

From Study Design to Executable Code: Automating OHDSI Study Implementation with Large Language Models

Hanjae Kim^{1,2}, Min Seong Kim¹, Seonji Kim^{1,2}, Seng Chan You^{1,2,3}

¹ Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Republic of Korea,

² Institute for AI and Social Innovation, Yonsei University, Seoul, Republic of Korea

³ Yonsei Institute for Digital Health, Yonsei University, Seoul, Republic of Korea

Background

Although the Observational Health Data Sciences and Informatics (OHDSI) provides standardized analytic tools for reproducible network studies, translating conceptual study designs into executable workflows can still require substantial familiarity with OHDSI-specific analytic conventions and R programming. Large language models (LLMs) have shown potential for bridging natural language and structured computation across various domains, and we hypothesized that the standardized nature of OMOP CDM and the well-defined structure of Strategus¹ analysis specifications make the OHDSI ecosystem suitable for LLM-assisted automation. We therefore developed the Text-guided Health-study Estimation and Specification Engine Using Strategus (THESEUS) as a framework using a two-step process: LLM-assisted standardization of free-text study descriptions into OHDSI-compatible analytic specifications, followed by generation of executable Strategus scripts.

Methods

Conceptual framework design

THESEUS uses a two-stage architecture consisting of standardization and code generation. In the standardization step, an LLM converts free-text study descriptions into structured JSON specifications using an OHDSI-compatible schema and field-level guidance for key settings. This schema encompasses the items required to specify a Cohort Method analysis and allows multiple entries for study period, time-at-risk (TAR), propensity score (PS) adjustment, and outcome model settings, enabling sensitivity analyses within the same structured specification. The validated JSON is then converted into Strategus R scripts through deterministic rule-based logic. This rule-based mapping is possible because HADES² modules are standardized and Strategus scripts follow recurring implementation patterns.

Evaluation of the framework

Previously published papers of comparative effectiveness studies conducted within the OHDSI community were used as natural language samples to evaluate the framework's ability to standardize free-text descriptions into structured analytic JSON specifications. To assess the generalizability of the framework beyond the OHDSI community, we additionally used non-OHDSI study papers. We manually extracted texts describing study period, TAR, PS adjustment, and outcome model settings, and evaluated LLMs under two input conditions: (1) providing primary analysis text only for each study; and (2) providing all text related to full analyses including sensitivity analyses, which allowed the generated JSON specifications to include multiple fields within each section of the schema (e.g., multiple risk windows within the TAR section for a single study). For the primary-analysis setting,

accuracy was assessed separately for each of the four specification sections. For the full-analyses setting, we conducted field-level evaluation, treating each subfield as an independent unit and aggregating them across all studies to compute sensitivity, precision, and the number of false positives (FPs) per study.

For OHDSI studies only, we further tested augmented descriptions of the primary settings. For each study, two augmented versions were generated from the primary-analysis description by preserving the original text's structure and wording as much as possible while recombining the study design differently, enabling evaluation of model robustness under varied study-design settings.

For all evaluations, we compared a diverse set of proprietary LLMs from 4 major vendors: OpenAI (gpt-5.5, gpt-5.4-mini), Google (gemini-3.1-pro, gemini-3.1-flash-lite), Anthropic (claude-opus-4-8, claude-haiku-4-5), and DeepSeek (deepseek-v4-pro, deepseek-v4-flash).

Prototype implementation and workflow integration

To provide the framework for different research workflows, THESEUS was implemented as an ATLAS³-like graphical user interface (GUI) for human-in-the-loop review and as command-line accessible tooling that can be integrated into diverse research workflows.

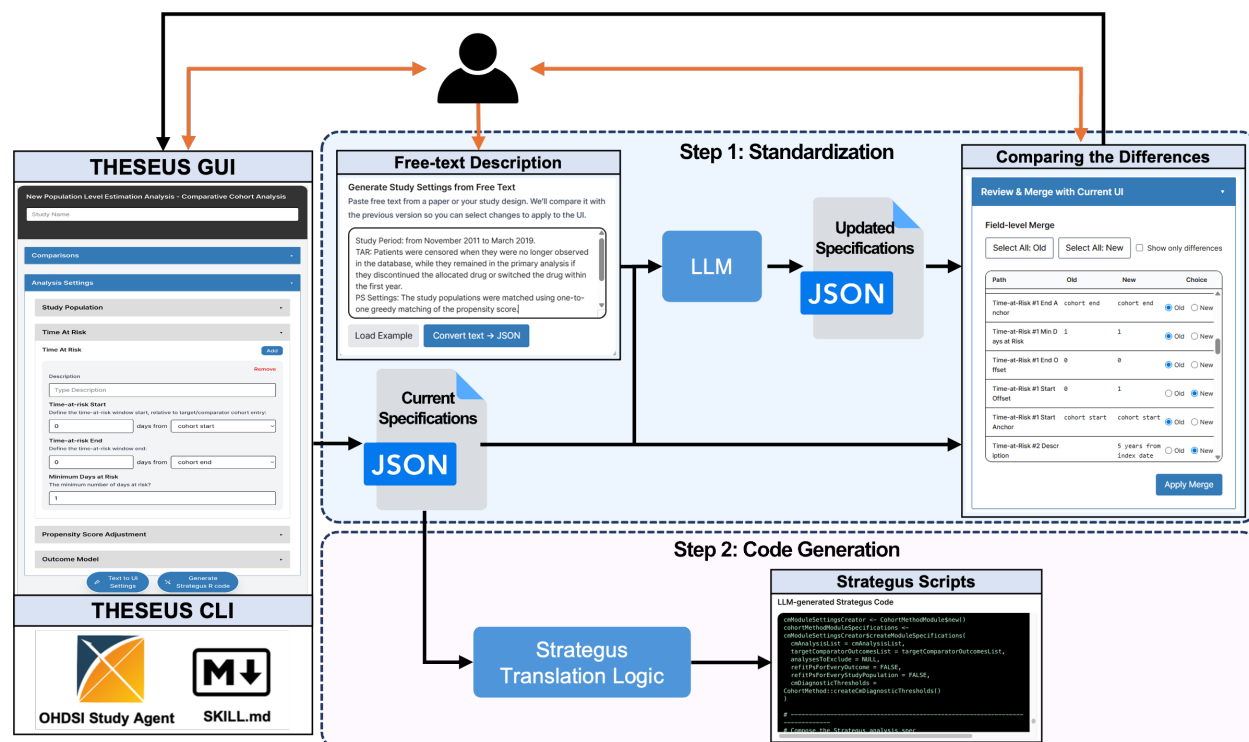


Figure 1. THESEUS as a GUI- and CLI-accessible bridge between free-text study descriptions and executable Strategus workflows. The orange colored-arrows represent the interaction between human and the system. GUI indicates graphical user interface; LLM, large language model; JSON, JavaScript Object Notation

Results

Resulting prototypes and public resources

The current prototype is available through both a web-based GUI (<https://theseus2.vercel.app/>) and CLI-oriented workflows (<https://github.com/dr-you-group/theseus-plugin.git>) (Figure 1). In the GUI workflow, users can enter

free-text study descriptions, compare proposed structured specifications with the current configuration, selectively accept changes, and export Strategus-ready code. In the CLI workflow, a Claude Code/Codex plugin exposes THESEUS functionality as Model Context Protocol tools and slash commands, supporting study description standardization, specification validation, and generation of Strategus-ready code from validated specifications. In parallel, we have also incorporated the framework into OHDSI's ongoing Study Agent initiative, an agent-style study design assistant that supports end-to-end OHDSI study workflows including Cohort Method analysis (<https://github.com/OHDSI/StudyAgent>).

Standardization Evaluation Results

15 OHDSI studies and 15 non-OHDSI studies were used for evaluating the performance of the standardization step in the framework. Under the primary analysis setting, overall accuracy within the OHDSI studies ranged from 0.83 to 0.97 across models, with gpt-5.5 and deepseek-v4-pro achieving the highest overall accuracy (gpt-5.5: study period 1.00, TAR 0.93, PS adjustment 0.93, outcome model 1.00; deepseek-v4-pro: study period 1.00, TAR 0.87, PS adjustment 1.00, outcome model 1.00) (Table 1). Within the non-OHDSI studies, overall accuracy ranged from 0.77 to 0.97, with gemini-3.1-pro achieving the highest (study period 0.93, TAR 0.93, PS adjustment 1.00, outcome model 1.00) (Table 1).

Models	Accuracy									
	OHDSI studies (N=15)					Non-OHDSI studies (N=15)				
	SP	TAR	PS	OM	Overall	SP	TAR	PS	OM	Overall
GPT-5.5	1.00	0.93	0.93	1.00	0.97	0.93	0.87	1.00	1.00	0.95
GPT-5.4-mini	1.00	0.73	0.60	1.00	0.83	0.87	0.33	0.87	1.00	0.77
Gemini-3.1-Pro	1.00	0.87	0.93	1.00	0.95	0.93	0.93	1.00	1.00	0.97
Gemini-3.1-Flash-lite	1.00	0.73	0.93	1.00	0.92	0.87	0.40	0.93	0.93	0.78
Claude-opus-4.8	1.00	0.80	0.93	1.00	0.93	0.93	0.60	1.00	0.93	0.87
Claude-haiku-4.5	1.00	0.67	0.93	1.00	0.90	0.93	0.60	1.00	1.00	0.88
DeepSeek-V4-pro	1.00	0.87	1.00	1.00	0.97	0.87	0.80	0.93	1.00	0.90
DeepSeek-V4-flash	1.00	0.80	0.93	1.00	0.93	0.93	0.67	0.93	1.00	0.88

Table 1. Study-level standardization accuracy across OHDSI and non-OHDSI studies under the primary analysis condition. OHDSI indicates Observational Health Data Sciences and Informatics; SP, Study Period; TAR, Time-at-Risk; PS, Propensity Score (adjustment); OM, Outcome Model

At the field-level evaluation for the full-analyses setting, precision ranged from 0.75 to 0.99, sensitivity from 0.72 to 0.97, and FPs per study from 0.07 to 1.53 in OHDSI studies (Table 2). Gpt-5.5 and claude-opus-4.8 achieved the strongest field-level performance (precision 0.99, sensitivity 0.97, false positives per study 0.07). In non-OHDSI studies, precision ranged from 0.75 to 0.92, sensitivity from 0.75 to 0.91, and false positives per study from 0.40 to 1.27. Gpt-5.5 and gemini-3.1-pro achieved the strongest field-level performance (precision 0.92, sensitivity 0.91, false positives per study 0.40).

In the augmented primary setting, two augmented texts were generated per OHDSI study, yielding 30 augmented descriptions. Overall accuracy ranged from 0.82 to 1.00, with gpt-5.5 achieving perfect accuracy in all four specification sections.

Models	OHDSI studies (n=99)			Non-OHDSI studies (n=83)		
	Precision	Sensitivity	FPs per study	Precision	Sensitivity	FPs per study
GPT-5.5	0.99	0.97	0.07	0.92	0.91	0.40
GPT-5.4-mini	0.75	0.72	1.53	0.75	0.75	1.27
Gemini-3.1-Pro	0.94	0.94	0.40	0.92	0.91	0.40
Gemini-3.1-Flash-lite	0.87	0.81	0.73	0.77	0.77	1.20
Claude-opus-4.8	0.99	0.97	0.07	0.83	0.81	0.87
Claude-haiku-4.5	0.92	0.88	0.47	0.83	0.82	0.87
DeepSeek-V4-pro	0.95	0.93	0.33	0.89	0.88	0.53
DeepSeek-V4-flash	0.91	0.87	0.53	0.84	0.83	0.80

Table 2. Field-level standardization accuracy across OHDSI and non-OHDSI studies under the multiple analysis condition. OHDSI indicates Observational Health Data Sciences and Informatics; FPs, False Positives

Conclusion

THESEUS suggests that LLMs can support comparative effectiveness study workflows most effectively when used to interpret and standardize study descriptions within a constrained analytic framework, rather than to generate analytic code in an unconstrained manner. Our evaluation of the standardization step showed promising performance across OHDSI and non-OHDSI study samples, while also indicating that complex descriptions of TAR specifications remain challenging. When sensitivity analyses were included, overall field-level performance showed high sensitivity with low false positive rates (less than one per study, except for gpt-5.3-mini).

Although the current evaluation was limited by the number of studies and focused on Cohort Method specifications, pairing LLM-assisted standardization with structured schemas and deterministic Strategus code generation may help lower technical barriers to OHDSI-based studies while preserving reproducibility and transparency. Further validation in real-world network study settings will be needed before THESEUS can be considered a mature production tool.

References

1. Sena AG, Knoll C, Gilbert J, Reps J, DeFalco F, Blacketer C, et al., editors. Strategus: Marching towards transparent, reproducible research. 2022 OHDSI Collaborator Showcase; 2022. Available from: https://www.ohdsi.org/wp-content/uploads/2022/10/13-Anthony-Sena_Strategus_2022symposium-Anthony-Sena.pdf
2. Schuemie M, Reps J, Black A, Defalco F, Evans L, Fridgeirsson E, et al. Health-Analytics Data to Evidence Suite (HADES): Open-Source Software for Observational Research. *Stud Health Technol Inform*. 2024;310:966–70.
3. OHDSI [Internet]. Software Tools. [cited June 5, 2026]. Available from: <https://www.ohdsi.org/software-tools/>