frontiers | Frontiers in Surgery

TYPE Original Research
PUBLISHED 18 December 2023
DOI 10.3389/fsurg.2023.1271775

OPEN ACCESS

EDITED BY
Mohammed Ali Alvi,
University Health Network (UHN), Canada

REVIEWED BY
Alessandro Boaro,
University of Verona, Italy
Redi Rahmani,
Barrow Neurological Institute (BNI), United States

*CORRESPONDENCE
Sayan Biswas
✉ sayan.biswas@nca.nhs.uk

†These authors have contributed equally to this work and share first authorship

RECEIVED 02 August 2023
ACCEPTED 27 November 2023
PUBLISHED 18 December 2023

CITATION
Biswas S, McMenemy L, Sarkar V, MacArthur J, Snowdon E, Tetlow C and George KJ (2023) Natural language processing for the automated detection of intra-operative elements in lumbar spine surgery.
Front. Surg. 10:1271775.
doi: 10.3389/fsurg.2023.1271775

Natural language processing for the automated detection of intra-operative elements in lumbar spine surgery

Sayan Biswas[1*†], Lareyna McMenemy[1†], Ved Sarkar[2], Joshua MacArthur[1], Ella Snowdon[1], Callum Tetlow[3] and K. Joshi George[4]

[1]Faculty of Biology, Medicine and Health, University of Manchester, Manchester, United Kingdom, [2]College of Letters and Sciences, University of California, Berkeley, CA, United States, [3]Division of Data Science, The Northern Care Alliance NHS Group, Manchester, United Kingdom, [4]Department of Neurosurgery, Manchester Centre for Clinical Neurosciences, Salford Royal Hospital, Manchester, United Kingdom

Background: The aim of this study was to develop natural language processing (NLP) algorithms to conduct automated identification of incidental durotomy, wound drains, and the use of sutures or skin clips for wound closure, in free text operative notes of patients following lumbar surgery.

Methods: A single-centre retrospective case series analysis was conducted between January 2015 and June 2022, analysing operative notes of patients aged >18 years who underwent a primary lumbar discectomy and/or decompression at any lumbar level. Extreme gradient-boosting NLP algorithms were developed and assessed on five performance metrics: accuracy, area under receiver-operating curve (AUC), positive predictive value (PPV), specificity, and Brier score.

Results: A total of 942 patients were used in the training set and 235 patients, in the testing set. The average age of the cohort was 53.900 ± 16.153 years, with a female predominance of 616 patients (52.3%). The models achieved an aggregate accuracy of >91%, a specificity of >91%, a PPV of >84%, an AUC of >0.933, and a Brier score loss of ≤0.082. The decision curve analysis also revealed that these NLP algorithms possessed great clinical net benefit at all possible threshold probabilities. Global and local model interpretation analyses further highlighted relevant clinically useful features (words) important in classifying the presence of each entity appropriately.

Conclusions: These NLP algorithms can help monitor surgical performance and complications in an automated fashion by identifying and classifying the presence of various intra-operative elements in lumbar spine surgery.

KEYWORDS

clips, dural tears, natural language processing, sutures, spine surgery, wound drains

Abbreviations

AUC, area under the receiver-operating curve; CI, confidence intervals; CPT, current procedural terminology; DCA, decision curve analysis; ICD, international statistical classification of diseases; NLP, natural language processing; PPV, positive predictive value; TRIPOD, transparent reporting of multivariable prediction models for individual prognosis or diagnosis; XGBoost, extreme gradient boosting.

1. Introduction

Administrative, billing, and coding tasks are a major source of financial and economic burden on healthcare systems worldwide (1). With the increase in healthcare and labour costs in recent years, major health systems are shifting towards minimising financial expenditure while maximising patient care. A key component in this process is optimising the clinical coding pipeline by reducing the burden on labour with limited manual review and intervention. The clinical coding process involves transforming medical records, usually presented as free text written by clinicians, into structured codes using the standardised Current Procedural Terminology (CPT) and the International Statistical Classification of Diseases (ICD) codes. The purpose of such clinical coding is to characterise the use of hospital services, document patient outcomes, and quantify clinical and surgical practices to allow for optimal financial reimbursement and to inform healthcare service planning and policy (2, 3).

Natural language processing (NLP) is a domain of machine learning that focuses on the analysis of structured and unstructured free text. NLP techniques are well suited for clinical coding due to their ability to analyse free text in real time with great precision. In the United Kingdom, the General Medical Council states that maintaining accurate and detailed clinical documentation is essential across all specialties for good medical practice (4), in addition to providing information for research, audits, and medicolegal records (5, 6). The current epidemic of defensive practice due to fear of medicolegal repercussions has had an extensive impact on neurosurgical documentation practices, resulting in more detailed documentation of procedures (7). Despite guidelines being available for the documentation of operative notes (8), many studies have demonstrated the inadequate quality of operative notes with much salient information missing, including the nature of the surgery, indication of surgery, estimated blood loss, incidence of complications, and postoperative instructions (6, 9–11). Such non-standardised documentation can lead to greater manual review times, making the extraction of relevant information more labour-intensive. The creation of accurate NLP algorithms trained on a large number of heterogeneous documents can be used to supplement the current clinical coding process, reducing the need for extensive and tedious manual reviews.

Spine surgery comprises the majority of operative cases in neurological surgery. Incidental durotomy, lumbar drains, and type of skin closure (sutures or clips) are important elements included in operative notes and are associated with patient outcomes, and therefore accurate documentation is vital to inform best clinical practice (12–16). At present, CPT and ICD-10 codes are used to identify incidental durotomies and "dural tears" within operative notes. However, these modalities have been shown to lack sensitivity, resulting in the underreporting of these complications (17–19). To the best of our knowledge, no such codes exist for the identification of the use of drains or wound closure technique used. Hence, the aim of this study is to develop NLP algorithms to conduct automated surveillance for identification of incidental durotomy, wound drains, and the use of sutures or skin clips for wound closure, in free text operative notes of patients following lumbar surgery. Towards this, in this study we attempted to evaluate if NLP techniques could be harnessed to analyse operative notes to detect the three important elements of spine surgery: incidental durotomy, the use of wound drains, and type of skin closure (suture or clips).

2. Materials and methods

2.1. Guidelines

The following guidelines were followed in this study: the *Journal of Medical Internet Research* (JMIR) Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research, and the Transparent Reporting of Multivariable Prediction Models for Individual Prognosis or Diagnosis (TRIPOD) checklist (20, 21).

2.2. Data source and outcome measure

A single tertiary neurosurgical centre retrospective case series analysis was conducted for all patients who underwent lumbar spine surgery between January 2015 and June 2022. The inclusion criteria for this study were as follows: (1) patient age more than or equal to 18 years, (2) patient underwent a primary lumbar discectomy and/or decompression at any lumbar level, and (3) availability of index surgery operation notes in our electronic health records. The exclusion criteria included any patients with incomplete data and patients who underwent primary lumbar discectomy and/or fusion. The hospital's electronic patient records were examined and a total of 1,177 patients were identified. Each patient's operation note was then blinded and extracted in an anonymised manner. Our study was approved by the local hospital's institutional review board because of the retrospective and anonymised operative note data collection method. The study was registered as a health improvement project with the requirement for patient consent being waived. All methods were conducted in accordance with local and national guidelines and regulations.

Along with the operation notes, the age (continuous) and gender of the patient (male or female) were also collected as independent variables. There were three primary outcomes for each operation note: (1) the presence of intra-operative durotomies (binary outcome), (2) the placement of wound drains (binary outcome), and (3) the use of clips or sutures for skin closure (binary outcome). The terms durotomies and dural tears are used interchangeably in this paper. Each patient's operation note was reviewed and annotated by blinded researchers (LM and SB) who were not involved in the care of these patients. The results of each outcome category were then verified by the senior author.

2.3. Data pre-processing

The data acquisition, pre-processing, model development, and evaluation pipeline have been highlighted in Figure 1. The dataset was initially cleaned with a custom data-cleaning function that consisted of the removal of special characters retrieved from the Natural Language Toolkit (NLTK) such as "@/{$#%&" and stopwords including "*and*", "*or*", and "the". These words do not carry significant meaning or information in text analysis tasks, hence their removal helps to de-noise the text data resulting in the better efficiency and performance of NLP models. Stemming and lemmatisation are two common techniques used in the data-cleaning function, both of which aim to normalise words by reducing them to their base or root forms. Stemming achieves this by removing any suffixes at the end of a word, while lemmatisation is the process of reducing a word to its base or dictionary form (known as the lemma) while taking into account the context and part of speech of the word.

Lastly, the CountVectorizer library function was used to pre-process the cleaned data. By default, CountVectorizer uses the "term frequency" weighting for single tokenisation, which means it represents each word by the number of times it appears in a document. This results in a document-term matrix where each element represents the frequency of a particular word in a specific document. The resulting matrix is then used as the input to various machine learning algorithms such as clustering, classification, and topic modelling. By representing text data in a numerical format, the CountVectorizer enables machine learning (ML) algorithms to process and analyse textual data, which would otherwise be difficult due to the unstructured nature of natural language.

2.4. Model development

An 80:20 training–testing split was carried out on the total cohort of 1,177 patients, with 942 patients in the training set and 235 patients in the testing set. The datasets were stratified for the outcome variables to account for class imbalances. An extreme gradient-boosting (XGBoost) NLP classifier was developed to predict each outcome category. XGBoost was selected as the classifier of choice owing to a number of factors: (1) its ability to handle high-dimensional feature spaces such as word to vector embeddings used in NLP, (2) the ability to handle and adjust for sparse and imbalanced datasets using weighted loss functions and subsampling, (3) its faster computational run time and scalability, and (4) explicit feature importance calculation for each input attribute (11, 22). Three individual models were created for identifying each outcome category, and the outputs from the models were concatenated to produce a multilabel ensemble output with the predicted probabilities for each outcome. The three ML models will be referred to as the dural tear, drains, and clips vs. sutures models in this paper. An iterative process termed Grid Search was used to optimise the model hyperparameters. In grid search, a predefined set of hyperparameter values is defined, and the model is trained and evaluated on all possible combinations of these values to achieve the highest level of accuracy.

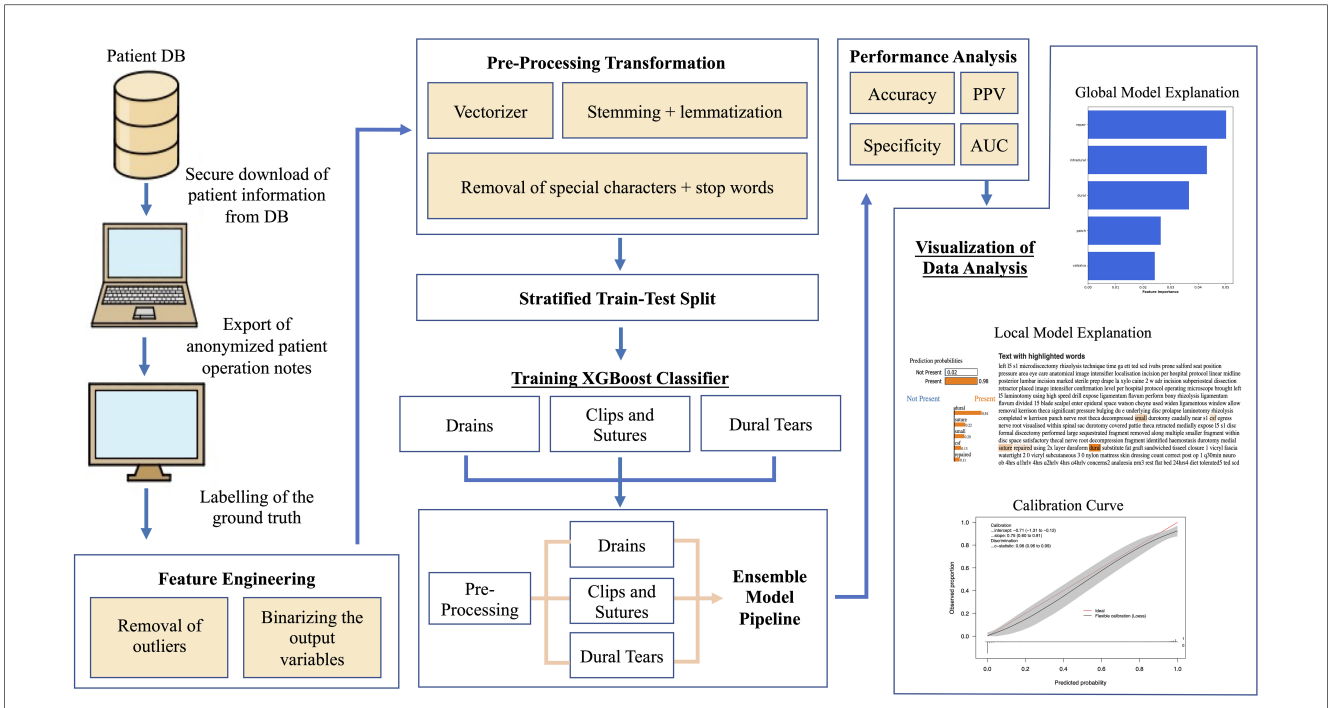


FIGURE 1
Data acquisition, processing, analysis, and visualisation pipeline. DB, database.

The models were trained on fivefold stratified K-fold cross-validation with five repeats on the training dataset. The training and testing datasets were stratified by each of the outcome categories to standardise the class imbalances within our outcome variables and provide us with the best overall performance results for the models. The performance of the models were evaluated via five performance metrics on the training and testing sets: accuracy, precision/positive predictive value (PPV), specificity, area under the receiver-operating curve (AUC)/discrimination, and the Brier score loss. All metrics were bootstrapped with 1,000 resamples to derive the associated 95% confidence intervals (CIs). Each model was then calibrated on the testing set. Calibration refers to how well a model's predicted probabilities align with the true observed probabilities in the study population. This is evaluated using a calibration curve, which is ideally a 45° straight line starting from the origin, with a slope of 1 (indicating the spread of the model's estimated probabilities over the observed probabilities), and an intercept of 0 (indicating how much the model tends to over- or underestimate the true probability). In this study, the preferred method of calibration was Platt scaling or sigmoid binned calibration, which involved dividing the probability range into 10 bins and evaluating the shape of the calibration curve, as well as its slope, intercept, and the Brier score loss metric. In addition, the decision curve analysis (DCA) was used to evaluate and plot the clinical benefit of using the NLP algorithms to predict the presence of each outcome variable over a wide range of predicted threshold probabilities. The DCA illustrates the net benefit defined as the number of true positives detected for each outcome class when using the NLP algorithms on individual patient operation notes.

A model-specific global feature importance analysis was conducted on the trained models via retrieval of each model's relative feature weights that were averaged across all training folds. Furthermore, the Local Interpretable Model-agnostic Explanation analysis was performed to predict and highlight the important features on an individual patient operation note level.

2.5. Statistical analysis

All statistical analyses were conducted using IBM SPSS software (Statistical Package for the Social Science; SPSS Inc., Chicago, IL, USA) Version 25 for Mac, Microsoft Excel (Office 365, Microsoft, Seattle, WA, USA), and the R coding language (R Foundation for Statistical Computing, Vienna, Austria). Histogram plots and the Kolmogorov–Smirnov test were utilised for tests of normality for the continuous variables. The chi-squared tests were used to compare all categorical variables, and the independent samples *t*-test was used to compare the means of the continuous variables. Temporal trend analysis with a linear line of best fit was conducted for all variables, within our retrospective observation time period. A *p*-value <0.05 was considered statistically significant.

TABLE 1 Cohort demographics of the total patient cohort.

	Total cohort ($n = 1,177$)
Age	53.900 ± 16.153
Sex	
Female	616 (52.3%)
Male	561 (47.7%)
Drain(s)	
Yes	373 (31.6%)
No	801 (67.8%)
Closure	
Clips	458 (38.9%)
Sutures	710 (60.3%)
Dural tear(s)	
Yes	117 (9.9%)
No	1,060 (90.1%)

3. Results

3.1. Cohort demographics

A total of 1,177 patients were included in the study, with 942 patients in the training set and 235 patients in the testing set. Table 1 demonstrates the total cohort demographics. The average age of the cohort was 53.900 ± 16.153 years, with a female predominance of 616 patients (52.3%). The rates of intra-operative durotomy and the use of wounds drains were 9.9% (117/1,177) and 31.6% (373/1,177), respectively. Overall, the use of sutures [710 (60.3%)] was more common for skin closure compared with the use of metal surgical clips [458 (38.9%)]. The inter-variable comparative analysis (Table 2) demonstrated a significant relationship between increasing patient age and the use of sutures (p-value = 0.001). We also noted that with an ageing population, the operative age of our patients significantly increased over our observation period (p-value = 0.013). There was also a statistically significant relationship between the use of sutures for skin closure in cases with intra-operative dural tears and wound drains (p-value < 0.001). However, there was no

TABLE 2 Inter-variable statistical correlation analysis using *t*-tests for continuous variables and Chi-square tests for categorical variables.

	p-value					
	Age	Sex	Drains	Closure	Dural tear (s)	Year of surgery
Age		0.137	0.283	0.001[a]	0.501	0.013[a]
Sex						
Female	0.137		0.278	0.294	0.217	0.322
Male						
Drain(s)						
Yes	0.283	0.278		<0.001[a]	0.554	0.906
No						
Closure						
Clips	0.001[a]	0.294	<0.001[a]		<0.001[a]	0.017[a]
Sutures						
Dural tear(s)						
Yes	0.501	0.217	0.554	<0.001[a]		0.853
No						

[a]Statistically significant *p*-value.

statistically significant relationship between the use of wound drains and the presence of intra-operative dural tears (p-value = 0.554).

3.2. Temporal trend analysis

The Mann–Kendall test was used to analyse the temporal trends of the variables across our observation time period as shown in Supplementary Figure S1. During the study period, the total number of lumbar discectomies and/or decompressions decreased significantly from 220 surgeries in 2015 to 56 in the first half of 2022 (−112 estimated in a year) (tau = −0.929, p-value = 0.002). This decline was observed in all the years with the exception of 2019, which saw an increase of one operation from the previous year. It was noted that there was also a decrease in all spinal procedures post COVID-19, which may account for the decrease. The frequency of intra-operative durotomies/dural tears did decrease over the study period; however, no statistical significance was observed (tau = −0.286, p-value = 0.386), with rates ranging from 5.8% to 14.2%. The frequency of intra-operative placement of wound drains also statistically significantly increased over the study period, rising from 18.6% in 2015 to 41.1% in 2022 (tau = 0.643, p-value = 0.035). The preferred method of skin closure also changed over the study period, demonstrating a preference for closure with sutures in later years (tau = 0.5, p-value = 0.108) with a rise from 54% in 2015 to 75% in 2022. We observed an exact yet complementary decrease in the use of surgical clips for skin closure over the years (tau = −0.5, p-value = 0.108).

3.3. Model performance

Table 3 provides the performance metrics for the three ML models on the testing dataset. The dural tears model achieved an accuracy of 91.7615 (95% CI: 88.636–94.602), a PPV of 84.211% (95% CI: 80.667–90.000), a specificity of 99.032% (95% CI: 96.959–99.750), and an AUC of 0.946 (95% CI: 0.917–0.970). The drains model achieved an accuracy of 94.894% (95% CI: 92.330–97.160), a PPV of 88.696% (95% CI: 82.308–94.000), a specificity of 94.694% (95% CI: 90.886–97.025), and an AUC of 0.950 (95% CI: 0.923–0.973). The clips vs. sutures model achieved an accuracy of 93.750% (95% CI: 91.193–96.307), a PPV of 94.495% (95% CI: 91.379–97.260), a specificity of 91.177% (95% CI: 84.770–95.153), and an AUC of 0.933 (95% CI: 0.923–0.973). Figure 2 shows the calibration curves for each of the models. The dural tears model had a propensity to underpredict the presence of a dural tear, with a Brier score loss of 0.082 (95% CI: 0.054–0.114), an intercept of 0.91 (95% CI: 0.46–1.36), and a slope of 0.99 (95% CI: 0.76–1.23). The drains model demonstrated excellent calibration across all predicted probabilities with a Brier score loss of 0.051 (95% CI: 0.028–0.076), an intercept of −0.71 (95% CI: −1.31 to −0.12), and a slope of 0.75 (95% CI: 0.60–0.91). The clips vs. sutures model demonstrated a tendency to overpredict the use of sutures for skin closure, with a Brier score loss of 0.063 (95% CI: 0.037–0.088), an intercept of −0.01 (95% CI: −0.61–0.60), and a slope of 0.65 (95% CI: 0.53–0.77). Lastly, the decision curve analysis on the testing set revealed that all NLP algorithms ensured greater clinical net benefit at all possible threshold probabilities relative to the default decisions of changes made for all or none patients (Figure 3).

3.4. Model explainability

The global feature importance calculations for the NLP algorithms are presented in Figure 4. These explanations highlight that for identification of an intra-operative durotomy, the five most meaningful features (words) are: “repair,” “intradural,” “dural,” “patch,” and “Valsalva.” The five most important features (words) for detecting the intra-operative placement of a lumbar drain are: “drain,” “fascial,” “scoliosis,” “clotting,” and “incision.” Similarly, for detecting whether surgical clips or traditional sutures were utilised for skin closure, the following five words were the most important: “clip,” “staple,” “warmer,” “lamina,” and “clamp.” In addition, the local feature importance analysis for an example patient level operation note demonstrates that the dural tear model is able to identify the five most important clinically meaningful features to detect the presence of an intra-operative dural tear (Figure 5). Interestingly, the local feature importance analysis for the drains and clips vs. sutures model demonstrated that the algorithm primarily searched only for the words “drain” and “clip,” respectively, to make the prediction, with the other aforementioned features possessing very little impact on the outcome.

4. Discussion

This study analysed the trends in the use of various intra-operative elements in spine surgery and developed NLP algorithms capable of reliably identifying these elements in operative notes. The automated identification of these elements

TABLE 3 Performance metrics of the machine learning model on the testing set with 95% confidence intervals.

Model	Accuracy (%)	Precision/PPV (%)	Specificity (%)	AUC	Brier score loss
Testing set (n = 235)					
Dural tears	91.761 (88.636–94.602)	84.211 (80.667–90.000)	99.032 (96.959–99.750)	0.946 (0.917–0.970)	0.082 (0.054–0.114)
Drains	94.894 (92.330–97.160)	88.696 (82.308–94.000)	94.694 (90.886–97.025)	0.950 (0.923–0.973)	0.051 (0.028–0.076)
Clips vs. sutures	93.750 (91.193–96.307)	94.495 (91.379–97.260)	91.177 (84.770–95.153)	0.933 (0.923–0.973)	0.063 (0.037–0.088)

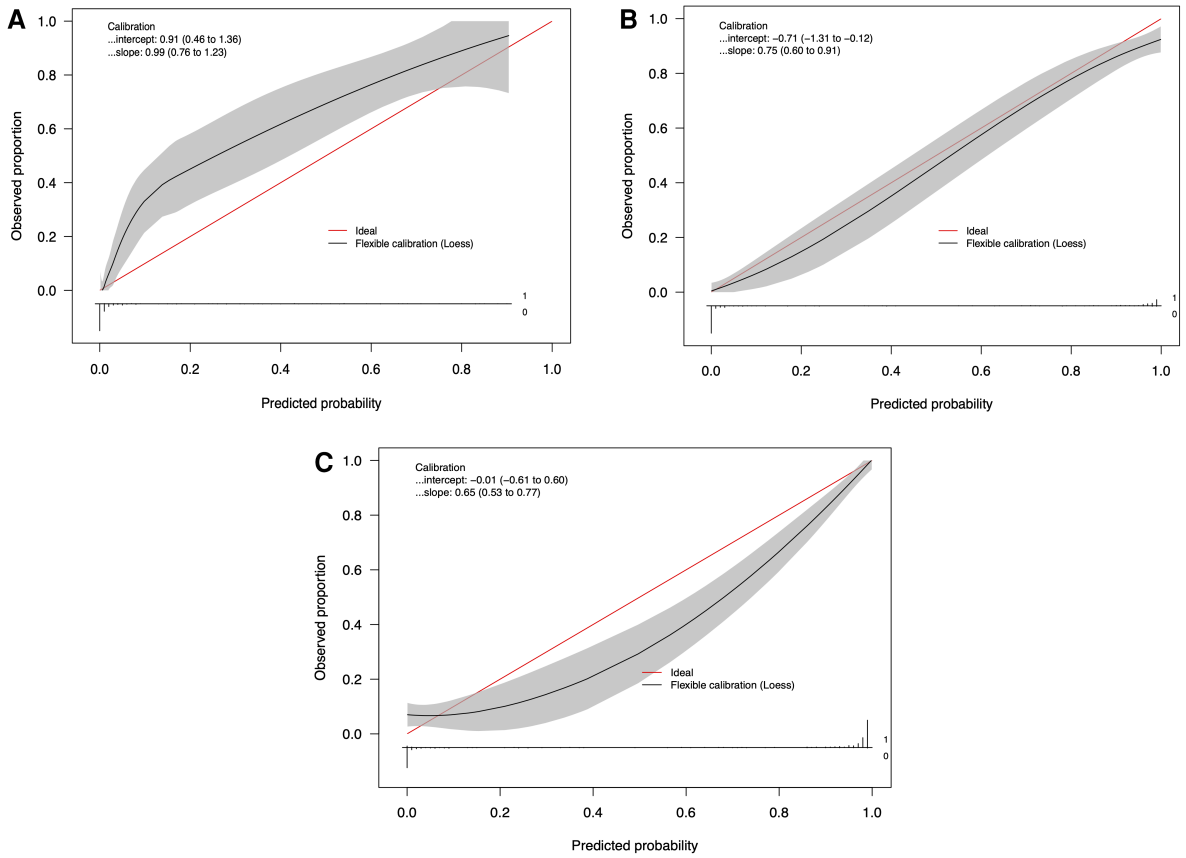


FIGURE 2
Calibration curves of the natural language processing models for identifying (A) dural tears, (B) wound drains, and (C) clips vs. sutures, in the testing set (n = 235).

can facilitate more efficient clinical coding and billing processes, help optimise hospital quality improvement and safety efforts, assist clinicians in auditing surgical practices, and guide overall resource allocation. This study demonstrates that our NLP algorithms are capable of reliably and accurately identifying the placement of intra-operative wound drains, the presence of incidental dural tears, and whether surgical clips or sutures were utilised for skin closure. This is the first ever study from a country with a public healthcare system that has demonstrated the feasibility of using automated NLP systems in operative notes to potentially guide both surgical practices and resource allocation.

The use of NLP techniques in spine surgery has seen a rise in the recent years and is projected to rapidly grow in the future (23, 24). The ability of NLP to perform precise automated surveillance of operative notes, to answer clinically relevant questions, serves to reduce the burden of time-intensive and error-prone reviews by clinical coders (25, 26). The delays in clinical coding within the National Health Service (NHS) impose a significant burden, with the potential for funding to be blocked if coding is not completed within a prerequisite timeframe (27). The average accuracy of this coding has been reported at approximately 83%, with large inter-study variability (28). Such problems exist in majority of healthcare systems worldwide and necessitate the development of automated techniques capable of facilitating these burdensome manual record review processes. Within this realm, Zaidat et al. have already developed an XLNet model capable of automatically generating CPT billing codes from operative notes for three specific surgical procedures: anterior cervical discectomy and fusion (ACDF), posterior cervical discectomy and fusion (PCDF), and cervical disc arthroplasty (CDA) (2). Such models have the potential to greatly reduce manual review/input, minimise errors in the coding process, and promote standardisation. Most recently, Shost et al. have also demonstrated a model capable of reliably identifying the type of spinal surgery performed via analysis of patient consent forms (29). The ability to rapidly classify surgical practices can be beneficial to both hospitals and the practicing surgeons. This will help track surgical volume, surgery-specific patient outcomes, and also provide trainees with a method of tracking individual surgical experience. In addition, NLP algorithms have also demonstrated predictive value in classifying lumbar spine imaging findings and in determining the need for surgical intervention in patients with low back pain via analysis of radiological and clinical reports (30). These examples highlight the importance of NLP techniques in improving the provision of patient care and demonstrate the clinical utility of such models in enhancing hospital and surgical practices.

Our NLP algorithms were developed to identify the presence of three important intra-operative factors that play a role in guiding

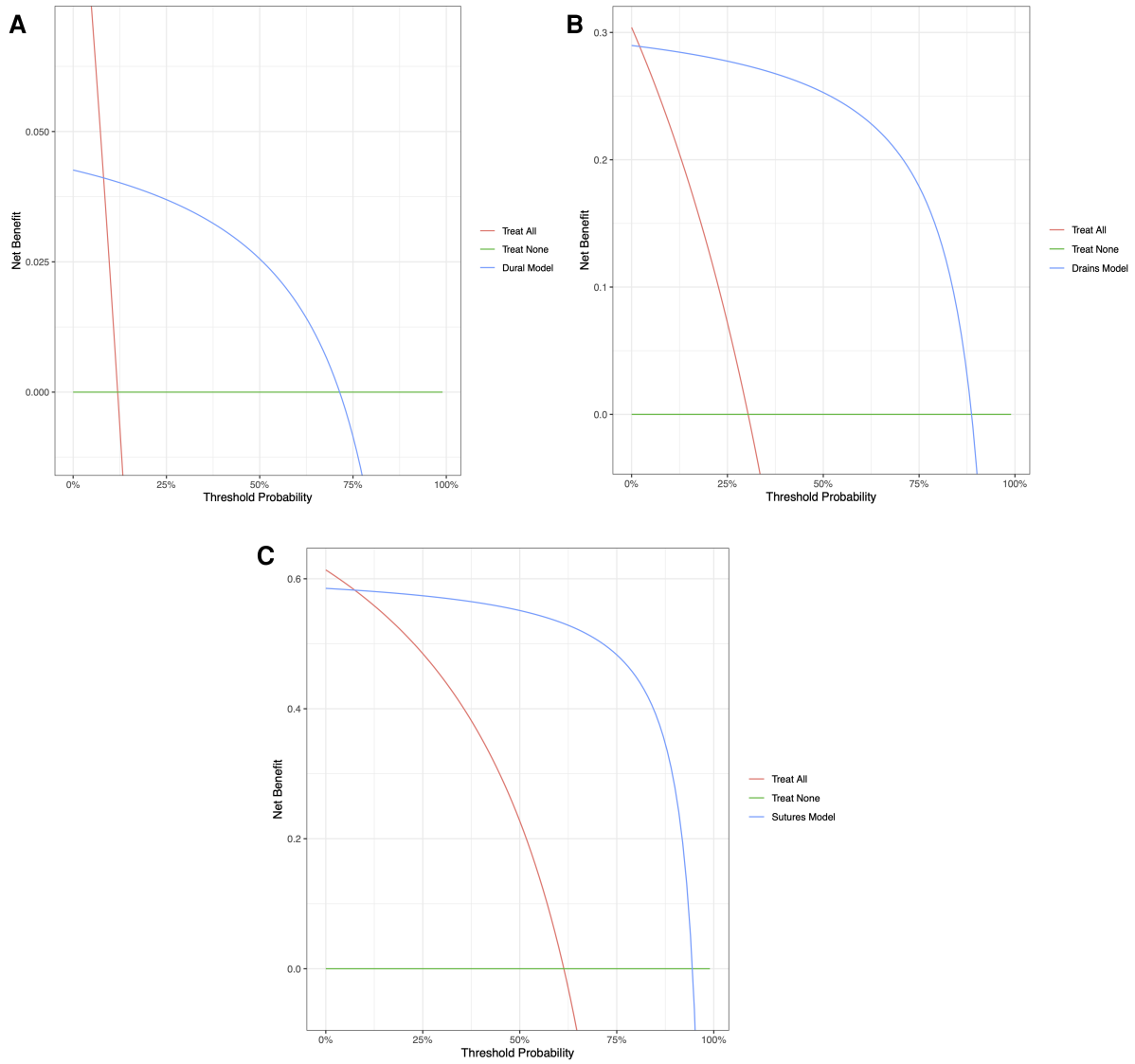


FIGURE 3
Decision curve analyses comparing expected clinical net benefit of the three models: (A) dural tears, (B) wound drains, and (C) clips vs. sutures, on the testing set.

the resource allocation and surgical practices of a neurosurgical department. In this study, the prevalence of incidental durotomy was 9.9%, in line with the recent literature on lumbar surgery (18, 31, 32). Our model demonstrated adequate discrimination and performance in identifying intra-operative dural tears and highlighted the use of clinically relevant features (words) to make its predictions. Previous studies by Karhade et al. have also been successful in the identification of incidental durotomy with an accuracy of 99%, surpassing the performance of CPT and ICD-10 codes, which demonstrated an accuracy of only 64% (18). Interestingly, however, the feature importance in their NLP algorithm showed different features compared with ours, further underscoring the potential variability in NLP algorithm performance across different cohorts that are geographically separated, and highlighting the need for broader validation studies (17). The importance of reliably identifying cases of intra-operative incidental durotomy is highlighted by evidence suggesting that patients with durotomies tend to have increased operative durations and inpatient length of stay (LOS) (33). Thus, accurate depiction of the rates of incidental durotomy can aid postoperative patient counselling, quantify surgical complication rates, and help track surgical performance.

For the wound drains model, our study demonstrated an accuracy of almost 95%. Previous studies have concluded that postoperative drains are currently being overused in spinal surgery, potentially imposing an increased risk of unnecessary complications, while not lending substantial benefit (34). Most notably, reports have suggested an elevated risk of surgical site infections (SSI) (35, 36), although this has been refuted by other papers (15, 37–39). Ho et al. interestingly report that both the absence of a wound drain and increased drainage when drains are used indicate an increased risk of delayed infection after posterior

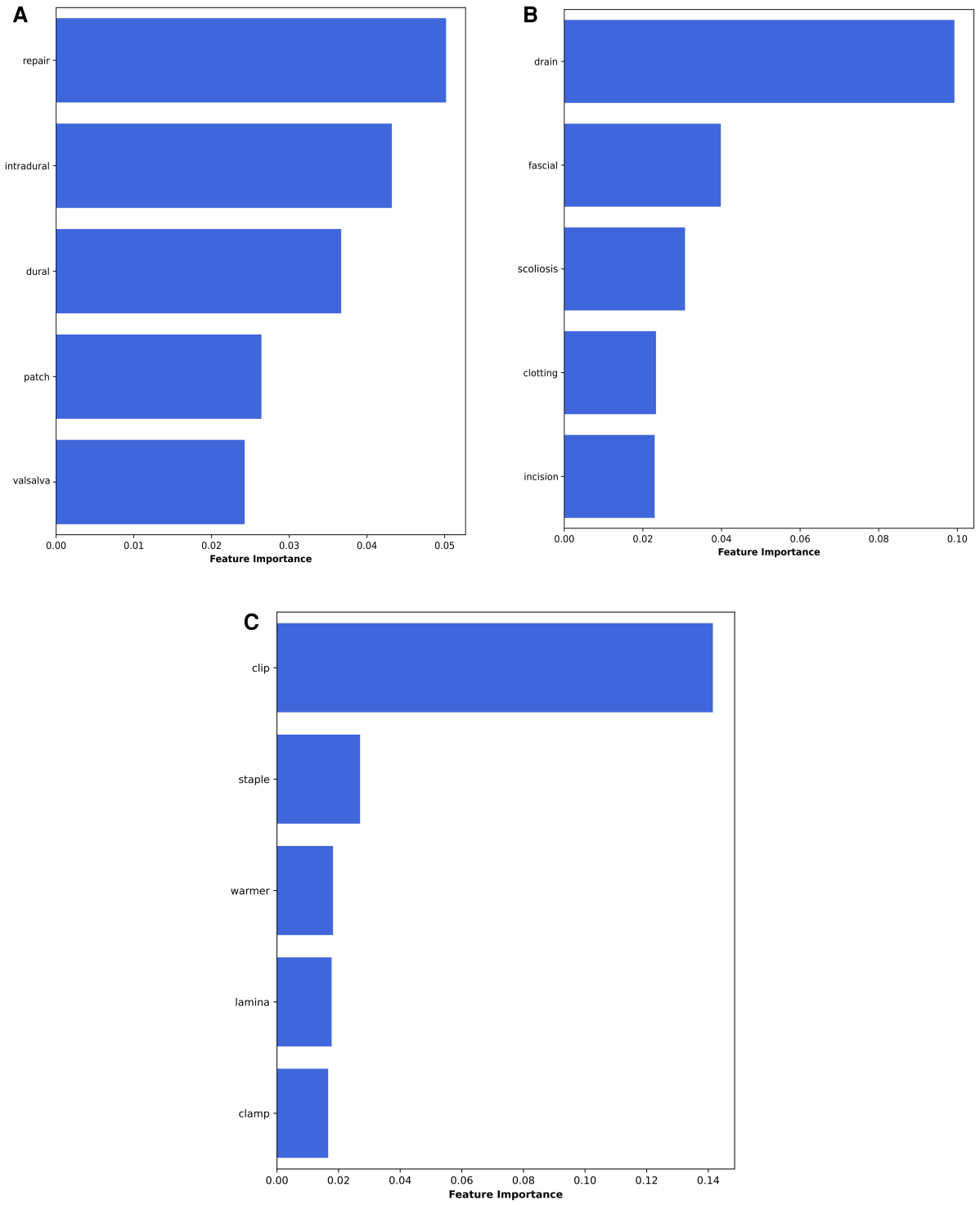
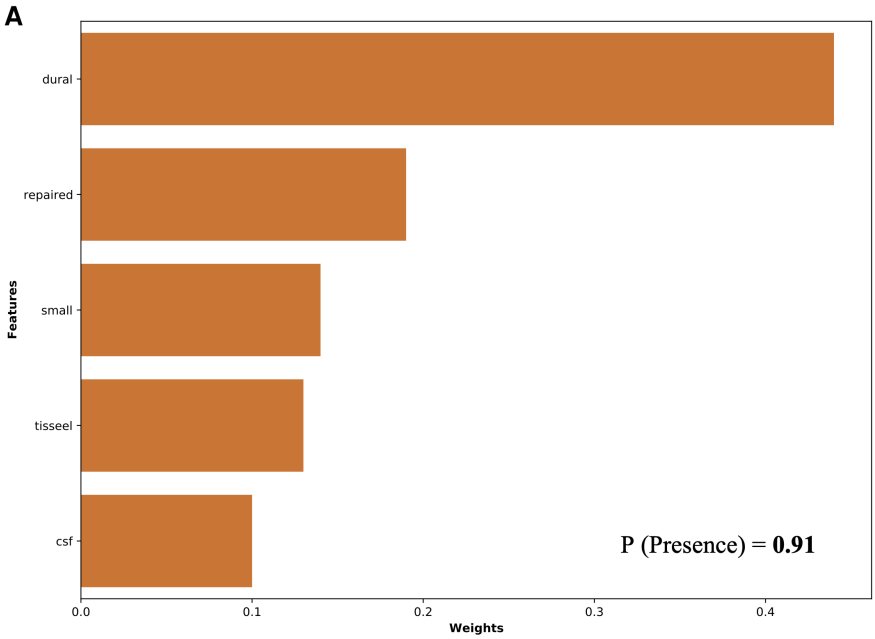


FIGURE 4
Global feature importance values for the NLP algorithms: (**A**) dural tears, (**B**) wound drains, and (**C**) clips vs. sutures.

spine surgery (36). Walid et al. additionally found that the use of postoperative drains was linked to increased post-haemorrhagic anaemia, and a subsequent requirement of allogenic blood transfusions (40), which may impose greater costs to the healthcare system. Adogwa et al. have also demonstrated that patients with postoperative drains have a significantly longer LOS compared with patients with no drains (37). The combination of such factors highlights the importance of tracking and quantifying the use of drains in spine surgery, and therefore the development of our NLP algorithm will allow for its automated and reliable detection. The future application of this algorithm in tracking wound drain use and the associated SSI rates remains to be investigated.



B

left l5 s1 microdiscectomy rhizolysis technique time ga ett ted scd ivabs prone salford seat position pressure area eye care anatomical image intensifier localisation incision per hospital protocol linear midline posterior lumbar incision marked sterile prep drape la xylo caine 2 w adr incision subperiosteal dissection retractor placed image intensifier confirmation level per hospital protocol operating microscope brought left l5 laminotomy using high speed drill expose ligamentum flavum perform bony rhizolysis ligamentum flavum divided 15 blade scalpel enter epidural space watson cheyne used widen ligamentous window allow removal kerrison theca significant pressure bulging du e underlying disc prolapse laminotomy rhizolysis completed w kerrison punch nerve root theca decompressed small durotomy caudally near s1 csf egress nerve root visualised within spinal sac durotomy covered pattie theca retracted medially expose l5 s1 disc formal discectomy performed large sequestrated fragment removed along multiple smaller fragment within disc space satisfactory thecal nerve root decompression fragment identified haemostasis durotomy medial suture repaired using 2x layer duraform dural substitute fat graft sandwiched tisseel closure 1 vicryl fascia watertight 2 0 vicryl subcutaneous 3 0 nylon mattress skin dressing count correct post op 1 q30min neuro ob 4hrs q1hrly 4hrs q2hrly 4hrs q4hrly concerns2 analgesia prn3 rest flat bed 24hrs4 diet tolerated5 ted scd tinzaparin 40mg c od tomorrow evening6 monitor wound csf leak7

FIGURE 5
Local feature importance analysis for detecting dural tears in an example individual patient operation note as generated by the NLP algorithm.

Further, our clips vs. suture model demonstrated an accuracy of >93% accuracy, and our temporal trend analysis showed a preference for using sutures for wound closure. Various studies have concluded that suturing is more efficient when compared with the use of clips for good wound closure, resulting in lower rates of separation, prevention of SSI, and ultimately shorter hospital LOS (41, 42). Contrastingly, postoperative analysis of visual analogue pain scores comparing the use of clips to non-absorbable sutures have also demonstrated a significantly quicker and pain-free experience for patients with stapled wounds (43). From an economic perspective as well, studies have demonstrated that staples/clips are less expensive than sutures and that the financial gain appears to increase as laceration length increases (44). However, conflicting literature exists on the impact of sutures and clips on patients postoperatively (22), with the need for future robust randomised control trials to further investigate their effects. Nevertheless, such single-use surgical items are the largest contributors to the surgical carbon footprint and hence precise quantification of such use can guide both financial and environmental practices (45). Therefore, such automated NLP techniques can facilitate accurate data collection and analysis of the use of clips and sutures in neurosurgery. Nevertheless, the

utility of this NLP algorithm in identifying and predicting postoperative LOS, risk of SSI, and the estimated carbon footprint after a surgery remains to be explored in a future study.

4.1. Limitations

Despite these results, the study has several limitations. First, this was a retrospective analysis at a single centre and therefore the development and testing of the NLP algorithms was geographically limited to a specific region. This raises questions about the algorithms' generalisability and their performance in diverse linguistic and clinical contexts. Furthermore, the surgeons affiliated with the healthcare entities in the study likely share practices that influence the specific terminology used to document the various intra-operative characteristics, which could bias the results. Hence, future prospective and external validation of the algorithms needs to be performed to validate the clinical utility of the algorithms. In addition, there are other approaches that can be utilised to adapt our general model to geographically distinct regions. Geographically customisable models can be implemented via techniques such as federated learning and transfer learning. Federated learning enables the collaborative training of models across multiple centres without data sharing, preserving both privacy and centre-specific relationships and trends in the data. Transfer learning further facilitates rapid fine-tuning, which can efficiently adapt a base model to new regions by learning from small local datasets, boosting model performance and reliability. Secondly, though these models are able to reliably identify the outcomes of interest, a further manual review by clinical coders will still be required to exclude any cases of false positives or false negatives. Thus, the need for manual review will still exist, though with a considerably lower level of burden. Hence, multicentre, linguistically different validation studies in hospitals with varying coding/billing practices are required to determine the reliability of these models. Lastly, with the advent of state-of-the-art large language models such as Bidirectional Encoder Representations from Transformers and Generative Pre-trained Transformer models, the need for manual annotation of unstructured, free text data may exponentially reduce. These models are capable of independently performing named entity recognition and can understand the contextual nuances of each outcome of interest. For example, these models would be able to interpret the reason/context for using a drain, or the reason for a durotomy. Thus, in the future the goal would be to develop such models capable of functioning independently without the need for any manual annotation or review.

5. Conclusion

In conclusion, this study evaluated the feasibility and reliability of NLP algorithms in determining the presence of three intra-operative elements in lumbar spine surgery. We demonstrate that these NLP models possess great discriminative ability and accuracy in predicting the presence of wound drains, incidental dural tears, and the use of clips or sutures for wound closure. These models can help automate the clinical coding process, help optimise hospital quality improvement, and monitor surgical performance and practices. This is the first ever study from a country with a primarily public healthcare system that has demonstrated the feasibility of using automated NLP systems in operative notes to potentially guide both surgical practices and resource allocation.

Data availability statement

The original contributions presented in this study are included in the article/**supplementary materials** and further inquires should be directed to the corresponding author.

Ethics statement

The studies involving humans were approved by Northern Care Alliance Research and Innovation Team. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Author contributions

SB: Conceptualization, Formal analysis, Methodology, Supervision, Writing – original draft, Writing – review & editing. LM: Data curation, Formal analysis, Writing – original draft. VS: Formal analysis, Software, Writing – review & editing. JM: Conceptualization, Writing – review & editing. ES: Writing – original draft, Writing – review & editing. CT: Data curation, Methodology, Supervision, Writing – review & editing. KG: Methodology, Supervision, Writing – review & editing.

Funding

The authors declare that no financial support was received for the research, authorship, and/or publication of this paper.

Acknowledgments

The authors wish to acknowledge the Department of Neurosurgery at Salford Royal Hospital for providing the data.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this paper can be found online at: https://www.frontiersin.org/articles/10.3389/fsurg.2023.1271775/full#supplementary-material

References

1. Chernew M, Mintz H. Administrative expenses in the US health care system: why so high? *JAMA*. (2021) 326(17):1679–80. doi: 10.1001/jama.2021.17318

2. Zaidat B, Tang J, Arvind V, Geng EA, Cho B, Duey AH, et al. Can a novel natural language processing model and artificial intelligence automatically generate billing codes from spine surgical operative notes? *Global Spine J*. (2023):21925682231164936. doi: 10.1177/21925682231164935. [Epub ahead of print]

3. Dong H, Falis M, Whiteley W, Alex B, Matterson J, Ji S, et al. Automated clinical coding: what, why, and where we are? *NPJ Digit Med*. (2022) 5(1):159. doi: 10.1038/s41746-022-00705-7

4. GMC. *Good medical practice* (2013). Available at: https://www.gmc-uk.org/ethical-guidance/ethical-guidance-for-doctors/good-medical-practice (accessed July 19, 2023).

5. Hamza A, Abdalrahim H, Idris S, Ahmed O. Evaluating the operative notes of patients undergoing surgery at Omdurman Teaching Hospital, Sudan. *Sch J App Med Sci*. (2013) 1(6):668–72. doi: 10.36347/sjams.2013.v01i06.003

6. Rogers A, Bunting M, Atherstone A. The quality of operative notes at a general surgery unit. *S Afr Med J*. (2008) 98(9):726–8.

7. Gadjradj PS, Ghobrial JB, Harhangi BS. Experiences of neurological surgeons with malpractice lawsuits. *Neurosurg Focus*. (2020) 49(5):E3. doi: 10.3171/2020.8.FOCUS20250

8. RCS. *Good surgical practice* (2019). Available at: https://www.rcseng.ac.uk/standards-and-research/gsp/ (accessed July 19, 2023).

9. Lefter LP, Walker SR, Dewhurst F, Turner RWL. An audit of operative notes: facts and ways to improve. *ANZ J Surg*. (2008) 78(9):800–2. doi: 10.1111/j.1445-2197.2008.04654.x

10. Malik S, Nogaro M, Shenoy R, Mitchell P. Improving trauma operation notes at an emerging trauma unit in the regional trauma network. *Bull R Coll Surg Engl*. (2013) 95(6):1–3. doi: 10.1308/003588413X13643054410386

11. Nyamulani N, Mulwafu W. The quality of hand-written operative notes in a surgical unit at Queen Elizabeth central hospital (QECH), Malawi: a prospective completed audit loop study. *Malawi Med J*. (2018) 30(2):86–9. doi: 10.4314/mmj.v30i2.6

12. Nandyala SV, Elboghdady IM, Marquez-Lara A, Noureldin MNB, Sankaranarayanan S, Singh K. Cost analysis of incidental durotomy in spine surgery. *Spine (Phila Pa 1976)*. (2014) 39(17):E1042–51. doi: 10.1097/BRS.0000000000000425

13. Saxler G, Krämer J, Barden B, Kurt A, Pförtner J, Bernsmann K. The long-term clinical sequelae of incidental durotomy in lumbar disc surgery. *Spine (Phila Pa 1976)*. (2005) 30(20):2298–302. doi: 10.1097/01.brs.0000182131.44670.f7

14. Strömqvist F, Sigmundsson FG, Strömqvist B, Jönsson B, Karlsson MK. Incidental durotomy in degenerative lumbar spine surgery—a register study of 64,431 operations. *Spine J*. (2019) 19(4):624–30. doi: 10.1016/j.spinee.2018.08.012

15. Kanayama M, Oha F, Togawa D, Shigenobu K, Hashimoto T. Is closed-suction drainage necessary for single-level lumbar decompression?: review of 560 cases. *Clin Orthop Relat Res*. (2010) 468(10):2690–4. doi: 10.1007/s11999-010-1235-6

16. Yilmaz E, Blecher R, Moisi M, Ankush C, O'Lynnger TM, Abdul-Jabbar A, et al. Is there an optimal wound closure technique for major posterior spine surgery? A systematic review. *Global Spine J*. (2018) 8(5):535–44. doi: 10.1177/2192568218774323

17. Karhade AV, Oosterhoff JHF, Groot OQ, Agaronnik N, Ehresman J, Bongers MER, et al. Can we geographically validate a natural language processing algorithm for automated detection of incidental durotomy across three independent cohorts from two continents? *Clin Orthop Relat Res*. (2022) 480(9):1766–75. doi: 10.1097/CORR.0000000000002200

18. Karhade A, Bongers M, Groot O, Kazarian E, Cha T, Fogel H, et al. Natural language processing for automated detection of incidental durotomy. *Spine J*. (2020) 20(5):695–700. doi: 10.1016/j.spinee.2019.12.006

19. Ehresman J, Pennington Z, Karhade A, Huq S, Medikonda R, Schilling A, et al. Incidental durotomy: predictive risk model and external validation of natural language process identification algorithm. *J Neurosurg Spine*. (2020) 33(3):342–8. doi: 10.3171/2020.2.SPINE20127

20. Collins GS, Reitsma JB, Altman DG, Moons K. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC Med*. (2015) 13(1):1. doi: 10.1186/s12916-014-0241-z

21. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res*. (2016) 18(12):e323. doi: 10.2196/jmir.5870

22. Selvadurai D, Wildin C, Treharne G, Choksy SA, Heywood MM, Nicholson ML. Randomised trial of subcuticular suture versus metal clips for wound closure after thyroid and parathyroid surgery. *Ann R Coll Surg Engl*. (1997) 79(4):303–6.

23. Huang BB, Huang J, Swong KN. Natural language processing in spine surgery: a systematic review of applications, bias, and reporting transparency. *World Neurosurg*. (2022) 167:156–64.e6. doi: 10.1016/j.wneu.2022.08.109

24. Agaronnik ND, Kwok A, Schoenfeld AJ, Lindvall C. Natural language processing for automated surveillance of intraoperative neuromonitoring in spine surgery. *J Clin Neurosci*. (2022) 97:121–6. doi: 10.1016/j.jocn.2022.01.015

25. Bacco L, Russo F, Ambrosio L, D'Antoni F, Vollero L, Vadalà G, et al. Natural language processing in low back pain and spine diseases: a systematic review. *Front Surg*. (2022) 9:957085. doi: 10.3389/fsurg.2022.957085

26. Groot O, Ogink P, Oosterhoff J, Beam A. Natural language processing and its role in spine surgery: a narrative review of potentials and challenges. *Semin Spine Surg*. (2021) 33(1):100877. doi: 10.1016/j.semss.2021.100877

27. Alonso V, Santos JV, Pinto M, Ferreira J, Lema I, Lopes F, et al. Problems and barriers during the process of clinical coding: a focus group study of Coders' perceptions. *J Med Syst*. (2020) 44(3):62. doi: 10.1007/s10916-020-1532-x

28. Burns EM, Rigby E, Mamidanna R, Bottle A, Aylin P, Ziprin P, et al. Systematic review of discharge coding accuracy. *J Public Health (Oxf)*. (2012) 34(1):138–48. doi: 10.1093/pubmed/fdr054

29. Shost MD, Meade SM, Steinmetz MP, Mroz TE, Habboub G. Surgical classification using natural language processing of informed consent forms in spine surgery. *Neurosurg Focus*. (2023) 54(6):E10. doi: 10.3171/2023.3.FOCUS2371

30. Krebs B, Nataraj A, McCabe E, Clark S, Sufiyan Z, Yamamoto SS, et al. Developing a triage predictive model for access to a spinal surgeon using clinical variables and natural language processing of radiology reports. *Eur Spine J*. (2023) 32(1):181–9. doi: 10.1007/s00586-023-07552-4

31. Mueller KB, Garrett CT, Kane S, Sandhu FA, Voyadzis JM. Incidental durotomy following surgery for degenerative lumbar disease and the impact of minimally invasive surgical technique on the rate and need for surgical revision: a case series. *Oper Neurosurg (Hagerstown)*. (2021) 21(5):351–5. doi: 10.1093/ons/opab282

32. Albayar A, Spadola M, Blue R, Saylany A, Dagli MM, Santangelo G, et al. Incidental durotomy repair in lumbar spine surgery: institutional experience and review of literature. *Global Spine J*. (2022):21925682221141370. doi: 10.1177/21925682221141368. [Epub ahead of print]

33. Desai A, Ball PA, Bekelis K, Lurie J, Mirza SK, Tosteson TD, et al. SPORT: does incidental durotomy affect longterm outcomes in cases of spinal stenosis? *Neurosurgery*. (2015) 76 Suppl 1(01):S57–63; discussion S63. doi: 10.1227/01.neu.0000462078.58454.f4

34. Reier L, Fowler JB, Arshad M, Siddiqi J. Drains in spine surgery for degenerative disc diseases: a literature review to determine its usage. *Cureus*. (2022) 14(3):312–23. doi: 10.7759/cureus.23129

35. Mirzai H, Eminoglu M, Orguc S. Are drains useful for lumbar disc surgery? A prospective, randomized clinical study. *J Spinal Disord Tech*. (2006) 19(3):171–7. doi: 10.1097/01.bsd.0000190560.20872.a7

36. Ho C, Sucato DJ, Richards BS. Risk factors for the development of delayed infections following posterior spinal fusion and instrumentation in adolescent idiopathic scoliosis patients. *Spine (Phila Pa 1976)*. (2007) 32(20):2272–7. doi: 10.1097/BRS.0b013e31814b1c0b

37. Adogwa O, Elsamadicy AA, Sergesketter AR, Shammas RL, Vatsia S, Vuong VD, et al. Post-operative drain use in patients undergoing decompression and fusion: incidence of complications and symptomatic hematoma. *J Spine Surg*. (2018) 4 (2):220–6. doi: 10.21037/jss.2018.05.09

38. Brown MD, Brookfield KFW. A randomized study of closed wound suction drainage for extensive lumbar spine surgery. *Spine (Phila Pa 1976)*. (2004) 29 (10):1066–8. doi: 10.1097/00007632-200405150-00003

39. Choi HS, Lee SG, Kim WK, Son S, Jeong TS. Is surgical drain useful for lumbar disc surgery? *Korean J Spine*. (2016) 13(1):20–3. doi: 10.14245/kjs.2016.13.1.20

40. Walid MS, Abbara M, Tolaymat A, Davis JR, Waits KD, Robinson JS, et al. The role of drains in lumbar spine fusion. *World Neurosurg*. (2012) 77(3–4):564–8. doi: 10.1016/j.wneu.2011.05.058

41. Zaman S, Mohamedahmed AYY, Peterknecht E, Zakaria RM, Mohamedahmed SYY, Hajibandeh S, et al. Sutures versus clips for skin closure following caesarean section: a systematic review, meta-analysis and trial sequential analysis of randomised controlled trials. *Langenbecks Arch Surg*. (2022) 407(1):37–50. doi: 10.1007/s00423-021-02239-0

42. Mostofi K, Peyravi M, Shirbacheh A, Shirbache K. A comparison between different suture techniques in lumbar spine surgery. *Int Wound J*. (2023) 20 (2):296–301. doi: 10.1111/iwj.13875

43. Meiring L, Cilliers K, Barry R, Nel CJ. A comparison of a disposable skin stapler and nylon sutures for wound closure. *S Afr Med J*. (1982) 62(11):371–2.

44. Orlinsky M, Goldberg RM, Chan L, Puertos A, Slajer HL. Cost analysis of stapling versus suturing for skin closure. *Am J Emerg Med*. (1995) 13(1):77–81. doi: 10.1016/0735-6757(95)90248-1

45. Rizan C, Lillywhite R, Reed M, Bhutta MF. The carbon footprint of products used in five common surgical operations: identifying contributing products and processes. *J R Soc Med*. (2023) 116(6):1410768231166135. doi: 10.1177/01410768231166135