Validation of natural language processing to determine the presence and size of abdominal aortic aneurysms in a large integrated health system

Myra McLenon, BS,[a] Steven Okuhn, MD,[b,c] Elizabeth M. Lancaster, MD, MAS,[c] Michaela M. Hull, MS,[d] John L. Adams, PhD,[d] Elizabeth McGlynn, PhD,[d] Andrew L. Avins, MD, MPH,[e,f,g] *and* Robert W. Chang, MD,[e,h]
Overland Park, Kan; and San Francisco, Pasadena, Oakland, and South San Francisco, Calif

ABSTRACT

Objective: Previous studies of the natural history of abdominal aortic aneurysms (AAAs) have been limited by small cohort sizes or heterogeneous analyses of pooled data. By quickly and efficiently extracting imaging data from the health records, natural language processing (NLP) has the potential to substantially improve how we study and care for patients with AAAs. The aim of the present study was to test the ability of an NLP tool to accurately identify the presence or absence of AAAs and detect the maximal abdominal aortic diameter in a large dataset of imaging study reports.

Methods: Relevant imaging study reports (n = 230,660) from 2003 to 2017 were obtained for 32,778 patients followed up in a prospective aneurysm surveillance registry within a large, diverse, integrated healthcare system. A commercially available NLP algorithm was used to assess the presence of AAAs, confirm the absence of AAAs, and extract the maximal diameter of the abdominal aorta, if stated. A blinded expert manual review of 18,000 randomly selected imaging reports was used as the reference standard. The positive predictive value (PPV or precision), sensitivity (recall), and the kappa statistics were calculated.

Results: Of the randomly selected 18,000 studies that underwent expert manual review, 48.7% were positive for AAAs. In confirming the presence of an AAA, the interrater reliability of the NLP compared with the expert review showed a kappa value of 0.84 (95% confidence interval [CI], 0.83-0.85), with a PPV of 95% and sensitivity of 88.5%. The NLP algorithm showed similar results for confirming the absence of an AAA, with a kappa of 0.79 (95% CI, 0.799-0.80), PPV of 77.7%, and sensitivity of 91.9%. The kappa, PPV, and sensitivity of the NLP for correctly identifying the maximal aortic diameter was 0.88 (95% CI, 0.87-0.89), 88.8%, and 88.2% respectively.

Conclusion: The use of NLP software can accurately analyze large volumes of radiology report data to detect AAA disease and assemble a contemporary aortic diameter-based cohort of patients for longitudinal analysis to guide surveillance, medical management, and operative decision making. It can also potentially be used to identify from the electronic medical records pre- and postoperative AAA patients "lost to follow-up," leverage human resources engaged in the ongoing surveillance of patients with AAAs, and facilitate the construction and implementation of AAA screening programs. (J Vasc Surg 2021;■:1-8.)

Keywords: Abdominal aortic aneurysm; Natural history; Natural language processing; Screening; Surveillance

Abdominal aortic aneurysm (AAA) disease is an important societal health issue in the United States, with an estimated 9928 rupture-related deaths in 2017.[1] Despite this, little is definitively known about the natural history of AAAs to guide diagnostic and therapeutic procedures. Most natural history studies of AAAs have been limited by small, selected datasets or heterogeneous pooled analyses owing to the difficulty of obtaining AAA diameters in large groups of patients and for individual AAA patients with multiple studies over time.[2-18] However, the number of cross-sectional abdominal imaging studies performed has been increasing, along with public

From the Softek Illuminate, Inc, Overland Park[a]; the Division of Vascular Surgery, Department of Surgery, Veterans Affairs San Francisco Healthcare System, San Francisco[b]; the Division of Vascular Surgery, Department of Surgery, University of California, San Francisco, San Francisco[c]; the Kaiser Permanente Center for Effectiveness and Safety Research, Pasadena[d]; the Division of Research, Kaiser Permanente Northern California, Oakland[e]; the Department of Medicine[f] and Department of Epidemiology and Biostatistics,[g] University of California, San Francisco, San Francisco; and the Division of Vascular Surgery, Department of Surgery, The Permanente Medical Group, South San Francisco.[h]

The present study was supported by The Permanente Medical Group Delivery Science and Physician Researcher Programs (Northern California) and the Kaiser Permanente Center for Effectiveness and Safety Research (Pasadena, Calif).

Author conflict of interest: S.O. has been a paid Clinical Board Advisor for, and is a stockholder in, Softek Illuminate Inc. M.M. is an employee of, and stockholder in, Softek Illuminate Inc. E.M.L., M.M.H., J.L.A., E.M., A.L.A., and R.W.C. have no conflicts of interest.

Additional material for this article may be found online at www.jvascsurg.org.

Correspondence: Robert W. Chang, MD, The Permanente Medical Group, 1200 El Camino Real, 3 Floor MOB, South San Francisco, CA 94080 (e-mail: robert.W.Chang@kp.org).

The editors and reviewers of this article have no relevant financial relationships to disclose per the JVS policy that requires reviewers to decline review of any manuscript for which they may have a conflict of interest.

0741-5214

https://doi.org/10.1016/j.jvs.2020.12.090

awareness of AAAs, which has led to an increase in the incidental reporting of small AAAs.[9,19-23]

Reported studies of abdominal aortic morphology and diameter have relied on laborious time-intensive manual data abstraction from unstructured, free-text radiology reports, which has limited the sample size that could be analyzed.[2,12,14,24,25] Natural language processing (NLP) software tools might present a more efficient option for extracting this information from archived radiology reports.[26-31] These tools could be used to streamline the human resources required to screen for AAAs and to monitor the increasing number of AAAs diagnosed that require long-term surveillance.

In the present study, we evaluated the accuracy of an NLP engine in analyzing 15 years of aortic morphology and diameter data from an AAA screening and surveillance program in a large integrated healthcare system to determine whether it could be used to facilitate both the natural history assessment of AAAs and patient care.

ARTICLE HIGHLIGHTS

- **Type of Research:** A blinded, randomized, multicenter retrospective cohort analysis of prospectively collected registry data from the Kaiser Permanente Northern California Population and Condition Tracking System
- **Key Findings:** In the present study of 18,000 abdominal imaging studies, a natural language processing algorithm accurately confirmed the presence of an abdominal aortic aneurysm with a positive predictive value of 95% and sensitivity of 88.5%.
- **Take Home Message:** The use of natural language processing software can accurately analyze large volumes of radiology report data to detect abdominal aortic aneurysm disease and assemble a contemporary aortic diameter-based cohort of patients to guide screening, surveillance, medical management, and operative decision making.

METHODS

Study setting. The Kaiser Permanente Northern California (KPNC) is an integrated healthcare delivery system that serves >4 million members across 21 medical centers and hospitals. At present, >9000 physicians are employed by the system, with >300 full-time radiologists. The Kaiser Foundation Research Institute institutional review board approved the present study, with a specified waiver of patient consent for an observational data-only study.

Imaging and clinical data sources. The subjects whose radiology reports were used for the present study were all patients enrolled in the KPNC population and condition tracking system, an arterial aneurysm registry (including, but not limited to, AAAs), and who had undergone abdominal imaging at a KPNC medical facility from June 22, 1992 to June 16, 2017. This aneurysm registry is an ongoing prospective program started in 2003 for the purposes of dedicated nurse-monitored surveillance and screening to ensure proper clinical care for a variety of diseases. The transcriptions of all radiology reports since 1992 are electronically accessible for both clinical and research purposes.

The study modalities included axial computed tomography, ultrasound studies, magnetic resonance imaging, magnetic resonance angiography, and positron emission tomography. For patients enrolled in the AAA registry, the relevant imaging reports were identified by searching the included imaging modalities for any 1 of 41 Current Procedural Terminology codes related to AAA screening or diagnosis (Appendix, Supplementary Tables I and II; and the Supplementary Fig, online only). The clinical indications for these studies included the following: (1) institution-directed AAA screening; (2) suspicion for AAA and an AAA identified; (3) non–AAA-related indication and an AAA was identified incidentally; and (4) follow-up imaging for any patient already enrolled in AAA surveillance.

Overall, 32,778 patients in the aneurysm registry had undergone 230,660 relevant imaging studies during the study period. The textual reports from the relevant imaging studies were extracted from the KPNC electronic medical records (EMRs). No program-wide templates were used for image reporting during the study period. No imaging review was performed as a part of the present study. Patients with a single relevant imaging study were excluded because no opportunity would be available for longitudinal assessment, a focus of a subsequent investigation. Additional demographic and clinical data were obtained from the EMRs.

Development of NLP algorithms for aortic diameter and disease detection. Our initial effort focused on two related, but distinct, processes: the detection of the presence of an AAA and the identification of the maximal abdominal aortic diameter, when reported. We used a previously developed solution from Softek Illuminate, Inc (Overland Park, Kan). This NLP search engine consists of an AAA classifier module to determine the presence of an AAA and a separate AAA diameter-extraction module. Using the combined data from the two modules, an algorithm was developed to confirm the absence of an AAA using the following criteria: (1) discrete classifiers describing "no AAA"; (2) a size <2.5 cm (from the extraction module); and (3) maximal diameter unknown, but abdominal aorta mentioned without describing an AAA or abnormal aortic morphology to suggest an AAA (assuming that if a radiologist had discussed the aorta but had not suggested the presence of aneurysmal disease, the likelihood of an AAA was very low).

The data processing pipeline for these modules was divided into three linear steps: (1) text preprocessing and normalization; (2) capturing appropriate measurements by sorting the reports into 53 diameter and 4 nondiameter bins; and (3) final AAA classification of the imaging document.

The first step to enable learning was the normalization of the report text. This process included converting various inflections of words to the same root, converting abbreviations to their full forms (ie, AP to anterior-posterior), and standardizing report formatting to account for bullet points, new lines, and so forth.

The measurement module then sorted a report into 1 of 57 mutually exclusive bins. These consisted of 53 diameter bins, starting with <2.0 cm, 2.0 cm, and additional bins in increments of 1 mm to ≤7.0 cm, with all sizes >7.0 cm grouped together. If an abdominal aortic diameter could not be reliably determined by the NLP algorithm, four mutually exclusive "nondiameter" bins would be used: (1) AAA present, diameter unknown; (2) confirmed absence of AAA; (3) no mention of aorta in text; and (4) abdominal aorta mentioned, but diameter unknown.

Although the length of an AAA is usually ignored when considering the risk of rupture, all three dimensions of aortic measurements were acknowledged by the NLP algorithm. Occasionally, a radiologist will use approximations such as "greater than 5.0 cm"; thus, comparative phrases such as these were also considered when determining the maximal aortic diameter.

The machine learning modules were trained on 75% of an expert-validated teaching dataset of 6000 radiology reports and 8000 isolated sentences. The machine learning features were restricted to data elements extractable from the textual report and excluded meta-data. Various shallow learning approaches were tested using the training data to find the best solution. For this module, an AAA was defined as a maximal aortic diameter of ≥2.5 cm. Additionally, if a radiologist used morphologic terms consistent with an AAA (eg, dilated, enlarged, saccular,) the NLP will consider the aorta positive for an AAA even for a maximal diameter <2.5 cm. Validation was then performed using the remaining 25% of the documents (1594 studies) with a positive predictive value (PPV or precision) of 98%, sensitivity (recall) of 98%, and an F1 score of 98% (not previously reported).

Statistical analysis validation. From the 230,660 included radiology reports, 18,000 reports were randomly selected for manual blinded review. An individual patient could have contributed more than one imaging report. Two reviewers, one board-certified vascular surgeon (S.O.) and one trained coinvestigator (M.M.), each reviewed 10,000 reports, recording the presence or absence of AAAs and separately placing each report into 1 of the 57 bins described. The interrater reliability was assessed using 2000 randomized overlapping reports. These 20,000 reviewer observations were then used as the reference standard to validate the NLP. When the two reviewers disagreed, the board-certified vascular surgeon's choice was considered the reference standard.

Table I. Demographics of patients included in aneurysm registry

Variable	Total cohort (n = 32,778)
Age, years	78.3 ± 11.7
Male gender	24,263 (74)
Race	
White	23,786 (73)
Asian/Pacific Islander	3517 (11)
Hispanic	2358 (7)
Black	2232 (7)
Other/unknown	885 (3)
Smoking status	
Never	8438 (26)
Former	16,004 (49)
Current	7287 (22)
Unknown	1049 (3)
History of CVD	21,962 (67)
History of diabetes	5319 (16)
Imaging studies, No.	7.0 ± 5.8

CVD, Cardiovascular disease.
Data presented as mean ± standard deviation or number (%).

Four measurements of agreement between the reviewers and the NLP and between the reviewers (interrater reliability) were considered: (1) kappa (measurement of chance-corrected agreement for multiple classifications); (2) PPV (precision); (3) sensitivity (recall); and (4) F1 score [F1 = 2 × (PPV × sensitivity)/(PPV + sensitivity)]. The F1 score is the harmonic mean of the PPV and sensitivity and is a better overall measure of incorrectly classified cases (false-positive and false-negative results) by the NLP. For tables with more than two outcomes, the PPV and sensitivity are the weighted average and the F1 score is the overall mean of the F1 scores for each individual outcome.

The NLP was evaluated in three areas: (1) the ability to correctly determine the presence of an AAA; (2) the ability to conclusively rule out an AAA when none is present; and (3) the ability to detect the presence of a maximal abdominal aortic diameter measurement (when provided) and record this diameter accurately. The exact methods and classifications used to evaluate the ability of the NLP to confirm the absence of an AAA are described in the Appendix (online only).

Table II. Comparison of AAA detection between NLP and reviewers and interrater reliability assessment

	Expert reviewers		
NLP	**AAA present**	**AAA absent or unknown**	**Total**
AAA present	7751	412	8163
AAA absent or unknown	1011	8826	9837
Total	8762	9238	18,000
	Interrater reliability assessment		
	Kappa (95% CI)	**PPV; sensitivity**	**F1 score**
NLP vs reviewers	0.842 (0.834-0.849)	0.950; 0.885	0.916
Reviewer 1 vs 2	0.939 (0.924, 0.954)	0.965; 0.972	0.968

AAA, Abdominal aortic aneurysm; *CI*, confidence interval; *NLP*, natural language processing; *PPV*, positive predictive value.

Table III. Comparison of AAA absence detection between NLP and reviewers and interrater reliability assessment

	Expert reviewers		
NLP	**AAA absent**	**AAA present or unknown**	**Total**
AAA Absent	3790	1090	4880
AAA present or unknown	333	12,787	13,120
Total	4123	13,877	18,000
	Interrater reliability assessment		
	Kappa (95% CI)	**PPV; sensitivity**	**F1 score**
NLP vs reviewers	0.790 (0.779-0.800)	0.777; 0.919	0.842
Reviewer 1 vs 2	0.992 (0.901-0.942)	0.922; 0.956	0.939

AAA, Abdominal aortic aneurysm; *CI*, confidence interval; *NLP*, natural language processing; *PPV*, positive predictive value.

RESULTS

Of the randomly selected 18,000 studies that underwent the reference standard expert review, 48.7% were positive for an AAA and 49.2% contained a maximal aortic diameter. The most frequently included imaging studies were computed tomography (n = 8978; 50%), followed by ultrasound (n = 6891; 38%), and magnetic resonance imaging (n = 1520; 8.5%). Of the included patients, 74% were men and 73% were white, with a mean age of 78 ± 12 years (Table I). The mean number of imaging studies per patient was 7.04 ± 5.79, and the mean follow-up time was 7.02 ± 5.28 years.

The interrater reliability between expert reviewer 1 and 2, as measured by the kappa coefficient, was 0.939 (95% confidence interval [CI], 0.924-0.954) for the presence of an AAA, 0.922 (95% CI, 0.901-0.942) for the confirmed absence of an AAA, and ≥0.966 (95% CI, 0.954-0.978) for the size detection in the imaging reports with an included maximal aortic diameter.

Presence and absence of AAA. The validation results showed a kappa value of 0.842 (95% CI, 0.834-0.849) between the expert reviewers and NLP for the correct identification of an AAA. The NLP demonstrated a PPV of 95.0% for accurately identifying studies positive for an AAA and a sensitivity of 88.5%. Thus, the studies positive for an AAA had been incorrectly reported as no AAA or an unknown AAA 11.5% of the time (Table II). The F1 score for the NLP correctly identifying the presence of an AAA was 91.6%.

The NLP algorithm showed similar results for confirming the absence of an AAA, with a kappa of 0.790 (95% CI, 0.799-0.800), PPV of 77.7%, sensitivity of 91.9%, and F1 score of 84.2% (Table III). Thus, the NLP did not identify a truly negative report only 8.1% of the time but had incorrectly reported the absence of an AAA for 22.3% of the cases.

Maximal aortic diameter. The interrater reliability kappa value between the NLP and expert reviewers for determining the presence of an aortic diameter measurement was 0.752 (95% CI, 0.743-0.762). The PPV, sensitivity, and F1 scores were 98.8%, 75.8%, and 85.8%, respectively (Table IV). The ability of the algorithm to correctly identify the detected maximal aortic diameter is presented in Table V. Evaluating with a 1-mm discrimination showed a kappa value of 0.877 (95% CI, 0.869-0.885), a PPV of 88.8%, a sensitivity of 88.2%, and an F1 score of 97.0%. The reliability measurements consistently improved when the aortic diameter was analyzed using clinically relevant subgroups (maximal

Table IV. Comparison of maximal AAA diameter reported between NLP and reviewers and interrater reliability assessment

	Expert reviewers		
NLP	**Diameter reported**	**Diameter unknown**	**Total**
Diameter reported	6720	79	6799
Diameter unknown	2143	9058	11,201
Total	8863	9137	18,000
	Interrater reliability assessment		
	Kappa (95% CI)	**PPV; sensitivity**	**F1 score**
NLP vs reviewers	0.752 (0.743-0.762)	0.988; 0.758	0.858
Reviewer 1 vs 2	0.977 (0.968-0.986)	0.992; 0.985	0.988

AAA, Abdominal aortic aneurysm; *CI*, confidence interval; *NLP*, natural language processing; *PPV*, positive predictive value.

Table V. NLP reporting metrics for reporting maximal AAA diameter stratified by size bin grouping

Diameter assignment	Kappa (95% CI)	PPV	Sensitivity	F1 score
1-mm Increments				
All patients	0.877 (0.869-0.885)	0.888	0.882	0.862
Only ≥2.5 cm	0.898 (0.890-0.906)	0.903	0.905	0.885
Only ≥3.0 cm	0.905 (0.897-0.913)	0.909	0.911	0.889
Large diameter increments[a]				
All patients	0.908 (0.900-0.916)	0.934	0.930	0.912
Only ≥2.5 cm	0.931 (0.924-0.939)	0.951	0.951	0.941
Only ≥3.0 cm	0.938 (0.930-0.946)	0.959	0.959	0.946

AAA, Abdominal aortic aneurysm; *CI*, confidence interval; *NLP*, natural language processing; *PPV*, positive predictive value.
[a]Large diameter increments: <2.5 cm, 2.5-2.9 cm, 3.0-3.9 cm, 4.0-4.9 cm, 5.0-5.4 cm, ≥5.5 cm.

aortic diameter >2.5 cm or >3.0 cm) and larger size bins (<2.5 cm, 2.6-2.9 cm, 3.0-3.9 cm, 4.0-4.9 cm, 5.0-5.4 cm, ≥5.5 cm).

DISCUSSION

In the present study of a large imaging dataset, we found excellent ability of a commercially available NLP tool to detect the presence of an AAA and a reasonable ability to confirm the absence of an AAA. For the cases for which a maximal abdominal aortic diameter had been reported, we found a high degree of certainty that the NLP tool will detect it and, more importantly, report it correctly. We also found outstanding interrater reliability, and the large size of our dataset enabled a precise estimation of our outcomes with narrow CIs.

NLP for natural history research of AAAs. Numerous limitations are present in the available AAA natural history literature that the use of NLP can help address. Early studies of the natural history of AAA were aimed simply at identifying when to operate on AAAs and when they could be observed.[4,5,7,14,24,32] These studies were single-center reports and had included highly selected patients.[5,14,15,24] Other AAA natural history studies had only included patients between specific age ranges or meeting certain AAA diameter criteria.[7,10,16,17,22,24,33] In attempts to overcome this selection bias, other early investigators studied >24,000 consecutive autopsies to identify the prevalence and risk of AAAs.[4] That analysis was subject to bias owing to the difficulty in measuring the size of an aorta (intact or ruptured) in a deceased patient and by the confounding variable of death for the vast majority of patients who had died of causes other than an AAA.

The most rigorous and contemporary study reporting on the natural history of AAAs was from the RESCAN group with a pooled cohort of 15,471 patients from 18 different data sets. However, although it is the largest study to date, this carefully conducted meta-analysis had several limitations.[2] Patients with an AAA outside the 3.0- to 5.4-cm diameter range were excluded, and 7 of the 18 studies had only included men. The imaging methods and maximal diameter definitions had also varied across the studies, none of which had used NLP techniques. The RESCAN group also found that each of the 18 data sets had reported different rates of AAA diameter changes. Thus, statistical techniques were used to merge them into their final cohort, at the cost of significant

heterogeneity. Ultimately, these consistent limitations have made it challenging to update contemporary surveillance guidelines and indications for AAA repair in a convincing fashion.

NLP represents a computerized method to analyze text. It is based on a series of human-derived programming rules that characterize reported phrases and words to simulate human-like interpretation. It has been used in medical language systems since the mid-1990s and reported in various specialties to select naturally occurring conditions for translational research.[26,29,31,34,35] For peripheral arterial disease, Afzal et al[36] reported their experience using an open-source NLP tool (MedTagger) to identify patients with disease from clinical documentation. They found greater accuracy using NLP than using previous billing claims data. Lee et al[29] reported that using NLP provided accurate reporting from colonoscopy reports with various formats.

In one of the earliest examples involving AAAs, van Walraven et al[22] used a less-advanced, hard coded (string search) algorithm to screen an institutional radiology database to find incidentally mentioned AAAs and determine whether these patients had received appropriate surveillance and treatment. Sohn et al[26] trained an NLP program to determine the presence of AAAs and extract a size when a AAA was present using a small (650-report) dataset. They found excellent PPV, sensitivity, and F1 scores.[26] Morioka et al[27] reported on the use of a more sophisticated, but still hard coded, approach for 1402 screening ultrasound reports of male ever-smokers. They also reported high PPV, sensitivity, and F1 scores.[27] Although the validation statistics reported in these studies were high, the results were possibly inflated because the testing and training were both performed using the same dataset. Furthermore, these approaches did not use true machine learning because their analytic approach was, in part, hard-coded and had involved simply scanning the free text for predetermined lists of character strings. For example, if a preceding space was missing from the text of a radiology report, one of these reported algorithms would have incorrectly placed the report into a category indeterminate for an AAA instead of a category indicating the presence of an AAA.[27]

Although we have reported kappa, PPV, sensitivity, and F1 scores comparable to those from previous AAA NLP studies, our results represent an advance because our derivation and validation processes were performed against a much larger data set (18,000 vs 1200 or 600 patients) than used in previous studies, with the added advantage of a "real world" setting of non-template, free text reports from multiple radiologists and hospitals and a variety of imaging modalities.[26,27] Our imaging reports included various imaging modalities of the chest, abdomen, and pelvis instead of ultrasound examinations of the abdomen alone. Thus, our NLP tool was developed to determine the correct abdominal aortic location for an AAA, to answer the question of whether an AAA was present.[27] For example, miscategorizing a thoracic aortic aneurysm as an AAA would have resulted in an accuracy demerit during validation. Thus, even if the diameter classification had been correct, this would have been considered an error. The NLP algorithm was trained to ignore metadata (eg, examination type: chest computed tomography vs abdominal ultrasound) and "hearsay" historical data (eg, "History:"; or "Indication:"), making its determinations more accurate. Finally, unlike others, our NLP tool is potentially adaptable to any EMR system. It was not necessary to manually classify reports to "teach" the NLP before running the program, and no time was required for developing and training the algorithm. For these reasons and based on our stringent validation process, we are confident that these findings are generalizable.

Because our NLP would be extremely unlikely to assign a maximal aortic diameter if one were not present (PPV, 99%) and missed the presence of a maximal aortic diameter in a dictated report only 24% of the time (sensitivity, 76%), it is very well suited for a natural history study application where understanding of the accurate diameters is more important than missing some. Perhaps more importantly, when reported, the maximal abdominal aortic diameter was correct 88.8% of the time. Also, these numbers improved when the size bins were combined into broader diameter ranges and when smaller diameter bins (<2.5 cm, <3.0 cm) were excluded, reflecting clinically relevant, actionable situations.

NLP for AAA surveillance and screening. A broad consensus has been reached regarding the efficacy of AAA screening in decreasing all-cause mortality in selected patients and AAA surveillance in preventing rupture and guiding medical therapy.[18,21,25,37-43] However, AAA screening and pre- and postoperative AAA surveillance programs have not been widely implemented. This is, in part, because, without automated software solutions, identifying surveillance-appropriate patients is very labor intensive. Our NLP correctly identified 88.5% of possible AAAs (sensitivity), and 95% of the identifications of an AAA were correct (PPV), substantially reducing the need for manual image report review (F1 score, 92%). Human oversight will be required for the foreseeable future in AAA surveillance and screening programs. However, because even the most highly trained case managers cannot intuit an AAA unreported in an imaging study, the sensitivity must be high in these clinical situations in which missing an AAA would be worse than falsely identifying an AAA. An NLP with these attributes will allow for screening an entire EMR system (ie, radiology reports, operative records, clinical notes) for patients with AAAs who have not otherwise had an AAA diagnosed. It could also use the diameter classifier to

search for abdominal aortic diameter changes and, thereby, optimally use limited case manager resources to establish whether patients are being followed up appropriately and whether patients with more rapidly expanding AAAs require more frequent monitoring. This could also be applied to repaired AAAs postoperatively.

Unpublished Kaiser data collected during the early stages of a systematic rollout of AAA screening in northern California in 2012 to 2013 indicated that ~20% of 65- to 75-year-old male ever-smokers will have undergone relevant cross-sectional abdominal imaging for nonscreening purposes between 65 and 75 years of age. An NLP program with the ability to confirm the absence of an AAA could similarly be used to review patient data (ie, >64-year-old male ever-smokers) within a database and eliminate those who will not require ultrasound screening based on relevant findings from previous imaging studies, thereby saving a healthcare system unnecessary screening studies and the associated costs. When the NLP algorithm confirmed the absence of an AAA, it was correct 78% of the time (PPV). However, it did not confirm the absence of an AAA in <8% of cases in which an AAA was truly absent. We believe these results are adequate (F1 score, 84%) for the clinical purposes of AAA EMR screening, especially because the sensitivity was quite high, and a case manager could review all confirmed cases of AAA absence to ensure accuracy.

The present study had limitations. Many patients in the registry had already been enrolled in a AAA screening program in accordance with the U.S. Preventive Services recommendations and, thus, by definition, had been preselected to have an elevated pretest probability of an AAA. Next, the NLP relies on a single measurement (outer wall to outer wall) assessment of the AAA, which has been the standard practice for decades but does not consider measurement error and the lack of standardization of the interpretation technique (eg, orthogonal, short axis, oblique, other definitions of aortic diameter) among radiologists. Rarely, the algorithm will mistake the size of a solid organ or other lesion as the aortic size. Finally, our NLP tool, like other text search mechanisms, can only evaluate machine-readable text.[29] It cannot interpret images, which might include scanned documents and/or the raw radiographic or ultrasound images. Moreover, the NLP depends on the radiologist dictating a visual description of the abdominal aorta and/or AAA. Thus, if these structures are not mentioned, no information regarding them can be captured.

CONCLUSION

The results from the present study have demonstrated that an AAA NLP search engine can analyze very large volumes of imaging report data and accurately detect the presence of, and confirm the absence of, an AAA to a clinically useful degree. In addition, when a maximal abdominal aortic diameter is reported, we found a high degree of certainty that the NLP tool will detect it and, more importantly, report it correctly. An NLP with this degree of accuracy allows for the assembly of a contemporary diameter-based cohort of AAA patients for longitudinal (natural history) analysis to guide surveillance, medical management, and operative decision making. It can also be used to identify pre- and postoperative AAA patients "lost to follow-up" in the EMR, leverage human resources engaged in the ongoing surveillance of AAA patients, and facilitate cost-effective implementation of AAA screening programs.

AUTHOR CONTRIBUTIONS

Conception and design: MM, SO, EM, AA, RC
Analysis and interpretation: MM, SO, EL, MH, JA, AA, RC
Data collection: MM, SO, MH, AA, RC
Writing the article: MM, SO, EL, MH, AA, RC
Critical revision of the article: MM, SO, EL, JA, EM, AA, RC
Final approval of the article: MM, SO, EL, MH, JA, EM, AA, RC
Statistical analysis: MM, SO, EL, MH, JA, EM, AA, RC
Obtained funding: EM
Overall responsibility: RC

REFERENCES

1. Centers for Disease Control and Prevention. CDC Wonder. Available at: http://wonder.cdc.gov/. Accessed January 12, 2020.
2. RESCAN Collaborators, Bown MJ, Sweeting MJ, Brown LC, Powell JT, Thompson SG. Surveillance intervals for small abdominal aortic aneurysms: a meta-analysis. JAMA 2013;309:806-13.
3. Li X, Zhao G, Zhang J, Duan Z, Xin S. Prevalence and trends of the abdominal aortic aneurysms epidemic in general population—a meta-analysis. PLoS One 2013;8:e81260.
4. Darling R, Messina C, Brewster D, Ottinger L. Autopsy study of unoperated abdominal aortic aneurysms: the case for early resection. Circulation 1977;56:II164-74.
5. Szilagyi DE, Smith RF, DeRusso FJ, Elliott JP, Sherrin FW. Contribution of abdominal aortic aneurysmectomy to prolongation of life. Ann Surg 1966;164:678-99.
6. Lederle FA. Rupture rate of large abdominal aortic aneurysms in patients refusing or unfit for elective repair. JAMA 2002;287:2968-72.
7. The UK Small Aneurysm Trial Participants. The U.K. small aneurysm trial: design, methods and progress. Eur J Vasc Endovasc Surg 1995;9:42-8.
8. McCarthy RJ, Shaw E, Whyman MR, Earnshaw JJ, Poskitt KR, Heather BP. Recommendations for screening intervals for small aortic aneurysms. Br J Surg 2003;90:821-6.
9. van Walraven C, Wong J, Morant K, Jennings A, Austin PC, Jetty P, et al. Radiographic monitoring of incidental abdominal aortic aneurysms: a retrospective population-based cohort study. Open Med 2011;5:e67-76.
10. Ahmad M, Mistry R, Hodson J, Bradbury AW. How quickly do asymptomatic infrarenal abdominal aortic aneurysms grow and what factors affect aneurysm growth rates? Analysis of a single centre surveillance cohort database. Eur J Vasc Endovasc Surg 2017;54:597-603.
11. Mell MW, Baker LC, Dalman RL, Hlatky MA. Gaps in preoperative surveillance and rupture of abdominal aortic

aneurysms among Medicare beneficiaries. J Vasc Surg 2014;59:583-8.
12. Chun KC, Schmidt AS, Bains S, Nguyen AT, Samadzadeh KM, Wilson MD, et al. Surveillance outcomes of small abdominal aortic aneurysms identified from a large screening program. J Vasc Surg 2016;63:55-61.
13. Schermerhorn ML, Bensley RP, Giles KA, Hurks R, O'Malley AJ, Cotterill P, et al. Changes in abdominal aortic aneurysm rupture and short-term mortality, 1995-2008: a retrospective observational study. Ann Surg 2012;256:651-8.
14. Nevitt MP, Ballard DJ, Hallett JW Jr. Prognosis of abdominal aortic aneurysms. N Engl J Med 1989;321:1009-14.
15. De Haro J, Acin F, Bleda S, Varela C, Medina FJ, Esparza L. Prediction of asymptomatic abdominal aortic aneurysm expansion by means of rate of variation of C-reactive protein plasma levels. J Vasc Surg 2012;56:45-52.
16. Powell JT, Sweeting MJ, Brown LC, Gotensparre SM, Fowkes FG, Thompson SG. Systematic review and meta-analysis of growth rates of small abdominal aortic aneurysms. Br J Surg 2011;98:609-18.
17. Lederle FA, Noorbaloochi S, Nugent S, Taylor BC, Grill JP, Kohler TR, et al. Multicentre study of abdominal aortic aneurysm measurement and enlargement. Br J Surg 2015;102:1480-7.
18. Thompson SG, Brown LC, Sweeting MJ, Bown MJ, Kim LG, Glover MJ, et al. Systematic review and meta-analysis of the growth and rupture rates of small abdominal aortic aneurysms: implications for surveillance intervals and their cost-effectiveness. Health Technol Assess 2013;17:1-118.
19. Smith-Bindman R, Miglioretti DL, Larson EB. Rising use of diagnostic medical imaging in a large integrated health system. Health Aff (Millwood) 2008;27:1491-502.
20. Chaikof EL, Dalman RL, Eskandari MK, Jackson BM, Lee WA, Mansour MA, et al. The Society for Vascular Surgery practice guidelines on the care of patients with an abdominal aortic aneurysm. J Vasc Surg 2018;67:2-77.e2.
21. Oliver-Williams C, Sweeting MJ, Jacomelli J, Summers L, Stevenson A, Lees T, et al. Safety of men with small and medium abdominal aortic aneurysms under surveillance in the NAAASP. Circulation 2019;139:1371-80.
22. van Walraven C, Wong J, Morant K, Jennings A, Jetty P, Forster AJ. Incidence, follow-up, and outcomes of incidental abdominal aortic aneurysms. J Vasc Surg 2010;52:282-90.e2.
23. van Walraven C, Wong J, Morant K, Jennings A, Austin PC, Jetty P, et al. The influence of incidental abdominal aortic aneurysm monitoring on patient outcomes. J Vasc Surg 2011;54:1290-7.e2.
24. Badger SA, Jones C, McClements J, Lau LL, Young IS, Patterson CC. Surveillance strategies according to the rate of growth of small abdominal aortic aneurysms. Vasc Med 2011;16:415-21.
25. O'Donnell TFX, Schermerhorn ML. Abdominal aortic aneurysm screening guidelines: United States Preventative Services Task Force and Society for Vascular Surgery. J Vasc Surg 2020;71:1457-8.
26. Sohn S, Ye Z, Liu H, Chute CG, Kullo IJ. Identifying abdominal aortic aneurysm cases and controls using natural language processing of radiology reports. AMIA Jt Summits Transl Sci Proc 2013;2013:249.
27. Morioka C, Meng F, Taira R, Sayre J, Zimmerman P, Ishimitsu D, et al. Automatic classification of ultrasound screening examinations of the abdominal aorta. J Digit Imaging 2016;29:742-8.
28. Wong A, Plasek JM, Montecalvo SP, Zhou L. Natural language processing and its implications for the future of medication safety: a narrative review of recent advances and challenges. Pharmacotherapy 2018;38:822-41.
29. Lee JK, Jensen CD, Levin TR, Zauber AG, Doubeni CA, Zhao WK, et al. Accurate identification of colonoscopy quality and polyp findings using natural language processing. J Clin Gastroenterol 2017;53:e25-30.
30. Cai T, Giannopoulos AA, Yu S, Kelil T, Ripley B, Kumamaru KK, et al. Natural language processing technologies in radiology research and clinical applications. Radiographics 2016;36:176-91.
31. Kang SK, Garry K, Chung R, Moore WH, Iturrate E, Swartz JL, et al. Natural language processing for identification of incidental pulmonary nodules in radiology reports. J Am Coll Radiol 2019;16:1587-94.
32. Lederle FA, Wilson SE, Johnson GR, Reinke DB, Littooy FN, Acher CW, et al. Immediate repair compared with surveillance of small abdominal aortic aneurysms. N Engl J Med 2002;346:1437-44.
33. Cao P, De Rango P, Verzini F, Parlani G, Romano L, Cieri E. Comparison of surveillance versus aortic endografting for small aneurysm repair (CAESAR): results from a randomised trial. Eur J Vasc Endovasc Surg 2011;41:13-25.
34. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994;1:161-74.
35. Al-Haddad MA, Friedlin J, Kesterson J, Waters JA, Aguilar-Saavedra JR, Schmidt CM. Natural language processing for the development of a clinical registry: a validation study in intraductal papillary mucinous neoplasms. HPB (Oxford) 2010;12:688-95.
36. Afzal N, Sohn S, Abram S, Scott CG, Chaudhry R, Liu H, et al. Mining peripheral arterial disease cases from narrative clinical notes using natural language processing. J Vasc Surg 2017;65:1753-61.
37. Kim LG, Scott RAP, Ashton HA, Thompson SG. A sustained mortality benefit from screening for abdominal aortic aneurysm. Ann Intern Med 2007;146:699-706.
38. Thompson S, Ashton H, Gao L, Buxton M, Scott R. Multicentre Aneurysm Screening Study Group. Final follow-up of the Multicentre Aneurysm Screening Study (MASS) randomized trial of abdominal aortic aneurysm screening. Br J Surg 2012;99:1649-56.
39. Ali MU, Fitzpatrick-Lewis D, Kenny M, Miller J, Raina P, Sherifali D. A systematic review of short-term vs long-term effectiveness of one-time abdominal aortic aneurysm screening in men with ultrasound. J Vasc Surg 2018;68:612-23.
40. Wild JB, Stather PW, Biancari F, Choke EC, Earnshaw JJ, Grant SW, et al. A multicentre observational study of the outcomes of screening detected sub-aneurysmal aortic dilatation. Eur J Vasc Endovasc Surg 2013;45:128-34.
41. Summers KL, Kerut EK, Sheahan CM, Sheahan MG III. Evaluating the prevalence of abdominal aortic aneurysms in the United States through a national screening database. J Vasc Surg 2021;73:61-8.
42. Guirguis-Blake JM, Beil TL, Senger CA, Coppola EL. Primary care screening for abdominal aortic aneurysm: updated evidence report and systematic review for the US Preventive Services Task Force. JAMA 2019;322:2219-38.
43. Chun KC, Anderson RC, Smothers HC, Sood K, Irwin ZT, Wilson MD, et al. Risk of developing an abdominal aortic aneurysm after ectatic aorta detection from initial screening. J Vasc Surg 2020;71:1913-9.

Submitted Aug 17, 2020; accepted Dec 23, 2020.

Additional material for this article may be found online at www.jvascsurg.org.

APPENDIX (online only).

confirmed absence of abdominal aortic aneurysms

In addition to the assessment for accurate size attribution, we evaluated the ability of the natural language processing (NLP) algorithm to correctly discriminate among four mutually exclusive nondiameter bins if no diameter measurement had been given (1, abdominal aortic aneurysm [AAA] present, diameter unknown; 2, confirmed absence of AAA; 3, no mention of aorta in text; and 4, abdominal aorta mentioned but diameter unknown). To develop an algorithm to confirm the absence of an AAA, the "diameter present" bin was then divided into diameters of $<$2.5 cm and $\geq$2.5 cm, yielding a total of six bins (Supplementary Table II, online only).

A confirmed absence of an AAA was defined as the combination of three bins: (1) maximal aortic diameter $<$2.5 cm; (2) a confirmed absence of an AAA; and (3) abdominal aorta mentioned but diameter unknown (Supplementary Fig, online only). The "abdominal aorta mentioned but diameter unknown" bin was logically included here because a board-certified radiologist had mentioned the abdominal aorta explicitly (usually describing other aortic pathology) but had not suggested the diagnosis of an AAA; thus, it could be reasonably assumed no AAA was present.

The presence of an AAA or an unknown AAA status was defined as the combination of the remaining three bins: (1) maximal aortic diameter $>$2.5 cm; (2) AAA present, diameter unknown; and (3) no mention of aorta in text.

We then collapsed this resulting 6$\times$6 table (Supplementary Table II, online only) into a 2$\times$2 table (Table III), representing the case of a confirmed absence of an AAA and an AAA present or unknown.

Supplementary Table I (online only). Current Procedural Terminology codes related to AAA screening or diagnosis

CPT code	CPT Description	Modality
4225	US retroperitoneal aorta screening	US
72131	CT lumbar spine, no contrast	CT
72132	CT lumbar spine, with contrast	CT
72133	CT lumbar spine without and with contrast	CT
72142	MRI spine cervical with contrast	MRI
72148	MRI lumbar spine, no contrast	MRI
72149	MRI of lumbar spine with contrast	MRI
72156	MRI cervical spine without and with contrast	MRI
72158	MRI lumbar spine, without and with contrast	MRI
72191	CT angiography, pelvis, without and with contrast	CT
72192	CT pelvis, no contrast	CT
72193	CT pelvis, with contrast	CT
72194	CT pelvis, without and with contrast	CT
72195	MRI pelvis, no contrast	MRI
72196	MRI pelvis, with contrast	MRI
72197	MRI pelvis, without and with contrast	MRI
72198	MRA of pelvis	MRA
74150	CT abdomen, no contrast	CT
74160	CT abdomen, with contrast	CT
74170	CT abdomen without and with contrast and further sections	CT
74174	CT angiography, abdomen and pelvis, with contrast, including without contrast	CT
74175	CT angiography, abdomen	CT
74176	CT abdomen and pelvis, no contrast	CT
74177	CT abdomen and pelvis, with contrast	CT
74178	CT abdomen and pelvis, without and with contrast	CT
74181	MRI abdomen, no contrast	MRI
74182	MRI abdomen, with contrast	MRI
74183	MRI abdomen without and with contrast	MRI
74185	MRA abdomen	MRA
75635	CT angiography, abdominal aorta and bilateral iliofemoral lower extremity	CT
76380	CT, follow-up	CT
76700	US abdomen	US
76705	US abdomen, B-scan, limited	US
76770	US B-scan retroperitoneal, complete	US
76775	US B-scan, retroperitoneal, limited	US
78812	PET tumor imaging, skull base to mid-thigh	PET
78813	PET study for tumor, whole body	PET
78814	PET with concurrent CT, tumor localization, limited area	PET
78815	PET with concurrent CT, tumor localization, skull base to mid-thigh	PET
78816	PET with concurrent CT, tumor localization, whole body	PET
93978	US duplex with color Doppler, aorta, IVC, and iliac vessels	US

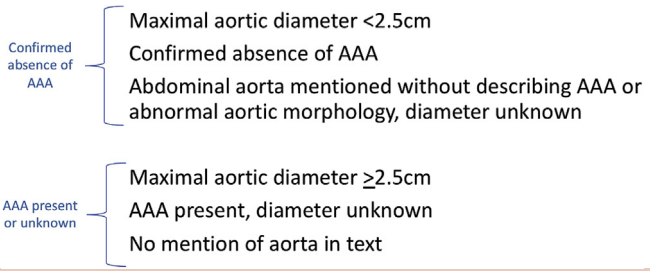
AAA, Abdominal aortic aneurysm; *CPT*, Current Procedural Terminology; *CT*, computed tomography; *IVC*, inferior vena cava; *MRA*, magnetic resonance angiography; *MRI*, magnetic resonance imaging; *PET*, positron emission tomography; *US*, ultrasound.

Supplementary Table II (online only). Expert reviewers vs NLP for categorization of AAA in two diameter size bins and four nondiameter size bins and interrater reliability assessment

	Expert reviewers					
NLP	Aorta <2.5 cm	Aorta ≥2.5 cm	AAA present, diameter unknown	Confirmed absence of AAA	No mention of aorta in text	Abdominal aorta mentioned, diameter unknown
Aorta <2.5 cm	405	141	6	20	2	7
Aorta ≥2.5 cm	26	6148	12	27	0	5
AAA present, diameter unknown	18	1295	278	118	13	72
Confirmed absence of AAA	96	218	85	1627	5	161
No mention of aorta in text	0	110	12	21	4919	46
Abdominal aorta mentioned, diameter unknown	26	380	222	341	31	1107

	Interrater reliability assessment			
	Kappa (95% CI)	PPV	Sensitivity	F1 score
NLP vs reviewers	0.734 (0.727-0.742)	0.878	0.790	0.690
Reviewer 1 vs 2	0.946 (0.934-0.958)	0.963	0.964	0.895

AAA, Abdominal aortic aneurysm; *CI*, confidence interval; *NLP*, natural language processing; *PPV*, positive predictive value.



Supplementary Fig (online only). Definition of confirmed absence of abdominal aortic aneurysms (*AAA*s).