AI-Driven Clinical Trial Optimization: Data Infrastructure, Recruitment And Documentation

Junyi Gao[124], Zifeng Wang[3], Jimeng Sun[3], Ewen M Harrison[1]

1 THE UNIVERSITY of EDINBURGH 2 HDRUK Health Data Research UK 3 UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN 4 IQVIA

Trial Panorama

Data Infrastructure

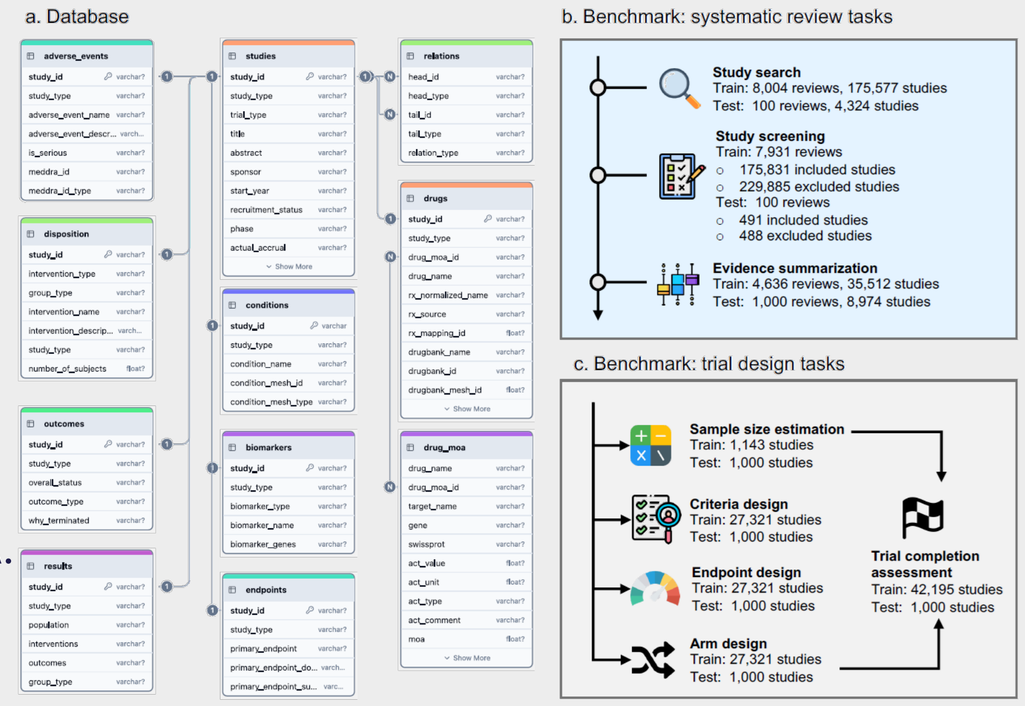
NeurIPS 2025

The Challenge: AI development is hindered by clinical trial data being **fragmented and unstructured** across many global sources.

Solution: A large-scale, structured database unifying **~2 million** trial records from **15 global sources**.

Key Features:

- Integrates trial protocols, conditions, interventions, and outcomes.
- Links to standard ontologies like DrugBank and MedDRA.
- Provides a suite of 8 benchmark tasks for AI development and evaluation.



COMPOSE

Patient-Trial Matching

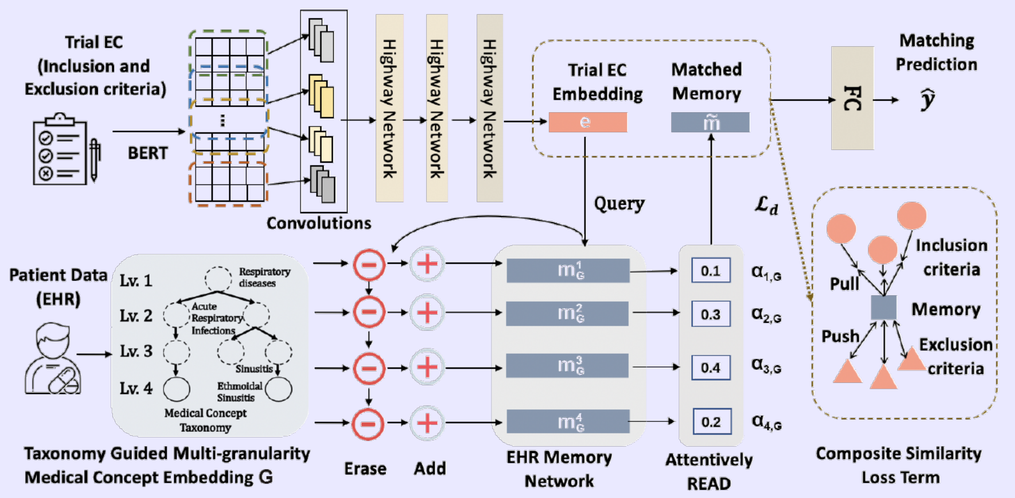
SIGKDD 2020

The Challenge: Patient recruitment is slow and expensive due to the difficulty of matching complex, textual eligibility criteria against structured Electronic Health Records (EHRs).

Solution: A Cross-Modal Pseudo-Siamese Network (COMPOSE) that dynamically matches a patient's longitudinal health record to specific trial criteria.

Key Features:

- Achieved 83.7% patient-trial matching accuracy, a 24.3% relative improvement over baselines on **billion-level real-world dataset**.
- Reached 98.0% AUC on the criteria-level matching task.



DocTr

Trial Site Recommendation

Nature Health

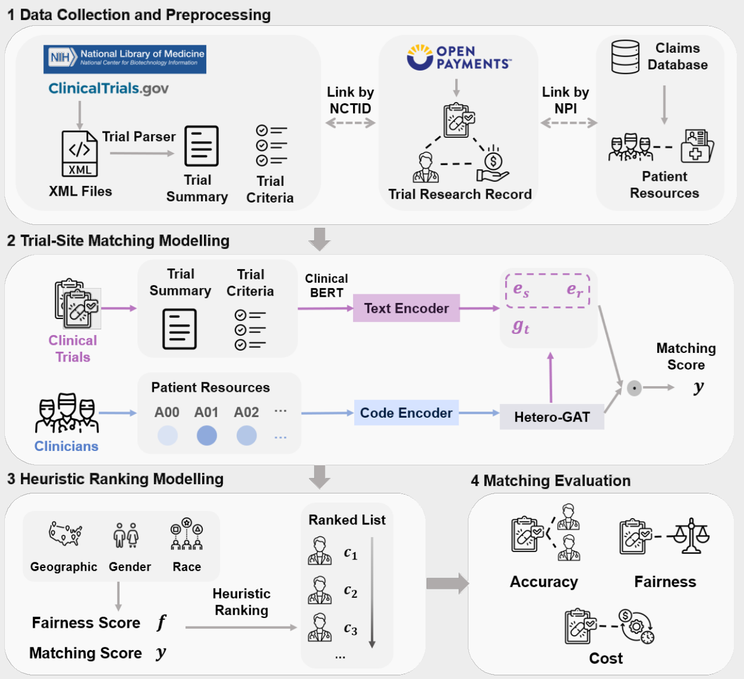
Under Review

The Challenge: Manual **clinician recruitment** is a major bottleneck, with 80% of trials failing to meet enrollment timelines

Solution: A cross-modal deep learning model that matches clinicians using real-world evidence from **trial documents**, **patient encounters**, and **historical enrollment data** from **OpenPayment**.

Key Features:

- **58% higher match similarity** than baselines on **new, unseen trials**.
- **Improved site fairness** scores by up to **25%**.
- Reduced **competing trials** among recommendations to **near-zero**.



InformGen

Patient Documentation

JAMIA

Under Review

The Challenge: Drafting high-stakes patient documents like **Informed Consent Forms (ICFs)** requires extreme factual accuracy and regulatory compliance, making it a resource-intensive task.

Solution: An **LLM-driven "copilot"** for drafting ICFs, combining optimized knowledge parsing from trial protocols with a **human-in-the-loop** framework for verification.

Key Features:

- Achieved near **100% compliance** with 18 core FDA regulatory rules.
- Attained over **90% factual accuracy** when used with human oversight.