Emergency Radiology (2019) 26:301–306
https://doi.org/10.1007/s10140-019-01673-4

ORIGINAL ARTICLE

A natural language processing algorithm to extract characteristics of subdural hematoma from head CT reports

Peter Pruitt[1,2] · Andrew Naidech[2,3] · Jonathan Van Ornam[4,5,6] · Pierre Borczuk[5,6] · William Thompson[7]

Received: 20 November 2018 / Accepted: 11 January 2019 / Published online: 28 January 2019

Abstract

Purpose Subdural hematoma (SDH) is the most common form of traumatic intracranial hemorrhage, and radiographic characteristics of SDH are predictive of complications and patient outcomes. We created a natural language processing (NLP) algorithm to extract structured data from cranial computed tomography (CT) scan reports for patients with SDH.

Methods CT scan reports from patients with SDH were collected from a single center. All reports were based on cranial CT scan interpretations by board-certified attending radiologists. Reports were then coded by a pair of physicians for four variables: number of SDH, size of midline shift, thickness of largest SDH, and side of largest SDH. Inter-rater reliability was assessed. The annotated reports were divided into training (80%) and test (20%) datasets. Relevant information was extracted from text using a pattern-matching approach, due to the lack of a mention-level gold-standard corpus. Then, the NLP pipeline components were integrated using the Apache Unstructured Information Management Architecture. Output performance was measured as algorithm accuracy compared to the data coded by the two ED physicians.

Results A total of 643 scans were extracted. The NLP algorithm accuracy was high: 0.84 for side of largest SDH, 0.88 for thickness of largest SDH, and 0.92 for size of midline shift.

Conclusion A NLP algorithm can structure key data from non-contrast head CT reports with high accuracy. The NLP is a potential tool to detect important radiographic findings from electronic health records, and, potentially, add decision support capabilities.

Keywords Subdural hematoma · Natural language processing · Cranial CT reports · Intracranial hemorrhage

✉ Peter Pruitt
peter.pruitt@northwestern.edu

[1] Department of Emergency Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

[2] Center for Healthcare Studies, Northwestern University Feinberg School of Medicine, 633 N St. Clair Street, Chicago, IL 60622, USA

[3] Department of Neurology, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

[4] Harvard Affiliated Emergency Medicine Residency, Boston, MA, USA

[5] Department of Emergency Medicine, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114, USA

[6] Department of Emergency Medicine, Harvard Medical School, Boston, MA, USA

[7] Center for Health Information Partnerships, Northwestern University Feinberg School of Medicine, 625 N Michigan Ave., Chicago, IL 60611, USA

Introduction

Subdural hematoma (SDH) is the most common form of traumatic intracranial hemorrhage, resulting in over 60,000 hospitalizations annually [1]. The computed tomography (CT) scan is a diagnostic for SDH, and the CT scan report typically contains key elements used in deciding the disposition and treatment plan for patients with SDH. The size and location of hematoma have also been shown to be predictive of outcomes [2]. Midline shift (brain herniation) can lead to coma and death. The ability to extract this important data from CT scans, or the accompanying report, is crucial to patient care and research for patients with SDH.

Radiology reports are typically entered into the electronic health record (EHR) as free text. While these reports are generally in a consistent format and use relatively similar terminology, the information is not captured as discrete data

elements. Because of this, these reports cannot be easily used in situations where structured data is needed, such as for clinical decision support tools or in observational studies. This presents a particular problem for SDH research, which has no verified severity of injury scale that includes an interpretation of neuroimaging data, in contrast to aneurysmal subarachnoid hemorrhage (e.g., Fisher grade), intracerebral hemorrhage (e.g., hematoma volume), or ischemic stroke (e.g., large vessel occlusion). Natural language processing (NLP) has been used in multiple prior instances to classify radiology reports [3–6]. The NLP for data element extraction has been successfully utilized in a few specific clinical research scenarios (e.g., colonoscopy reports) [7, 8]; however, we are unaware of prior use in neuroimaging. A functional, accurate NLP algorithm could allow researchers to extract data from a large number of head CT reports more quickly and precisely than manual extraction, leading to advances in prognosis, predictors of complications, and evaluation of potential treatments that might improve patient outcomes.

The primary objective of this investigation was to develop a NLP algorithm to structure SDH characteristics from cranial CT scan reports with known SDH and then to test this mechanism for accuracy when compared to the performance of a pair of trained physician abstractors reviewing the same CT scan reports.

Methods

Population, data collection, and coding

Radiographic data was collected from patients presenting to a single, academic level 1 trauma center. The sample was consecutive CT scan reports with isolated SDH retrieved from the electronic medical record. Reports were identified by searching patient records for discharge diagnoses consistent with intracranial hemorrhage and then narrowed to isolated SDH in a process that has been previously described [2]. Isolated SDH was defined as presence of SDH with no other type of intracranial hemorrhage (such as subarachnoid hemorrhage, cerebral contusion, epidural hematoma). However, patients could have more than one SDH present. Other types of intracranial hemorrhage were excluded in order to simplify algorithm development, as these other hemorrhages have different key characteristics. SDH count was determined by tallying all hemorrhages referenced in the radiology report. Hemorrhages that spanned more than one region but were confluent were counted as a single hematoma. All scans were interpreted and/or approved by attending radiologists.

After initial review, scans were coded by abstractors for four variables: total number of SDH, presence and degree of midline shift (in millimeters), thickness of largest SDH (in millimeters), and side of largest SDH. Variables were chosen as they make up the key components of several SDH risk stratification decision tools [2, 9]. Each scan was assessed twice and results were adjudicated by a senior investigator, an attending emergency medicine physician with substantial experience interpreting and adjudicating cranial CT scan reports in a research setting; the final result of this interpretation was considered the gold-standard value.

Model creation and system development

An algorithm was created to extract from the written CT scan reports using a pattern matching approach. This method was chosen due to the lack of a mention-level gold-standard corpus. The NLP algorithm consists of a pipeline of components that first detect section, sentence, and word token boundaries, breaking the text down into identifiable paragraphs, sentences, and phrases. Subsequent components execute pattern matching searches for mentions of SDH (e.g., “subdural hematoma” or “extra-axial collection”), mentions of location (e.g., “right frontal,” “left occipital lobe”), and measure phrases (e.g., “6 mm”). An example of the location of data extracted within a typical CT scan report is provided in Box 1. Measure phrases and location information are associated with SDH mentions using a heuristic of co-occurrence within the same sentence. These components of the algorithm are then integrated using the Apache Unstructured Information Management Archive (UMIA) environment (Apache Software Foundation, Wakefield, MA). The code of the algorithm was made publicly available after completion and is available for download and free, non-commercial use via the general public license at https://github.com/NUNLP/NeuroNLP.

Box 1. Example CT scan report

HISTORY:
TRAUMA, ASSESS FOR HEMORRHAGE

REPORT

HISTORY: As above.

COMPARISON: No relevant prior examinations available.

TECHNIQUE: Head CT without contrast.

FINDINGS:

There is a predominantly hyperdense extra axial collection layering over the left cerebral convexity and left tentorium, consistent with a subdural hematoma. The component overlying the left cerebral convexity measures **11-mm** and the tentorial component measures **15-mm**.

There is associated mass effect with effacement of the adjacent left cerebral sulci and left lateral ventricle and *6-mm of rightward midline shift.* There is mild subfalcine herniation. There is no uncal herniation. There is also a small amount of subdural hemorrhage along the falx and right tentorium.

There is hypodensity in the periventricular and subcortical white matter, a nonspecific finding that is often seen in patients with microangiopathic changes. The brain parenchyma is otherwise unremarkable. There is no evidence of intracranial mass lesion, acute infarction.

There is scattered mucosal thickening of the paranasal sinuses. The bilateral mastoid air cells are well aerated and clear. A prosthesis is noted in the right globe.

There is a large posterior subgaleal hematoma. No fracture is identified.

IMPRESSION:
Predominantly hyperdense subdural hematoma layering over the left cerebral convexity and left tentorium**, measuring up to 15 mm in greatest diameter**. There is also a small amount of subdural hemorrhage along the right tentorium and the falx. There is associated mass effect with *6-mm of rightward midline shift* and mild subfalcine herniation.

Nonspecific white matter hypoattenuation, likely representing microangiopathic changes.

Bold indicates data used for thickness. *Italics* indicate data used for midline shift. Red indicates data used for SDH count. Underline indicates data used for side of SDH

Outcomes and data analysis

Performance of the algorithm was measured using accuracy compared to data abstracted by two emergency physicians. Accuracy, defined as percentage of the NLP extracted values that were the same as the physician consensus gold standard, was calculated. Statistical analysis was performed using R v3.4 (R Foundation for Statistical Computing, Vienna, Austria) with the IRR and psych packages.

Results

A total of 612 CT scan results, each the first CT scan recorded in the EHR for unique patients, were extracted and used to create the training and test datasets. All scans in the corpus had all four of the key features described.

The created NLP algorithm was found to have 84–90% agreement with human abstractors for size of SDH, degree of midline shift, and side of largest SDH. The algorithm had less optimal performance when attempting to determine SDH count. Algorithm accuracy and Cohen's kappa comparing the NLP algorithm to agreement of two human abstractors (gold standard) are displayed in Table 1.

Discussion

In this investigation, an NLP algorithm was derived and validated to identify and extract key data elements (thickness of SDH, amount of midline shift, side of largest SDH) from radiology reports with performance comparable to that of physicians. This is the first described NLP algorithm to extract information from head CT reports, although cranial CT scans are one of the most commonly used radiographic tests. Additionally, this is one of the first applications of the NLP to extract data into a structured format from radiology reports, rather than to classify if a report contains a positive result (e.g., thickness of subdural hematoma, which is more informative than only mention of its presence).

The NLP algorithm had differing performance for each variable. Midline shift had the highest performance, likely because discussion of midline shift, including the amount of shift, is almost always isolated to a single sentence. Side of hematoma and thickness are also frequently described in close proximity to the primary mention of the hematoma, contributing to their similar accuracy. On the other hand, count is very rarely mentioned explicitly, which makes the process of counting the exact number of hematomas present more indirect, potentially explaining the algorithm's relatively less accurate performance when extracting this variable.

The use of the NLP techniques is consistently increasing in medicine, both related to clinical and research applications. While the use of the NLP on radiology reports has been extensively previously studied [3], and there have been prior investigations using the NLP to classify cranial CT scan report outcomes [10], this is the first algorithm that has been created to extract data from cranial CT scan reports. Outside of radiology, a previous study extracted data from colonoscopy reports and were able to obtain excellent accuracy [7], while another used the NLP to extract structured data from mammography reports, obtaining an F score of 86% [11]. The accuracy statistics from the mammography study are comparable in accuracy to what is reported here; this is one of the only other data extraction reports published so far which included accuracy statistics. There have been several other applications of the NLP to identify other types of radiography reports with positive findings, including pulmonary nodules, pulmonary emboli, and abdominal aortic aneurysms. The use of the NLP to extract clinically relevant characteristics from head CT reports is an innovative and logical next step that may have important implications for the diagnosis and management of neurological conditions.

Using the NLP to abstract data has several potential important research and clinical uses. It may be used to improve dataset creation for larger, more robust and more generalizable observational studies by allowing for the use of much larger datasets due to the reduced need for

Table 1 Performance of the NLP algorithm structuring data from CT scan reports

Variable	NLP to human gold-standard accuracy (%)	NLP to human gold-standard kappa (95% confidence interval)
Amount of midline shift	90.5	0.82 (0.78–0.86)
Maximum SDH thickness	87.4	0.87 (0.84–0.89)
Side of SDH	84.2	0.75 (0.71–0.79)
SDH count	71.9	0.29 (0.22–0.36)

human abstractors. In fact, while there is some investment needed to program the initial algorithm, the marginal cost of extracting data from additional records is practically nil. In addition to the benefits for research, accurate NLP technologies may allow for improved real-time decision support capabilities by allowing these tools to extract richer, more robust data from the medical record by providing access to free text fields instead of only structured entries [12].

This investigation does have several important limitations. Data was gathered from a single center with a single style of formatting. Additionally, reports were created and finalized by a single attending group of emergency radiologists and the performance of the algorithm was not compared based on the authoring radiologist. Hospital, physician group, and regional variations may limit the generalizability of this algorithm when applied to other reports, although the large size of the cohort makes generalizability more likely. Despite the excellent agreement between data extracted by the NLP algorithm and the gold standard, other NLP techniques might further improve accuracy. For example, more advanced information extraction algorithms could be applied given a mention-level gold-standard corpus. There are inherent limitations in human abstraction accuracy so the gold-standard, adjudicated records may not be 100% correct. However, multiple reviewers abstracted each record with a senior researcher adjudicating, so this error is likely minimized in this study. Finally, we referenced the NLP toward the CT scan report, which is taken as a faithful interpretation of the images; an error in the interpretation could lead to an unfaithful abstraction of what the CT actually shows, even if it accurately represents the report. Another group has recently reported acceptable accuracy of automated recognition of abnormalities on chest radiographs directly from a large repository of images, which is another potential avenue of research for SDH and brain injury [13]. Automated recognition typically requires thousands of images, more than is available in any known dataset of patients with SDH.

Future investigations should attempt to improve the accuracy of the NLP algorithms presented here. Substituting a program that uses named-entity recognition using the UMLS dictionary, such as cTAKES, will likely substantially improve system accuracy [14]. Additionally, adding a component to classify hematoma type in addition to extracting data would eliminate the need to manually determine the presence of an isolated subdural hematoma, making it a helpful step toward creating a unified NLP algorithm to interpret cranial CT scan reports.

Conclusion

An NLP algorithm can successfully abstract the side of SDH, thickness of SDH, and the degree of midline shift after SDH from head CT reports with SDH with excellent accuracy in a test cohort. The algorithm, available freely, may accelerate research and patient care for SDH, the most common form of traumatic intracranial hemorrhage.

Funding sources Dr. Pruitt was supported by a National Research Service Award postdoctoral fellow supported by the Agency for Healthcare Research and Quality (AHRQ) T-32 HS 000078 (PI: Jane L. Holl, MD, MPH). AHRQ was not involved in the design or execution of this research. Dr. Pruitt is now supported by a career development award from the Society for Academic Emergency Medicine Foundation.

Author contributions PP, AN, and WKT conceived of the study and designed the analysis. PB, JO, and PP participated in the abstraction and coding of data. WKT programmed the algorithm. PP performed the data analysis. PP drafted the manuscript and all authors contributed substantially to its revision. PP takes responsibility for the paper as a whole.

Compliance with ethical standards

Conflicts of interest The authors declare that they have no conflict of interest.

References

1. Marin JR, Weaver MD, Yealy DM, Mannix RC (2014) Trends in visits for traumatic brain injury to emergency departments in the United States. JAMA 311:1917–1919. https://doi.org/10.1001/jama.2014.3979
2. Pruitt P, Van OJ, Borczuk P (2017) A decision instrument to identify isolated traumatic subdural hematomas at low risk of neurologic deterioration, surgical intervention, or radiographic worsening. Acad Emerg Med 24:1377–1386. https://doi.org/10.1111/acem.13306
3. Pons E, Braun LMM, Hunink MGM, Kors JA (2016) Natural language processing in radiology: a systematic review. Radiology 279:329–343. https://doi.org/10.1148/radiol.16142770
4. Jain NL, Friedman C (1997) Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. Proc AMIA Annu Fall Symp: 829–833
5. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, Forshee R, Walderhaug M, Botsis T (2017) Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. J Biomed Inform 73:14–29. https://doi.org/10.1016/j.jbi.2017.07.012
6. Cai T, Giannopoulos AA, Yu S, Kelil T, Ripley B, Kumamaru KK, Rybicki FJ, Mitsouras D (2016) Natural language processing technologies in radiology research and clinical applications. RadioGraphics 36:176–191. https://doi.org/10.1148/rg.2016150080
7. Gawron AJ, Thompson WK, Keswani RN, Rasmussen LV, Kho AN (2014) Anatomic and advanced adenoma detection rates as quality metrics determined via natural language processing. Am J Gastroenterol 109:1844–1849. https://doi.org/10.1038/ajg.2014.147
8. Kuo T-T, Rao P, Maehara C, et al (2016) Ensembles of NLP tools for data element extraction from clinical notes. AMIA. Annu Symp proceedings AMIA Symp 2016:1880–1889
9. Orlando A, Levy AS, Rubin BA, Tanner A, Carrick MM, Lieser M, Hamilton D, Mains CW, Bar-Or D (2018) Isolated subdural hematomas in mild traumatic brain injury. Part 2: a preliminary clinical

decision support tool for neurosurgical intervention. J Neurosurg:1–8. https://doi.org/10.3171/2018.1.JNS171906
10. Yadav K, Sarioglu E, Choi HA et al (2016) Automated outcome classification of computed tomography imaging reports for pediatric traumatic brain injury. Acad Emerg Med 23:171–178. https://doi.org/10.1111/acem.12859
11. Esuli A, Marcheggiani D, Sebastiani F (2013) An enhanced CRFs-based system for information extraction from radiology reports. J Biomed Inform 46:425–435. https://doi.org/10.1016/j.jbi.2013.01.006
12. Demner-Fushman D, Chapman WW, McDonald CJ (2009) What can natural language processing do for clinical decision support? J Biomed Inform 42:760–772. https://doi.org/10.1016/j.jbi.2009.08.007
13. Lakhani P, Sundaram B (2017) Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology 284:574–582. https://doi.org/10.1148/radiol.2017162326
14. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG (2010) Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 17:507–513. https://doi.org/10.1136/jamia.2009.001560