Elucidata

White Paper

Enhance biomedical insight generation by improving data quality

How do you know if your bio-molecular data is high quality?

It is an axiom that bad data is costly. This is as true for bio-molecular data as it is for anything else. Drug discovery teams often complain about 'bad' data but do not have the language to describe it.

Alfred

Bioinformatics scientist at a big Pharma company

Alfred has spent 2 weeks analyzing data from a single cell experiment only to find out later that the right controls were never in place. What an unproductive activity!

Susan

Biologist at a mid-size Pharma company

Susan is looking for a study done by her colleague who has left. She is unable to retrieve the data. No keywords work!

Elina

Bioinformatics scientist at an early-stage company

Elina is looking into public datasets for an indication of their interest. Unfortunately, the studies she found had not been executed at the right level of precision. Three rounds of preliminary data cleaning had gone waste. And she could not find another study that is useful. Something must be out there!

Each of these scenarios represent different 'Bad data problems.' Luckily, these are all solvable. Just what is 'bad data' in biomedical research? Read on to learn about a framework to describe data quality and how it helps scientists in overcoming the 'bad data' roadblocks.

Data quality decisions in the current day context

Before the Digital Revolution, data generation and consumption happened generally in the same place (Eg. a research laboratory/ institute). Researchers shared their inferences in the form of research papers, conference articles, reports, etc. Now, with the advancement of technology, the data itself is being shared with different users across the globe. This provides enormous potential in terms of the inferences derived from this data because now, it can be explored from a different perspective than that of the data producer. This paradigm shift requires us to change the way we think about data quality to meet the needs of a more diverse group of data stakeholders.

Data stakeholders can be broadly divided into four categories (Giannoccaro et al.):

Data producers

Generate, collect, or prepare data (e.g. biologists, clinicians, data scientists, biocurators)

Data custodians

Design, develop and maintain the data and infrastructure (e.g., data engineers, cloud engineers)

Data managers

Manage the entire life cycle of data (e.g. head of data science, head of data monitoring and management, chief data officer)

Data consumers

Use the data in their day-to-day work (e.g. data scientists, bioinformatics scientists, AI/ML researchers, biologists)

Although data quality is important for all stakeholders, the quality criteria differ for each segment (Giannoccaro et al.). Therefore, it is critical to develop a framework that delivers for each stakeholders' needs

Ensuring data quality: The need to look beyond FAIR principles

FAIR stands for **Findability, Accessibility, Interoperability, and Reusability of data.** The FAIR Guiding Principles is a prominent example of quality control methods for data management, as described by Wilkinson et al. While improving these aspects of data quality can ensure that individual datasets meet defined community standards and promote data reuse, the FAIR metrics will not be sufficient to ensure data quality for use cases requiring more stringent data standards such as machine learning (ML) applications. There are several research consortiums and communities that have focused on ensuring the quality of publicly available biomedical data. Despite many such efforts, there is a lack of a framework for defining overall data quality, thus contributing to the variability of these data quality control methods across domains, data types, and use cases. This variability makes it difficult to compare the quality of one dataset with another.

Two ways to look at data quality

Data quality traits can be categorized into two groups based on whether those traits are inherent to the data (intrinsic) or not (extrinsic).

Intrinsic Data Quality

Intrinsic data quality traits are inherent to the data. Improving them is often left to the producers of biomedical data such as the biologists or clinicians who conduct experiments. Generally, intrinsic data quality cannot be improved after the data has been generated. Data with high intrinsic quality are more likely to be suitable for many use cases. Implementing quality controls during sample preparation and measurement of biomedical data can significantly improve intrinsic data quality. The variables of intrinsic data quality are used as validation criteria to determine whether data is suitable for analyses.

Key contributors to intrinsic data quality:

Experiment Design

- Clearly defined experimental variables
- Sufficient number of samples to statistically isolate the effect of variable(s) of interest from confounding biological or technical factors
- Sufficient number of replicates (technical and biological)
- Controls (negative and positive)

Metadata Records

- Annotations on the biological system and the samples being studied
- Experimental factors, observational variables, and confounding factors
- Instruments and technology used molecules etc.

Measurement

- Use of appropriate technology platforms that have been designed to measure the features of interest at the desired resolution
- Stringent quality controls.
- The measurements should be dependable for downstream analysis

Extrinsic Data Quality

Extrinsic data quality is affected by systems and processes that interact with the data after it has been generated. Examples include all factors that do not determine the intrinsic quality of data. Improving extrinsic data quality is often the mandate of data custodians and data managers as it can be improved through data curation. High extrinsic data quality makes it easier for users to assess relevant data.

Key contributors to extrinsic data quality:

Standardization

- Consistent field names that contain a specific type of metadata
- Permissibility of the values in metadata fields. (Eg. usage of accepted ontologies)

Accuracy

- Correctness of the values present in a metadata field
- Correctness of measurements

Data integrity

- Alteration of metadata fields - accidentally/ maliciously modified/ destroyed
- Retention of all metadata provided by data generators
- Availability of all eligible data from source
- Inclusion of measurements from all samples in a dataset

Breadth

- Presence of essential metadata fields for most use cases
- Conformation of the metadata to information standards defined by the community

Completeness

- Availability of all relevant metadata fields

Data Quality management framework

This section outlines the key steps involved in (extrinsic) data quality management of different types of biomedical data as derived from the theory of Total Data Quality Management (TDQM).

Choose & Define	• Choose the most important dimensions instead of including a maximum number of dimensions • Define the context of each dimension clearly. (i.e., which information is being measured and its intended outcome)
Measure	• Measure the data quality based on each of the chosen dimensions • View data quality assessment as an iterative process • Build processes that allow continuous measurement and monitoring of key quality dimensions
Analyze	• Analyze the measurements to prioritize the fixing of different quality issues • Analyze the reason behind recurring quality issues and rectify them • Probe whether the solution would involve more visibility, automation, or a better review process • Prioritize improvement based on user impact
Improve	• Implement the improvement strategies; measure the change in data quality dimensions • Implement improvement strategies as close as possible to the source of the problem

Data Quality management framework

While intrinsic data quality needs to be assured while the data is being generated, the extrinsic data quality can mostly be improved by taking certain steps iteratively. We deliver high-quality data by working on these steps which fall under the umbrella of ‘data curation.’

Data cleaning and structuring	Data reprocessing	Metadata annotations
• Especially important if data is coming from several sources, of several types and/or generated by different laboratories • Focuses on cleaning the metadata/data and structuring them into a consistent format • Aids in automating detection, measurement, and correction of data quality issues	• Defined data processing pipelines remove measurement artefacts to scale and normalize the data, and to detect outliers/ mislabeled samples • Improve comparability across datasets by harmonizing measurements	• Infer additional metadata information through associated literature or data records such as lab notebooks or associated documents. • Metadata associated with experiment made explicit through annotations to aid evaluation of the experimental design quality

Are you future proofing your data?

We live in exciting times where we can draw actionable insights from enormous amounts of data. Whether you are a data producer, custodian, or consumer, you have the power to improve the quality of data in your realm. The trick is to think proactively! For more information on how to improve your data quality please contact Elucidata at info@elucidata.io.