The End of Data Chaos: Modernising Data Infrastructure to Revolutionise Clinical Research

In the world of life sciences, data is everything. Better data leads to better research, which in turn leads to better outcomes that improve health and wellness. At least, that's how it's supposed to work. In reality, pharmaceutical companies and clinical research organisations (CROs) are often buried under massive amounts of data that is difficult to analyse effectively and rapidly. Compounding this problem is that data often comes from multiple sources that are not inherently compatible or standardised and require additional transformations, which adds expenses and delays to the development of new drugs and therapies. In many ways, it is a paradox that having more data sources has made research a less efficient process. However, there are a number of approaches that life sciences organisations can take to harmonise their data, eliminate delays, and achieve better outcomes.

Why is there a problem in the first place? With many trials using six or more data sources, and much of that information originating from external or non-EDC sources, data chaos in clinical studies is now the norm. The exponential increase in the volume, variety, and velocity of data being used in trials has led to data integration, reconciliation, and review challenges that contribute to delays in trial timelines. To improve efficiencies across clinical studies, the life sciences industry needs to reimagine the clinical data ecosystem in a way that focuses on interoperability, simplifying complexities, and streamlining the flow of all clinical data. What is needed are new technologies and processes to centralise and manage disparate data sources. Building a data infrastructure designed for clinical development that automates the flow of data from collection to analysis enhances data quality, improves efficiencies, and promotes the ability to use artificial intelligence and machine learning techniques to predict risk.

The Data Chaos Problem

In theory, having more data sources should create better insights. After all, isn't the goal to have as much data as possible to lead to richer views of the patient experience? The reality of clinical development is more complex. Data arrives from many different sources – real-world data, genomics, wearables, and more – each with its own formatting and standardisation. Each source has its own way of standardising its data, meaning that these sources are not ingested in a neat and organised way. This causes data managers and scientists – those responsible for ensuring the consistency, integrity and quality of clinical trial data streams – to devote significant time to cleaning and organising data and reviewing different sources in many different ways. In order to compare data with other similar trials to assess the safety of a new medicine, data must be standardised. That's the only way to draw valuable insights from the information. The process of mapping data to standard formats and then analysing it can be both time- and cost-intensive.

Despite industry standards, data across the product life cycle is highly varied, volumes are increasing, and the pace at which the data is being created is far faster than five years ago. For starters, information that is not easy to analyse requires multiple steps just to get it prepped adequately for researchers to work with it. Taking weeks or even months to do, it is a never-ending process because new data is always ingested into the pipeline. This is actually one of the major sources of delays in getting therapies and drugs to market. In simple terms, if scientists do not have adequate data management protocols, they cannot access or analyse important information in a reasonable timeframe. These kinds of inefficiencies directly affect clinical trial operations when valuable research time needs to be devoted to what is essentially an IT problem.

These increased cycle times are not only inefficient, they are expensive. Biopharmaceutical organisations increase their workforces just to stay ahead of the data onslaught. This directly affects cycle times, impedes speed to market, and also hits the bottom line. It is a lose-lose for everyone. In fact, there is a 40 per cent increase in last patient last visit to database lock cycle time for companies that incorporate data from five data sources, and most large trials have no fewer than eight sources. This creates extreme delays for important trials and therapies.

It's easy to think about this as a theoretical problem or a process issue, but it's actually a significant business problem for life sciences companies. Organisations that don't have effective data management tools in place cannot get important insights or analyse data in a reasonable timeframe. Both of these problems can lead to significant inefficiencies in clinical trial operations through increased cycle times and labour costs.

New Impacts on Effective Data Management

In recent years, with the advent of handheld computing, researchers have been trying to create new models for clinical trials that use technology to improve the patient experience in research, make participation in trials across the globe more accessible, and speed the pace of development. When the COVID-19 pandemic upended the clinical trial industry, shutting down trial sites and encouraging participants to stay home, the decentralised models of research accelerated. As it was nearly impossible for patients to actually visit hospitals and clinics for in-person visits, clinical development had to adapt – and quickly – to a new reality. Being an extremely controlled industry, the life sciences industry is notoriously slow to adopt technology. When it became a necessity, organisations came together to ensure that trials could continue remotely. Now, the industry is realising that the way forward is through decentralised and hybrid trials.

Elements that were never even on our radar a decade ago have become almost commonplace. For example, wearables have gone from a fringe technology to a "must-have" approach for many trials that depend on real-time telemetrics. When you mix in things like validated instruments, genomics, and sensors, it's clear that the current technology infrastructure and manual methods employed by many organisations to review these data is not advanced enough to manage the influx of new data sources. Today we are talking in terms of things like zettabytes and petabytes that weren't even in the lexicon a decade ago. It's not a snowstorm: it's an avalanche, and the tools to handle this data need to evolve.

The big question, of course, is whether or not we will return to the old model after the pandemic crisis is declared over. No one knows the answer, although the #NoGoingBack movement certainly has

its adherents. In fact, they correctly point out that no fewer than five vaccines for the novel coronavirus were developed in less than a year using an almost entirely decentralised approach.

Whatever the methodologies and tools are going to be moving forward, what is exceedingly clear is that the real problem entails data and analytics tools that have simply not kept pace with changing data resources. Whether trials are held in one location, multiple facilities, or virtually, the method of information collection and analysis does not work as well as it should. This is no reflection on the quality of clinical researchers, many of whom I've seen do heroic work over the last 15 months. Rather it is an indictment of an approach to technology that has changed very little despite the industry revealing its inadequacies.

If existing models aren't working, what is the right approach? It comes back to technology that can manage multiple data sources, deliver high-value analytics, and help clinical organisations implement artificial intelligence tools with predictive capabilities. In fact, Gartner identified the ability to support audit trails and machine learning as top imperatives for CROs. At the crux of this conversation is how to improve interoperability between data sources and technologies.

An Ecosystem for Clinical Data

There's no single answer adequate to address the incredibly complicated problem of data chaos, but there is a relatively straightforward first step to help clinical trial organisations reduce work cycles and improve data management: create a modern data infrastructure that is capable of dealing with today's large, complicated sets of information. The right infrastructure needs the ability to leverage artificial intelligence and machine learning to automate as much of the collection and analysis as possible, which in turn will reduce the time and cost associated with transforming clinical trial data into useful analytics.

As decentralised and hybrid trials open the door for improved recruitment and retention, clinical trials are able to easily reach diverse patient populations, as they no longer rely on patient proximity to trial sites. Diversity in patient populations lead to more thorough data sets, and in turn, drugs and therapies that work for everyone – which is why many pharmaceutical companies are building the promise of diversity into their clinical trials. With standardised data sets, companies can monitor the race and gender of patients and ensure that diversity standards are being met.

To truly achieve modernisation in clinical trials, two key factors are required. The first is building strong technology foundations that can handle the volume of data created in every trial. The second is a focus on automation to modernise the management of different components of the clinical trial data lifecycle.

An automated clinical data pipeline is the solution to many problems that the life sciences industry faces. A robust data pipeline starts with automating the ingestion of information while easily cleaning and standardising various formats for efficient downstream processing. The goal is to achieve scalability while also eliminating manual intervention.

The next key aspect is transforming and standardising the raw data for analyses with simplified workflows and self-service capabilities without the need for additional programming.

Clinical organisations also require tools to curate and standardise data with built-in governance, and they need to be able to publish data for near-real-time consumption with governed access to all stakeholders. To dig a bit further into this, the key elements of a good data management strategy are:

- The ability to automatically ingest and standardise data from any source in any format. This becomes more efficient if you have out-of-the-box connectors to different clinical systems and metadata repositories via APIs. Ingestion capabilities should conform to data-enhancing standards (ODM) and include a mechanism to check on data structures' compliance automatically while integrating with an MDR.
- The capacity to consolidate and organise raw data along with the metadata in flexible data stores to enable easier transformation and consumption.
- The ability to apply validation rules against the metadata to curate and enforce standardisation. Ideally, a data pipeline should also offer mature transformation and mapping capabilities to standardise the raw data for submission and publishing to data marts for end user consumption.
- Have interoperability capabilities for connecting to data-driven applications and analytics. Of course, the entire pipeline should support blinding/unblinding and audit trails — again in the context of data integrity.

Clinical trials have gotten infinitely more complex over the last five years, and they are going to get even more complicated over time, not less. A modern data infrastructure needs to enable data integrity and provide capabilities to maximise the benefits that come from data standardisation. Without technology and data infrastructure modernisation, companies will continue to struggle to take control of the data and extract its value. One thing we should all be able to agree on is that data is the currency of the life sciences industry; companies that lag with their investments into technology resources will not get the insights they need to drive the results they want to see.

The Future of Data Chaos

The clinical data deluge requires a modernised approach to clinical trials. This can be achieved by implementing a data and technology-focused strategy, with modernised data management as the foundation. An end-to-end platform-centric approach – versus point solutions – will expedite data integrity and usefulness. But none of this is possible without investments into tools and technologies that are interoperable, standards-based, and can fit well into the existing infrastructure. The ability to evolve will be crucial as the industry continues to shift research priorities based on current trends and future needs.

Sheila Rocchio

Sheila Rocchio leads global marketing for eClinical Solutions helping to inform clinical researchers on how cloud based technology and analytics can improve and accelerate their digital initiatives. She has twenty years of experience in product and marketing roles in software and services companies that help digitize the clinical development process.

Email: srocchio@eclinicalsol.com