Magna Scientia Advanced Research and Reviews
eISSN: 2582-9394
Cross Ref DOI: 10.30574/msarr
Journal homepage: https://msarr.magnascientiapub.com/

(RESEARCH ARTICLE)

Check for updates

DATA-DRIVEN APPROACHES TO IMPROVING PATIENT RECRUITMENT AND RETENTION IN U.S. CLINICAL TRIALS: CURRENT STRATEGIES AND FUTURE DIRECTIONS

Elizabeth Ngenia Kamau *

St. Cloud State University Plymouth, MN.
* Corresponding Author

Magna Scientia Advanced Research and Review, 2026, 18(01), 025–041

Publication history: Received on 25 July 2026; revised on 02 September 2026; accepted on 05 September 2026

Article DOI: https://doi.org/10.30574/msarr.2026.18.1.0177

ABSTRACT

Recruiting and retaining patient's remains one of the most important issues in clinical trials, where about 80% of the trials fail to achieve their enrolment schedule, and 30 percent of the recruited subjects do not see through the end of the trial period. The impact of these challenges is a delay in the development of drugs, a rise in the cost, and scientific invalidity. The appearance of data-centered strategies, such as artificial intelligence (AI), machine learning (ML), big data analytics, and integration of electronic health record (EHR), has given hope towards finding solutions to these time-old hindrances. The paper is a critical review of existing evidence-based practices for enhancing patient recruitment and retention in U.S. clinical trials and analyzes their successes and future research and practice directions. Some of the methodologies that we analyzed are natural language processing, predictive analytics, goal programming, deep learning models, and HER-based surveillance systems.

The current approaches indicate that their efficiency is surprisingly high in recruitment, and AI-driven systems yielded results of 40-70% faster screening time, 23-fold more valid patients, and varied enrollment rates up to 15-45%. The main technologies are NLP-based eligibility screening, enrollment success predictive modeling, automated systems of patient-trial matches, and Real-Time EHR. Nevertheless, equity and diversity challenges, interoperability of data, regulations, and practical use continue to be experienced. Data-driven approaches are a paradigm shift in the domain of clinical trial recruitment and retention, and they provide tangible advances in the areas of efficiency, cost-effectiveness, and identification of patients. In line with future developments, it is possible to estimate more profound integration of large language models, privacy-preserving analytics by using federated learning, adaptive recruitment policies in real-time, and frameworks to assure equitable representations. It needs interdisciplinary cooperation, effective regulation mechanisms, and further advancement of AI/ML approaches to achieve successful implementation.

Keywords: Clinical trials, Artificial intelligence, Machine learning, Electronic health records, Predictive modeling.

1 INTRODUCTION

Evidence-based medicine relies on clinical trials, which are the fundamental aspect of scientific research that allows assessing the safety and effectiveness of new therapeutic methods. Nevertheless, the clinical trial business suffers a chronic and long-studied crisis, the lack of opportunities to find and retain adequate amounts of participants within

budgets and timelines. About 80% of clinical trials are unable to achieve their enrollment targets on time, and almost 30% of participants enrolled in trials do not continue until the end (Wilkins et al., 2021). The indirect effects of these recruitment and retention failures include delayed development of therapies, inflated research expenses ranging from $600,000 to $8 million per day of delay, reduced statistical power, and, eventually, reduced access to potentially life-saving therapies (Huml et al., 2020).

A specific problem with patient recruitment and retention is exceptionally sharp in the United States, as the clinical trial landscape is becoming increasingly complex in terms of protocols, eligibility, geographic distribution of the trial site, and the underrepresentation of racial and ethnic minorities (Ahluwalia et al., 2022; Brooks et al., 2015). The conventional forms of recruitment, such as the physician referral system, community outreach, and mass media advertisement, have not been adequate to meet these complex challenges. In addition, the COVID-19 pandemic has demonstrated the vulnerability of traditional recruitment methods and increased the pace at which new, technology-enabling solutions are required (Bikou et al., 2023). Over the past few years, a combination of various technological advances has provided unprecedented opportunities to change the manner of clinical trial recruitment and retention. The introduction of electronic health records (EHRs) in large numbers led to the creation of enormous repositories of organized and unorganized patient data. With recent developments in artificial intelligence (AI) and machine learning (ML), specifically in natural language processing (NLP) and deep learning, complicated clinical data can be analyzed. Big data analytics solutions have given the tabulating power needed to compute and evaluate millions of patient records within seconds. A combination of these data-driven strategies provides the opportunity to transform the process of patient identification, screening, enrolling, and retaining into clinical trials (Murphy et al., 2017; Gong et al., 2026; Samuel et al., 2024).

This manuscript aims to cover a comprehensive, critical review of existing data-driven strategies to enhance patient recruitment and retention in U.S. clinical trials. The evidence literature will be synthesized, taking into consideration current published results focusing on the methods used, effectiveness, and implementation issues, and perspectives of AI, ML, big data analytics, and HER-based strategies. In the analysis, we used more than 90 peer-reviewed articles, such as empirical research, methodological, and systematic reviews. We put emphasis on the U.S. case, as the regulatory landscape, healthcare infrastructure, and diversity issues, which define American clinical research, are specific. Through the amalgamation of existing information and mapping the way ahead, this manuscript will inform and speed up the implementation of information-driven solutions that will eventually enhance the efficiency, equity, and scientificity of

2 THEORETICAL FOUNDATIONS AND BACKGROUND

2.1 Clinical Trial Recruitment Crisis

Patient recruitment and retention in clinical trials have been known to be a problem for decades, yet they remain one of the most important obstacles in the process of successful drug development and medical research. As it is always empirical evidence, most clinical trials do not fulfill their enrollment requirements within the initially planned period. The research shows that about 80% of trials are delayed because of recruitment issues, and most trials are extended by 6 months or years following their original schedule (Wilkins et al., 2021; Evans et al., 2021). The economic consequences of these delays are deep, and the industry estimates that every day of delaying the drug to the market costs between 600,000 and 8 million dollars in lost revenue and additional costs incurred in the development process (Huml et al., 2020). In addition to an economic factor, recruitment failures undermine the generalizability and scientific validity of trial findings. Weak studies due to lack of enrolment cannot identify clinically significant treatment effects; they draw false negative conclusions, discarding potentially excellent therapies. In contrast, those trials that ultimately achieve the target of enrolment following lengthy recruitment durations might be contaminated by time effects with the alteration in the standard of care, ongoing trials, or even alterations in patient populations that may impact the similarity of the initial and final enrollees (Anisimov, 2023). The level of retention is also a major problem. The average rate of participant dropout of approximately 30% by therapeutic areas not only decreases the statistical power, but it also may create a possible selection bias when being nonrandom and depending on the treatment effect or patient characteristics (Nejat et al., 2024). The high dropout rates require over-enrolment to assure sufficient sample sizes, and this again elevates trial costs and the recruitment dilemma. Retention may be even worse in certain fields of therapy, especially where trials are conducted over a long-term period, and the dropout rates may be more than 4050% (Milliken et al., 2023).

2.2 Traditional Recruitment problems

The conventional approaches to clinical trial recruitment have various constraints related to each other that restrict their performance. To begin with, the screening of potentially eligible patients is very ineffective. Doctors normally use memory and random opportunities to screen trial-eligible patients, leading to missed opportunities and bias in selection due to patients having more frequent clinical interactions (Meystre et al., 2023). Despite the use of systematic screening methods, manual review of the chart is time-consuming, labor-intensive, and subject to human error, especially in cases

where the eligibility criteria are complicated with multiple inclusion and exclusion criteria across diverse domains of data (Idnay et al., 2021). Second, patient access to clinical trials is severely restricted to patients due to geographic and logistical constraints. In the US, clinical trial locations are overrepresented in urban academic healthcare institutions, which means that rural and underserved communities face significant barriers to access (Milliken et al., 2023). Patients might be required to travel hundreds of miles to study locations, and there are costs involved in transportation, accommodation, and time cost (von Guthrie 1104). Such barriers in logistics have a disproportionate impact on patients with low incomes and lead to the continued underrepresentation of diverse groups of patients in clinical research (Ahluwalia et al., 2022).

Third, lack of specificity regarding recruitment due to a lack of awareness and information. Even when patients know about any clinical trials, many of them could still be oblivious of what they could be missing and, in some instances, when they know, they do not have enough information on what to do. They might be unaware of trial opportunities or simply do not have the resources and time to advise patients on trial opportunities, especially in a community practicing setting (Brathwaite et al., 2023). The identification and recruitment process is further complicated by the fact that the U.S. healthcare system is fragmented, and patients get care provided by various and unconnected services and health systems (Umoren, Korang, et al., 2025). Fourth, there are challenges of trust and cultural barriers especially among racial and ethnic minority groups. The history of abuse in medical research; the Tuskegee syphilis study, has caused abiding suspicion of clinical research in African American society. These populations may also be hindered in recruitment due to language barriers, health belief and decision-making differences based on diversity but research have also shown lack of diversity in research staff (Brooks et al., 2015; Ahluwalia et al., 2022). These obstacles add to the continued absence of diversity in clinical trials, which restricts the generalized application of results and could fuel the health difference.

2.3 The Promise of Data-driven Approaches

The information-based solutions that are more focused on the use of AI, ML, big data analytics, and EHR integration have the potential to overpower the numerous shortcomings of the traditional recruitment processes. On a more basic level, these methods allow large groups of patients to be screened in a systematic, automated manner so as to identify potentially eligible candidates with hitherto unparalleled speed and accurateness (Gong et al., 2026). Instead of creating a system that physicians have to memorize, or reviewing manual charts to sort out exceptions, AI-based systems can actively analyze EHR data, automatically executing complicated eligibility criteria and creating real-time alerts whenever some potentially eligible patients have been found (Meystre et al., 2023). The computing of vital clinical data in unstructured text in the sources of clinical notes, radiology report, and pathology report data is normally not available using the traditional structured query methods with the help of natural language processing technologies (Idnay et al., 2021; Fang et al., 2022). This feature comes in especially since it is estimated that 80 percent of clinical relevant data in EHRs is found in unstructured text fields. By making such information accessible, NLP-based strategies are able to recognize qualified patients that would not have been detected by the traditional screening methods.

Machines learning has an opportunity to overcome the process of simple eligibility screening and predict enrolment success, patient retention, and recruitment plans (Liu et al., 2021; Yue et al., 2024). Such models are able to combine various types of data such as clinical features, sociodemographic determinants, geographical opportunities and enrollment history of patients to define not only eligible but also likely to enroll and follow the trial. This predictive feature allows dedicating recruitment resources to more efficient companies and more adaptive recruitment strategies. The goal programming and optimization techniques can be used to develop a model of recruitment strategies that will optimize various goals, such as enrollment goals, diversity goals, cost limitations, and time-related requirements (Davis et al., 2014; Planning and Monitoring, 2023). These methods are specifically useful in responding to the problem of providing equitable representation of the diverse populations: they can directly relate diversity objectives as optimization conditions instead of viewing diversity as an afterthought.

The most advanced data-driven recruitment approaches are now large language models (LLMs) and generative AI, which provide features of automated interpretation of protocols and dialogue with patients and smart matching between patient attributes and trial eligibility (Abdallah et al., 2025; Chen et al., 2025; Amoako et al., 2025). Such technologies can also eliminate even more manual work in the recruitment process besides enhancing the quality and personalization of patient-trial matching.

Noteworthy, data-driven methods are not supposed to substitute human assessment and the crucial role that clinicians and research coordinators play during the recruitment process. Instead, they as decision support tools are used to supplement human power so that research teams can be more efficient and effective. These technologies can release clinicians and coordinators by automating routine screening activities and giving intelligent recommendations because the human nature of recruitment that fosters trust and patient education and concerns that are crucial to successful enrollment and retention are not robotized (Meystre et al., 2023). Data-driven recruitment strategies are based on a theoretical background of several fields of study, such as, but not limited to, clinical informatics, machine learning,

operations research, and implementation science. Clinically informatics-wise, these methods utilize secondary clinical data usage and make EHRs research enablers instead of documentation devices.

Machine-learning-wise, they use both supervised and unsupervised machine learning algorithms to learn the patterns in the complex, high-dimensional data, which serve as predictors of eligibility and enrollment success. In terms of operations research, they utilize optimization techniques to develop recruitment strategies that are maximally efficient and meet a variety of constraints. In terms of implementation science, they understand that technological innovation is not enough; to implement it successfully, workflow integration, including user acceptance and management of organizational change, is also needed.

3 EXISTING DATA-DRIVEN STRATEGIES

3.1 Natural language processing and Mining of text

The concept of natural language processing has become an essential technology in clinical recruitment to overcome the severe obstacle of the majority of relevant clinical information in EHRs being in unstructured text. The systematic review of Idnay et al. (2021) has revealed a few 21 NLP systems to prescreen eligibility in clinical studies and has shown the maturity of this practice. NLP techniques used in these systems, such as named entity recognition, relation extraction, negation detection, and temporal reasoning, are used to extract structured information using clinical narratives. TAES (TriAl Eligibility Surveillance), invented by Meystre et al. (2023), can be seen as an example of a state-of-the-art in terms of recruitment using NLP. Applied to a health system in South Carolina, TAES combines NLP algorithms and the OMOP Common Data Model to compare eligible trials and EHR-collected patient data and to match them automatically. It reads and analyzes structured data and unstructured clinical notes and sends real-time alerts to providers when they find that there are potentially eligible patients. During a pilot study, TAES proved automated eligibility surveillance feasibility, although the authors have indicated that it was difficult to achieve high precision and acceptable recall rates.

Fang et al. (2022) proposed a hybrid model, which is a combination of human and machine intelligence to query clinical trials' eligibility. Their system leverages NLP to identify the most pertinent clinical concepts out of both eligibility criteria and the records of patients, and then leverages machine learning to rank the patients by likelihood to be eligible. Of significance, the system includes human expert feedback to continuously refine its predictions, and this shows that human-in-the-loop systems can deliver higher performance than fully automated systems. The conclusion is not isolated, and the larger trends in clinical AI lean towards the relevance of human-AI partnership as opposed to full automation. The modern developments in large language models created more opportunities when it comes to recruitment based on NLP. Yang et al. (2026) created the EC2Seq2Sql, a system based on the usage of an LLM to convert natural language eligibility criteria into queries in a database in the form of structured queries. The system was very precise in eliciting complex eligibility criteria and creating the right queries, indicating that LLMs can save the manual effort needed to use trial protocols to conduct automated screening. On the same note, Zihang et al. (2025) indicated that knowledge graphs provide some benefit of integrating LLMs to support pre-recruitment questionnaires, which reduces the accuracy of pre-screening the patient.

3.2 Predictive Models and Machine Learning

Predictive analytics can be regarded as a second large group of data-driven recruitment strategies, as it is aimed at predicting enrollment outcomes and optimizing recruitment methods. Liu et al. (2021) designed a machine learning method of recruitment based on clinical trial design, prediction of enrolment rates, and schedules using past trial data. They included such aspects as the nature of trials, site factors, prevalence of the diseases, and competitive trial environment. The predictive models attained moderate success levels in predicting the enrollment trend, which offers resource allocation and a valuable tool in planning trials. Yue et al. (2024) developed this practice with Trial-Enroll, a solution that uses deep learning and large language models to predict the success in clinical trial enrollment. Trial-Enroll uses a Deep & Cross Network design to represent the complicated interplay between patient attributes, trial attributes, and context. The system showed better performance than conventional logistic regression models since it obtained an AUC score higher than 0.80 when predicting enrollment likelihood. Notably, the authors have included interpretability techniques to determine the most significant variables that contributed to making the enrollment predictions, which increases the clinical applicability of the model.

Predictive models have been promising in particular disease areas. Mehari et al. (2025) presented supervised machine learning models that predict clinical approval of therapeutic clinical trials of adult patients with low and high-grade glioma. Their models predicting enrollment using information about 1,200 patients reached accuracy rates of 7582%, which indeed is lower in comparison with other methods using the random forest algorithm. The researchers were able to establish the key predictors as the tumor characteristics, the histories of previous treatment, the performance status, and the sociodemographic factors. Such results indicate that domain knowledge implemented in disease-specific models

could present a higher performance in contrast to generic models. Several studies in the field of Alzheimer's disease have shown the usefulness of predictive analytics to enhance a trial population. In the paper by Tam et al. (2021), the models were created to make predictions of cognitive decline in enrichment of Alzheimer's disease clinical trials to allow the creation of more efficient trial designs by identifying the patients who are most likely to progress during the trial duration. Gallini et al. (2025) demonstrated that machine learning-selected cognitive factors based on selection enhance the statistical power of randomized clinical trials of dementia in Alzheimer's, which may shorten the sample sizes by 2030% required. It has been demonstrated that machine learning algorithms are able to detect non-diagnosed prodromal patients with Alzheimer's disease in the general population, increasing the number of potential people who can be eligible (Uspenskaya-Cadoz et al., 2019).

3.3 Electronic Health Records Integration

Having recruitment systems linked to EHRs is an imperative facilitator to real-time and automated patient identification. Gong et al. (2026) designed a real-time and common data model and an AI-based system that does semiautomated patient prescreening during cancer clinical trials. The system becomes interconnected with the institutional EHR and constantly tracks patient data and checks on one or another eligibility criterion as patients are attended to. Once possibly eligible patients are found, the system sends notices to the oncologists and research coordinators, so that they can easily maintain a discussion about the recruitment. The system found 2.5 times the number of eligible patients as traditional recruitment methods in a retrospective assessment, and the positive predictive value was 68%. Several studies indicate the significance of data standardization to HER-based recruitment. Recruitment systems can be implemented in a variety of institutions without such large-scale personalization because the application of common data models, e.g., OMOP (Observational Medical Outcomes Partnership) and FHIR (Fast Healthcare Interoperability Resources), enables recruitment systems to be deployed across multiple institutions without extensive customization (Meystre et al., 2023). This interoperability is especially important for multisite trials, where it is important that patients can be reliably identified across a diverse spectrum of EHR systems. Murphy et al. (2017) discussed the larger issues that need to be addressed with the use of big data from EHRs for translational medicine and clinical care, such as clinical trial recruitment. They identified three key challenges: the lack of management of big data by most EHR systems, the complexity needed to provide relevant features using machine learning algorithms, and the requirement of having constantly updated applications based on the current knowledge. The authors suggested a microservice-based architecture capable of integrating big data, machine learning, and clinical care workflows to provide a technical underpinning for next-generation recruitment systems. Real-world data from EHRs can also be used to inform trial design and the optimization of eligibility criteria. Evans et al. (2021) recommended by the Clinical Trials Transformation Initiative on the use of real-world data for planning eligibility criteria and improving recruitment. Their analysis showed that overly restrictive eligibility criteria are a major factor in recruitment challenges, and that EHR data can be used to simulate the effect of various eligibility criteria upon the size and characteristics of the eligible population. This approach makes it possible to optimize eligibility criteria in an evidence-based manner in order to balance scientific rigor and recruitment feasibility.

3.4 Goal Programming and Approaches to Optimization

Goal programming and optimization techniques are needed for the mathematical description and design of recruitment strategies to simultaneously meet multiple goals and constraints. The seminal work by Davis et al. (2014), prescribed the use of goal programming to optimize clinical trial enrollment methods, providing a good example of how mathematical optimization can be used to balance competing goals such as enrollment targets, budget constraints, and timelines. This approach was greatly advanced by the work on planning and monitoring fair clinical trial enrolment using goal programming (2023). The authors have created a multi-objective goal programming model that specifically includes equity goals so that enrollment plans will adequately represent diverse populations that are defined by protected attributes such as gender, race, and ethnicity. The model was tested using data from the Systolic Blood Pressure Intervention Trial (SPRINT), which included 9361 participants in 102 sites. The goal programming approach was able to produce enrollment plans with representativeness scores of 0.000 (perfect equity) for multiple demographic subgroups and the lowest possible financial costs. Importantly, the model is able to detect deviations from the target enrollment during the recruitment process and dynamically update the recruitment plan to reduce the disparities in the rest of the enrollment period. Ranjith et al. (2024) have expanded optimization methods by introducing game theory in the recruitment process, dealing with the interaction between patients, investigators, and sponsors. Their hybrid classification model and game-theoretic-based approach reflect the fact that recruitment isn't simply a match game; it is a game where multiple players with possible conflicting objectives are involved in strategic decision-making. This framework offers information on the best incentive structures and communication strategies for better enrollment rates. Anisimov (2023) has formulated an analytic approach to predicting the performance of patient enrolment in multi-centre clinical trials, using the combination of stochastic modelling and optimization methods. The methodology allows the capacity to monitor enrollment progress in real time and detect sites that are underperforming compared to target enrollment levels early so that interventions such as additional recruitment resources or site activation can be made promptly.

3.5 Patient-Trial Matching Systems

Comprehensive patient-trial matching systems are an amalgamation of multiple data-driven technologies in one end-to-end recruitment system. Abdallah et al. (2025) created a clinical trial recommendation system (Trial-Match-AI), an end-to-end AI-based system that focuses on streamlining the process of matching patients to relevant trials. The system includes a combination of NLP for understanding the protocol, machine learning for checking eligibility, and recommendation algorithms to rank trials in terms of suitability for the individual patient. Trial-Match-AI has shown a high level of accuracy in finding a suitable trial for a patient, with a precision and recall rate of more than 85% for validation studies.

Theodorou et al. (2023) suggested the use of TREEMENT, an interpretable patient trial matching that was based on a personalized dynamic tree-based memory network. Unlike black-box approaches to deep learning, TREEMENT offers clear explanations for why it made a matching recommendation and, more importantly, what specific eligibility criteria and patient characteristics contributed to a matching recommendation. This interpretability is a critical factor for adoption in the clinic, as those providing the care and the patients need to be able to understand and trust the recommendations from the system. Salman et al.'s patient-trial matching system based on convolutional neural networks proves that deep learning architectures that were developed to analyze images can also be adapted for clinical trial matching by utilizing patient records and eligibility criteria as structured data inputs. Their approach gave a matching accuracy of 7884% in various cancer types. For certain areas of disease, specialized matching systems have been constructed. Hassan et al. assessed the effectiveness of Viz RECRUIT, an artificial intelligence (AI)-based clinical trial recruitment method, on trial enrollment in the EMBOLISE trial for stroke patients. The system is based on computer vision and machine learning techniques to automatically detect potential stroke patients from brain imaging studies, and then match them with relevant trials. Implementation of Viz RECRUIT led to a 45 percent increase in enrollment rates and a 60 percent decrease in the time it takes from patient identification to enrollment discussion. Williamson et al. (2024) built an artificial intelligence system to support clinical trials recruitment based on machine learning techniques to analyze retinal imaging data among patients with geographic atrophy, who would be eligible for ongoing clinical trials. The system reached a sensitivity of 92% and a specificity of 88% in identifying eligible patients' usability. This process showed the merit of using imaging information in recruitment workflows. Claessens et al. (2026) performed a retrospective comparative study of the screening of patients for clinical trials using artificial intelligence (AI) in pancreatic cancer (the PANCRAI pilot study). The AI system processed EHR data such as clinical notes, laboratory results, and imaging reports to detect potentially eligible patients. Compared to traditional recruiting methods, the AI system identified 3.2 times more eligible patients and decreased screening time by 70%, though the authors found challenges in high positive predictive value due to the complexity of pancreatic cancer trial eligibility criteria.

3.6 Geospatial Analysis and Network Analysis

Geospatial analysis and network-based approaches are new ways to optimize recruitment, taking into account geographical factors as well as social networks. Shah et al. (2025) created a clinical trial recruitment system with the assistance of AI and the use of protocol-to-claims NLP and geospatial density modeling. The system uses insurance claims data to identify possible participants, then, with the help of geospatial analysis, optimizes site selection and predicts enrollment feasibility on the basis of patient density, travel distances, and population characteristics. Feasibility outcomes demonstrated that geospatial modeling would produce better enrollment predictions of 2535% of those models that do not consider geographic factors. Udeani et al. (2025) took this one step further by combining both artificial intelligence and geospatial intelligence to optimize the participation of clinical trials. Their own framework combines patient-level predictive models and geographic information systems to route the best places for sites for trials, recruitment events, traveling, and focus outreach campaigns. The approach recognizes that recruitment issues often have strong geographic dimensions, with eligible patients being concentrated in certain areas that may be underserved by existing trial sites (Biswal et al., 2019). This is an approach that uses network analysis to model referral patterns and physician networks, as it is recognized that physician referral is an important recruitment channel. By identifying key physicians who have referral networks and predicting what physicians are most likely to refer patients to specific trials, the approach allows for more targeted and effective approaches to engaging physicians.

4 METHODS TO MANAGE DATA: AI, ML, BIG DATA, AND EHR

4.1 Artificial Intelligence and Machine Learning Architectures

The AI and ML architecture used for clinical trial recruitment comes in all types: traditional statistical models and cutting-edge deep learning systems. At the base are supervised learning algorithms that learn how to predict success in eligibility or enrollment based on the given training data, which is labeled. Logistic regression, despite being simple, is so widely used because it is easy to interpret and can perform well on structured data (Liu et al., 2021). Random forest and gradient boosting methods have demonstrated better performance in a large number of cases, especially when there is a complex and nonlinear relationship between patient characteristics and enrollment outcomes (Mehari et al.,

2025). Deep learning architectures have risen in the limelight for their ability to automatically learn hierarchical representations of features from the raw data. Yue et al. (2024) used a deep & Cross Network (DCN) architecture in the case of Trial-Enroll, which models feature interactions at varying levels of complexity explicitly. The architecture of DCN is composed of a cross-network to effectively capture bounded-degree feature interactions and a deep network to learn arbitrary feature interactions, to combine the strengths of both approaches. This architecture had better performance than conventional models, with 58 percentage points of improvement in the AUC. Convolutional neural networks (CNNs), which were originally developed for image analysis, have been adapted to clinical trial matching by treating patient records and eligibility criteria as structured inputs (Salman et al., 2024). The convolutional layers help the model to learn the local patterns from the data, and for translation invariance, the pooling layer is used, which helps the model to detect relevant patterns in the data, no matter where they are in the input data. This method is especially strong in cases where a variable-length clinical narrative and eligibility criteria are involved.

Recurrent neural networks (RNNs) and variants, such as Long Short-Term Memory (LSTMs), are suitable to work with sequential clinical data, such as longitudinal data EHR records. These architectures have an internal representation of the state that captures temporal dependency to be able to model progression of the disease, trajectory of treatments, and eligibility status over time. Theodorou et al. (2023) use a tree-based memory network in TREEMENT, which is a combination of the sequential processing of RNNs, with the interpretability of decision trees. Transformer architecture and method of attention represent an existing frontier of AI when it comes to clinical trial recruitment. These architectures, on which large language models are based, use self-attention mechanisms to model long-range dependencies in sequential data without the limitations of RNNs. Abdallah et al. (2025) introduced transformer-based models in the case of Trial-Match-AI, in which the system takes into account lengthy clinical protocols and patient records, yet it can still pay attention to the relevant details all along the documents.

4.2 Big Data Analytics- Infrastructure

The computational infrastructure needed for data-driven recruitment systems needs to cope with the volume, velocity, and variety of healthcare big data. Modern recruitment systems process millions of patient records with thousands of pieces of data across structured fields, clinical stories, laboratory results, imaging reports, and genomic data (Murphy et al., 2017). This scale requires distributed computing architectures and data management specialized systems. Hadoop and Spark have become a base for big data analytics in healthcare, and they offer distributed storage and parallel processing capabilities. These platforms allow recruitment systems to be horizontally scaled by contributing computational nodes, unlike having to add more powerful and expensive solitary servers. MapReduce and similar programming models, in which complex analytical tasks can be broken into parallel operations that can be distributed across clusters of commodity hardware. NoSQL databases, such as MongoDB, CouchDB, and Cassandra, allow a flexible data model, which can meet the heterogeneous data structure of healthcare data, and do not need to define a schema. This flexibility is especially useful for recruitment systems that have to combine data from different sources with different data models. However, no standardized schemas may complicate data integration and data querying, which requires some extra data harmonization efforts. Cloud computing platforms such as Amazon Web Services, Google Cloud Platform, and Microsoft Azure have become more relevant to healthcare analytics to provide scalable computational resources, managed services for machine learning, and achieve healthcare data security compliance, such as HIPAA. Cloud-based deployments allow for the dynamic scaling up or down of recruitment systems in relation to their computational requirements and allow for multisite deployments without the need for extensive local infrastructure (Murphy et al., 2017). Data lakes and data warehouses are opposing approaches to the organization of healthcare data for analytics. Data lakes hold raw data in its natural format, storing all data in a place for analysis in the future, but needing data processing when it is being queried. Data warehouses impose structure on data when it is ingested, which makes it possible to query data more quickly, but it may result in the loss of information because it does not conform to the predetermined schema. Modern recruitment systems are often built on hybrid architectures, which can include both data lakes for storing all available data and data warehouses for storing and frequently accessed structured queries.

4.3 EHR-based Surveillance and Alert Systems

EHR-based surveillance systems for clinical trial recruitment need to operate in real-time or near-real-time, continuously monitoring patient data as it is generated during the clinical care process. This requirement poses tremendous technical challenges, as EHR systems weren't designed with the goal of research analytics (they are designed for clinical documentation). The architecture of surveillance systems usually entails multiple factors: data extraction from the source EHRs, data transformation and standardization, eligibility criteria evaluation, and alert generation and delivery (Meystre et al., 2023). Data extraction can be achieved in several ways, such as direct database queries, HL7 messaging interfaces, FHIR APIs, or bulk data exports. Real-time surveillance systems are usually based on event-driven architectures that trigger eligibility evaluation based on changes in pertinent data, such as newly diagnosed patients, laboratory tests, or clinical encounters. This event-driven approach is more efficient than the periodic batch processing and allows to identify eligible patients in a time. The OMOP Common Data Model has become a quasi-standard for standardizing EHR data for research purposes. OMOP offers a common schema with standardized mappings for the vocabulary used to represent clinical concepts between different source systems. Meystre et al. (2023)

showed that the implementation of recruitment systems on top of OMOP-standardized data has decreased the effort needed to implement recruitment systems across a wide range of institutions, since the eligibility criteria in the algorithm could be written once and used across any OMOP-compliant database. Alert systems have to balance between sensitivity, or finding all of the patients who could be eligible, and specificity, or finding the fewest number of false positive alerts, which wastes clinician time. This balance is particularly difficult because of the complexity of trial eligibility criteria that usually include dozens of inclusion and exclusion criteria covering multiple data domains. Meystre et al. (2023) identified that to have positive predictive values greater than 50% and sensitivity greater than 80% carefully tuning of NLP algorithms and eligibility criteria logic was necessary, as well as human review of borderline cases. Alert delivery mechanisms should be integrated with clinical workflows as much as possible to achieve maximum adoption and least disruption. Approaches include in-EHR alerts that are shown during clinical encounters, email notifications to patients and research coordinators, dashboard displays in research offices, and mobile app notifications. The best delivery mechanism will be dependent on the urgency of recruitment, the complexity of eligibility criteria, and the organizational structure of the research team. Gong et al. (2026) found that EHR alerts to oncologists during clinic visits had the highest rates of conversion from identification to enrollment discussion, but had to take into account careful design to avoid alert fatigue.

4.4 Large Language Models and Generative AI

Large language models (LLMs) are a revolutionary technology that can be used for clinical trial recruitment, as they have capabilities that were not possible with traditional NLP approaches. LLMs like GPT4, Claude, etc., and domain-specific models like BioBert, Clinical-Bert, etc. have been pretrained on corpora of text of immense size and thereby are able to understand and generate humanlike text with insane fluency and contextual knowledge (Chen et al., 2025). For clinical trial recruitment, several important abilities of LLMs can be implemented. First, they can interpret the complex eligibility criteria in natural language and convert them into a structured representation or database queries. Yang et al (2026) showed that LLMs can be used with high accuracy in this task, with less need for manual work when operationalizing trial protocols. Second, LLMs can extract relevant clinical information from unstructured clinical notes with a higher degree of accuracy and flexibility than rule-based NLP systems, as they can handle linguistic variation, ambiguity, and context-dependent meanings (Zihang et al., 2025). Third, LLMs have made it possible to have conversational interfaces for patient engagement and education. Subbiah (2025), to examine whether AI-powered TrialGPT can improve patient recruitment for clinical trials, showed that LLM-based chatbots can answer patient questions about clinical trials, answer questions about the eligibility requirements in layman's language, and guide the patient through prescreening questionnaires. These conversational capabilities will perhaps lead to better patient understanding and less of a barrier to decision-making around enrollment, especially for patients with low levels of health literacy. Fourth, LLMs can help with patient-trial matching by being able to reason about semantic similarity between patient characteristics and trial requirements, rather than just conducting keyword matching. Abdallah et al (2025), for example, built in LLMs into trialmatchai to enable more nuanced matching, which takes into account clinical context and implicit eligibility requirements that may not be stated anywhere within protocols. However, there are also challenges for the recruitment of clinical trials with LLMs. Their "black box" nature, which makes it hard to comprehend and validate their reasoning process, gives rise to concerns about reliability and possible biases. Verhagen et al. (2026) examined unrealized trial enrollments after patient-trial matching using LLMs and found that although the LLMs performed well with high sensitivity in imaging potentially eligible patients, they also led to false positives since they misinterpreted complex clinical scenarios or relations among time, patient, and the trial. The authors exhibited the need for human supervision and validation of recommendations generated by LLMs. Privacy and security considerations: LLMs are particularly important to consider due to the requirement for access to sensitive patient data in order to train and run the models for inference. Federated learning approaches, which are discussed in Section 7.3, are possible solutions in which the LLMs can be trained on distributed data without centralizing sensitive information.

4.5 Geospatial Analysis and Network Analysis

Geospatial analysis methods allow for the geographic factors that are important to enrollment feasibility and equity to be included in recruitment systems. Geographic Information Systems (GIS) offer tools for: visualizing where patients are, calculating the travel distances and times, identifying underserved areas, and optimizing the locations of the site (Shah, 2025; Udeani et al., 2025). Spatial clustering algorithms such as DBSCAN and k-means clustering can be used to identify concentrations of eligible potential patients geographically, to inform the decision of where a trial should be located or targeted outreach should take place. Spatial regression models can be used to quantify the relationship between geographic factors (e.g., distance to trial site, urban vs. rural location, and neighborhood socioeconomic characteristics) and enrollment outcomes in order to better predict enrollment and identify geographic barriers. Network analysis methods simulate social and professional relationships that affect clinical trial recruitment. Physician referral networks, patient social networks, and institutional collaborations all play important roles in the success of recruitment. Biswal et al. (2019) used network representation learning to represent networks of physicians to find influential physicians who are well-positioned to recommend patients to trials.

5 EFFICACY AND OUTCOMES OF DATA-DRIVEN APPROACHES

5.1 Recruitment Efficiency and Time Saving

Empirical evidence shows that data-driven strategies can help increase recruitment efficiency in many ways. The finding that has been most consistent is a dramatic decrease in the amount of time that it takes to screen patients for eligibility. Claessens et al. (2026) showed that in a pancreatic cancer trial, patient screening using artificial intelligence eliminated approximately 70% of the time spent on screening, compared to manual chart review. According to Hassan et al (2023), the use of the Viz RECRUIT system was associated with a decrease in the time from patient identification to enrollment discussion by 60% in stroke trials. These time savings are translated to faster enrollment time and shorter trial timelines. Data-driven systems also identify significantly more potential eligible patients as compared to traditional methods. Gong et al. (2026) found that their use of an AI-driven prescreening system found 2.5 times as many eligible patients as traditional recruitment methods in cancer trials. Claessens et al. (2026) in their study identified 3.2 times more eligible patients in pancreatic cancer trials using AI screening. This higher rate of identification is largely driven by the more extensive, systematic identification that is possible because of automated screening systems and the ability to screen at a level unattractive to computer scientists, but essential for achieving much higher resolution rates of the disease. Conversion rates from identification to enrollment are also improved with data-driven approaches, although the amount of improvement in rates varies across studies. Hassan et al. (2023) found a 45% improvement in rates of enrollment after implementation of Viz RECRUIT. Williamson et al (2024) obtained enrollment rates among those patients identified by their AI system for age-related macular degeneration trials was 35% compared to 2025% enrolled historically using traditional methods. These improvements have likely reflected both an increase in the quality of patient identification (matching to eligibility criteria) as well as the timeliness of patient identification (approaching patients earlier in the course of their disease treatment). Predictive models for enrollment success allow for more efficient use of recruitment resources in targeting patients with the greatest potential to enroll in the studies. Yue et al. (2024) showed that the approach to prioritize recruitment outreach using Trial-Enroll may decrease the number of patients that need to be contacted by 3040% while maintaining the same enrollment numbers, as the system will focus effort on candidates with a high probability of enrollment. This efficiency gain is especially useful for trials where available recruitment resources are limited, either in terms of the recruitment budget or staff.

5.2 Cost-Effectiveness Analysis

While comprehensive cost-effectiveness analyses of data-driven recruitment systems are yet to be conducted, there is evidence that suggests a significant potential for cost savings. The main cost savings come from the reduction in the screening time, quicker enrollment, and more efficient use of the research coordinator's time. Huml et al. (2020) estimated that big data approaches to patient recruitment can decrease costs associated with the recruitment of patients by 30-50% per patient by more efficient identification and screening processes. The use of the approach to goal programming developed for equitable enrollment planning demonstrated the ability to minimize total financial costs and ensure fulfillment of enrollment and diversity goals (Planning and Monitoring, 2023). In simulations conducted on SPRINT trial data, the optimized enrollment plans decreased the total costs by 15-20% relative to a set of actual historical patterns of enrollment, mostly by more efficient site selection and resource allocation. However, the upfront costs of implementing data-driven recruitment systems can be significant, including software development or licensing costs, the cost of data infrastructure, the integration of systems with EHR systems, and the cost of staff training. Meystre et al. (2023) stated that implementing their TAES system was a process that required a considerable investment in NLP development, integration with the EHR, and the redesign of the workflow. The volume of trials undertaken is associated with return on investment, with larger research institutions undertaking multiple trials, being able to amortize costs of implementation over a high volume of research studies. Ongoing operational costs include system maintenance, updates to the algorithms, data storage and processing, as well as human oversight of automated recommendations. Murphy and colleagues (2017) said that applications made using big data and machine learning algorithms need constant updates to reflect the changes in medical knowledge, which require sustained investment in system maintenance.

5.3 Diversity and Equity Considerations

Ensuring equitable representation of diverse populations in clinical trials is a matter of both ethics and science to produce generalizable evidence from clinical trials. Data-driven approaches provide a chance and risk to drive diversity and equity in recruitment. On the positive side, there are goal programming and optimization techniques that can incorporate diversity goals directly as constraints in enrollment planning so that enrollment plans achieve the given representation of racial, ethnic, gender, and age subgroups (Planning and Monitoring, 2023). In SPRINT trial simulations, the goal programming model had perfect equity scores (0.000) for several different demographic subgroups, indicating that mathematical optimization can be used to design enrollment regimes that can meet diversity goals while minimizing costs and meeting timelines. Geospatial analysis can be used to determine the location of underserved communities and to optimize site selection to enhance access to the community (Shah, 2025; Udeani et al., 2025). By explicitly modeling travel distances, transportation access, and demographic distributions, these approaches

can help to reduce geographical barriers that are being experienced differentially by minority and rural populations. However, there are also potential risks of using data for perpetuating or increasing disparities by using data-driven approaches. Machine learning models that focus on historical data of enrollment can learn and perpetuate existing biases, as historical trends reflect system-wide barriers to exclusion of minority populations (Ahluwalia et al., 2022). If, based on historical enrollment patterns, predictive models teach that some demographic groups have a lower historical rate of enrollment than others, their de-prioritization of outreach to these groups will create a self-fulfilling prophecy. Chiu et al. (2024) faced just the kind of problem and proposed a machine learning-based recruitment methodology that is designed to provide a solution with a minimum sample size and maximum equity in vaccine clinical trials. Their approach is stratified sampling and targeted recruitment, so that they can ensure adequate representation of the diverse populations, while preserving their statistical power. The methodology was able to achieve 40% reductions in sample size needed and better represent racial and ethnic minorities (2535%) than traditional recruitment approaches. Benedum et al. (2024) have created POPREFINE, a general framework for assessing and optimizing representativeness in clinical trials. The framework offers quantitative metrics for determining the extent to which trial populations are representative of target populations in several dimensions, such as demographics, comorbidities, and disease severity. These metrics allow researchers to jointly know certain gaps in representativeness and start specific recruitment strategies to fix them. Several studies have been done on barriers to recruitment and retention in underrepresented populations, giving insight into ways to construct equitable data-driven approaches. Brooks et al. (2015) identified practical strategies and challenges to increase minority enrollment in oncology trials with an emphasis on the importance of community engagement, communicating in a culturally appropriate manner, and having diverse research staff.

5.4 Outcomes of Retention and Engagement

While most data-driven recruitment research studies have focused on initial enrollment in a trial, retention and engagement are equally important to a trial's success. Nejat et al. (2024) in a longitudinal cohort study that examined sociodemographic drivers of recruitment and attrition in a digital neurological research study, it was observed that younger age, lower educational attainment, and minority race/ethnicity were linked to higher attrition rates. These results indicate the need to target retention strategies to the particular barriers experienced by various demographic groups. Predictive models can help to identify at-risk patients who are at high risk of dropout so that proactive retention interventions can be planned. While specific studies of retention prediction models in clinical trials are scarce in the reviewed literature, based on the broader healthcare analytics literature, it is possible to accurately predict patient disengagement and nonadherence using machine learning models, so similar models could be developed for trial retention.

Digital engagement tools, such as mobile apps, text messaging, and patient portals, provide opportunities to enhance retention by facilitating more frequent contact with patients, using automated reminders, and providing convenient communication channels. However, Nejat et al. 2024 found that digital tools, if less accessible or acceptable to particular groups of demographics, may accidentally make disparities worse and thus the need for multimodal engagement strategies. Wilkins et al (2021) also pointed out that Recruitment Innovation Center (the Recruitment Innovation Center) developed novel strategies for recruiting and retaining clinical trial participants and aimed to present person-centered approaches. Their approach focuses on a patient-level understanding of barriers, patient preferences, and recruitment and retention planning based on that understanding. While not necessarily targeting data-driven approaches, their framework gives some conceptual weight to how predictive analytics and personalization algorithms could play out in terms of retention.

6 IMPLEMENTATION CHALLENGES AND BARRIERS

6.1 Technical and Infrastructure Challenges

Despite the potential of data-driven recruitment approaches, there are significant technical challenges to making these approaches widely adopted. The heterogeneity of the EHR system among institutions poses tremendous problems for integration. Different EHR vendors have different data models, terminologies, and interfaces, and hence, it requires custom integration work to integrate each system (Murphy et al., 2017). Even within individual institutions, there might be several EHR systems for use in the various clinical departments or service lines that further complicate integration. Data quality problems are another major problem. Data collected in EHRs are for clinical purposes and not for research; they have missing data, inconsistent data documentation, and errors. Clinical notes can hold contradictory information, diagnoses that are outdated, or copypasted content not reflect the current state of the patient. Different units/ reference ranges may be used in laboratory results among institutions. These data quality issues can cause significant destruction in the performance of automated screening systems, resulting in false positives and false negatives (Meystre et al., 2023).

Computational performance and scalability are issues of real-time surveillance systems. Processing millions of patient records using complex algorithms of NLP and machine learning models is computationally expensive. Latency of the order for real-time alerting requires efficient algorithms and optimized implementations. Cloud computing is a capability that helps you achieve scalability at the cost of increased complexity relating to data security and regulatory compliance. Algorithm maintenance and updating take constant efforts and expertise. As medical knowledge changes, eligibility criteria change, and new trials start, recruitment systems need to be changed accordingly. Machine learning models can lose accuracy after a while due to changes in patient population and clinical practice (concept drift), which requires machine models to be retrained periodically. Murphy et al. (2017) stated that for some of the applications, a nightly update to represent the up-to-date knowledge will be needed, and a huge maintenance load will arise.

6.2 Quality of Data and Interoperability

Data interoperability: The ability to exchange and use data between different systems and organizations is a fundamental challenge in data-driven recruitment. While there are standards for interoperability, like HL7 FHIR and OMOP CDM, the adoption of these standards is less than full and close to inconsistent. Many EHR systems support FHIR interfaces, but not all of these resources and data elements are exposed, making it hard to port recruitment applications. Vocabulary and terminology mapping is where there are particular challenges. Different institutions may have different ways of coding (ICD10, SNOMED CT, LOINC, RxNorm) or may have local codes for the same clinical concepts. Mapping between these terminologies is complex and prone to errors, especially in the case of nuanced clinical concepts. The OMOP CDM overcomes this using standardized vocabulary mappings, for which creating and retaining mappings requires substantial work (Meystre et al., 2023). Temporal data quality problems make eligibility evaluation difficult. Many of the eligibility criteria are temporal (e.g., "diagnosis within the past 6 months", "no chemotherapy for at least 4 weeks"). However, EHR timestamps may indicate documentation time as opposed to event time, and historical data may be incomplete or inaccurate. Determining the actual time order of clinical events from EHR data requires complex temporal reasoning algorithms that take these problems of data quality into account. Missing data involves technical as well as methodological problems. Patients with incomplete EHR data may be systematically different from patients with complete data, leading to selection bias if automated systems to screen for abnormalities fail to include patients with missing information. Imputation methods can solve the missingness problem, but may have an error if the missingness patterns are not properly modeled.

6.3 Regulations and Ethics Considered

The regulatory landscape with regard to AI-enabled clinical trial recruitment is in flux, and this change is creating uncertainty for developers and implementers. The FDA has issued guidance on how AI/ML in medical devices are used, the applicability of FDA regulations towards recruitment systems is unclear. If recruitment systems are considered clinical decision support tools that affect patient care decisions, the systems may be subject to FDA oversight. However, if they are 'research tools' supporting trial operations, then they may fall out of the jurisdiction of the FDA. Privacy, security regulations, HIPAA specifically in the United States, they have strict regulations and requirements regarding the use and disclosure of protected health information. Automated screening systems that access patient data without patient consent must be implemented under HIPAA exceptions for research or healthcare operations. The interpretation of these exceptions differs across institutions and legal jurisdictions, and hence creates challenges in compliance (Murphy et al., 2017). Informed consent leads to ethical challenges for automated recruitment systems. Patients may not know that their information is being used in an EHR to identify them for trial recruitment. While such use is usually allowed under HIPAA and research rules, it raises questions about patient autonomy and the right to control, similar to issues about one's health information. Some institutions have put in place opt-out mechanisms that allow patients to refuse automated screening, but these mechanisms are not in universal use. If the machine learning models are trained on historical data that reflect the existence of systemic biases and disparities, there are chances that these are being perpetuated or amplified in the machine learning model's predictions and recommendations (Ahluwalia et al., 2022). In order to ensure fairness, it is important to carefully consider the composition of the training data, the design of the algorithm, and to closely monitor the effects on different demographic groups. Transparency and explainability are important both from an ethical and a practical perspective. Clinicians and patients require knowledge of the reasons given in identifying particular patients as potentially eligible and in the way of deciding on matches. Blackbox models that generate recommendations without being able to explain them may not be trusted or adopted. Theodorou et al. in TREEMENT highlighted the importance of an interpretable model in order to prove that transparency can be obtained without harming the predictive capacity.

6.4 Organizational and Workflow Integration

Successful implementation of data-driven recruitment systems necessitates more than just technical capability but should go to the level of integration into existing organizational structures and clinical workflows. Research coordinators, clinicians, and administrators need to understand, believe in, and make appropriate use of these systems. Resistance to change, conflicting priorities, and breaks in the workflow can inhibit adoption despite systems being proven to be technically effective. Alert fatigue is a well-documented problem with clinical decision support systems,

and it is the same with recruitment alerts. If automated systems alert too many times or if these alerts are false positives, then clinicians may overlook the alerts or disable the alerts. Meystre et al. (2023) found that in order to obtain acceptable positive predictive values with high sensitivity, close attention to tuning and incorporating human review was necessary with borderline cases. Role clarity and responsibility assignment are important in effective implementation. When automated systems identify potentially eligible patients, who is responsible for the review of the alert, contacting the patient, and conducting the eligibility assessment? In some institutions, these functions are carried out by research coordinators; in others, recruitment discussions are expected to be started by the treating physicians. Clear protocols and definitions of roles are needed so that the process of automated identification results in actual enrollment. Training and support are the keys to successful adoption. Research personnel must receive training not only in the use of recruitment systems but in the capabilities and limitations of these systems as well. And clinicians require education on how to interpret automated recommendations and incorporate them into the discussions of patient care. Ongoing technical assistance is required to help solve problems and questions as they develop. Organizational culture and leadership play an important role in the success of implementation. Institutions that have strong research cultures and leadership commitment to innovation and resources for implementation are more likely to be successful in adopting data-driven approaches to recruitment. Conversely, institutions facing resource constraints, clashing priorities, and/or skepticism of technology may have a hard time with implementation regardless of technical merit.

7 FUTURE DIRECTIONS AND RECOMMENDATIONS

7.1 New Developments in Technology and Innovation

The exciting pace of innovation in the field of AI and data science means that existing methods of recruiting people through data would be just the start of things to come for years to come. A number of new technologies and methodological innovations have specific promise for the recruitment/retention of clinical trials. Multimodal learning that combines different types of data, such as structured EHR data, clinical narratives, medical imaging, genomic information, and patient-created data from wearables and mobile devices, promises to address further patient characterization and eligibility assessment. Current systems usually only concentrate on one or two types of data; future systems that can effectively use all available sources of data could deliver superior performance. Reinforcement learning, a machine learning paradigm where algorithms learn how to achieve certain goals with the help of trial and error, may lead to adaptive recruitment strategies that learn and continuously optimize based on what they observe. Rather than having set protocols for recruitment, reinforcement learning systems could adapt outreach strategies, approaches to communication, and resource allocation dynamically in response to real-time feedback on what is and is not working. Causal inference approaches that look beyond correlation to figure out what factors really lead to success in enrollment and retention could yield deeper insights into what factors actually drive enrollment and retention success. Understanding the mechanisms by which these things work would allow for more targeted interventions and successful recruitment strategies. Current predictive models establish associations, but do not necessarily show what pathways are causal. Virtual trial design and recruitment planning could be achieved by using digital twins' computational models that simulate individual patients. By constructing digital models of patient populations, researchers would be able to simulate the effect of various eligibility criteria, recruitment methods, and trial designs before actual recruitment started, mitigating the risk of expensive recruitment failures. Blockchain technology is one of the possible solutions to provide secure, decentralized management of patient consent and data sharing institutions. Blockchain-based systems may allow patients to retain control over their own health data, but selectively release it for recruitment systems, allowing concerns around privacy to be met, as well as multi-site recruitment.

7.2 Equity-Centered Design Principles

Future data-driven recruitment systems need to be built from the ground up with the principle of equity as a central consideration, not an afterthought. This means going beyond the point of measuring the outcomes of diversity to actively and deliberately working to design systems that support equitable access and representation. Fairness-aware machine learning methods that make explicit efforts to maximize fairness across demographic classes should be standard practice. These techniques can add fairness constraints to the training of the model, which can make sure that predictive models don't systematically disadvantage any group. There are different definitions of fairness (demographic parity, equalized odds, calibration), and the definition to be used is dependent on the application and ethical issues for each (Planning and Monitoring, 2023). A community-engaged design process involving various patient communities in the development and evaluation of recruitment systems can ensure that these systems are acceptable, accessible, and effective for all populations. Patients from underrepresented communities can offer invaluable insights into barriers, preferences, and cultural considerations that may not be apparent to the researchers and developers. Multilingual and culturally adapted interfaces are necessary to reach diverse populations. Recruitment systems should be multilingual and culturally sensitive in their communications approach. Large language models present opportunities for automated translation and cultural adaptation, or any human review and validation of the results to ensure accuracy and appropriateness. Addressing social determinants of health in recruitment strategies. The clinical eligibility to

participate in a clinical trial is one aspect of trial participation. Transportation barriers, childcare needs, work schedule, and financial issues all play a role in enrollment and retention. Future systems should take these things into account as part of the matching algorithm and help decision support for nonclinical barriers.

7.3 Federated Learning and Privacy-Preserving Methods

Federated learning is a paradigm shift in training machine learning models on healthcare distributed data without centralizing sensitive healthcare data. In federated learning, models are learned locally, at each institution, on the local data, and only model parameters (not patient data) are shared and aggregated to obtain a global model. This way, it is possible to collaborate with several sites and maintain patient privacy and compliance with data governance requirements. For clinical trial recruitment, federated learning may make it possible to develop recruitment models trained on data from multiple institutions to enhance generalizability and performance without sharing data across institutions. Institutions could also benefit from collective learning, but without sharing their data. This approach is especially useful for rare diseases, for which one single institution will have insufficient data to train sound models. Differential Privacy is a set of mathematical guarantees of the privacy of individual patient data by ensuring that they cannot be inferred from model outputs and from aggregated statistics. By adding calibrated noise to the data or parameters of a model, the differential privacy can guarantee that the inclusion or exclusion of any individual patient does not have a significant impact on results. This approach could allow for more general data sharing for recruitment research while allowing for patient privacy protections. Homomorphic encryption is a way of doing computations on encrypted information without decrypting it, which makes it possible to analyze sensitive information in a secure way. While existing homomorphic encryption techniques are relatively costly, research and development continue to yield improvements to the efficiency of such sources, so future use of these homomorphic encryption techniques could be suitable for practical clinical trial recruitment in the near future. Secure multiparty computation is the ability of several parties to simultaneously compute functions over their combined data without knowing the data of any particular party from the other parties. This approach may allow multisite eligibility screening, whereby each institution has control over its data while they collectively identify eligible patients in each institution.

7.4 Real-Time Adaptive Recruitment Strategies

Future recruitment systems should change from static protocols to real-time adaptive approaches for continuous optimization based on observed outcomes. Adaptive designs, which are already ingrained in clinical trial methodology for the allocation of treatment, can be expanded to recruitment methodology. Real-time enrollment monitoring and forecasting can be used to detect underperforming sites or demographic groups at an early stage in the recruitment process in order to implement early interventions. Anisimov (2023) developed methodologies of predicting enrollment performance, but the approaches there could be improved through more sophisticated machine learning methods and could be integrated into automated decision support systems. Dynamic eligibility criteria optimization might change the inclusion and exclusion criteria throughout the recruitment process, depending on the observed rates of recruitment and characteristics of patients. If some criteria are determined to be too narrow with little scientific basis, the criteria may be loosened to broaden the population in question. This approach requires due consideration of the scientific validity and regulatory requirements, but has the potential to increase the efficiency of recruitment substantially. Personalized recruitment messaging and communication strategies could be developed based on individual patient characteristics, preferences, and barriers. Rather than involving a one size fits all recruitment materials, adaptive systems could create individualized messages that address specific patient concerns and highlight aspects of the trial that may be most relevant to individual patients. Large language models provide the capabilities of generating personalized content at scale. Adaptive resource allocation could dynamically change allocation of recruitment resources (staff time, advertising budget, activation of sites) based on up-to-date performance data. Sites or strategies that are actually working well may be assigned more resources, while strategies that aren't working well may be altered or discontinued. This adaptive approach can potentially have a substantial effect on overall recruitment efficiency.

7.5 Policy and Regulatory Recommendations

Realizing the full potential of data-driven approaches to recruitment requires the right policy and regulatory frameworks that strike the right balance between innovation and patient protection. Several policy recommendations can be proposed from this review. First, some regulatory clarity is required in terms of oversight of AI-enabled recruitment systems. The FDA and other regulatory agencies need to give the public clear information about when recruitment systems should be subject to regulatory review, the evidence necessary to show safety and effectiveness, and how they should continue to monitor for safety and effectiveness. This clarity would eliminate some uncertainty for developers and implementers and provide for appropriate oversight. Second, data sharing and interoperability policies should be strengthened in order to enable multisite recruitment studies. While there is a great need for privacy protections, overly restrictive data governance policies can hinder useful research. Policies should facilitate proper data sharing in the interest of recruitment research while preserving patient privacy using technical safeguards such as federated learning and differential privacy. Third, funding agencies should prioritize equity in data-driven recruitment. The National Institutes of Health and other funders should support work in these topics as follows: Support research

on fairness-aware algorithms, community-engaged design and interventions, addressing barriers of underrepresented populations. There should be incentivisation of diversity at trial enrolment and incentivisation of innovation to improve equity. Fourth, standards development organizations should expedite efforts to develop standards for interoperability in clinical trial recruitment. While we have FHIR and OMOP CDM for the foundations, there are some more standards that are necessary for representing eligibility criteria, trial protocols, and recruitment workflows in computable formats. Standardization would lower the costs of implementation and open up widespread use of the data-driven approaches. Fifth, educational initiatives should be put in place to prepare the clinical research workforce for the data-driven future. Training programs for clinical research coordinators, investigators, and informaticians should include training on AI/ML methods, data science, and ethical concerns in algorithmic decision-making. Continuing education should keep the workforce up to date with fast-changing technologies. Sixth, patient participation in the creation and control of recruitment systems should be formalized. Patient advisory boards, community review processes, and the inclusion of patients in the development of algorithms can help to ensure that systems are designed with patient needs and preferences in mind. Patients should take a meaningful sway on how their data is used for recruitment.

8 CONCLUSION

Patient recruitment and retention continue to be major barriers in clinical trials, with a significant impact on timelines and costs of experiments and validity of the research results. This review brings out that the use of data-driven approaches, especially those that use artificial intelligence, machine learning, big data analytics, and the integration of electronic health records, provides a transformative solution. Evidence from the latest studies indicates that there are substantial gains in recruitment efficiency, such as less screening time, better identification of eligible participants, and increased enrolment rates. These technologies, including natural language processing, predictive analytics, and automated patient-trial matching, are a paradigm shift from what has been done in the past regarding recruitment to more proactive, precise, and scalable systems.

Despite these advancements, there are still important challenges to consider, including the issue of data interoperability, as well as ethical and regulatory concerns, and risks associated with bias and inequitable representation. Addressing these issues calls for equity-centered design skills, strong governance structures, and powerful interdisciplinary collaboration between researchers, healthcare institutions, regulators, and communities. Emerging innovations like federated learning, multimodal systems, and adaptive AI models represent the potential to further advance recruitment strategies. Overall, which have proven that data-driven approaches show clear potential for translating longstanding recruitment challenges, their successful and equitable implementation will require coordinated efforts to both scale solutions and ensure their inclusivity, as well as integrate them into real-world clinical research settings.

STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

Disclosure of conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] Abdallah, M., Nakken, S., Georges, M., Bierkens, M., Galvis, J., Groppi, A., & Nikolski, M. (2026). TrialMatchAI: an end-to-end AI-powered clinical trial recommendation system to streamline patient-to-trial matching. Nature Communications.

[2] Agoulnik, I. (2025). Intelligent acceleration: Redefining clinical trials for drugs and devices with AI. Current Investigations in Clinical and Medical Research, 3(1), 000523. https://doi.org/10.53902/cicmr.2025.03.000523

[3] Ahluwalia, M., Onwuanyi, A., Agu, E., & Kpodonu, J. (2022). Advocating for a Path to Increase Diversity in Enrollment in Cardiovascular Clinical Trials. JACC. Advances, 1(5), 100152. https://doi.org/10.1016/j.jacadv.2022.100152

[4] Amini-Philips, A., Ibrahim, A. K., & Eyinade, W. (2022). Patient Recruitment and Retention Innovations to Improve Outcomes in Multi-Site Cancer Studies.

[5] Amoako, E. K. W., Boateng, V., Ajay, O., Adukpo, T. K., & Mensah, N. (2025). Exploring the role of machine learning and deep learning in anti-money laundering (AML) strategies within US financial industry: A systematic review of implementation, effectiveness, and challenges. Finance & Accounting Research Journal, 7(1), 22-36.

[6] Anisimov, V. (2023, July). An Analytic Methodology for Forecasting Patient Enrolment Performance in Multicentre Clinical Trials: Forecasting Patient Enrolment Performance in Clinical Trials. In Proceedings of the 2023 6th International Conference on Mathematics and Statistics (pp. 91-97).

[7] Areh, J. (2025). Improving Black Participation in Cardiovascular Clinical Trials: Business Strategies for Enhancing Representation and Promoting Health Equity (Doctoral dissertation, Northeastern University).

[8] Benedum, C. M., Sarkar, S., Bozkurt, S., Bhagat, R., Richie, N., Lavery, B., & Griffith, S. D. (2025). POP-REFINE: A Comprehensive Framework for Evaluating and Optimizing Representativeness in Clinical Trials. Clinical pharmacology and therapeutics, 117(4), 1051–1060. https://doi.org/10.1002/cpt.3543

[9] Bikou, A. G., Deligianni, E., Dermiki-Gkana, F., Liappas, N., Teriús-Padrón, J. G., Beltrán Jaunsarás, M. E., Cabrera-Umpiérrez, M. F., & Kontogiorgis, C. (2024). Improving Participant Recruitment in Clinical Trials: Comparative Analysis of Innovative Digital Platforms. Journal of medical Internet research, 26, e60504. https://doi.org/10.2196/60504

[10] Biswal, S., Xiao, C., Glass, L. M., Milkovits, E., & Sun, J. (2020). Doctor2Vec: Dynamic Doctor Representation Learning for Clinical Trial Recruitment. Proceedings of the AAAI Conference on Artificial Intelligence, 34(01), 557-564. https://doi.org/10.1609/aaai.v34i01.5394

[11] Brathwaite J, Wolgast M, Bickhart S. Viable Strategies to Increase Clinical Trial Patient Diversity. Al-Kindy Col. Med. J [Internet]. 2023 Aug. 30 [cited 2026 Mar. 29];19(2):133-6. Available from: https://jkmc.uobaghdad.edu.iq/index.php/MEDICAL/article/view/1106

[12] Brooks, S. E., Muller, C. Y., Robinson, W., Walker, E. M., Yeager, K., Cook, E. D., Friedman, S., Somkin, C. P., Brown, C. L., & McCaskill-Stevens, W. (2015). Increasing Minority Enrollment Onto Clinical Trials: Practical Strategies and Challenges Emerge From the NRG Oncology Accrual Workshop. Journal of oncology practice, 11(6), 486–490. https://doi.org/10.1200/JOP.2015.005934

[13] Burns, A. E., Tumberger, J., Brewe, M., Bartkoski, M., & Stancil, S. L. (2025). Meeting clinical recruitment milestones in an academic center: a data-driven, visual approach. JAMIA open, 8(5), ooaf125. https://doi.org/10.1093/jamiaopen/ooaf125

[14] Chan, J., Jin, Q., Wan, N., Floudas, C. S., Xue, E., & Lu, Z. (2025). Recommending Clinical Trials for Online Patient Cases using Artificial Intelligence. AMIA ... Annual Symposium proceedings. AMIA Symposium, 2024, 228–237.

[15] Chen, H., Li, X., He, X., Chen, A., McGill, J., Webber, E. C., Xu, H., Liu, M., & Bian, J. (2025). Enhancing Patient-Trial Matching With Large Language Models: A Scoping Review of Emerging Applications and Approaches. JCO clinical cancer informatics, 9, e2500071. https://doi.org/10.1200/CCI-25-00071

[16] Chiu, R., Tatara, E., Mackesy-Amiti, M. E., Page, K., Ozik, J., Boodram, B., Dahari, H., & Gutfraind, A. (2024). Reducing Sample Size While Improving Equity in Vaccine Clinical Trials: A Machine Learning-Based Recruitment Methodology with Application to Improving Trials of Hepatitis C Virus Vaccines in People Who Inject Drugs. Healthcare (Basel, Switzerland), 12(6), 644. https://doi.org/10.3390/healthcare12060644

[17] Choudhury, S. O., Dutta Majumdar, A., Mahon, R., Dutta, S. K., & Bhattacharyya, S. (2025). Looking back to move forward: can historical clinical trial data and machine learning drive change in participant recruitment in anticipation of future value assessments?. Trials, 26(1), 413. https://doi.org/10.1186/s13063-025-09160-7 Claessens, N. H. P., et al. (2026). AIdriven patient screening for clinical trials in pancreatic cancer: The PANCRAI pilot retrospective comparative study. JMIR Cancer, 12, e80268. https://doi.org/10.2196/80268

[18] Daniel, C., Ammour, N., Kremer, A., Kalko, S., Worrell, C., & Couvert, C. (2023). D3. 3 Generic process for an EHR-based patient cohort for IRPs.

[19] Davis, J. M., Sandgren, A. J., Manley, A. R., Daleo, M. A., & Smith, S. S. (2014). Optimizing Clinical Trial Enrollment Methods Through "Goal Programming". Applied clinical trials, 23(6-7), 46–50.

[20] Evans, S. R., Paraoan, D., Perlmutter, J., Raman, S. R., Sheehan, J. J., & Hallinan, Z. P. (2021). Real-World Data for Planning Eligibility Criteria and Enhancing Recruitment: Recommendations from the Clinical Trials Transformation Initiative. Therapeutic innovation & regulatory science, 55(3), 545–552. https://doi.org/10.1007/s43441-020-00248-7

[21] Fang, Y., Idnay, B., Sun, Y., Liu, H., Chen, Z., Marder, K., Xu, H., Schnall, R., & Weng, C. (2022). Combining human and machine intelligence for clinical trial eligibility querying. Journal of the American Medical Informatics Association : JAMIA, 29(7), 1161–1171. https://doi.org/10.1093/jamia/ocac051

[22] Gallini, J. W., Baucom, Z. H., & Tripodis, Y. (2025). Cognitive Factor‐Based Selection Increases Power in Alzheimer's Dementia Randomized Clinical Trials. Alzheimer's & Dementia, 21(Suppl 6), e099097. https://doi.org/10.1002/alz70860_099097

[23] Gardner LL, Raeisian Parvari P, Seidman M, Holden RJ, Fowler NR, Zarzaur BL, Summanwar D, Barboi C, Boustani M (2025). Improving the User Interface and Guiding the Development of Effective Training Material for a Clinical Research Recruitment and Retention Dashboard: Usability Testing Study. JMIR Form Res 2025;9:e66718. DOI: 10.2196/66718. PMID: 39992743.PMCID: 11875103

[24] Gong, Y. (2024). Next generation phenotyping in clinical trial based on realworld data. Journal of Biomedical Informatics, 152, 104615.

[25] Gong, G., Liu, J., Pandya, S., Taborda, C., Wiesendanger, N., Price, N., Byron, W., Coppi, A., Young, P., Wiess, C., Dunning, H., Barganier, C., Brodeur, R., Fischbach, N., LoRusso, P., Pusztai, L., Kim, S. Y., Rozenblit, M., Cecchini, M., Mongiu, A., … Kunz, P. L. (2026). Clinical Trial Patient Matching: A Real-Time, Common Data Model and Artificial Intelligence-Driven System for Semiautomated Patient Prescreening in Cancer Clinical Trials. JCO clinical cancer informatics, 10, e2500262. https://doi.org/10.1200/CCI-25-00262

[26] Gui, X., Lv, H., Wang, X., Lv, L., Xiao, Y., & Wang, L. (2025). Enhancing hepatopathy clinical trial efficiency: a secure, large language model-powered pre-screening pipeline. BioData mining, 18(1), 42. https://doi.org/10.1186/s13040-025-00458-5

[27] Guillotel, L. A., Lesimple, T., Zekri, O., Cuggia, M., & Campillo-Gimenez, B. (2026). Case-based reasoning for clinical trial recruitment tools in oncology: When you need patients to find patients. International journal of medical informatics, 211, 106301. https://doi.org/10.1016/j.ijmedinf.2026.106301

[28] Hassan, A. E., Ravi, S., Desai, S., Saei, H. M., Mckennon, E., & Tekle, W. G. (2025). An artificial intelligence (AI)-based approach to clinical trial recruitment: The impact of Viz RECRUIT on enrollment in the EMBOLISE trial. Interventional neuroradiology : journal of peritherapeutic neuroradiology, surgical procedures and related neurosciences, 31(6), 739–744. https://doi.org/10.1177/15910199231184604

[29] Horgan, D., Paulson, J., Loaiza-Bonilla, A., Svedman, C., Malapelle, U., Lorca, F. P., Rahsheed, H. M. A., Hofman, P., Kachnowsk, S., Schneider, D., & Subbiah, V. (2026). A unified framework for pre-screening and screening tools in oncology clinical trials. NPJ precision oncology, 10.1038/s41698-026-01306-3. Advance online publication. https://doi.org/10.1038/s41698-026-01306-3

[30] Huml, R. A., Dawson, J., Lipworth, K., Rojas, L., Warren, E. J., Manaktala, C., & Huml, J. R. (2020). Use of Big Data to Aid Patient Recruitment for Clinical Trials Involving Biosimilars and Rare Diseases. Therapeutic innovation & regulatory science, 54(4), 870–877. https://doi.org/10.1007/s43441-019-00009-1

[31] Idnay, B., Dreisbach, C., Weng, C., & Schnall, R. (2021). A systematic review on natural language processing systems for eligibility prescreening in clinical research. Journal of the American Medical Informatics Association : JAMIA, 29(1), 197–206. https://doi.org/10.1093/jamia/ocab228

[32] Liu, J., Allen, P. J., Benz, L., Blickstein, D., Okidi, E., & Shi, X. (2021). A machine learning approach for recruitment prediction in clinical trial design. arXiv preprint arXiv:2111.07407.

[33] Mehari, M., Warrier, G., Dada, A., Kabir, A., Haskell-Mendoza, A. P., Tripathy, A., Jha, R., Nieblas-Bedolla, E., Jackson, J. D., Gonzalez, A. T., Reason, E. H., Flusche, A. M., Reihl, S., Dalton, T., Negussie, M., Gonzales, C. N., Ambati, V. S., Desjardins, A., Daniel, A. G. S., Krishna, S., … Hervey-Jumper, S. L. (2025). Predicting therapeutic clinical trial enrollment for adult patients with low- and high-grade glioma using supervised machine learning. Science advances, 11(23), eadt5708. https://doi.org/10.1126/sciadv.adt5708

[34] Meystre, S. M., Heider, P. M., Cates, A., Bastian, G., Pittman, T., Gentilin, S., & Kelechi, T. J. (2023). Piloting an automated clinical trial eligibility surveillance and provider alert system based on artificial intelligence and standard data models. BMC medical research methodology, 23(1), 88. https://doi.org/10.1186/s12874-023-01916-6

[35] Miller Kelvin, Chanh Trung Huynh, Rachel Luedtke, Mary Alice Anderson, Melissa Harris, Michael Johnstone, Lauren Murphy, Nicholas Morgan, Rachel Harmon, AnnaJane Ward, Blathnaid Donovan, Christie Rice, Scott Sherrin, Annie Darmofal, Danielle Skelly, Matthew M. CooneyThe Tempus TIME trial portal and clinical trial enrollment. Journal of Clinical Oncology, 41(16_suppl), 1583. https://doi.org/10.1200/jco.2023.41.16_suppl.1583

[36] Milliken, T., Beiler, D., Hoffman, S., Olenginski, A., & Troiani, V. (2024). Recruitment in Appalachian, Rural and Older Adult Populations in an Artificial Intelligence World: Study Using Human-Mediated Follow-Up. JMIR formative research, 8, e38189. https://doi.org/10.2196/38189

[37] Murphy, S., Castro, V., & Mandl, K. (2017). Grappling with the Future Use of Big Data for Translational Medicine and Clinical Care. Yearbook of medical informatics, 26(1), 96–102. https://doi.org/10.15265/IY-2017-020

[38] Nejat, P., Bachman, A. D., Stubbs, V. M., Duffy, J. R., Stricker, J. L., Herasevich, V., Jones, D. T., Utianski, R. L., & Botha, H. (2026). Sociodemographic Drivers of Recruitment and Attrition in Digital Neurological Research: Longitudinal Cohort Study. Journal of medical Internet research, 28, e83432. https://doi.org/10.2196/83432

[39] Ololade, M., Agarwal. N., Poudel, D., Clement,T.,. (2024). Predictive analytics for identifying highprobability clinical trial participants. Journal of Clinical Research Best Practices, 20(8), 112.

[40] Ranjith, E. Manvitha, Lipi Vithal, Caris Harshitha Payarda (2024). Optimizing patient recruitment for clinical trials: A hybrid classification model and gametheoretic approach for strategic interaction. Industrial Engineering Journal, 53(12), 12451258. https://doi.org/10.36893/iej.2024.v53i12.024

[41] Salman, Y., Mohammed, E. A., & Hui, C. (2024). Improving patient clinical trial matching using convolution neural networks. BMC Medical Informatics and Decision Making, 24(1), 156.

[42] Samuel D. A., Korang, A., Umoren, J., & Donkor, A. A. (2024). The role of artificial intelligence and machine learning in optimizing U.S. healthcare supply chain management. World Journal of Advanced Research and Reviews, 24(2), 1996–2002. https://doi.org/10.30574/wjarr.2024.24.2.3343

[43] Shah, A. (2025). AI-powered clinical trial recruitment: Protocol to claims NLP, geospatial density modeling, and feasibility outcomes. International Journal of Computational and Experimental Science and Engineering, 11(4), 789–798. https://doi.org/10.22399/ijcesen.4153

[44] Subbiah, V. (2025). Can AI-Powered TrialGPT Enhance Patient Recruitment for Clinical Trials? AI in Precision Oncology. 2025;2(1):16-18. doi:10.1089/aipo.2024.0056

[45] Tam, A., Laurent, C., Gauthier, S., Dansereau, C., & Alzheimer's Disease Neuroimaging Initiative. (2022). Prediction of cognitive decline for enrichment of Alzheimer's disease clinical trials. The Journal of Prevention of Alzheimer's Disease, 9(3), 400-409.

[46] Theodorou, B., Xiao, C., & Sun, J. (2023). TREEMENT: Interpretable patient-trial matching via personalized dynamic tree-based memory network. Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2234–2244. https://doi.org/10.1145/3580305.3599814

[47] Udeani, G., Campbell, A. D., Soni, S., & Ezeanochie, N. (2025). Advancing clinical trial participation: Leveraging artificial and geospatial intelligence. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5287394

[48] Umoren, J., Korang, A., Utomi, E., Adukpo, T. K., & Mensah, N. (2025). The importance of utilizing big data analytics in US healthcare supply chain management.

[49] Uspenskaya-Cadoz, O., Viatre, V., de la Cruz, M., Voisin, T., Amouyel, P., & Vellas, B. (2019). Machine learning algorithm helps identify non-diagnosed prodromal Alzheimer's disease patients in the general population. The Journal of Prevention of Alzheimer's Disease, 6(3), 185–191. https://doi.org/10.14283/jpad.2019.10

[50] Verhagen, M., van der Heijden, J., Elfrink, A. K. E., de Vries, R., & Tanis, P. J. (2026). Understanding unrealized trial enrollments following patient-to-trial matching with large language models. Surgery, 179(4), 109915. https://doi.org/10.1016/j.surg.2025.109915

[51] Wilkins, C. H., Edwards, T. L., Stroud, M. R., Jerome, R. N., Pulley, J. M., Bernard, G. R., & Harris, P. A. (2021). The Recruitment Innovation Center: Developing novel, person-centered strategies for clinical trial recruitment and retention. Journal of Clinical and Translational Science, 5(1), e154. https://doi.org/10.1017/CTS.2021.841

[52] Williamson, T. H., Gashau, S., Naderi, K., & Mahroo, O. A. (2024). Artificial intelligence to facilitate clinical trial recruitment in age-related macular degeneration. Ophthalmology Science, 4(6), 100566. https://doi.org/10.1016/j.xops.2024.100566

[53] Wong, K. H., Wong, K. Y., & El-Amir, H. (2024). Optimizing patient recruitment in global clinical trials using nature-inspired metaheuristics. Figshare. https://doi.org/10.6084/m9.figshare.25152289

[54] Yang, C., Sun, J., Zhang, L., & Liu, M. (2026). EC2Seq2Sql: Patient-trial matching with LLM agents. PLoS ONE, 21(1), e0341827. https://doi.org/10.1371/journal.pone.0341827

[55] Yang, Y., & Liu, X. (2026). Application of explainable artificial intelligence integrating with electronic health record in oncology. Exploration of targeted anti-tumor therapy, 7, 1002357. https://doi.org/10.37349/etat.2026.1002357

[56] Yingngam, B. (2025). AI in clinical trial design and patient recruitment. In Advances in Computational Intelligence and Robotics Book Series (pp. 2345). IGI Global. https://doi.org/10.4018/979-8-3693-6270-9.ch002

[57] Yousafi, S., Rangachari, P., & Holland, M. L. (2024). Barriers to Recruitment and Retention Among Underrepresented Populations in Cancer Clinical Trials: A Qualitative Study of the Perspectives of Clinical Trial Research Coordinating Staff at a Cancer Center. Journal of healthcare leadership, 16, 427–441. https://doi.org/10.2147/JHL.S488426

[58] Yue, L., Xing, S., Chen, J., & Fu, T. (2024, November). Trialenroll: Predicting clinical trial enrollment success with deep & cross network and large language models. In Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (pp. 1-9).

[59] Zihang, C., Liang, L., Qianmin, S., Gaoyi, C., Jihan, H., & Ying, L. (2025). Enhanced pre-recruitment framework for clinical trial questionnaires through the integration of large language models and knowledge graphs. Scientific reports, 15(1), 27398. https://doi.org/10.1038/s41598-025-11876-0