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.

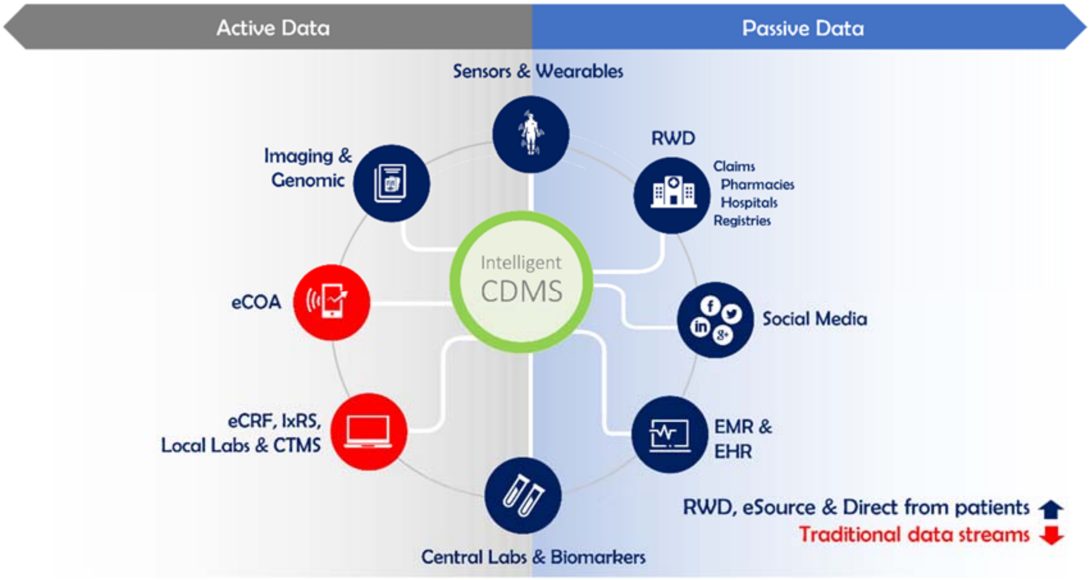

Fig 3. Clinical research technology ecosystem

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.