Word count for abstract: 299
Word count for text: 3173

Identifying and characterizing a chronic cough cohort through electronic health records

Running head: Chronic cough characterized via electronic records

Michael Weiner, MD[1,2,3]
Paul R. Dexter, MD[1,2,4]
Kim Heithoff, ScD[5]
Anna R. Roberts[1]
Ziyue Liu, PhD[2]
Ashley Griffith, MHA[1]
Siu Hui, PhD[1]
Jonathan Schelfhout, PhD[6]
Peter Dicpinigaitis, MD[7]
Ishita Doshi, PhD[6]
Jessica P. Weaver, PhD, MPH[6]

[1]Regenstrief Institute, Inc., Indianapolis, Indiana.
[2]Indiana University, Indianapolis, Indiana.
[3]Center for Health Information and Communication, U.S. Department of Veterans Affairs, Veterans Health Administration, Health Services Research and Development Service CIN 13-416, Richard L. Roudebush VA Medical Center, Indianapolis, Indiana.
[4]Eskenazi Health, Indianapolis, Indiana.
[5]Monument Analytics, Baltimore, Maryland.
[6]Merck & Co., Inc., Kenilworth, New Jersey.
[7]Albert Einstein College of Medicine and Montefiore Medical Center, Bronx, New York.

Correspondence:
Michael Weiner, M.D., M.P.H.
Regenstrief Institute, Inc. • 1101 West 10th Street
Indianapolis, IN 46202-4800 • U.S.A.
Tel. +1 317-274-9026 • mailto:mw@cogit.net

Jessica Weaver and Jonathan Schelfhout are employees of, and have stock ownership of, Merck & Co., Inc. Kim Heithoff and Ishita Doshi were employees of, and were stock holders of, Merck & Co., Inc. Michael Weiner's potentially relevant stock holdings are as follows: Abbvie, Inc., Accuray Inc., Allscripts Healthcare Solutions, Amgen, Inc., Boston Scientific Corp., Bristol Myers Squibb, Crispr Therapeutics Ag Com, Express Scripts Hldg Co., General Electric Co., Globus Med, Inc., Integer Hldgs Corp Com, Integra Lifesciences Holdings Corp., Intl Business Mach, Johnson & Johnson, Mallinckrodt PLC, Mead Johnson Nutrition, Medtronic PLC, Metlife Inc Com, Mylan N V SHS Euro, Novo-Nordisk A S ADR, Nuvasive, Inc., Orthofix Intl N.V., Perspecta Inc Com, Pfizer, Inc., Resmed, Inc., Roche Hldg Ltd., Seaspine Hldgs Corp., Senseonics Hldgs, Inc., Stryker Corp., Teva Pharmaceutical Industries, Varex Imaging Corp., Varian Med Sys, Inc., Walgreens Boots Alliance, Inc., Zimmer Biomet Hldgs, Inc., Zoetis, Inc. The study was sponsored by Merck, Sharp and Dohme. The following co-authors received

This is the author's manuscript of the article published in final edited form as:

Weiner, M., Dexter, P. R., Heithoff, K., Roberts, A. R., Liu, Z., Griffith, A., Hui, S., Schelfhout, J., Dicpinigaitis, P., Doshi, I., & Weaver, J. P. (2021). Identifying and Characterizing a Chronic Cough Cohort Through Electronic Health Records. Chest, 159(6), 2346–2355. https://doi.org/10.1016/j.chest.2020.12.011

funding for their work through Merck, Sharp and Dohme: Michael Weiner, Paul Dexter, Anna R. Roberts, Ziyue Liu, Ashley Griffith, and Siu Hui. Peter Dicpinigaitis has no conflicts.

Funding Source: Merck & Co., Inc.

Parts of this work were presented at the American College of Allergy, Asthma & Immunology 2018 Annual Scientific Meeting in Seattle, Washington on 17 November 2018. Results pertaining to referrals were accepted for presentation at the American Academy of Allergy, Asthma & Immunology Annual Meeting in Philadelphia, PA, 16 March 2020, but this meeting was cancelled due to coronavirus disease 2019.

ABBREVIATIONS

ACEI	Angiotensin-converting enzyme inhibitor
CC	Chronic cough
EHR	Electronic health record
ENT	Otolaryngology
NLP	Natural language processing
GERD	Gastroesophageal reflux disease
GI	Gastroenterology
ICD	International Classification of Diseases
INPC	Indiana Network for Patient Care
PPV	Positive predictive value
UACS	Upper airway cough syndrome

ABSTRACT

Background. Chronic cough (CC) of eight or more weeks affects about 10% of adults and may lead to expensive treatments and reduced quality of life. Incomplete diagnostic coding complicates identifying CC in electronic health records (EHRs). Natural language processing (NLP) of EHR text could improve detection.

Research Question. We assessed NLP in identifying cough in EHRs, and characterized adults and encounters with CC.Study Design and Methods. A Midwestern EHR system identified patients aged 18-85 during 2005-2015. NLP evaluated text notes except prescriptions and instructions, for mentions of cough. Two clinicians and a biostatistician reviewed twelve sets of 50 encounters each, with iterative refinements, until the positive predictive value for cough encounters exceeded 90%. NLP, ICD-10, or medication identified cough. Three encounters spanning 56 to 120 days defined CC. Descriptive statistics summarized patients and encounters, including referrals.

Results. Optimizing NLP required identifying and eliminating cough denials, instructions, and historical references. Of 235,457 cough encounters, 23% had a relevant diagnostic code or medication. Applying chronicity to cough encounters identified 23,371 patients (61% women) with CC. NLP alone identified 74% of these patients; diagnoses or medications alone identified 15%. The positive predictive value of NLP in the reviewed sample was 97%. Referrals for cough occurred in 3.0% of patients; pulmonary medicine was most common initially (64% of referrals).

Limitations. Some patients with diagnosis codes for cough, encounters intervals greater than four months, or multiple acute cough episodes may have been misclassified.

Interpretation. NLP successfully identified a large cohort with CC. Most patients were identified through NLP alone, rather than diagnoses or medications. NLP improved detection of patients

nearly seven-fold, addressing the gap in ability to identify and characterize CC disease burden. Nearly all cases appeared to be managed in primary care. Identifying these patients is important for characterizing treatment and unmet needs.

Chronic cough (CC) of eight or more weeks is common, affecting about 10% of adults (1). Prevalence estimates range from about 3% to 40%, depending on population (2–4). Patients with CC have reported frustration, irritability, or anger; frequent physician visits and testing; pain and social impact; and sleep disturbances (5). Many are treated empirically, though the presumptive diagnosis may be incorrect (6). The most common causes are gastroesophageal reflux disease (GERD), asthma, and postnasal drip (upper airway cough syndrome, UACS). Additional causes are respiratory infection, smoking, angiotensin-converting enzyme inhibitors (ACEI), and bronchitis. Refractory or unexplained CC, (7,8), defined by a CC that persists despite assessment and treatment according to guidelines (9–13), represents an important knowledge gap in the understanding of CC diagnosis and management. Although the diagnosis and treatment of chronic cough have been studied (9,14–18), electronic health records (EHRs) provide an opportunity to understand many aspects of CC without the expense of a prospective study. If valid EHR criteria for identifying CC can be identified, EHRs can be used to expand our ability to target patients for further treatment, and to measure outcomes. In addition, EHRs may inform us about the "natural history" in the course of medical care, in terms of characteristics, comorbidities, evaluations, treatment, and health services.

Developing a method to identify cough accurately and completely is problematic, because cough has been perceived as a symptom rather than a distinct medical condition. Therefore, it may, in the U.S., be underrepresented in structured data such as diagnostic codes, as a “rule out” diagnosis such as pharyngitis would instead often be used in outpatient settings. Unstructured, text-based notes from clinical encounters, however, are likely to refer to cough when patients seek treatment for it. Therefore, identifying ways to use both structured and unstructured data together to identify CC, which has no International Classification of Diseases

(ICD) code, may improve sensitivity of finding cases. We sought to develop and examine a technique of using structured and unstructured EHR data to identify cough encounters and patients with CC, and to apply the technique to characterize adults with CC, and their medical encounters.

METHODS

Indiana University's Institutional Review Board approved the study (protocol 1705384100). A waiver of consent was approved, due to the low risk of examining de-identified medical records retrospectively, and reporting anonymously.

Study design, setting, and participants

This was a retrospective cohort study. Patient data were available from medical records and a health information exchange. Part of a multi-institution regional health information exchange, the Indiana Network for Patient Care (INPC) contains more than 10 billion clinical data elements representing more than 100 hospitals in 38 Indiana health systems, with more than 36,000 providers and 13.5 million patients (19). Patients' INPC encounters were identified from Eskenazi Health and Indiana University Health. Eskenazi Health is a tax-supported institution serving a predominantly urban area in central Indiana. Indiana University Health is a statewide academic medical enterprise headquartered in central Indiana. The study population represents many socioeconomic and clinical dimensions.

The initial cohort included all patients of at least 18 and less than 86 years of age at the time of encounter or medication prescription, who received care at either institution between October 2005 and September 2015. We selected the upper age limit, because patients who develop their first CC after this age may have atypical etiologies.

Outcomes

We summarized the extent to which structured and unstructured EHR data identified patients with CC as confirmed via manual review of a random sample of records. We used descriptive statistics to summarize demographics, clinical encounters, comorbidity including Charlson index, and cough-related referrals to specialists in pulmonary medicine, otolaryngology

(ENT), gastroenterology (GI), and allergy. These referrals contained the term “cough” in the reason for referral.

Study procedures

Identifying cases of CC involved two steps: first, we identified medical encounters with cough (“cough encounter”)— defined as any outpatient encounter with structured or unstructured data indicating cough—and then used those encounters to identify CC based on chronicity (**Figure 1**).

Identifying encounters with cough. In identifying encounters, specificity was prioritized over sensitivity. To identify cough, two approaches were used. In the first approach, we used structured (coded) data. Cough was identified when an ICD code for cough—ICD-9 code 786.2 ("Cough") or ICD-10 code R05 ("Cough")—was present, or when benzonatate or dextromethorphan was prescribed in the outpatient setting, according to a National Drug Code for any form of the drug. These two drugs were selected based on their therapeutic specificity for cough. Although codeine, amitriptyline, pregabalin, gabapentin, certain inhalers or nasal sprays, and other drugs are sometimes used to treat cough, they are not specific for cough, and seemed unlikely to be used to treat cough in the absence of documentation of cough in the EHR.

Natural language processing (NLP). The second approach to identifying cough involved the use of unstructured, narrative ("free text") data, which were abstracted from clinical notes such as progress notes, encounter summaries, interpretations of diagnostic tests, summaries of medical procedures, or other written reports. To analyze text, we used a combination of inspection and nDepth, Regenstrief Institute’s tool for NLP and text mining (20). nDepth includes functions to determine the context and negation of terms. NLP is likely to identify cough that is not identified through structured data alone, because—as with many other syndromes—many encounters of patients with cough will be coded for other diagnoses instead; general coding guidance indicates that diagnostic coding for symptoms should occur only when no diagnosis classifiable elsewhere is identified. We used NLP to search documents of clinical encounters, for a text-based mention of "cough or "expectorat"—other letters could immediately follow either term—in any medical encounter. This would include terms such as “refractory cough”, since “cough” (in that case) is included in the phrase. We excluded encounters in which the term was negated (e.g., "Patient denied cough"), the person experiencing the cough was judged to be someone other than the patient (e.g., a family member with cough), or the entire document was an instructional document for patients, which might, therefore, indicate how to treat a cough that might not be present.

There was not a training set *per se*. Instead, a sample of matching clinical notes was then inspected, to identify additional relevant words, which were added and iteratively searched, to generate additional relevant records. A sample of records matching the search criteria was then examined manually, to confirm that the records referred to episodes of cough. In cases where an episode of cough was not confirmed (i.e., false-positive cases), the NLP criteria were revised to try to eliminate the false positives. Instructional phrases were also excluded. By validating a sample of notes, refinements of the NLP approach were made iteratively.

Identifying patients with CC. Using cough encounters as defined above, a patient with CC was then defined as a patient with any three cough encounters such that the gap between first and last cough was between 56 and 120 days. The use of 120 days as a maximum "gap" perhaps increased the chance that the three coughs shared a common etiology. The index date was then defined as the first of the three cough encounter dates. We excluded patients with no encounters within 183 days (approximately six months) before the index date, or with no encounters within

365 days after the index date, because these patients would be less likely to be receiving, around the index date, primary medical care within the institution. To identify additional patients who did not meet our criteria for CC but might have had a CC, we counted the number of additional patients with any mention of "chronic cough", "persistent cough", or "persistant cough" (a common spelling error). We excluded patients with an order for any angiotensin converting enzyme inhibitor on an encounter during the study period that preceded CC.

Additional refinements. Following the initial approach described above, we examined the findings and undertook additional steps. Using medical-record notes containing "cough" or "expectorat", we identified additional word forms and terms indicating cough. We excluded records reflecting general instructions to the patient, or denial of cough (i.e., negation). Twelve sets of 50 encounters at a time were reviewed by two or three reviewers, with refinements created between reviews, until the positive predictive value (PPV) for cough encounters exceeded 90%. Final criteria are shown in the **e-Appendix**.

Role of the Funding source

This project was funded by Merck & Co., Inc., which participated in study design and reporting. The project focused on no specific products of the sponsor.

RESULTS

The method’s positive predictive value for cough was 97% in sampled reviews, with identification of 235,457 encounters with cough by structured or unstructured data. Among these cough encounters, 19% had a diagnostic code or medication but did not mention cough in the note; an additional 4% had structured data and NLP indication of cough. The remaining 77% of encounters were identified only through NLP (**Figure 2**). The chronicity algorithm applied to these encounters identified 23,371 patients with CC. If NLP were not available, only 3393 (15%) of these patients would have been identified as having CC. Therefore, the use of NLP increased the cohort size by almost seven-fold.

Within the entire CC cohort, the number of cough encounters for each patient identified through NLP is shown in **Table 1**, which indicates that 99% of patients in the cohort had at least one cough encounter identified through NLP. A small number (N=233) of patients were identified through medications and diagnoses, without NLP evidence of cough. Among this CC cohort, 10,895 (47%) patients had a specific text mention of chronic or persistent cough, in addition to meeting the criteria for the cohort. We identified an additional 53,319 patients with such a text mention but without meeting the cohort eligibility criteria; demographic data were unavailable for 63 of them. We used a chronicity algorithm instead of simply text mentions of chronic or persistent cough, because clinicians and patients use varying definitions. Our chronicity algorithm provides a definition of CC that is consistent with guidelines.

Characteristics of the cohort are shown in **Table 2**. The mean age was 54. The table includes frequencies of mention of chronic or persistent cough among a larger cohort—with the union of that group and our main cohort in the final column—because these might reflect CC despite not having met our main criteria. We focused on the main cohort because a note’s indication of “chronic cough” or “persistent cough” does not, in itself, confirm a duration of at least eight weeks. **Figure 3** shows the chronological order of referrals. Among the referral types studied, cough-related referrals were seen in 711 (3.0%) of the patients. Pulmonary was the most common referral and *initial* referral (initial N=439), followed by ENT (N=208), allergy (N=38), and GI (N=26). Pulmonary referrals were followed by ENT in 21 cases, ENT was followed by pulmonary in 13, and pulmonary followed by allergy in 9. In the year following the index date, the median time from index date to first referral was 63 days. Referrals for two specialties were ordered among 53 patients, three specialties for five patients, and all four specialties for two patients.

In comparing, within the main cohort, patients with or without any NLP-based evidence of cough, patients without NLP-identified episodes had a greater likelihood of commercial insurance (46% vs. 30%), as well as fewer encounters. In comparing the main cohort to the larger, extended cohort of 76,627 patients, the main cohort was similar in most characteristics but was more likely to have referral to pulmonary medicine (2.0% vs. 1.3%).

Table 3 shows frequencies of the Charlson index and specific medical conditions. Most (54%) had a Charlson index above zero. **Table 4** shows the most common medical conditions based on diagnostic codes. Hypertension was most common, followed by cough, hyperlipidemia, tobacco use disorder, diabetes, and GERD. These tables omit 2,955 patients with unavailable diagnosis data.

DISCUSSION

To our knowledge, this is the first report of the use of NLP to identify and examine CC, and the largest assembled cohort of patients with CC. The study identified 23,371 patients with CC, using a combination of structured and unstructured data. The PPV (97%) appeared adequate, and builds confidence in the ability to identify many additional cases of true CC. As usual for NLP methods, achieving this required multiple iterations of refinement. NLP was ultimately instrumental in identifying the 74% of cohort members without structured evidence of CC. The "added value" of NLP tends to be observed especially when the investigated condition is a symptom or syndrome that may reflect any of a set of diseases, rather than a single disease; when it reflects a condition without a dedicated diagnostic code; or when it reflects a condition that is likely to be under-coded. Our findings suggest that studies of CC that rely only on structured data are likely to miss most cases.

Using the ICD to identify a chronic condition based on diagnostic codes and chronicity requires accurate and consistent coding across encounters, as well as occurrence of medical encounters themselves; patients without encounters will not be identified. Furthermore, if patients pursue evaluation and treatment but outside the definition's time frame, they might also be missed. Neither ICD-9 nor ICD-10 contain codes for CC. Our study provided evidence that a group of patients outside our CC cohort but with mentions of "chronic cough" or "persistent cough" shared many characteristics with the main cohort.

Many other studies of CC have examined only small cohorts, such as of 100 or fewer people. One systematic review of the diagnosis and treatment of CC in adults found 23 studies identifying only 3,636 patients (21). NLP can help, but the specific words and phrases that are most productive in identifying cases through NLP might differ among regions or systems due to various factors, such as geographic or cultural variations in terms used to describe cough, or potential for software-specific templates to generate predefined or system-unique passages of text repeatedly. Variations like this might explain certain differences that we found, such as in insurance and urbanicity, between patients identified through NLP and patients identified only through structured data. Further efforts to examine NLP among strata such as rural vs. urban areas, or academic vs. community-based practices, may be fruitful in identifying differences in NLP's performance. across strata

The automated approach used in this study to find CC is scalable to nearly any database containing the needed codes and text. Compared to prospective enrollment by a clinical expert, this method yields slightly reduced accuracy but has significant benefits in terms of generalizability to a population, and cohort size, which could facilitate examination of uncommon outcomes or complex referral patterns. It also allows for identifying potential CC earlier in the course of care, which is not possible if waiting for a diagnosis following one or more referrals.

Cough-related referrals were not common, but about one-fourth of the cohort had multiple primary care visits. Although we cannot easily confirm the visit's reason, the finding suggests that CC or its underlying causes are often initially managed in primary care. Referrals to pulmonary medicine were most common. This might indicate a prevalence of potentially pulmonary etiologies, or PCPs' greater need for evaluation and management of pulmonary issues, compared to others, for difficult cases. Although most referred patients were not referred to more than one specialty studied, the prevalence of multi-specialty evaluations suggests the possibility of initial or ongoing difficulty or uncertainty about CC's etiology. Turner and Bothamley reported that among 266 patients, the most frequent diagnoses were asthma (29%)

and GERD (22%), with 12% having unexplained CC (22). Good *et al.* reported, among 99 patients with CC referred to a respiratory center, that 55% had obstructive sleep apnea, and 32% had tracheobronchomalacia (6). Forty-two incorrect intake diagnoses were noted. Turner and Bothamley found that common diagnoses had often not been excluded in primary care, and estimated that 87% of patients could have been managed in primary care. Slovarp *et al.* reported that, among patients with refractory CC referred to receive behavioral therapy, the mean cough duration before the therapy, which improved 87% of patients, exceeded two years (23). Thus, selected patients may benefit from earlier referral. The need to refer patients must be carefully considered, as heterogeneity in "clinical profiles" can hinder referral decisions. Decision aids, as well as additional research about diagnostic accuracy, which patients to refer, and when to refer, may be useful in helping clinicians to identify such patients efficiently and accurately, and to refer them appropriately.

The study had several limitations. Patients with CC with encounters at intervals exceeding four months (i.e., less frequent) might not have been identified as having CC. Some cough encounters identified as one of a set of CC-defining encounters might have represented independent episodes of acute cough. Nonetheless, our approach, which relied on documentation in the medical record, matched a commonly used definition of CC. Diagnostic codes might not have been accurate in all cases. For example, some diagnosis codes are inadvertently "carried forward" despite resolution of the condition. Cough-related referrals may be underdetected due to missing or unavailable referral instructions. Referrals to speech pathology were not included. Narrative references to cough without using "cough" or "expectorat" as word stems may have been missed, but "cough" has few synonyms. Although the PPV was 97%, mentions of cough that might have been carried forward across encounters even as cough resolved might represent false positives.

INTERPRETATION

Of 235,457 cough encounters, 77% were identified only through NLP. Without NLP, only 15% of the patients with CC would have been identified. Although cough-related referrals to any of four cough-related specialties were uncommon, pulmonary medicine was the most common. Our study is the first of its kind to report the use of NLP in identifying a large cohort with CC. The method opens the door to robust population-based studies to characterize patients, practices, and outcomes, and to identify subgroups that may benefit from more intensive evaluations or treatments, as well as clinical decision aids.

ACKNOWLEDGEMENTS

The research reported here was supported by Merck & Co, Inc. Dr. Weiner is Chief of Health Services Research and Development at the Richard L. Roudebush Veterans Affairs Medical Center in Indianapolis, Indiana. The views expressed in this article are those of the authors and do not necessarily represent the views of the U.S. Department of Veterans Affairs. Parts of this work were presented at the American College of Allergy, Asthma & Immunology 2018 Annual Scientific Meeting in Seattle, Washington on 17 November 2018. We thank Vishal Bali for comments. Results pertaining to referrals were accepted for presentation at the American Academy of Allergy, Asthma & Immunology Annual Meeting in Philadelphia, PA, 16 March 2020, but this meeting was cancelled due to coronavirus disease 2019. Michael Weiner takes responsibility for the content of the article, including the data and analysis. Study design: MW, PRD, KH, ZL, SH, JS, PD, JPW. Acquisition and analysis: ARR, ZL. Drafting: MW. Interpretation, revisions, final approval, and accountability for accuracy and integrity: all authors.

REFERENCES

1. Song WJ, et al. The global epidemiology of chronic cough in adults: a systematic review and meta-analysis. Eur Respir J. 2015;45(5):1479–81. doi:10.1183/09031936.00218714
2. Chung KF, Pavord I d. Prevalence, pathogenesis, and causes of chronic cough. Lancet. 2008;371(9621):1364–74. doi:10.1016/S0140-6736(08)60595-4
3. Chamberlain SA, et al. The impact of chronic cough: a cross-sectional European survey. Lung. 2015;193(3):401–8. doi:10.1007/s00408-015-9701-2
4. Ford AC, d. Forman, Moayyedi P, Morice AH. Cough in the community: a cross sectional survey and the relationship to gastrointestinal symptoms. Thorax. 2006;61(11):975–9. doi:10.1136/thx.2006.060087
5. Kuzniar TJ, Morgenthaler TI, Afessa B, Lim KG. Chronic cough from the patient's perspective. Mayo Clin Proc. 2007;82(1):56–60. doi:10.4065/82.1.56
6. Good JT, Rollins DR, Kolakowski CA, Stevens AD, Denson JL, Martin RJ. New insights in the diagnosis of chronic refractory cough. Respir Med. 2018;141103–10. doi:10.1016/j.rmed.2018.06.024
7. Gibson PG, Vertigan AE. Management of chronic refractory cough. BMJ. 2015;351h5590. doi:10.1136/bmj.h5590.
8. Perotin J-M, et al. Managing patients with chronic cough: challenges and solutions. Ther Clin Risk Manag. 2018;141041–51. doi:10.2147/tcrm.s136036
9. Achilleos A. Evidence-based Evaluation and Management of Chronic Cough. Med Clin North Am. 2016;100(5):1033–45. doi:10.1016/j.mcna.2016.04.008

10. Brown KK. Chronic cough due to nonbronchiectatic suppurative airway disease (bronchiolitis): ACCP evidence-based clinical practice guidelines. Chest. 2006;129(1 Suppl):132S-137S. doi:10.1378/chest.129.1_suppl.132S

11. Gibson P, et al. Treatment of Unexplained Chronic Cough: CHEST Guideline and Expert Panel Report. Chest. 2016;149(1):27–44. doi:10.1378/chest.15-1496

12. Kardos P, et al. Guidelines of the German Respiratory Society for diagnosis and treatment of adults suffering from acute or chronic cough. Pneumologie. 2010;64(11):701–11. doi:10.1055/s-0030-1255526

13. Rosen MJ. Chronic cough due to bronchiectasis: ACCP evidence-based clinical practice guidelines. Chest. 2006;129(1 Suppl):122S-131S. doi:10.1378/chest.129.1_suppl.122S

14. Simpson CB, Amin MR. Chronic cough: state-of-the-art review. Otolaryngol Head Neck Surg. 2006;134(4):693–700. doi:10.1016/j.otohns.2005.11.014

15. Chung KF. Advances in mechanisms and management of chronic cough: The Ninth London International Cough Symposium 2016. Pulm Pharmacol Ther. 2017. doi:10.1016/j.pupt.2017.02.003

16. Birring SS, Floyd S, Reilly CC, Cho PSP. Physiotherapy and Speech and Language therapy intervention for chronic cough. Pulm Pharmacol Ther. 2017. doi:10.1016/j.pupt.2017.04.001

17. Satia I, Badri H, Al-Sheklly B, Smith JA, Woodcock AA. Towards understanding and managing chronic cough. Clin Med (Lond). 2016;16(Suppl 6):s92-s97. doi:10.7861/clinmedicine.16-6-s92

18. v. Poulose, Tiew PY, How CH. Approaching chronic cough. Singapore Med J. 2016;57(2):60–3. doi:10.11622/smedj.2016028

19. Indiana Health Information Exchange. Indiana Health Information Exchange [Internet]. 2019 [updated 2019; cited 2019 Jul 18]. Available from: https://www.ihie.org/

20. Regenstrief Institute, Inc. nDepth [Internet]. 2019 [cited 2019 Jun 21]. Available from: https://www.regenstrief.org/resources/ndepth/

21. French CT, et al. Assessment of Intervention Fidelity and Recommendations for Researchers Conducting Studies on the Diagnosis and Treatment of Chronic Cough in the Adult: CHEST Guideline and Expert Panel Report. Chest. 2015;148(1):32–54. doi:10.1378/chest.15-0164

22. Turner RD, Bothamley GH. Chronic cough and a normal chest X-ray - a simple systematic approach to exclude common causes before referral to secondary care: a retrospective cohort study. NPJ Prim Care Respir Med. 2016;2615081. doi:10.1038/npjpcrm.2015.81

23. Slovarp L, Loomis BK, Glaspey A. Assessing referral and practice patterns of patients with chronic cough referred for behavioral cough suppression therapy. Chron Respir Dis. 2018;15(3):296–305. doi:10.1177/1479972318755722

TAKE-HOME POINT

Study question. How much does natural language processing (NLP) contribute to identifying cases of chronic cough in electronic health records?

Results. With a positive predictive value of 97%, NLP alone identified 74% of a cohort of 23,371 patients with 235,457 cough encounters, while diagnoses or medications alone identified 15%, but optimizing NLP required identifying and eliminating cough denials, instructions, and historical references.

Interpretation. NLP successfully identified a large cohort with chronic cough, with most patients identified through NLP alone, rather than diagnoses or medications.

TABLES

Table 1. Patients with chronic cough (N=23,371), by total number of cough encounters identified through natural language processing (NLP)

Number of cough encounters identified through NLP	Number of patients	Percentage
0	233	1.00
1	708	3.03
2	2056	8.80
3	4220	18.06
4	3442	14.73
5	2605	11.15
6	1958	8.38
7	1525	6.53
8	1124	4.81
9	902	3.86
10	778	3.33
11 or more	3820	16.35

Table 2. Characteristics of patients with evidence of chronic cough

Characteristic	Main chronic-cough cohort (N=23,371)					Text included mention of "chronic cough" or "persistent cough" (some outside the main cohort)	Main cohort, OR text mention of "chronic cough" or "persistent cough"
	All in main cohort	Cough encounters by natural language processing					
		0	1	2	3 or more		
Number of patients	23,371	233	708	2,056	20,374	64,151	76,627
Age (N, %)							
18-19	404 (02)	3 (01)	22 (03)	46 (02)	333 (02)	1198 (02)	1505 (02)
20-29	1906 (08)	27 (12)	84 (12)	218 (11)	1577 (08)	3859 (06)	5275 (07)
30-39	2517 (11)	34 (15)	100 (14)	229 (11)	2154 (11)	5935 (09)	7584 (10)
40-49	3924 (17)	45 (19)	117 (17)	341 (17)	3421 (17)	10488 (16)	12613 (16)
50-59	5275 (23)	48 (21)	133 (19)	423 (21)	4671 (23)	15271 (24)	17900 (23)
60-69	4900 (21)	41 (18)	142 (20)	377 (18)	4340 (21)	14482 (23)	16735 (22)
70-79	3327 (14)	24 (10)	72 (10)	305 (15)	2926 (14)	9568 (15)	11091 (14)
80-85	1092 (05)	7 (03)	33 (05)	114 (06)	938 (05)	3174 (05)	3732 (05)
Gender (N, %)							
Female	15363 (66)	147 (63)	448 (63)	1296 (63)	13472 (66)	38478 (60)	46686 (61)
Male	8007 (34)	86 (37)	260 (37)	760 (37)	6901 (34)	25663 (40)	29931 (39)
Unknown	1 (00)				1 (00)	10 (00)	10 (00)
Race (N, %)							
White	16885 (72)	172 (74)	519 (73)	1476 (72)	14718 (72)	43045 (67)	52185 (68)
African American	3970 (17)	37 (16)	112 (16)	346 (17)	3475 (17)	9848 (15)	11932 (16)
Asian	137 (01)	4 (02)	8 (01)	18 (01)	107 (01)	291 (00)	367 (00)
Hispanic or Latino	126 (01)	0	2 (00)	9 (00)	115 (01)	343 (01)	418 (01)
American Indian or Alaska Native	17 (00)	0	3 (00)	2 (00)	12 (00)	41 (00)	48 (00)
Native	12 (00)	0	2 (00)	2 (00)	8 (00)	41 (00)	47 (00)

Hawaiian or Other Pacific Islander							
Other or unknown	2224 (10)	20 (09)	62 (09)	203 (10)	1939 (10)	10542 (16)	11630 (15)
Insurance (N, %)							
Commercial	7334 (31)	108 (46)	300 (42)	742 (36)	6184 (30)	19606 (31)	23782 (31)
Medicaid	4347 (19)	20 (09)	94 (13)	279 (14)	3954 (19)	12267 (19)	14633 (19)
Medicare	8400 (36)	53 (23)	197 (28)	696 (34)	7454 (37)	22687 (35)	26705 (35)
Other/unknown	2244 (10)	48 (21)	91 (13)	238 (12)	1867 (09)	6917 (11)	8150 (11)
Self-pay/uninsured	1046 (04)	4 (02)	26 (04)	101 (05)	915 (04)	2674 (04)	3357 (04)
Urban (N, %)	19456 (83)	213 (91)	625 (88)	1727 (84)	16891 (83)	52565 (82)	62874 (82)
Medical encounters, mean number per patient, within one year after index date (mean (SD), median (inter-quartile range))							
Inpatient	0.6 (1.6), 0 (1)	0.1 (0.5), 0 (0)	0.2 (0.7), 0 (0)	0.3 (0.9), 0 (0)	0.7 (1.7), 0 (1)	0.1 (0.8), 0 (0)	0.2 (1), 0 (0)
Outpatient	10.7(11.9), 7 (11)	4.9 (6), 3 (4)	5.3 (5.9), 4 (5)	6.1 (6.4), 4 (6)	11.5(12.3), 8 (11)	1.9 (6.4), 0 (0)	3.2 (7.8), 0 (2)
Emergency	1.7 (5), 0 (2)	0.3 (1.3), 0 (0)	0.7 (2.3), 0 (0)	1 (2.3), 0 (1)	1.9 (5.3), 0 (2)	0.2 (2), 0 (0)	0.5 (2.8), 0 (0)
Other outpatient	0.4 (1), 0 (0)	0.1 (0.6), 0 (0)	0.1 (0.5), 0 (0)	0.2 (0.7), 0 (0)	0.4 (1), 0 (0)	0.1 (0.4), 0 (0)	0.1 0.1 (0.5), 0 (0)

Table 3. Comorbidity of patients with evidence of chronic cough (N=20,416)

Characteristic	**Frequency, N (%)**
Charlson index	
0	9366 (46)
1	4975 (24)
2	2576 (13)
3+	3499 (17)
Medical condition	
Arthritis	2403 (12)
Chronic kidney disease	1240 (6.1)
COPD	2929 (14)
Dementia	76 (0.4)
Depression	2069 (10)
Diabetes mellitus	3120 (15)
Heart failure	971 (4.8)
Hyperlipidemia	5004 (25)
Hypertension	6510 (32)
Ischemic heart disease	1818 (8.9)

Table 4. Most common medical conditions of patients with evidence of chronic cough in the main cohort (N=20,416)

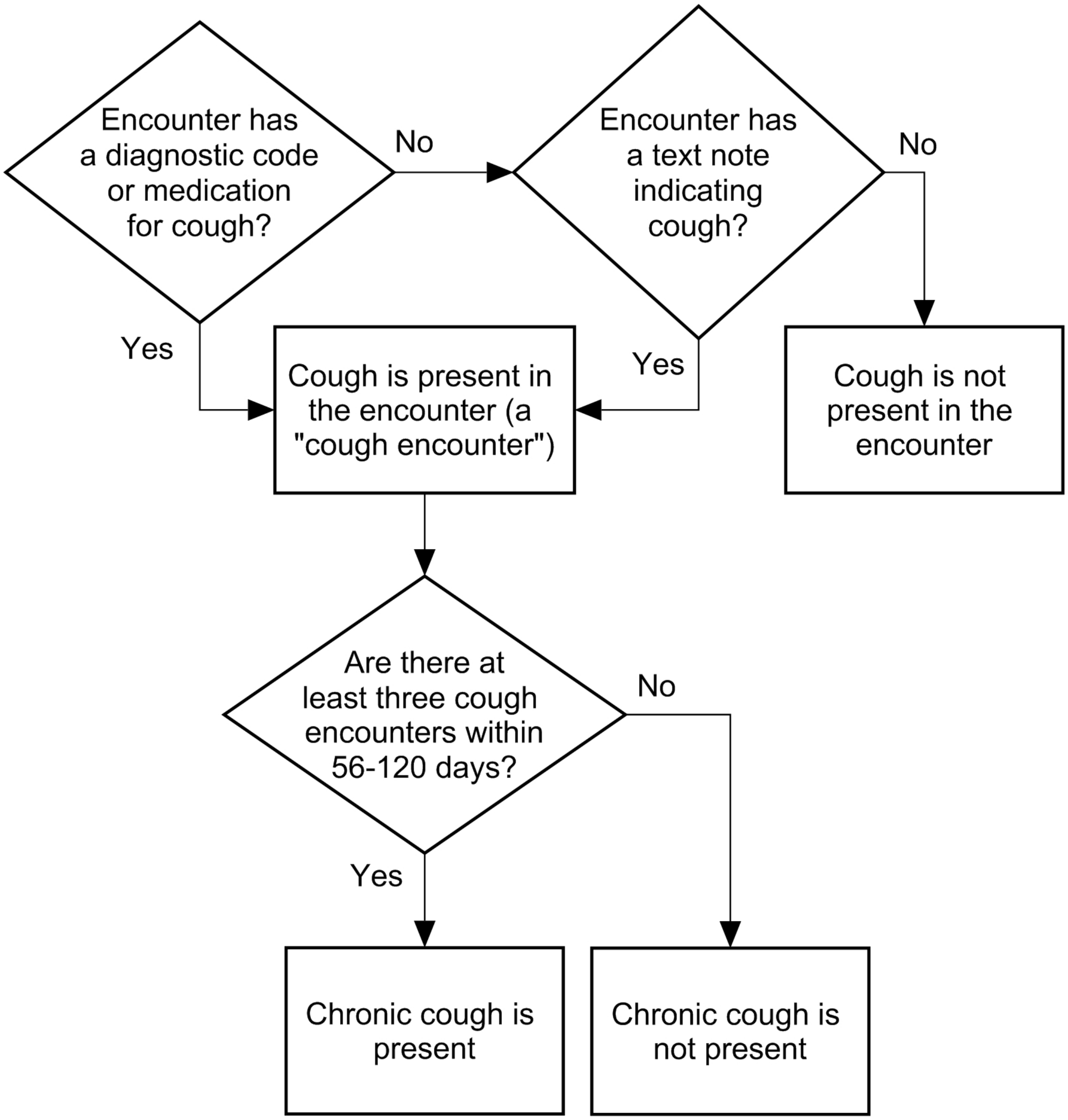
Medical condition	Number with ICD-9	ICD-9 rank	ICD-10 rank	Number with ICD-10
Hypertension, NOS	5581	1		
Cough	5311	2	2	3698
Hyperlipidemia NEC/NOS	5241	3	3	3512
Tobacco use disorder	3865	4	4	3368
Diabetes type 2, uncomplicated	3456	5	6	2566
GERD	3451	6	5	2733
Hypertension, benign	3323	7	1	6182
Long-term drug therapy	2865	8	7	2299
Chronic airway obstruction NEC	2660	9	10	1968
Chest pain NOS	2618	10		
Depressive disorder NEC	2558	11	11	1952
Hypothyroidism NOS	2419	12	13	1685
Respiratory abnorm NEC	2298	13		
Malaise and fatigue NEC	2290	14		
Anxiety state NOS	2227	15	12	1838
Abdominal pain-site NOS	2189	16		
Mammogram screening	2142	17	16	1534
Lumbago	2131	18	14	1573
Asthma without status asthmaticus	2099	19	18	1301
Anemia NOS	1999	20	19	1228
Bronchitis, acute, unspecified			20	1223
Immunization			8	2220
General adult exam w/o abnormal findings			9	2167
Atherosclerotic heart dz of native coronary artery w/o angina pectoris			15	1572
Hyperlipidemia, mixed			17	1305

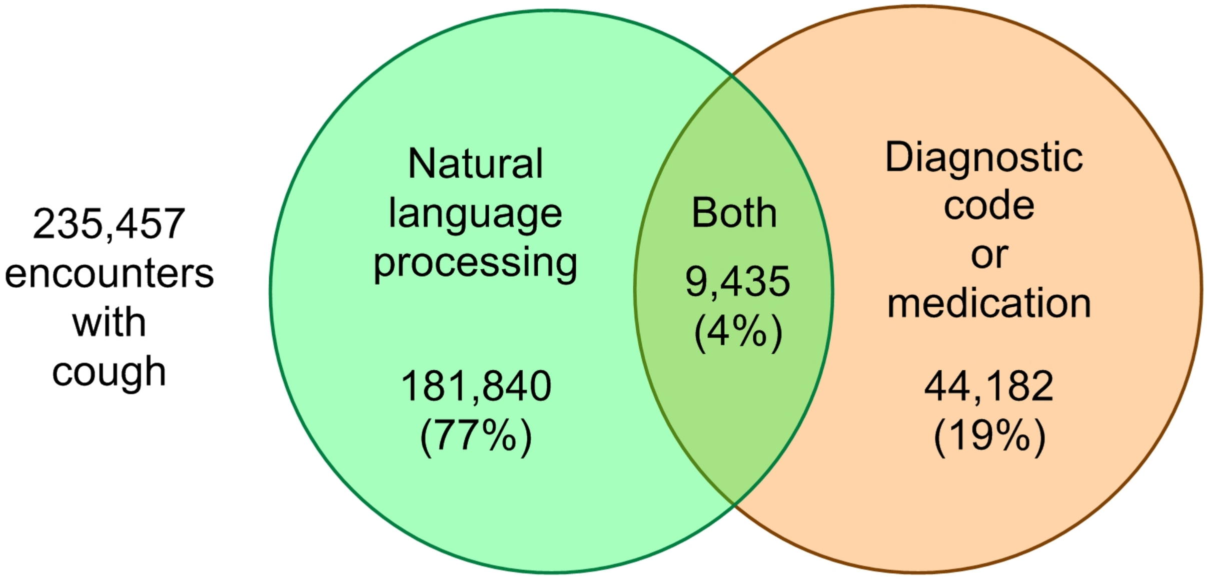
FIGURE LEGENDS

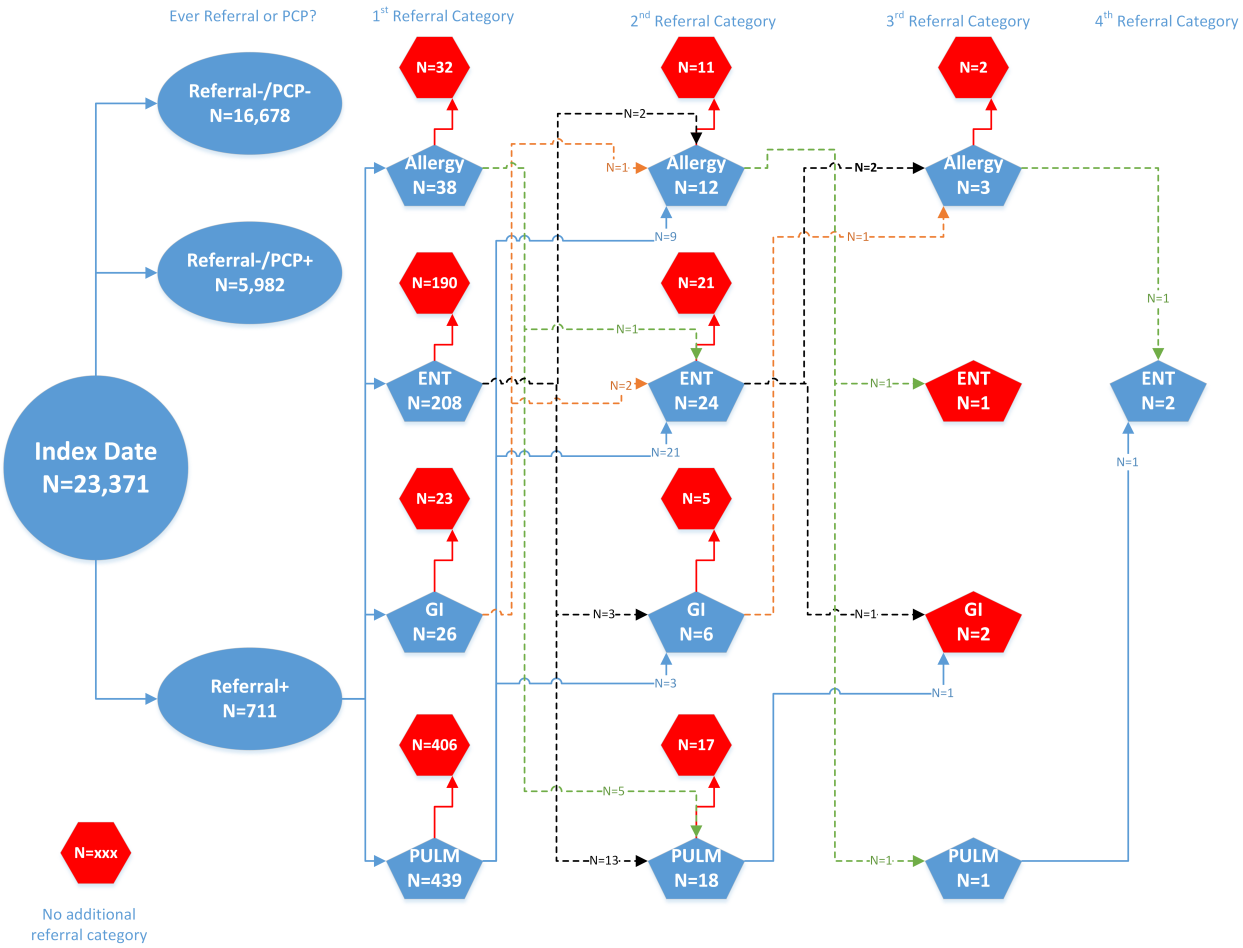
Figure 1. Identification of encounters with cough, and encounters with chronic cough. Diagnostic codes, medication orders, and narrative text in encounter notes identified cough encounters. Chronicity among multiple cough encounters identified chronic cough.

Figure 2. Structured and unstructured medical data indicating encounters with cough, among patients with chronic cough. Natural language processing analyzed unstructured, narrative text in notes of medical encounters. Diagnostic codes and medication records reflected structured data. Combined, the approaches identified 235,457 encounters with cough.

Figure 3. Cough-related referral patterns among four specialties: pulmonary medicine (pulm), allergy, otolaryngology (ENT), and gastroenterology (GI). Referrals are shown in columns from left to right, according to their chronological order. All frequencies preceded by “N=” indicate the number of patients following that pathway in the referral process.


Encounter has a diagnostic code or medication for cough?
No
Encounter has a text note indicating cough?
No
Yes
Yes
Cough is present in the encounter (a "cough encounter")
Cough is not present in the encounter
Are there at least three cough encounters within 56-120 days?
No
Yes
Chronic cough is present
Chronic cough is not present


235,457
encounters
with
cough
Natural
language
processing
181,840
(77%)
Both
9,435
(4%)
Diagnostic
code
or
medication
44,182
(19%)


Ever Referral or PCP?
1st Referral Category
2nd Referral Category
3rd Referral Category
4th Referral Category
Index Date N=23,371
Referral-/PCP- N=16,678
Referral-/PCP+ N=5,982
Referral+ N=711
Allergy N=38
ENT N=208
GI N=26
PULM N=439
N=32
N=190
N=23
N=406
Allergy N=12
ENT N=24
GI N=6
PULM N=18
N=11
N=21
N=5
N=17
Allergy N=3
ENT N=1
GI N=2
PULM N=1
N=2
ENT N=2
N=2
N=1
N=9
N=1
N=2
N=21
N=3
N=3
N=5
N=13
N=2
N=1
N=1
N=1
N=1
N=1
N=1
N=1
N=1
N=xxx
No additional referral category

ABBREVIATIONS

ACEI	Angiotensin-converting enzyme inhibitor
CC	Chronic cough
EHR	Electronic health record
ENT	Otolaryngology
NLP	Natural language processing
GERD	Gastroesophageal reflux disease
GI	Gastroenterology
ICD	International Classification of Diseases
INPC	Indiana Network for Patient Care
PPV	Positive predictive value
UACS	Upper airway cough syndrome