Summary: Use of natural language processing of unstructured data significantly increases the detection of peripheral arterial disease in observational data.

Background: In observational research, identifying patients with peripheral arterial disease (PAD) is typically done through diagnostic and procedure codes (e.g., ICD and CPT codes). Such approaches are feasible but challenging. Much of the critical information captured by physicians about PAD involves symptoms such as claudication or physical examination findings such as decreased pulses. These elements are rarely available as coded data points, but rather found as free text in clinical documents such as provider notes. Even the ankle-brachial index (ABI)—a common 'quantitative' data point for defining PAD— is generally embedded in the text of radiology reports and not found in structured datasets.

Natural language processing (NLP) of clinical documents has advanced considerably over the past decade (ref) and Regenstrief Institute has created a text-mining platform for the development and validation of NLP algorithms. In this study, we hypothesized that NLP could increase the identification of PAD patients in an electronic medical record compared with using structured data alone. We tested this hypothesis by developing and validating an NLP-based PAD detection model and comparing its output with that of an established code-based PAD definition run on the same population.

Methods:
Our target population was patients from two major medical systems in Indianapolis (Eskenazi Health and Indiana University Health) who were seen between January 1st 2009 and December 31st 2014 and had at least one clinical document stored in our system (n=1,977,827 patients). For our structured PAD definition, we used the algorithm developed Kullo et al [1,2]. Because our data warehouse does not include ankle-brachial indices in structured data, we separated the Kullo algorithm into two components: ICD9 / CPT-based criteria and ABI-based criteria. For the first component, we performed direct queries of our diagnosis and procedure tables. For the second component we developed a text-mining algorithm to cull ABI values from vascular laboratory reports. We set the PAD threshold as ABI < 0.9 (patients with ABIs > 1.4 were not included in our PAD cohort). The accuracy of our ABI extraction process was assessed by manual extraction of ABI values from 100 ABI reports by a blinded reviewer and comparing the manual and automated results.

To develop our algorithm for detecting PAD from unstructured data, we began by selecting 4 clinical concepts highly suggestive of PAD.

- Claudication
- Rest pain
- Diminished pulses
- Limb ischemia
- Peripheral arterial disease (direct documentation by provider)

For each of the above concepts, we utilized expert input to choose an initial set of words or phrases that commonly appear in clinical documentation to represent these concepts. These term sets were expanded through the use of clinical ontologies (MedDRA and SNOMED), to gather related phrases and synonyms. Finally, these synonyms were enriched through a process that creates a set of 'lexical variants' (e.g., ischemia → ischemic).

Once these initial stages were completed, we assigned a set of constraints regarding word distance (how far the words within each phrase can be separated while maintaining the same meaning). Where appropriate, we added other Boolean logic (e.g., claudication NOT (spinal OR jaw OR neurogenic)). Shown below is a partial example of the algorithm for detection of diminished pulses:

- {diminished}{pulse} (in either order, within a 2 word distance)
  - {diminished}
    - Diminish(ed/ment)
    - Decrease(d)
    - Faint
    - Absen(t/ce)
    - 1+
    - 0+
  - {pulse}
    - Pulse(s)
    - Tibialis anterior (TA)
    - Popliteal
    - Dorsalis pedis (DP)

We ran each algorithm on the following document types: provider notes, radiology reports, operative reports, vascular laboratory reports, admission notes, and discharge summaries. Each element of the algorithm was then checked for negation (e.g., "no evidence of decreased pulses") (NegEx ref). We performed 3 cycles of iteration, extracting 50 documents at a time, reviewing for any errors, and then making minor modifications to the algorithms.

For this study, our goal was to be conservative as possible in our definitions, maximizing the positive predictive value of our algorithms to ensure a high quality cohort. As a result, we avoided poorly specific phrases (e.g., 'PAD' itself was removed due to the challenge of disambiguating from of other concepts such as thumb pad, fat pad, sterile pad, etc). We also eliminated all 'hypothetical' mentions of our concepts such as 'may have claudication' or 'rule-out limb ischemia'.

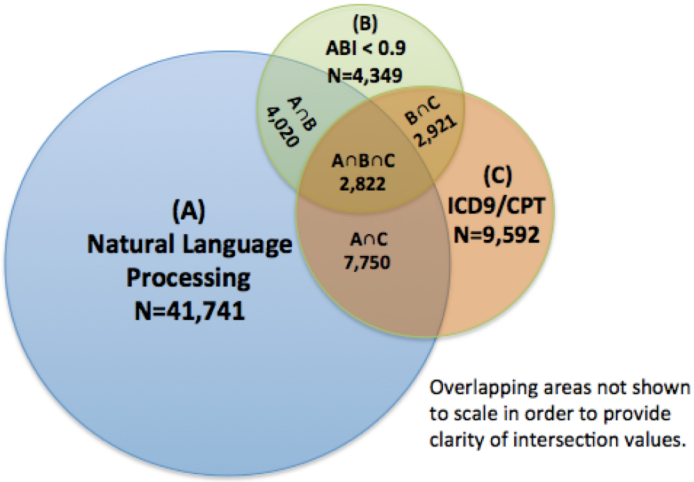
To measure the specificity of our algorithm, two of the authors (JD, NR) conducted manual review of 400 matching documents. For each document, reviewers were asked to assess the presence or absence of peripheral arterial disease. We then calculated the precision of the algorithm as follows: TP / (TP+FP).

Results:

Figure 1 summarizes the study results. As shown, the natural language processing algorithm identified significantly more PAD patients compared with the structured data algorithm (41,741 vs 9,592, $p<0.001$) and achieved a high level of specificity (98%). The individual components of our NLP algorithm varied in performance (Figure 1). The lowest performing component was diminished pulses, which had lower specificity (92%) due to non-PAD causes of decreased pulse including death, cardiac arrest, hypotension, and congenital anomalies. Otherwise, all NLP component algorithms exceeded 95% specificity. The errors that were seen were due to unusual negation patterns (e.g., "-claudication") or unexpected phrasings such as "at *rest pain* was minimal." Our ABI extraction process was 100% specific for identification of patients with ABIs < 0.9.

Overall 43,811 unique PAD patients were identified across all three detection methods. NLP identified 95.2% compared with 21.9% for ICD9/CPT codes and 9.9% for ABIs alone. Over 75% of patients with physician-documented evidence of peripheral arterial disease were not identifiable by structured data.

Figure 1. Venn diagram showing the number of patients identified by three PAD detection methods: A) natural language processing (NLP) of clinical notes, B) extraction of ABI values from vascular laboratory reports, C) queries of ICD9 and CPT codes. The performance of the NLP algorithms based on manual review is shown below.



NLP Algorithm	# Unique Patients	Specificity
Claudication	15337	96%
Rest Pain	2498	98%
Diminished pulses	5773	92%
Ischemic Limb	1339	99%
Peripheral Arterial Disease	31430	99%
Combined PAD Algorithm	**41741**	**98%**

Discussion:

The identification of patients with peripheral arterial disease in observational data is significantly enhanced by the use of natural language processing on unstructured data. We found 4 times as many patients using a conservative (i.e., high specificity) NLP algorithm than by using an established code-based strategy for PAD detection. A more expansive NLP approach, tolerating slightly reduced specificity and incorporating more radiologic and arteriogram findings, would likely yield a considerably larger cohort.

Our results suggest that PAD may be under-detected in traditional observational and epidemiological research. Population estimates of PAD prevalence based on large-scale claims databases may signficantly underrepresent the extent of this condition. While our study was performed in a single setting, our algorithms performed well across two large institutions with a large number of physicians using differing documentation styles. Nonetheless, this work should be validated in other environments to assess generalizability.

In summary, researchers using observational data to study peripheral arterial disease should incorporate unstructured data when possible to maximize the comprehensiveness and accuracy of their findings.

References:

1) 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. *J Am Med Inform Assoc*. 2010 Sep-Oct;17(5):568-74.
2) PheKB PAD Definition. https://phekb.org/phenotype/peripheral-arterial-disease
3) Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001;34:301-10.