SCDM
Society for Clinical Data Management
DATA DRIVEN

The evolution of Clinical Data Review

Version 1 (April 2022)

Authors

- Sanjay Bhardwaj, Head of Clinical Technology Strategy & Operations, Abbvie
- Patrick Nadolny, Global Head, Clinical Data Management, Sanofi

Reviewers and Contributors

- Joe Kim, Senior Delivery Principal, Slalom
- Jérémy Nadolny, Sr Business Associate, Trial Operations
- Demetris Zambas, Global Head, Data Monitoring & Management, Pfizer

The Society for Clinical Data Management (SCDM) would like to acknowledge the many volunteers who have contributed to the development of the SCDM reflection papers used as the primary basis for this topic brief.

Methodology

The SCDM Innovation Committee seeks to provide Thought-Leadership to our industry and support the SCDM vision of "*leading innovative clinical data science to advance global health research and development*". To that end, the SCDM Innovation Committee strives to demystify Clinical Data Science (CDS) and support the development of all Clinical Data Management (CDM) professionals, from subject matter experts (SMEs) working on clinical studies to CDM leaders setting the direction of their organizations.

The SCDM Innovation Committee is publishing topic briefs intended to serve as orientation guides on specific areas which are contributing directly or indirectly to the evolution of CDM into CDS. The content of those topic briefs is primarily an extract from the previously published SCDM Reflection Papers[1,2,3] which collectively provide a cohesive and comprehensive overview of CDS from the point of view of industry leaders. Due to the recent emergence of the CDS discipline and the absence of a comprehensive literature base regarding CDS within the Drug Development industry, this content was gathered from industry leaders through a consensus-based methodology. As CDS mature, it is anticipated that literature on this topic will blossom.

Copyright © 2022 Society for Clinical Data Management. All rights reserved

Introduction

Sometimes referred as data validation, clinical 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[4]. 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[4] (i.e., Clinical Data Managers).

Additionally, the regulators have also noticeably shifted their thinking over the past few years from expecting consistent levels of integrity across all data to focusing on what matters (e.g., critical data and processes) ensuring reliability of trials results (i.e., data quality). It means expanding the focus to be solely on data integrity by adding the data quality dimension. But without a doubt, while "the controls required for [data] integrity do not necessarily guarantee the quality of the data generated" [5], data integrity remains core and is expected to reach data quality. According to MHRA, data quality is "the assurance that data produced is exactly what was intended to be produced and fit for its intended purpose"[5]. Quality data also reflects the reality of what happened to the patients (e.g., The patient's blood pressure was indeed 132 over 83, the patient truly experienced an injection side reaction, etc.). ICH E6 (R2)[4] goes beyond integrity as well by expecting the ability to distinguish between reliable and potentially unreliable data.

In such a context, it would no longer make sense to continue applying a consistent level of data review across all data. By equally focusing attention to all data the same way, we cannot focus on what matters which really means that we are increasing the risk of not identifying data quality gaps on critical data during our data reviews.

It is also critical to realize that in some cases, it is possible that data integrity is reached for some data streams but not all. However, data quality can only be reached when all data streams together demonstrate the **credibility and reliability** of the trial results (i.e., outcome focused).

Last but not least, 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.

Topic brief

Sponsors should heed the call to focus on what matters as they revamp their data review strategy. This topic brief intends to highlight opportunities resulting from the evolution of our discipline in the context of the 5Vs of clinical data[2] (i.e., Volume, Variety, Velocity, Veracity and Value) as well as a risk-based CDM framework as articulated in the third reflection paper[3]. Last, we'll briefly explore the potential of emerging automation technologies.

The impact of the 5Vs to Clinical Data Review

A - Volume & Velocity

Review of large datasets generated continuously

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.

Being able to assess the individual data source's volume and velocity in tandem will directly inform the Clinical Data Scientist on what is the optimal approach to reviewing the data. High volumes will require automated solutions to assess the quality and integrity of the data as close as possible from its collection. It would also be advisable to consider implementing a study specific data risk assessment to identify these across all data sources and ensures specific mitigations are in place prior to enrolling the first patients. Additionally, high velocity data sources will require new approaches that drive action by detecting and promptly differentiating signals from background noise. CDS organizations must be able to develop processes and controls to identify appropriate signals when managing high volume and velocity data.

B – Variety

Reviews of more data sources

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 complex data 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

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[6]. Master protocols offer the opportunity to study **multiple IPs across multiple indications,** combining multiple traditional study phases together (e.g., Phase I & Phase IIa) which could potentially be integrating adaptive design components too.

This means that static and one-size-fits-all data review and reconciliation schemes would not work anymore. With the growing adoption of master protocols such as basket designs evaluating 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 mid-study adaptations to determine the applicability of specific data review technologies and adapt their data review plans specifically to the study variations. Finally, each adaptation inflection point may require "database lock" like strategies to ensure robust decision making.

Review of eSource and patient generated data

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 could mainly 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"[5].

C - Veracity and Value

While the veracity of the clinical data will largely be supported and controlled by the underlying technology, there will be instances where the sponsor's technology does not directly control the credibility and integrity of the data. Situations where the sponsor's CRO and/or sites are managing the underlying technology or the source of the data (e.g., EMR) which cannot be corrected requires the Clinical Data Scientist to be able to think critically about how to assess the credibility and integrity of the data. Equally important, the Clinical Data Scientist must effectively partner with the external parties to understand their controls, define what additional ones will be required and ultimately document all measures undertaken as well as the data flow and any transformations demonstrating the end-to-end chain of custody from data creation to analysis and reporting (e.g., in the Data Management Plan).

Reviews of metadata such as audit trail (Veracity)

With more data being collected as eSource and more complex data streams, our traditional safety nets such as Source Data Verification (SDV) and manual listing reviews are no longer applicable. Edit checks would not apply to data stream such as sensors and wearables. However, their criticality at the point of entry would become ever more important for site based direct data capture and ePROs. If an error at the point of initial entry is not identified, Clinical Data Scientist will not be able to request a correction as there is no other source to relate to. It is therefore not surprising that sponsors are expected by regulator to have additional procedures for risk-based routine audit trail reviews (ATRs). Without ATR, sponsors can miss important non-compliance issues that could only be evidenced through metadata. In addition, it is likely that GCP inspectors will be enquiring more about these procedures as well as requesting access to audit trail data during inspections. During the 2019 SCDM Conference, regulators shared issues identified by audit trail and metadata review during GCP inspections. Reviews by GCP inspectors revealed that in some cases:

- Data entry was not performed by authorized individuals
- ePRO data from an investigator led questionnaire was entered within an unreasonably short period of time: ePRO was entered in less time compared to the time it should take for the site personnel to 1) ask the questions to the patient and 2) enter their responses (i.e., data) in the ePRO system.
- Site entry occurring during the same time/day for a group of subjects.

- e-Source primary efficacy data was not entered directly into the application by sites per protocol (i.e., alternative source data existed). Additionally, discrepancies were detected when the inspectors retrieved the 'true' source data to verified ePRO data.

These issues could have easily been identified before the inspection with the right data review strategies. Such issues have the potential of delaying drug approval by months as inspection findings must be properly answered prior to approval. They can even question the reliability of trial results and lead to the exclusion of site data from statistical analysis.

So, we need to consider additional 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.

Fortunately, the e-Clinical Forum and SCDM Audit Trail Task Force to have published a position paper on ATR[7] providing process and technical considerations. The paper also provides example of use cases.

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) [4] expands the scope of data review to:

(a) identify unexpected lack of variability and protocol deviations

(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 be mitigated pro-actively. As predictions are by nature hypothesis on what could happen to data to be collected in the future, there would not be actual data to query nor to review. In such a case, the Clinical Data Scientist will need to investigate the root cause of the emerging signal leading to such a prediction. This means that Clinical Data Scientists 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. Others 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, as a Quality Tolerance

Limit (QTL) Threshold, 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.

Leverage the value of all industry standards to ease data review

Many companies still lack an end-to-end data standardization and integration strategy that considers all the dimensions of clinical data. It is critically important to understand that standards do not only apply to collection and transmission, but also to terminology and modeling. Failure to consider all data standardization dimension can result in process inefficiencies. Additionally, leveraging good standards not only facilitates the EDC and start-up process but also enables Clinical Data Review.

Standards can be classified in four layers providing synergetic values[8]:

Exchange standards: Schemas or file types for exchanging data between systems; the container for transporting data

Data model standards: Conceptual, logical or physical semantic descriptions of objects and their relationships

Metadata standards: Representations of individual data concepts useful for organizing data in a database or data exchange file

Terminology standards: Representations of individual data values

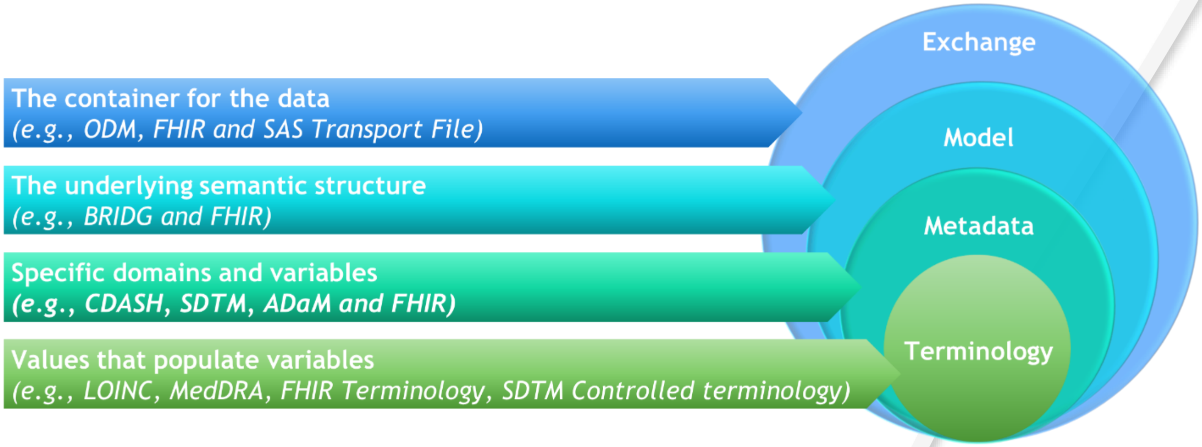
Exchange
Model
Metadata
Terminology
The container for the data
(e.g., ODM, FHIR and SAS Transport File)
The underlying semantic structure
(e.g., BRIDG and FHIR)
Specific domains and variables
(e.g., CDASH, SDTM, ADaM and FHIR)
Values that populate variables
(e.g., LOINC, MedDRA, FHIR Terminology, SDTM Controlled terminology)

Fig 1. Layers of clinical data standards

As advocated and championed by SCDM, it is also worth noting that healthcare standards can be leveraged in clinical research to ease data mapping, exchange, and data review. As an example, LOINC can be used to harmonize laboratory test terminology, therefore easing their consolidation and review.

However, the evolution of Clinical Data Review will not stop there.

Potential roadmap to the automation of Clinical Data Reviews

The growing maturity of automation technologies will impact its evolution further. Some CDM leaders foresee that AI will be driving the evolution of data review from traditional to supervised.

The evolution may include 4 distinct steps:

- **Stage #1 – Automation of traditional reviews:** Data trends and anomalies will continue to be identified via edit checks, listings, and dashboards. However, "simpler" and repetitive tasks such as query posting, and SAE Reconciliation will be automated with Robotic Process Automation (RPA).
- **Stage #2 – Actionable reviews:** In this context, systems identify data trends and anomalies via Machine Learning (ML) based automations and statistically based analytics to detect atypical data patterns[9]. The adoption of advanced analytics rose recently to support risk-based monitoring and has significant potential in reducing the need for complex and labor-intensive manual data review and trending.
- **Stage #3 – Guided reviews:** Guided review is the natural evolution of actionable cleaning. Once the automated detection of data trends and anomalies have matured, and when the volume of actions manually taken by CDS is meaningful, ML Tools will be able to learn from them and automate actions. In this first stage, CDS will review suggested actions prior to execution. Natural Language Generation (NLG) may also be implemented to automate the writing of queries.
- **Stage #4 – Supervised reviews:** In the end, systems will automatically detect and act. They will only escalate new scenarios to Clinical Data Scientists when they do not have enough knowledge to take a decision autonomously. In this context, Clinical Data Scientists will oversee the systems, support their training, and arbitrate complex data issues that systems cannot "judge". There is a high potential of introducing bias by training the initial algorithm on a "ground truths" not representative of all production scenarios. CDS will therefore need to monitor and supervise the entire ecosystem especially at the early stage of deployment and prevent the machine to learning from atypical scenarios, as well as manage privacy and ethical breaches.

As shown in figure 2, the value and more importantly the transformational impact of automation will grow as CDM evolves to CDS, from stage #1 to #4 automations. Stage #1 automation will mostly require very few the technology components. However, Stage #4 automation will be more complex and need the full power of all automation components combined (i.e., RPA, ML, AI & NLG).

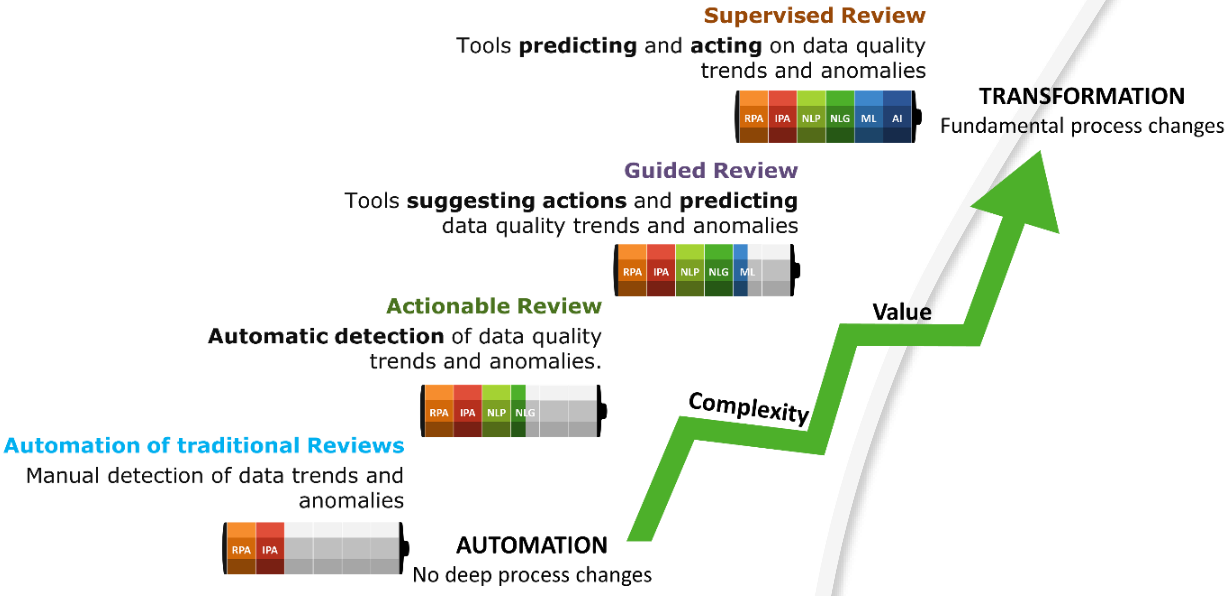
Supervised Review
Tools predicting and acting on data quality trends and anomalies
RPA
IPA
NLP
NLG
ML
AI
TRANSFORMATION
Fundamental process changes
Guided Review
Tools suggesting actions and predicting data quality trends and anomalies
RPA
IPA
NLP
NLG
ML
Actionable Review
Automatic detection of data quality trends and anomalies.
RPA
IPA
NLP
NLG
Value
Complexity
Automation of traditional Reviews
Manual detection of data trends and anomalies
RPA
IPA
AUTOMATION
No deep process changes

Fig 2. Evolution of data reviews powered by AI technologies

Conclusion

Considering all of these, we could summarize the evolution of data review scope from CDM to 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

Each CDM organization could pragmatically identify low hanging fruit opportunities based on their current landscape to initiate the modernization of their clinical data review strategy. From a practical standpoint, CDS competencies will ultimately need to align with the radical required changes mentioned in this topic brief in order to achieve this major shift in the scope of clinical data review.

References

[1] SCDM, June 2019, The Evolution of CDM to Clinical Data Science – A Reflection Paper on the impact of the Clinical Research industry trends on Clinical Data Management. Available at https://scdm.org/wp-content/uploads/2019/09/SCDM-Reflection-Paper-Evolution-to-Clinical-to-Data-Science.pdf

[2] SCDM, March 2020, The Evolution of CDM to Clinical Data Science (Part 2: The technology enablers) – A Reflection Paper on how technology will enable the evolution of Clinical Data Management to Clinical Data Science. Available at https://scdm.org/wp-content/uploads/2020/03/SCDM-Reflection-Paper-Evolution-to-Clinical-to-Data-Science-Part-2.pdf

[3] SCDM, August 2021, The Evolution of CDM to Clinical Data Science (Part 3: Evolution of the CDM role) – A Reflection Paper on the evolution of CDM skillsets and competencies. Available at https://scdm.org/wp-content/uploads/2020/08/SCDM-Reflection-Paper-CDM-Role-Evolution-Part-3-.pdf

[4] ICH, Integrated Addendum to ICH E6(R1): Guideline for Good Clinical Practice E6(R2). Available at https://database.ich.org/sites/default/files/E6_R2_Addendum.pdf

[5] MHRA, March 2018, 'GXP' Data Integrity Guidance and Definition, Available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/687246/MHRA_GxP_data_integrity_guide_March_edited_Final.pdf

[6] FDA, November 2019, Adaptive Design Clinical Trials for Drugs and Biologics Guidance for Industry Available at https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics

[7] SCDM and eCF, Audit Trail Review, An Industry Position Paper on the use of Audit Trail Review as a key tool to ensure data integrity Available at https://scdm.org/wp-content/uploads/2021/04/2021-eCF_SCDM-ATR-Industry-Position-Paper-Version-PR1-2.pdf

[8] SCDM Webinar, Shannon Labout, Beyond the Implementation Guides: Why and how to use Data Standards – when you don't *have to* – Available at https://www.pathlms.com/scdm/courses/13573

[9] Laura Trotta, and Al, 2019, Detection of atypical data in multicenter clinical trials using unsupervised statistical monitoring, Available at https://journals.sagepub.com/doi/abs/10.1177/1740774519862564?journalCode=ctja

Main abbreviations

AI	Artificial Intelligence
ATR	Audit Trail review
BRIDG	Biomedical Research Integrated Domain Group
CDASH	Clinical Data Acquisition Standards Harmonization
CDM	Clinical Data Management
CDS	Clinical Data Science
DCT	Decentralized Clinical Trials
EDC	Electronic Data Capture
EMR	Electronic Medical Record
ePRO	Electronic Patient Reported Outcome
FHIR	Fast Healthcare Interoperability Resources
HL7	Health Level Seven
IPA	Intelligent Process Automation
LOINC	Logical Observation Identifiers Names and Codes
MedDRA	Medical Dictionary for Regulatory Activities
ML	Machine Learning
MHRA	Medicines and Healthcare products Regulatory Agency
NLG	Natural Language Generation
NLP	Natural Language Processing
ODM	Operational Data Model
QTL	Quality Tolerance Limit
RPA	Robotic Process Automation
RWD	Real World Data
RWE	Real World Evidence
SCDM	Society for Clinical Data Management
SDTM	Standard Data Tabulation Model
SDV	Source Data Verification
SME	Subject Matter Expert
SNOMED	Systematized Nomenclature of Medicine