Emily Pfaff, PhD, MS
Research Assistant Professor, Dept. of Medicine
Co-Director, CTSA Informatics and Data Science Core
University of North Carolina, Chapel Hill

Session 6: Infrastructure and Data Quality

Welcome to the Insights to Inspire 2021 Informatics, The Journey to Interoperability webcast. I'm Emily Pfaff, Research Assistant Professor in the Department of Medicine at University of North Carolina, Chapel Hill. I co-direct our CTSA Informatics and Data Science core. And I'm excited to talk to you today about infrastructure and data quality in the context of CTSA informatics.

Objective - I'll be discussing how infrastructure, informatics expertise and smart governance practices can help to improve data quality at CTSA hubs. I'll be focusing on research data warehouses that contain EHR (electronic health record) data, but many of these practices are also applicable to other kinds of data collected and managed by CTSAs. In this session, I'll be detailing some common examples of data quality issues that we see in research data warehouses. I'll go over some steps that CTSA hubs can take to address and improve data quality. I'll also touch on the role of institutional governance in addressing and improving data quality.

What are some common examples of data quality issues in research data warehouses? - We transform data from one form to another to make it more useful, to simplify it, to restructure it, to optimize query performance, et cetera. However, each time data are transformed from one form to another, there is a chance that data quality may degrade The example here, we've taken this photograph and transformed it to make it smaller in size, but in doing so, we've degraded the quality of the image. In the next few slides, I'll walk through some common examples of how similar degradation can happen with data. Before I begin, I'll note that I'll be covering this topic in a brief and relatively informal way, but if you'd like a more thorough treatment of the subject, I highly recommend the paper linked on this slide by Khan et al 2016.

Mapping Errors - Our first common cause of data quality issues is errors in mapping. This can happen in a few different ways, which I'll illustrate with some fictitious, but realistic examples. You'll see an example of simple human error in the source data in purple, we have an ambulatory visit. However, during manual vocabulary mapping our extract transform load developer, mis-mapped ambulatory to ambulance a very similar word with very different meaning on the right. We have an example of content knowledge error. In this case, our well-meaning mapper was a bit overzealous when

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.

consolidating labs from the source data and mapped, both serum creatinine, and urine creatinine to the same concept in transformed data. Both of these errors will result in incorrect data in the transformed version.

Missing Data - Data transformation can also exacerbate the problem of missing data. It's important to inventory, which data domains and elements are and are not ETL'd (extracted, transformed, and loaded) into your research data warehouse so that you have a good idea of what is present in your source data, but missing in your warehouse right off the bat. Data that exists in your source system only as PDFs, as can be the case with many genetic tests, results are definitely candidates to be missing data in your warehouse. Likewise data that comes from a separate source system with a different refresh cycle like state death index data may appear to be missing in your warehouse, even if it's really just lagging behind. You may also find that certain individual variables are frequently missing in your data. BMI is a common example, but there are many others. Your ETL developers and informatics team are generally in the best position to understand what's missing and why it's missing. And it's really important that that information gets communicated to end users of the data. In many cases, missing data is not wrong and does not need to be fixed, but definitely requires explanation.

Granularity changes - Another transformation related issue involves granularity changes between the source data and the target database. In the example, we have a long list of discharge dispositions and the source data that are aggregated or rolled up to less granular categories in the target database. This aggregation simplifies the data and may make it easier to work with, but it has the downside that the target data can no longer be used to answer certain questions such as how many patients were discharged to a skilled nursing facility. Moreover, once aggregated, it can be difficult to trace back from the rolled up data to this source concept, if you need it, unless you have the transformation code at your disposal. Now this is not at all to say that aggregation should not be used rather it's important for data warehouse designers to have a good understanding of the use cases for the data. So that aggregations don't unintentionally obscure important data points.

Loss of Context – Loss of Context can happen when important data elements get dropped in the course of transformation. In the example, on the left, the diagnosis type variable is dropped between the source and transformed data, perhaps out of a desire to simplify the data. However, diagnosis type is an important piece of information for many use cases. And without it there, billing diagnosis will be seen as equivalent to patient reported diagnoses in the target database. The example on the right is perhaps a bit worse, dropping the benign sounding "status field" on this billing data between source and target means that voided charge lines will appear to be the same as billed charges. This could result in some pretty scary math errors. Now that you've seen some common data quality issues in research data warehouses, let's take a look at what steps can be taken to address them.

What steps can hubs take the address these issues and improve data quality? - Run Periodic quality audits - make it a priority to run periodic quality audits. Such checks often fall by the wayside when your team is already spread thin, but running these checks as an upfront investment to save time rather than chasing errors when they're discovered downstream by end users. Some of the common data models in use by many CTSAs for PCORnet and OHDSI/OMOP have pre-scripted data quality checks that run against their models and produce a really helpful report. If you don't have one of these models, writing your own set of test scripts, using some of the checks detailed in the Kahn paper is a smart idea, regardless of what checks you use note that reports formatted as graphs or charts make errors much easier to spot than looking at raw data. **Training and workforce development -** There are also some training resources you and your team can use to improve understanding of your data model or models, if you have multiple. I've listed some examples here, but even if you don't have one of these two data models, the important thing is to take a deep dive into the structure and documentation for your particular model and get to know it really well. This knowledge will be an immense help when troubleshooting data.

Trouble shooting Data - No matter how careful you are, it is almost inevitable that data quality issues will be discovered by end-users from time to time. When that happens, the first question to ask is always*: Are the data wrong or do they reflect the source?* Data that are wrong, got mistranslated somewhere during transformation. To solve such problems, you'll have to retrace your ETL steps through the transformation code, define where things got garbled. These are fixable problems. *If the data reflect the source?* If the data reflect the source, however, the problem is much harder to fix. We'll see an example on the next slide on the left, you'll see that the data are wrong. The unit of measure was changed on the way from source to target making a 60 centimeter long infant, 60 inches long. Instead in the example on the right, however, the clinician entering the data actually documented the wrong unit in the EHR itself. Now our 60-inch long infant is part of the legal medical record and not a transformation error rather than change the source data in the data warehouse. This type of error might be one where you let users of the data decide how they want to handle such outliers. In these cases, the hands-off approach is perhaps the safest lest you change data in unintended ways. We've seen some steps that CTSA hubs can use to address data quality issues. Let's now turn to how institutional governance can contribute to improve data quality.

Institutional Governance - One of the most important things for governance and leadership to recognize is that maintenance time and effort must be built into the scope of work for data warehouse builds. It is a huge amount of effort to stand up a new research data warehouse, but that effort does not end when you finally go to production. There is significant maintenance effort involved, not just to fix things that break. Rather proactive maintenance is necessary such as updating ETLs, or vocabularies to keep up with new codes, new flow sheets or data collection practices. Building in enough time and effort in budgets and staffing plans to take care of these tasks is essential. Institutional governance can also take a leadership role in communicating to end users about data quality. Not every research question is equally well suited to using EHR data. Some are not well suited at all. Without some kind of project vetting process in place, investigators may end up trying to use EHR data for inappropriate questions and get frustrated as a result. Note that technical should be encouraged to be part of this project vetting in communication pipeline. They are very well-positioned to understand what questions can and cannot be answered by the data in your warehouse. Involving them in the governance process also presents leadership opportunities for your technical staff.

There are situations where governance can get a bit too involved with data quality and overwhelm the technical staff with laundry lists of things to improve. It is much more realistic to address issues one or a few at a time in priority order. Governance should play a role in setting the priority order, considering factors like the importance of the variable to the largest number of use cases. Data that are used very infrequently should still be on the to-do list, but should not take precedent over data used in say 80% of requests.

Takeaway points:

- 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 the 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.

We encourage you to view the following webcasts as they provide foundational information for the rest of the series.

Session 1 – Introduction to Insights to Inspire
Session 2 – Language of Informatics
Session 3 – Introduction to Informatics
Session 4 – Introduction to Maturity Models
Session 5 – Importance of Interoperability

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.

After that, please view the remaining webcasts in the order of your choice. The webcasts are available on the CLIC website: https://clic-ctsa.org/ or you can access them by searching CLIC_CTSA on Vimeo or YouTube.

Thank you for viewing Session 6 **– *Infrastructure and Data Quality***.
Please join us for the rest of the ***Informatics: The Journey to Interoperability series***.

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.