Open Access Journal of Clinical Trials

Dovepress
Taylor & Francis Group

Open Access Full Text Article

REVIEW

Bridging the Past and Future of Clinical Data Management: The Transformative Impact of Artificial Intelligence

Szymon Musik[1,2], Joanna Sasin-Kurowska[2], Mariusz Panczyk[1]

[1]Department of Education and Research in Health Sciences, Medical University of Warsaw, Warsaw, Poland; [2]Clinical Data & Insights, Biopharmaceutical Clinical Operations, R&D, AstraZeneca, Warsaw, Poland

Correspondence: Szymon Musik, AstraZeneca Poland, Postępu 14A Street, 02-676, Warsaw, Poland, Tel +48 536-931-331,
Email szymon.musik@astrazeneca.com

Abstract: Effective clinical data management is fundamental to clinical research and regulatory submissions. Modern clinical trials have increasingly adopted web-based electronic data capture (EDC) systems, which enhance data collection efficiency but introduces challenges in data integration and quality. This scoping review explores the transformative role of artificial intelligence and machine learning in evolving CDM into clinical data science. In the review, we followed the PRISMA-ScR guidelines and analyzed the literature from 2008 to 2025 using Scopus, Web of Science, and PubMed databases. A total of 26 papers were included and categorized into those related to clinical data management, natural language processing, and general artificial intelligence/machine learning adoption in clinical data management. The integration shows promise in enhancing data analysis, automating data cleaning, and predicting critical outcomes. The key emerging trends include risk-based quality monitoring, blockchain technology, remote monitoring, and patient-centric approaches involving wearables and mobile applications. The results clearly indicate a substantial increase in data volume in Phase III trials, underscoring the need for advanced technologies natural language processing offers significant potential in interpreting unstructured text data, thereby improving the clinical data management processes. The review concludes on different artificial intelligence/machine learning techniques like natural language processing, predictive analytics, and automation technologies, and their applications in improving data quality and streamlining clinical data workflows.

Keywords: clinical data management, artificial intelligence, machine learning, natural language processing, clinical data science, electronic data capture

Introduction

Effective management of clinical data is crucial for the development of new drugs and medical devices, serving as a fundamental pillar of clinical research and playing a pivotal role in regulatory submissions, such as New Drug Applications (NDAs). The data collection process in clinical studies is typically divided into three key phases: start-up, conduct, and close-out. During these phases, clinical data management (CDM) teams are responsible for designing and constructing clinical databases, ensuring that all modules and edit checks are aligned with the requirements outlined in the clinical study protocol (CSP). This process is followed by user acceptance testing (UAT), after which essential documents, including the Data Management Plan and Case Report Form (CRF) Completion Guidelines, are generated. The primary objective of CDM is to ensure the timely and accurate capture of high-quality data from electronic case report forms (eCRFs) or paper-based CRFs, thereby supporting the integrity of the clinical research process.[1]

In modern clinical trials, most of the data are collected through web-based electronic data capture (EDC) systems. These systems have significantly improved the efficiency, cost-effectiveness, and overall quality of data management. However, they also present new challenges, particularly in the integration of data from multiple sources, which can potentially compromise data quality. EDC systems must not only meet regulatory compliance standards but also be flexible, scalable, and ready for audit. Despite their numerous advantages, issues such as the integration of disparate data

Received: 3 December 2024
Accepted: 25 February 2025
Published: 11 April 2025

sources continue to pose significant challenges to maintaining data quality.[1,2] Ensuring data integrity and quality is essential, with integrity referring to the correct and secure management of data, while quality denotes the credibility and reliability of the data collected. Common challenges in CDM include transcription errors from source data, integration of heterogeneous datasets from multiple sources, and ensuring real-time validation of incoming data streams. AI/ML technologies offer targeted solutions to address these challenges effectively. For instance, automation tools powered by AI, such as robotic process automation (RPA), accelerate the integration of diverse datasets by balancing structures and formats, through overcoming the obstacles of data diversity. RPA can support human repetitive tasks. For example, Virtual Robot can have access to EDC with dedicated account that can be simply recognized via audit trails and after all can be supplied with all details of external data reconciliation eg Third-Party Vendor or SAE. Once all issues are identified by RPA, virtual robots can post queries in EDC to maintain data integrity between systems.[2] In addition, common challenges include transcription errors from the source data, which can sometimes be overlooked due to the reliance on risk-based quality management (RBQM) approaches. RBQM is a strategy that focuses monitoring and quality assurance efforts on the most critical data and processes that directly impact trial outcomes and patient safety. Regulatory guidelines, such as the EMA reflection paper and ICH E6 (R2) guidance, emphasize the importance of identifying and managing “critical data points”, which are specific data elements essential for ensuring patient safety as well as the primary and secondary outcomes.[3]

The complexity of data management in clinical trials has escalated significantly, as evidenced by the fact that Phase III trials now involve an average of 3.6 million data points—three times the amount recorded in 2011.[3] Table 1 shows dramatic increase in data volume underscores the urgent need for advanced technologies to manage and analyze such vast quantities of information effectively. Artificial intelligence (AI) and machine learning (ML) technologies offer a promising solution in this context. AI refers to the broader field of developing systems capable of performing tasks that typically require human intelligence, such as decision-making, language understanding, and pattern recognition. ML, a subset of AI, specifically focuses on the development of algorithms that enable computers to learn from and make predictions based on data without being explicitly programmed for each task.[4]

A notably robust application of Artificial Intelligence (AI) and Machine Learning (ML) in Clinical Data Management (CDM) is Natural Language Processing (NLP), which simplifies the analysis and interpretation of large volumes of unstructured text data, such as Serious Adverse Events description, clinical reports etc. NLP exploit a range of techniques, including tokenization (dividing text into smaller components such as words or phrases), named entity recognition (NER, identifying entities like drug names, patient demographics, or adverse events), and word embeddings (eg, BERT Word2Vec, GloVe), which map words to dense vector spaces to capture semantic relations. These techniques enable the extraction of structured data from unstructured text, significantly improving data integration and reducing manual effort. For example, NER can identify adverse events in patient narratives, providing critical insights for safety monitoring and reporting.[5,6]

Machine Learning (ML), a subset of AI, underpins many NLP techniques and broader applications in CDM. Supervised learning, which relies on labeled datasets, is commonly used for tasks such as adverse event prediction and classification of patient outcomes. Unsupervised learning, which operates without labeled data, is valuable for clustering patients with similar characteristics or identifying hidden patterns in datasets. Deep Learning (DL), an

Table 1 Statistical Trends on the Volume of Clinical Trial Data Over the years Based on (https://www.ciscrp.org/patient-data-collection-101-webinar-overview-article/)

Typical Phase III Pivotal Trial (mean for all TAs)	2001–2005	2011–2015	2015–2020
Total Data Points Collected	494,236	929,203	3,560,201
Proportion of Data ‘Non-Core’ (supporting miscellaneous endpoints	18%	32%	34%
Number of Data Collection Solutions and Applications Used*	2	4	4

Notes: (*Includes wearable devices, smartphone applications, electronic clinical and patient reported outcome assessment, electronic patient health information.

https://doi.org/10.2147/OAJCT.S509921

advanced subset of ML, leverages neural network architectures for complex data processing. Architectures such as recurrent neural networks (RNNs) and long short-term memory (LSTM) networks are particularly effective for sequential data like clinical timelines or text. Additionally, transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers), have revolutionized NLP by enabling contextual understanding in tasks such as text summarization or question answering. These advanced techniques are critical for addressing the complexity and scale of modern clinical data workflows.

The integration of NLP with AI and ML technologies holds great potential for transforming the way clinical data are managed, making this process more efficient, scalable, and capable of handling the growing complexity and volume of data in clinical trials. These technologies are increasingly being adopted to address the expanding challenges associated with data management in clinical trials.

As the reliance on electronic solutions for data management intensifies, sponsors are tasked with the critical responsibility of evaluating and selecting appropriate tools, with a particular emphasis on those capable of seamless data integration. Current research indicates that approximately 50% of clinical trials involve between one and five distinct data sources, with data completeness, quality, and cleaning emerging as primary challenges in these settings.[3] To provide a clearer understanding of the current landscape in CDM, as well as the specific challenges and opportunities presented by the integration of AI/ML technologies, Table 2 outlines the crucial steps involved in the database release process and highlights the emerging trends in this field.[1–3]

This review aimed to thoroughly examine the development of Artificial Intelligence (AI) and Machine Learning (ML) systems and their transformative role in the evolution from traditional Clinical Data Management (CDM) to the more advanced paradigm of Clinical Data Science (CDS). Out of the 4325 records initially identified, only 26 studies met the inclusion criteria, which focused on ensuring the relevance and quality of the selected works. While this limited number of included studies may restrict the generalizability of the findings, it allowed us to make more in-depth analysis of studies that directly address the integration of AI and ML in CDM. As a result, the conclusions should be interpreted within the context of this limitation. Future research is encouraged to build on these findings by exploring a broader range of studies to validate and expand the insights presented here. By exploring these developments, we sought to provide a comprehensive roadmap for the future integration and innovation of AI/ML technologies in CDM, thereby paving the way for more efficient, accurate, and scalable solutions in the realm of clinical research.

Materials and Methods

Design

We designed a scoping review that searched, assessed, and synthesized information related to the topic of integrating CDM with AI and ML. This review followed the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines, which provide a framework to ensure methodological rigor and transparency in the conduct of scoping reviews.[7]

Eligibility Criteria

We included a wide range of article types in our review, such as original research articles, systematic reviews, narrative reviews, overviews, updates that synthesize the existing literature, reports and technical papers, comments and opinions, and methodological papers. The timeframe for the included publications spanned from 2008 to 2025. Papers not written in English and those in which CDM was not the main topic were excluded from the review.

Information Sources and Search Strategies

Our search strategy was driven by the following research question: what existing data address the integration of CDM with artificial intelligence and machine learning? We conducted a scoping literature search covering papers published between January 1, 2008, and April 31, 2023, across three databases: Scopus, Web of Science, and PubMed. The search was carried out in May 2023. Two reviewers (SM and MP) independently screened the databases using a standardized protocol. Upon analyzing the abstracts, we found no potentially eligible articles published before 2008.

Table 2 List of Crucial Steps for Database Release, Along With Description of Current Challenges Due to the introduction of AI/ML-Based Technologies

Activity	**Description**
Phases of data collection	Development of Study Protocol: Design case report forms (CRFs) and configuration of the database.
	Data Collection: Train clinical sites on how to collect data from study participants and enter data into CRFs or electronic data capture (EDC) systems.
	Data Cleaning and Validation: Regular review of data for completeness and accuracy and resolve discrepancies through query management.
	Database Lock (DBL): Conduct a final data review, resolve all queries, and lock the database to ensure data integrity for analysis. Collect signatures from the Principal Investigators (or designees).
	Data Analysis and Reporting: Analyze and interpret the collected data to determine study outcomes and prepare clinical study reports for regulatory submission (FDA, EMA, and PMDA).
	Archiving and Data Retention: Store all study documentation and data securely for future reference and ensure compliance with regulatory data retention policies.
Documents to be released	Data Management Plan (DMP): Outlining the data management processes and procedures for a clinical study.
	Case Report Form (CRF) Guidelines: Providing instructions for completing CRFs to ensure consistency and accuracy.
	Data Validation Plan (DVP): Detailing the validation checks and rules to ensure data quality and integrity.
	User Acceptance Testing (UAT) Plan: Describing the process for testing the database and EDC system before release (go-live).
	Database Specifications: Defining the structure and format of the database to be used for data collection.
	Third-Party Data Transfer Agreement (DTA): Specifying the terms and conditions for transferring data between parties, eg, central lab.
	Data Cleaning Plan: Outlining the procedures for identifying and resolving critical data discrepancies.
Emerging trends in CDM	Artificial Intelligence (AI) and Machine Learning (ML): Utilizing AI/ML to enhance data analysis, predict outcomes, and automate data cleaning checks.
	Risk-Based Quality Monitoring (RBQM): Focusing on critical data points and employing RBQM.
	Real-World Data (RWD) Integration: Incorporating RWD from various sources to provide comprehensive insights.
	Blockchain Technology: Using blockchain for secure and transparent data management.
	Remote Monitoring: Implementing remote monitoring tools to ensure continuous oversight of trial data, especially in decentralized trials.
	Data Standardization and Interoperability: Promoting the use of standardized data formats and ensuring interoperability across different systems.
	Patient-Centric Approaches: Enhancing patient engagement and data collection through wearable devices and mobile apps.
	Regulatory Compliance: Adapting to evolving regulatory requirements for data privacy and security.
	Cloud Computing: Leveraging cloud-based solutions for scalable and flexible data management.
	Advanced Analytics: Employing advanced analytics for real-time data insights and decision-making.

Abbreviations: FDA, Food and Drug Administration; EMA, European Medicines Agency; PMDA, Pharmaceuticals and Medical Devices Agency.

The queries used in the search engines of the databases were as follows:

(((“Artificial intelligence”[Title/Abstract] OR “AI”[Title/Abstract] OR “Natural language processing”[Title/Abstract] OR “expert system*”[Title/Abstract] OR “Knowledge engineering”[Title/Abstract] OR “neural network*”[Title/

Abstract] OR "Intelligent retrieval"[Title/Abstract]) OR ("machine learning"[Title/Abstract])) OR (algorithm [Title/Abstract])) OR ("deep learning"[Title/Abstract])

AND

((((((((("Data management"[Title/Abstract]) OR ("CDM"[Title/Abstract])) OR ("Research data management"[Title/Abstract])) OR ("Data management plan"[Title/Abstract])) OR ("Data Administration"[Title/Abstract])) OR ("Information Management"[Title/Abstract])) OR ("Data Collection"[Title/Abstract])) OR ("Information Storage and Retrieval"[Title/Abstract])) OR ("Data Science"[Title/Abstract])) OR ("Clinical data science"[Title/Abstract])

Study Selection

A three-stage approach was employed for the inclusion and exclusion of studies in the final review process. Initially, duplicate studies were removed. Subsequently, a comprehensive screening was conducted across three phases: first based on titles, then abstracts, and finally, full texts. Discrepancies in article selection were resolved through consensus within the reviewing team. The selection process began with 4325 unique entries. Studies that did not pertain to clinical trials, CDM, or the core focus of our review were excluded. Ultimately, 23 papers and 3 reflection papers created by recognized group of experts in area of clinical data management met the inclusion criteria and were included in the final review. The identified reflection papers were not directly accessible through standard scientific publication searches. However, their content falls within the scope of the review and addresses important issues related to the adaptation of AI and ML in CDM. These papers provide valuable insights and contribute to the discussion on the challenges and future directions in the field of CDM in the context of AI/ML.

Data Collection Process, Extraction, and Quality Assessment

Two researchers (SM and MP) conducted a full-text analysis of the 22 selected papers. A data extraction form was utilized to systematically record information, including authors, year of publication, study design where applicable, and the main objectives and findings of each study. Standardized checklist tools were not employed to assess the quality and risk of bias of individual studies, as this review focused instead on a comprehensive synthesis of the available data.

Synthesis of the Results

The content of the 22 papers included in the final analysis was heterogeneous. Therefore, we performed an aggregative narrative synthesis of the included papers.

Results

The results of our research in Figure 1 highlight the number of papers referring to the diverse potential applications of AI and ML across various domains in clinical trials, which were categorized by specific topics. The review categorized the included papers into three main areas: CDM, NLP, and AI/ML applications more broadly. The results indicate that the primary focus was on CDM, with the highest number of papers (n = 10), highlighting its significant relevance in clinical research. In comparison, there were eight papers discussing general AI/ML applications and four papers specifically focusing on NLP. This distribution of research areas emphasizes the prominence of CDM as a key area for the integration of advanced AI/ML techniques in clinical research.

The characteristics of the included papers are summarized in Table 3. This table provides detailed information on each paper, including the study design, objectives, main findings, and the integration of data management and AI or ML techniques. It highlights the diversity of approaches and methodologies employed across the reviewed literature, offering insights into the current state and challenges in clinical data management, as well as the evolving role of artificial intelligence and machine learning in this field.

Discussion

The integration of AI and ML into CDM marks a transformative shift in the landscape of clinical research. This discussion synthesizes the findings from a scoping review that explores how AI and ML technologies are redefining the processes of data collection, organization, and analysis in clinical trials. By utilizing advanced algorithms, these

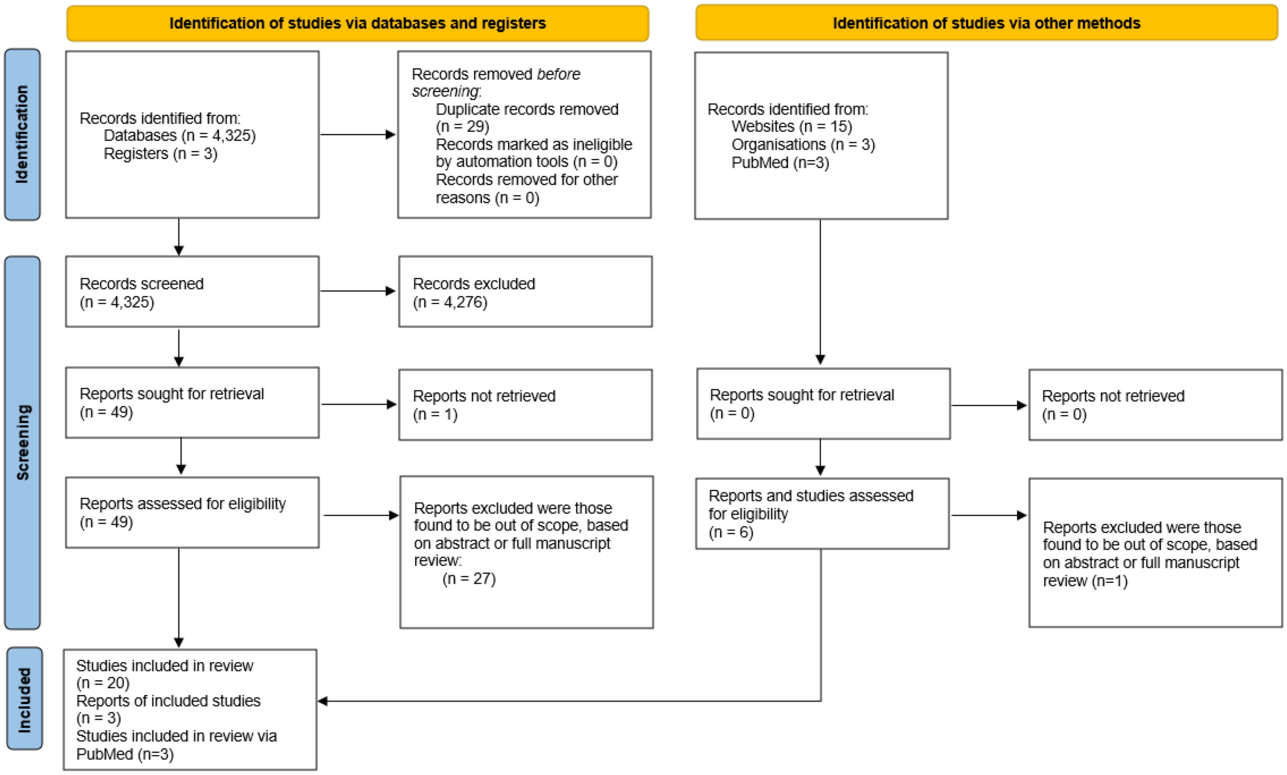
Identification of studies via databases and registers
Identification of studies via other methods
Identification
Records identified from:
Databases (n = 4,325)
Registers (n = 3)
Records removed *before screening*:
Duplicate records removed (n = 29)
Records marked as ineligible by automation tools (n = 0)
Records removed for other reasons (n = 0)
Records identified from:
Websites (n = 15)
Organisations (n = 3)
PubMed (n=3)
Screening
Records screened (n = 4,325)
Records excluded (n = 4,276)
Reports sought for retrieval (n = 49)
Reports not retrieved (n = 1)
Reports sought for retrieval (n = 0)
Reports not retrieved (n = 0)
Reports assessed for eligibility (n = 49)
Reports excluded were those found to be out of scope, based on abstract or full manuscript review:
(n = 27)
Reports and studies assessed for eligibility (n = 6)
Reports excluded were those found to be out of scope, based on abstract or full manuscript review (n=1)
Included
Studies included in review (n = 20)
Reports of included studies (n = 3)
Studies included in review via PubMed (n=3)

Figure 1 PRISMA Flow Diagram of Paper Researched.
Notes: PRISMA Figure adapted from Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. 2018,169(7):467-473. Common Creative.[7]

technologies offer substantial opportunities to streamline clinical workflows, enhance data accuracy, and accelerate the discovery of new medical insights. However, the adoption of AI and ML is not without its challenges. This discussion critically examines both the potential benefits, and the ethical and practical considerations associated with these emerging technologies, providing a comprehensive overview of their role in the evolution from traditional CDM to CDS. The discussion also addresses the integration of NLP and other AI-driven tools, emphasizing their impact on CDM and the broader implications for the future of clinical research.

The Role of AI and ML in Transforming CDM

Building on the discussion of the transformative impact of AI and ML on CDM, it is essential to recognize the diverse types of data handled in clinical trials, which are critical for assessing the efficacy and safety of medical interventions, including both drugs and medical devices. Traditionally, data collection has relied on methods such as paper-based forms and electronic data capture (EDC) systems, which provide structured frameworks for recording patient clinical data. While these methods have been foundational, the advent of emerging technologies has further revolutionized the process of data collection.

In recent years, the integration of wearables, mobile health applications, and remote monitoring devices has markedly enhanced the ability to collect data in real time.[8] These technologies not only allow for continuous tracking of patient outcomes and adherence to study protocols but also improve the granularity and accuracy of the collected data. Additionally, they play a crucial role in increasing patient engagement, which, in turn, contributes to more reliable and comprehensive clinical trial results.

The adoption of these innovative tools reflects a broader trend toward more dynamic and responsive CDM practices, where data are collected across multiple levels and integrated into corporate clinical data warehouses.[1] This multi-layered approach enables the efficient execution of clinical trials by facilitating the seamless aggregation and analysis of diverse data streams, ultimately driving more informed decision-making in healthcare research. Figure 2 illustrates the various

Table 3 Details of the Included Papers

Author(s) (Year)	Title	Study Design	Objective(s)	Main Findings	Data Management/ AI or ML Integration	Methodology	Conclusions and Limitations
Lu et al, (2010)[1]	Clinical data Management: Current Status, Challenges, and Future Directions from Industry Perspectives	Review article/ perspectives	Analyze the role, status, benefits, challenges, and future of clinical data management (CDM)	Effective CDM enhances drug development speed and quality; EDC technology offers significant benefits but also presents challenges	Discussion of EDC technology, CDMS, CTMS, and interoperability with eHR systems	Systematic overview of industry practices and technologies	Successful CDM requires technology adoption, process re-engineering, standardization, and skilled personnel; challenges include integration with other systems and maintaining high data quality
Nadolny P. et al, (2019)[2]	The Evolution of Clinical Data Management into Clinical Data Science	Reflection paper	Examine industry trends impacting CDM and provide insights into evolving practices	Emerging study designs, regulations, and technology innovations are reshaping CDM	AI, ML, NLP, RPA, blockchain, sensors, and wearables	Analysis of current practices and industry trends, risk-based approaches, and technology assessment	CDM must evolve into clinical data science to handle complex data and ensure trial reliability; limited by current technology adoption and regulatory challenges
Pharma Intelligence (2018)[3]	Challenges and Opportunities in Clinical Data Management	Report	Explore operational and quality issues in clinical data management	Manual data management is prevalent; real-time access to data is rare; data governance is critical for compliance	Limited mention of AI/ ML; focus on traditional data management methods	Online survey of 155 professionals in clinical research	Industry struggles with data quality and governance; confidence in data completeness is low; there is a need for improved data management processes
Jiang F. et al (2017)[5]	Artificial Intelligence in Healthcare: Past, Present and Future	Review article	Survey on the current status and future of AI applications in healthcare	AI can assist in early detection, diagnosis, treatment, and outcome prediction across various diseases, especially in cancer, neurology, and cardiology	AI techniques such as machine learning and natural language processing are applied to structured and unstructured healthcare data	Review of the existing literature and analysis of AI application trends	AI shows significant potential in improving healthcare outcomes but faces hurdles in real-life deployment, including regulatory challenges and data sharing barriers
Uzuner Ö. et al, (2007)[6]	Evaluating the State-of-the-Art in Automatic De-identification	Review article	Survey and evaluate automatic de-identification of PHI from medical discharge records	Best systems scored above 98% in F-measure; identifying ambiguous PHI remains challenging	Integration of CRFs, SVMs, HMMs, and decision trees with various features for de-identification	Token and instance-level evaluation using precision, recall, and F-measure	Systems perform well but struggle with ambiguous PHI; further research is needed to enhance robustness when dealing with heterogeneous data

(*Continued*)

Table 3 (Continued).

Author(s) (Year)	Title	Study Design	Objective(s)	Main Findings	Data Management/ AI or ML Integration	Methodology	Conclusions and Limitations
Rahman P. et al, (2020)[8]	Amplifying Domain Expertise in Clinical Data Pipelines	Review article	Present a framework for amplifying domain expertise in clinical data pipelines	Identification of challenges and solutions at each stage of the data pipeline to amplify domain expertise	Emphasis on making ML models more explainable and usable, while involving domain experts at every pipeline step	Review of the literature from database and visualization communities; presentation of a taxonomy for expertise amplification	Amplifying expertise optimizes task completion and directs experts to high-impact tasks, but excessive reliance on experts may reduce system reproducibility and scalability
Richesson Rachel L. et al, (2011)[9]	Data Standards for Clinical Research Data Collection Forms: Current Status and Challenges	Review article	Review existing CRF standards and discuss their limitations	Existing standards for CRFs encompass structure, content, and terminologies, but face limitations due to the protocol-specific nature of clinical research	Emphasis on the need for tools to support retrieval and reuse of existing items and standardized approaches for interoperability	Analysis of existing standards such as CDISC, ISO/IEC 11179, and OpenEHR	Future standards should bridge structural and content standards, with a focus on interoperability. Limitations include incomplete standardization and the challenge of harmonizing diverse research requirements
Barlow C. (2020)[10]	Oncology Research: Clinical Trial Management Systems, Electronic Medical Record, and Artificial Intelligence	Review of peer-reviewed articles, internet sources, book chapters, and white papers	Discuss the implications of electronic systems and regulations in clinical trials and their impact on oncology nurses	Electronic systems improve data transfer, remote enrollment, transparency, documentation, and audit trails in clinical trials	AI used for predictive analytics in clinical research; integration of EMR and CTMS enhances data accuracy and compliance	Review and analysis of the existing literature and regulatory guidelines	Electronic systems enhance clinical trial efficiency and participant safety; however, they require robust understanding and ongoing professional development for effective implementation. Limitations include the complexity of integration and privacy concerns
Mohseni M. et al, (2023)[11]	Electronic Patient-Reported Outcome (ePRO) Application for Patients with Prostate Cancer	Development and usability evaluation study	Develop a smartphone-based ePRO application for prostate cancer patients	The authors developed an ePRO app that received high satisfaction from both patients and specialists, facilitating patient–specialist communication and improving care	No specific AI or ML integration mentioned	Two-phase study: identifying user requirements through surveys and developing the app, followed by usability evaluation using PSSUQ	The app improves patient care and communication but requires further research on cost-effectiveness and integration with medical records

Olalekan Lee Aiyegbusi (2019)[12]	Key Methodological Considerations for Usability Testing of Electronic Patient-Reported Outcome (ePRO) Systems	Review article	Highlight key methodological issues in planning usability testing for ePRO systems	Emphasis on the importance of usability testing for ePRO systems to ensure that they are user-friendly, efficient, and acceptable to patients	Not specifically discussed in terms of AI or ML integration	Discussion of the complexity of ePRO systems, development stages, usability metrics, sample size, task scenarios, and moderating techniques	Comprehensive review of usability testing methodologies; further research is needed on cost-effectiveness and integration with medical records
Petrini C. et al, (2022)[13]	Decentralized clinical trials (DCTs): A few Ethical Considerations	Review article	Explore the ethical implications and requirements of DCTs	DCTs offer logistical benefits, increased inclusivity, and real-time data collection, but they face challenges in data integrity, patient relationship, and regulatory adaptation	Emphasis on the need for secure, transparent, and accurate data management, as well as potential use of blockchain technology	Literature review of DCT implementation and analysis of ethical issues	DCTs improve accessibility and convenience but present challenges in data quality, regulatory frameworks, and patient isolation; there is a need for rigorous guidelines and training
Nadolny P. et al, (2020)[14]	The Evolution of Clinical Data Management into Clinical Data Science (Part 2: The technology enablers)	Reflection paper	Provide insights into adopting emerging technologies for evolving CDM into CDS	New technologies such as AI, ML, and automation are essential for managing the increasing data volume, variety, and velocity in clinical trials	Emphasis on intelligent CDMS, real-time data integration, and AI-driven data review automation	Review and recommendations from industry leaders on technology enablers and fit-for-purpose data strategies	Effective technology adoption can enhance CDM to CDS; limitations include the need for validation and ethical considerations in AI usage
Nadolny P. et al, (2020)[15]	The Evolution of Clinical Data Management into Clinical Data Science (Part 3: The evolution of the CDM role)	Reflection paper	Provide insights into evolving CDM skillsets and competencies	Shift from data integrity to data quality, increased adoption of decentralized trials, and integration of AI/ML technologies	AI/ML used for data review, risk management, and continuous process improvement	Analysis of industry trends, regulatory guidelines, and expert opinions	Clinical data science is critical for future clinical research; rapid adaptation required due to COVID-19, with potential challenges in technology integration and training

(*Continued*)

Table 3 (Continued).

Author(s) (Year)	Title	Study Design	Objective(s)	Main Findings	Data Management/ AI or ML Integration	Methodology	Conclusions and Limitations
Sunyang Fu et al, (2023)[16]	Natural Language Processing for the Evaluation of Methodological Standards and Best Practices of EHR-based Clinical Research	Case study using EHR-based population studies	Leverage NLP techniques to discover reporting patterns and data abstraction methodologies for EHR-based clinical research	Upward trend in reporting EHR-related research methodologies and informatics methods, but overall ratio of reporting/adoption of methodologic standards is low	Development of an NLP algorithm using the MedTaggerIE framework to automate manual review processes	Data selection from REP, guidelines from EQUATOR Network, manual annotation, and NLP algorithm development	Methodologic standards are under-reported; there is high variation in clinical research reporting; the authors recommend developing frameworks, ontologies, and guidelines
Thanh-Dung Le et al, (2022)[17]	Detecting of a Patient's Condition from Clinical Narratives Using Natural Language Representation	Retrospective clinical study	Early diagnosis of heart failure in critically ill children using NLP	Multilayer perceptron neural network achieved 89% accuracy, 88% recall, and 89% precision	Clinical notes were preprocessed and analyzed using machine learning algorithms	Empirical experiments with learning algorithm, feature selection, and preprocessing of numeric values	This study successfully applied ML to detect heart failure; further work is needed for other languages and institutions
Prakash M Nadkarni et al, (2021)[18]	Natural Language Processing: An Introduction	Review/ tutorial	Provide an overview and tutorial of natural language processing (NLP) and modern NLP-system design	This article highlights the evolution of NLP, common sub-problems, current efforts in medical NLP, and future directions	This article summarizes machine-learning approaches, discusses the Apache UIMA framework, and considers IBM Watson's impact on medical NLP	Historical evolution, common sub-problems, machine-learning approaches, modern NLP architectures, and future directions	This article provides a comprehensive overview but may lack depth in its discussion regarding specific NLP applications and recent advancements
Weissler, E. H. et al, (2021)[19]	The Role of Machine Learning in Clinical Research: Transforming the Future of Evidence Generation	Narrative review and conference summary	Review the current and future state of ML in clinical research and identify priority areas for investigation	ML shows promise in improving efficiency and quality of clinical trials, but significant barriers remain	Various ML techniques applied to different phases of clinical trials, including data management and participant management	Review of the existing evidence and applications, stakeholder discussions, and identification of barriers and opportunities	ML can enhance clinical research efficiency and quality, but more peer-reviewed evidence and collaboration is needed; operational and philosophical barriers must be addressed to maximize benefits

Kelly H. Zou et al, (2022)[20]	Enhanced Patient-Centricity: How the Biopharmaceutical Industry is Optimizing Patient Care through AI/ML/DL	Review article	Explore AI/ML/DL applications in biopharma for patient-centric care	AI/ML/DL improves patient outcomes, but challenges exist in data interoperability, privacy, and regulatory frameworks	AI/ML/DL models applied to RWD and RCT data for predictive insights	Literature review of AI/ML/DL applications in diagnosis, treatment, and disease management	AI/ML/DL offers significant potential for enhancing patient care, but requires robust data governance and regulatory frameworks
Leiva, V. et al, (2025)[21]	Artificial intelligence and blockchain in clinical trials: enhancing data governance efficiency, integrity, and transparency	Perspective/ Analytical Study	To explore the integration of AI and blockchain in clinical trials to enhance data governance and transparency	AI and blockchain can revolutionize clinical trials by improving data integrity, process automation, and transparency	Highlighted use of AI for predictive modeling and real-time analytics; blockchain for data immutability and smart contracts	Review of bibliometric and network analysis; exploration of current advancements and strategic recommendations	Integration of AI and blockchain offers transformative opportunities in clinical trials, enhancing efficiency and stakeholder trust. Current limitations include scalability issues, integration with existing healthcare systems, and high implementation costs.
Hameed et al, (2023)[22]	The Changing Role of Data Management in Clinical Trials	Review article	Identify crucial tools used in CDM along with documentation and roles and responsibilities	CDM now has a strategic role, including protocol design, data validation, and ensuring data quality and patient safety	Use of advanced software tools such as ORACLE CLINICAL and RAVE for data management	Review of CDM processes including database design, data collection, data entry, data validation, and discrepancy management	CDM has evolved to meet regulatory and quality demands, with a shift to electronic systems and technological advancements, but it faces challenges in standardization and implementation
Askin S. et al, (2023)[23]	Artificial Intelligence Applied to Clinical Trials: Opportunities and Challenges	Literature review	Identify opportunities, challenges, and implications of AI in clinical trials	AI enhances recruitment, design, and analysis; challenges include data availability and regulatory guidance	AI/ML used for recruitment, trial design, patient monitoring, and data analysis	Review of 48 publications from 2017 to 2021, focusing on AI/ML in clinical trials	AI has significant potential but faces challenges in data quality, standardization, and regulatory acceptance
Gazali S. et al, (2017)[24]	Artificial Intelligence Based Clinical Data Management Systems: A Review	Review article	Review clinical data management systems and their integration with AI/ML	CDM systems improve data quality, speed, and regulatory compliance in clinical trials	This article highlights the use of AI/ML for data capture, validation, and management in CDM	Review of the existing literature and technologies in CDM	CDM systems enhance efficiency but face challenges such as learning curve and cost

(*Continued*)

Table 3 (Continued).

Author(s) (Year)	**Title**	**Study Design**	**Objective(s)**	**Main Findings**	**Data Management/ AI or ML Integration**	**Methodology**	**Conclusions and Limitations**
Liddicoat, J. E. et al, (2025)[25]	A policy framework for leveraging generative AI to address enduring challenges in clinical trials	Commentary/ Proposal	To propose the development and use of application-specific language models (ASLMs) for improving clinical trial design	ASLMs can enhance trial efficiency, inclusivity, and safety, offering significant improvements over general LLMs	Advocates for ASLMs, RAG capabilities, and smaller purpose-built models for improved clinical trial design	Proposes a three-step policy approach: development by regulatory bodies, customization by HTA bodies, deployment to stakeholders	ASLMs have potential to improve ethical and scientific value of clinical trials by enhancing design methodology and reducing flawed protocols. Limitations include the potential for AI systems to amplify biases, data confidentiality concerns, and risk of homogenization in trial protocols.
Richard F. Ittenbach (2023)[26]	From Clinical Data Management to Clinical Data Science: Time for a New Educational Model	Tutorial	Propose and provide a blueprint for a graduate-level curriculum in clinical data science	A structured curriculum integrating biostatistics, biomedical informatics, clinical operations, and regulatory affairs is necessary for advanced clinical data scientists	Emphasis on integrating interdisciplinary knowledge bases, but no specific AI or ML integration mentioned	Development of a theoretical framework, creation of core and research courses, and inclusion of pervasive skills and evaluation strategies	The curriculum prepares students for diverse professional settings; however, current limitations include the absence of formal degree programs and the need for further validation and refinement
Youssef, Alaa et al, (2024)[27]	Ethical Considerations in the Design and Conduct of Clinical Trials of Artificial Intelligence	Qualitative study with semi structured interviews	To explore the generalizability of NIH ethical principles for clinical trials to AI trials and identify unique ethical considerations.	The study confirmed the applicability of NIH's ethical principles but identified AI-specific challenges like social value measurement, fair participant selection, and risk-benefit evaluation.	Focus on data use terms in informed consent and transparency about patient data usage by developers during AI clinical trials.	Deductive approach using NIH's ethical principles and inductive approach to uncover broader ethical challenges in AI clinical trials.	Current ethical principles apply but need adaptation for AI-specific challenges in areas like social value, scientific validity, and informed consent. Limited by a small participant pool restricted to a specific AI application (DR screening) in the US, affecting generalizability to other regions or AI applications.

Abbreviations: CDM, clinical data management; EDC, electronic data capture; CDMS, clinical data management system; CTMS, clinical trial management systems; eHR, electronic health record; AI, artificial intelligence; ML, machine learning; CRF, case report form; CDISC, clinical data interchange consortium; ePRO, electronic patient-reported outcome; PSSUQ, post-study system usability questionnaire; DCT, decentralized clinical trial; NLP, natural language processing; DL, deep learning; RPA, robotic process automation; EMR, electronic medical record; SVM, support vector machine; HMM, hidden Markov model; PHI, protected health information; ASLM, application specific language models; RAG, retrieval augmented generation, DR, diabetic retinopathy.

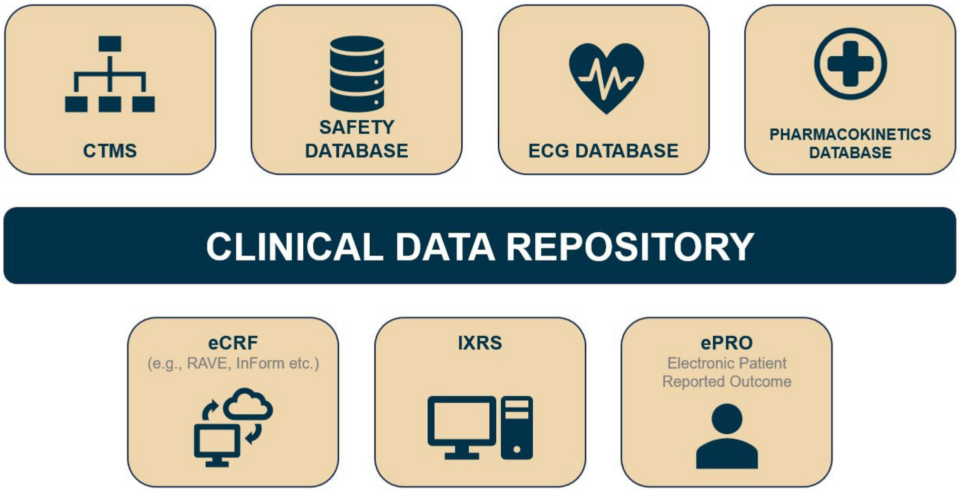
CTMS
SAFETY DATABASE
ECG DATABASE
PHARMACOKINETICS DATABASE
CLINICAL DATA REPOSITORY
eCRF
(e.g., RAVE, InForm etc.)
IXRS
ePRO
Electronic Patient Reported Outcome

Figure 2 A sample list of different sources of data covering all steps needed to support successful data collection within a clinical study, based on the study b y Lu et al.[1]

levels at which data are collected and integrated within a clinical data warehouse, highlighting the complexity and sophistication of modern CDM systems.

Case Report Form Systems

Electronic case report forms (eCRFs) have indeed transformed data collection in clinical research, offering significant improvements over traditional paper-based methods. As clinical trials have increasingly adopted electronic data capture (EDC) systems, eCRFs have emerged as a pivotal component in enhancing the efficiency, accuracy, and cost-effectiveness of data management processes. This shift is particularly relevant in the context of the evolving landscape of CDM, where AI, ML, and other advanced technologies are redefining how data are collected, organized, and analyzed.[9]

eCRFs streamline various aspects of data management, from collection to reporting and analysis. They provide a structured framework for data entry, with predefined fields that ensure consistency and standardization across global study sites. This standardization is critical in maintaining data integrity, particularly in large-scale, multi-site trials. While there are no universal standards for CRF design, best practices and conventions developed by organizations such as the Clinical Data Interchange Standards Consortium (CDISC) offer valuable guidance. However, it is important to recognize that these guidelines may not fully address the complexities of all types of clinical research, such as observational studies or those involving patient-reported outcomes (PROs).[10,11]

One of the key features of eCRFs is their ability to perform real-time data validation. Built-in validation rules, such as edit checks and queries, help minimize errors by flagging inconsistencies or missing data during entry. This immediate feedback mechanism is essential for ensuring data quality and adherence to clinical study protocols. Additionally, eCRFs maintain an audit trail of all data modifications, which is crucial for ensuring data integrity and traceability throughout the trial. This feature not only supports compliance with regulatory requirements but also aids data managers and medical monitors during data review, ensuring that the study is conducted according to the protocol.[1]

Another significant advantage of eCRFs is their integration capabilities with other systems, such as electronic health records (EHRs) and Interactive Voice/Web Response System (IxRS). This integration allows for seamless data transfer between eCRFs and third-party vendor systems, reducing the risk of duplicated data entry and improving data veracity. The ideal scenario involves extracting patient data directly from EHRs, thereby maximizing efficiency by reducing the time and resources required for data entry, verification, and query resolution. While there are few limitations in data integration, this process typically focuses on critical data points such as randomization data, demographics, and enrollment codes.[1] Overall, the adoption of eCRFs has led to significant improvements in clinical trials, enhancing not only the quality and consistency of data but also the overall efficiency of the trial process. These advancements align with the broader trend of integrating AI, ML, and other innovative technologies into CDM, underscoring the potential for continued improvements in this field.

Electronic Patient-Reported Outcomes

The adoption of electronic patient-reported outcomes (ePROs) has significantly advanced the collection of direct data from participants in clinical trials, aligning closely with the broader trend of integrating digital tools and AI-driven technologies into CDM. ePROs represent a key innovation in the shift toward more patient-centered approaches in clinical research, allowing for the real-time capture of patient-reported data through devices such as smartphones, tablets, and computers. This method of data collection offers a more streamlined and organized process compared to traditional paper-based surveys, which are often prone to errors and inefficiencies. Commonly used ePRO tools in clinical research include validated questionnaires such as SF-36, EQ-5D, and FACT-G, which assess various aspects of quality of life.[12]

The immediate capture of data through ePROs facilitates more accurate and timely evaluations of patient symptoms, treatment outcomes, and overall quality of life. This is particularly crucial in clinical trials where the precision of patient-reported data can directly influence the assessment of a drug's or a device's efficacy and safety. Additionally, ePROs enhance patient compliance by incorporating reminders and prompts, which help maintain adherence to data collection schedules and reduce the occurrence of incomplete datasets. This not only improves data quality but also minimizes the likelihood of inconsistencies, as validation checks can be built into the ePRO system to flag potential errors during data entry.[13]

Beyond improving data accuracy and compliance, ePROs also contribute to cost savings by reducing the need for manual data entry and transcription, thus increasing overall efficiency in data management. The interactive features of ePRO applications, such as multimedia content and visual displays, further engage patients, fostering a more active and informed participation in the clinical trial process. This shift toward a more patient-centric model in clinical research is particularly evident in the flexibility and convenience that ePROs offer, allowing for remote data collection and monitoring. This not only broadens the scope of patient participation but also reduces the necessity for frequent in-person visits, which can be a significant burden for participants.[8]

In conclusion, the introduction of ePROs into clinical research has marked a substantial advancement in trial methodology. By enabling more precise, timely, and patient-centered data collection, ePROs enhance the quality and effectiveness of clinical research. Their continued adoption is likely to further improve patient involvement and data reliability, solidifying their role as a valuable tool in the evolving landscape of CDM.

Decision Support Systems in Clinical Data Management

Decision support systems (DSSs) are increasingly becoming an integral component of CDM, offering sophisticated tools and insights that support clinical data managers, monitors, and medical coders in making informed decisions throughout the lifecycle of a study. As clinical trials grow in complexity, DSSs provide a critical layer of analysis and interpretation, leveraging advanced algorithms, ML, and data visualization techniques to transform raw data into actionable insights. These systems not only enhance the efficiency and accuracy of CDM but also contribute significantly to evidence-based decision-making in healthcare research.

In the context of CDM, DSSs encompass a wide range of software tools and analytical methods designed to assist stakeholders in effectively interpreting and utilizing clinical data. These systems are particularly valuable in integrating data from multiple sources, creating a comprehensive view of patient data that is essential for accurate analysis. For example, Rahman et al outlined the complex processes involved in maximizing the potential of DSSs, including data selection from EDC/EHR systems, annotation, curation, and rigorous data cleaning by domain experts. These steps are critical to ensure that the data used in a DSS are both accurate and relevant, ultimately leading to more robust statistical modeling and reliable outcomes.[14,15]

The key features of DSSs in CDM include advanced data analysis techniques, predictive modeling to forecast patient outcomes, and intuitive data visualization tools. These systems also offer clinical decision support functions, such as alerts and evidence-based guidelines, which are essential for timely and accurate decision-making. However, the implementation of a DSS is not without challenges. Factors such as data quality, integrity, interoperability, user training, and regulatory compliance must be carefully managed to fully reap the benefits of these systems. When properly implemented, DSSs can significantly enhance decision-making processes, improve the efficiency of clinical trials, and contribute to higher quality outcomes in healthcare research.[14–17]

Furthermore, the future of DSSs is poised to benefit from ongoing advancements in AI, ML, and NLP technologies. These developments are expected to lead to more advanced and sophisticated DSSs that are capable of providing personalized recommendations and deeper insights into patient data. By enabling more precise and tailored treatment strategies, these systems will play a pivotal role in the evolution toward personalized medicine. Continued investment in DSSs is crucial, as these systems will undoubtedly further enhance healthcare quality, efficiency, and patient outcomes in clinical research.[14–17]

Integration of Natural Language Processing in Clinical Data Workflows

NLP has become a pivotal tool in transforming CDM, particularly in how healthcare professionals extract, analyze, and interpret information from unstructured clinical texts. In clinical settings, vast amounts of data are generated daily from various sources, including patient encounters, electronic health records (EHRs), physician notes, and pathology reports. Traditionally, these unstructured data sources posed significant challenges for data management, as their complexity and lack of standardization often hindered efficient data analysis and effective decision-making processes.[14,18]

NLP addresses these challenges by utilizing advanced computational techniques to analyze and understand human language within clinical texts. By combining linguistics, ML, and AI, NLP algorithms can extract meaningful insights from free-text narratives, identify patterns, and categorize information in a way that structured data management systems typically cannot. For instance, in CDM, NLP plays a crucial role in medical coding, where it automates the extraction of relevant clinical codes, such as those used in MEDRA and WHODRUG, directly from textual documents, thereby improving both accuracy and efficiency in data processing.[19]

Beyond medical coding, NLP also supports global study teams by identifying potential risks at the study level, thus contributing to more informed decision-making throughout the clinical trial process. Additionally, NLP's ability to perform large-scale analysis of clinical texts across diverse patient populations makes it an invaluable tool in both clinical research and population health management. By aggregating and analyzing textual data from multiple EHRs, clinical trials, and research studies, NLP enables researchers to identify trends, discover new insights, and advance medical knowledge across various medical fields. However, despite its significant potential, the implementation of NLP in CDM is not without challenges. Issues such as data privacy concerns and algorithm accuracy must be carefully managed. This requires collaboration between data scientists, data curators, and regulatory bodies to develop robust NLP solutions that comply with privacy regulations, ensure transparency in algorithms, and promote seamless integration with existing data management systems.

In conclusion, while NLP offers tremendous promise for transforming CDM, its successful implementation depends on addressing several sub-problems, including sentence boundary detection, tokenization, shallow parsing, and problem-specific segmentation. By effectively harnessing the capabilities of NLP, healthcare organizations can unlock the full potential of their unstructured clinical data, enhance decision-making processes, and ultimately improve patient outcomes.[20]

Machine Learning Connected Systems

ML has emerged as a crucial tool in CDM, offering innovative solutions to the challenges associated with the vast amounts of data generated during clinical trials. Unlike traditional approaches, ML leverages advanced algorithms and statistical models to make predictions and decisions based on data, without the need for explicit programming. This capability is particularly valuable in the context of clinical trials, where the ability to efficiently analyze and interpret complex datasets is essential for optimizing clinical workflows and enhancing decision-making processes. The integration of ML with blockchain technology, which ensures that each piece of data entered is recorded in a way that cannot be altered without detection, improving the accuracy and reliability of the data. Additionally, blockchain can facilitate real-time access and auditability for authorized stakeholders, further enhancing transparency and regulatory compliance in clinical research. Furthermore, it provides a secure and tamper-proof framework for managing clinical data.[21,22]

ML techniques in CDM are primarily based on two approaches: supervised and unsupervised learning. Supervised learning involves training models using labeled input and output data, allowing the algorithm to make accurate predictions or decisions in tasks such as classification and regression. This method is particularly useful for evaluating patient outcomes and ensuring that clinical decisions are based on reliable data. In contrast, unsupervised learning does not rely on labeled outputs. Instead, it explores the data to uncover hidden patterns or structures, such as through clustering techniques, which can reveal significant, yet previously unidentified, trends in clinical data.[15]

One of the key applications of ML in CDM is predictive analytics. By analyzing historical patient data, including demographics, medical history, laboratory results, and treatment outcomes, ML algorithms can identify patterns that predict future events. For example, ML can be used to forecast disease onset, hospital readmissions, serious adverse events (SAEs), and adverse drug reactions (ADRs). These predictive capabilities are invaluable in clinical trials, as they allow researchers to anticipate potential issues and proactively address them by performing relevant data queries.[13–15] On the other hand during COVID-19 outbreak we have observed significant number of missing data due to multiple difficulties related to critical pandemic situation. ML can be used in such situation to impute not collected data within protocol mandated timeframes.[23]

Beyond predictive analytics, ML significantly contributes to the automation of routine tasks in CDM, such as data cleaning, coding, and documentation. Automating these processes not only streamlines clinical workflows but also reduces the administrative burden on clinical research professionals, allowing them to focus on higher-level decision-making. The enhanced accuracy and completeness of data resulting from ML-driven automation further improve the reliability of clinical trial outcomes, leading to more robust and actionable insights.[24]

In the broader context of clinical research and drug development, ML plays a pivotal role in analyzing large-scale datasets to identify novel insights and trends. By mining electronic health records (EHRs), clinical trial data, and biomedical literature, ML algorithms can accelerate the discovery of new biomarkers, therapeutic targets, and treatment modalities. These advancements have the potential to revolutionize medical research, paving the way for the development of innovative therapies and interventions that address unmet medical needs.[19,20] The use of generative AI in trial design further supports this by enabling the creation of optimized protocols that enhance trial efficacy and safety.[21]

However, the widespread adoption of ML in CDM is not without challenges. Concerns related to data privacy, algorithm transparency, and regulatory compliance must be addressed to ensure the ethical use of ML in healthcare. The complexity of ML models, particularly the opacity of their decision-making processes (often referred to as the "black box" problem), necessitates collaboration between healthcare providers, data scientists, and regulatory bodies. Developing robust ML solutions that adhere to privacy regulations, ensure interpretability, and promote ethical practices is essential to leverage the full potential of ML in clinical settings. Aligning these solutions with international standards and regulations, such as those proposed by regulatory bodies and Health Technology Assessment (HTA) agencies, is crucial for global applicability.[21,25]

In conclusion, ML represents a transformative tool for CDM in healthcare. By harnessing the capabilities of ML, healthcare organizations can unlock the full potential of their data, improve patient outcomes, and advance medical knowledge. This ultimately leads to better quality of care and improved outcomes for patients, particularly in the fast-paced environment of clinical trials.[4] Future advancements in ML, particularly when integrated with blockchain and ASLMs, will further enhance clinical trial processes, offering more effective and equitable healthcare solutions.[21,25]

Artificial Intelligence

AI has emerged as a transformative force in CDM, offering unprecedented opportunities to enhance decision-making, improve patient care, and drive innovation in drug development. By leveraging advanced algorithms and computational techniques, AI can extract actionable insights from complex datasets, including biomarkers, prognostic indicators, and morbidity and mortality rates. This capability allows healthcare providers and researchers to make more informed decisions, ultimately leading to better patient outcomes and more efficient clinical processes.[5]

However, the integration of AI in healthcare/clinical research presents uncommon ethical concerns, including difficulties in measuring social value, establishing scientific validity, ensuring fair participant selection, evaluating risk-benefit proportions, and addressing complexities in data use consent.[25] One of the most significant applications of AI in CDM is in medical imaging analysis, which plays a critical role in patient recruitment for clinical trials, particularly in oncology. AI-powered algorithms can analyze medical images, such as X-rays, MRIs, and CT/PET scans, with a high degree of accuracy, detect abnormalities, identify patterns, and assist in making precise diagnoses. An example of this is the Radiology and Enterprise Medical Imaging Extensions (REMIX) platform, which supports automatic image interpretation in large-scale, multi-disciplinary oncology trials. Such AI-driven technologies are crucial for ensuring that the right patients are selected for trials, thereby improving the efficiency and effectiveness of clinical research.[5,26]

Beyond imaging, AI plays a pivotal role in predictive modeling and risk stratification. By analyzing vast amounts of data from electronic health records, genomic profiles, and other clinical variables, AI algorithms can predict patient outcomes such as hospital readmissions, mortality, and disease progression. These predictive models allow healthcare providers as well as sponsors of clinical trials to identify high-risk patients early, enabling timely interventions and tailored treatment plans that mitigate risks and enhance patient outcomes.

AI also significantly enhances clinical decision support by synthesizing vast amounts of clinical data and generating evidence-based recommendations. AI-powered decision support systems can integrate patient-specific data with clinical guidelines/study protocol in clinical research and the medical literature to provide personalized treatment suggestions, alert clinicians to potential drug interactions, and assist in diagnostic decision-making. These tools empower healthcare providers to make more informed decisions, reduce medical errors, and deliver more effective and efficient patient care.[6,25]

Moreover, AI contributes to the advancement of precision medicine by analyzing genomic data and molecular profiles to identify targeted therapies and personalized treatment approaches. By examining genetic variations, biomarker expression patterns, and drug response data, AI algorithms can match patients with the most appropriate therapies and clinical trials. This personalized approach has the potential to improve treatment outcomes, reduce adverse effects, and enhance patient satisfaction, making it a cornerstone of modern healthcare.[6]

In addition to its applications in clinical care, AI is revolutionizing clinical research and data management by automating data collection, analysis, and interpretation. AI algorithms can mine electronic health records, biomedical literature, and clinical trial data to identify emerging research trends, uncover novel insights, and accelerate the development of new therapies and interventions. These AI-driven insights have the potential to significantly advance medical research and propel scientific discoveries in various disease areas.

However, the deployment of AI in clinical trials/healthcare raises critical ethical concerns, such as the potential for algorithmic bias, challenges in ensuring equitable access to AI-driven technologies, and the complexity of informed consent processes. These issues necessitate further guidance on where to focus empirical and normative ethical efforts to minimize unintended harm to trial participants and ensure that AI interventions are safe and effective especially if the decision are impacting multiple participants of clinical trials, notably in areas with high unmet clinical needs.[27] Addressing these issues requires close collaboration between healthcare providers, data scientists, and regulatory bodies to develop ethical AI solutions that prioritize patient privacy, ensure fairness in algorithmic decision-making, and adhere to stringent regulatory standards. Such efforts are essential to fully leverage the transformative potential of AI in CDM.

Conclusion

The integration of AI and ML into CDM offers significant opportunities to revolutionize clinical research. These technologies enhance decision-making, improve patient care, and accelerate drug development by extracting actionable insights from complex datasets, streamlining workflows, and enabling more personalized care approaches. We must also address pressing challenges in contemporary clinical trials, such as managing increasingly complex and diverse datasets/trials mentioned above and ensuring real-time oversight of data quality and consistency between different databases. For instance, AI-powered automation tools like robotic process automation (RPA) can reconcile heterogeneous data sources and reduce manual workload, offering scalable solutions for multi-center studies. However, to fully realize these benefits, AI governance teams should be established within pharmaceutical organizations. These teams would oversee compliance with regulatory frameworks, continuously improve AI/ML tools, and ensure ethical implementation practices across all workflows.

However, the implementation of AI and ML in CDM is not without challenges. The key issues include data privacy, algorithmic transparency, and regulatory compliance. The opaque nature of many AI models complicates efforts to ensure fairness and ethical standards, particularly in clinical settings where patient safety is paramount. Important research questions include how to optimize ML algorithms for anomaly detection in heterogeneous datasets, how to scale AI tools across global clinical sites/departments, and how to create standardized frameworks for data interoperability.[21]

Addressing these challenges requires collaboration among healthcare providers, data scientists, and regulatory bodies to develop AI and ML solutions that comply with privacy regulations, ensure transparency, and maintain ethical integrity. Partnership is also essential to create standardized groundworks for interoperability, scalability, and ethical AI practices across different stakeholders in the clinical research ecosystem. Furthermore, organizations should focus on structured implementation strategies, including phased rollouts and pilot projects that evaluate the scalability and regulatory compliance of AI/ML tools. Continued investment in research and development is also necessary to refine these technologies and integrate them effectively into clinical workflows.

The future of AI and ML in CDM is promising, with advancements in NLP and other AI-driven tools expected to further enhance system capabilities. These innovations will likely lead to more advanced decision support systems, thereby contributing to precision medicine. For instance, prognostic analytics and advanced NLP models can enable real-time risk stratification and safety monitoring, directly linking AI technologies to enhanced trial outcomes and regulatory compliance. Integrating AI governance teams into organizations will also ensure that these improvements align with long-term goals of data security, transparency, and standardization. However, attaining these benefits will depend on navigating regulatory challenges and ensuring ethical implementation.

In conclusion, while AI and ML have the potential to transform CDM, their successful adoption hinges on overcoming significant challenges related to privacy, transparency, and regulation. By establishing dedicated AI governance teams, defining clear research priorities, and adopting a staged implementation approach, organizations can ensure the ethical and effective deployment of AI/ML technologies. A phased approach to AI/ML integration, involving pilot studies followed by scalable implementations, can streamline this transformation while ensuring regulatory alignment and ethical oversight. By addressing these issues, the healthcare industry can fully leverage these technologies to improve patient outcomes, streamline clinical trials, and advance medical research. As these fields evolve, regulators will play a crucial role in shaping the frameworks that ensure these innovations are both safe and effective for all stakeholders.

Data Sharing Statement

Data will be available on the main site of study. Please contact the correspondence author for future access.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This study was supported by the Implementation Doctorate Programme entitled “Optimizing Clinical Data Management”, funded by the Ministry of Education and Science (Grant reference: DWD/6/0494/2022). The research was conducted at the Medical University of Warsaw in collaboration with AstraZeneca.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Zhengwu L, Jing S. Clinical data management: current status, challenges, and future directions from industry perspectives. *Open Access J Clin Trials*. 2010;2. doi:10.2147/OAJCT.S8172.
2. Patrick N. The evolution of CDM into clinical data science. A reflection paper on the impact of the clinical research industry trend on CDM. 2019.
3. Intelligence P. Challenges and opportunities in clinical data management. *Res Rep*. 2018.
4. Stuart Russell PN. Artificial Intelligence: a modern approach. 2016.
5. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *BMJ*. 2017;2:svn–2017. doi:10.1136/svn-2017-000101
6. Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc*. 2007;14(5):550–563. doi:10.1197/jamia.M2444

7. Tricco A, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Internal Med*. 2018;168 (3):169. doi:10.7326/M18-0850
8. Rahman P, Nandi A, Hebert C. Amplifying domain expertise in clinical data pipelines. *JMIR Med Inform*. 2020;8(11):e19612. doi:10.2196/19612
9. Richesson R, Nadkarni P. Data standards for clinical research data collection forms: current status and challenges. *J Am Med Inform Assoc*. 2011;18 (3):341–346. doi:10.1136/amiajnl-2011-000107
10. B C. Oncology research: clinical trial management systems, electronic medical record. *and Artificial Intelligence Sem Oncol Nurs*. 2020;36 (2):151005. doi:10.1016/j.soncn.2020.151005
11. Mohseni M, Ayatollahi H, Arefpour AM. Electronic patient-reported outcome (ePRO) application for patients with prostate cancer. *PLoS One*. 2023;18(8):e0289974. doi:10.1371/journal.pone.0289974
12. Aiyegbusi OL. Key methodological considerations for usability testing of electronic patient-reported outcome (ePRO) systems. *Qual Life Res*. 2020;29(2):325–333. doi:10.1007/s11136-019-02329-z
13. Petrini C, Mannelli C, Riva L, Gainotti S, Gussoni G. Decentralized clinical trials (DCTs): a few ethical considerations. *Front Public Health*. 2022;10:1081150. doi:10.3389/fpubh.2022.1081150
14. Patrick N. The evolution of CDM into clinical data science (Part 2: the technology enablers). A reflection paper on how technology will enable the evolution of CDM into clinical data science. 2020.
15. Patrick N. The evolution of CDM into Clinical Data Science (Part 3: The evolution of the CDM role). a reflection paper on the evolution of CDM skillsets and competencies. 2020.
16. Fu S, Carlson L, Peterson K, et al. Natural language processing for the evaluation of methodological standards and best practices of EHR-based clinical research. *AMIA Summits on Translational Sci Proc*. 2020;2020:171.
17. Le T-D, Noumeir R, Rambaud J, Sans G, Jouvet P. Detecting of a patient's condition from clinical narratives using natural language representation. IEEE Open Journal of Engineering in Medicine and Biology. 2022; 1–7. doi:10.1109/OJEMB.2022.3209900.
18. Nadkarni P, Ohno-Machado L, Chapman W. Natural language processing: an introduction. *J Am Med Inform Assoc*. 2011;18(5):544–551. doi:10.1136/amiajnl-2011-000464
19. Weissler E, Naumann T, Andersson T, et al. The role of machine learning in clinical research: transforming the future of evidence generation. *Trials*. 2021;22(1):22. doi:10.1186/s13063-021-05489-x
20. Zou K, Li J. Enhanced patient-centricity: how the biopharmaceutical industry is optimizing patient care through AI/ML/DL. *Healthcare*. 2022;10 (10):1997. doi:10.3390/healthcare10101997
21. Leiva V, Castro C. Artificial intelligence and blockchain in clinical trials: enhancing data governance efficiency, integrity, and transparency. *Bioanalysis*. 2025;1–16. doi:10.1080/17576180.2025.2452774
22. Fauziyya Shahul Hameed NN, Poldasari S. The changing role of data management in clicnical trials. *Int J Innov Res Technol*. 2023;9(10).
23. Askin S, Burkhalter D, Calado G, El Dakrouni S. Artificial intelligence applied to clinical trials: opportunities and challenges. *Health Technol*. 2023;13(2):203–213. doi:10.1007/s12553-023-00738-2
24. Gazali, Kaur S, Singh I. Artificial intelligence based clinical data management systems: a review. *Inf Med Unlocked*. 2017;9:219–229. doi:10.1016/j.imu.2017.09.003
25. Liddicoat J, Lenarczyk G, Aboy M, Minssen T, Mann S. A policy framework for leveraging generative AI to address enduring challenges in clinical trials. *Npj Digital Med*. 2025;8(1). doi:10.1038/s41746-025-01440-5
26. Ittenbach RF. From clinical data management to clinical data science: time for a new educational model. *Clin Transl Sci*. 2023;16(8):1340–1351. doi:10.1111/cts.13545
27. Youssef A, Nichol AA, Martinez-Martin N, et al. Ethical considerations in the design and conduct of clinical trials of artificial intelligence. *JAMA Network Open*. 2024;7(9):e2432482. doi:10.1001/jamanetworkopen.2024.32482