



## OPINION PAPER

# GSK Data Strategy & Management Whitepaper 2023: Transformation from a Data Management to a Data Science Enabled Organization – Where Are We Heading?

Miho Hashio, Michal Kolodziej, Mike Olds, Elizabeth Samuel, Suraj Ravindran, Arshad Mohammed and Marcin Makowski

We, just like the catalogue companies, are facing technological changes that clearly provide an opportunity. To efficiently accelerate pipeline delivery with high quality data, we believe we must develop a standardized data acquisition strategy with AI/ML-driven quality checking and advanced data analytics. This is very exciting chance for us to improve our performance by leveraging our rapidly advancing science and technology. However, we are seeing risks to our benefiting from these new technological advancements because of our own old routines, processes, and preconceptions. Are we progressing at a pace?

In this article we will look deep into the soul of our organization to define where we want to go, identify the anchors, and try to figure out how to break loose from them.

**Keywords:** Manage Clinical Research Data; Collect data; Define/document data handling process; Process Data

## 1. Introduction

Have you ever bought a piece of clothing from a paper catalogue? This business model of selling apparel was quite successful throughout the 20<sup>th</sup> century. When the internet revolution came, it seemed that companies with well-established paper catalogues would be the first to succeed in online retail. They had already established distribution and returns processes for remote commerce and had also somehow managed to solve the eternal problem of people thinking they are slimmer than they really. However, they failed to successfully adapt to this new online market and new firms took control of this new paradigm. When these cases are studied at business schools it becomes apparent that these companies saw the rise of the internet as an annoyance rather than an opportunity, and their business model and processes continued revolving around the twice-a-year paper catalogue. This old business model was an anchor that kept them in place and blocked them from successfully moving forward with this new means of commerce.

The authors of this article are leaders of a department that manages data in clinical trials at GSK – a multinational biopharma company. We, just like the catalogue companies, are facing technological changes that clearly provide an opportunity. To efficiently accelerate pipeline delivery with high quality data, we believe we must

develop a standardized data acquisition strategy with Artificial Intelligence/Machine Learning (AI/ML)-driven quality checking and advanced data analytics. This is a very exciting chance for us to improve our performance by leveraging our rapidly advancing science and technology.

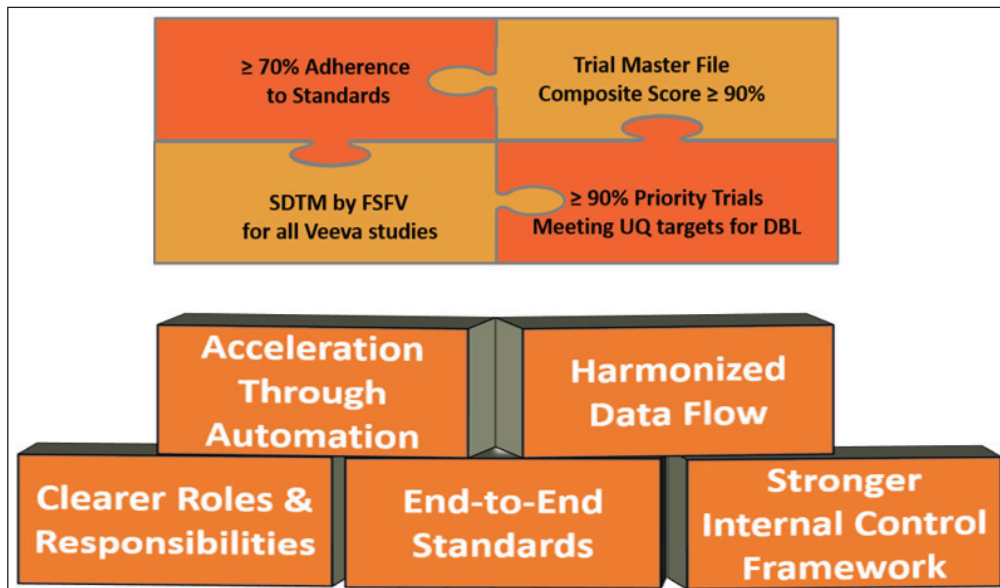
However, we see that old routines, processes and preconceptions are risks to our benefiting from these new technological advancements. Are we progressing at a pace?

In this article we look deep into the soul of our organization to define where we want to go, to identify the anchors, and to try to figure out how to break loose from them.

## 2. Current and Historical State of “Data Management” Organizations for Clinical Trials

Clinical Data Management (CDM) has historically been the team to design electronic Case Report Forms (eCRF) and to oversee the capture and cleaning of clinical trial data for analysis and reporting. While the relative stability of these processes has allowed CDM teams to master the day-to-day data validation and querying processes, the time taken to carry out these processes has prevented them from being able to leverage their time for additional value-added activities.<sup>1</sup>

At GSK, Data Strategy and Management (DS&M) is the organization within Global Clinical Operations that oversees the acquisition, management, and delivery of clinical trials’ data (**Figure 1**). There are seven sub-functions within DS&M at GSK as follows:



**Figure 1:** GSK Data Strategy & Management in Global Clinical Operations.

- Data Acquisition & Data Analytics
- Data Management
- Central Monitoring
- Study Data Tabulation Model (SDTM) Delivery & Programming
- Service Delivery for functional service partnerships (FSPs)/full-service outsourcing (FSO)
- Innovation
- Data Standards

Many data management (DM) organizations have been processing data the same way for a long time, despite the availability of newer data technologies and the changes to regulations.<sup>1</sup>

### 3. Future State of a “Data Science Organization” for Data Management

Data Science is a buzzword, and everyone is trying to move from traditional clinical development organizations to modernized data science clinical development organizations. However, what does it really mean for a DM organization?

At GSK we answered this question with three foundational statements; 1. data science is a method, 2. data quality means different things for different people, and 3. data from all sources matters. Below we discuss them one by one.

#### 3.1. Three Foundational Statements

##### 3.1.1. Data Science is a Method

By this statement we underline that data science is not a purpose. There are many definitions and models of data science. These new definitions and models define new roles or “personas”, such as data engineer, data analyst, AI/ML architect, and business translator. They also focus on the utilization of new analytical techniques to understand much larger datasets that have greater variability than data received in the past. Nevertheless, probably none of these definitions explicitly state the purpose of the

business activities that need to be addressed. Data science is a new way of achieving existing business goals, but with this conclusion we also realize that our business goals also need to be refined.

##### 3.1.2. Data Quality Means Different Things for Different People

With a degree of oversimplification one can say that Clinical Data Management organizations produce datasets for clinical trials. These are used for internal and external decision making on the progress of a compound to the next phase of development, registration, reimbursement, etc., as well as to monitor patient safety. These datasets need to be delivered in a timely manner and with quality. The term “quality”, though, is understood differently by different internal and external functions. An example of different views on an aspect of data quality – data completeness – is presented in **Figure 2**. To tackle this issue, we decided to separate the term “data quality” into two different concepts:

- Data consistency: Understood as compliance with standards, ability to pass through data consistency checks, such as P21, completeness understood as absence of missing fields.
- Data reliability: Understood as completeness of observation, lack of data manipulation, degree to which data reflects reality.

These two concepts reflect the needs and perceptions of different stakeholders and functions and are further explained in **Figure 3**. They correspond quite well with the terms “data integrity” and “data quality” presented across SCDM position papers.<sup>2</sup>

##### 3.1.3. Data From All Sources Matters

The primary historical business of CDM revolved around data collected via eCRFs, whether captured via paper CRF or electronically (eCRF). This was initially the sole method

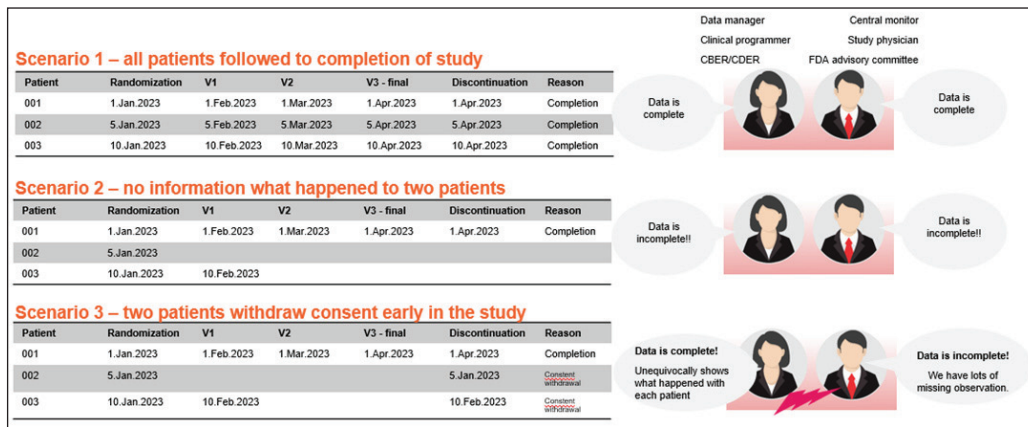


Figure 2: Example of Varying Perception of Data Quality Depending on Function/Stakeholder Group.

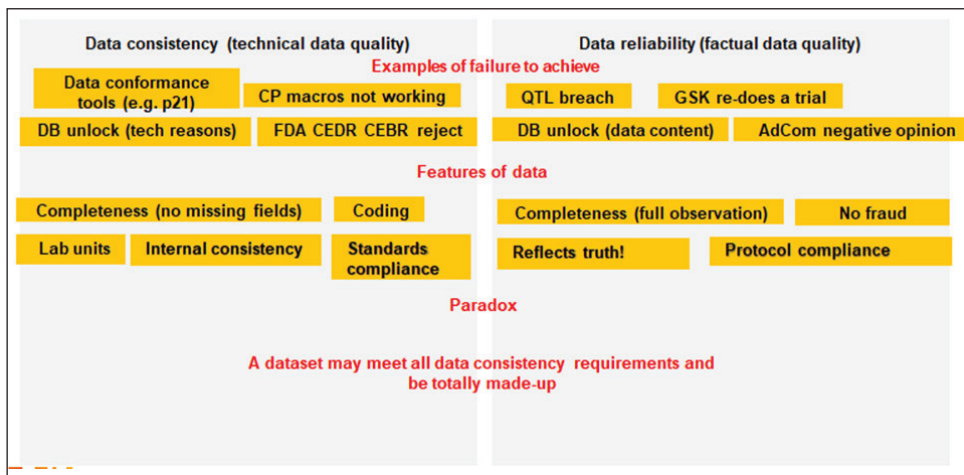


Figure 3: Examples of “Data consistency” and “Data reliability”.

of collecting clinical trial data, and even later remained the primary source of data collection for clinical research. But as technology evolved, new sources of data started to appear in the CDM landscape. Initially these were Randomization and Trial Supply Management (RTSM) systems and central laboratories. However, these days, non-eCRF sources of data usually surpass eCRF data in quantity, variety, and importance.

Let’s look at the variety. The median number of data sources in a clinical trial was four in 2019.<sup>3</sup> These non-eCRF sources included, but are not limited to: electronic Patient Reported Outcomes (ePRO), electronic Clinical Outcomes Assessment (eCOA), central readers of site procedures (eg, oncology tumor response data, ECGs, spirometry, endoscopies), medical devices data (e.g., vital signs sensors, actigraphy etc.); and multiple kinds of lab data, including standard safety labs, pharmacokinetics (PK), pharmacodynamics (PD), immunogenicity, genetic data, biomarkers, and many more.<sup>4</sup> In terms of volume, the non-eCRF sources can exceed the eCRFs’ quantity of data by many factors of magnitude.<sup>5</sup> Lastly, importance. In many trials’ designs, the primary endpoint data is collected outside of the eCRF, e.g., via external data transfers including central labs, eCOA devices, ePROs or more. In our practice, we see that in most situations in which a database unlock is considered or performed, the non-eCRF data are the primary reason.

Considering all the above, the importance of specific data for a study (and not source system) should define how quickly we need it, how often we refresh it, and how much effort we invest in cleaning and reviewing it.

### 3.2. Technology Ecosystem for Data Science Organization

Technology is the fuel that can supercharge the Clinical Data Science (CDS) rocket. With the right people at the steering wheel (People) and the right operating manuals (Process), it is time to fill up the tank with technological advancements that can better facilitate the collection, processing, review, and analyses of the data.

Clinical Data Science will never be possible without a cross-functional collaboration platform that enables planning, executing, and monitoring activities that lead to the end goal of clean and high-quality study data. This can include solutions for overall study risk and issue management, real-time instream reporting, dependencies between functions, quality checks, predictions for meeting study milestones, etc.

In today’s complex and evolving clinical trial landscape, it is important to have end-to-end data standards that can provide clarity to the clinical data being collected, which allows for downstream efficiency gains and increased automation. Standardization of unstructured and unstandardized data seems to be one of the main

challenges CDS organizations will need to find solutions to. Study-defined data points (**active data**) are already becoming a minority compared with data points acquired from the world (**passive data**) via Real World Data (RWD), Decentralized Clinical Trial (DCT), wearables, sensors, etc.<sup>5</sup> The technical platform for CDM of the future (CDS) will therefore need to have capabilities for data finding, curation, and transformations into usable formats, rather than for the perfect design of the data collection tools ahead of study start. Multiple technologies have emerged in recent years, including **Artificial Intelligence (AI)/Machine Learning (ML)**, that can lead towards effective management of enough *passive data* to significantly reduce the burden to patients (especially those from comparative groups).

### 3.3. Resulting Vision of Data Science Organization at GSK

Our resulting vision to transform our organization into one that is based on data science over data management is a combination of the statements above and data science paradigms that are not specific to our industry. Study-level elements of this vision are presented in **Figure 4**. Here are some key points:

- 1 As our main objective (data quality) has two aspects, we propose defining roles that are responsible for each of these: one that takes care of data consistency, and another that ensures data reliability. In general data science terminology, they would represent the business translator skillset.
2. As we treat all data sources equally, we need the people managing data acquisition to be able to

design study data flows in such a way that all in-stream reviews are fed with data from all source domains, abandoning the traditional “eCRF first, then for a very long time nothing” paradigm. This, in turn, represents the data engineer skillset.

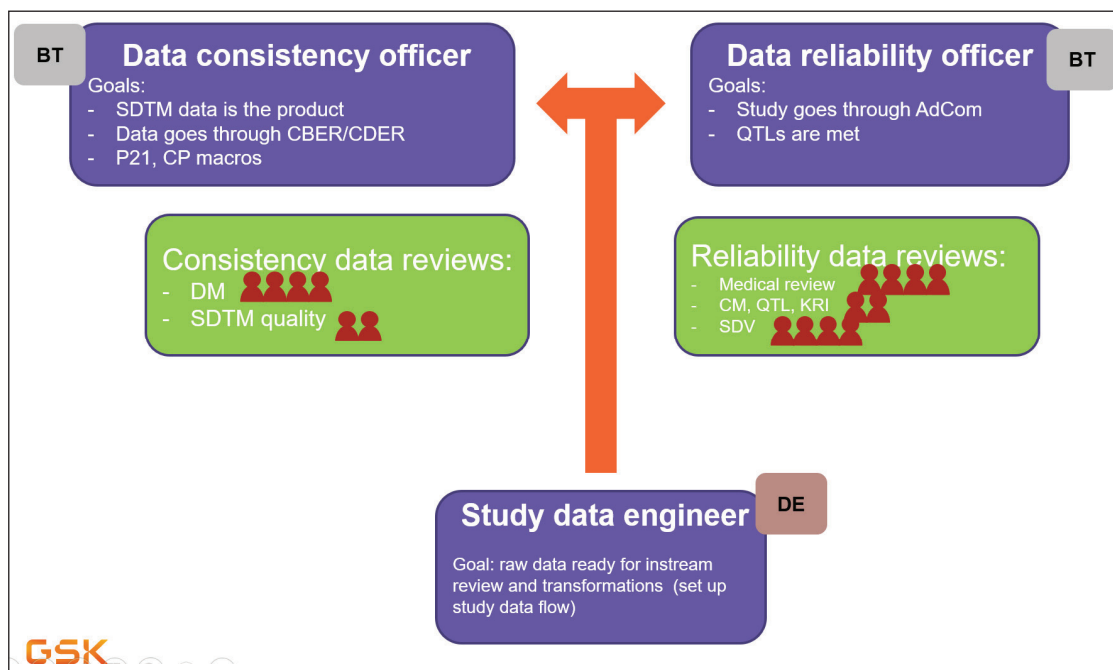
3. In the middle layer between acquisition of data and achieving a data product meeting both consistency and reliability needs, we currently have a set of data cleaning and in-stream review processes. We plan to align these clearly with consistency or reliability goals and, over time, to automate them using advanced analytics and machine learning. This requires the creation of an above-study data science office that is staffed by people that represent the data analysts, advanced programmers, data engineers, and data modelers, to understand and work for AI/ML technologies.

### 4. Potential Blockers – Challenges We Are Facing

What are the main key challenges we face that could hold us back from moving to a more Data Science based Data Management Organization? We describe several examples in this section, to look at the challenges more closely and to consider how we might leverage these challenges as opportunities.

#### 4.1. Evolving Technology of Data Acquisition and the Reality

Although the technology of assessments for clinical trials and data acquisition is evolving and the sources of clinical trial data is expanding, the eCRF still holds a central



**Figure 4:** Proposed Model – Data Science & Strategy for Clinical Development.

Acronyms:

BT: Business Translator DE: Data Engineer, SDTM: Study Data Tabulation Model.

CBER: Center for Biologics Evaluation and Research , CDER: Center for Drug Evaluation and Research.

P21: Pinnacle 21, CP: Clinical Programming, QTL: Quality Tolerance Limit, KRI: Key Risk Indicator, SDV: Source Data Verification.

position in our data collection and processing. Here are several observations to support this statement:

1. The data collected outside of the eCRF is called external data, ie, external to eCRF or third-party/vendor data. This causes the perception that this part of the data has a secondary or a supportive role.
2. In our company, the data acquisition group was historically organized into one group that handled the building of eCRFs and another group that handled all the other data sources.
3. A key milestone during the study set-up period that is visible to all stakeholders is the eCRF go-live, which does not reflect other data sources.
4. Even if an assessment subject to central review is the primary endpoint, sometimes additional medical review are conducted on eCRF data. For example, an oncology studies' review for progression data, which focuses on eCRF tumor response data.
5. ICH GCP requires the investigator or an authorized member of the investigator's staff to sign off on eCRF data to confirm the observations recorded, which is part of essential documentation for clinical trials.<sup>6</sup> This becomes a hectic, meaningless exercise at the end of many trials.

There are many reasons why the eCRF stays in the center of our minds. However, there is a possibility that many of these reasons can be circumvented or are a pure reflection of our routines. If this is the case, we can expect the appearance of new organizations that will totally abandon the eCRF as a source of data. We may then end up just like the paper catalogue companies.

**4.2. Data Flow**

We realized that we need to rethink our approach of data collection and processing considering the “5 Vs” of data (volume, variety, velocity, veracity, and value).<sup>7-9</sup> Our

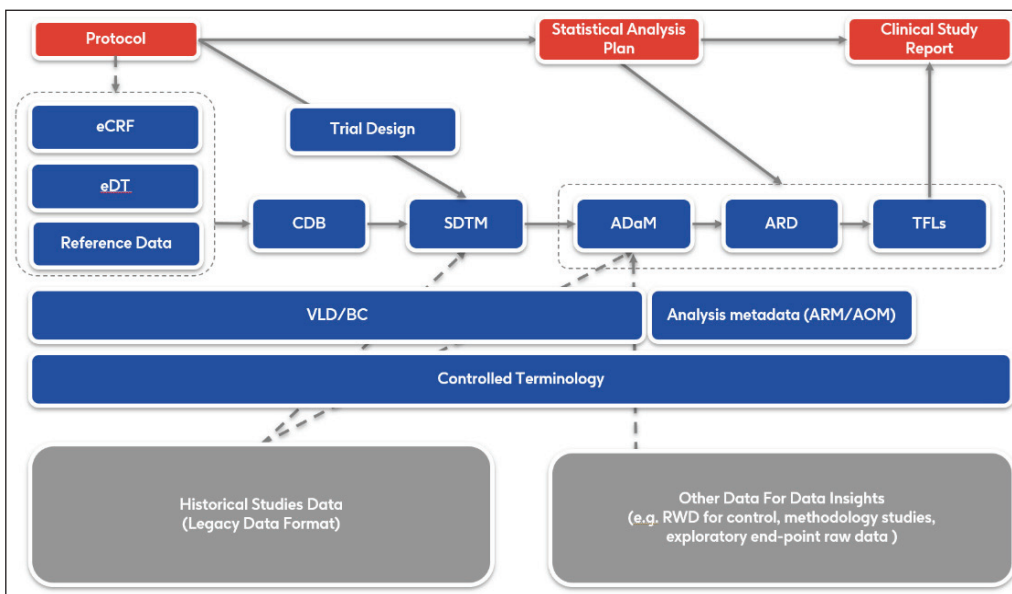
current data flow is fragmented; some of the key challenges to efficient data flow include system limitations (eg, lack of real time integration & interoperability between systems), mis-matched data structures, lack of clarity around multiple cross-functional handoffs and systems used throughout the data flow, using (or abusing) SDTM to identify problems with raw data, file conversions from data source to data destination, etc.

**4.3. Metadata Driven Automation/ Data Standards**

There are several long-term industry-wide initiatives ongoing in this area, such as CDISC 360, TransCelerate's Digital Data Flow, and ICH M11.<sup>6,10-11</sup> However, in many companies as well as in our own, the protocol is still a Word document. Because all studies have their own specific needs, there are frequently multiple “nice to have” or “just in case” end points included in the protocols. These needs are “special” and result in multiple variations of study protocols created for studies that should have similar data collection needs. Study teams' understanding of Data Standards is essential to be able to utilize available industry and corporate Data Standards across our studies, but this is not always the case. This mindset and behavior are slowing down our ability to truly apply data standards end-to-end.

Another challenge is that we lack metadata links throughout clinical trials' data flow. For example, in many cases, the protocol is not fully digitized and written as a document. It therefore does not have metadata links to downstream components e.g., eCRF, Statistical Analysis Plan (SAP), Clinical Study Reports (CSR), etc. Because of this, manual interpretation is required to set up a clinical trial database and transform the raw data into SDTM for each study and our current metadata definitions are not simply linked well enough to enable end-to-end digital data flow.

CDISC Standards are required by the US FDA, Japan's PMDA, and China's NMPA. Currently, the European



**Figure 5:** Clinical Data Standards – Metadata Link and Governance.

Medicines Agency (EMA) is also conducting a pilot project on analysis of raw data from clinical trials.<sup>12,13</sup> There are standard validation rules to ensure conformity to the data standards requirements, but these regulatory agencies' requirements are not always identical. It therefore frequently requires additional effort to conform to specific regulatory requirements, which sometimes means re-working data to create submission packages for a specific regulatory submission.

#### 4.4. "Focus on All" Rather Than Risk-Based Approach

ICH Good clinical practice (GCP) guidelines recommend risk-based approaches.<sup>14,15</sup> Other guidelines, such as ICH E8 R1, underline "fit for purpose" quality and "absence of errors that matter".<sup>16</sup> These statements have led to a proliferation of risk-based approaches to on-site monitoring over the last decade. Nevertheless, this proliferation of risk-based approaches has not fully penetrated the world of data management.

Our database lock checklist tracks metrics such as "all queries closed" and "all planned SDV completed". However, the focus on these aspects may sometimes result in activities that are quite illogical and are a waste of company time and resources. For example, a monitor might be asked to go to a site to SDV three forms to lock the database of a study with over 10,000 participants. We can easily assume that the probability of these three forms containing errors that impact the study's final results is close to zero. However, neither our operating procedures nor working practices are great in allowing acceptance of non-material errors in data.

### 5. Proposal of Potential Solutions

To address the challenges/blockers described in section 4, we would like to share potential solutions, focusing on the following aspects: people and technology.

#### 5.1. People – Soft Skills and Hard Skills

It is an essential foundation for CDM to have a deep understanding of the entire lifecycle of a data point (eg, clinical concept, data integrity, regulatory understanding, and some basic understanding of programming and statistics). Clinical Data Science (CDS) expands the scope of Clinical Data Management (CDM) by adding the data meaning and value dimensions (i.e., data is credible and reliable). CDS also needs to assume a key leadership role in a clinical trial as well as requires the ability to generate knowledge and insights from clinical or operational data to support clinical research, which requires additional expertise, approaches, and technologies.<sup>3</sup>

In today's rapidly evolving technical and scientific landscape, an important soft skill is a person's learning agility and curiosity; their willingness to learn and to adopt new technologies; as well as the communication skills needed to effectively engage with both technical and non-technical stakeholders to drive the adoption of new technologies and processes. To achieve this transformation of roles, we must come out of our comfort zones. There are also several combined roles that may help us move to the next steps (described in section 3.3).

- **Roles to facilitate data engineering:** These roles enable smooth data acquisition and data transformation for clinical trials' data. As the first step, we propose to have roles as follows:
  - **Technical Designer** that has scientific /biomedical background and understanding of clinical trials, study protocols and technology.
  - **Data Steward** that has technical and domain knowledge to support end-to-end data flow.
- **Roles to facilitate integrated data review:** Data Consistency officer and Data Reliability officer roles are desired for long-term conduct integrated data review.

We must be an agile, dynamic organization with key talents rotating through multiple roles relevant to the clinical data flow, each to be cross trained, exposed to various perspectives, and are encouraged to grow.

#### 5.2. Technology and Data Governance

At GSK, we are currently working with our strategic tech partners on the best means to adopt new technologies/platforms to apply automation, improve the user experience, and bring required study team members to access the single truth source of the data in timely manner. It will facilitate further collaboration, improve the same level of understanding of the clinical trial data, what could be predicted to potentially go wrong (risks), what happened in the study and how it was managed (issues), and how to avoid the same happening again (library of failures with mitigation proposals).

This area will also require industry-wide partnerships, although that is already happening under the umbrella of HL7 FHIR/Vulcan, TransCelerate, and others to agree on the best mechanisms of acquiring data from medical records already collected and available from national healthcare ecosystems.

Once the data is acquired and standardized, it is again time to deploy AI/ML to perform some initial "analysis" on the data. At GSK we are exploring Deep Learning solutions by scanning our study datasets in search of data anomalies, unusual patterns, and inconsistencies between various data domains; these are then highlighted to the Data Managers as potential data issues that could require follow up with the sites. Following this initial machine-human review from a data consistency perspective, we are exploring possibilities of deploying algorithms to support central monitors and medical reviewers with more complex and sophisticated analyses from a data reliability perspective. Our overall aim is to automate (either via simple RPA (Robotic Process Automation) solutions, or through high-end AI/ML algorithms) the repetitive or simple activities performed by humans, who can then focus on the critical evaluation of the outputs (or even more importantly on the **outcome**) of the machine pre-processed (or we are hesitant to say, *pre-analyzed*) data.

All of the above will not happen without establishing a foundational technological infrastructure that allows for multidimensional work on clinical data. The examples of key dimensions are not comprehensive, but examples are listed as follows:

- Operational data vs clinical data
- Data consistency vs. data reliability
- Instream data review vs statistical analysis of the study results
- Real-time data for clinical trials vs historical data vs other data for data insights
- Acquired/collected data vs output of analyses
- Traditional data conversion/mapping vs metadata-driven automation and data modeling

At GSK, we are looking into the best ecosystem that can support all the above dimensions, but we also realize it is not possible to have a one-size-fits-all solution. The key groups of potential solutions include an operational, day-to-day workbench that allows for instream, ongoing data reviews (Veeva CDMS) and a more static data lake, called a Development Data Fabric (DDF) that allows for broader data consumption, data mining, and other data re-use scenarios. The problem to overcome is to create a much more cohesive and functionally agnostic solutions ecosystem that allows for cross-functional oversight of the data we are collecting, transforming, and submitting for analyses.

Having fundamental harmonized data standards for clinical trial data and strong governance are pre-requisites for metadata driven automation, but having fundamental data standards itself is not sufficient. The standards should be compatible with the technology landscape to gain maximum efficiencies and return on investment. Continuous development and improvement are required for data standards to support emerging technologies and science, as well as various new data sources and direct data ingestions from these data sources.

It is also important to have the ability to standardize or to transform the data into standard structures for data aggregation and re-use when additional analyses are required for accelerating submission and pipeline delivery, and to give additional insights from historical data, within a proper data governance structure (see **Figure 5**).

### 5.3. Changes Management for Continuous Disruption

Continuous changes are the norm more than ever today, and we need to be resilient to these changes. To mitigate the risks of uncertainty and disruption, proper change management, communication and engagement are key. A goal is to provide stability where we can, and to try to ensure the psychological comfort of our teams, which will be needed for them to fully embrace these changes in a positive way.

## 6. Conclusion

We have described our perspective of how our current Data Management organization can become a futuristic and innovative Data Science organization, driven by disruptive thinking and courage. We believe it is key to embrace new technologies and science with open minds, as well as with proper focus and prioritization, to identify what our ultimate goals, roles, and responsibilities should be. Learning from others and being part of industry-wide collaborations and open-source initiatives will help

to sharpen our focus and will assist in accelerating this transformation. This is enabled by our organization's culture of respect and mutual support, along with the creation of an environment with dynamic opportunities that allow for expanding our teams' knowledge, taking additional responsibilities, rotations, and secondment opportunities to develop people continuously across the full spectrum of the clinical data flow.

## Competing Interests

The authors have no competing interests to declare.

## References

1. **Society for Clinical Data Management.** The Evolution of Clinical Data Management into Clinical Data Science. A Reflection Paper on the impact of the Clinical Research industry trends on Clinical Data Management. Published June, 2019. [https://scdm.org/wp-content/uploads/2024/03/2019\\_Evolution-of-CDM-to-CDS-Part-1-Drivers.pdf](https://scdm.org/wp-content/uploads/2024/03/2019_Evolution-of-CDM-to-CDS-Part-1-Drivers.pdf).
2. **Catherine Célingant, Lynne Cesario, Patrick Nadolny. Society for Clinical Data Management.** The adoption of risk-based CDM approaches Version 1. Published July, 2022. <https://scdm.org/wp-content/uploads/2024/03/SCDM-rb-CDM-July-6th.pdf>.
3. **Indupuri R, Rocchio S.** Enabling digital transformation: managing external clinical data sources to advance drug development. *Applied Clinical Trials*. Published November 13, 2020. <https://www.appliedclinicaltrials.com/view/enabling-digital-transformation-managing-external-clinical-data-sources-to-advance-drug-development>.
4. **US Food and Drug Administration.** Guidance for industry – Electronic source data in clinical investigations. Published September, 2013. <https://www.fda.gov/media/85183/download>.
5. Tufts CSDD Impact Report January/February 2021.
6. **Clinical Data Interchange Standards Consortium.** CDISC\_360\_Project\_White\_Paper. Published June, 2021. [https://www.cdisc.org/sites/default/files/2021-06/CDISC\\_360\\_Project\\_White\\_Paper.pdf](https://www.cdisc.org/sites/default/files/2021-06/CDISC_360_Project_White_Paper.pdf).
7. **Society for Clinical Data Management.** 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. Published March, 2020. [https://scdm.org/wp-content/uploads/2024/03/2020\\_Evolution-of-CDM-to-CDS-Part-2-Technology-Enablers.pdf](https://scdm.org/wp-content/uploads/2024/03/2020_Evolution-of-CDM-to-CDS-Part-2-Technology-Enablers.pdf).
8. **Society for Clinical Data Management.** The Evolution of Clinical Data Management into Clinical Data Science (Part 3: The evolution of the CDM role) A Reflection Paper on the evolution of CDM skillsets and competencies. Published Aug, 2020. [https://scdm.org/wp-content/uploads/2024/03/2020\\_Evolution-of-CDM-to-CDS-Part-3-Evolution-of-CDM-Role.pdf](https://scdm.org/wp-content/uploads/2024/03/2020_Evolution-of-CDM-to-CDS-Part-3-Evolution-of-CDM-Role.pdf).

9. **Chartie S, Nadolny P, Young R.** Society for Clinical Data Management. The 5Vs of Clinical Data Version 1. Published March 2022. <https://scdm.org/wp-content/uploads/2024/03/SCDM-The-5Vs-of-Clinical-Data-FINAL.pdf>.
10. **TransCelerate. Digital data flow asset – clinical trial process.** [www.transceleratebiopharmainc.com/initiatives/digital-data-flow/](http://www.transceleratebiopharmainc.com/initiatives/digital-data-flow/).
11. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). Clinical electronic structured harmonized protocol (CESHarP) M11. Draft version, 2022. [https://database.ich.org/sites/default/files/ICH\\_M11\\_draft\\_Guideline\\_Step2\\_2022\\_0904.pdf](https://database.ich.org/sites/default/files/ICH_M11_draft_Guideline_Step2_2022_0904.pdf).
12. **European Medicines Agency.** EMA launches pilot project on analysis of raw data from clinical trials | European Medicines Agency (europa.eu).
13. **European Medicines Agency.** European Medicines regulatory network data standardisation strategy. Published December 16, 2021. [https://www.ema.europa.eu/en/documents/other/european-medicines-regulatory-network-data-standardisation-strategy\\_en.pdf](https://www.ema.europa.eu/en/documents/other/european-medicines-regulatory-network-data-standardisation-strategy_en.pdf).
14. **International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH).** Integrated addendum to ICH E6(R1) Guideline for good clinical practice ICH E6 (R2). Published November, 2016. [https://database.ich.org/sites/default/files/E6\\_R2\\_Addendum.pdf](https://database.ich.org/sites/default/files/E6_R2_Addendum.pdf).
15. **International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH).** ICH harmonized guideline Good Clinical Practice (GCP) E6(R3). Draft version, May, 2022. [https://database.ich.org/sites/default/files/ICH\\_E6%28R3%29\\_DraftGuideline\\_2023\\_0519.pdf](https://database.ich.org/sites/default/files/ICH_E6%28R3%29_DraftGuideline_2023_0519.pdf).
16. **International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH).** General considerations for clinical studies ICH E8(R1). Published October 6, 2021. [https://database.ich.org/sites/default/files/E8-R1\\_Guideline\\_Step4\\_2021\\_1006.pdf](https://database.ich.org/sites/default/files/E8-R1_Guideline_Step4_2021_1006.pdf).

**How to cite this article:** Hashio M, Kolodziej M, Olds M, Samuel E, Ravindran S, Mohammed A, Makowski M. GSK Data Strategy & Management Whitepaper 2023: Transformation from a Data Management to a Data Science Enabled Organization – Where Are We Heading? *Journal of the Society for Clinical Data Management*. 2024; 4(1): 9, pp.1–8. DOI: <https://doi.org/10.47912/jscdm.349>

**Submitted:** 07 February 2024

**Accepted:** 13 June 2024

**Published:** 29 July 2024

**Copyright:** © 2024 SCDM publishes JSCDM content in an open access manner under a Attribution-Non-Commercial-ShareAlike (CC BY-NC-SA) license. This license lets others remix, adapt, and build upon the work non-commercially, as long as they credit SCDM and the author and license their new creations under the identical terms. See <https://creativecommons.org/licenses/by-nc-sa/4.0/>.



*Journal of the Society for Clinical Data Management* is a peer-reviewed open access journal published by Society for Clinical Data Management.

**OPEN ACCESS** The Open Access icon, which is a stylized 'O' with a circular arrow around it, indicating that the content is freely available.