Paper RE09

You Can't Just Cut the Gordian Knot: Approaches to Real-World Data Quality

S. Robert Collins, SAS Institute, Cary, NC

ABSTRACT

Data quality issues prevent a broader use of real-world data (RWD). We begin with an examination of factors that impact RWD in the health care setting. Next, we discuss how methods for extracting and transmitting data can impact data quality. Finally, we look at processes that preserve data fidelity while integrating sources and creating analytic data. Privacy issues complicate matters even further with varying standards between locations. We present an approach to evaluating and addressing RWD quality that begins with more traditional approaches and then look at ways to apply AI to these processes and examine the potential to use synthetic data to enhance privacy while also addressing data issues and bias. In the process, we will consider the challenges and benefits of integrating proprietary and open-source tools to address these issues while ensuring governance, transparency and compliance to standards while aligning with the goals of other RWD efforts.

BACKGROUND

Ancient Greece's version of "The Sword in the Stone" held that whoever could untie the Gordian knot would rule all of Asia. Legend says that, rather than trying to untie the knot, Alexander the Great simply sliced it open with his sword. Unfortunately, there is no simple shortcut that can solve the puzzle of RWD data quality - we must methodically determine how the knot was formed so that we can attempt to tease it apart.

Real-world data sources do not benefit from the same careful processes that have been developed and standardized to ensure the completeness and correctness of randomized controlled trials. This includes factors such as careful design of case report forms, standardized vocabulary and data dictionaries, site training, data monitoring and queries and more.

As researchers, we are typically unable to control the processes used to capture and record the source data or even the processes to extract and transfer the RWD. Our efforts in receiving and integrating the data and preparing data for analysis must be informed by a knowledge of potential issues upstream from our influence while ensuring our actions are transparent, logical and validated.

ISSUES AT DATA CAPTURE

HEALTHCARE SETTING

There are numerous factors that affect the quality of data at the time it is captured in a healthcare setting. The first and most obvious issues are human error in recording values accurately. A common example is transposing numbers during entry such as entering a height of 157.3cm when the measured value was 175.3cm. The wrong value in this case is subtle enough to pass unrecognized. Entering data using the wrong units can also cause problems, e.g., degrees Fahrenheit vs. degrees Celsius. Additionally, there may be differences in the significance of values recorded.

Terminology can also vary between individuals, departments and institutions. This is controlled in RCTs by imposing standardization that is not possible in the real-world setting.

A site may choose to implement free text instead of coded values for certain elements. This can exacerbate the standardization of entries across different users of the system.

Sometimes, data elements are missing. The problem with missing data in RWD is there is no indication for the reason the data element is not recorded. When I was collecting research data in an academic medical center, we defined and noted missing values as "not applicable," "technically inadequate," or "truly missing." "Not applicable" was for values that should not exist such as occupation for a pediatric patient. "Technically inadequate" was for when there were issues with the equipment or patient while a test was being performed. "Truly missing" was used for data where no value was collected but should have been such as when a patient's height was recorded but not weight was entered.

Unfortunately, missing data are often not "missing at random" and may represent some weakness in the process. Sometimes, it may be as simple as inter-observer differences in the way two clinical resources within the same institution record the same types of events. Even the same individual may skip some items in some circumstances and not others. Again, the standardization of data collection processes in the urgency of front-line patient care is different than that for RCTs.

Unfortunately, the source of missing values can introduce bias in the data. There is an entire body of literature on this topic and ways to identify the type of missingness and how to address it but we will return to this topic later.

ELECTRONIC HEALTH RECORD SYSTEMS

In medical records – manual or electronic, there is often no process to update older information as new information becomes available. Even something as simple as a patient intake history and physical exam could potentially include inaccurate historical information such as medications or previous diagnoses. Even if the newer data are correct and accurate, looking only at older vital signs will only reflect a snapshot of that point in time and may not be reflective of the current state. Obvious examples include age, weight and other factors that are expected to change over time. Other factors may be more subtle. But this can impact a retrospective study looking for pre-existing comorbidities that may be omitted or incorrectly recorded.

A similar issue is what happens when a new EHR is implemented? Are the older data migrated in to the new system and, if so, what is the process? This can introduce gaps or inconsistencies in the historical data trying to map data stored in the previous system to the way data are stored in the newer system.

Note that each instance of an EHR is typically customized to the institution. Different implementations of the same vendor's EHR will not be the same due to institution-specific choices. To further complicate matters, a single institution may multiple EHR systems or different implementations of the same system for different services, e.g., oncology vs. cardiology.

EXTRACTING AND TRANSMITTING DATA

When retrieving data from an EHR system, care must be taken to ensure the correct records are extracted. In addition, it is important to ensure complete records are extracted. If data are mistakenly censored during the extract process, it may be impossible to tell on the receiving end. Examples of this censoring could include a time range inadvertently eliminating records that should be included. Additionally, all of the relevant domains should be included in the extract. For instance, diagnosis codes could appear in several different areas of the EHR and if one of those sources is omitted, then the data for patients may be incomplete.

Unstructured data presents another concern for extracting data from an EHR. While free text data such as physician's notes are known to contain important information about the health status and treatment of patients, these are often omitted due to privacy concerns. A simple example of how free text can "leak" PII would be a note that includes text such as "Mr. Smith indicates…" Another example would be imaging data that includes labels with the patient's identifiers. Some data providers will work to identify specific types of information from unstructured data and implement processes to create structured data of those results. In such a case, the data provider should document the methods applied.

Finally, data providers typically offer specific formats for the data they provide. Historically, this has been delimited text files but bulk transfers of data via the cloud are becoming more common along with FHIR and other methods. If delimited text files are used, vertical bar or similar separators are highly recommended over comma- or tab-separated data. Recipients should be sure they request data in the format(s) that will ensure their success.

RECEIVING AND INTEGRATING DATA

As mentioned at the opening, researchers typically are not able to influence the steps we've discussed to this point. Once a transfer is received from the data provider is where the recipient's real work begins.

As a first step, any documentation received with the data transfer should be reviewed. This will become the touchstone against the received data will be measured.

If the data were received as delimited text files, it's worth a quick glance to see if the data appear as expected. If possible, check received row counts, patient counts and similar summaries against the transmission documents. Unix command line utilities are excellent at performing many of these activities on text data before the data are loaded into the target environment.

As data are loaded into the environment where they will be used by the researchers, it is critical to again check received row counts, patients counts and rows of look-up tables against the transmission documents. Before any attempt is made to use the data, you must next confirm for each table the numbers of data columns, column names, data types, values and ranges and other parameters of the data. It is amazingly common for documentation to list column names that don't actually match the received data.

It is also not uncommon for field values to not match the documented values. This could be as simple as issues with casing or result from categorical elements being exported as numeric instead of formatted text values. The latter can sometimes be seen with variables such as RACE, GENDER and others. It is recommended to also check for consistent casing of transmitted data – "YES" and "Yes" are not necessarily equal in your code. Especially if you are receiving text data, confirm separators, quotation marks, special characters. Quotations in quoted strings are a common source of issues as are commas embedded in text if the data are comma-separated. This is why vertical-bar delimited data or similar field separators are preferred over commas or tabs.

Once the data are loaded into the environment, additional checks should be applied to ensure the transmission was received and interpreted correctly. Exploratory data analyses can help ensure the reasonableness of the loaded data.

Note that SQL queries are a great way to quickly summarize the data once it has been loaded. Nested queries can perform grouped summaries without requiring the data to be sorted which can result in significant gains in performance. An excellent example of this would be to confirm row counts or issues in these types of situations:

- Confirm a table only has one row per unique patient ID.
- Summarize how many patients each have 1, 2, 3, etc., records in a table (including looking for outliers).
- Checking coded values match the documentation and obtaining a frequency count of each value.

Any discrepancies should be investigated, documented and, if necessary, resolved before proceeding to apply the data for analyses.

PREPARING ANALYTIC DATA

When preparing data for analysis, similar patterns apply to RWD as are used with RCTs. Use care to write clear, documented code and look for ways to reuse code that is known to be reliable for frequent operations.

When selecting records from a larger set or performing SQL joins, it is often beneficial to keep both the records of interest and the ones you omit so that you can ensure the process completed as expected. SQL queries on the results can, again, be useful here. The omitted records can be discarded after your quality checks are completed to conserve storage.

Watch for unobserved categories in subsetted data, especially with SQL joins. As an example, if the area you are working on can have values of “North,” “East,” “West,” and “South” but your subset does not contain any observations for “South,” you may want to do an outer join with the distinct categories to ensure all four values appear in your summary list (with “South” having zero counts).

Handling missing data is, again, a well-published topic but there are some general recommendations. Avoid making blanket decisions on how to handle missing data. Instead, make all decisions at analysis time in coordination with the statisticians and other stakeholders. In some cases, it may be appropriate to omit or impute the individual missing values. In other cases, it may be permissible to omit the observation from analyses. Omitting the entire patient can be justified in certain circumstances but must be carefully considered for the potential impact on the analysis.

MITIGATING RISKS

Remember that there are many factors that can influence the quality of real-world data sources. These issues are usually addressed in RCT data by established, validated processes. Unfortunately, those processes are not applicable to RWD.

For data source issues, researchers can only react. Fortunately, with the heightened interest in RWD, additional attention and efforts are coming to bear on methods to ensure quality RWD. AI may be able to help, but any corrections need to be flagged and made in real-time, otherwise, they are just new potential data issues.

For data transmission and receipt, be sure to employ transparent, standardized, reusable processes including data summarization. This greatly simplifies subsequent transmissions for the same data source and will speed the process for new data sources.

When preparing analytic data, check and recheck work. Again, where applicable, the reuse of processes and code can reduce effort while improving results.

CONCLUSION

RWD come with a variety of potential data quality issues that cannot be resolved using the same methods as RCTs. Except in rare cases, RWD is anonymized meaning there is no way to “reach back” to confirm suspected data issues. Even if we could, there is no “golden record” to verify information is correct such as when comparing CRF data against medical charts in an RCT.

Watch for and participate in harmonization, guidance and methods efforts across stakeholders but, in the interim, carefully apply methods to the portion of the data lifecycle that you can impact to ensure minimizing any issues with RWD.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: S. Robert Collins
Company: SAS Institute
Email: robert.collins@sas.com
LinkedIn: https://www.linkedin.com/in/srobertcollins/
Work Phone: 919.531.0051
Website: https://www.sas.com