5.2. Impact of regulations on Data Quality and CDM Practices

The evolving regulatory landscape is directly impacting CDM due to its increased focus on data privacy, lineage, processing and technology. The list below is not exhaustive but represents of a steady trend:

- FDA Guidance and EMA Reflection paper on Risk-based Monitoring (2013)
- FDA Guidance on Electronic Source Data in Clinical Investigations (2013)
- ICH E6 (R2) (Nov 2016)
- Chinese Reform on Leading PIs for Medical Device (Oct 2017)
- German Regulations on eCRF data review (January 2018)
- MHRA 'GXP' Data Integrity (March 2018)
- GDPR (Effective May 2018)
- FDA Use of Electronic Health Record Data in Clinical Investigations (July 2018)
- EMA Consultation on eSource Direct Data Capture (November 2018)

Even if details may differ across regulations, some fundamental principles like Data Privacy are typically well managed by CDM organizations. While not new to the industry, other concepts such as risk-based approaches are just being adopted by some CDM organizations. The traditional 100% cleaning and one-size-fits-all approach remains the most commonly applied method. In this section, we will focus on the regulatory changes that are prompting a reexamination of these established CDM practices.

a) How do regulations define Data Quality

Surprisingly, Data Quality is not universally understood within CDM (& beyond) and is often confused with Data Integrity. As stated in the latest MHRA guidance, data integrity is **not** data quality as "the controls required for integrity do not necessarily guarantee the quality of the data generated"[8].

It is somewhat easy to demonstrate and understand Data Integrity as many regulations such as 21 CFR Part 11 and the 2018 MHRA guidance on 'GXP Data Integrity' have explicitly defined data integrity. Data integrity is often associated with ALCOA which is defined as **A**ttributable, **L**egible, **C**ontemporaneous, **O**riginal and **A**ccurate. Now, let's use an extreme case to differentiate Data Integrity from Data Quality. Let's assume that some data entered in the eCRF can be:

- **A**ttributable as being entered by the site personnel and confirmed by the audit trail,
- **L**egible in the site source,
- **C**ontemporaneously collected at the time where the activity was performed at the site,
- **O**riginal to the source as confirmed by SDV and
- **A**ccurate (i.e. free from errors, complete and within ranges).

The data from the example above meets the ALCOA requirements demonstrating the core attributes of data integrity. But theoretically, it could have been collected from non-calibrated instruments or by non-medically qualified personnel. Data could have also been collected from sites not adhering to expected good research practices such as propagating the same non-critical data from visit to visit (e.g. copying same vital signs data from one visit to another instead of collecting vital signs at each visit). These scenarios are rare but have happened. The corresponding data would clearly not be considered quality data. They could lead to the exclusion of the site data in the Clinical Study Report, endangering the statistical power of the population and therefore negatively impact the study outcome. So, meeting the key criteria of Data Integrity (e.g. ALCOA) is not enough to ensure Data Quality.

Data Quality is somewhat more “subjective”. In 1999, the Institute of Medicine defined high-quality data *as “data strong enough to support conclusions and interpretations equivalent to those derived from error-free data*”. In 2016, ICH E6 (R2) focused on activities essential to ensuring the reliability of trial results with capabilities to distinguish between reliable and potentially unreliable data. In 2018, MHRA defined data quality as “the assurance that data produced is exactly what was intended to be produced and fit for its intended purpose. This incorporates ALCOA”[8]. All of those suggests that Data Quality is reached when data support the right decision-making (i.e. fit for purpose).

As a general guidance, we can define Quality vs. Integrity as follow:

Data Integrity means that the **Data is managed the right way**

Data Quality means that the **Data is credible and reliable**

Many CDM organizations have historically strived to achieve data integrity as a primary outcome. Processes have been designed to ensure that 100% of the expected data has been collected, missing data retrieved, inconsistent data cleaned, and external data reconciled (i.e. data validation). While Data Integrity is a mandatory attribute of data quality, reaching data credibility and reliability must become our priority (i.e. reach data quality though fit for purpose data reviews).

To set the direction and help in distinguishing between reliable and potentially unreliable data, ICH E6 (R2) is suggesting reviewing data differently to identify and evaluate:

- Data outliers and unexpected lack of variability,
- Data trends such as the range, consistency, and data variability within and across sites,
- Systematic or significant errors in data collection and reporting at a site or across sites,
- Potential data manipulation or data integrity problems

ICH E6 (R2) scope goes beyond traditional data cleaning processes and requires reviews combining patient data from all sources including but not limited to safety data, protocol deviations, audit trails, metadata and operational data. At the end of the day, the data need to reliably support the evaluation of the objectives set in the protocol.