

# **The Evolution of Clinical Data Management to Clinical Data Science**

## *A Reflection Paper on the impact of the Clinical Research industry trends on Clinical Data Management*

**4 June 2019**



## **Society for Clinical Data Management**

© 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 -

1. Foreword .....	3
2. Abstract .....	3
3. Acknowledgements .....	3
4. State of the Clinical Data Management discipline in the industry .....	4
5. Drivers of change and their impact on CDM .....	5
5.1. Clinical Research approaches and their rising complexities .....	5
a) Adaptive Study Design .....	6
b) Study designs to accommodate multiple investigational products and/or indications .....	7
c) Study design leveraging Synthetic Arms .....	8
d) Decentralized Clinical Trials .....	8
e) Impact of emerging study designs on Clinical Data Management.....	9
5.2. Impact of regulations on Data Quality and CDM Practices .....	10
a) How do regulations define Data Quality .....	10
b) Risk-based Clinical Data Management Approaches.....	11
c) Impact of local regulations to CDM.....	12
5.3. Evolution of technologies.....	13
a) Reduction of EDC centricity moving forward .....	13
b) Artificial Intelligence (AI) based applications.....	14
c) Blockchain.....	17
d) Sensors and Wearables.....	18
e) Impact of new technologies on Clinical Data Management .....	19
6. Evolution of the Clinical Data Management Role .....	19
a) CDM at the Cross-section of Risk-Based Study Execution .....	19
b) Foundational Clinical Data Management Competencies.....	20
c) New and Refined skillset.....	20
7. Conclusion .....	21
References.....	22
Main abbreviations .....	23

---

## 1. Foreword



As SCDM is celebrating its 25<sup>th</sup> year anniversary, the SCDM Innovation Committee seeks to raise awareness on the upcoming industry trends affecting Clinical Data Management (CDM) and prepare for its evolution toward Clinical Data Science. In the context of this reflection paper, Clinical Data Science is defined as the strategic discipline enabling data driven Clinical Research approaches and ensuring subject protection as well as the reliability and credibility of trial results. Clinical Data Science encompasses processes, domain expertise, technologies, data analytics and Good Clinical Data Management Practices essential to prompt decision making throughout the life cycle of Clinical Research.

The target audience for this reflection paper is broad-based clinical data management professionals - from Subject Matter Experts (SMEs) executing CDM activities for their clinical studies to CDM Leaders setting the direction for their organizations.

## 2. Abstract

The main objective of this paper is to provide a forward-looking and pragmatic view on why and how emerging study designs, regulations and technology innovations are reshaping the role and profile of CDM. This paper particularly examines the industry drivers and trends in Clinical Research and their direct impact on CDM. The paper concludes by providing organizations, leaders and SMEs initial insights on the evolution of the CDM role and associated best practices in our journey from traditional Clinical Data Management to Clinical Data Science.

The SCDM Innovation Committee intends to release subsequent papers and/or articles which will provide deeper insights on specific areas of change summarized or introduced in this paper such as technology enablers and the CDM role evolution.

## 3. Acknowledgements

***Disclaimer: Not all the views expressed in this reflection paper may be those of the individual companies or entities for which the authors are employed or affiliated.***

Authors and Contributors:

- Patrick Nadolny, Global Head CDM & Programming, Allergan, SCDM Innovation Committee Chair
- Demetris Zambas, Global Head Data Monitoring & Management, Pfizer, Former SCDM Board Chair
- François Torche, CEO, CluePoints
- Sanjay Bhardwaj, Global Head Clinical Data Management, Allergan, SCDM Board Member

Acknowledgments:

The reflection paper has been reviewed and/or been based on the work of SCDM sub-teams. Main SCDM leaders include:

- Alice Phung, Global Head Clinical Programming, Allergan, SCDM Technology Innovation Team Lead
- Joe Pascua, Sr Manager, Data Management, Genentech, SCDM CDM Role Evolution Team Lead
- Linda King, Global eCOA Capability Lead, Astellas Pharma, SCDM Board Chair
- Michael Goedde, Global Head Data Operations, PRA, SCDM Board Vice-Chair

SCDM also acknowledges the many volunteers who participated in the SCDM Innovation Committee since January 2018 and contributed in forming the thoughts expressed in this reflection paper.

---

## 4. State of the Clinical Data Management discipline in the industry

Many CDM organizations have been processing data the same way for a long time despite the availability of newer data technologies and the changes to regulations. While process stability has allowed them to master the day-to-day data validation and querying processes, this has prevented them from leveraging their precious resources for additional value-added activities. The excessive focus on established technology and traditional data cleaning processes resulted in a disconnect with the true desired outcome. Today, many operational models and processes focus on “outputs” (e.g. all data is collected, all queries resolved as quickly as possible after Last Patient Last Visit, all external data reconciled) and not necessarily “outcomes” (e.g. reliable, trustworthy and scientifically sound data to gain regulatory approval).

To date, the introduction of technologies has often resulted in increasing complexity and cost rather than enabling the efficient development of new therapies. Many CDM leaders have acknowledged that the full potential of EDC has not been realized and that most implementations ended-up converting existing inefficient paper processes inside an electronic tool. As examples, EDC data has not consistently been used to accelerate site payments or drive SAE reporting in safety databases. Overall, changes to downstream processes and systems were not transformational enough to take advantage of faster data availability from EDC. Additionally, despite the faster availability of data from the sites compared to paper-based processes, the cycle time from Last Patient Last Visit to database lock has been hardly reduced over the last decade, decreasing from about nine (9) weeks in 2008 to seven (7) weeks in 2018<sup>1</sup>. Additionally, it has been shown that on average only 3.7% of the data entered in the eCRF is changed after entry in EDC with only 2.6% of changes not attributed to Source Data Verification (SDV)<sup>2</sup>. As a result, some leaders are questioning the true impact and efficiency of traditional data review and CDM processes. This perception is re-enforced by the release of risk-based regulations such as ICH E6 (R2) opening the door to different approaches.

Lastly, Mergers & Acquisitions (M&A) activities and constant re-engineering have created complex operational scenarios. It is not rare to observe multiple versions of similar processes, across multiple operational models (e.g. In-House, Outsourced, Off-shore) and different technologies utilized (Inc. legacy systems not retired yet). This largely depends on the M&A strategy, the compliance of merging companies with industry standards such as CDISC, the stage of integration of the acquired companies and/or the implementation phase of new data collection and/or data management solutions.

At the end of the day, no matter the state of a CDM organization, the accelerating pace of change is calling for action. The volume of data collected outside EDC has already eclipsed the volume of data collected on eCRFs and is growing at a much faster pace every year, fueled by the cries for patient centricity leading to the rapid adoption of eCOA, wearables, sensors and other eSource solutions. Additionally, the increasing cost of Clinical Development and the need for greater predictability of outcome requires the use of more complex study designs and risk-based approaches to clinical study execution. Finally, solutions based on Natural Language Processing (NLP), Artificial Intelligence (AI) and Machine Learning (ML) are maturing rapidly in other industries and opening the door to meaningful and life changing opportunities. There is no turning back. CDM needs to seize the opportunities offered by recent technological advances and regulatory changes to emerge as a strengthened discipline making a positive and meaningful impact on Clinical Development.

## 5. Drivers of change and their impact on CDM

### 5.1. Clinical Research approaches and their rising complexities

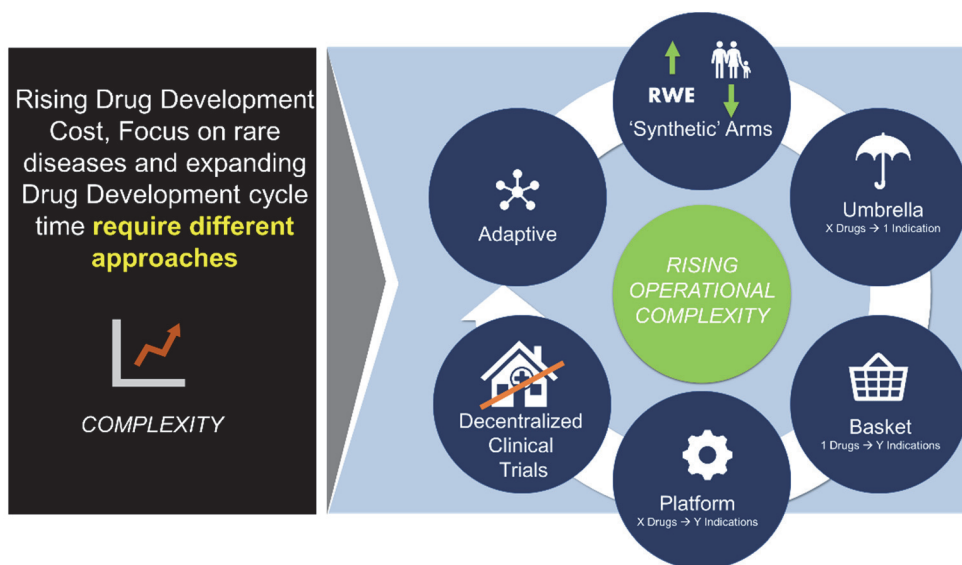
The rising cost of healthcare has an unsustainable trajectory and the costs of developing **one** new drug have also skyrocketed to \$2.6B to “complete the journey from initial discovery to the marketplace”<sup>3</sup>. In addition, “Sponsors spent an estimated \$86 billion on all contracted R&D services during 2018, surpassing internal staff and infrastructure spending by nearly \$20 billion”<sup>4</sup>. In addition, the pressure from patients, payers and sovereign governments to reduce the cost of health care is driving the need for outcome-based pricing with clear evidence (i.e., data) on the clinical benefits to patient vs. the cost of therapies.

At the same time, the Science continues to evolve. Some bio-pharmaceutical companies are finding innovative precision therapies treating conditions based on individual genetic structures (i.e. Gene Therapies). “As more diseases are being redefined based on genomic subtype, researchers have more novel targets and more opportunities to precisely modulate or even repair the basic biological drivers of illness”<sup>5</sup>. “Between 1989 and 2015, 2,335 clinical trials related to gene therapies”<sup>6</sup> have been initiated requiring complex study designs and different approaches to manage clinical data.

So, our industry cannot continue to operate in the same traditional manner and needs “to identify innovative trial designs, evaluate the role of decentralized clinical trials ... that can enable trials to generate reliable evidence needed to assess product safety and efficacy more efficiently”<sup>5</sup>.

Additionally, the diminishing pool of trial participants in certain therapeutic areas requires new and innovative approaches to engage, recruit and retain patients. So, as individualized therapies become the norm, there will be pressures to bring trials to the patients versus the traditional approach of bringing patients to the clinical sites. Patient-centric Clinical Development is no longer just a buzzword.

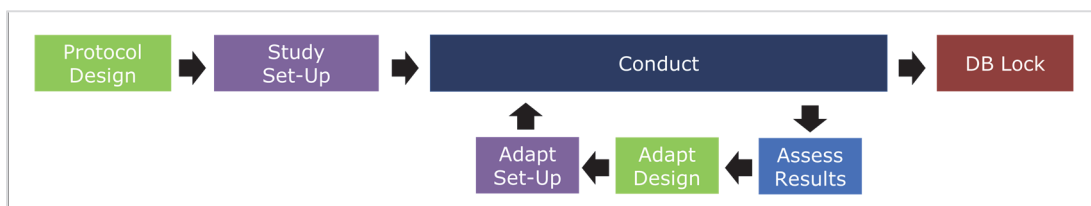
In this section, we’ll explore the evolving study designs and how they impact CDM practices.



**Fig. 1 Rising Operational Complexity**

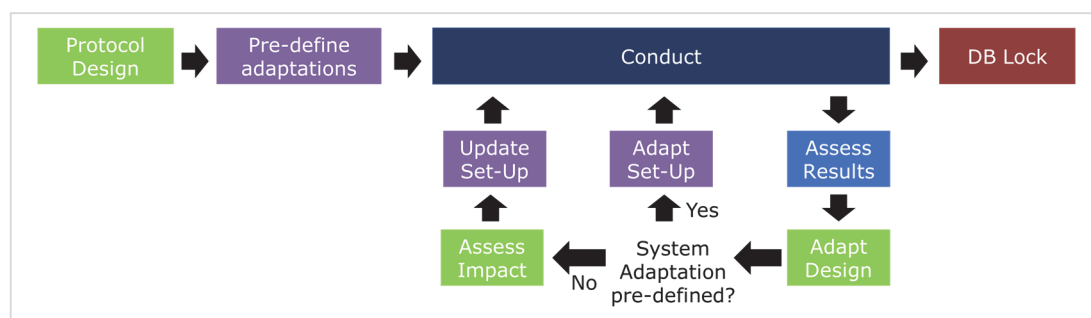
## a) Adaptive Study Design

In its simplest form, **adaptive design** may consist of combining study phases into one protocol (Phase I/II, Phase II/III). However, its true intent is to use data collected in the earlier stages of the study to adapt its design moving forward in accordance with pre-specified rules defined in the study protocol. Examples of adaptations include changes to the dose regimen, study arms, sample size, sub-populations and study duration. Adaptations may apply to new patients being enrolled or all patients retrospectively. In some cases, the outcome of data reviews could result in stopping the study.



**Fig 2. Adaptive Study Design**

Theoretically, Adaptive Design applies to all clinical development phases (i.e. Phase I through IV) and all types of studies. In reality, the current limitations of the traditional clinical development processes and technologies are restricting its adoption due to their lack of flexibility. As an example, most IxRS systems cannot change treatment arms, dose and/or block allocations without going through lengthy programming changes. As a consideration to cope with these many adaptation scenarios, CDM may proactively pre-program all possible adaptations defined in the protocol before study start. This seemingly “proactive” approach would however require designing all data collection tools including IxRS with every complex branching logic which risks extending study initiation cycle times. As an alternative, CDM could minimize the risk to start-up timelines as well as optimize the design of the data collection methods by applying a risk-based study execution (a.k.a. RBx) approach and only pre-program the most likely adaptations. However, it is still possible that data reviewed during the study leads to unanticipated adaptations. In these cases, it may become necessary to put the study on hold until these systems are updated.



**Fig 3. Example of workaround to processes and systems limitation**

Overall, the study and data flow set-up of adaptive design protocols require a deep understanding of the strengths and weaknesses of the processes and systems to anticipate adaptations. CDM needs to lead the study team in assessing how to implement data collection tools allowing patients, sites or countries to follow different Schedule of Visits and Procedures.

---

b) Study designs to accommodate multiple investigational products and/or indications

Traditionally, Clinical Development consists of clinical studies going through sequential phases (I to IV) where **one Investigational Product (IP)** such as a Drug, Biologic or Device is evaluated to assess its safety and efficacy in **one indication**. This is a costly and lengthy endeavor. It often takes more than a decade to gain marketing approval for a new prescription medicine<sup>3</sup>. So, to expedite Clinical Development and reduce cost, new study designs are emerging as alternatives to the traditional linear study phase approach.



The **Umbrella** design where **multiple IPs** are being tested in **one indication** may be perceived as the least disruptive to current CDM practices. It is potentially possible to collect similar Safety and Efficacy data across IPs for the same indication. However, some variations of the eCRF may arise from differences across the IPs being tested. As an example, the use of different routes of administration where local site reactions need to be collected for injections and topical IPs, which is not required for other modes of administration. Additionally, the safety profile and duration of effect of each of the IPs may also require longer safety follow-up for one IP compared to the others. Many variations could be handled with traditional data capture systems with flexible designs leveraging branching logics, dynamic forms, derivations, etc. Additionally, to avoid unintentional unblinding, CDM must perform a thorough protocol operationalization assessment (e.g. how to handle local site reactions for a study with Injectables and Systematic Drugs) that would require specific data handling (e.g. use of an unblinded data scientists or unblinding evaluating investigators) and system configurations.



**Basket** design where **one IP** is tested across **multiple indications** addresses some of the challenges of the Umbrella design (e.g. the variation in IP route of administration). However, it introduces new complexities such as handling of different efficacy endpoints. This would affect the design of the eCRF, eCOA and external data streams.



The **Platform** design where **multiple IPs** are tested across **multiple indications** will inherit the complexity of both Umbrella and Basket designs. This design may be more beneficial in Phase I allowing fast screening of IPs.

Additionally, to leverage the strengths to the above designs, some Clinical Research organizations may foresee benefits (e.g. economically, to reduce site burden, etc.) in designing roll-over trials where patients completing pre-defined milestones across all studies within a program are “rolled over” into one single Umbrella, Basket or Platform long term follow-up study.

In all cases, careful attention needs to be given early in the process when performing the risk assessment, to identify critical processes and data as well as designing and testing the end to end data collection and processing tools. It is also important to realize that these three designs will dramatically impact the set-up of the randomization systems with corresponding emergency unblinding procedures.

---

### c) Study design leveraging Synthetic Arms

The use of “**Synthetic Arms**” is an emerging study design where one or multiple study arms are replaced by previously collected data from either clinical studies or Real-World Evidences (RWE). In this scenario, there is a need to generate derived data, sometimes referred as secondary data assets from existing data sources to avoid exposing patients unnecessarily to the study experiments. This design also helps in expediting and potentially saving costs of Clinical Development by reducing the sample size of patients to be physically enrolled. The derived arm is often used to replace the comparator arm.

The challenge of this design resides in generating data to be compared to the remaining/enrolled study arm. To do so, the “Synthetic arm” data need to have similar variables collected at similar timepoints with the quality required to meet regulatory scrutiny. This “Synthetic arm” data needs to be “clean” and complete enough to power statistical analysis. Some additional coding may be required when using RWE previously coded with SNOMED and ICD-10. As no new queries can be generated on previously collected data assets, the role of CDM would evolve from traditional “Data Cleaning” to “identifying, filtering and curating” existing data to make it fit for use.

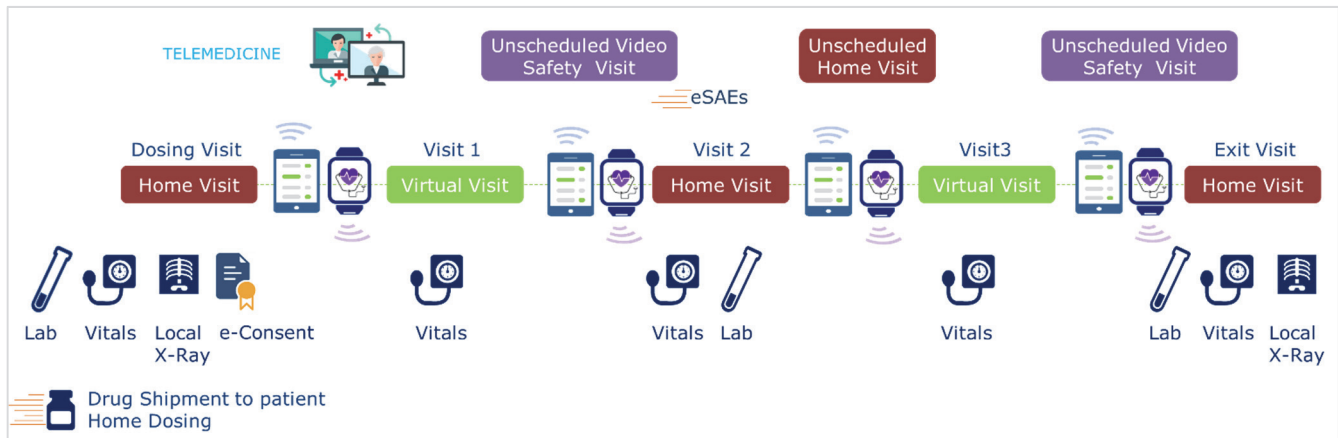
For now, this study design is primarily used for reimbursement studies but could be later expanded to submission studies when this design has been further proven as well as access to reliable RWE has been made possible.

### d) Decentralized Clinical Trials

The **Decentralized Clinical Trials (DCTs)** model is also referred as Site-Less or Virtual Study model. This study design places the patients at the center of the trial with the aim to limit or eliminate the need for patients to travel to an investigational site. This patient-centric model is likely the most disruptive for CDM especially when vastly decentralized (i.e. very few to no site visits). Data processing would be focused on data consolidation from diverse technologies and sources rather than data cleaning. So, CDM should proactively assess data mastering and data reconciliation keys across modalities. Additionally, a risk assessment tailored to DCT could allow the implementation of a pragmatic risk-based study execution approach to data processing.

In this context, traditional EDC is not applicable as all data are collected directly from the patients utilizing multiple eSource data collection modalities such as eCOA, Devices, Wearable, Sensors etc. All patient visits are either remote (i.e. Telemedicine) or conducted through home nursing. A pool of investigators may be available remotely to answer patient’s emergencies. Figure 4 depicts an example of a visit schedule that a patient may follow in a fully Decentralized Clinical Trial. In this context, most of the traditional data cleaning processes are not applicable. Additionally, there are no site monitors to address data related issues on CDM’s behalf. Due to the limitation of addressing data issues after collection, this model heavily relies on technologies collecting error free data with quality checking at the time of data collection. Lastly, CDM needs to wrangle and consolidate the data collected from the myriad of sources to enable the monitoring of data remotely. Pre-defined data handling strategies would be advisable to address illogical data (e.g. incompatible with life) since the source cannot be updated. CDM may also need to define a process to “disqualify and/or flag” such implausible data.





**Fig 4. Example of design of a fully Patient Centric / Decentralized Clinical Trials**

Today DCT is used in a limited number of clinical trials. Some indications are more appropriate than others. Some protocols requiring complex procedures (e.g. Implant Procedures) or necessitating large diagnostic equipment (e.g. MRI) would naturally not be implementable at patient’s home.

Additionally, likely due to the immaturity of current technologies and risk aversion, many companies are only piloting some aspects of fully Decentralized Clinical Trials while keeping core visits and assessments at investigator sites (e.g. Dosing Visit and Exit visits). This may look simpler than a fully decentralized trial, but it is adding complexity by multiplying technologies, data sources and stakeholders. The set-up of such hybrid studies (i.e. partially decentralized) would require careful planning leading to the set-up of multiple systems with complex data flows and integrations.

#### e) Impact of emerging study designs on Clinical Data Management

These emerging and fast adopted study designs are leading to many data flow and study execution complexities that most CDM organizations are poorly equipped to handle today. The challenges are compounded by the lack of adaptive technologies and processes. Many data systems available on the market such as EDC, eCOA, CTMS and IxRS have been designed to handle traditional clinical trials and cannot rapidly adapt to design changes. Many cannot even implement multiple Schedules of Visits and Procedures. In that context, CDM needs to think critically to leverage available technologies, processes and work within the regulations to operationalize the clinical study protocol for data acquisition and management. It is important to anticipate all scenarios as early as possible in the process and proactively identify mitigation strategies. Per ICH E6 Rev. 2, mitigation strategies must include *“the design of efficient clinical trial protocols, **tools and procedures for data collection and processing**, as well as the collection of information that is essential to decision making.”*<sup>7</sup>

This is requiring fundamental changes in skillsets, technologies and processes. CDM must lead the way for study teams to efficiently operationalize these new study designs. CDM must implement “fit for purpose” and “end to end” data strategies to prevent the critical risks introduced by the adoption of these innovative study designs.

---

## 5.2. Impact of regulations on Data Quality and CDM Practices

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

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

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

### a) How do regulations define Data Quality

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

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

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

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

---

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

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

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

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

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

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

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

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

#### b) Risk-based Clinical Data Management Approaches

Risk-based Clinical Data Management finds its origin in the Risk-Based Monitoring (RBM) FDA guidance and EMA reflection paper and is reinforced by ICH E6 (R2). While the first two are suggesting resource optimization in some of the on-site monitoring activities by adopting a Risk-based approach (e.g. focus on what really matters: critical data and processes), the ICH E6 (R2) is reinforcing and broadening the perspective to better address Data Quality expectations. It is offering an interesting and innovative perspective on how we should be conducting clinical trials by inviting sponsors (and CROs) to consider risk-based study execution to enable operational efficiency.

ICH E6 (R2) starts with a brand-new section dedicated to Risk Identification and Assessment (section 5.0). A mission expected not to be just focused on operational matters but on any aspect of the clinical trial execution that may put the entire submission at risk. As a result, risk controls and risks mitigations are expected to be defined, monitored and documented. CDM is at the core of some of these controls

---

and mitigations. So, it is essential that CDM contributes to the Risk Identification and Assessment together with the other key study team players.

It is also worth mentioning that the Risk Assessment process may offer an opportunity to help frame and author the cross functional Quality Management Plan (QMP) and most of the other functional plans (e.g., Data Management Plan and Clinical Monitoring Plan). Indeed, as the study team goes through the risks and defines controls and mitigations (together with responsibilities and frequency of actions), it is likely that the outcome of such assessment provides most of the elements that are present in all plans such as the Data Management and Data Review plans. Technologies might help to facilitate the authoring and synchronization of these essential documents and avoid duplication of work while ensuring alignment with the overall QMP.

Across the ICH E6 (R2) guidance, we also see the growing role of data analytics to support the detection of atypical patterns. Sections 5.0.4 and 5.0.7, related to Risk Control and Review, are suggesting the need to use analytical techniques such as Key Risk Indicators (KRIs) and/or Quality Tolerance Limits (QTLs) to measure the risk during the conduct of the trial and detect anomalies. The CRO oversight addition to section 5.2 and the new central monitoring role supported by statistical techniques defined in section 5.18.3 are suggesting that we should go beyond just a “supervised” approach and enlarge the spectrum of data to be monitored. In other words, the risk controls are there to monitor the risks that the study team thought may happen, however their limited number make it impossible to address every possibility. So, even though challenging to implement, an approach where analytics are used to monitor **all collected data** near real time, may help “guarantee” that most if not all issues are detected.

To support these new analytical methods, the study team needs new skills and critical thinking. Skills that likely reside, as indicated in section 5.18.3, in Data Management or Biometrics functions. A new role, the centralized monitor, that complements the site monitor during his/her investigations is also emerging in some companies.

### c) Impact of local regulations to CDM

Changes to global regulations such as ICH E6 generate a lot of attention from Clinical Development stakeholders as Good Clinical Practice defines our foundational principles. While less visible, local regulations may still have significant impact to Clinical Development including CDM activities with the risk of local inspection findings affecting global studies and potentially submissions.



As a first example, new **local German requirements** were introduced in January 2018 regarding the “timely” review of the CRF data by “medically qualified” personnel and the endorsement of the data by the Primary Investigator (PI) at regular/critical timepoints. Some companies may decide to apply these requirements to German sites only; others to all sites. Whatever the decision is, the ability to implement, monitor and demonstrate adherence to these requirements may depend on the flexibility of the EDC system. CDM would likely have to adjust the set-up of eCRF signatures in EDC, document the approach in the Data Management Plan and potentially create custom reports to help monitor the “timeliness” of the review by the Investigator as attested by their signatures on the eCRF. The implementation strategy must be tailored by each company and CDM has a key role to play in ensuring the compliance to those requirements.



As a second example, the **Chinese Reform on Leading PIs** for Medical Device released in October 2017 further illustrates the impact of local regulations to CDM. The requirement is that Chinese device studies involving more than three (3) sites must have a leading Site/PI assigned. Per local regulations, the leading PI is considered responsible for “clinical data administration and analysis” which includes CDM, Statistical Analysis and Report Writing. Even though individual company interpretations may differ on this, there is no dispute of the fact that the lead PI needs some level of oversight of CDM activities across all Chinese sites. To an extreme, oversight may be understood by some companies as the need for the lead PI to approve CDM Documents (e.g. eCRF Specifications, DM Plan) and/or to manage a CRO performing the CDM activities. To complicate the situation further, this local regulation does not apply to Drug studies conducted in China.

The goal of this reflection paper is not to provide guidance on how to implement local regulations but to raise awareness of the rising complexity for CDM. Local regulations are now going beyond the typical data privacy regulations that CDM has been mostly addressing to date. CDM needs to work closely with regional regulatory and operational teams to assess and implement solutions that comply to local regulations because study level processes and data handling conventions may differ by country.

### 5.3. Evolution of technologies

There are strong desires in the industry to leverage innovative study designs to bring life changing therapies to patients as well as scaling-up patient centric clinical development. These significant shifts have dramatic impacts on the operational aspects of study conduct including the underlying technologies used today. Unfortunately, organizations are struggling to operationalize and scale-up emerging technologies. Reasons include risk aversion, lack of maturity of these technologies as well the additional costs of change management and validation. Additionally, data flow architectures have become highly complex. On a positive note, automation capabilities and opportunities to extract meaningful operational value from data are demonstrating unprecedented potential and the prospects of an impactful change is within reach. At the end of the day, technology must be meaningful and enable what the new era of clinical development demands!

#### a) Reduction of EDC centricity moving forward

Traditional EDC is becoming one data capture modality amongst many. The need to collect source data directly from the patient is already leading to the rapid adoption of eCOA, wearables, sensors and other eSource solutions. As an extension to the 2018 Tuft survey on “Strategies from Data Management Leaders to Speed Clinical Trials”, *Veeva completed an EDC Survey showing that “97% of companies expect to use more clinical data from a wider variety of sources and 70% plan to use a data source that they are not currently using today”*<sup>9</sup>. Furthermore, several companies informally surveyed during the SCDM 2018 Leadership forum stated that over 70% of the data volume (e.g. lab, eCOA, etc.) is **not** coming from EDC. Additionally, more protocols are utilizing objective and measurable endpoints (e.g. automated imaging reading) instead of relying on investigator or rater assessments. People dependent assessments are more expensive and riddled with variability due to subjective interpretations and/or lack of experience. As an example, “discrepancy rates between 2 radiologists have been found to be around 30%”<sup>10</sup>. As data sources are becoming more diverse and prevalent, the 3<sup>rd</sup> party data volume is outpacing site-based data capture in EDC. With that, CDM priorities and focus need to be adjusted.

---

## b) Artificial Intelligence (AI) based applications

The use of “Artificially Intelligent” applications leveraging Natural Language Processing (NLP), Machine Learning (ML) and Robotics is fast growing outside our industry. These AI “*based technologies have the potential to transform healthcare by deriving new and important insights from the vast amount of data generated during the delivery of healthcare every day*”<sup>11</sup>. Intelligent applications have long left the Silicon Valley computer labs and have entered many homes. Interactive systems like Alexa, Apple Siri and Google Home have gained adoption and trust. This evolution is primarily fueled by the accumulation of huge amounts of data associated with strong computing power. This is enabling the training and validation of complex ML algorithms.



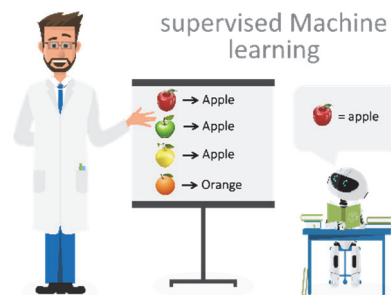
It is technically possible today to automate repetitive and simple tasks as the cost and ease of implementation of AI technologies is becoming attractive.

**Chatbots** are intelligent “virtual assistants”. Leveraging text messaging, Chatbots can partially replace an actual support person. Based on ML and NLP, they offer a “human-like” interaction between people and machines. As an example, it is technologically conceivable to develop chatbots to provide updates on study activities (e.g. Ask a CDM Chatbot to “*provide the counts of sites with more than 5 pending queries on study A*”). You could also easily improve the user experience by adding voice recognition instead of typing text (e.g. using Siri for Apple devices, Cortana for Microsoft, etc.). Chatbots could radically change the study team dynamics and reduce the time many CDM professional spend in relaying information to their peers in the Clinical Trial Team.

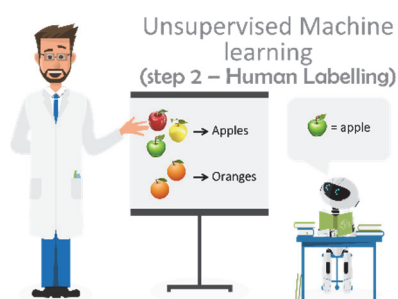
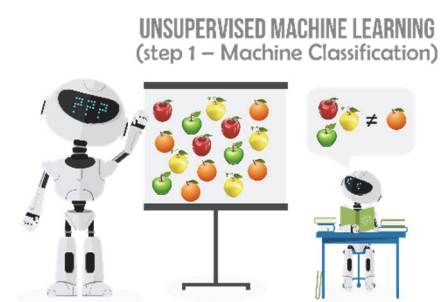
**Robotic Process Automation (RPA)** is not new but is now becoming simpler to implement and a cost-effective solution. RPA enables automation through configurable software that simply mimics well defined human actions. Unlike using scripts on the backend to “automate” manual tasks, RPA is enabling virtual robots to do predictable and repetitive human activities. As an example, a Virtual Robot could be fed with the details of external data reconciliation errors. The Virtual Robot could then login in EDC with its own account (e.g. Login: CDM Robot) and post the corresponding queries. The advantage of such a method is that it does not require changes to EDC (e.g. no integration needed). It leverages EDC traceability (Login logs, Audit Trail, etc.). You could have as many virtual robots as you need working 24 hours a day and 7 days a week to manage the study workload. The same technology could apply to other simple CDM tasks. Virtual robots can become an unlimited virtual CDM workforce. However, RPA requires programming, full understanding of the domain being automated and may be somewhat limited in its scope of applicability. Last, as RPAs are sensitive to User Interfaces (UI) layouts (i.e. RPAs must be adjusted if the UI is changed), CDM need to consider the stability of the systems involved in task automation or anticipate updates to RPAs when systems are upgraded.

**Machine Learning (ML)** uses techniques that learn to make predictions or decisions from data without explicit programming. The ability for “ML software to learn from real-world feedback (training) and improve its performance (adaptation) makes these technologies uniquely situated among software”<sup>11</sup>. There are different types of ML including supervised and unsupervised learning.

- **Supervised learning** tools are initially fed with training data (i.e. the input / question) and their interpretation (i.e. the output / answer). The algorithm can learn from human experience through “training data” and make future decisions and predictions (i.e. answer similar questions from new data). Once the system is trained, its accuracy can be verified using test data (i.e. data without the answer to the problem). Training data can be adjusted as necessary until the desired learning is achieved. Overall, the easier the problem is to solve, the less training data you need. As an example, consider decisions on when to raise a query. Most cases are simple such as when data is out of range, then raise a query. So, you can train the system to do so with limited training data. From our CDM experience most scenarios are not as simple. If data is out of range at a visit, you may want to check if the same issue was queried at a previous visit. If it was, did the site confirm the data “as is” due to a patient’s medical condition? If yes, you would not query again. You would likely close the query and enter a comment. This behavior could be trained too but would require more training data and time.



- **Unsupervised learning** tools are fed with data without their interpretation. In that case, the algorithm can only determine which data are more similar to another. Unsupervised learning is trying to understand how things work. The system can learn through observations (not through experience). In the scenario of query generation, CDM could feed the system with query history without expressing the ideal behaviors. The system would process all the information to determine the different query handling scenarios. The system would require enough cases (i.e. data) to correctly learn by itself. Scenarios would then be confirmed and labelled by humans (e.g. one scenario being about out of range data).



In simplistic terms, a human can train a system to recognize apples and oranges using supervised learning. When unsupervised, a machine can learn by itself to differentiate apples from oranges without knowing what an apple or an orange is. From here, a human has to “label” what classified group defines oranges and which one defines apples to complete the learning process. The unsupervised process is ultimately easier to implement as it does not require the pre-definition of training data which could be complex in Clinical Research. Today, a robust form of unsupervised

learning, called “deep learning” can successfully classify patterns. As an example, deep learning can robustly analyze images to make medical diagnosis (e.g. Classify images with vs. without disease characteristics such as nodules).

---

Without a doubt, unsupervised learning systems especially those using deep learning, have proven their ability to make sound predictions. Unfortunately, they cannot yet provide the reasoning behind their predictions. This is making their use for regulated activities challenging. These techniques need careful review of the classification generated by the algorithm. It is however expected that learning algorithms will be able to explain reasoning back to humans in an understandable way in the very near future. At that stage, CDM will then be able to consider more regulated scenarios where humans would only verify the prediction(s) determined by machines.

Lastly, it is very important to understand that ML is designed to learn and adapt. Like humans, ML Tools will make decisions tomorrow that are better than the one they made today. So, CDM needs to carefully anticipate the implication of evolving systems in our regulated environment when reproducibility of results is expected. This has prompted FDA *“to reimagine an approach to premarket review for AI/ML-driven software modifications”* <sup>11</sup>. While focused on Software as Medical Device (SaMD), this FDA discussion paper highlights the need for new approaches with AI. We also need to realize that ML learning could be biased by the training datasets or if human supervisors are ‘just’ following the decisions suggested by the algorithm which would re-enforce the biased behavior. CDM will need to leverage its risk-based data management strategy when implementing such disruptive technology to ensure reliable and ethical (i.e. unbiased) decision making.

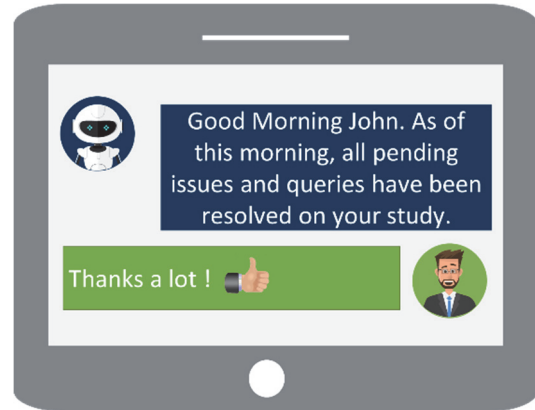
**Natural Language Processing (NLP)** is used by machines to process human language to extract specific information from documents, to generate text from voice (or vice versa), to understand meaning, etc. Today, machines can accurately carry out simple requests, ease web searches, summarize documents or translate languages. NLP is one of the key enabling technologies behind chatbots and spam email detection. While NLP is almost 70 years old, it has tremendously progressed over the last decade by adopting machine learning techniques instead of solely trying to automate grammar rules typically full of exceptions. Apple Siri released in 2010 and Amazon Echo (i.e. Alexa) released in 2015 are two examples demonstrating this technological breakthrough, which are now part of the daily routines of many people across the globe and can manage many different languages. Looking beyond these “simple” use cases, NLP can offer many opportunities to Pharmaceutical Companies.

- NLP can be leveraged to *“extract the names of drugs, diseases, patients, and pharma companies using rule-based or statistical method”*<sup>12</sup> in order to support Pharmacovigilance activities as a vast amount of information about Adverse Events (AEs) reside in unstructured narratives. Similar methodologies could be applied to [identify potential AEs from comments](#) captured in EDC, eCOA or other systems which is currently labor intensive and subject to human errors.
- NLP could also automate [data extraction from documents](#) (including from legacy Clinical Study Reports of compounds acquired from other companies or even Monitoring Visit Reports) and load them into structured databases. NLP has the potential to ultimately replace traditional Extract Transform and Load (ETL) technologies and provide capabilities never seen before.
- Finally, some industries are using NLP to [support regulatory compliance](#). Organizations are extracting key concepts from regulations and verifying that those concepts are reflected in contracts and Standard Operating Procedures. NLP can also be used to review large volumes of documents and search for potential liabilities which could prevent issues in the context of Mergers and Acquisitions.



---

In conclusion, there are numerous opportunities for technology-enabled Intelligence combining “humans” and “machines”. Augmented Intelligence is often the preferred term rather than Artificial Intelligence. We could foresee a Clinical Data Management world built on a *working model that includes virtual Clinical Data Managers working alongside Human Clinical Data Scientists*. AI could enable humans to achieve a different level of reasoning not otherwise possible. In this context, computers and humans working together could make better predictions than either group of humans or group of machines could do on their own.



### c) Blockchain

**Blockchain** is a promising emerging technology yet to be used meaningfully in Clinical Research. As its name suggests, a blockchain is a chain of interconnected blocks (i.e. data). Each block contains a time stamped and non-modifiable version of data. Blocks can store different data types including files, images and time sequenced data from sensors. Each block has its own encryption key and a link to its previous version of the data. Users can therefore retrieve the full version history of the data whenever it is needed. However, Blockchain offers more than the functionalities of a bullet proof audit trail.

By nature, a blockchain is distributed in a peer to peer network where each peer joining the network has a full contemporaneous version of the blockchain (i.e. data sharing). If a member of the network is changing the data, a new block is created and sent to all peers in the network so that all members have the same synchronized data with its full history. There is no more original vs. copy of the data. Blockchain allows for a systems agnostic data management approach. All systems can read or modify the same data and all systems are constantly aligned. We could theoretically trace data from its inception in an EMR system to its review by a regulatory agency in a submission package.

In addition, each block could be linked to a digital contract. The contract would enforce what can or cannot be done to the data. One could consider attaching the Informed Consent to the data and limit the access and use of that data to only people authorized by the Informed Consent (e.g. patients, sites, Contract Research Organizations, Central Laboratories). Any member of the contract could be given access to all the data or just a subset of it. No one could modify the data without a contract or without other data subscribers being unaware of it.

Pharmaceutical companies are just starting to understand the potential of blockchain in Clinical Research. In addition, its implementation would require significant changes to the technology infrastructure that most systems use today. Nonetheless, Blockchain and AI are believed to be two of the most impactful upcoming technologies which could radically transform how we do Clinical Data Management.

---

#### d) Sensors and Wearables

The growing availability of affordable and reliable mobile Health (mHealth) technologies including wearables and sensors is driving the interest of many pharmaceutical companies. The need and the appetite to collect data directly from patients is rising. The recent evolutions include the desire to:

1. Support the implementation of remote trials,
2. Ensure ongoing safety oversight,
3. Collect different objective endpoints, and
4. Reduce patient burden by leveraging available data and technologies (e.g. applying BYOD beyond mobile phone to include smart watches and home intelligent devices).

Some CDM organizations may see the collection of continuous and large volumes of data as a costly and unsurmountable challenge. Other organizations see it as a robust way to collect real-time, untampered and high-quality data eliminating issues from error prone manual processes. Regardless, CDM should establish a clear and proactive strategy. Below are some considerations:

- Sensors and Wearables data are **eSource** by nature since they are collected directly from the patient and stored electronically first. As stated in the SCDM eSource white paper, one of the seven principles of eSource is “Control for Quality”. While “in the world of electronic records, it is possible to control the integrity of data seen or modified by many parties over the life time of the records” it “requires a very thoughtful implementation of systems of controls working together”<sup>13</sup>. CDM must establish “the data chain of custody (e.g., understanding how the data are generated), how the data are connected to other devices or networks, and who has access to the data after it is generated and stored on a device or server before it reaches the data management team”<sup>14</sup>.

It is also worth acknowledging that not all eSource are equal. As an example, sensors and wearables are very different from eSource collected from eCOA. Like EDC, eDiaries and eCOA offer online edit checks upon entry. Some companies are using web-based forms as eDiary and eCOA back-ups in case the device is malfunctioning. None of that is applicable for sensors and wearables. There is no way to go back in time to collect lost data. In cases where sensors and wearables are critical to the study, it may be advisable to ensure that patients have rapid access to back-up device(s).

- Sensors and wearables generate **high volume of data** (millions to trillions of times more than EDC) at **high velocity** (i.e. generated continuously multiple times per second). In this context, **traditional CDM processes would not be viable**. No organization would agree to provide thousands of times more resources and budget to manage cleaner data collected directly from the patients. However, it is expected that CDM monitors the data flow efficiently to rapidly identify safety and data signals.

An implausibly low pulse of 12 recorded on the eCRF would be handled by edit checks. Needless to say, CDM must implement more sophisticated data monitoring tools to identify sudden changes in data patterns from wearables. One missing wearable value out of hundreds of thousand on a given day is likely insignificant. A sudden rise in a resting pulse up-to 120 beats per minute that was previously recorded as 85 may be a safety risk or may be explained by the fact that the patient ran across the parking lot prior to arriving for the visit. It is therefore important to define up-front escalation paths and alert levels when managing fast paced and high data volume.

While those are a few considerations, it highlights the need for different approaches and technologies.

---

## e) Impact of new technologies on Clinical Data Management

Opportunities offered by emerging technologies have the potential to revolutionize Clinical Development and dramatically change CDM at its core. Those will enable Clinical Data Scientists to proactively contribute to study designs via creative and innovative data capture methods. These opportunities will fuel the evolution of Clinical Data Management to Clinical Data Science by automating repetitive tasks like query management and surfacing impactful data issues truly compromising the scientific integrity of the clinical trials. Clinical Data Scientists will have to lead root-cause analysis and come-up with potentially complex remediations. Eventually, Clinical Data Scientist SMEs will continually “train” and monitor expert systems to improve their accuracy in detecting issues. Eventually, some systems might autonomously act to prevent further risks. Clinical Data Scientists powered by intelligent systems and assisted by Virtual Clinical Data Managers will handle the rising complexities stemming from the new study designs and omni channel data collections systems (i.e. from diverse sources and systems).

## 6. Evolution of the Clinical Data Management Role

The people aspect of the changing landscape cannot be underestimated. As the new data sources and supporting technological advances are taking a foothold in the Clinical Research domain, the role of the Clinical Data Manager will become much more complex. This change is at the same time a blessing and a challenge. It offers new opportunities to data experts who have been viewed as the central steward of clinical data quality - the Clinical Data Manager. However, it is a dramatic shift leading to the need to upskill and prepare CDM professionals. Overall, this is a unique opportunity to re-shape our identity for years to come.

### a) CDM at the Cross-section of Risk-Based Study Execution

With the adoption of risk-based principles articulated in ICH E6 (R2), CDM is at the cross-section of fit for purpose (i.e. Risk-Based) study execution. As examples, Clinical Data Scientists will need to:

- **Proactively Manage Risk:** The Risk Assessment must go beyond the typical risks impacting study timelines, deviation from data standards or identifying critical data and processes. Risk Management must start with risk prevention (i.e. Quality by Design) by identifying threats prior to the first patient being enrolled into the study. So, Clinical Data Scientists must assess risks associated with protocol design, study-set, country involved, profile of sites selected, deviations from Standard Of Care and any other study execution activities which have the potential of leading to errors that could negatively impact the credibility and reliability of the trial results.
- **Identify systematic or process driven data issues** including those stemming from trial-design and study conduct factors such as rate of enrollment, technologies used, etc. The key will be to efficiently and reliably monitor such risks through the holistic review of **all** clinical and operational data (i.e. finding data patterns and anomalies across studies, countries, sites, and patients).
- **Manage risk associated with rather complex data flows** resulting from disparate data sources.
- Understand **risk-based approaches** when conducting clinical studies in a **patient centric and decentralized environment**.

---

## b) Foundational Clinical Data Management Competencies

While the role is evolving, several of the foundational competencies for a Clinical Data Manager remain the same. For example, the following are essential building blocks:

- Attention to detail
- Therapeutic area knowledge
- Communication skills in articulating complex data findings to the trial teams
- Systematic data review and trending
- Project Management
- Design of data collection tools

## c) New and Refined skillset

CDM Leadership and Subject Matter Experts must proactively guide the transition of their staff from Clinical Data Manager to Clinical Data Scientist. Organizations must consider how to best combine or split responsibilities across role(s) to leverage internal expertise and talents. Additionally, while this already exists in some companies, CDM organizations must consider the best approaches to centralized data monitoring. This will require the use of data review and analytical skills to aid site monitoring teams to efficiently run onsite and remote site monitoring activities.

Ultimately, to remain the most effective and relevant, CDM professionals will need to build on the core CDM skillsets and focus on emerging opportunities offered by technology, regulations and Clinical Development strategies.

Some of the emerging skillsets include:

- **Robust critical thinking and process knowledge.** The nature of issues identified through advanced analytical capabilities will lead to root cause analysis and corrective actions. These will mostly entail adjusting processes moving forward to prevent re-occurrence as opposed to just correcting data.
- **Broader cross functional collaboration.** For example, Clinical Data Scientists will have to consider the recruitment strategy and outcome of study, country and site feasibility. Clinical Data Scientists will have to tailor data collection systems, data review strategies, and training requirements (at study, countries, sites and patients level) that consider the risks associated with the expected patients and sites diversities (e.g. geographical, cultural, experience, etc.).
- Ability to align the flows of data with the need of the **next generation clinical protocol** so that:
  1. Data can be collected to “ensure human subject protection and the reliability of trial results”<sup>4</sup>,
  2. Existing data can be made fit for use to meet protocol endpoint such as generating trustworthy evidences from Real-World Data (RWD).
- **Deep knowledge of data** including the characteristics of different types of data, such as EHR data from inpatient vs. outpatient from a biorepository. Understanding the implications of data context, quality, source, amount, and workflow<sup>14</sup>.
- **Advanced analytical and technical skills** to interrogate and mine high volumes of data from a variety of data sources. Clinical Data Scientist must drive the development of tools extracting meaningful insights to detect potentially unreliable data threatening the validity of the trial results.
- High level **understanding of Artificial Intelligence methods** and scope of applicability.

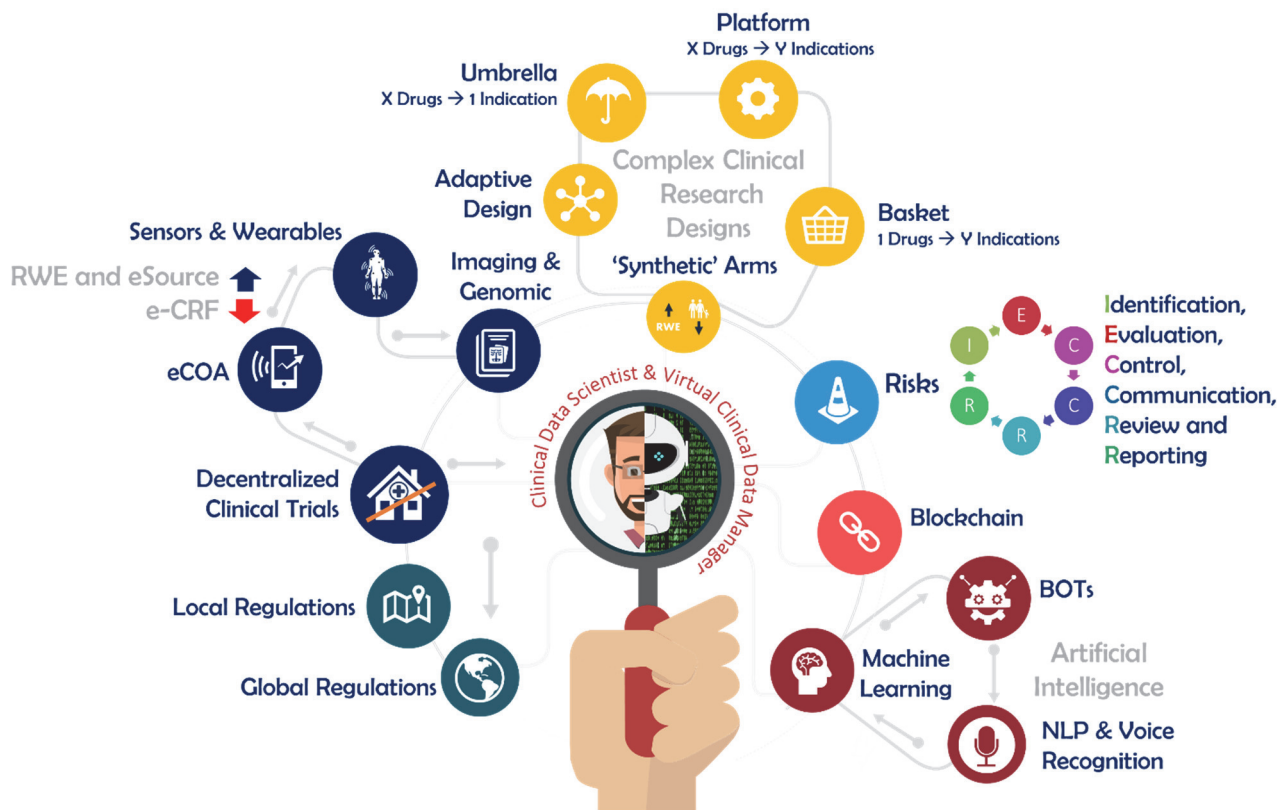
- Ability to help build and test machine-learning algorithms to detect patterns of missing data, outliers, trends and/or lack of variability which can be applied as a standard within their therapeutic area or set of studies with or without human supervision.

## 7. Conclusion

The drivers for change explored in this reflection paper clearly illustrate the necessity for CDM to keep pace with the Clinical Research industry evolution and anticipate the downstream impact on the overall CDM and health development processes down to the study level. The rise of big and complex data stream, the availability of innovative technologies, the maturity of Artificial Intelligence, the adoption of new study designs and the evolutions of regulations are already starting to reshape what CDM means today. The divide between clinical trial data and Real-World Evidence is collapsing quickly. These changes are no longer buzzwords used to attract people at conferences, but a reality we all must tackle.

It is no longer possible to blindly apply a “one size fits all” approach and continue with our current approaches. No one could expect SOPs to pre-define all data processing variations resulting from the factors highlighted in figure 5. We are entering an exciting era where quality by design, critical thinking, risk-based and fit for purpose approaches will prevail. Our mission is to set a roadmap toward **Clinical Data Science** which requires the evolution of our skillsets, processes, technologies and best practices.


In this data and patient centric framework, CDM will play a strategic role in ensuring the reliability of the trial results and support the transformation that Clinical Research needs.



**Fig 5. The rising complexity of the CDM Role**

---

## References

- 1 KMR, 2018 Data, <https://kmrgroup.com/>, CYCLE TIME METRIX 
- 2 TransCelerate, Sep 2014, Evaluating Source Data Verification as quality Control Measure in Clinical Trials. Available at <https://journals.sagepub.com/doi/pdf/10.1177/2168479014554400>
- 3 Biopharmaceutical Research & Development, May 2015, The Process Behind New Medicines, Available at <https://www.phrma.org/report/biopharmaceutical-research-and-development-the-process-behind-new-medicines>
- 4 Tufts, IMPACT Report March 2019, Drug Development Outsourcing Outpaces Internal Spending but Remains Tactical and Reactive, According to Tufts Center for the Study of Drug Development. Available at <https://csdd.tufts.edu/impact-reports/>
- 5 FDA, March 2019, “Statement by FDA Commissioner Scott Gottlieb, M.D., on new strategies to modernize clinical trials to advance precision medicine, patient protections and more efficient product development”. Available at [https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm633500.htm?utm\\_campaign=031419\\_Statement\\_FDA%20issues%20statement%20on%20new%20strategies%20to%20modernize%20clinical%20trials&utm\\_medium=email&utm\\_source=Eloqua](https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm633500.htm?utm_campaign=031419_Statement_FDA%20issues%20statement%20on%20new%20strategies%20to%20modernize%20clinical%20trials&utm_medium=email&utm_source=Eloqua)
- 6 Eve Hanna, Cécile Rémuzat, Pascal Auquier and Mondher Toumi, Jan 2017. Gene therapies development: slow progress and promising prospect, Available at <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5328344/>
- 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](https://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6/E6_R2_Step_4_2016_1109.pdf)
- 8 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](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/687246/MHRA_GxP_data_integrity_guide_March_edited_Final.pdf)
- 9 Veeva EDC Survey, 2018, Available at <https://www.veeva.com/edc-survey/>
- 10 Rajeshwari Sridhara, Sumithra J. Mandrekar and Lori E. Dodd, May 2013, “Missing Data and Measurement Variability in Assessing Progression-Free Survival Endpoint in Randomized Clinical Trials”. Available at: <http://clincancerres.aacrjournals.org/content/19/10/2613>
- 11 FDA, 2019, Discussion and Request for Feedback, “Proposed Regulatory Framework for Modifications to Artificial/Machine Learning [AI/ML]- Based Software as Medical Device (SaMD)
- 12 Marco Lagi, Natural Language Processing – Business Applications, February 2019. Available at <https://emerj.com/ai-sector-overviews/natural-language-processing-business-applications/>
- 13 SCDM, 201, eSource Implementation in Clinical Research. Available at <https://www.scdm.org/publications/white-papers/>
- 14 SCDM, September 2016, eSource in Clinical Research: A Data Management Perspective on the use of Mobile Health Technology. Available <https://www.scdm.org/publications/white-papers/>

---

## Main abbreviations

AI	Artificial Intelligence
BYOD	Bring Your Own Device
CDM	Clinical Data Management
CTMS	Clinical Trial Management System
DCTs	Decentralized Clinical Trials
eCOA	electronic Clinical Outcome Assessment
EHR	Electronic Health Records
EMA	European Medicines Agency
FDA	Food and Drug Administration
GDPR	General Data Protection Regulation
IP	Investigational Product
IxRS	Interactive Response System (x being any type including Voice or Web)
KRIs	Key Risk Indicators
MHRA	Medicines and Healthcare products Regulatory Agency
ML	Machine Learning
QTLs	Quality Tolerance Limits
NLP	Natural Language Processing
RBM	Risk-Based Monitoring
RPA	Robotic Process Automation
RWD	Real-World Data
RWE	Real-World Evidence
SCDM	Society for Clinical Data Management
SME	Subject Matter Expert