Original Paper

Large Language Model Symptom Identification From Clinical Text: Multicenter Study

Andrew J McMurry[1,2], PhD; Dylan Phelan[1], MS; Brian E Dixon[3,4], MPA, PhD; Alon Geva[1,5], MPH, MD; Daniel Gottlieb[1,6], MPA; James R Jones[1], MPhil; Michael Terry[1], BS; David E Taylor[4], BS; Hannah Callaway[4], MS; Sneha Manoharan[4], MS; Timothy Miller[1,2], PhD; Karen L Olson[1,2], PhD; Kenneth D Mandl[1,2], MPH, MD

[1]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, United States
[2]Department of Pediatrics, Harvard Medical School, Boston, MA, United States
[3]Department of Health Policy and Management, Fairbanks School of Public Health, Indiana University, Indianapolis, IN, United States
[4]Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, IN, United States
[5]Department of Anesthesia, Harvard Medical School, Boston, MA, United States
[6]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, United States

Corresponding Author:

Kenneth D Mandl, MPH, MD
Computational Health Informatics Program
Boston Children's Hospital
401 Park Drive, LM5506, Mail Stop BCH3187
Boston, MA 02215
United States
Phone: 1 617-355-4145
Email: Kenneth.Mandl@Childrens.Harvard.edu

Abstract

Background: Recognizing patient symptoms is fundamental to medicine, research, and public health. However, symptoms are often underreported in coded formats even though they are routinely documented in physician notes. Large language models (LLMs), noted for their generalizability, could help bridge this gap by mimicking the role of human expert chart reviewers for symptom identification.

Objective: The primary objective of this multisite study was to measure the accurate identification of infectious respiratory disease symptoms using LLMs instructed to follow chart review guidelines. The secondary objective was to evaluate LLM generalizability in multisite settings without the need for site-specific training, fine-tuning, or customization.

Methods: Four LLMs were evaluated: GPT-4, GPT-3.5, Llama2 70B, and Mixtral 8×7B. LLM prompts were instructed to take on the role of chart reviewers and follow symptom annotation guidelines when assessing physician notes. Ground truth labels for each note were annotated by subject matter experts. Optimal LLM prompting strategies were selected using a development corpus of 103 notes from the emergency department at Boston Children's Hospital. The performance of each LLM was measured using a test corpus with 202 notes from Boston Children's Hospital. The performance of an *International Classification of Diseases, Tenth Revision* (*ICD-10*)–based method was also measured as a baseline. Generalizability of the most performant LLM was then measured in a validation corpus of 308 notes from 21 emergency departments in the Indiana Health Information Exchange.

Results: Symptom identification accuracy was superior for every LLM tested for each infectious disease symptom compared to an *ICD-10*–based method (F_1-score=45.1%). GPT-4 was the highest scoring (F_1-score=91.4%; P<.001) and was significantly better than the *ICD-10*–based method, followed by GPT-3.5 (F_1-score=90.0%; P<.001), Llama2 (F_1-score=81.7%; P<.001), and Mixtral (F_1-score=83.5%; P<.001). For the validation corpus, performance of the *ICD-10*–based method decreased (F_1-score=26.9%), while GPT-4 increased (F_1-score=94.0%), demonstrating better generalizability using GPT-4 (P<.001).

Conclusions: LLMs significantly outperformed an *ICD-10*–based method for respiratory symptom identification in emergency department electronic health records. GPT-4 demonstrated the highest accuracy and generalizability, suggesting that LLMs may augment or replace traditional approaches. LLMs can be instructed to mimic human chart reviewers with high accuracy. Future work should assess broader symptom types and health care settings.

J Med Internet Res 2025;27:e72984; doi: 10.2196/72984

Keywords: natural language processing; artificial intelligence; large language models; symptom recognition; clinical text mining; medical informatics; infectious disease surveillance; epidemiologic methods; emergency medical services; electronic health records

Introduction

To practice medicine, accurate identification and interpretation of symptoms are paramount. Symptoms are primary indicators of patient health, underpinning diagnostic processes [1] and choice of therapeutic interventions [2]. Identifying symptoms is also fundamental to public health [3,4], medication safety [5,6], clinical research [7,8], and clinical trials [9-13]. Though symptoms are routinely documented in physician notes, coded formats such as the *International Classification of Diseases, Tenth Revision* (*ICD-10*) [14] often underreport patient symptoms [4,15-18]. The gap between medical coding practices and richer phenotyping has motivated many efforts to develop natural language processing (NLP) of physician notes [17].

Traditional NLP methods for symptom identification [15,18,19] typically target specific note sections [20,21] such as the chief complaint [19,22-24] and often struggle to interpret if or when symptoms are positive [25-27]. The context [18,21,28] surrounding infectious respiratory diseases includes symptoms pertaining to acute infections, noninfectious conditions, treatment side effects [6], indications for treatment, or patient instructions (eg, "Use albuterol inhaler as needed for difficulty breathing").

Large language models (LLMs) hold potential to overcome such limitations [29,30]. As LLMs are derived from population scale examples, they may better infer symptoms from internet text such as articles about symptom checklists [1,31], disease progression [32], and medical decision-making [2]. Unlike traditional clinical NLP models, LLMs are not trained to any specific domain, which means that LLMs should be more generalizable to documentation variation across health care locations and may not require site-specific training [20,33] to achieve state-of-the-art accuracy.

We sought to measure the accuracy of LLMs for symptom identification, with a focus on infectious respiratory disease symptoms [4]. The code and results are available free of charge with the Apache open-source license 2.0 [34].

Methods

Study Design

This is a multisite retrospective study of infectious respiratory disease symptoms documented in electronic health records. Ground truth symptom labels were annotated by human expert chart reviewers. Two symptom identification methods were compared to ground truth labels: (1) an *ICD-10*–based method using coded data and (2) an LLM-based method using unstructured emergency department (ED) notes. LLM prompting strategies were developed for Llama 2 70B Chat [35], Mistral AI Mixtral 8×7B Instruct [36], GPT-3.5 turbo (version 0125) [37] and GPT-4 turbo (version 0125) [37]. The selection of LLMs at the time of experimentation represented the state of the art available in our Health Insurance Portability and Accountability Act (HIPAA)–authorized environments.

Setting

Boston Children's Hospital (BCH), a large Northeastern urban pediatric academic medical center, and the Indiana Health Information Exchange (IHIE) [38,39], a Midwestern statewide health information exchange network, were the study sites. Notes from BCH ED patients (aged 21 years and younger) and from IHIE ED patients (any age) with a COVID-19 diagnosis between March 1, 2020, and May 31, 2022, were eligible for inclusion into the study corpus.

Study Corpus

A study corpus of 613 notes was selected to ensure that it contained examples of rare symptoms. Apache cTAKES [40] was used to first identify positive symptoms in each note. At BCH, notes were then selected to include at least 30 positive examples for each of the 11 symptoms, as well as notes with no positive symptoms. These were used for a development corpus (103 notes) to select optimal strategies for each LLM, and a test corpus (202 notes) to measure accuracy. At IHIE, a validation corpus (308 notes) was randomly selected from a larger sample of 300 positive notes for each symptom and used to assess multisite generalizability in a setting comprising many health care locations.

Ground Truth

Three BCH experts collaboratively defined inclusion and exclusion criteria for symptom annotation guidelines [4]. They performed iterative cycles of independent chart review, collaborative adjudication of disagreements, and collaborative refinement of symptom annotation guidelines until a consensus was reached. Expert pairs reviewed notes from their own site. Interrater reliability was assessed with the kappa statistic [41,42] (overall mean 0.96, SD 0.07; details in Multimedia Appendix 1).

Measures

Eleven symptoms related to infectious respiratory disease were measured: congestion or runny nose, cough, diarrhea, dyspnea (shortness of breath), fatigue, fever or chills, headache, loss of taste or smell, muscle or body aches, nausea or vomiting, and sore throat.

F_1-scores, precision, and recall were calculated for each symptom and for all symptoms combined [42]. Micro F_1-scores were used, rather than macro F_1-scores, to allow for stronger competition from *ICD-10*–based metrics, which were quite poor for some symptoms. McNemar tests were used to evaluate LLM versus *ICD-10*–based performance.

With an overall α of .05, a Bonferroni adjustment for 12 comparisons (11 symptoms plus no symptoms) set the threshold at $P<.0042$.

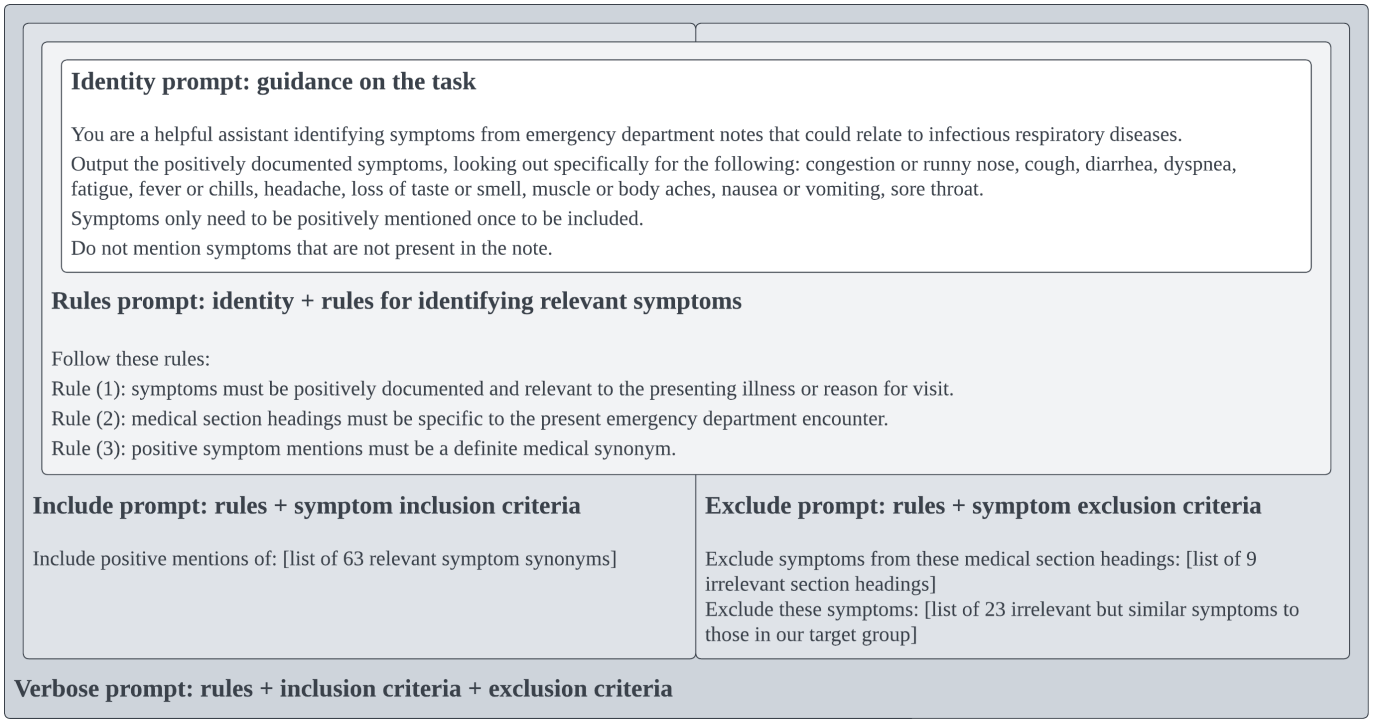
Comparator

ICD-10 codelists (Multimedia Appendix 2) [4] for each symptom were compiled by 3 experts at BCH using online resources [43,44]. The panel collaboratively reviewed whether each candidate code met the inclusion or exclusion criteria defined in the symptom annotation guidelines. *ICD-10* codes recorded at the time of ED discharge were matched against the final symptom codelists.

Prompt Engineering

For each LLM, 5 chart review prompts [45] were developed to follow symptom annotation guidelines. An overview is shown in Figure 1. Prompts ranged in complexity from an identity prompt, where LLMs were instructed to assume the identity of a chart reviewer, to a verbose prompt containing symptom-specific synonyms and inclusion and exclusion criteria. The 5 prompts were evaluated across 4 output parsing pipelines, yielding 20 prompting strategies for each LLM (Multimedia Appendix 3). All pipelines normalized LLM output into a structured CSV format containing symptoms identified in each note. Of the 4 LLM output parsing pipelines, 2 handled text and 2 handled JSON.

Figure 1. Large language model prompts intended to reproduce chart review criteria. The identity prompt contains text present in every type of prompt. The rules prompt extends the identity prompt with basic chart review criteria. Include and exclude prompts extend the rules prompt with symptom-specific criteria. The verbose prompt combines all prompts to approximate the same chart review criteria used by human subject matter experts.



Ethical Considerations

The BCH Committee on Clinical Investigation (BCH IRB-P00043392) and the Indiana University Institutional Review Board (IU IRB 24673) each determined the study to be exempt from full human participant oversight. Waivers of consent were obtained to allow corpus extraction and chart review of ED notes for institutional review board–approved study personnel. Notes were not shared between sites and not anonymized prior to LLM processing. All analyses were conducted in HIPAA-secure environments. Open-source LLMs were hosted on premises. OpenAI models were hosted by Azure under a Business Associates Agreement for HIPAA compliance. Clinical notes and patient data have been omitted from figures, tables, and appendices; only aggregate statistics are reported.

Results

Demographic characteristics of patients with notes in the study corpus are presented in Multimedia Appendix 4. Frequencies for each symptom are in Multimedia Appendix 5. Figure 2 shows symptom identification F_1-scores in the development corpus using the optimal prompting strategy for each LLM. Optimal LLM instructions for chart review varied considerably among LLMs (Multimedia Appendix 6). Every LLM was optimized using the JSON output parsing pipeline.

The performance of each symptom identification method was evaluated with the test corpus using the F_1-score statistic. The *ICD-10*–based method performed worst (F_1-score=45.1%) compared to each LLM method. GPT-4 was the highest-scoring LLM (F_1-score=91.4%; P<.001), followed by GPT-3.5 (F_1-score=90.0%; P<.001), Llama2 (F_1-score=81.7%; P<.001), and Mixtral (F_1-score=83.5%; P<.001). Figure 3 shows symptom accuracy for the optimal prompting strategy of each LLM as well as the *ICD-10*–based method. Multimedia Appendix 7 contains method details and statistical results.

Using the validation corpus from IHIE, GPT-4 accuracy was measured with no further model training or fine-tuning of the BCH model. Accuracy improved for GPT-4 (F_1-score=94.0%; an absolute increase of 2.6%) but accuracy for the *ICD-10*–based method was worse (F_1-score=26.9%; an absolute decrease of 18.2%). Generalizability from the BCH to IHIE corpus was better for GPT-4 than the *ICD-10* method (P<.001). Figure 4 shows that GPT-4 accuracy was higher than the *ICD-10*–based method for all symptoms at both sites. Details and results are in Multimedia Appendix 8.

Figure 2. F_1-scores for large language model (LLM) optimal prompt strategies using the development corpus. Each color denotes a symptom identification method using its optimal prompting strategy: Llama2 (identity prompt), Mixtral (exclude prompt), GPT-3.5 (identity prompt), and GPT-4 (include prompt). Each of the 11 infectious disease symptoms are shown as well as a summary score for all symptoms. Overall, GPT-4 performed best, with a micro F_1-score of 90.8% for all symptoms combined.

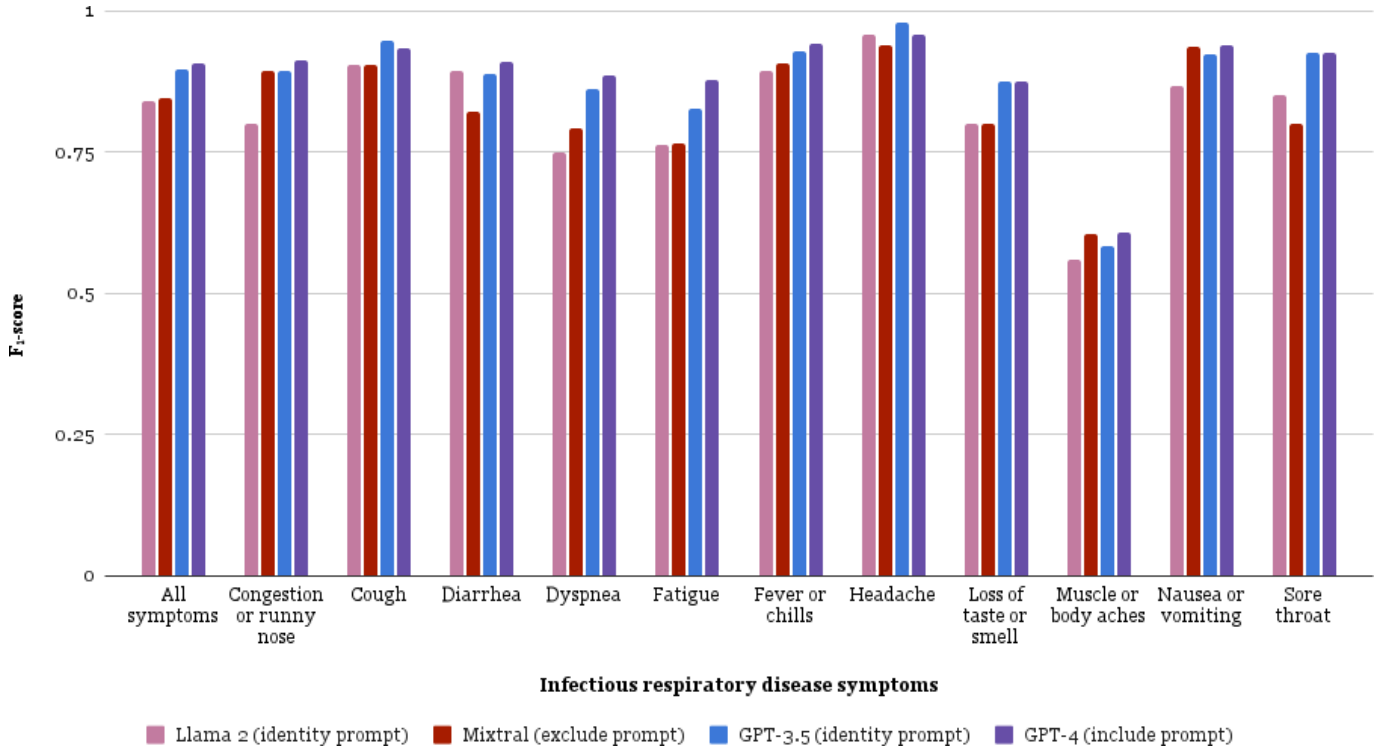


Figure 3. F_1-scores for large language model (LLM) optimal prompt strategies using the test corpus. Each color denotes a symptom identification method: Llama2, Mixtral, GPT-3.5, GPT-4, and the *ICD-10*–based method. Each of the 11 infectious respiratory disease symptoms are shown as well as a summary score for all symptoms. GPT-4 performed best, with an overall micro F_1-score of 91.4% for all symptoms combined. *ICD-10*: *International Classification of Diseases, Tenth Revision*.

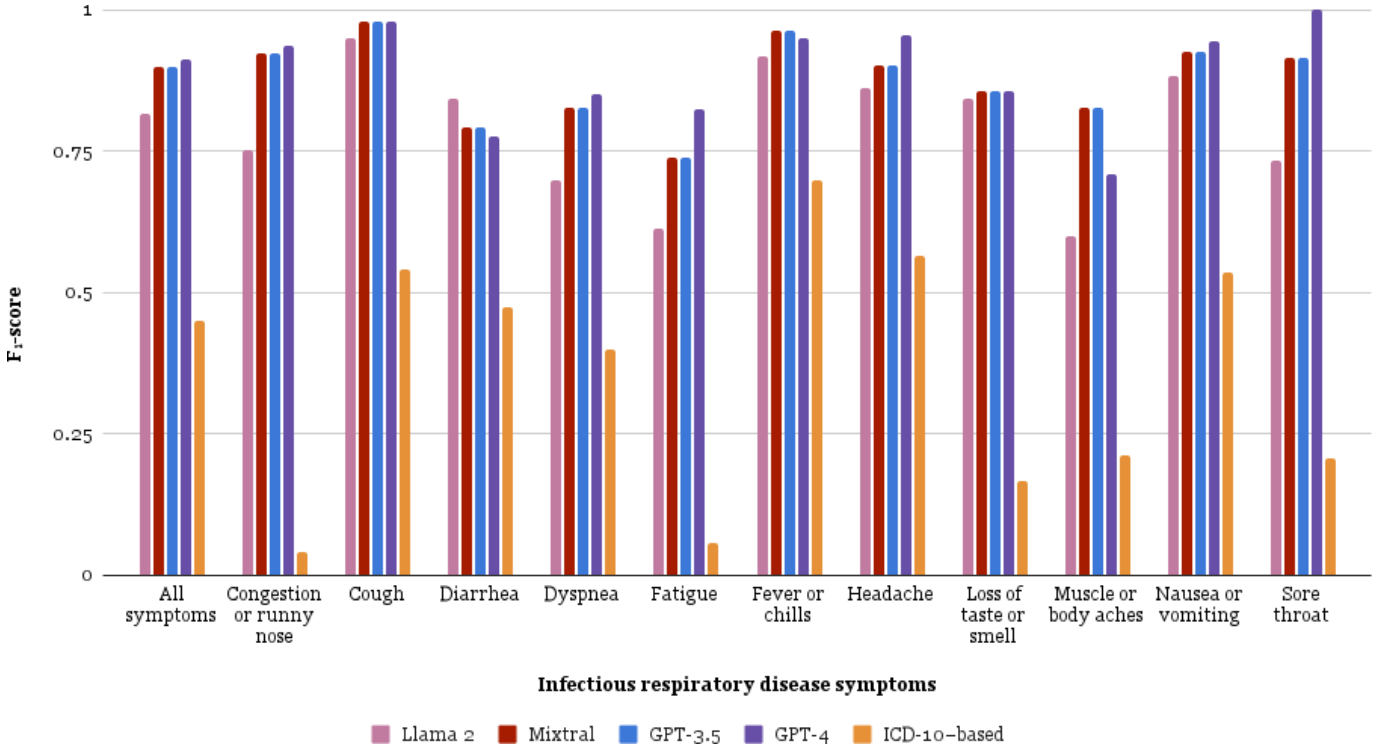
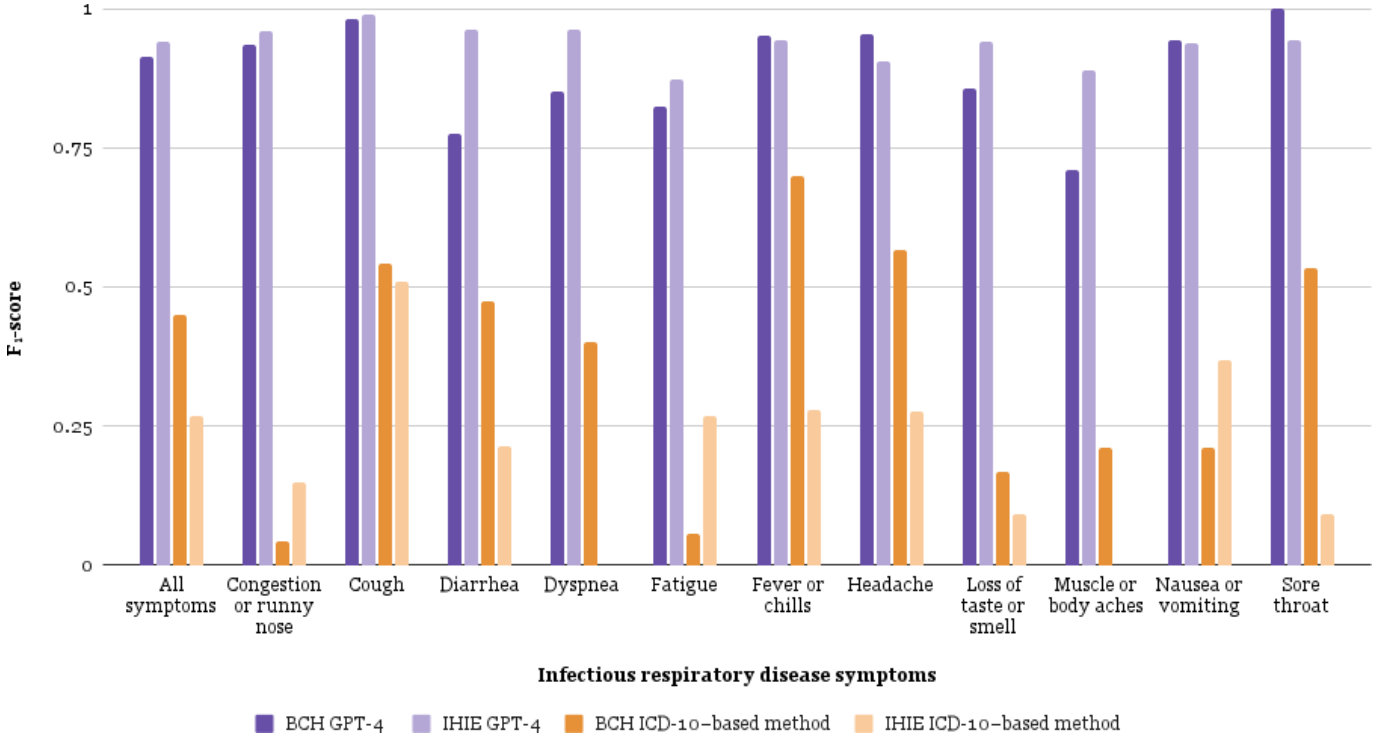


Figure 4. Generalizability of symptom identification accuracy across sites. One site is a large Northeastern urban pediatric academic medical center (BCH). The other is a Midwestern statewide health information exchange (IHIE) that provided data from 21 emergency departments. GPT-4 and the *ICD-10*–based method were compared. F_1-scores are shown for each method and symptom benchmarked against ground truth labels from chart reviews in test and validation corpora. BCH: Boston Children's Hospital; *ICD-10*: *International Classification of Diseases, Tenth Revision*; IHIE: Indiana Health Information Exchange.



Discussion

Principal Results

In this multisite study, LLM-based symptom identification consistently outperformed *ICD-10*–based methods for each infectious respiratory disease symptom evaluated. GPT-4 achieved the highest F_1-score, and results generalized well to an external validation corpus without customization. Low accuracy for *ICD-10*–based symptom identification and variability in multisite studies are consistent with prior literature [16,18,46].

Importantly, LLM strategies all used "zero-shot" prompts and required no site-specific artificial intelligence training, fine-tuning, or ground truth examples. The potential to reduce human labor represents a major advantage of LLM methods over traditional NLP methods that require human labor to curate symptom concept dictionaries, annotate ground truth examples, and calibrate at each health care site.

Limitations and Future Work

This study focused specifically on identifying symptoms of infectious respiratory diseases. However, generalizability of LLMs to other clinical domains and broader symptom categories remains to be validated. Furthermore, while GPT-4 performance was excellent in a validation corpus from 21 EDs, other settings, including primary care, should be studied. Other LLM models such as Google Gemini, Anthropic Claude, and DeepSeek R1 were not available for use in our HIPAA-secure settings. Future work should explore recent LLM developments. For example, the latest agentic methods could generalize to new symptom sets dynamically through multistage interactions with users.

It was beyond the scope of this study to estimate symptom prevalence in the study population. However, given outstanding LLM performance, one could approximate true prevalence from apparent prevalence in electronic health records [47]. Future work is needed to incorporate LLM-assisted chart review and pattern recognition. Doing this in real time, at a national scale, would truly improve public health efforts [3,47,48].

Conclusions

Our findings underscore the potential of LLMs to address gaps in traditional methods to identify symptoms in health records, paving the way for advancements in syndromic biosurveillance and other use cases. LLMs can be instructed to mimic human chart reviewers with high accuracy. Future work should assess broader symptom types and health care settings.

Acknowledgments

Support for this study was provided by the Advanced Research Projects Agency for Health (ARPA-H) and the National Center for Advancing Translational Sciences (NCATS; 75N95023D00001, 75N95023F00019, and 75N95024F00013), National Institutes of Health (U01TR002623), the Office of the National Coordinator for Health Information Technology (ONC; 90AX0031 and 90C30007), and the Centers for Disease Control and Prevention (CDC) of the US Department of Health and Human Services as part of a financial assistance award. Generative artificial intelligence was not used to design or conduct this study or prepare the manuscript.

Data Availability

The emergency department (ED) notes analyzed during this study are protected under privacy and confidentiality regulations and cannot be shared openly. However, the prompts, supporting datasets (excluding ED notes) and detailed methodological descriptions are available to facilitate reproducibility from the corresponding author or from the repository on GitHub [34]. Access will be granted in accordance with ethical and institutional guidelines. All code, including large language model prompts, and results are freely available on GitHub [34].

Authors' Contributions

As per guidelines of the International Committee of Medical Journal Editors, all authors contributed to the conceptualization or design of the study and the acquisition, analysis, or interpretation of the data as follows: conceptualization (KDM, AJM, DP, AG, TM, JRJ, and DG), data curation (AJM, AG, HC, and SM), formal analysis (AJM and DP), funding acquisition (KDM), investigation (AJM, DP, AG, DET, HC, and SM), methodology (KDM, AJM, DP, DG, TM, JRJ, and KLO), project administration (JRJ, BED, and DET), software (DP, AJM, MT, and DG), supervision (KDM and TM), validation (BED, DET, HC, and SM), and visualization (DP and JRJ). In terms of manuscript preparation, drafts of the manuscript were written by AJM, DP, KDM, and KLO; critical input was solicited from all authors and incorporated. All authors reviewed, edited, and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Kappa agreement scores are shown for human expert chart reviewers at 2 sites (BCH and IHIE) At BCH, a third reviewer (AG) was available for measurement. IHIE had 2 reviewers. BCH: Boston Children's Hospital; IHIE: Indiana Health Information Exchange.

[XLSX File (Microsoft Excel File), 146 KB-Multimedia Appendix 1]

Multimedia Appendix 2

ICD-10 codes for symptoms of infectious respiratory disease. *ICD-10*: *International Classification of Diseases, Tenth Revision*.
[XLSX File (Microsoft Excel File), 12 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Prompting templates and strategies, including verbatim prompting templates used across different large language models to conform to their instruction tuning specifications, as well as all 20 prompting strategies examined using our development corpus.
[PDF File (Adobe File), 569 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Patient demographics across sites and corpora. Demographics reported include binned age groups, administrative sex, and patient-reported race.
[XLSX File (Microsoft Excel File), 210 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Frequency of suspected symptoms at the time of corpus construction. To ensure decent distribution of symptoms across each corpus, samples were based on cTAKES-annotated symptom mentions. This aimed to guarantee that, even for rare symptoms, a bare minimum of symptoms likely to be positive was included in all corpora.
[XLSX File (Microsoft Excel File), 140 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Large language model (LLM) symptom identification performance using the development corpus. F_1-scores are provided for all 80 combinations of models and strategies. Detailed performance results are provided for each LLM using their best performing LLM strategy.
[XLSX File (Microsoft Excel File), 106 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Symptom identification performance using the test corpus from BCH and the best strategy identified for each LLM. Metrics include F_1-score, sensitivity, specificity, positive predictive value, negative predictive value, as well as raw counts of true positives, false negatives, true negatives, and false positives across all symptoms individually and aggregated. McNemar significance tests compare *ICD-10*–based symptom identification to LLM-based symptom identification. BCH: Boston Children's Hospital; *ICD-10*: *International Classification of Diseases, Tenth Revision*; LLM: large language model.
[XLSX File (Microsoft Excel File), 228 KB-Multimedia Appendix 7]

Multimedia Appendix 8

LLM symptom identification performance using the validation corpus. Sheets provided show detailed results for GPT-4 and *ICD-10* (including performance metrics and raw counts) as well as tables comparing the performance of both the validation and test corpora. McNemar significance tests compare *ICD-10*–based symptom identification to LLM-based symptom identification. *ICD-10*: *International Classification of Diseases, Tenth Revision*; LLM: large language model.
[XLSX File (Microsoft Excel File), 149 KB-Multimedia Appendix 8]

References

1. Chen A, Chen DO, Tian L. Benchmarking the symptom-checking capabilities of ChatGPT for a broad range of diseases. J Am Med Inform Assoc. Sep 1, 2024;31(9):2084-2088. [doi: 10.1093/jamia/ocad245] [Medline: 38109889]
2. Harskamp RE, De Clercq L. Performance of ChatGPT as an AI-assisted decision support tool in medicine: a proof-of-concept study for interpreting symptoms and management of common cardiac conditions (AMSTELHEART-2). Acta Cardiol. Mar 15, 2024;79(3):358-366. [doi: 10.1080/00015385.2024.2303528]
3. Mandl KD, Gottlieb D, Mandel JC, et al. Push button population health: the SMART/HL7 FHIR Bulk Data Access application programming interface. NPJ Digit Med. Nov 19, 2020;3(1):151. [doi: 10.1038/s41746-020-00358-4] [Medline: 33299056]
4. McMurry AJ, Zipursky AR, Geva A, et al. Moving biosurveillance beyond coded data using AI for symptom detection from physician notes: retrospective cohort study. J Med Internet Res. Apr 4, 2024;26:e53367. [doi: 10.2196/53367] [Medline: 38573752]
5. Matheny ME, Yang J, Smith JC, et al. Enhancing postmarketing surveillance of medical products with large language models. JAMA Netw Open. Aug 1, 2024;7(8):e2428276. [doi: 10.1001/jamanetworkopen.2024.28276] [Medline: 39150707]

6. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. J Am Med Inform Assoc. Jan 1, 2020;27(1):3-12. [doi: 10.1093/jamia/ocz166] [Medline: 31584655]
7. Clark-Cutaia MN, Rivera E, Iroegbu C, Arneson G, Deng R, Anastasi JK. Exploring the evidence: symptom burden in chronic kidney disease. Nephrol Nurs J. 2022;49(3):227-255. [doi: 10.37526/1526-744X.2022.49.3.227] [Medline: 35802361]
8. Stubbs A, Kotfila C, Xu H, Uzuner Ö. Identifying risk factors for heart disease over time: overview of 2014 i2b2/UTHealth shared task Track 2. J Biomed Inform. Dec 2015;58 Suppl(Suppl):S67-S77. [doi: 10.1016/j.jbi.2015.07.001] [Medline: 26210362]
9. Ni Y, Kennebeck S, Dexheimer JW, et al. Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. J Am Med Inform Assoc. Jan 2015;22(1):166-178. [doi: 10.1136/amiajnl-2014-002887] [Medline: 25030032]
10. A study to compare two formulations of xylometazoline/dexpanthenol nasal spray for the treatment of nasal congestion. ClinicalTrials.gov. URL: https://clinicaltrials.gov/study/NCT03439436 [Accessed 2025-05-19]
11. Open trial of biofeedback for respiratory symptoms. ClinicalTrials.gov. URL: https://clinicaltrials.gov/study/NCT05973513 [Accessed 2025-05-19]
12. Gulden C, Mate S, Prokosch HU, Kraus S. Investigating the capabilities of FHIR search for clinical trial phenotyping. Stud Health Technol Inform. 2018;253:3-7. [Medline: 30147028]
13. Yarlas A, Maher S, Bayliss M, et al. The Inflammatory Bowel Disease Questionnaire in randomized controlled trials of treatment for ulcerative colitis: systematic review and meta-analysis. J Patient Cent Res Rev. 2020;7(2):189-205. [doi: 10.17294/2330-0698.1722] [Medline: 32377552]
14. ICD-10-CM. Classification of Diseases, Functioning, and Disability. 2024. URL: https://www.cdc.gov/nchs/icd/icd-10-cm/index.html [Accessed 2025-05-19]
15. Malden DE, Tartof SY, Ackerson BK, et al. Natural language processing for improved characterization of COVID-19 symptoms: observational study of 350,000 patients in a large integrated health care system. JMIR Public Health Surveill. Dec 30, 2022;8(12):e41529. [doi: 10.2196/41529] [Medline: 36446133]
16. Crabb BT, Lyons A, Bale M, et al. Comparison of International Classification of Diseases and Related Health Problems, Tenth Revision codes with electronic medical records among patients with symptoms of coronavirus disease 2019. JAMA Netw Open. Aug 3, 2020;3(8):e2017703. [doi: 10.1001/jamanetworkopen.2020.17703] [Medline: 32797176]
17. Koleck TA, Dreisbach C, Bourne PE, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J Am Med Inform Assoc. Apr 1, 2019;26(4):364-379. [doi: 10.1093/jamia/ocy173] [Medline: 30726935]
18. Hardjojo A, Gunachandran A, Pang L, et al. Validation of a natural language processing algorithm for detecting infectious disease symptoms in primary care electronic medical records in Singapore. JMIR Med Inform. Jun 11, 2018;6(2):e36. [doi: 10.2196/medinform.8204] [Medline: 29907560]
19. Karagounis S, Sarkar IN, Chen ES. Coding free-text chief complaints from a Health Information Exchange: a preliminary study. AMIA Annu Symp Proc. 2020;2020:638-647. [Medline: 33936438]
20. Zhou W, Dligach D, Afshar M, Gao Y, Miller TA. Improving the transferability of clinical note section classification models with BERT and large language model ensembles. Proc Conf Assoc Comput Linguist Meet. Jul 2023;2023:125-130. [Medline: 37786810]
21. Zhang F, Laish I, Benjamini A, Feder A. Section classification in clinical notes with multi-task transformers. In: Lavelli A, Holderness E, Jimeno Yepes A, Minard AL, Pustejovsky J, Rinaldi F, editors. Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI). 2022:54-59. [doi: 10.18653/v1/2022.louhi-1.7]
22. Gould DW, Walker D, Yoon PW. The evolution of BioSense: lessons learned and future directions. Public Health Rep. 2017;132(1_suppl):7S-11S. [doi: 10.1177/0033354917706954] [Medline: 28692386]
23. Reis BY, Kirby C, Hadden LE, et al. AEGIS: a robust and scalable real-time public health surveillance system. J Am Med Inform Assoc. 2007;14(5):581-588. [doi: 10.1197/jamia.M2342] [Medline: 17600100]
24. McMurry AJ, Gilbert CA, Reis BY, Chueh HC, Kohane IS, Mandl KD. A self-scaling, distributed information architecture for public health, research, and clinical care. J Am Med Inform Assoc. 2007;14(4):527-533. [doi: 10.1197/jamia.M2371] [Medline: 17460129]
25. Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. J Biomed Inform. Oct 2009;42(5):839-851. [doi: 10.1016/j.jbi.2009.05.002] [Medline: 19435614]
26. Lin C, Bethard S, Dligach D, Sadeque F, Savova G, Miller TA. Does BERT need domain adaptation for clinical negation detection? J Am Med Inform Assoc. Apr 1, 2020;27(4):584-591. [doi: 10.1093/jamia/ocaa001] [Medline: 32044989]

27. Miller T, Bethard S, Dligach D, Savova G. End-to-end clinical temporal information extraction with multi-head attention. Proc Conf Assoc Comput Linguist Meet. Jul 2023;2023:313-319. [Medline: 37780680]
28. Wang H, Gao C, Dantona C, Hull B, Sun J. DRG-LLaMA: tuning LLaMA model to predict diagnosis-related group for hospitalized patients. NPJ Digit Med. Jan 22, 2024;7(1):16. [doi: 10.1038/s41746-023-00989-3] [Medline: 38253711]
29. He K, Mao R, Lin Q, et al. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. Inform Fusion. Jun 2025;118:102963. [doi: 10.1016/j.inffus.2025.102963]
30. Workman TE, Ahmed A, Sheriff HM, et al. ChatGPT-4 extraction of heart failure symptoms and signs from electronic health records. Prog Cardiovasc Dis. 2024;87:44-49. [doi: 10.1016/j.pcad.2024.10.010] [Medline: 39442600]
31. Pugliese G, Maccari A, Felisati E, et al. Are artificial intelligence large language models a reliable tool for difficult differential diagnosis? An a posteriori analysis of a peculiar case of necrotizing otitis externa. Clin Case Rep. Sep 2023;11(9):e7933. [doi: 10.1002/ccr3.7933] [Medline: 37736475]
32. Maillard A, Micheli G, Lefevre L, et al. Can chatbot artificial intelligence replace infectious diseases physicians in the management of bloodstream infections? A prospective cohort study. Clin Infect Dis. Apr 10, 2024;78(4):825-832. [doi: 10.1093/cid/ciad632] [Medline: 37823416]
33. Nori H, Lee YT, Zhang S, et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. arXiv. Preprint posted online on Nov 28, 2023. [doi: 10.48550/arXiv.2311.16452]
34. Smart-on-fhir/infectious-symptoms. GitHub. 2025. URL: https://github.com/smart-on-fhir/infectious-symptoms-llm-study [Accessed 2025-05-19]
35. Meta Llama 2. Meta Llama. URL: https://llama.meta.com/llama2/ [Accessed 2025-05-19]
36. Mixtral of experts. Mistral AI. 2023. URL: https://mistral.ai/news/mixtral-of-experts/ [Accessed 2025-05-19]
37. GPT-4. OpenAI. URL: https://openai.com/index/gpt-4-research/ [Accessed 2025-05-19]
38. Overhage JM, Kansky JP. The Indiana Health Information Exchange. In: Health Information Exchange. Elsevier; 2023:471-487. [doi: 10.1016/B978-0-323-90802-3.00022-8] ISBN: 9780323908023
39. Williams KS, Rahurkar S, Grannis SJ, Schleyer TK, Dixon BE. Evolution of clinical Health Information Exchanges to population health resources: a case study of the Indiana network for patient care. BMC Med Inform Decis Mak. Feb 24, 2025;25(1):97. [doi: 10.1186/s12911-025-02933-9] [Medline: 39994604]
40. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010;17(5):507-513. [doi: 10.1136/jamia.2009.001560] [Medline: 20819853]
41. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276-282. [Medline: 23092060]
42. Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. J Am Med Inform Assoc. 2005;12(3):296-298. [doi: 10.1197/jamia.M1733] [Medline: 15684123]
43. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. Jan 1, 2004;32(Database issue):D267-70. [doi: 10.1093/nar/gkh061] [Medline: 14681409]
44. ICD-10-CM. URL: https://icd10cmtool.cdc.gov/ [Accessed 2025-05-19]
45. Wang L, Chen X, Deng X, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. NPJ Digit Med. Feb 20, 2024;7(1):41. [doi: 10.1038/s41746-024-01029-4] [Medline: 38378899]
46. Nelson SJ, Yin Y, Trujillo Rivera EA, et al. Are ICD codes reliable for observational studies? Assessing coding consistency for data quality. Digit Health. 2024;10:20552076241297056. [doi: 10.1177/20552076241297056] [Medline: 39493629]
47. Mandl KD, Kohane IS. Federalist principles for healthcare data networks. Nat Biotechnol. Apr 2015;33(4):360-363. [doi: 10.1038/nbt.3180] [Medline: 25850061]
48. McMurry AJ, Gottlieb DI, Miller TA, et al. Cumulus: a federated electronic health record-based learning system powered by Fast Healthcare Interoperability Resources and artificial intelligence. J Am Med Inform Assoc. Aug 1, 2024;31(8):1638-1647. [doi: 10.1093/jamia/ocae130] [Medline: 38860521]

Abbreviations

BCH: Boston Children's Hospital
ED: emergency department
HIPAA: Health Insurance Portability and Accountability Act
***ICD-10*:** *International Classification of Diseases, Tenth Revision*
IHIE: Indiana Health Information Exchange
LLM: large language model
NLP: natural language processing

Edited by Andrew Coristine; peer-reviewed by Karthik Sarma, Michael Dohopolski, varun kumar nomula; submitted 23.02.2025; final revised version received 17.06.2025; accepted 18.06.2025; published 31.07.2025

Please cite as:
McMurry AJ, Phelan D, Dixon BE, Geva A, Gottlieb D, Jones JR, Terry M, Taylor DE, Callaway H, Manoharan S, Miller T, Olson KL, Mandl KD
Large Language Model Symptom Identification From Clinical Text: Multicenter Study
J Med Internet Res 2025;27:e72984
URL: https://www.jmir.org/2025/1/e72984
doi: 10.2196/72984