THESEUS: An LLM-powered Research Assistant Bridging the OHDSI Ecosystem for OMOP-CDM Observational Studies

PRESENTER: Hanjae Kim

INTRODUCTION

Background

- OHDSI supports a comprehensive ecosystem of open-source analytics tools
  - ATLAS: a GUI tool for designing OHDSI studies
  - HADES: a collection of R packages commonly used to ensure reproducibility in multi-institutional studies
  - Strategus: a single R package that orchestrates HADES modules
- The current best practice involves **defining cohorts in ATLAS and configuring the remaining components with Strategus**
- However, there is currently no system that automatically converts study designs defined in ATLAS into executable Strategus R scripts

Objective

- To develop an LLM-powered software designed to assist researchers **by automatically translating ATLAS-style designs into Strategus R scripts**

METHODS

Development of a Software Prototype: THESEUS

"Text-guided Health-study Estimation and Specification Engine Using Strategus"

THESEUS: a prototype GUI that resembles the 'population-level estimation' tab of ATLAS with two additional key functionalities (Figure 1):

1. **Text2JSON**: converting free-text into GUI-based analysis specifications with human-in-the-loop approach
2. **JSON2Strategus**: converting GUI-based analysis specifications into Strategus scripts

Evaluation of the LLM Modules

Data: 15 published target trial emulation study paper

Text2JSON

- Key idea: assess the module's ability to configure 3 study sections - (1) study period, (2) time-at-risks (TAR), (3) propensity score (PS) adjustment - from the papers
- 3 input conditions:
  1) Primary analysis text only
  2) Full protocol text (all texts related to any possible analyses, including sensitivity analyses)
  3) Full methods section
- Outputs: study specifications in JSON format
- Metrics:
  - Primary analysis: accuracy for each section
  - Full protocol & Full methods:
    * a single section could contain several subfields
    - (Section-level) accuracy* for each section
      * considered accurate if all subfields are correct
    - (Field-level*) sensitivity, precision, FPs per study
      * treating each subfield of all sections as an independent unit

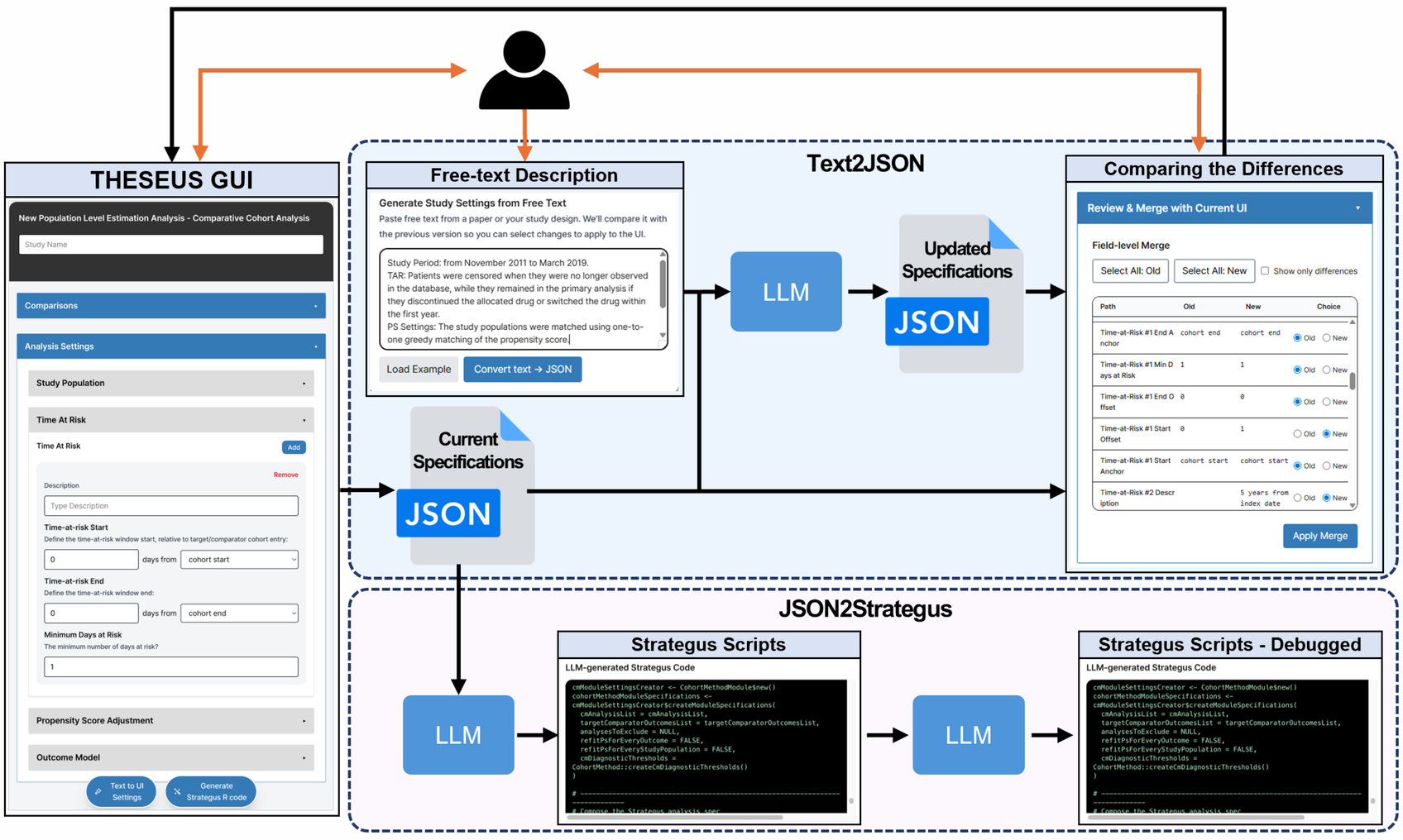
JSON2Strategus

- Inputs: gold standards from the Text2JSON evaluation
- Outputs:
  1) Initial R scripts based on Strategus
  2) R scripts after the debugging step
- Metrics: accuracy*
  * considered accurate if executed without an error

LLMs:

- OpenAI: GPT-5, GPT-5-mini
- Google: Gemini-2.5-Pro, Gemini-2.5-Flash
- Anthropic: Claude-sonnet-4-5, Claude-haiku-4-5

LLMs can bridge the gap between natural language study designs, GUI-based configuration, and executable code scripts within the OHDSI ecosystem



THESEUS - App

Demo Video

Text2JSON

- The module takes text descriptions of study designs together with the current JSON representation of the GUI as inputs, and generates an updated JSON specification
- The system provides a side-by-side comparison of the original and revised specifications, allowing users to selectively accept or reject modifications

JSON2Strategus

- The module takes the JSON representation of GUI as inputs and generates Strategus R scripts
- Additional debugging step can correct errors when the initial script fails

Figure 1. Overall workflow of THESEUS powered by LLMs with human-in-the-loop configuration.
The orange colored-arrows represent the interaction between human and the system.

Results

Table 1. Section-level evaluation results of the Text2JSON

Models	Accuracy											
	Primary analysis				Full protocol				Full methods			
	SP (n=10)	TAR (n=15)	PS (n=15)	Overall	SP (n=10)	TAR (n=15)	PS (n=15)	Overall	SP (n=10)	TAR (n=15)	PS (n=15)	Overall
GPT-5	**1.00**	0.67	**1.00**	0.89	0.80	0.60	**0.93**	0.78	0.70	0.67	0.87	**0.75**
GPT-5-mini	0.90	0.73	0.80	0.81	0.90	0.40	0.33	0.54	0.60	0.27	0.53	0.47
Gemini-2.5-Pro	**1.00**	**0.80**	**1.00**	**0.93**	0.90	**0.80**	0.80	**0.83**	0.60	0.67	0.73	0.67
Gemini-2.5-Flash	**1.00**	0.73	**1.00**	0.91	0.90	0.60	0.73	0.74	0.70	**0.73**	0.67	0.70
Claude-sonnet-4-5	**1.00**	0.53	**1.00**	0.84	0.90	0.60	**0.93**	0.81	0.70	0.33	0.67	0.57
Claude-haiku-4-5	**1.00**	0.67	**1.00**	0.89	**1.00**	0.67	0.80	0.82	**0.80**	0.53	**0.93**	**0.75**

SP indicates study period; TAR: time-at-risk; PS: Propensity score adjustment

Table 2. Field-level evaluation results of the Text2JSON

Models	Full protocol (n=84)			Full methods (n=84)		
	Precision	Sensitivity	FP per study	Precision	Sensitivity	FP per study
GPT-5	0.86	0.86	0.73	0.88	0.84	**0.6**
GPT-5-mini	0.6	0.59	2.07	0.77	0.61	0.93
Gemini-2.5-Pro	**0.92**	**0.91**	**0.4**	**0.92**	**0.88**	1.13
Gemini-2.5-Flash	0.88	0.84	0.6	0.91	**0.88**	1.27
Claude-sonnet-4-5	0.86	0.87	0.73	0.86	0.84	1.93
Claude-haiku-4-5	0.91	0.86	0.47	0.88	0.86	1.8

FP indicates false positive

Figure 2. Evaluation results of the JSON2Strategus

Conclusion

- This study suggests an LLM-powered system that bridges the gap between natural language study designs, GUI-based configuration, and executable code scripts within the OHDSI ecosystem
- The current prototype is limited to population-level estimation, but it can be expanded to support a broader range of study designs

Hanjae Kim[1,2], Min Seong Kim[2], Seng Chan You[1,3]

[1]Department of Biomedical Systems Informatics, Yonsei University College of Medicine

[2]Department of Material Sciences and Engineering, Yonsei University College of Engineering

[3]Yonesi Institute for Digital Health, Yonsei University