5.2) The evolution of data reviews

Sometimes referred as data validation, data review is part of the overall study monitoring strategy. It should not be confused with, or limited to, on-site monitoring because it is much broader – it is the act of overseeing the clinical trial, not just the investigational sites.

ICH E6 is clear: the sponsor should determine the appropriate extent and nature of monitoring and should develop a systematic, prioritized, risk-based approach to monitoring clinical trials. The sponsor may choose on-site monitoring, a combination of on-site and centralized monitoring, or, where justified, centralized monitoring[7]. Clinical data review fits into that context. It is a remote evaluation of accumulating data, performed in a timely manner, supported by appropriately qualified and trained persons[7] (i.e., Clinical Data Managers).

The regulators have also noticeably shifted their thinking over the past few years from requiring consistent levels of quality across all data to focusing on critical data and ensuring **that the data produced is exactly what was intended to be produced** and **fit for its intended purpose**[6]**.**

Sponsors should heed the call to focus on what matters as they revamp their data review strategy. The scope of data review within a risk-based CDS study execution goes beyond patient data and includes the interrogation of the audit trails which contain precious information on how the protocol is being operationalized and the way in which data is being collected. This information is relevant to both the integrity and quality of the study data. As emphasized by regulators at the 2019 SCDM annual conference[10], sponsors need to leverage audit trail data during the quality control stage to ensure and be able to demonstrate the integrity of the data used to support product submissions. Audit trail review may be most critical with regards to third party data including eSource.

To be successful, CDS organizations must first leverage the right tools. Part 2[2] suggested two major technologies enabling the transformation of data reviews. First, CDM needs intelligent Clinical Data Management Systems (CDMS) to consolidate, interrogate and reconcile complex data streams. Second, embarking on the AI journey could help CDM move from traditional to supervised and actionable data reviews.

The summary below provides the list of core changes to expect in the context of the 5Vs of clinical data[2] (i.e., Volume, Variety, Velocity, Veracity and Value) assuming a risk-based CDM Framework as articulated in the previous section.

Change	Impact on CDM Role
Review of large datasets generated continuously (Volume & Velocity)	With the increased use of m-Health solutions including sensors and wearables, the volume and velocity of data is exploding. This means that it is no longer possible to use manual processes based on listings or patient profile to review such a large volume of disparate data. It is necessary to implement different strategies moving beyond data filtering and trending to strategies based on story telling visualizations, statistical and Machine Learning (ML) models as well as leveraging intelligent automations. Interrogating such data may require different technology expertise such as non-SQL.

Reviews of more data sources (Variety)	The number and complexity of sources including real world data (RWD) and those coming from decentralized clinical trials (DCTs) makes it impossible to centrally manage them into technology solutions like EDC or traditional CDMS. Additionally, many of those sources do not comply with clinical research standards. For example, they may not be coded with the medical dictionary for regulatory activities (MedDRA) nor follow CDISC standards. This means that data reviews solely centered around EDC and edit checks are not comprehensive enough anymore. It also means that CDM needs to integrate different types of data such as sequenced data from sensors and data from electronic medical records (EMR). Some data are structured, others are not. CDS experts will also need to understand data standards beyond CDISC such as the fast healthcare interoperability resources (FHIR) standards, consider new technologies such as intelligent CDMS and leverage medical terminologies beyond MedDRA including the international classification of diseases (ICD) and the systematized nomenclature of medicine (SNOMED).
Reviews of data from studies with adaptive and/or master protocol designs (Variety)	According to the FDA, an adaptive design is one that allows for prospectively planned modifications to **one or more aspects of the study design** based on accumulating data from subjects in the trial. Patient populations, sample size, treatment arms, etc. could be adapted, as necessary[11]. Master protocols offer the opportunity to study **multiple IPs across multiple indications** which could potentially include adaptive design too. This means that static data review and reconciliation schemes would not work anymore. With evolving protocol requirements potentially including multiple indications, the data being captured could differ from patient to patient and even from visit to visit which is complicating the detection of missing data, procedures and visits. Additionally, variations in patient population characteristics may lead to a different focus in safety and efficacy reviews. To tailor data review strategies accordingly, Clinical Data Scientists must understand the downstream impact of protocol variations and amendments to determine the applicability of specific data review technologies. Additionally, they must pay attention to the contemporaneousness of the data as design adaptations are often only triggered if data is up to date. Finally, each adaptation inflection point may require database lock like strategies to ensure robust decision making.
Review of eSource and patient generated data (Variety)	Patient centric data collected from e-COA, m-Heath solutions, EMR, sensors and wearables are considered eSource. It is almost impossible to modify eSource data once it has been generated. This means that feedback on the data quality and integrity needs to be provided at the time of data generation. After data is generated, CDM will rarely be able to send a query to request a correction. So, data anomalies will be tagged and explained for the most part. Beyond data tagging, MHRA introduced the concept of "data exclusion" based on a "valid scientific justification, that the data are not representative of the quantity measured". Also, "all data (even if excluded) should be retained with the original data and be available for review in a format that allows the validity of the decision to exclude the data to be confirmed"[6].

Reviews of metadata such as audit trail (Veracity and Volume)	With more data being collected as eSource and more complex data streams, our traditional safety nets such as Source Data Verification (SDV), edit checks and manual listing reviews are no longer applicable. So, we need to consider new data review strategies leveraging metadata such as audit trails to ensure data validity. Unfortunately, audit trail format is not standardized across technologies and only a few technologies such as EDC typically export audit trail through CDISC ODM. This means that custom data integrations and reviews strategies need to be conducted. Additionally, the volume of audit trails will impact data integration and review strategies. Note that the e-Clinical Forum and SCDM will jointly publish an industry position paper on audit trail review later in 2020.
Centralized Data Reviews based on advanced trends and signals detection (Veracity and Value)	Historically the focus of CDM reviews was limited to the identification of missing, inconsistent and outlying data. ICH E6 (R2) expands the scope of data review to: (a) identify unexpected lack of variability and protocol deviations[7] (b) examine data trends such as the range, consistency, and variability of data within and across sites (c) evaluate for systematic or significant errors in data collection and reporting at a site or across sites; or potential data manipulation or data integrity problems (d) analyze site characteristics and performance metrics (e) select sites and/or processes for targeted on-site monitoring This requires advanced analytics solutions based on statistical and ML methodologies that will generate complex data trends and signals going beyond the scope of edit checks or straightforward data reconciliation tools. Those may detect propagated, fabricated and intentionally altered data (e.g., to falsify inclusion/exclusion criteria). Additionally, predictive algorithms may indicate the emergence of a risk to mitigate pro-actively. This means that Clinical Data Scientists will require a deeper knowledge of the end to end data flow to investigate signals highlighting atypical patient, site and country behaviors. Some might be indicative of a systematic process error, sloppiness or deliberate bias. Other could be false positives. As a result, Clinical Data Scientists need a comprehensive understanding of the clinical research processes and systems including those related to other internal and external stakeholders such as sites and patients.
Review of RWD (i.e., Curation of passive data) (Value)	Passive data refers to data generated as a by-product of real-world medical care processes or other patient activities[2]. This data is usually not collected for clinical research purposes but can be curated and utilized in research such as a synthetic control arm, for protocol optimization, as a benchmark, etc. Typically, this data is not modifiable, not anonymized at its source, not matching clinical research standards and scattered across multiple unmastered systems. This means that Clinical Data Scientists will need to **curate passive data** (i.e., anonymize, integrate, organize and assess the data collected from various RWD sources). They need to implement objective methodologies to confirm its integrity and quality to generate the appropriate secondary data assets and real word evidences (RWE) from RWD to be used in the context of clinical research.

From a practical standpoint, CDS competencies will need to align with the radical technology changes in order to support this major shift in the scope of data review. This includes the following:

- Managing intelligent CDMS
- Using new data interrogation techniques (e.g., non-SQL)
- Using analytic tools leveraging statistical methodologies
- Implementing robotic and intelligent process automations (RPA and IPA)
- Implementing intelligent solutions powered by AI methodologies such as ML

Considering all of these, we could compare the data review scope of the CDM vs. CDS as:

CDM Data Review Scope	CDS Data Review Scope
Focused on EDC	Focused on DCT technologies
Low volume of data and sources	High volume of data and sources
Simple data flows	Complex data flows
Focused on logical thinking (Output)	Focused on critical thinking (Outcome)
Standard processes across studies	Risk-based processes tailored for each study
Focused on data integrity	Focused on data quality (i.e., data reliability)
Data cleaning	Data review, tagging, exclusion and curation
Clinical research data	Clinical research and healthcare data
Traditional programming (SQL, C#, SAS, etc.)	ML (Python, R, etc.), non-SQL

This evolution would require the following roles requirements to support new data review approaches:

Best Practices	Soft Skills
• Risk-based study execution • KRIs and QTLs life cycle • Story telling visualizations • Audit trail reviews • Data tagging, exclusion & curation	• Critical thinking • Ability to understand complex data flows
Competencies	**Foundational Knowledge**
• Advanced analytics • Advanced data interrogation methods • ML methodologies	• New research methodology (adaptive, master protocols) • Decentralized clinical trials approaches & technologies • Risk-based methodologies and regulations • Understanding of new concepts such as sequenced data, unstructured data, data mining, ML, etc.