

specificity, and discovered that structured data (ICD-9 and Current Procedural Terminology (CPT) [11] codes) can only identify less than 11% of the cases. Liao et al [12] demonstrated in a rheumatoid arthritis study that utilizing narrative data in the EMR resulted in a significantly higher positive predictive value (0.94) than using codified data alone (0.88). Fisman et al [13] compared the performance of an NLP system, designed to detect pneumonia-related concepts and deduce the presence and absence of acute bacterial pneumonia, to that of human experts, and concluded that their system performed similarly to the experts and better than lay persons and keyword searches. The Informatics for Integrating Biology to the Bedside (i2b2) has organized a series of challenges to identify patient smoking status [14], obesity and its co-morbidities [15] [16], medication, and assertion. Xu et al [17] applied NLP techniques to extract medication information from discharge summaries and achieved 0.90 to 0.96 F-measure.

The goal of this paper is to apply, extend and evaluate an open source clinical NLP system for the task of phenotype extraction from radiology notes for a specific clinical research problem, that of identifying PAD cases.

Methods

Classification categories and gold standard

The annotation guidelines used to create the gold standard were developed by a cardiovascular specialist (IJK). There are 4 classification categories: positive PAD (POS), negative PAD (NEG), probable PAD (PROB) and unknown (UNK). POS is indicated by the presence of severe stenosis or occlusion in a lower extremity artery (including and distal to the iliac artery; prior revascularization by stenting or balloon angioplasty, or surgical endarterectomy or graft placement in the absence of aneurysmal disease as an indication). PROB is defined as a moderate stenosis in a lower limb artery. NEG is the absence of a moderate or severe stenosis or occlusion in a lower extremity artery or presence of only mild stenosis. UNK reflects the lack of information thus the inability to bin into any of the above categories.

Manual annotations to create the gold standard were performed by a cardiovascular research fellow. A set of 135 radiology notes was used for the guidelines development and training of the fellow and the system. The final gold standard consists of 455 radiology notes each labeled with one of the four categories. The distribution of the final gold standard is $N(\text{POS}) = 223$, $N(\text{NEG}) = 19$, $N(\text{PROB}) = 63$, $N(\text{UNK}) = 150$. This set of 455 notes was the dataset for the final system evaluation.

Natural Language Processing toolset

We used Mayo's clinical Text Analysis and Knowledge Extraction System (cTAKES) [18]. In summary, cTAKES is envisioned as a comprehensive NLP toolset for processing the clinical narrative and extracting information from it. Current cTAKES annotators constitute a pipeline of NLP components such as sentence boundary detector, tokenizer, part-of-speech tagger and shallow parser. Currently, the highest-level component discovers clinical named entities (NEs) of type diseases/disorders, signs/symptoms, anatomical sites, procedures and drugs. Each discovered NE has attributes for (1) the text span, (2) the terminology/ontology code the NE maps to, (3) the negation attribute to indicate whether the NE is negated, and (4) the status with a value of *current*, *history of*, *family history of*, or *possible*. cTAKES is being extended with modules for coreference resolution, temporal relations, semantic role labeling and clinical question answering, all of which will be made available open source.

Technical approach

Radiology notes for the following procedures were assessed: lower extremity angiograms (conventional, magnetic resonance, or computed tomography) and lower extremity ultrasound. Radiology notes describing exams of other body parts, e.g. abdomen, chest or pelvis, were filtered out in a pre-processing step. These filtered out documents are classified as UNK. Exams are coded in CPT and these codes are contained in the header of each radiology note; their usage is being referred to in the "ultrasound or vascular interventional radiology report" parts of the pseudo code below. Each radiology note describing a relevant exam after the CPT code filtering was processed through cTAKES and then classified as follows based on the discovered evidence:

```
//this function returns a document label
if (POS evidence exists)
    return POS;
else if (PROB evidence exists)
    return PROB document label;
else if (NEG evidence exists)
    if (report type == ultrasound
        || report type == vascular interventional
        radiology)
        return UNK;
    else
        return NEG;
else
    return UNK;
```

The negative evidence in Ultrasound and Vascular Intervention Radiology tests is not considered strong enough as these two tests are localized. The precedence for multiple evidence within a radiology note was POS, PROB, NEG, UNK.

The PROB, POS and NEG evidence was extracted as follows. The first step was to discover the relevant NEs. cTAKES components for sentence boundary detection, tokenization, part-of-speech tagging and shallow parsing were applied without any modifications. We used cTAKES NE recognition (NER) module with dictionaries tailored to the specific task of PAD discovery as created by the cardiovascular specialist (IJK) and the CV fellow. Table 1 summarizes these dictionaries. For terms with Unified Medical Language System (UMLS) [19] concept unique identifiers (CUI), the ontology code attribute for that concept is populated with the CUI, otherwise that attribute gets a *null* value. The list of modifiers associated with POS PAD evidence includes *extensive*, *complete*, *high-grade*, *severe*, *significant*, *moderate*, *moderate focal*, *multi-focus*. That for PROB PAD evidence includes *diffusely*, *small amount of*, *some*. A signal for PROB PAD evidence is also associated with a *null* modifier for *stenosis*, *plaque* and *atheromatous*.

Dictionary type	Number of dictionary entries	Examples
Anatomical sites	71 (29)	<i>common femoral artery</i> <i>posterior tibial artery</i> <i>femoropopliteal artery</i>
Disorders	60 (51)	<i>complete occlusion</i> <i>atherosclerosis</i> <i>stenosis</i>
Procedures	22 (13)	<i>balloon angioplasty</i> <i>artery bypass</i>

Table 1. Summary of dictionaries used with cTAKES NER module for PAD case discovery. Numbers in brackets are terms without UMLS CUIs.

To discover PAD evidence, a relevant disorder and/or procedure must occur with a vascular anatomical site associated with the lower extremities. Thus, an asserted relation between the two NEs needs to be extracted which constitutes the second step in the evidence-level algorithm. For that, we developed a new cTAKES module for two types of relations which correspond to the UMLS relations of *location_of* and *degree_of*. An asserted relation is annotated if a procedure or a disorder term occurs with an anatomical site with an optional modifier within a given window size. We defined the window size as the sentence. For example, in the sentence “Aortogram and pelvic angiography was obtained, revealing **moderate to high-grade stenosis** within the proximal external **iliac arteries** bilaterally.”, the PAD disorder **moderate to high-grade stenosis** was related to the anatomical site of **iliac arteries**. We also used cTAKES negation detection module to

discover negated NEs. For example, in the sentence “The internal and external iliac arteries are well seen and the **common femoral artery** is pristine without focal changes of **atherosclerosis**.”, the disorder term **atherosclerosis** was negated thus leading to a negated *location_of*(*atherosclerosis*, *common femoral artery*) relation. The term “patent” was added as a negation word. Some of the dictionary entries are stand-alone terms, i.e. they do not need to participate in an asserted relation to signal PAD, e.g. *severe atherosclerotic disease*.

Evaluation

Accuracy is used to report the overall agreement between the system output and the gold standard:

$$(1) \text{ accuracy} = \frac{\text{systemCorrectLabels}}{\text{totalGoldStandardLabels}}$$

We also report results per category in terms of sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV):

$$(2) \text{ sensitivity} = \frac{\text{truePositives}}{\text{truePositives} + \text{falseNegatives}}$$

$$(3) \text{ specificity} = \frac{\text{trueNegatives}}{\text{falsePositives} + \text{trueNegatives}}$$

$$(4) \text{ PPV} = \frac{\text{truePositives}}{\text{truePositives} + \text{falsePositives}}$$

$$(5) \text{ NPV} = \frac{\text{trueNegatives}}{\text{trueNegatives} + \text{falseNegatives}}$$

The baseline method is NER without the relation assertion.

Results and Discussion

The contingency matrix for the four categories is in Table 2. The results for the sensitivity, specificity, PPV and NPV per category are in Table 3. Overall accuracy agreement between the system and the gold standard was 0.930 as compared to the baseline system accuracy of 0.461.

		System				Total
		POS	NEG	UNK	PROB	
Gold standard	POS	221 (50)	1 (49)	0 (84)	1 (40)	223
	NEG	16 (13)	1 (6)	1 (6)	2 (0)	19
	UNK	1 (0)	2 (2)	147 (147)	0 (1)	150
	PROB	15 (0)	3 (9)	5 (25)	40 (29)	63
Total		237 (50)	22 (73)	153 (262)	43 (70)	455

Table 2. Contingency matrix for system and gold standard agreement. Results in brackets are for the baseline system.

Sensitivity for POS, PROB and UNK ranged 0.93-0.96 and was lower for NEG (0.72). Specificity for all categories is in the 90's. The PPV for the POS and

UNK categories is very strong (high 90's), however it dropped for NEG (0.84) and was lower for PROB (0.63). The NPV for all categories was strong. The performance of the baseline is explained by the discovery of stand-alone terms as it lacks the relation assertion component.

	Sensitivity	Specificity	PPV	NPV
POS	0.932 (1.000)	0.990 (0.573)	0.991 (0.224)	0.931 (1.000)
NEG	0.727 (0.178)	0.993 (0.980)	0.842 (0.684)	0.986 (0.829)
UNK	0.960 (0.826)	0.990 (0.990)	0.960 (0.987)	0.980 (0.860)
PROB	0.930 (0.420)	0.944 (0.944)	0.634 (0.558)	0.992 (0.907)

Table 3: Evaluation results per category. Results in brackets are for the baseline system.

The error analysis of the NEG and PROB results showed several main sources of errors. The first one was in the negation and certainty detection, for example the document containing the sentence

*Example 1: The left **posterior tibial artery** cannot be identified and could be **occluded***

was classified by the system as NEG, while the gold standard was POS. The system's label was due to the incorrect negation of *occluded*. Another example is the document containing the evidence

*Example 2: Both **popliteal arteries** are patent with mild to **moderate atheromatous plaque**.*

which had a gold standard label of PROB. The system incorrectly assigned the NEG category based on the presence of "patent".

A similar example is

*Example 3: The **popliteal artery** aneurysms bilaterally are patent and contain **moderate atheromatous plaque** and/or mural thrombus*

where the system label was NEG, while the gold standard was PROB.

The second source of errors was in the gold standard. For example, the document containing the sentence

*Example 4: **Focal mild stenosis** of the proximal right external **iliac artery***

had a gold standard PROB label. However when we asked the cardiovascular fellow annotator what was the motivation for this classification, she said that the combination of "mild" and "iliac artery" is very unlikely to cause flow limitation to the legs and thereby unlikely to cause clinically manifest PAD, which led to a reclassification to a NO. That source of errors emphasizes the degree of cognitive burden the manual abstraction task poses to the expert, thus leading to occasional errors.

The main source of system errors for the PROB category classification was in the discovery of the disease modifiers and correctly asserting the relation between the disease and the severity indicator. The indicators include traditional severity terms like "mild", "moderate", "diffuse" and "some" as well as procedures like "graft" and "stent" and additional qualifying terms like "ulcerating plaques".

Another area for improvement targets our strategies for managing multiple evidences and higher level discourse phenomena such as coreference. An example text is:

*Example 5: Scattered plaque in both the **left and right common femoral arteries** and upper **superficial femoral arteries**. No evidence of high-grade **stenosis** in these vessels.*

The first sentence provides candidate evidence for a PROB category. The second sentence, however, clearly states a NEG case by mentioning "these vessels" which refers to the arteries described in the previous sentence. That reference link is to be established by a coreference resolution module which our extensions currently lack. Because of this deficiency the note is classified incorrectly as PROB.

The language of the radiology notes posed some challenges. For example, one note consisted of just these two sentences

*Example 6: Both **popliteal arteries** are patent with normal flow. No evidence of aneurysm or Baker's cyst.*

with a gold standard label of NEG. The human annotator actually referred to the CPT code to make the final judgment. This additional knowledge gleaned from the CPT code can be incorporated in the algorithm to provide more context.

Abbreviations are known to present challenges to any NLP system. For example, an abbreviation like "fem-pop" for "femoropopliteal" is likely to be parsed as three tokens which affects the downstream parsing. Because NER is performed on a noun phrase window, the abbreviation will not be considered a candidate for that window which will lead to a missed NE. Thus, the final system classification will be the incorrect UNK. Possible abbreviations of interest need to be included in the dictionary.

A limitation of the current study is the inclusion of pre-coordinated terms in our dictionaries which do not have mappings to UMLS. A more elegant solution would decompose each term into its basic units. Composite terms could then be assembled

based on relations specified in an ontology using a combination of rule-based and machine learning methods.

In this paper, we described our first step towards a long-term goal of relation discovery from the clinical narrative, a topic that we are actively pursuing, e.g. temporal and coreference relations between patient's events [20]. The cTAKES extensions, dictionaries and list of CPT codes will be released on the eMERGE website and www.ohnlp.org in fall of 2010.

Conclusion

In this paper, we applied, extended and evaluated a comprehensive clinical natural language processing system, cTAKES, to the discovery of peripheral artery disease cases from radiology notes. Our next steps will be (1) improving the PROB and NEG group classification, (2) scaling up to processing 700,000 radiology notes, and (3) merging with information from other EMR components to enable EMR-wide phenotype extraction. We also plan to evaluate the algorithm on data provided by eMERGE consortium sites to test its portability.

Acknowledgements

This work was funded by grant U01-HG04599 as part of the Mayo eMERGE study (PI Chute).

References

1. eMERGE, https://www.mc.vanderbilt.edu/victr/dcc/projects/acc/index.php/Main_Page.
2. Hirsch, A., Criqui, M., Treat-Jacobson, D., Regensteiner, J., Creager, M., Olin, J., Krook, S., Hunninghake, D., Comerota, A., Walsh, M., McDermott, M., and Hiatt, W. *Peripheral arterial disease detection, awareness, and treatment in primary care*. JAMA, 2001. **286**(11): p. 1317-24.
3. Crowley, R., Castine, M., Mitchell, K., Chavan, G., McSherry, T., and Feldman, M. *caTIES- a grid based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research*. JAMIA, 2010. **17**: p. 253-264.
4. Friedman, C. *A broad-coverage natural language processing system*. Proc AMIA. 2000.
5. *Health Information Text Extraction (HITEx)*: https://www.i2b2.org/software/projects/hitex/hitex_manual.html.
6. *Open Health Natural Language Processing consortium (OHNLP)*: www.ohnlp.org.
7. Li, L., Chase, H., Patel, C., Friedman, C., and Weng, C. *Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening: a case study*. Proc AMIA. 2008.
8. Zeng, Q., Goryachev, S., Weiss, S., Sordo, M., Murphy, S., and Lazarus, R. *Extracting principal diagnosis, comorbidity, and smoking status for asthma research: evaluation of a natural language processing system*. BMC Med Inform Decis Mak, 2006: p. 2006:30.
9. Himes, B., Dai, Y., Kohane, I., Weiss, S., and Ramoni, M. *Prediction of chronic obstructive pulmonary disease (COPD) in asthma patients using electronic medical records*. JAMIA, 2009. **16**: p. 371-379.
10. Penz, J., Wilcox, A., and Hurdle, J. *Automated identification of adverse events related to central venous catheters*. JBI, 2007. **40**: p. 174-182.
11. *Current Procedural Terminology (CPT)*: <http://www.ama-assn.org/ama/pub/physician-resources/solutions-managing-your-practice/coding-billing-insurance/cpt.shtml>.
12. Liao, K., Cai, T., Gainer, V., Goryachev, S., Zeng-Treitler, Q., Raychaudhuri, S., Szolovits, P., Churchill, S., Murphy, S., Kohane, I., Karlson, E., and Plenge, R. *Electronic medical records for discovery research in rheumatoid arthritis*. Arthritis care & research, In Press. doi:10.1002/acr.20184.
13. Fiszman, M., Chapman, W., Aronsky, D., Evans, R., and Haug, P. *Automatic detection of acute bacterial pneumonia from chest X-ray reports*. JAMIA, 2000: p. 593-604.
14. Uzuner, Ö., Goldstein, I., Luo, Y., and Kohane, I. *Identifying Patient Smoking Status from Medical Discharge Records*. JAMIA, 2008. **15**(1): p. 14-24.
15. Uzuner, O. *Second i2b2 workshop on natural language processing challenges for clinical records*. Proc AMIA. 2008.
16. Uzuner, Ö. *Recognizing Obesity and Comorbidities in Sparse Data*. JAMIA., 2009. **16**(4): p. 561-570.
17. Xu, H., Stenner, S., Doan, S., Johnson, K., Waitman, L., and Denny, J. *MedEx: a medication information extraction system for clinical narratives*. JAMIA, 2010(17): p. 19-24.
18. Savova, G., Masanz, J., Ogren, P., Zheng, J., Sohn, S., Kipper-Schuler, K., and Chute, C. *Mayo Clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications*. JAMIA, In press.
19. National Library of Medicine, *Unified Medical Language System*: <https://login.nlm.nih.gov/cas/login?service=http://umlsks.nlm.nih.gov/uPortal/Login>. 2010.
20. Savova, G., Bethard, S., Styler, W., Martin, J., Palmer, M., Masanz, J., and Ward, W. *Towards temporal relation discovery from the clinical narrative*. Proc AMIA. 2009. San Francisco, CA.