# Investing in Data Science to Unlock Clinical Data Value

Richard Young

The need for biopharma companies to equip data managers with the training and resources necessary to capitalize on new digital health tools.

DATA MANAGEMENT

When data sources are stored and managed separately, it is difficult to see patterns and reconcile differences. Each source creates and stores data in a different format, with a different schema. Mapping data across sources is labor intensive and introduces traceability challenges. Data managers use data science to ensure data from diverse sources are gathered, harmonized, and formatted for the research scientists to analyze. Organizations that want to analyze big data should equip data managers with the technology and training necessary to ensure the new data is usable.

## Good data science

Quality issues that slip through data management impact downstream users such as medical monitors, potentially the most expensive resource in our organizations. These are precious resources, and the financial impact of their decisions is substantial. Poor data quality increases the likelihood of inaccurate decisions, and both false positives and false negatives can cost a pharmaceutical company millions of dollars.[1]

Data management is largely manual today, but that is beginning to change. At the 2019 Society for Clinical Data Management (SCDM) Annual Conference, leading organizations discussed technologies to improve efficiency, such as machine learning (ML) to help automate mapping into the study data tabulation model (SDTM) format required by FDA and robotic process automation to help code incoming data. Equipping data managers with tools that assist in cleaning and harmonizing data is the most cost-effective means to supply downstream stakeholders with critical information faster.

Currently, data managers spend a significant portion of their day on relatively menial and manual tasks, with a heavy focus on managing electronic data capture (EDC) data. Technologies can streamline, automate, and even obviate many of those tasks and help data managers provide the organization with greater visibility and insights into the data.

Lotus Clinical Research, a CRO and research site network specializing in pain management, recently invested in training for one of its data managers and quickly reaped the benefit. "I highly recommend investing in a reporting power-user. We have a new EDC platform and our reporting guru created reports and dashboards that automated hours of work previously spent on manual tracking and reporting," said Jeanne Strain, vice president of data services, Lotus Clinical Research. "Sharing those reports with colleagues on other trials has saved time across the organization and gives sponsors faster visibility into their data."

There are numerous examples of how real-time visibility into patient data can help an organization run more effective and efficient clinical trials.

- Identifying compliant patients and decreasing loss-to-follow-up reduces the number of patients needed and speeds database lock.
- Enrollment dashboards showing site performance help surface what is or is not working earlier so other sites can course correct.
- Data visualizations of patient data such as blood pressure readings can surface outliers or potential adverse events.

Clean, trustworthy data is a prerequisite to extracting those insights. Getting clean data to downstream decision-makers faster results in higher-quality outcomes.

Artificial Intelligence (AI), ML, sensors, and wearables each hold significant promise to help design more effective trials, improve patient safety, and identify which patients benefit from a specific investigational product.

However, these technologies have thus far failed to live up to their hype. A major impediment to the success of any AI or ML project is the heterogeneity and discord in their data.

When Vas Narasimhan, CEO of Novartis, was asked about AI and ML, he spoke first about the challenge of getting clean data to work with. "The first thing we've learned is the importance of having outstanding data to actually base your machine learning on," said Narasimhan. "In our own shop, we've been working on a few big projects, and we've had to spend most of the time just cleaning the data sets before you can even run the algorithm. That's taken

us years just to clean the datasets. I think people underestimate how little clean data there is out there, and how hard it is to clean and link the data."[2]

The promise of AI will remain out of reach until organizations find a sustainable and repeatable means to manage and harmonize data from disparate sources. A critical component will be data managers applying the practice of data science to master the idiosyncrasies of biologic and real-world data. Clean data allows greater confidence in patterns and ability to predict outcomes. Once achieved, organizations can affordably and productively apply AI to produce the much-anticipated benefits.

## Technology to equip a data scientist

Case report forms (CRFs) in EDC systems, which for years have been the primary source and store of clinical data, can no longer be the center of a data manager's universe. Industry leaders at the SCDM 2018 Leadership Forum estimated that less than 30% of the data volume comes from the EDC. In most organizations, the EDC is the sole system designed to manage clinical data. According to a 2017 report by Tufts Center for the Study of Drug Development, most organizations (77%) reported difficulty loading external data sources into their EDC and 66% attributed those difficulties to EDC system limitations and integration issues.

Organizations ready to invest in their clinical data infrastructure should consider three main technologies:

**1) A clinical data platform.** Data managers will work more efficiently with a purpose-built platform to aggregate, clean, and normalize data. While a dedicated platform for managing all data sources was historically a luxury, they could now be considered core. Two of the most important steps for data management are selecting the appropriate data aggregation platform and learning to use the query and reporting capabilities provided. Capabilities that automate or eliminate manual data cleaning tasks will free the data manager's time, while advanced reporting equips them to deliver high-value insights.

**2) Scripting and visualization.** Visualization tools help you look at data and extract meaning from what the data is showing. Usability is key to enabling data managers to move quickly and explore different aspects of a problem. Visualizations help you explore different ideas and formulate the right questions to ask.

**3) Augmented intelligence.** Augmented intelligence applications such as ML, natural language processing, and robotic process automation are valuable tools for a data scientist and should be considered in partnership with the data management team, not as a replacement of the team. Data scientists play an important role in training the learning algorithms and are the natural partners enabling these systems to deliver value.

## Skills development for data science

Pragmatically, what does this mean for data managers today? The following are four specific recommendations for how data managers can educate and position themselves to help their organization leverage today's data, and enable data managers to contribute as trusted advisors.

- Researching the specific ALCOA (attributable, legible, contemporaneous, original, accurate) risks and limitations associated with each real-world data (RWD) source. For example, patient registries provide extensive observational data on patients that is of good quality and relatively inexpensive. However, confounding (i.e., the inability to isolate the impact of relationships between dependent and independent variables) is a significant problem and makes it difficult to attribute trends in registry data to particular therapies.
- Familiarizing oneself with FDA guidance on using electronic health record (EHR) data in clinical investigations, which includes recommended practices for common situations such as handling data modifications. EHR-EDC integrations reduce the need for data entry by sites and associated source data verification (SDV), but numerous complications exist. Data managers should advise the organization on integration decisions that help preserve data lineage, protect personal identifiable data, and preserve identity masking.
- Advising study teams on data source selections. Selecting the appropriate data source requires making informed decisions and balancing trade-offs —i.e., which sources contain the desired information with the greatest reliability, fewest gaps, and in high enough volume to balance across treatment arms? Educational courses from SCDM, such as mastering mobile and digital technologies, provide a framework for evaluating new sources in clinical trials.
- Advising organizations on data management technology selections. Many established systems were architected before the recent expansion in data sources. Data managers should research the capabilities of modern tools and platforms to identify those designed with diverse, heterogeneous data sources in mind. Industry conferences such as the DIA and SCDM annual meetings are most valuable during times of change. For anyone dealing with requests for new data sources, it is worth attending an industry conference to speak firsthand with the early adopters and supporting-technology providers.

## Conclusion

Challenges with data quality were the third most-cited barrier to completing clinical trials in 2018[3] and data management is becoming even more difficult with each new source. And yet, innovation in the tools and training to support data managers has been stagnant. To truly capitalize on AI and RWD, biopharmaceutical companies must invest in the data science skills and resources of their data management teams.

***Richard Young,*** *Vice President of Strategy, Veeva Vault CDMS; email: richard.young@veeva.com*

## References

1. The Burden of the "False-Negatives" in Clinical Development: Analyses of Current and Alternative Scenarios and Corrective Measures. *Clinical and Translational Science*. July 2017.
2. Novartis CEO Who Wanted To Bring Tech Into Pharma Now Explains Why It's So Hard. *Forbes*. Published online Jan. 16, 2019.
3. Global Top Health Industry Issues: Defining the Healthcare of the Future. PwC Health Research Institute. 2019.