European Spine Journal
https://doi.org/10.1007/s00586-023-07552-4

ORIGINAL ARTICLE

Developing a triage predictive model for access to a spinal surgeon using clinical variables and natural language processing of radiology reports

Brandon Krebs[1] · Andrew Nataraj[2] · Erin McCabe[1] · Shannon Clark[3] · Zahin Sufiyan[3] · Shelby S. Yamamoto[4] · Osmar Zaïane[3] · Douglas P. Gross[5]

Received: 12 October 2022 / Revised: 17 January 2023 / Accepted: 22 January 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract

Purpose To utilize natural language processing (NLP) of MRI reports and various clinical variables to develop a preliminary model predictive of the need for surgery in patients with low back and neck pain. Such a model would be beneficial for informing clinical practice decisions and help reduce the number of unnecessary surgical referrals, streamlining the surgical process.

Methods A historical cohort study was conducted using de-identified data from patients referred to a spine assessment clinic. Various demographic, clinical, and radiological variables were included as potential predictors. Full-text radiology reports of patients' MRI findings were vectorized using NLP before applying machine learning algorithms to develop models predicting who underwent surgery. Outputs from these models were then entered into a logistic regression model with clinical variables to develop a preliminary model predictive of surgical recommendations.

Results Of the 398 patients assessed, 71 underwent spine surgery. NLP variables were significant predictors in univariate analysis but did not remain in the final logistic regression model. An outcome of receiving surgery was predicted by a primary symptom of low back and leg pain (adjusted odds ratio 2.81), distal pain indicated by a pain diagram (adjusted odds ratio 2.49) and self-reported difficulties walking (adjusted odds ratio 2.73).

Conclusion A logistic regression model was created to predict which patients may require spine surgery. Simple clinical variables appeared more predictive than variables created using NLP. However, additional research with more data samples is needed to validate this model and fully evaluate the usefulness of NLP for this task.

Keywords Surgical outcome · Predictive factors · Back and neck pain · Spinal surgery

Introduction

Low back and neck pain are common health conditions and major causes of disability that affect quality of life and participation in daily activities [1, 2]. Many people are referred to spine surgeons to address their spinal conditions; however, only a small proportion of individuals are surgical candidates [3]. A large volume of non-surgical referrals to surgeons burdens the healthcare system and deprives deserving patients of timely access to surgical decision-making.

Spinal surgery triage clinics are one solution to this issue, where patient referrals to spine surgeons are reviewed by another healthcare provider [4]. The healthcare provider triages patients, directing them toward the most appropriate management for their symptoms (e.g., multidisciplinary pain clinic, physiotherapy or other conservative care options, or consultation with a spine surgeon) [5]. This reduces the

✉ Douglas P. Gross
dgross@ualberta.ca

1 Faculty of Rehabilitation Medicine, University of Alberta, Edmonton, Canada

2 Department of Surgery, University of Alberta, Edmonton, Canada

3 Department of Computing Science, University of Alberta, Edmonton, Canada

4 School of Public Health, University of Alberta, Edmonton, Canada

5 Department of Physical Therapy, University of Alberta, 2-50 Corbett Hall, Alberta, Edmonton T6G 2G4, Canada

Published online: 06 February 2023

number of patients needed to be seen by the surgeon but often still involves a long waitlist for patients. If patients who are highly likely to require surgery could be identified based on self-reported patient characteristics, clinical features or diagnostic imaging reports, the triage process could be further optimized.

Predictive models based on patient clinical and psychosocial characteristics exist for predicting spine surgery outcomes [6–9]; however, fewer predictive models have been developed using self-report patient characteristics and/or diagnostic imaging textual reports for triaging potential surgery candidates. One earlier attempt to improve the decision-making and selection of patients for lumbar fusion surgery was unsuccessful [10]. Willems concluded that "*currently used tests do not improve the results of fusion by better patient selection, these tests should not be recommended for surgical decision making in standard care*."[10] Since then, advancements in computer technology and machine learning methods have opened the door for more sophisticated modeling. [11] Accuracy has proven to be fair to good [12, 13], with the most promising models incorporating data from diagnostic imaging (area under the curve = 0.88) [14].

In the current study, we examined patients referred to a spine assessment clinic in Edmonton, Alberta, to determine whether we could develop a preliminary model for predicting the need for surgical intervention. We used a combination of traditional regression analyses and more advanced machine learning techniques to construct this model. Particularly, the current study aimed to use both natural language processing (NLP) of textual reports of patients' magnetic resonance imaging (MRI) and logistic regression analysis to develop a preliminary model predictive of surgical decision-making outcomes. This study aimed to address a gap in the current literature, as limited research has utilized NLP from textual reports of patients' MRIs to predict surgical decision-making outcomes. However, due to the sample size and exploratory nature of this study, the resulting model will need further validation before it can be readily applied in clinical settings. We hypothesized that we would be able to develop a model that would be moderately accurate for identifying who requires spine surgery. Specifically, we hypothesized that based on patient-reported symptoms and MRI reports, a model using both logistic regression and NLP could be developed to predict patients who would be offered surgery. We also hypothesized that predictive model accuracy would be substantially improved (R^2 improvement of > 0.2) by adding NLP of MRI textual reports to the preliminary logistic regression model.

Methods

Design

A historical cohort study was conducted, with predictor variables collected prior to the surgical decision. Historical cohort designs are a form of observational research allowing the investigation of exposures (i.e., predictive factors) and outcomes, such as surgical outcomes, of participants [15]. Observational studies can be limited by their inability to derive causal results; however, they are particularly useful for testing associations between predictor variables and an outcome. This research was approved by the University of Alberta's Health Research Ethics Board.

Study sample

Data were collected from the electronic medical record of patients presenting to the Kaye Edmonton Spine Assessment Clinic between November 1, 2019, and June 30, 2021. Patients with a wide variety of spinal conditions are referred to the clinic. Referrals include a spinal MRI, a description of the patients' primary symptoms and a self-report questionnaire. A registered nurse reviews these documents and does an assessment with the patient, either through telehealth or in-person. Patients who could possibly benefit from surgery are then seen by a spine surgeon who makes the final decision with the patient about surgery. We obtained textual MRI reports from radiologists for 763 patients referred to the spine assessment clinic, 133 (17.4%) of whom were assessed as surgical patients, and 630 (82.6%) were not surgical (see Fig. 1). Further, we obtained patient self-report data from a clinical assessment questionnaire completed by 398 of these patients prior to being seen at the spine assessment clinic. Of the 398 patients, 71 (17.8%) were surgical, and 327 (82.2%) were not. The predictive model was based on data from these 398 patients (see Fig. 1).

Measures: independent variables

The dataset included various descriptive variables extracted from the patient self-assessment questionnaire, including demographic factors (e.g., age, sex, height, weight), health/condition-related factors (e.g., primary symptoms, pain ratings, symptom duration, mechanism of

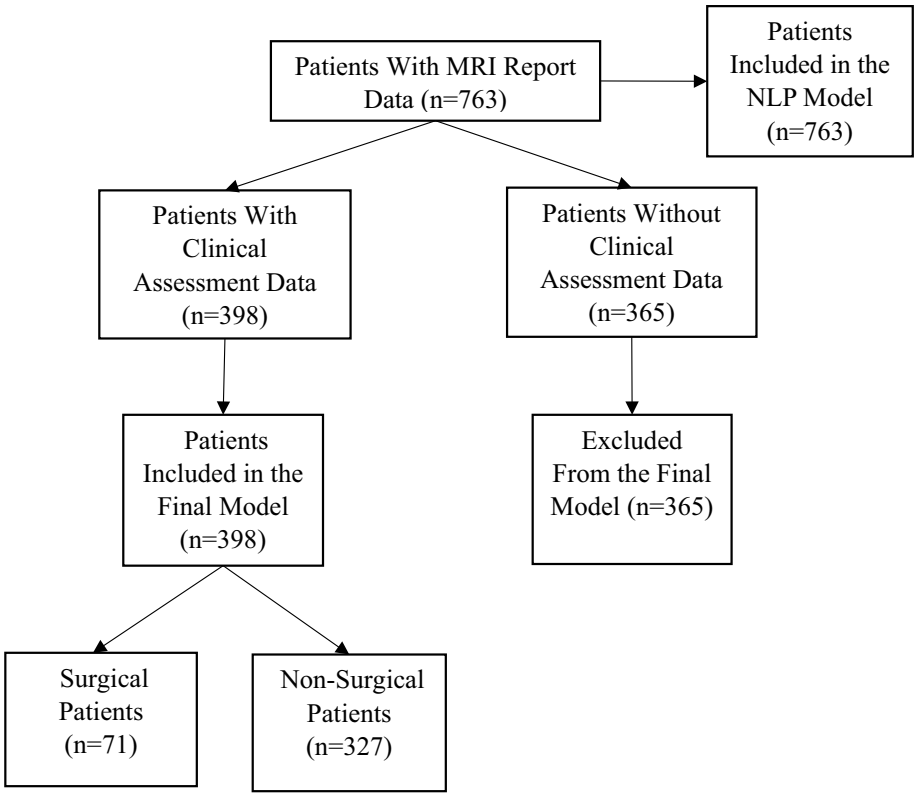

Fig. 1 Participant flowchart

injury, pain location, walking difficulties, previous spinal surgery, attending allied healthcare professionals, medications used, injury impact, past conditions), a measure of disability (i.e., Oswestry Disability Index (ODI)) [16], and a global health rating (numerical rating scale 0–100).

Machine learning applications

We explored predictive natural language processing (NLP), a form of machine learning, to predict outcomes from radiologists' textual report of patients' MRI findings [17]. Machine learning is concerned with the design of algorithms to discriminate between classes (outcomes) based on empirical data [18]. NLP is a subfield of machine learning that provides a solution for analysis of narrative electronic data, such as that usually found in radiology reports, rather than discrete data points [19]. In NLP, textual data are converted to structured (discrete) data for further analysis.

Using the larger dataset of 763 participants with MRI and outcome data, we used NLP to create a model capable of reviewing individual patient radiology reports, then making a recommendation of whether that patient requires spine surgery. The text from the reports was processed (removed stop words, conversion of all characters to lowercase, and exploration of three different methods for text vectorization: bag-of-words, term frequency-inverse document frequency (TF-IDF), and word2vec text vectorization) [20]. The terms severe central stenosis, severe spinal cord compression, or large, extruded disc fragment were prioritized as being indicative of a patient requiring surgery. Next, machine learning classifiers were used to generate models classifying cases as surgical or not surgical. As it proved challenging to develop a model with balanced performance due to class imbalance in the dataset, three models (an ensemble of logistic regression and support vector machine with TF-IDF text vectorization, logistic regression with word2vec text vectorization, and logistic regression with TF-IDF text vectorization) were created prioritizing specificity, recall, and precision, respectively. Each of these models was used to generate a variable that was then entered into the final logistic regression model to explore their utility in the final prediction task.

Outcome (dependent) variable

The outcome variable for this study was the final clinical decision made by the neurosurgeon who saw the patient at the spine assessment clinic (e.g., surgery or no surgery). The surgical assessments conducted at the spine assessment clinic were completed by one of five experienced neurosurgeons. All five neurosurgeons hold the qualification of fellow of the Royal College of Surgeons of Canada (in neurosurgery), a designation recognized by the health community as proof of world-class expertise.

Data analysis

Descriptive statistics for the 398 patients with a patient self-assessment questionnaire were conducted. We compared the 71 surgical patients to the 327 non-surgical patients using appropriate statistical tests (i.e., *t* tests and Chi-square tests). Logistic regression was used to examine associations between all potential predictor variables and the surgical recommendation. This included calculating univariate odds ratios for the 763 patients with NLP models of MRI textual report results.

To address our first hypothesis, we used logistic regression to identify factors predictive of surgery. We used a purposeful modeling strategy [21], where independent variables identified as significantly different at $p < 0.10$ between surgical and non-surgical cases in the descriptive comparison were sequentially entered into a multivariable regression model. Besides key demographic variables such as age and sex, variables that were at $p > 0.10$ were removed to form the preliminary logistic regression analysis. To test our second hypothesis, the most predictive NLP variable identified through univariate logistic regression analysis was forced into the preliminary logistic regression model. The final multivariable logistic regression model included all variables significantly associated with the outcome at $p < 0.05$. Change in Nagelkerke R^2 value was examined. The relevant assumptions for multivariable regression analysis (e.g., multicollinearity, linearity of independent variables) were tested. [22] All analyses were conducted with IBM SPSS v28.

Results

Patient characteristics

Patient demographic and clinical characteristics according to whether they were recommended for surgery following their visit to the spine assessment clinic ($n = 71$) or were deemed non-surgical ($n = 327$) are shown in Table 1. Surgical patients were more likely to report low back combined with leg pain (78.9% vs. 47.1%, $p < 0.001$), difficulties walking (63.4% vs. 36.1%, $p < 0.001$), had previous X-ray imaging completed (73.2% vs. 60.0%,$p = 0.04$), and used daily narcotic (35.2% vs. 16.8%, $p = 0.001$) and neuroleptic (38.0% vs. 19.3%, $p < 0.001$) medication. When asked to allocate the proportion of pain experienced in either their spine or extremity, surgical patients were more likely to report a proportion of back pain of less than 60% (62.0% vs. 33.3%, $p < 0.001$) and a proportion of leg pain greater than 60% (46.5% vs. 26.3%, $p = 0.05$). Similarly, on the bodily pain diagram, surgical patients were more likely to represent their pain as radiating distally into their extremities (66.2% vs. 37.0%, $p < 0.001$). Total ODI percentage (50.2% vs. 44.1%, p 0.01) and the ODI subscores of pain intensity (3.25 vs. 2.78, $p = 0.004$), personal care (1.49 vs. 1.24, p 0.04), walking (2.25 vs. 1.64, $p < 0.001$), standing (2.85 vs. 2.41, $p = 0.01$), sex (2.67 vs. 2.33, $p = 0.02$), and social (2.92 vs. 2.48, $p = 0.01$) were higher in surgical patients.

There were missing data on the self-reported measures, with 107 (26.9%) patients not completing at least one of the questionnaire measures. Patients with missing data on the self-reported measures were significantly less likely to have a primary symptom report of low back and leg pain (26.0% versus 62.9%, $p < 0.001$), less likely to report difficulty walking (17.8% versus 49.5%, $p < 0.001$), and less likely to report a secondary symptom of numbness in the arms or legs (34.6% versus 47.4%, $p < 0.001$). They were also less likely to be recommended for surgery (10.3% versus 20.6%, $p = 0.02$).

NLP variables

The univariate odds ratios (OR) and 95% confidence intervals (CI) predicting surgery for the NLP variables are shown in Table 2. The model prioritizing specificity (OR 1.56, 95% CI 0.88, 2.79) was not significantly associated with surgery. In contrast, the output from the models prioritizing recall (OR 1.68, 95% CI 1.15, 2.44) and precision (OR 1.97, 95% CI 1.33, 2.93) were significantly associated with the surgical recommendation.

Factors predictive of surgery

Univariate associations, preliminary, and final regression models displaying crude and adjusted ORs (95% CIs) predicting a surgical recommendation are shown in Table 3. Various health/injury-related variables were found to be significantly associated with a surgical recommendation in the final regression model, with a Nagelkerke R^2 of 0.2. In the final model, patients had higher odds of being in the surgical group if they reported low back and leg pain (AOR 2.81, 95% CI 1.21, 6.52), distal pain (AOR 2.49, 95% CI 1.22, 5.08), and difficulties walking (AOR 2.73, 95% CI 1.39, 5.37). Age (AOR 1.01, 95% CI 0.99–1.03), female sex (AOR 0.78, 95% CI 0.40–1.51), the proportion of back pain (AOR 0.54, 95% CI 0.28, 1.05), and the NLP precision variable (AOR 0.84, 95% CI 0.42, 1.69) did not enter the final model predicting surgical recommendation.

Discussion

Our initial hypothesis was partly supported as numerous variables were associated with a need for surgical intervention, and a moderately predictive logistic regression model was produced. However, in contrast to our initial

Table 1 Descriptive characteristics of patients assessed in the spine assessment clinic ($n=398$)

Variables	Full sample ($n=398$)	Surgery ($n=71$)	No surgery ($n=327$)	p value
Age	55.5 (± 14.8)	57.9 (± 15.6)	55.0 (± 14.6)	0.13
Sex				0.71
Male	222 (100%)	41 (18.5%)	181 (81.5%)	
Female	176 (100%)	30 (17.0%)	146 (83.0%)	
Height in cm ($n=341$)	168.98 (± 15.14)	169.63 (± 13.02)	168.83 (± 15.01)	0.70
Weight in Kg ($n=343$)	85.60 (± 21.00)	84.19 (± 20.38)	85.90 (± 21.16)	0.57
Smoke				0.28
Yes	80 (100%)	11 (13.8%)	69 (86.2%)	
No	225 (100%)	43 (19.1%)	182 (81.9%)	
Missing	93 (100%)	17 (18.2%)	76 (81.8%)	
Referral primary symptom				< 0.001*
Neck pain with arm pain	31 (100%)	2 (6.5%)	29 (93.5%)	
Low back pain with leg pain	210 (100%)	56 (26.7%)	154 (73.3%)	
Other	154 (100%)	12 (7.8%)	142 (92.2%)	
Difficulty walking				< 0.001*
Yes	163 (100%)	45 (27.6%)	118 (72.4%)	
No	235 (100%)	26 (11.1%)	209 (88.9%)	
Missing	0	0	0	
Primary pain rating ($n=181$)				0.35
Low–moderate pain (1–6)	65 (100%)	9 (13.8%)	56 (86.2%)	
Severe pain (7–10)	326 (100%)	61 (18.7%)	265 (81.3%)	
Missing	7 (100%)	1 (14.3%)	6 (85.7%)	
Referral secondary symptom				
Neck pain				0.14
Yes	75 (100%)	9 (12.0%)	66 (88.0%)	
No	323 (100%)	62 (19.2%)	261 (80.8%)	
Missing	0	0	0	
Neck pain with arm symptoms				0.28
Yes	49 (100%)	6 (12.2%)	43 (87.8%)	
No	349 (100%)	65 (18.6%)	284 (81.4%)	
Missing	0	0	0	
Mid back pain				0.12
Yes	70 (100%)	8 (11.4%)	62 (88.6%)	
No	328 (100%)	63 (19.2%)	265 (80.8%)	
Missing	0	0	0	
Low back pain				0.70
Yes	149 (100%)	28 (18.8%)	121 (81.2%)	
No	249 (100%)	43 (17.3%)	206 (82.7%)	
Missing	0	0	0	
Low back pain with leg symptoms				0.18
Yes	157 (100%)	33 (21.0%)	124 (79.0%)	
No	241 (100%)	38 (15.8%)	203 (84.2%)	
Missing	0	0	0	
Arm pain only				0.29
Yes	5 (100%)	0	5 (100%)	
No	393 (100%)	71 (18.1%)	322 (81.9%)	
Missing	0	0	0	
Leg pain only				0.82
Yes	15 (100%)	3 (20.0%)	12 (80.0%)	

Table 1 (continued)

Variables	Full sample ($n=398$)	Surgery ($n=71$)	No surgery ($n=327$)	p value
No	383 (100%)	68 (17.8%)	315 (82.2%)	
Missing	0	0	0	
Numbness/tingling in arms/legs				0.32
Yes	175 (100%)	35 (20.0%)	140 (80.0%)	
No	223 (100%)	36 (16.1%)	187 (83.9%)	
Missing	0	0	0	
Weakness in arms/legs				0.15
Yes	128 (100%)	28 (21.9%)	100 (78.1%)	
No	270 (100%)	43 (15.9%)	227 (84.1%)	
Missing	0	0	0	
Clumsiness of hands				0.37
Yes	46 (100%)	6 (13.0%)	40 (87.0%)	
No	352 (100%)	65 (18.5%)	287 (81.5%)	
Missing	0	0	0	
Balance difficulties				0.19
Yes	94 (100%)	21 (22.3%)	73 (77.7%)	
No	304 (100%)	50 (16.4%)	254 (83.6%)	
Missing	0	0	0	
Other				1.00
Yes	56 (100%)	10 (17.9%)	46 (82.1%)	
No	342 (100%)	61 (17.8%)	281 (82.2%)	
Missing	0	0	0	
Secondary pain rating ($n=167$)				0.60
Low–moderate pain (1–6)	97 (100%)	20 (20.6%)	77 (79.4%)	
Severe pain (7–10)	248 (100%)	45 (18.1%)	203 (81.9%)	
Missing	53 (100%)	6 (11.3%)	47 (88.7%)	
Primary symptom duration				0.40
12 Weeks or less	30 (100%)	7 (23.3%)	23 (76.7%)	
3–12 months	126 (100%)	26 (20.6%)	100 (79.4%)	
1 Years plus	238 (100%)	38 (16.0%)	200 (84.0%)	
Missing	4 (100%)	0	4 (100%)	
Mechanism of injury				0.57
Trauma/injury	49 (100%)	8 (16.3%)	41 (83.7%)	
Work-related injury	34 (100%)	5 (14.7%)	29 (85.3%)	
Fall	23 (100%)	7 (30.4%)	16 (69.6%)	
Vehicle collision	23 (100%)	4 (17.4%)	19 (82.6%)	
Unknown cause	211 (100%)	40 (19.0%)	171 (81.0%)	
Other	48 (100%)	6 (12.5%)	42 (87.5%)	
Missing	10 (100%)	1 (10.0%)	9 (90.0%)	
Claims				0.91
No claims	338 (100%)	60 (17.8%)	278 (82.2%)	
Yes claims present	60 (100%)	11 (18.3%)	49 (81.7%)	
Missing	0	0	0	
Change in symptoms				0.70
Yes	353 (100%)	63 (17.8%)	290 (82.2%)	
No	39 (100%)	6 (15.4%)	33 (84.6%)	
Missing	6 (100%)	2 (33.3%)	4 (66.7%)	
Reported symptom Change				0.16
Worsening	298 (100%)	57 (19.1%)	241 (80.9%)	

Table 1 (continued)

Variables	Full sample ($n=398$)	Surgery ($n=71$)	No surgery ($n=327$)	p value
Improving	40 (100%)	4 (10.0%)	36 (90.0%)	
Missing	60 (100%)	10 (16.7%)	50 (83.3%)	
Similar symptoms in past				0.76
Yes	156 (100%)	27 (17.3%)	129 (82.7%)	
No	210 (100%)	39 (18.6%)	171 (81.4%)	
Missing	32 (100%)	5 (15.2%)	27 (81.8%)	
Past spine surgery				0.97
Yes	44 (100%)	8 (18.2%)	36 (81.8%)	
No	351 (100%)	63 (17.9%)	288 (82.1%)	
Missing	3 (100%)	0	3 (100%)	
Agree to surgery				0.08
Yes	330 (100%)	67 (20.3%)	263 (79.7%)	
No	44 (100%)	4 (9.1%)	40 (90.9%)	
Missing	24 (100%)	0	24 (100%)	
Goals of surgery				0.10
Relief of neck/back pain	174 (100%)	26 (14.9%)	148 (85.1%)	
Relief of arm/leg pain	117 (100%)	29 (24.8%)	88 (75.2%)	
N/A	1 (100%)	0	1 (100%)	
Missing	106 (100%)	16 (15.1%)	90 (84.9%)	
Reported proportion of neck pain				0.24
59 or less	63 (100%)	5 (7.9%)	58 (92.1%)	
60 or more	69 (100%)	10 (14.5%)	59 (85.5%)	
Missing	266 (100%)	56 (21.0%)	210 (79.0%)	
Reported proportion of arm pain				0.11
59 or less	77 (100%)	13 (16.9%)	64 (83.1%)	
60 or more	45 (100%)	3 (6.7%)	42 (93.3%)	
Missing	276 (100%)	55 (19.9%)	221 (80.1%)	
Reported proportion of back pain				<0.001*
59 or less	153 (100%)	44 (28.8%)	109 (71.2%)	
60 or more	156 (100%)	21 (13.5%)	135 (86.5%)	
Missing	89 (100%)	6 (6.7%)	83 (93.3%)	
Reported proportion of leg pain				0.05*
59 or less	183 (100%)	33 (18.0%)	150 (82.0%)	
60 or more	119 (100%)	33 (27.7%)	86 (72.3%)	
Missing	97 (100%)	5 (5.2%)	91 (93.8%)	
Pain diagram				<0.001*
Central	164 (100%)	17 (10.4%)	147 (89.6%)	
Distal	168 (100%)	47 (28.0%)	121 (72.0%)	
Whole body	40 (100%)	3 (7.5%)	37 (92.5%)	
Missing	26 (100%)	4 (15.4%)	22 (84.6%)	
Allied healthcare providers				0.09
Yes	287 (100%)	57 (19.9%)	230 (80.1%)	
No	111 (100%)	14 (12.6%)	97 (87.4%)	
Missing	0	0	0	
Imaging				
X-ray				0.04*
Yes	248 (100%)	52 (21.0%)	196 (79.0%)	
No	150 (100%)	19 (12.7%)	131 (87.3%)	
Missing	0	0	0	

Table 1 (continued)

Variables	Full sample ($n=398$)	Surgery ($n=71$)	No surgery ($n=327$)	p value
CT scan				0.39
Yes	64 (100%)	9 (14.1%)	55 (85.9%)	
No	334 (100%)	62 (18.6%)	272 (81.4%)	
Missing	0	0	0	
MRI				0.20
Yes	346 (100%)	65 (18.8%)	281 (81.2%)	
No	52 (100%)	6 (11.5%)	46 (88.5%)	
Missing	0	0	0	
Bone scan				0.24
Yes	36 (100%)	9 (25.0%)	27 (75.0%)	
No	362 (100%)	62 (17.1%)	300 (82.9%)	
Missing	0	0	0	
Nerve test				0.97
Yes	34 (100%)	6 (17.6%)	28 (82.4%)	
No	218 (100%)	39 (17.9%)	179 (82.1%)	
Missing	146 (100%)	26 (17.8%)	120 (82.2%)	
Spinal injections				0.66
Yes	82 (100%)	16 (19.5%)	66 (80.5%)	
No	316 (100%)	55 (17.4%)	261 (82.6%)	
Missing	0	0	0	
Work impact				0.69
Working normal/restricted hours	152 (100%)	28 (18.4%)	124 (81.6%)	
Not working/on leave due to current condition	179 (100%)	30 (16.8%)	149 (83.2%)	
Missing	67 (100%)	13 (19.4%)	54 (80.6%)	
Impact—able to do most activities				0.08
Yes	88 (100%)	10 (11.4%)	78 (88.6%)	
No	163 (100%)	33 (20.2%)	130 (79.8%)	
Missing	147 (100%)	28 (19.0%)	119 (81.0%)	
Impact—minor difficulty				0.97
Yes	147 (100%)	25 (17.0%)	122 (83.0%)	
No	128 (100%)	22 (17.2%)	106 (82.8%)	
Missing	123 (100%)	24 (19.5%)	99 (80.5%)	
Impact—major difficulty				0.23
Yes	259 (100%)	50 (19.3%)	209 (80.7%)	
No	69 (100%)	9 (13.0%)	60 (87.0%)	
Missing	70 (100%)	12 (17.1%)	58 (82.9%)	
Impact—social impact				0.67
Yes	235 (100%)	45 (19.1%)	190 (80.9%)	
No	99 (100%)	17 (17.2%)	82 (82.8%)	
Missing	64 (100%)	9 (14.1%)	55 (85.9%)	
Psychological impact on safety				0.21
Agree	213 (100%)	42 (19.7%)	171 (80.3%)	
Disagree	163 (100%)	24 (14.7%)	139 (85.3%)	
Missing	22 (100%)	5 (22.7%)	17 (77.3%)	
Psychological impact on worry				0.31
Agree	287 (100%)	49 (17.1%)	238 (82.9%)	
Disagree	97 (100%)	21 (21.6%)	76 (78.4%)	
Missing	14 (100%)	1 (7.1%)	13 (92.9%)	
Psychological impact on Hope				0.11

Table 1 (continued)

Variables	Full sample ($n=398$)	Surgery ($n=71$)	No surgery ($n=327$)	p value
Agree	257 (100%)	52 (20.2%)	205 (79.8%)	
Disagree	120 (100%)	16 (13.3%)	104 (86.7%)	
Missing	21 (100%)	3 (14.3%)	18 (85.7%)	
Psychological impact on enjoyment				0.25
Agree	358 (100%)	68 (19.0%)	290 (81.0%)	
Disagree	29 (100%)	3 (10.3%)	26 (89.7%)	
Missing	11 (100%)	0	11 (100%)	
Medication use				0.16
Yes	356 (100%)	69 (19.4%)	287 (80.6%)	
No	29 (100%)	2 (6.9%)	27 (93.1%)	
Choose not to answer	4 (100%)	0	4 (100%)	
Missing	9 (100%)	0	9 (100%)	
Medication use duration				0.82
Less than 3 months	70 (100%)	16 (22.9%)	54 (77.1%)	
3 months to 1 year	89 (100%)	19 (21.3%)	70 (78.7%)	
Over 1 year	150 (100%)	29 (19.3%)	121 (80.7%)	
Missing	89 (100%)	7 (7.9%)	82 (92.1%)	
Use of over the counter medications				0.14
Never/intermittent	163 (100%)	26 (16.0%)	137 (84.0%)	
Daily	160 (100%)	36 (22.5%)	124 (77.5%)	
Missing	75 (100%)	9 (12.0%)	66 (88.0%)	
Non-steroidal anti-inflammatory Medication				0.27
Never/intermittent	208 (100%)	42 (20.2%)	166 (79.8%)	
Daily	64 (100%)	9 (14.1%)	55 (85.9%)	
Missing	126 (100%)	20 (15.9%)	106 (84.1%)	
Muscle relaxant Medication				0.43
Never/intermittent	220 (100%)	40 (18.2%)	180 (81.8%)	
Daily	57 (100%)	13 (22.8%)	44 (77.2%)	
Missing	121 (100%)	18 (14.9%)	103 (85.1%)	
Narcotic medication				0.001*
Never/intermittent	196 (100%)	28 (14.3%)	168 (85.7%)	
Daily	80 (100%)	25 (31.2%)	55 (68.8%)	
Missing	122 (100%)	18 (14.8%)	104 (85.2%)	
Anti-depressant Medication				0.13
Never/intermittent	190 (100%)	30 (15.8%)	160 (84.2%)	
Daily	76 (100%)	18 (23.7%)	58 (76.3%)	
Missing	132 (100%)	23 (17.4%)	109 (82.6%)	
Neuroleptic medication				< 0.001*
Never/intermittent	185 (100%)	24 (13.0%)	161 (87.0%)	
Daily	90 (100%)	27 (30.0%)	63 (70.0%)	
Missing	123 (100%)	20 (16.3%)	103 (83.7%)	
Past medical history				
Physical conditions				0.14
Yes	232 (100%)	47 (20.3%)	185 (79.7%)	
No	166 (100%)	24 (14.5%)	142 (85.5%)	
Missing	0	0	0	
Psychological Conditions				0.54
Yes	143 (100%)	24 (16.8%)	119 (83.2%)	
No	233 (100%)	45 (19.3%)	188 (80.7%)	

Table 1 (continued)

Variables	Full sample ($n=398$)	Surgery ($n=71$)	No surgery ($n=327$)	p value
Missing	22 (100%)	2 (9.1%)	20 (90.9%)	
Global health rating				0.09
Score of 0–59	187 (100%)	40 (21.4%)	147 (78.6%)	
Score of 60–100	197 (100%)	29 (14.7%)	168 (85.3%)	
Missing	14 (100%)	2 (14.3%)	12 (85.7%)	
Oswestry disability index (ODI) Total percentage ($n=321$)	45.1% (±18.3%)	50.2% (±15.1%)	44.1% (±18.7%)	0.01*
ODI pain intensity ($n=380$)	2.86 (±1.32)	3.25 (±1.06)	2.78 (±1.35)	0.004*
ODI personal care ($n=379$)	1.28 (±1.03)	1.49 (±0.80)	1.24 (±1.07)	0.04*
ODI lifting ($n=376$)	2.98 (±1.35)	3.22 (±1.17)	2.93 (±1.38)	0.054
ODI walking ($n=381$)	1.75 (±1.34)	2.25 (±1.28)	1.64 (±1.33)	<0.001*
ODI sitting ($n=382$)	2.06 (±1.38)	2.06 (±1.33)	2.06 (±1.40)	0.50
ODI standing ($n=383$)	2.49 (±1.46)	2.85 (±1.40)	2.41 (±1.46)	0.01*
ODI sleep ($n=366$)	1.90 (±1.40)	1.89 (±1.28)	1.90 (±1.43)	0.46
ODI sex ($n=367$)	2.39 (±1.18)	2.67 (±1.03)	2.33 (±1.20)	0.02*
ODI social ($n=369$)	2.56 (±1.39)	2.92 (±1.09)	2.48 (±1.44)	0.01*
ODI travel ($n=368$)	2.08 (±1.35)	2.32 (±1.42)	2.03 (±1.33)	0.12

Means are compared using independent sample t tests comparing surgery and no-surgery groups

Frequencies are compared using Chi-square tests of association comparing surgery and no-surgery groups

Table 2 Univariate regression analysis showing associations between output from the natural language processing models and future surgery ($n=763$)

	Unadjusted odds ratio (95% confidence interval)	p value	Nagelkerke R squared
Variable (output for each case)			
Model prioritizing specificity	1.56 (0.88–2.79)	0.13	0.005
Model prioritizing recall	1.68 (1.15–2.44)	0.007*	0.016
Model prioritizing precision	1.97 (1.33–2.93)	<0.001*	0.024

hypothesis, the results of the NLP analysis did not improve model accuracy and did not enter the final logistic regression model predicting surgical recommendation. In fact, the three variables in the final predictive model (low back and leg pain, distal pain, and difficulties walking) are commonly collected clinical variables that are easily measured using a simple self-report questionnaire. This suggests that the variables most predictive of surgical eligibility are actually self-report data easily obtained through an intake questionnaire. In comparison, the NLP analysis results suggest that the MRI data provided little additional predictive value above what the self-report data offer. Limited research has used machine learning techniques such as NLP to model the predictive value that textual reports of patients' MRI findings have on surgical decision-making. As such, the results of the current study add to our current knowledge regarding factors associated with surgical decision-making. Results from models like this could be used to screen patients referred for surgical consultation or shared with primary care physicians to help them determine appropriate surgical referrals. However, due to the small sample size in the surgical outcome group, this preliminary model requires further validation before it can be readily applied in clinical settings.

The combination of low back and leg pain, as well as distal pain, is most likely indicative of radiculopathy due to lumbar spinal stenosis or disc pathology [1]. These indicators are commonly used by clinicians to distinguish between mechanical back pain (i.e., non-specific) and more specific pathology that may be remedial to surgery [23, 24]. Difficulty walking is likely an indicator of the severity of the condition as well as potential neurologic involvement due to compression of dorsal nerve roots, which again would be remediated by surgery [24]. The empirically derived predictive model, therefore, makes conceptual and clinical sense. In fact, these variables are already most likely the key factors the clinician in the spine assessment clinic uses to determine

Table 3 Logistic regression analysis predicting whether the patient underwent surgery ($n = 281$)

	Unadjusted odds ratio (95% confidence interval)	*p* value	Adjusted odds ratio (95% confidence interval)	*p* value
Variables				
Block 1				
Age	1.01 (1.00–1.03)	0.38	1.01 (0.99–1.03)	0.34
Sex				
Female	0.87 (0.47–1.59)	0.65	0.83 (0.45–1.54)	0.55
Nagelkerke R^2			0.006	
Block 2				
Age	1.01 (1.00–1.03)	0.38	1.01 (0.99–1.03)	0.37
Sex				
Female	0.87 (0.47–1.59)	0.65	0.79 (0.42–1.49)	0.47
Primary symptom				
Other	1.0		1.0	
Neck and arm pain	2.75 (0.50–15.21)	0.25	2.81 (0.50–15.77)	0.24
Low back and leg pain	3.70 (1.66–8.23)	0.001*	3.71 (1.66–8.28)	0.001*
Nagelkerke R^2			0.076	
Block 3				
Age	1.01 (1.00–1.03)	0.38	1.01 (0.99–1.03)	0.28
Sex				
Female	0.87 (0.47–1.59)	0.65	0.82 (0.43–1.56)	0.54
Primary symptom				
Other	1.0		1.0	
Neck and arm pain	2.75 (0.50–15.21)	0.25	3.30 (0.55–19.62)	0.19
Low back and leg pain	3.70 (1.66–8.23)	0.001*	3.10 (1.37–7.02)	0.007*
Pain diagram				
Central	1.0		1.0	
Distal	2.83 (1.46–5.51)	0.002*	2.43 (1.23–4.80)	0.01*
Whole body	0.64 (0.17–2.34)	0.50	0.58 (0.15–2.19)	0.42
Nagelkerke R^2			0.13	
Block 4				
Age	1.01 (1.00–1.03)	0.38	1.01 (0.99–1.03)	0.39
Sex				
Female	0.87 (0.47–1.59)	0.65	0.78 (0.40–1.52)	0.47
Primary symptom				
Other	1.0		1.0	
Neck and arm pain	2.75 (0.50–15.21)	0.25	4.24 (0.70–25.53)	0.12
Low back and leg pain	3.70 (1.66–8.23)	0.001*	2.73 (1.19–6.29)	0.02*
Pain diagram				
Central	1.0		1.0	
Distal	2.83 (1.46–5.51)	0.002*	2.27 (1.21–5.02)	0.01*
Whole body	0.64 (0.17–2.34)	0.50	0.66 (0.17–2.53)	0.54
Walking difficulties				
Yes	2.67 (1.42–5.01)	0.002*	2.72 (1.38–5.33)	0.004*
Proportion of back pain				
60% or greater	0.43 (0.23–0.79)	0.007*	0.54 (0.28–1.05)	0.07
Nagelkerke R^2			0.19	
Block 5				
Age	1.01 (1.00–1.03)	0.38	1.01 (0.99–1.03)	0.38
Sex				
Female	0.87 (0.47–1.59)	0.65	0.78 (0.40–1.51)	0.46

Table 3 (continued)

	Unadjusted odds ratio (95% confidence interval)	*p* value	Adjusted odds ratio (95% confidence interval)	*p* value
Primary symptom				
Other	1.0		1.0	
Neck and arm pain	2.75 (0.50–15.21)	0.25	4.12 (0.68–24.88)	0.12
Low back and leg pain	3.70 (1.66–8.23)	0.001*	2.81 (1.21–6.52)	0.02*
Pain diagram				
Central pain	1.0		1.0	
Distal pain	2.83 (1.46–5.51)	0.002*	2.49 (1.22–5.08)	0.01*
Whole body pain	0.64 (0.17–2.34)	0.50	0.65 (0.17–2.50)	0.53
Walking difficulties				
Yes	2.67 (1.42–5.01)	0.002*	2.73 (1.39–5.37)	0.004*
Proportion of back pain				
60% or Greater	0.43 (0.23–0.79)	0.007*	0.54 (0.28–1.05)	0.06
Natural language processing variable				
Precision	1.17 (0.62–2.20)	0.63	0.84 (0.42–1.69)	0.63
Nagelkerke R^2			0.20	

whether surgery is needed, which may be why these variables had less missing data in those selected for surgery.

Given that only three clinical variables were predictive of the surgical decision, the comprehensive questionnaire completed before clinical assessment at the spine assessment clinic could likely be reduced in length. Overall, it appears that not all the commonly collected information is useful for surgical triage decision-making.

One strength of the current study is the large number of important clinical variables that were available for the predictive modeling. The use of NLP to analyze textual MRI reports was also novel. However, while significant univariate ORs were observed for two of the NLP models, these were not predictive after including other clinical measures. A limitation of the NLP analysis was a fairly low sample size for this type of analysis. A larger sample could help eliminate the class imbalance issue (i.e., higher number of non-surgical than surgical cases). Collecting more data samples would increase the range of potentially feasible models and allow the training of state-of-the-art deep learning and transformer models (e.g., BERT, RoBERTa, ALBERT, etc.) [25, 26]. Consequently, the developed model would be expected to produce a more accurate prediction of the surgical recommendation. Data augmentation techniques could also be used to tackle the issue of imbalance within the dataset, and a more sophisticated approach to data preprocessing could be developed to generate statistically significant text vectors. Additionally, data came from one spine assessment clinic, and the model has not been externally validated; thus, results may not be widely applicable to other clinics. However, the predictor variables in the model are commonly used clinical factors for selecting spine surgery. Another limitation of the current study is that by design, it relies on retrospective data, which can be prone to misclassification bias. However, since this study only aimed to develop a preliminary model requiring further validation at a later point, utilizing retrospective data is justifiable as an initial step. Future research with larger sample sizes and a prospective methodological approach should be done to validate the model established in this study and overcome limitations inherent to retrospective data. Lastly, the surgical outcomes of the patients in our study are unknown. Thus, we do not know whether the surgical decisions made were optimal.

Conclusion

In this study, a preliminary model was created to predict who may require spine surgery. Simple clinical variables appeared more predictive than variables created using NLP machine learning. However, additional research with more data samples is needed to fully evaluate the usefulness of NLP for this task.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00586-023-07552-4.

Acknowledgements Data were provided by the Kaye Edmonton Spinal Assessment Clinic.

Funding Funding for this project was provided by the Alberta Spine Foundation and Alberta Health Services.

Declarations

Conflict of interest Dr. Nataraj is a partner in a company developing an online data collection tool for spine surgery outcomes. The other investigators declare no conflicts of interest.

References

1. Hartvigsen J, Hancock MJ, Kongsted A, Louw Q, Ferreira ML, Genevay S, Hoy D, Karppinen J, Pransky G, Sieper J, Smeets RJ, Underwood M (2018) What low back pain is and why we need to pay attention. Lancet 391:2356–2367. https://doi.org/10.1016/S0140-6736(18)30480-X
2. GBD Disease Injury Incidence and Prevalence Collaborators (2016) Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the global burden of disease study 2015. Lancet 388:1545–1602. https://doi.org/10.1016/S0140-6736(16)31678-6
3. Foster NE, Anema JR, Cherkin D, Chou R, Cohen SP, Gross DP, Ferreira PH, Fritz JM, Koes BW, Peul W, Turner JA, Maher CG (2018) Prevention and treatment of low back pain: evidence, challenges, and promising directions. Lancet 391:2368–2383. https://doi.org/10.1016/S0140-6736(18)30489-6
4. Hall H, Prostko ER, Haring K, Fischer M, Cheng BC (2021) A successful, cost-effective low back pain triage system: a pilot study. N Am Spine Soc J 5:100051. https://doi.org/10.1016/j.xnsj.2021.100051
5. Wilgenbusch CS, Wu AS, Fourney DR (2014) Triage of spine surgery referrals through a multidisciplinary care pathway a value-based comparison with conventional referral processes. Spine 39:S129–S135. https://doi.org/10.1097/Brs.0000000000000574
6. Khor S, Lavallee D, Cizik AM, Bellabarba C, Chapman JR, Howe CR, Mohit AA, Oskouian RJ, Roh JR, Shonnard N, Dagal A, Flum DR (2018) Development and validation of a prediction model for pain and functional outcomes after lumbar spine surgery. JAMA Surg 153:634–642. https://doi.org/10.1001/jamasurg.2018.0072
7. Asher AL, Devin CJ, Archer KR, Cotai S, Parker SL, Bydon M, Nian H, Harrell FE, Speroff T, Dittus RS, Philips SE, Shaffrey CI, Foley KT, McGirt MJ (2017) An analysis from the quality outcomes database, part 2. predictive model for return to work after elective surgery for lumbar degenerative disease. J Neurosurg Spine 27:370–381. https://doi.org/10.3171/2016.8.SPINE16527
8. Muller D, Haschtmann D, Fekete TF, Kleinstuck F, Reitmeir R, Loibl M, O'Riodan D, Porchet F, Jeszenszky D, Mannion AF (2022) Development of a machine-learning based model for predicting multidimensional outcome after surgery for degenerative disorders of the spine. Eur Spine J 31:2125–2136. https://doi.org/10.1007/s00586-022-07306-8
9. Staartjes VE, Stumpo V, Ricciardi L et al (2022) FUSE-ML: development and external validation of a clinical prediction model for mid-term outcomes after lumbar spinal fusion for degenerative disease. Eur Spine J. https://doi.org/10.1007/s00586-022-07135-9
10. Willems P (2013) Decision making in surgical treatment of chronic low back pain: the performance of prognostic tests to select patients for lumbar spinal fusion. Acta Orthop Suppl 84:1–35. https://doi.org/10.3109/17453674.2012.753565
11. Saravi B, Hassel F, Ulkumen S, Zink A, Shavlokhova V, Couillard-Despres S, Boeker M, Obid P, Lang MG (2022) Artificial intelligence-driven prediction modeling and decision making in spine surgery using hybrid machine learning models. J Pers Med 12:509. https://doi.org/10.3390/jpm12040509
12. Broida SE, Schrum ML, Yoon E, Sweeney AP, Dhruv NN, Gomolay MC, Yoon ST (2022) Improving surgical triage in spine clinic: predicting likelihood of surgery using machine learning. World Neurosurg 163:e192–e198. https://doi.org/10.1016/j.wneu.2022.03.096
13. Boden LM, Boden SA, Premkumar A, Gottschalk MB, Boden SD (2018) Predicting likelihood of surgery before first visit in patients with back and lower extremity symptoms: a simple mathematical model based on more than 8000 patients. Spine 43:1296–1305. https://doi.org/10.1097/BRS.0000000000002603
14. Wilson B, Gaonkar B, Yoo B, Salehi B, Attiah M, Villaroman D, Ahn C, Edwards M, Laiwalla A, Ratnaparkhi A, Li L, Cook K, Beckett J, Macyszyn L (2021) Predicting spinal surgery candidacy from imaging data using machine learning. Neurosurgery 89:116–121. https://doi.org/10.1093/neuros/nyab085
15. Klebanoff MA, Snowden JM (2018) Historical (retrospective) cohort studies and other epidemiologic study designs in perinatal research. Am J Obstet Gynecol 219:447–450. https://doi.org/10.1016/j.ajog.2018.08.044
16. Fairbank JC, Pynsent PB (2000) The oswestry disability index. Spine 25:2940–2952. https://doi.org/10.1097/00007632-200011150-00017
17. Tan WK, Hassanpour S, Heagerty PJ, Rundell SD, Suri P, Huhdanpae HT, James K, Carrell DS, Langlotz CP, Organ NL, Meier EN, Sherman KJ, Kallmes DF, Luetmer PH, Griffith B, Nerenz DR, Jarvik JG (2018) Comparison of natural language processing rules-based and machine-learning systems to identify lumbar spine imaging findings related to low back pain. Acad Radiol 25:1422–1432. https://doi.org/10.1016/j.acra.2018.03.008
18. Gross DP, Steenstra IA, Harrell FE Jr, Bellinger C, Zaiane O (2020) Machine learning for work disability prevention: introduction to the special series. J Occup Rehabil 30:303–307. https://doi.org/10.1007/s10926-020-09910-1
19. Turchin A, Florez Builes LF (2021) Using natural language processing to measure and improve quality of diabetes care: a systematic review. J Diabetes Sci Technol 15:553–560. https://doi.org/10.1177/19322968211000831
20. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv:1301.3781. Available at https://arxiv.org/abs/1301.3781
21. Hosmer DW, Lemeshow S, Sturdivant RX (2013) Applied logistic regression. Hoboken, New Jersey
22. Harrell FE, Lee KL, Mark DB (1996) Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 15:361–387. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4%3C361::AID-SIM168%3E3.0.CO;2-4
23. Kreiner DS, Hwang SW, Easa JE, Resnick DK, Baisden JL, Bess S, Cho CH, DePalma MJ, Dougherty P 2nd, Fernand R, Ghiselli G, Hanna AS, Lamer T, Lisi AJ, Mazanec DJ, Meagher RJ, Nucci RC, Patel RD, Sembrano JN, Sharma AK, Summers JT, Taleghani CK, Tontz WL Jr, Toton JF (2014) An evidence-based clinical guideline for the diagnosis and treatment of lumbar disc herniation with radiculopathy. Spine J 14:180–191. https://doi.org/10.1016/j.spinee.2013.08.003
24. Kreiner DS, Shaffer WO, Baisden JL, Gilbert TJ, Summers JT, Toton JF, Hwang SW, Mendel RC, Reitman CA (2013) An evidence-based clinical guideline for the diagnosis and treatment of degenerative lumbar spinal stenosis (update). Spine J 13:734–743. https://doi.org/10.1016/j.spinee.2012.11.059
25. Gupta P, Gandhi S, Chakravarthi BR (2021) Leveraging transfer learning techniques BERT, RoBERTa, ALBERT and

DistilBERT for fake review detection. In: Proceedings of the 13th annual meeting of the forum for information retrieval evaluation. https://doi.org/10.1145/3503162.3503169
26. Qasim R, Bangyal WH, Alqarni MA, Almazroi AA (2022) A fine-tuned bert-based transfer learning approach for text classification. J Healthc Eng 2022:1–17. https://doi.org/10.1155/2022/3498123

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.