Big Data Applications for Clinical Trials

Jacob Tibenkana
Indiana University
107 S. Indiana Avenue
Bloomington, Indiana 47405-7000
jtibenka@iu.edu

ABSTRACT

This paper will provide a brief overview of how Pharmaceutical companies use big data applications to track and analyze data for clinical trials. This paper will also explore the relationship between data quality and the integrity of the clinical studies.

1 INTRODUCTION

Today, many trails are conducted across the world in many fields such as oncology, rare diseases, central nervous diseases, etc. These trails aim to further advance treatments for diseases in the these fields. Clinical trials have a long history. "The world's first clinical trial is recorded in the Book of Daniel in the Bible. During his rule in Babylon, Nebuchadnezzar ordered his people to eat only meat and drink only wine, but several young men of royal blood, who preferred to eat vegetables, objected. The king allowed these rebels to follow a diet of legumes and water fi!? but only for 10 days. When Nebuchadnezzar's experiment ended, the vegetarians appeared better nourished than the meat-eaters." [1]. Now this maybe a good example for a simple straight forward trial, however today's trails are more complicated and can last longer than 10 days, usually a period of 12 months to 60 months. This is only the tip of the iceberg, because they are all sorts of criteria (inclusive and exclusive) that have to be adhered to for subject selection. The study design and documentation have to be approved by the FDA and IRB, contracts between sites, CROs, sponsors, vendors, and licensors that have to be executed in order for the trial to proceed. Subject recruitment has to occur and raters have to be trained to conduct assessments with the subjects on a specified study schedule, etc. As one can see there are many factors to consider and much work that goes into running and completing a clinical trial today. Thankfully, with the advancement in technology, these factors are becoming easier to manage, because of data management, from a perspective of data tracking and analysis using, TMF, EDC, and other vendor databases/services.

2 TMF, CTMS, EDC, AND IN-HOUSE VENDOR SYSTEMS

If an entity (person, company, organization, etc.) plans to conduct a clinical trial, how do they go about addressing the factors mentioned above. Let's exclude the labor part such as the day to day tasks, and deliverables of the clinical study, but think of each factor above as various information that can be collected, tracked, and most importantly stored. This is where the use of big Data applications and analysis comes into the play. According to the National Institute for Health Research (NHS), a trial master file (TMF) should be made at the initial phase of the trial. "TMF is the collection of documentation that allows the conduct of the clinical trial, the integrity of the trial data and the compliance of the trial with GCP to be evaluated. It is also essential to allow the trial to be effectively managed by the sponsor as it allows the appropriate individuals access to the necessary trial documentation." (Pearce, Apr 18, 2016) There is a large amount of documentation that must be collected, stored (for audit purposes), and submitted to the government for approval for a trial to happen. Thanks to the big data applications field, electronic TMFs (eTMF) have been created to host all these documents for the sponsor and CROs. This information can be accessed and retrieved for various purposes depending on which phase the study is in (start up, mid, data lock, and close out). Some of these eTMFs are through MasterControl, "a document control solution provider" (MasterControl, 2000). Other eTMs are in-house platforms provided by vendor companies for the trials. Some of these vendors companies include Phlex Global, SureClinical, etc. Another area that falls into the big data applications aspect for clinical trials is the software Clinical Trial Management System (CTMS). According to an article called CTMS: What You Should Know, sponsors use CTMS to exceed organizational limits, preparation, reporting, timelines, be diligent, manage distribution of drug supplies, tracking, and offer data to business intelligence systems that are used as digital dashboards for clinical trial managers. (Dagalur, Mar 16, 2016) Sponsors can have their own CTMS or typically they will contract other companies (for example, Simpletrials, Allegro, Parexel, etc.) with their own in-house CTMS to manage this aspect of the trial. Before computers, one would assume that subject data from trails was recorded on paper, however today, databases play a huge role. For example, the electronic data capture. According to Kristina Lopienski artilce, The Beginnerfis Guide to an Electronic Data Capture (EDC) System, the EDC is a database that stores subject data collected in a clinical study; more and more studies are using EDC software and replacing traditional paper records. The article goes on to say that different EDC systems are used by sponsors, CROs, sites, and vendors to manage data in the aspects of tracking, quick access, security, and organization, and compliance. (Lopienski,February 2, 2016) EDCs don't only serve as a storage system, but they can be important for individuals who work with the clinical data stored in these systems such as clinicians, data managers, and data analysts. For subject selection and validation, clincians review subject data that is electronically captured from screen visits to determine if a subject is eligible to be randomized into the study for the clinical trial or if the subject should not be included in the clinical trial. This is important, because if we are conducting a trial for a drug to treat depression, we want subjects with indications of depression or those who have been diagnosed

with depression in the past, but not subjects who do not fall in this criteria. The screen visit will assess for these parameters and the data collected, which is electronically captured can be reviewed by a clinician; in order to determine if the scores and history of the subject meet the inclusive criteria into the study or if the subject falls into the exclusive criteria and should not be included in the study. In the long term, this protects the integrity of the study, because if you have the right subjects, one should expect the subject data to be approximately valid. Approximately is used here, because there other outside factors that will influence subject scores, but the goal is to help produce quality data by selecting the right subjects. Because electronic data capture systems store subject data for each visit throughout study, granted the subject does not drop out, sponsors, CROs, and vendors can use this as a tracking system for progress towards the result they wish to accomplish (end point reliability). Working together, clinicians, data managers, and data analysts can access and extract data from EDCs in form of CSV files or excel sheets to review and analyze the data. Here Data analysts and maybe Data Scientist can use R or Python to investigate the data and create reports via Latex or visuals with tableau for the sponsors on the performance of the raters during the subject visits, or a trend over time in score responses of the subject visits such as are the scores showing improvement for a subject during the course of the study, or are they the same, or are they getting worse, etc.

3 CHALLENGES

According to Dr. Steven Targum in his article, The Distinction between Clinical and Research Interviews in Psychiatry, published by Innovations in Clinical Neuroscience Journal, "Clinical research attempts to minimize any extraneous factors that might affect assessments and adversely influence trial outcomes, including the potential for high placebo responses." (pg.44, March 8, 2011) The Placebo effect is a factor that exists in clinical trials today, and can impact the results of a trail. In order for sponsors to prove that their drug can be used as a treatment for a certain indication, the trial results of the drug have to be more effective than placebo. Unfortunately, placebo cannot be eliminated in a trial, but it can be mitigated. Other challenges include adverse events such as wrong dosage of the drug is given to the subject, protocol deviations or an event of death due to the trial, which can terminate the trail.

4 CONCLUSION

Clinical Trials are utilizing big data applications and analysis to store and track a multitude of documentation from the sites, vendors, sponsors, licensors, and CROs for regulatory purposes and operation of the trail itself. In addition, massive data is collected throughout the trial durations in different applications such as databases and is then investigated and analyzed for various reporting and purposes. Even with this technological progress in clinical trials, there is still more innovations to discover. Advancement in artificial Intelligence, and applications of Data Mining will change clinical trials of the future.

REFERENCES
[1] Dr Arun Bhatt. 2010. Evolution of Clinical Research: A History Before and Beyond James Lind. 1(1) (January 2010), 6–10. hhttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3149409/

5 BIBTEX ISSUES

Warning–no journal in Bhatt10

(There was 1 warning)

6 ISSUES

DONE:

Example of done item: Once you fix an item, change TODO to DONE

6.1 Formatting

Incorrect number of keywords or HID and i523 not included in the keywords

6.2 Writing Errors

Spelling errors

Do not use the phrase *In this paper/report we show* instead use *We show*. It is not important if this is a paper or a report and does not need to be mentioned

6.3 Citation Issues and Plagiarism

Claims made without citations provided

Need to paraphrase long quotations (whole sentences or longer)

6.4 Character Errors

Erroneous use of quotation marks, i.e. use "quotes" , instead of " "

If you see a ffigure and not a figure in text you copied from a text that has the fi combined as a single character

6.5 Structural Issues

The paper has less than 2 pages of text, i.e. excluding images, tables and figures