



association for **clinical data management**

Challenge ChatGPT: CDMs vs AI

Authors: ACDM Newsletter Committee



Challenge ChatGPT: CDMs vs AI

Context

The recent ACDM Hot Topic 'Getting started with AI in CDM' (February 14, 2024), delivered by Mike Collinson and Derek Lawrence of Saama, stripped AI of its scary mask and let CDMs see that it is just another tool we can use in our quest for data perfection. Mike and Derek explained how machine learning models are trained using large datasets and need a human in the feedback loop to catch any errors and make the models more accurate.

Furthermore, whilst the machines can learn and provide recommendations based on the input, it is ultimately the human's responsibility to make the decisions and take the necessary actions using the data that AI produces. For example, say the output produced by the AI flags that PRO data for multiple participants is being entered by the same user – an obvious red flag for fraudulent data. Upon manual review of the output, it is discovered that the username of the user is EPROVENDOR1. Since the PRO data is being imported directly to an EDC system via an ePRO vendor for this project, this is expected for this type of data and not fraudulent data. Possible action to be taken here would be to update the ML model to advise that EPROVENDOR1 is the only legitimate username for multiple participants across sites in the project. If the AI tool was empowered to take action, the data collected would be incorrectly labelled as fraudulent and could have been excluded from the analysis

During the presentation, Mike and Derek explained that many Cloud providers offer training in Machine Learning and some platforms allow the use of their ML tools for free, such as OpenAI who created ChatGPT. The Newsletter Committee decided to challenge one of these platforms, pitting the Artificial Intelligence brain against the Actual Intelligence of the committee members.

The question: 'How can data managers ensure the integrity of the data being collected in the clinical trial?' was first put to the Committee Members. Once their article was finalized, the same question was typed into ChatGPT v3.5.

How can data managers ensure the integrity of the data being collected in the clinical trial? – by the ACDM Newsletter Committee

One of the Clinical Data Manager's responsibilities is to make sure that the data collected for the clinical trial report is comprehensive, accurate and readable. Clean data is the essential ingredient needed to report accurate results and therefore CDMs build in multiple cycles of reviews and checks into their data plans. At each step of the data collection process, checks are put in place to ensure the integrity of the information being collected. Processes are defined for performing the reviews in order to be able to quantify the cleanliness of the data they are responsible for. Below is a summary of the checks DMs can incorporate to ensure the integrity of the data being collected:

- Fields in EDC systems are built to collect certain data types (e.g. numeric, character).
- Programmed edit checks are built into EDC to run across the entered data, including checks for:
 - > Data not within an expected range
 - > Date and times of assessments not performed when expected
 - > Missing data
- Reports are programmed to enable identification of errors in data which cannot be identified via EDC
 - > Vendor Data, such as lab results, ePRO, IXRS, imaging data, etc.
- Reports are programmed to compare the data in the different repositories
 - > Data collected in EDC is compared with Header Data from Labs, Imaging, ePRO, Wearables, etc.
 - > Integrated data is compared with the source, any changes in the source should be reflected in the data repository
- Data Transfer Plans are created to ensure that external data conform to the required specifications
- Outputs of datasets are produced for functional experts to perform manual review of the data as a whole
 - > Centralized Monitoring, Medical Review, Adjudication of Adverse Events, etc.
- Audit Trails of the systems used in the data collection are reviewed
 - > Check the audit trails for unauthorised users performing inappropriate actions on the data. For example:
 - >> Site user at one site editing another site's data
 - >> User with DM role inactivating a data point
 - >> Check the timing of the entry of the data
 - >> Data entered before the event occurred
 - >> ePRO data expected to be collected over several days is entered on the same day
 - >> Check for unusual/unexpected activity
 - >> Multiple changes of the same datapoint
 - >> Wearable data showing device had been idle/switched off
 - >> Significant delay between the assessment performed and the data being entered

Challenge ChatGPT: CDMs vs AI

- Reports are programmed to flag protocol deviations that have occurred in the clinical trial
- QC is performed on the data and actions on the data
 - > All updates to systems and programs are reviewed and quality checked at the time of action
 - > Anomalies in the data are checked to ensure that the data has been queried and the anomaly has been correctly addressed as per process.

Clinical Data Managers must consider all the tools in their arsenal. They must look at the project as a whole and choose which tools are the best fit. For projects with millions of data points, it is not feasible to perform manual review on each and every piece of data collected. Conversely, for small projects yielding lesser amounts of data, it may not cost effective to program numerous complicated edit checks, when it would take less hours to review the data manually.

Clinical Data Managers must start the project with the end in mind. They must consider the data sources, consider how the data will be formatted and collated, consider how the data will be cleaned and how they can show the data collected is complete and accurate. The integrity of the clinical trial results depends on the integrity of the data collected; the results should be able to be replicated, therefore it is essential that the data and all actions on the data are transparent and beyond reproach.

How can data managers ensure the integrity of the data being collected in the clinical trial? – A response produced by ChatGPT v3.5

Ensuring the integrity of data collected in a clinical trial is crucial for maintaining the reliability and credibility of the trial results. Data managers play a vital role in this process. Here are several strategies they can employ:

1. **Data Collection Procedures:** Implement standardized data collection procedures to ensure consistency and accuracy in data collection across different sites and researchers involved in the trial.
2. **Data Validation:** Implement data validation checks at the point of data entry to identify and correct errors or inconsistencies in real-time. This includes range checks, format checks, and logical checks.
3. **Training and Education:** Provide comprehensive training to staff involved in data collection to ensure they understand the importance of data integrity and are proficient in data collection procedures.
4. **Quality Control Checks:** Conduct regular quality control checks on the collected data to identify any discrepancies or anomalies that may indicate issues with data integrity.
5. **Documentation and Record Keeping:** Maintain detailed documentation of all data collection procedures, including any deviations or corrections made during the trial. This helps ensure transparency and traceability of the data.
6. **Data Monitoring and Auditing:** Implement regular data monitoring and auditing processes to assess the accuracy and completeness of the collected data. This may involve independent review by external auditors or data monitoring committees.
7. **Use of Electronic Data Capture (EDC) Systems:** Implement EDC systems that offer features such as built-in validation checks, audit trails, and data encryption to enhance data integrity and security.
8. **Adherence to Regulatory Standards:** Ensure compliance with regulatory standards and guidelines, such as Good Clinical Practice (GCP) guidelines, which outline requirements for data integrity and management in clinical trials.
9. **Risk Management:** Identify potential risks to data integrity early in the trial planning phase and develop strategies to mitigate these risks throughout the trial.
10. **Data Encryption and Security Measures:** Implement robust data encryption and security measures to protect the confidentiality and integrity of the collected data, especially when transmitting or storing sensitive information.

By implementing these strategies, data managers can help ensure the integrity of the data collected in a clinical trial, thereby enhancing the reliability and credibility of the trial results.

Conclusion

First thing we noticed was the speed at which the article from ChatGPT was generated. Lines of text appeared on screen faster than the average reading speed and the article was completed in less than 10 seconds. The format of the response from ChatGPT was also note-worthy; an introduction, list of points and a brief conclusion, any schoolteacher would be proud of it for giving the answer in a fully structured essay format. Each of the ten points ChatGPT makes is relevant to the question, but in a broader context than the article produced by the committee. The committee article focusses on the specific tasks involved in data cleaning, whereas ChatGPT explores DM concepts, highlighting how these practices feed into data integrity assurance.

The AI revolution is here and it's not going away. Instead of an apocalyptic take over by machines, Terminator style, the machines will still need the humans to execute the programs and feedback on the outputs. As CDMs we can utilize this technology to build in efficiencies with data collection and processing. We can upskill and learn how to use the ML tools to design better data collection tools, streamline data flows, interrogate the data as a whole, rather than nit-pick at individual data points. As the above articles show, neither the Artificial Intelligence or the Actual Intelligence has the edge; the ChatGPT article was generated much quicker than the newsletter committee's but the newsletter committee has real life experience of clinical data management and provided more detail in their article. A writer often dreads the blank page, therefore ChatGPT could be used as a starting point for a written document, it could be used as a framework for the human to loop their own ideas around.

acdm 

association for **clinical data management**

www.acdmglobal.org