Artificial intelligence and machine learning in clinical data management: applications, challenges, and future perspectives

Miss. Shweta D. Dhikale[1], Prof. Rohan R. Kherud[2], Payal M. Donagre[3], Radhika D. Aher[4], Dnyaneshwari G. Aher[5]

[1,3,4,5]Student, [2]Professor
SND College of Pharmacy Babhulgaon, Yeola, Dist. Nashik, Maharashtra, India- 423401.
Corresponding Author-Miss.Shweta D. Dhikale*
Email ID- shwetadhikale2710@gmail.com

1] Abstract:
Clinical Data Management (CDM) is a key element to ensure the accuracy, completeness, reliability and integrity of the data collected during the clinical trials. However, the increasing volume and complexity of clinical data collected from electronic data capture (EDC) systems, electronic health records (EHRs), laboratory systems, wearable devices and patient-reported outcomes have presented challenges for traditional data management approaches. Artificial Intelligence (AI) and Machine Learning (ML) offer promising opportunities for improving the efficiency and quality of clinical data management. This review covers the applications, challenges and future prospects of AI and ML in CDM. It highlights the potential for these technologies to be applied in automated data cleaning and validation, intelligent query generation, medical coding, data integration and reconciliation, anomaly detection, missing-data management, predictive analytics, and continuous data quality assessment. Natural Language Processing (NLP) and Generative AI may also further assist in the interpretation of unstructured clinical information, automated documentation, and data-review activities. Such technologies can reduce repetitive manual work, improve consistency and enable timely locking of databases. However, there are issues such as patient privacy and data security, algorithmic bias, low quality training data, model transparency, system integration and regulatory compliance, in implementing AI. This requires the right model validation, data governance, audit trails and performance monitoring. And the more important it gets there to be a human-in-the-loop to make sure that a qualified professional looks at the AI-generated outputs. Generally, When used responsibly and with proper oversight by humans, AI and ML have great potential to make the CDM more efficient, reliable and data-driven.

Keywords: Clinical Data Management, Artificial Intelligence, Machine Learning, Clinical Trials; Data Quality, Data Cleaning, Natural Language Processing, Generative AI, Predictive Analytics, Human-in-the-Loop.

2] Introduction:

2.1 Clinical Data Management [CDM]: Clinical research refers to research carried out with human participants in order to produce knowledge that improves health, health care, and health systems by evaluating health conditions, diseases, and medical interventions.[1] One question that research trainees frequently ask is "What is Clinical Data Management (CDM) and why is it important?" CDM is an essential and integral component of clinical trials, and in practice nearly all researchers engage in CDM tasks during their projects,

regardless of whether they consciously recognize these activities as CDM.[2] Clinical data management (CDM) plays a vital role in making sure that clinical research data is dependable.This is accomplished by managing the whole workflow, from assisting with planning how data are captured, to supervising collection, applying quality assurance, and getting the data ready for analysis. The overarching aim of CDM is to produce quality data by reducing errors and missing information and by increasing the amount of quality data available for analysis . This function underpins the overall quality of Clinical Data Management in Clinical research.[3] It is also in line with the ICH E6 (R3) Good Clinical Practice guideline, which sets out the fundamental values for data governance, including data integrity, traceability, confidentiality, reliability and fitness for purpose.[4]

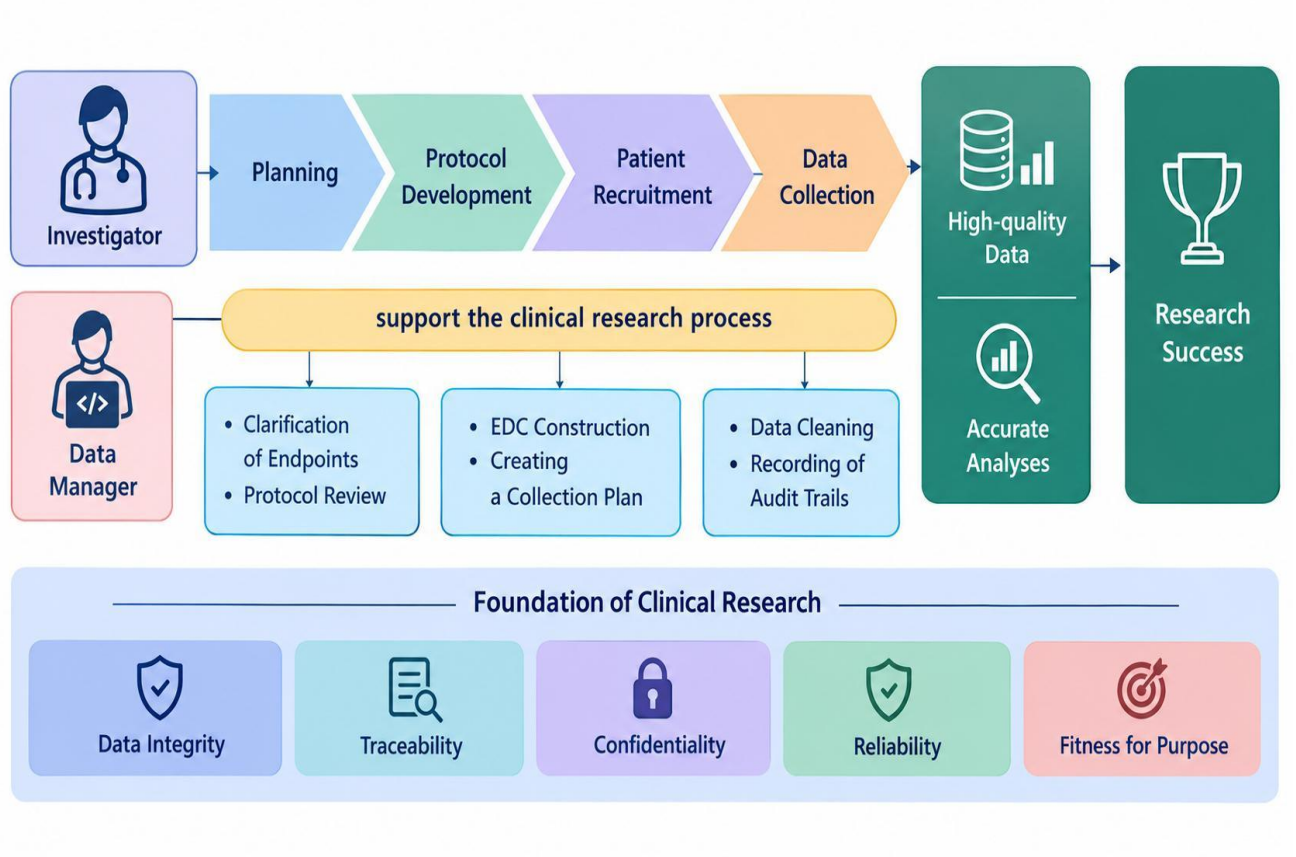


Figure 1: Clinical Research Process And Data Management:

Clinical research is conducted in multiple phases, and every phase requires substantial support for data management. Data managers collaborate with investigators to plan studies, create protocols, enroll patients and gather data, helping to ensure that the information obtained is accurate, complete and appropriate for analysis. Their efforts contribute to the overall quality and success of the research. In the end, the five principles of data integrity, traceability, confidentiality, reliability and fitness for purpose constitute the basis of clinical data management and correspond with the requirements of ICH E6 (R3).The field of CDM has changed significantly over recent years. Nowadays, data from clinical trials is gathered from numerous sources beyond conventional case report forms (CRFs), including electronic health records (EHRs), wearable devices, patient reported outcomes (PROs) and medical images. As a result, the amount of data grew and the complexity of data structures rose.[5] These shifts have heightened the workload of CDM and made it increasingly necessary to preserve data quality as much as possible. Meanwhile, qualified data managers are in short supply, particularly in organizations lacking dedicated CDM departments.[6] Given this situation, the rapid advancement of artificial intelligence (AI), especially in large language models, has drawn interest as a potential tool to support CDM activities such as data cleaning, medical coding, and query generation.[7] This review describes examples of AI use in core CDM-related tasks, including data cleaning, medical coding, and

query generation, and considers the regulatory requirements and risk management factors necessary for the responsible integration of AI in clinical research.[8]

2.2 Clinical Data Life cycle: In the context of systematic clinical data quality management, the life cycle is a set of quality assurance activities of the healthcare data during the construction process as well as in its daily use. In other words, it is a period which comprises a number of procedures rather than one particular activity. According to the source, life cycle is defined as an umbrella term that includes "all steps in the process" (collection) of quality management.[9]

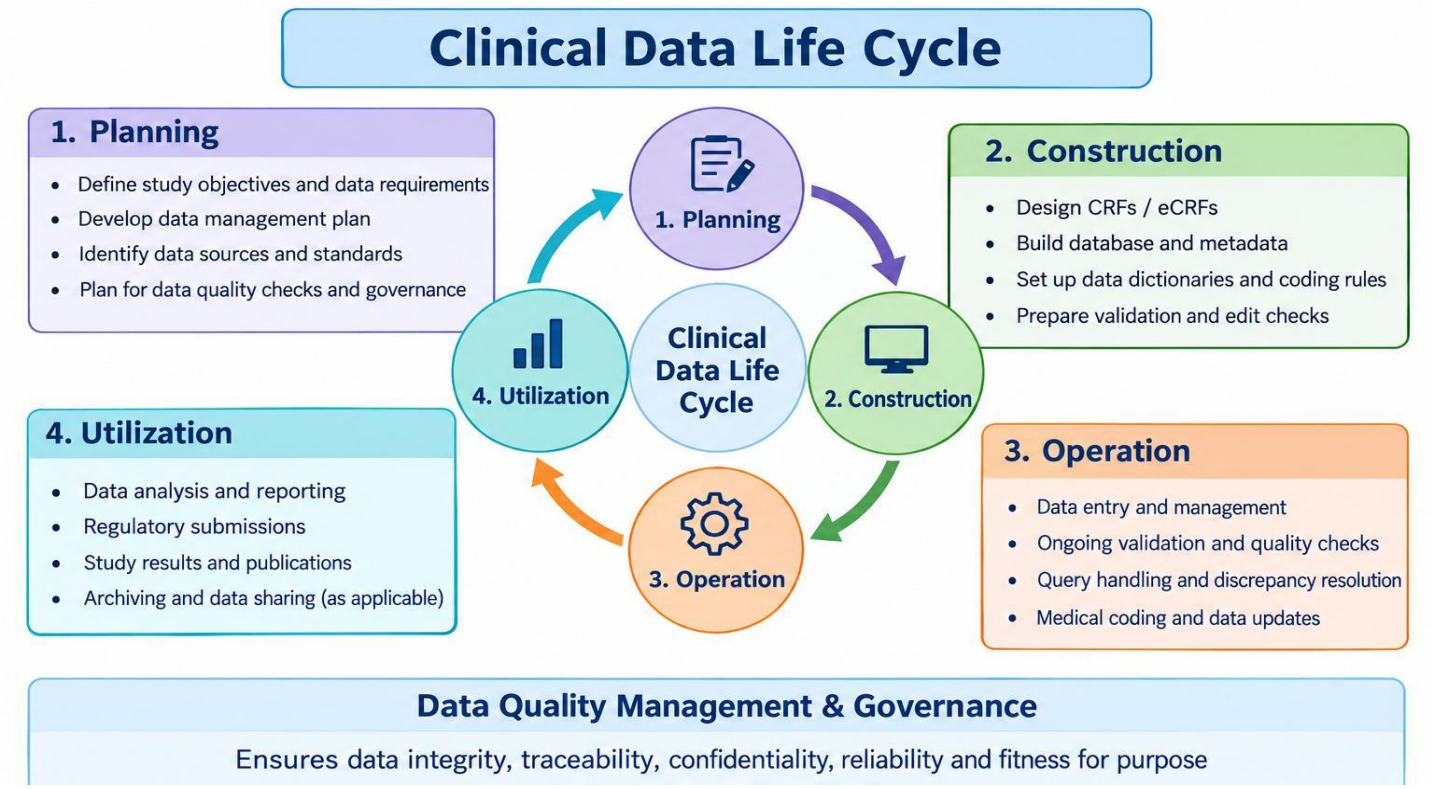


Figure 2: Presents the clinical data life cycle and outlines the data quality management methods applicable to each stage. The life cycle comprises four stages: planning, construction, operation, and utilization. To ensure high-quality data, data management should align with the relevant life-cycle stage and follow established governance principles.

Clinical data management follows a structured life cycle that begins with study planning and continues through the collection, monitoring, and use of clinical data. Rather than treating data quality as a final checking activity, this approach integrates quality measures throughout the entire process. It supports consistent data handling, effective governance, reliable quality assessment, and the responsible use and reuse of clinical information. Recent literature commonly describes four connected stages: planning, construction, operation, and utilization.

Stages Of Clinical Data Life Cycle:

1. Planning stage:

The planning stage lays the framework for managing clinical data throughout the study. It is during this time that the study objectives are set, the elements or sources of the data are identified, and the data standards are chosen. Moreover, a comprehensive data-management plan is prepared. Planning should involve

documentation, storage and security, stewardship, accessibility, and the potential for data reuse in the future. By establishing quality expectations at an early stage it becomes possible to implement suitable controls and procedures before data collection starts, reducing the risk of issues arising later during the study.

2. Construction stage:

The construction stage comprises the creation of systems and processes required to collect and organize clinical data. It is often related to designing case report forms (CRFs or eCRFs) in addition to database designs, metadata, data dictionaries, coding standards and validation or edit checks.Once collected, the data may require cleansing, transformation, or standardization to meet the specific requirements set. Quality-control activities should be incorporated into this stage, not delayed until the data is ready for analysis. Addressing issues early reduces the possibility of errors propagating throughout the study-database.[10]

3. Operation stage:

In the operation stage the clinical data environment is actively monitored and maintained. Data is assessed through predefined checks that detect missing values, inconsistencies, implausible outcomes and discrepancies. These commonly include data validation, query handling and resolution, discrepancy management, medical coding reviews, updates, and routine quality-assurance assessments. This stage is important because it enables data issues to be identified and resolved prior to data being used for research, analysis, or regulatory purposes.[10,11]

4. Utilization stage:

The utilization stage involves the use of clinical data after appropriate quality checks are completed. According to the study and any approvals, the data can be used for statistical analysis, clinical research, regulatory submissions, scientific publications, or approved secondary research. This can also reveal shortcomings in previous collection or management processes, which can contribute to improved future studies, data collection practices, and quality assurance procedures. For this reason utilization does not necessarily mean the end of the data lifecycle; the insights gained can inform future planning and continuous improvement.[10]

2.3 Need for Advanced Technologies in CDM:

1. Increase in Volume of Clinical Data:

The management of data in clinical trials is becoming increasingly challenging due to the significant growth in the number of data points collected per trial. For example, Phase III trials are now averaging 3.6 million data points, compared to three times this number collected by late-stage trials a decade ago. Table -1 shows the dramatic increase in data volume, indicating the critical need for advanced technologies to manage and analyze such vast amounts of information efficiently.[12]

Typical Phase III Pivotal Trial (mean for all TAs)	2001-2005	2011-2015	2015-2020	2025-2026 (Recent benchmark study)
Total Data Points Collected	494,236	929,203	3,560,201	≈5,957,122
Proportion of Data "Non-Core" (supporting miscellaneous endpoints)	18%	32%	34%	16.2%
Number of Data Collection Solutions and Applications Used	2	4	4	Not reported

Table 1: (Data adapted from Rising protocol design complexity is driving rapid growth in clinical trial data volume, by Tufts Center for the Study of Drug Development, 2021.)

2.Multiple Data Sources :

The current clinical trials involve the collection of data from a wide range of sources, such as EDC platforms, EHRs, laboratories, patient-reported outcomes, imaging, and various digital health tools, including wearables and Apps. One of the major challenges in clinical data management is integrating these diverse data points, as discrepancies in format, structure, and standards can undermine interoperability and data quality. As your review points out, contemporary EDC systems are particularly ill-suited to addressing this challenge of multi-source data integration, threatening the overall integrity of trial data.[7] Furthermore, eSource technologies enable clinical data to be collected from numerous sources, such as EHRs, mobile Apps, wearables and sensors, laboratory results, imaging studies, ECGs, and other electronic records, which are directly sent into clinical research repositories or data warehouses.[13] The expanding wealth of diverse, heterogeneous data necessitates integration and harmonization as essential requirements for AI/ML applications. Machine-learning approaches have been explored to harmonize health data at syntactic, structural, and semantic levels, which has the potential to reduce the manual effort required to map data from various sources into standardized formats.[14]

3.Data Quality Issues:

Clinical data are often characterized by a lack of completeness, plausibility, concordance, security, currency, and interoperability. These data quality dimensions must be assessed throughout the clinical data quality–management life cycle, involving planning, construction, operation, and utilization. Problems such as missing or inconsistent entries, unreliable values, varying data structures or standards, security gaps, delayed data access, and difficulty linking data between systems can undermine data quality and utility. Therefore, implementing systematic data-quality assessments at every stage of the life cycle is crucial to maintaining accuracy, reliability, reproducibility, and validity. The review also underscores the importance of standardized quality-control practices and automation for effective clinical data quality assessment and management.[10]

4.Manual Data-Management Limitations:

Conventional CDM includes a range of activities, such as CRF review, data entry, data validation, discrepancy management, medical coding, data extraction, and database lock. When these activities require manual review, the increasing volume of data can lead to substantial workloads, risking consistency and human error. Previous studies have demonstrated that complete automation is absent in various aspects of clinical data management. Therefore, advanced technologies are capable of automating repetitive tasks and focusing CDM professionals on complex cases and decisions that require clinical or contextual judgment.[3,15]

5.Need for Faster and More Accurate Data Processing:

Clinical data must be processed efficiently to ensure timely database locking for statistical analysis in clinical trials. Data cleaning, validation, reconciliation, query handling, and final review must be completed before database lock. Timely and integrated data reviews support rapid progress toward database lock while maintaining data completeness and quality.[16]

AI and ML can enhance clinical data management by facilitating tasks such as data cleaning, medical coding, query creation, and detecting potential data issues. The current research highlights the importance of human-in-the-loop (HITL) approaches whereby AI performs preliminary processing or flagging but leaves decision-making and final review to qualified personnel. This approach employs automation to improve efficiency while ensuring human oversight of data quality and critical CDM judgments.[17]

2.4 Emergence of AI and ML in Clinical Research:

➢ **Basic Concept of Artificial Intelligence (AI):** Artificial Intelligence (AI) represents a computer-based approach that enables machines to perform tasks that require human-like abilities, including sensing, reasoning, learning, and decision-making. The development of AI has been facilitated by advances in computing power, increased digital data, and improved algorithms and software. These techniques are increasingly being applied in healthcare in both research and clinical settings. Automated approaches can help mitigate some of the challenging elements of designing, executing, and analyzing large-scale clinical trials. The growing complexity of clinical trials and the availability of larger clinical datasets present expanding opportunities for leveraging AI and other data science techniques in clinical research.[18]

➢ **Basic Concept of Machine Learning (ML):** Machine learning (ML) refers to a set of computational methodologies that are leveraged to achieve artificial intelligence (AI) and to enhance clinical research through learning from patient data. In this context, raw clinical inputs such as medical images, physiological signals, and similar data are transformed into numerical forms that algorithms can process. The ML model explores patterns or relationships in this data and applies them to make predictions or generate other useful outputs.[19]

•Key Concepts and Workflow of Machine Learning:

- The data gathered from patients are pre-processed and encoded into mathematical representations to be computed.[20]
- The model is trained to learn the underlying patterns or mappings between inputs and desired outputs.[19,21]
- Once trained, the model applies these learned patterns to new data to make predictions or derive insights.[19,21]

•Main ML approaches/ Paradigms:

1) Supervised learning: The model is trained using labeled examples (known results). It improves by minimizing the gap between its predictions and the ground-truth labels.[19]

2) Self-supervised learning: The model learns structure directly from unlabeled data, uncovering internal patterns without requiring explicit labels for each sample.[22]

3) Reinforcement learning: The model learns by taking actions in an environment and adapting its behavior based on the rewards or penalties it receives for those actions.[23]

4) Generative models: These models can generate new data samples or produce enhanced versions of existing data (e.g. improved images or augmented records).[24]

•Choosing an ML method:

The appropriate approach is determined by :

-The specific research or clinical task (e.g. diagnosis, prognosis, image enhancement).

-The type and amount of data available (labeled vs. unlabeled data, image vs. tabular data).

-Computational Resources and Infrastructure

-The degree of interpretation and transparency needed in model decisions.[21,25]

➢ **Difference Between Conventional and AI Data-Processing Methods:**

Aspects	Conventional Data Processing	AI-Based Data Processing	References
Basic principle	Adhere to a set of predetermined instructions or procedures.	Uses computational techniques to accomplish tasks such as perception, reasoning, learning, and decision making.	Conventional: [27,28] AI: [26,27]

Decision logic	Uses well-defined logical expressions for example IF – THEN conditions.	AI systems can use mathematical or computational frameworks to help with decision-making, and learning-based systems can find meaningful patterns in data.	Conventional: [28] AI: [26,27]
Source of knowledge	Such rules can be the knowledge that experts tell us.	One branch of AI is machine learning, which finds relationships in data, instead of relying on manually written instructions.	Conventional: [28] AI/ML: [26]
Learning from data	It does not find new patterns from data, automatically.	The more examples they train on , the better they can get at making accurate predictions .	Conventional: [27] AI/ML: [26]
Processing of complex data	The decisions are taken based on the logic built in the system.	AI is able to handle difficult clinical and research applications, especially when dealing with large and complex datasets.	Conventional: [28] AI/ML: [26,27]
Programming requirement	The rules and reasoning to be applied must be pre-entered.	Depending on the approach, direct programming is necessary or not, as traditional AI can be instructed by specific rules, whereas ML learns the patterns of the model during the training process.	Conventional: [28] AI/ML: [26]
Output	It checks for the conditions and gives results.	Artificial intelligence can be used for prediction, classification, logical analysis, and decision support.	Conventional: [28] AI/ML: [26,27]
Adaptability / learning		The performance of ML is evaluated in terms of training the model on one part of the	Conventional: [28]

	Changes usually involve editing of the underlying rules or system.	data and testing it on unseen examples.	AI/ML: [3]

3] Clinical Data Management Work flow:

3.1 Major Stages:

1. Study database design:

The clinical database is designed to reflect the study protocol and all the pertinent fields of data to be captured. This includes protocol-specific case report forms (CRFs)/ electronic case report forms (eCRFs) fields and variables, database design and program- or manual edit checks, and other applicable data management elements. The data management plan describes the database design and related data management specifications. [29,30]

2. Data Collection:

Clinical data are collected by adhering to the study protocol using the CRFs/eCRFs and electronic data-capture (EDC) systems. The type of data collected using the EDC systems includes demographic data, medical history, concomitant medications, study visits, laboratory tests, clinical procedures, and other variables specified in the protocol. [29,31] EDC systems can be used to capture and manage data from all sites involved with the trial.[32]

3. Data Integration:

Clinical trial data can originate from multiple sources. In general, data integration refers to the process by which information is incorporated into a common database by combining data from different systems or sources so that they can be used consistently [32]. Recent studies emphasize the need for data integration involving the combination of disparate data sources and the use of data transformation/data extraction, transformation, and loading (ETL). [33]

4. Data Validation:

This step involves checking if the data meet specified criteria and identifying missing, inconsistent, invalid, or out-of-range values. The process can involve automated edit checks and manual checks to identify potential issues and improve data quality. [29,30]

5. Data Cleaning:

Data cleaning focuses on finding and correcting errors, inconsistencies, missing values, and other data quality issues[10,29]. Contemporary clinical data management (CDM) practices are emphasizing quality control, programmed checks and queries instead of relying primarily on data cleaning.[34]

6. Query Management:

Queries are raised to resolve inconsistencies and discrepancies in the data. Queries are reviewed and responded to by the appropriate study personnel and then closed after verification and review. Query management is a significant aspect of clinical-trial data management. [29,34]

7. Medical Coding:

Medical coding involves the use of standard medical terminologies to encode the narrative information in the CRFs/eCRFs. This process can apply to adverse events, medical history, concomitant medications, and other clinical data items. [29,30]

8. Data Reconciliation:

Data reconciliation involves reconciling discrepancies that may arise while capturing or managing data from different sources. It can involve reconciling laboratory data, safety data, imaging data, and other vendor-supplied data with those in the clinical database.[32]Data reconciliation helps ensure the veracity of the study database. [4]

9. Database Lock: Database lock occurs after the completion of the data capture, cleaning, coding, reconciliation, validation, and query resolution processes. After database lock, no further changes to the database can be made, and the data are ready for statistical analysis.[29,32]

10. Data Export/reporting: Following database lock, the final data set can be exported or transferred to another location for statistical analysis and reporting. Overall, clinical data-management processes contribute to establishing a reliable data set for analysis and reporting. [29,33]

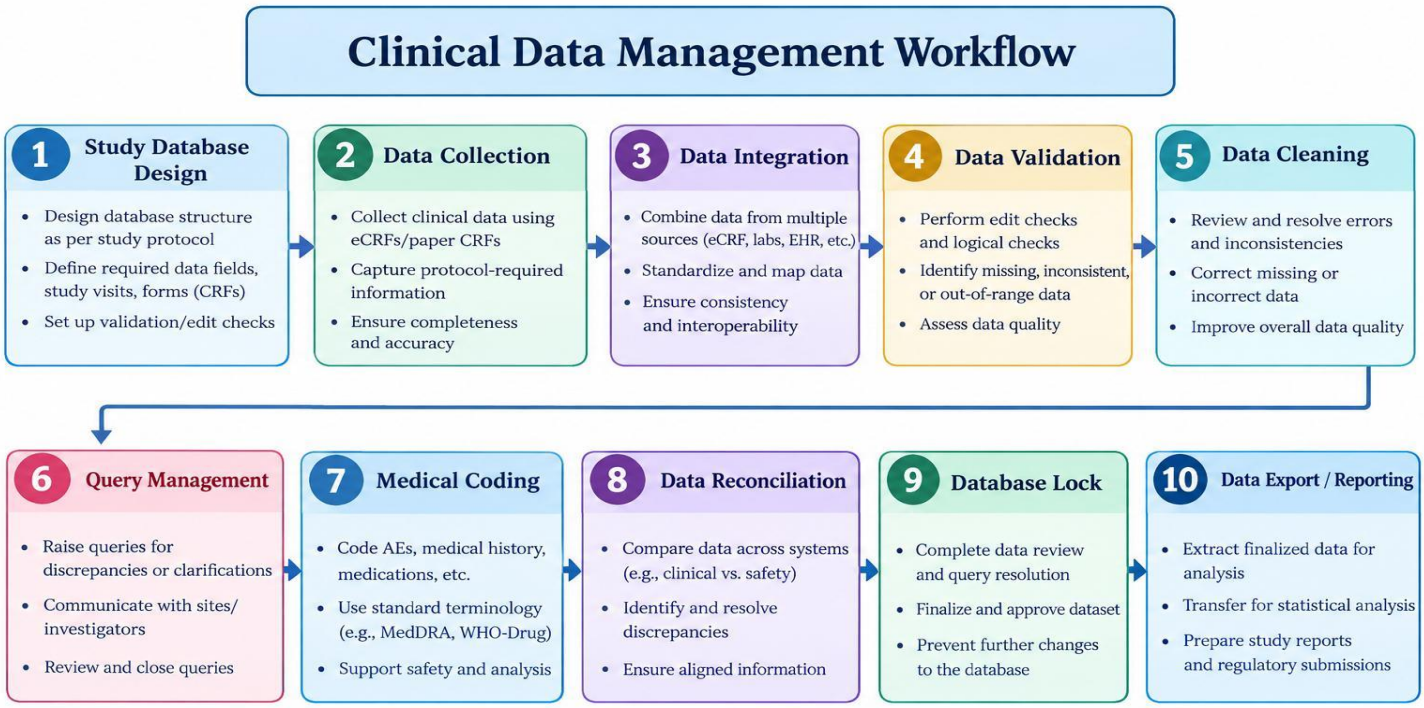


Figure 3 : Clinical Data Management Work flow

3.2 Traditional vs. AI/ML-Enabled Clinical Data Management:

Aspects	Traditional CDM	AI/ML-enabled CDM	Reference
Manual Processes	Traditional clinical data management is a very manual process. It involves data entry, quality review, identifying inconsistencies and raising queries. Instead of simply using modern automation, data teams often manually dig through clinical records, frequently using spreadsheets to find problems and ask follow-up questions.	Artificial intelligence can help clean the data and generate queries and minimize the need for manual review without losing the human touch where it is needed.	[17,25,29]

Automated Processes	The traditional approach to managing clinical data includes manual data review, data cleaning, managing discrepancies, and programd validation checks. Traditional EDC and CDM systems can automate some of these processes with predefined rules.	Systems enabled with AI/ML can identify potential discrepancies, predict data queries and auto-generate the query text.	[17,25,29]
Human AI-Collaboration	Clinical data-management staff review data, identify discrepancies and manage query resolution. In classic workflows, these tasks can be manual checks and fixed criteria for validation.	In human-in-the-loop workflows, AI is used for initial processing or detection, but the final decision is still made by human personnel.	[17,25,29]

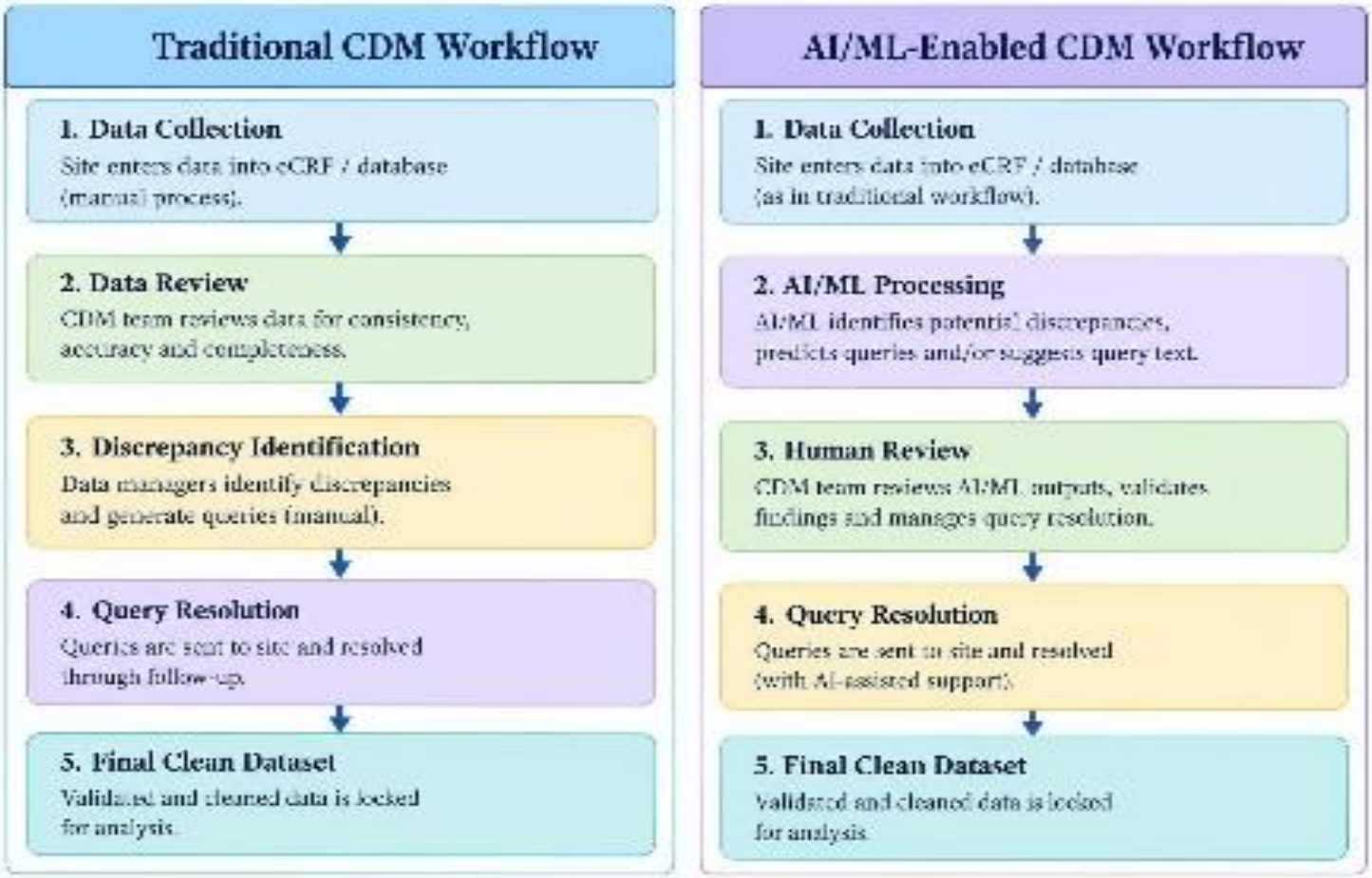


Figure 4 : Traditional Clinical Data Management workflow VS. AL/ML-Enabled Clinical Data Management workflow

4]Artificial Intelligence and Machine Learning Technologies Used in CDM:

Artificial Intelligence (AI) and Machine Learning (ML) are transforming Clinical Data Management (CDM) into an automated Clinical Data Science (CDS) by enabling it to perform rule-based processing of data and predictions on clinical development data. Their application represents a strategic decision-support layer, a "force-multiplier," that executes the processes of data gathering, cleansing, anomaly detection, and validation in real-world settings[7,33,35]

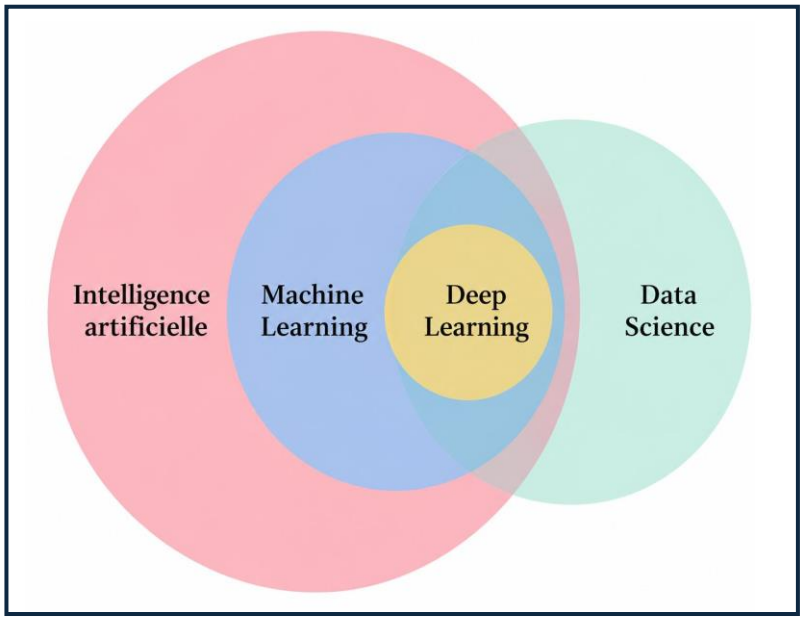


Figure 5 : Artificial Intelligence and Machine Learning Technologies Used in CDM:

4.1 Artificial Intelligence (AI):

Artificial Intelligence (AI) denotes the capability of computer systems to execute tasks that typically necessitate human intelligence, including learning, reasoning, making decisions, recognizing patterns, and solving problems. In Clinical Data Management (CDM), AI can analyze extensive amounts of clinical trial data and aid data managers in detecting errors, inconsistencies, missing data, and unusual patterns.

The fundamental AI framework employed in CDM generally comprises data input, data processing, AI/ML algorithms, decision-making or predictions, and output/feedback. Clinical data sourced from Electronic Data Capture (EDC) systems, Electronic Health Records (EHRs), laboratories, imaging systems, and wearable devices can be processed through AI-supported systems.[7,36]

4.2 Machine Learning (ML):

Machine Learning (ML) is a branch of AI that enables computer systems to learn patterns from data and make predictions or decisions without being explicitly programmed for every individual task. ML is particularly useful in CDM because clinical trials generate large and complex datasets.

➢ **Types of Machine Learning:**

1. Supervised Learning: The algorithm learns from labelled datasets in which the correct outcome is already known. It can be used for predicting data errors, classifying queries, and identifying potentially incorrect clinical data.

2. Unsupervised Learning: The algorithm identifies hidden patterns or groups within data without predefined labels. It can be useful for detecting unusual data patterns, patient-data clusters, and potential anomalies.

3. Semi-Supervised Learning: This approach combines a small amount of labelled data with a large amount of unlabelled data. It can be useful when manually labelled clinical datasets are limited.

4. Reinforcement Learning: The algorithm learns through interaction with an environment by receiving rewards or penalties for its actions. Its potential applications include optimization of decision-support and adaptive data-management processes.

ML can improve automated data review, anomaly detection, query prioritization, missing-data assessment, and predictive data-quality monitoring.

4.3 Deep Learning:

Deep Learning is a sophisticated subfield of Machine Learning that employs artificial neural networks composed of multiple layers to identify intricate patterns within large datasets. In contrast to conventional ML methods, deep-learning models are capable of learning key features on their own from raw or barely processed data.

Within clinical data management, deep learning can be utilized for complex, high-volume clinical datasets sourced from EDC systems, EHRs, laboratory systems, medical imaging, and wearable devices. It can assist with anomaly detection, pattern recognition, forecasting data-quality issues, and uncovering intricate relationships among clinical variables.

Deep neural networks prove especially valuable when datasets include a large number of variables or unstructured information. Yet deep-learning models might demand substantial volumes of high-quality training data and can be hard to interpret, which raises challenges concerning explain ability and validation.[7,36]

4.4 Natural Language Processing (NLP):

Natural Language Processing (NLP) is an area of AI which enables computers to comprehend, interpret, process, and produce human language. NLP is exceedingly relevant to CDM, because a significant amount of clinical information exists in unstructured text, for example, clinical notes, adverse-event descriptions, medical histories, laboratory comments.

NLP has the ability to automatically extract relevant information from clinical notes, transforming unstructured information into structured data. It can also help with medical coding, terminology mapping, adverse-event identification, and classification of clinical information.

For instance, systems based upon NLP can identify adverse-event descriptions and assist with mapping to standardized medical terminology such as MedDRA. NLP can further aid with medication terminology and detect inconsistencies between free-text information and structured database fields.

Thus, NLP can reduce manually review demands and improve efficiency in processing large amounts of clinical text.[37,38,39,40]

4.5 Generative AI and Large Language Models (LLMs):

Generative Artificial Intelligence (Generative AI) refers to AI systems that can generate new content, such as text, summaries, responses, and structured information, according to the data it is given and the instructions it is asked for. Large Language Models (LLMs) are sophisticated AI models that are trained on vast amounts of text and can understand and generate natural language.

Within Clinical Data Management, Generative AI and LLMs can have applications in automated documentation, summarization of clinical data, drafting of query, assisting with review of data, and preparing SOP support materials.

Generative AI can assist CDM professionals with summary of large datasets, drafting of data queries, detection of potentially important discrepancy, and interacting with clinical databases via natural-language instruction. It can also function as an intelligent CDM assistant for routine documentation and review of data.

However, Generative AI can generate inaccurate or fabricated information, commonly known as AI hallucinations. Thus, its usage in CDM should be accompanied by human review, appropriate validation, data-

security measures, audit trails, and regulatory compliance. AI should mainly function as decision-support tool instead of completely replacing qualified clinical data-management professionals.[7,36,38]

5. Applications of AI and ML in Clinical Data Management:

AI and ML are becoming more widely used in Clinical Data Management to enhance the accuracy, consistency, efficiency, and speed of clinical trial data processing. Traditional CDM activities frequently require repetitive manual work, including data cleaning, validation, query handling, reconciliation, and medical coding. AI and ML can handle many of these tasks, enabling clinical data managers to concentrate on complex discrepancies, oversight, and decision-making. The primary applications are outlined below.

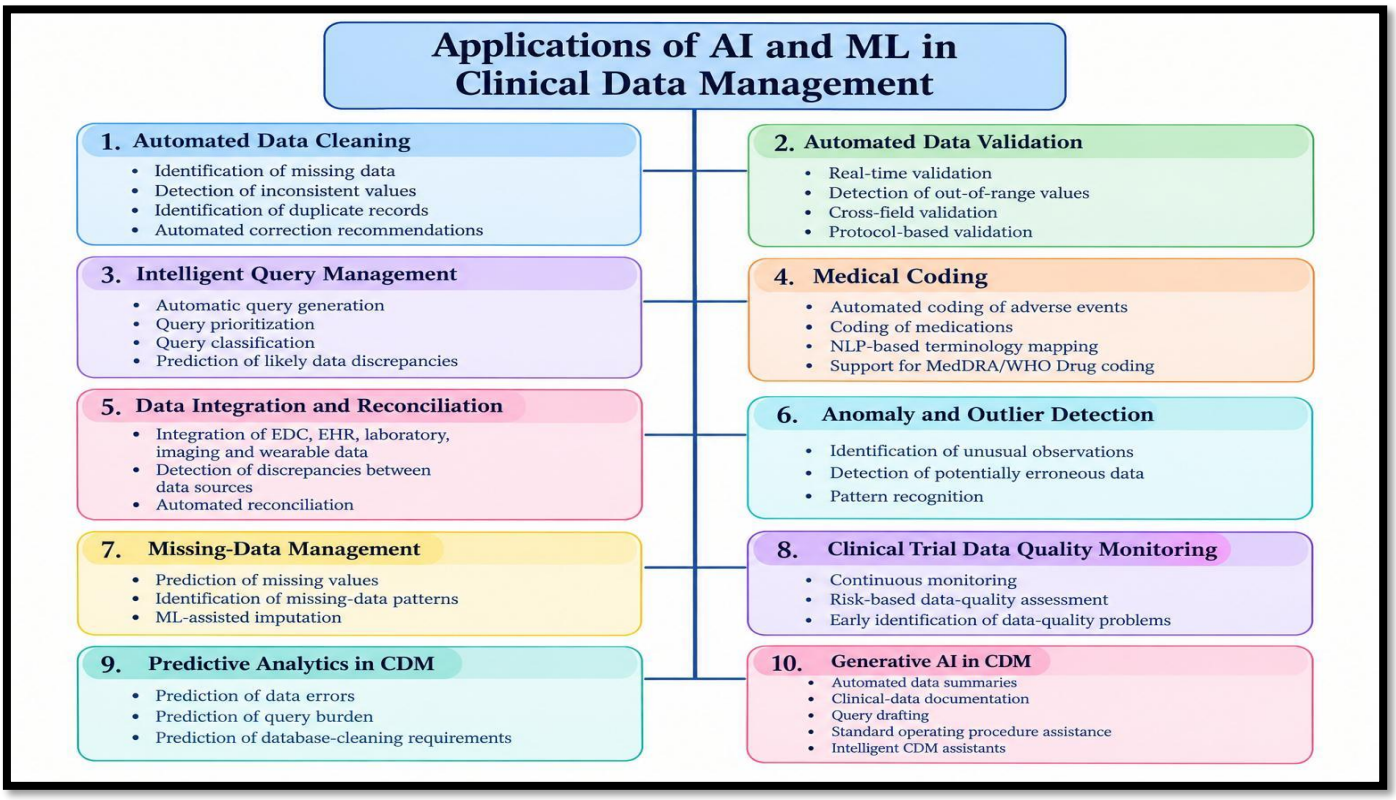


Figure 6 : Applications of AI and ML in Clinical Data Management.

5.1 Automated Data Cleaning:

Automated data cleansing is a critical use of AI and ML in Clinical Data Management (CDM). Clinical trials produce substantial amounts of data from sources such as electronic data capture (EDC) systems, electronic health records (EHRs), labs, imaging technologies, patient-reported outcomes, and more. Reviewing these datasets manually can be labor-intensive and may lead to inconsistent detection of errors.

AI and ML techniques can detect missing, inconsistent, duplicate, and potentially faulty data by learning from established rules and historical data patterns.

Identification of missing data: AI systems have the capability to automatically detect missing values in key clinical variables, including laboratory results, treatment dates, information about adverse events, and visit assessments.

Detection of inconsistent values: Machine learning models are capable of detecting values that do not align with other variables or with typical clinical patterns. For instance, a combination of treatment dates or lab results that are not possible can be highlighted for further examination.

Identification of duplicate records: Automated algorithms are capable of comparing patient identifiers, visit details, dates, and other relevant attributes to identify possible duplicates in patient records.

Automated correction recommendations: Instead of altering clinical data directly, AI systems can suggest potential corrections using established rules, past trends, or connections between different variables.The final decision on these corrections can still be made by qualified human experts and the study sponsor.
This approach helps decrease the amount of manual work involved while promoting more consistent and reliable data-quality procedures.[41]

5.2 Automated Data Validation:

Data validation ensures that the information collected during clinical trials meets specific standards, follows the agreed-upon study plans, and adheres to the rules for managing data. Artificial intelligence and machine learning can enhance standard validation methods by constantly checking new data as it comes in and spotting any inconsistencies or unexpected patterns.
Real-time validation: Real-time validation can take place as data are input into an EDC system.For instance, an algorithm might instantly detect an age value that falls outside the allowed range for the study or a date that does not align with another recorded event.
Detection of out-of-range values: Detection of laboratory results, vital signs, demographic data, or other measurements that fall outside established or statistically expected limits.
Cross-field validation: Examination of relationships between multiple fields. For example, the system may identify inconsistencies between treatment dates, visit dates, and adverse-event dates.
Protocol-based validation: Automated checking of data against study-specific inclusion criteria, visit schedules, treatment requirements, and protocol-defined rules.
Combining conventional validation rules with ML-based pattern recognition may allow CDM systems to identify both predefined errors and less obvious anomalies.[42]

5.3 Intelligent Query Management:

Query management is an important part of clinical data cleaning. In the past, data managers would look for inconsistencies and manually create queries for clinical sites.AI and machine learning can help with this task by finding possible inconsistencies and supporting the creation and prioritization of queries.
AI systems can examine past queries and how they were resolved to spot repeating issues.
Using these insights, they can spot records that might have errors even before a human notices a problem.
Automatic query generation: AI can draft queries describing a detected discrepancy using standardized and understandable language.
Query prioritization: Queries can be prioritized according to factors such as clinical importance, patient safety implications, database-cleaning timelines, or probability of being a true discrepancy.
Query classification: Queries can be categorized according to their type, such as missing data, inconsistent dates, laboratory abnormalities, coding issues, or protocol-related discrepancies.
Prediction of likely data discrepancies: ML models can identify records with a higher probability of requiring future queries based on historical data patterns.
Human review remains important because automatically generated queries should be checked for accuracy, relevance, and appropriate wording before being communicated to clinical sites.[43,44]

5.4 Medical Coding:

Medical coding is the process of translating clinical information into standardized terms.This is especially crucial when dealing with adverse events, medications, medical history, and other clinical details.
Natural Language Processing, which is a part of artificial intelligence focused on understanding and processing human language, can aid in automated or semi-automated coding.
It helps by recognizing clinical concepts in free-text notes and linking them to standardized terminology.

Automated coding of adverse events: Natural language processing models can identify the clinical concept in an adverse event description and recommend a suitable standardized term.
Coding of medications: Artificial intelligence can recognize drug names, including abbreviations, brand names, and misspellings, and link them to standardized medication terminology.
NLP-based terminology mapping: Free-text clinical notes can be translated into standardized medical concepts using natural language processing.
Support for MedDRA and WHO Drug coding: AI-assisted systems can suggest appropriate terminology from reference sources like the Medical Dictionary for Regulatory Activities (MedDRA) and the WHO Drug Dictionary.
AI-based coding can enhance consistency and decrease the need for repetitive manual tasks.
However, terms that are unclear or involve complex clinical situations may still require review by trained medical coders to ensure that the chosen terminology accurately reflects the original data.[45]

5.5 Data Integration and Reconciliation:

Modern clinical trials often incorporate various data sources such as EDC systems, EHRs, central laboratories, imaging platforms, wearable devices, electronic patient-reported outcomes, and other specialized systems. Combining and aligning data from these different sources presents a major challenge in data management and cleansing.
AI and machine learning can help in finding connections between records from different systems and in spotting inconsistencies between them.
For instance, an algorithm can compare laboratory results from a central lab with the corresponding values entered in the EDC system.
It can also detect differences in patient identifiers, dates, treatment details, or clinical measurements.

- Integration of heterogeneous data sources
- Detection of discrepancies between systems
- Automated record matching
- Automated reconciliation of corresponding records
- Identification of potentially missing data transmissions

Such approaches can reduce the amount of manual reconciliation required while improving the consistency of integrated clinical datasets.[46]

5.6 Anomaly and Outlier Detection:

Anomaly detection involves identifying observations that differ substantially from expected patterns. Unlike conventional validation rules, which generally identify predefined errors, ML algorithms can detect unusual patterns that may not have been anticipated during database design.
For example, an algorithm may identify a laboratory value that is technically within an acceptable range but is highly unusual compared with the patient's previous measurements or with other observations from the same study.
Identification of unusual observations: AI and ML algorithms can identify clinical data that significantly differ from expected patterns. They can detect unusual laboratory results, vital signs, treatment responses, or measurements across visits. These observations can be flagged for further investigation, helping clinical data managers identify potential errors while distinguishing them from genuine clinical findings.
Detection of potentially erroneous clinical data: AI and ML algorithms can identify clinical data that may contain errors by analyzing patterns, relationships, and historical records. They can detect inconsistent laboratory results, incorrect dates, implausible measurements, and mismatched patient information. Flagging such records enables clinical data managers to investigate and resolve potential errors efficiently.

Recognition of unusual temporal patterns: AI and ML algorithms can analyze clinical data over time to identify unusual changes or trends. They can detect unexpected variations in laboratory results, vital signs, treatment responses, or patient visits. Identifying these temporal patterns helps data managers recognize potential errors, inconsistencies, or clinically relevant events requiring further investigation.

Pattern recognition across patients or visits:AI and ML algorithms can compare clinical data across multiple patients and study visits to identify recurring patterns, similarities, or differences. This helps detect inconsistencies, unusual trends, and potential data-quality issues. Such analysis enables clinical data managers to identify problems efficiently and focus review on records requiring further investigation.

Identification of unexpected relationships between variables: AI and ML algorithms can analyze relationships among different clinical variables to identify unexpected associations or inconsistencies. For example, they may detect unusual relationships between laboratory results, treatment exposure, vital signs, or adverse events. This helps clinical data managers identify potential data errors and investigate clinically relevant patterns.[47,48]

5.7 Missing-Data Management :

Missing data is a prevalent occurrence in clinical trials, and it can arise from various factors such as missed visits, incomplete assessments, technical difficulties, withdrawal from the study, or other circumstances. CDM teams can benefit from the integration of AI and machine learning (ML) to comprehend and effectively manage missing-data patterns. ML algorithms have the ability to analyze both historical and contextual data in order to identify variables or patient visits that have a higher likelihood of missing data.

Prediction of missing values: Models can estimate the likelihood that a particular data point will be missing.

Identification of missing-data patterns: AI can determine whether patient populations, variables, treatment groups, sites, or specific visits are the focus of missingness.

ML-assisted imputation: Machine learning techniques may be able to assist with statistical imputation methods when they are suitable.[49,50]

5.8 Clinical Trial Data Quality Monitoring:

AI and machine learning can change how clinical trial data quality is monitored, moving from occasional manual checks to a more ongoing and risk-focused approach. In traditional methods, data quality is usually assessed at set times. With AI systems, data can be analyzed in real time as it comes in, helping to spot quality issues as they appear.

Continuous monitoring: Continuous monitoring involves the ongoing automatic observation of systems, networks, and data.It helps identify potential security risks, performance issues, and operational problems early, allowing for quick action to address incidents and reduce disruptions throughout your entire digital business environment.

Risk-based data-quality assessment: Risk-based data-quality assessment focuses on validating data by concentrating on the most important assets that could have the biggest impact.This method helps find errors more effectively, reduces risks related to compliance, and supports better decision-making in the areas that truly matter.

Early identification of data-quality problems: Detecting data quality issues early on helps catch errors as they first appear.This forward-thinking method stops incorrect data from moving further along the process, lowers the expenses needed to fix problems later, and provides reliable information for both day-to-day operations and long-term planning.

Site-level pattern analysis: Site-level pattern analysis looks closely at trends, how operations are carried out, and any performance issues at each specific physical location.By focusing on each site individually, this

method finds unique insights related to that location, improves efficiency there, reduces risks, and helps make better, more accurate decisions about how the operations run.[51,52]

5.9 Predictive Analytics in CDM:

Predictive analytics uses data from the past and present to estimate how likely future events or issues are to happen. In the context of Clinical Data Management (CDM), machine learning models can examine past clinical trial data and operational details to forecast possible needs related to data management.

Prediction of data errors: Predicting data errors involves using sophisticated machine learning techniques and analyzing past trends to foresee potential issues before they affect business operations. This proactive method prevents incorrect data from occurring, improves the accuracy of data, and helps avoid expensive problems that arise later in the process.

Prediction of query burden: The ability to predict the workload involves estimating the amount and difficulty of incoming data requests based on past trends. This forward-looking approach helps in efficiently managing system resources, avoiding performance issues, and maintaining smooth operations when there is a high volume of requests.

Prediction of database-cleaning requirements: Predicting when database cleaning is needed helps figure out when data sets will start to get worse or have more mistakes over time. This method helps plan maintenance better, uses resources wisely, stops problems before they happen, and keeps data clean and reliable, which supports the business running smoothly.

Prediction of recurring data-quality issues: Looking at past data, we can expect similar quality problems to keep happening, which shows there are ongoing mistakes and problems that aren't being fixed. This forward-looking approach stops repeated mistakes, stops future mistakes from happening, and keeps all important business data dependable over time.[7,53]

5.10 Generative AI in CDM:

Generative AI is a more recent advancement in managing clinical data. Generative AI is different from regular machine learning systems that mainly focus on categorizing or forecasting results. Instead, generative AI can create things like text, summaries, explanations, and other kinds of content using the information given.

Automated data summaries: Automated data summaries quickly turn big sets of information into easy-to-understand insights by using smart algorithms. This helps businesses make decisions faster, shows important trends in how well they're performing, and saves time for analysts throughout the company.

Clinical-data documentation: Clinical data documentation means carefully writing down a patient's health details, the treatments they receive, and the results of those treatments into their electronic health records. This helps follow rules and regulations, keeps care going smoothly, and makes medical decisions better overall.

Query drafting: Clinical data documentation means writing down accurate information about a patient's health, the treatments they receive, and the results of those treatments in electronic health records. This helps follow rules, keeps care going smoothly, and makes medical decisions better.

Standard operating: A standard operating procedure (SOP) is a detailed list of steps made by a company to guide workers in doing regular or complicated tasks in a clear and efficient way.It helps make sure that everyone does things the same way every time, keeps the work to a high standard, follows the rules, and reduces mistakes.

Intelligent CDM assistants: Smart CDM helpers use smart AI to do regular clinical data management jobs automatically. They make data cleaning faster, help write queries, and find problems in the data, which speeds up the clinical trial process a lot and keeps the data very accurate.

AI and ML are used in CDM throughout the whole process of handling clinical data, starting from entering and checking the data, then cleaning it, assigning codes, matching information, keeping track of data quality, and getting everything ready for the final database lock. Their main contribution isn't just replacing manual tasks but the ability to handle big and complicated data sets all the time, spot patterns, figure out possible issues, and help clinical data management experts make better decisions.

At the same time, carrying out the implementation needs careful focus on data privacy, checking if the algorithms work correctly, making sure the decisions can be explained, avoiding bias, keeping the system secure from cyber threats, following the rules set by regulations, being able to review and check the process, and having people watch over things. So, AI and machine learning are most useful when they help support clinical data management skills, instead of taking over the whole decision-making process by themselves.[39,54]

6. Benefits of AI/ML in Clinical Data Management:

The incorporation of Artificial Intelligence (AI) and Machine Learning (ML) into Clinical Data Management (CDM) has completely transformed the way clinical trial data is processed, validated, and analyzed. Conventional CDM approaches depend largely on manual entry, fixed rule-based validation, and time-consuming query management. Through the adoption of intelligent automation, sponsor organizations and Contract Research Organizations (CROs) can reach unmatched levels of speed, accuracy, and efficiency.

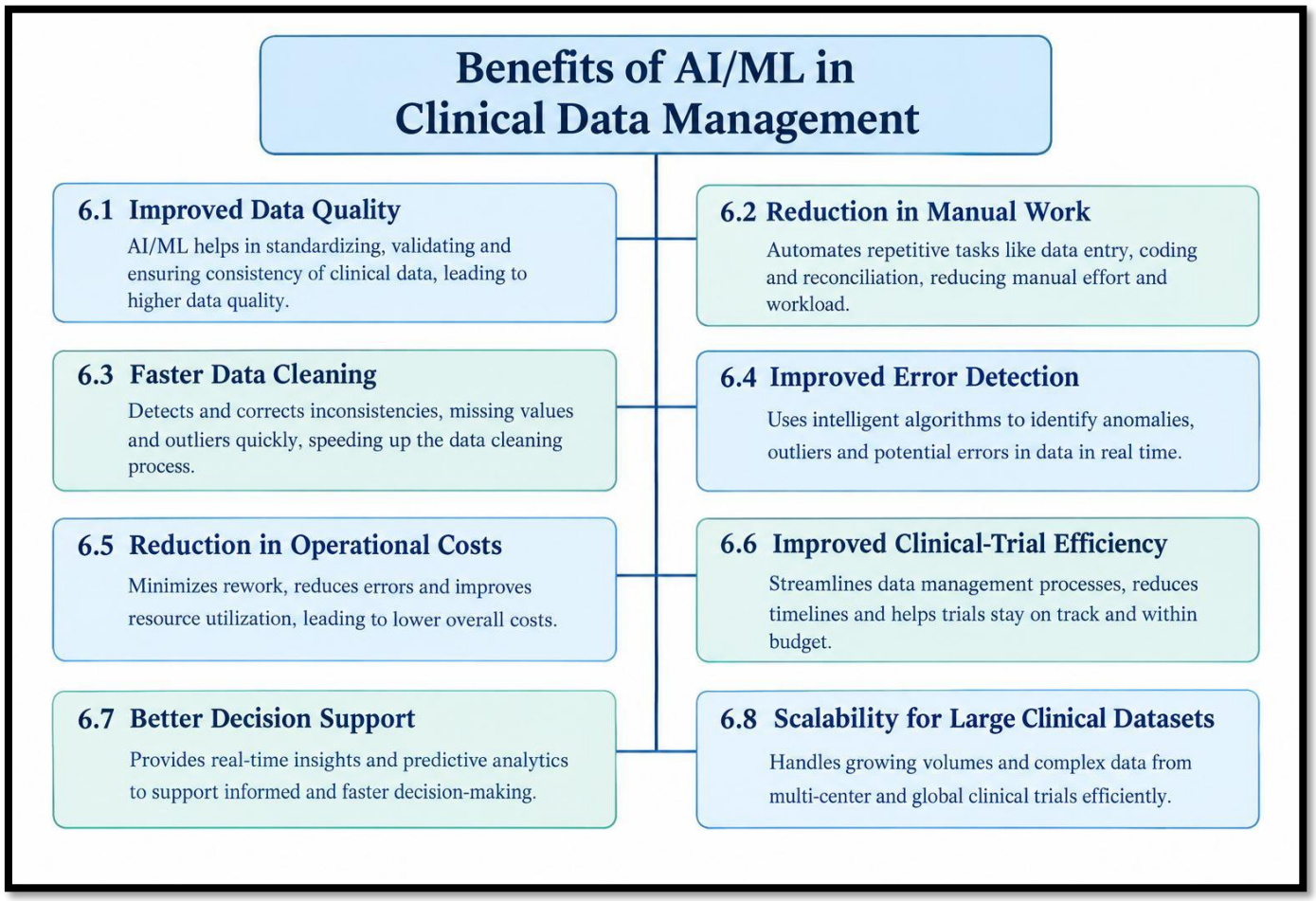


Figure 7: Benefits of AL/ML in Clinical Data Management.

6.1 Improved Data Quality:

AI and ML algorithms continually learn from past clinical trial datasets in order to spot subtle inconsistencies, missing fields, or conflicting patient records. In contrast to inflexible static rules, adaptive machine learning

models strengthen data integrity by detecting complex data patterns across varied sources, such as electronic health records (EHRs), wearable devices, and central laboratory systems.

6.2 Reduction in Manual Work:

Data managers have traditionally devoted countless hours to repetitive data entry checks and cross-source reconciliation. ML-driven automation takes care of routine data handling tasks, allowing clinical data management professionals to concentrate on higher-value risk-based monitoring and strategic trial oversight.

6.3 Faster Data Cleaning:

Data cleaning is frequently the most time-consuming bottleneck in clinical trials. Natural Language Processing (NLP) and predictive algorithms automate the review of unstructured clinical narratives and medical coding, greatly reducing database lock timelines and speeding up study completion.

6.4 Improved Error Detection:

Advanced anomaly detection models can identify outliers and fraudulent data points—such as fabricated patient logs or impossible physiological value combinations—much earlier and more dependably than standard rule-based checks, protecting trial integrity and regulatory compliance.

6.5 Reduction in Operational Costs:

Automating repetitive workflows sharply reduces the labor overhead and human error rates tied to manual query resolution. This decrease in manual man-hours translates directly into lower overall operational expenses for clinical development programs.

6.6 Improved Clinical-Trial Efficiency:

By streamlining data flow and cutting query cycles, AI-enabled CDM shortens the overall drug development lifecycle. Faster data turnaround enables sponsors to make timely, data-driven operational decisions.

6.7 Better Decision Support:

AI equips clinical teams with real-time analytics, predictive dashboards, and actionable insights. This strengthens risk assessment, protocol compliance monitoring, and patient retention strategies across the trial lifecycle.

6.8 Scalability for Large Clinical Datasets:

Contemporary decentralized and hybrid clinical trials produce enormous volumes of high-frequency data from wearable devices, mobile apps, and electronic data capture (EDC) systems. AI/ML architectures scale effortlessly to process these enormous streaming datasets without any loss of performance.[55,56]

7. AI/ML and Clinical Data Standard:

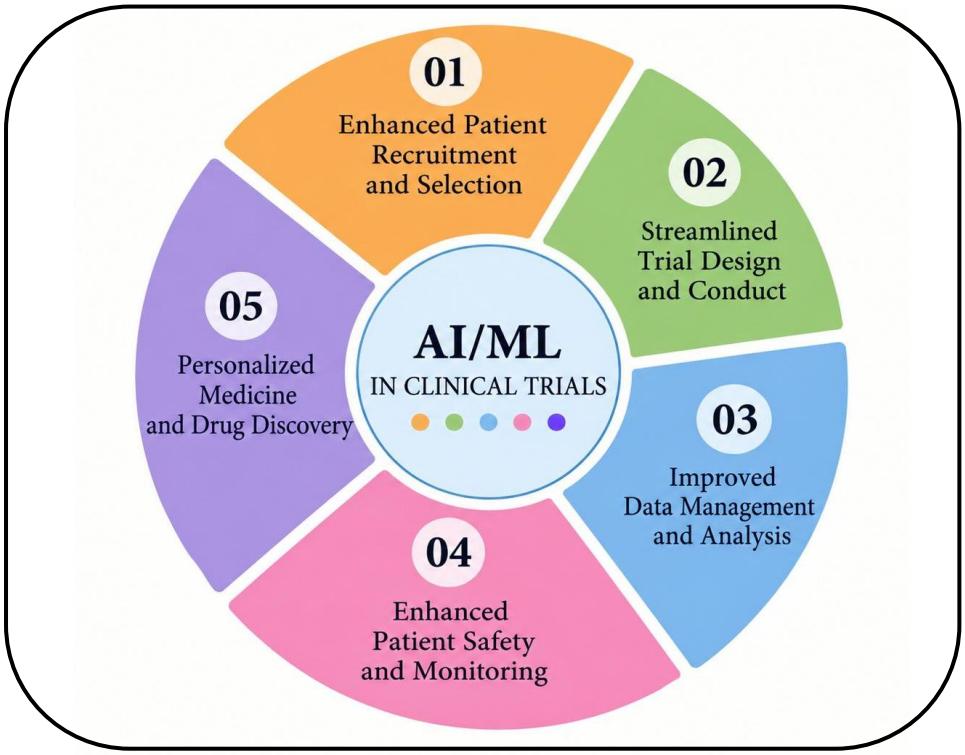


Figure 8 : AI/ML in Clinical Trials.

The Clinical Data Interchange Standards Consortium (CDISC) offers internationally accepted standards for organizing and sharing clinical research data. It is important to standardize data because clinical trials produce large amounts of varied data from different sources, such as electronic data capture systems, laboratories, medical records, and patient-reported outcomes. Keeping data in consistent formats helps improve the ability of different systems to work together, makes data easier to combine, reduces confusion, and supports efficient regulatory submissions and reviews. CDISC standards therefore offer a shared framework for presenting clinical trial information across different studies and organizations**.**

7.1 CDISC Standards

The Study Data Tabulation Model (SDTM) is a CDISC standard used to structure clinical trial data into predefined categories.

7.2 SDTM

It changes collected data into a consistent format, making it easier to understand, compare, and review information from different studies. SDTM organizes observations like demographics, adverse events, lab results, and vital signs into specific categories. This helps ensure that data can be traced back and improves the efficiency of analysis and regulatory processes.

7.3 ADaM

The Analysis Data Model (ADaM) provides standardized structures for creating datasets that are ready for statistical analysis. ADaM datasets are designed to support analysis while allowing clear tracking back to the original source data and SDTM data. By using standardized datasets, repetitive data preparation tasks are minimized, and analysts and programmers can perform analysis more consistently.

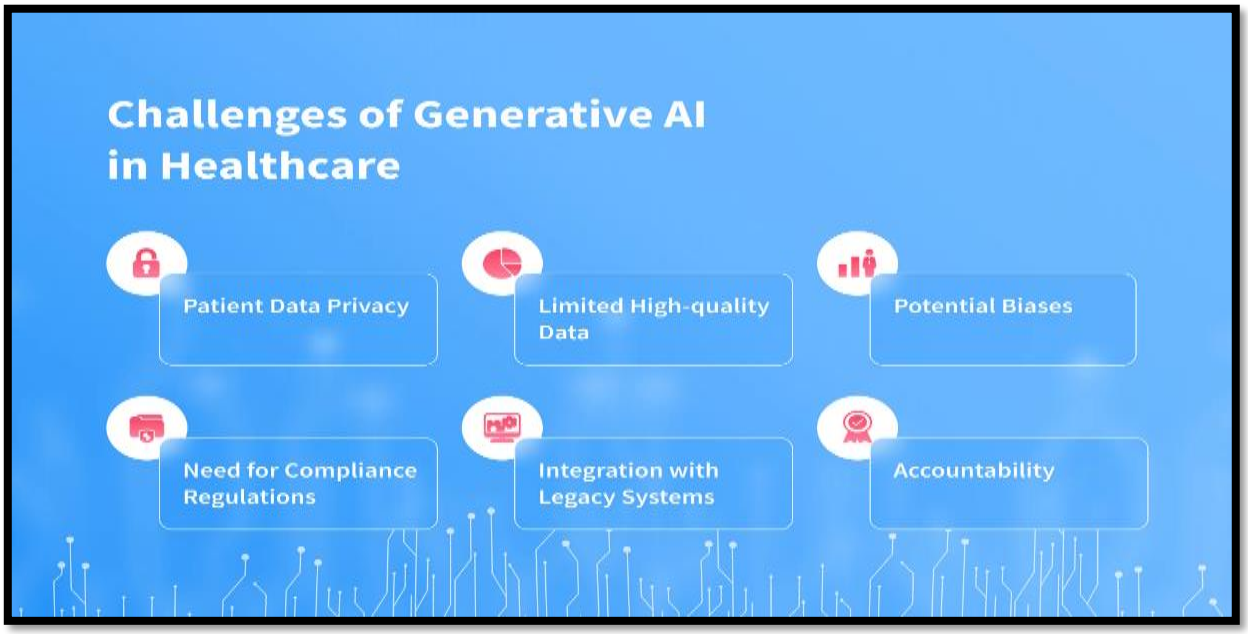
7.4 Controlled Terminologies

Controlled terminology uses standardized terms, codes, and definitions to consistently represent clinical concepts. Examples include standard descriptions for adverse events, lab tests, units of measurement, and medical conditions. Using controlled terminology reduces inconsistencies that may arise from synonyms, abbreviations, spelling variations, or inconsistent coding. It is especially important when integrating data from multiple studies or organizations.

7.5 Role of AI/ML in Standardization

AI and machine learning can help with implementing and maintaining clinical data standards by automating routine data management tasks. Automated mapping models can identify connections between data from source systems and appropriate CDISC domains or variables. AI-assisted data transformation can help convert diverse source data into standardized formats while decreasing the need for manual programming. Machine learning can also help detect issues such as missing data, inconsistent formats, unexpected relationships, and errors in standardization. Additionally, natural language processing (NLP) can assist with terminology matching, such as identifying equivalent clinical terms and suggesting suitable controlled terminology codes. Human review is still important because AI-generated mappings and terminology suggestions need to be validated, especially when clinical meaning is unclear. Overall, AI and machine learning can support CDISC processes by increasing automation, consistency, and scalability, while ensuring proper human oversight.[57,58,59]

8] Challenges and Limitations:



Clinical data management (CDM) can be greatly enhanced by artificial intelligence (AI) and machine learning (ML), but putting these technologies into practice presents enormous operational, ethical, technological, and legal obstacles. To guaranty that AI/ML-supported CDM procedures continue to be precise, safe, transparent, and clinically dependable, these restrictions must be removed.

8.1Data privacy and confidentiality: Sensitive patient data, such as medical histories, test results, genetic data, and personally identifiable information, can be found in clinical databases. Appropriate procedures for data anonymization, access control, consent, and adherence to relevant data-protection laws are necessary when using AI/ML.

8.2 Patient Data Security: Clinical data processing may be done on more platforms and interfaces thanks to AI/ML systems. Inadequate cybersecurity can leave data vulnerable to misuse, illegal access, and breaches. Thus, it is crucial to have secure data storage, encryption, authentication, and ongoing monitoring.

8.3 Data Bias: AI/ML models have the potential to replicate or exacerbate biases found in clinical datasets. Insufficient representation of certain demographic or clinical populations can diminish model performance for those groups, possibly leading to disparities in data quality or decision-support outcomes.

8.4 Poor-Quality Training Data: The effectiveness of an AI/ML model is heavily influenced by the quality of its training dataset. Data issues such as missing values, inconsistent coding, duplicate entries, incorrect data, and differences in collection methods can undermine model reliability. As a result, comprehensive data purification and uniformity are necessary.

8.5 Algorithmic Errors: When AI/ML models come across data that deviates from their training circumstances, they may generate inaccurate recommendations, forecasts, or classifications. If automated data processing errors are not identified through the proper validation processes, they may spread to downstream CDM operations.

8.6 Explain ability and Interpretability: Certain sophisticated machine learning models function as "black boxes," making it challenging for CDM experts to comprehend why a specific result was produced. Inadequate interpretability can undermine confidence and make regulatory review, auditing, validation, and error inquiry more difficult.

8.7 Regulatory Uncertainty: The legal landscape pertaining to AI/ML in clinical research is still developing. Organizations using AI-based CDM solutions may face difficulties due to jurisdictional variations in standards as well as ambiguity surrounding software classification, model modifications, validation, and accountability.

8.8 Integration with Existing CDM Systems: Electronic data capture (EDC), clinical trial management, electronic health record (EHR), and clinical data storage solutions are used by many clinical organizations. Because various systems have different data formats, standards, interfaces, and workflows, integrating AI/ML tools with them can be technically challenging.

8.9 High Implementation Cost: Significant investments in software, cybersecurity, data preparation, system validation, computing infrastructure, and staff training may be necessary for implementation. Institutions or smaller businesses with fewer technical resources may find these expenses especially difficult.

8.10 Need for Skilled Professionals: Professionals who comprehend both clinical data management and AI technologies are necessary for the successful adoption of AI/ML. Successful adoption may be hampered by a lack of employees with knowledge of data science, clinical research, legal requirements, and AI validation.

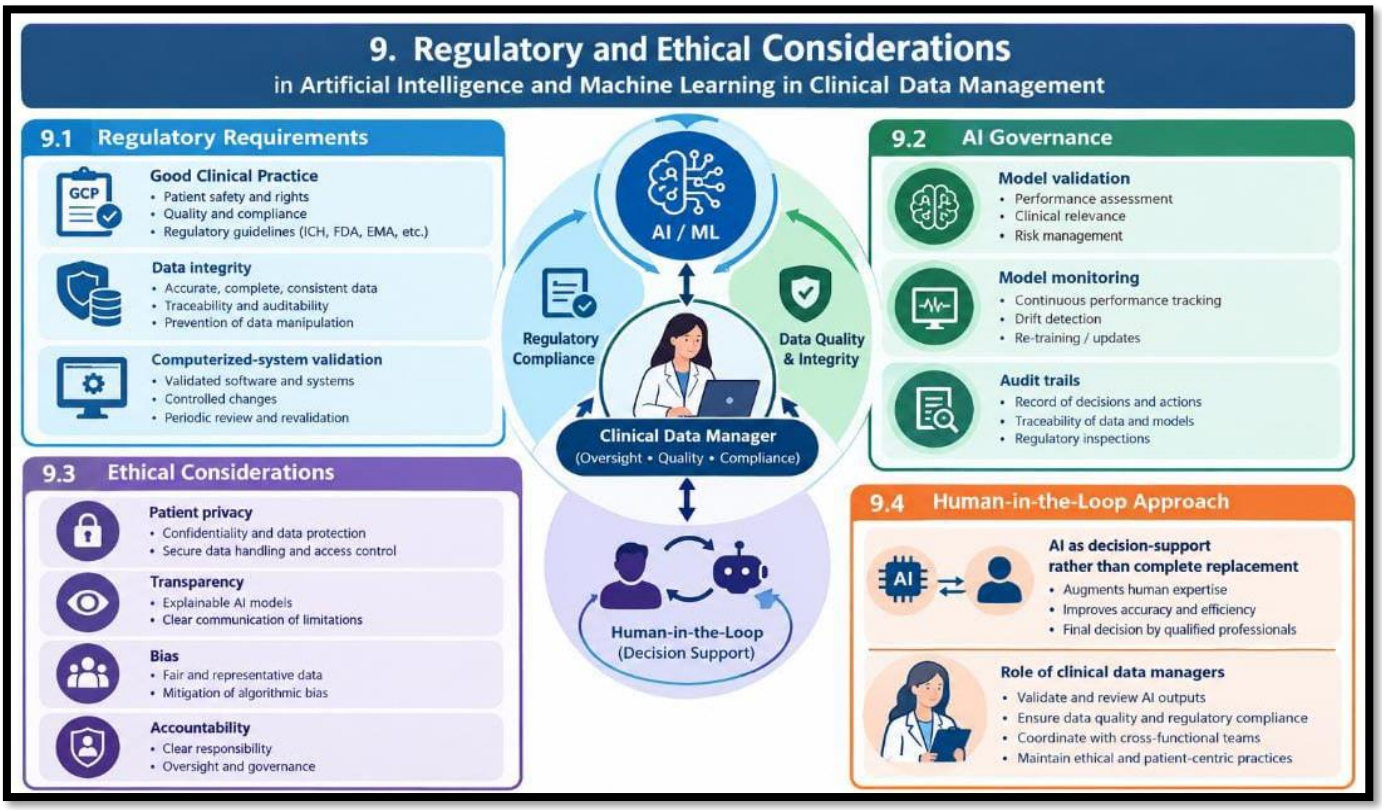
8.11 Human Oversight: In healthcare data management, AI/ML should supplement human judgment rather than totally replace it. Validating automated outputs, looking into anomalies, handling exceptional cases, and making sure automated procedures adhere to study protocols and data-management criteria all still require human inspection.

8.12 Risk of Generative-AI Hallucination: Generative AI systems are capable of producing what are known as "hallucinations," which are plausible but factually inaccurate information. If outputs in CDM are accepted without verification, fake values, erroneous interpretations, or unsubstantiated summaries may jeopardize data quality. As a result, regulated use cases, source verification, validation processes, and suitable human review should all be incorporated into generative-AI applications.

In general, a balance between automation and human monitoring is necessary for the effective integration of AI/ML in CDM. To guaranty that AI/ML applications deliver dependable benefits without jeopardizing patient safety, privacy, or clinical data integrity, strong data governance, cybersecurity, strict validation, transparent model development, regulatory compliance, and ongoing monitoring are crucial.[4,60,61,62]

9. Regulatory and Ethical Considerations:

The integration of Artificial Intelligence (AI) and Machine Learning (ML) into Clinical Data Management (CDM) can improve data review, query management, anomaly detection, coding, data cleaning, and clinical-trial analytics. However, because clinical data directly influence patient safety, trial integrity, and regulatory decision-making, AI/ML systems must operate within established regulatory, quality, ethical, and governance frameworks. Current regulatory thinking emphasizes a risk-based, lifecycle-oriented and human-centric approach to AI/ML. The European Medicines Agency (EMA), for example, recommends that AI/ML use throughout the medicines lifecycle comply with existing legal requirements while addressing ethics, fundamental rights, data quality, bias, and model reliability.



9.1 Regulatory Requirements

Good Clinical Practice (GCP): Good Clinical Practice provides the fundamental framework for conducting clinical trials in a manner that protects the rights, safety, and well-being of participants and ensures that clinical-trial data are reliable. When AI/ML is used for clinical data management, its application should not compromise protocol compliance, participant protection, data reliability, or the ability to reconstruct how results were generated. AI-supported processes such as automated data cleaning, anomaly detection, medical coding, and data review should therefore be appropriately documented, validated, and controlled. Regulatory agencies increasingly emphasize lifecycle management and appropriate oversight for AI/ML applications used in medicine development.

Data Integrity: Data integrity is essential for clinical research because regulatory conclusions depend on trustworthy and traceable data. AI/ML systems should preserve the principles of accurate, complete, consistent, attributable, contemporaneous, original and reliable records. Training and operational datasets should have defined sources, quality controls, appropriate access restrictions, and documented transformations. Data preprocessing, missing-data handling, algorithmic decisions, corrections, and model outputs should be traceable so that the history of clinically relevant information can be reconstructed.

Computerized-System Validation: AI-enabled CDM platforms should be appropriately validated according to their intended use and associated risks. Validation should demonstrate that the system performs consistently and accurately under defined conditions. It may include verification of data inputs, algorithms, interfaces, automated edit checks, data transformations, outputs, security controls, and error-handling mechanisms. Because some ML systems can change or be updated over time, validation should be considered a lifecycle activity rather than a one-time exercise. FDA/IMDRF Good Machine Learning Practice principles similarly emphasize safe, effective, high-quality development and lifecycle management of AI/ML systems.

9.2 AI Governance:

Model Validation: Before implementation in clinical data management, an AI/ML model should undergo appropriate technical and clinical validation. Important aspects include accuracy, precision, sensitivity, specificity where applicable, robustness, reproducibility, generalizability, and performance in the intended population and context. Training, validation, and independent testing datasets should be appropriately separated to reduce risks such as overfitting and data leakage. For high-impact clinical-trial applications, EMA recommends testing with prospectively generated data representative of the intended context and documenting models and data-processing pipelines in a traceable manner.

Model Monitoring: Model performance should continue to be monitored after deployment because changes in patient populations, clinical practices, data sources, or data quality may cause performance degradation. Monitoring can include predefined performance thresholds, periodic quality reviews, evaluation of false-positive and false-negative results, and assessment for data or model drift. A documented procedure should specify when a model requires recalibration, retraining, suspension, or replacement. Regulatory guidance emphasizes monitoring deployed models and managing risks associated with retraining or model changes.

Audit Trails: Audit trails are important for maintaining traceability in computerized clinical-trial systems. They should capture relevant activities such as data entry, modifications, automated processing, model versions, user actions, approvals, and system-generated outputs. Proper audit trails allow clinical data managers and auditors to determine what happened, when it happened, and, where appropriate, who or what system performed the action. This is particularly important when AI automatically identifies discrepancies or modifies data-processing workflows.

9.3 Ethical Considerations:

Patient Privacy: Clinical datasets may contain highly sensitive personal and health information. AI/ML systems should therefore use appropriate privacy and security safeguards, including controlled access, secure storage and transmission, data minimization, de-identification or pseudonymization where appropriate, and appropriate governance of secondary data use. Privacy requirements should be incorporated throughout the AI lifecycle rather than addressed only after model development.

Transparency: AI systems used in clinical data management should provide sufficient information about their purpose, intended use, limitations, inputs, outputs, and performance to enable appropriate human oversight. Highly complex models may be difficult to interpret; therefore, documentation and understandable explanations of AI-supported decisions are important. FDA, Health Canada, and MHRA guidance on ML-enabled medical devices emphasizes transparency and the performance of the human-AI team.

Bias: AI models can reproduce or amplify biases present in training data. Bias may arise from under-representation of particular demographic groups, differences in data quality between sites, historical clinical practices, or inappropriate model design. In clinical data management, biased algorithms could result in unequal identification of data anomalies, inconsistent patient classification, or unreliable predictions. Consequently, datasets should be assessed for representativeness, and model performance should be evaluated

across relevant subgroups. EMA specifically highlights the need for active measures to prevent bias and promote trustworthy AI/ML applications.

Accountability: Clear responsibility should be assigned for AI-supported clinical data processes. AI should not be treated as an independent accountable entity. Sponsors, investigators, clinical data managers, technology providers, and other responsible personnel should have clearly defined roles regarding system validation, monitoring, data quality, security, documentation, and decision-making. Accountability is particularly important when an AI-generated recommendation is incorrect or when an automated process produces an unexpected result.

9.4 Human-in-the-Loop Approach:

AI as decision-support rather than complete replacement: A Human-in-the-Loop (HITL) approach is particularly important when AI/ML is applied to clinical data management. Under this approach, AI functions primarily as a decision-support and automation tool rather than a complete replacement for qualified human professionals. AI can rapidly identify discrepancies, prioritize queries, detect unusual data patterns, classify information, and support data review; however, important decisions should remain subject to appropriate human review.

Role of clinical data managers: Clinical data managers play a central role in this framework. They are responsible for reviewing AI-generated outputs, investigating unusual findings, confirming important data corrections, evaluating model performance, maintaining documentation, and ensuring compliance with the clinical data management plan and applicable regulatory requirements. Human experts can also recognize contextual information that an algorithm may not adequately capture. A human-centric approach is specifically emphasized in current EMA thinking on AI/ML in the medicines lifecycle.

The HITL model can therefore be viewed as a collaborative system in which AI provides speed, scalability, and pattern-recognition capabilities while clinical data professionals provide contextual judgment, oversight, accountability, and regulatory interpretation. This approach helps balance the efficiency benefits of AI with the need for patient protection, data integrity, transparency, and scientifically reliable clinical-trial results.

➢ **Key Points for Clinical Data Management:**

Area	Major Requirement
GCP	Protect participants and maintain reliable clinical-trial data
Data integrity	Ensure data remain accurate, complete, consistent, traceable, and reliable
System validation	Validate AI-enabled computerized systems according to intended use and risk
Model validation	Demonstrate appropriate accuracy, robustness, reproducibility.
Model monitoring	Continuously monitor performance, drift, and retraining risks
Audit trails	Maintain traceability of data changes, user actions, model versions, and outputs
Privacy	Protect patient confidentiality and sensitive clinical information

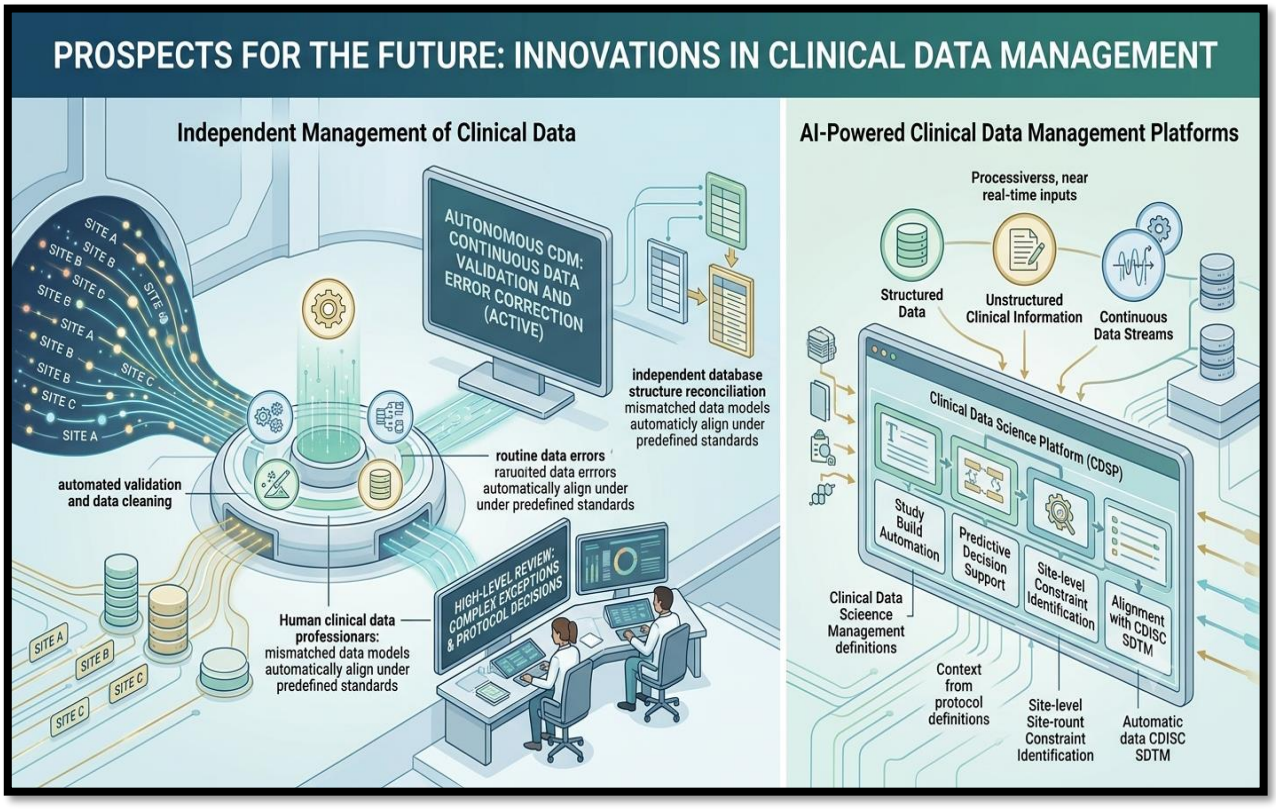
Transparency	Document AI purpose, limitations, inputs, outputs, and performance
Bias control	Identify and mitigate dataset and algorithmic bias
Accountability	Define responsibility for AI-supported processes and decisions
Human-in-the-loop	Keep qualified clinical data professionals involved in important decisions

Table 4 : Key Points for Clinical Data Management

Overall, regulatory and ethical governance should not be considered an additional step after AI implementation; it should be incorporated throughout the entire AI lifecycle—from data acquisition and model development to validation, deployment, monitoring, updating, and retirement. This approach supports the responsible use of AI/ML in Clinical Data Management while maintaining patient privacy, data integrity, scientific validity, and regulatory compliance.[4,36,63,64]

10. Future Perspectives:

Artificial Intelligence (AI) and Machine Learning (ML) are expected to significantly influence the future of Clinical Data Management (CDM). The next generation of clinical data management will move beyond conventional data cleaning and monitoring toward **intelligent, automated, predictive, and continuously monitored systems**.



10.1 Independent Management of Clinical Data:

It is anticipated that highly automated and autonomous data-management settings would be the direction of future progress in clinical data management (CDM). Intelligent CDM systems may continuously gather and analyze data from many clinical sites, carry out context-sensitive validation, and automatically detect and fix routine data errors rather than heavily relying on operator initiation and intervention. Additionally, sophisticated algorithms can apply predetermined data standards and reconcile disparities in database structures without requiring manual programming for each specific transformation.

Clinical data professionals' duties will probably change to include evaluating complex clinical exceptions, crucial data-quality issues, and study protocol decisions as routine data-processing, formatting, validation, and administrative tasks become more automated.

10.2 Clinical Data Management Platforms Powered by AI:

New Clinical Data Science Platforms (CDSPs) are anticipated to go beyond the traditional roles of Electronic Data Capture (EDC) systems. These advanced platforms can act as unified intelligence layers that can handle structured data, unstructured clinical information, and continuously produced data almost in real time. By utilizing flexible and modular software architectures, platforms powered by AI can autonomously assist in configuring studies by deciphering information related to protocols. They can also recognize potential site-level data-management constraints, assist in predictive decision-making, and ensure alignment with existing clinical data standards and frameworks, such as CDISC SDTM.

10.3 Clinical Data Management Assistants Powered by Generative AI:

Generative Artificial Intelligence and specialized Large Language Models (LLMs) can serve as smart aids for professionals managing clinical data. These systems can analyze intricate clinical-trial protocols and help transform protocol requirements into operational elements like electronic Case Report Forms (eCRFs), specifications for edit checks, and documents related to monitoring.

Upcoming AI assistants could enable clinical data managers to engage with extensive and intricate clinical databases through natural-language commands. A data manager might request the system to pinpoint clinical sites exhibiting atypical variations in adverse-event reporting. The system could then deliver pertinent statistical overviews, visual illustrations, and possible data inquiries, thus minimizing the time needed for manual data exploration.

10.4 Monitoring the Quality of Clinical Data in Real Time:

Clinical data management is slowly transitioning from occasional, retrospective data-cleaning tasks to ongoing and real-time quality evaluation. Rather than relying on predetermined review periods, smart monitoring systems can assess incoming patient data immediately upon its availability. Real-time processing allows machine learning systems to detect irregular data-entry patterns, potential protocol deviations, site-level irregularities, and variations in equipment-related information. Prompt detection of these issues can minimize the corrective efforts needed around database lock and enhance the overall efficacy of clinical data management.

10.5 Combining EHR, EDC, Wearables, and Real-World Data:

Future advanced CDM systems are anticipated to operate as cohesive data ecosystems that can consolidate information from various clinical and digital origins. These can encompass conventional EDC data, Electronic Health Records (EHRs), wearable gadgets, Digital Health Technologies (DHTs), electronic patient-reported outcomes, registries, and additional sources of RWD.

Interoperability technologies, like standards such as HL7 FHIR, can enable the sharing of information across various healthcare and clinical research systems. Data pipelines powered by AI can enhance automated data mapping, eliminate duplicate records, resolve inconsistencies, and extract valuable insights from massive amounts of constantly generated data.

10.6 Transparent and Reliable Artificial Intelligence:

With the growing involvement of AI and ML systems in clinical data management, the significance of transparency and interpretability will rise. Explainable AI (XAI) methods are being created to offer comprehensible justifications for automated classifications, anomaly notifications, forecasts, and coding suggestions. Methods like SHAP and LIME assist in determining the elements that influence a model's result. When integrated into clinical data-management interfaces, these techniques can enable data managers and regulatory reviewers to comprehend the reasons behind a particular record being flagged or why an AI system produced a specific recommendation. The advancement of clear and understandable AI will be crucial for building trust in automated systems and aiding regulatory assessment.

10.7 Digital Twins and Predictive Clinical Studies:

Digital Patient Twins (DPTs) are a new idea where computational models are created to depict individual trial subjects by utilizing data from various sources. These models could integrate historical clinical trial data, Electronic Health Records, genetic traits, Real-World Evidence, and information gathered from wearable devices.In clinical research, digital twins can aid in simulating and assessing trial protocols prior to the actual execution of the study. They might also aid in the creation of Synthetic Control Arms (SCAs), possibly decreasing reliance on traditional control groups in certain clinical trial scenarios. In addition, predictive models can help in forecasting participant attrition, detecting possible safety issues, and assessing probable trial results, especially in research concerning small or rare disease groups.

10.8 Future Responsibilities of Clinical Data Managers:

The growing use of AI and ML is not expected to remove the necessity for clinical data managers. Rather, it is anticipated that the essence of their duties will evolve. Conventional tasks like developing manual edit-checks, tracking routine queries, and reviewing data repetitively could see greater support from automation.[36,63,66,67]

11] Conclusion:

Artificial Intelligence (AI) and Machine Learning (ML) are driving a fundamental paradigm shift in Clinical Data Management (CDM), transitioning the field from a historically manual, labor-intensive, and rule-bound process into an automated, proactive, and highly intelligent ecosystem. Modern clinical trials generate vast volumes of complex, heterogeneous data from diverse sources, making traditional data cleaning and monitoring methods increasingly inefficient and prone to human error. By embedding AI/ML algorithms across the clinical trial data lifecycle—from initial data collection and multi-source integration to validation, cleaning, and advanced analytics—organizations can dramatically enhance both operational efficiency and overall data integrity.

Key practical applications of AI/ML in CDM include automated data validation, rapid identification of missing or conflicting entries, real-time anomaly and outlier detection, smart query generation and management, and advanced Natural Language Processing (NLP) to extract unstructured information from clinical notes. Furthermore, predictive analytics and risk-based clinical data monitoring enable trial managers to identify potential operational risks, operational bottlenecks, and protocol deviations well before they compromise trial outcomes.

By catching errors early in the trial timeline, AI/ML implementation reduces repetitive manual tasks, accelerates database cleaning and lock schedules, and empowers cross-functional teams to make faster, evidence-based decisions. Consequently, these capabilities significantly elevate the benchmark for clinical data quality while reducing operational overhead and accelerating time-to-market for novel therapeutic interventions.

However, the widespread adoption of AI/ML in CDM is not without substantial challenges. Key hurdles include handling non-standardized or low-quality source data, mitigating algorithmic bias, ensuring model interpretability ("black box" issues), protecting patient privacy, maintaining robust cybersecurity, and achieving seamless interoperability with legacy electronic data capture (EDC) systems. Additionally, the industry faces a critical skills gap, necessitating continuous upskilling for clinical research personnel to effectively collaborate with intelligent software systems.

Regulatory compliance and human-in-the-loop oversight remain paramount for the ethical and legal deployment of AI/ML technologies. Machine learning predictions and automated outputs must undergo rigorous validation and structured review by qualified clinical data managers, particularly when decisions directly impact participant safety, trial integrity, or formal regulatory submissions to authorities such as the FDA or EMA. AI should serve to augment human expertise rather than replace clinical judgment.

Looking ahead, future CDM systems will become increasingly self-learning, adaptive, and capable of real-time continuous data surveillance. As technologies mature, the integration of generative AI and automated decision-support frameworks will further streamline routine workflows. Supported by strong data governance, transparent model architectures, rigorous validation, and expert human supervision, AI/ML holds immense potential to establish a highly efficient, accurate, transparent, and patient-centric framework for next-generation clinical trial data management.

- **Summarize the transformation of CDM through AI/ML:**

Artificial Intelligence (AI) and Machine Learning (ML) are fundamentally restructuring Clinical Data Management (CDM), converting labor-intensive, human-driven workflows into continuous, intelligent systems. While traditional CDM relied heavily on post-hoc batch processing and manual data checks, modern AI/ML architectures process electronic health records (EHR) and trial inputs in real time. Algorithms automatically detect anomalies, clean messy data pipelines, auto-flag protocol deviations, and streamline query resolution, significantly shortening database lock timelines and reducing operational costs. Beyond mere automation and error detection, AI/ML introduces novel capabilities that elevate trial data reliability and execution:

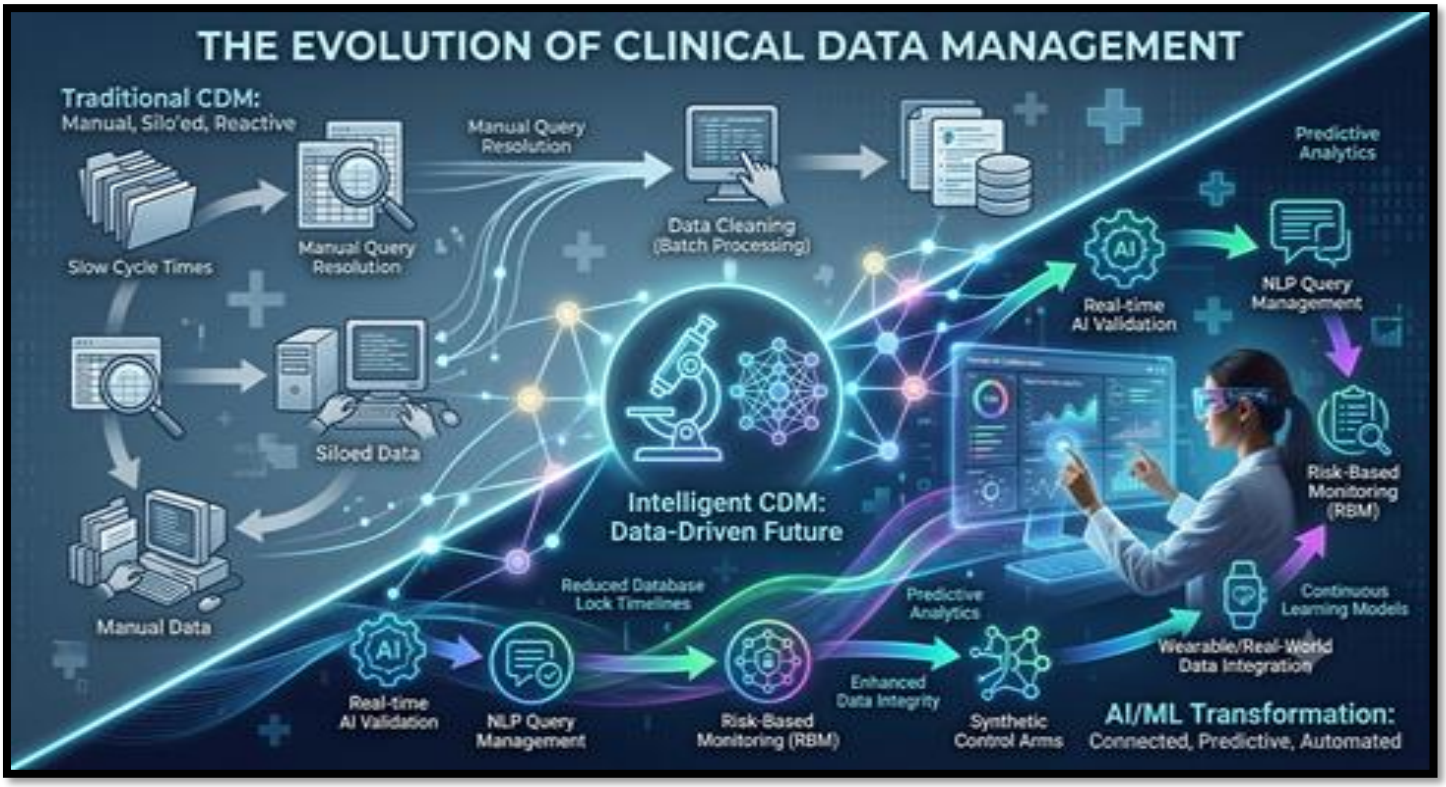
Predictive Query Resolution & Intelligent Automation: Advanced natural language processing (NLP) models automatically parse unstructured clinical notes, lab reports, and doctor logs. By learning from historical trial databases, AI can draft, route, and even self-resolve routine queries without human intervention.

Risk-Based Monitoring (RBM) & Synthetic Control Arms: ML models continuously analyze site-level performance metrics to predict operational bottlenecks or data fabrication risks before they impact the trial. Furthermore, AI-generated synthetic data and historical trial controls are increasingly used to augment small control groups in rare-disease studies.

Continuous Cross-Source Data Harmonization: Modern trials integrate disparate real-world data (RWD) streams, including wearable sensor telemetry, mobile apps, and multi-center labs. Deep learning models dynamically harmonize heterogeneous data formats into standardized models (such as CDISC SDTM) instantly, ensuring seamless data interoperability.

Ultimately, AI/ML redefines the clinical data manager's role—shifting focus from repetitive, reactive manual review to high-level data governance, safety analysis, and protocol optimization. By predicting compliance

risks before they escalate, AI-powered CDM ensures faster regulatory submissions and higher data integrity across complex global trials.



➢ **Major Applications:**

Automated Data Validation and Anomaly Detection: AI and ML algorithms can continuously evaluate data collected through Electronic Data Capture (EDC) systems. These technologies can identify unusual patterns, inconsistencies between variables, outliers, protocol deviations, and other potential data-quality issues that may not be detected through conventional rule-based edit checks.

NLP-Based Processing of Unstructured Data: Natural Language Processing (NLP) techniques can analyze unstructured clinical information such as medical narratives, discharge summaries, investigator notes, and other textual records. NLP can extract clinically relevant information, including adverse events, medical history, and indicators of protocol compliance.

Automated Medical Coding: Machine learning can assist in converting free-text medical terms, including adverse events and concomitant medications, into standardized terminology using dictionaries such as MedDRA and WHO Drug. The system can provide appropriate coding suggestions, which can subsequently be reviewed and confirmed by qualified medical coders.

Intelligent Query Management: AI-based systems can analyze clinical data and historical query-resolution patterns to identify potential data queries, recommend appropriate actions, and assist in drafting and prioritizing queries. This can improve query resolution and reduce the accumulation of unresolved queries before database lock.

EHR-to-EDC Data Integration: AI-enabled integration technologies can facilitate the mapping, reconciliation, and transfer of relevant patient information from Electronic Health Records (EHRs), wearable devices, and other external data sources into EDC platforms. This reduces manual data transcription and supports more efficient integration of multiple clinical data streams.

➢ **Benefits for Clinical-Trial Data Quality and Efficiency:**

Faster Database Lock: Continuous AI-assisted data review enables errors and inconsistencies to be identified throughout the study rather than being addressed mainly during the final stages. This can reduce the time required for database cleaning and potentially accelerate database lock.

Improved Data Integrity: AI/ML systems can identify complex relationships and unusual data patterns at an early stage, helping to minimize manual errors, inconsistencies, and gaps in data review. Consequently, the overall reliability and integrity of clinical-trial datasets can be improved.

Risk-Based Allocation of Resources: AI-driven analytics can help prioritize data-management activities according to risk. Instead of manually reviewing all data with equal intensity, clinical data-management teams can concentrate their expertise on high-risk sites, important safety information, and significant or recurring data anomalies.

Integration of Multiple Data Sources: AI/ML technologies can assist in combining and harmonizing information obtained from different sources, including wearable devices, electronic patient-reported outcomes (ePRO), central laboratories, EHRs, and hospital databases. This supports the development of a more comprehensive and consistent clinical-trial dataset.

➢ **Challenges Requiring Attention:**

Limited Explainability of Complex Models: Some advanced AI and deep-learning models operate as relatively complex or "black-box" systems, making it difficult to understand how particular predictions, classifications, or data-quality alerts are generated. Limited interpretability may create challenges during regulatory review and inspection.

Integration and Interoperability Issues: Implementing AI solutions within existing clinical data-management platforms can be technically challenging, particularly when organizations use older systems or when clinical-trial data are generated across different sites, platforms, and geographical regions.

Data Bias and Model Performance Changes: AI/ML models depend heavily on the quality and representativeness of training data. Models developed using historical datasets may reproduce existing biases or may not perform consistently when applied to different populations, new protocols, or rare-disease clinical trials. Continuous monitoring may therefore be required to identify changes in model performance.

Implementation Costs: The initial adoption of AI/ML may require substantial investment in software, infrastructure, system validation, data preparation, cybersecurity, and employee training. Organizations must therefore carefully evaluate the financial and operational requirements before large-scale implementation.

➢ **Importance of Regulatory Compliance and Human Oversight:**

Human-in-the-Loop Approach: AI should primarily support, rather than completely replace, qualified clinical data-management professionals. Human experts should retain responsibility for reviewing important AI-generated outputs, including data queries, medical coding decisions, and potentially significant data modifications.

GxP and 21 CFR Part 11 Compliance: AI-assisted clinical data-management systems should operate in accordance with applicable Good Practice (GxP) requirements and electronic-record regulations such as 21 CFR Part 11. Appropriate audit trails should document relevant system activities, outputs, and changes to ensure traceability and data integrity.

Validation of AI-Based Systems: AI/ML software used in regulated clinical-trial environments requires appropriate validation and ongoing performance assessment. Validation should establish that the system operates consistently, produces reliable results, and remains suitable for its intended use across relevant clinical-trial conditions.

➢ **Future Potential of Intelligent CDM Systems:**

Automated Trial Setup and Configuration: Future AI systems may be capable of interpreting clinical-trial protocols and assisting in the automatic development of electronic Case Report Forms (eCRFs), data structures, and study-specific edit checks. This could reduce the time and manual effort required during initial study configuration.

Synthetic Control Arms and Generative AI: Machine learning and generative approaches may increasingly support the development of synthetic control groups using appropriately selected historical clinical-trial data. Such approaches could potentially reduce the requirement for conventional control participants in selected trial designs while supporting efficient study planning.

Adaptive and Self-Correcting Data Pipelines: Future intelligent systems may be able to identify and resolve minor data-structure or schema discrepancies across different clinical sites and data sources. They may also identify early indicators of protocol-compliance problems and alert study teams before these issues become significant.

Predictive Clinical-Trial Risk Management: Advanced AI systems may use real-time data streams to predict potential problems such as site-performance issues, participant dropout, data-quality concerns, and emerging safety signals. This could shift clinical data management from a primarily reactive approach toward a more proactive and predictive model of trial oversight.

REFERENCES:

1. National Institutes of Health, Office of Research on Women's Health. (n.d.). *Clinical research and trials*. https://orwh.od.nih.gov/clinical-research-and-trials
2. Gerritsen, M. G., Sartorius, O. E., van der Veen, F. M., & Meester, G. T. (1993). Data management in multi-center clinical trials and the role of a nation-wide computer network: A 5 year evaluation. *Proceedings of the Annual Symposium on Computer Applications in Medical Care*, 659–662.
3. Krishnankutty, B., Bellary, S., Kumar, N. B. R., & Moodahadu, L. S. (2012). Data management in clinical research: An overview. *Indian Journal of Pharmacology, 44*(2), 168–172. https://doi.org/10.4103/0253-7613.93842
4. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. (2025). *ICH harmonised guideline: Good clinical practice E6(R3)*. https://database.ich.org/sites/default/files/ICH_E6%28R3%29_Guideline_2025_0116.pdf
5. Mahadik, S., Sen, P., & Shah, E. J. (2025). Harnessing digital health technologies and real-world evidence to enhance clinical research and patient outcomes. *Digital Health, 11*, Article 20552076251362097. https://doi.org/10.1177/20552076251362097
6. Mitchell, E. J., Goodman, K., Wakefield, N., et al. (2022). Clinical trial management: A profession in crisis? *Trials, 23,* Article 357. https://doi.org/10.1186/s13063-022-06286-4
7. Musik, S., Sasin-Kurowska, J., & Panczyk, M. (2025). Bridging the past and future of clinical data management: The transformative impact of artificial intelligence. *Open Access Journal of Clinical Trials, 17*, 15–33. https://doi.org/10.2147/OAJCT.S509921
8. Cole, S. (n.d.). *What is human-in-the-loop?* IBM. https://www.ibm.com/think/topics/human-in-the-loop
9. Lee, S., Roh, G., Kim, J., Lee, Y. H., Woo, H., & Lee, S. (2023). Effective data quality management for electronic medical record data using SMART DATA. *International Journal of Medical Informatics, 180,* Article 105262. https://doi.org/10.1016/j.ijmedinf.2023.105262
10. An, D., Lim, M., & Lee, S. (2025). Challenges for data quality in the clinical data life cycle: Systematic review. *Journal of Medical Internet Research, 27*, Article e60709. https://doi.org/10.2196/60709

11. Weng, C. (2020). Clinical data quality: A data life cycle perspective. *Biostatistics and Epidemiology, 4*(1), 6–14. https://doi.org/10.1080/24709360.2019.1572344
12. Tufts Center for the Study of Drug Development. (2021). *Rising protocol design complexity is driving rapid growth in clinical trial data volume*. Tufts University.
13. Parab, A. A., Mehta, P., Vattikola, A., Denney, C. K., Cherry, M., Maniar, R. M., & Kjaer, J. (2020). Accelerating the adoption of eSource in clinical research: A TransCelerate point of view. *Therapeutic Innovation & Regulatory Science, 54*(5), 1141–1151. https://doi.org/10.1007/s43441-020-00138-y
14. Müller, A., Christmann, L.-S., Kohler, S., Eils, R., & Prasser, F. (2023). Machine learning for medical data integration. *Studies in Health Technology and Informatics, 302*, 691–695. https://doi.org/10.3233/SHTI230241
15. Nourani, A., Ayatollahi, H., & Solaymani Dodaran, M. (2019). A review of clinical data management systems used in clinical trials. *Reviews on Recent Clinical Trials, 14*(1), 10–23. https://doi.org/10.2174/1574887113666180924165230
16. Lobanov, V. S., & Agrafiotis, D. K. (2019). A cross-source, system-agnostic solution for clinical data review. *Database, 2019*, Article baz017. https://doi.org/10.1093/database/baz017
17. Ohi, S., Iwamoto, T., Ikeda, D., Kawanishi, Y., Kitajima, K., & Ohyanagi, H. (2026). Artificial intelligence (AI)-aided clinical data management: Applications, human-in-the-loop workflows, and regulatory considerations. *Global Health & Medicine, 8*(3), 182–192. https://doi.org/10.35772/ghm.2026.01051
18. Han, R., Acosta, J. N., Shakeri, Z., Ioannidis, J. P. A., Topol, E. J., & Rajpurkar, P. (2024). Randomised controlled trials evaluating artificial intelligence in clinical practice: A scoping review. *The Lancet Digital Health, 6*(5), e367–e373. https://doi.org/10.1016/S2589-7500(24)00047-5
19. Weissler, E. H., Naumann, T., Andersson, T., Ranganath, R., Elemento, O., Luo, Y., Freitag, D. F., Benoit, J., Hughes, M. C., Khan, F., Slater, P., Shameer, K., Roe, M., Hutchison, E., Kollins, S. H., Broedl, U., Meng, Z., Wong, J. L., Curtis, L., Huang, E., & Ghassemi, M. (2021). The role of machine learning in clinical research: Transforming the future of evidence generation. *Trials, 22*(1), Article 537. https://doi.org/10.1186/s13063-021-05489-x
20. Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics, 19*(6), 1236–1246. https://doi.org/10.1093/bib/bbx044
21. de Hond, A. A. H., Leeuwenberg, A. M., Hooft, L., Kant, I. M. J., Nijman, S. W. J., van Os, H. J. A., Aardoom, J. J., Debray, T. P. A., Schuit, E., van Smeden, M., Reitsma, J. B., Steyerberg, E. W., Chavannes, N. H., & Moons, K. G. M. (2022). Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: A scoping review. *npj Digital Medicine, 5*(1), Article 2. https://doi.org/10.1038/s41746-021-00549-7
22. Krishnan, R., Rajpurkar, P., & Topol, E. J. (2022). Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering, 6*(12), 1346–1352. https://doi.org/10.1038/s41551-022-00914-1
23. Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). The MIT Press.
24. van Breugel, B., Liu, T., Oglic, D., & van der Schaar, M. (2024). Synthetic data in biomedicine via generative artificial intelligence. *Nature Reviews Bioengineering, 2*(12), 991–1004. https://doi.org/10.1038/s44222-024-00245-7
25. Cho, H., She, J., De Marchi, D., El-Zaatari, H., Barnes, E. L., Kahkoska, A. R., Kosorok, M. R., & Virkud, A. V. (2024). Machine learning and health science research: Tutorial. *Journal of Medical Internet Research, 26*, e50890. https://doi.org/10.2196/50890

26. Miller, M. I., Shih, L. C., & Kolachalama, V. B. (2023). Machine learning in clinical trials: A primer with applications to neurology. *Neurotherapeutics, 20*(4), 1066–1080. https://doi.org/10.1007/s13311-023-01384-2pubmed.ncbi.nlm.nih
27. Thomas, L. B., Mastorides, S. M., Viswanadhan, N. A., Jakey, C. E., & Borkowski, A. A. (2021). Artificial intelligence: Review of current and future applications in medicine. *Federal Practitioner, 38*(11), 527–538. https://doi.org/10.12788/fp.0174pubmed.ncbi.nlm.nih
28. Papadopoulos, P., Soflano, M., Chaudy, Y., Adejo, W., & Connolly, T. M. (2022). A systematic review of technologies and standards used in the development of rule-based clinical decision support systems. *Health and Technology, 12*(4), 713–727. https://doi.org/10.1007/s12553-022-00672-9
29. Oronsky, B., Burbano, E., Stirn, M., Brechlin, J., Abrouk, N., Caroen, S., Coyle, A., Williams, J., Cabrales, P., & Reid, T. R. (2023). Data management 101 for drug developers: A peek behind the curtain. *Clinical and Translational Science, 16*(9), 1497–1509. https://doi.org/10.1111/cts.13582pmc.ncbi.nlm.nih
30. Lebedys, E., Famatiga-Fay, C., Bhatkar, P., Johnson, D., Viswanathan, G., & Zozus, M. N. (2021). Data management plan. *Journal of the Society for Clinical Data Management, 4*. https://doi.org/10.47912/jscdm.116jscdm
31. Buckley, M. T., Vattikola, A., Maniar, R., & Dai, H. (2021). Direct data extraction and exchange of local labs for clinical research protocols: A partnership with sites, biopharmaceutical firms, and clinical research organizations. *Journal of the Society for Clinical Data Management, 1*(1), 1–5. https://doi.org/10.47912/jscdm.21jscdm
32. Montano, O., Pestronk, M., Johnson, D., Muthanna, M., Redkar-Brown, D., Russo, R., Eade, D., & Zozus, M. N. (2021). Electronic data capture—Study conduct, maintenance and closeout. *Journal of the Society for Clinical Data Management, 1*(1). https://doi.org/10.47912/jscdm.31
33. Wang, Z., Craven, C., Syed, M., Greer, M., Seker, E., Syed, S., & Zozus, M. N. (2024). Clinical data warehousing: A scoping review. *Journal of the Society for Clinical Data Management, 4*(2). https://doi.org/10.47912/jscdm.320
34. Stokman, P. G., Ensign, L., Langeneckhardt, D., Mörsch, M., Nuyens, K., Herrera, D., Hochgräber, G., Cassan, V., Beineke, P., Kwock, R., Voortman, A., Vogelgesang, S., Boussetta, S., & Bitzer, B. (2021). Risk-based quality management in CDM: An inquiry into the value of generalized query-based data cleaning. *Journal of the Society for Clinical Data Management, 1*(1). https://doi.org/10.47912/jscdm.20
35. Geraci, J., Rao, P., Grandinetti, C., Qorri, B., Nadolny, P., Ayalew, K., Bregnhøj, L., Edwards, L., Hofmann, K., Khozin, S., Schaltenbrand, N., Stemmler, T., Yeomans, A., Zambas, D., & Khin, N. A. (2025). Current opportunities for the integration and use of artificial intelligence and machine learning in clinical trials: Good clinical practice perspectives. *Journal of the Society for Clinical Data Management, 5*(2). https://doi.org/10.47912/jscdm.426
36. U.S. Food and Drug Administration. (2025, January). *Considerations for the use of artificial intelligence to support regulatory decision-making for drug and biological products: Draft guidance for industry and other interested parties* [Draft guidance]. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological
37. Lin, A., Wang, Z., Jiang, A., Chen, L., Qi, C., Zhu, L., Mou, W., Gan, W., Zeng, D., Xiao, M., Chu, G., Peng, S., Wong, H., Zhang, L., Zhang, H., Deng, X., Wang, Y., Zhang, J., Cheng, Q., … Luo, P. (2025). Large language models in clinical trials: Applications, technical advances, and future directions. *BMC Medicine, 23*(1), Article 563. https://doi.org/10.1186/s12916-025-04348-9pmc.ncbi.nlm.nih

38. Abdellaoui, C., Redjdal, A., & Seroussi, B. (2025). Generative-AI-based approaches for information extraction from clinical notes: A scoping review. *Studies in Health Technology and Informatics, 328*, 193–197.
39. Omar, M., Nadkarni, G. N., Klang, E., & Glicksberg, B. S. (2024). Large language models in medicine: A review of current clinical trials across healthcare applications. *PLOS Digital Health, 3*(11), Article e0000662. https://doi.org/10.1371/journal.pdig.0000662
40. Yan, C., Yang, R., Yan, H., Wei, X., Zhou, E., Zhao, S., & Zhuang, Y. (2026). Application of artificial intelligence in MedDRA coding: A practical exploration from clinical data management perspective. *Therapeutic Innovation & Regulatory Science, 60*(4), 1059–1067. https://doi.org/10.1007/s43441-026-00953-9
41. Kreimeyer, K., Foster, M., Pandey, A., Arya, A., Halford, G., Jones, S. F., … & Hurdle, J. F. (2017). Natural language processing of clinical notes on secondary use of electronic health records: A systematic review. *Journal of Biomedical Informatics, 73*, 14–22. https://doi.org/10.1016/j.jbi.2017.07.008
42. Getz, K. A., & Campo, R. A. (2018). Trends in clinical trial design, management, and technology adoption. *Nature Reviews Drug Discovery*, *17*(8), 534–535. https://doi.org/10.1038/nrd.2018.107
43. El Emam, K., Jonker, E., Arbuckle, L., & Malin, B. (2015). A systematic review of re-identification attacks on health data. *PLOS ONE, 10*(12), Article e0144331. https://doi.org/10.1371/journal.pone.0144331
44. Vayena, E., Salathé, M., & Madoff, L. C. (2015). Ethical challenges of big data in public health. *PLOS Computational Biology, 11*(2), Article e1003904. https://doi.org/10.1371/journal.pcbi.1003904
45. Tavabi, N., Singh, M., Pruneski, J., & Kiapour, A. M. (2024). Systematic evaluation of common natural language processing techniques to codify clinical notes. *PLOS ONE, 19*(3), Article e0298892. https://doi.org/10.1371/journal.pone.0298892
46. Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics, 42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010
47. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys, 41*(3), Article 15. https://doi.org/10.1145/1541880.1541882
48. Goldstein, M., & Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multidimensional data. *PLOS ONE, 11*(4), Article e0152173. https://doi.org/10.1371/journal.pone.0152173
49. Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). John Wiley & Sons. https://doi.org/10.1002/9781119482260
50. van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429492259
51. U.S. Food and Drug Administration. (2018). *Oversight of clinical investigations—A risk-based approach to monitoring*. U.S. Department of Health and Human Services. https://www.fda.gov/media/116754/download
52. Society for Clinical Data Management. (2020). *Good clinical data management practices (GCDMP)*. https://scdm.org/
53. Kantor, K., & Morzy, M. (2024). Machine learning and natural language processing in clinical trial eligibility criteria parsing: A scoping review. *Drug Discovery Today, 29*(10), Article 104139. https://doi.org/10.1016/j.drudis.2024.104139

54. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems 33* (pp. 1877–1901). Curran Associates, Inc. https://papers.nips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html
55. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7
56. Schwabe, D., Becker, K., Seyferth, M., Klaß, A., & Schaeffter, T. (2024). The METRIC-framework for assessing data quality for trustworthy AI in medicine: A systematic review. *npj Digital Medicine, 7*, Article 203. https://doi.org/10.1038/s41746-024-01196-4
57. Clinical Data Interchange Standards Consortium. (n.d.). *Standards*. https://www.cdisc.org/standards
58. Clinical Data Interchange Standards Consortium. (n.d.). *Study Data Tabulation Model (SDTM)*. https://www.cdisc.org/standards/foundational/sdtm
59. Clinical Data Interchange Standards Consortium. (n.d.). *Foundational standards*. https://www.cdisc.org/standards/foundational-standards
60. World Health Organization. (2021). *Ethics and governance of artificial intelligence for health: WHO guidance*. https://www.who.int/publications/i/item/9789240029200
61. National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1
62. U.S. Food and Drug Administration, Health Canada, & Medicines and Healthcare products Regulatory Agency. (2021, October). *Good machine learning practice for medical device development: Guiding principles*. https://www.fda.gov/media/153486/download
63. European Medicines Agency. (2024). *Reflection paper on the use of artificial intelligence (AI) in the medicinal product lifecycle* (EMA/CHMP/CVMP/83833/2023). https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-use-artificial-intelligence-ai-medicinal-product-lifecycle_en.pdfema.europa
64. International Medical Device Regulators Forum. (2025). *Good machine learning practice for medical device development: Guiding principles* (IMDRF/AIML WG/N88 FINAL:2025). https://www.imdrf.org/documents/good-machine-learning-practice-medical-device-development-guiding-principlesimdrf
65. U.S. Food and Drug Administration. (2024). *Electronic systems, electronic records, and electronic signatures in clinical investigations: Questions and answers*. https://www.fda.gov/
66. Health Level Seven International. (n.d.). *FHIR: Fast Healthcare Interoperability Resources*. https://www.hl7.org/fhir/
67. Shen, M.-D., Chen, S.-B., & Ding, X.-D. (2024). The effectiveness of digital twins in promoting precision health across the entire population: A systematic review. *npj Digital Medicine, 7*, Article 145. https://doi.org/10.1038/s41746-024-01146-0