The Spine Journal 22 (2022) 272–277

Clinical Study

Natural language processing for prediction of readmission in posterior lumbar fusion patients: which free-text notes have the most utility?

Aditya V. Karhade, MD, MBA[a,b], Ophelie Lavoie-Gagne, MD[a], Nicole Agaronnik, BA[a], Hamid Ghaednia, PhD[a], Austin K. Collins, BS[a], David Shin, BS[a], Joseph H. Schwab, MD, MS[a,b,*]

[a] *Department of Orthopaedic Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA*
[b] *Harvard Combined Orthopaedic Residency Program, Boston, MA, USA*
Received 24 May 2021; revised 19 July 2021; accepted 9 August 2021

Abstract

BACKGROUND CONTEXT: The increasing volume of free-text notes available in electronic health records has created an opportunity for natural language processing (NLP) algorithms to mine this unstructured data in order to detect and predict adverse outcomes. Given the volume and diversity of documentation available in spine surgery, it remains unclear which types of documentation offer the greatest value for prediction of adverse outcomes.

STUDY DESIGN/SETTING: Retrospective review of medical records at two academic and three community hospitals.

PURPOSE: The purpose of this study was to conduct an exploratory analysis in order to examine the utility of free-text notes generated during the index hospitalization for lumbar spine fusion for prediction of 90-day unplanned readmission.

PATIENT SAMPLE: Adult patients 18 years or older undergoing lumbar spine fusion for lumbar spondylolisthesis or lumbar spinal stenosis between January 1, 2016 and December 31, 2020.

OUTCOME MEASURES: The primary outcome was inpatient admission within 90-days of discharge from the index hospitalization.

METHODS: The predictive performance of NLP algorithms developed by using discharge summary notes, operative notes, nursing notes, physical therapy notes, case management notes, medical doctor (MD) (resident or attending), and allied practice professional (APP) (nurse practitioner or physician assistant) notes were assessed by discrimination, calibration, overall performance.

RESULTS: Overall, 708 patients were included in the study and 83 (11.7%) had 90-day inpatient readmission. In the independent testing set of patients (n=141) not used for model development, the area under the receiver operating curve of NLP algorithms for prediction of 90-day readmission using discharge summary notes, operative notes, nursing notes, physical therapy notes, case management notes, MD/APP notes was 0.70, 0.57, 0.57, 0.60, 0.60, and 0.49 respectively.

CONCLUSION: In this exploratory analysis, discharge summary, physical therapy, and case management notes had the most utility and daily MD/APP progress notes had the least utility for prediction of 90-day inpatient readmission in lumbar fusion patients among the free-text documentation generated during the index hospitalization. © 2021 Elsevier Inc. All rights reserved.

Keywords: Artificial intelligence; Machine learning; Natural language processing; Prediction; Readmission; Spine

FDA device/drug status: Not applicable.

Author disclosures: ***AVK***: Nothing to disclose. ***OL-G***: Nothing to disclose. ***NA***: Nothing to disclose. ***HG***: Nothing to disclose. ***AKC***: Nothing to disclose. ***DS***: Nothing to disclose. ***JHS***: Scientific Advisory Board: Chordoma Foundation (Nonfinancial); Speaking and/or Teaching Arrangements: AO Spine (Travel Expense Reimbursement, Outside 12-Month Requirement), Stryker Spine (B, Outside 12-Month Requirement).

*Corresponding author. Department of Orthopedic Surgery, Massachusetts General Hospital, Harvard Medical School, 55 Fruit St, Boston, MA 02114, USA. Tel.: (617) 543-5227; fax: (617) 726-7587.

E-mail address: jhschwab@mgh.harvard.edu (J.H. Schwab).

https://doi.org/10.1016/j.spinee.2021.08.002
1529-9430/© 2021 Elsevier Inc. All rights reserved.

Introduction

Growing healthcare expenditures have prompted national cost containment efforts, including policy changes that incentivize reduction in hospital readmission rates. The Hospital Readmissions Reduction Program (HRRP) established by the Centers for Medicare and Medicaid Services (CMS) enforces payment penalties for hospitals with excess readmissions for six common procedures including elective primary total hip arthroplasty (THA) and total knee arthroplasty (TKA) [1]. The success of HRRP in reducing readmission rates [2] will likely lead to a continued expansion of high-cost procedures used for payment adjustment models, including spine surgery. Elective spine surgery for lumbar stenosis, an increasingly common procedure, has an estimated 90-day readmission rate of 7.2% [3], closely following estimates for TKA. The 90-day readmission rate for lumbar spinal fusion appears to exceed TKA, with some estimates as high as 24.8% [4]. These statistics underscore the importance of timely identification of salient preoperative characteristics and risk factors for adverse events and readmission following spine surgery.

Efforts to reduce cost and improve care quality have prompted development of automated machine learning methods for extraction of risk factors from electronic health records (EHRs).These methods process unstructured free text notations in a fraction of the time required by burdensome manual chart review, and have previously automated identification of incidental durotom [5], wound infection [6], and intraoperative vascular injury [7]. One study using machine learning classification of over 75 predischarge variables identified age, comorbidity burden, surgical duration, intraoperative morphine equivalents, total direct cost, and length of stay to be important predictors for readmission following spine surgery [8]. Despite the advantages of such automated methods for extraction of risk factors, speed and efficiency remain limited by the volume of information that must be processed in EHRs. Furthermore, efficiency of real-time adverse event monitoring requires identification of information categories (eg, specific note types) with high probability for pertinent documentation. The purpose of this study was to explore the utility of various note types for prediction of 90-day readmission following lumbar spinal fusion. Identification of the most high-yield note types for prediction of adverse outcomes can facilitate timely identification of risk factors.

Materials and methods

Data source

Retrospective review of electronic medical records two academic medical centers and three community hospitals was approved by our institutional review board. Inclusion criteria for the study were: (1) age 18 years or older undergoing; (2) posterior lumbar fusion (open or minimally invasive, with or without interbody) between January 1, 2016 and December 31, 2020; (3) in the inpatient setting for; (4) lumbar spinal stenosis or spondylolisthesis; and (5) with discharge summary notes, operative notes, nursing notes, physical therapy notes, case management notes, medical doctor (MD) (resident or attending) and allied practice professional (APP) (nurse practitioner or physician assistant) notes available for review. Patients were excluded if they underwent concurrent anterior intervention, concurrent surgery at other locations in the spine, surgery for tumor, trauma, infection, and/or adult spinal deformity.

Outcome

The primary outcome was defined as unplanned readmission within 90 days of discharge from the index hospitalization.

Data analysis

The available patient population was divided by a stratified split into training and testing cohorts. The free-text notes were preprocessed to standardize the input for the machine learning–based NLP algorithms. This process included: (1) removal of extraneous white space; (2) conversion of all words to lowercase; (3) removal of stop words such as “the” “and”; and (4) conversion to stems of all words (eg, “wheezing” to “wheez”). Next the processed free-text notes were converted into matrix form with the bag-of-words method. This technique generates a matrix (m by n) where m, the number of rows is equal to the number of patients in the population. N, the number of columns, represents the processed “words” or “features.” The contents of the matrix are the frequencies of the features in the free-text notes. Subsequently, the term-frequency inverse document frequency (TF-IDF) method was used to regularize the matrix to account for words with very low frequency and very high frequency. This process was repeated to generate bag-of-words matrices for each note type.

Finally, the processed matrices were used as input for an extreme gradient boosting (XGBoost) supervised machine learning algorithm. Performance of the XGBoost algorithm was assessed in the independent testing set not used for model development. Algorithm performance was assessed by discrimination (area under the receiver operating curve [AUROC], area under the precision-recall curve [AUPRC]), calibration (calibration intercept, calibration slope), and overall performance (Brier score). The null model Brier score was calculated as the Brier score for a model that would predict a predicted probability equal to the observed incidence of 90-day readmission for every patient.

To determine the features used by the algorithms for prediction of 90-day readmission, relative variable importance plots were generated. The Anaconda Distribution (Anaconda, Inc., Austin, TX), Python (Python Software Foundation, Wilmington, DE), R (The R Foundation, Vienna, Austria), and RStudio (RStudio, Boston, MA) were used for data analysis.

Results

Overall, 708 patients underwent posterior lumbar fusion and 83 (11.7%) had 90-day inpatient readmission. In the independent testing set of patients (n=141) not used for model development, the AUROC of NLP algorithms for prediction of 90-day readmission using discharge summary notes, operative notes, nursing notes, physical therapy notes, case management notes, MD/APP notes were 0.70, 0.57, 0.57, 0.60, 0.60, and 0.49, respectively (Table). The AUROC of an NLP algorithm using all note types as input data was 0.70. The highest AUPRC for prediction of 90-day readmission was achieved using all notes, physical therapy, and discharge summary notes, with 0.30, 0.24, and 0.23, respectively. The null model Brier score was 0.10. The lowest Brier scores were achieved using all notes (0.10) or physical therapy notes (0.10).

Relative variable importance showed that the most important features for the discharge summary notes were: "ipratropium," "type 2," "inci_clean," "ml_mouth," "known_ultram," "week," "pulm," "l4_5," "influenza_quadriv," "4_mg' (Fig. 1). The most important features for the operative notes were: "l3_5," "s1_foramin," "10_ml," "scar," "local," "l5_spinal." The most important features for the nursing notes were: "will_continu," "wheez," "walker," "thought," "issu," "notifi." The most important features for the physical therapy notes were: "will," "chair," "abil," "transfer," "home_without," "follow_1," "can_use," "cant," "bed." The most important features for the case management notes were: "rehab," "notifi," "will," "clear," "care_plan." The most important features for the MD/APP notes were: "metformin," "necessari," "pain_control," "void_trial."

Discussion

The increasing volume of free-text notes available in EHRs has led to an opportunity for application of NLP algorithms to detect and predict adverse outcomes. In this study, NLP algorithms were applied to free-text notes generated during the index hospitalization for lumbar spine fusion to identify which free-text notes had the greatest utility for prediction of unplanned readmission in this cohort. The AUROC of NLP algorithms predicting 90-day readmission was highest for discharge summary notes (0.70), with moderate AUROC values for physical therapy (0.60), case management (0.60), operative (0.57), and nursing (0.57) notes. On the other hand, MD/APP notes were the least useful for predicting 90-day readmission (0.49). Thus, the predictive utility of NLP algorithms applied to free-text documentation in EHRs is dependent on both the author and type of EHR documentation. In addition to rigorous quality improvement investigations utilizing NLP algorithm applications, future examination of NLP algorithms incorporating other medical information available in EHRs (ie, laboratory values, radioimaging, etc.) may identify further value of NLP applications in quality and care optimization for patients undergoing elective lumbar spinal fusion (Fig. 2).

The 90-day readmission rate in this study (11.7%, n=83), is within previously reported 90-day readmission rates ranging from 4.2 to 24.8% [4,9]. Risk factors for readmission following posterior lumbar spinal fusion have included demographic, socioeconomic, preoperative, and perioperative factors [3,4,8–14]. Readmission after elective spine surgery has been associated with total charges of approximately $7,300 [15], notwithstanding the unquantifiable toll on individual patient's overall biopsychosocial health [16]. As such, accurate and efficient identification of risk for readmission can optimize both quality and cost of care for patients undergoing elective spine surgery.

Prior investigations of readmission risk following elective posterior lumbar spinal fusion have examined categorized EHR data either in national databases codified by trained staff or institutional patient cohorts [3,8,13,14]. Inherent limitations of these methods include the requirement of resources for EHR data entry as well as codified EHR data suitably formatted for traditional statistical models. Codification of EHR data is achieved either by providers directly entering patient data into predefined EHR categories (ie, activity status, smoking status, problem lists, etc.), manual chart review, or data entry into specified databases by trained clinical reviewers (such as in the Healthcare Cost and Utilization Project). On the other hand, NLP algorithm application to free-text documentation offers the

Table
Predictive utility of free-text notes generated during the index inpatient hospitalization assessed in the testing set, n=141

Notes	AUROC	AUPRC	Calibration intercept	Calibration slope	Brier
Discharge	0.70 (0.53, 0.83)	0.23 (0.10, 0.44)	2.42 (1.82, 3.02)	0.44 (0.11, 0.76)	0.11 (0.07, 0.17)
Operative	0.57 (0.43, 0.71)	0.14 (0.07, 0.31)	3.33 (2.74, 3.93)	0.13 (-0.20, 0.47)	0.11 (0.07, 0.18)
Nursing	0.57 (0.41, 0.71)	0.14 (0.07, 0.32)	1.86 (1.22, 2.50)	0.12 (-0.13, 0.37)	0.11 (0.07, 0.17)
Physical therapy	0.60 (0.43, 0.75)	0.24 (0.10, 0.52)	1.72 (1.13, 2.31)	0.27 (-0.09, 0.62)	0.10 (0.06, 0.16)
Case management	0.60 (0.44, 0.75)	0.15 (0.08, 0.27)	2.26 (1.62, 2.90)	0.17 (-0.11, 0.44)	0.11 (0.07, 0.18)
MD or APP	0.49 (0.32, 0.66)	0.13 (0.05, 0.28)	1.87 (1.20, 2.54)	0.01 (-0.23, 0.25)	0.11 (0.07, 0.17)
All notes	0.70 (0.52, 0.84)	0.30 (0.13, 0.55)	1.54 (0.87, 2.21)	0.34 (0.09, 0.59)	0.10 (0.06, 0.15)

APP, advanced practice provider; AUPRC, area under the precision recall curve; AUROC, area under the receiver operating curve; MD, doctor of medicine.

Null model Brier score=0.10.

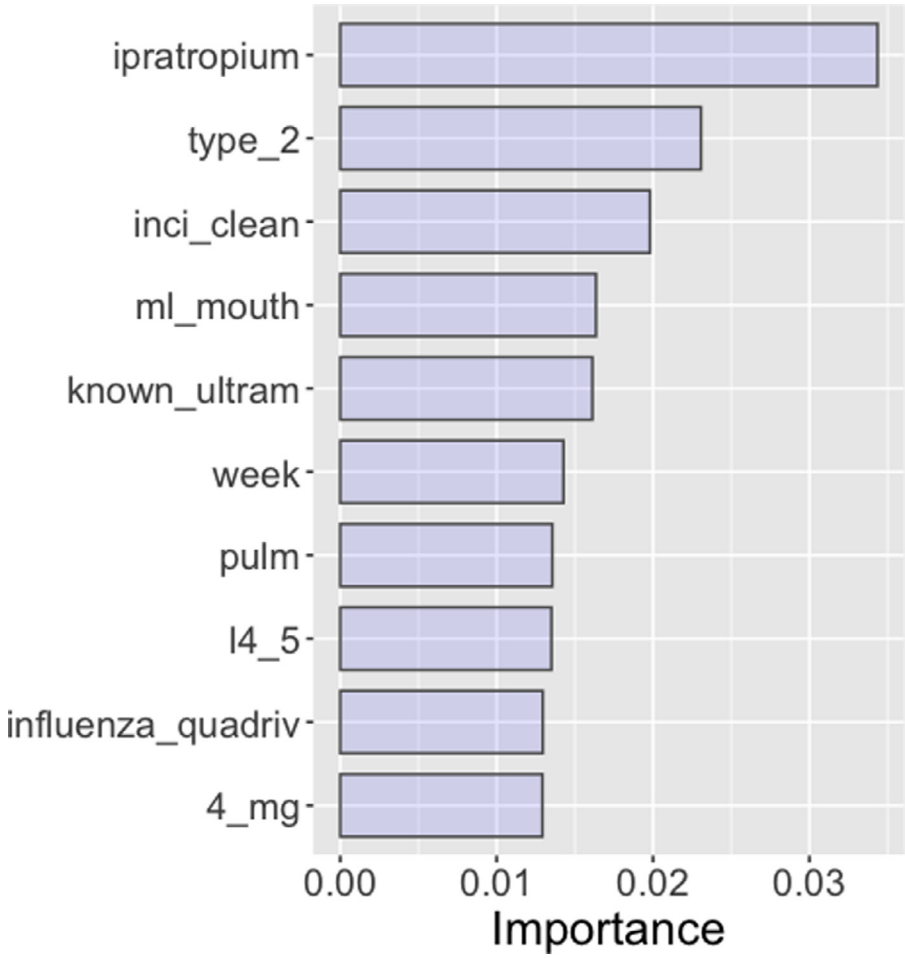

Fig. 1. Global variable importance plots for discharge notes.

advantage of minimizing resources required for model development and deployment while maximizing model consideration of EHR information already available for legal and billing purposes.

Clinical intuition would suggest consideration of both pre and perioperative factors to provide the most comprehensive evaluation of a patient's risk for readmission. Unsurprisingly, prior investigations seeking to identify the relative importance of demographic, preoperative, or perioperative factors in risk of readmission have shown mixed results [3,8,13,14,17]. Similar to clinical decisions, the present investigation applied NLP algorithms to a variety of free-text documentations encompassing both pre- and perioperative clinical data. Discharge, physical therapy, and case management notes demonstrated the most utility for prediction of 90-day inpatient readmission whereas daily MD/APP progress notes demonstrated the least utility. Consistent with prior literature [8,9,14,18], highly weighted key words in NLP model prediction of 90-day readmission included comorbidities such as type 2 diabetes ("type_2") and home without assistance ("home_without").

NLP algorithms in the present investigation were developed utilizing a machine learning approach. Supervised machine learning development of NLP algorithms does not require explicit programming of linguistic structures that remain static but rather recruits the advantage of adaptive rules during the training process. As such, maintenance and external applicability of machine learning–based NLP algorithms is more flexible and less prone to paradoxical static rules in complex algorithms. On the other hand, machine learning–based NLP algorithms require a sufficient volume of data for training to extract clinically relevant features. In the context of the present investigation, the machine learning –based NLP approach identified several clinically relevant features such as a clean incision ("inci_clean") and home without assistance ("home_without"). However, lack of explicit programming in machine learning approaches to NLP also allows for consideration of less clinically relevant features, evident in model selection of features such as date of service ("date_servic"). Overall, rule-based approaches to NLP remain relatively common due to interpretability and ease of troubleshooting, however machine learning and hybrid approaches are more likely to lead to development of more readily adaptable models [19].

There are several limitations that must be acknowledged to appropriately interpret results of the present investigation.

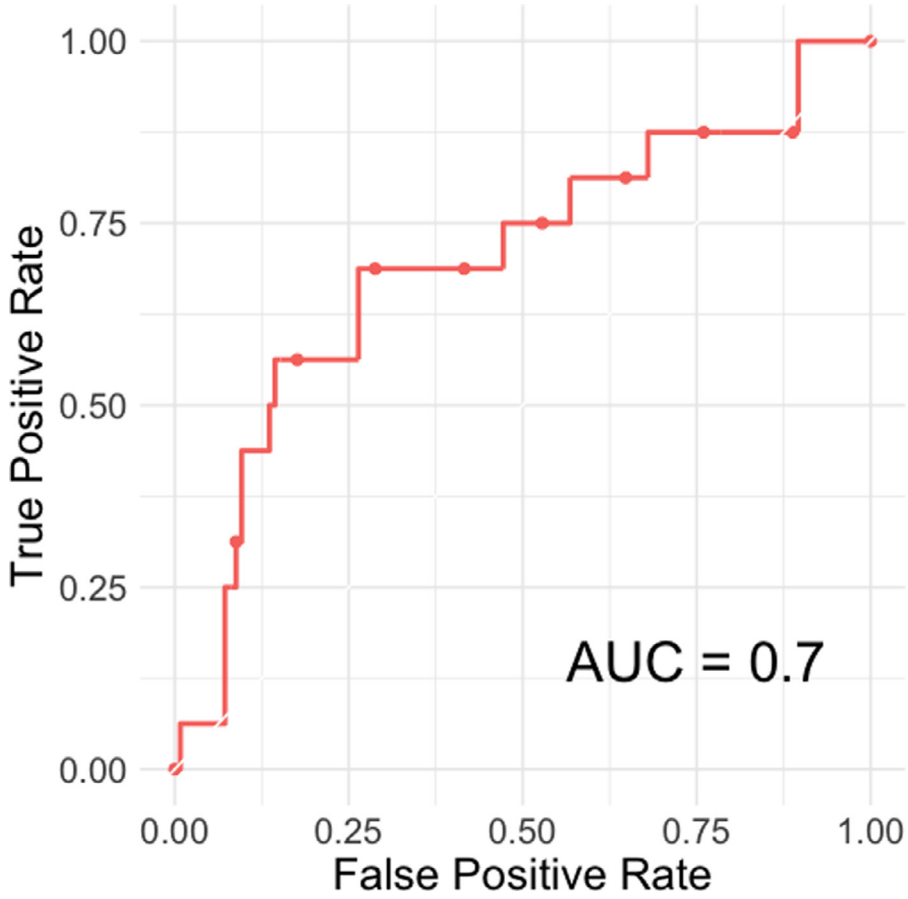

Fig. 2. Area under the receiver operating curve for NLP algorithms using discharge summary notes for prediction of 90-day readmission, n=141.

Firstly, this study was designed as a retrospective analysis of patients receiving surgical care within a single healthcare entity with a purpose of exploring the internal validity of NLP applications to EHR free-text documentation. Thus, the utility of NLP algorithms may not be applicable in alternate practice settings and algorithms developed herein have not been externally validated for multi-institutional application. Second, a supervised approach was utilized for development of NLP algorithms involving manual identification of patients readmitted to the single healthcare entity within 90 postoperative days. Patients seeking care outside of the study's healthcare entity may thus not have been accurately captured within the dataset, possibly contributing to under-reporting of readmission rates. Finally, 90-day readmission risk prediction models were limited to documentation-specific NLP models (ie, operative note, physical therapy note, etc.) but is more likely a complex interaction among several risk factors. Compilation of NLP applications to free-text documentation, patient biopsychosocial factors, laboratory, and radioimaging data is likely to confer a superior predictive performance and is an interesting future direction for machine learning–assisted quality improvement. The sample size of this study prevented the stratification of the study population by clinical variables such as advanced age that impact risk of readmission. Further studies with larger populations are needed to investigate the results of this initial exploratory study in subgroups such as patients over the age of 65. The focus of this study was documentation generated during the index hospitalization alone; as a result, free-text documentation such as clinic consult notes and preoperative anesthesia clearance notes were not assessed. Further studies with more comprehensive assessment of free-text notes at multiple time points will extend the work presented here. Finally, this study was an exploratory analysis that focused on a single variable—the type of free-text documentation during the index hospitalization. More comprehensive assessment of NLP data preprocessing steps, feature selection, free-text representation methods, population sample size, and types of supervised ML algorithms are likely to yield greater insights that will aid healthcare systems in building the most optimal algorithms for predicting adverse events in spinal disorders.

Despite these limitations, the exploratory application of NLP algorithms to free-text documentation in the present study demonstrates a promising avenue for optimized detection of patients at risk for 90-day readmission following posterior lumbar spinal fusion. Incorporation of future NLP applications into HRRP policy decisions may prove useful to minimize the risk of selection bias against surgical

candidates with higher medical complexity and associated risk of HRRP penalizations. Current systems for minimization of 90-day readmission risk following elective spine surgery rely on resource intensive quality improvement initiatives, dedicated clinical documentation, or retrospective query. On the other hand, a rigorously validated NLP algorithm has the potential to dynamically optimize quality and cost of care in future informatics-assisted spinal healthcare.

Conclusion

In this exploratory analysis, discharge summary, physical therapy, and case management notes had the most utility and daily MD/APP progress notes had the least utility for prediction of 90-day inpatient readmission in lumbar fusion patients among the free-text documentation generated during the index hospitalization.

Ethics statement

This study was approved by our institutional review board.

Declarations of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors report no funding disclosures for this study

References

[1] Hospital Readmissions Reduction Program. 2020; Available at: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/HRRP/Hospital-Readmission-Reduction-Program.

[2] Ibrahim AM, Dimick JB. A decade later, lessons learned from the hospital readmissions reduction program. JAMA network open. 2019;2(5):e194594-e.

[3] Ilyas H, Golubovsky JL, Chen J, Winkelman RD, Mroz TE, Steinmetz MP. Risk factors for 90-day reoperation and readmission after lumbar surgery for lumbar spinal stenosis. Journal of Neurosurgery: Spine. 2019;31(1):20-6.

[4] Baaj AA, Lang G, Hsu W-C, Avila MJ, Mao J, Sedrakyan A. 90-day readmission after lumbar spinal fusion surgery in new york state between 2005 and 2014. Spine. 2017;42(22):1706-16.

[5] Karhade AV, Bongers ME, Groot OQ, Kazarian ER, Cha TD, Fogel HA, et al. Natural language processing for automated detection of incidental durotomy. The Spine Journal. 2020;20(5):695-700.

[6] Karhade AV, Bongers ME, Groot OQ, Cha TD, Doorly TP, Fogel HA, et al. Can natural language processing provide accurate, automated reporting of wound infection requiring reoperation after lumbar discectomy? The Spine Journal. 2020;20(10):1602-9.

[7] Karhade AV, Bongers ME, Groot OQ, Cha TD, Doorly TP, Fogel HA, et al. Development of machine learning and natural language processing algorithms for preoperative prediction and automated identification of intraoperative vascular injury in anterior lumbar spine surgery. The Spine Journal. 2020.

[8] Martini ML, Neifert SN, Oermann EK, Gal J, Rajan K, Nistal DA, et al. Machine learning with feature domains elucidates candidate drivers of hospital readmission following spine surgery in a large single-center patient cohort. Neurosurgery. 2020;87(4):E500-E10.

[9] Bernatz JT, Anderson PA. Thirty-day readmission rates in spine surgery: Systematic review and meta-analysis. Neurosurgical focus. 2015;39(4):E7.

[10] Kurian SJ, Yolcu YU, Zreik J, Alvi MA, Freedman BA, Bydon M. Institutional databases may underestimate the risk factors for 30-day unplanned readmissions compared to national databases. J Neurosurg Spine 2020;33(6):845–53.

[11] Malpani R, Gala RJ, Adrados M. Galivanche AR, Clark MG, Mercier MR, et al. High, as well as low, preoperative platelet counts correlate with adverse outcomes after elective posterior lumbar surgery. Spine (Phila Pa 1976) 2020;45(5):349–56.

[12] Phan K, Ranson W, White SJ. Cheung ZB, Kim J, Shin JI, et al. Thirty-day perioperative complications, prolonged length of stay, and readmission following elective posterior lumbar fusion associated with poor nutritional status. Glob Spine J 2019;9(4):417–23.

[13] Ranti D, Mikhail CM, Ranson W. Cho B, Warburton A, Rutland JW, et al. Risk factors for 90-day readmissions with fluid and electrolyte disorders following posterior lumbar fusion. Spine (Phila Pa 1976) 2020;45(12):E704.. e12.

[14] Rubel NC, Chung AS, Wong M, Lara NJ, Makovicka JL, Arvind V, et al. 90-day readmission in elective primary lumbar spine surgery in the inpatient setting: A nationwide readmissions database sample analysis. Spine. 2019;44(14):E857-E64.

[15] Wiley MR, Carreon LY, Djurasovic M, Glassman SD, Khalil YH, Kannapel M, et al. Economic analysis of 90-day return to the emergency room and readmission after elective lumbar spine surgery: A single-center analysis of 5444 patients. Journal of Neurosurgery: Spine. 2020;34(1):89-95.

[16] Bekeris J, Wilson LA, Fiasconaro M. Poeran J, Liu J, Girardi F, et al. New onset depression and anxiety after spinal fusion surgery: incidence and risk factors. Spine (Phila Pa 1976) 2020;45(16):1161–9.

[17] Hopkins BS, Yamaguchi JT, Garcia R. Kesavabhotla K, Weiss H, Hsu WK, et al. Using machine learning to predict 30-day readmissions after posterior lumbar fusion: an NSQIP study involving 23,264 patients. J Neurosurg Spine 2020;32(3):399–406.

[18] Jain D, Singh P, Kardile M, Berven SH. A validated preoperative score for predicting 30-day readmission after 1–2 level elective posterior lumbar fusion. Eur Spine J 2019;28(7):1690–6.

[19] Kreimeyer K, Foster M, Pandey A. Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. J Biomed Inform 2017;73:14–29.