



Richard Young,
Vice President,
Vault CDMS,
Veeva Systems

Data First: Building a Foundation for Digital and Decentralized Clinical Trials

Without a strategy for managing clinical data centrally, sponsors lack the visibility needed to optimize trial outcomes

Innovation and adaptation around COVID-19 restrictions have brought a “can do” spirit to the industry, particularly in clinical trials. This led to an increase in adoption of decentralized models that allow patients to participate in trials without repeated site visits, offering sponsors a way to increase enrollment and prevent attrition.

But the buzz surrounding patient centricity and decentralized trials has drowned out an inconvenient truth. Sponsors have no clear-cut way to aggregate and review the huge volume of patient data being gathered from disparate sources. This problem is not new - the industry has been working with decentralized trials, in some form, for over a decade. And looking ahead, to truly run a connected and digital clinical trial that enables decentralized approaches, companies will need complete and concurrent clinical data.

Sponsors may be able to capture patient data remotely and in real-time, but they cannot verify and reconcile it in anything close to real-time. Instead, most organizations use tedious, manual methods to aggregate and clean each silo individually.

We’ve heard from one Top 20 sponsor that, for a single trial, such efforts required 27 people working in shifts, 24 hours a day, for six weeks. Considering that most big pharma companies run hundreds of trials each year, the time and costs required to do this for every trial are prohibitive.

An Explosion in third-party sources and non-static formats

The roots of this problem go back to the earliest days of traditional electronic data capture (EDC), where we saw data sources fragmented and isolated as we opted for speed of collection and sacrificed speed of analysis. With novel trial designs creating

a drive for faster decisions, the ongoing explosion of data required for both traditional and decentralized clinical trials is changing our thinking. Today, a typical Phase III trial uses close to 10 data sources and generates an average of 3.6 million data points, three times the level that was seen 10 years ago.¹ One study found the cost to support data transfers between systems or companies in life sciences is \$156 million annually.²

Increased use of third-party data sources and non-static patient data have added complexity. Where, 10 years ago, most key clinical data came from physicians and was stored in the EDC, only about one-quarter of that data is now stored in the EDC. The remaining 75%—third-party data from smartphones and other sources—is managed independently and must be reconciled against the EDC. More and more frequently we see primary efficacy and safety data coming from outside of the eCRF, creating increased pressure on data integration strategies.

Further complications are introduced by non-static data formats, such as readings from wearable devices and monitors. Data formats differ between different systems, even for the same type of data—patient heart rate, for example. If measured via a stethoscope in the clinic, it will be recorded as a single data point, but a wearable device will produce continuous, high-frequency data. In the end, huge volumes of data collected at different frequencies will need to be managed, reconciled, and interpreted.

The added complexity of digital and decentralized trials

With decentralized trials, data that was once solely collected at the site may also be collected during telehealth visits, in-home visits, through ePRO apps, and more, each creating its own silo. Frequently, these

scenarios require additional systems to collect the data and one size rarely fits all.

To get a clear view of the patient, this data must be aggregated and harmonized. For example, to see a patient's heart rate, there are multiple places to look—the EDC, the in-home visit log, or the iWatch reading. The source for each patient can also vary by visit—a trip to the clinic one day and an iWatch reading the next.

The same data is collected by different systems, at different times, and structured in different formats for decentralized trials, which makes synchronization significantly harder. This makes traditional data management capabilities such as querying important, yet it is not included in newer data collection tools.

One sponsor shared the challenges it faced after installing a system designed to handle patient data from remote nursing visits. The application did not include a querying tool. As a result, when data discrepancies emerged, the company faced three ugly choices: querying the data via email outside of the system; re-keying the data into its EDC; or paying for an expensive one-off integration. As this example illustrates, sponsors need an infrastructure that brings data together in an effective and scalable way.

When trials are fully digital with no paper, sites must use eSource, and an incremental challenge emerges. In these cases, data are not anchored to an EDC, which has traditionally served as the backbone for trial data, providing a reference point for other sources to be checked against. In a world of eSource that is managed by sites and where practices vary by site, working without the EDC means losing a fixed anchor against which to reconcile data. As a result, the data is much more difficult to clean.

When eSource and EDC co-exist in separate systems, sponsors must bear the enormous effort and cost required to reconcile the two. What will happen when we create truly patient-centric processes, whereby the patient will decide which visit they attend in person, versus perform remotely?

Mixing and matching at the data point level will become the norm, and current linear solutions are not designed with that in mind. Add to that the impact of protocol amendments and adaptive designs, and the conclusion is simple: companies can make it work for their important trials, but costs are prohibitive for the average study.

An integrated platform that connects the patients and sites with the sponsor's infrastructure for one point of cleaning and review would eliminate many

of these challenges. Unfortunately, such an infrastructure doesn't yet exist, although many technology providers are working to deliver such solutions.

Time lag prevents data-driven decision making

In decentralized trials, the time lag between data collection and the availability of clean data makes it more difficult to make informed decisions during the trial. Many new data-collection instruments are stand-alone tools that lack data review capabilities. When data needs to be reviewed, it must be transferred to the sponsor or CRO and imported into a separate system. When discrepancies are found, data managers must resort to disconnected email exchanges to issue queries, adding further delays and more manual work.

In 2020 and 2021, after COVID-19 restrictions took effect, some pharma companies invested considerably in decentralized trial technologies, only to find themselves with data that could not be connected or verified. They've waited months to extract, clean, and reconcile it with their EDC data, finding unexpected anomalies, such as different dates for a patient's adverse event. There has been no easy way to query data sources and ensure validity.

Delays viewing the data prevent trial practitioners from making data-driven decisions in a timely manner or being able to assure regulators that the data represents a completely accurate account of each patient's experience. For example, investigators may need to determine why a sensor reading appears out of range, such as when one patient's blood pressure suddenly spikes. Currently, there is no rapid way to verify these type of root causes.

Consider a rare disease trial where each patient's outcomes are potentially meaningful to the others. If one patient's diagnostic readings trigger a change in the treatment plan, sponsors should be able to make that change instantly. If there is a delay before the data can be cleaned and checked, it leaves them liable for failing to stop potentially harmful treatments.

Establishing end-to-end data flows will be crucial if decentralized trials—and not just data collection—are to run in real-time. Not only will it be key to clinical trial agility and ensuring the validity of results, but it will also be a pre-requisite for adaptive trials.

Ironically, the industry's push for clinical innovation has only compounded the data management challenge. New technologies are being overlaid on a data management foundation that hasn't changed in decades. Designing a trial that is adaptive, digi-

tal and connected, and allows for decentralized execution (all in one protocol) with systems and solutions that support it must be the long-term goal.

Standards help but will never fully solve the challenge

There is a clear and understandable desire for greater development and use of standards to address challenges. The industry has spent over a decade working to define data standards, yet still struggles with overall diversity and complexity.

“Standardization is very important if you want to simplify trials. Right now, you have 10 to 15 different external data providers, and everyone uses a different way to ingest data. Imagine what happens when you move to 20 or 30 different data sources. How will you standardize that data?” asks Mayank Anand, vice president and global head of data strategy and management at GSK.³

But there is no easy way to standardize. The needs for standards are diverse, and there are different standards for how data should be collected, moved, analyzed, and submitted. In addition, the clinical environment is dynamic: data changes; needs change; the understanding of science, and the human body, changes. Considering the pace of innovation in life sciences and the speed at which new data sources are introduced, it is not realistic to rely solely on standards.

The path toward complete and concurrent data

Company strategies for digital and decentralized trials must incorporate plans to connect the myriad sources of patient data into a single clinical data management system. With decentralized trials, the same data for different patients will be collected in different ways. Aggregating data in a central clinical data management system (CDMS) is crucial to achieving the visibility and timeliness that we know contribute to more effective trials.

Technology providers are exploring different ways to achieve this connection, but some options scale better than others. One approach is to use a clinical database or data workbench that stores clinical data in one place, allowing it to be cleaned and harmonized.

Veeva is working on such an approach, Veeva CDB, with several of the top 20 pharma companies.

Veeva CDB includes a data lake that holds disparate datasets in their native structure.

These datasets are mapped to the study backbone using five metadata fields as a simple “key,” rather than requiring a complete mapping, transformation, or adherence to a standard. This auto-mapping helps make data for the same point from different sources more equivalent (e.g., a blood pressure reading from a site visit vs. one taken from a device at home). Data can be ingested, aggregated, cleaned, and made readily accessible to other stakeholders in the organization.

One thing is clear, sponsors seeking decentralized data collection on the front end need centralized data management at the back end to prevent fragmented, heterogeneous data from slowing trials down. Automating the ingestion and harmonization steps will eliminate the time lag between data collection and access to clean data. Aggregating and cleaning sources simultaneously also address the need for patients to be treated consistently. For example, patient data from digital sources do not receive preferential treatment, which could otherwise result in questions of bias.

An additional approach to this problem will be to unify EDC and eSource solutions so that they capture data into the same system, which would allow data queries to be handled via EDC tools. Sponsors should also consider the benefit of having a single platform and data model extending from the patient to the sponsor. These requirements leave data experts with a strategic question: should they extend patient-facing data collection tools into the sponsor data environment, or move in the opposite direction?

Whatever approaches are used in the future, the industry clearly needs to make it easier to work with external data. “Pharma is already struggling to manage the volume of data we have for trials today,” said Anand. “Over the next few years, as the speed of data ingestion increases, the industry will expect clean data output to be faster too.” ☺

References

1. Tufts CSDD Impact Report, Rising protocol design complexity is driving rapid growth in clinical trial data volume. Volume 23, Number 1, 2021
2. Liaison Healthcare Informatics, Managing and Integrating Clinical Trials Data: A Challenge for Pharma and their CRO Partners
3. Applied Clinical Trials, Addressing Digital Trials, 2021