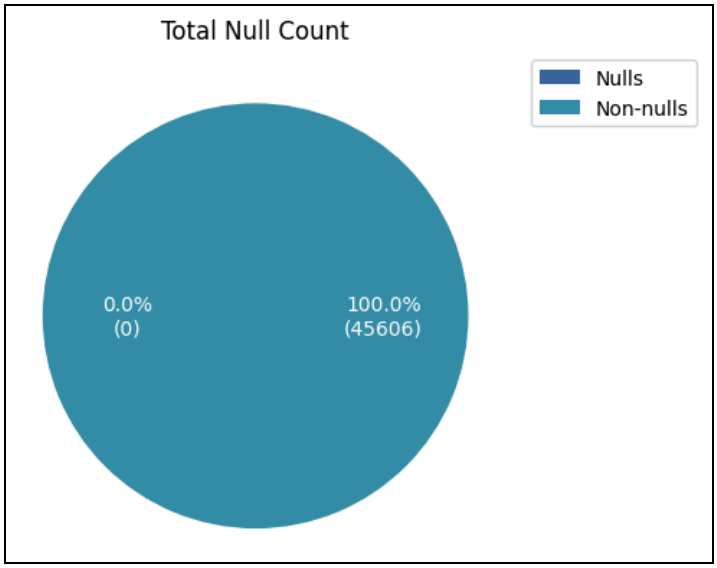

Figure 2. Null and non-null percentages for diagnosis and medication domains.

Analysis of duplicate values revealed a significant number of duplicates within the Diagnosis and Medication domains. 35.8% of Diagnosis entries and 5.3% of Medication entries were duplicates. 33.1% of Diagnosis and 25.8% of Medication domain entries were unique values. Total duplicate values within the Diagnosis and Medication domains was 41.1% while the remaining 58.9% were unique entries.

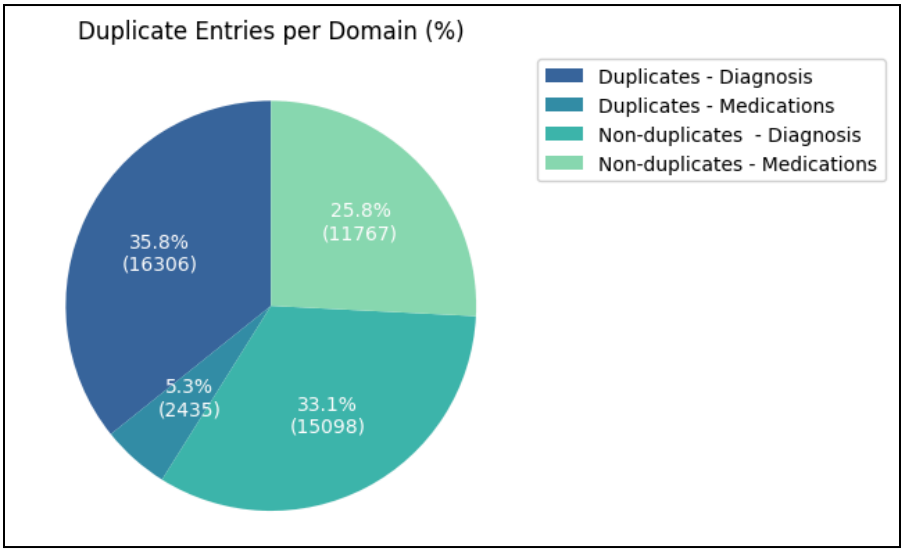

Figure 3. Duplicate and non-duplicate percentages for Diagnosis and Medication domains.

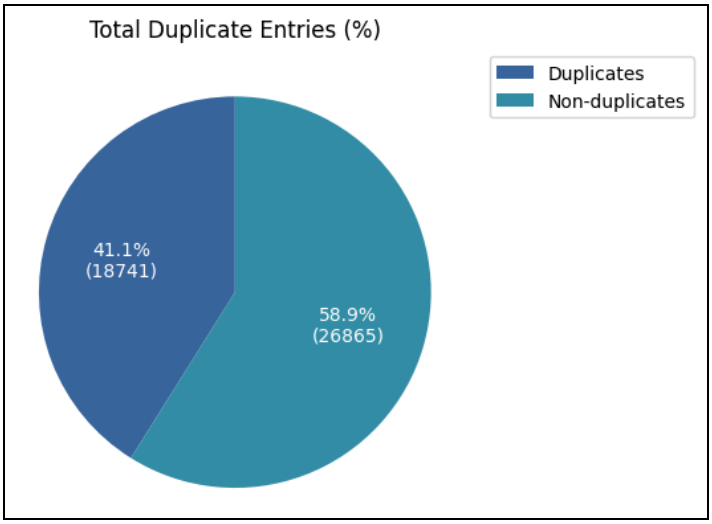

Figure 4. Total duplicate and non-duplicate percentage for diagnosis and medication domains.

Validity Check

In order to check data validity, the route options from the Medication domain were plotted and were manually reviewed to make sure that they make sense in the context of the data. This analysis revealed that all entries make sense.

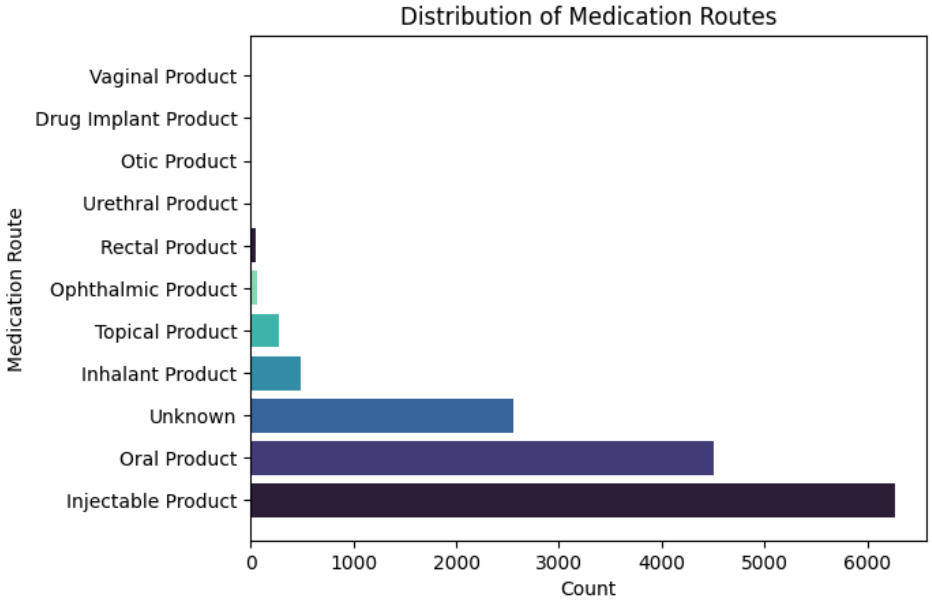

Figure 5. Distribution of medication routes in the Medication domain.

Log Review - date range 5/5

In order to perform data log review, the dates within the Diagnosis and Medication domains were graphed and manually reviewed to make sure that the range makes sense within the context of the data. The dates range from 2011 to 2020 without any outliers present, which is an acceptable range for this data.

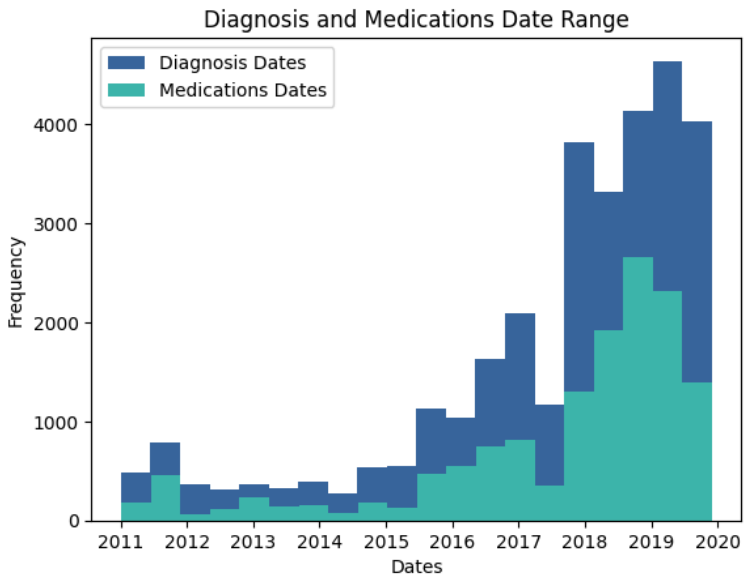

Figure 6. Range of dates for Diagnosis and Medication domains.

Discussion

Data quality dimensions were given a score of 1-5 for each applicable assessment method. The majority of assessment methods passed their assigned tests and were given a score of 5/5. The only method to receive less than a 5/5 was the element presence aspect of the completeness domain. This was given a score of 3/5 due to the high number of duplicate values found within the dataset.

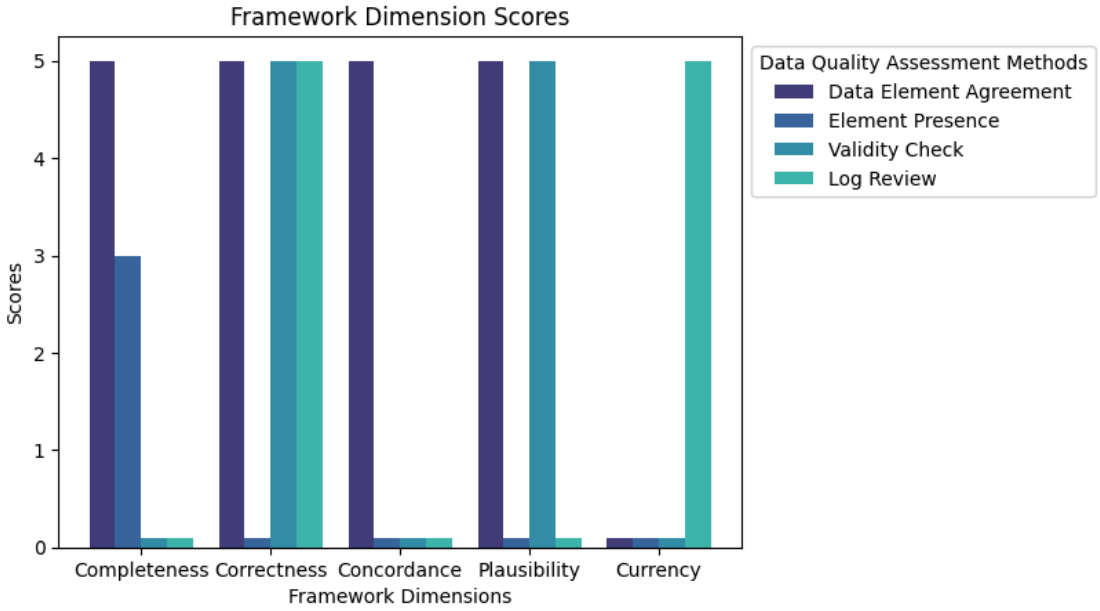

Figure 7. Framework dimension scores for each data quality assessment method.

The average for each framework dimension was calculated to give a normalized framework dimension score. Data correctness, concordance, plausibility, and currency received normalized scores of 5/5. Data completeness received a normalized score of 4/5. Overall, the data within the Diagnosis and Medications domains scored a 4.8/5.

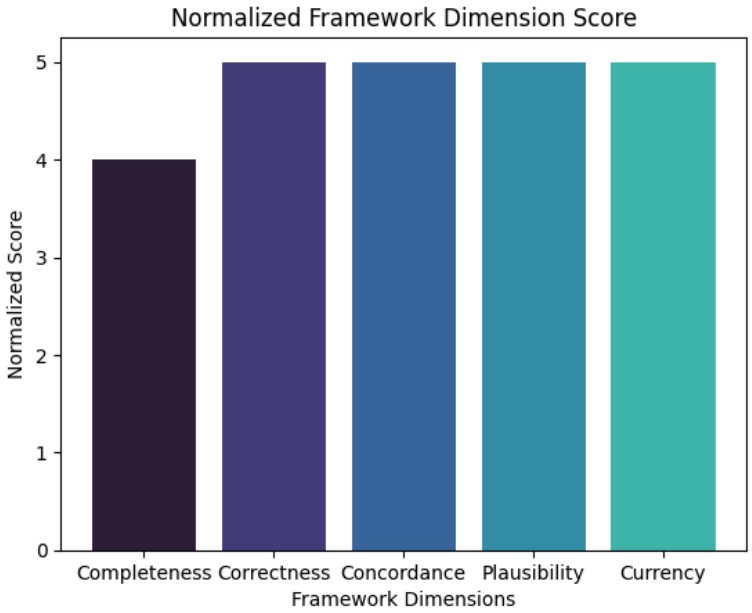

Figure 8. Normalized data scores for each data quality dimension.

References

Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association, 20(1), 144–151. https://doi.org/10.1136/amiajnl-2011-000681

Data Assessment Code:
https://github.com/logancorrea/Data-Wrangling/blob/6e7db879ad2678ee2e641c967c7b185ce0eeac7c/data-quality-assessment/data-quality-assessment.ipynb