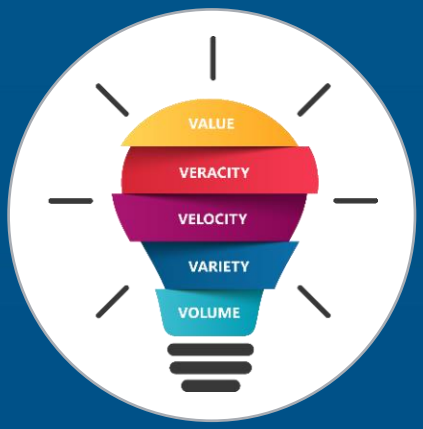

The 5Vs of Clinical Data

Version 1 (March 2022)

Authors

- Steve Chartier, Sr. Director of Engineering, CALYX
- Patrick Nadolny, Global head, Clinical Data Management, Sanofi
- Richard Young, Vice President, Vault CDMS Strategy, Veeva

Reviewers and Contributors

- Sanjay Bhardwaj, Head of Clinical Technology Strategy & Operations, Abbvie
- Joe Kim, Senior Delivery Principal, Slalom
- Jérémy Nadolny, Sr Business Associate, Trial Operations

SCDM would like to acknowledge the many volunteers who have contributed to development of the SCDM reflection papers used as the primary basis for this topic brief.

Methodology

The Society for Clinical Data Management (SCDM) Innovation Committee seeks to provide Thought-Leadership to our industry and support the SCDM vision of "*leading innovative clinical data science to advance global health research and development*". To that end, the SCDM Innovation Committee strives to demystify Clinical Data Science (CDS) and support the development of all Clinical Data Management (CDM) professionals, from subject matter experts (SMEs) working on clinical studies to CDM leaders setting the direction of their organizations.

The Innovation Committee is publishing topic briefs intended to serve as orientation guides on specific areas which are contributing directly or indirectly to the evolution of CDM into CDS. The content of those topic briefs is primarily an extract from the previously published SCDM Reflection Papers[1,2,3] which collectively provide a cohesive and comprehensive overview of CDS from the point of view of industry leaders. Due to the recent emergence of the CDS discipline and the absence of a comprehensive literature base regarding CDS within the Drug Development industry, this content was gathered from industry leaders through a consensus-based methodology. As CDS mature, it is anticipated that literature on this topic will blossom.

Introduction

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 data are re-shaping CDM.

First and foremost, 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 V dimensions.

Volume - Variety – Velocity - Veracity - Value

To successfully move from the current CDM to next generation of CDS paradigms, the way we collect, interact with, and gain insights from different subsets of data must evolve.

	Current CDM Paradigm	Desired CDS Paradigm
Volume	• 10's to 100's of datapoints per patient per week (i.e., few datapoints per eCRF)	• Thousands to billions 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), other structured and unstructured data
Velocity	• Days, weeks and months • Entered in the eCRF from days to weeks after patient's visits	• Near real-time • RESTful APIs providing interoperability between computer systems
Veracity	• Exact copy of source, ALCOA • Mainly confirmed by SDV and queries • Perfect (i.e., 100% error free) • Manual (clinical and data management), statistical and scientific Reviews	• Focused on what matters (e.g., critical data and processes) • Risk-based data strategies • AI-driven automation detection and resolution of issues • Fit-for-purpose (i.e., scientifically plausible and strong enough to support reliability of trial results)
Value	• Focused on regulatory submissions	• Broader secondary use (e.g., synthetic arms, patient engagement, machine learning training datasets, 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 support the implementation of emerging data strategies and enable the aggregation, integration, and interpretation of high volumes of data, continually generated from many sources with complex data structures. CDM must also derive relevant secondary data assets to extract the full value of our data.

Topic brief

Let's explore the 5 Vs of clinical data and understand their meaning!

Volume (Terabytes, petabytes, exabytes to yottabytes)

In 2012, Tufts[4] estimated that on average, phase III studies collected close to 1 million data points. Today we measure m-Health data points in the billions. This dramatic increase 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 1 shows the expected volume of actigraphy data generated by wearables (in blue) compared to data generated from site visits (in orange) which is barely visible on the figure by comparison. 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 throughout the study with a spike at almost 375,000 readings per day. In comparison, pulse would only be generated 3,380 times through site visits, assuming patient's visits every 2 weeks with at most 260 readings in a week across patients.

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.

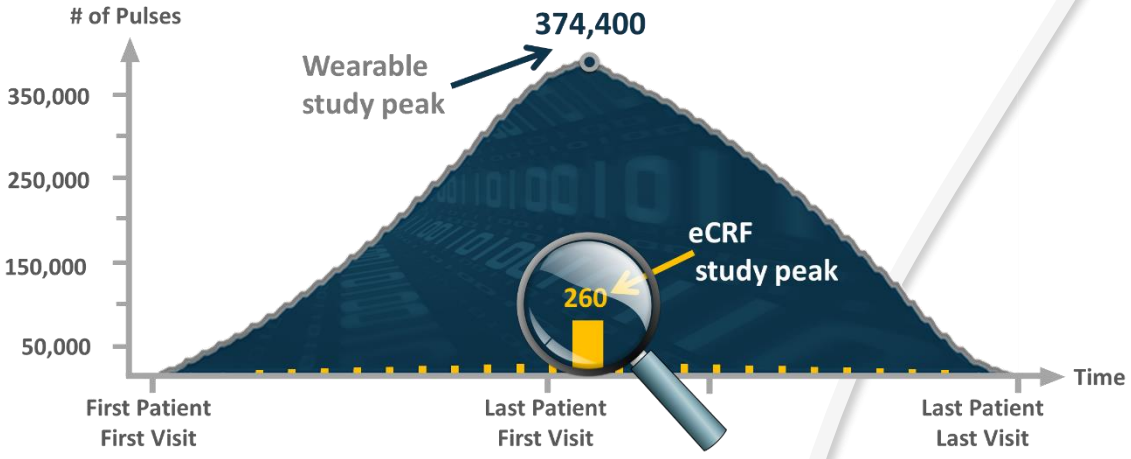

Fig 1. Daily volume of actigraphy data from wearable vs. weekly e-CRF pulse data

Another example of data expansion is the increase in focus on wellbeing, and the volume of passive data sources now being made available for research. According to the IQVIA[5] 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.

Variety

With more than 200 new health apps added to app stores every day[5], it is not surprising that sponsors are increasingly using digital health technologies in clinical research and leveraging apps to collect a variety of data including 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. As clearly seen during the COVID-19 pandemic, patient centric solutions such as telemedicine and home nursing also benefit sponsors by reducing on-site activities, optimizing site selection, speeding up enrollment, easing data collection and supporting rapid decision-making through immediate access to data.

To achieve our CDS goal of managing this growing volume and variety of data, we must develop new clinical data science principles, data collection tools, data review and data analytics strategies. As an example, patient centric data collected from eCOA, m-Heath solutions, EMR, sensors and wearables are considered eSource. It is almost impossible to modify them once they have been generated. This means that feedback on the data quality and integrity of this variety of eSources needs to be provided at the time of data generation. After data is generated, CDM will rarely be able to send a query to request its correction. So, data anomalies will likely need to be tagged and explained for the most part. However, will data tagging be enough to deliver reliable data to reach sound conclusions required for regulatory approval? Beyond data tagging, MHRA introduced the concept of "data exclusion" [6]. This mean that "unreliable data" with a potential of impacting the reliability of the trial results could be excluded based on a "valid scientific justification, that the data are not representative of the quantity measured" [6].

Additionally, in accordance with good record keeping and to allow the inspection and reconstruction of the original data, "all data (even if excluded) should be retained with the original data and be available for review in a format that allows the validity of the decision to exclude the data to be confirmed"[6]. Even if not widely use yet, data tagging, and exclusion may become a standard practice within CDS in years to come to support the generalization of eSource and decentralized clinical trials (DCTs).

Furthermore, 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 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 interpretation?

Additionally, even though not new, the implementation of Audit Trail Review (ATR) is gaining momentum in supporting study monitoring. This is also fueled by the more frequent focus on audit trails from GCP Inspectors. Those can provide critical insights on how the data is being collected, leading to the identification of process improvements or lack of understanding of the protocol instructions, up to the rare cases of manipulation from data originators[7]. The support of ATR requires the acquisition and integration of audit trails from all sources which is therefore contributing to the increased volume and variety of data.

Velocity

To support real-time data access, we need to understand, prioritize, and synchronize data transactions into appropriate data storage at increased volume, speed, and frequency. Data from wearables for instance, can be generated 24 hours a day, 7 days a week. Taking the example in figure 1, 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". The integration of high volume and variety of data at high velocity 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 Wi-Fi enabled, therefore 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.

Additionally, regardless of the data acquisition and integration technology being used, we need to synchronize the data flow velocity to our needs across all data streams. As patient's data is highly related to one another, we need to review and correlate multiple data sources simultaneously. As an example, it would not make sense to reconcile two data sources extracted months apart.

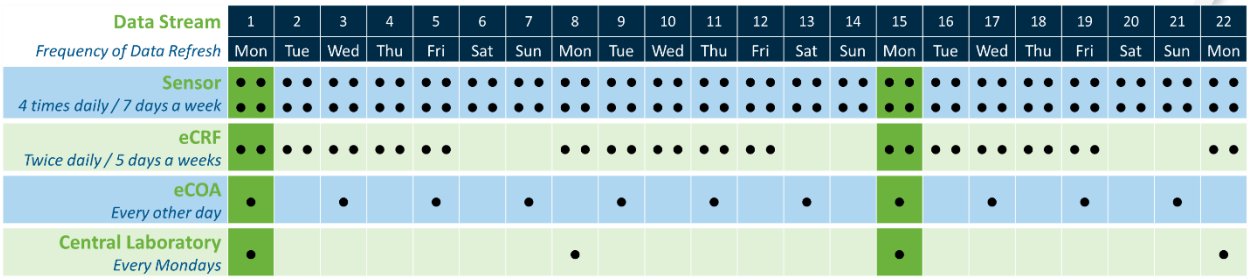

Fig 2a. Data transfer frequency VS. data reconciliation / consolidation

Referring to the simple theoretical in Fig 2a, the data could be synchronized and therefore optimally reconciled only every two weeks. As shown in Fig 2b, changing the eCOA transfer frequency from every other day to 3 times a week on Mondays, Wednesdays and Fridays would at least enable weekly optimum data reconciliations and reduce data refreshes workload throughout the study life cycle.

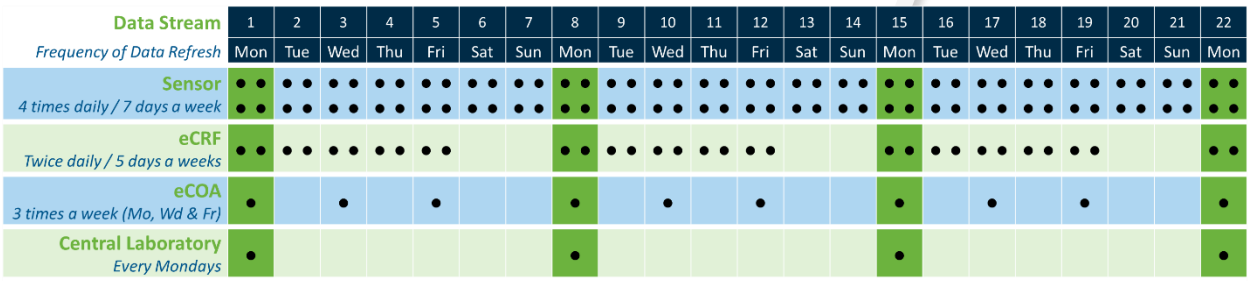

Fig 2b. Data transfer frequency VS. data reconciliation / consolidation

The need for data synchronization would be true for all other data driven activities including ongoing data and safety reviews, risk assessments, etc. Synchronizing data flows would prevent rework resulting from data refresh misalignments. Additionally, moving forward, synchronizing data velocity will be more frequently driven by remote working practices. How we integrate data cleaning into site driven workflows is playing a critical role in our ability to be agile. As an example, performing SDV during the COVID-19 pandemic forced us to look at alternative solutions to remotely synchronize data, documents and processes with the sudden loss of physical access to sites. These new remote practices will in future require CDS to explore new data review capabilities with sites, beyond simple query-based clarifications.

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 conformity, 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.

First, we must not let perfection become the enemy of the good, especially where "good" is fit for purpose and good enough. If veracity maps a journey towards fit-for-purpose, we must assess how far we pursue perfection for each data type. Directionally, the concept of Quality Tolerance Limits (QTLs) is a good example of a fit-for-purpose and measurable quality framework that can be used across data streams. 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 also determine and enforce fit-for-purpose data quality standards. In this setting, we can define a positive and a negative goal for data veracity. Positively, we can attain to deliver data veracity not to exceed a set tolerance limit (e.g., not exceed x% of missing data). We can also assign a negative target, where we attain to remove any issues (e.g., address all missing data) that would alter the end analysis. It is often the case that this latter goal ("negative") will be easier to define in our cleaning strategy as defining quality target requires historical information and may be perceived as subjective. However, trying to eliminate all data issues may be neither attainable nor desired for non-critical data. So, CDS must learn how to set-up measurable and objective quality target by truly representing our data veracity objectives.

We must, most importantly, focus on Quality by Design (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.

It is no longer possible to use manual processes based on listings or patient profile to confirm data veracity from such a large volume of disparate data coming as such velocity. It is necessary to implement different strategies moving beyond data filtering and trending to strategies based on story telling visualizations, statistical and Machine Learning (ML) models as well as leveraging intelligent automations. Interrogating such data may require different technology expertise such as non-SQL.

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.

Value

Importantly, 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 current CDM context, we value quality data enabling the reliable interpretation of the trials results. In the context of CDS, the value of data goes beyond integrity and quality to ensure its interpretability. 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. It means that we need to seek patient authorization up-front for using their data for purposes other than the scope of the protocol through unambiguous informed consent forms. Some companies are beginning to re-use clinical and health data in new ways influencing others to seriously consider it.

Examples include:

- Creating synthetic arms either from past clinical trials or from RWD,
- Engaging and retaining patients by feeding them back with study wide data summaries during study conduct,
- Creating machine learning training datasets to improve operational processes such as automating query detection or enhancing the reliability and accuracy of endpoints assessments,
- Extracting real world evidence from real world data to gain insights on how to improve standard of care or better understand drug effectiveness in a real-world setting.

While those are a few examples, they highlight the potential of maximizing the value of clinical data.

Conclusion

At the end of the day, through the application of proven data strategies, we can leverage emerging technologies to extract the full value of data for all stake holders (i.e., patients, sites, sponsors, regulators, caregivers, and payers) and defeat data silos, the enemy of our value extraction journey.

As depicted in figure 3 below, the value extraction journey begins with generating insights from raw data, leading to knowledge, which in turns generates intelligence which can be used for automations, predictions, scientific conclusion and optimization of processes and reduction of efforts.

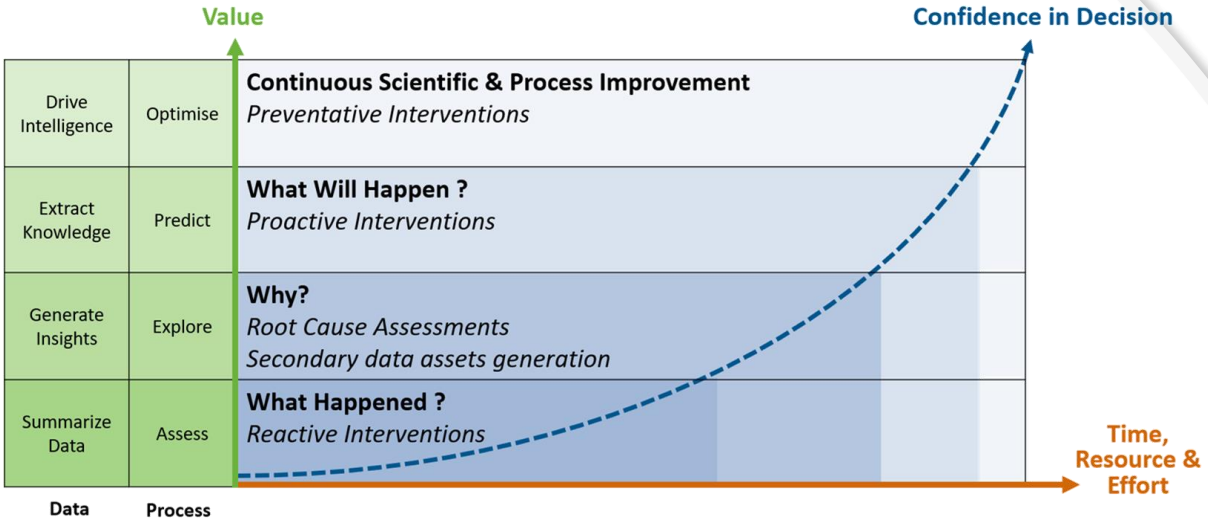

Fig 3. The data value extraction journey

The nature of data will not stop evolving complicating the data value extraction. The 5 V's will grow and in future we should consider discussing variability, validity, virality, visualization and viscosity, to name a few of them. It is important that CDM owns these data decisions, not only to optimize the current trials but also to set the agenda for future trials. Using data to change our approach, to enable more effective trials is incredibly important to the future of clinical research, not just to CDS.

Ultimately, CDS's remit must ensure the veracity of data coming from a variety of sources with high volume and high velocity. We can provide the technology enablement that our Clinical Data Scientists need to extract full current and future value of our treasured clinical research and healthcare data.

Fig 4. The 5Vs data journey from collection to value generation

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/wp-content/uploads/2019/09/SCDM-Reflection-Paper-Evolution-to-Clinical-to-Data-Science.pdf

[2] SCDM, March 2020, The Evolution of CDM to Clinical Data Science (Part 2: The technology enablers) – A Reflection Paper on how technology will enable the evolution of Clinical Data Management to Clinical Data Science. Available at https://scdm.org/wp-content/uploads/2020/03/SCDM-Reflection-Paper-Evolution-to-Clinical-to-Data-Science-Part-2.pdf

[3] SCDM, August 2021, The Evolution of CDM to Clinical Data Science (Part 3: Evolution of the CDM role) – A Reflection Paper on the evolution of CDM skillsets and competencies. Available at https://scdm.org/wp-content/uploads/2020/08/SCDM-Reflection-Paper-CDM-Role-Evolution-Part-3-.pdf

[4] Tufts, November 2012, Clinical Trial Complexity, available at https://www.nationalacademies.org/documents/embed/link/LF2255DA3DD1C41C0A42D3BEF0989ACAECE3053A6A9B/file/D774AC7AEFEE0E80425328941C08CF3E6CE97D0BF748

[5] 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

[6] MHRA, March 2018, 'GXP' Data Integrity Guidance and Definition, Available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/687246/MHRA_GxP_data_integrity_guide_March_edited_Final.pdf

[7] SCDM and eCF, Audit Trail Review, An Industry Position Paper on the use of Audit Trail Review as a key tool to ensure data integrity Available at https://scdm.org/wp-content/uploads/2021/04/2021-eCF_SCDM-ATR-Industry-Position-Paper-Version-PR1-2.pdf

Main abbreviations

API	Application Programming Interface
ATR	Audit Trail review
CDM	Clinical Data Management
CDS	Clinical Data Science
DCT	Decentralized Clinical Trial
eCOA	electronic Clinical Outcome Assessment
eCRF	electronic Case Report Form
EDC	electronic Data Capture
IxRS	Interactive Response System (x being any type including Voice or Web)
ML	Machine Learning
QbD	Quality by Design
QTL	Quality Tolerance Limit
RWD	Real World Data
SDV	Source Data Verification
SME	Subject Matter Expert