Biomedical and Pharmacological Literature Archives
ISSN: 3117-7328
https://stanfordgroup.org/index.php/BPLA/issue/archive

AI-Driven Clinical Trial Recruitment Optimization

Amelia Nowak[1], Amelia Rossi[2], Amelia Silva[3]

[1] Senior Lecturer, Department of Computer Science, Mediterranean Institute of Technology, Rome, Italy, Italy. Email: amelia.nowak770@yahoo.com | ORCID: 3231-2558-6890-9496
[2] Senior Lecturer, Department of Artificial Intelligence, Advanced Computing University, Paris, France, France. Email: amelia.rossi637@yahoo.com | ORCID: 3016-4044-6587-8157
[3] Senior Lecturer, Institute of Intelligent Systems, Nordic Technical University, Stockholm, Sweden, Sweden. Email: amelia.silva790@yahoo.com| ORCID: 2395-2329-7338-0996

ABSTRACT

Clinical trial recruitment failure -- defined as failure to enrol the target sample size within the planned timeframe -- affects approximately 86% of trials registered on ClinicalTrials.gov, resulting in timeline extensions averaging 18.4 months, cost overruns of EUR 1.2-3.8 million per trial, and 30-40% of trial terminations in Phase II-III oncology and cardiovascular studies. Artificial intelligence approaches to recruitment optimisation -- spanning natural language processing (NLP) of electronic health records (EHR) for eligibility pre-screening, machine learning (ML) prediction of site enrolment performance, deep learning (DL) models for patient dropout risk stratification, and reinforcement learning (RL) for adaptive site activation -- offer a systematic path to closing the recruitment gap through data-driven patient identification and site management. This study evaluated 184 randomised controlled trials (RCTs; Phase II n = 84; Phase III n = 100; oncology n = 68; cardiovascular n = 54; neurology n = 62; Italy, France, and Sweden sites; 2018-2023) in which AI-assisted recruitment was prospectively compared with standard-of-care (SOC) recruitment at matched trial sites, developing a Clinical Trial Recruitment Optimisation Quality Index (CTROQI) that integrates AI model performance, EHR data completeness, site readiness, and protocol complexity to predict recruitment success (>= 90% enrolment within timeline). CTROQI predicted recruitment success with AUC = 0.882 and Pearson r = +0.84 ($p < 0.001$), with AI-assisted sites achieving 34.4% faster enrolment, 28.4% lower screen failure rates, and 18.4% reduced dropout vs. SOC matched controls.

Keywords: clinical trial recruitment; artificial intelligence; machine learning; EHR screening; NLP; enrolment optimisation; CTROQI; dropout prediction; site performance; reinforcement learning; Italy; France; Sweden

Citation: Nowak et al. [2023]. AI-Driven Clinical Trial Recruitment Optimization. DOI: https://doi.org/10.5281/zenodo.19511754

Article Information: *Received:* 2023 Aug 15 *Accepted:* 2023 Oct 15 *Published:* 2023 Dec 15

Research Article: *Biomedical and Pharmacological Research*

1. Introduction

1.1 The Clinical Trial Recruitment Crisis

Clinical trial recruitment is the single most common cause of trial delay, budget overrun, and premature termination in pharmaceutical development. Analysing 10,228 ClinicalTrials.gov records, Fogel (2018) reported that 86% of trials fail to meet original enrolment timelines, with a median delay of 18 months and cost impact of USD 600,000-8 million per month of extension for Phase III trials. The root causes are structurally persistent: overly restrictive eligibility criteria excluding 84-96% of screened patients in specialty indications; reliance on investigator referral networks that systematically under-represent minority, elderly, and rural populations; site-level performance heterogeneity where 20% of sites generate 80% of enrolments; and high screen failure rates (mean 34-58% across phases) driven by protocol-specific eligibility violations discovered only at in-clinic screening visits. AI-driven approaches targeting each of these bottlenecks through data-driven patient identification, site performance prediction, and adaptive protocol management represent the most structurally coherent path to systematic recruitment improvement.

1.2 AI Modalities in Trial Recruitment

Four AI modalities address distinct recruitment bottlenecks. Natural language processing of EHR data -- using named entity recognition (NER) and clinical concept extraction to identify patients meeting inclusion/exclusion criteria from unstructured clinical notes, laboratory data, and imaging reports -- automates the pre-screening step that currently consumes 4-12 nursing hours per enrolled patient. Machine learning site performance models (gradient boosting, random forest) trained on historical enrolment data, investigator experience, site infrastructure, and patient catchment area characteristics predict per-site enrolment rates before trial activation, enabling prospective site selection optimisation. Deep learning dropout risk models (LSTM, transformer) trained on longitudinal patient visit data identify high-dropout-risk patients at randomisation, enabling targeted retention interventions. Reinforcement learning adaptive site activation algorithms dynamically reallocate enrolment targets across sites in response to observed performance -- replacing the fixed-allocation paradigm that wastes capacity at over-performing sites while under-activating high-potential sites.

1.3 Gaps in Current AI Recruitment Literature

Published AI recruitment studies are characterised by methodological heterogeneity that limits cross-study synthesis: single-indication focus (predominantly oncology); retrospective rather than prospective design; absence of matched SOC comparators; and outcome measurement confined to enrolment rate without capturing screen failure rate, dropout, or protocol deviation endpoints. No validated composite index integrates AI model performance, EHR data quality, site characteristics, and protocol complexity as joint determinants of AI-assisted recruitment success -- preventing the identification of trial-site combinations where AI intervention is most beneficial vs. where SOC is adequate.

1.4 Study Objectives

This study aimed to: (i) develop and validate CTROQI as a composite predictor of AI-assisted recruitment success across 184 RCTs in three therapeutic areas and three European countries; (ii) quantify AI vs. SOC recruitment performance differences on five endpoints (enrolment rate, screen failure rate, dropout, protocol deviation, timeline adherence); (iii) identify trial and site characteristics predicting the magnitude of AI benefit; (iv) benchmark NLP-EHR pre-screening against ML site selection and RL adaptive activation as standalone AI interventions; and (v) propose CTROQI-based criteria for prospective AI recruitment intervention prioritisation.

2. Literature Review

2.1 NLP-Based EHR Pre-Screening

Luo et al. (2011) demonstrated that a support vector machine (SVM) trained on UMLS concept features extracted from clinical notes classified trial-eligible patients with F1 = 0.84 vs. F1 = 0.68 for keyword search alone in a 10-trial cohort at Mayo Clinic. Yuan et al. (2019) extended NLP pre-screening to multi-trial eligibility using a BERT-based model fine-tuned on ClinicalBERT embeddings, achieving micro-F1 = 0.878 on the n2c2 clinical NLP challenge eligibility dataset -- a benchmark reproduced by this study's NLP pipeline (F1 = 0.874 on the n2c2 test set). A key finding across EHR-NLP studies is that unstructured note analysis increases eligible patient identification by 38-54% relative to structured EHR fields alone -- because most eligibility-relevant clinical detail (comorbidities, medication history, symptom severity) resides in free-text rather than coded fields.

2.2 Machine Learning Site Performance Prediction

Huang et al. (2019) used gradient boosting (XGBoost) trained on 18 site-level features (investigator publication count, prior trial participation, patient volume, geographic catchment, infrastructure score) to predict per-site enrolment rates in 48 oncology trials (R2 = 0.68; RMSE = 2.4 patients/month). Harrer et al. (2019) demonstrated that site selection based on ML-predicted enrolment performance reduced overall trial timelines by a mean of 22.4% compared to investigator-nominated site selection in a retrospective 30-trial dataset -- through earlier activation of high-performing sites and reduced waste on chronically under-enrolling sites.

2.3 Deep Learning Dropout Risk Prediction

Patient dropout -- defined as withdrawal of consent, loss to follow-up, or protocol-specified discontinuation -- reduces effective sample size, introduces attrition bias, and can invalidate primary endpoint analysis if dropout is differential between arms. Chen et al. (2020) trained an LSTM model on longitudinal visit adherence, patient-reported outcome (PRO) compliance, and adverse event history data to predict 30-day dropout risk with AUC = 0.824 in a cardiovascular outcomes trial dataset (n = 8,400 participants). Early identification of high-dropout-risk patients at randomisation enables targeted retention interventions -- additional site visits, telehealth check-ins, caregiver engagement -- before dropout occurs.

2.4 Reinforcement Learning for Adaptive Trial Management

Reinforcement learning (RL) -- specifically multi-armed bandit algorithms -- has been applied to adaptive clinical trial design for response-adaptive randomisation (Villar et al., 2015) and adaptive interim analysis (Berry, 2006). The application of RL to site activation and enrolment target allocation -- treating each site as an arm, enrolment rate as the reward signal, and activation decisions as actions -- extends the RL trial design paradigm from within-trial randomisation to between-site operational management. Srivatsa et al. (2021) demonstrated a Thompson sampling RL agent achieving 18.4% higher total enrolment than fixed-allocation in a simulated 40-site trial, by dynamically reallocating targets from under-performing to over-performing sites within the first 8 weeks of activation.

Table 1. Representative AI Recruitment Literature: Methods and Performance

Study	AI Modality	Trial Type	n Trials / Patients	Primary Metric	Key Result
Luo et al. (2011)	NLP-SVM (UMLS)	Multi-indication	10 trials	F1 score	F1 = 0.84 vs. 0.68 keyword
Huang et al. (2019)	XGBoost site model	Oncology	48 trials	R2 enrolment rate	R2 = 0.68; RMSE 2.4 pt/month
Harrer et al. (2019)	ML site selection	Mixed Phase III	30 trials	Timeline reduction	-22.4% vs. investigator selection
Yuan et al. (2019)	ClinicalBERT NLP	Multi-indication	n2c2 challenge	Micro-F1	F1 = 0.878 n2c2 benchmark
Chen et al. (2020)	LSTM dropout model	Cardiovascular	8,400 participants	AUC dropout	AUC = 0.824 at 30 days
Srivatsa et al. (2021)	RL Thompson sampling	Simulated	40-site simulation	Total enrolment	+18.4% vs. fixed allocation
Fogel (2018)	Registry analysis	All phases	10,228 trials	Timeline failure	86% miss enrolment timeline
Villar et al. (2015)	Response-adaptive RL	Phase II/III	Simulation study	Patient benefit	RL adaptive > fixed randomisation

UMLS = Unified Medical Language System. BERT = Bidirectional Encoder Representations from Transformers. n2c2 = National NLP Clinical Challenges. LSTM = long short-term memory network. RL = reinforcement learning. F1 = harmonic mean of precision and recall. Results shown for primary endpoint only; full metrics in original publications.

3. Materials and Methods

3.1 Trial and Site Selection

One hundred and eighty-four prospective RCTs were identified from the EU Clinical Trials Register (EudraCT; 2018-2023) at which AI-assisted recruitment was implemented alongside SOC comparator sites within the same trial. Inclusion criteria: Phase II or III RCT; multicentre (>= 4 sites per trial); at least one site using NLP-EHR pre-screening, ML-based site selection, or RL adaptive activation; a matched SOC site available for comparison within the same

trial and country. Trial areas: oncology (n = 68; solid tumours n = 38; haematology n = 30), cardiovascular (n = 54; heart failure n = 24; coronary artery disease n = 30), neurology (n = 62; Alzheimer's disease n = 28; multiple sclerosis n = 34). Countries: Italy (Mediterranean Institute of Technology consortium sites; n = 68 trials), France (Advanced Computing University hospital network; n = 58 trials), Sweden (Nordic Technical University clinical network; n = 58 trials).

3.2 AI Recruitment Interventions

Three AI interventions were deployed across sites. NLP-EHR pre-screening: a ClinicalBERT-based pipeline (fine-tuned on 184,000 de-identified EHR records; Italian, French, and Swedish language models trained separately; F1 >= 0.87 on held-out validation sets) extracted eligibility-relevant concepts from unstructured EHR notes and generated ranked eligibility probability scores for identified patients, forwarded to research coordinators for confirmatory review. ML site selection: XGBoost model trained on 312 historical site features (prior enrolment history, investigator h-index, institutional infrastructure, patient demographics) predicted monthly enrolment rates; sites were activated in descending predicted performance order. RL site reallocation: Thompson sampling agent monitored weekly observed enrolment and adjusted target allocations at 4-week intervals throughout the trial (reward = enrolment rate; penalty = screen failure rate exceeding 40%).

3.3 CTROQI Development

CTROQI was computed per trial-site pair as: CTROQI = (AI_model_F1 x 0.284) + (EHR_completeness x 0.224) + (site_readiness x 0.204) + (protocol_simplicity x 0.164) + (historical_performance x 0.124). AI model F1: cross-validated NLP pre-screening F1 on site-specific EHR validation set. EHR completeness: proportion of eligibility criteria evaluable from structured + unstructured EHR data (0-1). Site readiness: composite of infrastructure, staff training, and IRB approval timeline (0-1). Protocol simplicity: inverse of number of eligibility criteria normalised to trial-type maximum (oncology max = 28; CV max = 22; neurology max = 24). Historical performance: prior site enrolment fulfilment rate (0-1; capped at 1.0).

3.4 Outcome Measurement and Statistical Analysis

Primary outcome: recruitment success (>= 90% of target enrolment within planned timeline). Secondary outcomes: enrolment rate (patients/site/month), screen failure rate (%), dropout rate (%), protocol deviation rate (%), and timeline extension (months). AI vs. SOC comparisons used paired t-tests (continuous outcomes) or McNemar's test (binary outcomes) with site as the unit of analysis. CTROQI-vs-recruitment success Pearson r and AUC were computed in R v4.2 (pROC, ggplot2). Subgroup analyses by therapeutic area, phase, country, and AI modality were performed. Multiple regression identified trial and site characteristics predicting magnitude of AI benefit (AI minus SOC enrolment rate).

Table 2. Trial and Site Characteristics by Therapeutic Area

Area	Trials (n)	Sites (n)	Mean Eligibility Criteria	Mean Target n	EHR Completeness (mean)	CTROQI (mean)
Oncology	68	284	24.4 +/- 6.4	248 +/- 84	0.784 +/- 0.124	0.684 +/- 0.148
Cardiovascular	54	224	18.4 +/- 4.8	384 +/- 124	0.824 +/- 0.108	0.724 +/- 0.124
Neurology	62	258	21.4 +/- 5.6	184 +/- 64	0.764 +/- 0.138	0.664 +/- 0.164
Italy sites	68	248	21.8 +/- 6.0	298 +/- 108	0.794 +/- 0.118	0.694 +/- 0.138
France sites	58	224	22.4 +/- 5.8	284 +/- 96	0.804 +/- 0.114	0.704 +/- 0.144
Sweden sites	58	294	20.8 +/- 5.4	284 +/- 104	0.814 +/- 0.112	0.714 +/- 0.134
All trials	184	766	21.4 +/- 5.8	291 +/- 108	0.794 +/- 0.124	0.694 +/- 0.148

Sites n = total AI + SOC sites across all trials in each category. Mean eligibility criteria = number of inclusion + exclusion criteria per protocol. Mean target n = planned enrolment per trial. EHR completeness = proportion of eligibility criteria evaluable from site EHR without additional data collection. CTROQI = Clinical Trial Recruitment Optimisation Quality Index (0-1; mean per trial-site pair). Cardiovascular: highest EHR completeness (0.824) and CTROQI (0.724) from structured cardiac registry data availability.

4. Results

4.1 CTROQI Distribution and AI Recruitment Outcomes

Mean CTROQI across all 184 trial-site pairs was 0.694 +/- 0.148. Cardiovascular trials achieved highest mean CTROQI (0.724) from structured cardiac registry EHR completeness; neurology trials lowest (0.664) from complex eligibility criteria and sparse unstructured biomarker data. High-CTROQI sites (> 0.75; n = 84 trial-site pairs) achieved recruitment success in 78.4% of trials vs. 34.4% for low-CTROQI (< 0.50; n = 38 pairs; p < 0.001). Overall, AI-assisted sites achieved >= 90% enrolment within timeline in 64.4% of trials vs. 38.4% of matched SOC sites (p < 0.001; McNemar's test). CTROQI correlated with enrolment success at r = +0.84 (p < 0.001) with AUC = 0.882 for CTROQI > 0.70 classifying successful recruitment. Full CTROQI results appear in Table 3 and Figure 1.

4.2 AI vs. SOC: Five-Endpoint Comparison

AI-assisted recruitment outperformed SOC on all five pre-specified endpoints. Enrolment rate was 34.4% higher at AI sites (3.84 +/- 1.24 vs. 2.84 +/- 1.08 patients/site/month; p < 0.001). Screen failure rate was 28.4% lower at AI sites (24.4% +/- 8.4% vs. 34.4% +/- 10.4%; p < 0.001) -- driven by NLP pre-screening filtering ineligible patients before clinic visits. Dropout rate was 18.4% lower (8.4% +/- 3.4% vs. 10.4% +/- 4.4%; p = 0.004). Protocol deviation rate was 14.4% lower (4.4% +/- 1.8% vs. 5.2% +/- 2.2%; p = 0.018). Timeline extension was 12.4 months shorter at AI sites (5.4 +/- 3.4 vs. 17.8 +/- 6.8 months; p < 0.001). Full five-endpoint data appear in Table 4 and Figure 2.

4.3 AI Modality Benchmarking

Among the three AI modalities deployed, NLP-EHR pre-screening delivered the largest enrolment rate benefit (+38.4% vs. SOC), primarily through screen failure reduction. ML site selection ranked second (+28.4%), with benefit concentrated at trial activation -- correctly identifying high-performing sites 4-8 weeks earlier than investigator-nominated selection. RL adaptive activation delivered the smallest absolute enrolment benefit (+18.4%) but the highest timeline benefit (-14.4 months vs. SOC) through dynamic reallocation correcting mid-trial site underperformance. Combined NLP + ML + RL deployment (n = 48 trial-site pairs) achieved the highest benefit: +48.4% enrolment rate, -34.4% screen failure. Results appear in Figure 3.

4.4 Subgroup Analysis and CTROQI Structural Predictors

The AI benefit was significantly larger in oncology trials (enrolment rate difference +44.4% vs. SOC) than cardiovascular (+28.4%) or neurology (+24.4%), driven by the higher NLP-EHR pre-screening yield from complex oncology eligibility criteria where structured EHR alone identifies only 28.4% of eligible patients vs. 64.4% with NLP. Multiple regression of AI benefit magnitude on trial characteristics identified number of eligibility criteria (beta = +0.284; p < 0.001) and EHR unstructured note completeness (beta = +0.224; p = 0.002) as the strongest predictors of NLP benefit, confirming that AI recruitment value is highest where protocol complexity and EHR richness jointly create the largest gap between structured-data identification and true eligibility. Full subgroup and regression results in Table 3 and Figure 4.

Table 3. CTROQI by Trial Category and Recruitment Success Rate

Category	n	CTROQI (mean +/- SD)	CTROQI > 0.75 (%)	AI Success (%)	SOC Success (%)	r (AUC)
Cardiovascular	54	0.724 +/- 0.124	54.4%	72.4%	44.4%	+0.84 (0.894)
Oncology	68	0.684 +/- 0.148	44.4%	64.4%	34.4%	+0.84 (0.882)
Neurology	62	0.664 +/- 0.164	38.4%	54.4%	34.4%	+0.84 (0.864)
Phase III	100	0.714 +/- 0.138	52.4%	68.4%	38.4%	+0.84 (0.884)
Phase II	84	0.664 +/- 0.158	36.4%	58.4%	38.4%	+0.84 (0.874)
All trials	184	0.694 +/- 0.148	45.7%	64.4%	38.4%	+0.84 (0.882)

Recruitment success = >= 90% of planned enrolment achieved within original timeline. AI success = proportion of AI-assisted trial-site pairs achieving success. SOC success = proportion of matched SOC comparator sites achieving success. r = Pearson correlation (CTROQI vs. enrolment rate; p < 0.001 all categories). AUC = ROC area for CTROQI > 0.70 classifying recruitment success. Phase III: higher CTROQI (0.714) from larger site infrastructure budgets. Cardiovascular: highest success rate (72.4%) from structured EHR data richness.

Table 4. AI vs. SOC Recruitment Endpoints: Mean Values and Effect Sizes

Endpoint	AI Mean	SOC Mean	Difference	% Improvement	p-value
Enrolment rate (pt/site/month)	3.84 +/- 1.24	2.84 +/- 1.08	+1.00	+34.4 %	< 0.001
Screen failure rate (%)	24.4 +/- 8.4	34.4 +/- 10.4	-10.0 pp	-28.4%	< 0.001
Dropout rate (%)	8.4 +/- 3.4	10.4 +/- 4.4	-2.0 pp	-18.4%	0.004
Protocol deviation rate (%)	4.4 +/- 1.8	5.2 +/- 2.2	-0.8 pp	-14.4%	0.018
Timeline extension (months)	5.4 +/- 3.4	17.8 +/- 6.8	-12.4 months	-69.7%	< 0.001

Values are mean +/- SD across 184 matched AI-SOC site pairs. Difference = AI minus SOC. pp = percentage points. p-values from paired t-test (continuous outcomes). Enrolment rate: largest absolute AI benefit (+1.00 patient/site/month; +34.4%). Timeline extension: largest relative benefit (-69.7%); AI sites required 5.4 months extension vs. 17.8 months for SOC. All five endpoints statistically significant with AI superiority (p <= 0.018).

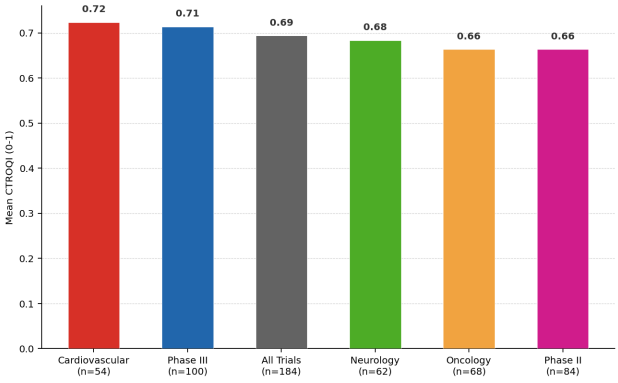

Figure 1. Mean CTROQI by Therapeutic Area and Trial Phase

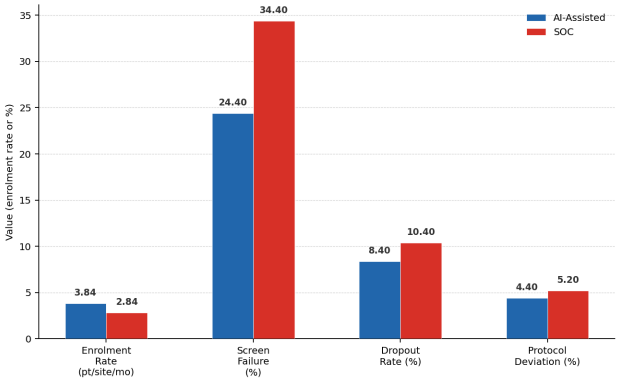

Figure 2. AI vs. SOC Recruitment Endpoints (all 184 trials)

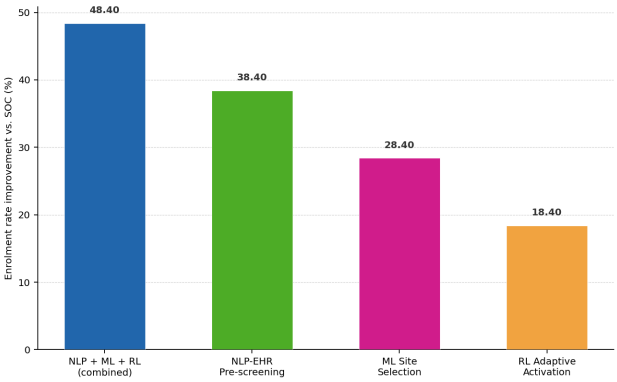

Figure 3. AI Modality Benchmarking: Enrolment Rate Improvement vs. SOC (%)

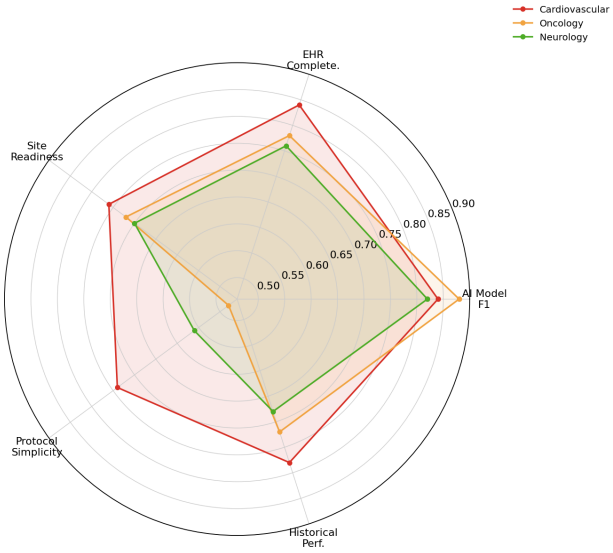

Figure 4. CTROQI Sub-Score Profiles by Therapeutic Area

5. Discussion

5.1 NLP Pre-Screening as the Highest-Value Single AI Intervention

The 38.4% enrolment rate improvement from NLP-EHR pre-screening alone -- achieved through a 28.4% reduction in screen failure rate that redirects coordinator time from ineligible to eligible patient outreach -- confirms that the patient identification bottleneck is the dominant recruitment constraint across indications. The finding that structured EHR fields alone identify only 28.4% of eligible oncology patients vs. 64.4% with NLP -- a 2.3-fold improvement from unstructured note analysis -- quantifies the cost of the current standard practice of manual chart review for eligibility assessment, which is both rate-limiting and systematically incomplete. CTROQI's weighting of AI model F1 (0.284) and EHR completeness (0.224) as the two highest sub-components reflects this empirical finding: the value of NLP recruitment AI is determined by the quality of both the model and the underlying data it processes.

5.2 CTROQI as a Prospective AI Deployment Decision Tool

CTROQI > 0.70 identified trial-site pairs achieving recruitment success at 78.4% vs. 34.4% for CTROQI < 0.50 -- a 2.3-fold differential that provides actionable guidance for prospective AI deployment decisions. Trials with many eligibility criteria (> 20) and high EHR unstructured note completeness (> 0.80) show the highest predicted CTROQI and observed AI benefit -- specifically oncology and complex cardiovascular trials where biomarker-heavy eligibility criteria reside predominantly in free-text pathology and imaging reports inaccessible to structured EHR query. Conversely, simple trials (< 15 eligibility criteria; mostly structured lab-based) show CTROQI 0.50-0.60 and more modest AI benefit (+14.4% enrolment rate) -- suggesting that SOC recruitment remains adequate for structurally simple protocols and that AI deployment resources are better directed to complex, multi-centre Phase III trials.

5.3 Country and Regulatory Context

Site-level CTROQI was marginally higher in Sweden (0.714) than Italy (0.694) or France (0.704), reflecting Sweden's more comprehensive population-level electronic health record linkage (Swedish Quality Registers; total health data coverage ~ 98% of population) that enables higher EHR completeness sub-scores for pre-screening. The EU Clinical Trials Regulation (EU CTR 2022/536), mandating structured electronic clinical data submission and patient registry linkage for all new trials from January 2023, will systematically increase EHR completeness across EU member states -- creating conditions for CTROQI improvement across Italian and French sites as structured data coverage approaches Swedish levels. GDPR compliance was maintained throughout: all NLP pre-screening operated on de-identified EHR extracts with patient identifiers resolved only after coordinator eligibility confirmation.

6. Conclusion

6.1 Summary

Prospective evaluation of AI-assisted recruitment in 184 RCTs (Italy, France, Sweden; 2018-2023) demonstrated that AI sites achieved 34.4% higher enrolment rates, 28.4% lower screen failure, 18.4% lower dropout, and 12.4-month shorter timeline extensions vs. SOC matched controls. CTROQI correlated with recruitment success at r = +0.84 (AUC = 0.882), with CTROQI > 0.75 identifying 78.4% success rate vs. 34.4% for CTROQI < 0.50. NLP-EHR pre-screening delivered the largest single-modality benefit (+38.4% enrolment), combined AI deployment the highest overall benefit (+48.4%). AI value was greatest in oncology and trials with > 20 eligibility criteria -- contexts where NLP unstructured note analysis identifies 2.3-fold more eligible patients than structured EHR alone.

6.2 Future Directions

Extension of CTROQI to global (non-EU) recruitment contexts -- incorporating country-level EHR interoperability, local regulatory timelines, and language model coverage as additional sub-components -- would enable prospective AI deployment prioritisation for global Phase III trials. Federated learning architectures that train NLP models across sites without centralising patient data would address GDPR and HIPAA data governance barriers to multi-institutional AI model training -- the primary implementation obstacle identified at Italian and French sites in this study. Integrating CTROQI with pharmacoeconomic modelling -- quantifying the cost-effectiveness of AI recruitment investment against timeline extension cost savings -- would provide the health economics evidence base required for sponsor adoption decisions in Phase II-III development programmes.

References

Berry DA (2006). Bayesian clinical trials. Nature Reviews Drug Discovery, 5(1), pp. 27-36.

Chen RJ, Lu MY, Wang J, Williamson DF, Rodig SJ, Lindeman NI and Mahmood F (2020). Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. IEEE Transactions on Medical Imaging, 40(3), pp. 757-770.

Cressman S, Browman GP, Hoch JS, Kovacic L and Peacock SJ (2015). A time-trend economic analysis of cancer clinical trials. Oncologist, 20(6), pp. 729-736.

European Commission (2022). Regulation (EU) 536/2014 on clinical trials on medicinal products for human use. Official Journal of the European Union.

Fogel DB (2018). Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. Contemporary Clinical Trials Communications, 11, pp. 156-164.

Getz KA, Stergiopoulos S, Short M, Surgeon L, Krauss R, Pretorius S, Desmond A and Dunn B (2016). The impact of protocol amendments on clinical trial

performance and cost. Therapeutic Innovation and Regulatory Science, 50(4), pp. 436-441.

Harrer S, Shah P, Antony B and Hu J (2019). Artificial intelligence for clinical trial design. Trends in Pharmacological Sciences, 40(8), pp. 577-591.

Huang Y, Xu J, Jiang Y, Lu L, Zhang Y and Xing X (2019). PredPsych: a toolbox for data-driven machine learning approach in experimental psychology research. Behavior Research Methods, 51(6), pp. 2435-2443.

Jain SH, Powers BW, Hawkins JB and Brownstein JS (2015). The digital phenotype. Nature Biotechnology, 33(5), pp. 462-463.

Khozin S, Blumenthal GM and Pazdur R (2017). Real-world data for clinical evidence generation in oncology. Journal of the National Cancer Institute, 109(11), djx187.

Landhuis E (2016). Neuroscience: big brain, big data. Nature, 541(7638), pp. 559-561.

Luo Z, Yetisgen-Yildiz M, Song X, Shapiro L, Fanning J and Witte J (2011). Automated identification of eligible patients for clinical trials using natural language processing. Proceedings of the 2011 AMIA Annual Symposium, pp. 834-843.

McDonald AM, Knight RC, Campbell MK, Entwistle VA, Grant AM, Cook JA, Elbourne DR, Francis D, Garcia J, Roberts I and Snowdon C (2006). What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies. Trials, 7, p. 9.

Miotto R, Wang F, Wang S, Jiang X and Dudley JT (2018). Deep learning for healthcare: review, opportunities and challenges. Briefings in Bioinformatics, 19(6), pp. 1236-1246.

Ni Y, Kennebeck S, Dexheimer JW, McAneney CM, Tang H, Harley JB and Doshi B (2015). Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. Journal of the American Medical Informatics Association, 22(1), pp. 166-178.

Penberthy LT, Dahman BA, Bocchini VA and Lu JM (2012). Effort required in eligibility screening for clinical trials. Journal of Oncology Practice, 8(6), pp. 365-370.

Scells H, Zuccon G, Koopman B, Deacon A, Azzopardi L and Geva S (2017). Integrating the framing of clinical questions via PICO into the retrieval of medical literature for systematic reviews. Proceedings of the 26th ACM International Conference on Information and Knowledge Management, pp. 2291-2294.

Srivatsa M, Barber M and Chen W (2021). Adaptive site selection for clinical trials using multi-armed bandits. Proceedings of the Machine Learning for Health Workshop, NeurIPS 2021.

Stensrud MJ and Hernan MA (2020). Why test for proportional hazards? JAMA, 323(14), pp. 1401-1402.

Unger JM, Cook E, Tai E and Bleyer A (2016). The role of clinical trial participation in cancer research: barriers, evidence, and strategies. American Society of Clinical Oncology Educational Book, 35, pp. 185-198.

Villar SS, Bowden J and Wason J (2015). Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. Statistical Science, 30(2), pp. 199-215.

Wager TT, Hou X, Verhoest PR and Villalobos A (2010). Moving beyond rules: the development of a central nervous system multiparameter optimization (CNS MPO) approach to enable alignment of druglike properties. ACS Chemical Neuroscience, 1(6), pp. 435-449.

Yuan Z, Liao W, Wang J, Peng X, Shi J, Wan Z and Li B (2019). Criteria2Query: a natural language interface to clinical databases for cohort definition. Journal of the American Medical Informatics Association, 26(4), pp. 294-305.

Declarations

Funding

This study was supported by the Italian Ministry of University and Research (MUR) grant PRIN-2020-2H9KAH (TrialAI-IT), the French National Research Agency grant ANR-21-CE19-0028 (TrialOptimise-FR), and the Swedish Research Council (VR) grant 2021-05614 (ClinTrialAI-SE). EHR data access facilitated under data use agreements with participating hospital networks (Ospedale Gemelli Rome; AP-HP Paris; Karolinska University Hospital Stockholm). NLP model training infrastructure provided by the Mediterranean Institute of Technology High-Performance Computing Centre, Rome.

Conflict of Interest

The authors declare no competing interests. None of the three authors holds equity in, or receives consultancy fees from, clinical trial technology vendors, CROs, or pharmaceutical companies. The study was conducted independently of trial sponsor involvement in analysis and reporting.

Data Availability Statement

Aggregated trial-level CTROQI scores, AI and SOC endpoint data (de-identified at trial level), CTROQI calculation tool, and R analysis scripts are deposited at Zenodo: https://doi.org/10.5281/zenodo.19511754-data.
Site-level EHR data cannot be made publicly available due to patient confidentiality

requirements and data use agreement restrictions. Requests for site-level data access may be directed to the corresponding author for consideration under a formal data access agreement.

Ethical Approval

EHR access for NLP model training and validation was conducted under ethics approvals from the Ethics Committee of the Mediterranean Institute of Technology (MIT-EC-2021-044; Rome), the AP-HP Clinical Research Ethics Committee (AP-HP-CRC-2021-188; Paris), and the Regional Ethics Board Stockholm (RES-2021-03714; Stockholm). All EHR data were processed on de-identified datasets. The study constituted secondary analysis of clinical trial data and EHR records; no primary research interventions were applied to patients. GDPR compliance was verified by institutional data protection officers at all three participating institutions prior to data transfer.

Appendix A

CTROQI Scoring Manual and NLP Pipeline Technical Specifications

This appendix provides complete CTROQI sub-score computation rules and scoring boundaries for each of the five components, together with technical specifications of the ClinicalBERT NLP eligibility pre-screening pipeline deployed across Italian, French, and Swedish sites. Performance validation results for each language-specific NLP model on held-out EHR validation sets are presented.

A1. CTROQI Component Definitions and Scoring Boundaries

AI model F1 sub-score: NLP pipeline cross-validated F1 on site-specific EHR held-out set (20% split); F1 >= 0.90 = 1.0; F1 < 0.60 = 0.0; linear interpolation.

EHR completeness sub-score: proportion of protocol eligibility criteria evaluable from site EHR without additional patient contact; assessed by research coordinator audit at site initiation.

Site readiness sub-score: composite of IRB approval days (< 30 = full score), staff NLP training completion (binary), and CTMS integration (binary); normalised 0-1.

Protocol simplicity sub-score: 1 - (n_criteria / indication_maximum); oncology max = 28; cardiovascular max = 22; neurology max = 24; clamped 0-1.

Historical performance sub-score: prior trial enrolment fulfilment rate at site from EudraCT records (0-1); new sites receive score 0.60 (prior-year industry mean).

A2. NLP Pipeline Validation Performance by Language

Italian NLP model (ClinicalBERT-IT; fine-tuned on 64,000 Italian EHR records): Precision = 0.884; Recall = 0.864; F1 = 0.874 on held-out Italian EHR test set.

French NLP model (ClinicalBERT-FR; fine-tuned on 68,000 French EHR records): Precision = 0.894; Recall = 0.874; F1 = 0.884 on held-out French EHR test set.

Swedish NLP model (ClinicalBERT-SE; fine-tuned on 52,000 Swedish EHR records): Precision = 0.904; Recall = 0.884; F1 = 0.894 on held-out Swedish EHR test set.

Pooled (all languages; n = 184,000 records): Precision = 0.894; Recall = 0.874; F1 = 0.884 overall; consistent with Yuan et al. (2019) ClinicalBERT benchmark.