Addressing Missing Data in Clinical Trials – The Data Science Approach

Digital Health Technologies (DHTs) have revolutionized clinical trial data collection, while also promising to make research more efficient and more patient centric. However, shifting the power to input data from clinicians to participants increases the risk of missed datapoints. This can compromise the ability to draw inferences or lead to incomplete submissions, threatening the success of otherwise highly promising clinical trials.

Digital Data Sources and Missed Datapoints

Recent years have seen adoption of a wide range of digital health technologies in clinical trials, from wearables and sensors to electronic patient reported outcomes (ePROs) and diaries. The potential benefits of this shift are well documented and include expanded patient accessibility and inclusivity to increased real-world data.

However, there are also potential pitfalls. It moves control of data generation and entry from highly trained staff to clinical trial participants who may not have the same precision focus or pay the same attention to detail as their professional counterparts. This is compounded by the fact that decentralized trials (DCTs), where the frequency of data collection may be hourly or even daily, vastly increase the number of data points being collected. In fact, phase III clinical trials currently generate an average of 3.6 million data points – three times the data collected by late-stage trials 10 years ago.[1]

The ability to capture data remotely without supervision can all add up to missing data or values within records or time series. A complete sample size and variability per study protocol, with complete datapoints, is required for efficient analysis. Missed data points can lead to diminished statistical power and affect analysis, thus negatively impacting a sponsor's ability to demonstrate product efficacy. If participants with missing values are omitted from analysis, misleading results might be obtained regarding the effect of treatment, unreliable P values may be obtained, and assessments of the importance of prognostic factors may be inaccurate.[2]

It follows, then, that it is of utmost importance to have complete data as much as possible.

Power of Data Science

Data science, which combines the power of statistics, advanced analytics, and artificial intelligence (AI) techniques such as machine learning (ML) to uncover actionable insights in large datasets, is at the forefront of many innovations in clinical research.

It is being used in the collection, management, and analysis of clinical data, automating the processes and reducing error rates. This is important because securing the overall quality of clinical data is paramount to ensuring quality care and appropriate decision-making in the medical and healthcare fields. Importantly, it also offers possible solutions to the missing data problem.

Risk-based Monitoring through Real-time Metrics

The first option is risk-based monitoring, a form of centralized monitoring in which sponsors can review the study-wide data in near real time as it accumulates, allowing early identification of pre-defined risks. Data science tools build visualizations and provide an oversight of the trial data. These data driven insights optimize subject and site monitoring by enabling early notice of trends such as missing data, outliers, and unusual behaviors. Monitors are then able to quickly spot missing datapoints and untimely data entry and therefore take timely action.

The approach also allows data scientists to compare the data being collected by individual sites, as it accumulates, to identify outliers. With each site following the same protocol, all datasets should be roughly equivalent. As such, those logging fewer changes than their counterparts could be at risk of missing datapoints. Using analytics and visualization, researchers can quickly spot such outliers then use additional visualizations, such as form-by-form comparisons, to investigate the root of the anomaly. They can then take any necessary corrective action, such as providing additional site training, before the issue is able to impact on data quality.

Other Solutions

This is not the only way data science tools are being used to address the missing data issue in clinical trials.

ML algorithms, for example, can recognize patterns from previous trials, and use that information to highlight areas in which missing data may occur in the future. Likewise, predictive analysis can be used to analyze results from historical studies, and forecast potential future issues. It can also lead to the formation of new hypotheses to be validated by trial data. In addition, predictive modelling can be used to identify patients who will respond favorably to treatments, based on their clinical history, thus reducing occurrences of missing data due to patient drop out.

Big data analytics is particularly suited to DHT-driven trials, which generate a high volume, velocity and variety of data. The technology can collect these large amounts of unstructured, real-world data, organize, analyze, and visualize the results, to explore them for unexpected patterns. This can result in more accurate missing data predictions and insights than traditional data management solutions.

Another potential solution to missing data is data augmentation, or artificially increasing the amount of data by generating new data points from the existing data. In simple terms, it extends the data and generates more records, thereby making up for the missing information. However, this approach should be used with caution in clinical research. As it extrapolates from existing data, it therefore runs the risk of inducing bias.

Some approaches are best suited to particular types of missing data. Estimation equations, which replace missing data with averages from across the data set, for example, can be useful when dealing with a minimal number of empty fields.

Conclusion

Missing data can undermine the scientific credibility of conclusions and threaten the success of drug development efforts. When it comes to solutions, prevention is better than cure. Along with the usual validation procedures of ensuring accuracy and completeness of data, efforts should be targeted to minimizing missing data during the design, planning, conduct, and analytic stages. The regular monitoring of real-time metrics throughout the duration of the study, is a highly effective preventive measure.

In circumstances where missing data does occur, possible corrective actions include use of estimation equations and data augmentation, though both have their limitations. As data science continues to develop, the industry expects these solutions to evolve, helping researchers to produce the high-quality clinical evidence needed to make reliable clinical trial inferences, and give studies the very best chance of success.

REFERENCES

1. https://www.globenewswire.com/news-release/2021/01/12/2157143/0/en/Rising-Protocol-Design-Complexity-Is-Driving-Rapid-Growth-in-Clinical-Trial-Data-Volume-According-to-Tufts-Center-for-the-Study-of-Drug-Development.html
2. Ibrahim JG, Chu H, Chen MH. Missing data in clinical studies: issues and methods. J Clin Oncol. 2012 Sep 10;30(26):3297-303. doi: 10.1200/JCO.2011.38.7589. Epub 2012 May 29. PMID: 22649133; PMCID: PMC3948388. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3948388/#B5

Pamela Adede

Data Operations Programmer at Phastar where her core work entails programming databases for data capture, validation, and extraction for clinical trials. Pamela holds a BSc degree in Computer Science from Kabarak University in Kenya.