a) Clinical research concepts and strategies

Alternative protocol design is not new. Adaptive trials have been in use for many years and the adoption of master protocol designs increased almost nine-folds between 2010 and 2019[12]. According to the FDA, a master protocol is a protocol designed with multiple sub-studies, which may have different objectives and involves coordinated efforts to evaluate one or more investigational drugs in one or more disease subtypes within the overall trial structure. The three types of master protocols which could include adaptive design elements are umbrella, basket, and platform designs, each bringing benefits but leading to the challenges described below:

Adaptive

- Increase to the onset planning times to cater for possible adaptation scenarios
- Dynamic design increases the complexity of data review and data flows especially as most aspects of study design could be adapted including endpoints, treatment arms and data collection
- Complex simulations required to model trial design and adaptations

Umbrella
X Drugs → 1 Indication

- Multiple drugs with potentially different routes of administration increasing complexities with DCI design, investigational product (IP) distribution and safety review for concomitant medications and adverse events
- Complicates drug management and masking procedures

Basket
1 Drug → Y Indications

- Multiple diseases requiring increased domain expertise of CDS team
- Each indication would likely require specific endpoints
- Greater variation of participant population increases complexity for data reviews

Platform
X Drugs → Y Indications

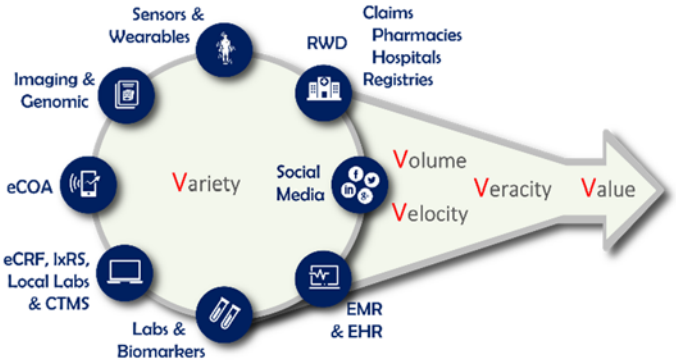
- Inherits complexities of umbrella and basket designs
- Additional upfront planning to consider all data related scenarios
- Increased number of interim analysis
- Multiple sub-trials may require dedicated or specialized CDS teams to manage

A common challenge associated with these four designs is the need for flexible DCIs requiring complex and dynamic branching logics in EDC, eCOA and IRT. Such designs could be implemented within the context of traditional bricks and mortar sites or in a decentralized setting.

Study set-up is no longer a matter of translating an approved protocol into eCRFs. It is evolving as a specialized activity where the Clinical Data Scientist needs to master scientific and operational concepts and being able to drive the set-up of an integrated and flexible technology centric data ecosystem. It is also important to know that scenario planning is multidimensional as adaptive design, master protocol and decentralization of the clinical trial are not mutually exclusive (i.e., they could all happen concurrently in a single study).

While the overarching benefits of these trial designs can accelerate the overall development of a new drug, they will challenge Clinical Data Scientists in terms of design, controls, data review, and increased frequency of interim analysis. All of this will require careful planning to execute properly. In addition to the need to think critically and no longer logically, this does imply that the foundational knowledge required to do so is substantially increased.

b) Clinical data acquisition, standards and modeling



For Clinical Data Scientists, the most critical of the 5Vs is understanding the value of the data as not all data are created equal and will not bring the same value. Assessing how data sources will contribute to the primary objectives of the protocol is and will remain a core competency, however assessing the secondary value of each source and extracting the right evidence from it, whether it be for synthetic arm, translational research, and patient engagement will drive the focus of how that data is managed and the effort allocated to it. This is already being done for reimbursement by extracting real world evidence (RWE) from real world data (RDW).

Assessing the variety of the data sources collected in the clinical trial, where they are going to be integrated (e.g., in EDC or not), how they are transformed into useful information to generate insights, and how they will drive action is the second highest priority for Clinical Data Scientists. Is the data structured or unstructured? What applicable data standards can be used? How will the data be interpreted, by whom, and when is it needed by? All of these questions lead directly to how the data should be modeled. Data modelling for CDISC standards, such as CDASH, SDTM and to a lesser extent BRIDG and ODM, are already part of our core competencies, however we will need to expand experience and proficiency in other standards such as HL7, FHIR, in order to fuse RWD sources such as EMR/EHR into our intelligent CDMS and enable approaches like eSource and direct data capture.

Being able to assess the individual data source's volume and velocity in tandem will directly inform the Clinical Data Scientist on what is the optimal approach to managing the data. High volumes will require automated solutions to assess the quality and integrity of the data. On the other hand, high velocity data sources will require new approaches that drive action by detecting and promptly differentiating signals from background noise. Clinical Data Scientists will need to be able to develop processes and controls to identify appropriate signals when managing high volume and velocity data.

While the veracity of the clinical data will largely be supported and controlled by the underlying technology, there will be instances where the sponsor's technology does not directly control the credibility and integrity of the data. Situations where the sponsor's CRO and/or sites are managing the underlying technology or the source of the data (e.g., EMR) which cannot be corrected requires the Clinical Data Scientist to be able to think critically about how to assess the credibility and integrity of the data and equally important, effectively partner with the external parties to understand their controls and what additional ones will be required.

c) Automation technologies

First and foremost, organizations must anticipate the implication of using AI based solutions leading to the set-up of digital workforces that can work 24 hours a day and be "hired" (i.e., implemented) and on-boarded quickly and at a decreasing cost. This digital evolution will profoundly transform the workplace in years to come. Also, as we leverage advanced capabilities such as ML, RPA and IPA, we need to augment our approaches to testing, deployment, and management. It is easier to automate chaos than it is to automate order.

Conceptually, RPA has similarities with edit checks. RPA scripts using simple branching logics and Boolean operators (e.g., and, or, not, etc.) can be applied to automate repetitive and transitional processes. Using those scripts, an RPA bot will act very much like an end user, capable of logging in and out of systems to perform basic data-oriented tasks. The Clinical Data Scientist will need to not only be able to identify the appropriate use cases where RPA can be applied but be able to assess the value of automating it given the underlying cost and time involved. Freeing up part of an FTE per year, may not offset the cost of design, development, and testing of an RPA bot, whereas freeing up several FTEs per year will result in a positive return on investment (ROI). The cost needs to be balanced with quality as the automation of large scale repetitive manual tasks could result in meaningful process accuracy and reliability improvements. Additionally, new procedures will be required to test and manage the identity and deployment of bots, given they will perform steps like any human, and their actions will be recorded in systems audit trails. Auditors will undoubtedly seek to understand how RPA bots were developed, tested, and managed in production environments.

Intelligent solutions powered by AI require the most significant transformation for CDM. Overall, ML based solutions will act as virtual Clinical Data Managers assisting Clinical Data Scientists. As a result, expert Clinical Data Scientists will mentor virtual Clinical Data Managers to accurately perform data reviews and other CDS activities. This means that when ML models are established, Clinical Data Scientists will need to define the objectives of the intelligent solution and identify the datasets required for training and testing. These datasets will need to cover all expected data review scenarios (i.e., be complete) and be truly representative of the use cases anticipated in production (e.g., include data review scenarios across all study phases and therapeutic areas).

The failure to define the right datasets could bias the system behaviors and lead to inconsistent data review accuracy. Clinical Data Scientists also need to play a key role in assessing the performance of the solution by defining the right testing strategies (e.g., by setting a minimum ratio of the number of correct predictions compared to the total number of inputs in the test dataset[2]).

The figure below is an example of a process for the "mentoring" (i.e., training) of an ML-based solution.

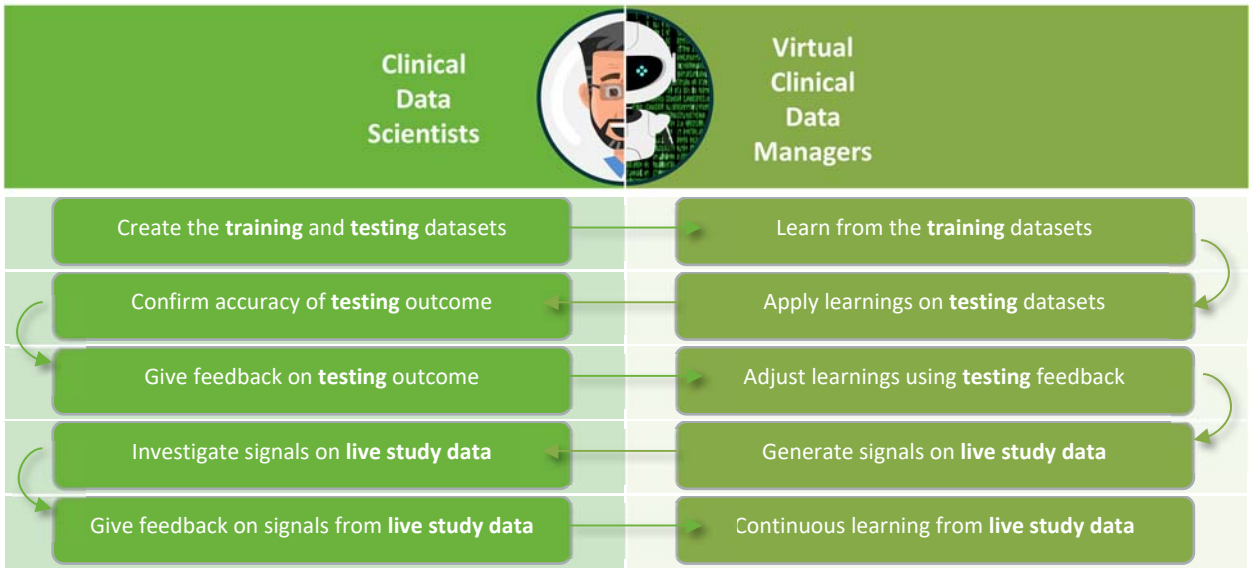


Fig 8. Example of learning process for an ML-Based solution

ML technologies require fundamentally different approaches to their development and validation including the use of methods like recall (i.e., ability to predict in all cases), precision (i.e., accuracy of the predictions) and F-measures (i.e., combining recall and precision) to determine the reliability of models.

First, the model must be trained on a controlled dataset of known quality. Then, it can be applied to the testing dataset to assess the precision and recall in a simulated "real-world" environment to understand its precision and recall.

Its acceptable level of accuracy should be one that is better than the accuracy of the current process. If not, its use needs to include commensurate human supervision and training to monitor its accuracy. Lastly, once a machine learning model is deployed, it will need to be retrained at specific intervals in order to maintain and ultimately improve its accuracy over time.

d) Vendor oversight

Almost every trial initiated by a sponsor today is enabled by a multitude of vendors, providing services ranging from strategic to tactical and global to localized. These partnerships manifest themselves in varying forms, from multi-years relationships to one-off contracts.

In recent years, health authorities have placed more responsibility on sponsors in terms of vendor oversight, and this has been further solidified by the latest addendum of ICH E6 (R2) [7]. Clinical Data Scientists, who oversee external vendors, will need to transition from managing status and timelines to overseeing quality delivery through data-driven insights on vendor performance that are tailored to the services rendered.

Facilitating this will require direct access to a vendor's operational data, centralizing raw data internally, and developing capabilities to analyze quality and performance on a continuous basis, rather than the more traditional, passive approaches. Additionally, the Clinical Data Scientists will also need to strengthen their relationships and understanding of interdependencies with Clinical Operations, Quality Assurance and Procurement functions to effectively manage a vendor with a single, integrated voice from the sponsor.

e) Influential leadership

Given the dynamic and evolving nature of the clinical research landscape, Clinical Data Scientists will have a central role in leading cross-functional teams as well as driving complex decision making. CDS is not a support function but a key stakeholder responsible for the most critical asset: the study data. Clinical Data Scientists must therefore demonstrate influential and leadership skills. Doing so requires business acumen, technical capabilities, and the ability to manage continuous change within their organizations and teams. As discussed in previous sections, Clinical Data Scientists will need to meaningfully expand their core competencies and foundational skills.

Additionally, there will be an increased emphasis on the soft skills such as:

- Understanding the points of view of a wider range of stakeholders including sites and patients
- Critically assessing risks and their impacts to determine best mitigation strategies
- Suggesting alternative and operationally sound solutions (i.e., being a pragmatic innovator)
- Articulating complex technological and scientific concepts
- Understanding the ramifications and rational behind study team decisions

Ultimately, Clinical Data Scientists must be tactful and empathic listeners able to drive consensus around complex scenarios and if necessary, demonstrate decisiveness by taking and owning decisions.

In conclusion, we could compare the scope of the CDM vs. Advanced CDS scope as:

CDM Scope	CDS Scope
one and two dimensional trials	Three or more dimensional trials
Project management	Cross-functional leadership
Randomized controlled trials	Adaptive & master protocols
Vendor management	Vendor oversight
Clinical research standard	Clinical research and healthcare standards
Logical thinking	Critical thinking

This evolution would require the following CDS roles requirements:

Best Practices	Soft Skills
• Intelligent systems management • Generation of secondary data assets (e.g., synthetic arms) • Data and system integrations	• Critical thinking • Pragmatism • Influential leadership • Ability to manage ambiguities and dynamic environments
Competencies	**Foundational Knowledge**
• Advanced analytics • Vendor oversight • Patient centric technologies	• New research methodology (adaptive, master protocols) • Decentralized clinical trials approaches and technologies • Risk-based methodologies and regulations • Understanding of new data concepts such as sequenced data and unstructured data • Health care standard models and terminologies • Automation and AI concepts (e.g., supervised vs. unsupervised ML)