Autonomous Phenotyping for Agentic EHR-Based Clinical Research Workflows

Project Proposal (Week 4)

Irith Katiyar Benet Fité Abril

Instructions (do not delete this box). This proposal is at most **2 pages of main text** (single-spaced, 12pt font, 1" margins). You may use a **3rd page for figures/tables** and have **unlimited references**. Keep the section headings below and replace the placeholder text with your own writing.

Hypothesis or Research Question

This project aims to answer the research question: Can multimodal and longitudinal phenotyping algorithms, namely KOMAP and LATTE, be operationalized in a phenotyping-as-a-service framework via an MCP server to enable autonomous and reproducible end-to-end EHR-based cohort studies, particularly in cases where codified EHR data alone does not reliably capture true disease status?

Background and Significance

In EHR-based studies, defining accurate patient cohorts is often challenging because diagnostic codes alone often do not reflect true clinical phenotypes. For example, in an MGB cohort study, only 58% of patients with multiple ICD codes for rheumatoid arthritis were confirmed as actual cases upon clinical review [1]. There have been numerous developments in EHR phenotyping algorithms to improve the accuracy of cohort selection, including MAP [2], which generates weakly supervised phenotypes from structured data; KOMAP [1], which builds on MAP by automatically selecting features from both structured data and clinical notes via online narrative and codified feature search engine (ONCE); and LATTE [3], which performs incident phenotyping by annotating the timing of clinical events. These tools allow for robust phenotyping without human-labeled, gold-standard sets.

In recent months, M4 [4] has emerged as an MCP server that allows agents to autonomously query EHR datasets such as MIMIC-IV [5] and eICU [6] and perform cohort selection and basic statistical analyses. Cohort selection with M4 currently defaults to simple code-based heuristics (Figure 1), and attempts to use LLMs to draft phenotyping algorithms have shown insufficient logical precision for accurate cohorts [7]. While KOMAP and LATTE provide robust cohort definitions, using them currently requires manual configuration and computational expertise in their libraries. This project will build an MCP server that integrates these tools into an autonomous end-to-end clinical research pipeline based on M4, enabling accurate and reproducible cohort selection and downstream analyses.

Dataset

The project will utilize MIMIC-IV [5], a comprehensive EHR repository from Beth Israel Deaconess Medical Center that contains de-identified longitudinal data for approximately 365,000 patients and 331,000 clinical notes, which are essential for multimodal phenotyping that relies on extracting clinical concepts (CUIs) from free text. Any data preprocessing steps required by the phenotyping algorithms (e.g., rolling up ICD codes to Phecodes, extracting summary covariance matrices, etc.) will be programmed as tools and skills in our infrastructure so the agent can use them as needed.

Preprint.

Baseline Model

The baseline for this project is the current M4 system, which performs cohort selection using simple code-based heuristics without access to advanced phenotyping algorithms. This baseline is reasonable because it is representative of the autonomous workflows that our MCP server aims to enhance.

Proposed Methodology

This project will develop an end-to-end clinical research agent, with its main contribution being a novel MCP server that exposes KOMAP and LATTE as callable tools for LLM-based agentic workflows that can benefit from EHR phenotyping. Claude Code will be used as the orchestrator that manages the agentic workflow and coordinates interactions across multiple servers. It will integrate directly with M4, which serves as the data access layer for automatically querying MIMIC-IV. This architecture enables a modular data flow where Claude Code retrieves multimodal and longitudinal patient features from M4's tools and then pipes them into our server's tools for automated training and inference of weakly supervised phenotyping models. Each phenotyping algorithm is registered as a tool within the FastMCP framework and has a defined profile, enabling Claude Code to select the appropriate tool based on the user's query. For example, Claude Code should invoke LATTE for phenotypes involving temporality, such as the onset of a disease. Agent Skills will define the preprocessing steps needed to go from raw tabular data and notes to the opinionated input formats of these tools. These include extracting CUIs from notes, mapping ICD codes to standardized vocabularies, and validating input schemas to ensure correctly structured tool invocations. By exposing KOMAP and LATTE as MCP tools, these high-throughput libraries are transformed into interoperable modules that enable agents to execute complex phenotyping logic dynamically.

Training will just involve the automated calibration of KOMAP/LATTE where they learn from silver-standard surrogates. Evaluation will involve an independent LLM-as-a-judge to analyze the agent's logged traces to audit its clinical reasoning, tool selection accuracy, and technical correctness. We will further validate the agent by tasking it with reproducing cohorts or basic results from prior research studies on diseases like RA with not as clear-cut phenotypes. Note that our goal is not to evaluate the performance of the phenotyping algorithms themselves but to demonstrate that they can be integrated into an agentic workflow.

Ultimately, this project will deliver a scalable, open-source infrastructure for agentic clinical research, with the novel contribution of making state-of-the-art phenotyping algorithms accessible to LLM-based agents for the first time. Beyond autonomous cohort selection, this system can enable advanced downstream analyses and surveillance systems, such as a clinical coding drift detector that monitors changes in coding practices over time or other tools for incident disease monitoring.

Resources

Implementation will use Python, FastMCP, Claude Code, the existing KOMAP and LATTE phenotyping libraries, a Python-R bridge for KOMAP, and Agent SDK or LangChain if Claude Code alone is insufficient for building our agentic pipeline (see below).

Challenges and Contingency Plans

If integration challenges arise, we will substitute MAP for KOMAP and, if necessary, restrict the scope to non-longitudinal phenotyping rather than adding LATTE. If vanilla Claude Code proves insufficient for stable multi-step orchestration, we will transition to Anthropic's Agent SDK or LangGraph to enable tailored planning. To mitigate context rot in long research sessions, we will implement memory optimization agents for selective storage and retrieval of information, and summarization agents for compressing intermediate reasoning to maintain stable context windows. Performance limitations of the underlying phenotyping algorithm will be treated as out of scope, as this project focuses on the agentic orchestration and infrastructure rather than improving the phenotyping method itself.

References

[1] Xin Xiong, Sara Morini Sweet, Molei Liu, Chuan Hong, Clara-Lea Bonzel, Vidul Ayakulangara Panickan, Doudou Zhou, Linshanshan Wang, Lauren Costa, Yuk-Lam Ho, Alon Geva, Kenneth D. Mandl, Suchun Cheng, Zongqi Xia, Kelly Cho, J. Michael Gaziano, Katherine P. Liao, Tianxi Cai, and Tianrun Cai. Knowledge-driven online multimodal automated phenotyping system. Pages: 2023.09.29.23296239.

[2] Katherine P. Liao, Jiehuan Sun, Tianrun A. Cai, Nicholas Link, Chuan Hong, Jie Huang, Jennifer E. Huffman, Jessica Gronsbell, Yichi Zhang, Yuk-Lam Ho, Victor Castro, Vivian Gainer, Shawn N. Murphy, Christopher J. O'Donnell, J. Michael Gaziano, Kelly Cho, Peter Szolovits, Isaac S. Kohane, Sheng Yu, Tianxi Cai, and with the VA Million Veteran Program. High-throughput multimodal automated phenotyping (map) with application to phewas. *bioRxiv*, 2019.

[3] Jun Wen, Jue Hou, Clara-Lea Bonzel, Yihan Zhao, Victor M. Castro, Vivian S. Gainer, Dana Weisenfeld, Tianrun Cai, Yuk-Lam Ho, Vidul A. Panickan, Lauren Costa, Chuan Hong, J. Michael Gaziano, Katherine P. Liao, Junwei Lu, Kelly Cho, and Tianxi Cai. LATTE: Label-efficient incident phenotyping from longitudinal electronic health records. 5(1):100906.

[4] Rafi Al Attrach, Pedro Moreira, Rajna Fani, Renato Umeton, and Leo Anthony Celi. Conversational LLMs simplify secure clinical data access, understanding, and analysis. original-date: 2025-10-27T18:02:44Z.

[5] Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. 10(1):1.

[6] Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. The eICU collaborative research database, a freely available multi-center database for critical care research. 5(1):180178.

[7] Chao Yan, Henry H Ong, Monika E Grabowska, Matthew S Krantz, Wu-Chen Su, Alyson L Dickson, Josh F Peterson, QiPing Feng, Dan M Roden, C Michael Stein, V Eric Kerchberger, Bradley A Malin, and Wei-Qi Wei. Large language models facilitate the generation of electronic health record phenotyping algorithms. 31(9):1994–2001. _eprint: https://academic.oup.com/jamia/article-pdf/31/9/1994/59920040/ocae072.pdf.

Figures

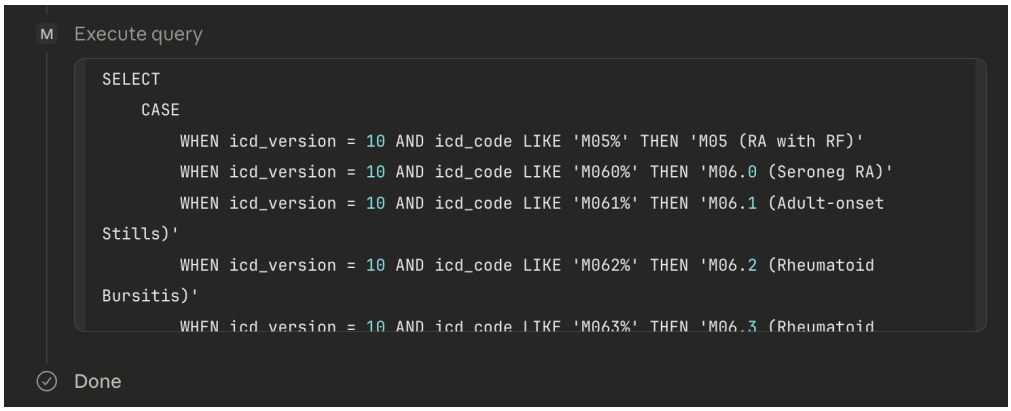

Figure 1: SQL query result after prompting Claude Code integrated with M4 for MIMIC-IV: "Create a cohort of patients with rheumatoid arthritis." This illustrates that M4 as-is relies on diagnostic billing codes (ICD-based) filtering for cohort selection, which has proven to be incomplete for RA.