Leveraging GPT-4 for Identifying Clinical Phenotypes in Electronic Health Records: A Performance Comparison between GPT-4, GPT-3.5-turbo and spaCy's Rule-based & Machine Learning-based methods

Kriti Bhattarai[1,2], Inez Y. Oh[1], Jonathan Moran Sierra[3], Philip R.O. Payne[1,2], Zachary B. Abrams[1], Albert M. Lai[1,2]

[1] Institute for Informatics, Data Science & Biostatistics, Washington University School of Medicine, St. Louis, Missouri, USA
[2] Department of Computer Science, Washington University in St. Louis, Missouri, USA
[3] Medical Scientist Training Program, Washington University School of Medicine, St. Louis, Missouri, USA

Corresponding Author:
Kriti Bhattarai
Department of Computer Science
Institute for Informatics, Data Science & Biostatistics
Washington University in St. Louis
660 S. Euclid Ave, 6th Floor
St. Louis, MO, 63110, USA
kriti.bhattarai@wustl.edu

Keywords: generative pre-trained transformer (GPT), natural language processing, clinical informatics, clinical phenotype extraction, electronic health records

Word Count:3181

ABSTRACT

Objective: Accurately identifying clinical phenotypes from Electronic Health Records (EHRs) provides additional insights into patients' health, especially when such information is unavailable in structured data. This study evaluates the application of OpenAI's transformer-based Generative Pre-trained Transformer (GPT)-4 model to identify clinical phenotypes from EHR text in non-small cell lung cancer (NSCLC) patients. The goal is to identify disease stages, treatments and progression utilizing GPT-4, and compare its performance against GPT-3.5-turbo, and two rule-based and machine learning-based methods, namely, scispaCy and medspaCy.

Materials and Methods: Phenotypes such as initial cancer stage, initial treatment, evidence of cancer recurrence, and affected organs during recurrence were identified from 13,646 records for 63 NSCLC patients from Washington University in St. Louis, Missouri. The performance of the GPT-4 model is evaluated against GPT-3.5-turbo, medspaCy and scispaCy by comparing precision, recall, and weighted F1 scores.

Results: GPT-4 achieves higher F1 score, precision, and recall compared to medspaCy and scispaCy's models. GPT-3.5-turbo performs similar to that of GPT-4. GPT models are not constrained by explicit rule requirements for contextual pattern recognition. SpaCy models rely on predefined patterns, leading to their suboptimal performance.

Discussion and Conclusion: GPT-4 improves clinical phenotype identification due to its robust pre-training and remarkable pattern recognition capability on the embedded tokens. It demonstrates data-driven effectiveness even with limited context in the input. While rule-based models remain useful for some tasks, GPT models offer improved contextual understanding of the text, robust clinical phenotype extraction, and improved ability to provide better care to the patients.

BACKGROUND AND SIGNIFICANCE

Introduction

Extracting clinical phenotypes from unstructured Electronic Health Records (EHRs) is a critical task in natural language processing (NLP). Accurately identifying relevant phenotypes from unstructured text utilizing NLP techniques provides additional insights into patients' health, especially when such information is unavailable in structured data. NLP extraction techniques facilitate this process by mapping unstructured text to a structured representation, making it easier to evaluate patients' disease progression, treatment modalities, and treatment effectiveness. This is particularly evident when analyzing data from non-small cell lung cancer patients, where unstructured text is abundant. Accurately identifying disease stage, treatments and progression from cancer text will contribute to continued research efforts aimed at improving treatment strategies for non-small lung cancer patients, assessing disease progression, and improving lung cancer-related outcomes.

Background

Clinical phenotype extraction is an ongoing research area where the type of extraction tasks and target phenotypes vary across different clinical domains. Several rule-based, machine learning-based, and deep-learning models have been applied to phenotype extraction.[1-6] While rule-based models extract phenotypes based on pre-defined patterns, most machine learning and deep-learning approaches are trained on sentences or documents labeled with the relevant phenotypes and the model subsequently classifies texts into these phenotypes.[5,7] MedspaCy[6] and scispaCy[8] are two recent and extensively-used hybrid frameworks that utilize statistical and machine-learning methods in conjunction with rule-based NLP to identify clinical phenotypes. Various studies have utilized medspaCy and scispaCy to identify specific sections within EHR

text for NER, extract phenotypes from relation extraction documents, and generate text embeddings.[9-13]

Although extracting clinical phenotypes is essential, several problems and gaps remain in the existing literature. Firstly, there is a lack of consensus on the most effective technique for clinical phenotype extraction, with studies showing contrasting results.[14,15] Furthermore, current methods for handling information extraction tasks in clinical NLP often lack robustness, leading to suboptimal performance.[16] This is especially evident in cancer-related data, where phenotype extraction from public clinical data does not translate to cancer text sourced from proprietary institutional data, which tend to be unprocessed compared to the cleaned public data . In addition, the limited availability of labeled, publicly accessible cancer EHR text leaves an important domain unexplored for NLP.

Pre-trained transformer-based language models have recently been studied for tasks such as question answering, text generation, and machine translation.[17,18] Despite the success of transformer-based language model in such tasks, their application in the context of clinical phenotype extraction remain underexplored, opening up numerous avenues of research. It is essential to investigate these recent transformer-based methods in specific clinical domains and compare their performance to previous machine learning and rule-based models to generate insights into their potential benefits for clinical phenotype extraction.

OBJECTIVES

The aim of this study was to investigate the most recent transformer-based language models as they remain underexplored for clinical phenotype extraction from EHR text. We evaluated the application of OpenAI's Generative Pre-trained Transformer (GPT)-4 model[19] for clinical phenotype extraction in an EHR retrospective study focusing on non-small cell lung

cancer patients as a specific case study. In particular, we used GPT-4 to identify individual words or tokens in a data sequence as distinct phenotypes. Specifically, we measure the prevalence of specific lung cancer phenotypes, including cancer stage, treatment modalities, cancer recurrence, and organs affected by cancer recurrence. These phenotypes are important for informing treatment decisions and assessing disease progression in non-small cell lung cancer patients.

We built the model framework using a clinical dataset from Washington University in St. Louis, Missouri, for a patient population diagnosed with non-small cell lung cancer. To evaluate the effectiveness of GPT-4, we compared its results against a subject matter expert's manual annotation. We also conducted a comparative analysis with GPT-3.5-turbo[20], and medSpacy and sciSpacy, currently recognized as some of the effective rule-based and machine learning approaches in clinical phenotype extraction.

Our comparison between scispaCy, medspaCy, GPT-3.5-turbo and GPT-4 aims to highlight the strengths and weaknesses of each approach for phenotype extraction, providing valuable insights into their effectiveness and potential use for cases in clinical phenotype extraction from EHR. In evaluating these current approaches for phenotype extraction, we also note their limitations.

MATERIALS AND METHODS

To extract a detailed representation of specific lung cancer phenotypes, we used GPT-4, available through Microsoft's Azure OpenAI Service. To perform a direct comparison with the annotated data, recorded at the patient-level, we mapped encounter-level details to patient-level details. We compared and evaluated the performance of the current models by comparing true positives (recall) and false positives at the phenotype-level. The following subsections discuss

the datasets, annotation methods, and methodologies used for extracted information, baseline comparison techniques, and evaluation metrics used to quantify differences in results. **Figure 1** illustrates the pipeline we followed for extraction.

Patient Population and Data Sources/Corpus Creation

Retrospective outpatient and inpatient EHR data were obtained from Washington University Physicians / BJC Healthcare in St. Louis, Missouri, for all patient encounters with a non-small cell lung cancer diagnosis between 2018-2023. In total, we extracted 13,646 narratives from the EHR for 63 patients. The unstructured texts for these patients included progress reports, physician notes, and nursing reports. The texts primarily describe patients' disease trajectory during their visit, ranging from primary cancer diagnosis, cancer stage, treatment type, treatment completion, and cancer recurrence (**Figure 2**).

Lung cancer phenotypes extracted from the clinical narratives

Our extraction pipeline currently targets four types of phenotypes: cancer stage, cancer treatment (chemotherapy, radiation, surgery), evidence of cancer recurrence, and organs affected by cancer recurrence. The variations extracted for each phenotype are listed in **Table 1**. We attempted to search for all variations of the targeted phenotypes from the corpus.

Table 1: Variations of the relevant phenotypes used in the search for phenotype extraction. All strings were case-insensitive.

Phenotype	Variations
Initial treatment	Chemotherapy Chemo-radiation Radiation Surgery Lobectomy Segmentectomy Wedge Resection
Initial stage	Stage 0 Stage 1 Stage 2 Stage 3 Stage 4
Cancer recurrence instances	Relapsed Recurred Recurrence Recurrent
Organs affected by cancer recurrence	Liver Kidney Bone Brain Lymph Local Lung Adrenal glands Pleura Pericardium

Gold-standard data annotation

The results from the phenotype extraction pipeline for each model were evaluated against a gold-standard manual annotation from a subject matter expert at the same institution, containing expert determination of cancer type, treatment, recurrence instances, and organs affected by cancer recurrence for a subset of the same patient cohort. The annotation served as a benchmark against which the performance of the model pipelines was evaluated.

A Research Electronic Data Capture (REDCap) form was designed to collect responses from the annotator to capture comprehensive coverage of the phenotypes and consistently accurate results across all patients. The annotated dataset consisted of 63 unique patients from the BJC EHR, comprising a total of 13,646 records for all patients. The annotator annotated the narratives, including copying evidence from the narratives. This evidence guided their interpretation or choice, and they recorded it in a free-text field. All the phenotypes mentioned in **Table 1** were identified in the annotator's annotation, with some phenotypes being identified more frequently than others, depending on the nature of the patient's disease trajectory. Some patients show cancer recurrence in multiple organs, and the percentage is inclusive of each affected organ. **Table 2** summarizes the frequency of annotations corresponding to each phenotype variation.

Table 2. Frequency of annotations corresponding to each phenotype variation identified for each patient within the cohort, based on the available annotations.

Phenotype	Variations	Number of Annotations	Percentage of Patients with Annotations
Initial treatment	Chemotherapy	9	11.11%
	Chemo-radiation	34	9.52%
	Radiation	7	14.29%
	Surgery	6	53.97%
Initial stage	Stage 0	1	1.61%
	Stage 1	9	12.90%
	Stage 2	3	6.45%
	Stage 3	33	51.61%
	Stage 4	15	24.11%
Relapse instances	Relapsed	21	33.33%
	Not Relapsed	42	66.67%
Organs affected by cancer relapse	Liver	2	15.79%
	Kidney	1	5.26%
	Bone	3	26.31%
	Brain	9	47.37%
	Lymph	1	5.26%
	Local Lung	1	5.26%
	Adrenal glands	1	10%
	Pleura	1	10%
	Pericardium	1	5.26%

Non-small cell lung cancer phenotype extraction and model comparison

We implemented GPT-4 and compared its performance with GPT-3.5-turbo, medspaCy and scispaCy. For all models, the input to the models were the phenotypes and its variations. For GPT, we implemented the default zero-shot model where the model input was the text together with the prompt to guide the model for phenotype extraction. We used the same phenotype variations for extraction across all spaCy model implementations. GPT models did not require inclusion of all phenotype variations.

Development of the GPT pipeline as an information extractor to extract each phenotype

GPT-3.5-turbo and GPT-4 are a transformer-based language model trained on a large unspecified corpus for multiple NLP tasks, including natural language generation developed by OpenAI. It has been used for natural language generation tasks using their chatCompletion and translations endpoint. Our setup is an adaptation of the sequence labeling task from chatCompletion framework for phenotype extraction. The sequence labeling setup requires providing context to the model, where the model generates responses that include labeled phenotypes from the clinical notes. The model outputs are the expected phenotypes we are trying to extract. The core idea involves assigning specific labels to individual words or tokens in the clinical notes, capturing the relevant information while retaining the original context.

To build the GPT framework, we used Microsoft's Azure OpenAI Service, which provides REST API access to OpenAI's language models. We deployed the OpenAI API endpoint into a HIPAA-compliant subscription within Washington University's Azure tenant. This enabled us to study the performance of GPT in a secure and HIPAA-compliant manner. Additionally, we applied for and received an exemption from content filtering, abuse monitoring, and human review of our use of the Azure OpenAI service, which removes the ability of Microsoft employees to perform any form of data review. At the time of our experiments, GPT-3.5-turbo Version 0301 and GPT-4 Version 0613 were the most recent GPT models available.

For phenotype extraction, the model identifies treatment procedures, stage information, and recurrence information within the clinical notes mentioned in **Table 1**. Each word or token in a clinical note are categorized contextually through prompts with the relevant phenotype categories (e.g., treatment, staging) or their sub-categories (e.g., surgery, radiation, chemotherapy, stage numbers) to extract desired information. Due to the probabilistic design of

GPT models, the output may include extra words or phrases around the actual phenotypes, which are then parsed in the post-processing step (**Supplementary Table 1**). The variations of the prompts and the sample results from each prompt are provided in **Supplementary Figure 1** and **Supplementary Table 1**. Optimized hyperparameters of the model are listed in **Supplementary Table 2**.

Development of the spaCy-based NLP pipelines to extract each clinical phenotype using hybrid techniques

In our study, we implemented spaCy's rule-based and machine learning-based approaches. ScispaCy is a rule-based and Named Entity Recognition (NER)-based Python library for biomedical text processing, which has demonstrated robust results on several Named Entity Recognition (NER) tasks compared to the neural network models of the time.[5] It is trained on gene data, PubMed articles, medications datasets, and one of their proprietary datasets. We implemented scispaCy version 0.5.2 following the code structure specified in their documentation. We added specific phenotypes of interest and their corresponding string variations as rules in the pipeline that were then extracted as a result of the model. We incorporated scispaCy's built-in functions to handle negation and NER. The results are strings extracted from the text and the position of the characters in that text. If a string was not present in the text, the output was null. Finally, the output was mapped into their specific phenotype categories.

MedspaCy is also a rule-based and NER-based Python library that includes UMLS (Unified Medical Language System)[21] mappings for clinical phenotype extraction. A similar approach was applied for medspaCy (version 1.0.1) as scispaCy. The output from medspaCy was similar to scispaCy, with strings extracted from the text and the position of the characters from

that text. The final result from the pipeline were all the strings that medspaCy extracted. Our current implementation focused on capturing exact matches of the phenotype variations mentioned in **Table 1**.

For medspaCy and scispaCy, each existing output string from the clinical notes that matched with phenotype variations was later assigned to the relevant phenotype categories on a patient-level, which were then analyzed as the final extracted phenotypes.

RESULTS AND EVALUATION

We evaluated the performance of each model in identifying the targeted cancer phenotypes (staging, treatment, recurrence, and organs) using precision, recall, and weighted F1 scores to collectively assess the effectiveness of each model in capturing the phenotypes. The results for all models are reported in **Table 3**.

Table 3: Phenotype extraction performance results for all models.

Approach	Phenotype	F1-Score	Precision	Recall
GPT-4	Staging	0.92	0.93	0.91
	Treatment	0.92	0.95	0.89
	Recurrence	0.96	0.94	0.98
	Organs	0.68	0.67	0.70
GPT-3.5-turbo	Staging	0.90	0.93	0.88
	Treatment	0.91	0.94	0.89
	Recurrence	0.96	0.93	1.00
	Organs	0.62	0.59	0.65
scispaCy	Staging	0.66	0.61	0.71
	Treatment	0.60	0.58	0.63
	Recurrence	0.61	0.54	0.71
	Organs	0.55	0.57	0.54
medspaCy	Staging	0.66	0.63	0.69
	Treatment	0.61	0.58	0.65
	Recurrence	0.59	0.56	0.63
	Organs	0.53	0.55	0.51

Comparison of Models

The GPT-4 model demonstrated higher F1 scores with high precision and recall, indicating its ability to correctly identify all instances of recurrence, staging, treatment, and organs in the clinical text better than scispaCy and medspaCy (**Table 3**). GPT-3.5-turbo and GPT-4 had similar performance across most phenotypes with one phenotype showing identical performance. Although scispaCy had lower F1 scores than GPT-3.5-turbo and GPT-4, it outperformed medspaCy in most phenotype extraction tasks. MedspaCy had the lowest F1 score, suggesting it is less effective at information extraction than scispaCy, GPT-3.5-turbo and GPT-4. This is potentially due to its less advanced NER techniques than scispaCy and GPT models. The model-generated output of GPT-3.5-turbo and GPT-4 varied across each run but maintained the underlying meaning of the result across all runs (**Supplementary Table 1**).

Qualitative Analysis of the Results

We performed a qualitative analysis of the results made by each model in phenotype extraction to better understand their strengths and weaknesses.

GPT-4 was better able to correctly identify cancer phenotypes while minimizing misclassifications, leading to a higher F1 score compared to GPT-3.5-turbo, medspaCy and scispaCy. When comparing GPT-3.5-turbo and GPT-4, we identified that both models captured contextual information accurately. However, the generated text from GPT-4 is more specific to the prompt than the text generated from GPT-3.5-turbo (**Supplementary Table 4**). Upon examining the errors, we observed that GPT models sometimes mislabeled phenotypes when the context was ambiguous, especially when the same sentence discussed multiple phenotypes (Details about GPT's treatment extraction results are in **Supplementary Table 3).**

MedspaCy and scispaCy could not identify contextual phenotypes or phenotypes mentioned in a negated context, synonyms not part of the rules, and spelling errors. GPT-3.5-turbo and GPT-4 were far better in these cases. For example, GPT-3.5-turbo and GPT-4 were able to identify "T1c N0 M0" as an indication of a cancer stage, whereas the other models could not identify without significant further pipeline engineering. (**Supplementary Table 4-5**). This could be due to spaCy's inability to learn contextual information.

Across all models, the more specific we defined a phenotype in the prompt or by rules, the better were our chances of identifying it correctly. Details of all error analyses are described in **Supplementary Tables 3, 5, and 6**.

DISCUSSION

Our study highlights GPT-4's remarkable performance in identifying phenotypes with minimal preprocessing and postprocessing steps compared to rule-based or traditional machine-learning-based algorithms. This aligns well with the established notion that large language models are data-driven and highly effective even with limited contextual information, unlike rule-based or traditional machine learning algorithms that rely solely on predefined patterns or rules known to researchers or clinicians.[22]

GPT-3.5-turbo performs similar to that of GPT-4 for some phenotypes. The choice of GPT-3.5-turbo versus GPT-4 would depend on scalability and cost-effectiveness. While GPT-4 is more scalable as its results are more specific to prompts, GPT-3.5-turbo is cost-effective (**Supplementary Table 4**). Overall, GPT models, with its robust unsupervised pre-training and remarkable pattern recognition capability on the tokens, outperforms other models as it extracts relevant patterns and relationships without being constrained by the need for prior knowledge of explicit patterns, rules, or meaning. Based on the context provided in the prompt, GPT is able to

capture variations in the representation of the clinical phenotypes, making it well suited for information extraction tasks that could extend beyond our focus on its application in oncology.

Our analyses also revealed that GPT demonstrated significant performance improvement than the other models, even in its default zero-shot setup. We implemented the GPT models, not specifically fine-tuned on clinical text. Fine-tuning with clinical text requires additional labeled clinical text, which is not readily available and would have been time-consuming to procure.

For the GPT model outputs, we also obtained varying texts from the API across multiple iterations of the same query despite using the same prompt, suggesting that GPT model might not provide identical results across multiple iterations of the same query. This could be due to its probabilistic design. After analyzing the output texts, we found that all the extracted phenotypes were correctly identified within the text, with only differences in the words and language used (**Supplementary Table 7**).

The comparative analysis also revealed that scispaCy performed better than medspaCy in our study, possibly because of the additional NER components and diverse data sources that it is trained on, in addition to handling the specific type of data that medspaCy is trained on. However, both approaches exhibited limitations in handling complex patterns and context-specific phenotypes. Results from medspaCy and scispaCy also indicate that rule-based models do not handle speculation, negation, and context ambiguity adequately (**Supplementary Table 5-6**).

Furthermore, while medspaCy and scispaCy offer deterministic results based on predefined rules, they fall short of capturing the contextual information required for effective information extraction in clinical text. Because of that, researchers must also have comprehensive knowledge of the phenotypes and variations of the phenotypes for extraction.

Finally, it is worth considering the interpretability aspect of these models. While medspaCy and scispaCy's rule-based nature allows for more straightforward interpretability, there might be some challenges in interpreting the results of the GPT model due to its unknown internal parameters.

LIMITATIONS

Despite these promising results, we acknowledge some limitations in this study. We evaluated our results using F-1 metrics, which have proven effective in comparing the performance of large language models (LLMs) to that of rule-based and machine learning-based models. However, it is important to re-consider the utility of traditional evaluation metrics, especially when comparing LLM-generated text with human-generated reference text. This is crucial due to the potential discrepancies in reference texts and variations in the representation of results across different LLMs, suggesting that traditional information retrieval metrics may not be well suited for all LLM tasks (**Supplementary Table 3**). Addressing these limitations will be a key focus in our future research.

Our second limitation pertains to the models evaluated against GPT. We initially investigated two other transformer-based language models, T5[23] and ClinicalBERT[18], to compare their results with the GPT language model. However, since the extracted results did not include the necessary phenotypes, we opted against their inclusion in the main manuscript and made comparisons with spaCy's rule-based and machine learning-based methods. Results utilizing a subset of clinical text for T5 and ClinicalBERT are included in **Supplementary Table 7-8**.

Additionally, we note that our random selection of a subset of patients may introduce bias and affect model performance. While the dataset was extracted from a 5-year cohort, the evaluation was based on a random subset of patients. Biases in the data could also lead to limitations in handling diverse clinical text or phenotypes and affecting model performance. Future research could address these limitations by including a larger dataset. Including a larger dataset in future research would address this limitation.

CONCLUSION

In conclusion, the study highlights the potential of GPT-4 for accurate phenotype recognition in clinical text. GPT-3.5-turbo model demonstrates performance similar to that of GPT-4. Both GPT models seem to be effective not only for already well-known text generation tasks but also surprisingly for information extraction tasks. While medspaCy and scispaCy offer deterministic results and have utility for some tasks, they exhibit limitations in handling complex patterns and context-specific phenotypes. Therefore, leveraging data-driven and contextually aware advanced language models like GPT-3.5-turbo and GPT-4 and addressing their current limitations opens up new possibilities for robust clinical phenotype extraction, ultimately leading to additional insights into patients' health and improved care.

ACKNOWLEDGEMENTS

This work was supported by Centene Corporation contract (P19-00559) for the Washington University-Centene ARCH Personalized Medicine Initiative. Jonathan Moran Sierra was supported by NIH/NIGMS T32GM007200.

CONFLICTS OF INTEREST STATEMENT

The authors do not have conflicts of interest to disclose.

REFERENCES

1. Cronin RM, Fabbria D, Denny JC, et al. A comparison of rule-based and machine learning approaches for classifying patient portal messages. *International Journal of Medical Informatics* 2017; 105: 110–20.
2. Oh IY, Schindler SE, Ghoshal N, et al. Extraction of clinical phenotypes for Alzheimer's disease dementia from clinical notes using natural language processing. *JAMIA Open* 2023; 6: ooad014.
3. Tome E, Seljak BK, Korosec P. A Rule-Based Named-Entity Recognition Method for Knowledge Extraction of Evidence-Based Dietary Recommendations. *PLoS One* 2017; 12: e0179488.
4. Peng Y, Torii M, Wu CH, et al. A Generalizable NLP Framework for Fast Development of Pattern-Based Biomedical Relation Extraction Systems. *BMC Bioinformatics* 2014; 15: 1.
5. Lee S, Shin J, Kim HS, et al. Hybrid Method Incorporating a Rule-Based Approach and Deep Learning for Prescription Error Prediction. *Drug Safety* 2022; 45: 27-35.
6. Eyre H, Chapman AB, Peterson KS, et al.Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. *In: AMIA Annual Symposium proceedings* 2021; 438-447.
7. Kocaman V, Talby D. Accurate Clinical and Biomedical Named Entity Recognition at Scale. *Software Impacts* 2022; 13: 100373.

8. Neumann M, King D, Beltagy I, et al. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. *In: Proceedings of the 18th BioNLP Workshop and Shared Task* 2019; 319-327.
9. Sorbello A, Haque SA, Hasan R, et al. Artificial Intelligence-Enabled Software Prototype to Inform Opioid Phramacovigilance from Electronic Health Records: Development and Usability Study. *JMIR AI* 2023*; 2:* e45000.
10. Gururaja S, Dutt R, Liao T, et al. Linguistic representations for fewershot relation extraction across domains. *In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* 2023; 1: 7502–7514.
11. Li J, Wang Y, Zhang S, et al. Rethinking document-level Relation Extraction: A Reality Check. *In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* 2023; 1: 5715-5730.
12. Shibayama S, Yin D, Matsumoto K.Measuring novelty in science with word embedding. *PLoS One* 2021; 16: e0254034
13. Yin D, Wu Z, Yokota K, et al. Identify novel elements of knowledge with word embedding. *PLoS One* 2023; 18: e0284567
14. Yang Y, Wu Z, Yang Y, et al. A Survery of Information Extraction Based on Deep Learning. *Applied Sciences* 2022; 12: 9691
15. Landolsi MY, Hlaoua L, Romdhane LB. Information extraction from electronic medical documents: state of the art and future research directions. *Knowledge and Information Systems* 2023; 65: 463-516

16. Lossio-Ventura JA, Sun R, Boussard S, et al. Clinical concept recognition: Evaluation of existing systems on EHRs. *Frontiers Artificial Intelligence*. 2023; 5:1051724.

17. Radford A, Narasimhan K, Salimans T, et al. Improving Language Understanding by Generative Pre-Training. 2018

18. Alsentzer E, Murphy J, Boag W, et al. Publicly available clinical BERT embeddings. *In: Proceedings of the 2nd Clinical Natural Language Processing Workshop* 2019*;* 72-78

19. OpenAI. 2023. Retrieved from https://arxiv.org/abs/2303.08774.

20. OpenAI. 2023. Retrieved from https://platform.openai.com/docs/api-reference/completions. Accessed July 2023.

21. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research* 2004; 32: D267-D270

22. Brown TB, Mann B, Ryder N, et al. Language Models are Few-Shot Learners. *Neural Information Processing Systems* 2020.

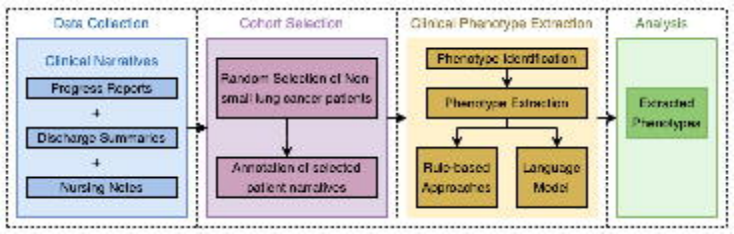
23. Raffel C, Shazeer N, Roberts A, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 2020; 1-67

Figure Legends:

Figure 1. Step-by-step approach to extracting phenotypes.

Clinical narratives from the EHR were extracted as part of for the data collection process. A subset of the narratives was randomly selected for manual annotation. ScispaCy, medspaCy, GPT-3.5-turbo and GPT-4 models were implemented for phenotype extraction. Extracted phenotypes were compared with the annotations.

Figure 2. Sample text from unstructured narratives of non-small cell lung cancer patients. The text highlighted in red are the targeted phenotypes for extraction.


Data Collection
Clinical Narratives
Progress Reports
+
Discharge Summaries
+
Nursing Notes
Cohort Selection
Random Selection of Non small lung cancer patients
Annotation of selected patient narratives
Clinical Phenotype Extraction
Phenotype Identification
Phenotype Extraction
Rule-based Approaches
Language Model
Analysis
Extracted Phenotypes

Date of Diagnosis: X/X/XX AJCC 8th Edition - Clinical stage from : Stage IIIB (cT4, cN1, cM0)

Stage

received definitive SBRT to 5500 cGy completed on XX/XX/XX

Treatment - Radiation

Relapse/Recurrence

CT on XX/XX/XX showed tumor relapse, and PET on XX/XX/XX showed suspicious liver lesions in addition to a left upper lobe pulmonary nodule.