5. The role and skillsets of Clinical Data Scientists

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

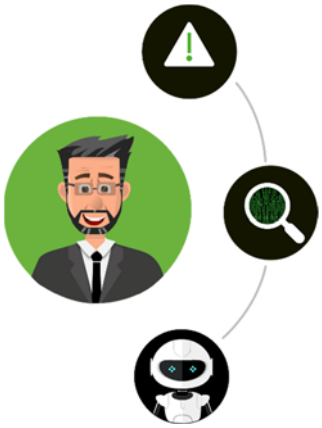
The first fundamental change in our CDS journey is the shift in focus from data integrity to data quality. But without a doubt, while "the controls required for [data] integrity do not necessarily guarantee the quality of the data generated" [6], data integrity remains core and is expected to reach data quality. According to MHRA, data quality is "the assurance that data produced is exactly what was intended to be produced and fit for its intended purpose"[6]. Quality data also reflects the reality of what happened to the patients (e.g. The patient's blood pressure was indeed 132 over 83, the patient truly experienced an injection side reaction, etc.). ICH E6 (R2)[7] goes beyond integrity as well by expecting the ability to distinguish between reliable and potentially unreliable data and by driving focus on critical data.

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

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

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

CDS role evolution framework



Risk based CDM approaches aligned with new regulations focused on

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

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

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

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

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

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

7. Impact of the role evolution

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

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

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

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

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

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

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

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

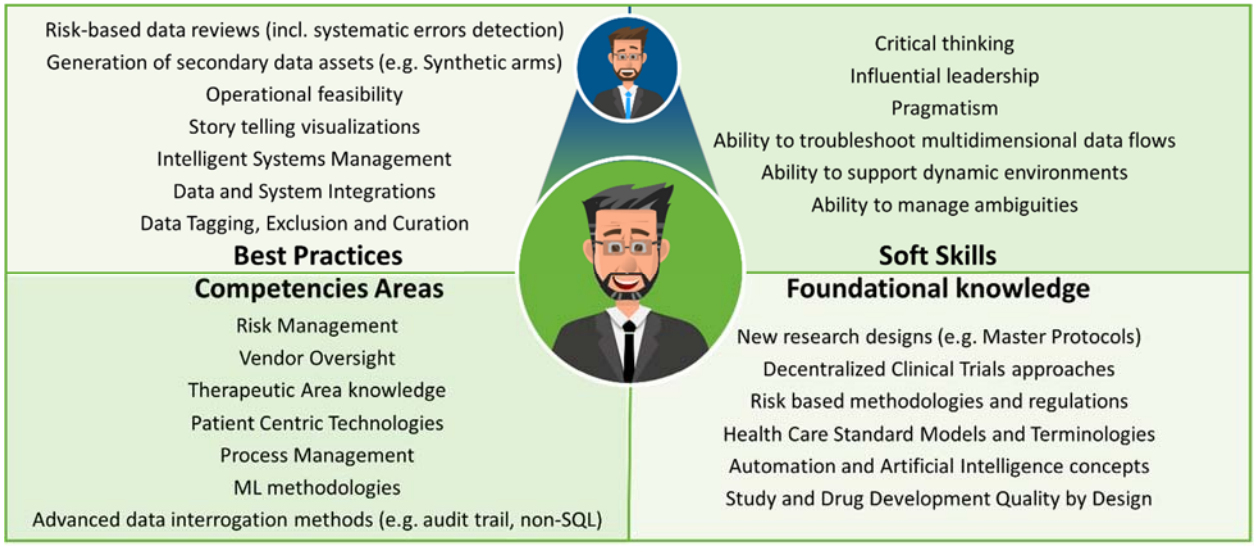


Fig 10. CDS role framework

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

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

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

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

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

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

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

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

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

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

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

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

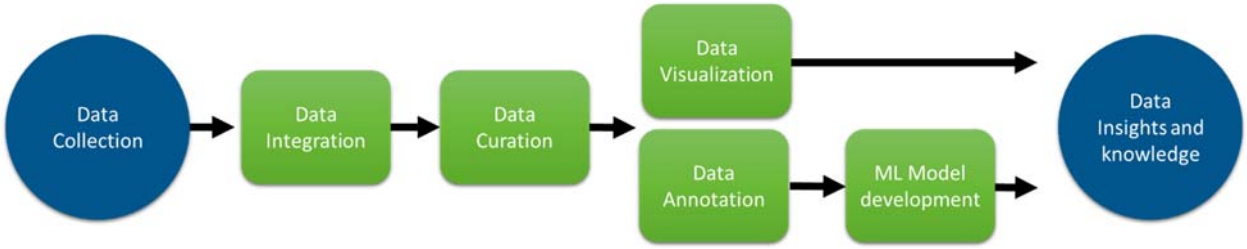


Fig 11. From data to knowledge data flow

8. Conclusion

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

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

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

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