Journal of Clinical Epidemiology 200 (2026) 112501

COMMENTARY

Artificial intelligence and target trial emulation: toward scalable and credible real-world evidence

Tianyi Liu*, Yanzhong Wang

School of Life Course & Population Sciences, King's College London SE1 1UL, London, UK

Accepted 3 September 2026; Published online 9 September 2026

Abstract

Background and Objectives: Randomized controlled trials remain the gold standard for causal inference but are often resource-intensive, time-consuming, and operationally complex to conduct. The growing availability of real-world data has created new opportunities for generating real-world evidence (RWE), and target trial emulation (TTE) provides a principled framework for designing observational studies that approximate a target trial. At the same time, artificial intelligence (AI) is increasingly applied in health research, yet its application within TTE has received limited systematic attention.

Conclusion: In this Commentary, we examine how AI may support different components of TTE to enable more scalable and credible RWE generation, while considering the methodological risks such applications may introduce. We highlight potential failure mechanisms that may threaten causal validity and the safeguards needed to preserve methodological validity, transparency, and human oversight. We also outline priorities for empirical evaluation and methodological development to support AI-enhanced RWE generation aligned with clinical and regulatory expectations. © 2026 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Keywords: Artificial intelligence; Target trial emulation; Real-world data; Real-world evidence; Causal inference; Observational research

Plain Language Summary

TTE is a way of using routinely collected health-care data to design observational studies that more closely resemble clinical trials. In this Commentary, we discuss how AI could support different components of TTE, from defining study populations to analyzing outcomes. We also highlight important risks, including bias, lack of transparency, and over-reliance on automated methods. AI should support, rather than replace, researchers. With appropriate validation, human oversight, and safeguards, it may help make RWE more efficient, reproducible, and trustworthy.

1. Introduction

Randomized controlled trials (RCTs) are widely regarded as the gold standard for establishing causal relationships in clinical and epidemiological research [1]. However, the conduct of RCTs is often resource-intensive, time-consuming, and operationally complex [2]. In some clinical settings, ethical considerations may also limit the feasibility of conducting concurrent RCTs [3]. The rigorous design required for regulatory acceptance and scientific validity inevitably entails substantial financial investment, extended timelines, and complex logistical coordination, including participant recruitment, clinical site preparation, and training of research staff [4]. Moreover, trial protocols must be carefully predefined, and unexpected amendments during the study may further increase both cost and duration [5].

In parallel, the rapid expansion of real-world data (RWD), including electronic health records, clinical registries, claims data, and administrative health databases,

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

* Corresponding author. School of Life Course & Population Sciences, King's College, Addison House, Guy's Campus, Newcomen Street, London SE1 1UL, UK.

E-mail address: tianyi.5.liu@kcl.ac.uk (T. Liu).

https://doi.org/10.1016/j.jclinepi.2026.112501

0895-4356/© 2026 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

What is new?

Key findings

- Within TTE, AI should be understood as a set of methodological tools that may support specific components rather than a single downstream analytical model.
- The use of AI within TTE may introduce specific causal validity risks, including data-generation distortion, temporal misalignment, estimation bias, representation risks, and failures in transparency and governance.

What this adds to what is known?

- Existing applications of AI in clinical research and RWE generation are often disconnected from the explicit causal design structure of TTE.
- This Commentary links AI applications to the different components of TTE and frames their value around causal validity, transparency, reproducibility, and regulatory credibility.

What is the implication and what should change now?

- Future work should empirically evaluate AI-assisted approaches within TTE against conventional approaches and, where possible, relevant RCT evidence.
- Methodological safeguards, validation procedures, audit trails, and human oversight should be embedded when AI is applied within TTE rather than added after analysis.
- Shared repositories, reusable analytical components, and multidisciplinary governance are needed to support credible AI-supported RWE generation.

has created new opportunities for evidence generation outside highly controlled clinical trials [6]. Although these data sources were not originally collected for research purposes, they offer unprecedented scale and richness [7]. Regulatory agencies have increasingly recognized the value of RWD, issuing guidance and frameworks to support its appropriate use in clinical research, including those from the US Food and Drug Administration (FDA) [8] and the UK Medicines and Healthcare products Regulatory Agency (MHRA) [9]. Real-world evidence (RWE), which is typically derived from RWD, has also been incorporated into decision-making frameworks by health technology assessment (HTA) agencies such as the UK National Institute for Health and Care Excellence [10].

Within this evolving landscape, target trial emulation (TTE) has emerged as a prominent methodological framework for generating causal evidence from observational data, including RWD [11]. Building on the seminal work of Hernán and colleagues [12], TTE requires investigators to make explicit the target trial underlying a causal observational analysis and use that specification to guide the design and analysis of the observational study. This trial-anchored approach makes key causal design decisions explicit while also highlighting where limitations of the available RWD may require refinement of the target trial specification or observational study design. In recent years, the framework has gained increasing attention and methodological refinement. In particular, methodological guidance such as the TARGET reporting guideline [13] has further clarified best practices for TTE specification, implementation, and reporting.

At the same time, artificial intelligence (AI), including traditional machine learning (ML), deep learning (DL) models such as neural networks, and natural language processing (NLP), is increasingly being applied across medicine and health care [14]. Reflecting this growing role, regulatory bodies have begun to articulate more specific expectations for the use of AI in drug development and evidence generation. The FDA has proposed a risk-based framework for assessing AI model credibility according to its intended context of use [15], while the MHRA has emphasized complementary principles of safety, robustness, transparency, fairness, accountability, and appropriate governance across the AI lifecycle [16]. Although challenges remain in terms of interpretability, robustness, generalizability, and governance, the continued integration of AI into clinical research and evidence generation appears inevitable [17]. Against this backdrop, we argue that the current moment presents a timely opportunity to consider how AI, with its natural compatibility with large-scale RWD, could be systematically integrated into the TTE framework. In this Commentary, we discuss how AI may support TTE throughout its iterative process, with the potential to improve the scalability, efficiency, and reliability of RWE generation.

2. AI in clinical research and RWE

When referring to AI in this context, we do not restrict the discussion to any specific model or algorithm. Rather, AI is considered broadly as a collection of computational approaches that may enhance different components of the research process.

AI methods have already been widely applied across clinical research tasks. ML models have been used for clinical risk prediction, risk stratification, and time-to-event (survival) analysis using structured clinical data [18]. DL approaches have also been applied to multimodal biomedical data sources, including medical imaging,

genomic data, and physiological signals [19]. In addition, NLP has increasingly incorporated transformer-based models, including large language models (LLMs) and other generative AI approaches. These methods have been applied to information extraction, clinical coding, and cohort identification from unstructured clinical text [20,21], while emerging LLM-based applications are extending toward broader generative and reasoning-oriented tasks [22,23].

Beyond individual clinical prediction tasks, AI has also been increasingly applied within the broader RWE ecosystem. In observational health data analyses, AI methods have been used for tasks such as phenotyping, patient clustering, and causal modeling [24]. In parallel, within traditional RCT settings, AI approaches have been explored to support recruitment optimization, eligibility screening, and trial monitoring [25].

Despite this rapid growth, these applications have not yet been systematically integrated into the methodological structure of TTE. Although AI-driven components have been incorporated in some applications of TTE, most relevant AI methods have been developed and applied more broadly across clinical and observational research, independently of the explicit design framework that underpins TTE. Nevertheless, many of these methods may be transferable to specific components of TTE when appropriately aligned with its causal and design principles.

Because TTE provides an explicit, trial-anchored structure for causal study design, the potential roles of AI may differ across its interconnected and iterative components, involving distinct methodological requirements, data modalities, and outputs. However, these potential roles have not yet been systematically articulated.

Therefore, a structured and integrated discussion of how AI may support the interconnected components of TTE is needed. Such a framework could help organize existing evidence; distinguish established applications from emerging opportunities; identify methodological gaps; and inform future standards, regulatory guidance, and validation strategies for AI-supported RWE generation.

3. AI across TTE components

The foundational framework for TTE was originally articulated by Hernán and colleagues [12], who described seven core protocol components required to emulate a target randomized trial: eligibility criteria, treatment strategies, assignment procedures, follow-up period, outcomes, causal contrasts of interest, and the analysis plan. Together, explicit specification of these components provides a rigorous conceptual basis for guiding observational study design and analysis according to the target trial.

In practice, implementing TTE requires operationalizing these components within RWD, with target trial specification, data feasibility, study design, and analysis often informing one another iteratively. The present framework therefore considers these components as interconnected, reflecting a general progression from cohort definition and treatment assignment to outcome evaluation and causal estimation, while allowing for refinement across components as the study develops.

Two additional components are incorporated to reflect common practices in contemporary data-driven research and clinical trial methodology. First, TTE typically begins with problem formulation, in which the causal research question is defined and translated into a target trial specification, often informed by frameworks such as population, intervention, comparator, outcome. Second, validation and reproducibility assessments are incorporated to support transparency, methodological robustness, and alignment with emerging reporting standards and regulatory expectations.

Taken together, these components provide a structured basis for iterative observational study design and evidence generation. Table 1 summarizes how AI may support different components of this process.

The representative tasks shown in Table 1 illustrate how existing and emerging AI applications may map onto specific TTE components. Different components involve distinct analytical tasks and may therefore draw on different families of AI methods and types of data. Many of these approaches have already been explored in health and biomedical data research or broader RWE applications, while their use within TTE remains at varying stages of development. Their relevance to TTE therefore lies not in being unique to this framework, but in how they may be applied to specific tasks while remaining aligned with explicit causal design principles. Table 1 brings these applications together within the TTE framework to provide a structured overview of where AI may contribute and the types of methods that may be relevant. The approaches are illustrative rather than prescriptive, and their selection should be guided by the specific task, data characteristics, methodological assumptions, and intended context of use. In some settings, simpler and more interpretable methods may be preferable to more complex models. Their application within TTE therefore requires fit-for-purpose evaluation and appropriate domain and causal expertise, particularly where AI-supported outputs may influence causal validity.

4. Failure modes and causal validity risks

Rather than reiterating the many potential opportunities of AI, which have already been widely discussed across health research [53], we focus here on a more critical issue:

Table 1. Mapping representative AI applications across target trial emulation components

Component (description)	Representative task	AI approach (example models)	Input → output
Problem formulation Define the causal research question and translate it into a target trial specification.	Candidate causal research question formulation from the literature [26,27]	RAG using foundation LLMs (eg, GPT, Claude, Gemini)	Clinical literature → Candidate PICO and causal research question
	Automated PICO extraction from trial reports [28,29]	Biomedical NLP transformers (eg, BioLinkBERT, SciBERT)	Clinical trial free text → Structured PICO fields
	Assessment of target trial feasibility using RWD [30]	NLP-based eligibility modeling (eg, BiLSTM-CRF)	Eligibility criteria and RWD → Feasibility assessment of eligibility criteria
Eligibility criteria Define the characteristics of individuals eligible for inclusion in the study population.	Automated phenotyping of eligible patient populations [31]	Supervised ML phenotyping models (eg, RF, XGBoost, DNN)	Structured EHR data (diagnoses, laboratory results, prescriptions) → Eligibility phenotype labels
	Extraction of eligibility criteria from trial protocols [32,33]	Transformer-based NER models (eg, PubMedBERT, SciBERT, RoBERTa)	Trial protocol free text → Structured eligibility criteria
Treatment strategies Define the interventions or treatment strategies to be compared between study arms.	Identification of longitudinal treatment trajectories [34,35]	Unsupervised trajectory clustering (eg, longitudinal k-means, LSTM-VAE clustering)	Longitudinal medication records → Treatment trajectory clusters
Assignment procedures Define how individuals are assigned to treatment strategies.	Propensity score estimation [36,37]	ML-based propensity score models (eg, RF, boosted trees, DNN)	Baseline covariates → Estimated probability of treatment assignment
	Identification of potential confounders from high-dimensional data [38,39]	ML-based high-dimensional covariate selection (eg, Group Lasso, GBCEE)	High-dimensional EHR covariates → Candidate confounders and proxy variables
Follow-up period Define the start and end of follow-up for outcome assessment.	Reconstruction of patient event timelines [40,41]	LLM-based temporal relation extraction (eg, Llama, Gemma, Mistral)	Clinical reports →Structured patient timelines
Outcomes Define the clinical end points used to evaluate treatment effects.	Automated clinical outcome phenotyping [42]	Supervised ML phenotyping models (eg, RF, SVM, DNN)	Structured EHR data (diagnoses, procedures, laboratory results) → Outcome phenotype labels
	Clinical outcome information extraction [43]	Clinical NLP transformers (eg, BioClinicalBERT, RoBERTa)	Clinical notes → Structured outcome variables
Causal contrasts of interest (estimand) Define the estimand representing the causal effect of interest.	AI-assisted causal question formalization to inform estimand specification [44]	LLM-based causal copilots (eg, GPT)	Research question and contextual knowledge → Candidate causal specification
	Causal structure assessment to support estimand specification [45,46]	Causal discovery algorithms (eg, NOTEARS, DAG-GNN)	Observational variables → Candidate causal graphs
Analysis plan Define the analytical methods used to estimate the causal effect.	Estimation of average and heterogeneous treatment effects [47,48]	Causal ML methods (eg, causal forests, T-/X-learners, Super Learner for TMLE)	Covariates, treatment, and outcomes → Average and heterogeneous treatment-effect estimates
	Survival modeling for time-to-event outcomes [49]	ML-based survival models (eg, random survival forests, DeepSurv)	Covariates, event/censoring times, and event indicators → Predicted hazards or survival probabilities
Validation and reproducibility assessment Assess the robustness, reproducibility, and credibility of the study findings.	AI-assisted reproducibility assessment of analytical pipelines [50]	LLM-based coding and analysis agents (eg, GPT, Claude, Gemini)	Analysis scripts and workflow metadata → Reproducibility assessment

(*Continued*)

Table 1. Continued

Component (description)	Representative task	AI approach (example models)	Input → output
	Cross-database protocol replication and transportability assessment [51]	LLM-assisted study specification and code generation (eg, GPT, Claude, Gemini)	TTE protocol and RWD specifications → Adapted executable analyses
	Benchmarking observational results against RCT evidence [52]	AI-assisted trial retrieval and evidence matching (eg, GPT, MedCPT)	Observational study specifications and RCT literature → Concordance assessment

RCT, randomized controlled trial; AI, artificial intelligence; BioClinicalBERT, BioClinical Bidirectional Encoder Representations from Transformers; BiLSTM-CRF, bidirectional long short-term memory–conditional random field; DAG-GNN, directed acyclic graph neural network; DNN, deep neural network; EHR, electronic health record; GBCEE, generalized Bayesian causal effect estimation; GPT, generative pretrained transformer; LLM, large language model; LSTM-VAE, long short-term memory variational autoencoder; ML, machine learning; NER, named entity recognition; NLP, natural language processing; NOTEARS, Noncombinatorial Optimization via Trace Exponential and Augmented lagRangian for Structure learning; PICO, population, intervention, comparator, outcome; RAG, retrieval-augmented generation; RF, random forest; RWD, real-world data; RWE, real-world evidence; SVM, support vector machine; TMLE, targeted maximum likelihood estimation; TTE, target trial emulation; XGBoost, extreme gradient boosting.

The AI approaches and example models shown are illustrative rather than prescriptive. Many are adapted from broader clinical and RWE applications and may be transferable to analogous tasks within TTE. Where LLMs are considered, they would typically be integrated into task-specific, appropriately grounded pipelines rather than used as standalone general-purpose models. Their suitability depends on the specific task, data characteristics, methodological assumptions, interpretability requirements, intended context of use, and appropriate domain and causal expertise, and should be assessed on a fit-for-purpose basis.

the potential mechanisms through which AI-enabled processes may introduce distortions during TTE.

Because TTE aims to emulate the causal structure of an RCT, errors introduced across its interconnected components may directly compromise causal validity. AI-assisted processes may therefore inadvertently trigger well-recognized methodological biases if they are not carefully designed and monitored [12]. For example, incorrect specification of time zero may lead to immortal time bias [54]. Similarly, errors in cohort construction, treatment or outcome definition, confounder identification or adjustment, or analytical specification may introduce selection bias, measurement bias, or residual confounding [55]. If unrecognized, such mechanisms may distort the underlying data-generating process or causal structure, thereby undermining the credibility of AI-enabled RWE and eroding trust in observational causal inference.

To better illustrate these methodological risks, the major categories of potential causal distortions, referring broadly to errors or biases that may compromise causal validity, and their underlying failure mechanisms within the TTE framework are summarized in Table 2.

At first glance, these potential risks may appear substantial. However, the application of AI within TTE may in fact present fewer fundamental concerns than many other AI applications in health care, particularly predictive modeling. In predictive settings, model outputs often function as final products that may directly influence clinical or policy decisions, while the interpretation of complex models may remain challenging [53].

By contrast, within TTE, AI is primarily used to assist intermediate research tasks, such as protocol specification, cohort identification, or variable extraction, rather than to replace core causal design decisions. Time-zero alignment and estimand specification, in particular, require substantive and methodological judgment. For time zero, eligibility, treatment assignment, and the start of follow-up must be appropriately aligned to preserve causal time ordering. Similarly, estimand specification requires investigators to determine the treatment strategies to be compared, the causal contrast of interest, and how intercurrent events, such as treatment discontinuation, switching, or rescue therapy, should be handled. AI may support these processes by reconstructing temporal information, structuring protocol elements, or supporting causal reasoning, but the resulting design decisions ultimately require scientific, clinical, and methodological judgment. Importantly, AI-supported outputs remain directly inspectable by investigators: protocols, constructed cohorts, and extracted variables can be systematically reviewed and validated before downstream causal analyses are conducted.

In this sense, AI within TTE functions less as an autonomous decision-making system and more as a methodological tool that improves efficiency by supporting labor-intensive procedural steps while maintaining human oversight. The critical methodological question is therefore not whether AI should be used, but how it should be embedded within TTE with appropriate safeguards. With appropriate methodological safeguards and validation procedures implemented across components of the TTE framework, AI-enabled processes may enhance efficiency without undermining causal validity.

5. Future directions for AI-enhanced TTE

The applications of AI discussed above represent only a preliminary set of opportunities that appear particularly promising across different components of TTE. These examples should not be interpreted as an exhaustive catalog

Table 2. Mapping failure mechanisms to causal distortions in AI-enabled target trial emulation

Category of causal distortion	Underlying failure mechanisms	Implications for causal validity	Potential safeguards
Data-generation distortion	Measurement error, exposure or outcome misclassification, informative missingness, data leakage, coding inconsistencies, linkage errors, harmonization or ontology mapping errors, and rare-event instability	Distortion of the observed data-generating process before causal estimation	Data-quality audits, phenotype validation against reference standards, missing-data diagnostics, leakage checks, linkage validation, sensitivity analyses
Temporal-structural misalignment	Incorrect time-zero specification, immortal time bias, look-ahead bias, misalignment between eligibility and treatment assignment, inappropriate exposure episode construction, and mishandling of time-varying confounders	Violation of causal time ordering required for valid treatment contrasts	Prespecified temporal definitions, time-zero alignment checks, exposure-episode validation, assessment of time-varying confounding, emulation review
Estimation and structural bias	Model misspecification, residual or unmeasured confounding, inappropriate adjustment for colliders or mediators, positivity violations, lack of covariate support, and data-driven model selection or hyperparameter tuning bias	Biased or unstable treatment effect estimates	DAG-informed covariate selection, positivity and overlap diagnostics, prespecified analysis plans, alternative model specifications, sensitivity and quantitative bias analyses
AI-system and representation risks	Representation distortion in learned embeddings, generative errors in phenotype or rule construction, nondeterministic model outputs, deployment mismatch between training and application environments, and sensitivity to input formulation or prompts	Silent alteration of AI-derived data representations despite apparent compliance with causal assumptions	Human verification against reference standards, model and prompt version control, repeatability testing, external validation, reproducible pipelines, audit trails
Validation, transparency, and governance failure	Inadequate robustness testing, post hoc optimization, selective reporting, insufficient internal or external validation, governance opacity in analytic pipelines, and lack of benchmarking against reference trials	Reduced credibility, reproducibility, and regulatory confidence	Prespecified validation criteria, independent replication, transparent reporting, provenance and version documentation, benchmarking against reference evidence

AI, artificial intelligence; DAG, directed acyclic graph.

The categories, failure mechanisms, implications, and safeguards presented in Table 2 provide an illustrative overview of potential methodological risks within AI-enabled TTE and should be interpreted in the context of the specific study and AI application.

of potential AI applications. As AI methods continue to evolve, additional capabilities are likely to emerge that may support different components of TTE. The next step is therefore to move beyond isolated examples and more formally conceptualize the coordinated use of AI across different TTE components. This includes systematically mapping current practices, identifying methodological gaps, and evaluating where AI-assisted approaches can provide the greatest benefit while preserving causal validity.

An important priority will be the development of empirical demonstrations of AI-assisted TTE in clinical areas where both RWD and RCT evidence are well established. Domains such as cardiology [56] and nephrology [57] provide natural starting points, as many interventions in these fields have already been evaluated in both observational studies and RCTs. Such case studies could illustrate how AI-assisted components may be integrated across the TTE components, including target trial specification, assessment of feasibility within available RWD, cohort construction, AI-assisted phenotyping and cohort selection, generation of exposure and outcome definitions, and evaluation of study validity and reproducibility. These studies would also enable benchmarking of AI-supported TTE implementations against conventional manually conducted approaches and against corresponding RCT evidence, allowing systematic evaluation of effect estimate consistency, reproducibility, and computational efficiency.

In parallel, methodological work will be needed to develop reusable analytical components that can support different aspects of TTE. These may include tools for target

trial specification and protocol formulation, automated cohort identification from RWD, AI-assisted phenotyping and variable extraction, construction of propensity score models and other confounding adjustment procedures, and consistent implementation of time-zero definitions. Although motivated by the needs of TTE, many of these components may also be applicable across broader clinical research workflows [25].

Another important direction involves the development of shared repositories of studies applying TTE conducted using AI-assisted pipelines. Such repositories could store reusable trial specifications, cohort definitions, exposure-outcome pairs, and analytical pipelines derived from previous studies. Initial entries may be derived from early AI-assisted TTE implementations in cardiometabolic diseases, where interventions such as statins [58], glucagon-like peptide-1 receptor agonists [59], and metformin [60] have been extensively evaluated in both RCTs and RWD studies. Over time, these resources could support methodological transparency, facilitate reproducibility, and accelerate the accumulation of empirical evidence.

Finally, the responsible implementation of AI-assisted TTE will require sustained domain expertise and human oversight throughout the research process. Trial methodologists, statisticians, clinical researchers, data scientists, and software engineers each contribute to the design, implementation, and validation of different components of TTE. Participation from regulatory and HTA stakeholders is also required to ensure that emerging AI-supported approaches remain aligned with established methodological standards and evidence evaluation frameworks [61]. Such multidisciplinary oversight supports the scientific rigor, transparency, and credibility of AI-supported TTE within clinical research and RWE generation.

6. Conclusion

The integration of AI within the principled framework of TTE may offer a promising pathway toward more scalable and credible RWE generation. Realizing this potential requires careful alignment of AI-assisted methods with the interconnected and iterative components of TTE, together with explicit consideration of potential failure modes that may threaten causal validity. By embedding AI within a transparent, expert-led causal design framework and applying appropriate methodological safeguards, AI-supported processes may enhance efficiency while preserving the transparency, reproducibility, and scientific rigor required for reliable causal inference from RWD.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT to support language editing and improve readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

CRediT authorship contribution statement

Tianyi Liu: Writing − original draft, Conceptualization. **Yanzhong Wang:** Writing − review & editing, Supervision, Conceptualization.

Declaration of competing interest

There are no competing interests for any author.

Data availability

No data was used for the research described in the article.

References

[1] Bothwell LE, Greene JA, Podolsky SH, Jones DS. Assessing the Gold Standard–Lessons from the history of RCTs. N Engl J Med 2016;374:2175−81. https://doi.org/10.1056/NEJMms1604593.

[2] Yusuf E, Rosendaal FR. Valid evidence for diagnosis and treatment of infections in the intensive care unit: beyond randomized control trial Study design (trial emulation and machine learning). J Clin Med 2022;11. https://doi.org/10.3390/jcm11133600.

[3] Matthews AA, Danaei G, Islam N, Kurth T. Target trial emulation: applying principles of randomised trials to observational studies. BMJ 2022:e071108. https://doi.org/10.1136/bmj-2022-071108.

[4] Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. Contemp Clin Trials Commun 2018;11:156−64. https://doi.org/10.1016/j.conctc.2018.08.001.

[5] Getz KA, Stergiopoulos S, Short M, Surgeon L, Krauss R, Pretorius S, et al. The impact of protocol amendments on clinical trial performance and cost. Ther Innov Regul Sci 2016;50:436−41. https://doi.org/10.1177/2168479016632271.

[6] Dang A. Real-world evidence: a primer. Pharmaceut Med 2023;37:25−36. https://doi.org/10.1007/s40290-022-00456-6.

[7] Wilson BE, Booth CM. Real-world data: bridging the gap between clinical trials and practice. EClinicalMedicine 2024;78:102915. https://doi.org/10.1016/j.eclinm.2024.102915.

[8] U.S. Food and Drug Administration. Real-world data: assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products. Silver Spring, MD: U.S. Food and Drug Administration; 2024.

[9] Medicines and Healthcare products Regulatory Agency. MHRA guideline on randomised controlled trials using real-world data to support regulatory decisions. London: Medicines and Healthcare products Regulatory Agency (MHRA); 2025.

[10] National Institute for Health and Care Excellence. NICE real-world evidence framework. London: National Institute for Health and Care Excellence (NICE); 2022.

[11] Fu EL. Target trial emulation to improve causal inference from observational data: what, why, and how? J Am Soc Nephrol 2023;34:1305−14. https://doi.org/10.1681/ASN.0000000000000152.

[12] Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. Am J Epidemiol 2016;183:758−64. https://doi.org/10.1093/aje/kwv254.
[13] Cashin AG, Hansford HJ, Hernán MA, Swanson SA, Lee H, Jones MD, et al. Transparent reporting of observational studies emulating a target trial: the TARGET Statement. BMJ 2025. https://doi.org/10.1136/bmj-2025-087179.
[14] Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J 2021;8:e188−94. https://doi.org/10.7861/fhj.2021-0095.
[15] U.S. Food and Drug Administration. Considerations for the use of artificial intelligence to support regulatory decision-making for drug and biological products. Silver Spring, MD: U.S. Food and Drug Administration (FDA); 2025.
[16] Medicines and Healthcare products Regulatory Agency. Impact of AI on the regulation of medical products. London, UK: Medicines and Healthcare products Regulatory Agency (MHRA); 2024.
[17] Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. A review of challenges and opportunities in machine learning for health. AMIA Jt Summits Transl Sci Proc 2020;2020:191−200.
[18] Liu T, Krentz A, Lu L, Curcin V. Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis. Eur Heart J Digital Health 2025;6:7−22. https://doi.org/10.1093/ehjdh/ztae080.
[19] Stahlschmidt SR, Ulfenborg B, Synnergren J. Multimodal deep learning for biomedical data fusion: a review. Brief Bioinform 2022;23. https://doi.org/10.1093/bib/bbab569.
[20] Guan Z, Zeng D, Li H, Wong TY, Sheng B. Can generative artificial intelligence empower target trial emulations? Lancet Digit Health 2026:100950. https://doi.org/10.1016/j.landig.2025.100950.
[21] Yuan K, Yoon CH, Gu Q, Munby H, Walker AS, Zhu T, et al. Transformers and large language models are efficient feature extractors for electronic health record studies. Commun Med 2025;5:83. https://doi.org/10.1038/s43856-025-00790-1.
[22] Li J, Deng Y, Sun Q, Zhu J, Tian Y, Li J, et al. Benchmarking large Language models in evidence-based medicine. IEEE J Biomed Health Inform 2025;29:6143−56. https://doi.org/10.1109/JBHI.2024.3483816.
[23] Hu Y, Zuo X, Zhou Y, Peng X, Huang J, Keloth VK, et al. Information extraction from clinical notes: are we ready to switch to large language models? J Am Med Inform Assoc 2026;33:553−62. https://doi.org/10.1093/jamia/ocaf213.
[24] Ayala Solares JR, Diletta Raimondi FE, Zhu Y, Rahimian F, Canoy D, Tran J, et al. Deep learning for electronic health records: a comparative review of multiple deep neural architectures. J Biomed Inform 2020;101:103337. https://doi.org/10.1016/j.jbi.2019.103337.
[25] Olawade DB, Fidelis SC, Marinze S, Egbon E, Osunmakinde A, Osborne A. Artificial intelligence in clinical trials: a comprehensive review of opportunities, challenges, and future directions. Int J Med Inform 2026;206. https://doi.org/10.1016/j.ijmedinf.2025.106141.
[26] Asai A, He J, Shao R, Shi W, Singh A, Chang JC, et al. Synthesizing scientific literature with retrieval-augmented language models. Nature 2026;650:857−63. https://doi.org/10.1038/s41586-025-10072-4.
[27] Gosak L, Štiglic G, Pruinelli L, Vrbnjak D. PICOT questions and search strategies formulation: a novel approach using artificial intelligence automation. J Nurs Scholarsh 2025;57:5−16. https://doi.org/10.1111/jnu.13036.
[28] Ghosh M, Mukherjee S, Santra P, Na G, Basuchowdhuri P. BLINKtextsubscriptLSTM: BioLinkBERT and LSTM based approach for extraction of PICO frame from clinical trial text. In: Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD). New York, NY, USA: ACM; 2024:227−31. https://doi.org/10.1145/3632410.3632442.
[29] Beltagy I, Lo K, Cohan A. SciBERT: a pretrained Language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA, USA: Association for Computational Linguistics; 2019:3613−8. https://doi.org/10.18653/v1/D19-1371.
[30] Lee K, Liu Z, Mai Y, Jun T, Ma M, Wang T, et al. Optimizing clinical trial eligibility design using natural Language processing models and real-world data: algorithm development and validation. JMIR AI 2024;3:e50800. https://doi.org/10.2196/50800.
[31] Yang S, Varghese P, Stephenson E, Tu K, Gronsbell J. Machine learning approaches for electronic health records phenotyping: a methodical review. J Am Med Inform Assoc 2023;30:367−81. https://doi.org/10.1093/jamia/ocac216.
[32] Li J, Wei Q, Ghiasvand O, Chen M, Lobanov V, Weng C, et al. A comparative study of pre-trained language models for named entity recognition in clinical trial eligibility criteria from multiple corpora. BMC Med Inform Decis Mak 2022;22:235. https://doi.org/10.1186/s12911-022-01967-7.
[33] Tian S, Erdengasileng A, Yang X, Guo Y, Wu Y, Zhang J, et al. Transformer-based named entity recognition for parsing clinical trial eligibility criteria. In: Proceedings of the 12th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. New York, NY, USA: ACM; 2021:1−6. https://doi.org/10.1145/3459930.3469560.
[34] Mullin S, Zola J, Lee R, Hu J, MacKenzie B, Brickman A, et al. Longitudinal K-means approaches to clustering and analyzing EHR opioid use trajectories for clinical subtypes. J Biomed Inform 2021;122:103889. https://doi.org/10.1016/j.jbi.2021.103889.
[35] Qiu J, Hu Y, Li L, Erzurumluoglu AM, Braenne I, Whitehurst C, et al. Deep representation learning for clustering longitudinal survival data from electronic health records. Nat Commun 2025;16:2534. https://doi.org/10.1038/s41467-025-56625-z.
[36] Westreich D, Lessler J, Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. J Clin Epidemiol 2010;63:826−33. https://doi.org/10.1016/j.jclinepi.2009.11.020.
[37] Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. Stat Med 2010;29:337−46. https://doi.org/10.1002/sim.3782.
[38] Benasseur I, Talbot D, Durand M, Holbrook A, Matteau A, Potter BJ, et al. A comparison of confounder selection and adjustment methods for estimating causal effects using large healthcare databases. Pharmacoepidemiol Drug Saf 2022;31:424−33. https://doi.org/10.1002/pds.5403.
[39] Wyss R, Yanover C, El-Hay T, Bennett D, Platt RW, Zullo AR, et al. Machine learning for improving high-dimensional proxy confounder adjustment in healthcare database studies: an overview of the current literature. Pharmacoepidemiol Drug Saf 2022;31:932−43. https://doi.org/10.1002/pds.5500.
[40] Andrew JJ, Potier J, Garcelon N, Burgun A, Vincent M. Using large language models for temporal relation extraction from pediatric clinical reports. JAMIA Open 2025;8. https://doi.org/10.1093/jamiaopen/ooaf121.
[41] Knez T, Žitnik S. Multimodal learning for temporal relation extraction in clinical texts. J Am Med Inform Assoc 2024;31:1380−7. https://doi.org/10.1093/jamia/ocae059.
[42] Gibson TB, Nguyen MD, Burrell T, Yoon F, Wong J, Dharmarajan S, et al. Electronic phenotyping of health outcomes of interest using a linked claims-electronic health record database: findings from a machine learning pilot project. J Am Med Inform Assoc 2021;28:1507−17. https://doi.org/10.1093/jamia/ocab036.
[43] Xie K, Gallagher RS, Conrad EC, Garrick CO, Baldassano SN, Bernabei JM, et al. Extracting seizure frequency from epilepsy clinic notes: a machine reading approach to natural language processing. J Am Med Inform Assoc 2022;29:873−81. https://doi.org/10.1093/jamia/ocac018.
[44] Petersen M, Alaa A, Kıcıman E, Holmes C, van der Laan M. Artificial intelligence−based copilots to generate causal evidence. NEJM AI 2024;1. https://doi.org/10.1056/AIp2400727.

[45] Zheng X, Aragam B, Ravikumar P, Xing EP. DAGs with NO TEARS: continuous optimization for structure learning. In: Advances in Neural Information Processing Systems (NeurIPS), 31. Red Hook, NY: Curran Associates, Inc.; 2018.
[46] Yu Y, Chen J, Gao T, Yu M. DAG-GNN: DAG structure learning with graph neural networks. Proceedings of the 36th International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research97 2019:7154−63.
[47] Abécassis J, Dumas É, Alberge J, Varoquaux G. From prediction to prescription: machine learning and causal inference for the heterogeneous treatment effect. Annu Rev Biomed Data Sci 2025;8:381−404. https://doi.org/10.1146/annurev-biodatasci-103123-095750.
[48] Möller M, Wild E-M, Tan W, Schreyögg J. Estimating heterogeneous treatment effects with real-world health data: a scoping review of machine learning methods. Value Health 2026;29:905−13. https://doi.org/10.1016/j.jval.2026.01.013.
[49] Wiegrebe S, Kopper P, Sonabend R, Bischl B, Bender A. Deep learning for survival analysis: a review. Artif Intell Rev 2024;57:65. https://doi.org/10.1007/s10462-023-10681-3.
[50] Wang Z, Danek B, Yang Z, Chen Z, Sun J. Making large language models reliable data science programming copilots for biomedical research. Nat Biomed Eng 2026. https://doi.org/10.1038/s41551-025-01587-2.
[51] Kim H, Kim M, Kim S, You SC. From study design to executable code: automating target trial emulation with large language models. JAMIA Open 2026;9. https://doi.org/10.1093/jamiaopen/ooag131.
[52] Jin Q, Wang Z, Floudas CS, Chen F, Gong C, Bracken-Clarke D, et al. Matching patients to clinical trials with large language models. Nat Commun 2024;15:9074. https://doi.org/10.1038/s41467-024-53081-z.
[53] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25:44−56. https://doi.org/10.1038/s41591-018-0300-7.
[54] Suissa S. Immortal time bias in pharmacoepidemiology. Am J Epidemiol 2008;167:492−9. https://doi.org/10.1093/aje/kwm324.
[55] Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc 2014;21:221−30. https://doi.org/10.1136/amiajnl-2013-001935.
[56] Ciofani JL, Allahwala UK, Ekmejian A, Pepe S, Bhindi R. The pursuit of truth: revisiting causation versus Association in contemporary cardiovascular research. Heart Lung Circ 2025;34:1150−5. https://doi.org/10.1016/j.hlc.2025.10.005.
[57] Zoccali C, Tripepi G. Clinical trial emulation in nephrology. J Nephrol 2025;38:11−23. https://doi.org/10.1007/s40620-024-02158-5.
[58] Caniglia EC, Rojas-Saunero LP, Hilal S, Licher S, Logan R, Stricker B, et al. Emulating a target trial of statin use and risk of dementia using cohort data. Neurology 2020;95:e1322−32. https://doi.org/10.1212/WNL.0000000000010433.
[59] Xu Y, Huang T, Zhang Y, Ji D, Tuttle KR, Carrero J-J, et al. Risk of heart failure hospitalization for GLP-1 receptor agonists versus DPP-4 inhibitors or SGLT-2 inhibitors in patients with type 2 diabetes: a target trial emulation. Circulation 2026. https://doi.org/10.1161/CIRCULATIONAHA.125.075157.
[60] Dickerman BA, García-Albéniz X, Logan RW, Denaxas S, Hernán MA. Evaluating metformin strategies for cancer prevention: a target trial emulation using electronic health records. Epidemiology 2023;34:690−9. https://doi.org/10.1097/EDE.0000000000001626.
[61] Castanon A, Bray BD, Ramagopalan SV. RWE ready for reimbursement? A round up of developments in real-world evidence relating to health technology assessment: part 15. J Comp Eff Res 2024;13. https://doi.org/10.57264/cer-2024-0033.