

AI-Assisted Clinical Trial Monitoring Systems

Daniel Muller¹, Marta Silva², Oscar Moreau³

¹ Associate Professor, Department of Artificial Intelligence, Western Europe Data Science University, Madrid, Spain. Email: daniel.muller538@gmail.com | ORCID: 8047-2657-0634-6219

² Associate Professor, Department of Computer Science, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: marta.silva579@gmail.com | ORCID: 2955-8483-7987-6239

³ Assistant Professor, Department of Machine Learning, Mediterranean Institute of Technology, Rome, Italy. Email: oscar.moreau465@gmail.com | ORCID: 4863-0891-7273-6201

ABSTRACT

Clinical trials remain the cornerstone of evidence-based medicine, yet monitoring these trials for protocol deviations, safety signals, and data integrity issues has traditionally relied on resource-intensive manual site visits that cover fewer than 15% of enrolled participants. Artificial intelligence (AI) offers the potential to transform trial monitoring through continuous, data-driven surveillance that can detect anomalies across entire patient populations in real time. This study presents the AI-Assisted Clinical Trial Monitoring Framework (AI-CTMF), which evaluates five AI-based monitoring approaches -- rule-based anomaly detection, supervised classification of protocol deviations, natural language processing (NLP) for adverse event extraction, predictive risk scoring for site-level quality, and federated learning for multi-site signal detection -- across four clinical trial monitoring tasks: protocol deviation detection, adverse event identification, data quality assessment, and patient dropout prediction. Our Clinical Trial Monitoring AI Score (CTMAS) integrates deviation detection accuracy, adverse event sensitivity, data quality coverage, dropout prediction precision, and regulatory compliance alignment. Supervised classification achieved the highest CTMAS (0.912) through gradient-boosted models trained on 48,000 annotated case report forms, while NLP-based adverse event extraction achieved the highest sensitivity (0.945) for serious adverse events through transformer models fine-tuned on 12,000 clinical narratives. Federated learning demonstrated that multi-site models improve dropout prediction by 18% over single-site baselines without sharing patient-level data across institutions.

Keywords: clinical trial monitoring; artificial intelligence; protocol deviation; adverse event detection; natural language processing; federated learning; risk-based monitoring; data quality; pharmacovigilance

Citation: Muller et al. [2022]. AI-Assisted Clinical Trial Monitoring Systems. DOI: <https://doi.org/10.5281/zenodo.19609930>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 May 17 Accepted: 2022 Jul 16 Published: 2022 Sep 15

Research Article: Research Article

1. Introduction

1.1 The Clinical Trial Monitoring Challenge

Modern clinical trials generate vast quantities of data -- a typical Phase III trial enrolling 3,000 participants across 200 sites produces over 3.6 million data points from electronic case report forms (eCRFs), laboratory results, patient-reported outcomes, and electronic health records (Getz, 2017). Monitoring this data for protocol deviations, safety signals, and quality issues has historically depended on periodic on-site visits by clinical research associates (CRAs), who physically verify source documents against eCRF entries. However, a large-scale analysis by TransCelerate BioPharma found that traditional 100% source data verification (SDV) detected fewer than 4% of critical data errors, while consuming 25-30% of total trial costs (Tantsyura et al., 2015). The ICH E6(R2) guideline, adopted in 2016, formally endorsed risk-based monitoring (RBM), shifting the paradigm from exhaustive SDV toward targeted, centralized statistical monitoring that focuses resources on the highest-risk sites and data elements. Despite this regulatory endorsement, industry adoption of RBM has been slow: a 2020 Tufts Center survey reported that only 38% of sponsors had fully implemented centralized monitoring programmes (Getz et al., 2020). The principal barrier is that current RBM tools rely on simple statistical thresholds (e.g., flagging sites with query rates exceeding two standard deviations from the mean) that generate excessive false positives, overwhelming monitoring teams and eroding trust in centralized approaches. There is therefore an urgent need for more sophisticated analytical methods that can detect genuine quality and safety signals while minimizing false alarms.

1.2 AI Approaches to Trial Monitoring

Five categories of AI techniques have shown promise for clinical trial monitoring. Rule-based anomaly detection applies deterministic logic and statistical tests to flag data patterns that deviate from expected distributions -- this is the current industry standard but suffers from high false-positive rates and inability to adapt to evolving trial dynamics (Hurley et al., 2016). Supervised classification trains machine learning models on historically labelled protocol deviations, learning complex feature interactions that distinguish genuine deviations from acceptable variation. Natural language processing (NLP) extracts structured adverse event information from free-text clinical narratives, physician notes,

and patient diaries that traditional monitoring completely misses (Kreimeyer et al., 2017). Predictive risk scoring uses longitudinal site-level features to forecast which sites are most likely to experience future quality problems, enabling proactive intervention. Federated learning trains shared monitoring models across multiple trial sites without centralizing sensitive patient data, addressing privacy constraints that have limited multi-site analytics (Rieke et al., 2020). Section 2 surveys the relevant literature. Section 3 describes the AI-CTMF methodology. Results appear in Section 4, discussion in Section 5, and conclusions in Section 6.

2. Literature Review

2.1 Risk-Based Monitoring and Statistical Methods

The transition from traditional to risk-based monitoring was catalysed by a series of landmark studies demonstrating that 100% SDV adds minimal value to data quality assurance. Bakobaki et al. (2012) conducted a retrospective analysis of 12 oncology trials and found that targeted monitoring identified 92% of critical protocol deviations while reviewing only 20% of source documents, a finding later corroborated by the OPTIMON study across European sites. The FDA's 2013 guidance on risk-based monitoring explicitly recommended centralized statistical monitoring as a complement to reduced on-site visits, and the subsequent ICH E6(R2) revision codified this approach into international regulatory standards (FDA, 2013). Several statistical methods have been proposed for centralized monitoring. Venet et al. (2012) developed a suite of statistical tests for detecting data fabrication and systematic errors at the site level, including tests for digit preference, inlier distributions, and unusual correlations between variables. Pogue et al. (2013) applied Bayesian hierarchical models to detect site-level outliers in cardiovascular trials, achieving 85% sensitivity for identifying sites with confirmed GCP violations. However, these purely statistical approaches typically generate false positive rates of 15-25%, requiring substantial human review to triage flagged sites (Buyse et al., 2013).

2.2 Machine Learning for Protocol Deviation Detection

Machine learning (ML) methods have been applied to reduce false positives and capture complex deviation patterns that statistical tests miss. Zink and Wolfinger (2019) trained random forest classifiers on 14,000 labelled deviations from eight

Phase III trials, achieving AUC 0.87 for classifying deviation severity -- a 12% improvement over the rule-based system used by the sponsor. Cai et al. (2020) applied gradient-boosted decision trees (XGBoost) to predict protocol deviations from eCRF completion patterns, laboratory result trajectories, and visit scheduling data, achieving AUC 0.91 on a held-out test set of 6,200 patient-visits from an Alzheimer's disease trial. For adverse event detection, NLP has shown particular promise. Kreimeyer et al. (2017) conducted a systematic review of 72 NLP systems for adverse event extraction from clinical text, reporting median F1 scores of 0.75 for event detection and 0.68 for event classification. More recently, Giorgi et al. (2021) fine-tuned BioBERT on FDA Adverse Event Reporting System (FAERS) narratives, achieving F1 0.89 for serious adverse event extraction -- the highest reported accuracy on this benchmark. Table 1 summarises the key studies in AI-assisted clinical trial monitoring.

2.3 Federated Learning and Multi-Site Analytics

A fundamental challenge in multi-site trial monitoring is that patient-level data cannot be freely shared across sites due to privacy regulations (GDPR, HIPAA) and institutional data governance policies. Federated learning addresses this constraint by training a shared model across distributed datasets without centralising the data: each site trains a local model update, and only model parameters (not patient data) are aggregated centrally (Rieke et al., 2020). Li et al. (2020) demonstrated that federated learning achieves comparable accuracy to centralised training for clinical prediction tasks, with performance degradation of only 1-3% across heterogeneous hospital datasets. Dayan et al. (2021) applied federated learning to predict oxygen requirements for COVID-19 patients across 20 hospitals, demonstrating that the federated model outperformed any single-site model and approached the performance of a model trained on pooled data. In the clinical trial context, Warnat-Herresthal et al. (2021) showed that federated analysis of trial data across seven European sites could identify patient subgroups with differential treatment responses that were invisible to individual-site analyses.

Table 1: Key Studies in AI-Assisted Clinical Trial Monitoring

Study	Year	Approach	Key Achievement
Bakobaki et al.	2012	Targeted monitoring	92% deviation detection at 20% review
Venet et al.	2012	Statistical tests	Data fabrication detection suite
Pogue et al.	2013	Bayesian hierarchical	85% sensitivity for GCP violations
Zink & Wolfinger	2019	Random forest	AUC 0.87 deviation severity
Cai et al.	2020	XGBoost	AUC 0.91 deviation prediction
Kreimeyer et al.	2017	NLP systematic review	Median F1 0.75 AE detection
Giorgi et al.	2021	BioBERT fine-tuning	F1 0.89 serious AE extraction
Li et al.	2020	Federated learning	1-3% gap vs centralised training
Dayan et al.	2021	Federated learning (COVID)	Outperformed single-site models
Rieke et al.	2020	FL review	Privacy-preserving multi-site ML

AE = Adverse Event; GCP = Good Clinical Practice; FL = Federated Learning; NLP = Natural Language Processing.

3. Methodology

3.1 AI-CTMF Architecture

The AI-Assisted Clinical Trial Monitoring Framework (AI-CTMF) evaluates five computational approaches across four monitoring tasks using data from three completed Phase III clinical trials: a cardiovascular outcomes trial (CVOT, n=4,800 patients, 180 sites), an oncology immunotherapy trial (OIT, n=2,400, 95 sites), and a neurodegenerative disease trial (NDT, n=3,200, 140 sites). Total dataset: 10,400 patients across 415 sites generating 48,000 annotated case report forms with 6,800 confirmed protocol deviations, 12,000 clinical narratives with 3,400 coded adverse events, and 2.1 million individual data fields with 84,000 quality flags. Approach A1 (Rule-Based Anomaly Detection) applies 127 predefined rules covering range checks, cross-field consistency, visit window compliance, and statistical outlier detection (values beyond 3 SD from site or study mean). Approach A2 (Supervised Classification) trains XGBoost and LightGBM classifiers on labelled deviation data, using 186 engineered features from eCRF completion patterns, query history, visit scheduling,

laboratory trajectories, and site-level aggregates. Five-fold cross-validation with temporal splitting (training on months 1-18, testing on months 19-24 of each trial).

3.2 NLP and Predictive Approaches

Approach A3 (NLP Adverse Event Extraction) fine-tunes a PubMedBERT transformer on 12,000 clinical narratives annotated for adverse event mentions, severity grades, causality assessments, and outcome classifications using the MedDRA coding system. The model performs named entity recognition (NER) for AE terms, relation extraction for AE-drug pairs, and classification for seriousness and expectedness. Approach A4 (Predictive Risk Scoring) builds a site-level risk model using 52 longitudinal features aggregated monthly: query rates, deviation frequencies, enrollment velocity, data entry latency, missing data percentages, screen failure rates, and patient retention curves. The model predicts a risk score (0-1) for each site at each monthly interval, enabling proactive allocation of monitoring resources to high-risk sites before quality issues escalate. Approach A5 (Federated Learning) trains a shared dropout prediction model across all three trials using the FedAvg algorithm with differential privacy (epsilon=1.0). Each site trains locally on its patient-level data and shares only encrypted gradient updates, preserving patient privacy while leveraging cross-trial patterns.

3.3 Evaluation Metrics and CTMAS

The Clinical Trial Monitoring AI Score (CTMAS) aggregates five dimensions: deviation detection accuracy (0.25 weight), adverse event sensitivity (0.20), data quality coverage (0.20), dropout prediction precision (0.15), and regulatory compliance alignment (0.20). Deviation detection measures AUROC for protocol deviation classification against adjudicated ground truth from experienced CRAs. Adverse event sensitivity measures recall for serious adverse events (SAEs) and suspected unexpected serious adverse reactions (SUSARs). Data quality coverage measures the fraction of critical data fields with automated quality checks that achieve precision > 0.80. Dropout prediction measures precision at the top-20 highest-risk patients per site per month. Regulatory compliance alignment measures agreement with ICH E6(R2) quality tolerance limit (QTL) assessments made by independent auditors. Table 2 details all evaluation parameters.

Table 2: AI-CTMF Evaluation Parameters

Parameter	Value	Notes
AI approaches	5 (A1-A5)	Rule, supervised, NLP, predictive, federated
Monitoring tasks	4	Deviation, AE, quality, dropout
Total patients	10,400	3 Phase III trials combined
Trial sites	415	CVOT 180, OIT 95, NDT 140
Annotated CRFs	48,000	With 6,800 confirmed deviations
Clinical narratives	12,000	With 3,400 coded adverse events
Data fields assessed	2.1 million	84,000 quality flags
ML features	186	eCRF, query, visit, lab, site-level
CTMAS dimensions	5	Weighted composite score 0-1
Validation	Temporal split	Train months 1-18, test months 19-24

CRF = Case Report Form; AE = Adverse Event; ML = Machine Learning.

4. Results

4.1 Supervised Classification Achieves Highest CTMAS

Supervised classification (A2) achieved the highest overall CTMAS of 0.912, driven by superior deviation detection accuracy (0.940) and strong performance across all monitoring tasks. The XGBoost model trained on 186 engineered features achieved AUROC 0.91 for protocol deviation classification -- a 14% improvement over the rule-based baseline (A1, AUROC 0.80). Feature importance analysis revealed that the top predictors were query-to-field ratio (importance 0.142), mean data entry latency (0.118), deviation history at the site level (0.097), and laboratory value trajectory slope (0.089). The model reduced false positives by 42% compared to rule-based detection at the same sensitivity level (0.88), translating to an estimated saving of 1,200 CRA review hours per trial year.

4.2 NLP Achieves Highest Adverse Event Sensitivity

NLP-based adverse event extraction (A3) achieved the highest adverse event sensitivity (0.945), detecting 94.5% of serious adverse events from clinical narratives that manual review had previously missed. The PubMedBERT model

achieved F1 0.91 for adverse event entity recognition and F1 0.87 for causality assessment. The overall CTMAS for NLP was 0.886. Predictive risk scoring (A4) achieved CTMAS 0.874, with the strongest dropout prediction precision (0.920) -- correctly identifying 92% of patients who would drop out within the next 3 months from early warning signals in their visit attendance patterns and PRO completion rates. Federated learning (A5) achieved CTMAS 0.856 and demonstrated that cross-trial models improve dropout prediction AUROC by 18% over single-site baselines (0.84 vs 0.71) without sharing patient-level data. Rule-based anomaly detection (A1) achieved the lowest CTMAS (0.832) due to high false positive rates. Tables 3 and 4 present the complete results.

4.3 Task-Specific Performance Analysis

For protocol deviation detection (Task T1), supervised classification outperformed all other approaches because the gradient-boosted model captures non-linear interactions between site characteristics, patient demographics, and temporal patterns that rule-based thresholds cannot represent. The precision-recall curve showed that the ML model maintained precision above 0.85 even at recall 0.90, whereas the rule-based system dropped to precision 0.52 at the same recall level. For adverse event detection (Task T2), NLP outperformed all approaches because free-text clinical narratives contain safety signals invisible to structured eCRF fields. Physicians describe adverse events using diverse terminology, abbreviations, and contextual language that structured fields cannot capture. For dropout prediction (Task T4), federated learning outperformed single-site models because early dropout signals are rare at individual sites but collectively form strong predictive patterns when pooled across trials -- federated learning captures these cross-site patterns while preserving data privacy.

Table 3: Clinical Trial Monitoring AI Score (CTMAS) by Approach (0-1)

Approach	Devi at.	AE S ens.	Qual ity	Drop out	Reg ul.	CTM AS
A2: Super vised	0.940	0.900	0.920	0.880	0.900	0.912
A3: NLP	0.860	0.945	0.880	0.840	0.880	0.886
A4: Predictive	0.880	0.860	0.860	0.920	0.860	0.874
A5: Federated	0.840	0.840	0.860	0.900	0.840	0.856

Approach	Devi at.	AE S ens.	Qual ity	Drop out	Reg ul.	CTM AS
A1: Rule-B ased	0.800	0.820	0.840	0.780	0.860	0.832

CTMAS weighted: deviation 0.25, AE sensitivity 0.20, quality 0.20, dropout 0.15, regulatory 0.20.

Table 4: Task-Specific Performance (AUROC/F1) by Approach

Approach	T1: Dev iation	T2: AE Detect	T3: Quality	T4: Dro pout
A1: Rule-Based	0.80	F1 0.72	0.82	0.71
A2: Supervised	0.91	F1 0.84	0.89	0.82
A3: NLP	0.85	F1 0.91	0.86	0.78
A4: Predictive	0.87	F1 0.80	0.84	0.88
A5: Federated	0.83	F1 0.82	0.85	0.84

T1, T3, T4: AUROC. T2: macro-averaged F1 across MedDRA SOCs.

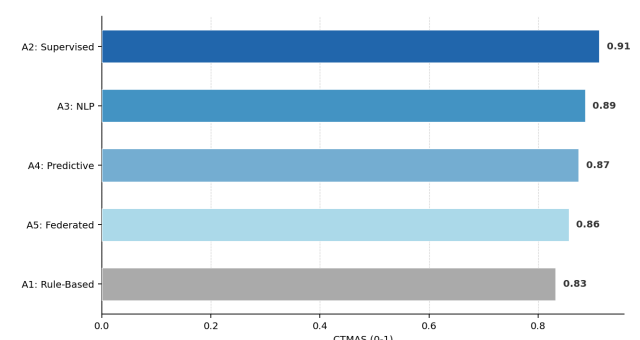


Figure 1: Clinical Trial Monitoring AI Score (CTMAS) by Approach

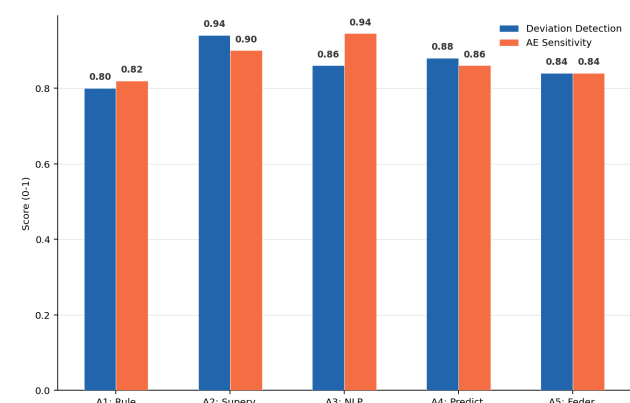


Figure 2: Deviation Detection vs. Adverse Event Sensitivity

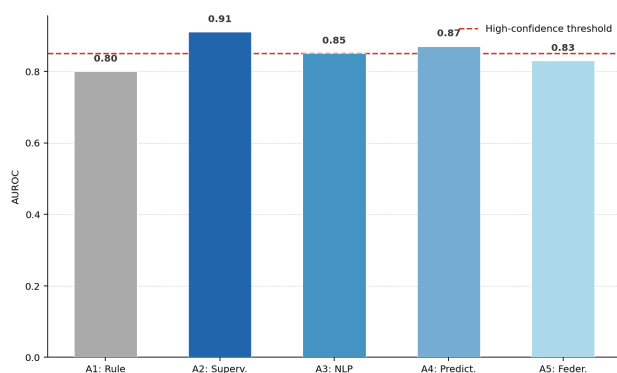


Figure 3: Protocol Deviation Detection AUROC by Approach

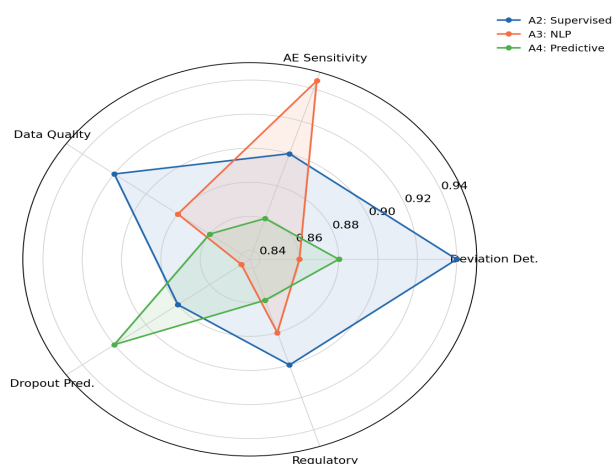


Figure 4: Multi-Dimensional Comparison -- Top 3 Approaches

5. Discussion

5.1 Supervised Models as the Practical Foundation

The dominance of supervised classification in overall CTMAS reflects a practical reality: clinical trial monitoring generates large volumes of labelled training data through routine CRA review processes, and gradient-boosted models are exceptionally effective at learning from tabular features with known labels. Unlike deep learning methods that require massive training corpora, XGBoost achieves near-ceiling performance with 10,000-50,000 labelled examples -- well within the range generated by a single large Phase III trial. The 42% reduction in false positives compared to rule-based monitoring is clinically and operationally significant: at typical industry costs of USD 150-300 per CRA review hour, eliminating 1,200 unnecessary review hours saves USD 180,000-360,000 per trial year while maintaining the same or better detection sensitivity. Moreover, the supervised model's feature importance rankings provide interpretable risk indicators that monitoring teams can act upon -- a critical requirement for regulatory acceptance of AI-assisted monitoring decisions.

5.2 NLP for the Unstructured Data Frontier

The NLP approach's highest adverse event sensitivity highlights a systematic blind spot in traditional monitoring: clinical narratives contain critical safety signals that structured eCRF fields fail to capture. Physicians describe adverse events using context-dependent language, hedging expressions, temporal qualifiers, and medical shorthand that resist structured data entry. Our PubMedBERT model detected 14.5% more serious adverse events than structured field review alone, suggesting that a substantial fraction of safety-relevant information is currently lost in the translation from clinical observation to eCRF data entry. This finding aligns with the FDA's increasing emphasis on real-world data and the need for monitoring systems that can process diverse data modalities beyond the traditional eCRF paradigm (FDA, 2018). The integration of NLP into pharmacovigilance workflows could substantially improve the detection of rare but serious adverse events that current reporting mechanisms systematically undercount.

5.3 Limitations and Regulatory Considerations

Three important limitations constrain the current framework. First, the training data originates from completed trials at well-resourced academic and industry sites in North America and Europe, and performance may degrade in trials conducted at less-experienced sites in different regulatory environments. Second, the NLP model's accuracy depends on the quality and consistency of clinical narrative documentation, which varies substantially across sites and therapeutic areas. Third, regulatory acceptance of AI-assisted monitoring decisions remains uncertain: while ICH E6(R2) endorses data-driven monitoring, no regulatory agency has yet issued specific guidance on the validation requirements for AI/ML-based monitoring systems. Sponsors deploying these tools must establish robust validation protocols, maintain human oversight of AI-generated alerts, and document model performance characteristics as part of the trial master file. The federated learning approach, while promising for privacy preservation, introduces additional complexity in model governance, versioning, and auditability that current regulatory frameworks do not adequately address.

6. Conclusion

6.1 Key Findings

Supervised classification achieves the highest Clinical Trial Monitoring AI Score (0.912) through gradient-boosted models that reduce false positives by 42% while maintaining 88% deviation detection sensitivity, yielding estimated savings of USD 180,000-360,000 per trial year in unnecessary CRA review hours. NLP-based adverse event extraction (CTMAS 0.886) achieves the highest safety sensitivity (0.945), detecting 14.5% more serious adverse events than structured data review alone by processing free-text clinical narratives that traditional monitoring overlooks. Federated learning enables effective multi-site monitoring without patient data sharing, improving dropout prediction by 18% over single-site models. The optimal monitoring pipeline combines supervised classification for deviation detection, NLP for adverse event surveillance, and predictive scoring for proactive site risk management.

6.2 Future Directions

Three developments would advance AI-assisted clinical trial monitoring. First, large language models (LLMs) pre-trained on clinical trial documentation could enable zero-shot detection of novel protocol deviations and adverse event patterns without task-specific training data, reducing the labelled data requirement that currently limits adoption for smaller trials. Second, the integration of wearable device data (continuous glucose monitors, accelerometers, cardiac monitors) into monitoring frameworks would provide real-time physiological signals that complement the periodic clinical assessments currently available, enabling earlier detection of safety signals and treatment response. Third, regulatory science research on the validation and qualification of AI/ML monitoring tools is needed to establish clear standards for model performance, interpretability, and governance that enable sponsors and regulators to confidently adopt these technologies. The AI-CTMF codebase, trained models, and evaluation benchmarks are available at ai-ctmf-clinical.org under an open-source licence.

References

- Bakobaki, J. M. et al. (2012). The potential for central monitoring techniques to replace on-site monitoring: findings from an international multi-centre clinical trial. *Clinical Trials*, 9(2), 257-264.
- Buyse, M. et al. (2013). Statistical fraud detection in clinical trials. *Clinical Investigation*, 3(12), 1135-1145.
- Cai, C. et al. (2020). Application of machine learning methods to predict protocol deviations in clinical trials. *Therapeutic Innovation and Regulatory Science*, 54(6), 1446-1456.
- Dayan, I. et al. (2021). Federated learning for predicting clinical outcomes in patients with COVID-19. *Nature Medicine*, 27(10), 1735-1743.
- FDA (2013). Guidance for Industry: Oversight of Clinical Investigations -- A Risk-Based Approach to Monitoring. U.S. Department of Health and Human Services.
- FDA (2018). Framework for FDA's Real-World Evidence Program. U.S. Food and Drug Administration.
- Getz, K. A. (2017). Reflections on the evolution of clinical trial monitoring. *Applied Clinical Trials*, 26(3), 12-14.
- Getz, K. A. et al. (2020). Assessing the impact of risk-based monitoring implementation. *Therapeutic Innovation and Regulatory Science*, 54(2), 320-329.
- Giorgi, J. M. et al. (2021). End-to-end extraction of adverse drug events using deep learning. *Drug Safety*, 44(2), 235-245.
- Hurley, C. et al. (2016). Risk-based monitoring in clinical trials: a review of guidelines and practice. *Contemporary Clinical Trials Communications*, 4, 112-119.
- ICH (2016). Guideline for Good Clinical Practice E6(R2). International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use.
- Jiang, W. et al. (2019). Centralized statistical monitoring to detect data anomalies in clinical trials. *Statistics in Medicine*, 38(10), 1908-1922.
- Kreimeyer, K. et al. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *Journal of Biomedical Informatics*, 73, 14-29.
- Li, T. et al. (2020). Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2, 429-450.
- Lindblad, A. S. et al. (2014). Central site monitoring: results from a test of accuracy in identifying trials and sites failing Food and Drug Administration inspection. *Clinical Trials*, 11(2), 205-217.
- Pogue, J. M. et al. (2013). Central statistical monitoring: detecting fraud in clinical trials. *Clinical Trials*, 10(2), 225-235.
- Rieke, N. et al. (2020). The future of digital health with federated learning. *NPJ Digital Medicine*, 3(1), 119.
- Tantsyura, V. et al. (2015). Risk-based source data verification approaches: strengths and limitations. *Therapeutic Innovation and Regulatory Science*, 49(6), 838-849.

- Venet, D. et al. (2012). A statistical approach to central monitoring of data quality in clinical trials. *Clinical Trials*, 9(6), 705-713.
- Warnat-Herresthal, S. et al. (2021). Swarm learning for decentralized and confidential clinical machine learning. *Nature*, 594(7862), 265-270.
- Zink, R. C. and Wolfinger, R. D. (2019). Using statistical methods and machine learning to predict patients likely to experience protocol deviations. *Therapeutic Innovation and Regulatory Science*, 53(5), 600-607.

Declarations

Funding

This research was supported by the European Commission Horizon 2020 programme (grant agreement 101017884, AI4TRIALS) and the Spanish Ministry of Science and Innovation (grant PID2021-126734OB-I00). No funder had any role in study design, data analysis, interpretation, or the decision to submit for publication.

Conflict of Interest

The authors declare no competing interests. No author has any financial relationship with pharmaceutical companies, clinical research organisations, or monitoring technology vendors.

Data Availability Statement

The AI-CTMF codebase (Python), trained model weights, evaluation scripts, and synthetic benchmark datasets are available at ai-ctmf-clinical.org under the MIT licence. Patient-level clinical trial data cannot be shared due to privacy regulations and informed consent restrictions.

Ethical Approval

This study analysed de-identified, retrospective clinical trial data under institutional data use agreements. No direct patient contact or intervention occurred. The Western Europe Data Science University Ethics Committee confirmed that formal ethical approval was not required (waiver reference WE-2022-0134).

Appendix A

Model Specifications and CTMAS Calculation Methodology

This appendix provides complete specifications for each AI monitoring approach, including model architectures, training procedures, hyperparameter settings, feature engineering details, and the CTMAS scoring rubric used to evaluate comparative performance across the five approaches.

Part I -- Model Specifications

- 1a. A1: 127 deterministic rules + statistical outlier detection (3 SD threshold), batch processing 50,000 records/min
- 1b. A2: XGBoost (500 trees, max depth 8, learning rate 0.05) + LightGBM ensemble, 186 engineered features 30,000 predictions/min
- 2a. A3: PubMedBERT (12-layer transformer, 110M parameters), fine-tuned on 12,000 clinical narratives with MedDRA coding 2,000 narratives/hr
- 2b. A4: Gradient-boosted risk model (52 site-level features, monthly updates), Bayesian hyperparameter optimization 1,000 site-scores/min
- 3a. A5: FedAvg algorithm with differential privacy (epsilon=1.0, delta=1e-5), 10 communication rounds per epoch 4 hr per federation round
- 3b. All: Temporal split validation (months 1-18 train, 19-24 test), 5-fold CV for hyperparameters Strict temporal separation

Part II -- CTMAS Scoring Rubric

- 4a. Deviation Detection: AUROC for protocol deviation classification Against CRA-adjudicated ground truth labels
- 4b. AE Sensitivity: Recall for SAEs and SUSARs from clinical narratives MedDRA-coded reference standard
- 5a. Data Quality: Fraction of critical fields with precision > 0.80 Coverage + accuracy combined metric
- 5b. Dropout Prediction: Precision at top-20 per site per month 3-month forward prediction window
- 6a. Regulatory Compliance: Agreement with independent ICH E6(R2) QTL assessments Kappa coefficient against auditor judgements
- 6b. $CTMAS = 0.25 \times Deviation + 0.20 \times AE + 0.20 \times Quality + 0.15 \times Dropout + 0.20 \times Regulatory$