Towards automated phenotype definition extraction using large language models

Ramya Tekumalla[1],Juan M. Banda[2,3]
[1] Mercer University, Atlanta, Georgia, USA,
[2] Stanford Health Care, Stanford, California, USA,
[3] Observational Health Data Sciences and Informatics, New York, New York, USA

Background

Electronic phenotyping, the process of extracting meaningful health characteristics from digital health data, is a cornerstone of modern biomedical research and personalized medicine. With the advent of electronic health records (EHRs) and the explosion of digital health data, the potential to leverage this information for improved patient care and innovative medical research has grown exponentially. Traditionally, phenotyping relied on manual methods, involving comprehensive literature reviews and collaborative efforts among clinicians and researchers to define specific health outcomes, diseases, or conditions [1,2] . This process, although thorough, is time-consuming and not easily scalable [3,4].

The integration of structured and unstructured data has given rise to advanced electronic phenotyping methods [5,6]. These methods utilize rule-based systems, machine learning (ML), and natural language processing (NLP) to analyze vast datasets, offering more precise and comprehensive phenotypic insights[7]. While rule-based systems involved predefined criteria and logical conditions to identify phenotypes, ML techniques included supervised, unsupervised, and weakly supervised methods[8] allowing for data-driven identification and classification of phenotypes. The expansion of NLP further enhances this capability by enabling the extraction of relevant information from free-text clinical notes, thereby expanding the scope of data that can be analyzed.

Recent advancements in machine learning, particularly the development of large language models (LLMs) such as PhenoBCBERT and PhenoGPT[9], have revolutionized electronic phenotyping. These models, equipped with hundreds of billions of parameters, leverage few-shot learning to achieve high performance with minimal training examples. This capability significantly reduces the time and effort required to define and refine phenotypes, making the process more scalable and adaptable to the fast-paced advancements in medical research and emerging health crises .

Methods

In this work, we propose an innovative approach to address the scalability challenge in electronic phenotyping. Our work is anchored in two main objectives: first, to define a standard evaluation task/set specifically tailored for this domain, and second, to evaluate various prompting approaches for extracting phenotype definitions from LLMs. The establishment of a standard evaluation task is crucial as it serves as a benchmark to ensure that the outputs produced by LLMs are not only useful but reliable. To create an evaluation set we used 10 professionally created phenotypes: five from PheKB[10] and five from the OHDSI phenotype library[11]. Extracting phenotypes from sources like OHDSI is easier due to their structured format and use of the OMOP Common Data Model. In contrast, PheKB offers flexibility without a mandated data model, requiring more effort to adapt algorithms across different systems. To streamline this, we manually curated and developed automated code to extract and format elements from PheKB phenotypes for automatic evaluation.

To evaluate prompting approaches to extract phenotype definitions from LLMs, we experimented with several methods (Zero-shot, One-shot, Iterative prompting, Seeding) and finalized a prompt for consistent results. The final prompt was: "Provide a computational phenotype for <INSERT_PHENOTYPE> with codes, their names, logical conditions, and code counts in tabular format." We evaluated LLM efficiency in two scenarios: comparing definitions from GPT-3.5 and GPT-4, and comparing GPT-4 definitions with human-curated ones. Metrics included code overlap, string overlap, logical matching, and analysis of inconsistencies and incorrect definitions.

Results

Key findings indicate that GPT models excel at generating precise codes but struggle with textual strings, showing variability in outputs across iterations. Interestingly, LLMs effectively extract logical conditions for including or excluding codes in phenotype definitions. This variability in code and string overlap is partly due to the diverse code systems used in literature and the definitions[12]. Table 1 presents the results of comparison between GPT 3.5 versus GPT 4.

GPT-4 generates codes with marginally higher reliability than textual strings or concept names. Despite generating fewer codes overall, GPT-4's codes may have a higher positive predictive value (PPV) for accurately identifying intended phenotypes, suggesting their specificity and relevance are high. This implies that GPT-4 might be averaging codes from sources and surfacing the most popular ones. However, hallucinations were present in both models, with GPT-3.5 exhibiting a higher tendency for inaccuracies than GPT-4. Table 2 presents the results of comparison between human definitions versus GPT models.

Metric	Average %	Minimum %	Maximum %
Codes overlap	41.26	0.00	75.00
Logic overlap	80.00	50.00	100.00
Strings overlap	28.52	0.00	50.00

Table 1. Comparison between GPT 3.5 vs GPT 4

Using Biomedical Content Explorer[13] linked with PubDictionaries, ICD10, and ICD10-CM dictionaries, we compared GPT-3.5 and GPT-4 in generating phenotype codes. The results highlight the models' weaknesses, particularly their inaccuracies and hallucinations. These issues were more pronounced for less-documented phenotypes, underscoring the need for cautious use and meticulous verification of LLM-generated data. Enhancing training methodologies to address literature scarcity on specific phenotypes is crucial to improving model accuracy. Figure 1 presents the comparisons of GPT hallucinations when producing codes.

Model	Metric	Average %	Minimum %	Maximum %
GPT 4	Codes overlap	50.94	20.00	88.89
	Logic overlap	90.00	50.00	100.00
	Strings overlap	48.59	0.00	100.00
GPT 3.5	Codes overlap	27.51	10.00	85.20
	Logic overlap	70.20	0.00	90.00
	Strings overlap	41.28	0.00	75.12

Table 2. Comparison between human definition vs GPT models

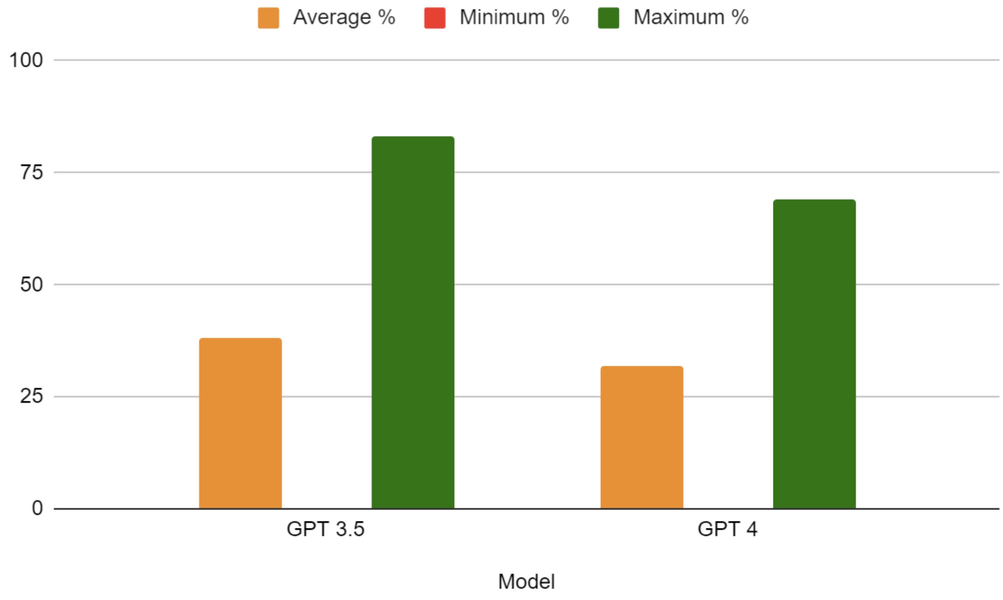

Figure 1. Comparisons of GPT hallucinations when producing codes

Conclusion

Our exploration of LLMs for automating phenotype definition extraction highlights their potential to enhance scalability and efficiency in digital healthcare. While GPT-3.5 and GPT-4 show promise in generating medically relevant codes, challenges remain in achieving consistent textual output and avoiding inaccuracies. The study underscores the need for robust evaluation and validation frameworks to ensure LLM reliability. Despite hallucinations and inconsistencies, GPT models can serve as valuable initial steps or augmentation tools, significantly streamlining and improving electronic phenotyping methodologies.

References

1. Nadkarni, G. N. *et al.* Development and validation of an electronic phenotyping algorithm for chronic kidney disease. *AMIA Annu. Symp. Proc.* **2014**, 907–916 (2014).
2. Smoller, J. W. The use of electronic health records for psychiatric phenotyping and genomics. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **177**, 601–612 (2018).
3. Rasmussen, L. V. *et al.* Considerations for Improving the Portability of Electronic Health Record-Based Phenotype Algorithms. *AMIA Annu. Symp. Proc.* **2019**, 755–764 (2019).
4. Huckvale, K., Venkatesh, S. & Christensen, H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. *NPJ Digit Med* **2**, 88 (2019).
5. Banda, J. M., Seneviratne, M., Hernandez-Boussard, T. & Shah, N. H. Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. *Annu Rev Biomed Data Sci* **1**, 53–68 (2018).
6. Swerdel, J. N., Hripcsak, G. & Ryan, P. B. PheValuator: Development and evaluation of a phenotype algorithm evaluator. *J. Biomed. Inform.* **97**, 103258 (2019).
7. Luo, L. *et al.* PhenoTagger: a hybrid method for phenotype concept recognition using human phenotype ontology. *Bioinformatics* **37**, 1884–1890 (2021).
8. Agarwal, V. *et al.* Learning statistical models of phenotypes using noisy labeled training data. *J. Am. Med. Inform. Assoc.* **23**, 1166–1173 (2016).
9. Yang, J. *et al.* Enhancing phenotype recognition in clinical notes using large language models: PhenoBCBERT and PhenoGPT. *Patterns (N Y)* **5**, 100887 (2024).
10. Kirby, J. C. *et al.* PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J. Am. Med. Inform. Assoc.* **23**, 1046–1052 (2016).
11. Banda, J. M., Halpern, Y., Sontag, D. & Shah, N. H. Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network. *AMIA Jt Summits Transl Sci Proc* **2017**, 48–57 (2017).
12. Brandt, P. S. *et al.* Characterizing variability of electronic health record-driven phenotype definitions. *J. Am. Med. Inform. Assoc.* **30**, 427–437 (2023).
13. Kim, J. Biomedical Content Explorer. https://chat.openai.com/g/g-wdWOSr2gs-biomedical-content-explorer (2023).