# Applying a Text-Search Algorithm to Radiology Reports Can Find More Patients With Pulmonary Nodules Than Radiology Coding Alone

**Rolando Sanchez, MD, MS; George Bailey; Peter J. Kaboli, MD, MS; Steven B. Zeliadt, PhD, MPH; Julie A. Lang, RN, BSN, MBA; and Richard M. Hoffman, MD, MPH**

**Introduction:** Chest imaging often incidentally finds indeterminate nodules that need to be monitored to ensure early detection of lung cancers. Health care systems need effective approaches for identifying these lung nodules. We compared the diagnostic performance of 2 approaches for identifying patients with lung nodules on imaging studies (chest/abdomen): (1) relying on radiologists to code imaging studies with lung nodules; and (2) applying a text search algorithm to identify references to lung nodules in radiology reports.

**Methods:** We assessed all radiology studies performed between January 1, 2016 and November 30, 2016 in a single Veterans Health Administration hospital. We first identified imaging reports with a diagnostic code for a pulmonary nodule. We then applied a text search algorithm to identify imaging reports with key words associated with lung nodules. We reviewed medical records for all patients with a suspicious radiology report based on either search strategy to confirm the presence of a lung nodule. We calculated the yield and the positive predictive value (PPV) of each search strategy for finding pulmonary nodules.

**Results:** We identified 12,983 imaging studies with a potential lung nodule. Chart review confirmed 8,516 imaging studies with lung nodules, representing 2,912 unique patients. The text search algorithm identified all the patients with lung nodules identified by the radiology coding (n = 1,251) as well as an additional 1,661 patients. The PPV of the text search was 72% (2,912/4,071) and the PPV of the radiology code was 92% (1,251/1,363). Among the patients with nodules missed by radiology coding but identified by the text search algorithm, 130 had lung nodules > 8 mm in diameter.

**Conclusions:** The text search algorithm can identify additional patients with lung nodules compared to the radiology coding; however, this strategy requires substantial clinical review time to confirm nodules. Health care systems adopting nodule-tracking approaches should recognize that relying only on radiology coding might miss clinically important nodules.

Author affiliations can be found at the end of the article.
**Correspondence:** Rolando Sanchez (rolando-sanchez@uiowa.edu)

Rapid advances in imaging technology have led to better spatial resolution with lower radiation doses to patients. These advances have helped to increase the use of diagnostic chest imaging, particularly in emergency departments and oncology centers, and in screening for coronary artery disease. As a result, there has been an explosion of incidental findings on chest imaging—including indeterminate lung nodules.[1,2]

Lung nodules are rounded and well-circumscribed lung opacities (≤ 3 cm in diameter) that may present as solitary or multiple lesions in usually asymptomatic patients. Most lung nodules are benign, the result of an infectious or inflammatory process. Nodules that are ≤ 8 mm in diameter, unless they show increase in size over time, often can be safely followed with imaging surveillance. In contrast, lung nodules > 8 mm could represent an early-stage lung cancer, especially among patients with high-risk for developing lung cancer (ie, those with advanced age, heavy tobacco abuse, or emphysema) and should be further assessed with close imaging surveillance, either chest computed tomography (CT) alone or positron-emission tomography (PET)/CT, or tissue biopsy, based on the underlying likelihood of malignancy.

Patients who receive an early-stage lung cancer diagnosis can be offered curative treatments leading to improved 5-year survival rates.[3,4] Consequently, health care systems need to be able to identify these nodules accurately, in order to categorize and manage them accordingly to the Fleischner radiographic and American College of Chest Physicians clinical guidelines.[5,6] Unfortunately, many hospitals struggle to identify patients with incidental lung nodules found during diagnostic chest and abdominal imaging, due in part to poor adherence to Fleischner guidelines among radiologists for categorizing pulmonary nodules.[7,8]

The Veterans Health Administration (VHA) system is interested in effectively detecting patients with incidental lung nodules. Veterans have a higher risk of developing lung cancer when compared with the entire US population, mainly due to a higher incidence of tobacco use.[6] The prevalence of lung nodules among veterans with significant risk factors for lung cancer is about 60% nationwide, and up to 85% in the Midwest, due to the high prevalence of histoplasmosis.[7] However, only a small percentage of these nodules represent an early stage primary lung cancer.

Several Veterans Integrated Service Networks (VISNs) in the VHA use a radiology diagnostic code to systematically identify imaging

studies with presence of lung nodules. In VISN 23, which includes Minnesota, North Dakota, South Dakota, Iowa, and portions of neighboring states, the code used to identify these radiology studies is 44. However, there is high variability in the reporting and coding of imaging studies among radiologists, which could lead to misclassifying patients with lung nodules.[8]

Some studies suggest that using an automated text search algorithm within radiology reports can be a highly effective strategy to identify patients with lung nodules.[9,10] In this study, we compared the diagnostic performance of a newly developed text search algorithm applied to radiology reports with the current standard practice of using a radiology diagnostic code for identifying patients with lung nodules at the Iowa City US Department of Veterans Affairs (VA) Health Care System (ICVAHCS) hospital in Iowa.

## METHODS

Since 2014, The ICVAHCS has used a radiology diagnostic code to identify any imaging studies with lung nodules. The radiologist enters "44" at the end of the reading process using the Nuance Powerscribe 360 radiation reporting system. The code is uploaded into the VHA Corporate Data Warehouse (CDW), and it is located within the radiology exam domain. This strategy was created and implemented by the Minneapolis VA Health Care System in Minnesota for all the VA hospitals in VISN 23. A lung nodule registry nurse was provided with a list of radiology studies flagged with this radiology diagnostic code every 2 weeks. A chart review was then performed for all these studies to determine the presence of a lung nodule. When detected, the ordering health care provider was alerted and given recommendations for managing the nodule.

We initially searched for the radiology studies with a presumptive lung nodule using the radiology code 44 within the CDW. Separately, we applied the text search strategy only to radiology reports from chest and abdomen studies (ie, X-rays, CT, magnetic resonance imaging [MRI], and PET) that contained any of the keyword phrases. The text search strategy was modeled based on a natural language processing (NLP) algorithm developed by the Puget Sound VA Healthcare System in Seattle, Washington to identify lung nodules on radiology reports.[9] Our algorithm included a series of text searches using Microsoft SQL. After several simulations using a random group of radiology reports, we chose the keywords: "lung AND nodul"; "pulm AND nodul"; "pulm AND mass"; "lung AND mass"; and "ground glass". We selected only chest and abdomen studies because on several simulations using a random group of radiology reports, the vast majority of lung nodules were identified on chest and abdomen imaging studies. Also, it would not have been feasible to chart review the approximately 30,000 total radiology reports that were generated during the study period.

From January 1, 2016 through November 30, 2016, we applied both search strategies independently: radiology diagnostic code for lung nodules to all imaging studies, and text search to all radiology reports of chest and abdomen imaging studies in the CDW (Figure). We also collected demographic (eg, age, sex, race, rurality) and clinical (eg, medical comorbidities, tobacco use) information that were uploaded to the database automatically from CDW using *International Statistical Classification of Diseases, Tenth Edition* and demographic codes. The VHA uses the Rural-Urban Commuting Areas (RUCA) system to define

**TABLE 1** Sociodemographic and Clinical Characteristics of Patients With Lung Nodules

| Variables | Radioloy Coding and Text Search[a] (n = 1,251) | Text Search (n = 1,661) | Total (N = 2,912) | *P* Value |
|---|---|---|---|---|
| Age, No. (%) | | | | < .001 |
| < 50 y | 32 (2.6) | 103 (6.3) | 135 | |
| 50-59 y | 127 (10.2) | 184 (11.1) | 311 | |
| 60-69 y | 495 (39.6) | 616 (37.1) | 1,111 | |
| 70-79 y | 466 (37.3) | 494 (29.7) | 960 | |
| > 80 y | 131 (10.5) | 264 (15.9) | 395 | |
| Gender, male, No. (%) | 1,209 (96.6) | 1,589 (95.7) | 2,798 | .18 |
| RUCA, highly rural, No. (%) | 906 (72.4) | 1,178 (70.9) | 2,084 | .37 |
| Comorbidities | | | | |
| Mean Charlson comorbidity score (SD) | 3.5 (3.0) | 4.1 (3.3) | | < .001 |
| COPD, No. (%) | 707 (56.5) | 905 (54.5) | 1,612 | .28 |
| Smoking status, No. (%) | | | | |
| Current smoker | 468 (37.4) | 555 (33.4) | 1,023 | |
| Nonsmoker | 783 (62.6) | 1,106 (66.6) | 1,889 | .02 |

Abbreviations: COPD, chronic obstructive pulmonary disease; RUCA, Rural Urban Commuting Areas.
[a]The text search algorithm identified all the patients with radiology code 44.

**TABLE 2** Patients With Radiology Studies Identified as Abnormal by Each Search Strategy

| Criteria | Radiology Code (n = 1,363) | Text Search (N = 4,071) |
|---|---|---|
| Patients with confirmed lung nodule, No. | 1,251 | 2,912 |
| Patients with lung nodule > 8 mm, No. (%) | 302 (24) | 432 (11) |
| Nodule imaging source, No. (%) | | |
| Chest computed tomography | 906 (72) | 1,705 (59) |
| Other imaging types[a] | 345 (28) | 1,207 (41) |

[a]Includes chest and abdomen X-rays, magnetic resonance imaging, and positron emission tomography.

rurality, which takes into account population density and how closely a community is linked socioeconomically to larger urban centers.[11] The protocol was reviewed and approved by the institutional review board of ICVAHCS and the University of Iowa.

The presence of a lung nodule was established by having the lung nodule registry nurse manually review the charts of every patient with a radiology report identified by either code 44 or the text search algorithm. The goal was to ensure that our text search strategy identified all reports with a code 44 to be compliant with VISN expectations. Cases in which a lung nodule was described in the radiology report were considered true positives, and those without a lung nodule description were considered false positives.

We compared the sociodemographic and clinical characteristics of patients with lung nodules between those identified with both code 44 and the text search and those identified with the text search alone. We used $\chi 2$ tests for categorical variables (eg, age, gender, RUCA, chronic obstructive pulmonary disease (COPD), smoking status) and *t* tests for continuous variables (eg, Charlson comorbidity score). A $P$ value $\leq .05$ was considered statistically significant. To assess the yield of each search strategy, we determined the number of patients with lung nodules detected by the text search and the radiology diagnostic code. We also calculated the positive predictive value (PPV) and 95% CI of each search strategy.

## RESULTS

We identified 12,983 radiology studies that required manual review during the study period. We confirmed that 8,516 imaging studies had lung nodules, representing 2,912 patients. Subjects with lung nodules were predominantly male (96%), aged between 60 and 79 years (71%), and lived in a rural area (72%). More than 50% of these patients had COPD and over a third were current smokers (Table 1). The text search algorithm identified all of the patients identified by the radiology diagnostic code (n = 1,251). It also identified an additional 1,661 patients with lung nodules that otherwise would have been missed by the radiology code. Compared with those identified only by the text search, those identified by both the radiology coding and text search were older, had lower Charlson comorbidity scores, and were more likely to be a current smoker.

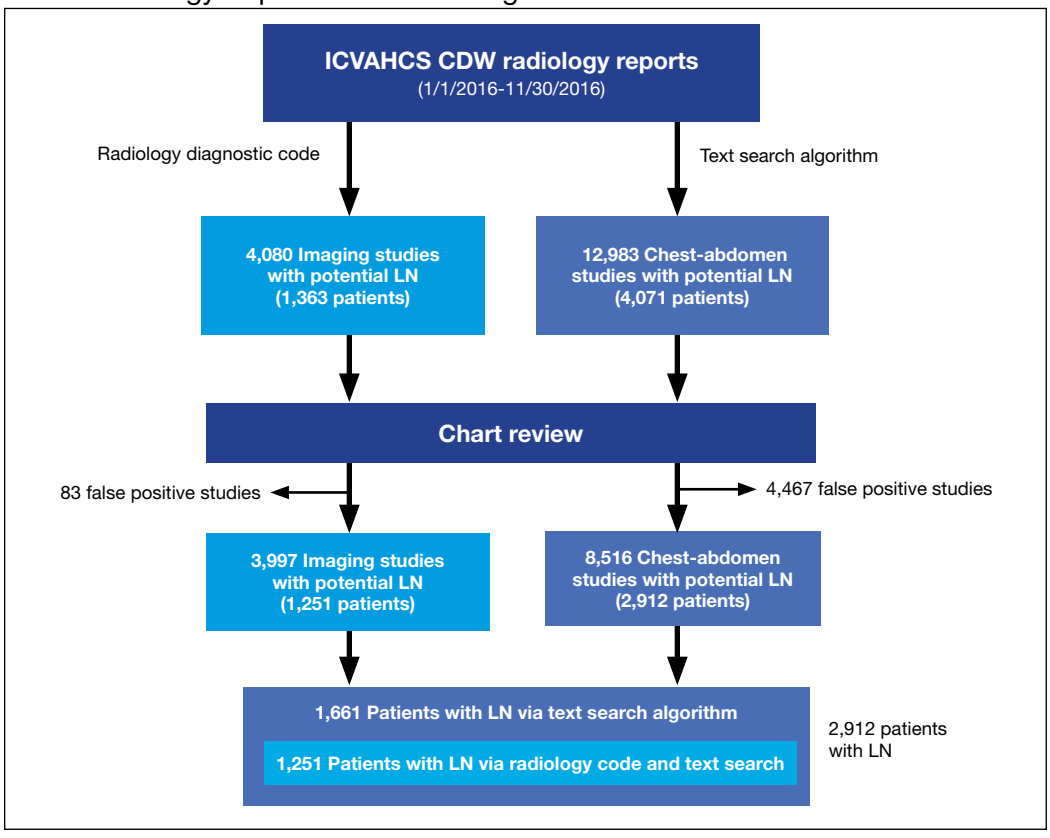
The text search algorithm identified more than twice as many patients with potential lung nodules compared with the radiology diagnostic code (4,071 vs 1,363) (Table 2). However, the text search algorithm was associated with a much higher number of false positives than was the diagnostic code (1,159 vs 112) and a lower PPV (72% [95% CI, 70.6-73.4] vs 92% [95% CI, 90.6-93.4], respectively). The text search algorithm identified 130 patients with lung nodules of moderate to high risk for malignancy (> 8 mm diameter) that were not identified by the radiology code. When the PPV of each search strategy was calculated based on imaging studies with nodules (most patients had > 1 imaging study), the results remained similar (98% for radiology code and 66% for text search). A larger proportion of the lung nodules detected by code 44 vs the text search algorithm were from CT chest studies.

## DISCUSSION

In a population of predominantly older male veterans with significant risk factors for lung cancer and high incidence of incidental lung nodules, applying a text search algorithm on radiology reports identified a substantial number of patients with lung nodules, including some with nodules > 8 mm, that were missed by the radiologist-generated code.[9,10] Improving the yield of detection for lung nodules in a population with high risk for lung cancer would increase the likelihood of detecting patients with potentially curable early-stage lung cancers, decreasing lung cancer mortality.

The reasons for the high number of patients with lung nodules missed by the radiology code are unclear. Potential explanations may include the lack of standardization of imaging reports by the radiologists (ie, only 21% of chest CTs used

**FIGURE** Radiology Reports Search Strategies




Abbreviations: CDW, Corporate Data Warehouse; ICVAHCS, Iowa City Veterans Affairs Health Care System; LN, lung nodule.

a standardized template describing a lung nodule in our study), a problem well recognized both within and outside VHA.[8,12]

The text search algorithm identified more patients with lung nodules but had a higher rate of false positives when compared with the diagnostic code. The high rate of false positives resulted in more charts to review and an increased workload for the lung nodule registry team. The challenges presented by an increased workload should be balanced against the potential harms of missing nodules that develop into advanced cancer.

## Text Search Adjustments

Refining the text search criteria algorithm and the chart review process may decrease the rate of false positives significantly without affecting detection of lung nodules. In subsequent simulations, we found that by adding an exclusion criteria to text search algorithm to remove reports with specific keywords we could substantially reduce the number of false positive reports without affecting the detection rate of the lung nodules. These exclusion criteria would exclude any reports that: (1) contain "nodul" within the next 8 words after mentioning "no"; (2) contain "clear" within the next 8 words after mentioning "lung" in the text (eg, "lungs appear to be clear"); (3) contain "clear" within the next 4 words after mentioning "otherwise" in the text (eg, "otherwise appear to be clear"). Based on our study results, we further refined the text search strategy by limiting the search to only chest imaging studies. When we applied the revised algorithm to a random sample of imaging reports, we found all the code 44 radiology reports were still captured, but we were able to reduce the number of radiology reports needing review by about 80%.

Although classification approaches are being refined to improve radiology performance in multiple categories of nodules, this study suggests that alternative approaches based on text algorithms can improve the capture of pulmonary nodules that require surveillance. These algorithms also can be used to augment radiologist reporting systems. This represents an investment

in resources to build a team that should include a bioinformatics specialist, lung nodule registry personnel (review charts of the detected imaging studies with lung nodules, populating the lung nodule database, and determining and tracking the need of imaging follow up), a lung nodule clinic nurse coordinator, and a dedicated lung nodule clinic pulmonologist.

Radiology departments could employ this text search approach to identify missed nodules and use an audit and feedback system to train radiologists to code lung nodules consistently at the time of the initial reading to avoid delays in identifying patients with nodules. Alternatively, the more widespread use of a standardized CT chest radiology reports using Fleischner or the American College of Radiology Lung Imaging Reporting and Data System (Lung RADS) templates might improve the detection of patients with lung nodules.[5,13,14]

The VHA system should have an effective strategy for identifying incidental lung nodules during routine radiology examinations. Relying only on radiologists to identify and code pulmonary nodules can lead to missing a significant number of patients with lung nodules and some patients with early stage lung cancer who could receive curative therapy.[12,14-16] The use of a standardized algorithm, like a text search strategy, might decrease the risk of variation in the execution and result in a more sensitive detection of patients with lung nodules. The text search strategy might be easily implemented and shared with other hospitals both within and outside the VHA.

### Limitations

This study was performed in a single VHA hospital and the findings may not be generalizable to other settings of care. Second, our study design is susceptible to work-up bias because the results of a diagnostic test (eg, chest or abdomen imaging) affected whether the chart review was used to verify the test result. It was not feasible to review the patient records of all radiology studies done at the facility during the study period, consequently complete 2 × 2 tables could not be created to calculate sensitivity, specificity, and negative predictive value.

## CONCLUSION

A text search algorithm of radiology reports increased the detection of patients with lung nodules when compared with radiology diagnostic coding alone. However, the improved detection was associated with a higher rate of false positives, which requires manually reviewing a larger number of patient's chart reports. Future research and quality improvement should focus on standardizing the radiology reporting process and improving the efficiency and reliability of follow up and tracking of incidental lung nodules.

## Acknowledgments

The work reported here was supported by a grant from the Office of Rural Health (N32-FY16Q1-S1-P01577), US Department of Veterans Affairs, Veterans Health Administration. We also had the support from the Veterans Rural Health Resource Center-Iowa City, and the Health Services Research and Development (HSR&D) Service through the Comprehensive Access and Delivery Research and Evaluation (CADRE) Center (REA 09-220).

## Author affiliations

**Rolando Sanchez** is a Clinical Assistant Professor of Pulmonary and Critical Care Medicine; **Peter Kaboli** is a Professor of Internal Medicine; and **Richard Hoffman** is a Professor of Internal Medicine, all at the University of Iowa Carver College of Medicine in Iowa City. **George Bailey** is a Research Data Manager; **Julie Lang** is a Registered Nurse and Research Coordinator; and Peter Kaboli is an Associate Investigator, all in the Center for Access and Delivery Research and Evaluation (CADRE) at the Iowa City VA Healthcare System. **Steven Zeliadt** is a Research Professor of Public Health at the Seattle-Denver Center of Innovation for Veteran-Centered and Value-Driven Care, VA Puget Sound Health Care System and the University of Washington School of Public Health in Seattle.

## Author disclosures

The authors report no actual or potential conflicts of interest with regard to the article.

## Disclaimer

The opinions expressed herein are those of the authors and do not necessarily reflect those of *Federal Practitioner*, Frontline Medical Communications Inc., the U.S. Government, or any of its agencies.

## References

1. Jacobs PC, Mali WP, Grobbee DE, van der Graaf Y. Prevalence of incidental findings in computed tomographic screening of the chest: a systematic review. *Journal of computer assisted tomography.* 2008;32(2):214-221.
2. Frank L, Quint LE. Chest CT incidentalomas: thyroid lesions, enlarged mediastinal lymph nodes, and lung nodules. *Cancer Imaging.* 2012;12(1):41-48.
3. National Institutes of Health, National Cancer Institute, Surveillance, Epidemiology, and End Results Program. Cancer stat facts: lung and bronchus cancer. https://seer.cancer.gov/statfacts/html/lungb.html. Accessed April 8, 2020.
4. Alberg AJ, Brock MV, Ford JG, Samet JM, Spivack SD. Epidemiology of lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest.* 2013;143(5 Suppl):e1S-e29S.
5. MacMahon H, Naidich DP, Goo JM, et al. Guidelines for Management of Incidental Pulmonary Nodules Detected on CT Images: From the Fleischner Society 2017. *Radiology.* 2017;284(1):228-243.
6. Zullig LL, Jackson GL, Dorn RA, et al. Cancer incidence among patients of the U.S. Veterans Affairs Health Care System. *Mil Med.* 2012;177(6):693-701.
7. Kinsinger LS, Anderson C, Kim J, et al. Implementation of lung cancer screening in the Veterans Health Administration. *JAMA Intern Med.* 2017;177(3):399-406.
8. Iqbal MN, Stott E, Huml AM, et al. What's in a name?

Factors associated with documentation and evaluation of incidental pulmonary nodules. *Ann Am Thorac Soc.* 2016;13(10):1704-1711.
9. Farjah F, Halgrim S, Buist DS, et al. An automated method for identifying individuals with a lung nodule can be feasibly implemented across health systems. *Egems (Wash DC).* 2016;4(1):1254.
10. Danforth KN, Early MI, Ngan S, Kosco AE, Zheng C, Gould MK. Automated identification of patients with pulmonary nodules in an integrated health system using administrative health plan data, radiology reports, and natural language processing. *J Thorac Oncol.* 2012;7(8):1257-1262.
11. US Department of Veterans Affairs, Office of Rural Health. https://www.ruralhealth.va.gov/aboutus/ruralvets.asp. Updated January 28, 2020. Accessed April 8, 2020.
12. Blagev DP, Lloyd JF, Conner K, et al. Follow-up of incidental pulmonary nodules and the radiology report. *J Am Coll Radiol.* 2016;13(2 suppl):R18-R24.
13. Eisenberg RL, Fleischner S. Ways to improve radiologists' adherence to Fleischner Society guidelines for management of pulmonary nodules. *J Am Coll Radiol.* 2013;10(6):439-441.
14. Aberle DR. Implementing lung cancer screening: the US experience. *Clin Radiol.* 2017;72(5):401-406.
15. Gould MK, Donington J, Lynch WR, et al. Evaluation of individuals with pulmonary nodules: when is it lung cancer? Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest.* 2013;143(5 Suppl):e93S-e120S.
16. Callister ME, Baldwin DR. How should pulmonary nodules be optimally investigated and managed? *Lung Cancer.* 2016;91:48-55.