AI in Drug Discovery and Development

Prof. Rajnish Kumar
Dept. of Pharmaceutical Engineering and Technology
IIT-(BHU), Varanasi

Week-08
Lecture-40

Welcome to the course "AI in Drug Discovery and Development." In today's session, we will talk about data collection and monitoring in regulatory submissions. So, by the end of this lecture, you will be able to understand the importance of accurate and standardized data collection in clinical trials. Identify key tools and technologies used for data capture, explain the role of data standards in ensuring regulatory-ready data, and distinguish between traditional. And risk-based monitoring approaches in clinical trials recognize the structure and components of regulatory data submission packages. And explore how AI is being integrated into clinical data monitoring and regulatory processes to improve efficiency and decision-making.

So let us see the importance of data in clinical trials because the whole concept of a clinical trial is actually based on data. So, data is, you know, one of the most important pillars of clinical trials. So, we talked about the study design and protocol development. So, we need to define the objectives.

So the accurate data. It helps researchers to clearly define trial objectives such as primary and secondary endpoints. And the data can help us in, you know, determining the sample size as well as the statistical power. and sample size calculations they rely on preliminary data to ensure that the trial can detect meaningful differences between the treatment groups. So, the data is important for deciding the inclusion and exclusion criteria as well because it guides the selection of appropriate participants to ensure that the study population reflects the target patient population.

Talking about the quality of data collection, there are standardized procedures, such as rigorous data collection protocols like electronic data capture systems. They minimize the errors and variability and thus ensure consistency across the study sites. So, in case we have, for example, multiple sites from which the data is coming. So, this sterilization procedure is quite important to make sure that all the data is reliable and is usable in the trial. And then the real-time monitoring, like the continuous data monitoring, allows early identification of issues such as protocol deviations.

or adverse events, which are crucial for patient safety as well as validation verification processes like data cleaning and validation. They help ensure that the data is accurate, complete, and reliable. So, let's talk about the data analysis and interpretation. So, the efficacy and safety evaluation. So, robust statistical analysis of trial data is key to determining whether a treatment is effective or safe.

We can only know the efficacy and safety if we have enough data for the trial. And then handling variability, properly managed data allows for subgroup analysis. Adjustments for confounding factors increase the reliability of the conclusions, as well as adaptive designs wherein ongoing data analysis in some clinical trials enables adaptive modifications, such as those adjustments. or sample size re-estimation to optimize the trial's outcomes. Talking about regulatory compliance and decision-making, data acts as evidence for approval.

The regulatory bodies, such as the FDA or AMA, require comprehensive, high-quality data to make informed decisions on drug or device approval. As well as helping to improve the transparency and reproducibility, detailed data records support the reproducibility of trial results and provide a transparent audit trail for regulatory review. And then, data is also important for post-market surveillance, where the data collected during and after clinical trials helps in monitoring long-term safety. Effectiveness once a treatment is available to the public, discussing the ethical considerations related to data. So, accurate and timely data collection is critical for protecting trial participants by ensuring that adverse events are promptly identified and managed.

So, that is how it ensures patient safety as well as reliable data, which reinforces the trust between participants. And researcher ensuring that participants are fully informed about the risks and benefits of their involvement, and that is what you know informed consent is all about. So, we talk about the types of data, so there can be multiple types of data; basically, we can divide it into quantitative and qualitative data, where the quantitative data can be categorical. or numerical, where the categorical will be like nominal or ordinal, and the numerical will be kind of discrete. or continuous, while the qualitative data will be in the form of descriptions, or it can be in the form of diagrams, or in the form of maps.

Okay, so what are the sources of data in clinical trials, actually? So, in the modern trial, they use a combination of manual and digital tools depending on the study scale, complexity, and location. So, these can be different sources of clinical trial data. So, these can be case report forms, short-named as CRFs, where all the details are actually filled into these forms. And then these can be electronic health records of those participants in the clinical trial. This can be clinical monitoring reports, or it can be laboratory and imaging data, which is coming from, you know, all those, you know, biomarker identification.

Evaluation studies can come from wearable devices and sensors to collect, for example, vitals and other parameters for use in the clinical trial. So, these are different sources from which the data can come in a clinical trial. So, if we talk about the case report forms, these are usually paper case report forms, which are traditional physical forms completed manually at the study site. Nowadays, it has been replaced by electronic CRFs, eCRFs. So, where these digital forms are filled using electronic data capture systems.

And then these electronic data capture systems or electronic health records are kind of software platforms that allow investigators to enter, store, manage, and transfer trial data electronically. And then we have wearable technology and remote monitoring devices, like those worn by or used near patients to collect physiological or behavioral data. In real time or near real time. Some examples of wearable technologies are smartwatches, continuous glucose monitors, etc. So now we have talked about those, you know, different sources of data and why data is important in clinical trials.

So let us talk about some of the standards in data collection. So, data in clinical trials must be consistent, high-quality, and regulatory-ready. So, these three characteristics are very important for the data which we are creating in the clinical trials. So, the standardization enables faster regulatory reviews, improved data sharing, and greater efficiency in analysis and submission. So, there are various regulations.

Clinical Data Interchange Standards Consortium (CDSC) provides global data standards for clinical research. And then you have 21 CFR Part 11, which is a regulation of the US FDA. Defining criteria under which electronic records and electronic signatures are considered trustworthy and equivalent to paper records. And then you have the ICH E6 R2 guideline, which discusses the good clinical practice guidelines. And then, a set of good documentation practices and principles endorsed by the FDA, WHO, and AMA are known as the ALCOA principle and the ALCOA plus principle.

So originally it was like Alcoa, and then it was extended to AlCOA Plus to emphasize data integrity. So, what are those, AlCOA? So, the first principle is the attributable, which means who has performed the action and when the action was performed. The second principle is legibility, which means the data must be readable and understandable. Third is the contemporaneous record, which is made at the time of the activity; it should be the original first capture of the data or a certified copy, and it shall be accurate, error-free, and truthful. So, earlier it was like Alcoa, and then a plus was added, like it should be complete, with no missing data or unexplained gaps.

It shall be consistent. It shall be uniform throughout the trial. It shall be enduring. Data must be preserved over time. And then it shall be available and readily accessible for audits, inspections, or analysis. And then, coming to the data monitoring.

So data monitoring in clinical trials is the ongoing oversight of trial data and processes to ensure accuracy, completeness, and compliance with the protocol, good clinical practices, and regulatory requirements. So, there are basically two types of monitoring techniques. One is traditional monitoring, which involves routine on-site visits with extensive source data verification for all data points. And then there is risk-based monitoring, where a targeted approach is used that focuses monitoring efforts on the most critical trial risks and data using centralized and remote tools. So, if we compare these methods, like on-site monitoring, clinical research associates physically visit the trial site to perform source data verification, review the records, and check the drug accountability.

And it is, you know, used mainly in complex trials, first-in-human trials, or high-risk trials. And then you have the centralized monitoring system where the data is reviewed from a central location using real-time analytics, dashboards, and automated checks. So usually, it is used in large multicenter trials, and then you have remote monitoring where the clinical research associate performs data review and source checks remotely using the electronic systems. So usually these are used in hybrid trials or decentralized trials, especially in pandemic settings as well. Okay, once we have generated the data.

So next is the data submission package. So, there is this CTD, a common technical document. So, it is a harmonized format developed by the ICH for submitting information to the regulatory authorities. So, it has, you know, module 1, which talks about the regional admin information. And then you have module two, which contains the summaries of modules three to five.

And then you have module three, which is about the quality, CMC, chemistry, manufacturing and control, all the details about the synthesis of those molecules, manufacturing of those molecules, and quality control data for all those molecules. And then you have module four, which is, you know, non-clinical study reports like toxicology or PKPD studies. And then you have Module Five, which includes the clinical study reports and the datasets. The clinical study report (CSR) is a comprehensive document that presents the methods, conduct, and results of a clinical trial, typically for phase 1 to 3 trials; the format and content are guided by the ICH E3 guidelines. So, the CSR includes trial objectives and design, patient demographics, efficacy and safety results, protocol deviations and adverse events, as well as statistical analysis.

And then you have the integrated summaries. So, the integrated summary of safety (ISS), which combines safety data from multiple clinical trials, evaluates the overall safety profile. And then you have an integrated summary of efficacy, which aggregates efficacy data across trials to support clinical claims. So, these both are submitted as part of NDA, BLA module 5 and requires standardized data sets and clear traceability. So, if you see this,

this is the clinical data management process where the review and finalization of the study documents are followed by database design and data collection.

Followed by CRF tracking, followed by data entry, data validation, discrepancy management, and medical coding. Database locking, extraction, and archival. So, this is, you know, the clinical data management process: how the data is generated, processed, and stored in a clinical trial and used in a clinical trial. Talking about the database design, the idea is to create a robust electronic database that will house all the clinical trial data. So, define the structure like table field data type according to the study protocol and CRF, which are case report form requirements.

And set up validation rules, like range checks or format checks, to reduce errors at the point of data entry and ensure the design meets regulatory requirements, such as 21 CFR Part 11 compliance in the US. The next step is the review and finalization of the study documents. where we confirm that all study related documents, protocols, CRFs, data management plans etc. are consistent and ready to use. So, we finalize the study protocol to outline objectives, endpoints and methods.

We finalize the CRFs or eCRFs that specify exactly what data will be collected, and we obtain the necessary approval from regulatory authorities and the ethics committees. The next step is the data collection, where we gather the participant data in accordance with the protocol and CRF specifications. So, the data can be collected from multiple sources like patient visits, laboratory tests, imaging, and electronic health records, and we have to ensure that the data is recorded accurately and completely, whether on paper. Or it is in the electronic data capture systems and maintains the participants' privacy and follows GCP guidelines. So then, with CRF tracking, we monitor the flow of CRFs to ensure all required data points are accounted for and that the log of received CRFs is maintained.

Check for missing pages or sections, track the status of CRFs across different sites to identify delays or incomplete submissions, and communicate with clinical sites to resolve the missing or late data. The next step is data entry, where the collected data is input into the clinical database accurately, and for paper-based trial data entry, staff transcribe CRF information into the database. And for EDC-based trials, site staff or coordinators enter data directly at the point of collection, and we need to do a double data entry, or built-in system checks may be used to minimize transcription errors. The next step is the data validation, where we apply quality checks to identify any errors or inconsistencies in the entire dataset. We run automated validation rules like range checks, data consistency checks, or missing fields, and generate the queries for data points that fall outside expected parameters or appear incomplete.

and use the validation reports to quickly spot trends or repeated errors. And then we need to, you know, manage the discrepancy as well. So, we resolve any data discrepancies or queries raised during the validation process. We send those queries back to the site personnel, such as the investigators or study coordinators, for clarification or correction. And then document all changes and justifications to maintain a clear audit trail and ensure timely resolution so that the data set remains as clean and up to date as possible.

So that is medical coding. The next thing is medical coding, where standardized medical terms like adverse events, medications, or diagnoses for consistent analysis and reporting are used. And then we use standardized coding dictionaries like MedDRA for adverse events and the WHO Drug Dictionary for medications. And we map verbatim terms reported by investigators or participants to standardized codes and maintain consistency across all study sites and time points. The next topic is database locking. It frees the database to prevent any further changes once the data is deemed clean and complete.

Confirm that all queries are resolved, all data points are verified, and obtain sign-off from relevant stakeholders, such as the data manager, biostatistician, and project leads. Log the database so the data set can be used for final analysis, ensuring no further edits can occur. The last step is the data extraction archival, where we prepare the final data set for analysis and securely store all trial-related data and documents. Export or extract the log data set for statistical analysis and regulatory submission. Archive the database, CRFs, and associated documents in compliance with regulatory requirements, which are often 5 to 15 years.

or more for specific studies and ensure secure long-term storage with controlled access for the audits and inspections. So, there are multiple gaps in the current system. So let us discuss them one by one. So, the first thing is inconsistent or non-standardized data formats. So, there are multiple sites that may use different data capture methods, such as paper versus EDC systems, leading to variations in the data format.

Inconsistencies in quality can delay data cleaning and validation, making it difficult to compile a cohesive data set for regulatory review. Another limitation is the limited integration with electronic health records. So many clinical trial databases are not seamlessly integrated with EHR systems, requiring manual data entry or cumbersome data transfer. So, the manual process increases the risk of transcription errors. Data duplication, missing information, and the lack of real-time or near real-time monitoring.

Where traditional data monitoring often relies on periodic site visits or retrospective reviews, it may not catch errors or issues early. So, the delayed detection of data discrepancies, protocol deviations, or adverse events can compromise patient safety and data integrity. So, another issue is the inefficient query resolution process, where the discrepancies identified during data validation can be slow to resolve due to the manual process. Limited site resources and the prolonged query turnaround time can delay data backlog and final regulatory submissions. And then you have the underutilization of risk-based monitoring as well.

Although risk-based monitoring approaches are recommended to focus resources on high-risk sites or data points, many trials still rely on 100% source data verification. And this can be both time-consuming and costly. Diverting attention from critical data issues that may truly impact trial outcomes. And then you have the complexities of decentralized or remote trials. The rise of virtual or hybrid clinical trials introduces new data sources, like wearable devices or telehealth platforms, that are not always well integrated.

And in data fragmentation technology, interoperability challenges can lead to incomplete data sets and complicated monitoring workflows. And then the issues related to data privacy and security concerns. So increasingly stringent regulations like GDPR or HIPAA require robust systems and processes to protect the participant data. Ensuring compliance can be complex, especially across multiple regions, and can slow data sharing and monitoring efforts if not managed properly. And then we have a limited use of advanced analytics for quality oversight.

Many organizations still rely on manual or basic statistical methods to identify outliers and trends in clinical data. So, without using advanced, you know, analytics and ML-based anomaly detection, some data quality issues or safety signals may go unnoticed until the later stage. Related to the resource constraints and site burden, the clinical sites often have limited staff and resources, leading to backlogs in data entry, query resolution, and monitoring activities. And the under-resourced sites can cause delays in data availability and reduce the overall data quality. Okay, so now we have talked about those gaps in data collection and processing.

So, where can AI come in? How can we use AI? So, it can be used for, you know, automated data capture integration. So, AI-driven tools can automatically extract data from unstructured sources like paper records or PDFs to populate clinical databases, reducing manual error. So, you can have, you know, the AI systems that can immediately flag inconsistencies or missing data. as it is captured, enabling prompt error correction.

So these are intelligent data-quality checks. So, on the fly, they can check whether there is some problem with this data or there is an error in this data. So that can be done. You can use NLP for unstructured data. NLP is very good at it. So, it automatically maps free text descriptions of adverse events to standardize medical codes, enhancing consistency across data sets.

Then you can use AI for risk-based monitoring and predictive analytics. So, AI identifies and prioritizes high-risk sites or data points, ensuring that monitoring efforts are focused where they are most needed. And then you can use it for continuous safety surveillance, where AI continuously analyzes safety data in near real-time to detect emerging risks or trends, improving patient safety oversight. And then you can use it for streamlined regulatory reporting as well. AI can compile and format data from multiple sources into submission-ready reports, facilitating a smoother regulatory review process.

So you can see, like, in the data analysis and data curation. So, you can use AI for at almost every step and very efficiently. So let us see some of the notable AI and ML models or tools that are being used. So, for automated data capture integration, there is this layout lm. It is a transformer-based model designed for document understanding, capable of extracting structured information from forms and scanned documents. Then, for the intelligent data quality checks, we have the isolation forest, which is an unsupervised learning algorithm that efficiently detects anomalies and data quality issues by isolating outliers in a data set.

And for risk-based monitoring and predictive analytics, you have XGBoost, which is a powerful predictive modeling tool that includes risk stratification and performance forecasting in clinical trials. For natural language processing, you can use BioBERT. And for continuous safety surveillance, you can use LSTM networks. And then, for streamlined regulatory reporting, you can use T5, which is a text-to-text transfer transformer and a versatile model. Can be fine-tuned for tasks such as text summarization and automated report generation, streamlining the creation of submission-ready documents.

So let us have a look at some real-world examples, like the Medidata AI Simulants, which is a synthetic data generator. So, it uses generative AI to create synthetic data sets that mimic real-world clinical trial data while preserving privacy. So, these data sets helps optimize protocols and predict regulatory compliance risk. And then, the impact of using Medidata is that it reduced protocol amendments by 30% by simulating trial outcomes during the design. So, we have the Deep6 AI, which is a tool for EHR analysis and eligibility screening.

It uses NLP, which extracts unstructured data from over 30 million electronic health records to identify eligible patients and streamline data collection for regulatory submission. So, the impact was that it cut screening time by 42% in an Alzheimer's trial and improved diversity in submissions by identifying the underrepresented cohorts. And then we have Certa's compliance automation, where the AI tracks GDPR or CCPA compliance in data storage and automates consent management for global trials. So, the impact is that it reduces the compliance cost by almost 30% via automated workflows and eliminates 50% of audit risks through real-time regulatory updates.

And then you have the MAD Institute's AI for data de-identification. So, it automates the de-identification of medical records and imaging data to meet privacy regulations such as HIPAA or GDPR. So, the impacts is like it shortened the data submission timeline by 40% in cardiovascular studies and also enabled rapid response to safety alerts by FDA reporting. Benevolent AI is target validation. So, AI identifies high potential drug targets and validates them using the historical trial data for regulatory filing. So, the impact was that it advanced you know a fibrosis drug candidate to phase 3 with robust preclinical data and it reduced the target validation time by 40% accelerating the IND submissions.

Okay, in the end, I would like to discuss this tool called PyTrial. So, it's a kind of Python package that implements various clinical trial tasks supported by AI algorithms. So, it can implement six essential track development tasks, such as patient outcome prediction, trial site selection, trial outcome prediction, patient trial matching, trial similarity search, and synthetic data generation. So, this is how you know: you can see that it has 23 AI/ML-ready data sets and is using 34 ML algorithms; thereby, it can work on these 6 AI4 trial tasks that we have discussed now. So, it can be used, you know, starting from phase 1 to phase 4, and then it can be, as I said, used for patient trial matching, for trial search, for patient data simulation, for trial outcome prediction, for trial site selection, and for trial outcome prediction as well.

So, this is the architecture of how this PI trial is actually built. So, I would say that it's one of the wonderful tools which have been developed. and this is not the only tool actually there are plenty of tools which are being developed for their application in the clinical trial Most of them are, you know, open source, actually, so those can be used for free without paying any subscription charges. So, then we talk about the Consort AI guidelines for regulatory submissions. So, Consort AI is a kind of consolidated standard for reporting trials.

AI is the AI-specific extension of the widely accepted CONSORT 2010 guidelines. So, it ensures that the results of AI-based trials are reported clearly and transparently, and it enables readers, such as clinicians, researchers, and regulators, to understand what was done. Why and what was found to assess the validity of the results and avoid bias? So, coming to the summary, the data collection and monitoring form the foundation of high-quality clinical research, ensuring that trial data is reliable, complete, and regulatory compliant. So, the standardization through CDISC formats like SDTM, ADaM along with adherence to 21 CFR part 11 and AlCOA+ plus principles is essential for submission readiness and auditability. Monitoring strategies, whether traditional or risk-based, play a pivotal role in safeguarding data integrity and participant safety across trial sites. And data submission packages, including CTDs, CSRs, and integrated summaries, streamline regulatory reviews and facilitate the approval process.

And compliance with HIPAA and other privacy frameworks ensures the protection of participant information throughout the trial lifecycle. And AI-driven tools are increasingly integrated into data monitoring and submission workflows, enabling real-time insight, error detection, and enhanced decision making. So, with that, I have an open question for you. As AI continues to reshape how we collect, monitor, and submit clinical trial data, could future clinical trials become fully decentralized and self-monitoring, minimizing human oversight while maintaining regulatory compliance and data integrity? So, these are some of the references that you can go through to get more information about this topic. And with that, thank you.