Towards rapid augmented phenotyping using large language models

Juan M. Banda[1], Azza Shoaibi[2], Gowtham Rao[2], Evan Minty[3], Christophe Lambert[4] ,Joel Swerdel[2], Christian Reich[5], George Hripcsak[6], Patrick Ryan[2]

[1] Georgia State University, Atlanta, Georgia, USA [2] Janssen Research & Development, Titusville, New Jersey, USA, [3] O'Brien Institute for Public Health, Department of Medicine, University of Calgary, Calgary, Canada, [4] University of New Mexico, Department of Internal Medicine, Albuquerque, NM, USA [5] Northeastern University, Bouvé College of Health Sciences, Boston, Massachusetts, USA, [6] Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, New York, USA

Background

Large language models (LLM), such as GPT-4(1), have emerged as transformative tools with immense potential in various fields, including the medical domain. Potentially, LLMs can significantly enhance clinical decision-making as they have the capacity to process and analyze vast amounts of medical data, including research papers, clinical guidelines, electronic health records, and patient histories. By synthesizing this information, LLMs can provide clinicians with up-to-date, evidence-based insights and recommendations, aiding in accurate diagnoses and treatment plans. This can lead to improved patient outcomes, reduced errors, and more efficient healthcare delivery.

In this work, we try to address a very pressing need of the OHDSI community: **the scalability of the electronic phenotyping task**, a cornerstone of almost every single study carried out by the community. Electronic phenotyping is the process of using electronic health records (EHRs) and other digital health data to identify and classify patient phenotypes. It involves analyzing structured and unstructured data through rule-based approaches, machine learning techniques, natural language processing (NLP), and hybrid methods. Rule-based approaches use predefined algorithms, while machine learning employs models trained on labeled datasets to predict phenotypes. In the OHDSI community we have ATLAS(2) for rule-based definitions and APHRODITE(3) for machine learning/probabilistic definitions. However, in order to create a good phenotype definition, a considerable amount of hours are poured into performing literature reviews and achieving alignment between multiple clinical and domain experts to converge on a single phenotype definition. This makes phenotyping not scalable in its current approach. While machine learning approaches have tried to bridge this gap (4,5), they are not widely used in practice.

With the intention to leverage the abstraction capabilities of LLMs, we aim to leave the literature review and condensation work to the model and have a domain expert assess the output. While this approach is not high-throughput or is intended to be fully automatic, it would greatly simplify and streamline the phenotyping task. An additional gap we fill with this study is the need to have domain-specific benchmarks (phenotyping in this case) to assess LLM performance within the wide variety of potential medical applications that exist. As shown by Dash et al. (6), there is a need to have fine-grained assessments of LLMs that go beyond having them answer United States Medical Licensing Examination (USMLE) questions or solve research benchmarks (7,8).

Methods

We selected 25 target phenotypes compiled from various sources, such as OHDSI Phenotype Phebruary 2022 and 2023, and some provided by our domain-expert reviews. Table 1 shows all definitions. We then prompted GPT-4 with several different prompts and, based on an internal evaluation, the prompt that

released the most consistent, and similarly formatted results was: "*Provide a computable phenotype definition for <insert phenotype name here>*".

Table 1. List of phenotypes evaluated.

Inflammatory Bowel Disease
Ulcerative Colitis
Crohn's disease
Rheumatoid Arthritis
Plaque Psoriasis
Lupus
High Blood Pressure
Metastatic Cancer
Anticoagulation
Atopic dermatitis
Sepsis
Hospital Inpatient Stay
Outpatient Visits
Pregnancy
Allergic Reaction to Medication
Depression
Requiring High dose immunosuppression
Acquired Neutropenia
Appendicitis
Parkinson's Disease
Acute Pancreatitis
Anaphylaxis
Acute Hepatic Failure
Idiopathic Inflammatory Myopathies
Dementia

Lastly, we had four domain experts evaluate the computable phenotype definitions generated by GPT-4. The evaluation was qualitative, as in, we wanted to evaluate if the automatically generated definitions would be harmful, useful in their present form, or need additional work to be useful. The questions and possible responses to our reviewers were: *"Would this definition misguide somebody with its use?"* with three possible answers: No, Yes, Not completely, and *"Is this definition good enough to code a computable phenotype definition?"* with possible answers: Yes, Needs some work, Needs a lot of work, or Not useful at all. Note that we are not counting the clinical codes produced or the implementation of such computable definitions, just the usability of what was produced by the LLM. That quantitative work will be further evaluated in a follow-up analysis.

Results

In table 2 we see the frequency of the responses by our domain-expert reviewers. For question 1 “Would this definition misguide somebody with its use?” It is an interesting finding that 98% of the responses are not deemed completely irrelevant to the point of misguiding somebody. However, it is clear that the evaluators do not think the generated responses are fully useful out of the box, as 73% of them are labeled as not completely. Only 25% of them are good enough, to their standard to even begin considering using them.

Table 2. Evaluation counts for each question for the evaluated phenotypes.

Question 1: Would this definition misguide somebody with its use?			***Question 2: Is this definition good enough to code a computable phenotype definition?***	
Not Completely	73		Needs a lot of work	68
No	25		Needs some work	27
Yes	2		Not useful at all	4
			Yes	1

The response patterns to the second question “Is this definition good enough to code a computable phenotype definition?” do showcase that these generated responses would still need considerable work from domain experts to be potentially used. With 68% of them categorized as “needing a lot of work” and 27% of them needing additional work. It is interesting to note that only four responses were deemed not a good start for creating a computable definition. However, the grading of these did not overlap between reviewers, meaning there are clearly different patterns as to how domain-experts see these generated responses. As we have more than two reviewers, we calculated the Fleiss' Kappa coefficient (9) for both questions. For question 1 we have a of 0.755, with a p value > 0.0001, meaning there is very strong agreement between our reviewers. For question 2, the coefficient value is 0.318 with a p-value > 0.0001, which shows that there is fair agreement between reviewers, but far from general consensus.

Conclusion

While LLMs are a very promising technology that can, and will, make advances in the medical domain, the need to have appropriate benchmarks and evaluations like this one is vital. In the field of electronic phenotyping, we have shown that while the responses of GPT-4 are mostly coherent, they still need considerable amounts of work to be useful to create definitions that humans spend large amounts of time doing. We also showed that even when evaluating, there is considerable variability as to how phenotype definitions are considered useful and how much work they need, elucidating the need to have better alignment as a community for this. However, there is promise that LLMs with additional prompting and instruction tuning (10) for the task at hand, such limitations can be overcome and lead us steps closer to scalable phenotyping with humans in the loop.

References

1. OpenAI. GPT-4 Technical Report [Internet]. arXiv [cs.CL]. 2023. Available from: http://arxiv.org/abs/2303.08774

2. Atlas: ATLAS is an open source software tool for researchers to conduct scientific analyses on standardized observational data [Internet]. Github; [cited 2023 Jun 15]. Available from:

https://github.com/OHDSI/Atlas

3. Banda JM, Halpern Y, Sontag D, Shah NH. Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network. AMIA Jt Summits Transl Sci Proc [Internet]. 2017 Jul 26;2017:48–57. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28815104

4. Agarwal V, Podchiyska T, Banda JM, Goel V, Leung TI, Minty EP, et al. Learning statistical models of phenotypes using noisy labeled training data. J Am Med Inform Assoc [Internet]. 2016 Nov;23(6):1166–73. Available from: http://dx.doi.org/10.1093/jamia/ocw028

5. Halpern Y, Horng S, Choi Y, Sontag D. Electronic medical record phenotyping using the anchor and learn framework. J Am Med Inform Assoc [Internet]. 2016 Jul;23(4):731–40. Available from: http://dx.doi.org/10.1093/jamia/ocw011

6. Dash D, Thapa R, Banda JM, Swaminathan A, Cheatham M, Kashyap M, et al. Evaluation of GPT-3.5 and GPT-4 for supporting real-world information needs in healthcare delivery [Internet]. arXiv [cs.AI]. 2023. Available from: http://arxiv.org/abs/2304.13714

7. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. N Engl J Med [Internet]. 2023 Mar 30;388(13):1233–9. Available from: http://dx.doi.org/10.1056/NEJMsr2214184

8. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature [Internet]. 2023 Apr;616(7956):259–65. Available from: http://dx.doi.org/10.1038/s41586-023-05881-4

9. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull [Internet]. 1971 Nov;76(5):378–82. Available from: https://psycnet.apa.org/fulltext/1972-05083-001.pdf

10. Peng B, Li C, He P, Galley M, Gao J. Instruction Tuning with GPT-4 [Internet]. arXiv [cs.CL]. 2023. Available from: http://arxiv.org/abs/2304.03277