Quality Assurance Practices in Managing Electronic Health Records Datasets for Research

Ursula A. Rogers¹, Shelley A. Rusincovitch¹ and Pamela Barth, MRP²
¹Duke Translational Research Institute, Durham, NC, ²Duke Health Technology Solutions, Durham, NC

Background and Introduction

Data from electronic health records (EHRs), which have been collected during routine healthcare, provide a wealth of information from which to conduct research. EHR dataset collections from multiple data partners have been an important basis for the Southeastern Diabetes Initiative (SEDI). Retrieving data from EHRs, however, is not an elementary task. When it comes to Quality Assurance (QA), it is imperative that certain considerations are taken into account to identify potential mistakes in extracting, organizing and representing data.

Findings

The SEDI analysts adopted three QA processes that were successful in addressing the challenges:

• Identifying the “source of truth” was extremely important for end-to-end validation. Data as far upstream as possible was compared with final values in the dataset.

• Utilizing analysts with root cause analysis skills was critical, specifically for being able to trace and sleuth data inconsistencies through various levels of transformation.

• Decisions about which data to extract and why were discussed and well documented so that there were no misunderstandings about the content of the data.

Recommendations

These methods are generalizable to a variety and breadth of data extraction projects. We encourage all QA teams to consider these topics and build adequate time in the project scope to address these concerns.

Methods & Challenges

After assessing several iterations of QA on SEDI EHR extractions, we generalized six challenges with high potential impact to quality of a final dataset:

1. Extraction bias: Extracting EHR data is heavily dependent on the person extracting data. It is influenced by the analyst’s background and their work infrastructure; for example, clinical versus financial operations groups.

2. Diluted data: Often, the data are pulled from downstream systems, such as data warehouses, and not directly from production. Downstream data sources may involve unknown “black box” logic or may only contain a subset of data.

3. Distributed domains: Data for given domains may be spread out across many tables and/or schemas. If the analyst is not well versed in the source tables, there is potential that data can be missed or data with different meaning is pulled instead.

4. Inappropriate mapping: Most analysts pulling the data do not have direct insight into how the data points were originally created within the workflow of clinicians or coders, potentially leading to inaccurate data mapping & assumptions based on variable names.

5. Data element selection: Technical teams must recognize and address decisions about which data to extract. These decisions may be influenced by the team’s perception of the customer’s expertise or intention.

6. Patient/encounter “drift”: Drift occurs naturally in EHRs when patients have multiple accounts that are later rectified and merged. The impact on datasets can be unanticipated when a query run against the same observation period has different results, due to underlying changes in the source data.

Background and Introduction

Data from electronic health records (EHRs), which have been collected during routine healthcare, provide a wealth of information from which to conduct research. EHR dataset collections from multiple data partners have been an important basis for the Southeastern Diabetes Initiative (SEDI). Retrieving data from EHRs, however, is not an elementary task. When it comes to Quality Assurance (QA), it is imperative that certain considerations are taken into account to identify potential mistakes in extracting, organizing and representing data.

Findings

The SEDI analysts adopted three QA processes that were successful in addressing the challenges:

• Identifying the “source of truth” was extremely important for end-to-end validation. Data as far upstream as possible was compared with final values in the dataset.

• Utilizing analysts with root cause analysis skills was critical, specifically for being able to trace and sleuth data inconsistencies through various levels of transformation.

• Decisions about which data to extract and why were discussed and well documented so that there were no misunderstandings about the content of the data.

Recommendations

These methods are generalizable to a variety and breadth of data extraction projects. We encourage all QA teams to consider these topics and build adequate time in the project scope to address these concerns.

Methods & Challenges

After assessing several iterations of QA on SEDI EHR extractions, we generalized six challenges with high potential impact to quality of a final dataset:

1. Extraction bias: Extracting EHR data is heavily dependent on the person extracting data. It is influenced by the analyst’s background and their work infrastructure; for example, clinical versus financial operations groups.

2. Diluted data: Often, the data are pulled from downstream systems, such as data warehouses, and not directly from production. Downstream data sources may involve unknown “black box” logic or may only contain a subset of data.

3. Distributed domains: Data for given domains may be spread out across many tables and/or schemas. If the analyst is not well versed in the source tables, there is potential that data can be missed or data with different meaning is pulled instead.

4. Inappropriate mapping: Most analysts pulling the data do not have direct insight into how the data points were originally created within the workflow of clinicians or coders, potentially leading to inaccurate data mapping & assumptions based on variable names.

5. Data element selection: Technical teams must recognize and address decisions about which data to extract. These decisions may be influenced by the team’s perception of the customer’s expertise or intention.

6. Patient/encounter “drift”: Drift occurs naturally in EHRs when patients have multiple accounts that are later rectified and merged. The impact on datasets can be unanticipated when a query run against the same observation period has different results, due to underlying changes in the source data.