Querying the Queries – An AI Approach to Manage Clinical Data Quality

High quality clinical trial data is essential for a successful clinical trial. This data is the foundation for the analysis, submission, approval, labelling and marketing of a compound. A focus on data cleaning, an essential process in the collection and management of clinical data, ensures that the data collected is consistent and accurate.

However, it is not unusual for data errors to occur during data entry. Some of the most common are spelling or transcription errors, range errors and text errors, which impact coding. While automated edit checks exist to prevent the entry of inaccurate information, they are not able to detect all potential data entry issues. As data quality is at the crux of clinical trial success, clinical data management teams also use a manual approach to data cleaning by raising queries to the clinical trial sites to resolve any potential safety issues or inconsistencies in the data collected.

Often, on some studies, there can be numerous manually generated queries, which are time consuming and costly. By applying AI (artificial intelligence) techniques to understand the context of these queries, it may help improve automated edit checks or even offer opportunities to put additional checks or processes in place to help identify issues earlier in the studies. By applying machine learning to historic manual queries across different studies to understand common issues across and within studies, it could enable a more targeted approach to process optimisation for clinical trial data cleaning.

The Process of Data Collection

Clinical data management teams at each trial site work collaboratively to ensure that the data collected is managed in a conscientious way and then is reported clearly, accurately, and delivered securely to the data repository for access by a CRO or sponsor. An essential part of this process is data cleaning to ensure the data is consistent and accurate.

Generally, to ensure quality, data is checked for inconsistencies such as missing data. This step in the process can be automated as data is entered (edit checks). It can also be done manually after data has been entered through a query to the trial site.

Steps to Reducing Manual Queries via AI

In looking to determine how AI approaches could be utilized, it was necessary to review several clinical trial situations.

For example, in Study 1, a higher than expected number of queries were identified. In this case out of 21,103 queries, 7,560 or 36% were manual. Considering that the average cost of a manual query from start to close is about €150, the high percentage of manual queries added significant cost to the study.

Taking a closer look at the specifics of the manual queries, the data included the form and variable the query was raised on, the row the query was raised on and the query message.

That information provided the basis for further study with the aim of reducing the number of manual queries. However, there were questions to be researched, including whether or not we could identify themes in the manual queries without subjecting them to human bias.

LDA offered an opportunity to identify these themes. As a type of topic modeling algorithm, LDA is used as a pre-processing step in ML and applications of pattern classification. Its purpose is to reduce the number of data features to a more manageable number before the process of classification. And, secondly to learn the topic distribution of each document in a collection of documents.

Looking across a set of documents, for example, there may be commonalities in words which appear in each document, but these documents may have different numbers of these common words, which makes it difficult to identify and define the appropriate category for each document.

Via LDA, the mission was to determine if the number of manual inquiries could be reduced through the understanding of problematic forms. Could this technology result in more focused edit checks and could queries be auto-generated?

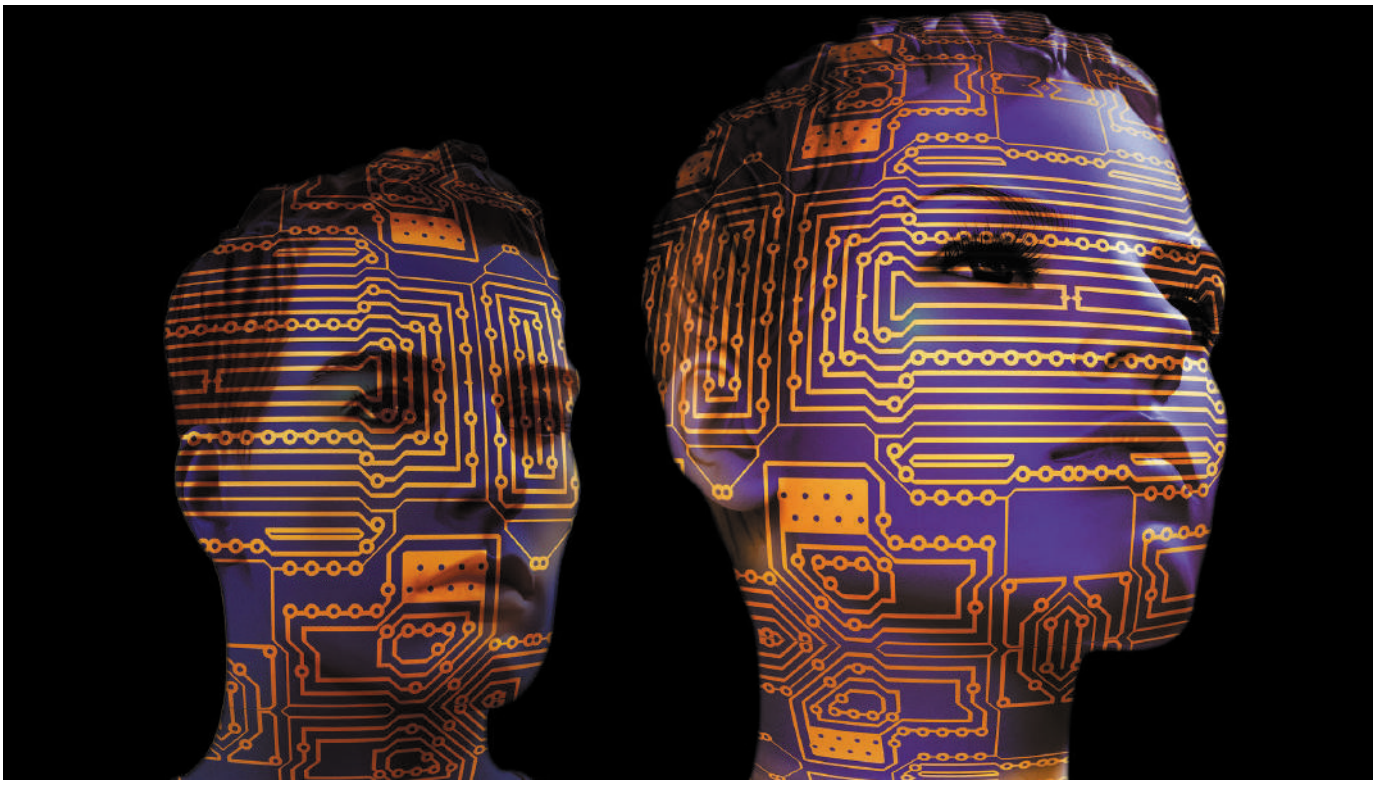
First, each query was tokenized or split into words with common data management words removed, i.e., confirm, verify, check, please, thank you, etc. and LDA applied to the queries to create a number of common topics for further review. The results of the LDA were visualised in the context of the forms used to collect the data and study experts consulted to bring deeper understanding of the context.

Study Results Provide Efficacy of AI Approach

The visualisations provided the study experts with insights into the different topics, identifying the most common words in each topic together with a summary of the context, such as the form and variable that the queries within a topic were raised against. This enabled the team to really understand what was driving queries within a topic. This understanding provided the basis to explore how these queries may be reduced going forward, for example through enhanced edit checks or, as described below, through a rules-based approach to speed up the discovery of the potential data issue.

Rules Approach to Further Study Research

Based on the efficacy of the AI approach and interpretation of the different topics, next steps included investigating whether the generation of rules could identify some of the queries in the 'top' topics identified. This would be a critical foundation for progressing



onward, to applying this approach to more studies, and would uncover comparative overlap and differences in each study.

To implement a rules-based scenario, however, a dataset would have to be created to assess the impact of the implemented rules using a historical snapshot from Study 1. For example, work with a topic result and an expert to generate rules around the AE and sick day medications (as described above). And finally apply these rules to the data snapshot and come to an understanding of how many manual queries are aligned with the results.

Study 2 was an extension study building on the results of Study 1, plus results from an earlier study in the same clinical indication that would help understand overlap and differences. As expected, there was a large amount of overlap. Interesting to note, however, that the differences in Study 1 that had been addressed were different in Study 2. For example, issues arose in central labs were in Study 1 but not Study 2.

Study 3 proved even more interesting. It was a completely different phase, study, sponsor, indication, and population and yet it also revealed some differences and some overlap.

Potential of the AI Approach as Part of the General Data Management Process.

A final step would be – if the investigation proved successful – exploring the potential of this approach as part of the general data management process.

Via LDA, the mission determined that the number of manual inquiries could be reduced through the understanding of problematic forms. Additionally, this technology would result in more focused edit checks and would allow queries to be auto-generated.

The exercise very definitely demonstrated the benefits of applying AI to the manual data queries generated during the data cleaning process. It proved that this technology provides the opportunity to significantly reduce the number of manual queries during a trail and thus increase efficiencies while reducing costs.

Although automation and AI techniques play a key role, managing and distributing clinical trial data will, for the foreseeable future be a human-machine endeavour. While machines may data-driven and more accurate than manual approaches, it will always require human attributes to provide the critical interpretation for better understanding of this data.

Jennifer Bradford

Jennifer Bradford, PhD is Head of Data Science for the CRO PHASTAR. She previously worked for the Advanced Analytics Group at AstraZeneca, leading the development of the REACT clinical trial monitoring tool, which she later customised and delivered to other sponsors as part of Cancer Research UK (CRUK). Within CRUK and in close collaboration with the Christie hospital she worked on EDC, app development and wearables data analytics in the context of clinical trials. She has a degree in Biomedical Sciences from Keele University and a bioinformatics Masters and PhD from Leeds University.

Sheelagh Aird

Sheelagh Aird, PhD is Senior Director, Data Operations for the CRO PHASTAR (www.phastar.com). She has more than 30 years of experience in clinical data management, Sheelagh has directed and delivered projects in all phases of clinical trials across numerous therapeutic areas and data collection platforms. Sheelagh holds a BSc in pharmacology and doctorate in pharmacokinetics from the University of Bath. She has led PHASTAR's Data Operations group since 2016.