Identifying Cases of Shoulder Injury Related to Vaccine

Administration (SIRVA) Using Natural Language Processing

Chengyi Zheng, PhD*¹, Jonathan Duffy, MD², In-Lu Amy Liu, MS¹, Lina S. Sy, MPH¹, Ronald A. Navarro, MD³, Sunhea S. Kim, MPH¹, Denison Ryan, MPH¹, Wansu Chen, PhD¹, Lei Qian, PhD¹, Cheryl Mercado, MPH¹, Steven J. Jacobsen, MD, PhD¹

¹ Department of Research & Evaluation, Kaiser Permanente Southern California, Pasadena, California, USA

² Immunization Safety Office, Centers for Disease Control and Prevention, Atlanta, GA, USA

³ Kaiser Permanente South Bay Medical Center, Harbor City, California

Corresponding Author:

Chengyi Zheng, PhD

Department of Research and Evaluation, Kaiser Permanente Southern California

100 S Los Robles Ave, 2nd Floor,

Pasadena, CA 91101

Email: Chengyi.X.Zheng@KP.ORG

Background:

Shoulder injury related to vaccine administration (SIRVA) accounts for more than half of all claims received by the National Vaccine Injury Compensation Program. However, there is a lack of population-based studies due to the challenge of identifying SIRVA cases in large health care databases.

Objective:

To develop a natural language processing (NLP) method to identify SIRVA cases from clinical notes.

Methods:

We conducted the study among members of a large integrated health care organization who were vaccinated between 04/1/2016 and 12/31/2017 and had subsequent diagnosis codes indicative of shoulder injury. Based on a training dataset with a chart review reference standard of 164 individuals, we developed an NLP algorithm to extract shoulder disorder information, including prior vaccination, anatomic location, temporality and causality. The algorithm identified three groups of positive SIRVA cases (definite, probable and possible) based on the strength of evidence. We compared NLP results to a chart review reference standard of 100 vaccinated individuals. We then applied the final automated NLP algorithm to a broader cohort of vaccinated individuals with a shoulder injury diagnosis code and performed manual chart confirmation on a random sample of NLP-identified definite cases and all NLP-identified probable and possible cases.

Results:

In the validation sample, the NLP algorithm had 100% accuracy for identifying 4 SIRVA cases and 96 individuals without SIRVA. In the broader cohort of 53,585 individuals, the NLP

algorithm identified 291 definite, 124 probable, and 52 possible SIRVA cases. The chart-confirmation rates for these groups were 95.3%, 67.7% and 18.9%, respectively.

Conclusions:

The algorithm performed with high sensitivity and reasonable specificity in identifying positive SIRVA cases. The NLP algorithm can potentially be used in future population-based studies to identify this rare adverse event, avoiding labor-intensive chart review validation.

KEYWORDS

shoulder injury related to vaccine administration (SIRVA); natural language processing (NLP); causal relation; temporal relation; pharmacovigilance; electronic health records (EHR); vaccine safety; artificial intelligence; big data; population health; real-world data

Abbreviations

CI: confidence interval

EHR: electronic health records

ICD-10-CM: International Classification of Diseases 10th Revision Clinical Modification

KPSC: Kaiser Permanente Southern California

NLP: natural language processing

NPV: negative predictive value

PPV: positive predictive value

SIRVA: shoulder injury related to vaccine administration

VICP: National Vaccine Injury Compensation Program

Introduction

In 2017, shoulder injury related to vaccine administration (SIRVA) was officially added to the Vaccine Injury Table by the National Vaccine Injury Compensation Program (VICP).[1-3] The VICP defined SIRVA as shoulder pain and limited range of motion occurring after the administration of a vaccine intended for intramuscular administration in the upper arm. SIRVA is caused by an injury to the musculoskeletal structures of the shoulder (e.g., tendons, ligaments, bursae). In 2019, the number of claims related to SIRVA rose to 55% of all claims received by VICP, which resulted in an annual payout of more than \$200 million.[4] Meanwhile, there has been increasing debate on whether vaccination or vaccine can cause subsequent shoulder problems.[5-7]

The debate is fueled by the lack of high-quality evidence from population-based studies. The majority of SIRVA publications are limited to case reports.[1, 8-11] Because SIRVA is a rare event but shoulder disorder is one of the most common musculoskeletal conditions, identifying SIRVA cases requires manual chart review of a large number of medical records.[11, 12]

Compared with manual chart review of medical records, natural language processing (NLP) is more efficient and produces more consistent results.[13, 14] For clinical research, NLP facilitates the identification and extraction of information that is unavailable or incomplete in structured data.[15-17] In vaccine safety studies, we have used NLP to identify two vaccine-related adverse events, anaphylaxis and local reaction.[18, 19]

Objective

Our objective was to develop an efficient SIRVA case-finding strategy using an NLP algorithm.

Methods

Setting

This study was conducted at Kaiser Permanente Southern California (KPSC), an integrated health care system that provides prepaid comprehensive health care to over 4.7 million racially, ethnically, and socioeconomically diverse members.[20] KPSC's electronic health record (EHR) system stores medical information about socio-demographics, utilization, diagnoses, laboratory tests, pharmacy utilization, membership history, and vaccination. This study was carried out using structured data and free-text clinical notes from the EHR. The Institutional Review Board at KPSC approved this study.

Vaccinated Population With Presumptive Shoulder Injury

The study was conducted among KPSC members ≥ 3 years of age who had at least one intramuscular vaccine administered in the arm between 04/1/2016 and 12/31/2017 within a KPSC facility (Figure 1). Each vaccination was specified by the enrollee's unique identifier, the vaccination date (index date, i.e., Day 0), and the laterality of vaccination. Membership was required for the 180 days before and after the index date.

Among the vaccinated population described above, we identified individuals with a presumptive shoulder injury using International Classification of Diseases 10th Revision Clinical Modification (ICD-10-CM) codes (Multimedia Appendix 1) within 180 days after the index date; the laterality of the shoulder injury code had to match that of the vaccination. We excluded

individuals who had a shoulder-related appointment or had a shoulder injury code within 180 days before the index date.

On Day 0, members could have had clinical visits with pre-existing shoulder conditions and subsequently receive vaccinations. To exclude these Day 0 pre-existing conditions, we required at least two encounters on Day 0, of which at least one of the latter encounters had to be an urgent care, emergency department, or virtual visit. Also, Day 0 encounters were excluded if the first encounter on Day 0 had a shoulder injury code or if the encounter occurred before vaccination. In order to exclude vaccine-related local reactions, one of the most common adverse events occurring shortly after vaccination, a shoulder injury code also needed to appear during Days 31-180 post-vaccination.

Figure 1. Flow Chart Showing the Selection of Eligible Vaccinated Individuals With Presumptive Shoulder Injuries, Application of The NLP Algorithm, and SIRVA Case Confirmation Results.

The index date is the vaccination date.

ICD (ICD-10-CM) = International Classification of Diseases, 10th Revision, Clinical Modification.

The complete list of ICD-10 codes is provided in Multimedia Appendix 1.

SIRVA Case Definition

The VICP's SIRVA case definition was created for medical-legal purposes. To meet this case definition, a vaccine recipient must manifest all of the following[3]:

- 1) Pain and reduced range of motion are limited to the shoulder in which the intramuscular vaccine was administered;
- 2) Pain occurs within 48 hours of vaccination;
- 3) No history of pain, inflammation or dysfunction of the affected shoulder prior to vaccination that would explain the alleged condition;
- 4) No other condition or abnormality is present that would explain the patient's symptoms.

To be eligible to file a VICP claim, symptoms must last more than 6 months after vaccination.[21]

Based on the VICP SIRVA case definition and other publications,[1, 8, 12, 22, 23] we created a SIRVA case definition suitable for a population-based study using EHR data. A valid SIRVA case needed to meet five criteria:

- 1) Damage to the shoulder region occurred and was confirmed by signs and symptoms (i.e., pain, limited range of motion, weakness, stiffness) and clinical diagnosis.
- 2) Shoulder injury occurred in the same arm in which a vaccine was injected.
- 3) Shoulder injury started within seven days after vaccination.
- 4) Vaccination was a possible cause of the shoulder injury and no other known causes were associated with the shoulder injury.
- 5) Shoulder injury lasted more than 30 days post-vaccination.

Sub-population for Training and Validation of NLP Algorithm

To increase the likelihood of including true SIRVA cases in the datasets used for training and validating the NLP algorithm, we applied additional criteria to the presumptive cases to define a sub-population (n=517) (Figure 2). The criteria were as follows: 1) exclusion of individuals with an external shoulder injury (e.g., accident) code during Days -180 to 180, 2) exclusion of individuals with a shoulder injury code on Day 0, 3) requirement of a shoulder injury code during Days 1-30, and 4) requirement of a shoulder injury code on at least two different dates during Days 31-180. The criteria were based on characteristics of chart-confirmed SIRVA cases from a prior study.[12]

Training Dataset (n=164)

From the sub-study population described above, we selected a random sample for chart review. The NLP algorithm was built and refined based on incremental releases of training data.[24] Once the NLP algorithm stabilized and achieved good performance, we stopped the training process. The final training dataset had 164 cases.

Validation Dataset (n=100)

From the remaining individuals in the sub-study population (n=353), we randomly selected another 100 cases to form the validation dataset. The chart review results were used to evaluate the performance of the final NLP algorithm.

Figure 2. Flow Chart to Create the Dataset for Training and Validation Data Sampling

ICD-10 code group A = shoulder disorder diagnoses reported in the SIRVA literature

ICD-10 code group B = other shoulder disorder diagnoses not previously reported in the SIRVA literature

ICD-10 code group C = shoulder symptom codes

ICD-10 code group D = shoulder injury codes (ICD-10-CM chapter 19: Injury, poisoning and certain other consequences of external causes)

The complete list of ICD-10 codes is provided in Multimedia Appendix 1.

Manual Chart Review

We created a chart review form based on the SIRVA case definition. Chart abstractors reviewed the medical records and recorded information in the abstraction form (Multimedia Appendix 2) using the REDCap system[25]. The abstraction form was derived from a previous study of subdeltoid bursitis following vaccination, but was expanded to include other shoulder disorder diagnoses.[12] The chart abstraction and adjudication processes were similar to those used in past vaccine safety studies.[12, 19] An ascertainment period of 180 days after vaccination was used for both NLP and chart abstraction, allowing individuals sufficient time to seek medical care.[12] A second person reviewed each completed abstraction form for quality. A KPSC physician adjudicated the potential cases according to the SIRVA case definition for cases in which the chart reviewers had difficulty making a final assessment.

NLP Terminology Development

We built customized terminologies using a set of tools, including Linguamatics I2E[26], fastText[27], Gensim[28], GloVe[29], and NLTK[30]. The terminologies were derived from

various data sources, including the clinical notes of the study subjects, Vaccine Adverse Event Reporting System (VAERS) reports[31], ontologies (e.g., Unified Medical Language System (UMLS)[32]), semantic lexicons (e.g., WordNet[33]), and other online resources.

NLP Indexing

The pre-processing steps included section detection, sentence separation, and tokenization (i.e., segmenting text into linguistic units such as words and punctuation). For each token, the indexing process added annotations for matched concepts and general linguistic entities (e.g., lexical chunks like noun or verb phrases). Additional annotations captured linguistic variations such as wildcard, substring, spelling correction, and morphological variation.

NLP Search

We used a rule-based NLP algorithm for this study.[13, 24, 34, 35] The NLP algorithm was developed to search each indexed note at different levels: section (e.g., “Past Medical History”), intra-sentence, and cross-sentence. A distance-based relationship detection algorithm was applied to relate terms to other terms based on the number of words or sentences between them, thereby associating shoulder injury with information on vaccination site, temporality, or causality. The relationship detection algorithm also allowed for terms to be specified as ordered, or nested (e.g., an inner relation is an element of an outer relation). We used negation algorithms similar to pyConText/NegEx[36] to identify negated, uncertain and hypothetical statements. The relationship search identified the three types of information associated with shoulder injury which are described below.

1. Anatomic site

The NLP algorithm extracted the body location and laterality of the shoulder injury. For example, “left” and “arm” were identified as the laterality and body location of the shoulder injury, respectively, in the sentence “Patient has persistent pain in his left arm.”

2. Temporality

Temporal information was extracted by the NLP algorithm (Multimedia Appendix 3). Incomplete temporal information was inferred based on the note creation date. For example, dates with missing year information in clinical notes were assumed to occur near the note creation date. Linguistic terms, such as prepositions, were used to extract temporal relationships such as the onset date and duration linked with the vaccination event (e.g., “**for** 2 months”, “**over** the past 2 weeks”, “**since** last Thursday”).

3. Causality

Within each note, the NLP algorithm searched for possible causes of shoulder injury and classified them into seven types (Table S4.1 in Multimedia Appendix 4). The determination of causal relationships between cause and shoulder injury was made by lexical-syntactic rules based on over 70 trigger terms (Table S4.2, Multimedia Appendix 4). The terminologies for causes of shoulder injury other than vaccination are listed in Table S4.3, Multimedia Appendix 4. Moreover, for each relationship search above, we also extracted the vaccine name if available because multiple vaccines could be administered concomitantly or during follow-up.

NLP Case Classification

The final classification was based on the case definition described in the section “SIRVA case definition” by integrating vaccine, anatomic location, temporality, and causality information.

Because our algorithm emphasized sensitivity, we captured additional probable and possible cases identified by NLP with weaker evidence as defined by the following three criteria:

- Cross-sentence causality: The vaccination cause was identified only by cross-sentence causal relationship search. For example, shoulder injury and vaccination were described in separate sentences:

“Patient requesting an appointment for evaluation for left arm pain.

States experiencing pain x 1 month s/p flu vaccine.”

- Vaccination was identified as a cause of shoulder injury ≤ 30 days after vaccination. Because causality was less likely to be documented when the visit date was farther away from the onset date, vaccination might only be established as the cause of the shoulder injury within ≤ 30 days of vaccination, but not > 30 days after vaccination.
- Vaccine mismatch: The vaccine associated with shoulder injury that was documented in the clinical note did not match the vaccine recorded in the vaccination file.

Positive cases which met the SIRVA case definition were further classified into three groups:

- “definite” if they met none of the three criteria;
- “probable” if they met only one of the three criteria;
- “possible” if they met ≥ 2 of the three criteria.

NLP algorithm performance

We evaluated the NLP algorithm's accuracy in identifying SIRVA cases compared to the chart review reference standard in the validation dataset. We calculated sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) and their 95% confidence intervals (CIs). Since the NLP algorithm could potentially be accurate in determining a case not to be

SIRVA, but based on an incorrect assessment of an individual component of the SIRVA case definition not being met, we also conducted an error analysis of cases in which there were discrepancies between the NLP algorithm and chart review for individual components of the case definition.

Application of NLP Algorithm to Study Population and Chart Confirmation

The final NLP algorithm was applied to the broader study population of vaccinated individuals with presumptive shoulder injury (based on codes) to identify potential SIRVA cases. We performed manual chart confirmation on a random sample of NLP-identified definite cases and all NLP-identified probable and possible cases, and calculated chart confirmation rates and their 95% CIs.

We assembled the final group of SIRVA cases based on the chart review results. We calculated the time between vaccination and the first visit for a shoulder disorder in these SIRVA cases. We also examined the vaccination-related temporal and causal statements in their clinical notes.

Results

Out of 3,758,764 eligible vaccinations, we identified 77,819 records with a shoulder injury code (Figure 1). Among them, 16,048 had a code on Day 0. After applying the Day 0 inclusion criteria, the number of Day 0 records remaining was 100. The NLP algorithm was applied to 53,585 vaccinated individuals with presumptive shoulder injury.

Validation Results

The NLP algorithm achieved perfect accuracy (100%) in identifying the 4 SIRVA cases from the validation dataset (n = 100). However, the small number of positive cases resulted in wide CIs for sensitivity and PPV (39.6-100.0%) (Table 1). Meanwhile, the CIs for specificity and NPV remained narrow (95.2-100.0%).

Table 1. Accuracy Measurements of the NLP Algorithm on Identifying SIRVA, as Compared With the Chart-confirmed Validation Data

Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
100.0	100.0	100.0	100
39.6-100.0	95.2-100.0	39.6-100.0	95.2-100.0

Note: 95% confidence intervals are displayed beneath point estimates

PPV = positive predictive value; NPV = negative predictive value

Discrepancies between the NLP algorithm and chart review were investigated by component (Multimedia Appendix 5). For laterality, discrepancies were typically due to conflicting evidence or documentation errors in the clinical notes themselves. For temporality, the NLP algorithm incorrectly assigned symptom onset when performing cross-sentence searches, and incorrectly assigned injury duration based on incorrect laterality or capture of a resolved shoulder injury. For causality, the NLP algorithm missed causes such as daily activity and accident.

NLP-identified Potential SIRVA Cases (n=467)

We applied the final NLP algorithm to the clinical notes of 53,585 vaccinated individuals with presumptive shoulder injury. Among them, 99.9% (n=53,530) had at least one clinical note on Days 0-180 following vaccination. The total number of clinical notes searched by NLP was 4,292,610. The average number of clinical notes per individual was 80. The NLP algorithm identified shoulder injury in 46,086 records, and 96.5% of them had matched laterality compared

to the vaccination files. The NLP algorithm identified at least one cause for 55.5% of the NLP-identified shoulder injury cases. The temporal relation search identified the onset date for 99.2% of the NLP-identified shoulder injury cases. About 76% of these NLP-identified shoulder injury cases had symptom duration > 30 days post-vaccination. The number of potential SIRVA cases identified by the NLP algorithm was 467, classified into 291 definite, 124 probable, and 52 possible SIRVA cases.

Final SIRVA Cases After Chart Review (n=378)

We performed chart review on 129 definite and all probable and possible NLP-identified SIRVA cases (Table 2). The chart confirmation rates were 95.3% [95% CI, 90.2% to 97.9%], 67.7% [95% CI, 59.1% to 75.3%], and 18.9% [95% CI, 8.7% to 30.8%] for the definite, probable, and possible groups, respectively. We did not chart review the remaining 162 NLP-identified definite cases due to a lack of resources and high confirmation rate for this group of cases. After including these 162 NLP-definite cases and 216 chart-confirmed cases, the final number of SRIVA cases was 378.

Table 2. Number of NLP-identified Cases and Chart-confirmed Cases

NLP-identified group	NLP Identified	Chart Abstracted	Chart Confirmed	Confirmation Rate, %
Definite	291	129	123	95.3%
Probable	124	124	84	67.7%
Cross-sentence causality	64	64	46	71.9%
Vaccination cause identified ≤ 30 days after vaccination	41	41	26	63.4%
Vaccine mismatch	19	19	12	63.2%
Possible	52	52	9	18.9%
Total	467	305	216	70.8%

Among these 378 cases, the median time from vaccination to the first visit with a shoulder injury code was 43 days (interquartile range: 21–80 days, range 0-180 days). The majority of cases (361, 95.5%) had explicit temporal statements on symptoms onset in relation to vaccination. For example, “L shoulder pain that started the day she got a flu shot”, “Right shoulder pain and neck stiffness since immunizations.” The symptom onset for the remaining cases (17, 4.5%) could be derived based on the date of clinical visit, symptom duration, and causality statement. For example, “Reports having R shoulder pain for last 2 months. Thought related to vaccine she received in R arm.” There were 147 additional cases that had explicit causal statements regarding the shoulder condition and the vaccination, e.g., “status post vaccination - suspect rotator cuff irritation from vaccination itself.” Of those, 40 cases had mention of incorrect vaccine administration.

Discussion

SIRVA is a rare outcome following vaccination without a specific diagnosis code, and it is impractical to conduct manual chart review to identify all SIRVA cases. We developed and validated an NLP algorithm to identify potential SIRVA cases with high accuracy. The only past population-based study on SIRVA[12] was limited to shoulder bursitis after influenza vaccination. In that study, a random sample of 526 out of 1098 presumptive cases was chart reviewed to identify 12 incident subdeltoid bursitis cases attributed to vaccination. In this study, we included individuals with all types of shoulder disorder diagnoses after vaccinations. Out of 53,585 presumptive cases, the NLP algorithm combined with manual chart review yielded 378 SIRVA cases. Among 3.8 million vaccinations, the rate of SIRVA in this study was around 1 per

10,000 vaccinations. It should be noted that our SIRVA case definition was different from that of the VICP and other studies in regards to symptom onset, duration, and severity.

Although the NLP algorithm's overall accuracy was high, some challenges remained with the laterality component, despite the addition of laterality information in ICD-10-CM coding. First, descriptions of symptom location might not be precise. For example, the arm could refer to the region from the shoulder joint to the elbow joint (upper arm) or further down to the wrist. Second, the laterality recorded in the vaccine file or documented in the clinical notes could be incorrect. These issues must be considered when conducting studies using anatomic and laterality information.

There were several lessons learned from the temporality component of the NLP algorithm. First, there could be documentation of multiple onset dates during the 180 days after vaccination. Second, the disease onset information was more likely to be incomplete or inaccurate when the onset date was in the distant past. It could make it difficult to determine the onset date if the clinical visit date was further away from the vaccination date. In this study, to maximize sensitivity, any potential case with an onset falling within the predefined onset window satisfied the onset criteria.

In our study, the causality component worked reasonably well in identifying vaccination-related causality statements. While the provider or patient may have stated that the shoulder injury was vaccination-related, such statements do not provide definitive proof of causality. Because shoulder symptoms could have an insidious onset with multiple contributing factors, it was

difficult to draw definitive conclusions about cause and effect. To improve specificity, we excluded cases with non-vaccination causes of shoulder injury. However, it was still challenging to identify non-vaccination causes. First, there were numerous causes of shoulder injuries. Second, some of the causes could also be the treatment for the shoulder problem. For example, exercise could be both the cause and the therapy plan for shoulder injuries. Third, the cause of shoulder injury was often not mentioned in the clinical notes. In this study, the NLP algorithm could not identify the cause in about half of the cases. Lastly, the cause of shoulder injury was often not described in the same sentence as the shoulder symptom. The cross-sentence relationship search increased the sensitivity but decreased specificity. Causal relations have been studied extensively in the NLP field [37], but only a few studies focused on health-related causal relations and were conducted using Twitter messages [38] and literature [39-41]. One study extracted causal relations from clinical text using three causal key phrases (because, due, secondary to) and discontinuation key phrases to detect adverse drug reactions in ambulatory notes and achieved high specificity (98%) but low sensitivity (31%) and PPV (45%).[42]

SIRVA-related shoulder symptoms are common for other acute or chronic medical conditions with many possible causes. Correctly integrating the NLP-identified laterality, temporality, and causality information is nontrivial. For the same patient, different clinical encounters could attribute the shoulder injury to different causes. In this study, we made patient-level classifications by utilizing the information identified from all the components from all the notes. The combination of information across multiple notes increased the sensitivity of finding SIRVA cases but reduced the specificity since the NLP algorithm could misinterpret unrelated information extracted from multiple notes.

Because we tailored the NLP algorithm to emphasize sensitivity, the confirmation rates were low in the probable (67.7%) and possible (18.9%) groups. However, since SIRVA is a rare event, manual review of all the probable and possible cases was feasible in this study. In future studies, instead of categorizing the NLP output based on the strength of evidence, one could build a machine learning model on top of the NLP outputs[13] to further improve accuracy and develop thresholds. The SIRVA cases identified in this study could also serve as training data for a machine learning algorithm.

Limitations

This study had some potential limitations. Our study population was limited to vaccinated individuals with a diagnosis code for shoulder injury. However, loss of sensitivity is expected to be minimal since we used a comprehensive list of codes. Additionally, shoulder injuries can last a long time and are often accompanied by repeated visits. The 6-month lookback window used in this study might not have been sufficient to remove pre-existing shoulder conditions. Failure to exclude prior shoulder conditions could reduce the specificity of the NLP algorithm. In our vaccine-related local reaction study[20], most people diagnosed with a presumptive code of interest on Day 0 had symptom onset before vaccination. In this study, we excluded most cases with a shoulder injury code on Day 0. Further research is needed to study the association between SIRVA and Day 0 shoulder injury codes. Finally, because our method was tailored to this specific outcome after vaccination, its generalizability for use with other outcomes is unclear.

Conclusion

We developed and validated an NLP algorithm to identify potential SIRVA cases among vaccinated individuals with presumptive shoulder injury. The algorithm achieved high sensitivity and reasonable specificity. The NLP algorithm can potentially be used in future population-based studies to identify this rare adverse event, avoiding labor-intensive chart review validation.

Acknowledgments

We want to thank the following persons for their contributions to data collection and medical record abstraction: Anna Lawless, Bernadine Dizon, Claire Park, Denison Ryan, Jose Pio, Joy Gelfond, Karen Schenk, Kerresa Morrisette, Melena Taylor, Nancy Canul-Jauriga, Radha Bathala, and Sunhea Kim. This study was funded through the Vaccine Safety Datalink under contract 200-2012-53580 from the Centers for Disease Control and Prevention (CDC). The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Conflicts of Interest

Lina Sy received research support from GlaxoSmithKline, Dynavax, Seqirus, and Novavax for studies unrelated to this paper. All other authors report no conflicts of interest related to the submitted work.

References

1. Atanasoff S, Ryan T, Lightfoot R, Johann-Liang R. Shoulder injury related to vaccine administration (SIRVA). *Vaccine*. 2010 Nov 29;28(51):8049-52. PMID: 20955829. doi: 10.1016/j.vaccine.2010.10.005.
2. Institute of Medicine. *Adverse Effects of Vaccines: Evidence and Causality*. Stratton K, Ford A, Rusch E, Clayton EW, editors. Washington, DC: The National Academies Press; 2012. 894 p. ISBN: 978-0-309-21435-3.
3. Health Resources and Services Administration. Vaccine Injury Table. 2017 [cited 2020 10/26/2020]; Available from: <https://www.hrsa.gov/sites/default/files/vaccinecompensation/vaccineinjurytable.pdf>.
4. National Vaccine Injury Compensation Program Data Report. 2020; Available from: <https://www.hrsa.gov/vaccine-compensation/data/index.html>.
5. American Academy of Orthopaedic Surgeons. Rotator cuff tendinopathy, adhesive capsulitis, and arthritis cannot be caused by vaccine administration. Position statement 1190. June 2019 [updated June 2019; cited 2020 9/12/2020]; Available from: <https://www.aaos.org/contentassets/1cd7f41417ec4dd4b5c4c48532183b96/1190-rotator-cufftendinopathy-adhesive-capsulitis-and-arthritis-can-not-be-caused-by-vaccine-administration.pdf>.
6. Department of Health and Human Services. National Vaccine Injury Compensation Program: Revisions to the Vaccine Injury Table. 2020; 12]. Available from: <https://www.federalregister.gov/documents/2020/07/20/2020-15673/national-vaccine-injury-compensation-program-revisions-to-the-vaccine-injury-table>.

7. Gonzalez AI, Kortlever JTP, Moore MG, Ring DC. Influenza Vaccination Is Not Associated with Increased Number of Visits for Shoulder Pain. Clin Orthop Relat Res. 2020 Mar 2. PMID: 32141910. doi: 10.1097/CORR.0000000000001215.
8. Martin Arias LH, Sanz Fadrique R, Sainz Gil M, Salgueiro-Vazquez ME. Risk of bursitis and other injuries and dysfunctions of the shoulder following vaccinations. Vaccine. 2017 Sep 5;35(37):4870-6. PMID: 28774564. doi: 10.1016/j.vaccine.2017.07.055.
9. Bodor M, Montalvo E. Vaccination-related shoulder dysfunction. Vaccine. 2007 Jan 8;25(4):585-7. PMID: 17064824. doi: 10.1016/j.vaccine.2006.08.034.
10. Salmon JH, Geoffroy M, Eschard JP, Ohl X. Bone erosion and subacromial bursitis caused by diphtheria-tetanus-poliomyelitis vaccine. Vaccine. 2015 Nov 17;33(46):6152-5. PMID: 26458794. doi: 10.1016/j.vaccine.2015.09.090.
11. Erickson BJ, DiCarlo EF, Brause B, Callahan L, Hannafin J. Lytic Lesion in the Proximal Humerus After a Flu Shot: A Case Report. JBJS Case Connect. 2019 Jul-Sep;9(3):e0248. PMID: 31274645. doi: 10.2106/JBJS.CC.18.00248.
12. Hesse EM, Navarro RA, Daley MF, Getahun D, Henninger ML, Jackson LA, et al. Risk for Subdeltoid Bursitis After Influenza Vaccination: A Population-Based Cohort Study. Ann Intern Med. 2020 Jun 23. PMID: 32568572. doi: 10.7326/M19-3176.
13. Zheng C, Rashid N, Wu YL, Koblick R, Lin AT, Levy GD, et al. Using natural language processing and machine learning to identify gout flares from electronic clinical notes. Arthritis Care Res (Hoboken). 2014 Nov;66(11):1740-8. PMID: 24664671. doi: 10.1002/acr.22324.
14. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc. 2011 Sep-Oct;18(5):544-51. PMID: 21846786. doi: 10.1136/amiajnl-2011-000464.

15. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc.* 2013 Jan 01;20(1):144-51. PMID: 22733976. doi: 10.1136/amiajnl-2011-000681.
16. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc.* 2013;21(2):221-30. doi: 10.1136/amiajnl-2013-001935.
17. Cohen KB, Demner-Fushman D. *Biomedical Natural Language Processing*: John Benjamins Publishing Company; 2014. ISBN: 9789027271068.
18. Yu W, Zheng C, Xie F, Chen W, Mercado C, Sy LS, et al. The use of natural language processing to identify vaccine-related anaphylaxis at five health care systems in the Vaccine Safety Datalink. *Pharmacoepidemiol Drug Saf.* 2020 Feb;29(2):182-8. PMID: 31797475. doi: 10.1002/pds.4919.
19. Zheng C, Yu W, Xie F, Chen W, Mercado C, Sy LS, et al. The use of natural language processing to identify Tdap-related local reactions at five health care systems in the Vaccine Safety Datalink. *Int J Med Inform.* 2019 2019/07/01/;127:27-34. doi: 10.1016/j.ijmedinf.2019.04.009.
20. Koebnick C. Sociodemographic Characteristics of Members of a Large, Integrated Health Care System: Comparison with US Census Bureau Data. *The Permanente Journal.* 2012 2012/08/01/:37-41. doi: 10.7812/TPP/12-031.
21. Health Resources & Services Administration. What You Need to Know About the National Vaccine Injury Compensation Program (VICP). 2019 [updated April 2019; cited 2021 03/23/2021]; Available from: www.hrsa.gov/sites/default/files/hrsa/vaccine-compensation/resources/about-vaccine-injury-compensation-program-booklet.pdf.

22. Hibbs BF, Ng CS, Museru O, Moro PL, Marquez P, Woo EJ, et al. Reports of atypical shoulder pain and dysfunction following inactivated influenza vaccine, Vaccine Adverse Event Reporting System (VAERS), 2010-2017. *Vaccine*. 2020 Jan 29;38(5):1137-43. PMID: 31784231. doi: 10.1016/j.vaccine.2019.11.023.
23. Hesse EM, Atanasoff S, Hibbs BF, Adegoke OJ, Ng C, Marquez P, et al. Shoulder Injury Related to Vaccine Administration (SIRVA): Petitioner claims to the National Vaccine Injury Compensation Program, 2010-2016. *Vaccine*. 2020 Jan 29;38(5):1076-83. PMID: 31771864. doi: 10.1016/j.vaccine.2019.11.032.
24. Zheng C, Yu W, Xie F, Chen W, Mercado C, Sy LS, et al. The use of natural language processing to identify Tdap-related local reactions at five health care systems in the Vaccine Safety Datalink. *Int J Med Inform*. 2019 Jul;127:27-34. PMID: 31128829. doi: 10.1016/j.ijmedinf.2019.04.009.
25. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009 Apr;42(2):377-81. PMID: 18929686. doi: 10.1016/j.jbi.2008.08.010.
26. Cormack J, Nath C, Milward D, Raja K, Jonnalagadda SR. Agile text mining for the 2014 i2b2/UTHealth Cardiac risk factors challenge. *J Biomed Inform*. 2015 Dec;58 Suppl:S120-7. PMID: 26209007. doi: 10.1016/j.jbi.2015.06.030.
27. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*. 2017;5:135-46.

28. Rehurek R, Sojka P. Software Framework for Topic Modelling with Large Corpora. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks; Valletta, Malta. vrehruvrek2010software: ELRA; 2010. p. 45--50.
29. Pennington J, Socher R, Manning C, editors. Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP); 2014.
30. Loper E, Bird S. NLTK: the Natural Language Toolkit. Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics - Volume 1; Philadelphia, Pennsylvania. 1118117: Association for Computational Linguistics; 2002. p. 63-70.
31. Shimabukuro TT, Nguyen M, Martin D, DeStefano F. Safety monitoring in the Vaccine Adverse Event Reporting System (VAERS). Vaccine. 2015 Aug 26;33(36):4398-405. PMID: 26209838. doi: 10.1016/j.vaccine.2015.07.035.
32. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267-70. PMID: 14681409. doi: 10.1093/nar/gkh061.
33. Miller GA. WordNet: a lexical database for English. Commun ACM. 1995;38(11):39-41. doi: 10.1145/219717.219748.
34. Zheng C, Luo Y, Mercado C, Sy L, Jacobsen SJ, Ackerson B, et al. Using natural language processing for identification of herpes zoster ophthalmicus cases to support population-based study. Clin Exp Ophthalmol. 2018 Jun 19. PMID: 29920898. doi: 10.1111/ceo.13340.

35. Zheng C, Sun BC, Wu YL, Lee MS, Shen E, Redberg RF, et al. Automated Identification and Extraction of Exercise Treadmill Test Results. *Journal of the American Heart Association*. 2020 Mar 3;9(5):e014940. PMID: 32079480. doi: 10.1161/JAHA.119.014940.
36. Chapman BE, Lee S, Kang HP, Chapman WW. Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. *J Biomed Inform*. 2011 Oct;44(5):728-37. PMID: 21459155. doi: 10.1016/j.jbi.2011.03.011.
37. Asghar N. Automatic extraction of causal relations from natural language texts: a comprehensive survey. *arXiv preprint arXiv:160507895*. 2016.
38. Doan S, Yang EW, Tilak SS, Li PW, Zisook DS, Torii M. Extracting health-related causality from twitter messages using natural language processing. *BMC Med Inform Decis Mak*. 2019 Apr 4;19(Suppl 3):79. PMID: 30943954. doi: 10.1186/s12911-019-0785-0.
39. Khoo CS, Chan S, Niu Y, editors. Extracting causal knowledge from a medical database using graphical patterns. *Proceedings of the 38th annual meeting of the association for computational linguistics*; 2000.
40. Mihaila C, Ohta T, Pyysalo S, Ananiadou S. BioCause: Annotating and analysing causality in the biomedical domain. *BMC Bioinformatics*. 2013 Jan 16;14:2. PMID: 23323613. doi: 10.1186/1471-2105-14-2.
41. Bakal G, Talari P, Kakani EV, Kavuluru R. Exploiting semantic patterns over biomedical knowledge graphs for predicting treatment and causative relations. *J Biomed Inform*. 2018 Jun;82:189-99. PMID: 29763706. doi: 10.1016/j.jbi.2018.05.003.
42. Cantor MN, Feldman HJ, Triola MM. Using trigger phrases to detect adverse drug reactions in ambulatory care notes. *Qual Saf Health Care*. 2007 Apr;16(2):132-4. PMID: 17403760. doi: 10.1136/qshc.2006.020073.