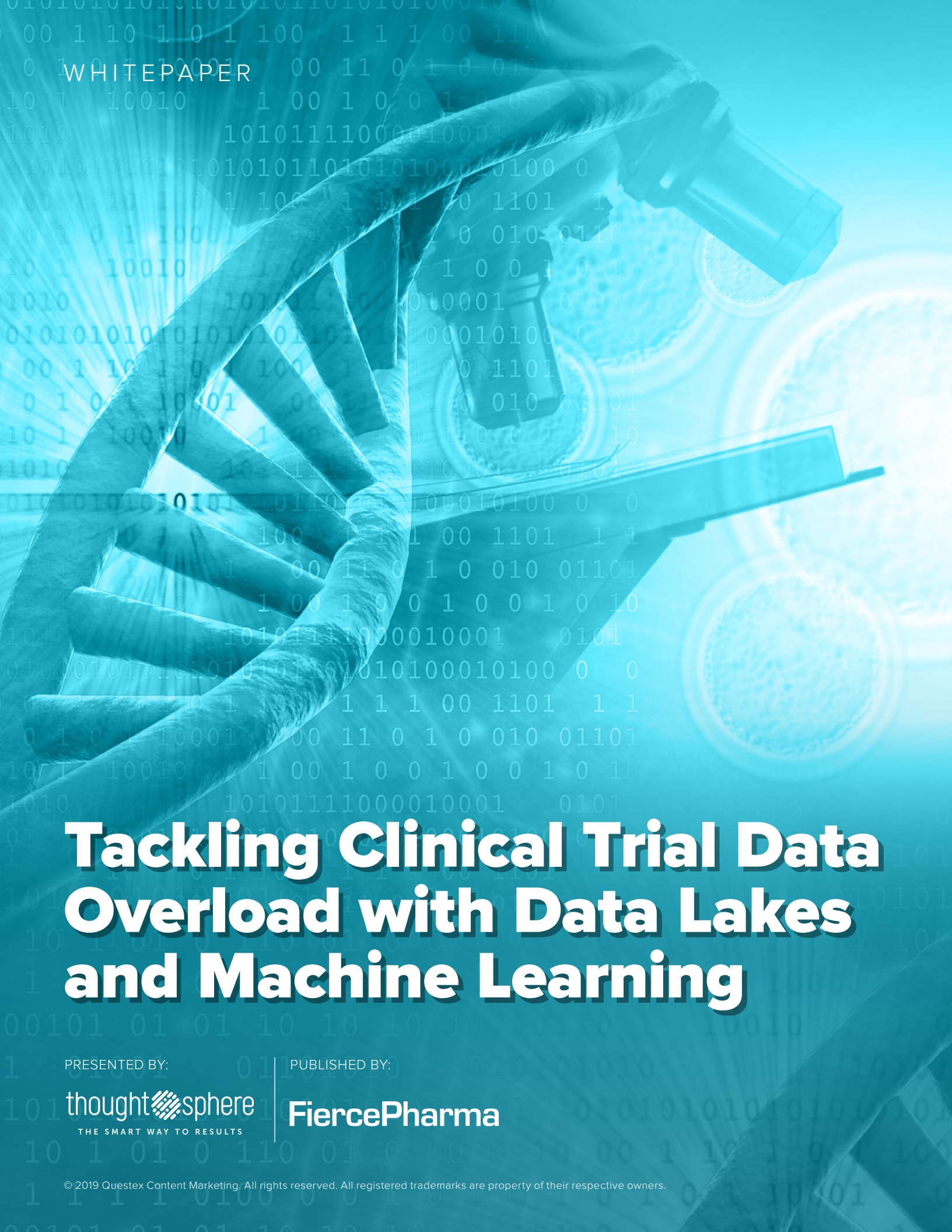

WHITEPAPER

Tackling Clinical Trial Data Overload with Data Lakes and Machine Learning

PRESENTED BY:

thoughtsphere
THE SMART WAY TO RESULTS

PUBLISHED BY:

FiercePharma

CONTENTS

THE STATE OF CLINICAL TRIAL DATA TODAY

Data is central to all aspects of clinical research. Sponsors run trials to generate safety and efficacy data. In doing so, they gather a wealth of data that hides insights into study performance. Viewed in light of today's data-rich environment, those facts suggest clinical research may be on the cusp of a golden age in which a multitude of technologies feed data into advanced analytics. Such a system would give study teams in-depth, real-time oversight of metrics, making trials safer, faster and more efficient.

That vision is underpinned by significant increases in the volume and variety of data generated by clinical trials. Today, study teams can use wearables to monitor subjects 24/7, track shipments of study drugs across the supply chain and get real-time updates on study metrics such as recruitment and screen failures. For most of the history of clinical research, none of those types of study oversight were possible.

The proliferation of data is occurring alongside a related increase in analytic capabilities. Study teams can apply advanced analytics to study data to both understand what worked in earlier trials and ensure active studies are proceeding to plan. The emergence of artificial intelligence (AI) and machine learning (ML) as part of the toolkit available to study teams is creating further opportunities, pointing to a future in which automated systems rapidly analyze large datasets to extract benefits, value and insights.

That is the potential of big data and advanced analytics. However, that potential is largely unrealized today. While sponsors, CROs and vendors all see scope for data to reshape the design and execution of clinical trials, the reality on the ground is very different. Rather than benefiting study teams, the rapid increase in the volume and variety of data is actually creating challenges.

WHY DATA CREATES CHALLENGES

Some of the challenges facing study teams stem from the explosion in the availability of data. Clinical trials run in 2015 had almost 90% more endpoints than those conducted a decade earlier.[1] Another study found 50% of clinical data managers handle six or more data sources in a typical study.[2]

The additional data collected in modern clinical trials can reveal new details about the effects of a drug and yield insights into the operational performance of studies. These benefits can only be realized if the data is ingested, aggregated and standardized in near real time. Yet, many trial sites and study teams are forced to tackle their roles in these key, data-intensive tasks using technologies from another era.

A 2017 survey found 41% of trial sites use paper charts to capture patient medical information.[3] The figure increases to 56% among sites dedicated to clinical research. More than half of CROs still use paper case report forms.[4]

Even fully digital trials face significant data challenges. A typical study features an alphabet soup of data systems including EDC, eTMF, IRT and QMS supplied by different vendors that use different data models. Study teams may lack a way to view all the data generated by a trial, as it is generated, forcing them to navigate multiple systems to establish even an outdated, partial overview of study metrics.

There is little public research into the impact of these limitations in the handling of study data. However, researchers have quantified the effect of the explosion in patient data. From 2007 to 2017, time from last patient, last visit to database lock increased by three days.[4] The Tufts Center for the Study of Drug Development, which performed the analysis, said the rise was “due in large part to the rapid growth in eClinical data volume and diversity of data captured.”

Organizations that develop the ability to manage rapid data growth will gain a competitive advantage. Study teams could be analyzing data in real time to improve the operational performance of their clinical trials. Instead, they are drowning in a torrent of data that creates as many problems as it solves. The era of taking a hands-on, convoluted approach to ingesting, aggregating and standardizing data must end.

56%

of dedicated clinical research sites use paper charts to capture patient medical information

HOW DATA LAKES CLEAR ROADBLOCKS

As it stands, storing all the data generated by clinical trials in one place in a way that is both cost-efficient and productive is a challenge. Data warehouses have supported clinical research successfully for years, but their limitations are being exposed by the variety of data generated in studies today and the velocity at which it must be processed.

The most common enterprise data warehouse structure was established to support preprocessed and structured data. Clinical trials, in contrast, generate data from wide diversity of sources such as EDC and EMR systems, central laboratories, patient-reported outcomes and wearables. That means working with data that is structured in a variety of ways. To further complicate the situation, no one clinical trial is the same as the next, preventing companies from applying the same template to all studies.

Study teams need to be able to access and analyze the diverse data generated in their trials in real time. That task is beyond the original remit of data warehouses but is right in the wheelhouse of data lakes.

The agile architecture of a data lake is designed to hold various types of raw data in their original form as they are received from the source. This allows for an ecosystem with little to no change to the original data format and removes the need for content enrichment or integration that takes up valuable space. With schema-on-read access to information in all kinds of formats, data lakes allow data scientists to understand and process data in innovative ways.

This fluid data structure is integral to scientists who need to rapidly analyze and interpret information as it comes in, without sacrificing valuable time and resources for organization and conversion. This also means that there is no hierarchy to the way that the data lake stores and prioritizes its contents: the data is held in an unstructured way, making it easier to access and analyze when it is ready to be used.

A few key data lake advantages include:

- Raw, real-time data storage.
- Accelerated time-to-value data delivery.
- Low-cost storage for high volumes of data.
- The ability to remain agile and reconfigurable.
- Improved search and knowledge of content.

Data lake storage solutions also provide scalability for organizations. The threshold of a cloud-based data lake scales up or down based on load. All the different components and software used in big data are deployable in a way that eliminates the need to search or integrate with other solutions.

Study teams need to be able to access and analyze the diverse data generated in their trials in real time.

DATA LAKES IN ACTION

The key differentiator of data lakes is their ability to quickly ingest, aggregate and standardize diverse sets of data with minimal manual effort on the part of the user. That differentiator opens up a variety of use cases by giving study teams real-time access to a centralized repository of all their data.

Equipped with such a repository, study teams can make informed decisions about their clinical trials. A team can, for example, define and monitor KRIs—key risk indicators—to ensure that their trials are going to plan. As the data that informs the KRIs is available in real time, study teams can take corrective actions as a potential problem arises.

The same data lake can enable study teams to identify sites that are enrolling slowly or experiencing a high rate of screen failures, thereby generating insights to inform targeted support and training. Again, this is only possible if the study team has ready access to timely, comprehensive study metrics.

Information on safety parameters kept in data lakes provides study teams with another important source of actionable insights. If, for example, the study team detects a worrying trend in vital sign data, it can try to find out whether it is a genuine safety signal, or an anomaly caused by site staff errors. Either way, the team needs to know about the problem as soon as possible. Data lakes facilitate such timely insights.

A SMARTER SET OF DATA

All data lakes share the core concept of storing raw data. Yet, not all data lakes are created equal. The speed at which data is ingested, aggregated and standardized, plus the manual effort those tasks involve, varies from data lake to date lake.

Market-leading data lakes use AI/ML to streamline the ingestion and standardization process. This often entails building an ontology of how previous clinical trials were mapped. That done, an advanced AI/ML algorithm can automatically map data as it flows into the data lake, cutting the time it takes to get the data into an analyzable form.

The model even works in the absence of an ontology. In that situation, the AI/ML identifies certain words or expressions to map the study and standardize the data without the existence of the ontology.

AI/ML plays a different role once the data is standardized. From the moment the data begins to flow into the lake, the AI/ML starts to identify potential problems, correlations and trends that analysts instantly access and begin to use and incorporate into analytics. For example, users can review the performance, safety and compliance metrics for a study using analytics and easily detect cross-site issues by comparing one site to the rest. Users can drill down into advanced analytics on adverse events, recruitment rates, protocol deviations and data cleanliness.

A smart data lake monitors movement every step of the way, as specific statistical algorithms contextualize and define certain triggers based on system behaviors. Using these powerful tools, AI/ML-equipped data lakes work the data straight into a study data tabulation model, which becomes an imperative tool for regulators to understand and analyze long after research is complete.

The clinical trial space often feels like a race against the clock. What truly gives industry leaders a tangible advantage is time-cutting tools that increase the speed at which they identify, test and deliver innovative solutions. Incorporating a data storage system that lets you set it, forget it and learns on its own is an important step toward making your organization faster, smarter and future proofed.

“The beauty of a modernized data lake architecture that incorporates AI/ML is that it’s not just restricted to what’s happening in the current clinical data landscape,” ThoughSphere CEO Sudeep Pattnaik said. “Having a data platform that scales to accommodate data in real time provides a connected view that cuts down future risk. Adopting AI and ML as a healthcare professional will be instrumental in years to come and those who do are the experts who will eventually influence industry protocol and become thought leaders in the space.”

> “Incorporating a data storage system that lets you set it, forget it and learns on its own is an important step toward making your organization faster, smarter and future proofed.”

LEVERAGING YOUR DATA LAKE

Equipped with a thorough understanding of the industry's pain points, ThoughtSphere set out to create a clinical data and analytics solution that helps healthcare leaders focus on what they know best: their research. The result is an innovative platform that leverages data science expertise and AI/ML to increase R&D efficiency by up to 30%.

ThoughtSphere's unique cloud-based solution accomplishes this by:

- Leveraging big data, data lake architecture (NoSQL) to ingest and aggregate both structured and unstructured data with relative ease.
- Being source system-agnostic and highly configurable, loading data from multiple sources and formats, including ODM, line listings, XML, JSON, SAS export and SAS data files.
- Using smart mapping for rapid transformation and standardization, all done via configuration with an evolving reusable mapping library based on machine learning algorithms.
- Learning data ontology over time using statistical and ML methods and thus identifying risks (also known as smart risks).

With ThoughtSphere Cloud, clients have cut costs by up to 30% through a reduction in manual data aggregation and have saved up to 50% of their time mapping highly complex studies. Those datapoints show the business value that the ThoughtSphere Cloud provides to sponsors and CROs.

While some technological advances offer little business value, ThoughtSphere has used its data science and clinical research expertise to strike at the heart of the biggest problems facing sponsors and CROs, thereby ensuring its platform makes a meaningful difference to their operations.

thoughtsphere
THE SMART WAY TO RESULTS

ThoughtSphere is a leading cloud-based clinical data hub and analytics company founded by clinical trial experts. Our mission is to help life science companies develop and deliver treatments to patients faster and smarter using data science. With our innovative platform these companies can reduce clinical development costs, optimize and enhance the effectiveness of clinical trial processes, and gain actionable insights.

1. Tufts Center for the Study of Drug Development. Rising Protocol Complexity Is Hindering Performance while Driving Up Cost of Clinical Trials, According to the Tufts Center for the Study of Drug Development. *GlobeNewswire News Room* (2018). Available at: https://www.globenewswire.com/news-release/2018/07/17/1538332/0/en/Rising-Protocol-Complexity-Is-Hindering-Performance-while-Driving-Up-Cost-of-Clinical-Trials-According-to-the-Tufts-Center-for-the-Study-of-Drug-Development.html. (Accessed: 25th September 2019)
2. Challenges And Opportunities In Clinical Data Management. Available at: https://www.oracle.com/a/ocom/docs/dc/oracle-clinical-data-report-1809-final-26-sept.pdf?elqTrackId=a3c3795787d24ddb905a0872489fcbd8&elqaid=75274&elqat=2. (Accessed: 25th September 2019)
3. | The need for—and barriers to—adopting eSource. Available at: https://www.centerwatch.com/news-online/2017/02/15/need-barriers-adopting-esource/. (Accessed: 25th September 2019)
4. Tufts Center for the Study of Drug Development. eClinical Data Volume and Diversity Pose Increasing Challenges and Delays, According to the Tufts Center for the Study of Drug Development. *GlobeNewswire News Room* (2018). Available at: https://www.globenewswire.com/news-release/2018/01/09/1286051/0/en/eClinical-Data-Volume-and-Diversity-Pose-Increasing-Challenges-and-Delays-According-to-the-Tufts-Center-for-the-Study-of-Drug-Development.html. (Accessed: 25th September 2019)