Research Article
Volume 13:09, 2022

Journal of
Health & Medical Informatics

ISSN: 2157-7420 Open Access

Cohort Identification for Trampoline-associated Traumatic Dental Injuries among Pediatric Patients from Clinical Notes Using Machine Learning

Joseph W. Sirrianni[1]*, Jin Peng[1], Yungui Huang[1] and Homa Amini[2]

[1]Nationwide Children's Hospital, IT Research and Development, 4321 S 18th St, Columbus, OH, 43205-2664, USA
[2]Nationwide Children's Hospital, Department of Dentistry, 4321 S 18th St, Columbus, OH, 43205-2664, USA

Abstract

Background: Cohort identification is a crucial task for performing retrospective clinical analysis. The utilization of natural language processing, especially the modern and advanced approaches using deep learning modeling, may improve this task by allowing for improved classification of patients by cohort status. However, this utilization has not been applied in the dentaldomain.

Objective: We aim to identify patients that suffer trampoline-associated traumatic dental injuries among all trampoline-associatedinjuries.

Methods: We develop and apply a natural language processing cohort identification pipeline, consisting of text filtering rules and a machine learning model trained using historic data. The pipeline processes a patient's clinical notes for a series of temporally related encounters and produces a binary prediction of whether the patient has suffered a trampoline-injury or not. We experimented with six different machine learning models: logistic regression, random forest, decision tress, linear-SVM, naïve bayes, and a fine-tuned ClinicalBERT model.

Results: The fine-tuned ClinicalBERT model had the best performance of the models on our evaluation data with a PPV of 0.836 and a sensitivity of 0.898. The application of the pipeline on our data increased the cohort size for all trampoline injuries from an initial 7454 patients to 15,010 patients and the trampoline-associated traumatic dental injuries cohort from an initial 102 patients to 140 patients.

Conclusion: We present a novel natural language processing powered pipeline for identifying a trampoline-associated injury cohort for dental research. Our results demonstrate the superiority of deep learning over traditional machine learning models on our specific task. Our process for identifying patient encounters by activity type is generalizable to several different types of injuries and applicable to other research cohorts.

Keywords: Cohort Identification • Natural Language Processing • Machine Learning • Deep Learning

Introduction

The wide adoption of electronic health records (EHRs) for patient care has also offered greater opportunity for researchers to conduct quantitative and qualitative retrospective analysis through secondary use of EHR [1]. These EHRs contain information in both structured fields, such as time of visit, billing codes, and diagnosis codes, and unstructured fields, such as clinical notes and images. Typical research process starts with identifying a patient cohort meeting inclusion and exclusion criteria, often with manual chart reviews, then moves to getting data on the cohort by querying their EHR structured data fields. This approach works for much research, but it restricts the types of patient cohorts that can be reliably constructed at scale, as majority of patient information is not captured in structured data field but are buried in unstructured fields [2].

The unstructured clinical notes offer a more descriptive account of patient information and are a powerful resource for deriving more specific patient cohorts [3]. However, extracting desired information from unstructured fields in an EHR is difficult.

Natural Language Processing (NLP), a subfield of artificial intelligence (AI), is a powerful approach that can be used to enhance cohort identification by allowing more sophisticated querying of unstructured data in patient EHRs [4]. Indeed, NLP has been applied extensively in the clinical domain for cohort identification using EHRs [5-9]. These NLP-powered cohort identifications systems have leveraged rule-based systems [8], traditional machine learning models [6,7,9], and deep learning models [3,5,6,7,9]. Leveraging NLP for improved for cohort identification has been observed in several clinical subdomains, including radiology [6], ICU patients [7], cardiology [9], and clinical trials [5]. However, to our knowledge, NLP has not been utilized for identifying dental patient cohorts. Our work aims to fill this research gap.

In this work, we are interested in examining the proportion of trampoline-associated injuries that result in a traumatic dental injury among pediatric patients at Nationwide Children's Hospital (NCH). In this case, identifying documented traumatic dental injuries is very straightforward; existing traumatic dental injury ICD codes were consistently used in the patient EHRs and are easily query-able. However, identifying trampoline-related injuries is considerably more difficult; trampoline injuries do have query-able ICD codes (ICD-9: E005.3 and ICD-10: Y93.44), but they are not reliably applied to all the patients in our historic records. Thus, to identify patients that suffered trampoline-related injuries, we cannot rely solely on the structured data, but must also consider the unstructured, clinical note fields for explicit mentions of trampoline injuries.

Objectives

Our objective is to identify uncoded instances of trampoline-associated

***Address for Correspondence:** Joseph W. Sirrianni, Nationwide Children's Hospital IT Research and Development, 4321 S 18th St, Columbus, OH, 43205-2664, USA, Tel: 16147226009; E-mail: joseph.sirrianni@nationwidechildrens.org*

Date of Submission: 10 August, 2022, Manuscript No: jhmi-22-71656; **Editor assigned:** 12 August, 2022, PreQC No: P-71656; **Reviewed:** 15 September, 2022, QC No: Q-71656; **Revised:** 20 September, 2022, Manuscript No: R-71656; **Published:** 30 September, 2022, DOI: 10.37421/2157-7420.2022.13.436

injuries using NLP to expand our existing cohort, by leveraging patient clinical notes. Specifically, we seek to expand two cohorts: 1) a cohort of all trampoline-associated injuries, and 2) a cohort of all trampoline-associated dental injuries.

Methods and Materials

We developed a two phased NLP pipeline to predict if a given set of patent encounter notes contains a trampoline injury. The first phase applies text filtering rules to the note text. The second phase is a predictive model that will take in the note text as input and will produce a binary prediction about whether the note contains a trampoline injury. The entire development and application process is outlined in Figure 1.

Patient population

The patient population was extracted from EHR data at Nationwide Children's Hospital. The repository consists of all digital records of patients dating back as early as 2006 and contains over 93 million clinical notes.

Data querying and dataset construction

We collected our data into three datasets: positive data, negative data, and unknown data. Each instance was initially defined as a patient encounter, however, early on we ran into an issue handling follow-up encounters. These encounters would often contain the same language as an initial encounter for the injury but may not necessarily have the corresponding ICD code associated with it. This is an issue, because our pipeline only processes the clinical note text and will miss classify follow-up encounters.

To address this issue, we aggregated encounters for the same patient that occurred within six months of one another into encounter series. Each encounter series is a collection of encounters that ideally all relate to the same injury and will include both the patient's initial encounter at NCH and any subsequent follow-up encounters.

To gather encounter data into encounter series, we queried our in-house clinical note search engine that contains NCH's historic clinical note data. The search engine allows for search notes by keywords as well as filtering results by a note's metadata, such as note type, note author, and department.

The positive dataset consists of encounter series that have at least one trampoline ICD code associated with it. These are treated as true positives. The negative dataset consists of encounter series that contain encounters that either don't mention the word "trampoline" or only mention the word "trampoline" in a non-injury context note. The unknown dataset consists of encounter series where at least one encounter mentions the word "trampoline" in an injury context note. In total, the positive dataset had 7454 encounter series, the negative set had 16,700, and the unknown set had 16,228.

The definition of a non-injury and injury context note is based on a meta-data filter applied to the search queries. Please refer to the appendix for a full explanation of the data querying process.

NLP pipeline

Text pattern filtering: A manual review of notes revealed that many note templates mention the word "trampoline" as part of a list of prohibited activities for patients following their assessment. We applied a simple text pattern filter looking for two phrases that appear as part of note templates: 1) "<word> use, trampoline use" and 2) "jumping on a trampoline or rough house play should also be avoided". If the phrases were found that portion of the text was removed. If that was the only mention of "trampoline" in the note it was removed from the unknown dataset.

Model development: Five traditional machine learning models (Logistic Regression, Linear-SVM, Decision Tree, Random Forest, and Naïve Bayes) and one deep learning model (ClinicalBERT) [10] were trained on our dataset.

Data Pre-processing and feature extraction: Each instance's text was lower-cased, and the numbers, punctuations, and English stop words were removed. Then, the text is tokenized into 1-3 grams and weighted using term frequency-inverse document frequency. Tokens that appeared in fewer than 5% of the instances or more than 95% were removed from the vocabulary. In total, each instance had 1376 features that served as the input for the traditional machine learning models.

Model training: The training dataset was comprised of the positive and negative instances. The unknown instances were set aside in an Unknown set that was used to find the unlabeled positives to add to our final cohort. Each model was trained and analyzed using 5-fold cross validation across the entire training set. Grid search selection was used to select parameters for each of the models. The logistic regression model was trained using elastic net regularization. The ClinicalBERT model is a pre-trained deep neural network model fine-tuned on a large corpus of clinical notes 10. The model trained using a 90-10 split for training-validation for 4 epochs, after triggering early stopping.

Manual annotation of the unknown dataset

The positive and negative datasets were used to train and validate the predictive models. However, due to procedure used to create the positive, negative, and unknown datasets, there are distinct differences in their construction. So, the evaluation of the models on the training data may not reflect the performance of the models when applied to the unknown dataset.

To address this difference in datasets, we sampled 273 encounters. Two annotators labeled each of the 273 encounters labeling them as either positive, negatives, or unknown if they describe trampoline-associated injuries. In total, there were 108 positives, 163 negatives, and 2 unknowns in the sample. The unknowns were discarded. The inter-annotator reliability had a Cohen kappa score of 0.840.

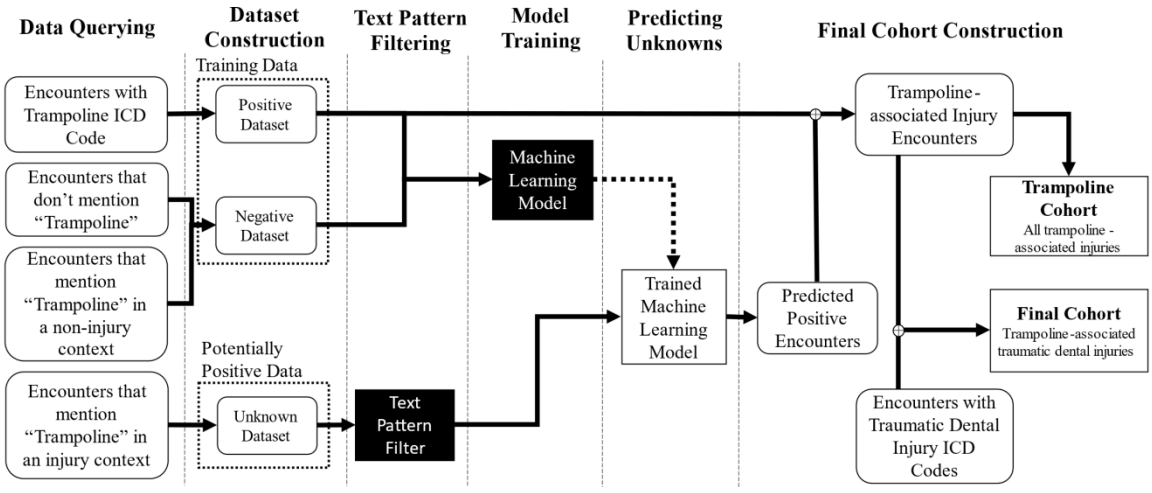

Figure 1. High-level framework for the trampoline prediction NLP pipeline.

Ethical considerations

This study received Nationwide Children's Hospital IRB approval (ID: STUDY00001450), chaired by Karen White, PhD.

Results

The results of each model's performance on the training data with 5-fold cross validation and the manually annotated unknown dataset are presented in Table 1, along with the threshold values for the four models that produce continuous logit outputs.

On the training data the fine-tuned ClinicalBERT model has the highest performance across sensitivity, PPV, and F1-score. Linear SVM had the next highest performance, followed by logistic regression.

On the unknown data, the fine-tuned ClinicalBERT model has the highest performance. Logistic Regression was the next highest performing model, followed by random forest and linear SVM models.

Final cohort construction

We applied the fine-tuned ClincalBERT model to the unknown set. The model predicted 7556 encounter series as trampoline-related injuries out of a possible total of 16,151. We combined these predictions with our initial labeled set of encounters to produce a total population size of trampoline related injuries as 15,010 encounter series.

In total, across the positive and unknown datasets there were 179 encounter series with a traumatic dental injury code. Of these dental injury encounter series, 102 had both a trampoline and dental injury ICD code. Of the remaining 77 in the unknown set, the text pattern filter eliminated all but 44 encounter series. The ClinicalBERT model identified 39 positives among these. A manual evaluation found that three instances where misclassified (one false negative and two false positives), bringing the total to 38 positives (Figure 2).

Discussion and Analysis

The deep learning model, ClinicalBERT, had superior performance than the traditional machine learning models. This result is consistent with the broader field of NLP and with the prior literature of applying deep learning for cohort identification. Gehrmann, et al. (2018) [7] compared the performance of convolutional neural networks (CNN) to logistic regression for cohort identification and found that CNNs outperformed logistic regression in all their tests. Likewise, Wu, et al. (2020) [9] compared CNNs with Recurrent Neural Networks (RNNs) to logistic regression and found that the deep learning models performed significantly better in terms of specificity and the RNN had greater F1-score and accuracy for one of their tasks. Ong, et al. (2020) [6] compared

Table 1. Performance of the models on the 5-fold cross validation training set and the set of manually annotated unknown data.

Model	Threshold	Training Dataset			Manually Annotated Dataset		
		Sensitivity	PPV	F1-score	Sensitivity	PPV	F1-score
Logistic Regression	0.45	0.950	0.861	0.905	0.824	0.824	0.824
Random Forest	0.55	0.942	0.895	0.904	0.815	0.800	0.807
Decision Tree	0.5	0.938	0.864	0.900	0.852	0.730	0.786
Linear SVM	-	0.958	0.903	0.930	0.815	0.800	0.807
Naïve Bayes	-	0.910	0.764	0.830	0.769	0.776	0.772
ClinicalBERT	0.5	0.960	0.926	0.943	0.898	0.836	0.866

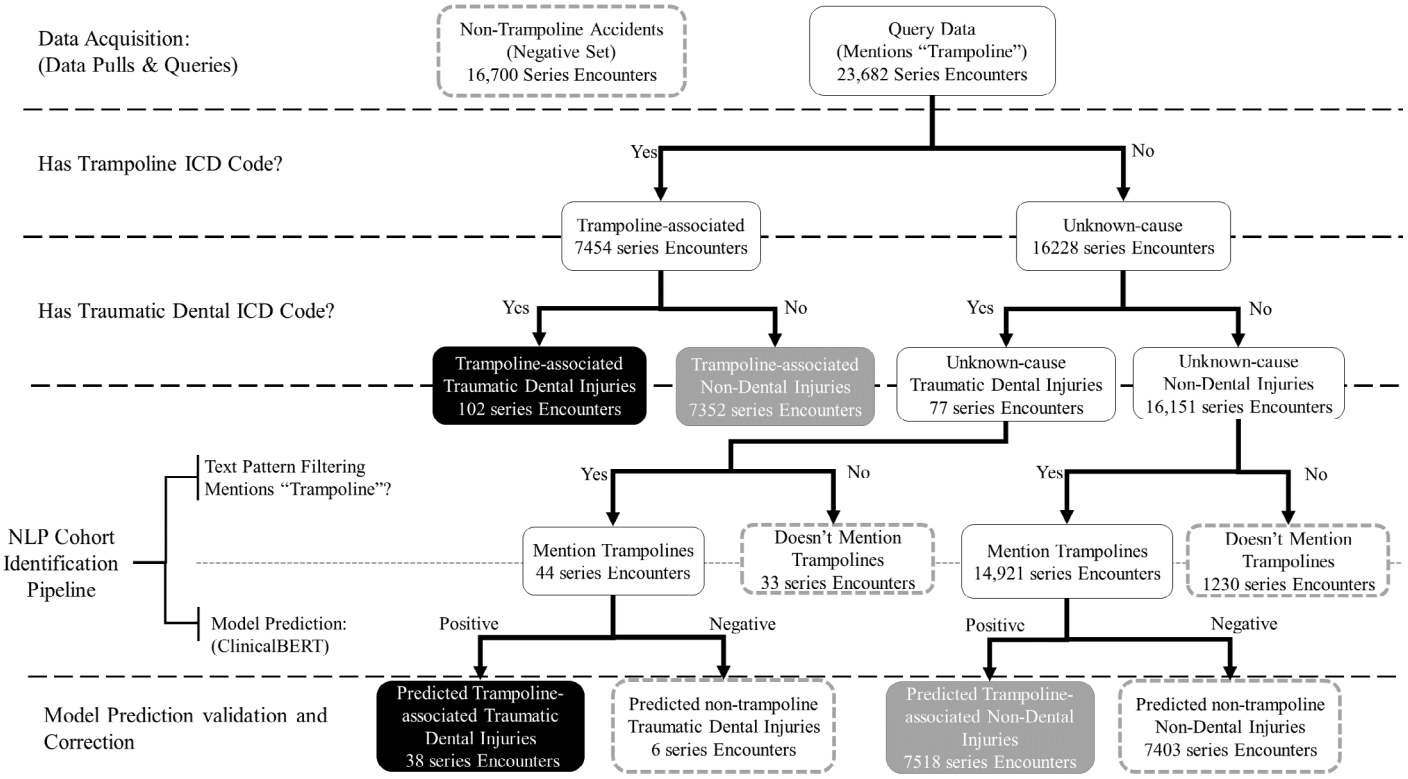

Figure 2. Breakdown of the cohort encounter series as different phases of the NLP cohort identification pipeline. Black rectangles indicate that the encounters are part of the final cohort, while grey rectangles indicate membership with the trampoline-injury cohort.

several models; include logistic regression, k-nearest neighbors, and several tree-based models. They found that the RNNs consistently outperform other models when paired with word-embedding features.

Another major observation of our results is the distinct drop in performance from the training dataset to the unknown dataset across all models. This drop is likely due to the differences between the training data and the unknown data. Our training dataset was constructed using a set of known positives encounters (encounters with the trampoline ICD code) and a collection of known/probable negative encounters from two sources. Thus, our training data is not perfectly representative of the unknown data, as for example, some negative cases come from a context that the unknown data does not. This discrepancy, while important for model training to cover more cases, is reflected in the model performance.

A manual examination of some of the false positive predictions shows that many specific cases can appear in the notes which can confuse the model. For example, several encounters that mentions historic trampoline injuries that may be relevant with ongoing injury, but not the cause of the immediate encounter. Likewise, some encounters mention trampoline injuries in passing, but are not linked to their current injury. Encounters such as these, highlight the difficulty of predicting trampoline-injuries using clinical note text alone.

We were able to identify a total of 6612 additional trampoline-associated injury encounters, bringing the trampoline cohort from 7336 initially to 13,949 encounters, a 1.9 times increase, and the trampoline-associated traumatic dental injury cohort from 100 to 139, a 1.39 times increase. As demonstrated by our results, this is a powerful method for increasing the size of cohorts in the dental domain.

While our application area in this paper is specific, our methodology is general enough to apply to a whole host of dental cohorts. The ability to intelligently process unstructured EHR fields at scale, will allow dental clinical research to examine a wider breath of patient encounters and outcomes than possible with prior methods.

Conclusion

Using NLP powered supervised machine learning; we developed a predictive model to determine if clinical note encounters are describing trampoline-associated injuries. Our model achieves a high level of predictive power and we were able to increase our cohort sizes by 1.9 and 1.39 times compared to only using the ICD codes from the structured fields. While we were focused on trampoline injuries, our methodology can be easily augmented to identify other dental cohorts. Any cohort with a sizable number of pre-identified positive and negative examples can utilize our supervised machine learning pipeline to develop a predictive model that can be applied to large scale, unknown data.

Limitations

The data used came from Nationwide Children's Hospital facilities and network, which is highly region specific and tends to share a set of standard clinical note types and templates. It is possible our models' performance varies when applied to clinician notes from other organizations.

Acknowledgements and Funding

The project described was supported by Award Number UL1TR002733 from the National Center for Advancing Translational Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Center for Advancing Translational Sciences or the National Institutes of Health. We would like to acknowledge Sarah Wagner for her work in pulling the data for this project.

Conflict of Interest

None Declared.

References

1. "Secondary Analysis of Electronic Health Records." *Cham (CH): Springer* (2016).
2. Wang, Yanshan, Ahmad Tafti, Sunghwan Sohn and Rui Zhang. "Applications of Natural Language Processing in Clinical Research and Practice." In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials, 22–25. Minneapolis, Minnesota: Association for Computational Linguistics (2019).
3. Gehrmann, Sebastian, Franck Dernoncourt, Yeran Li and Eric T. Carlson, et al. "Comparing Rule-Based and Deep Learning Models for Patient Phenotyping." ArXiv: 1703.08705 [Cs, Stat] (2017).
4. Sarmiento, Raymond Francis and Franck Dernoncourt. "Improving Patient Cohort Identification Using Natural Language Processing." In Secondary Analysis of Electronic Health Records. Cham (CH): Springer (2016).
5. Segura-Bedmar, Isabel and Pablo Raez. "Cohort Selection for Clinical Trials Using Deep Learning Models." *J Am Med Inform Assoc* 26 (2019): 1181–1188.
6. Ong, Charlene Jennifer, Agni Orfanoudaki, Rebecca Zhang and Francois Pierre M. Caprasse, et al. "Machine Learning and Natural Language Processing Methods to Identify Ischemic Stroke, Acuity and Location from Radiology Reports." *PLoS One* 15 (2020): e0234908.
7. Gehrmann, Sebastian, Franck Dernoncourt, Yeran Li and Eric T. Carlson, et al. "Comparing Deep Learning and Concept Extraction Based Methods for Patient Phenotyping from Clinical Narratives." *PLoS One* 13 (2018): e0192360.
8. Chen, Long, Yu Gu, Xin Ji and Chao Lou, et al. "Clinical Trial Cohort Selection Based on Multi-Level Rule-Based Natural Language Processing System." *J Am Med Inform Assoc* 26 (2019): 1218-1226.
9. Wu, Xiao, Yuzhe Zhao, Dragomir Radev and Ajay Malhotra. "Identification of Patients with Carotid Stenosis Using Natural Language Processing." *Eur Radiol* 30 (2020): 4125-4133.
10. Huang, Kexin, Jaan Altosaar and Rajesh Ranganath. "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission." ArXiv E-Prints 1904 (2019): arXiv: 1904.05342.

How to cite this article: Sirrianni, Joseph W., Jin Peng, Yungui Huang and Homa Amini. "Cohort Identification for Trampoline-associated Traumatic Dental Injuries among Pediatric Patients from Clinical Notes Using Machine Learning." J Health Med Informat 13 (2022): 436.