The Evolution of Clinical Data Management into Clinical Data Science (Part 3: The evolution of the CDM role)

A Reflection Paper on the evolution of CDM skillsets and competencies

31 Aug 2020

Society for Clinical Data Management

Our Vision

"*Leading innovative clinical data science to advance global health research and development"*

Our Mission

"Connect and inspire professionals managing global health data with global education, certification and advocacy"

- TABLE OF CONTENTS -

Contents

1. Foreword

In this paper, the Society for Clinical Data Management (SCDM) Innovation Committee seeks to build upon the two previous SCDM publications on the evolution of Clinical Data Management toward Clinical Data Science (Part 1[1] and Part 2[2]).

In Part 1, the SCDM Innovation Committee addressed the industry drivers contributing to the evolution of our discipline. In Part 2, the committee focused on the adoption of emerging technology required to enable Clinical Data Science (CDS). This third and last reflection paper on this series is providing insights on the evolution of Clinical Data Management (CDM) skillsets and competencies.

Since the release of the first two reflection papers, the world faced the unprecedented COVID-19 pandemic forcing organizations to reconsider many standard processes and clinical research approaches. In a span of few months, study teams had to evaluate and mitigate operational and scientific risks never faced before. This moved teams into a complex problem-solving situation requiring critical thinking and pragmatism to look beyond our traditional approaches.

One striking outcome is the increased adoption of decentralized research which the industry only reluctantly considered implementing before. Many companies quickly implemented some aspect of clinical trials decentralization and anticipate continuing once the COVID-19 pandemic is over. The result of a poll conducted during the SCDM webinar in response to the pandemic on the 14th of June 2020 indicated that 65% of companies leveraged telehealth during COVID-19 and 46% of them anticipated to scale-up the use of processes and systems enabling the decentralization of clinical trials moving forward. So, Decentralized Clinical Trials[1,2] are no longer a hypothetic future, it became a reality that many CDM organizations had to adapt to.

Additionally, 71% of respondents to our poll said they expected to scale-up risk-based CDM strategies, and 67% their centralized monitoring strategies, in the near future.

The path toward CDS is accelerating and the need to take urgent and decisive action has never been as critical. We hope that the three reflection papers will help all CDM professionals, from subject matter experts (SMEs) working on clinical studies to CDM leaders to better understand what CDS is and lead to the expansion of the scope of CDM by adding the data meaning and value dimensions (i.e., data is credible and reliable) and therefore contributing to the evolution of our discipline.

2. Abstract

The main objective of this paper is to provide insights on how CDM professionals who have successfully and passionately contributed to the credibility of CDM can evolve their skillsets and competencies to cope with the increasing complexities of clinical research. This demands novel approaches maximizing the potential of available technologies. In the context of this paper, we would define a **skill** as a learned ability and a **competency** as the capacity to successfully apply those skills to perform a specific task.

We will also explore the impact of this evolution on organizations and on operating business models. Even though Clinical Data Managers have been efficiently supporting clinical studies for over three decades now, defining their role still remains "*complex, as it encompasses walking the very fine line of the fundamental (old school) CDM mindset of data integrity as the highest maxim on the one hand, and the exponentially increasing landscape of technology and its potential opportunity on the other hand*"[3].

While recognizing the significance of what CDM has achieved thus far, it is important to reflect on our role's evolution and anticipate the rising needs highlighted in Part 1[1] while leveraging the technologies mentioned in Part 2[2]. Moving forward, Clinical Data Scientists must leverage their core CDM knowledge to best understand how to apply technology to drive process improvements. It is by combining their deep subject matter expertise with technical literacy that real improvements can be made.

Ultimately, the three reflection papers provide a comprehensive view of what CDS is and will help you create a future proof roadmap both for your organization and for your career.

3. Acknowledgements

Disclaimer: The views expressed in this reflection paper do not necessarily reflect those of the companies or entities the authors are employed by or affiliated with.

Lead authors:

- Patrick Nadolny, Global Head, CDM & Programming, Allergan, SCDM Innovation Committee Chair
- Sanjay Bhardwaj, Global Head, CDM, Abbvie, SCDM Board Vice-Chair
- Catherine Celingant, Sr Director, Data Monitoring & Management – Oncology, Pfizer
- Ian Shafer, Principal, PwC
- Inder Sachdeva, Portfolio Delivery Leader, Clinical Data Sciences & Safety, Cognizant

Content contributors and reviewers:

- Paige Pearson, Associate Director, Data Science, Abbvie
- Joshua Wilson, Executive Director, Biometrics, Strategic Technology Advancement, Syneos Health
- Mrunalini Jagtap, Data Management Lead, Singapore Institute for Clinical Sciences (SICS)
- Nechama Katan, Director, Data Science Lead, Pfizer
- Nicole Rodrigues, Sr Clinical Data Specialist, Medtronic
- Catherine Hall, Vice President, Product Strategy, endpoint
- Anita Walden, Associate Director, Oregon Health and Science University
- Kishore Kumar, PhD, Consilx Digital
- Inderbir Singh, Sr Director, Clinical Data Management, Biometrics at Syneos Health, Member of SCDM India Steering Committee
- Sarvesh Singh, VP, Data Management, ICON Clinical Research
- Deven Babre, Director, Life Sciences Digital Operation, Cognizant Technology Solutions India, Member of SCDM India Steering Committee

Reviewers:

- Linda King, Global eCOA Capability Lead, Astellas Pharma, immediate past SCDM Board Chair
- Demetris Zambas, Global Head, Data Monitoring & Management, Pfizer, Former SCDM Board Chair
- Michael Goedde, Global Head Data Operations, PRA, SCDM Board Chair

SCDM would also like to acknowledge the many volunteers who have participated in the SCDM Innovation Committee and have contributed to forming the thoughts expressed in this reflection paper.

4. Current state of the Clinical Data Management role

As the industry's leading CDM organization, the SCDM has created anchor points for our discipline such as the Good Clinical Data Management Practice[4] (GCDMP©) and the certification program for Clinical Data Managers[5]. Together, they have been defining for almost two decades, the expected competencies, foundational knowledge and best practices of today's CDM roles.

First published in 2000, the GCDMP© provides a reference for CDM organizations in their implementation of high quality CDM processes for paper and EDC based studies. Its twenty-eight (28) chapters guide Clinical Data Managers preparing for CDM training and education.

GCDMP Chapters	
Data Privacy	External Data Transfers
Data Management Plan	Patient-Reported Outcomes
Project Management for the Clinical Data Manager	CDM Presentation at Investigator Meetings
Vendor Selection and Management	Training
Data Management Standards in Clinical Research	Metrics in Clinical Data Management
Design and Development of DCIs	Assuring Data Quality
Edit Check Design Principles	Measuring Data Quality
EDC - Concept and Study Start-up	Data Storage
EDC - Conduct	Data Entry Processes
EDC - Study Closeout	Coding Dictionary Management & Maintenance
CRF Completion Guidelines	Safety Data Management and Reporting
CRF Printing and Vendor Selection	Serious Adverse Event Data Reconciliation
Database Validation, Programming & Standards	Database Closure
Laboratory Data Handling	Clinical Data Archiving

Fig 1. List of GCDMP Chapters

The SCDM certification program launched in 2004, identifies seventy (70) competencies organized into the eight (8) core domains (see figure 2). The number of competencies in the **Design, Project Management, Data Processing and Programming** domains represent over 85% of all those identified in the certification program which aligns by design well with the GCDMP© chapters.

Competency domain	# of competencies	% of total competencies
Design	21	*30.0%*
Project Management	16	*22.9%*
Data Processing	15	*21.4%*
Programming	8	*11.4%*
Testing	2	*2.85%*
Training	2	*2.85%*
Personnel Management	3	*4.3%*
Review	3	*4.3%*

Fig 2. List of SCDM Certification domains

The SCDM Task Analysis Survey conducted in September 2018 re-confirmed the relevance of all competencies included in the survey. At a high level, the competencies included in each of the eight (8) domains cover the following areas:

Study Design Identification and set-up of all data collection instruments (DCIs) such as EDC and eCOA, data handling and reporting tools leveraging clinical data standards. It also includes core CDM documents such as the Data Management Plan (DMP) and Case Report Form (CRF) Completion Guideline

Programming Creation of the required tools defined during study design. Scope includes programming of the eCRF (Screens and Edit Checks), reports, ad-hoc querying, data imports, transformations and extracts

Data Processing Data Lifecycle from collection to archival. Includes the collection, transfer, import, cleaning, coding, reconciliation and quality assessment of clinical study data

Testing Definition and execution of testing strategies for required tools

Training Ensuring understanding of CDM processes across the organization

Personnel Management Ensuring CDM staff oversight

Project Management Ensuring oversight of CDM activities from study initiation to study close-out including vendor management

Review Expert review of study and CDM deliverables

In addition to the competencies themselves, twenty-five (25) foundational knowledge topics have been confirmed by the 2015 and 2018 Task Analysis Surveys as necessary to the performance of the CDM competencies. Those include but are not limited to the topics listed below:

- Therapeutic development and clinical research fundamentals
- Scientific method
- Good Clinical Practices (GCP) and other guidance
- Software Development Life Cycle (SDLC) concepts
- Audit methodologies
- Project management fundamentals
- Basic statistical concepts
- Data and metadata models, standards and terminologies (incl. medical terminology)
- Workflow design, analysis, and control fundamentals

Last, beyond those competencies, foundational knowledge and best practices, the following are commonly expected **soft skills** considered as essential building blocks for Clinical Data Managers.

- Attention to details
- Logical thinking
- Adaptability
- Ability to articulate complex concepts to the trial teams
- Ability to investigate and troubleshoot complex data trends
- Ability to work with cross-functional teams

While there are some variations across companies, the CDM role framework below, based on core competency domains, foundational knowledge, best practices and soft skills represents the core expectations from Clinical Data Managers today.

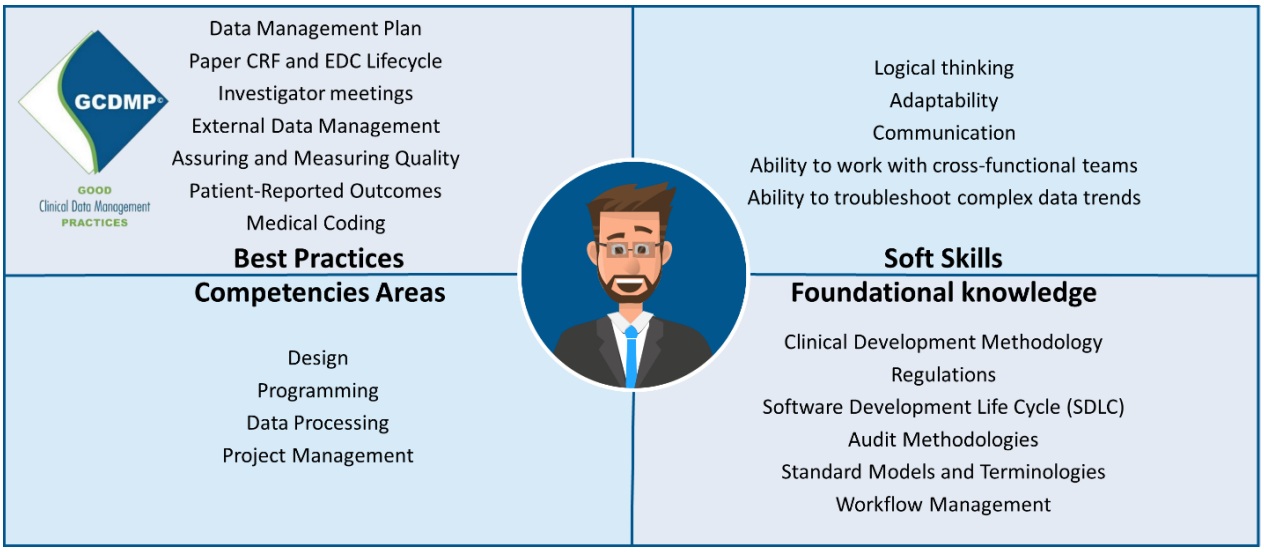


Fig 3. CDM role framework

In summary, even if some companies may have started to expand the scope of their CDM responsibilities, they are still revolving their activities around the following:

- The lifecycle of DCIs and Clinical Data Standards
- The end to end data flow from collection to archival
- The review and reconciliation of clinical study data
- The management of third-party data and related vendors
- One harmonized set of best practices (i.e., GCDMP© Chapters)
- One main CDM tool (i.e., EDC)
- The project management and documentation for all the responsibilities above

This SCDM framework has robustly anchored the CDM discipline for many years. While those will remain critical for years to come, the SCDM has initiated the journey toward CDS and is preparing our discipline to successfully support the evolving needs of clinical research. The following sections will expand on the two first reflection papers and address the impact of this evolution to CDM roles.

5. The role and skillsets of Clinical Data Scientists

As mentioned in Part 2[2] and reinforced in the section above, CDM is responsible for the lifecycle of clinical data from collection to delivery for statistical analysis in support of regulatory activities. CDM primarily focuses on data collection, data flow and data integrity (i.e., ensuring that data is managed the right way). CDS expands the scope of CDM by adding the data meaning and value dimensions (i.e., data is credible and reliable). CDS also requires the ability to generate knowledge and insights from clinical data to support clinical research which requires additional expertise, approaches and technologies.

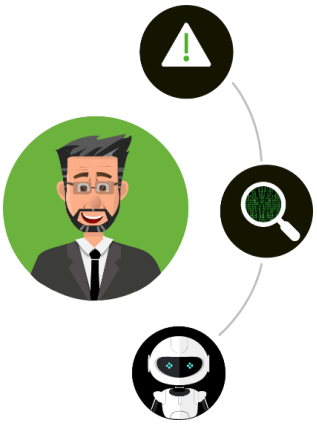
The first fundamental change in our CDS journey is the shift in focus from data integrity to data quality. But without a doubt, while "the controls required for [data] integrity do not necessarily guarantee the quality of the data generated" [6], data integrity remains core and is expected to reach data 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"[6]. 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)[7] goes beyond integrity as well by expecting the ability to distinguish between reliable and potentially unreliable data and by driving focus on critical data.

It is 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).

The second fundamental change is the end of the one-size-fit all approach based on **one** set of processes and **one** EDC centric data flow. As accelerated by the COVID-19 pandemic, truly adapting to patients by leveraging the capabilities at each site will generate study, country and site-specific data flows. It means that Clinical Data Scientists will have to drive the study team through all potential scenarios to optimize operational study execution while minimizing risks to patients' safety and reliability of the trials results. This represents a change **from logical to critical thinking** which is at the core of the role evolution.

Summarizing the insights from the previous two reflection papers[1,2], Clinical Data Scientists will need to deliver quality data and adapt to new concepts which are framed around three major themes explored in this section of the reflection paper and summarized in the CDS role evolution framework below.

CDS role evolution framework



Risk based CDM approaches aligned with new regulations focused on

1. Quality by Design (QbD)
2. Critical to Quality (CtQ) factors
3. Critical data and processes
4. Risks lifecycle management (Incl. Assessment, root cause analysis, etc.)

New ways of conducting **data reviews** to ensure **data quality** adapting to the

1. Decline of EDC centricity and increase in data variety
2. Decentralization of Clinical Trials
3. Focus on data reliability
4. Volume, Variety and Velocity of data and metadata
5. Oversight of increasingly complex and study specific data

Advanced CDS Competencies stemming from the evolution of **clinical research** and **technologies** supporting

1. New protocol designs such as adaptive and master protocol
2. The increasing use of Real-World Data (RWD)
3. The increasing reliability and affordability of m-Health solutions
4. The adoption of Artificial Intelligence (AI)

While the speed of change is overwhelming, the opportunity to re-shape clinical research is unprecedented. It is therefore crucial to act now and define a strategy enabling our Clinical Data Managers to evolve into Clinical Data Scientists fully equipped to embark on the CDS journey.

5.1) Risk-based CDM approaches

Over the past decade, regulators have issued several guidance documents such as ICH E6 (R2)[7] which define risk-based principles and advocate for the use of risk-based approaches. Embracing these methods, the industry has already successfully implemented risk-based approaches in the site monitoring and system validation spaces for several years. As a result, our traditionally risk-averse industry has become more comfortable with strategies that match efforts and focus commensurately to the risks.

Fig 4. CDS RBQM framework

We now have a meaningful opportunity to expand the use of risk-based quality management (RBQM) principles to encompass all study design and execution aspects within CDM. As articulated in Part 1[1], CDS must redefine its processes and roles to control risks to activities essential to ensuring human subject protection and the reliability of trial results[7]. Learning from the evolution of traditional to risk-based site monitoring, we must realize that this is a fundamental culture and role change. This means moving from a one-size-fits all process based on fixed standards to a new paradigm where quality is infused at the design stage to proactively prevent risks to arise as much as possible (see figure 4).

a) Study Quality by Design

ICH E6 (R2) dedicated a new section on quality management (section 5.0) focusing on risks lifecycle management. It covers the quality controls used from the identification to the reporting of the risks. But rather than controlling the risks by implementing mitigation and monitoring strategies, we should simply use QbD to avoid them in the first place which means we should start with the end in mind.

QbD stands on the assumptions that quality should be planned proactively and not be the act of retrospectively perform Quality Controls (QC). First, QbD must rely on the appropriate foundation including the company culture, policies, systems, processes and people. Next, it relies on optimal protocol design, a critical step with a huge influence on the ultimate success or failure of the trial.

So, consistent with the QbD principles, risks and mitigations should be identified prior to the protocol finalization. If possible, the protocol should be adjusted to prevent the risks. If de-risking the study protocol is not possible, the study team must implement timely mitigation strategies to manage risks and prevent issues from occurring. Of particular importance as part of this proactive risk mitigation process is the need to ensure that "all aspects of the trial are operationally feasible" and "avoid unnecessary complexity, procedures, and data collection" [7]. The Clinical Data Scientist must steer the study team to only perform procedures that are essential to the outcome of the clinical trial. Ultimately, the study team must proactively confirm that the planned protocol is operationally acceptable for all sites and countries involved.

As a key member of the study team, the role of the Clinical Data Scientist is to drive the conversation and proactively manage risks that matter most, for example CtQ factors associated to critical data and processes. The Clinical Trials Transformation Initiative (CTTI)[8] introduced the CtQ factors in 2015 and organized them around the six (6) major categories (see figure 5) with strong emphasis on protocol design. Subsequently, they became a central theme in the ICH E8 draft guidance on the general considerations for clinical trials. The CtQ factors included in Annex 3 of the draft guidance are almost identical to those from the CTTI with one notable difference (i.e., changing the focus from data quantity to data quality during protocol design). In the draft guidance these factors are "*considered to be critical because, if their integrity were to be undermined by errors of design or conduct, the reliability or ethics of decision-making would also be undermined*"[9].

Lastly, the ICH E8 draft guidance is reemphasizing the need to ensure the scientific and operational feasibility of the protocol and fit-for-purpose processes considering the diversity of data sources.

CtQ Categories	CtQ factors
Protocol Design	*Eligibility Criteria* *Randomization* *Masking* *Types of Controls* *Data Quantity (CTTI) – Data Quality (ICH E8)* *Endpoints* *Procedures Supporting Study Endpoints and Data Integrity* *Investigational Product (IP) Handling and Administration*
Feasibility	*Study and Site Feasibility* *Accrual (i.e., Enrollment Strategy)*
Patient Safety	*Informed Consent* *Withdrawal Criteria and Trial Participant Retention* *Signal Detection* *Safety Reporting* *Data Monitoring Committee (DMC) / Stopping Rules (if applicable)*
Study Conduct	*Training* *Data Recording and Reporting* *Data Monitoring and Management* *Statistical Analysis*
Study Reporting	*Dissemination of Study Results*
Third-party Engagement	*Delegation of Sponsor Responsibilities and Collaborations*

Fig 5. CTTI and ICH E8 CtQ categories and factors

Considering this frame of reference and to ensure QbD, the Clinical Data Scientist must strongly contribute to, if not lead, the mitigation of some the risks associated with CtQ factors.

There are many risk areas associated with the CtQ factors, including but are not limited to the:

- Complexity of protocol designs such as umbrella, basket, platform and adaptive
- Vulnerability of the patient population (e.g., elderly, pediatric)
- Complexity of enrollment procedures (e.g., consent, eligibility, stratification and randomization)
- Deviations from standard of care

- Characteristics of the participating countries (e.g., Standard of care, customs, dialects)
- Planned rate and distribution of enrollment
- Number, profile and experience of the study sites and countries
- Nature of the protocol-required procedures, with specific emphasis on the burden they may place on patients and sites
- Organization of the trial (e.g., site-centric vs. decentralized) with telemedicine and home nursing
- Planned technologies used to collect data including when patients bring their own device (BYOD)
- Complexity of the data flow, including variety of the data sources
- Oversight of the capture and modification of the eSource data owned by the sites
- Number and experience of the data and operational vendors
- Any other study execution activities which may lead to data errors that could negatively impact the credibility and reliability of the trial results

While many risks would be evaluated and accounted for in the overall risk mitigation plans from the multidisciplinary study team, some risks such as the complexity of the data, data flows, third party vendors and planned technologies used for data collection would be the primary focus of CDS.

To foster study QbD, Clinical Data Scientists and study teams must understand the advanced concepts introduced in Part 1[1] and 2[2]. This may require changes in both the composition of the protocol review team as well as the process for developing protocols. To ease this evolution, CDS organizations could pre-define guidance for the mitigations of risk associated with standard CtQ factors by leveraging historical information on process and data issues. As the example in figure 6 suggests, the Clinical Data Scientists will need to plan for a robust risk monitoring strategy using analytics tools including key risk indicators (KRIs) and quality tolerance limits (QTLs) to proactively identify trends resulting from known risks such as enrollment speed. Those monitoring strategies could be documented in the sponsor's Integrated Quality Management Plan (IQMP) or in the Data Management Plan (DMP).

	Scenarios for enrollment speed Very fast or very slow enrollment impacting data review strategies	
	Fast Enrollment (Scenario #1)	**Slow Enrollment (Scenario #2)**
Potential Data Quality Risks	Unable to match the speed / frequency of data reviews with speed/volume of data collection	Higher risks of sites closure prior to DB Lock impacting data review and reconciliation strategies
Mitigations	• Leverage technologies (e.g., EDC, IRT, eCOA) to identify error as fast and as close as possible from the data source • Focus review on critical data points impacting eligibility • Prioritize the readiness of data review tools that matter most (Incl. KRIs & QTLs)	• Trend Screen Failure metrics • Monitor site closures • Fine-tune data review frequency considering early sites closure risks

Fig 6. Example of CDS risk mitigation guidance

Moreover, those mitigation measures should be pre-identified and implemented to the extent possible prior to study start to allow for the optimum quality controls during study conduct.

b) Risk-based study execution (i.e., The Quality Control stage)

During study conduct, the Clinical Data Scientist should continuously monitor the data related risks by:

- Monitoring the risks identified during the study design phase using the defined KRIs and QTLs
- Performing holistic data reviews including review of the various data audit trails
- Monitoring for the possible emergence of any new risks, including but not limited to:
  - Risk to database availability which could delay study start
  - Risk to study timelines which could negatively impact the availability of study results for safety reviews, the potential submission and product approval
  - Impact of protocol amendments
- Assessing the effectiveness of the implemented risk mitigations
- Adjusting or augmenting risk mitigations as necessary

The identification of outliers and atypical data patterns frequently rely on KRIs and statistical methodology tools. This means the resolution of "issues" is no longer as simple as using SDV and queries from edit checks to verify the accuracy of the data. Instead, the study team must understand the signal generated and apply critical reasoning to analyze the likely root causes. Once a signal is determined to be an issue, the underlying process or data issue needs to be addressed. Lastly, to close the loop, teams must follow-up to make sure the issue has been fully resolved.

Below are some examples of signals that can be found with the potential responses made by teams.

1. All patients at a site in Puerto Rico are Hispanic: An atypical proportion of one ethnicity at the site may be statistically outlying compared to other study sites outside South America but not unexpected in this case. The team does not need to act on the signal but should follow up until the site has finished recruiting to see if the pattern evolves.
2. Many patients at a site have the same respiratory rate: Rather than questioning if the value was correctly entered into the source document, teams should think about how this lack of variability occurred. It is possible, but highly unlikely, that many patients at a site have the same respiratory rate. It is more likely that something was wrong with how the measurements were taken and/or recorded. In this case, the process used to collect and record the rate should be reviewed, and the importance of accurate data collection and recording reiterated to the site personnel. Since the current data is not going to change, any issue with the process in taking measurements should be addressed, fixed, and monitored.
3. Patients on an oncology trial have either no or a very low number of adverse events (AEs): This is statistically unlikely. The study team should ensure the site personnel understand how to collect AEs, and possibly use source data review (SDR) to check for unreported AEs. The site personnel may need retraining, and the study team must follow up to make sure the situation is resolved. Current data might not change, but the process must be fixed and then tracked for ongoing correctness.

To address the examples above, the Clinical Data Scientists and study team must dig deep into the data to understand the root cause of the issues. They need to perform detailed analysis of KRIs and data review findings to resolve them. Occasionally, the team will need to go through multiple iterations of analysis and follow-up to fully understand the root cause. This requires a focus on details and strong communication skills as most findings will not result in queries but rather in addressing systematic process issues and site behaviors.

c) Continuous CDS process improvement

Clinical Data Scientists and study teams must leverage the lessons learned during study execution and adapt the processes to prevent reoccurrence of the issue moving forward. For systematic issues, the mitigation of a specific risk may involve a corrective and preventive action (CAPA) to do so.

Although the CAPA process is usually driven centrally by the quality organization, Clinical Data Scientists should be familiar with it. They, as SMEs in the risk management lifecycle, should also be comfortable contributing to the process through the characterization of the risk and suggesting pragmatic and robust remediations and preventive actions.

d) Impact of Risk-based approaches to CDS role

CDM must evolve substantially if it is to support QbD and risk-based study execution. This will have a dramatic impact on the CDS roles. Overall, the end to end management of the operational and scientific risks must be embedded throughout the entire process. Figure 7 below is an example of a risk-based CDS process where new tasks depicted in green are added to the traditional CDM steps depicted in blue:

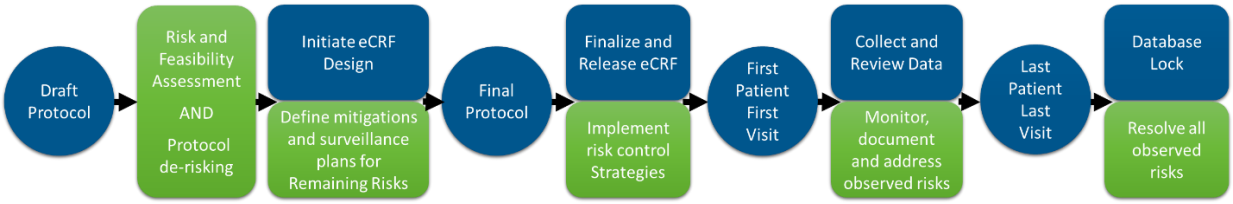


Fig 7. Example of a risk-based CDS process flow

To support such process, the scope of traditional CDM vs. risk-based CDS would be as follow:

Traditional CDM Scope	Risk-based CDS Scope
Focused on logical thinking (Output)	Focused on critical thinking (Outcome)
Study set-up upon protocol finalization	Quality by Design
Standard processes across studies	Risk-based processes tailored for each study
Focused on data integrity	Focused on data quality (i.e., data reliability)
Reviews of data after their collection	Risk-based data monitoring

This evolution would require the following CDS roles requirements:

Best Practices	Soft Skills
• Risk-based study execution • KRIs and QTLs life cycle	• Critical thinking and root cause analysis • Adaptability • Pragmatism • Influential leadership
Competencies	**Foundational Knowledge**
• Risk lifecycle management • Advanced analytics • Process management	• New research methodology (adaptive, master protocols) • Decentralized clinical trials approaches and technologies • Risk-based methodologies and regulations • Strong data flow and system literacy to investigate multifaceted issues

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 datafrom 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.

5.3) Advanced CDM competencies

Finally, let's look beyond risk-based CDS approaches and the evolution of data reviews. The new world of CDM (i.e., CDS) must be able to continuously improve capabilities across multiple dimensions, including regulatory, operations, technology, and data. Clinical Data Scientists need to rely on effective and efficient processes and controls, enabled by technology to support novel and increasingly complex trial designs, across multiple delivery modalities, all while adapting to evolving global and local regulations. As a result, Clinical Data Scientists will need advanced competencies to generate the high quality and high integrity data needed to drive the expected study outcome.

a) Clinical research concepts and strategies

Alternative protocol design is not new. Adaptive trials have been in use for many years and the adoption of master protocol designs increased almost nine-folds between 2010 and 2019[12]. According to the FDA, a master protocol is a protocol designed with multiple sub-studies, which may have different objectives and involves coordinated efforts to evaluate one or more investigational drugs in one or more disease subtypes within the overall trial structure. The three types of master protocols which could include adaptive design elements are umbrella, basket, and platform designs, each bringing benefits but leading to the challenges described below:

- Increase to the onset planning times to cater for possible adaptation scenarios
- Dynamic design increases the complexity of data review and data flows especially as most aspects of study design could be adapted including endpoints, treatment arms and data collection
- Complex simulations required to model trial design and adaptations

- Multiple drugs with potentially different routes of administration increasing complexities with DCI design, investigational product (IP) distribution and safety review for concomitant medications and adverse events
- Complicates drug management and masking procedures

- Multiple diseases requiring increased domain expertise of CDS team
- Each indication would likely require specific endpoints
- Greater variation of participant population increases complexity for data reviews

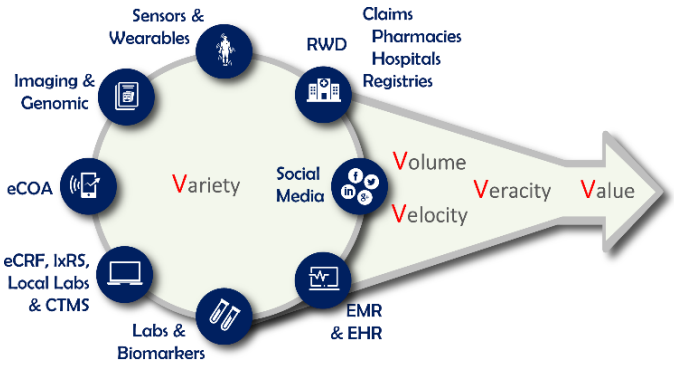
- Inherits complexities of umbrella and basket designs
- Additional upfront planning to consider all data related scenarios
- Increased number of interim analysis
- Multiple sub-trials may require dedicated or specialized CDS teams to manage

A common challenge associated with these four designs is the need for flexible DCIs requiring complex and dynamic branching logics in EDC, eCOA and IRT. Such designs could be implemented within the context of traditional bricks and mortar sites or in a decentralized setting.

Study set-up is no longer a matter of translating an approved protocol into eCRFs. It is evolving as a specialized activity where the Clinical Data Scientist needs to master scientific and operational concepts and being able to drive the set-up of an integrated and flexible technology centric data ecosystem. It is also important to know that scenario planning is multidimensional as adaptive design, master protocol and decentralization of the clinical trial are not mutually exclusive (i.e., they could all happen concurrently in a single study).

While the overarching benefits of these trial designs can accelerate the overall development of a new drug, they will challenge Clinical Data Scientists in terms of design, controls, data review, and increased frequency of interim analysis. All of this will require careful planning to execute properly. In addition to the need to think critically and no longer logically, this does imply that the foundational knowledge required to do so is substantially increased.

b) Clinical data acquisition, standards and modeling



For Clinical Data Scientists, the most critical of the 5Vs is understanding the value of the data as not all data are created equal and will not bring the same value. Assessing how data sources will contribute to the primary objectives of the protocol is and will remain a core competency, however assessing the secondary value of each source and extracting the right evidence from it, whether it be for synthetic arm, translational research, and patient engagement will drive the focus of how that data is managed and the effort allocated to it. This is already being done for reimbursement by extracting real world evidence (RWE) from real world data (RDW).

Assessing the variety of the data sources collected in the clinical trial, where they are going to be integrated (e.g., in EDC or not), how they are transformed into useful information to generate insights, and how they will drive action is the second highest priority for Clinical Data Scientists. Is the data structured or unstructured? What applicable data standards can be used? How will the data be interpreted, by whom, and when is it needed by? All of these questions lead directly to how the data should be modeled. Data modelling for CDISC standards, such as CDASH, SDTM and to a lesser extent BRIDG and ODM, are already part of our core competencies, however we will need to expand experience and proficiency in other standards such as HL7, FHIR, in order to fuse RWD sources such as EMR/EHR into our intelligent CDMS and enable approaches like eSource and direct data capture.

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 managing the data. High volumes will require automated solutions to assess the quality and integrity of the data. On the other hand, high velocity data sources will require new approaches that drive action by detecting and promptly differentiating signals from background noise. Clinical Data Scientists will need to be able to develop processes and controls to identify appropriate signals when managing high volume and velocity data.

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 and equally important, effectively partner with the external parties to understand their controls and what additional ones will be required.

c) Automation technologies

First and foremost, organizations must anticipate the implication of using AI based solutions leading to the set-up of digital workforces that can work 24 hours a day and be "hired" (i.e., implemented) and on-boarded quickly and at a decreasing cost. This digital evolution will profoundly transform the workplace in years to come. Also, as we leverage advanced capabilities such as ML, RPA and IPA, we need to augment our approaches to testing, deployment, and management. It is easier to automate chaos than it is to automate order.

Conceptually, RPA has similarities with edit checks. RPA scripts using simple branching logics and Boolean operators (e.g., and, or, not, etc.) can be applied to automate repetitive and transitional processes. Using those scripts, an RPA bot will act very much like an end user, capable of logging in and out of systems to perform basic data-oriented tasks. The Clinical Data Scientist will need to not only be able to identify the appropriate use cases where RPA can be applied but be able to assess the value of automating it given the underlying cost and time involved. Freeing up part of an FTE per year, may not offset the cost of design, development, and testing of an RPA bot, whereas freeing up several FTEs per year will result in a positive return on investment (ROI). The cost needs to be balanced with quality as the automation of large scale repetitive manual tasks could result in meaningful process accuracy and reliability improvements. Additionally, new procedures will be required to test and manage the identity and deployment of bots, given they will perform steps like any human, and their actions will be recorded in systems audit trails. Auditors will undoubtedly seek to understand how RPA bots were developed, tested, and managed in production environments.

Intelligent solutions powered by AI require the most significant transformation for CDM. Overall, ML based solutions will act as virtual Clinical Data Managers assisting Clinical Data Scientists. As a result, expert Clinical Data Scientists will mentor virtual Clinical Data Managers to accurately perform data reviews and other CDS activities. This means that when ML models are established, Clinical Data Scientists will need to define the objectives of the intelligent solution and identify the datasets required for training and testing. These datasets will need to cover all expected data review scenarios (i.e., be complete) and be truly representative of the use cases anticipated in production (e.g., include data review scenarios across all study phases and therapeutic areas).

The failure to define the right datasets could bias the system behaviors and lead to inconsistent data review accuracy. Clinical Data Scientists also need to play a key role in assessing the performance of the solution by defining the right testing strategies (e.g., by setting a minimum ratio of the number of correct predictions compared to the total number of inputs in the test dataset[2]).

The figure below is an example of a process for the "mentoring" (i.e., training) of an ML-based solution.

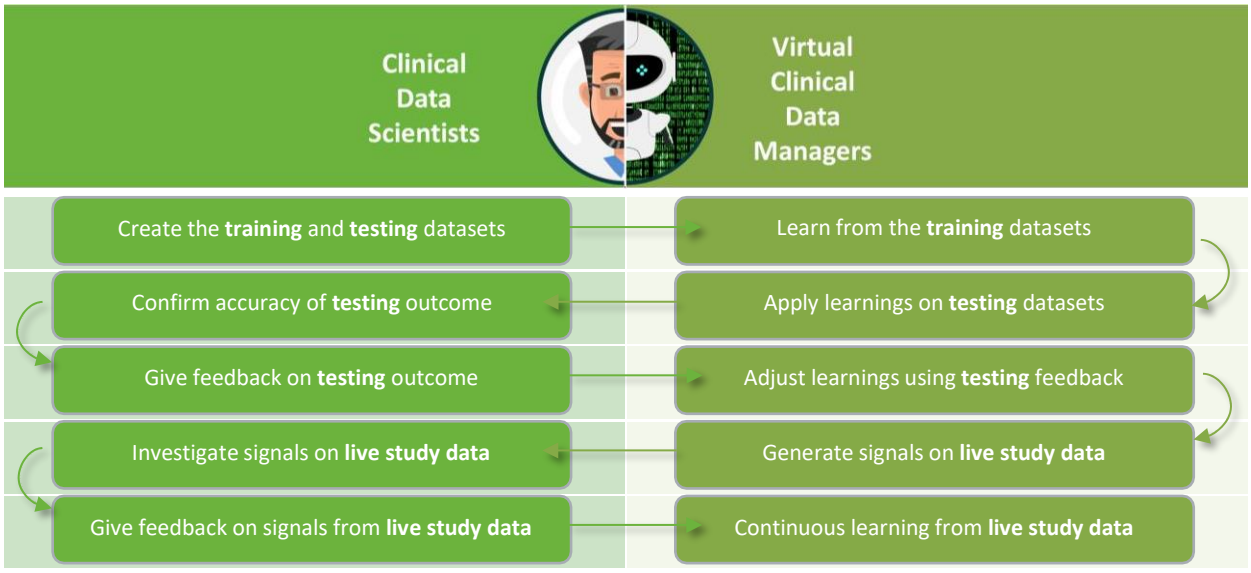


Fig 8. Example of learning process for an ML-Based solution

ML technologies require fundamentally different approaches to their development and validation including the use of methods like recall (i.e., ability to predict in all cases), precision (i.e., accuracy of the predictions) and F-measures (i.e., combining recall and precision) to determine the reliability of models.

First, the model must be trained on a controlled dataset of known quality. Then, it can be applied to the testing dataset to assess the precision and recall in a simulated "real-world" environment to understand its precision and recall.

Its acceptable level of accuracy should be one that is better than the accuracy of the current process. If not, its use needs to include commensurate human supervision and training to monitor its accuracy. Lastly, once a machine learning model is deployed, it will need to be retrained at specific intervals in order to maintain and ultimately improve its accuracy over time.

d) Vendor oversight

Almost every trial initiated by a sponsor today is enabled by a multitude of vendors, providing services ranging from strategic to tactical and global to localized. These partnerships manifest themselves in varying forms, from multi-years relationships to one-off contracts.

In recent years, health authorities have placed more responsibility on sponsors in terms of vendor oversight, and this has been further solidified by the latest addendum of ICH E6 (R2) [7]. Clinical Data Scientists, who oversee external vendors, will need to transition from managing status and timelines to overseeing quality delivery through data-driven insights on vendor performance that are tailored to the services rendered.

Facilitating this will require direct access to a vendor's operational data, centralizing raw data internally, and developing capabilities to analyze quality and performance on a continuous basis, rather than the more traditional, passive approaches. Additionally, the Clinical Data Scientists will also need to strengthen their relationships and understanding of interdependencies with Clinical Operations, Quality Assurance and Procurement functions to effectively manage a vendor with a single, integrated voice from the sponsor.

e) Influential leadership

Given the dynamic and evolving nature of the clinical research landscape, Clinical Data Scientists will have a central role in leading cross-functional teams as well as driving complex decision making. CDS is not a support function but a key stakeholder responsible for the most critical asset: the study data. Clinical Data Scientists must therefore demonstrate influential and leadership skills. Doing so requires business acumen, technical capabilities, and the ability to manage continuous change within their organizations and teams. As discussed in previous sections, Clinical Data Scientists will need to meaningfully expand their core competencies and foundational skills.

Additionally, there will be an increased emphasis on the soft skills such as:

- Understanding the points of view of a wider range of stakeholders including sites and patients
- Critically assessing risks and their impacts to determine best mitigation strategies
- Suggesting alternative and operationally sound solutions (i.e., being a pragmatic innovator)
- Articulating complex technological and scientific concepts
- Understanding the ramifications and rational behind study team decisions

Ultimately, Clinical Data Scientists must be tactful and empathic listeners able to drive consensus around complex scenarios and if necessary, demonstrate decisiveness by taking and owning decisions.

In conclusion, we could compare the scope of the CDM vs. Advanced CDS scope as:

CDM Scope	CDS Scope
one and two dimensional trials	Three or more dimensional trials
Project management	Cross-functional leadership
Randomized controlled trials	Adaptive & master protocols
Vendor management	Vendor oversight
Clinical research standard	Clinical research and healthcare standards
Logical thinking	Critical thinking

This evolution would require the following CDS roles requirements:

Best Practices	Soft Skills
• Intelligent systems management • Generation of secondary data assets (e.g., synthetic arms) • Data and system integrations	• Critical thinking • Pragmatism • Influential leadership • Ability to manage ambiguities and dynamic environments
Competencies	**Foundational Knowledge**
• Advanced analytics • Vendor oversight • Patient centric technologies	• New research methodology (adaptive, master protocols) • Decentralized clinical trials approaches and technologies • Risk-based methodologies and regulations • Understanding of new data concepts such as sequenced data and unstructured data • Health care standard models and terminologies • Automation and AI concepts (e.g., supervised vs. unsupervised ML)

6. Impact of the CDS evolution on business models

6.1) Historical background on business models

Newer technologies, evolving processes, and innovative clinical research strategies are impacting roles irrespective of CDM business models including in-house and outsourced teams in both FSP and full-outsourcing service models.

In the case of the **outsourced model**, the service provider (i.e., the CRO) must adapt its processes, technologies, and resource capability to meet the industry's evolving CDS expectations to remain competitive. These expectations are not separate or distinct between sponsor and CRO—this evolution is required for all organizations.

However, in the case of the **FSP model**, the system, process, and role dependencies between the service provider and the sponsor are integral to the business model itself. The evolution by the sponsor on any or all of these three dimensions has a direct impact on the service provider and its ability to operate in an FSP model. As a result, the sponsor and FSP provider (e.g., CRO, BPO, technology service provider) need to carefully plan and align their evolution toward CDS together to account for changes to systems, processes, and roles. It is also important to note that large scale FSP services may be provided by traditional CROs which have diversified through both the outsourcing and FSP models.

Traditional FSPs were established to flexibly augment resource capacities with experienced staff and, in some cases, as a way to reduce operational costs by converting fixed internal costs to discretionary lower costs. Unlike the outsourcing CRO model, the sponsor keeps the control over the data by having FSP staff using its systems and processes.

Small-scale FSP models used primarily for staff augmentation will require adjusting the staff selection to account for the evolving CDS responsibilities. The onboarding and training of FSP staff will likely be similar to the onboarding of sponsor's own employees.

Larger scale FSP models will require more significant adaptations. They are predominantly offshore to take advantage of lower cost of resources. The prospects of such savings led some sponsors to engage service providers to establish large offshore centers in a variety of countries and regions including India, South Africa, Mexico, Asia, and others. The model often followed the strategy of hiring and training a mix of fresh graduates (i.e., junior staff) and experienced resources to lower the overall wage costs and by delegating high volume and/or repetitive tasks to them to realize meaningful short-term ROI. As shown in figure 9, the scope initially included activities such as data entry of paper CRFs and diaries, discrepancy management and data reconciliation. In some cases, the scope included technically driven tasks such as database set-up, report programming, dataset creation and upgrade from legacy systems to newer technologies requiring migration of data.

Though the overall trend is still to shift toward offshore locations, there are instances where the high ratio of offshore junior FSP or CRO staff has resulted in an experience gap leading some sponsors to insource activities back to their own higher skilled resources. Additionally, in recent years, many sponsors have invested in their own operational CDM centers in low-cost locations, called "captive sites" leading to the insourcing of many of key roles.

Lastly, as the industry evolves, and as new options emerge, the viability of the lower cost and lower skill business models are being challenged for a number of reasons, including but not limited to the following:

- Increasing wages in established low cost locations
- Rising clinical research complexities requiring to up-skill R&D staff to a level where the training of existing FSP or CRO resources is not enough
- Increasing competition for the recruitment of advanced degrees in data sciences and statistics
- Reliability and cost-effectiveness of automation solutions based on RPA and IPA eliminating simple and repetitive manual tasks
- Variability in clinical study designs where one-size-fits all and predictable processes are becoming the exception
- Shift from reducing operational cost to eliminating the cost of non-quality by refocusing on first-time quality

So, business models will need to be adapted to newer perceptions in order to evolve their offering and reassess their model to ensure they subsist long term.

6.2) Business models of the future

While some providers will continue to focus their model on the transfer of non-core roles to them (i.e., resource enabled model), others are more radically reconsidering the model of the future in order to

invert the insourcing trend leading to the decline of the current models. Providers may benefit in delivering technology driven innovations that leverage their expertise and deep know-how of the sponsor's processes and systems (i.e., technology enabled model). The expectation is that the investment cost in technology enabled models and processes will be offset by long-term savings.

Traditional resource enabled models

As an example, the figure below illustrates the evolution of the tasks delegated by sponsors to offshore FSPs over the last decade. Information was gathered through a survey of FSP delivery leaders in India on the evolution of the span and scope of tasks managed by large FSPs. During this time, FSPs were able to upskill their staff through training to meet the increase in responsibilities. Some tasks requiring cross-collaboration like database lock and analytics have not been yet transitioned to FSPs by all sponsors.

Task Assignments to FSPs			
Tasks	**2010**		**2020**
	FSPs	*In-house & Captive*	*FSPs*
Database Development	Most	Most	All
Specification Writing & UAT	Most	All	All
Data Entry (CRF or Paper diaries)	All	Most	All
Study Planning	None	All	All
Discrepancy Management	All	All	All
Study Management	Few	All	All
Dataset Creation	Most	All	All
Coding	Most	All	All
Database Lock	Many	All	Most
Analytics and Reporting	None	All	Many
Quality Control	All	All	All
External Data Reconciliation	Many	All	All
Mainly Out of Scope or Future Planned		**In scope or Limited Scope**	

Fig 9. The FSP Task assignments

The speed of change is expected to accelerate as a result of the trends highlighted in Part 1[1]. So, for resource enabled service providers, the opportunity lies in aligning resources needs to the sponsor's shift toward CDS. The transition requires a paradigm shift as simply up-skilling resources to the new clinical research approaches and emerging technologies may not be enough to adapt to changes in competencies, foundational knowledge, and soft skills.

Service providers will have to ensure their staff roles evolve to keep-up with the pace of change of sponsor's technologies and processes. It may become challenging for providers that only focus on resource availability. A parallel pace will require that providers continue to invest into their training to ensure alignment with the sponsor requirements. This is expected to put an ongoing burden on the providers to manage a long-term transformation while ensuring no impact on delivery. The demand from sponsors to source experts in technology like AI/ML solutions may lead to recruitment and

retention challenges. As sponsors start looking for future proof partners, providers will need appropriate recruitment approaches and investment in talent development and retention (i.e., training, reskilling and career development) to remain relevant.

Transition to technology enabled business models

In contrast, a few providers originally supporting resource enabled models are now delivering some of their services with higher predictability and quality through their own technologies as opposed to solely rely on the sponsor's technologies. Those providers are leveraging software as a service (SaaS) solutions which do not require complex integrations and can be used in addition to the sponsor systems.

Some examples include:

- ML based SDTM Mapping solutions
- SDTM compliance QC tools
- Metadata based eCRF Design creation tools
- EDC design QC tools

This approach allows the sponsor to leverage innovative technologies without the implementation costs. In return, the technology enabled provider can cater to higher quality services, minimize the reliance on hard to find experts and ultimately maximize revenues when technology ROI is realized. In addition to low footprint SaaS technologies, some providers are also offering the sponsors end-to-end services around more complex third-party solutions such as IRT, EDC and eCOA including the licensing of the solution from its technology providers. This has been facilitated by the fact that some technology providers have sub-contracted their services to large providers who gained expertise by delivering them.

In comparison, some technology providers are becoming service providers by leveraging the knowledge of their own technologies and the expertise of their delivery teams. So, we have seen both backward and forward integration of technology-to-service and service-to-technology models. This new paradigm seems to be in favor of all parties - sponsor can focus on their core R&D priorities while providers deliver the latest technology requirements and focus on innovations. In this model, the sponsor has the flexibility to retain core technologies and data in house and transfer the full management of operational study specific solutions to the technology enabled providers. This also potentially allows the sponsor to test emerging solutions flexibly prior to investing in a lengthy and costly implementation project.

The segment which will still remain out of the ambit of the above expectations will be highly specialized players in both technology and services. As mentioned in the reflection paper Part 1, there are several technology organizations developing RBM, DCT, eSource, supervised cleaning, future proof platforms and newer models capable of utilizing AI/ML. Such organizations may solely remain as innovators and providers will be required to collaborate with them by bringing in their services strength as well as technology support.

While these evolutions are not certain, we foresee that all models will need to transform to either scaled-up resource-enabled services or technology-enabled services by acquiring the desired knowledge and capabilities either internally or through relationships with partners. But regardless of the model, the CDS requirements and role expressed in the reflection papers do not change. It comes down to clear communication pathways and even clearer assignments of accountability.

7. Impact of the role evolution

The evolution from CDM to CDS summarized in this paper results from evolving regulations, technologies, and clinical research approaches. This represents a major shift in focus, not only for CDM but for all clinical research stakeholders.

Summary of the CDM focus	Summary of the CDS focus
Achieve data integrity	Achieve data quality
Quality controls	Quality by Design
Focused on logical thinking (Output)	Focused on critical thinking (Outcome)
Randomized controlled trials	Adaptive and master protocols
Focused on site generated data	Focused on eSource data from DCTs
Standard processes across studies (one size-fits-all)	Risk-based processes tailored for each study (focus on what matters)
Low volume of data and sources	High volume of data and sources
Simple data flows	Complex data flows
Vendor management	Vendor oversight
Data cleaning	Data review, tagging, exclusion and curation
Project Management	Cross-functional leadership
Clinical research standard	Clinical research and healthcare standards
Clinical research data	Clinical research and healthcare data
Traditional programming (SQL, C#, SAS, etc.)	ML (Python, R, etc.)
Standard data interrogation (e.g., SQL)	Advanced data interrogation (e.g., non-SQL)

While taking different pathways, many CDM leaders will gradually evolve their organization toward their own tailored CDS future. To initiate such a change management endeavor, they must clearly define their own ultimate destination and value proposition for their organization considering the evolution of the industry toward a digital and patient centric future.

This path will be highly influenced by their **current** company landscape including:

- Size (From small biotech to top 10 pharmaceutical companies)
- Geographical footprint
- CDM roles, scope and structure (e.g., flat, hierarchical or matrixed)
- Culture (incl. digital literacy, tolerance to mistakes, agility, silos, innovators vs. followers)
- Merger and Acquisition strategies
- Study team composition
- Cross functional dependencies
- Technologies (e.g., availability of a metadata repository (MDR) or not)
- Talent pool
- Emerging functions (e.g., Start-Up, Design Center, Digital Innovation) and roles (e.g., Head of Clinical Data Science, Chief Digital Officer, Digital Integration Specialist, Trial Innovation Lead)

Beyond those, the roadmap and change management plan must integrate aspects such as:

- Human resources strategies: Job classification, career ladders, talent acquisition, compensation, onboarding, training, upskilling, mentoring and evaluation
- CDS operating models (e.g., in-house, outsourcing and FSP models)
- Internal and external stakeholder relationship management
- Organizational change including culture

For CDM itself, this leads to the evolution of its competencies, foundational knowledge, best practices and soft skills requiring the following expectations to be added on top of the existing CDM.

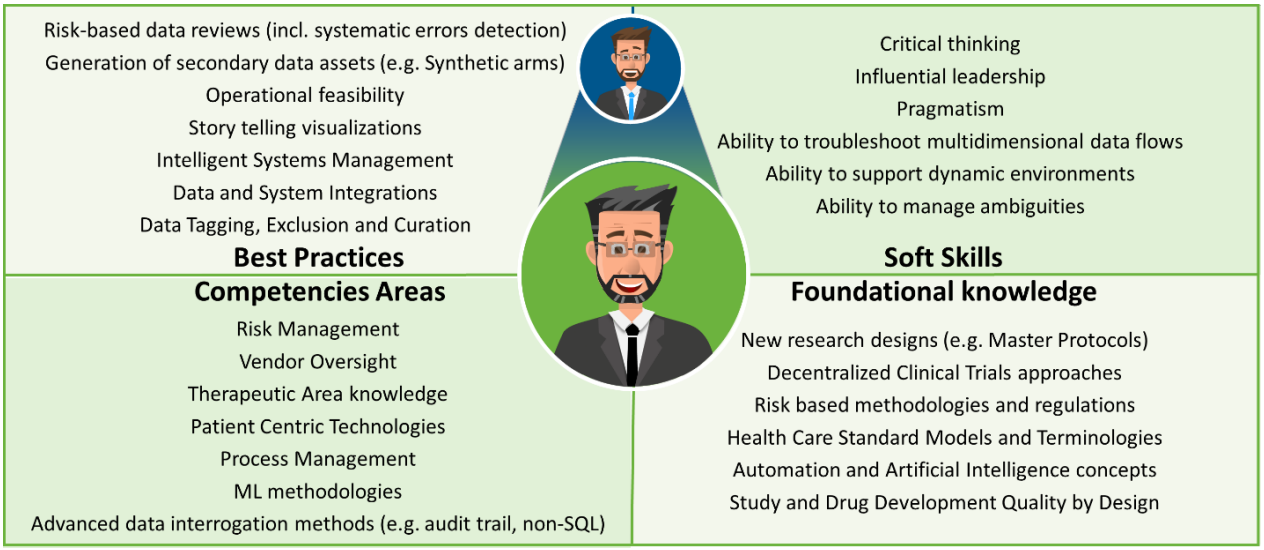


Fig 10. CDS role framework

While this framework will need to adapt in the coming years with further evolutions in technology and regulations, it could be leveraged as a starting point to support the evolution of the CDM roles toward CDS. The **soft skills** and **foundational knowledge** expectations will likely be added to the job descriptions and hiring requirements aligned with each organization strategies.

Furthermore, the need to know how to apply those skills to specific tasks (i.e., **competencies**) aligned with new **best practices** will guide training and up-skilling approaches to enable Clinical Data Scientists to take on new roles.

Below are a few examples of the definition of technical and non-technical roles that may emerge:

- **Data Integration Specialist:** The integration of data and knowledge from several sources is also known as data fusion. New data types are entering clinical studies regularly. CDS needs to evaluate new technologies including wearable devices using sensors. Further, they need to liaise with scientific and technology experts and be willing to explore new data types. As an example, time sequenced data generated by sensors and wearables at high velocity and volume cannot be integrated in the same manner as EDC data which requires complementary knowledge and technologies.

- **Data Mining and Profiling Specialist:** Data mining and profiling are the initial steps in data analysis, where users explore a large dataset, structured or not, to uncover initial patterns, characteristics, and points of interest.

  This process is *not* meant to reveal every bit of information a dataset holds, but rather to help create a broad picture of important trends and major points to study in greater detail. Data profiling can also assist by reducing work time and finding more useful and actionable insights from the start alongside to presenting clear paths to perform better analysis.

- **Data Curator:** The curation of data includes its anonymization, integration, organization and exploration. The intent is to objectively confirm its integrity and quality to generate the appropriate secondary data assets such as RWE from RWD.

- **Data Annotator:** The annotation of ideally curated data is the process of labeling the data available in various formats like text, video or images. For supervised ML labeled data sets are required, so that machine can easily and clearly understand the input patterns.

- **Data Visualization Expert ("Storyteller"):** Data visualization is the graphical representation of information and data. Too often, visualizations have been limited to interactive but still basic descriptive statistics using simple graphs. Being able to tell a clear story from a large volume of data is crucial as insights are difficult to discover otherwise.

  CDS must discover data trends and signals threatening the reliability of the trial results in an actionable way. Data Visualization Experts must design solutions combining and transforming diverse and complex data sources into insightful visualizations.

- **ML Model Builder:** ML Models are developed and trained by leveraging statistical and programming methodologies. The developer must also lead the selection of the appropriate curated and if necessary annotated datasets for ML model training and testing.

Those are just examples of potential interdependent CDS roles supporting a subset of the overall CDS data flow starting from data integration to data interpretation, through mining, curation and annotation to generate knowledge from data.

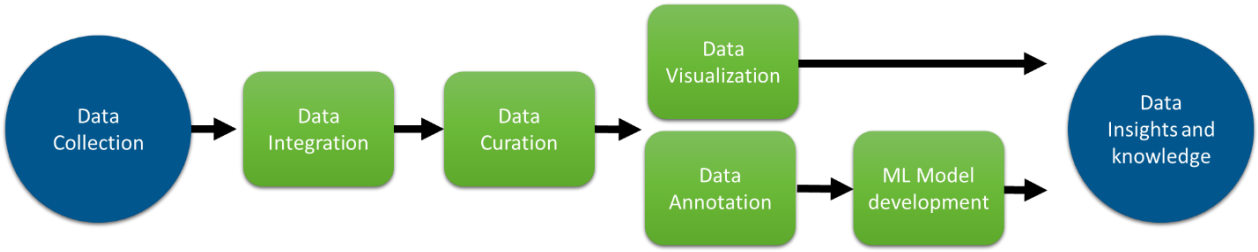


Fig 11. From data to knowledge data flow

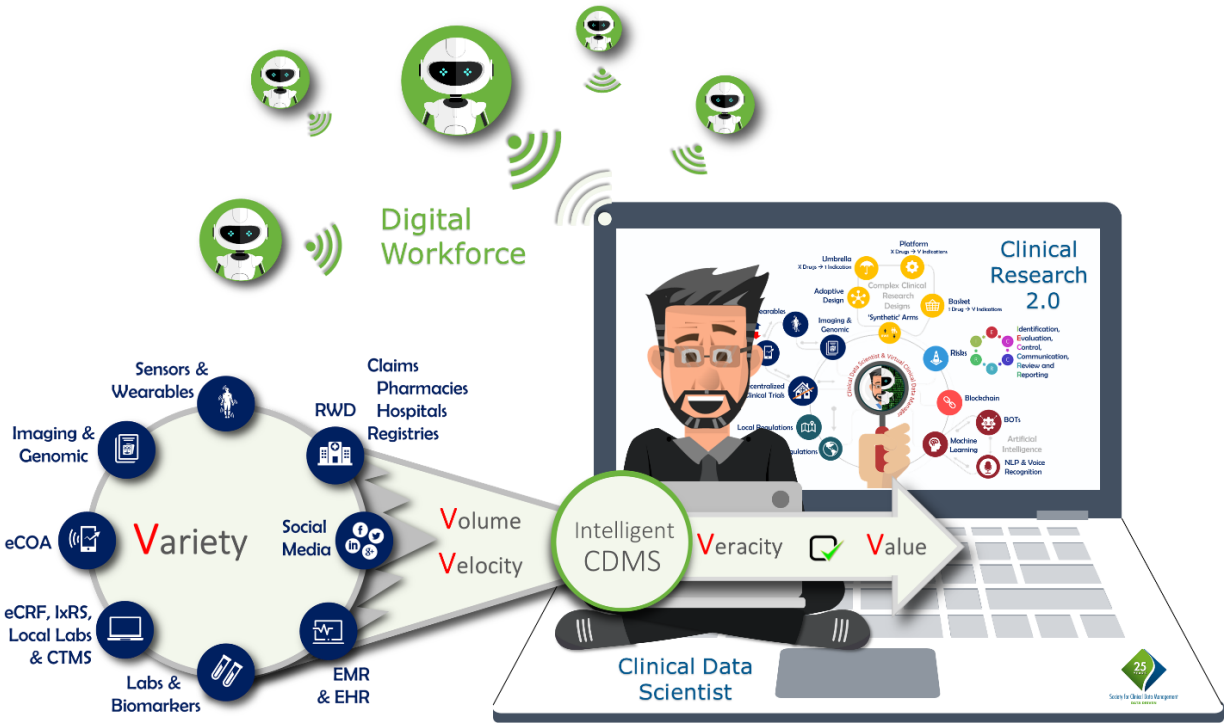
8. Conclusion

The evolution toward CDS has started and is unavoidable especially as the COVID-19 pandemic has accelerated the decentralization of clinical trials at a scale never seen before. On the one hand, this has led to stronger support from leadership and regulators. It also removed many of the traditional adoption barriers across all stakeholders. On the other hand, the need to evolve quicker is adding pressure to adapt without much pro-active organizational readiness.

So, CDM must transform into CDS rapidly to emerge as a true clinical research enabler. To seize this meaningful opportunity, CDM leaders must take advantage of the recent changes in the clinical research landscape, the significant investment in DCT related infrastructures as well as the growing maturity of automation technologies.

This is a complex task requiring thoughtfulness and a clear strategy. To support our community, the SCDM Innovation Committee has released three reflection papers on our evolution toward CDS providing insights on drivers, regulations, technologies, and roles. We hope that these will guide experts embarking on their CDS journey to develop their own strategies leveraging their current CDM expertise as a foundation to meet the demand of clinical research and regulations by leveraging novel approaches and maximizing the potential of available technologies.

Last, per its vision, SCDM will continue to *lead innovative clinical data science to advance global health research and development* and as such intends to release further helpful information on its CDS website as we anticipate the evolution of our industry and technology to continue to influence our CDS destination.



References

[1] SCDM, June 2019, The Evolution of Clinical Data Management into Clinical Data Science – A Reflection Paper on the impact of the Clinical Research industry trends on Clinical Data Management. Available at https://cdn.fs.pathlms.com/a1AnYc9xSiu8GbVj4bHi?cache=true

[2] SCDM, March 2020, The Evolution of Clinical Data Management into Clinical Data Science (Part 2: Thetechnology enablers) – A Reflection Paper on the impact of the Clinical Research industry trends onClinical Data Management. Available at https://scdm.org/wp-content/uploads/2020/03/SCDM- Reflection-Paper-Evolution-to-Clinical-to-Data-Science-Part-2.pdf

[3] Applied Clinical Trials, Sony Salzman, March 2020, Clinical Data Manager: A Roadmap for the future. Available at http://www.appliedclinicaltrialsonline.com/clinical-data-manager-roadmap-future

[4] SCDM Good Clinical Data Management Practice Guidelines, Available at https://scdm.org/gcdmp/

[5] SCDM Certification program for Clinical Data Managers, Available at https://scdm.org/certification/

[6] 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

[7] ICH, Integrated Addendum to ICH E6(R1): Guideline for Good Clinical Practice E6(R2). Available at https://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6/E6_R2_Step_4_2016_1109.pdf

[8] CTTI, May 2015, Quality by Design (QbD) Project, Critical to Quality Factors Principles document https://www.ctti-clinicaltrials.org/files/principles_document_finaldraft_19may15_1.pdf

[9] ICH, ICH E8, General Considerations for Clinical Trials Draft Guidelines , Available at https://database.ich.org/sites/default/files/E8-R1_EWG_Draft_Guideline.pdf

[10] SCDM Conference, Sep 2019, Ni A. Khin, M.D. (FDA), Kassa Ayalew, M.D., M.P.H (FDA), Dr Jean Mulinde (FDA) and Lisbeth Bregnhøj (Danish Medicines Agency), Regulatory Panel Presentation, Available at https://bit.ly/39Am8t6

[11] 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

[12] BMC, 2019, Systematic review of basket trials, umbrella trials, and platform trials: a landscape analysis of master protocols. Available at https://trialsjournal.biomedcentral.com/articles/10.1186/s13063-019-3664-1

Main abbreviations

AI	Artificial Intelligence
BPO	Business Process Outsourcing
BYOD	Bring Your Own Device
CAPA	Corrective Action/Preventive Action
CCDM	Certified Clinical Data Manager
CDM	Clinical Data Management
CDMS	Clinical Data Management System
CDS	Clinical Data Science
CRF	Case Report Form
CRO	Clinical Research Organization
CtQ	Critical to Quality
CTTI	Clinical Trials Transformation Initiative
DCI	Data Collection Instrument
DCTs	Decentralized Clinical Trials
DMP	Data Management Plan
eCOA	electronic Clinical Outcome Assessment
eCRF	Electronic Case Report Form
EDC	Electronic Data Capture
EHR	Electronic Health Records
EMA	European Medicines Agency
EMR	Electronic Medical Records
FDA	Food and Drug Administration
FHIR	Fast Healthcare Interoperability Resources
FSPs	Functional Service Providers
GCDMP	Good Clinical Data Management Practices
GCP	Good Clinical Practices

ICD	International Classification of Diseases
IP	Investigational Product
IPA	Intelligent Process Automation
IQMP	Integrated Quality Management Plan
IRT	Interactive Response Technology
KRIs	Key Risk Indicators
MDR	Metadata repository
MedDRA	Medical Dictionary for Regulatory Activities
mHealth	Mobile Heath
MHRA	Medicines and Healthcare products Regulatory Agency
ML	Machine Learning
QbD	Quality by Design
QC	Quality Control
QTLs	Quality Tolerance Limits
RBM	Risk-Based Monitoring
RBQM	Risk-Based Quality Management
ROI	Return on Investment
RPA	Robotic Process Automation
RWD	Real-World Data
RWE	Real-World Evidence
SCDM	Society for Clinical Data Management
SDLC	Software Development Life Cycle
SDR	Source Data Review
SDV	Source Data Verification
SME	Subject Matter Expert
SNOMED	Systematized Nomenclature of Medicine