Utility of Large Language Models for Concept Set Curation

Adit Anand, BS[1], Anna Ostropolets, MD, PhD[2], Patrick Ryan, PhD[1,2], George Hripcsak, MD, MS[1]
[1]Department of Biomedical Informatics, Columbia University, New York, NY, USA
[2]Janssen Research & Development, Raritan, NJ, USA

Background

Constructing phenotypes from electronic health records (EHR) is critical to conduct observational health research(1), as seen through initiatives such as OHDSI(2). Substantial informatics and clinical expertise are necessary to construct rule-based phenotypes(3), and iterative validation of phenotyping algorithms assists with refining a phenotype's criteria (i.e., diagnosis codes, procedures, medications)(4,5). To assist with building phenotyping algorithms, OHDSI develops a wide range of software tools; in particular, PHOEBE is a medical concept recommender system that facilitates curating clinical concepts(6). Large language models (LLMs) have been leveraged to generate rule-based phenotyping algorithms, but such work recognizes the need for a "human-in-the-loop" approach to curating medical concepts to use in phenotyping algorithms(7). In this study, we explore the utility of LLMs to curate clinical concepts for EHR phenotyping from PHOEBE recommendations.

Methods

For each phenotype, we input its name (i.e., "acute myocardial infarction") into PHOEBE(6), a concept recommender system that provides recommendations based on clinical relevance and computational methods, to produce an initial recommended list of OMOP concepts from the condition table(8). We then iteratively prompt zero-shot ChatGPT-3.5-Turbo and ChatGPT-4o(9) to generate a True/False answer for whether a recommended concept is specific to the phenotype. Concepts ChatGPT identifies as specific are included in the phenotype's concept set, which PHOEBE uses to generate a second list of recommended condition concepts. We then leverage the same prompt engineering strategy to assess these concepts' specificity and include those identified as specific in the final concept set. Table 1 details the prompt we use along with sample input.

In this study, we appraise our pipeline on 4 phenotypes: type 1 diabetes mellitus, acute myocardial infarction, pulmonary hypertension, and rheumatoid arthritis. We compute the sensitivity, specificity, and AUROC (average of sensitivity and specificity). We manually curate, with clinical insight, a collection of true positive concepts and a collection of true negative concepts that are used to compute the summary statistics. We obtain PHOEBE recommendations using the Columbia University CUMC ATLAS instance. The code used to produce the results are in the linked repository (https://github.com/adit-anand/chatgpt-concept-set-curation/tree/ohdsi-submission).

Example Malady	Example Malady Description	Example Condition	Prompt Template
Acute Myocardial Infarction	Acute myocardial infarction (AMI), commonly known as a heart attack, is a life-threatening medical emergency characterized by the sudden occlusion of a coronary artery, resulting in ischemia and necrosis of cardiac tissue. It is defined as the abrupt interruption of blood flow to a portion of the myocardium, leading to myocardial cell death and subsequent release of cardiac biomarkers such as troponin. AMI typically presents with severe chest pain or pressure, often radiating to the left arm, jaw, or back, along with accompanying symptoms such as shortness of breath, diaphoresis, nausea, and vomiting. Diagnosis is confirmed by clinical history, electrocardiography (ECG) findings indicative of ST-segment elevation or new-onset Q waves, and elevated cardiac biomarkers. Treatment involves immediate reperfusion therapy to restore blood flow to the ischemic myocardium, utilizing thrombolytics or percutaneous coronary intervention (PCI). Additional therapies include antiplatelet agents, anticoagulants, beta-blockers, angiotensin-converting enzyme (ACE) inhibitors, and statins to prevent recurrent ischemic events and reduce mortality. Prognosis varies depending on the extent of myocardial damage, timely intervention, and the presence of comorbidities. Exclusions for AMI include other causes of acute chest pain such as unstable angina, aortic dissection, and pulmonary embolism.	Atrial cardiopathy	You are a helpful medical expert. Your task is to assess whether an inputted condition is specific to an inputted malady. Specific means that if you have the condition, then you definitely have the malady, but if you have the malady, you may or may not have the condition. A condition should be considered records of 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. A description of the malady is provided to assist you, but please use your extensive medical knowledge in addition to this description when performing the task at hand. Start by providing a “yes” or “no” answer, and then explain your rationale for the answer you provide. Here is the malady: {{ENTER THE MALADY}} Here is a brief description of the malady: {{ENTER MALADY DESCRIPTION}} Here is the condition: {{ENTER CONDITION}}

Table 1. The prompt template provided to both versions of ChatGPT along with example content for each field

Results

Table 2 shows the sensitivity, specificity, and AUROC of each phenotype's corresponding concept set when using ChatGPT-3.5-Turbo and ChatGPT-4o. We observe that the pipeline using ChatGPT-4o produces concept sets with negligible sensitivity and thus insignificant AUROC for each phenotype. These low sensitivity values indicate ChatGPT-4o rejects an overwhelming majority of PHOEBE's recommendations, which consist of a mixture of phenotypically-relevant clinical concepts and unwanted clinical concepts. These findings highlight that ChatGPT-4o performs worse relative to ChatGPT-3.5-Turbo. To better understand why this discrepancy occurs, Table 3 presents 3 PHOEBE-recommended medical concepts ChatGPT-4o falsely identifies as not specific to their corresponding phenotype.

	ChatGPT-3.5-Turbo			**ChatGPT-4o**		
	Sensitivity	**Specificity**	**AUROC**	**Sensitivity**	**Specificity**	**AUROC**
Acute Myocardial Infarction	0.879	0.479	0.679	0.094	0.985	0.539
Pulmonary Hypertension	0.000	0.693	0.347	0.000	1.000	0.500
Rheumatoid Arthritis	0.941	0.430	0.686	0.185	1.000	0.593
Type 1 Diabetes Mellitus	0.952	0.684	0.819	0.000	1.000	0.500

Table 2. Sensitivity, specificity, and AUROC of the concept set generated by PHOEBE recommendations and zero-shot ChatGPT prompting for each phenotype

In Table 3, ChatGPT-4o justifies its assessment of "Type 1 diabetes mellitus uncontrolled" by emphasizing that if a patient with type 1 diabetes mellitus receives that diagnosis, no new information is revealed about the patient's state of health. This implies ChatGPT-4o places greater value in diagnoses that contribute to understanding new aspects of a patient's health. Furthermore, ChatGPT-4o explains its decision for "Bilateral deformity of hands due to rheumatoid arthritis" by stating "deformity of hands" is not exclusively caused by rheumatoid arthritis. This indicates ChatGPT-4o assesses whether new clinical information ("bilateral deformity of hands" in this case) contributes to the phenotype, as opposed to accounting for the semantic meaning of the condition concept. Finally, ChatGPT-4o rationalizes its assessment of "First myocardial infarction" by assuming it is a different concept from RA. In reality, they represent the same medical condition, which illustrates how ChatGPT fails to account for clinical concepts' semantic similarity.

Phenotype	Concept Name	ChatGPT-4o Explanation
Type 1 Diabetes Mellitus	Type 1 diabetes mellitus uncontrolled	The condition "Type 1 diabetes mellitus uncontrolled" is not specific to Type 1 Diabetes Mellitus (T1DM) in the sense that having uncontrolled T1DM does not definitively mean you have T1DM. The term "uncontrolled" simply indicates that the diabetes, which is already diagnosed as T1DM, is not well-managed or regulated, leading to poor blood glucose control. However, the presence of "uncontrolled" diabetes does not provide new diagnostic information about the type of diabetes; it only describes the state of management of an already diagnosed condition. Therefore, it does not meet the criteria of being specific to T1DM because it does not independently confirm the diagnosis of T1DM. It merely describes the status of a pre-existing diagnosis.
Rheumatoid Arthritis	Bilateral deformity of hands due to rheumatoid arthritis	While bilateral deformity of the hands can be a consequence of rheumatoid arthritis (RA), it is not specific to RA. This condition can also be seen in other forms of arthritis and joint diseases, such as osteoarthritis or psoriatic arthritis. Therefore, having bilateral deformity of the hands does not definitively mean that a person has RA, even though it is a common manifestation in advanced cases of RA.
Acute Myocardial Infarction	First myocardial infarction	A "first myocardial infarction" refers to the initial occurrence of a myocardial infarction (heart attack) in an individual. While having a first myocardial infarction means that the person has experienced an acute myocardial infarction (AMI), it is not specific to AMI in the sense that it does not exclusively indicate the presence of AMI. The term "first myocardial infarction" simply denotes the initial event of a heart attack and does not imply that every instance of AMI is a first myocardial infarction. Therefore, the condition "first myocardial infarction" is not specific to the malady "acute myocardial infarction" because it does not meet the criterion that if you have the condition, you definitely have the malady.

Table 3. The concept names and LLM-generated explanations for three phenotypically-relevant clinical concepts assessed by ChatGPT-4o to be not specific to their corresponding phenotypes

Conclusion

Using PHOEBE to recommend medical concepts and filtering these recommendations with ChatGPT-4o or ChatGPT-3.5-Turbo is a promising method to curate clinical concepts for phenotyping algorithms. One avenue to further explore is how well different prompt engineering strategies or few-shot techniques perform. Additionally, it is necessary to evaluate the phenotypes that arise from our current methodology's concept sets using established OHDSI tools. Finally, reproducing this analysis with clinical domains such as procedures and medications would assist with assessing the methodology's generalizability across the OMOP CDM.

References

1. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. 2013 Jan 1;20(1):117–21.
2. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216:574–8.
3. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. Annu Rev Biomed Data Sci. 2018 Jul 20;1(Volume 1, 2018):53–68.
4. Zheng NS, Feng Q, Kerchberger VE, Zhao J, Edwards TL, Cox NJ, et al. PheMap: a multi-resource knowledge base for high-throughput phenotyping within electronic health records. J Am Med Inform Assoc. 2020 Nov 1;27(11):1675–87.
5. Wu P, Gifford A, Meng X, Li X, Campbell H, Varley T, et al. Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. JMIR Med Inform. 2019 Nov 29;7(4):e14325.
6. Ostropolets A, Hripcsak G, Knoll C, Ryan P. PHOEBE 2.0: selecting the right concept sets for the right patients using lexical, semantic, and data-driven recommendations. [cited 2024 Jun 7]; Available from: https://www.ohdsi.org/wp-content/uploads/2022/10/6-Ostropolets_Phoebe2.0-abstract.pdf
7. Yan C, Ong HH, Grabowska ME, Krantz MS, Su WC, Dickson AL, et al. Large language models facilitate the generation of electronic health record phenotyping algorithms. J Am Med Inform Assoc. 2024 Apr 13;ocae072.
8. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012 Jan 1;19(1):54–60.
9. ChatGPT [Internet]. [cited 2023 Dec 13]. Available from: https://chat.openai.com