TrialGenie: Empowering Clinical Trial Design with Agentic Intelligence and Real World Data

Haoyang Li[1†], Weishen Pan[1†], Suraj Rajendran[2†], Chengxi Zang[1], Fei Wang[1*]

[1]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA
[2]Tri-Institutional Computational Biology & Medicine Program, Cornell University, NY, USA

[†]Equal contribution
*Corresponding author: few2001@med.cornell.edu

Abstract: Clinical trial design (CTD) is a time-consuming process that requires substantial domain expertise. Large-scale real-world data (RWD), such as electronic health records (EHR), encodes practice-based evidence that is of tremendous value to CTD. In recent years, many machine learning methods have been developed to extract such real-world evidence (RWE) from the RWD to inform CTD, but they still need to be communicated with the domain experts extensively in an iterative manner to be further refined and ultimately useful. In this paper, we introduce TrialGenie, an agentic framework that derives RWE for helping with CTD. Through the iterative conversation and analysis across agents with different roles, TrialGenie can autonomously refine trial protocols and finally generate a robust report containing insights that inform better CTD. We applied TrialGenie on the CTD process of several acute diseases including septic shock, acute heart failure, acute pulmonary edema, and acute kidney injury using the MIMIC-IV data. The results demonstrate TrialGenie's capabilities in facilitating and accelerating the CTD process.

Introduction

Randomized controlled trials (RCTs) remain the gold standard for evaluating the efficacy and safety of medical interventions. The time and costs, as well as ethical considerations of conducting a full RCT have led researchers and practitioners to seek approaches to improve the clinical trial design (CTD) process to achieve efficiency and success rate of the corresponding RCT. Real world data (RWD), such as electronic health records (EHRs) and insurance/pharmaceutical claims, contain tremendous practiced based evidence that are insightful for informing CTD. Numerous statistical and machine learning models have been developed in the past decade for extracting such real world evidence (RWE),[1–5] among which target trial emulation (TTE)[6] is a representative framework with the goal of emulating an RCT with RWD. By explicitly mirroring the protocol of an RCT—defining eligibility criteria, specifying treatment assignment strategies, identifying relevant start and end times for follow-up, and selecting appropriate analytic strategies—TTE aims to estimate causal treatment effects and produce results that can closely approximate what might have been derived from an RCT. Several recent research works have demonstrated the great potential of TTE.

Despite the promise, there are several challenges of implementing TTE. First, the elements in the target trial protocol (including eligibility criteria, treatment strategies, and outcome, etc.) need to be matched to the real world data. Usually these elements are described as natural language in the trial protocol, but the EHR data include a large portion of standardized structured information (e.g., encoded with the Observational Medical Outcomes Partnership (OMOP[1,7–9])). Rigorous computable phenotyping process[10] is needed to build such mappings. Second, there is information in the target trial protocol that may not exist in RWD, such as the special biomarkers associated with particular diseases. In this case, we need domain expertise to determine if the information can be dropped or effective surrogate information can be constructed from the RWD. Third, RWD are observational in nature, where the patients are not randomized. This introduces complex selection and confounding biases. Thus, appropriate covariate balancing and causal inference methods are needed to estimate the “true” treatment effects from the RWD.

In addition to replicating RCTs with RWD, TTE also offers an effective tool for extracting evidence from RWD. For example, Zang et al.[9] used it as a hypotheses generation tool for identifying repurposable drug candidates for Alzheimer’s disease. Liu et al.[11] leveraged it to assess the impact of different eligible criteria on real world treatment effectiveness estimation for non-small cell lung cancer trials. Rajendran et al.[12] designed a stratified TTE study for investigating the heterogeneous response for corticosteroid treatment for sepsis patients. In these studies, TTE produced insights even though a target trial may not exist. These insights can help with the design of a real RCT, through extensive conversations with domain experts and iterative adjustments, and this process is still time-consuming.

The rapid development of multi-agent systems (MAS) provides the opportunity for building an autonomous system to derive RWE and inform clinical trial design. MAS are computational frameworks where multiple autonomous entities, or agents, interact and collaborate to achieve a set of goals.[13] Each agent typically has its own specialty, knowledge base, and decision-making logic. Agents are often implemented as combinations of large language models (LLMs) with access to executable tools through APIs, enabling them to carry out diverse and complex tasks. The interactions among the agents have proven effective for coordinating diverse expertise, integrating heterogeneous data sources, as well as enabling parallel and iterative problem-solving.[14–20] The complexity of clinical trial design made MAS a promising approach for improving its efficiency and effectiveness.

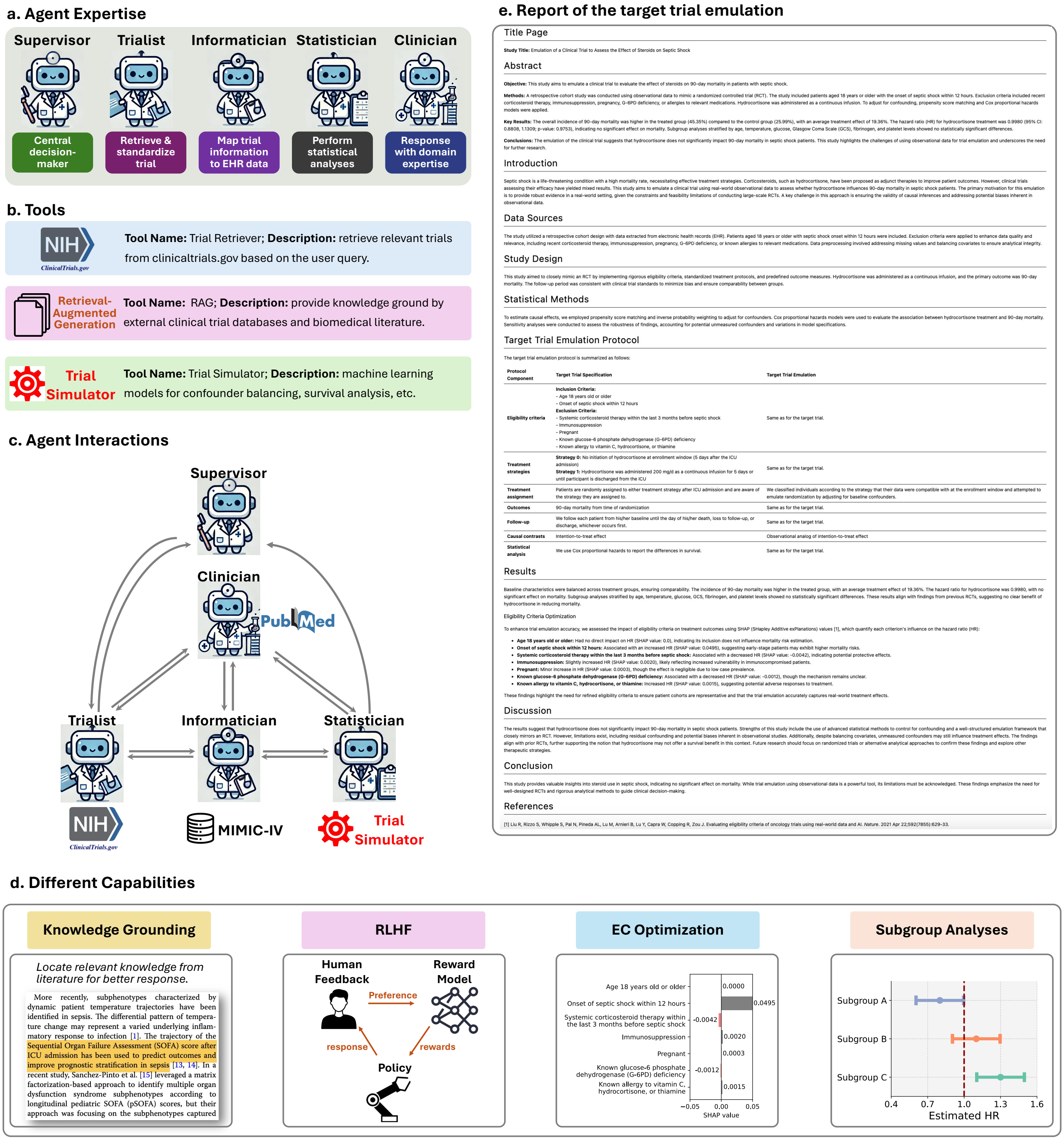
In this paper, we present TrialGenie, a multi-agent framework designed to facilitate and accelerate clinical trial design through the automated extraction and refinement of RWE from EHRs with human in the loop. The agents in TrialGenie collaborate together to finish the following procedures: (1) collect and standardize the information from existing clinical trials and literatures to provide the knowledge for the design of a target clinical trial; (2) generate the protocol of a target clinical trial; (3) map the trial related knowledge to EHR through computable phenotypes and create cohorts; (4) conduct statistical analysis to derive RWE that can inform the clinical trial design; (5) iteratively refine the trial protocol until satisfactory. TrialGenie also leverages Reinforcement Learning from Human Feedback (RLHF[21]) to iteratively improve the quality of the entire MAS with expert feedback. We further demonstrate TrialGenie's capability through case studies on trial design for acute conditions in the ICU setting with the MIMIC-IV[22] data set.

TrialGenie Architecture

Overview

TrialGenie is a modular multi-agent framework designed to support efficient, automated, and expert-aligned clinical trial design. It comprises five specialized agents—Supervisor, Trialist, Informatician, Clinician, and Statistician—each powered by large language models (LLMs) and equipped with distinct domain-specific capabilities (Figure 1a). These agents collaborate through structured conversations to perform key steps in the trial design workflow, including protocol specification, data extraction, covariate selection, statistical modeling, and interpretation. The architecture not only includes a core sequential pipeline, i.e., Supervisor → Trialist → Informatician → Clinician → Statistician → Supervisor (as shown in Figure 1c), which reflects the natural progression from initial trial specification to analysis and reporting, but also extends beyond this traditional sequential execution by supporting dynamic interactions among agents. For instance, the Informatician can consult the Clinician when facing data sparsity or missing covariates, prompting iterative refinements in eligibility criteria or variable/outcome substitutions. This flexible agent communication strategy allows the system to adapt to appropriately accommodate the challenges in real world data. Note that TrialGenie does not require manual specification of the workflow, allowing agents to dynamically respond to one another's outputs and adapt to new constraints or evolving objectives. For instance, if the Informatician identifies high levels of missingness for a key variable or detects poor covariate balance, the system can autonomously initiate a feedback loop with the Clinician to assess alternative variable definitions or biomedically appropriate surrogates. In addition, TrialGenie integrates tool-augmented reasoning capabilities of each agent (Figure 1b). The Trial Retriever helps the Trialist identify relevant protocols from clinical trial registries; the RAG module empowers the Clinician to ground its decisions in biomedical literature by performing semantic searches over large corpora such as PubMed; and the Trial Simulator provides the Statistician with access to statistical and machine learning libraries for confounder adjustment, outcome analysis, and treatment effect estimation. These tools are invoked automatically within the agent workflow to convert natural language insights into executable code, structured queries, and interpretable analytics. Beyond these core utilities, TrialGenie supports advanced reasoning functions (Figure 1d), including knowledge grounding from literature to improve factual accuracy, RLHF to align agent outputs with expert preferences, eligibility criteria (EC) optimization using Shapley-based attribution methods to quantify the influence of inclusion rules on outcomes, and subgroup analyses to uncover heterogeneous treatment effects that may be masked in the aggregate population. These capabilities are modular yet synergistic, enabling TrialGenie to continuously refine trial protocols with both methodologically robust and clinical interpretability.

The final output of this multi-agent system is a comprehensive trial design report (Figure 1e), which synthesizes contributions from all agents into a unified document. This report includes standardized sections—such as abstract, introduction, methods, protocol specifications, results, and discussion—and is enriched with protocol tables, statistical summaries, and visualizations (e.g., hazard ratios with confidence intervals, covariate balance diagnostics). By automating the generation of such high-quality outputs, TrialGenie not only accelerates the design cycle but also ensures transparency and reproducibility. Together, these components reflect TrialGenie’s capacity to transform the traditionally manual, expert-driven process of clinical trial design into an efficient, intelligent, and collaborative workflow.



Tool Name: Trial Retriever; **Description:** retrieve relevant trials from clinicaltrials.gov based on the user query.

Tool Name: RAG; **Description:** provide knowledge ground by external clinical trial databases and biomedical literature.

Tool Name: Trial Simulator; **Description:** machine learning models for confounder balancing, survival analysis, etc.

Locate relevant knowledge from literature for better response.

More recently, subphenotypes characterized by dynamic patient temperature trajectories have been identified in sepsis. The differential pattern of temperature change may represent a varied underlying inflammatory response to infection [1]. The trajectory of the Sequential Organ Failure Assessment (SOFA) score after ICU admission has been used to predict outcomes and improve prognostic stratification in sepsis [13, 14]. In a recent study, Sanchez-Pinto et al. [15] leveraged a matrix factorization-based approach to identify multiple organ dysfunction syndrome subphenotypes according to longitudinal pediatric SOFA (pSOFA) scores, but their approach was focusing on the subphenotypes captured

Title Page

Study Title: Emulation of a Clinical Trial to Assess the Effect of Steroids on Septic Shock

Abstract

Objective: This study aims to emulate a clinical trial to evaluate the effect of steroids on 90-day mortality in patients with septic shock.

Methods: A retrospective cohort study was conducted using observational data to mimic a randomized controlled trial (RCT). The study included patients aged 18 years or older with the onset of septic shock within 12 hours. Exclusion criteria included recent corticosteroid therapy, immunosuppression, pregnancy, G-6PD deficiency, or allergies to relevant medications. Hydrocortisone was administered as a continuous infusion. To adjust for confounding, propensity score matching and Cox proportional hazards models were applied.

Key Results: The overall incidence of 90-day mortality was higher in the treated group (45.35%) compared to the control group (25.99%), with an average treatment effect of 19.36%. The hazard ratio (HR) for hydrocortisone treatment was 0.9980 (95% CI: 0.8808, 1.1309; p-value: 0.9753), indicating no significant effect on mortality. Subgroup analyses stratified by age, temperature, glucose, Glasgow Coma Scale (GCS), fibrinogen, and platelet levels showed no statistically significant differences.

Conclusions: The emulation of the clinical trial suggests that hydrocortisone does not significantly impact 90-day mortality in septic shock patients. This study highlights the challenges of using observational data for trial emulation and underscores the need for further research.

Introduction

Septic shock is a life-threatening condition with a high mortality rate, necessitating effective treatment strategies. Corticosteroids, such as hydrocortisone, have been proposed as adjunct therapies to improve patient outcomes. However, clinical trials assessing their efficacy have yielded mixed results. This study aims to emulate a clinical trial using real-world observational data to assess whether hydrocortisone influences 90-day mortality in septic shock patients. The primary motivation for this emulation is to provide robust evidence in a real-world setting, given the constraints and feasibility limitations of conducting large-scale RCTs. A key challenge in this approach is ensuring the validity of causal inferences and addressing potential biases inherent in observational data.

Data Sources

The study utilized a retrospective cohort design with data extracted from electronic health records (EHR). Patients aged 18 years or older with septic shock onset within 12 hours were included. Exclusion criteria were applied to enhance data quality and relevance, including recent corticosteroid therapy, immunosuppression, pregnancy, G-6PD deficiency, or known allergies to relevant medications. Data preprocessing involved addressing missing values and balancing covariates to ensure analytical integrity.

Study Design

This study aimed to closely mimic an RCT by implementing rigorous eligibility criteria, standardized treatment protocols, and predefined outcome measures. Hydrocortisone was administered as a continuous infusion, and the primary outcome was 90-day mortality. The follow-up period was consistent with clinical trial standards to minimize bias and ensure comparability between groups.

Statistical Methods

To estimate causal effects, we employed propensity score matching and inverse probability weighting to adjust for confounders. Cox proportional hazards models were used to evaluate the association between hydrocortisone treatment and 90-day mortality. Sensitivity analyses were conducted to assess the robustness of findings, accounting for potential unmeasured confounders and variations in model specifications.

Target Trial Emulation Protocol

The target trial emulation protocol is summarized as follows:

Protocol Component	Target Trial Specification	Target Trial Emulation
Eligibility criteria	**Inclusion Criteria:** - Age 18 years old or older - Onset of septic shock within 12 hours **Exclusion Criteria:** - Systemic corticosteroid therapy within the last 3 months before septic shock - Immunosuppression - Pregnant - Known glucose-6 phosphate dehydrogenase (G-6PD) deficiency - Known allergy to vitamin C, hydrocortisone, or thiamine	Same as for the target trial.
Treatment strategies	**Strategy 0:** No initiation of hydrocortisone at enrollment window (5 days after the ICU admission) **Strategy 1:** Hydrocortisone was administered 200 mg/d as a continuous infusion for 5 days or until participant is discharged from the ICU	Same as for the target trial.
Treatment assignment	Patients are randomly assigned to either treatment strategy after ICU admission and are aware of the strategy they are assigned to.	We classified individuals according to the strategy that their data were compatible with at the enrollment window and attempted to emulate randomization by adjusting for baseline confounders.
Outcomes	90-day mortality from time of randomization	Same as for the target trial.
Follow-up	We follow each patient from his/her baseline until the day of his/her death, loss to follow-up, or discharge, whichever occurs first.	Same as for the target trial.
Causal contrasts	Intention-to-treat effect	Observational analog of intention-to-treat effect
Statistical analysis	We use Cox proportional hazards to report the differences in survival.	Same as for the target trial.

Results

Baseline characteristics were balanced across treatment groups, ensuring comparability. The incidence of 90-day mortality was higher in the treated group, with an average treatment effect of 19.36%. The hazard ratio for hydrocortisone was 0.9980, with no significant effect on mortality. Subgroup analyses stratified by age, temperature, glucose, GCS, fibrinogen, and platelet levels showed no statistically significant differences. These results align with findings from previous RCTs, suggesting no clear benefit of hydrocortisone in reducing mortality.

Eligibility Criteria Optimization

To enhance trial emulation accuracy, we assessed the impact of eligibility criteria on treatment outcomes using SHAP (SHapley Additive exPlanations) values [1], which quantify each criterion's influence on the hazard ratio (HR):

- **Age 18 years old or older:** Had no direct impact on HR (SHAP value: 0.0), indicating its inclusion does not influence mortality risk estimation.
- **Onset of septic shock within 12 hours:** Associated with an increased HR (SHAP value: 0.0495), suggesting early-stage patients may exhibit higher mortality risks.
- **Systemic corticosteroid therapy within the last 3 months before septic shock:** Associated with a decreased HR (SHAP value: -0.0042), indicating potential protective effects.
- **Immunosuppression:** Slightly increased HR (SHAP value: 0.0020), likely reflecting increased vulnerability in immunocompromised patients.
- **Pregnant:** Minor increase in HR (SHAP value: 0.0003), though the effect is negligible due to low case prevalence.
- **Known glucose-6 phosphate dehydrogenase (G-6PD) deficiency:** Associated with a decreased HR (SHAP value: -0.0012), though the mechanism remains unclear.
- **Known allergy to vitamin C, hydrocortisone, or thiamine:** Increased HR (SHAP value: 0.0015), suggesting potential adverse responses to treatment.

These findings highlight the need for refined eligibility criteria to ensure patient cohorts are representative and that the trial emulation accurately captures real-world treatment effects.

Discussion

The results suggest that hydrocortisone does not significantly impact 90-day mortality in septic shock patients. Strengths of this study include the use of advanced statistical methods to control for confounding and a well-structured emulation framework that closely mirrors an RCT. However, limitations exist, including residual confounding and potential biases inherent in observational studies. Additionally, despite balancing covariates, unmeasured confounders may still influence treatment effects. The findings align with prior RCTs, further supporting the notion that hydrocortisone may not offer a survival benefit in this context. Future research should focus on randomized trials or alternative analytical approaches to confirm these findings and explore other therapeutic strategies.

Conclusion

This study provides valuable insights into steroid use in septic shock, indicating no significant effect on mortality. While trial emulation using observational data is a powerful tool, its limitations must be acknowledged. These findings emphasize the need for well-designed RCTs and rigorous analytical methods to guide clinical decision-making.

References

[1] Liu R, Rizzo S, Whipple S, Pal N, Pineda AL, Lu M, Arnieri B, Lu Y, Capra W, Copping R, Zou J. Evaluating eligibility criteria of oncology trials using real-world data and AI. *Nature*. 2021 Apr 22;592(7855):629-33.

Figure 1. **(a)** Role of five agent in TrialGenie. **(b)** Tools used in TrialGenie. **(c)** Agent interactions. **(d)** Several Capabilities supported in TrialGenie. **(e)** Generated final report, which mainly includes the trial emulation protocol and key results of the septic shock case.

Agents

Each agent in TrialGenie is designed to perform a specific set of tasks as summarized below, contributing to the overall efficiency and accuracy of the process to inform clinical trial design.

- *Supervisor*: This agent acts as the central decision-maker, synthesizing inputs from the human user and determining the start or end of the workflow.
- *Trialist*: This agent retrieves and standardizes trial information from ClinicalTrials.gov and PubMed and then derives protocols of clinical trials to be simulated.
- *Informatician*: The Informatician bridges trial information with real-world data, generating dataset ready for analysis. This agent matches trial eligibility criteria to EHR data, performs quality assurance, and finally constructs datasets for statistical analysis.
- *Clinician*: The Clinician provides domain expertise by analyzing literature and answering clinical questions. This agent is responsible to identify related covariates and outcomes, validate trial design and interpret statistical analysis results.
- *Statistician*: The Statistician conducts trial emulations and statistical analyses. This agent selects statistical methods, conducts outcome analyses, and generates results.

TrialGenie for Clinical Trial Design

Parse and Summarize Trial Information

Given the user's interest, it is critical to search for relevant information from historical trials, which is the main job of the trialist agent. It extracts, annotates, and standardizes the relevant information including eligibility criteria, treatment strategies, and outcomes from related trials. Figure 2 illustrates the main functionalities of the Trialist.

First, the Trialist uses an LLM concept extraction prompt, designed and refined based on the Criteria2Query3.0[23] framework, to extract key components of the relevant trials, including eligibility criteria, treatment, and outcomes. Relevant clinical concepts are identified and annotated into domains such as Demographics, Condition, Device, Procedure, Drug, Measurement, Observation, and Visit, and their associated values along with the temporal information are extracted as well. We added additional instructions to the prompt for handling the components with multiple concepts (e.g. ""Allergy to vitamin C, hydrocortisone, or thiamine"") or omitted concepts (e.g. "Patients < 18 years" with concept age omitted in the text). The original prompt is from Criteria2Query3.0[23] and our modified version is provided in Supplementary Table 1.

Since identical concepts might be expressed differently across trials (e.g., "ICU" vs. "Intensive Care Unit"), the Unified Medical Language System (UMLS)[24] dictionary and Observational Health Data Sciences and Informatics (OHDSI[25]) APIs are used to standardize concepts. Then, each component is matched with design patterns defined in the existing clinical trial ontologies.[26,27] The mathematical and temporal operators (e.g., "less than" from "less than 24 hours"), as well as the related number and units, are identified by the temporal and value normalization modules from the Criteria2Query3.0 framework[23].

All the concepts involved in the target trials are mapped to their concept ids via the standardized vocabulary in OMOP common data model. Specially, the concepts were represented based on the standard of OMOP CDM: condition (ICD-9/ICD-10), drug (RxNorm), measurement (LOINC), procedure (SNOMED CT). This provided the necessary information for the informatician to generate the SQL queries for retrieving relevant information from the EHR data.

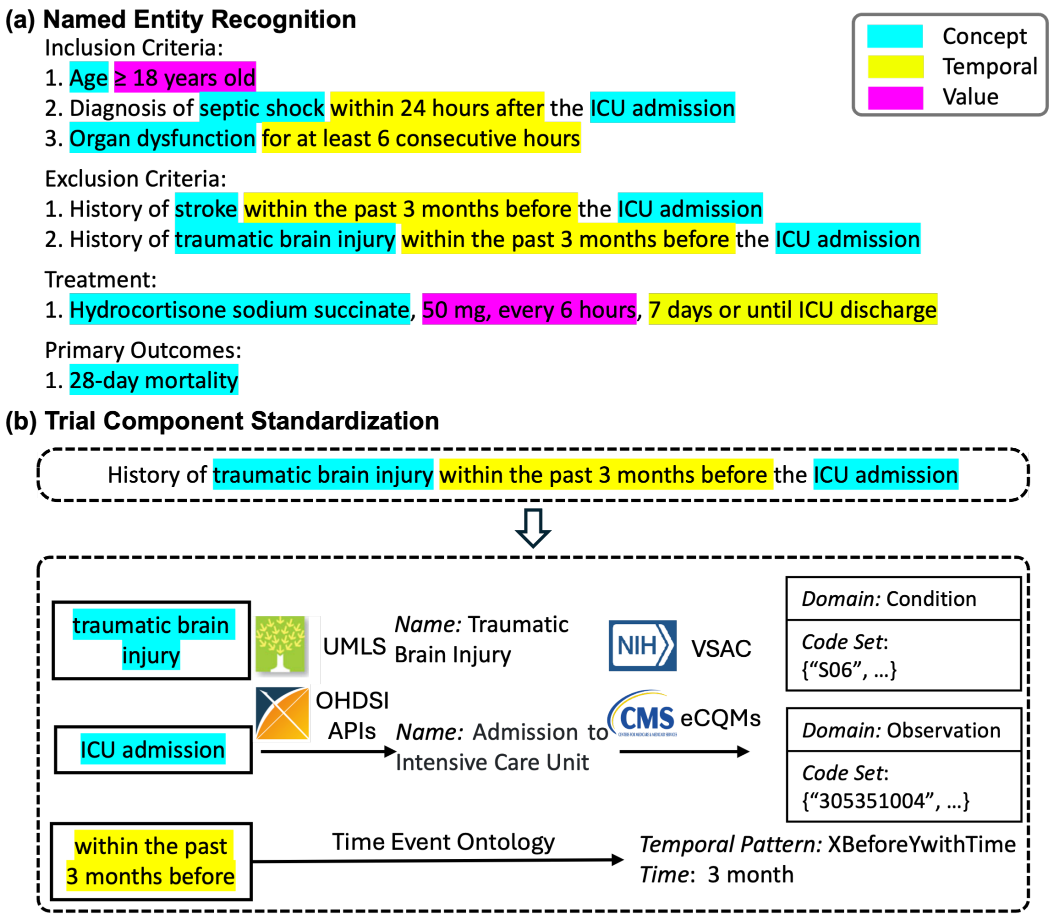

(a) Named Entity Recognition
Concept
Temporal
Value
Inclusion Criteria:
1. Age ≥ 18 years old
2. Diagnosis of septic shock within 24 hours after the ICU admission
3. Organ dysfunction for at least 6 consecutive hours
Exclusion Criteria:
1. History of stroke within the past 3 months before the ICU admission
2. History of traumatic brain injury within the past 3 months before the ICU admission
Treatment:
1. Hydrocortisone sodium succinate, 50 mg, every 6 hours, 7 days or until ICU discharge
Primary Outcomes:
1. 28-day mortality
(b) Trial Component Standardization
History of traumatic brain injury within the past 3 months before the ICU admission
traumatic brain injury
UMLS
Name: Traumatic Brain Injury
NIH
VSAC
Domain: Condition
Code Set: {"S06", ...}
OHDSI APIs
ICU admission
Name: Admission to Intensive Care Unit
CMS eCQMs
Domain: Observation
Code Set: {"305351004", ...}
within the past 3 months before
Time Event Ontology
Temporal Pattern: XBeforeYwithTime
Time: 3 month

Figure 2. Functionalities of Trialist.

Build Dataframe for Analysis

Once the trial related information has been retrieved and standardized, the Informatician agent constructs a high-quality analytical dataframe that operationalizes the trial specifications using real-world EHR data. The Informatician translates these trial related information, including eligibility criteria, treatment assignments, and outcome measures, into executable SQL queries, inspired by Criteria2Query3.0.[23] Each criterion is rendered as a Common Table Expression (CTE),[28] allowing for modular and sequential cohort construction. Simple eligibility rules—such as "Age ≥ 18 years"—are directly translated into SQL filters, while more complex conditions, such as time-sensitive interventions or compound dosage rules, are handled through nested logic and temporal joins.[29,23,30,31] This structured querying ensures that the resulting cohort faithfully represents the target population defined by the trialist agent.

To collect necessary covariates for downstream analysis, the Informatician collaborates with the Clinician agent, who provides advice on clinically relevant variables based on the disease context and trial objectives. These typically include demographic features, laboratory values, vital signs, diagnosis and medications. The selected covariates are then mapped to OMOP-compliant fields and integrated into the dataframe.

With the cohort and covariates defined, the Informatician can extract data from EHR warehouse to build a comprehensive, analytics-ready dataframe. Treatment data includes drug administration records, dosage schedules, and timing; outcome data focuses on time-to-event endpoints such as

28-day mortality; and covariate data spans the pre-intervention period, capturing relevant clinical measurements. All variables are aligned temporally and semantically with the trial protocol.

Data quality assurance is a critical part of this pipeline. The Informatician performs rigorous checks for completeness, logical consistency, and clinical plausibility.[32–34] For example, outcome durations are validated to ensure non-negativity, and missingness is assessed across key variables. In cases of high missing rates, the agent may consult the Clinician to identify validated surrogate markers such as base excess. Outliers are addressed adaptively through imputation or exclusion strategies,[35,36] depending on clinical context. The finalized output is a clean, structured dataframe containing patient-level rows with identifiers, eligibility flags, treatment indicators, outcome metrics, and baseline covariates—fully aligned with the elements needed for a trial protocol and ready for causal inference and statistical analysis.

Statistical Analysis

The Statistician agent in TrialGenie is responsible for translating the study protocol into a reproducible analytic workflow, applying appropriate statistical analysis, especially causal inference techniques, and summarizing the results. This agent's methodology comprises five main components: (1) selecting balancing and modeling methods, (2) performing covariate balancing, (3) conducting survival analyses, and (4) generating a final report of the findings.

First, the Statistician selects the best covariate balancing strategy and outcome analysis method. In TrialGenie, the agent evaluates across multiple options for balancing including Propensity Score Matching (PSM),[37–39] Inverse Probability of Treatment Weighting (IPTW),[40–42] or no balancing. The Statistician bases this selection on factors such as sample size, the distribution of covariates, and the research objective of estimating causal effects, which are determined by analyzing the input from the Informatician. If the dataset is moderately sized and sufficiently rich in covariates, PSM is often chosen and preferred for its intuitive design and interpretability.[43] For each outcome analysis step, the Statistician can select from Cox Proportional Hazards,[44–46] Kaplan-Meier estimation,[47] parametric survival models, random survival forests (RSFs),[48] or doubly robust methods.[49] The final choice depends on checks of proportional hazards assumptions, the nature of the endpoints, and the need to balance interpretability (e.g., hazard ratio estimates) with predictive performance (e.g., random survival forests for high-dimensional data).

Beyond the main analytical workflow of building trial protocols, balancing covariates, and conducting survival analyses, the Statistician agent in TrialGenie is able to deepen the investigation of treatment effects and refine the design of the emulated trial. Specifically, the Statistician can perform subgroup analysis and EC optimization to explore treatment heterogeneity and systematically evaluate the influence of different conditions on overall results.

Following an initial survival analysis (for instance, via Cox Proportional Hazards), the Statistician agent may detect that the estimated treatment effect is not statistically significant in the overall sample. In such a scenario, the Statistician can perform subgroup analyses, wherein it prompts the Clinician to propose a covariate and threshold for splitting the cohort into two clinically relevant subgroups (e.g., patients with a severity marker below or above a certain level). The Statistician then reruns the survival model for each subgroup, comparing the hazard ratios and confidence intervals separately. This helps identify any sub-populations where the treatment might be more (or less) effective—an important consideration in critical care settings, where interventions can yield heterogeneous responses. If no subgroup exhibits a significant effect or if multiple subgroup splits fail to uncover meaningful patterns, the process stops to avoid excessive data-driven exploration.

Trial emulation also often requires iterative adjustments to eligibility criteria to balance sample size, comparability with the original trial, and clinical relevance. To systematically evaluate which inclusion or exclusion conditions exert the greatest impact on the estimated hazard ratio, the Statistician implements Trial Pathfinder.[50] First, the agent enumerates different combinations (or "subsets") of eligibility rules, computing a hazard ratio for each subset. Next, it treats the presence or absence of each criterion within these subsets as a contribution game, where Shapley values quantify how adding a particular criterion changes the measured treatment effect. If a criterion consistently shifts the hazard ratio toward or away from a significant result, it receives a larger absolute Shapley value, indicating a strong influence on outcomes. This is then relayed back to the Clinician or Supervisor, who can decide whether to retain, modify, or remove certain criteria considering both statistical impact and clinical imperatives. By systematically exploring all—or a carefully selected range of—criterion combinations, the Statistician helps ensure that trial emulation decisions are informed by quantitative evidence as well as domain expertise.

Finally, the Statistician synthesizes all findings into a cohesive report. This write-up includes baseline summaries of the matched or weighted populations, adjusted hazard ratios (or other metrics such as risk differences, depending on the selected model), confidence intervals, and p-values. Where relevant, the Statistician compares the emulated trial estimates to published results from the original randomized trial, highlighting potential sources of discrepancy such as sample size limitations, subtle differences in inclusion criteria (due to the limitations of EHRs), or unresolved confounding. This report is then sent to the Clinician agent for further interpretation, ensuring that any remaining clinical or methodological concerns are addressed before finalizing conclusions on the treatment's effectiveness.

Request Clinician Suggestions

The Clinician agent in TrialGenie is designed to incorporate domain-specific medical expertise into the trial emulation workflow, ensuring that decisions about eligibility criteria, covariates, and study design remain clinically valid. Its methodology is structured around three core functions: retrieving and synthesizing medical literature, communicating clinical insights in a structured format, and collaborating iteratively with other agents to refine trial design and analysis.

The Clinician agent operates through two primary tasks: (1) reviewing reports generated by the Statistician agent and either recommending modifications or approving the analysis, and (2) providing evidence-based recommendations to other agents at various stages of the trial emulation process. These tasks are facilitated by a retrieval-augmented generation (RAG)[51] approach to gather relevant medical literature. Upon receiving a query—such as a request to confirm clinical plausibility or identify alternative covariates—the agent performs semantic searches over a biomedical knowledge base. Currently, this knowledge base comprises full-text PDFs or extracted abstracts stored in a FAISS index.[52] By leveraging embeddings generated by a sentence-transformer, the Clinician agent identifies the most pertinent sections of the literature, such as published guidelines on sepsis management or prior studies on corticosteroids, and synthesizes the retrieved passages into concise, evidence-based responses.

To ensure interoperability with other agents, the Clinician agent delivers its responses in standardized, machine-readable formats. For instance, when specifying eligibility criteria, it produces annotations with tags such as <Condition>, <Drug>, or <Measurement> to ensure consistent handling of clinical concepts. If recommending a relaxation of the time window for septic shock diagnosis, the agent might return an output explicitly identifying the condition domain (<Condition>septic shock</Condition>) and the temporal modifier (<Temporal>within 48 hours after</Temporal>). This structured communication allows other agents, particularly the Informatician, to directly map the recommendations onto EHR queries without requiring manual interpretation.

The Clinician agent collaborates iteratively with other agents throughout the trial emulation process. For example, when the dataset assembled by the Informatician is too small or when covariate balance is deemed insufficient by the Statistician, the Clinician proposes clinically acceptable modifications, such as relaxing exclusion criteria or substituting missing measurements with clinically equivalent variables. Similarly, for the task of report reviewing, if the Statistician identifies unexpected findings, the Clinician reviews surveys medical literature to determine whether the results align with established knowledge, propose additional adjustments, or investigate potential explanations for the discrepancies.

Optimization

RLHF[53] is deployed in TrialGenie to iteratively improve the quality of the entire MAS with expert feedback. Its workflow includes three steps: (1) human feedback collection, (2) preference modeling, and (3) policy optimization with PPO[54] and DPO.[55]

Human Feedback Collection. The outputs generated by each agent are first reviewed by human experts. Then the experts provide either: *Ratings* (e.g., 1–5) based on task-specific criteria such as correctness, clinical validity, or alignment with research goals; or *Rankings*, in which they compare multiple candidate outputs and indicate preference order.

Preference Modeling. Within the dataset aggregated from the human feedback, rating data is used to train reward models that predict scalar-valued quality scores $r(x)$ for outputs x and ranking data is converted into pairwise preferences (x^+, x^-), indicating that output x^+ is preferred over x^- for a given task.

Policy Optimization with PPO and DPO. TrialGenie fine-tunes the policies of its LLMs using the following methods: (1) Proximal Policy Optimization (PPO[54]): PPO is employed when expert feedback provides scalar ratings, especially in tasks with well-defined correctness signals (e.g., SQL validity). Then the policy $\pi_\theta(x)$ is optimized to maximize the expected reward:

$$\max_\theta \mathbb{E}_{x\sim\pi_\theta}[r(x)].$$

To ensure stability and prevent policy collapse, PPO uses a clipped surrogate objective:

$$L^{PPO}(\theta) = \mathbb{E}_t[\min(r_t(\theta)\hat{A}_t, \mathrm{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)],$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\mathrm{old}}(a_t|s_t)}$ is the importance sampling ratio, $\hat{A}_t$ is the advantage estimate, often approximated as $r(x_t) - b$ (b is a baseline). (2) Direct Preference Optimization (DPO[55]): DPO is employed when feedback is provided in the form of pairwise comparisons. Rather than estimating a separate reward model, DPO directly optimizes the policy using the relative likelihoods of preferred vs. non-preferred samples. The objective is to maximize:

$$L^{DPO} = \sum_{(x^+,x^-)} \log\, \sigma\big(\beta(\log \pi(x^+) - \log \pi(x^-))\big),$$

where $\pi(x)$ is the policy's probability of output x, β is a temperature parameter controlling preference sharpness, and $\sigma(\cdot)$ is the sigmoid function.

Meetings

In TrialGenie, meetings are designed as collaborative sessions that enable agents to share expertise, refine specific tasks, and ensure high-quality output, inspired by Swanson et al.[56] Depending on the specific task, meetings are categorized into two types, including team meetings and individual meetings.

Team Meetings

Team meetings in TrialGenie bring together all the agents to address complex, high-level issues requiring interdisciplinary expertise. These meetings are organized by the Supervisor, who sets the meeting agenda and synthesizes inputs. The discussions typically revolve around broad questions, such as optimizing trial eligibility criteria or selecting outcome analysis methods. For instance, consider a meeting focused on addressing high missing rates in lactate levels within the dataset. The Supervisor initiates the discussion by outlining the agenda: assessing whether surrogate variables like base excess can replace lactate levels. Each agent contributes based on their expertise. The Informatician presents data on the extent of missingness and the feasibility of implementing surrogate measures. The Clinician evaluates the clinical validity of base excess as a substitute, referencing medical literature from PubMed or other relevant sources. The Statistician weighs the statistical implications, particularly the impact on covariate balancing. The discussion unfolds over multiple rounds, with agents refining their responses based on feedback. The Supervisor consolidates the insights, approves the use of base excess, and assigns follow-up tasks to the Informatician for implementation. These meetings are essential for resolving ambiguities and achieving consensus on critical decisions. By organizing open discussions among the agents, team meetings ensure that TrialGenie leverages diverse expertise to navigate the inherent complexities of clinical trial design.

Individual Meetings

Individual meetings focus on task-specific activities, typically assigned to a single agent, with optional feedback from other agents or the Supervisor. These meetings allow for in-depth execution and refinement of specialized tasks, such as coding SQL queries or running outcome analysis models. An example of an individual meeting could involve the Informatician tasked with generating a dataset based on the eligibility criteria. The agenda specifies the need to construct SQL queries that incorporate relaxed temporal conditions for septic shock diagnosis. The Informatician writes the initial SQL queries, reviews and refines the queries iteratively, and finally outputs the dataset, accompanied by a summary of modifications and explanations. These individual meetings are important for maintaining the quality of the agent's outputs.

Results

We evaluated TrialGenie's performance by assessing the specialized capabilities of each agent using MIMIC-IV[57] database. We selected a diverse evaluation set of clinical trials covering various diseases, interventions, and trial phases. Our evaluation spanned four core dimensions: (1) entity extraction and trial parsing by the Trialist agent, (2) SQL query generation by the Informatician, (3) causal inference and outcome analysis by the Statistician, and (4) clinical reasoning and recommendation quality by the Clinician. Across these tasks, we benchmarked several LLMs, including GPT-4o[58] and three locally deployed LLMs Phi-4,[59] DeepSeek-R1:14b[60] (hereafter referred to as DeepSeek-R1), and Gemma-3:12b[61] (hereafter referred to as Gemma 3), to evaluate their impact on agent performance.

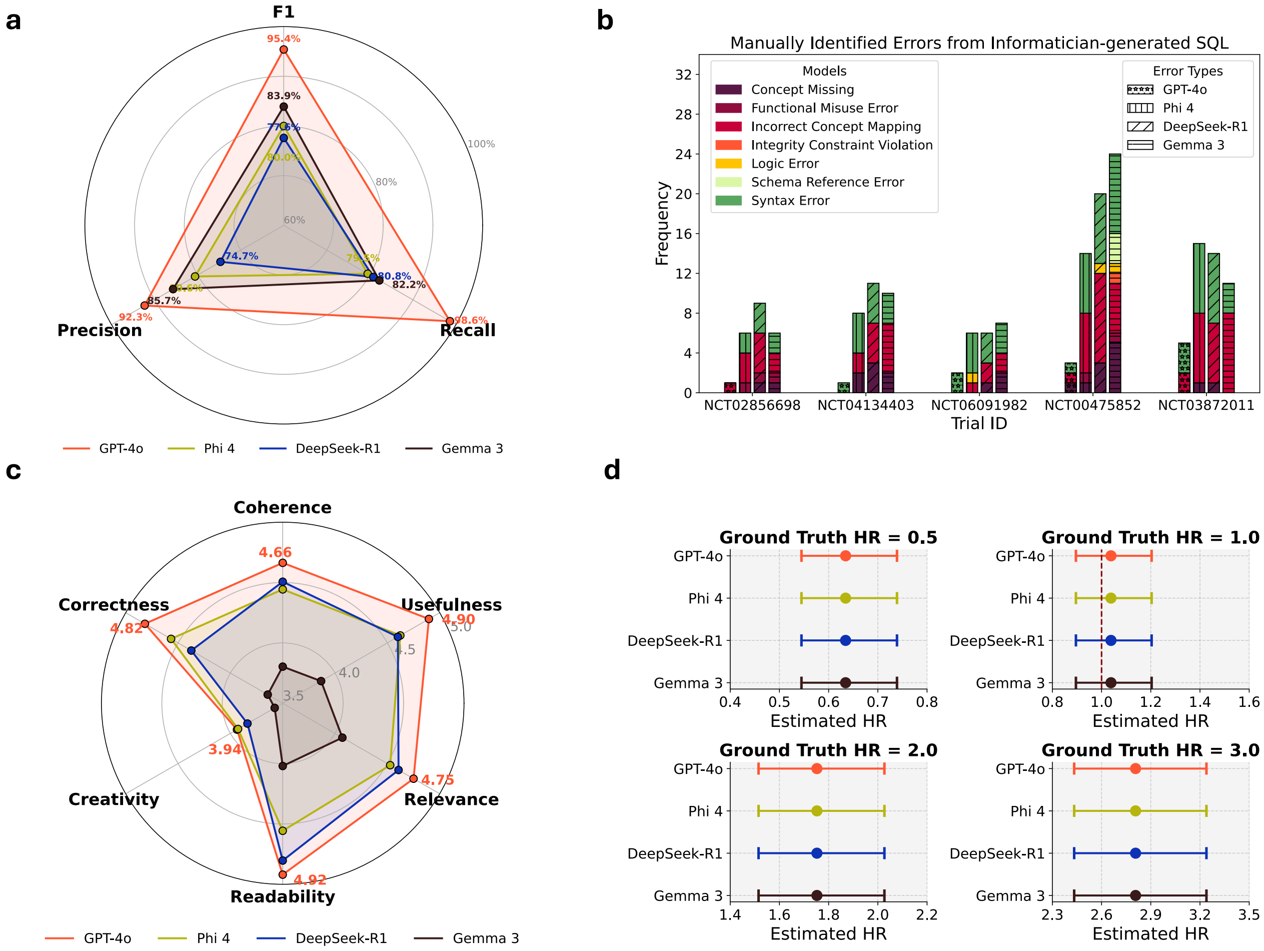


Figure 3. Evaluation of the agents in TrialGenie. (a) Comparison of entity parsing performance among four LLMs: GPT-4o, Phi-4, DeepSeek-R1, and Gemma 3, using precision, recall, and F1-score. GPT-4o achieved the highest recall (98.6%) and precision (92.3%), while Gemma 3 demonstrated the lowest performance. (b) Frequency of error types in SQL generation by the Informatician agent across different trials. GPT-4o had the fewest errors, whereas DeepSeek-R1 and Gemma 3 exhibited more errors, particularly in concept mapping and syntax formulation. (c) Evaluation of Clinician agent-generated responses based on readability, correctness, relevance, coherence, creativity, and usefulness. GPT-4o consistently outperformed other models, achieving the highest scores across all dimensions. (d) Performance comparison of estimated hazard ratios (HR) against ground truth HR values (0.5, 1.0, 2.0, 3.0) of the Statistician agent. All agents produced estimates closest to the ground truth across all scenarios.

Evaluations on Trialist

For the Trialist, evaluations concentrate on data preparation and entity parsing. Eligibility criteria of the 5 selected trials were manually annotated by two biomedical informatics experts to create a gold standard, and inter-annotator agreement is measured using Cohen's Kappa,[62] targeting a score of ≥0.7. Metrics for entity parsing include: (1) Precision: The proportion of correctly identified concepts among all extracted concepts. (2) Recall: The proportion of correctly identified concepts out of all concepts in the gold standard. (3) F1-score: The harmonic mean of Precision and Recall.

Results are presented in Figure 3(a). A total of 43 concepts were annotated across the five selected clinical trials. GPT-4o achieved the best performance in identifying these concepts, with a recall of 98.6% and precision of 92.3%. The high recall indicates that most of the ground-truth concepts (42 out of 43) can be correctly recognize. In comparison, the highest performance among the other models was achieved by Gemma, with a recall of 82.2% and precision of 85.7%. Notably, GPT-4o significantly outperformed other models in handling components that involve multiple concepts or omitted scopes, especially when enhanced with our prompt augmentation strategy. For example, in the criterion "Allergy to vitamin C, hydrocortisone, or thiamine," GPT-4o accurately extracted the concepts "allergy to vitamin C," "allergy to hydrocortisone," and "allergy to thiamine," while other models returned fragmented or incomplete extractions such as "allergy," "vitamin C," "hydrocortisone," and "thiamine." In another case, for the criterion "patients < 18 years," GPT-4o correctly inferred the omitted concept "age," which other models failed to recognize.

We selected a use case as effect of hydrocortisone in septic shock patients to further evaluate the performance within more clinical trials. Besides NCT03872011 and NCT04134403, 15 more clinical trials of this use case were selected and annotated to evaluate the trialist, with the same procedure and metrics. The results are shown in Supplementary Table 2. Similarly, GPT-4o outperforms the other models in identifying the 340 concepts, with a recall of 95.7% and precision of 86.8%.

Evaluations on Informatician

For the Informatician, evaluations focus on the SQL generation following Criteria2Query3.0,[23] using five clinical trials parsed by Trialist. Each generated SQL query is reviewed to identify potential errors manually. To ensure query correctness, the generated SQL statements were executed on the database, and their output was compared with expected cohort retrieval results.

Errors identified in the SQL queries were classified into seven categories (Supplementary Table 3), broadly falling into two types: semantic errors and structural errors.[23] Semantic errors included logic errors, where relational, temporal, or numerical expressions were misinterpreted (e.g., incorrect handling of age restrictions), concept omissions, where extracted clinical concepts were not incorporated into the query, and incorrect concept mapping, where the linkage between extracted criteria and database standard terminologies was inaccurate. Structural errors encompassed function misuse, integrity constraint violations, schema reference errors, and syntax errors, which affected the overall execution and validity of the queries.

We analyzed all four LLMs, with a focus on how accurately each model translated extracted trial information into executable SQL queries. The summarized results are shown in Figure 3b while detailed results are provided in Supplementary Table 4. Among all error types, incorrect concept mapping was the most frequent (41.18%), followed by syntax errors (31.18%), and concept missing errors (14.71%). These patterns suggest that even when clinical concepts are correctly identified, they are often misrepresented or omitted in the final query construction. Logic errors were relatively rare (2.35%) but still indicative of deeper inconsistencies in model reasoning.

Among the four LLMs, GPT-4o demonstrated the best performance, with only 12 total errors across all trials. The majority of its issues were syntax-related (58.33%, n=7), followed by incorrect concept mapping (33.33%, n=4), and a single instance of concept omission (8.33%, n=1). Importantly, GPT-4o committed no logic or integrity constraint errors, underscoring its strong consistency in both semantic interpretation and query structure. These results highlight GPT-4o's superior capacity to preserve the fidelity of trial information. Phi-4 showed moderate performance with a total of 47 errors.

It exhibited a similar error distribution pattern: 48.94% syntax errors (n=23), 40.43% incorrect concept mappings (n=19), and 8.51% concept omissions (n=4). Only one logic error (2.13%) was observed. Although its structural outputs were mostly valid, the relatively high proportion of concept mapping issues suggests that Phi-4 occasionally struggled to align extracted entities with the underlying OMOP schema. DeepSeek-R1 had the highest total error count (n=59), reflecting substantial challenges in both semantic and syntactic reliability. Incorrect concept mappings accounted for 42.37% (n=25) of its errors, and syntax issues made up 40.68% (n=24), alongside 9 concept omissions (15.25%) and one logic error. These results indicate difficulty in both capturing nuanced medical terminology and composing valid, executable queries, particularly for complex eligibility definitions. Gemma 3, while slightly outperforming DeepSeek-R1 in total error count (n=52), exhibited a similar pattern. Its errors were predominantly due to incorrect concept mapping (42.31%, n=22) and syntax issues (36.54%, n=19), along with 10 concept omissions (19.23%) and one logic error. Compared to Phi-4 and GPT-4o, Gemma 3 showed greater difficulty in preserving the completeness of clinical intent during SQL translation, especially in trials involving hierarchical or multi-part eligibility conditions. In conclusion, GPT-4o consistently outperformed the other models, making it the most reliable choice in the Informatician.

The evaluation results of GPT-4o further demonstrated a correlation between error frequency and eligibility criteria complexity[63] which reflects the number of intricate clinical concept patterns within a single phrase (Table 1). The trial NCT03872011 with the highest complexity score[63] (0.64) exhibited the most errors (n=5), while the trial NCT02856698 with the lowest complexity score (0.25) had the fewest errors (n=1). This trend suggests that query accuracy decreases as eligibility criteria become more complex, highlighting challenges in parsing and translating intricate clinical conditions into SQL queries. Overall, Informatician showed promising SQL generation capabilities but faced difficulties in handling high-complexity criteria. The results are consistent with previous findings in Criteria2Query3.0.[23]

Table 1. Complexity of eligibility criteria of these five clinical trials.

Trial ID	Number of criteria	Number of simple criteria	Number of complex criteria	Complexity score
NCT00475852	7	3	4	0.57
NCT02856698	4	3	1	0.25
NCT03872011	11	4	7	0.64
NCT04134403	10	7	3	0.30
NCT06091982	5	3	2	0.40

Evaluations on Statistician

For the Statistician, we primarily evaluate the causal inference methods. Covariate balance is measured using standardized mean differences (SMD),[64] aiming for values below 0.1. The accuracy of survival estimates is compared against published literature, while consistency with known trial results is assessed. Generated reports are also reviewed for clarity and alignment with research objectives. We also performed several evaluations of the various components of the Statistician using synthetic datasets.

For the first evaluation, we created a synthetic dataset with 1,000 subjects and 10 covariates sampled from a standard normal distribution. Treatment assignment was imbalanced using a logistic function based on sofa and age. Survival times were simulated under an exponential proportional hazards model with a baseline hazard of 0.1 and treatment effects corresponding to ground truth hazard ratios of 0.5, 1, 2, and 3. Random censoring was incorporated using an exponential distribution, and the ground truth average treatment effect (ATE) was defined as the risk difference at a 10-time unit horizon. Four different large language model (LLM) bases (GPT-4o, Phi-4, DeepSeek-R1, and Gemma 3) were used to drive agent decision-making and model selection (Table 2).

Across all LLMs, propensity score matching was consistently chosen as the balancing method by the Statistician. For a ground truth HR of 0.5, the Statistician estimated a hazard ratio of 0.6345 (95% CI: 0.5448–0.7389) and an ATE of –0.0997, compared with a ground truth ATE of –0.1759. When the true effect was null (HR = 1), the estimated HR was 1.0383 (95% CI: 0.8958–1.2034) with an ATE of 0.0244 (ground truth ATE = 0). For a moderate effect (HR = 2), the estimated HR was 1.7524 (95% CI: 1.5152–2.0267) and the ATE was 0.1461 (ground truth ATE = 0.1633). Finally, for HR = 3, the Statistician produced an estimated HR of 2.8074 (95% CI: 2.4330–3.2394) with an ATE of 0.2225, versus a ground truth ATE of 0.2832. These estimates were robust across all LLMs and outcome model selections, demonstrating that the Statistician can reliably recapitulate the ground truth parameters in trial emulation.

In the second evaluation, we generated a synthetic survival dataset with 1,000 subjects containing two covariates (sofa and age) and a binary treatment. In the overall (unstratified) analysis, the treatment effect was engineered such that the average hazard ratio (HR) was null (i.e., no significant treatment effect). However, the treatment effect was designed to interact with the sofa score such that, when subjects were grouped by a clinically meaningful cutoff, significant subgroup effects would emerge. Specifically, for treated subjects, the log hazard effect was set to +1 when the sofa score was below a threshold (e.g., <8.0) and –1 when the score was above that threshold; a constant was subtracted to force the marginal (unstratified) HR to be 1. We evaluated the ability of our multi-agent framework (incorporating Statistician and Clinician agents) to (1) detect the absence of an overall treatment effect and (2) uncover significant subgroup effects by stratifying on sofa. Three different large language model (LLM) bases (GPT-4o, Phi-4, and DeepSeek-R1) were deployed to drive agent decisions regarding subgroup creation and subsequent survival analysis using Cox proportional hazards models.

Table 2. Evaluation of Statistician's ability to recapitulate ground truth effect sizes from synthetic datasets.

LLM	Balancing Method	Chosen Outcome Models	Estimated HR	Estimated CI	Ground Truth HR	Estimated ATE	Ground Truth ATE
GPT-4o	Propensity Score Matching (PSM)	Cox Proportional Hazards, Kaplan-Meier Estimator	0.6345	(0.5448, 0.7389)	0.5	-0.0997	-0.175934
Phi-4	Propensity Score Matching (PSM)	Cox Proportional Hazards, Random Survival Forest	0.6345	(0.5448, 0.7389)	0.5	-0.0997	-0.175934
DeepSeek-R1	Propensity Score Matching (PSM)	Cox Proportional Hazards, Random Survival Forests	0.6345	(0.5448, 0.7389)	0.5	-0.0997	-0.175934
Gemma 3	Propensity Score Matching (PSM)	Cox Proportional Hazards Regression	0.6345	(0.5448, 0.7389)	0.5	-0.0997	-0.175934
GPT-4o	Propensity Score Matching (PSM)	Cox Proportional Hazards, Kaplan-Meier Estimator	1.0383	(0.8958, 1.2034)	1	0.0244	0

Phi-4	Propensity Score Matching (PSM)	Cox Proportional Hazards, Random Survival Forest	1.0383	(0.8958, 1.2034)	1	0.0244	0
DeepSeek-R1	Propensity Score Matching (PSM)	Cox Proportional Hazards, Random Survival Forests	1.0383	(0.8958, 1.2034)	1	0.0244	0
Gemma 3	Propensity Score Matching (PSM)	Cox Proportional Hazards Regression	1.0383	(0.8958, 1.2034)	1	0.0244	0
GPT-4o	Propensity Score Matching (PSM)	Cox Proportional Hazards, Kaplan-Meier Estimator	1.7524	(1.5152, 2.0267)	2	0.1461	0.16326836
Phi-4	Propensity Score Matching (PSM)	Cox Proportional Hazards, Random Survival Forest	1.7524	(1.5152, 2.0267)	2	0.1461	0.16326836
DeepSeek-R1	Propensity Score Matching (PSM)	Cox Proportional Hazards, Random Survival Forests	1.7524	(1.5152, 2.0267)	2	0.1461	0.16326836
Gemma 3	Propensity Score Matching (PSM)	Cox Proportional Hazards Regression	1.7524	(1.5152, 2.0267)	2	0.1461	0.16326836
GPT-4o	Propensity Score Matching (PSM)	Cox Proportional Hazards, Kaplan-Meier Estimator	2.8074	(2.4330, 3.2394)	3	0.2225	0.28317249
Phi-4	Propensity Score Matching (PSM)	Cox Proportional Hazards, Random Survival Forest	2.8074	(2.4330, 3.2394)	3	0.2225	0.28317249
DeepSeek-R1	Propensity Score Matching (PSM)	Cox Proportional Hazards, Random Survival Forests	2.8074	(2.4330, 3.2394)	3	0.2225	0.28317249
Gemma 3	Propensity Score Matching (PSM)	Cox Proportional Hazards Regression	2.8074	(2.4330, 3.2394)	3	0.2225	0.28317249

Table 3 shows the results. Unadjusted Cox modeling of the full dataset revealed no significant treatment effect. However, after stratification by sofa, subgroup analyses produced statistically significant differences. For analyses driven by GPT-4o and Phi-4, subjects with “SOFA < 8.0” had an estimated HR of 1.3849 [95% CI: 1.2101, 1.5850] while those with “SOFA ≥ 8.0” had an estimated HR of 0.7280 [95% CI: 0.5921, 0.8950]. In contrast, the DeepSeek-R1 and Gemma 3 driven analysis identified a slightly different cutoff (SOFA < 7.0 vs. SOFA ≥ 7.0) yielding HRs of 1.3443 [95% CI: 1.1760, 1.5366] and 0.8286 [95% CI: 0.6853, 1.0019], respectively. These findings demonstrate that while the overall analysis recapitulated the engineered null effect, appropriate stratification by the effect modifier sofa uncovered significant heterogeneity in treatment response. The consistency across different LLM bases supports the robustness of our multi-agent evaluation framework.

Table 3. Evaluation of Statistician’s ability to recapitulate ground truth effect sizes after subgroup stratification.

LLM	Number of Subgroup Analyses Done	Significant Stratification Levels	Estimated HR	Estimated CI
GPT-4o	1	SOFA < 8.0 SOFA ≥ 8.0	1.3849 0.7280	(1.2101, 1.5850) (0.5921, 0.8950)
Phi-4	2	SOFA < 8.0 SOFA ≥ 8.0	1.3849 0.7280	(1.2101, 1.5850) (0.5921, 0.8950)
DeepSeek-R1	1	SOFA < 7.0 SOFA ≥ 7.0	1.3443 0.8286	(1.1760, 1.5366) (0.6853, 1.0019)
Gemma 3	1	SOFA < 7.0 SOFA ≥ 7.0	1.3443 0.8286	(1.1760, 1.5366) (0.6853, 1.0019)

We evaluated our EC optimization algorithm by creating synthetic datasets with known ground-truth importances for each EC and then measuring how closely the Monte Carlo-based Shapley estimator recovered these importances under varying numbers of ECs. For each run, we drew a specified number of ECs (from 2 to 20), sampled true importances uniformly from a predefined range, and combined these with a baseline hazard-ratio (HR) value. We enumerated every possible subset of the ECs and generated noisy HRs by adding Gaussian noise (standard deviation = 0.5) to each subset's baseline plus its summed importances. Our algorithm then performed randomized permutations of the ECs to compute incremental differences in HR from the empty to the full set of rules, up to a maximum of 1000 iterations or until the standard error of the mean (SEM) reached a small tolerance (1e-3). We compared each criterion's estimated Shapley value against its ground-truth importance using mean absolute error (MAE). Supplementary Table 5 shows the MAE for varying number of ECs.

Supplementary Table 5 shows that, even in the presence of moderate noise, the algorithm converges effectively and typically produces low MAE values, indicating good agreement between the estimated Shapley values and the known true importances.

Evaluations on Clinician

The Clinician agent is evaluated for its ability to integrate medical expertise into the workflow. A questionnaire based on a 5-point Likert scale[65] assesses the relevance, clarity, and accuracy of the Clinician's recommendations.[23] Additionally, the agent's efficiency in refining trial designs and resolving discrepancies is measured through qualitative feedback from domain experts.

Table 4: Questionnaire and mean scores with standard deviations on Clinician-generated responses.

Category	Question	Mean score (standard deviation)			
		GPT-4o	Phi-4	DeepSeek-R1	Gemma 3
Readability	The overall writing format is easy to understand.	4.92 (0.13)	4.59 (0.35)	4.80 (0.17)	4.02 (0.25)
	The response is well-structured and logically organized (e.g., proper use of paragraphs, bullet points).	4.92 (0.13)	4.49 (0.37)	4.80 (0.17)	4.02 (0.25)
	The terminology and language style are appropriate for the intended audience.	4.92 (0.13)	4.59 (0.35)	4.81 (0.18)	4.02 (0.25)
Correctness	The response accurately conveys factual information and is free from hallucinations.	4.85 (0.26)	4.57 (0.58)	4.41 (0.32)	3.61 (0.45)
	The logical reasoning in the response is sound and follows expected principles.	4.80 (0.23)	4.55 (0.54)	4.35 (0.33)	3.63 (0.44)
	The response correctly uses technical or domain-specific terms when applicable.	4.80 (0.23)	4.58 (0.53)	4.36 (0.32)	3.69 (0.41)
Relevance	The response directly addresses the question or task given.	4.88 (0.12)	4.74 (0.33)	4.73 (0.18)	4.07 (0.04)

	The information provided is sufficiently detailed without unnecessary content.	4.88 (0.12)	4.60 (0.49)	4.76 (0.17)	4.07 (0.04)
	The response includes relevant supporting details, examples, or citations if required.	4.49 (0.70)	4.24 (0.65)	4.33 (0.81)	4.07 (0.04)
Coherence	The response maintains internal logical consistency throughout.	4.63 (0.52)	4.40 (0.72)	4.46 (0.70)	3.80 (0.40)
	If multiple responses are generated for the same prompt, they are consistent with each other.	4.68 (0.56)	4.43 (0.75)	4.53 (0.75)	3.81 (0.39)
	The response avoids contradictions or conflicting statements.	4.68 (0.56)	4.50 (0.81)	4.53 (0.75)	3.80 (0.40)
Creativity	The response demonstrates originality and innovative reasoning.	3.74 (0.95)	3.74 (0.97)	3.69 (1.00)	3.58 (0.51)
	The answer provides novel insights beyond common knowledge.	3.89 (0.89)	3.88 (0.91)	3.81 (0.93)	3.57 (0.50)
	The response presents multiple perspectives when appropriate.	4.18 (1.02)	4.16 (1.00)	4.01 (0.93)	3.58 (0.51)
Usefulness	The response is useful for the intended application (e.g., SQL generation, summarization, question answering).	4.92 (0.13)	4.56 (0.30)	4.54 (0.26)	3.87 (0.35)
	The generated response is easy to modify or refine for further use.	4.90 (0.12)	4.58 (0.33)	4.57 (0.22)	3.86 (0.34)
	The response format aligns with the requirements of the target application.	4.87 (0.18)	4.73 (0.32)	4.69 (0.08)	3.87 (0.35)
Average Score		**4.66**	**4.44**	**4.45**	**3.83**

Figure 3c presents the evaluation of Clinician-generated responses across five key dimensions: readability, correctness, relevance, coherence, and usefulness, with detailed subcategory scores provided in Table 4. Overall, GPT-4o outperformed all other models, achieving the highest average score of 4.66. DeepSeek-R1 and Phi-4 showed comparable overall performance (4.45 and 4.44, respectively), while Gemma 3 lagged behind with a significantly lower average score of 3.83.

- Readability: GPT-4o achieved perfect consistency in readability, scoring 4.92 across all aspects, including writing clarity, logical structure, and appropriate terminology. DeepSeek-R1 followed with solid scores around 4.80–4.81, showing well-organized responses with minor gaps in structure. Phi-4 scored slightly lower (4.49–4.59), suggesting occasional inconsistencies in formatting or flow. In contrast, Gemma 3 consistently underperformed in this dimension, with all readability items scoring 4.02, indicating challenges in producing clearly formatted and accessible clinical content.
- Correctness: GPT-4o again led in this category, scoring 4.85 for factual accuracy and 4.80 in logical reasoning and domain-specific usage, reflecting both precision and reasoning strength. Phi-4 was close behind (4.55–4.58), though with occasional factual or terminology slips. DeepSeek-R1 showed greater variability (4.35–4.41), pointing to sporadic hallucinations[66] or flawed reasoning. Gemma 3 trailed significantly, especially in factual correctness (3.61) and logical soundness (3.63), with further struggles in technical accuracy (3.69), underscoring its relative unreliability in clinical contexts.
- Relevance: GPT-4o achieved near-perfect scores in addressing tasks directly (4.88) and delivering the right level of detail (4.88), though its supporting citations could still improve (4.49). DeepSeek-R1 remained competitive (4.73–4.76), while Phi-4 had a wider spread (4.24–4.74), with occasional digressions or missing justifications. Gemma 3, despite being consistent, scored only 4.07 across all subcategories, suggesting generally on-topic but overly generic responses with limited depth or support.
- Coherence: GPT-4o performed strongly again (4.63–4.68), maintaining logical consistency across individual and multiple responses. DeepSeek-R1 showed fairly coherent outputs (4.46–4.53), but Phi-4 was less stable (4.40–4.50), occasionally presenting contradictions.

Gemma 3's coherence scores (3.80–3.81) point to structural fragility, with frequent internal inconsistencies or mismatches between ideas.

- Creativity: All four models showed room for growth in this aspect. GPT-4o displayed relatively higher creativity (3.74–4.18), especially in offering multiple perspectives. Phi-4 and DeepSeek-R1 were comparable (~3.7–4.1), while Gemma 3 again lagged with a narrow range (3.57–3.58).
- Usefulness: GPT-4o stood out as the most practically useful model, scoring 4.92 for application fit and 4.90 for modifiability. DeepSeek-R1 (4.54–4.69) and Phi-4 (4.56–4.73) were also rated as useful, but showed occasional formatting or specificity limitations. Gemma 3 scored the lowest in this category (3.86–3.87), indicating its output often lacked polish or adaptability.

In summary, GPT-4o delivered the most accurate, relevant, and clinically useful responses, making it the clear leader among all evaluated LLMs. DeepSeek-R1 and Phi-4 demonstrated decent usability with minor flaws, while Gemma 3 consistently underperformed across nearly all dimensions, indicating that its clinical reasoning and formatting still require significant improvement for integration in expert workflows.

Showcases

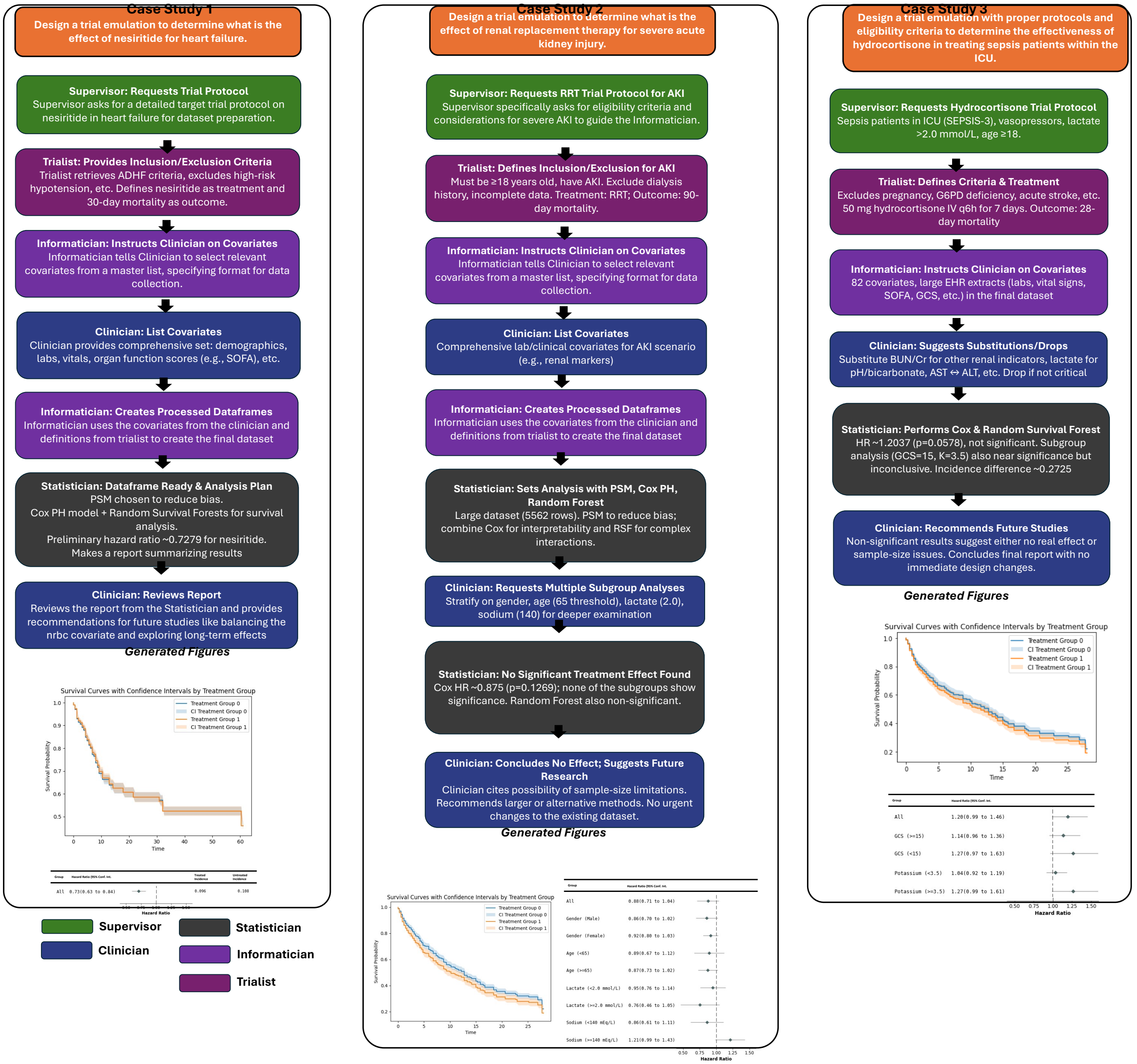

Case Study 1
Design a trial emulation to determine what is the effect of nesiritide for heart failure.
Supervisor: Requests Trial Protocol
Supervisor asks for a detailed target trial protocol on nesiritide in heart failure for dataset preparation.
Trialist: Provides Inclusion/Exclusion Criteria
Trialist retrieves ADHF criteria, excludes high-risk hypotension, etc. Defines nesiritide as treatment and 30-day mortality as outcome.
Informatician: Instructs Clinician on Covariates
Informatician tells Clinician to select relevant covariates from a master list, specifying format for data collection.
Clinician: List Covariates
Clinician provides comprehensive set: demographics, labs, vitals, organ function scores (e.g., SOFA), etc.
Informatician: Creates Processed Dataframes
Informatician uses the covariates from the clinician and definitions from trialist to create the final dataset
Statistician: Dataframe Ready & Analysis Plan
PSM chosen to reduce bias.
Cox PH model + Random Survival Forests for survival analysis.
Preliminary hazard ratio ~0.7279 for nesiritide.
Makes a report summarizing results
Clinician: Reviews Report
Reviews the report from the Statistician and provides recommendations for future studies like balancing the nrbc covariate and exploring long-term effects
Generated Figures
Survival Curves with Confidence Intervals by Treatment Group
Treatment Group 0
CI Treatment Group 0
Treatment Group 1
CI Treatment Group 1
Survival Probability
1.0
0.9
0.8
0.7
0.6
0.5
0
10
20
30
40
50
60
Time
Group
Hazard Ratio (95% Conf. Int.
Treated Incidence
Untreated Incidence
All
0.73(0.63 to 0.84)
0.096
0.108
0.50
0.75
1.00
1.25
1.50
Hazard Ratio
Case Study 2
Design a trial emulation to determine what is the effect of renal replacement therapy for severe acute kidney injury.
Supervisor: Requests RRT Trial Protocol for AKI
Supervisor specifically asks for eligibility criteria and considerations for severe AKI to guide the Informatician.
Trialist: Defines Inclusion/Exclusion for AKI
Must be ≥18 years old, have AKI. Exclude dialysis history, incomplete data. Treatment: RRT; Outcome: 90-day mortality.
Informatician: Instructs Clinician on Covariates
Informatician tells Clinician to select relevant covariates from a master list, specifying format for data collection.
Clinician: List Covariates
Comprehensive lab/clinical covariates for AKI scenario (e.g., renal markers)
Informatician: Creates Processed Dataframes
Informatician uses the covariates from the clinician and definitions from trialist to create the final dataset
Statistician: Sets Analysis with PSM, Cox PH, Random Forest
Large dataset (5562 rows). PSM to reduce bias; combine Cox for interpretability and RSF for complex interactions.
Clinician: Requests Multiple Subgroup Analyses
Stratify on gender, age (65 threshold), lactate (2.0), sodium (140) for deeper examination
Statistician: No Significant Treatment Effect Found
Cox HR ~0.875 (p=0.1269); none of the subgroups show significance. Random Forest also non-significant.
Clinician: Concludes No Effect; Suggests Future Research
Clinician cites possibility of sample-size limitations. Recommends larger or alternative methods. No urgent changes to the existing dataset.
Generated Figures
Survival Curves with Confidence Intervals by Treatment Group
Treatment Group 0
CI Treatment Group 0
Treatment Group 1
CI Treatment Group 1
Survival Probability
1.0
0.8
0.6
0.4
0.2
0
5
10
15
20
25
Time
Group
Hazard Ratio (95% Conf. Int.
All
0.88(0.71 to 1.04)
Gender (Male)
0.86(0.70 to 1.02)
Gender (Female)
0.92(0.80 to 1.03)
Age (<65)
0.89(0.67 to 1.12)
Age (>=65)
0.87(0.73 to 1.02)
Lactate (<2.0 mmol/L)
0.95(0.76 to 1.14)
Lactate (>=2.0 mmol/L)
0.76(0.46 to 1.05)
Sodium (<140 mEq/L)
0.86(0.61 to 1.11)
Sodium (>=140 mEq/L)
1.21(0.99 to 1.43)
0.50
0.75
1.00
1.25
1.50
Hazard Ratio
Case Study 3
Design a trial emulation with proper protocols and eligibility criteria to determine the effectiveness of hydrocortisone in treating sepsis patients within the ICU.
Supervisor: Requests Hydrocortisone Trial Protocol
Sepsis patients in ICU (SEPSIS-3), vasopressors, lactate >2.0 mmol/L, age ≥18.
Trialist: Defines Criteria & Treatment
Excludes pregnancy, G6PD deficiency, acute stroke, etc. 50 mg hydrocortisone IV q6h for 7 days. Outcome: 28-day mortality
Informatician: Instructs Clinician on Covariates
82 covariates, large EHR extracts (labs, vital signs, SOFA, GCS, etc.) in the final dataset
Clinician: Suggests Substitutions/Drops
Substitute BUN/Cr for other renal indicators, lactate for pH/bicarbonate, AST ↔ ALT, etc. Drop if not critical
Statistician: Performs Cox & Random Survival Forest
HR ~1.2037 (p=0.0578), not significant. Subgroup analysis (GCS=15, K=3.5) also near significance but inconclusive. Incidence difference ~0.2725
Clinician: Recommends Future Studies
Non-significant results suggest either no real effect or sample-size issues. Concludes final report with no immediate design changes.
Generated Figures
Survival Curves with Confidence Intervals by Treatment Group
Treatment Group 0
CI Treatment Group 0
Treatment Group 1
CI Treatment Group 1
Survival Probability
1.0
0.8
0.6
0.4
0.2
0
5
10
15
20
25
Time
Group
Hazard Ratio (95% Conf. Int.
All
1.20(0.99 to 1.46)
GCS (>=15)
1.14(0.96 to 1.36)
GCS (<15)
1.27(0.97 to 1.63)
Potassium (<3.5)
1.04(0.92 to 1.19)
Potassium (>=3.5)
1.27(0.99 to 1.61)
0.50
0.75
1.00
1.25
1.50
Hazard Ratio
Supervisor
Clinician
Statistician
Informatician
Trialist

Figure 4. Showcases on three clinical trials.

To demonstrate how TrialGenie operates on diverse clinical questions, we present three separate disease-based examples using Phi-4, each reflecting distinct trial emulation scenarios. These examples illustrate how the Supervisor, Trialist, Informatician, Clinician, and Statistician agents coordinate to parse protocols, handle eligibility criteria, select covariates, balance confounders, and produce final analyses.

Case 1: Impact of Nesiritide in Acute Heart Failure Patients

In this first demonstration, TrialGenie emulated a clinical trial (NCT00475852) assessing nesiritide's impact on heart failure outcomes (Supplemental File 1). The Supervisor initially requested a detailed target trial protocol, and the Trialist retrieved inclusion criteria specifying patients hospitalized for acute decompensated heart failure or diagnosed within 48 hours after admission, while excluding those at high risk of hypotension, those with systolic blood pressure above 180 mmHg, and individuals presenting severe structural heart conditions or prior nesiritide enrollment. The Supervisor and Clinician then collaborated to identify a comprehensive set of covariates, including age, gender, lactate, arterial blood gases, renal function markers, coagulation parameters, and neurological and cardiovascular metrics. These covariates were provided in a structured format to the Informatician, who constructed a dataset containing 6971 rows aligned with the specified trial criteria.

Upon receiving this dataset, the Statistician applied propensity score matching to reduce confounding between patients treated with nesiritide versus controls. There was one persistently unbalanced covariate (nrbc), but balance between the two treatment groups was improved for the remaining variables. Survival analyses then employed two methods. A Cox proportional hazards model indicated a hazard ratio near 0.73 (95% confidence interval approximately 0.63–0.84), suggesting a statistically significant reduction in adverse event risk among nesiritide-treated patients. A random survival forests approach supported these findings by demonstrating good predictive accuracy (C-index of about 0.79), although the short-term hazard ratio at 30 days remained near neutral (about 0.98).

The final report, titled "Emulating the Impact of Nesiritide in Heart Failure Patients: A Propensity Score-Matched Analysis," (Supplemental File 1) documented a slightly lower incidence of adverse events in the nesiritide group (9.57%) compared to controls (10.80%). The Clinician reviewed the results and recommended refining matching to address the unbalanced nrbc covariate, as well as conducting subgroup analyses to identify patient subsets that might derive particular benefit. Despite these limitations, the study's main conclusion—namely that nesiritide may offer a clinically meaningful decrease in adverse event risk—was consistent with existing evidence, illustrating TrialGenie's capacity to reproduce and refine insights from established clinical trials.

Case 2: Effect of Renal Replacement Therapy for Severe Acute Kidney Injury

In this second example, TrialGenie was tasked with emulating a study to determine whether renal replacement therapy (RRT) affects 90-day mortality among patients diagnosed with severe acute kidney injury (AKI) (Supplemental File 2). The Supervisor directed the Trialist to parse a target trial protocol requiring patients to be at least 18 years old with AKI, excluding those who had prior dialysis history, ongoing renal failure in dialysis, or incomplete data. The Clinician then selected covariates ranging from age, lactate, electrolytes, and coagulation parameters to scores like SOFA, which the Informatician used to assemble and clean a dataset of 5562 rows. Once the data was prepared, the Statistician applied propensity score matching to balance those who received RRT against those who did not, selecting Cox proportional hazards and random survival forests for survival analysis. Neither model indicated a significant difference in outcomes: the Cox model's hazard ratio (about 0.875) was not statistically significant, and the random survival forests analysis, despite demonstrating a high C-index (~0.806), yielded a hazard ratio near unity. Subgroup analyses stratified by gender, age ≥65, lactate ≥2.0, or sodium ≥140 similarly found no statistically significant treatment effects.

These findings suggest either that RRT had no meaningful impact within the MIMIC-IV study population or that factors such as sample size and unmeasured confounding limited detection of a true effect (Supplemental File 3). The Clinician considered additional covariates or larger cohorts to refine the emulation but concluded that the current dataset could not provide conclusive evidence. Overall, the final report emphasizes the need for expanded patient samples or alternative methods to clarify RRT's role in severe AKI management, demonstrating how TrialGenie manages null results and negative findings as systematically as it does positive ones.

Case 3: Effect of Hydrocortisone in Septic Shock Patients

In the third example, TrialGenie emulated the design of a trial (NCT04134403) evaluating hydrocortisone therapy in patients with septic shock (Supplemental File 3). The Supervisor first requested the Trialist to parse a protocol that stipulated a 24-hour window since ICU admission, vasopressor dependence (for maintaining adequate mean arterial pressure), lactate >2.0 mmol/L, and age ≥18 years. Exclusions encompassed pregnancy, G6PD deficiency, acute stroke or coronary syndrome, major hemorrhage, burn, trauma, or prolonged vasopressor use (>24 hours before randomization). The Clinician then selected an extensive set of covariates—spanning blood gases, electrolytes, organ function markers, and severity scores—to be curated by the Informatician. The resulting dataset contained 1153 patients meeting the inclusion and exclusion criteria.

Upon finalizing data cleaning, the Statistician chose propensity score matching to reduce confounding between patients receiving hydrocortisone (50 mg every 6 hours for up to 7 days) and controls. Two primary approaches—Cox Proportional Hazards and Kaplan-Meier estimation—were used to assess 28-day mortality. After identifying several unbalanced covariates, the Clinician proposed substituting specific measurements (e.g., ALT for AST, bicarbonate for anion gap, GCS total score for its individual components) to improve data balance and maintain clinical relevance.

Post-matching analyses indicated that hydrocortisone did not show a statistically significant benefit on survival when controlling for confounders (Supplemental File 3). While not significant, slight harm was shown from the use of steroids. These findings did not align with the original trial's results, suggesting that hydrocortisone may differ in treatment effect dependent on the cohort. Future efforts could explore refining eligibility criteria, as indicated by the SHAP-based optimization results. Overall, this showcase highlights how TrialGenie can rigorously reproduce an existing protocol, refine data quality, and generate robust causal inferences for critical-care interventions.

Discussions

TrialGenie represents an advancement in the use of agentic systems for empowering clinical trial design. By integrating role-specialized LLM agents, TrialGenie transforms a traditionally manual, expertise-intensive and time-consuming CTD process into an autonomous pipeline.

Among the five agents, the Trialist demonstrated its capability in retrieving and standardizing trial information, achieving a high F1-score of 95.4% using GPT-4o. The Informatician agent translated these parsed elements into SQL queries, showing that LLMs can map trial information into structured cohort definitions. The Statistician agent effectively executed causal inference workflows—matching gold-standard hazard ratio estimates across synthetic and real-world data—while the Clinician agent ensured medical validity and interpretability through structured reasoning grounded in biomedical literature.

Notably, GPT-4o consistently outperformed locally deployed models (Phi-4, DeepSeek-R1, and Gemma 3) across tasks, particularly in logical accuracy, syntactic correctness, clinical relevance. Our analysis of error types—most notably incorrect concept mapping and concept omission—points to persistent challenges in bridging trial information with data operations, especially in settings involving complex logic. These findings reinforce the need for systems that can reason not only across language and code, but also across clinical abstraction layers.

By modularizing clinical trial design into distinct roles, i.e., Supervisor, Trialist, Informatician, Clinician, and Statistician, TrialGenie reflects the natural division of efforts in clinical research teams. The architecture enables agents to iterate collaboratively with refining eligibility definitions, proposing surrogate covariates, or adjusting statistical models, which helps mirror the dynamic, interdisciplinary nature of clinical trial design in practice. The integration of Reinforcement Learning from Human Feedback (RLHF) further aligns model behavior with human experts, allowing agents to improve the response quality. These designs allow TrialGenie to function not merely as a pipeline, but as a learning ecosystem that continuously refines both knowledge and execution.

From a broader perspective, TrialGenie offers a new paradigm for integrating real-world evidence (RWE) into clinical research. Traditional pipelines often require substantial human intervention to iteratively reconcile protocol design, making scalability difficult. TrialGenie's agentic intelligence bridges this gap by autonomously surfacing data issues (e.g., missingness), querying domain knowledge via RAG, and adjusting the trial design accordingly. Such capability is particularly powerful in the domains where real-time insights are essential and trial feasibility is often constrained.

Several important directions remain for future exploration. First, many clinical trials incorporate not just structured EHR data, but also clinical notes, imaging, genomic profiles, or wearable data. Extending TrialGenie to handle multi-modal inputs would unlock new applications in precision medicine and rare disease research. This would require enhancing the reasoning capabilities of agents like the Clinician and Statistician to synthesize evidence across modalities. Second, although TrialGenie was evaluated on the MIMIC-IV dataset, its broader applicability across other healthcare systems remains an open question. Extending TrialGenie to function reliably across diverse databases or in federated environments with privacy constraints would enhance its value for multicenter trial design. Third, Clinician agent currently relies on RAG over curated literature corpora. Future iterations could enrich this by integrating biomedical knowledge graphs (e.g., iBKH[67]) to support relational reasoning over structured knowledge. This would enhance the agent's ability to infer indirect associations (e.g., comorbidities, mechanistic pathways), improve covariate selection, and provide more context-aware recommendations.

References

1. Hernán, M. A. & Robins, J. M. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *Am. J. Epidemiol.* **183**, 758–764 (2016).

2. Wang, S. V., Sreedhara, S. K. & Schneeweiss, S. Reproducibility of real-world evidence studies using clinical practice data to inform regulatory and coverage decisions. *Nat. Commun.* **13**, 5126 (2022).

3. Austin, P. C. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivar. Behav. Res.* **46**, 399–424 (2011).

4. Orcutt, X., Chen, K., Mamtani, R., Long, Q. & Parikh, R. B. Evaluating generalizability of oncology trial results to real-world patients using machine learning-based trial emulations. *Nat. Med.* **31**, 457–465 (2025).

5. Feuerriegel, S. *et al.* Causal machine learning for predicting treatment outcomes. *Nat. Med.* **30**, 958–968 (2024).

6. Hernán, M. A., Wang, W. & Leaf, D. E. Target Trial Emulation: A Framework for Causal Inference From Observational Data. *JAMA* **328**, 2446–2447 (2022).

7. Data Standardization – OHDSI. https://www.ohdsi.org/data-standardization/.

8. Hernán, M. A., Wang, W. & Leaf, D. E. Target Trial Emulation: A Framework for Causal Inference From Observational Data. *JAMA* **328**, 2446–2447 (2022).

9. Zang, C. *et al.* High-throughput target trial emulation for Alzheimer's disease drug repurposing with real-world data. *Nat. Commun.* **14**, 8180 (2023).

10. He, T. *et al.* Trends and opportunities in computable clinical phenotyping: A scoping review. *J. Biomed. Inform.* **140**, 104335 (2023).

11. Liu, R. *et al.* Evaluating eligibility criteria of oncology trials using real-world data and AI. *Nature* **592**, 629–633 (2021).

12. Rajendran, S. *et al.* Corticosteroids for infectious critical illness: A multicenter target trial emulation stratified by predicted organ dysfunction trajectory. *MedRxiv Prepr. Serv. Health Sci.* 2024.03.07.24303926 (2024) doi:10.1101/2024.03.07.24303926.

13. Li, X., Wang, S., Zeng, S., Wu, Y. & Yang, Y. A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth* **1**, 9 (2024).

14. The Virtual Lab: AI Agents Design New SARS-CoV-2 Nanobodies with Experimental Validation | bioRxiv. https://www.biorxiv.org/content/10.1101/2024.11.11.623004v1.

15. Li, J. *et al.* Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents. Preprint at https://doi.org/10.48550/arXiv.2405.02957 (2025).

16. Yue, L., Xing, S., Chen, J. & Fu, T. ClinicalAgent: Clinical Trial Multi-Agent System with Large Language Model-based Reasoning. Preprint at https://doi.org/10.48550/arXiv.2404.14777 (2024).

17. MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making. https://arxiv.org/abs/2404.15155.

18. Shi, W. *et al.* EHRAgent: Code Empowers Large Language Models for Few-shot Complex Tabular Reasoning on Electronic Health Records. Preprint at https://doi.org/10.48550/arXiv.2401.07128 (2024).

19. Cui, H. *et al.* LLMs-based Few-Shot Disease Predictions using EHR: A Novel Approach Combining Predictive Agent Reasoning and Critical Agent Instruction. Preprint at https://doi.org/10.48550/arXiv.2403.15464 (2024).

20. Natural Language Programming in Medicine: Administering Evidence Based Clinical Workflows with Autonomous Agents Powered by Generative Large Language Models. https://arxiv.org/abs/2401.02851.

21. Kaufmann, T., Weng, P., Bengs, V. & Hüllermeier, E. A Survey of Reinforcement Learning from Human Feedback. *arXiv.org* https://arxiv.org/abs/2312.14925v2 (2023).

22. Johnson, A. E. W. *et al.* MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* **10**, 1 (2023).

23. Park, J. *et al.* Criteria2Query 3.0: Leveraging generative large language models for clinical trial eligibility query generation. *J. Biomed. Inform.* **154**, 104649 (2024).

24. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**, D267–D270 (2004).

25. OHDSI – Observational Health Data Sciences and Informatics. https://www.ohdsi.org/.

26. Crowe, C. L. & Tao, C. Designing Ontology-based Patterns for the Representation of the Time-Relevant Eligibility Criteria of Clinical Protocols. *AMIA Summits Transl. Sci. Proc.* **2015**, 173–177 (2015).

27. Li, F. *et al.* Time event ontology (TEO): to support semantic representation and reasoning of complex temporal relations of clinical events. *J. Am. Med. Inform. Assoc. JAMIA* **27**, 1046–1056 (2020).

28. Harris, D. R., Henderson, D. W., Kavuluru, R., Stromberg, A. J. & Johnson, T. R. Using common table expressions to build a scalable Boolean query generator for clinical data warehouses. *IEEE J. Biomed. Health Inform.* **18**, 1607–1613 (2014).

29. Kim, W. On optimizing an SQL-like nested query. *ACM Trans Database Syst* **7**, 443–469 (1982).

30. Datta, S. *et al.* AutoCriteria: a generalizable clinical trial eligibility criteria extraction system powered by large language models. *J. Am. Med. Inform. Assoc. JAMIA* **31**, 375–385 (2023).

31. Weng, C., Tu, S. W., Sim, I. & Richesson, R. Formal representation of eligibility criteria: a literature review. *J. Biomed. Inform.* **43**, 451–467 (2010).

32. Lee, K., Weiskopf, N. & Pathak, J. A Framework for Data Quality Assessment in Clinical Research Datasets. *AMIA. Annu. Symp. Proc.* **2017**, 1080–1089 (2018).

33. Schmidt, C. O. *et al.* Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. *BMC Med. Res. Methodol.* **21**, 63 (2021).

34. Bernardi, F. A. *et al.* Data Quality in Health Research: Integrative Literature Review. *J. Med. Internet Res.* **25**, e41446 (2023).

35. Gress, T. W., Denvir, J. & Shapiro, J. I. Effect of removing outliers on statistical inference: implications to interpretation of experimental data in medical research. *Marshall J. Med.* **4**, 9 (2018).

36. Kazijevs, M. & Samad, M. D. Deep imputation of missing values in time series health data: A review with benchmarking. *J. Biomed. Inform.* **144**, 104440 (2023).

37. Kane, L. T. *et al.* Propensity Score Matching: A Statistical Method. *Clin. Spine Surg.* **33**, 120–122 (2020).

38. Caliendo, M. & Kopeinig, S. Some Practical Guidance for the Implementation of Propensity Score Matching. *J. Econ. Surv.* **22**, 31–72 (2008).

39. ROSENBAUM, P. R. & RUBIN, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55 (1983).

40. Chesnaye, N. C. *et al.* An introduction to inverse probability of treatment weighting in observational research. *Clin. Kidney J.* **15**, 14–20 (2022).

41. Bettega, F., Mendelson, M., Leyrat, C. & Bailly, S. Use and reporting of inverse-probability-of-treatment weighting for multicategory treatments in medical research: a systematic review. *J. Clin. Epidemiol.* **170**, 111338 (2024).

42. Austin, P. C. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Stat. Med.* **35**, 5642–5655 (2016).

43. Chesnaye, N. C. *et al.* An introduction to inverse probability of treatment weighting in observational research. *Clin. Kidney J.* **15**, 14–20 (2022).

44. Kumar, D. & Klefsjö, B. Proportional hazards model: a review. *Reliab. Eng. Syst. Saf.* **44**, 177–188 (1994).

45. Royston, P. & Parmar, M. K. B. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Stat. Med.* **21**, 2175–2197 (2002).

46. McLernon, D. J. *et al.* Assessing Performance and Clinical Usefulness in Prediction Models With Survival Outcomes: Practical Guidance for Cox Proportional Hazards Models. *Ann. Intern. Med.* **176**, 105–114 (2023).

47. RICH, J. T. *et al.* A PRACTICAL GUIDE TO UNDERSTANDING KAPLAN-MEIER CURVES. *Otolaryngol.--Head Neck Surg. Off. J. Am. Acad. Otolaryngol.-Head Neck Surg.* **143**, 331–336 (2010).

48. Pickett, K. L., Suresh, K., Campbell, K. R., Davis, S. & Juarez-Colunga, E. Random survival forests for dynamic predictions of a time-to-event outcome using a longitudinal biomarker. *BMC Med. Res. Methodol.* **21**, 216 (2021).

49. Funk, M. J. *et al.* Doubly Robust Estimation of Causal Effects. *Am. J. Epidemiol.* **173**, 761–767 (2011).

50. Liu, R. *et al.* Evaluating eligibility criteria of oncology trials using real-world data and AI. *Nature* **592**, 629–633 (2021).

51. Gao, Y. *et al.* Retrieval-Augmented Generation for Large Language Models: A Survey. Preprint at https://doi.org/10.48550/arXiv.2312.10997 (2024).

52. Douze, M. *et al.* The Faiss library. Preprint at https://doi.org/10.48550/arXiv.2401.08281 (2024).

53. Bai, Y. *et al.* Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. Preprint at https://doi.org/10.48550/arXiv.2204.05862 (2022).

54. Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. Proximal Policy Optimization Algorithms. Preprint at https://doi.org/10.48550/arXiv.1707.06347 (2017).

55. Rafailov, R. *et al.* Direct Preference Optimization: Your Language Model is Secretly a Reward Model. Preprint at https://doi.org/10.48550/arXiv.2305.18290 (2024).

56. Swanson, K., Wu, W., Bulaong, N. L., Pak, J. E. & Zou, J. The Virtual Lab: AI Agents Design New SARS-CoV-2 Nanobodies with Experimental Validation. 2024.11.11.623004 Preprint at https://doi.org/10.1101/2024.11.11.623004 (2024).

57. MIMIC-IV, a freely accessible electronic health record dataset | Scientific Data. https://www.nature.com/articles/s41597-022-01899-x.

58. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/.

59. Abdin, M. *et al.* Phi-4 Technical Report. Preprint at https://doi.org/10.48550/arXiv.2412.08905 (2024).

60. DeepSeek-AI *et al.* DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. Preprint at https://doi.org/10.48550/arXiv.2501.12948 (2025).

61. Introducing Gemma 3: The most capable model you can run on a single GPU or TPU. *Google* https://blog.google/technology/developers/gemma-3/ (2025).

62. McHugh, M. L. Interrater reliability: the kappa statistic. *Biochem. Medica* **22**, 276–282 (2012).

63. Ross, J., Tu, S., Carini, S. & Sim, I. Analysis of Eligibility Criteria Complexity in Clinical Trials. *Summit Transl. Bioinforma.* **2010**, 46–50 (2010).

64. Andrade, C. Mean Difference, Standardized Mean Difference (SMD), and Their Use in Meta-Analysis: As Simple as It Gets. *J. Clin. Psychiatry* **81**, 20f13681 (2020).

65. Likert, R. A technique for the measurement of attitudes. *Arch. Psychol.* **22 140**, 55–55 (1932).

66. Huang, L. *et al.* A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* **43**, 1–55 (2025).

67. Su, C. *et al.* Biomedical discovery through the integrative biomedical knowledge hub (iBKH). *iScience* **26**, 106460 (2023).

Supplemental Materials

Supplementary Table 1. Overall prompt designs used by Trialist. The prompt is mainly adapted from Criteria2Query3.0[23] with our modifications highlighted in red.

Task	Prompt
Concept extraction (Original from Criteria2Query 3.0)	Annotate clinical concepts from the given text using the following rules: 1) Annotate concepts with the domains 'Demographic', 'Condition', 'Device', 'Procedure', 'Drug', 'Measurement', 'Observation', 'Visit', 'Value', 'Negation_cue', 'Temporal', and 'Quantity'. If you cannot annotate with the given domains, you can name a new one (e.g., Drug_cycle, Visit, Provider, etc.). 2) Split the concepts as detail as possible. Each concept can be annotated only once with a single domain. 3) Normalize clinical abbreviation and acronyms and attached behind the original abbreviation with parenthesis. 4) Return your response under [Annotation] section. Following is not allowed examples: 1) <Measurement>EGFR <Value>triple postive</Value></Measurement> 2) <Condition>Hypertension, diabetes, heart failure, and dementia</Condition> Below is allowed examples: 1) <Measurement>EGFR</Measurement> <Value>triple positive</Value> 2) <Condition>hypertension</Condition>, <Condition>T2DM (Type 2 Diabetes Mellitus)</Condition>, <Condition>heart failure</Condition>, and <Condition>dementia</Condition> 3) Patient <Demographic>aged<Demographic> > <Value>65 years old</Value> 4) <Drug>Metformin</Drug> <Value>500 mg</Value> <Temporal>daily</Temporal> Following is information for each domain: 1) Condition is events of a Person suggesting the presence of a disease or medical condition stated as a diagnosis, a sign, or a symptom, which is either observed by a Provider or reported by the patient. 2) Drugs include prescription and over-the-counter medicines, vaccines, and large-molecule biologic therapies. Radiological devices ingested or applied locally do not count as Drugs. 3) Procedure is records of activities or processes ordered by, or carried out by, a healthcare provider on the patient with a diagnostic or therapeutic purpose. Lab tests are not a procedure, if something is observed with an expected resulting amount and unit then it should be a measurement. 4) Devices include implantable objects (e.g. pacemakers, stents, artificial joints), medical equipment and supplies (e.g. bandages, crutches, syringes), other

	instruments used in medical procedures (e.g. sutures, defibrillators) and material used in clinical care (e.g. adhesives, body material, dental material, surgical material). 5) Measurement contains both orders and results of such Measurements as laboratory tests, vital signs, quantitative findings from pathology reports, etc. OBSERVATION captures clinical facts about a Person obtained in the context of examination, questioning or a procedure. Any data that cannot be represented by any other domains, such as social and lifestyle facts, medical history, family history, etc. are recorded here. 6) Observations differ from Measurements in that they do not require a standardized test or some other activity to generate clinical fact. Typical observations are medical history, family history, the stated need for certain treatment, social circumstances, lifestyle choices, healthcare utilization patterns, etc. 7) Demographic can include factors of patient such as age, gender, race, ethnicity, education level, income, occupation, geographic location, marital status, and family size. Age term can be demographic but the specific age criteria should be annotated as value. 8) Negation_cue includes all information that negates clinical concepts. 9) Value is the numeric value or string test result of clinical concepts. Typicall values can be the results of Measurements such as Lab test, vital signs, and quantitative findings from pathology reports. It can also be the dosage of drugs, the frequency of drugs, positive/negative of Gene test or lab test, the duration of drugs or numeric criteria of age, weight, height, etc.
Concept extraction (Our modified version)	Annotate clinical concepts from the given text using the following rules: 1) Annotate concepts with the domains 'Demographic', 'Condition', 'Device', 'Procedure', 'Drug', 'Measurement', 'Observation', 'Visit', 'Value', 'Negation_cue', 'Temporal', and 'Quantity'. If you cannot annotate with the given domains, you can name a new one (e.g., Drug_cycle, Visit, Provider, etc.). 2) Split the concepts as detail as possible. Each concept can be annotated only once with a single domain. 3) Normalize clinical abbreviation and acronyms and attached behind the original abbreviation with parenthesis. 4) Return your response under [Annotation] section. Following is not allowed examples: 1) <Measurement>EGFR <Value>triple postive</Value></Measurement> 2) <Condition>Hypertension, diabetes, heart failure, and dementia</Condition> 3) <Observation>allergy</Observation> to <Drug>X</Drug>, <Drug>Y</Drug>, or <Drug>Z</Drug> Below is allowed examples: 1) <Measurement>EGFR</Measurement> <Value>triple positive</Value>

	2) <Condition>hypertension</Condition>, <Condition>T2DM (Type 2 Diabetes Mellitus)</Condition>, <Condition>heart failure</Condition>, and <Condition>dementia</Condition> 3) Patient <Demographic>aged<Demographic> > <Value>65 years old</Value> 4) <Drug>Metformin</Drug> <Value>500 mg</Value> <Temporal>daily</Temporal> 5) <Observation>allergy to X</Observation>, <Observation>allergy to Y</Observation>, or <Observation>allergy to Z</Observation> Following is information for each domain: 1) Condition is events of a Person suggesting the presence of a disease or medical condition stated as a diagnosis, a sign, or a symptom, which is either observed by a Provider or reported by the patient. 2) Drugs include prescription and over-the-counter medicines, vaccines, and large-molecule biologic therapies. Radiological devices ingested or applied locally do not count as Drugs. 3) Procedure is records of activities or processes ordered by, or carried out by, a healthcare provider on the patient with a diagnostic or therapeutic purpose. Lab tests are not a procedure, if something is observed with an expected resulting amount and unit then it should be a measurement. 4) Devices include implantable objects (e.g. pacemakers, stents, artificial joints), medical equipment and supplies (e.g. bandages, crutches, syringes), other instruments used in medical procedures (e.g. sutures, defibrillators) and material used in clinical care (e.g. adhesives, body material, dental material, surgical material). 5) Measurement contains both orders and results of such Measurements as laboratory tests, vital signs, quantitative findings from pathology reports, etc. OBSERVATION captures clinical facts about a Person obtained in the context of examination, questioning or a procedure. Any data that cannot be represented by any other domains, such as social and lifestyle facts, medical history, family history, etc. are recorded here. 6) Observations differ from Measurements in that they do not require a standardized test or some other activity to generate clinical fact. Typical observations are medical history, family history, the stated need for certain treatment, social circumstances, lifestyle choices, healthcare utilization patterns, etc. 7) Demographic can include factors of patient such as age, gender, race, ethnicity, education level, income, occupation, geographic location, marital status, and family size. Age term can be demographic but the specific age criteria should be annotated as value. Demographic term only includes the above factors, the word as 'patients' or 'patient' should not be annotated. 8) Some demographic factors are not explicitly included in the text, such as 'patients who are at least 18 years old' or 'less than 18 years old'. In such cases, the factor 'age' should be additional annotated. 9) Negation_cue includes all information that negates clinical concepts.

	10) Value is the numeric value or string test result of clinical concepts. Typicall values can be the results of Measurements such as Lab test, vital signs, and quantitative findings from pathology reports. It can also be the dosage of drugs, the frequency of drugs, positive/negative of Gene test or lab test, the duration of drugs or numeric criteria of age, weight, height, etc. The relational operator such as "<", ">", ">=", "<=" should be recognized as part of the VALUE information.

Supplementary Table 2. Performance comparison on entity parsing on the use case of effect of hydrocortisone in septic shock patients

Use Cases	Model	Precision (%)	Recall (%)	F1 (%)
Effect of Hydrocortisone in Septic Shock Patients	GPT-4o	**86.8**	**95.7**	**91.1**
	Phi 4	76.3	77.9	77.1
	DeepSeek-R1	79.0	77.6	78.3
	Gemma 3	80.7	83.6	82.1

Supplementary Table 3. Identified error types and definitions from the generated SQL query.[23]

Error type	**Definition**
Concept missing	Concept missing error means a clinical concept is detected at the concept extraction stage but not used in query formulation.
Functional misuse error	Functional misuse error means that SQL functions are incorrectly used in the query and eventually cause the error.
Incorrect concept mapping	Incorrect concept mapping means that extracted clinical concepts are mapped to the wrong concept IDs.
Integrity constraint violation	Integrity constraint violation means that the query violates the data consistency, such as inserting string value into the columns expecting integer concept IDs.
Logic error	Logic error means that Informatician incorrectly defines the logic, relation, or temporality of the concepts in the query.
Schema reference error	Schema reference error means that Informatician incorrectly referenced OMOP-CDM tables or columns.
Syntax error	Syntax error is an error that has incorrect SQL, syntax such as misspelled words or incorrect use of punctuation in the query.

Supplementary Table 4. The frequency of seven categories of SQL errors across five clinical trials for each evaluated LLM.

Trial ID	**LLM**	**Concept Missing**	**Functional Misuse Error**	**Incorrect Concept Mapping**	**Integrity Constraint Violation**	**Logic Error**	**Schema Reference Error**	**Syntax Error**
NCT00475852	GPT-4o	1	0	1	0	0	0	1
	Phi-4	1	1	6	0	0	0	6
	DeepSeek-R1	3	0	9	0	1	0	7
	Gemma 3	5	1	5	1	1	3	8
NCT02856698	GPT-4o	0	0	1	0	0	0	0
	Phi-4	0	1	3	0	0	0	2
	DeepSeek-R1	1	1	4	0	0	0	3
	Gemma 3	1	1	2	0	0	0	2

NCT03872011	GPT-4o	0	0	2	0	0	0	3
	Phi-4	1	0	7	0	0	0	7
	DeepSeek-R1	1	0	6	0	0	0	7
	Gemma 3	0	0	8	0	0	0	3
NCT04134403	GPT-4o	0	0	0	0	0	0	1
	Phi-4	2	0	2	0	0	0	4
	DeepSeek-R1	3	0	4	0	0	0	4
	Gemma 3	2	0	5	0	0	0	3
NCT06091982	GPT-4o	0	0	0	0	0	0	2
	Phi-4	0	0	1	0	1	0	4
	DeepSeek-R1	1	0	2	0	0	0	3
	Gemma 3	2	0	2	0	0	0	3

Supplementary Table 5. Evaluation of Statistician's ability to obtain accurate EC importances after optimization pipeline.

Number of Rules	**Mean Absolute Error**
2	0.220465
4	0.147269
6	0.210378
8	0.102421
10	0.079794
12	0.059300
14	0.039253
16	0.053420
18	0.053167
20	0.042415