The Evolution of Clinical Data Management into Clinical Data Science (Part 2: The technology enablers)

A Reflection Paper on how technology will enable the evolution of Clinical Data Management into Clinical Data Science

5 March 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 previous SCDM publication, **The Evolution of Clinical Data Management into Clinical Data Science (Part 1)**[1]**.**

In Part 1, the SCDM primarily addressed technology contributions to the evolution of our discipline. This second reflection paper elaborates further and shares insights from leaders, pioneers and early adopters of these emerging technologies.

We hope that it will help all Clinical Data Management (CDM) professionals, from subject matter experts (SMEs) working on clinical studies to CDM leaders setting the direction of their organizations, to adopt new clinical research approaches and technologies while ensuring compliance with evolving regulations.

2. Abstract

CDM is responsible for the life cycle of clinical data from collection to delivery for statistical analysis in support of regulatory activities. CDM primarily focuses on dataflows and data integrity (i.e., data is managed the right way). Clinical Data Science (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. This requires different expertise, approaches and technologies.

Part 1 defined CDS as the strategic discipline that enables data-driven clinical research approaches, provides subject protection and ensures the reliability and credibility of trial results. It encompasses processes, domain expertise, technologies, data analytics and good clinical data management practices, all of which are essential to prompt decision making throughout the life cycle of clinical research[1].

The Evolution of Clinical Data Management into Clinical Data Science (Part 2) provides CDM professionals with pragmatic insights by outlining lessons learned and recommending some tried and tested ways to adopt emerging technologies enabling our evolution toward CDS. These include digital-health technologies which, according to FDA[2], cover categories such as mobile health (m-Health), health information technology (IT), wearable devices, telehealth, telemedicine and personalized medicine. However, only those which have a direct impact on CDM will be covered in this paper.

In each section of this paper, we define the challenges facing us today and focus on approaches to tackle them. We will also seek to reflect and expand on the clinical research industry drivers and trends, and their direct impact on CDM identified in Part 1.

The SCDM Innovation Committee intends to release subsequent papers, including one on the evolution of the CDM role. This subsequent paper will provide insights on how CDM will evolve their skillsets (e.g., technical, soft skills, etc.) and processes to meet the demand of clinical research and regulations by maximizing the potential of available technologies.

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
- Catherine Hall, Vice President, Product Strategy, endpoint
- Joshua Wilson, Executive Director, Biometrics, Strategic Technology Advancement, Syneos Health
- Richard Young, Vice President, Vault CDMS Strategy, Veeva

Content contributors:

- Sanjay Bhardwaj, Global Head, CDM, Allergan, SCDM Board Vice-Chair
- Alice Phung, Global Head, Clinical Programming, Allergan, SCDM Technology Innovation Team Lead
- Laura Trotta, R&D Manager, CluePoints
- Shannon Labout, CEO & Principal Consultant, Data Science Solutions LLC, Former SCDM Board Chair
- Sabrina Steffen, Sr. Director, Innovation, Data Management & Connected Devices, IQVIA
- Mike Novotny, CEO, Medrio
- Lynne Cesario, Global Risk Based Monitoring Program Lead, Pfizer
- Prasanna Rao, Head, Artificial Intelligence and Data Science, Pfizer

Reviewers:

- Linda King, Global eCOA Capability Lead, Astellas Pharma, immediate past SCDM Board Chair
- Jill Notte, Assoc. Director, Strategic Product Solutions, IQVIA
- Grzegorz Cinciała, Director, Clinical Data Management, MDS
- Steve Chartier, Senior Director Engineering, PAREXEL
- 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 CluePoints for providing writing and document QC support as well as the many volunteers who have participated in the SCDM Innovation Committee and have contributed to forming the thoughts expressed in this reflection paper.

4. The technology inflection point

Today, CDM processes and systems find themselves at an inflection point. The age of paper-based processes is reaching near extinction, and the centricity of traditional EDC is in rapid decline.

Twenty-five years ago, clinical research was largely one-dimensional. Patients entering a study were assumed to follow a one straight-line path to completion. Our drug development processes and technologies were designed with that simplicity in mind: One main data source (site generated data), a simple drug development path (Phase I through Phase III), one main data collection tool (the Case Report Form (CRF)) and one key target data consumer (regulators). This was the era of one-size-fits-all and predictable simplicity.

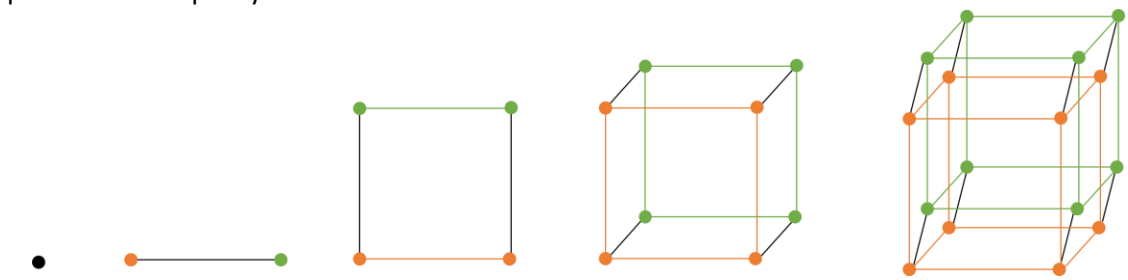


Fig 1: from 0D to 4D

Today, clinical trials are two-, three-, or even four-dimensional and we find ourselves trying to execute complex clinical trials with one-dimensional thinking using the same limited technologies we have always used. But, we must design the study we need, not the study our technology limits us to.

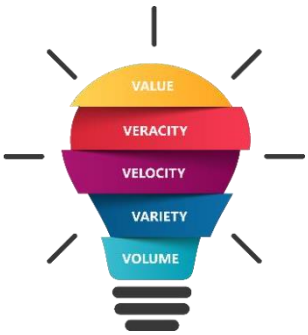
The current impact on CDM is simple to understand: We are finding new and innovative ways to work around the limitations of those technologies. This often results in fragmenting the process and creating niche solutions. As a result, we have further fragmentated systems and processes by building artificial functional silos within organizations, as well as outside (through greater outsourcing) by implementing front and back office strategies often associated with large scale offshoring.

Simply explained, we are bending solutions in ways that they were not designed for, and this approach is no longer viable. The traditional way of collecting and cleaning data can't keep up with the increasing complexity coming from the variety, velocity and volume of information we are collecting.

Perhaps the most important component to consider in the future will be risk-based data strategies. We need to accept that not all data is created equal, and that not all datasets need to be subject to the same levels of scrutiny. It's not about cutting corners but understanding that achieving perfection may require a level of effort that exceeds the rewards. In a dataset of a few hundred observations, a small number of errant recordings can soon change an analysis. But does the same hold true in a dataset measured in the millions, or even billions? Probably not.

Our mission is to reverse this pattern by identifying fit-for-purpose data strategies, optimizing processes and nurturing true technology innovation. If we follow this path, instead of continuing the way we have for the last few decades, we will find our solution.

5. Fit-for-purpose clinical data strategies (the 5 Vs)



The evolution of clinical research practices and supporting regulations, as well as massive advances in technology have fundamentally changed what clinical data is. As we define the future beyond traditional EDC, we need to rethink our approaches and understand how the "5 Vs" of clinical data are re-shaping CDM.

5.1 Volume (Terabytes, petabytes, exabytes to yottabytes)

In 2012, Tufts[3] estimated that in average, phase III studies collected close to 1 million data points. Today we measure m-Health data points in the billions. This demands the adoption of new strategies to improve the collection, processing and archiving of data supporting this new scale. CDM must re-imagine its practices to efficiently move **from a few datapoints per CRF to more than tens of thousands of datapoints generated per patient per week**.

Figure 2 shows the expected volume of actigraphy data generated by wearables (in blue) compared to data generated from site visits (in orange). The protocol requires 260 patients to be treated for 6 months. The enrollment period is estimated to last 6 months. With wearable device set to transmit data every minute, wearables would generate a pulse reading more than 68 million times. In comparison, pulse would only be generated 3,380 times through site visits, assuming patient's visits every 2 weeks.

With the incredible increase in data volume, CDM must be diligent and secure Quality by Design (QbD) by defining what really needs to be collected to support the protocol hypothesis vs. all data that can be generated through new technologies. Not all data generated by devices may be useful for statistical or further exploratory analysis. In the case of wearables, CDM may consider retaining the 68 million pulse readings as e-Source data while only retrieving data summaries at regular intervals (e.g., every hour or day). Data collected may only include key data characteristics (e.g., min, max, average, standard deviation, number of observations generated, etc.), aggregated (e.g., by time reference such as epoch) to better support downstream activities such as safety monitoring, data review and statistical analysis.

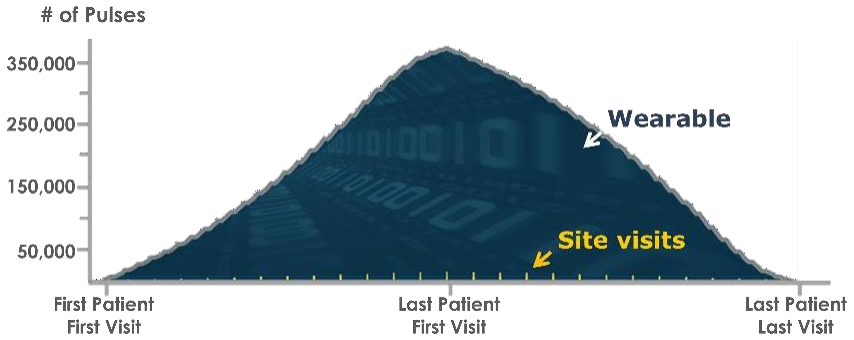


Fig 2. Daily volume of actigraphy data from wearable vs. e-CRF pulse data

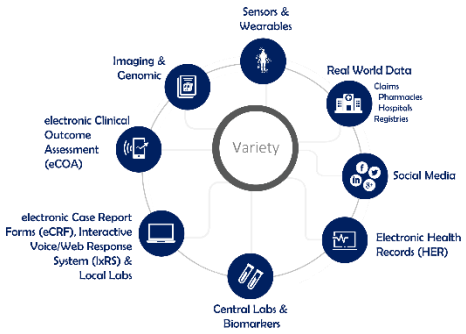
Another example of data expansion is the rapid increase in focus on wellbeing, and the array of passive data sources now being made available for research. According to the IQVIA[4] Institute for Human Data Science, there are currently over 318,000 health apps and more than 340 consumer wearable devices tracking, measuring, monitoring and connecting healthcare stakeholders. Additionally, there are more than 200 new health apps added to app stores every day.

It is therefore not surprising that sponsors are increasingly using digital health technologies in clinical research and leveraging apps to collect reported outcomes and other real-world data (RWD). However, most experiments with digital health have been confined to Phase IV trials, reflecting the perceived risk of incorporating digital measures into pivotal trials until they are validated and pressure tested.

This is unfortunate as those technologies can improve the efficiency of clinical research in many ways. Solutions for identifying sites, targeting and recruiting the right patients, collecting reported outcomes, gaining digital consent, screening patients remotely and conducting decentralized trials have all proven to be effective and useful. First and foremost, they benefit patients by removing enrollment barriers and enabling breakthrough medical advances especially for rare diseases. Solutions such as telemedicine can benefit sponsors by reducing resources, optimizing site selection, speeding up enrollment, easing data collection and supporting rapid decision-making through immediate access to data. Additionally, biosensors, predictive analytics and novel patient assessment media are leading to new discoveries, reducing time to insight and optimizing patient identification. The FDA[5] has released extensive guidance on the appropriate use of such technologies, as well as developing new endpoints vs. establishing equivalence with existing endpoints. To achieve our CDS goal of managing this growing volume of data, we must develop new data science principles, data collection tools and data analytics strategies.

5.2 Variety

CDM is tasked with integrating both structured and unstructured data from a wide range of sources and transform them into useful information. Integrating, managing and interpreting new data types such as genomic, video, RWD and sequenced information from sensors and wearables, requires new data technology strategies and questions the centricity of "traditional" EDC systems. The key questions are where and how, both logically and physically, should these disparate data sources be orchestrated to shorten the gap between data generation to data consumption?



5.3 Velocity

To support real-time data utilization, CDM needs to understand, prioritize and synchronize data transactions into appropriate data storage at increased volume, speed and frequency. Wearables data, for instance, is being generated 24 hours a day, 7 days a week. Taking the example in figure 2, up to 375,000 pulse readings could be generated in a day assuming data transmitted every minute. This would grow up to 22.5 million pulses with data transmitted every second. In a world where real time data is expected, it is not surprising that connectivity has become a core component of software development.

Application Programming Interfaces (API), used for web-based systems, operating systems, database systems, computer hardware, m-Health and software libraries, are enabling automated connectivity in new ways. This is moving focus from "data transfer" to "data integration". Such wide-ranging integration is technically possible, but is it necessary? So, CDM should evaluate the pros and cons for every data integration. We also need to stretch our thinking and expectations, because APIs do not just connect researchers, they provide a platform for automation. Beyond API, attention needs to be given to device selection. Some devices still do not include technology that facilitates real-time data flow. Some may not be Bluetooth or WiFi enabled, requiring the device to be docked, brought back to the site, or even sent back to the device provider to extract data from it and then transfer it to the sponsor.

5.4 Veracity

We can associate veracity with the key attributes of data integrity and particularly ALCOA+ (**A**ttributable, **L**egible, **C**ontemporaneous, **O**riginal, **A**ccurate, **C**omplete, **C**onsistent, **E**nduring and **A**vailable). Veracity can also be associated with some of the attributes of data quality such as data credibility and reliability. In this context, CDM needs to establish pro-active measures to secure the authenticity and security of the data. This is becoming critical in the world of e-Source and RWD where data can rarely be corrected, and where anonymization is increasingly challenging and critical.

Additionally, with the adoption of risk-based approaches, not all data may be subject to the same level of scrutiny. Different quality targets may be acceptable across different data types and sources. CDM will need to not only manage data, but determine and enforce fit-for-purpose data quality standards.

We must, most importantly, focus on QbD. Preventing issues at the source while ensuring the integrity and security of data will require new processes, tools and governance models. As an example, solutions like blockchain will ultimately enable undisputed veracity by ensuring that data cannot be tampered with. The origin of the data will be 100% guaranteed, access controlled by its owner (i.e., the patient) and any change documented in an unalterable audit trail. Unfortunately, the use of blockchain is in its infancy within clinical research. In the meantime, we must implement process and technology strategies that ensure full traceability, security and transparency of the flow of data.

Eventually, we will also need to secure the veracity of data on systems that we do not directly control, such as EHR with their disparate and complex data structures, like genomic data, medical imaging, unstructured data and documents, metadata and sequenced data.

5.5 Value

Importantly, expanding on the IBM 4Vs of Big Data[6], CDM needs to maximize the relative value of any one data point in this ocean of data. Given the rapid influx of information from such a large volume of highly diverse sources in today's trials, achieving this requires management and oversight.

In the context of CDS, value goes beyond integrity and quality. To leverage the full potential of the data we have, we must look beyond its original purpose. During a clinical trial, we collect data to validate the hypothesis of the protocol and, ultimately, obtain market authorizations. Once databases have been locked, most pharmaceutical companies will only re-use them for regulatory purposes (e.g., annual safety updates, integrated efficacy and safety summaries, market authorization in other countries, etc.). However, to unleash the full value of clinical trial data, sponsors must pro-actively anticipate what will be needed in the future and seek patient authorization up-front, through unambiguous informed consent forms.

Some companies are beginning to re-use clinical data in new ways and this has influenced others to seriously consider it. Examples include creating synthetic arms, identifying trends from audit trails, feeding back data to patients during study conduct to boost retention, creating machine learning training datasets, and extracting real world evidence from real world data.

As these examples show, emerging technologies need to be leveraged to extract the full value of data for all stake holders – patients, sites, sponsors, regulators, caregivers and payers – and stop creating data silos.

5.6 Technology impact of the 5Vs

When considering all dimensions of the 5Vs, it is evident that not all data is created equal. Therefore, our data strategies need to be commensurate with the risks, complexity, and value of the data collected. Additionally, data security and personal data protection are key elements that must be strategically anticipated and addressed prior to trial start. If the true value of this data is to be realized, it must be collected and captured in a consistent and timely manner that considers all 5 dimensions.

The table below summarizes the evolution of the 5 Vs from the current CDM to the desired CDS paradigm to support the contemporaneous clinical research needs:

	Current CDM Paradigm	Desired CDS Paradigm
Volume	• 10's to 100's of datapoints per patient per week (i.e., few datapoints per CRF)	• Thousands to millions of datapoints per patient per week
Variety	• EDC centric including local labs and PK • External data mostly limited to IxRS, central labs and eCOA	• Scope expanded to RWD, biomarkers, genomics, imaging, video, sensors and wearables (i.e., sequenced data), structured and unstructured data
Velocity	• Days, weeks and months • Data entered in eCRF days after patient's visits	• Near real time • RESTful APIs providing interoperability between computer systems
Veracity	• Exact copy of source, ALCOA • Mainly confirmed through SDV and queries • Perfect (100% error free) • Manual/scientific Reviews	• Focused on what matters (i.e., critical to quality factor) • Risk-based data strategies • AI-driven automation of issue detection and resolution • Fit-for-purpose (scientifically plausible and strong enough to support reliability of trial results)
Value	• Focused on regulatory submissions	• Broader secondary use (synthetic arms, patient engagement, machine learning training sets, etc.) • Continuous data insights on patients (e.g., safety, behavior, etc.) helping study design by improving sensitivity in measurements and better understanding of the disease to treat.

Ultimately, technologies must allow the implementation of emerging data strategies and enable the consolidation of high volumes of data, continually generated from many sources with complex data structures. CDM will also need to derive the required secondary data assets to extract the full value of our data. The value extraction journey begins with generating insights leading to knowledge, which generates intelligence which can be used for automations, predictions and scientific conclusion.

6. Intelligent Clinical Data Management Systems (CDMS)

To support the 5Vs and be future proof, CDM needs a source and technology-agnostic data collection, consolidation and management strategy looking beyond the transfer of source data to EDC. This demands a new generation of CDMS (including data platforms, workbenches, reporting framework, etc.) which are able to interact with an end-to-end ecosystem of technologies supporting all emerging needs (see Fig. 3). CDMS must also manage ***active*** data from clinical research as well as ***passive*** data from medical care and personal health devices.

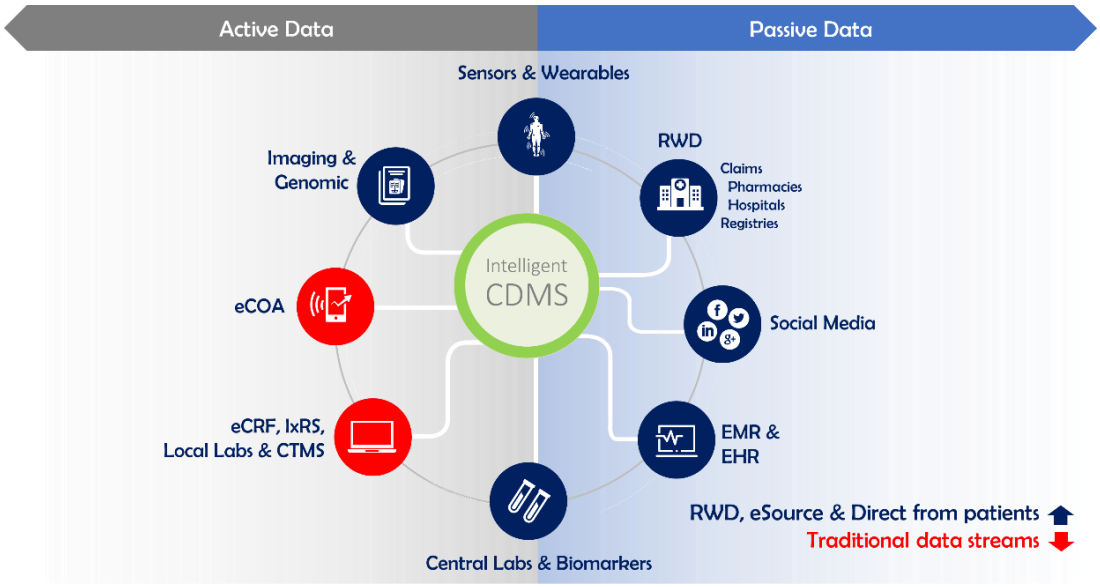


Fig 3. Clinical research technology ecosystem

Active data collection includes case-report forms, instruments and other means of data capture. It is data specifically collected for clinical research purposes and the patients are usually actively involved in its generation. Active data is collected, reviewed, and cleaned using the more traditional methods. However, in a decentralized clinical trial model, the act of querying active data will be different. There may be a need to use patient engagement portals or applications to reach patients directly in some instances. The query process in a decentralized model may not always allow for data corrections, especially if the data collected is being collected directly and considered source data. The data cleaning process may end up focusing less on correcting missing and inconsistent data and more on correcting the presumed behaviors that led to data issues to avoid them in the future.

Passive data collection refers to data that exists in a myriad of systems and generated as a by-product of real-world medical care processes or other patient activities. Often, this data is not collected for clinical research purposes but can be curated and utilized in research. While patients are generally not actively involved in this process, they must consent, either prospectively or retrospectively, to their data being used. Passive data will often not be cleaned using traditional methods since it cannot usually be modified once it is retrieved. Instead, methods such as analytics tools that use statistical algorithms to identify trends and anomalies are more viable ways to interrogate these types of data. This can highlight issues with the way data is collected, device malfunctions, patient behaviors and potentially more. This may then lead to follow-up actions to avoid future reoccurrences of the same issues.

Intelligent CDMS must enable real-time, data-driven and confident decision making from passive and active data to meet the CDS needs. To do so, they will need to manage many data formats, leverage a multitude of APIs and comply with all research and healthcare standards. Only then, they will be able to appropriately handle the data coming directly from all systems (e.g., RWD from health care systems, EHR storing both study-specific and patient healthcare data) and from patients (e.g., ePRO, mobile apps, sensors and wearables whether or not they have been designed to collect only protocol-specific data).

Additionally, with the volume, velocity and variety of data to manage, Clinical Data Scientists will need intelligent CDMS that enable them to interact with the data, rather than just collect and integrate them. Unfortunately, our current CDM views of data are not easily actionable. Today, we commonly look at our CDMS data through analytics, and then must often go back to the source system to perform additional data investigation and ultimately issue a query or correct the data. This is clearly not a sustainable situation. In addition, intelligent CDMS should either offer deep linking into source systems or offer ways to send feedback to the systems of records without additional login-in steps. While doing so, CDMS should retain a complete audit trail of all requests and data changes.

Furthermore, the use of RWD and e-Source at scale in the world of randomized clinical trials will require a fundamental process shift for CDM. The traditional data cleaning and discrepancy management processes will have to be re-imagined to align with these new sources. CDMS will require enhanced technical capabilities to automate intelligent extraction of real-world evidence (i.e., insights) from real-world data. Another important question is how to respond to the detection of a valid data anomaly. The MHRA provided the potential solution of 'data exclusion' in its March 2018 "GxP" Data Integrity guidance. The document states that data may be *"excluded where it can be demonstrated, through valid scientific justification, that the data are not representative of the quantity measured, sampled or acquired. 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"*[7].

It means that platforms built for source-agnostic data consolidation and management must allow for data tagging as well as means to capture data exclusion reasons beyond existing audit trail capabilities. Those platforms must ensure end-to-end traceability of data regardless of the data origin and format. Here, CDS will need to ensure that a valid scientific justification for data exclusion is captured, rather than relying on source data confirmation from the site to close queries. Intelligent CDMS also need to support the smart mapping of disparate data structures and data terminologies (e.g., support ML-based automapping of source to SDTM and MedDRA to ICD 10, etc.). Ideally, CDMS will have an inherent and flexible data schema supporting upstream and downstream data without extensive study specific set-up.

Lastly, we need to recognize that even today, some trials are still executed using paper CRFs. In fact, 32% of companies responding to the 2017 Tufts survey[8] still use paper CRFs. It means technology strategies should cater for ways to integrate and manage all data, including those from the remaining paper-based legacy studies. Ultimately, to shorten the gap between data generation to data consumption, Clinical Data Scientists must develop an ecosystem whereby patients, caregivers and researchers can appropriately share data, regardless of source and format.

7. Maximizing the value of traditional EDC

Overall, EDC was successful in displacing paper CRFs, speeding-up access to data from patient visits and streamlining the query process. Despite these great achievements, its full potential has not been fully

realized. For example, the challenges surrounding the transcription of site source data remains. Source data verification (SDV) functionalities have been implemented to better manage the symptoms, but not address the root cause which would eliminate the need to transcribe data altogether. Most EDC systems have not reduced the need for extensive, resource-intensive data checking processes, such as SDV and complex edit checks, and many implementations simply converted existing, inefficient paper processes to an electronic tool[1]. Additionally, its centricity is being challenged by the rapid adoption of direct-to-patient data captures (e.g., eCOA, sensors, wearables, etc.).

Furthermore, while the pharmaceutical industry concentrated on EDC, the healthcare industry invested heavily in its own electronic health record (EHR) systems, further increasing the technology gap between clinical research and medical care. Fortunately, sophisticated EHR systems, built on common healthcare data standards (i.e., Health Level Seven (HL7)) and interfaces, are now available. They can enable EHR-to-EDC data transfer, and ultimately eliminate the need for transcription and associated SDV[9]. While it is challenging to implement at large scale, this direct connection to EHR to ease clinical research is a real opportunity. Yet, CDM professionals must consider the increased level of complexity involved in integrating and managing these data streams, clearly documenting the data lineage from "collection to analysis" and addressing security and privacy requirements.

Regardless, EDC still has a critical role to play in today's clinical research and its use must be maximized. According to the 2017 Tufts study on the e-Clinical landscape[8], EDC remains the most prevalent clinical data application used, followed by randomization and trial supply management system (RTSM). It means CDM needs to be pragmatic about positioning EDC in a broader spectrum of data sources and maximize its value through a fit-for-purpose EDC Strategy while investing in future-proof solutions.

Below are lessons learned for identifying quick wins to maximize the value of traditional EDC.

EDC is not the optimum place to load external data: EDC systems were not designed to be the central study data repository and should not typically be viewed as the place to load all external data. CDM just used them this way, in the absence of any other viable solution. A total of 77% of companies surveyed by Tufts had issues loading data into their EDC application and 66% identified EDC systems themselves or integration issues as the primary reasons for being unable to load study data in EDC[8]. Additionally, EDC systems have been designed to meet specific requirements such as SDV or manual entry of data that do not apply to external data. It is therefore not surprising that less than 22.5% of the total volume of data in the EDC came from external sources, including but not limited to local lab data (5%), central lab data (5%), quality of life data (4%), and ePRO data (3%). Genomic data and m-Health data made-up the smallest portion of data in EDC systems at 0.4% and 0.3% respectively[8]. Lastly, as mentioned in the initial SCDM reflection paper[1], several companies informally surveyed during the SCDM 2018 Leadership forum stated that over 70% of the data volume (e.g., lab, eCOA, etc.) does **not** come from EDC.

In the absence of alternatives (e.g., intelligent CDMS allowing real-time data aggregation), CDM should limit the loading of external data in EDC to data critical to timely decision making. This may include investigator review, key safety reviews by sponsor physicians, site compliance (e.g., support drug reconciliation by monitor). Companies should investigate future-proof data integration and reporting platforms that are compatible with current and future data streams, including sensors and wearables.

Release EDC before First Patient First Visit (FPFV): The biggest challenge identified by most Tufts survey respondents was the journey from protocol to database release[8]. This often involves an extensive specifications process and, in some cases, the generation of a paper CRF. All this effort is to avoid mistakes in the build process, but at what cost? It takes, on average, more than 10 weeks to build a database. In some cases, it takes more than 14 weeks, with many companies blaming protocol complexity for delays. While complexity is a factor, misalignments within organizations which lead to multiple rounds of changes to specifications and protocol also play an important role. Overall, the time it takes to actually build an EDC is relatively short but much of the startup time is eaten up with the specification alignment process. As a result, more than 30% of companies said they often release the EDC after FPFV[8] preventing sites from entering data shortly after patient's visits. Most large companies stipulate in their processes that the EDC must be released before FPFV, or even before the first site is initiated.

While there is no regulation specifically requesting EDC to be released prior to FPFV, its availability before FPFV is critical, especially when EDC is to be used as an e-Source system for direct data capture. Sponsors may expose themselves to serious inspection findings if the EDC is determined to be a critical process enabler such as real-time safety assessment, verification of patient eligibility, drug reconciliation, etc. Companies releasing EDC after FPFV may also underestimate the downstream operational and potential data quality impact of such decisions. The study showed that, on average, companies releasing EDC post-FPFV observe an extra 5 days of delay from patient visit to the site entering the data in EDC, and up to 22 days longer database lock cycle times. Longer database lock cycle times are likely attributable to late data availability delaying data cleaning and increasing the volume of queries due to the lack of early feedback to sites from automated edit checks. Late and higher volumes of queries can also have a negative impact on site satisfaction. Perhaps most importantly, in early phase studies, this can delay the identification of safety concerns and expose patients to unnecessary risks.

Establish more flexible and efficient EDC build processes: Some reviewers of the 2017 Tufts study[8] challenged the need to build databases faster. They pointed to the frequency and prevalence of protocol changes prior to FPFV being a major risk. This has led some organizations to wait until the protocol is final before starting the EDC build process. As outlined above, the late release of EDC systems impacts the sites and overall study conduct. CDM ideally needs to develop processes that embrace flexibility, rather than adopt a process that starts with the final and approved protocol. With so many protocols subject to amendments in the age of adaptive designs, CDM has really no way out. Good practices include:

- Using standard and flexible EDC libraries accommodating multiple study designs
- Ensuring simple and user-friendly EDC design tuned to site workflows to foster direct data capture
- Leveraging industry standards (e.g., CDASH)
- Leveraging planning tools to manage dependencies and monitor activities on the critical path
- Targeting edit checks on what matters (i.e., being risk-based and focusing on critical data and processes)
- Using streamlined, fit-for-purpose and risk-based approaches to improve the costly and lengthy post-production eCRF change process

Leverage the value of all industry standards: Many companies still lack an end-to-end data standardization and integration strategy that considers all the dimensions of clinical data. It is critically important to understand that standards do not only apply to collection and transmission, but also to terminology and modeling. Failure to consider all data standardization dimension can result in process inefficiencies. Additionally, leveraging good standards not only facilitates the EDC and start-up process but also facilitates the creation of the datasets required for analyses and reporting.

Standards can be classified in four layers providing synergetic values[10]:

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

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

Terminology standards: Representations of individual data values

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

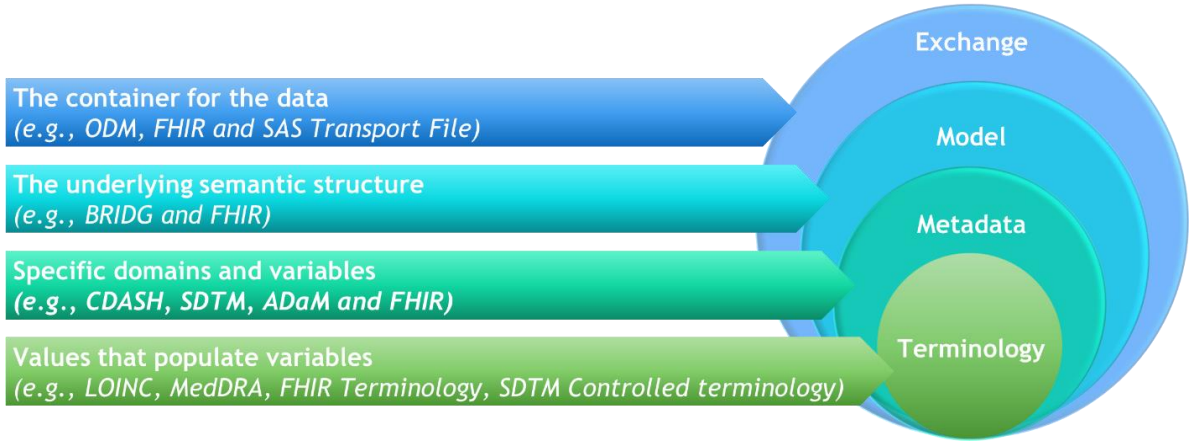


Fig 4. Layers of clinical data standards

As advocated and championed by SCDM, it is also worth noting that healthcare standards can be leveraged in clinical research to ease data mapping and exchange:

- Despite currently being at a low maturity level, the HL7 FHIR clinical research resources can be used to transfer EHR data to EDC. Aspects of these standards are interesting, as they support e-Source connections using modern technology and including the 4 layers outlined above.
- LOINC can be used to harmonize laboratory test terminology.

Few companies are exploring the use of metadata repository (MDR) systems to manage and apply all layers of standards to study-level processes to shorten the time to drug approval and drive automation efforts. Whether companies are using an MDR or other tools to manage standards, it is important to highlight that one of the key goals in managing set-up and post-production changes is to perform automated impact analysis. This allows for informed decision making, by providing more information on post-production changes and the impact of deviating from standards. Nowadays, much of the impact analysis is performed manually, which increases risk and extends timelines.

8. Impact of new clinical research approaches on technology

8.1 Adaptive design

The previous SCDM Reflection Paper[1] discussed adaptive design as means to optimize clinical research. Its adoption requires changes to CDM systems and processes according to the study adaptations. It means we need flexible systems and processes to support the requirements of adaptive trials with their associated statistical data inputs and outputs.

These advanced trial designs require fully integrated e-Clinical technologies to manage the overall scientific and operational complexities. Technology platforms must be able to adapt to rapid changes within the study, from the management of sites to the reported outcome data. They also need to account for all potential changes (e.g., sample size, dosing arms, regimen, patient population, endpoint, duration, and schedule) and provide real-time data for statistical modeling driving adaptations. Unlike simple traditional clinical trials, technologies that support adaptive trials need to be considered as one integrated ecosystem to efficiently surface the data and drive timely clinical decisions during the study. Moreover, data-related components of adaptive design go beyond collection. They include all aspects of clinical trial conduct:

- centralized lab management
- biomarker screening
- diagnostics and readers management
- centralized safety and medical monitoring
- advisory monitoring committees
- clinical oversight and monitoring
- site management
- screening and randomization management
- visit scheduling
- dispensing management
- drug supply, sourcing and logistics

Throughout all components, the Clinical Data Scientist will need to consider the 5Vs of the data itself, its format, its metadata, its authoritative source, its dependencies on other data points and shared sources, and how any changes to the data will be controlled and managed in addition to what integrations and transformations will be required of the data before and after it is stored in the intelligent CDMS.

IxRS is a good place to start with technology that supports adaptive trials, as it can manage the core design elements of the trial. Unless embedded with EDC, IxRS is usually the first system of data entry and generation. It identifies the subject ID, the site associated to the subject and will go on to identify the randomization, stratification and treatment assignment. Typically, it provides real-time status of the subjects.

Whether CDM, clinical operations or clinical drug supply function manages the system, the Clinical Data Scientist will be a key player in setting up IxRS for an adaptive trial. IxRS should likely be considered as an authoritative data source within the integrated ecosystem. This would avoid the need for any data reconciliation resulting from IxRS data being entered manually into other systems, such as EDC and safety databases.

IxRS systems would also benefit from other authoritative data within the integrated ecosystems. As an example, consider the case of a dose-ranging study where the objective is to most precisely estimate the dose-response curve. The first N patients will be randomized to one of 4 dose levels. At that point, the clinical endpoint (captured in the eCRF, laboratory, or connected device) will be used to update the

estimate of the dose-response curve. Given the new curve, new doses will be used going forward with the trial. These new doses will need to be pushed into the IxRS, along with their randomization ratios, so that subsequent patients are randomized appropriately. For this work, this integrated ecosystem needs to be connected, automated, and scalable.

Next is **EDC** which remains today the main data collection platform, and therefore must also provide the flexibility required. Specific recommendations for data collection systems include:

1. The system must enable fast mid-study changes – that means being able to build, test, validate and push live adaptations out within days. This is essential given the uncertainties surrounding study adaptations. As adaptations may only apply to a subset of the study participants, the EDC system must be flexible enough to push changes to targeted countries, sites or patients.
2. The database design concept of "groups" (though other names could be used such as "cohort" "arm", "sub-study", etc.) must exist within the system, not only to enable the separation of participants at the beginning of the study (for example, participants with biomarker 1 versus biomarker 2) but also as new criteria are added.
3. Participant IDs must be managed across these groups to maintain uniqueness, blinding, etc.
4. Defining unique and evolving paths must also be possible. Branching logics may also be based on treatment arm, study phase, etc. Therefore, the rules capabilities must be able to route participants based on new paths which are defined as the study evolves.

Note: Most EDC flexibility expectations are applicable to **eCOA**.

IxRS, EDC and eCOA requirements must be well defined and tested for accuracy and performance against expected adaptations and large amounts of data. Traditionally, UATs are conducted with a small sample of data based on simple scenarios to ensure the high-risk functional requirements are met. When supporting adaptive, complex and high-volume trial structures, large sample datasets should be used to test study technologies to ensure volume does not compromise the overall performance of the ecosystem. For a highly integrated platform, not only do the functional requirements of the system need to be tested, but so do the various integrations with other e-Clinical systems. In addition to data related systems changes, CDS will need to systematically assess, document and remediate the impact of adaptations to interim locks and data monitoring committee requirements (e.g., assess unmasking risks, update the Data Management Plan (DMP), implement remediation strategies prior to operationalization of the adaptations, etc.). This is important, as adaptive designs, by their very nature, are likely to increase the number of interim locks and require close data monitoring. Also, any changes to data processing systems have the potential to impact downstream digital and non-digital processes, as well as systems and tools such as dashboards, SDTM datasets, data flow, data review checks and protocol deviations. It is recommended to perform a complete risk assessment of process and technology changes driven by the study adaptations, and develop decision trees to guide study decisions.

Finally, it's important to note that the knowledge of accumulating study data can affect the conduct of a trial by influencing the behavior of its sponsor, investigators, and participants[11]. As such, data masking to sponsor, sites, and patients as well as the sharing of interim analysis results must be thoroughly controlled and documented. This can be challenging if the results of an interim analysis are driving adaptations potentially giving clues to trial stakeholders on the safety and efficacy of the IP.

8.2 Master Protocols

Today's trial design has moved past the adaptive design to Master Protocols. According to the FDA draft guidance[12], "*a master protocol is defined as a protocol designed with multiple sub studies, which may have different objectives and involves coordinated efforts to evaluate one or more investigational drug in one or more disease subtypes within the overall trial structure".* Master protocols can be designed with a fixed or adaptive design. They can also incorporate basket and umbrella designs as introduced in our reflection paper Part 1[1]. Master protocols are intended to improve the probability of matching the right treatment to the right patient with the right disease type to maximize a positive outcome. Although founded in oncology research, Master Protocols apply wherever a high level of heterogeneity of the disease exists – this could include infectious disease, mental health, or where multiple therapeutic agents need to be studied quickly.

With this change, data moved firmly away from being a collection effort to a design element, and with it the role of CDM is also fundamentally changing. CDS will need to carefully consider all data dimensions for key trial parameters in such complex and flexible trial design based on the population the protocol wants to target, the dose to select and the outcome to reach.

Let's consider an example. A sponsor is developing a series of molecules to treat one specific disease area. They will be available at various times, and the desire is to create a Master Protocol to test the efficacy and safety of each molecule. The individual molecules will be run under sub-protocols, which will inherit many features from the Master Protocol while maintaining a certain degree of uniqueness. There will be a common, shared control group to reduce the burden of control patients. The endpoints will be consistent, with similar designs, so handling them together will yield savings on start-up times and costs, as well as standardizing the analyses and data handling. However, it is possible that some molecules may require different evaluation times (e.g., the half-life of some molecules is 3 hours, others 3 weeks). The same sites can be used for all molecules, or different sites for different molecules. As the patients are enrolled, data is collected and analyzed. Promising molecules are exited from the program and advanced to the next stage of development, whereas those that fail to show efficacy are exited and development halted. Equivocal molecules will continue until a decision point is reached.

From this example we see generalities:

- Master Protocols combine and expand the concept of adaptive, basket and umbrella designs
- Individual sub-protocols may have differing timelines
- Arms may start or stop during the trial, and may not necessarily continue throughout
- Many molecules can share a common control group
- Designs can be used to evaluate multiple objectives in one trial:
    - In this case, multiple molecules for a single indication (umbrella trial)
    - But this can be flipped, where a single molecule is studied in multiple indications (basket trial)
- Master Protocols allow for efficiencies that are not possible with more standard adaptive designs, in terms of time, cost, and patient burden.

These studies are built on the premise of using a common ecosystem for seemingly disparate study parameters. As such, they demand a strong infrastructure where a high level of coordination and communication lays a foundation to apply complex statistical modeling to the targeted clinical space. Where successful, Master Protocols are truly patient-centric trials that improve trial participation, have a higher probability of demonstrating therapeutic effect, deliver better options and accelerate the drug development process.

Matching the right treatment to the right patient is highly complex and requires robust technology that supports a comprehensive data plan. CDS will need to consider more that the basic study parameters such as treatment arms, doses, duration, ratio, primary/secondary and futility endpoints, eligibility criteria and study power, diagnostics and analyses. The number of sub-studies, the criteria needed to resolve patients matching multiple sub-studies, confirmatory and exclusionary biomarkers, the evolving treatment landscape, and what criteria may potentially trigger a sub-study to be added or dropped must also be considered.

8.3 Decentralized Clinical Trials (DCT)

Decentralized trials are often referred to as if they are a distinct clinical trial category, but this is a fallacy. The concept of a decentralized trial is essentially predicated on decoupling clinical research activities from a physical site location.

This approach involves the use of technologies and processes, some of which we have been using for years, such as eCOA and web-based data collection tools, as well as remote data monitoring. Other technologies are newer to clinical research, but are often already being used by consumers, such as wearables, and telehealth tools.

As highlighted below, there are fundamental differences in the way the 5Vs apply to data collected in traditional trials when compared to DCT.

Traditional Approach		DCT Approach
Via observation / measurement	**How?**	Via connected devices and/or patient engagement tools
Brick and mortar study sites	**Where?**	Wherever patients roam or dwell
Patients and study personnel together	**Who?**	Patients alone physically (generally), sometimes visited at their home, but potentially with support through technology
Research-specific data only	**What?**	Research-specific and/or personal health-related data
Pre-specified intervals per protocol	**When?**	Pre-specified intervals, on demand, and/or continuously
For research purposes only	**Why?**	Sometimes for research, sometimes for patient's care

Overall, direct patient data collection supporting DCT is becoming more intelligent, bi-directional and interactive. Systems can share masked data with the patients to boost motivation while improving subject safety, compliance and ultimately, retention. While intriguing, this also highlights new elements and challenges around mobile device (e.g., BYOD or not), patient safety, ethical conduct, patient privacy, and shines a light on the quality of the clinical data being shared with patients including those from our CDMS. If we share data points that a patient (with or without their caregiver) uses to make a care decision, what are the implications and mitigations to consider within our clinical trials? CDS will need to consider how patients interact with different types of clinical trial technologies. More critically, CDS will need to flexibly establish data sharing, integration and cleaning strategies in an arena where ethical and strategic decisions play an even more vital role in patient safety and privacy.

Furthermore, the volume of data collected using these tools can be vast, and the desired use of this data must be considered prior to collecting it. For instance, if actigraphy devices will be used to measure the number of steps and hours of sleep, is the intent to compare a single participant's day-to-day activity, or to compare activity across participants? While it might be tempting to compare activity from patient to patient, with the relative variance of the data from these devices due to the device itself and user habits, it may not be possible to compare absolute measurements. Trends may be observed within a patient as to whether activity is decreasing, increasing, or remaining the same over time.

This is just one simple example of the need to rationalize the data collection methods as well as how the data will be used. In that context, what will the CDS role in reviewing this data be? Clearly, the data from a device can't be queried in most instances. An elderly patient will likely not know if they really walked 3,235 steps yesterday when they may only have walked from 1,500 to 2,000 steps a day before. But, if there is wide variability in the data day-to-day, there may be value in sending questions or reminders to the patient through an engagement app, such as, "*how many hours a day do you wear your device?*", "*did you forget to wear your device yesterday?*", or "*Please, remember to wear your device tomorrow?*".

The use of apps and devices is a critical technology component that must be leveraged effectively to achieve the goals of data collection in decentralized clinical trials. Connecting to a patient without the use of physical clinical sites does require utilizing patient-centric applications and devices that are easy to use and, ideally, require little training and maintenance. The value of data collected using a device is only realized if the patient is using the device consistently, correctly, and if the data is transmitted regularly to be monitored and analyzed.

Best practices currently suggest that the use of clinical grade devices is most appropriate for clinical trials[13]. However, the use of commercial-grade devices has been accepted in some instances[14] such as the Apple Watch to detect atrial fibrillation which have been approved by the FDA[15]. The larger sample set that can be achieved in a decentralized trial may allow for a greater tolerance of variability than in a traditional trial, which is important due to the higher likelihood of missing data[16].

The need to benefit from decentralized trial methods has never been greater, and the industry is likely to continue pursuing ways to introduce more aspects of decentralization into trials. Being prepared for the downstream impact to CDS will ensure that once other obstacles to adoption are overcome, we will be ready to handle the data regardless of how it is collected.

9. Automations

This paper will intentionally focus on automation technologies as these are mature and their applicability to CDM is immediate. As covered in the previous reflection paper, opportunities offered by emerging technologies have the potential to revolutionize clinical development and dramatically change CDM at its core[1]. While the automation of many CDM activities is unavoidable, the future of CDM has never been so bright. In a world where people and intelligent machines will collaborate, CDM leaders need to take proactive steps and articulate a clear capability strategy toward CDS that considers all aspects: people, process, regulations, standards and technologies. Each organization's CDS roadmap may depend on its data assets and technology maturity.

9.1 Emerging technologies driving automation

Before diving into automation, it is important to clarify the underlying technologies enabling the development of "artificially intelligent" applications and the differences between artificial intelligence (AI), robotic process automation (RPA), intelligent process automation (IPA) and other related technologies. Figure 5 shows the key components necessary to power CDM automations. These components have different roles to play and could be used either individually or combined to deliver the desired CDM automations. They start with basic process automation (i.e., RPA) and evolve toward the simulation of human cognitive behaviors such as reasoning and learning by machines (i.e., true AI). They collectively enable a wide range of opportunities for CDM that will impactfully transform our discipline toward CDS. CDM Leaders need to set a clear vision and roadmap to unleash this CDS potential.

Robotic Process Automation (RPA)	Intelligent Process Automation (IPA)	Natural Language Processing (NLP)	Natural Language Generation (NLG)	Machine Learning (ML)	Artificial Intelligence (AI)
Basic Automation of repetitive and predictable tasks	**Advanced Automation** of non-routine tasks	**Language and text analysis**	**Generate text and voice**	**Pattern recognition**	**Human like behaviors** Self optimization and Learning
No / simple decision making	Some decision making and problem solving	Knowledge gathering	Knowledge sharing	Prescriptive and predictive analysis	Deductive analysis

Fig 5. Technologies powering automations

First, **AI** is an evolving field and umbrella term which typically includes natural language processing (NLP), natural language generation (NLG), machine learning (ML) and computer vision (CV). It also includes intelligent process automation (IPA) which is the natural evolution of RPA. With IPA, robots can perform non-routine tasks requiring more complex problem solving.

RPA is process driven, meaning it focuses on process automation, whereas AI, which focuses on thinking and learning, is data driven. RPA is sometimes considered by some to be the simplest form of AI. The fact is that RPA solutions do not require problem solving and are at best using decision tree.

IPA combines both RPA and components of AI. IPA is often powered by ML and NPL which were both covered in the reflection paper Part 1[1]. Smart IPA workflows monitor and track information exchanged between people and systems in real time, and ML empowers intelligent, rules-based decision making. NLG converts and interprets text/voice to allow human language-based communications.

9.2 Roadmap to the automation of data reviews

As an example to show the evolution of CDM to CDS, some CDM leaders foresee that AI will be driving the evolution of data review from traditional to supervised. The evolution may include 4 distinct steps:

- **Stage #1 – Automation of traditional reviews:** Data trends and anomalies will continue to be identified via edit checks, listings and dashboards. However simple and repetitive tasks will be automated with RPA (see Section 9.3 on RPA).
- **Stage #2 – Actionable reviews:** In this context, systems identify data trends and anomalies via ML-based automations (see Section 9.4 on IPA) and statistically based analytics to detect atypical data patterns[17]. The adoption of advanced analytics rose recently to support risk-based monitoring and has significant potential in reducing the need for complex and labor-intensive manual data trending.
- **Stage #3 – Guided reviews:** Guided review is the natural evolution of actionable cleaning. Once the automated detection of data trends and anomalies have matured, and when the volume of actions manually taken by CDS is meaningful, ML Tools will be able to learn from them and automate actions. In this first stage, CDS will review suggested actions prior to execution. NLG may also be implanted to automate the writing of queries.
- **Stage #4 – Supervised reviews:** In the end, systems will automatically detect and act. They will only escalate new scenarios to Clinical Data Scientists when they do not have enough knowledge to take a decision autonomously. In this context, Clinical Data Scientists will oversee the systems, support their training, and arbitrate complex data issues that systems cannot "judge". They'll oversee the entire ecosystem to prevent bias, privacy and ethical breaches.

As shown in figure 6, the value and more importantly the transformational impact of automation will grow as CDM evolves to CDS, from stage #1 to #4 automations. Stage #1 automation will mostly require very few of the technology components from figure 5. However, Stage #4 automation will be more complex and need the full power of all automation components combined.

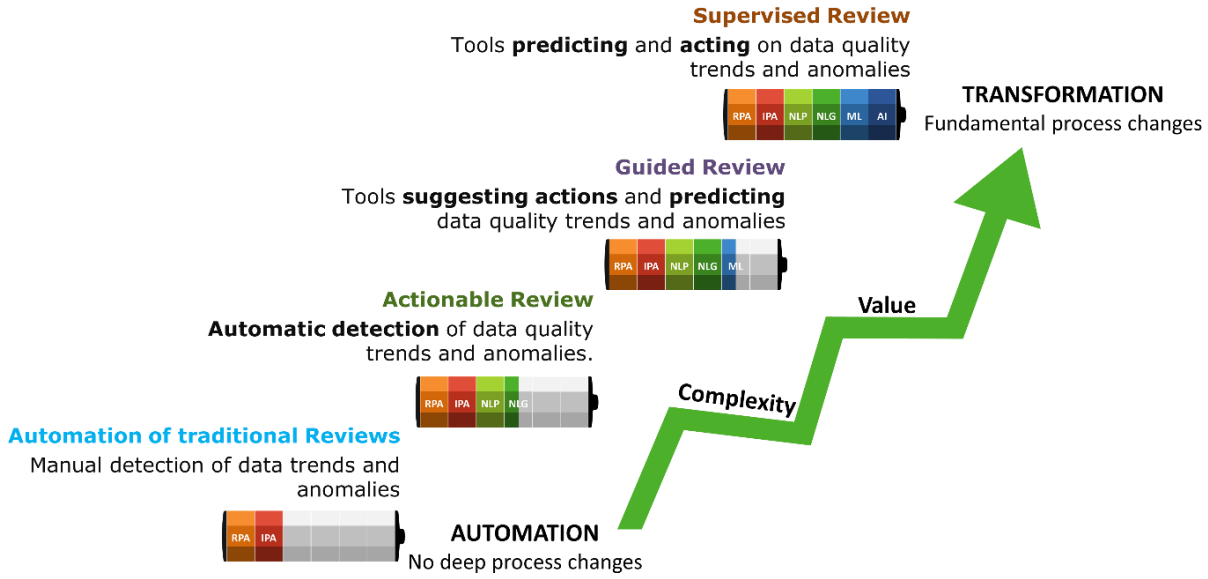
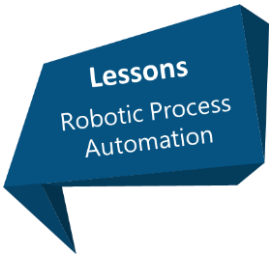


Fig 6. Evolution of data reviews powered by AI technologies

9.3 Robotic Process Automation

RPA is the simplest form of automation technology which is implemented through software applications, or robots, that perform predictable and routine tasks that do not require complex decision making. RPA is a mature technology which has been widely used across industries for years. It can mimic human interactions across multiple applications and is usually implemented to improve efficiency without the need to deeply change existing processes.

Typically, RPA applications take the form of a "bot", which can be thought of as digital employees (e.g., Virtual Clinical Data Manager). Implementing RPAs helps reduce process time, increases throughput, enhances accuracy and frees up human employees by helping them to focus on the tasks that bring the most value back to the organization. To unlock its value, it is imperative to begin by setting a framework to educate all clinical research stakeholders and evaluate prospective opportunities for automation.



As a ***first step***, organizations should establish an AI-focused team that is responsible for meeting the organization's vision and delivering its high-level strategic roadmap. The team should be cross-functional, with representation from all roles supporting clinical studies as well as enabling functions (e.g., IT, finance, etc.). This team should be chartered and provided the means to deliver impactful business automations. Each automation must set clear objectives and measurable outcomes (e.g., cycle times / resources reduction, accuracy increase by X%, etc.).

The ***second step*** is to educate the team on AI principles and their scope of applicability. As previously mentioned, AI is a combination of capabilities that allow machines to replicate human decision-making and/or interactions. Each capability offers different opportunities and is associated with different levels of complexity. Once these different principles are clear, it is easier to strategize and focus on specific RPA opportunities.

The ***third step*** is to collect potential ideas for automation. There are many ways to collate ideas, however, onsite workshops focused within each business line can accelerate the process by bringing together functional SMEs to review ideas and opportunities in real time, and provide a forum for others to discuss and build on them. Each SME should be prepared to present a description of the automation and where it fits in the high-level process, including any potential changes that may be required to the process to enable the automation. For instance, data may need to be stored in a different location to accommodate the bot, or access to a data source might need to be opened to a different business group that was historically not considered. The business value should also be clearly articulated. The expected outcome of these sessions is not only to identify opportunities for automation, but to prioritize the opportunities for delivery and estimate potential time and/or cost efficiencies and cycle-time reductions.

The ***fourth step*** for successful RPA implementation is to establish clear short-term and long-term automation roadmap goals. After assembling and reviewing all of the automation use cases, the team should prioritize the projects with the highest alignment with the company's strategic goals, and use those to set the foundation for how to implement AI/RPA with your department. As CDM review each idea, they can clearly define which components of the AI principles would be utilized to roll out the solution. Projects only requiring RPA capabilities, rather than those requiring hybrid solutions, should be

prioritized for easy and rapid implementation. Each project should be assigned a team lead, SMEs and project manager. The team should also align on a metric that will be reported out once the automation goes live and training/change management is in place to ensure full adoption of the new process. Once the organization has been trained on the RPA, it is important to track utilization of each bot to ensure the full value of the automation is truly realized. Low bot utilization could also point to issues with how it was programmed or implemented, and that should be addressed via programmed updates over time to ensure that the bots grow with the business process changes. As CDM optimizes and automates core processes, it must define target end-to-end processes and operational plans to ensure that opportunities are prioritized to enhance processes in lockstep with the company's technology strategy.

The **final step** is to build a sustainable automation development process, or an RPA Factory. This means developing a process with input from all relevant stakeholders, testing the process and validating each bot to ensure the workflow is correct and efficient. Potential areas of automation could include:

- Serious adverse event reconciliation
- Automated clean patient tracking
- Auto-generation of edit check specifications
- Automated loading of 3rd party data
- Automation of clinical data reviews
- Analysis and reporting system process triggers (scheduler)
- Automated quality review

Once a stable process has been developed, organizations should establish a global governance model to ensure alignment on prioritization and the funding of future automations.

9.4 Intelligent Process Automation

One large sponsor organization conducted a 'hackathon' which focused on the intelligent automation of the query management process. It involved the review of thousands of clinical data points to determine if a data discrepancy should be queried with a site. Many studies have thousands of manual queries raised over a long period of time (lasting from several months to many years). CDM must reconcile data across several CRFs by reviewing audit trails manually to determine if a query must be raised or not.

The hypothesis of the hackathon was that smart machines can learn from historical data by associating multiple clinical datapoints to specific manual queries, and establish clear data to query association patterns. When a similar pattern is presented to this trained model, the expectation was that it could predict discrepancies and, possibly, raise queries.

Data from several historical studies from various EDC systems were provided to the hackathon participants in a secure cloud environment. The studies had multiple data standards and millions of clinical data points. ML models were developed by four companies using various state-of-the-art techniques including deep learning. Vendors were evaluated on various aspects including innovation, accuracy and their ability to scale to a production solution. The outcome was successful in establishing a method for intelligent machines to perform CDM activities and proving that the hypothesis was correct. A semi-supervised learning approach was determined as the most successful method. Some labeled data were provided for supervised machine training by CDM SMEs and combined with an unsupervised learning algorithm where machines predicted similar patterns through observations.

As a next step beyond the hackathon, the selected ML will be trained with data beyond those provided for the hackathon to ensure strong accuracy and consistency of the predictions for critical data domains such as adverse events (AEs) and concomitant medications. Once perfected, the model could be operationalized by integrating it directly into EDC or intelligent CDMS systems. At this stage, these models can reliably detect complex cross-panel data discrepancies and save valuable time. The 'human in the loop' method will enable CDM to review all predictions, determine if they are correct, and later automate the generation of queries.

Natural language generation (NLG) will ultimately evolve to the point where machines can compose queries depending on the discrepancy context. The approach of using historical discrepancies and their associated queries will only train machines on known patterns. Future work to expand this use case to inference and clinical reasoning can move the AI needle more toward a CDS. For example, could machines detect AEs that are not reported? Can we take a subject holistically, analyze and reason whether a concomitant medication should have been taken when there was a reported AE etc.

Overtime, the model learnings from other clinical studies and SMEs will expand the horizons of AI and CDS. The technology can look across TAs, identify more complex patterns and adapt over time, while humans will assess machine inferences and provide invaluable feedback. In this context, computers and humans working together will make better predictions than either could do on their own.

Lastly, to successfully unleash the potential of AI-based automations, people will need to validate that this method is reliable, and that the IPA-based bot is a dependable 'digital assistant' that interacts with employees and partners to make important decisions. If an AI application is not well designed and managed, it may 'misbehave', with significant quality and ethical ramifications. These could range from damaged site relationships, negative impacts on patient safety and trial reliability, to missteps in drug manufacturing that affect quality and discriminatory decisions that elicit regulatory scrutiny.

Organizations need to consider establishing AI governance to build transparency and trust[18]. We should also consider ethical aspects, including bias that could potentially be introduced from data.

9.5 Considerations for the validation of ML based solutions

Validation of new ML Based solutions goes beyond the healthcare industry and has been extensively discussed in the literature[19]. Good practices recommend the use of three independent datasets in the development and validation of a new ML based solution. A model is trained on a specific dataset called the **training set** and is further optimized on an independent dataset, the **development set**. This optimization is an iterative process where the model parameters are tuned at each iteration to maximize specific performance criteria. This step is critical to ensure that models can generalize to new data while maintaining sufficiently high levels of performance. Finally, the performance of the optimized model is assessed on a third dataset, **the test set**. This step ensures that the validation of the model is performed independently from the model building process and that performances are not over-estimated.

Stakeholders need to play a key role in the definition of performance criteria and need to be trained in properly interpreting those metrics. This will leverage their power in making informed decisions regarding the release of new ML based solutions in a production environment. As an example, the accuracy, i.e., the ratio of the number of correct predictions to the total number of input samples, is a common performance criterion use to optimize ML solutions. However, the optimization of this unique

criterion may fail to produce models that are relevant for specific clinical use cases. Let us consider the optimization of a ML based solution, trained on the overall US population, to detect patients that will develop a rare disease. The US population is an unbalanced training set where the number of subjects with the disease (positive cases) is much lower than the number of subjects without the disease (negative cases). The optimization of the accuracy criterion alone in that setting would likely favor the detection of negative cases at the cost of missing true positive cases.

Once the model is released to production, several factors can lead to the degradation of its performance over time including a modification of the user's behavior or to the input data. A monitoring process should be put in place to actively assess the model performance over time and trigger specific actions when the model performance drops below a predefined threshold. These actions include the training and optimization of the model on new datasets and adaptation of the model architecture.

10. Emerging regulatory expectations for technologies

With the ever-evolving technology landscape, there is a greater need for CDM to ensure the effective management of data quality and integrity through those multiple technologies and data sources. As presented at the 2019 SCDM Annual Conference in Baltimore by regulators from the FDA and the Danish Medicines Agency[20], there are regulatory expectations from sponsors to ensure clarity of the dataflow and ongoing review of data and metadata. In a world where e-Source is becoming the norm, those reviews are critical to identify potential issues regarding data quality, completeness and overall compliance to protocol and regulations.

10.1 Audit Trail

Sponsors are expected to have procedures for risk-based routine audit trail reviews. As more trial data are collected as e-Source across numerous systems, this is becoming increasingly important. Without these procedures, sponsors can miss important non-compliance issues that could only be evidenced through metadata. In addition, it is likely that GCP inspectors will be enquiring more about these procedures as well as requesting access to audit trial data during inspections.

During the SCDM Conference, regulators shared issues identified by audit trail and metadata review during GCP inspections. Reviews revealed that:

- Data entry was not performed by authorized individuals
- ePRO data was entered within an unreasonably short period of time: ePRO was entered in less time compared to the time it should take for the site to complete 1) the assessment with a trial subject and 2) entering data in the ePRO system.
- e-Source primary efficacy data was not entered directly into the application by sites per protocol (i.e., alternative source data existed). Additionally, discrepancies were detected when the inspectors retrieved the 'true' source data to verified ePRO data.
- Site entry occurring during the same time/day for a group of subjects.

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

Please note that efforts are underway from the e-Clinical Forum and SCDM Audit Trail Task Force to publish an industry position paper on audit trail processes.

10.2 Inspection readiness considerations

The evolution of technologies and regulations have a direct impact on inspections. The multiplicity of systems used in clinical trials and the increasing volume of data collected beyond EDC are driving inspectors to focus on system and data lineage. During inspections, CDM organizations need to be able to clearly articulate their data flow, risk-based data review plan, system access and validations strategies. Inspectors may want to review processes, study documentation, systems validation packages and get systems demonstrations. They may even request access to systems, the eTMF, and data to freely assess compliance with protocol and regulations.

Beyond traditional inspection readiness, CDM organizations may want to consider the following:

- Sponsors need to simply and clearly illustrate the **end-to-end data flow** for a trial. CDMs should consider including a data flow diagram**,** with supporting explanations, in their DMP. Alternatively, they could create a separate data lineage plan, describing all the elements of data, from acquisition and data delivery to statistical analysis, that make up the study. Ideally, this would include all transformations and derivations.
- Proactively adjust **system decommissioning** strategies: Due to an increasing need for inspectors to access dynamic data and audit trails in the data collection tool, there should be consideration of a fit-for-purpose approach to decommissioning systems. For example, consider decommissioning non-submission studies within 4 months of database lock or upon CSR completion, whichever occurs first. For Phase III submission/pivotal studies, consider decommissioning EDC, eCOA, IxRS and similar systems after the first regulatory approval or first sponsor inspection, whichever occurs first.
- Anticipate **requests from inspectors to have access to systems**: Processes should be in place to describe access during an inspection whether it is at site or sponsor level. This should include an auditor read-only role, an expectation of training, how system access requests are submitted, and what roles are involved, as well as a process to revoke system access after the inspection. Training should be fit-for-purpose (e.g., 10-15 mins) and access should be granted as soon as possible: within 24 hours, where feasible, for example.
- Consider **system access for site vs sponsor inspections**: The process for granting inspectors a level of access should be relevant to the type of inspection being undertaken. For sponsor-level inspection, the sponsor should be ready to provide access to all sites and studies, whereas site-level inspection access should be restricted to the relevant site.
- **Anticipate site data archival challenges** (No more CD readers and USB Keys failing): As traditional CDs are being replaced by USBs, site archival continues to be a challenge:
    1. Are sites storing and retaining physical media appropriately?
    2. Is security placed on archives by sponsors making site access easy?
    3. Where are sites managing cross study archive credentials?
    4. Will data archived on physical media be readable over time?
    5. Is PDF the right way to archive data?

The list of challenges is long. Access to searchable data and audit trails from site archives is crucial during inspections and inspectors are growing tired of searching for audit trails in patient specific PDFs. There should be consideration of a cloud-based solution allowing sites to access to all data sources (e.g., EDC, eCOA, central labs, sensors, etc.) without the challenges of traditional hard media. This would enable sites working with multiple sponsors on many studies to have one centralized method of retrieving **dynamic** data and audit trail archives during an inspection without requesting last-minute support from the sponsor.

- **Readiness for site inspections**: CDM involvement in site level inspections has been growing over the last few years. Those include the review of systems and data flows. CDM organizations need to increase the scope of inspection readiness activities and consider site-level inspection readiness plans which include site specific dataflows.

While inspection readiness is important, the best readiness strategy is to foster QbD in all aspects of our work. If all processes, study and system documents are clear and well organized in the eTMF, inspection readiness activities will be more focused and less labor intensive.

11. Conclusions

Ultimately, CDM must ensure the veracity of data coming from a variety of sources with high volume and high velocity. Technology must allow Clinical Data Scientists, supported by Virtual Clinical Data Managers, to ultimately extract the full value of clinical research and health care data as illustrated by figure 7 below.

Fig 7. The 5Vs data journey from collection to value generation

In such a context, technology must become the enabler to a true and scalable change allowing CDM to meet the demand of clinical research while remaining compliant to increasing regulatory requirements.

CDM needs to leverage technology to:

- Unleash the data potential by generating valuable insights from raw data
- Find the needle in the haystack to identify data issues threatening data interpretability with advanced capabilities such as ML and AI
- Use automated and metadata-driven solutions for data aggregation and transformations
- Use RPA and IPA solutions to automate simple and repetitive tasks
- Leverage direct data capture and be DCT-proof (true e-Source, means no SDV and no queries)
- Be ready to support non-traditional protocol design (e.g., Master Protocol, Adaptive, etc.)
- Integrate and unify data coming from a growing number of systems to intelligent CDMS
- Ease inspections

In conclusion, technology is the key enabler of our evolution to CDS. CDM organizations should consider the elements outlined in this paper, as well in those in the first SCDM reflection paper (Part 1), to set their organization's vision and its corresponding roadmap to success.

References

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

[2] FDA, Digital Health Innovation Action Plan, available through the FDA Digital health webpage located at https://www.fda.gov/medical-devices/digital-health

[3] Tufts, November 2012, Clinical Trial Complexity, available at http://www.nationalacademies.org/hmd/~/media/34D1A23404A8492998AD2DF0CB6CD4D1.ashx …

[4] IQVIA, November 2017, The Growing Value of Digital Health: Evidence and Impact on Human Health and the Healthcare System. Available at https://www.iqvia.com/insights/the-iqvia-institute/reports/the-growing-value-of-digital-health

[5] FDA, December 2018, Real World Evidence Program. Available at https ://www.fda.gov/media/120060/download

[6] IBM, 2015, The 4 Vs of Big Data. Available at https://www.ibmbigdatahub.com/infographic/four-vs-big-data

[7] 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

[8] Tufts & Veeva, July 2017, eClinical Landscape Study. Available at https://www.veeva.com/wp-content/uploads/2017/11/Tufts-Veeva-2017-eClinical-Landscape-Study.pdf

[9] Clinical Pipe, Why Pharma Needs an EHR to EDC Connection, Available at https://www.clinicalpipe.com/blog/why-pharma-needs-an-her-to-edc-connection

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

[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] FDA, September 2018, Draft Guidance, Master Protocols: Efficient Clinical Trial Design Strategies to Expedite Development of Oncology Drugs and Biologics Available at https://www.fda.gov/regulatory-information/search-fda-guidance-documents/master-protocols-efficient-clinical-trial-design-strategies-expedite-development-oncology-drugs-and

[13] US National Library of Medicine, 2018, Elena S. Izmailova, John A. Wagner and Eric D. Perakslis, Wearable Devices in Clinical Trials: Hype and Hypothesis, Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6032822/

[14] Clinical Researcher, April 2019, An Overview of the Prospects for Using Wearables to Improve Clinical Trials, Available at https://acrpnet.org/2019/04/25/an-overview-of-the-prospects-for-using-wearables-to-improve-clinical-trials/

[15] Marco V. Perez, and Al, Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation, New England Journal of Medicine. Available at https://www.nejm.org/doi/full/10.1056/NEJMoa1901183

[16] Jenifer L. Hick, and Al, Best practices for analyzing large-scale health data from wearables and smartphone apps, NPJ Digital Medicine, Available at https://www.nature.com/articles/s41746-019-0121-1

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

[18] Cognizant, Jan 2019, Making AI responsible and effective, Available at https://www.cognizant.com/whitepapers/making-ai-responsible-and-effective-codex3974.pdf

[19] Andrew Ng, 2018, Machine Learning Yearning, available at https://www.deeplearning.ai/machine-learning-yearning

[20] 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

Main abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
BRIDG	Biomedical Research Integrated Domain Group
BYOD	Bring Your Own Device
CDASH	Clinical Data Acquisition Standards Harmonization
CDM	Clinical Data Management
CDMS	Clinical Data Management System
CTMS	Clinical Trial Management System
CDS	Clinical Data Science
DCTs	Decentralized Clinical Trials
eCOA	electronic Clinical Outcome Assessment
DMP	Data Management Plan
eCRF	electronic Case Report Form
EDC	Electronic Data Capture
m-Health	Mobile Heath
EHR	Electronic Health Records
EMA	European Medicines Agency
ePRO	electronic Patient Reported Outcome
FDA	Food and Drug Administration
FHIR	Fast Healthcare Interoperability Resources
HL7	Health Level Seven
IP	Investigational Product
IPA	Intelligent Process Automation
IxRS	Interactive Response System (x being any type including Voice or Web)
IRT	Interactive Response Technology (Synonymous to IxRS)
QbD	Quality by Design
MHRA	Medicines and Healthcare products Regulatory Agency
LOINC	Logical Observation Identifiers Names and Codes
ML	Machine Learning
NLP	Natural Language Processing
REST	REpresentational State Transfer
ODM	Operational Data Model
RBM	Risk-Based Monitoring
RPA	Robotic Process Automation
RTSM	Randomization and Trial Supply Management ((Synonymous to IxRS and IRT)
RWD	Real-World Data
RWE	Real-World Evidence
SCDM	Society for Clinical Data Management
SDTM	Standard Data Tabulation Model
SDV	Source Data Verification
SME	Subject Matter Expert