The Journal of Arthroplasty 34 (2019) 2216–2219

Contents lists available at ScienceDirect

The Journal of Arthroplasty

journal homepage: www.arthroplastyjournal.org

Artificial Intelligence and Machine Learning

Use of Natural Language Processing Tools to Identify and Classify Periprosthetic Femur Fractures

Meagan E. Tibbo, MD [a], Cody C. Wyles, MD [a], Sunyang Fu, MHI [b], Sunghwan Sohn, PhD [b], David G. Lewallen, MD [a], Daniel J. Berry, MD [a], Hilal Maradit Kremers, MS, MD [a, b, *]

[a] Department of Orthopedic Surgery, Mayo Clinic, Rochester, MN
[b] Department of Health Sciences Research, Mayo Clinic, Rochester, MN

ARTICLE INFO

Article history:
Received 8 July 2019
Accepted 18 July 2019
Available online 24 July 2019

Keywords:
total hip arthroplasty
natural language processing
periprosthetic femur fractures
machine learning
Vancouver classification

ABSTRACT

Background: Manual chart review is labor-intensive and requires specialized knowledge possessed by highly trained medical professionals. The cost and infrastructure challenges required to implement this is prohibitive for most hospitals. Natural language processing (NLP) tools are distinctive in their ability to extract critical information from unstructured text in the electronic health records. As a simple proof-of-concept for the potential application of NLP technology in total hip arthroplasty (THA), we examined its ability to identify periprosthetic femur fractures (PPFFx) followed by more complex Vancouver classification.

Methods: PPFFx were identified among all THAs performed at a single academic institution between 1998 and 2016. A randomly selected training cohort (1538 THAs with 89 PPFFx cases) was used to develop the prototype NLP algorithm and an additional randomly selected cohort (2982 THAs with 84 PPFFx cases) was used to further validate the algorithm. Keywords to identify, and subsequently classify, Vancouver type PPFFx about THA were defined. The gold standard was confirmed by experienced orthopedic surgeons using chart and radiographic review. The algorithm was applied to consult and operative notes to evaluate language used by surgeons as a means to predict the correct pathology in the absence of a listed, precise diagnosis. Given the variability inherent to fracture descriptions by different surgeons, an iterative process was used to improve the algorithm during the training phase following error identification. Validation statistics were calculated using manual chart review as the gold standard.

Results: In distinguishing PPFFx, the NLP algorithm demonstrated 100% sensitivity and 99.8% specificity. Among 84 PPFFx test cases, the algorithm demonstrated 78.6% sensitivity and 94.8% specificity in determining the correct Vancouver classification.

Conclusion: NLP-enabled algorithms are a promising alternative to manual chart review for identifying THA outcomes. NLP algorithms applied to surgeon notes demonstrated excellent accuracy in delineating PPFFx, but accuracy was low for Vancouver classification subtype. This proof-of-concept study supports the use of NLP technology to extract THA-specific data elements from the unstructured text in electronic health records in an expeditious and cost-effective manner.

Level of Evidence: Level III.

Total hip arthroplasty (THA) is a common surgical procedure. Over 500,000 THA procedures are performed each year in the United States and around 2.5 million Americans are currently living with THA [1]. Further rises in THA procedure volumes are expected in coming decades [2].

Periprosthetic femur fractures (PPFFx) are a rare but devastating complication of THA. They occur in about 1.5%-2% of primary and 12%-15% of revision THA procedures, and according to some estimates, the cumulative incidence of PPFFx reaches 5% by 15-20 years following primary THA [3–8]. As a result of the rise in THA procedures over time, and the expansion of indications to sicker and elderly patients, the burden of PPFFx is expected to increase over time [9–11]. Information on clinical risk factors and outcomes of

One or more of the authors of this paper have disclosed potential or pertinent conflicts of interest, which may include receipt of payment, either direct or indirect, institutional support, or association with an entity in the biomedical field which may be perceived to have potential conflict of interest with this work. For full disclosure statements refer to https://doi.org/10.1016/j.arth.2019.07.025.

* Reprint requests: Hilal Maradit Kremers, MS, MD, Mayo Clinic, 200 First Street SW, Rochester, MN 559505.

https://doi.org/10.1016/j.arth.2019.07.025
0883-5403/

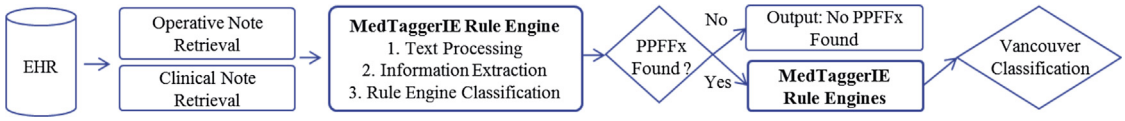

Fig. 1. PPFFx detection and Vancouver type classification workflow.

PPFFx are limited due to small single institution cohorts and difficulties associated with identifying and classifying PPFFx in large multicenter datasets. In most cases, manual chart review is labor-intensive and requires specialized knowledge possessed by highly trained medical professionals. The cost and infrastructure challenges required to implement this is currently prohibitive for most hospitals. Natural language processing (NLP) is a field in machine learning with the ability of the computer to understand, analyze, and retrieve data. NLP tools are distinctive in their ability to extract critical information from the raw, unstructured text in the electronic health records (EHR). As a simple proof-of-concept, for the potential application of this technique, we examined the potential use of NLP methods to identify and classify PPFFx in EHR.

Methods

Study Setting and Population

Study population comprised all primary THA procedures performed at a single academic institution (Mayo Clinic, Rochester, MN) between 1998 and 2016. During the study time period, the institution had an annual volume of around 800-1300 THA procedures performed by 35 orthopedic surgeons and all clinical orthopedic operative notes were contained in the EHR. An institutional total joint arthroplasty registry was maintained as part of routine care of all THA patients. Registry data collection was performed through manual chart review of EHR by trained registry personnel using standardized definitions for THA-specific data elements and PPFFx. Therefore, gold standard data for validation was readily available for all THA surgeries and PPFFx.

Orthopedic operative notes and clinical notes (surgical consult notes) were used for PPFFx detection and Vancouver classification. The training dataset for PPFFx detection comprised 1538 THA procedures with 89 PPFFx positive cases. The mean age of the 89 PPFFx positive cases was 78 years and women comprised 52%. The test dataset for PPFFx detection comprised 2982 THA procedures with 84 PPFFx positive cases. The mean age of the 84 PPFFx cases was 72 years and women comprised 59%.

NLP Algorithm Development

The workflow for algorithm development consisted of EHR data retrieval (operative and clinical notes), PPFFx detection, and Vancouver classification. The process started with the determination of surgical encounter type (PPFFx or non-PPFFx). The output contained positive PPFFx detection which was then sent further for Vancouver classification. Since the prevalence of PPFFx was low (0.1%-3.2%) [12], detecting PPFFx was prioritized to optimize the screening efficiency by removing a large number of non-PPFFx-related THA procedures before applying the Vancouver classifier. The operative notes were prioritized over the clinical notes due to its high reliability and validity. The detailed process workflow is shown in Figures 1 and 2.

The NLP system for PPFFx classification was developed on a training dataset and validated on a test dataset. Our NLP system was based on expert rules—target "textual markers" (ie, keywords related to PPFFx) were specified in the clinical narratives defined by orthopedic surgeons. The NLP system had 3 main components: text processing, concept extraction, and classification (Fig. 2).

The key components of the text processing pipeline were sentence segmentation, assertion identification, and temporal extraction. Assertion of each concept includes certainty (ie, positive, negative, and possible) along with experiencer (ie, patient, associated with someone else), while temporality identifies historical or present. For example, from the sentence "It was left a few mm proud to avoid causing a probable periprosthetic femur fracture," "periprosthetic femur fracture" would be extracted as a fracture concept, along with corresponding assertion status "probable," temporality "present," and experiencer "associated with the patient." Concept extraction is a knowledge-driven annotation and indexing process to identify phrases referring to concepts of interests in an unstructured text [13] ("periprosthetic femur fracture" extraction from the previous example). After concepts are extracted from operative reports and clinical notes, they are normalized to the specific categories. For instance, the concept "fracture" will be mapped to the category "finding – fracture." Patients with non-negated and present fracture findings were classified as fracture positive. The full list of concepts, keywords, modifiers, and diseases categories are listed in Table 1.

Our NLP system was implemented using the open-source NLP pipeline MedTaggerIE [14], a resource-driven open-source Unstructured Information Management Architecture [15] based IE framework. The MedTaggerIE separates domain-specific NLP knowledge engineering from the generic NLP process, which enables words and phrases containing clinical information to be directly coded by subject matter experts. The tool has been utilized in the eMERGE consortium to develop NLP-based phenotyping algorithms [16].

Statistical Analysis

The performance of the NLP algorithm was assessed using the gold standard registry data. We calculated the validation statistics of sensitivity and specificity [13].

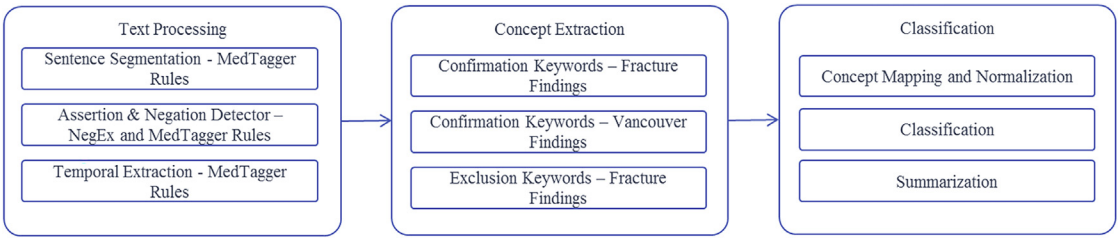

Fig. 2. MedTaggerIE rule engine.

Table 1
PPFX Keywords for Concept Extraction.

Confirmation keywords – finding – fractures
Periprosthetic hip fracture; Periprosthetic femur fracture; Fracture hip replacement; Fracture hip arthroplasty
Confirmation keywords – finding – Vancouver – A_G
Vancouver A_G; Vancouver A_G type; Greater trochanter fracture; Greater trochanteric; Greater trochanter; Greater troch; Vancouver A
Confirmation keywords – finding – Vancouver – A_L
Vancouver Al; Vancouver Al type; Lesser trochanteric
Confirmation keywords – finding – Vancouver – B1
Vancouver B1; Vancouver B1 type; Vancouver B type 1; Vancouver B-1; Vancouver 1B
Confirmation keywords – finding – Vancouver – B2
Vancouver B2; Vancouver B2 type; Vancouver B type 2; Vancouver B-2; Vancouver 2B
Confirmation keywords – finding – Vancouver – B3
Vancouver B3; Vancouver 3B; Vancouver B3 type; Vancouver B type 3; Vancouver B-3
Confirmation keywords – finding – Vancouver – C
Vancouver C; Vancouver C type; Vancouver type C; Vancouver C; Interprosthetic; Distal fixed; Spinal
Confirmation keywords – finding – Vancouver – Bone
Bone poor; Bone Osteopenic; Bone osteoporosis; Bone osteoporotic; Bone osteopenia; Bone loss
Exclusion – fracture
Conversion hip fracture; Conversion femur fracture; Conversion total femoral replacement

PPFFx, periprosthetic femur fracture.

Results

PPFFx Detection

In the test dataset of 2982 THA procedures with 84 cases of PPFFx, the PPFFx detection NLP algorithm demonstrated sensitivity of 100% and specificity of 99.8% (Table 2). All of the 6 cases that were falsely detected as PPFFx were complex cases that would be hard to classify even by manual chart review. For example, one case involved periprosthetic fracture about a total knee prosthesis in a patient who also has a THA on the same side. Such cases are universally difficult as they can be classified as a Vancouver C or an "interprosthetic fracture" or a periprosthetic fracture around TKA.

Vancouver Classification

Based on chart review, the majority of the 84 PPFFx cases were Vancouver B-class (68 cases, 81%). The overall Vancouver classification algorithm achieved a sensitivity of 78.6% and specificity of 94.8% in classifying PPFFx into the 6 Vancouver types (Table 3). The performance of only Vancouver B class was higher with a sensitivity of 88.2% and a specificity of 94.0%. Except for operative notes, poor documentation in radiology reports and surgical consult notes is the primary reason for the low sensitivity. The primary reason for the false negative cases were no confirmation keyword found in both operative and orthopedic consult notes. The majority of false negatives cases only have orthopedic consult notes. The information that will be used for the prediction is embedded in the section of "Image Interpretation," which is much less descriptive and informative than the information from the operative reports. The primary reason for the false positive cases was confirmation keywords from 2 categories. For example, an orthopedic surgeon said "As such, I would classify this most likely as a Vancouver B2, although this certainly could be a Vancouver B1 type fracture." Cases like this were much more challenging to predict.

Discussion

NLP algorithms are a promising alternative to the current gold standard of manual chart review for evaluating outcomes of THA. Despite their immaturity with respect to orthopedic applications, NLP algorithms applied to surgeon notes and orthopedic consult notes demonstrated excellent performance in delineating a simple binary outcome, in this case the presence or absence of PPFFx. However, accuracy of the algorithm was lower when trying to classify Vancouver subtypes given the wide variability in documentation in EHR, surgeon dictation styles, and precision of language. Nevertheless, this study provides a proof-of-concept for use of this technology in clinical research and registry development endeavors that reliably obtains data of interest in an expeditious and cost-effective manner. The orthopedics community currently rely on laborious manual review of unstructured orthopedic consult notes and operative reports to extract PPFFx information. NLP algorithms can therefore bridge the gap between free-text and structured data fields, such as diagnostic codes.

Limitations

Our findings must be interpreted in light of potential limitations. Although we performed the study at a large-volume institution with readily available gold standard data from the institutional total joint arthroplasty registry, PPFFx was a relatively rare event and the frequency of some Vancouver fracture classes was very low. Therefore, our Vancouver classification algorithm had limited keywords for A_G-A_L and C class PPFFx. Larger datasets with more advanced keywords and potential addition of machine learning are needed to further improve the accuracy of Vancouver classification. Second, as addressed in previous retrospective studies, documentation of Vancouver class in EHR is particularly poor, especially for conservatively managed cases, and this in turn limits capability of an NLP algorithm applied on EHR documentation. Therefore, most studies rely on review of radiographs. In our study, the majority of false negative cases only had orthopedic consultation notes, and no operative reports. Furthermore, it was difficult to classify PPFFx cases where there was uncertainty about the type of fracture in the notes (eg, Vancouver B1 vs A_L in some cases). Developing NLP algorithms in settings without pre-established gold standard data would be difficult and time-consuming as it would require manual

Table 2
Performance of the THA PPFFx Detection Algorithm.

	Gold PPFFx Cases	Gold Controls	Total
Predicted PPFFx cases	84	6	90
Predicted controls	0	2892	2892
Total	84	2898	2982

Performance of PPFFx detection algorithm: sensitivity = 100% and specificity = 99.8%.
THA, total hip arthroplasty; PPFFx, periprosthetic femur fracture.

Table 3
Performance of the THA PPFFx Vancouver Classification Algorithm.

	Gold Standard Classification					
	A_G	A_L	B1	B2	B3	C
Predicted class						
A_G	3	0	0	0	0	0
A_L	0	0	0	0	0	0
B1	0	2	24	2	1	4
B2	2	0	5	34	0	2
B3	0	0	0	0	2	0
C	0	0	0	0	0	3

Performance of Vancouver classification algorithm: sensitivity: 78.6% and specificity: 94.8%.
THA, total hip arthroplasty; PPFFx, periprosthetic femur fracture.

chart review for data validation. What seems most logical is to have algorithms created at institutions with registries and then disseminate them for refinement to institutions interested in applying the technology to their own operative reports. Additional studies are needed to assess the transportability of NLP algorithms to other institutions.

Despite the sometimes sparse documentation and therefore decreased accuracy of Vancouver classification, we achieved almost perfect accuracy in identifying 84 PPFFx in a cohort of almost 3000 THA surgeries. We anticipate that accuracy will be lower when the algorithms are applied to outside facilities, primarily due to differences in documentation. Clinical documentation variations are known section and affect an NLP system portability [17]. Otherwise, this study leveraged the operative and orthopedic consult notes of 29 different surgeons, and therefore comprises a wide range of terms. Algorithm development is an ongoing iterative process and we encourage other institutions to reach out to our group for testing in their EHR. The algorithms will certainly improve as efforts are undertaken to apply them to operative reports from several institutions.

In conclusion, NLP-enabled algorithms are a promising alternative to the current gold standard of manual chart review for evaluating outcomes in large THA datasets. Despite their immaturity with respect to orthopedic applications, NLP algorithms applied to surgeon notes demonstrated excellent accuracy in delineating PPFFx. However, accuracy of the algorithm was attenuated when attempting to predict the Vancouver classification subtype given the wide variability in surgeon dictation styles and precision of language. Nevertheless, this study provides a proof-of-concept for use of this technology with clinical research and registry development endeavors that reliably obtains data of interest in an expeditious and cost-effective manner.

Acknowledgment

We would like to acknowledge the National Institutes of Health (NIH) under Award Number R01 AG060920. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

[1] Kremers HM, Larson DR, Crowson CS, Kremers WK, Washington RE, Steiner CA, et al. Prevalence of total hip and knee replacement in the United States. J Bone Joint Surg Am 2015;97A:1386–97.
[2] Kurtz SM, Ong KL, Lau E, Bozic KJ. Impact of the economic downturn on total joint replacement demand in the United States: updated projections to 2021. J Bone Joint Surg Am 2014;96:624–30.
[3] Lewallen DG, Berry DJ. Periprosthetic fracture of the femur after total hip arthroplasty—treatment and results to date. J Bone Joint Surg Am 1997;79A:1881–90.
[4] Berry DJ. Epidemiology—hip and knee. Orthop Clin North Am 1999;30:183–90.
[5] Meek RMD, Norwood T, Smith R, Brenkel IJ, Howie CR. The risk of periprosthetic fracture after primary and revision total hip and knee replacement. J Bone Joint Surg Br 2011;93-B:96–101.
[6] Lindahl H, Malchau H, Herberts P, Garellick G. Periprosthetic femoral fractures—classification and demographics of 1049 periprosthetic femoral fractures from the Swedish National Hip Arthroplasty Register. J Arthroplasty 2005;20:857–65.
[7] Abdel MP, Watts CD, Houdek MT, Lewallen DG, Berry DJ. Epidemiology of periprosthetic fracture of the femur in 32 644 primary total hip arthroplasties: a 40-year experience. Bone Joint J 2016;98-B:461–7.
[8] Abdel MP, Houdek MT, Watts CD, Lewallen DG, Berry DJ. Epidemiology of periprosthetic femoral fractures in 5417 revision total hip arthroplasties: a 40-year experience. Bone Joint J 2016;98-B:468–74.
[9] Ricci WM. Periprosthetic femur fractures. J Orthop Trauma 2015;29:130–7.
[10] Lindahl H. Epidemiology of periprosthetic femur fracture around a total hip arthroplasty. Injury 2007;38:651–4.
[11] Della Rocca GJ, Leung KS, Pape HC. Periprosthetic fractures: epidemiology and future projections. J Orthop Trauma 2011;25:S66–70.
[12] Parvizi J, Rapuri VR, Purtill JJ, Sharkey PF, Rothman RH, Hozack WJ. Treatment protocol for proximal femoral periprosthetic fractures. J Bone Joint Surg Am 2004;86A:8–16.
[13] Manning CD, Schütze H. Foundations of statistical natural language processing. 2nd ed. Cambridge, MA: MIT Press; 1999.
[14] Liu H, Bielinski SJ, Sohn S, Murphy S, Wagholikar KB, Jonnalagadda SR, et al. An information extraction framework for cohort identification using electronic health records. AMIA Jt Summits Transl Sci Proc 2013;2013:149–53.
[15] Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. Nat Lang Eng 2004;10:327–48.
[16] Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. Bioinformatics 2010;26:1205–10.
[17] Sohn S, Wang Y, Wi CI, Krusemark EA, Ryu E, Ali MH, et al. Clinical documentation variations and NLP system portability: a case study in asthma birth cohorts across institutions. J Am Med Inform Assoc 2018;25:353–9. https://doi.org/10.1093/jamia/ocx138.