SIGNANTHEALTH
THE EVIDENCE GENERATION COMPANY™

WHITEPAPER

CLINICAL DATA AGGREGATION

BILL BYROM, PHD
Vice President, Product Intelligence & Positioning,
Signant Health UK

BACKGROUND

We are collecting more sources of data in our clinical trials than ever before. There are several factors influencing the rise in both variety and volume of data collected.

First, aside from breakthrough therapies, new drug treatments face fierce competition, and gaining more information and greater insights in clinical drug development is important to better distinguish and differentiate new treatments, not to mention enable faster, more accurate decision making. All of this can be achieved by measuring more and using more sources of data.

We also see a rise in decentralization of clinical trials, enabling more study components to be conducted locally or at home to simplify study participation. With this new study model comes an increasing number of new sources for the data needed.

Finally, the continued miniaturization of sensors and circuitry provides innovative methods to measure aspects of health not previously possible using new and novel data sources.

This drive to collect greater volumes and variety of data is alongside our drive to speed and reduce the cost of drug development – estimated to average $2.6 billion for each newly approved drug[1].

INCREASING DATA VOLUME AND VARIETY

Currently, sponsors and CROs leverage various disparate data sources in each clinical trial.

It is estimated that 50% of studies collect data using up to 5 disparate data sources, with 37% leveraging 6 to 10 different data sources, and an additional 13% of studies using 11 or more data sources[2].

The number of data sources is projected to continue to grow. In 2018, 97% of sponsor companies expected to use more clinical data from a wider variety of sources over the next 3 years[3], and this trend continues today.

While electronic data collection (EDC) has traditionally collected and managed the data, the volume of data collected outside EDC has already eclipsed the amount collected in eCRFs and continues to grow[4].

50% of studies collect data using up to 5 disparate data sources, 37% of studies use 6 to 10 different data sources, 13% of studies use 11 or more data sources.

Reconciliation of the multiple data sources can create a lot of activity late in the process.

THE PAIN POINT

Rising volumes of external data sources bring challenges to those involved in study monitoring, data review, and data management. Data from clinical trials comes from many different systems stored in many different source formats including EDC systems, clinical trial management systems (CTMS), eCOA, randomization and trial supply management (RTSM) systems, clinical laboratories, medical imaging, safety systems, sensors and wearables, as well as unstructured data sources. Unstructured data can come from electronic health records and social media for example.

Completing a clinical trial is not as simple as data-locking the EDC system. Reconciliation of the multiple data sources can create a lot of activity late in the process.

In addition, ongoing holistic medical review and monitoring of the clinical trial data is made increasingly complex and inefficient with disconnected data sources that require access and inspection through different solutions.

To enable actionable insights from data to drive effective medical review, risk-based monitoring and better decision making, study sponsors and CROs are spending increasing amounts of time and resource aggregating data and integrating disparate data sets. Without effective and efficient tools, this process can be error prone, time consuming, lead to delays in surfacing data for timely decision making, and extend data management cycle times.

THE GOAL OF DATA AGGREGATION SOLUTIONS

By harnessing the power of data, those running clinical trials can improve aspects of trial management, including risk management, early signal detection, medical oversight, and quality management, in addition to reducing data management cycle times from last patient visit to analysis-ready data.

The goal of a data aggregation solution is to continuously and automatically combine data from all the disparate sources in real time, so that the integrated data can be leveraged to drive actionable data-driven insights throughout the study and speed data management processes to reach analysis-ready data faster.

THE IMPORTANCE OF FLEXIBLE DATA ARCHITECTURE

Previously, sponsors and CROs used data warehouses to address the challenge of clinical data aggregation. Data warehouses use relational databases and pre-defined, hard-coded, data schemas to enable the consolidation of data from multiple sources. These schemas take careful planning and time-consuming construction, because the schema must be defined upfront and designed to enable the data to be reported for pre-defined use cases.

Data warehouses are constrained in the type of data they accept though. Data properties are defined before data is ingested, and only data that meets those standards can be added to the system.

One problem with this approach is when data sources change or are added (e.g., a specific system is replaced, a software update affecting data formats is released by an eClinical vendor, or fields in an eCRF are added or changed), the relational database schema must be updated and the clinical data warehouse must be re-validated.

This inflexibility makes it difficult to keep up with the growing volumes and variety of data we collect. A fundamental novel approach is needed.

Some of the world's most innovative companies have instead turned to data lake-based systems. A data lake-based system ingests data as is, in its raw format – whether structured, semi-structured, or unstructured. This big data architecture easily accommodates large volumes, variety, and velocity of raw data in their original forms as they're received from the source in their entirety. These systems can be constructed rapidly because there is no requirement to define the data formats and schema upfront.

When data are accessed from a data lake, for example when visualizing data for the purposes of medical data review, the schema required by the target solution is defined "on-read". This is where the power of the data lake architecture is realized. The data lake is not only able to ingest all data formats without significant up-front investment in data schema definition (a schema-on-write approach for a data warehouse), but once the data are ingested this architecture can flex to the variety of use cases for the aggregated data throughout the clinical trial by transforming data to the target schema required by each target system on an as needed basis (schema-on-read).

THE VALUE OF AGGREGATED DATA

Ingesting and aggregating data in real time enables sponsors and CROs to harness the value of the data throughout each clinical trial or program of studies, as well as to leverage tools more effectively to reduce data management cycle times.

DATA VISUALIZATION

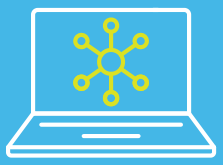
Real-time visualization of the aggregated data provides a 360-degree view of all patient and trial data at every step throughout the clinical trial. This enables sponsors and CROs to gain important, actionable insights into trial data and performance, in a timely manner, to better monitor and manage their studies.

Medical monitoring can be powered by real-time data aggregation, while solutions enable visualization of the combined data sources to provide a holistic view of individual patient data, overall data trends, and signals. Medical monitoring solutions, such as SmartSignals™ Study Oversight, also contain functionality to manage and record the workflow of the medical monitor using checklists, dashboards, and reports.

Risk-based monitoring (RBM) can leverage the aggregated clinical and trial performance data to power effective risk-based monitoring. SmartSignals Study Oversight contains comprehensive RBM capabilities, including selecting and defining quality indicators, visualizing these in easy-to-use study dashboards, and using these insights to direct and drive targeted monitoring activity.

DOWNSTREAM DATA MANAGEMENT

Real-time aggregated data in a single place provides the opportunity to reimagine data management. SmartSignals Data Workbench speeds data management activities by using machine learning to auto-generate data checks, surface discrepancies and reconciliation issues, and automate data cleaning and validation. Its interfaces enable effective curation and management of the data cleaning processes and speed data management cycle times, bringing cleaner data, faster.



SMARTSIGNALS CLINICAL DATA HUB

The good news is that the technology needed to facilitate a better, more holistic approach to data aggregation exists and is already in use. Our SmartSignals Clinical Data Hub brings standardized data together, in real time, in one place.

Using our cloud-based SaaS platform, we ingest data from any source format to enable sponsors and CROs to cleanse, map, transform, and aggregate their data in near real time throughout each study. Unstructured data is classified using natural language processing.

The inherent data lake architecture enables rapid setup and provides easy access to information in all kinds of formats for data scientists to understand and process data for multiple use cases.

The built-in advanced mapping and transformation engine uses smart mapping, incorporating machine learning algorithms for rapid data transformation and standardization.

The toolset within our Clinical Data Hub gives users the freedom to focus on high-value tasks to enhance data quality and operational efficiency: Teams can concentrate on analyzing clinical and operational data in real time, monitor risks, visualize outliers and trends, all while quickly automating data management activities to reduce cycle times.

REFERENCES

01 Tufts Center for the Study of Drug Development (2014). https://csdd.tufts.edu/cost-study

02 Pharma Intelligence / Oracle, 2018. Challenges And Opportunities In Clinical Data Management: Research Report. https://www.oracle.com/a/ocom/docs/dc/oracle-clinical-data-report-1809-final-26-sept.pdf?elqTrackId=a3c3795787d24ddb905a0872489fcbd8&elqaid=75274&elqat=2

03 Tufts Centre for the Study of Drug Development 2018, Examining Causes of and Potential Solutions to Clinical Data Management Cycle Time Challenges.

04 SCDM 2019, The Evolution of Clinical Data Management to Clinical Data Science: A Reflection Paper on the impact of the Clinical Research industry trends on Clinical Data Management.

05 Accenture 2019, The Future of Clinical Trials. https://www.accenture.com/gb-en/insights/life-sciences/future-clinical-trials

WHO IS SIGNANT HEALTH?

Signant Health is the evidence generation company. We are focused on leveraging software, deep therapeutic and scientific knowledge, and operational expertise to consistently capture, aggregate, and reveal quality evidence for clinical studies across traditional, virtual, and hybrid trial models. For more than 20 years, over 400 sponsors and CROs of all sizes – including all Top 20 pharma – have trusted Signant solutions for remote and site-based eCOA, eConsent, RTSM, supply chain management, and data quality analytics. Learn more at www.signanthealth.com.

SIGNANT HEALTH
THE EVIDENCE GENERATION COMPANY™