CLIC Center for Leading Innovation & Collaboration

Insights to Inspire 2021

Informatics: Journey to Interoperability

Emily Pfaff, PhD, MS

CTSA Clinical & Translational Science Awards Program

The University of Rochester Center for Leading Innovation and Collaboration (CLIC) is the coordinating center for the Clinical and Translational Science Awards (CTSA) Program, funded by the National Center for Advancing Translational Sciences (NCATS) at the National Institutes of Health (NIH), Grant U24TR002260.

University of Rochester Medical Center

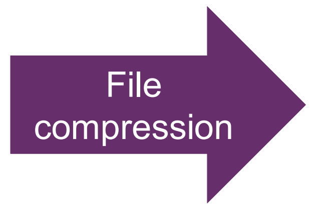
Infrastructure & Data Quality

Objective: Describe the role of infrastructure in improving data quality

During the Session

- What are some common examples of data quality issues in research data warehouses?
- What steps can hubs take to address these issues and improve data quality?
- How can institutional governance address data quality?

What are some common examples of data quality issues in research data warehouses?



Data transformation can make data more useful; however, each time data are transformed, there is a chance that data quality may degrade.

P.S.—For a more formal treatment of this subject, see Kahn, et al. (2016).

Mapping Errors

Simple human error
The concept of “ambulatory” visits in the source system gets mis-mapped to a similar-sounding word during ETL.

VISIT_ID	VISIT_TYPE	VISIT_DATE
34547	AMBULATORY	6/5/2004

Source data

VISIT_ID	VISIT_TYPE	VISIT_DATE
34547	AMBULANCE	6/5/2004

Transformed data

Content knowledge error
Serum and urine creatinine get mapped to the same lab identifier despite being very different tests.

PATIENT_ID	LAB_CD	LAB_NAME
29834723	Y77A89	CREATININE, SER
29834723	B212P0	CREATININE, UR

Source data

PATIENT_ID	LAB_CD	LAB_NAME
29834723	39452	CREATININE
29834723	39452	CREATININE

Transformed data

Missing Data

- It's important to know what data are and are not ETLed from your source database to your warehouse, so you can account for what's missing.
  - *Ex*: If genetic test data are stored in PDFs in the source system, they're unlikely to make it to your data warehouse as structured data.
  - *Ex*: Death data may lag a few months behind the present date in your source system.
- Individual values may also have a high rate of missingness.
  - *Ex*: If your EHR calculates BMI but does not store the calculated value, it may look like no one has a recorded BMI.
- In these cases the transformation is not *wrong*; but because the warehouse is far removed from the source, it can be *misleading* to users.
- Does not require "fixing"—rather, requires explanation.

Granularity Changes

DISCHG_DISP_CD	DISCHG_DISP_NAME
01	HOME
02	EXPIRED
03	TRANSFERED
04	LEFT AGAINST MED ADVICE
05	SKILLED NURS. FAC.
06	HOSPICE
07	REHAB

DISCHG_DISP_CD	DISCHG_DISP_NAME
H	HOME
D	DECEASED
OT	OTHER

- Transformations often “roll up” long lists of codes from a source system into a more manageable list.
- Can be helpful for analysis; aggregated categories should be guided by use case.
- Resulting aggregation may not be granular enough for all use cases.
- Source concepts can be grouped incorrectly—hard to trace back.

Loss of Context

All diagnosis codes are not the same—they have a type. If the type is lost through oversimplification, the data can be used incorrectly in analysis.

PATIENT_ID	DX_CD	DX_TYPE
29834723	E11.3	PATIENT REPORTED
29834723	U07.1	BILLING

Source data

PATIENT_ID	DX_CD
29834723	E11.3
29834723	U07.1

Transformed data

Losing a "status" flag on billing transactions can cause us to mix voided transactions in with non-voided transactions!

BILL_ID	BILL_AMT	STATUS
55476	3255.67	FINAL
55476	546.20	VOID

Source data

BILL_ID	BILL_AMT
55476	3255.67
55476	546.20

Transformed data

What steps can hubs take to address these issues and improve data quality?

Run periodic quality audits.

- Some common data models (e.g., PCORnet, OHDSI/OMOP) have pre-scripted data quality checks that can be run against local data.
- If you do not use a model with ready-made checks, writing and running your own is a great idea. (Use the checks listed in Kahn, et al!)
- Visualizations make errors easier to spot.

Training and workforce development.

- Excellent free training on the OMOP/OHDSI data model at EHDEN academy.
- i2b2 has an annual conference, which is a great place to interact with the community.

What steps can hubs take to address these issues and improve data quality?

Know the limitations of your data.
Once a problem is reported, you must answer a fundamental question:

Are the data “wrong”? Or do the data reflect the source?

If the data are wrong, an error occurred somewhere in transformation. You should retrace your ETL steps (by looking at the code) to find the step where failure occurred. *These problems are fixable.*

If the data reflect the source, we have the classic problem of “garbage in, garbage out.” *These problems are harder to fix.*

What steps can hubs take to address these issues and improve data quality?

The data are wrong.
You have mis-mapped your units of measure during transformation.

VISIT_ID	HEIGHT	HEIGHT_UNIT
34547	60	CM

Source data

VISIT_ID	HEIGHT	HEIGHT_UNIT
34547	60	IN

Transformed data

The data reflect the source.
The clinician thought she was entering centimeters, but the EHR was set to inches.

VISIT_ID	HEIGHT	HEIGHT_UNIT
34547	60	IN

Source data

VISIT_ID	HEIGHT	HEIGHT_UNIT
34547	60	IN

Transformed data

How can institutional governance address data quality?

Build in maintenance time/effort for infrastructure projects.

- Research warehouses are not “set it and forget it.”
- Maintenance is not just break/fix, but also updates.

Communicate to stakeholders about data quality.

- Not every study is equally well suited to use EHR data.
- Investigators should be made aware of known issues with data required for their study, preferably at time of approval.
- Encourage data SMEs to be a part of governance/request approval committees.

How can institutional governance address data quality?

Prioritize data quality improvements based on use cases.

- It is usually not feasible to address all data quality issues at once.
- Help technical staff by clearly prioritizing data quality fixes based on pressing use cases.
    - *Ex:* Admission date is frequently null, when we know it should not be.
        - This is a commonly requested and important variable; this is a high-priority fix.
    - *Ex:* Free-text flowsheet values are coming through ETL with odd ASCII characters.
        - Probably lower priority.

Takeaway Points from Webcast

- Data quality is a journey, not a destination.
- There are several common types of data quality errors that frequently pop up in research data warehouses.
- Periodic auditing and getting involved with a data model community are both important tools to improve quality.
- Institutional governance can improve data quality by giving technical staff protected time to concentrate on data quality, and by prioritizing the most important data quality issues.

What's Next…

- We encourage you to view these webcasts as they provide foundational information for the rest of the series:
  - Introduction to Insights to Inspire 2021
  - Language of Informatics
  - Introduction to Informatics
  - Introduction to Maturity Models

- After that, please view the remaining webcasts in the order of your choice by searching CLIC_CTSA on Vimeo or YouTube

Thank you for viewing
Infrastructure & Data Quality