DH05

Artificial Intelligence Potential Use Cases in Clinical Data Engineering

Mounika Pillamarapu, Parexel International, Bengaluru, India
Deepak Pakalapati, Parexel International, Hyderabad, India

ABSTRACT

Clinical data engineering deals with aggregation, cleaning, reconciliation, and standardization of data from various sources to ensure data veracity and readiness for use in clinical trial reporting. Artificial intelligence (AI) as a technology has just begun its logarithmic phase and can contribute immensely to almost all stages of drug development, more so for those involving data engineering, processing, and reporting. Applying AI techniques to data engineering requirements can speed up the processes, saving both time and money to the sponsors. Deduplication, data augmentation, real world data sanitization, outlier detection, data standardization, missing data and error correction are just a few examples of areas where AI techniques like machine learning, deep learning, feature extraction and natural language processing can have a focused impact. These use cases are discussed in considerable detail.

INTRODUCTION

AI is revolutionizing clinical trials, enhancing efficiency, accuracy, and patient care. The key applications of AI are optimization of protocols based on historical data, collecting real-time patient data through AI-powered devices, extracting information from unstructured sources using natural language processing (NLP), patient recruitment and retention, data analysis, optimizing molecule designs for better drug-target interactions, remote monitoring in decentralized trial, forecasting trial outcomes and enabling adaptive designs. However, challenges exist, including data privacy concerns, regulatory acceptance, and result interpretability.
In this paper, the following examples are discussed in detail.

Handling missing data in clinical trials
Deduplication
Data augmentation
Error Correction
Real world data sanitization

HANDLING MISSING DATA IN CLINICAL TRIALS

Missing data is a challenging issue in clinical trials that can significantly impact the validity, and interpretation of study results. It occurs when expected observations are not available for analysis, potentially introducing bias and reducing the statistical power of the study. Understanding the nature, causes, and implications of missing data is crucial for statisticians, and regulatory bodies involved in clinical trial design, conduct, and evaluation.

The different reasons for data missing in clinical trials are: Patient dropout or loss to follow-up, Adverse events leading to treatment discontinuation and, missed study visits or procedures.

The different types of missing data are given in the table below [1]:

Type of Missing Data	Description	Example
Missing Completely at Random (MCAR)	The missingness is unrelated to any observed or unobserved variables	A lab sample is accidentally dropped and cannot be analyzed; patients move to another city due to non-health related reasons
Missing at Random (MAR)	The probability of missingness depends on observed data but not on unobserved data	Older patients are more likely to miss follow-up visits, but age is recorded
Missing Not at Random (MNAR)	The probability of missingness depends on unobserved data	Patients discontinue treatment due to lack of efficacy or side effects

The most common statistical methods of handling the missing data are complete-case analysis (CCA), single imputation methods (e.g., last observation carried forward), multiple imputation. Traditional statistical methods bring certain disadvantages such as generation of biased results, reduced statistical power due to decreased sample size, increased complexity in statistical analyses and difficulty with interpreting the results. Using AI will offset these disadvantages as machine learning algorithms, can identify and model complex, non-linear relationships between

variables that traditional methods might miss. AI can handle large volumes of data making them well-suited for large-scale clinical trials or meta-analyses. Many AI algorithms can automatically identify the most relevant features for predicting missing values, potentially using information that might be overlooked in traditional approaches.

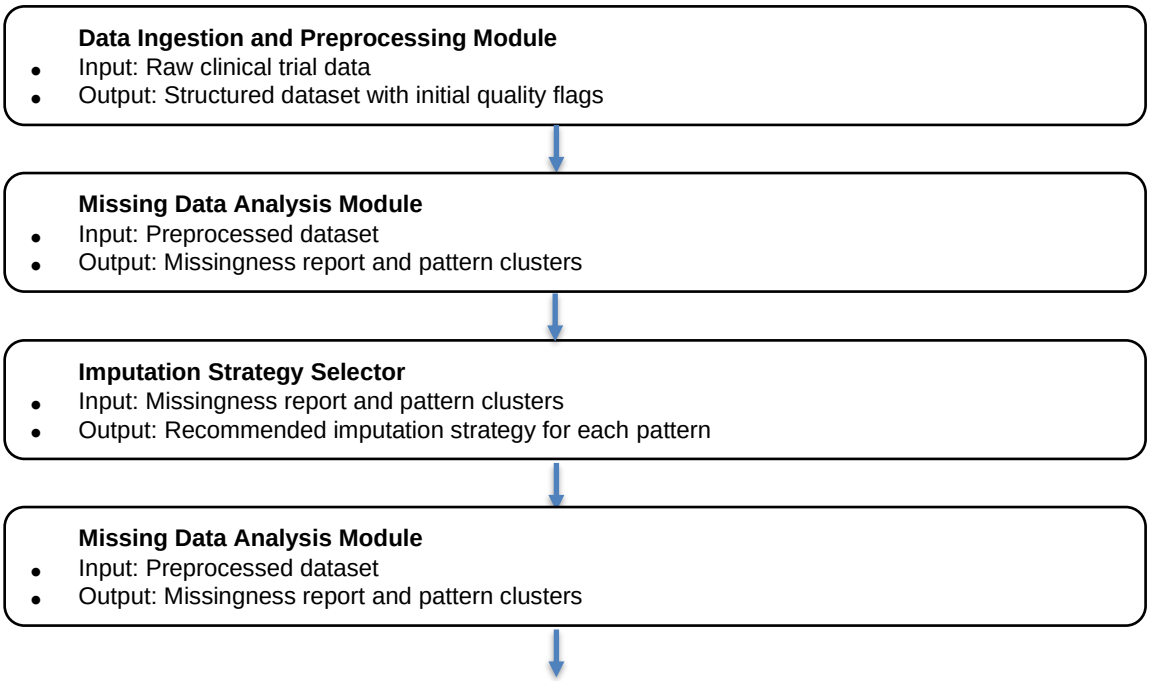
THE DIFFERENT APPROACHES IN AI THAT CAN BE IMPLEMENTED TO HANDLE THE MISSING DATA ARE:

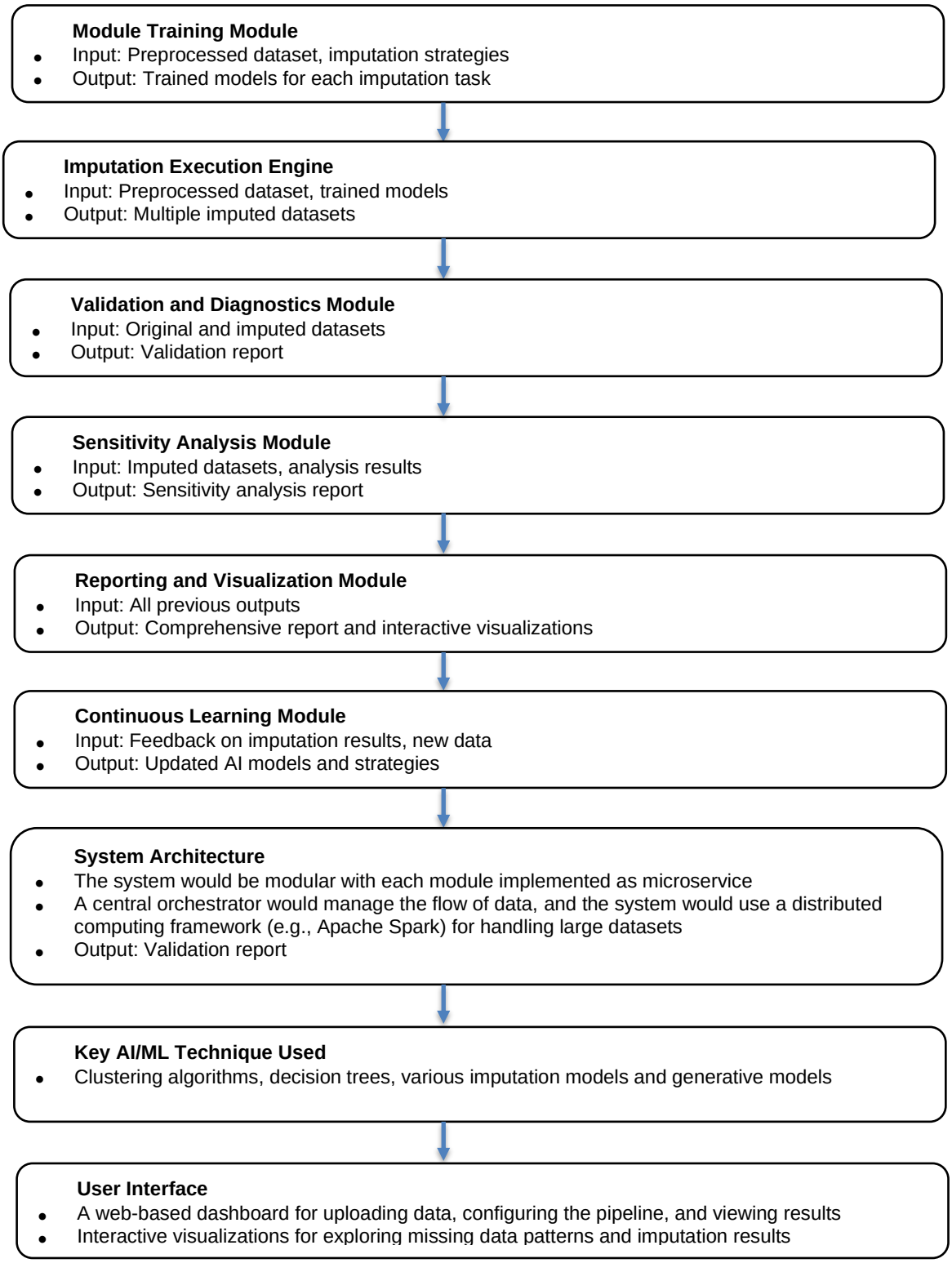
1. Advanced Machine learning models (imputation methods):
   These methods can capture complex relationships between variables that traditional statistical methods might miss. The examples for this ML models are:
   a. Random Forest Imputation used in Alzheimer's Disease Study [2].
   b. Neural Network Imputation for Laboratory Data
   c. Recurrent Neural Networks for Longitudinal Data
   d. Generative Adversarial Networks (GANs) [3]
2. Automated Imputation Method Selection
3. Natural Language Processing (NLP) for Unstructured Data
4. Anomaly Detection in Data Collection
5. Predictive Modeling for Missingness
6. Synthetic Data Generation
7. Time Series Forecasting

FRAMEWORK FOR THE AI MODEL TO IDENTIFY AND RECTIFY MISSING DATA

The framework begins with data exploration and preprocessing, analyzing the structure of input data and identifying variables and data quality issues. Missing data identification follows, scanning the dataset for missing values and identifying patterns of missingness. Missingness pattern analysis uses clustering algorithms to group similar patterns. Imputation strategy selection employs decision trees or ensemble methods to choose appropriate techniques based on variable types and missingness patterns. Model training involves training AI models like random forests or neural networks on complete cases for each type of missing data. Imputation execution applies these trained models to impute missing values, considering time-series specific models for longitudinal data. Uncertainty quantification implements multiple imputation to account for uncertainty. Validation and diagnostics assess imputation accuracy through cross-validation and checks for implausible values. Sensitivity analysis evaluates the impact of imputations on study conclusions. The process concludes with documentation and reporting, generating comprehensive reports on missing data patterns, imputation methods, and their impact. Finally, continuous learning updates the AI system based on feedback and new data to improve future imputations.

INTERGRATION BETWEEN AI AND EDC




Module Training Module
Input: Preprocessed dataset, imputation strategies
Output: Trained models for each imputation task
Imputation Execution Engine
Input: Preprocessed dataset, trained models
Output: Multiple imputed datasets
Validation and Diagnostics Module
Input: Original and imputed datasets
Output: Validation report
Sensitivity Analysis Module
Input: Imputed datasets, analysis results
Output: Sensitivity analysis report
Reporting and Visualization Module
Input: All previous outputs
Output: Comprehensive report and interactive visualizations
Continuous Learning Module
Input: Feedback on imputation results, new data
Output: Updated AI models and strategies
System Architecture
The system would be modular with each module implemented as microservice
A central orchestrator would manage the flow of data, and the system would use a distributed computing framework (e.g., Apache Spark) for handling large datasets
Output: Validation report
Key AI/ML Technique Used
Clustering algorithms, decision trees, various imputation models and generative models
User Interface
A web-based dashboard for uploading data, configuring the pipeline, and viewing results
Interactive visualizations for exploring missing data patterns and imputation results

POSSIBILITY OF USING AI IN REAL WORLD

The implementation of advanced AI models in clinical trials, particularly for handling missing data, shows significant promise. These models leverage existing technologies and modular design, offering potential solutions to longstanding challenges in data management. There is a real possibility for practical application through a phased implementation approach. This could begin with non-critical analyses, employing a hybrid approach that combines AI with traditional methods. Initially focusing on early-phase trials or specific therapeutic areas could pave the way for broader adoption. The success of these AI models likely depends on their continuous improvement and ability to provide valuable insights in a supporting role. As they prove their reliability and value in the clinical trial landscape, their use could gradually expand, potentially revolutionizing data handling in clinical research..

DEDUPLICATION

Deduplication in clinical trials is a crucial process for maintaining data quality and integrity by identifying and removing duplicate or redundant data entries from datasets. It serves to ensure accuracy, prevent result skewing, and enhance overall data quality. This process is commonly applied to patient records, adverse event reports, laboratory results, and medication logs. Duplicates can be exact (identical entries) or near-duplicates (entries with slight variations). Deduplication methods include automated systems, manual review, or a combination of both. Challenges include distinguishing true duplicates from legitimate repeated events, handling data from multiple sources, and dealing with data entry errors. Best practices involve implementing robust data entry systems, establishing clear guidelines, documenting decisions, and conducting regular quality checks. Deduplication is vital for data analysis as it ensures accurate patient counts and event frequencies, prevents overestimation of treatment effects or safety concerns, and contributes to the overall reliability of study results. Ultimately, deduplication is essential for maintaining data integrity in clinical trials, ensuring that final analyses are based on accurate and clean data.

DEDUPLICATION CATEGORIES AND APPLICABLE AI MODELS

Category	AI Models	Data type	Nature of Duplication	Examples
Structured Numerical and Categorical Data	ML Classifiers (Random Forest, SVM, XGBoost)	Patient demographics, lab results, vital signs	Exact matches, near-matches with slight variations	Random Forest can handle mix of numerical and categorical data e.g.: age, weight
Text-based Data	NLP Models (BERT, GPT variants, Word2Vec) [4]	Medical histories, adverse event descriptions, free-text fields	Semantic duplicates, paraphrased information	BERT can understand context and semantics, making it ideal for identifying duplicates in AE reports for same event with different words [5]
Time Series Data	Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks	Continuous patient monitoring data, medication adherence logs	Repeated patterns, overlapping time periods	LSTM networks can detect duplicates in time-series data like ECG
Image Data	Convolutional Neural Networks (CNNs), Siamese Neural Networks	Medical imaging (X-rays, MRIs, CT scans)	Visually similar images, same image with different metadata	CNNs can identify duplicate or highly similar medical images, even if they've been slightly modified or saved with different file names
Network or Relationship Data	Graph Neural Networks (GCN, GraphSAGE)	Patient-doctor interactions, medication prescription patterns	Similar patterns in relationship networks	Graph Neural Networks can identify duplicate patterns in patient-medication networks, where the same prescription pattern might be recorded multiple times

DATA AUGMENTATION

Data augmentation in clinical trials is a set of techniques used to increase data quantity and diversity without collecting new patient data. It aims to enhance statistical power, improve analysis robustness, potentially reduce required patient numbers, and explore additional hypotheses. Common techniques include synthetic data generation, bootstrapping, data simulation, and imputation. Data augmentation is applied in exploratory analyses, hypothesis generation, sample size calculations, sensitivity analyses, and machine learning model training. While it offers advantages like potential cost and time reduction in patient recruitment and allows exploration of rare events or subgroups, it has limitations. Augmented data should not replace real patient data for primary analyses, requires careful validation, and must comply with regulatory requirements and ethical considerations. Emerging trends include the use of AI for more sophisticated techniques, integration with real-world data sources, and application in adaptive trial designs. Overall, data augmentation should be used judiciously to ensure the validity and integrity of clinical trial results.

AI MODELS TO HANDLE DATA AUGMENTATION

- Generative Adversarial Networks (GANs) [6]: GANs consist of two neural networks: a generator and a discriminator. The generator creates synthetic data, while the discriminator tries to distinguish between real and synthetic data. Example: Generating synthetic patient profiles for rare diseases, including complex interactions between variables.
- Variational Autoencoders (VAEs): VAEs learn a compressed representation of the data and can generate new samples from this representation. Example: Creating synthetic medical imaging data, such as MRI scans or X-rays, to augment limited datasets in imaging studies.
- Sequence Models: Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks: Excellent for generating time-series data and can capture temporal dependencies in clinical data. Example: Simulating patient trajectories over time, including disease progression and treatment response.
- Transformer models: This model is best for NLP tasks, can be adapted to complex, long-range dependencies. Example: Generating synthetic electronic health records (EHRs) that maintain realistic temporal relationships between events.
- Conditional Generation: This involves generating synthetic data based on specific conditions or attributes. Example: Creating synthetic patient data for a specific demographic group or disease subtype that may be underrepresented in the original dataset.
- Multi-modal Data Augmentation: can handle multiple types of data simultaneously (e.g., numerical, categorical, text, and images). Example: Generating synthetic patient profiles that include both structured data (lab results, vital signs) and unstructured data (medical notes, imaging reports).
- Reinforcement Learning for Data Augmentation: The model learns to generate data that maximizes certain metrics. Example: Creating synthetic clinical trial data that maintains the statistical properties of the original data while increasing diversity.
- Transfer Learning and Pre-trained Models: Leveraging models pre-trained on large datasets and fine-tuning them for specific clinical trial data, particularly useful when working with limited data. Example: Using a model pre-trained on a large EHR dataset to generate synthetic data for a specific rare disease study.
- Differential Privacy in Data Augmentation: Incorporating differential privacy techniques into the data generation process to ensure patient privacy, crucial when working with sensitive medical data. Example: Generating synthetic patient data that maintains statistical utility while providing strong privacy guarantees.
- Explainable AI for Data Augmentation: Developing AI models that not only generate synthetic data but also provide explanations for their decisions, can help in validating the generated data and building trust in the augmentation process[6]. Example: A model that generates synthetic patient data and provides explanations for why certain values were chosen, based on learned patterns from the original data.
- Active Learning for Data Augmentation: Using active learning techniques to intelligently select which data points to augment. Example: Identifying underrepresented patient subgroups in a clinical trial and selectively generating synthetic data for these subgroups.
- Federated Learning for Data Augmentation: Leveraging federated learning techniques to create augmented data across multiple institutions without sharing raw patient data. Example: Multiple hospitals collaboratively training a data augmentation model without sharing their patient data directly. These advanced AI and machine learning techniques for data augmentation offer powerful tools for clinical researchers.

However, it's crucial to remember that:

1. The use of these techniques should be transparent and well-documented.
2. Rigorous validation of the augmented data is necessary to ensure its quality and relevance.
3. Regulatory considerations must be considered, especially regarding the use of synthetic data in regulatory submissions.
4. Ethical considerations, particularly around patient privacy and consent, should always be at the forefront.

ERROR CORRECTION

Error correction in clinical trials is a critical process that ensures data integrity, regulatory compliance, patient safety, and scientific validity of study results. It involves identifying, documenting, and rectifying various types of errors, such as data entry mistakes, protocol deviations, and inconsistencies. The process employs multiple detection methods, including automated checks, manual reviews, and statistical analyses. Best practices include implementing robust data management plans, using electronic data capture systems, conducting regular quality reviews, and following Good Clinical Practice guidelines. While challenges exist, such as balancing timely corrections with data integrity and managing errors across multiple sites, the use of advanced technologies like artificial intelligence and real-time monitoring tools is helping to enhance error correction processes in clinical trials.

AI MODELS TO IDENTIFY AND RECTIFY THE ERRORS

An overview of the key components of modern error correction strategies, from detection and classification to implementation and documentation, while highlighting the integration of cutting-edge technologies and best practices to ensure data integrity and minimize human intervention are mentioned below.

Comprehensive Error Correction Process:

- Error Detection through automated data validation checks and cross-referencing multiple data sources
- Error Classification into critical, major, and minor errors
- Root Cause Analysis to identify the source and determine if systematic or isolated
- Correction Planning based on error type and root cause
- Implementation of Corrections in the database and source documents if necessary
- Quality Control to verify accurate implementation
- Documentation of all changes, rationale, and impact

Advanced Error Correction Methodologies:

- Risk-Based Approach prioritizing based on potential impact
- Statistical Methods for imputation and systematic error correction
- Machine Learning Algorithms for predictive error flagging
- Natural Language Processing for free-text field analysis.

Best Practices:

- Standardization through SOPs
- Training and Education for staff
- Technology Integration using CTMS
- Real-Time Monitoring
- Cross-Functional Collaboration
- Transparency with regulatory authorities

Regulatory Considerations:

- Compliance with ICH GCP E6(R2)
- Data Integrity preservation
- Audit Trail maintenance

Advanced Technologies:

- Artificial Intelligence for pattern recognition and anomaly detection
- Blockchain for data immutability and traceability
- Internet of Things for real-time data collection and error detection

Minimize Human Intervention:

- Implement confidence scoring for AI-suggested corrections
- Automate high-confidence, low-risk error corrections
- Use AI to prioritize errors for human review.

REAL WORLD DATA SANITIZATION

Real world data sanitization in clinical trials is a crucial process that prepares real-world data for research use while safeguarding patient privacy and ensuring data integrity. It involves de-identification of personal information, data cleaning to correct errors and inconsistencies, data transformation to standardize formats, and quality control measures. This process enables researchers to utilize real-world evidence in clinical trials, providing a more comprehensive view of patient outcomes and treatment effectiveness. However, it faces challenges such as balancing data utility with privacy protection and managing diverse data sources. The process employs advanced tools like anonymization algorithms and machine learning techniques, while adhering to regulatory requirements such as GDPR and HIPAA. As the importance of real-world data in clinical research grows, effective sanitization processes become increasingly vital for maintaining ethical standards and research validity.

AI-DRIVEN AUTOMATATION OF DATA MASKING

AI-driven data masking involves several key steps and technologies

Identification of Data to Mask using Natural Language Processing

- Pattern Recognition
- Contextual Understanding
- Semantic Analysis
- Anomaly Detection

Determining How to Mask

- Risk Assessment
- Data Utility Preservation
- Contextual Masking
- Synthetic Data Generation

Implementing Masking Techniques

- Automated Redaction
- Data Transformation
- Consistent Pseudonymization
- Dynamic Masking

Continuous Learning and Improvement

- Feedback Loops
- Adaptive Algorithms
- Performance Monitoring

Compliance and Audit

- Regulatory Alignment
- Generating Automated Audit Trails
- Handling Edge Cases and Human Oversight with Confidence Scoring and Exception Handling

These AI techniques enable sophisticated identification of sensitive data, context-aware masking decisions, automated implementation of masking techniques, ongoing improvement of the masking process, regulatory compliance, and appropriate human intervention when needed. The AI system can analyze unstructured text, recognize patterns, understand context, assess risks, preserve data utility, apply various masking techniques, adapt to new privacy risks, align with regulations, and flag complex cases for human review, all while maintaining detailed audit trails.

CHALLENGES OF USING AI IN THE REAL WORLD

- Despite the potential benefits, the implementation of AI in clinical trials faces numerous significant challenges. A key hurdle is gaining regulatory approval. Bodies like the FDA and EMA are traditionally cautious towards new methods, especially concerning the "black box" nature of some AI algorithms. This necessitates extensive validation efforts to ensure compliance with Good Clinical Practice guidelines
- Data privacy and protection present another major concern. Ensuring patient confidentiality while utilizing vast amounts of data for AI models is crucial. Additionally, the lack of interpretability in advanced AI models poses challenges in explaining decisions to regulatory bodies and stakeholders, which is essential for transparency and trust.
- Data quality and consistency pose significant challenges in clinical trials, particularly when implementing AI systems. The issues of inconsistent data formats, missing or incomplete information, data entry errors, and variations in recording methods (such as differing date formats or name spellings) can severely impact the accuracy and reliability of AI models, potentially leading to flawed analyses and compromised research outcomes
- Technical challenges are also substantial. These include the integration of AI systems with existing infrastructure, ensuring scalability, and the need for significant computational resources. Standardizing data across diverse sources and maintaining data quality are critical for the effective functioning of AI models.
- The implementation process is further complicated by the need to ensure model generalizability across different therapeutic areas and patient populations. Continuous learning and updates are necessary to keep the models relevant and accurate, adding another layer of complexity.
- Navigating legal challenges related to data ownership and intellectual property rights of AI-generated data presents a complex hurdle in the implementation of AI systems for clinical trials. Additionally, the potential liability issues associated with using AI-augmented data in clinical decision-making raise important questions about responsibility and accountability, requiring careful consideration and robust legal frameworks to protect all stakeholders involved in the research process.
- From an organizational perspective, gaining acceptance from various stakeholders, justifying the costs associated with AI implementation, and bridging the knowledge gap between AI experts and clinical trial professionals present significant hurdles. Extensive training and expertise are required to effectively utilize these advanced systems.

- Lastly, handling rare events, ensuring auditability and reproducibility of results, and addressing the ethical implications of AI-driven decision-making in clinical trials are challenges that need careful consideration and resolution.

CONCLUSION

AI is revolutionizing various aspects of clinical trials, offering advanced solutions for data deduplication, sanitization, augmentation, and handling missing data. These AI-driven approaches enhance data quality, improve efficiency, and accelerate research outcomes while addressing complex privacy challenges. Through sophisticated algorithms and machine learning techniques, AI enables more accurate and consistent processing of large volumes of diverse data types, facilitates the generation of high-quality synthetic data, and implements advanced privacy-preserving techniques. Moreover, AI's ability to capture complex, non-linear relationships in datasets allows for more accurate imputations and real-time data processing, potentially preventing missing data before it occurs.

While the potential benefits of AI in clinical trials are compelling, significant challenges remain across technical, ethical, and regulatory domains. These include ensuring regulatory compliance, model interpretability, and robust validation. Successfully implementing AI in clinical research requires a multidisciplinary approach, combining expertise in data science, clinical research, and regulatory compliance. As the field evolves, continuous innovation in AI algorithms, data handling techniques, and privacy-preserving methods will be crucial. Equally important is the development of clear regulatory frameworks and industry standards to guide the responsible use of AI in this context. The future likely lies in hybrid approaches that combine AI's power with traditional statistical rigor, gradually integrating AI methods into clinical trials. With ongoing collaboration between data scientists, clinicians, statisticians, and regulatory bodies, AI has the potential to become a standard tool in clinical trial data management, ultimately contributing to more efficient and effective medical research.

REFERENCES

1. Inference and missing data https://doi.org/10.1093/biomet/63.3.581
2. Random forest model for feature-based Alzheimer's disease conversion prediction from early mild cognitive impairment subjects doi: 10.1371/journal.pone.0244773
3. Cancer diagnosis using generative adversarial networks based on deep learning from imbalanced data https://doi.org/10.1016/j.compbiomed.2021.104540
4. Increased Confidence in Deduplication of Drug Safety Reports with Natural Language Processing of Narratives at the US Food and Drug Administration https://doi.org/10.3389/fdsfr.2022.918897
5. medBERT.de: A comprehensive German BERT model for the medical domain https://doi.org/10.1016/j.eswa.2023.121598
6. GANerAid: Realistic synthetic patient data for clinical trials https://doi.org/10.1016/j.imu.2022.101118
7. Efficient automated error detection in medical data using deep-learning and label-clustering
8. ParexelGPT, an AI assistant developed by the Clinical Research Organization Parexel.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Mounika Pillamarapu & Deepak Pakalapati
Company: PAREXEL International (Bengaluru), PAREXEL International (Hyderabad)
Email: mounikaramesh.pillamarapu@parexel.com, deepak.pakalapati@parexel.com