Cohort-Pilot: Collaborative AI for Translating Natural Language into Actionable Cohort Analytics

Alvaro A. Alvarez[1,2], Farnoosh Sheikhi[1,2], Priya Desai[1,2]
[1] Stanford Health Care , [2]Stanford School of Medicine

Background

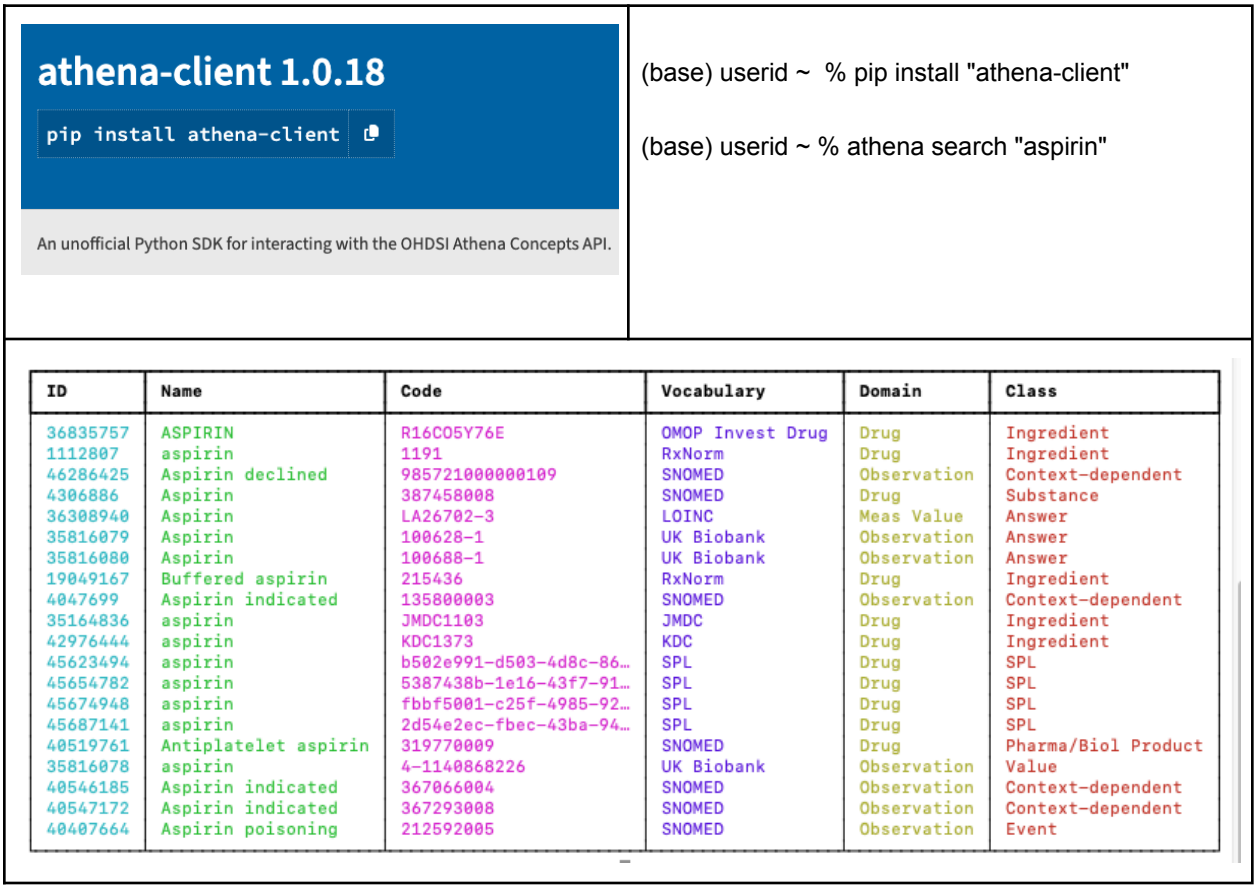
Existing tools within the OHDSI ecosystem, such as ATLAS and CohortGenerator, facilitate cohort construction but often require extensive manual concept curation and SQL adjustments tailored for local environments [2,3]. Similarly, commercial platforms often lack seamless end-to-end capabilities, forcing researchers to rely on iterative manual searches, scripting, and manual SQL editing, which leads to inefficiencies and inaccuracies [2].

Building on the feedback from of our recent research software demo, "Bridging the Language Gap: Generative Models for Efficient Medical Concept Discovery,"(#OHDSI2024 Collaborator Showcase honorees [5]), we aim to expand our contributions by addressing more complex cohort-building workflows.We present **Cohort-Pilot**, an innovative pipeline that collaboratively translates natural language cohort definitions into executable, validated BigQuery SQL. Cohort-Pilot emphasizes user-driven interactions and iterative refinements, merging human expertise with AI-driven automation to improve accuracy, transparency, and flexibility in cohort construction.
We believe **Cohort-Pilot** will enhance the user experience with OMOP while enabling real-time domain coverage profiling, allowing users to quickly identify data completeness across OMOP domains.

Methods

Cohort-Pilot utilizes structured orchestration through Pydantic models to enhance its functionality. It features a multi-agent architecture where specialized AI agents manage tasks such as natural language interpretation, SQL generation, syntax correction, and heuristic validation. The system employs a typed pipeline context for comprehensive management of pipeline states, including concept mappings, SQL drafts, and user feedback. An iterative refinement loop consisting of cycles of: Generate, Correct, and Review ensures continuous improvement in output quality. Additionally, modular service abstraction allows for clearly defined database operations and concept vocabulary management, promoting modularity and ease of maintenance.

To address the lack of an official API for Athena, we have developed and publicly released *athena-client* (Figure 1), which allows precise and iterative exploration of OMOP standard concept hierarchies, which is critical for the accuracy of Cohort-Pilot (https://github.com/aandresalvarez/bqcohortgen). In addition, the Cohort-Pilot facilitates dynamic SQL generation by integrating validated concepts into pre-optimized SQL templates specifically designed for the OMOP CDM v5.4 in BigQuery, while also utilizing domain-coverage queries to produce comprehensive person-level indicators across core OMOP domains.



ID	Name	Code	Vocabulary	Domain	Class
36835757	ASPIRIN	R16CO5Y76E	OMOP Invest Drug	Drug	Ingredient
1112807	aspirin	1191	RxNorm	Drug	Ingredient
46286425	Aspirin declined	985721000000109	SNOMED	Observation	Context-dependent
4306886	Aspirin	387458008	SNOMED	Drug	Substance
36308940	Aspirin	LA26702-3	LOINC	Meas Value	Answer
35816079	Aspirin	100628-1	UK Biobank	Observation	Answer
35816080	Aspirin	100688-1	UK Biobank	Observation	Answer
19049167	Buffered aspirin	215436	RxNorm	Drug	Ingredient
4047699	Aspirin indicated	135800003	SNOMED	Observation	Context-dependent
35164836	aspirin	JMDC1103	JMDC	Drug	Ingredient
42976444	aspirin	KDC1373	KDC	Drug	Ingredient
45623494	aspirin	b502e991-d503-4d8c-86…	SPL	Drug	SPL
45654782	aspirin	5387438b-1e16-43f7-91…	SPL	Drug	SPL
45674948	aspirin	fbbf5001-c25f-4985-92…	SPL	Drug	SPL
45687141	aspirin	2d54e2ec-fbec-43ba-94…	SPL	Drug	SPL
40519761	Antiplatelet aspirin	319770009	SNOMED	Drug	Pharma/Biol Product
35816078	aspirin	4-1140868226	UK Biobank	Observation	Value
40546185	Aspirin indicated	367066004	SNOMED	Observation	Context-dependent
40547172	Aspirin indicated	367293008	SNOMED	Observation	Context-dependent
40407664	Aspirin poisoning	212592005	SNOMED	Observation	Event

Figure 1: Athena Client Package Search Function

Current capabilities include comprehensive support for OMOP domains, rapid and reliable generation of executable SQL queries with detailed data completeness profiles, user-centered iterative refinement for high accuracy and relevance, and flexible AI model configuration optimized with OpenAI GPT-4o, adaptable for other models like Google Gemini, LLaMA, or local configurations. Ongoing development and validation efforts aim to enhance cohort definition processes, leveraging iterative refinement and collaborative interactions to gather extensive community feedback for further validation and refinement of the system's functionalities (Figure 2).

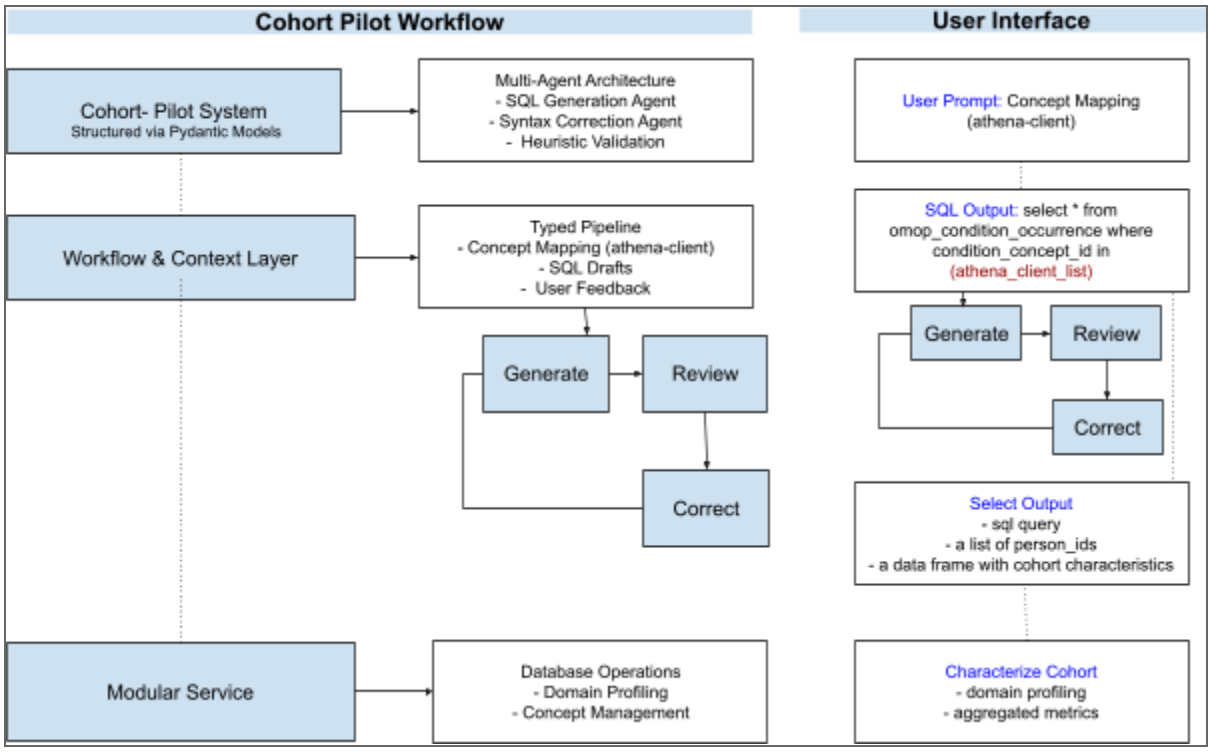


Figure 2: Intended Cohort Pilot Workflow

Results

Cohort-Pilot, currently in active development and validation stages, demonstrates significant potential in enhancing cohort definition processes. The implementation of task specialization through specialized AI agents has led to improved reliability and overall system performance. Iterative refinement, facilitated by explicit feedback loops, has dramatically enhanced the quality of outputs. Additionally, structured state management has streamlined pipeline management and debugging processes. Furthermore, providing aggregated outputs with detailed contextual information has significantly reduced errors and increased the relevance of results.
However, further evaluation is needed to compare Cohort-Pilot with established tools such as ATLAS to assess accuracy, efficiency, and completeness. Cohort-Pilot aims to gather extensive community feedback to validate and further refine its functionalities through iterative refinement and collaborative interactions.

Conclusion

Building upon our successful demonstration at the OHDSI Global Symposium [5], Cohort-Pilot represents a significant advancement in collaborative cohort construction methodologies. By integrating human insight with specialized AI capabilities, Cohort-Pilot enhances precision, usability, and efficiency in observational health research. We believe it will improve the user experience with OMOP and provide high-level aggregated metrics for defined cohorts.

References

1. athena-client
2. Hripcsak G, et al. Characterizing treatment pathways at scale using the OHDSI network. *PNAS.* 2016.
3. Voss EA, et al. Feasibility and utility of applications of the OMOP CDM. *AMIA J.* 2017.
4. OHDSI Athena Vocabulary Search API Docs
5. Bridging the Language Gap: Generative Models for Efficient Medical Concept Discovery