Identifying Abdominal Aortic Aneurysm Cases and Controls using Natural Language Processing of Radiology Reports

Sunghwan Sohn, PhD1, Zi Ye, MD, PhD2, Hongfang Liu, PhD1, Christopher G. Chute, MD, DrPH1, Iftikhar J. Kullo, MD2
1Division of Biomedical Statistics and Informatics, 2Division of Cardiovascular Diseases
Mayo Clinic, Rochester MN

Abstract
Prevalence of abdominal aortic aneurysm (AAA) is increasing due to longer life expectancy and implementation of screening programs. Patient-specific longitudinal measurements of AAA are important to understand pathophysiology of disease development and modifiers of abdominal aortic size. In this paper, we applied natural language processing (NLP) techniques to process radiology reports and developed a rule-based algorithm to identify AAA patients and also extract the corresponding aneurysm size with the examination date. AAA patient cohorts were determined by a hierarchical approach that: 1) selected potential AAA reports using keywords; 2) classified reports into AAA-case vs. non-case using rules; and 3) determined the AAA patient cohort based on a report-level classification. Our system was built in an Unstructured Information Management Architecture framework that allows efficient use of existing NLP components. Our system produced an F-score of 0.961 for AAA-case report classification with an accuracy of 0.984 for aneurysm size extraction.

Introduction
Abdominal aortic aneurysm (AAA) is present in about 10% of men older than 65 years 1,2. Most cases are asymptomatic and detected on imaging studies accidentally. Rupture of an AAA is the most severe complication and associated with a high mortality rate of 90% - the 14th leading cause of death in the U.S.A. 3. The US Preventive Services Task Force recommends to perform one time screening in all men older than 65 years who ever smoked and those older than 60 years with family history of AAA 3,4. The commonly used threshold for diagnosis of AAA is >=3cm and >=5.5cm for surgical repair 5,6. It is generally accepted that larger AAA tends to grow more rapidly than small AAA and at higher risk for rupture, but not in all cases 7,8. Recent studies have shown the influence of genetic factors on AAA 9,10. Several genetic variants have been reported to be associated with AAA 11-13. Identifying genetic determinants associated with AAA can facilitate understanding of underlying pathophysiology and modifiers of abdominal aortic size. As the first step to this end, we described the development of a tool to identify patients with AAA and extract information on progression by extracting sequential AAA size changes using natural language processing (NLP). The study of aneurysm progression may identify novel biomarkers of disease development.

In this paper, we implemented a rules-based system that identifies AAA-cases from radiology reports and extracts corresponding AAA sizes with date information. Our system utilized the NLP components from MedTagger 14 to process radiology reports and was built in the Apache Unstructured Information Management Architecture (UIMA) framework that provides an efficient way to add new components as well as reuse the existing NLP modules.

Background
AAA is typically diagnosed by physical examination, ultrasound, or CT scan and its outcomes are recorded in radiology reports. Medical experts can manually review radiology reports to identify AAA patients and extract corresponding AAA sizes for clinical studies and patient history summarization. Manual review, however, is time consuming and often impractical for a routine practice and large-scale clinical studies. In order to overcome these drawbacks, NLP techniques, which process unstructured text and convert it to a structured format, can be used to automatically extract AAA-related information and identify patient cohorts. Over the past decade, advances in NLP have produced promising results in information extraction from clinical text 15 and have been successfully applied in various clinical applications including patient medical status extraction 16,17, sentiment analysis 18, decision support 19,20, genome-wide association studies 21,22, and diagnosis code assignment 23,24. Recently, Mayo Clinic developed MedTagger 14, a NLP pipeline with a fast dictionary lookup, to process clinical text and annotate clinical concepts.

Our system used basic NLP components including dictionary lookup in MedTagger to process unstructured text and find medical concepts from radiology reports. The main annotations include: Sentence Detection parses sentence; Tokenization finds word token boundaries; Normalization generates one form for the various morphological variants
of the word through the NLM’s Lexical Variant Generation (LVG) tool (http://SPECIALIST.nlm.nih.gov), which make it possible to use normalized terms for dictionary lookup; Size Annotation extracts aneurysm size; AAA NE (Named Entity) Detection discovers AAA related concepts based on dictionary lookup; Negation identifies negated NEs; AAA Identification identifies AAA-case reports and extracts the corresponding aneurysm sizes.

Methods

This study used Mayo Clinic radiology reports—including ultrasound, CT, MRI, and angiography report—from eMERGE 25 patient cohorts. However, many radiology reports are not related to AAA. In order to maximize system efficiency, a hierarchical approach was used to determine the AAA patient cohort. First, potential AAA radiology reports were selected using keywords. Secondly, reports were classified into AAA-case vs. non-case using manually-crafted rules based on keywords and aneurysm size. Lastly, the AAA patient cohort was determined based on a report-level classification. The detailed methods are as follows:

A. Selection of potential AAA reports

Mayo’s radiology reports consist of multiple fields including test codes and test descriptions. Initially, we selected radiology reports based on specific CPT (Current Procedural Terminology) codes and code descriptions. However this approach missed many AAA-related reports. A better alternative was to use keywords – i.e., select potential AAA reports that contain both “aorta” and “abdominal” relevant terms because AAA-related reports must include these terms. This keyword-based search was able to catch those reports missed by a code-based search and retrieved a much higher number of potential AAA reports—i.e., out of 180K reports, the code based approach retrieved 3,370 reports and the keyword-based approach retrieved 11,420 reports. Table 1 shows the keywords we used. Those terms were expanded through both UMLS (http://www.nlm.nih.gov/research/umls/) concepts and the most frequent terms used in Mayo clinical notes.

B. AAA report classification

Figure 1 shows the pseudo code of the algorithm. After we selected the potential AAA-case report, each report was classified as AAA-case vs. non-case as follows:

AAA-case: Reports that contain “abdominal aorta” or “abdominal aorta aneurysm” related terms and aneurysm size at the examination date is equal to or greater than 3 cm.

Non-case: 1) Reports that contain status post indications (e.g., aortic/aorto + endograft, abdominal aortic endograft, repair of AAA, s/p AAA repair, endovasc repair AAA, etc).
2) Reports that contain only AAA related terms without the size information or “ectasia of abdominal aorta.”
3) Reports that contain explicit terms indicating “normal” AAA condition (e.g., normal caliber abdominal aorta, normal distal aorta), negated AAA (e.g., negative for abdominal aortic aneurysm), or the aneurysm size is less than 3 cm.
4) Reports that do not contain any AAA related information

It should be noted that our definition of AAA-case in this study excludes patients with open surgery or endovascular repair although they are AAA case in clinical perspective. This is because our system focuses on abstracting aneurysm sizes longitudinally and a track for aneurysm growth.

Size extraction: In radiology reports, AAA sizes are basically expressed as one-, two-, or three-dimension (AP, width/transverse, and length) and described in numerous ways (e.g., 4.4cm, measuring 4.4×5.3 cm, 4.4×5.3×6.1 cm,
maximum AP diameter of 3.7cm and a transverse diameter of 3.7cm, etc.). Although there can be more than one dimension for the aneurysm size description, only one value (i.e., maximum size of either AP or width/transverse) is used to determine an AAA-case. Some radiology reports contain the size(s) from a previous examination, but we only considered the size of the current examination.

The size annotator used regular expressions to extract the size description from free text. Then it selects the maximum value from AP and width/transverse and then normalized the value to “cm” (some values are in “mm”). The sizes that are not associated with the given examination date (i.e., sizes from previous examinations) were excluded based on description patterns as follows:

a. Select the size that comes with current indication word (e.g., “now measures/measuring”).
b. Exclude the size that comes with previous indication words (e.g, “previously/earlier/prior” “previous measurement(s) was/were” “prior exam” “compared to/with” “increased/decreased from” etc.).

AAA-related keywords: Table 2 includes keywords for abdominal aorta (AA), abdominal aorta aneurysm (AAA), status post (S/P), and normal. They were initially provided by a medical expert and expanded through UMLS concepts and frequent terms used in Mayo clinical notes. They were also normalized into canonical forms through NLM’s LVG in order to match variations (e.g., “aneurysm abdominal” can match with “aneurysm abdominals”). We used the dictionary-lookup in MedTagger to find those keywords.

C. AAA Patient cohort identification

Generally, patients have more than one examination and therefore have more than one report. After we classified all of the report class (i.e., AAA-case vs. Non-case), we can finally determine the AAA patients. If any report for a given patient is an AAA case, then we determine this patient as an AAA patient. Our AAA patient cohort also includes information of corresponding aneurysm size and examination date.

Table 2. AAA related keywords (normalized through LVG)

| AA                                      | AAA                                   | S/P                                      | Normal                                      |
|-----------------------------------------|---------------------------------------|------------------------------------------|---------------------------------------------|
| infrarenal aorta                        | abdominal aorta                       | post a.a.a. repair                       | normal caliber abdominal aorta              |
| abdominal aorta                         | infrarenal aortic aneurysm             | s/p a.a.a. repair                        | normal distal aorta                        |
| aorta                                   | aneurysm abdominal aorta               | endograft                                | abdominal aorta normal caliber             |
| abdominal aneurysm                      | infrarenal aortic aneurysm             | endovascular                            | aorta normal caliber                       |
| infrarenal location                     | infrarenal abdominal aorta             | aneurysm sac                             |                                             |
|                                        | infrarenal aortic aneurysm             | bifurcate endograft                      |                                             |

Results

Figure 2 shows the annotation types and values of a de-identified sample report in UIMA CAS Visual Debugger. The right window shows a radiology report snippet processed to populate annotations as they appear in the left window. The bottom left window shows annotation types and values including the report class and the AAA size. For this example, there are two AAA-related terms, “aaa” and “infrarenal aorta” and two AAA size information (3.6 x 3.4cm and 3.5 x 3.3cm). However, we only consider the AAA size of the current examination date (3.6 x 3.4cm). As the final size, we extract the larger one (3.6 cm).

A medical expert manually examined 650 radiology reports and classified them as AAA-case vs. non-case. If the report was an AAA case, the corresponding size was also extracted. We used 400 reports to train the system and held-out 250 reports to test. Our system was able to catch most AAA-case reports with a high F-score of 0.961 (61 TPs, 4 FPs, and 1 FNs) on the test set. Table 3 shows the corresponding evaluation performance.

The performance of AAA patient cohort identification is in Table 4. The test set contains 25 AAA-case patients and the system identified 27 patients as an AAA-case which led 2 FPs and 0 FN.

Our system was also able to generate sequential size variations in time that are required to build a sophisticated AAA phenotype algorithm in the future. For example:

PatientID(2.5cm:**/**/1999)/3.2cm:**/**/2008)/3.5cm:**/**/2009)/3.6cm:**/**/2010)/3.8cm:**/**/2011
Table 3. AAA-case report classification in the test set

| Evaluation     | value |
|----------------|-------|
| precision      | 0.939 |
| recall         | 0.984 |
| F-score        | 0.961 |
| size accuracy* | 0.984 |

* # correct sizes / #TP AAA cases

Discussion

Our system was able to classify most AAA-case reports with a high F-score. There was one false negative case due to the S/P relevant term “endovascular.” This report contained the term “pre-endovascular,” indicating it is a report before surgery and should not be treated as S/P. False positive cases were due to: incorrect negation (e.g., abdominal aorta negative for aneurysm - “aneurysm” is negated, but “abdominal aorta” is not), incorrect size determination (e.g., “under 3cm” was not treated as < 3cm), incorrect association with other than “abdominal” aorta (e.g., “a fusiform 5.5cm aneurysm of the distal thoracic and upper abdominal aorta extending...”).

A radiology report could contain more than one AAA size description, mainly due to a size description from the previous examination. We effectively eliminated the previous size by filtering out the size associated with words that indicate “previous” and achieved an accuracy of 0.984 (size accuracy in Table 3).

The AAA patient cohort identification was based on a simple rule—i.e., examining report-level AAA classification. Although report-level classification is not perfect, it is possible to identify an AAA patient if a patient has more than one AAA-case report and one of them is correctly classified.

Our results show that a rule-based system using NLP techniques could effectively identify an AAA patient cohort and extract aneurysm size from radiology reports. There is a potential role for an NLP-based size extractor to generate an electronic alert that will notify the referring physician about an AAA that exceeds a certain size threshold. Our approach may be helpful in ascertaining the presence of pathologies from radiology reports, which have size-based criteria, by adjusting pattern matching rules; for example, cerebral or other arterial aneurysms. The automated system for AAA patient cohort identification enables large-scale clinical study. Currently, our system is being applied to a larger patient cohort to identify AAA patients with size and date information for the eMERGE II phenotype study.

Acknowledgements

This work was supported by eMERGE II (HG06379) and Strategic Health IT Advanced Research Projects (SHARP) Program (90TR002).

References

1. Guirguis-Blake J, Wolff TA. Screening for abdominal aortic aneurism. *Am Fam Physician*. Jun 1 2005;71(11):2154-2155.
2. Lederle FA. Ultrasonographic screening for abdominal aortic aneurysms. *Ann Intern Med.* Sep 16 2003;139(6):516-522.

3. Roger VL, Go AS, Lloyd-Jones DM, et al. Executive summary: heart disease and stroke statistics--2012 update: a report from the American Heart Association. *Circulation.* Jan 3 2012;125(1):188-197.

4. Harris R, Sheridan S, Kinsinger L. Time to Rethink Screening for Abdominal Aortic Aneurysm? Comment on "Impact of the Screening Aneurysm Act on Abdominal Aneurysms Very Efficiently (SAAAVE) Act on Abdominal Ultrasonography Use Among Medicare Beneficiaries". *Arch Intern Med.* Sep 17 2012:1-2.

5. Lederle FA, Johnson GR, Wilson SE, et al. Rupture rate of large abdominal aortic aneurysms in patients refusing or unfit for elective repair. *Jama.* Jun 12 2002;287(22):2968-2972.

6. Solberg S, Forsdahl SH, Singh K, Jacobsen BK. Diameter of the infrarenal aorta as a risk factor for abdominal aortic aneurysm: the Tromso Study, 1994-2001. *Eur J Vasc Endovasc Surg.* Mar 2010;39(3):280-284.

7. Li ZY, Sadat U, J UK-I, et al. Association between aneurysm shoulder stress and abdominal aortic aneurysm expansion: a longitudinal follow-up study. *Circulation.* Nov 2 2010;122(18):1815-1822.

8. Wilson KA, Lee AJ, Lee AJ, et al. The relationship between aortic wall distensibility and rupture of infrarenal abdominal aortic aneurysm. *J Vasc Surg.* Jan 2003;37(1):112-117.

9. Ogata T, MacKean GL, Cole CW, et al. The lifetime prevalence of abdominal aortic aneurysms among siblings of aneurysm patients is eightfold higher than among siblings of spouses: an analysis of 187 aneurysm families in Nova Scotia, Canada. *J Vasc Surg.* Nov 2005;42(5):891-897.

10. Wahlgren CM, Larsson E, Magnusson PK, Hultgren R, Swedenborg J. Genetic and environmental contributions to abdominal aortic aneurysm development in a twin population. *J Vasc Surg.* Jan 2010;51(1):3-7; discussion 7.

11. Bown MJ, Jones GT, Harrison SC, et al. Abdominal aortic aneurysm is associated with a variant in low-density lipoprotein receptor-related protein 1. *Am J Hum Genet.* Nov 11 2011;89(5):619-627.

12. Gretarsdottir S, Baas AF, Thorleifsson G, et al. Genome-wide association study identifies a sequence variant within the DAB2IP gene conferring susceptibility to abdominal aortic aneurysm. *Nat Genet.* Aug 2010;42(8):692-697.

13. Helgadottir A, Thorleifsson G, Magnusson KP, et al. The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. *Nat Genet.* Feb 2008;40(2):217-224.

14. Wagholikar K, Torii M, Jonnalagadda S, Liu H. Feasibility of pooling annotated corpora for clinical concept extraction. Paper presented at: Proceedings AMIA CRI 2012; San Francisco, CA.

15. Savova G, Masanz J, Ogren P, et al. Mayo Clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc.* 2010;17(5):507-513.

16. Sohn S, Kocher J-PA, Chute CG, Savova GK. Drug side effect extraction from clinical narratives of psychiatry and psychology patients. *J Am Med Inform Assoc.* 2011;18(Suppl 1):144-149.

17. Sohn S, Savova GK. Mayo Clinic Smoking Status Classification System: Extensions and Improvements. Paper presented at: AMIA Annual Symposium2009; San Francisco, CA.

18. Sohn S, Torii M, Li D, Wagholikar K, Wu S, Liu H. A Hybrid Approach to Sentiment Sentence Classification in Suicide Notes. *Biomedical informatics insights.* 2012(Suppl. 1):43-50.

19. Demner-Fushman D, Chapman WW,зд Chapman W. Using natural language processing in decision support. *Journal of Biomedical Informatics.* 2009;42(5):760-772.

20. Aronsky D, Fiszman M, Chapman WW, Haug PJ. Combining decision support methodologies to diagnose pneumonia2001.

21. Kullo IJ, Fan J, Pathak J, Savova GK, Ali Z, Chute CG. Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. *Journal of the American Medical Informatics Association.* 2010;17(5):568-574.

22. Kullo IJ, Ding K, Jouni H, Smith CY, Chute CG. A genome-wide association study of red blood cell traits using the electronic medical record. *PLoS One.* 2010;5(9):e13011.

23. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association.* 2004;11(5):392.

24. Pakhomov SVS, Buntrock JD, Chute CG. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association.* 2006;13(5):516-525.

25. McCarty C, Chisholm R, Chute C, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC medical genomics.* 2011;4(1):13.