A systematic review of natural language processing applications in Trauma & Orthopaedics

From University of Aberdeen, Aberdeen, UK

Correspondence should be sent to L. Farrow luke.farrow@doctors.org.uk

Cite this article:
Bone Jt Open 2025;6(3): 264–274.

DOI: 10.1302/2633-1462.63.BJO-2024-0081.R1

L. Farrow,[1,2] A. Raja,[3] M. Zhong,[1] L. Anderson[1]

[1]Institute of Applied Health Sciences, University of Aberdeen, Aberdeen, UK
[2]Grampian Orthopaedics, Aberdeen Royal Infirmary, Aberdeen, UK
[3]School of Medicine, University of Edinburgh, Edinburgh, UK

Aims

Prevalence of artificial intelligence (AI) algorithms within the Trauma & Orthopaedics (T&O) literature has greatly increased over the last ten years. One increasingly explored aspect of AI is the automated interpretation of free-text data often prevalent in electronic medical records (known as natural language processing (NLP)). We set out to review the current evidence for applications of NLP methodology in T&O, including assessment of study design and reporting.

Methods

MEDLINE, Allied and Complementary Medicine (AMED), Excerpta Medica Database (EMBASE), and Cochrane Central Register of Controlled Trials (CENTRAL) were screened for studies pertaining to NLP in T&O from database inception to 31 December 2023. An additional grey literature search was performed. NLP quality assessment followed the criteria outlined by Farrow et al in 2021 with two independent reviewers (classification as absent, incomplete, or complete). Reporting was performed according to the Synthesis-Without Meta-Analysis (SWiM) guidelines. The review protocol was registered on the Prospective Register of Systematic Reviews (PROSPERO; registration no. CRD42022291714).

Results

The final review included 31 articles (published between 2012 and 2021). The most common subspeciality areas included trauma, arthroplasty, and spine; 13% (4/31) related to online reviews/social media, 42% (13/31) to clinical notes/operation notes, 42% (13/31) to radiology reports, and 3% (1/31) to systematic review. According to the reporting criteria, 16% (5/31) were considered good quality, 74% (23/31) average quality, and 6% (2/31) poor quality. The most commonly absent reporting criteria were evaluation of missing data (26/31), sample size calculation (31/31), and external validation of the study results (29/31 papers). Code and data availability were also poorly documented in most studies.

Conclusion

Application of NLP is becoming increasingly common in T&O; however, published article quality is mixed, with few high-quality studies. There are key consistent deficiencies in published work relating to NLP which ultimately influence the potential for clinical application. Open science is an important part of research transparency that should be encouraged in NLP algorithm development and reporting.

Take home message

- This study highlights some of the key potential uses of natural language processing in Trauma & Orthopaedics.
- It also identifies some methodological concerns with the currently available literature on the subject.

Bone & Joint Open

Introduction

There has been a massive influx of publications regarding artificial intelligence (AI) applications in the domain of Trauma & Orthopaedics (T&O).[1] One AI technique is natural language processing (NLP), which enables processing and analysis of large amounts of natural language or 'free-text' (for example, written information contained within a clinical letter) data.

It is estimated that approximately 80% of healthcare data are in an unstructured or 'free-text' format.[2] These data have the potential to provide a veritable wealth of useful information to guide clinical practice and research. NLP allows users to turn these unstructured data into meaningful material for analysis.

NLP is not without its challenges, in particular risk concerning potential identification of protected healthcare information contained within the free-text resource. Techniques such as 'Hidden In Plain Sight' (HIPS) methods have been developed to attempt to maintain free-text structure while ensuring anonymity,[3] but this in itself requires dedicated health data science infrastructure. Ethical concerns have also been raised about granting access to large volumes of anonymized free-text healthcare data without consent, although previous evidence has suggested that this is supported if particular safeguarding structures are in place.[4]

Despite these challenges, there has been evidence of successful use of NLP applications within the healthcare setting. Examples include delirium detection in the intensive care unit,[5] surveillance of patients at high risk of upper gastrointestinal cancer,[6] and predicting outcomes of critical care patients.[7]

Development of NLP applications has been reported within T&O, such as development of an arthroplasty database[8] and fracture identification.[9] No study to date has, however, methodically assessed the available NLP literature, including an evaluation of study quality and analysis of reported performance metrics. We therefore set out to perform a systematic review of NLP applications within T&O to better appraise current applications and guide future use.

Methods

This systematic review was performed and reported according to the PRISMA statement.[10] Registration prior to study commencement was undertaken on the Prospective Register of Systematic Reviews (PROSPERO) no. CRD42022291714.

Search strategy

Relevant articles were identified through a search of MEDLINE, Allied and Complementary Medicine Database (AMED), Excerpta Medica Database (EMBASE), and Cochrane Central Register of Controlled Trials (CENTRAL). An additional search of the grey literature was also undertaken using OrthoSearch (an orthopaedic-specific database which contains abstracts, articles, and associated media information).[11] All electronic searches were undertaken from database inception to 31 December 2023. Full electronic search terms are shown in Supplementary Table i. Reference lists from all extracted studies were reviewed for potentially eligible manuscripts.

Eligibility criteria

All studies that involved research related to the use of NLP in the setting of T&O and associated subspecialities were included. Exclusion criteria were studies involving other surgical or medical specialities, use of other AI techniques that were not specifically identified as NLP, publications in relation to generative AI, and non-English language publications.

Study identification

Two assessors (LF, AR) independently screened search output titles and abstracts for articles which met the eligibility criteria. Full-text review was undertaken to determine eligibility.

Data extraction

Data extraction was undertaken using a prespecified proforma by two independent assessors (LF, AR). Fields included: 1) Design overview: author, year, subspeciality, and NLP domain (e.g. online reviews/social media, or clinical/operation notes); 2) Introduction reporting: study aims; 3) Methods reporting: data source, data quality, data pre-processing, missing data, testing/training/internal validation, external validation, model type, and sample size calculation; 4) Results reporting: sample reporting, performance metrics, model evaluation, and model explanation; 5) Conclusions reporting: clinical practice interpretation, limitations, and future research; and 6) Open science: code and data availability.

Quality assessment

To our knowledge, there are no current globally defined reporting guidelines that relate specifically to NLP. We therefore used assessment of compliance to the reporting guidelines outlined by Farrow et al,[1] with each domain categorized as either complete, incomplete, or absent. Code and data availability were assessed separately. The reporting guidelines were chosen due to their specific relation to AI applications in T&O, with inclusion of reporting quality across several domains for the introduction, methods, results, and conclusions separately. An overall cumulative score (total/34) was derived from score tertiles to allow for better interpretability of the final score. Manuscripts with scores < 11, 11 to 22, and 23 to 34 were deemed poor, average, and good quality, respectively. Any disagreement regarding individual scoring of domains between data extractors was resolved by discussion.

Pooled performance metrics

Where feasible, according to study reporting, pooled performance metrics (mean and range) were assessed. This included: model accuracy; sensitivity (recall); specificity; precision (positive predictive value); area under the receiver operating curve (AUROC); F1 score; and calibration. Where scores for multiple cohorts were reported the highest performing model output was chosen for inclusion. All scores were defined by the individual study authors, including decisions around ground-truth labels.

Statistical analysis

Given the nature of the included data, meta-analysis was not feasible and therefore reporting was performed according to the Synthesis Without Meta-Analysis (SWiM) criteria.[12] Studies have been grouped by NLP domain, with assessment of study heterogeneity and evidence certainty determined by the variability of study validity/bias within each domain.

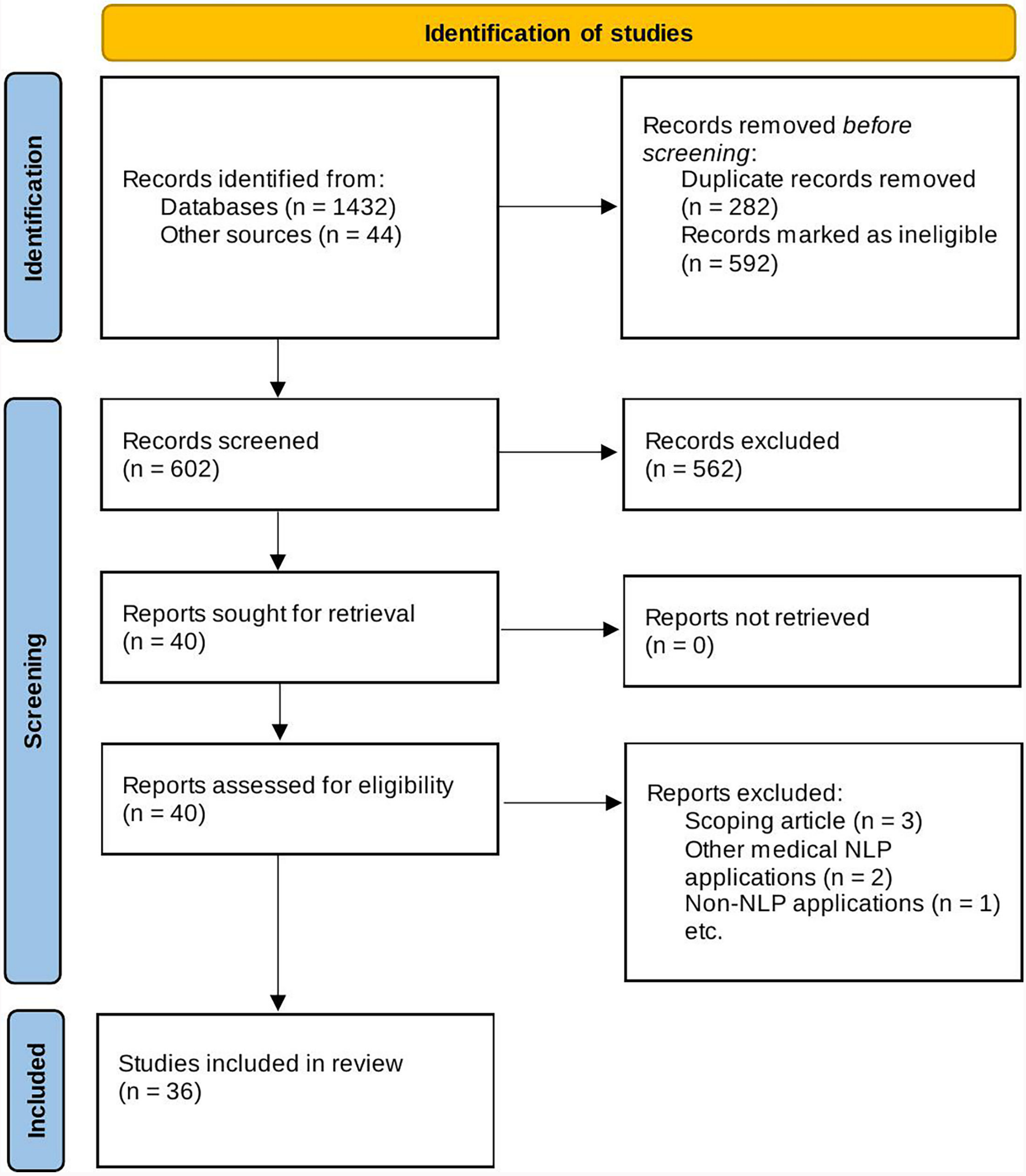
Identification of studies
Identification
Screening
Included
Records identified from:
Databases (n = 1432)
Other sources (n = 44)
Records removed *before screening*:
Duplicate records removed (n = 282)
Records marked as ineligible (n = 592)
Records screened (n = 602)
Records excluded (n = 562)
Reports sought for retrieval (n = 40)
Reports not retrieved (n = 0)
Reports assessed for eligibility (n = 40)
Reports excluded:
Scoping article (n = 3)
Other medical NLP applications (n = 2)
Non-NLP applications (n = 1)
etc.
Studies included in review (n = 36)

Fig. 1

Study selection process. NLP, natural language processing.

Results

Research results

Using the pre-specified search criteria, 602 potentially eligible records were included. Following full-text assessment, 36 manuscripts were included.[8,9,13–46] Figure 1 depicts the flow diagram of the full search process. The number of articles published per year increased from one between 2012 to 2017 to a peak of 13 in 2021 alone.

Characteristics of included studies

Study characteristics, incorporating the quality assessment scoring for each manuscript, are detailed in Table I.

Of the included articles, ten related to trauma, ten to arthroplasty, nine to spinal surgery, three to general orthopaedics, one to foot and ankle surgery, one to shoulder and elbow surgery, one to sports surgery, and one to tumour surgery.

With regards to NLP domains, the most commonly used were clinical or operation notes (50%) and radiology reports (36% each). Use in assessment of online reviews/social media and systematic reviews were less common (11% and 3%, respectively).

Table I. Summary of included studies including reporting assessment.

Design overview				Introduction reporting	Methods reporting								Results reporting			Conclusions reporting			Open science		
first author	Year	Sub-speciality	NLP domain	Study aims	Data source	Data quality	Data pre-processing	Missing data	Test, train, and validation methods	External validation	ML Output	Sample size calculation	Sample reporting	Model evaluation	Model explanation	Clinical practice interpretation	Limitations	Future research	Code availability	Data availability	Overall (/34)
Shah	2020	Arthroplasty	Clinical notes/ operation notes	1	2	0	2	0	2	0	2	0	0	2	0	2	2	2	0	0	17
Mohammadi	2020	Arthroplasty	Clinical notes/ operation notes	2	2	0	2	1	2	0	2	0	2	2	0	2	2	2	0	1	21
Blaker	2021	Trauma	Clinical notes/ operation notes	2	2	2	1	1	1	0	1	2	1	1	0	2	1	0	0	0	15
Karhade	2020	Spine	Clinical notes/ operation notes	2	2	1	2	0	2	0	2	0	1	2	2	2	2	1	0	0	21
Karhade	2020	Spine	Clinical notes/ operation notes	2	2	2	2	1	2	0	1	0	2	2	2	1	2	2	0	0	23
Sagheb	2021	Arthroplasty	Clinical notes/ operation notes	1	2	1	2	0	1	0	1	0	2	1	2	1	2	1	0	0	17
Wyles	2019	Arthroplasty	Clinical notes/ operation notes	2	2	2	2	0	0	2	1	0	2	1	1	2	2	1	2	0	22
Tibbo	2019	Trauma	Clinical notes/ operation notes	2	2	2	2	0	1	0	2	0	1	1	0	2	2	1	0	0	18
Fu	2021	Arthroplasty	Clinical notes/ operation notes	2	2	2	2	0	2	0	2	0	1	2	0	2	2	1	0	0	20
Karhade	2021	Spine	Clinical notes/ operation notes	2	2	0	2	0	0	0	1	0	1	2	2	2	2	2	0	0	17
Thirukumaran	2019	General	Clinical notes/ operation notes	2	2	2	2	0	2	0	1	0	2	2	2	2	2	2	0	0	23
Borjali	2021	Arthroplasty	Clinical notes/ operation notes	2	2	2	2	0	2	0	2	0	0	1	1	2	2	1	0	0	19
Karhade	2020	Spine	Clinical notes/ operation notes	2	2	2	2	0	2	0	1	0	2	2	2	2	2	1	0	0	22
Wyles	2022	Arthroplasty	Clinical notes/ operation notes	1	2	2	2	1	1	2	2	0	1	1	0	2	2	1	0	0	20
Karhade	2022	Spine	Clinical notes/ operation notes	1	2	2	2	0	2	2	2	0	1	2	2	2	2	1	0	0	23
Flores-Balado	2023	Arthroplasty	Clinical notes/ Operation notes	1	2	1	1	0	1	1	2	0	1	1	2	1	1	0	0	0	15
Tavabi	2022	Sports	Clinical notes/ operation notes	1	2	1	1	1	2	1	2	0	0	1	1	1	1	1	0	0	16
Kita	2022	Arthroplasty	Clinical notes/ operation notes	2	2	1	2	0	2	1	1	0	1	1	1	1	2	1	0	0	18
Langerhuizen	2021	General	Online reviews/ social media	2	1	0	2	1	1	0	1	0	1	0	0	1	1	0	0	2	13
Bovonratwet	2021	Arthroplasty	Online reviews/ social Media	2	2	1	1	0	0	0	1	0	2	1	0	2	1	0	0	0	13

(*Continued*)

(*Continued*)

Design overview				Introducti on reporting	Methods reporting								Results reporting			Conclusions reporting			Open science		
Menendez	2019	Shoulder and Elbow	Online reviews/ social media	2	2	0	0	0	0	0	1	0	2	0	0	2	1	0	0	0	10
Dominy	2021	Spine	Online reviews/ social media	2	1	0	1	0	0	0	1	0	1	1	0	2	1	0	0	0	10
Groot	2020	Tumour	Radiology reports for feature detection/classification	1	2	0	2	2	2	0	2	0	2	2	2	2	2	2	0	0	23
dos Santos	2019	Foot and Ankle	Radiology reports for feature detection/classification	2	2	0	1	0	0	0	1	0	2	2	0	2	2	0	0	1	16
Wang	2018	Trauma	Radiology reports for feature detection/classification	1	2	2	2	0	2	0	1	0	0	2	0	1	1	1	0	0	16
Wagholikar	2013	Trauma	Radiology reports for feature detection/classification	2	1	2	1	0	0	0	2	0	0	2	2	2	1	2	0	0	17
Grundmeier	2016	Trauma	Radiology reports for feature detection/classification	2	2	2	2	0	2	0	2	0	1	2	2	2	2	1	0	0	22
Do	2012	Trauma	Radiology reports for feature detection/classification	2	1	0	2	0	1	0	1	0	0	2	0	2	2	2	0	0	15
Kolanu	2021	Trauma	Radiology reports for feature detection/classification	2	2	1	2	0	1	2	2	0	1	2	1	2	2	1	0	0	21
Olthof	2021	Trauma	Radiology reports for feature detection/classification	2	2	2	2	0	2	0	2	0	1	2	2	1	1	0	0	0	19
Foufi	2018	Trauma	Radiology reports for feature detection/classification	1	1	1	1	0	0	0	1	0	0	1	0	1	0	1	0	0	8
Galbusera	2021	Spine	Radiology reports for feature detection/classification	2	2	2	2	0	1	0	2	0	1	2	1	1	2	0	0	0	18
Jungman	2021	Trauma	Radiology reports for feature detection/classification	2	2	2	2	0	1	0	2	0	2	2	0	2	2	0	0	0	19

(*Continued*)

(*Continued*)

Design overview				Introducti on reporting	Methods reporting								Results reporting			Conclusions reporting			Open science		
Tan	2018	Spine	Radiology reports for feature detection/classification	1	2	2	2	0	2	0	2	0	1	2	2	2	2	1	0	0	21
Huhdanpaa	2017	Spine	Radiology reports for feature detection/classification	2	2	2	1	0	2	0	2	0	2	2	1	1	1	0	0	0	14
Buchlak	2021	Arthroplasty	Systematic reviews	2	2	1	2	0	0	0	2	0	2	1	2	1	2	2	1	2	23

For the reporting assessment: 0 indicates domain absent from the manuscript, 1 indicates partial completion, and 2 indicates full completion. Total score across 17 domains = 34 points. Reporting criteria adapted from Farrow et al.[1]

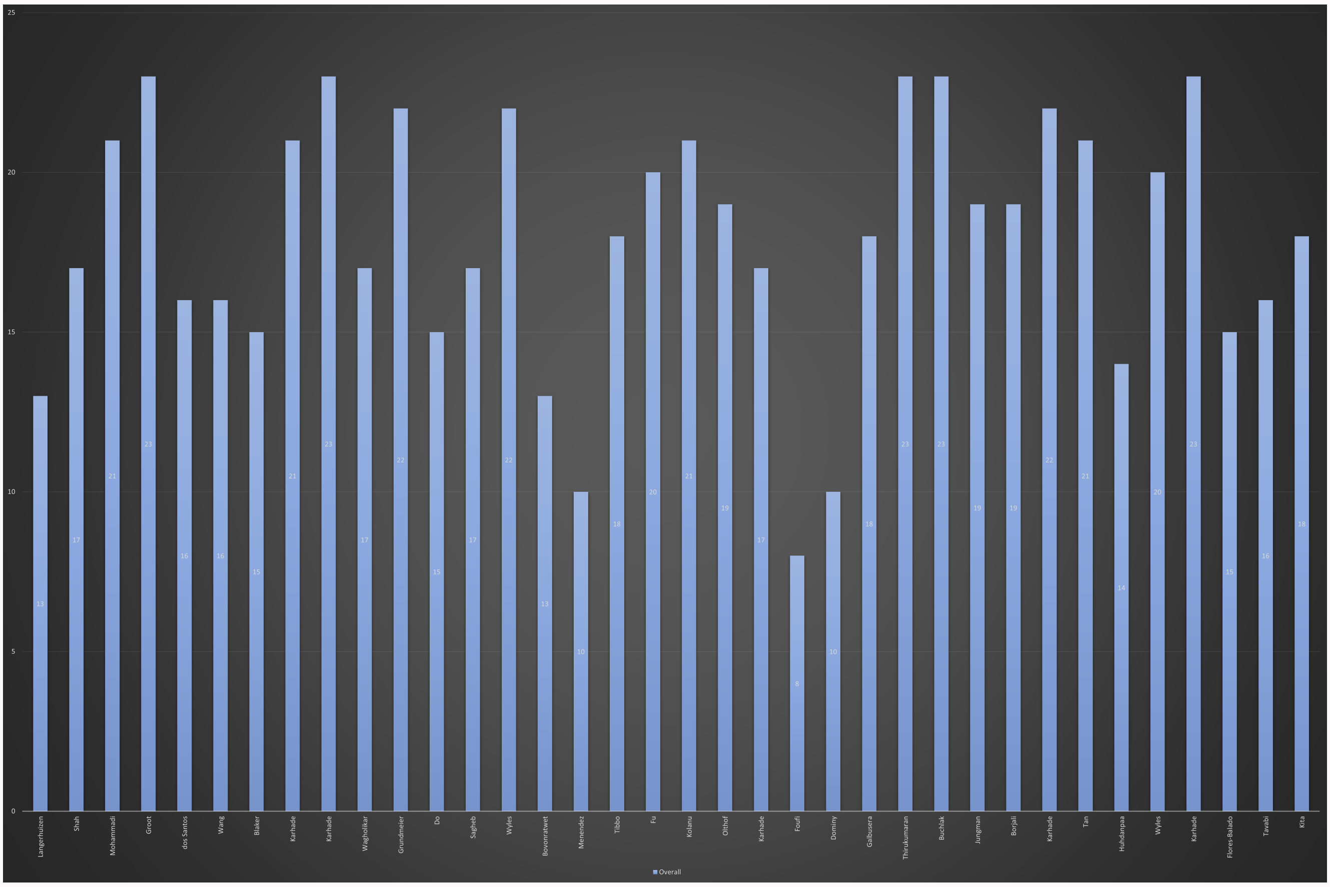

Fig. 2
Summary of overall results.

Overall assessment of study reporting

Of the 36 included studies, the median quality score was 18/34 (IQR 16 to 21); 11% were categorized as good quality, 83% average quality, and 6% poor quality. The most common incomplete study reporting fields were evaluation of missing data, external validation, and a sample size calculation. The top three most frequently completed reporting criteria were study aims, data source, and data pre-processing. Figure 2 demonstrates the bar plot of overall study reporting outcomes.

Reporting domains

Full details of the reporting domains for each individual study are demonstrated in Table I. These are taken from the study by Farrow et al.[1]

Introduction reporting: 26/36 (72% of included studies) had clear documentation of the study aims, with the remainder having at least partial completion.

Methods reporting: All studies at least partially identified their data source, with only 5/36 (14%) providing no details regarding quality of the supplied data. A total of 25/36 (92%) studies fully indicated the preprocessing steps undertaken prior to model training and testing, with only one study providing no preprocessing information. Both missing data and external validation were, however, poorly documented in the majority of studies, with this domain absent from 29/36 (81%) for both fields, respectively. Overall, 9/36 (25%) studies did not provide indication of their testing, training, and validation methods. All studies at least partially reported the type of NLP algorithm output. Only one study provided any form of sample size calculation for model development.

Results reporting: Reporting regarding the sample population was fully performed in 13/36 (36%), with model evaluation fully performed in 21/36 (58%). In all, 15/36 (42%) cases did not provide any reference to explainability of the developed model.

Conclusions reporting: All studies made some reference to potential clinical practice interpretation, with the vast majority (35/36; 97%) describing the study limitations. A total of 11/36 (31%) did not provide any reference regarding requirements for potential future research in their manuscript.

Open Science: Only one study provided the code for algorithm development and testing, with two studies providing the data in an open-source forum.

NLP domain: clinical notes/operation notes

Of the identified studies, 18 related to NLP analysis of clinical or operation notes,[8,14,18–20,24,25,28,29,31,35,38,39,42–46] nine studies related to arthroplasty, five to spinal surgery, two to trauma, one to general orthopaedics, and one to sports surgery. The most common application was to identify adverse outcomes, for example re-admission or surgical complications. Automated database/registry creation was also featured. The median quality assessment for studies in this domain was 19/34 (IQR 17 to 22); 17% were considered good quality and 83% average

Table II. Reported performance metrics.

Study details	Accuracy	Sensitivity (recall)	Specificity	Precision (positive predictive value)	Area under the receiver operating curve	F1 score	Calibration
Shah et al[8]	0.94						
Mohammadi et al[14]		0.79		0.27	0.82		
Groot et al[15]		0.94	0.82	0.97	0.97	0.96	0.73
Dos Santos et al[16]	0.77	0.63	1.00	1.00	0.85		
Wang et al[17]		0.93	1.00	1.00			
Blaker et al[18]				0.77			
Karhade et al[39]		0.89	0.99	0.89	0.99	0.89	1.17
Karhade et al[39]		0.86	0.93	0.51	0.92	0.64	0.61
Wagholikar et al[21]	0.92						
Grundmeier et al[22]	0.95	0.97		0.92		0.95	
Do et al[23]	0.79	0.90	0.95	0.90			
Sagheb et al[24]	0.98	1.00		1.00		1.00	
Wyles et al[25]	0.99						
Tibbo et al[28]		1.00	1.00				
Fu et al[29]		0.89	0.99	1.00		0.91	
Kolanu et al[9]		0.99	1.00	0.97			
Olthof et al[30]	0.96	0.95	0.98		0.99	0.95	
Karhade et al[20]					0.70		1.54
Foufi et al[32]	0.97						
Galbusera et al[34]	0.98	0.95	0.99			0.95	
Thirukumaran et al[35]		0.97		0.97	0.96	0.97	
Buchlak et al[36]					0.68		
Jungman et al[37]		0.81		0.83		0.82	
Borjali et al[38]		1.00		1.00			
Karhade et al[39]		0.94	1.00	0.83			
Tan et al[40]		0.94	0.95		0.98		
Huhdanpaa et al[41]		0.70	0.99	0.90		0.79	
Wyles et al[42]	1.00						
Karhade et al[31]		0.83	0.98	0.79	0.95	0.81	3.08
Flores-Balado et al[44]		0.99	0.91	0.19	0.99	0.32	
Tavabi et al[45]	1.00	1.00	1.00		1.00		
Kita et al[46]	1.00	1.00		0.99			
Mean values	0.94	0.91	0.97	0.84	0.91	0.86	1.43

quality. No study relating to clinical notes or operation notes was identified as poor quality.

NLP domain: radiology reports for feature detection/classification

Several studies (n = 13) related to application of NLP to radiology reports for feature detection and classification.[9,15–17,21–23,30,32,34,37,40,41] Eight related to trauma, three to spinal surgery, one to foot and ankle, and one to tumour. The most common application was the identification of presence or absence of a fracture (± classification). Median quality assessment was 18/34 (IQR 16 to 21); 8% were considered good quality, 84% average quality, and 8% poor quality.

NLP domains: online reviews/social media and systematic review

Four online reviews/social media reports were included,[13,26,27,33] with one study concerning the use of NLP to perform a systematic review (evaluating arthroplasty).[36] One each of the online reviews/social media studies related to general orthopaedics, arthroplasty, shoulder and elbow surgery, and spinal surgery. The main application of NLP to online reviews/social media was automated assessment of the patient experience/feedback using sentiment analysis. Median quality assessment was 12/30 (IQR 10 to 13); 20% were considered good quality, 40% average quality, and 20% poor quality.

Pooled performance metrics

In all, 20/36 studies (56%) reported at least one performance metric outcome. No single study reported results across all the domains assessed. Only five studies reported model calibration. The mean (range) performance metric outcomes for included studies (where reported) are detailed in Table II.

Discussion

The application of NLP to T&O represents a significant opportunity to use the vast quantities of unstructured free-text data generated from routine healthcare interactions, for example in providing summaries of electronic health records or automated analysis of radiology reports. We identified three key domains of current NLP use: clinical/operation notes; radiology reports; and social media/online review posts. Reported performance measure outcomes were almost universally positive (average scores > 80% across all domains); however, there were relatively few high-quality studies identified according to the used reporting criteria. The most problematic areas related to reporting of missing data assessment, external validation, and sample size calculation. Many studies also failed to share the code used as part of the NLP algorithms and report data availability, in accordance with open science principles. Development and widespread use of specific reporting standards related to the application of NLP to healthcare is essential to the appropriate development and reporting of future work in this area.

Our study is, to our knowledge, the first systematic review to focus on applications of NLP in relation to T&O. The results are consistent with reviews of NLP applications in other fields. For example, Davidson et al[47] examined NLP applications in radiology and identified that the key reporting domains that were poorly represented in studies were external validation, data availability, and code availability. The domains of missing data assessment and sample size assessment were not part of the reporting criteria used in that study, but are areas of critical importance to the correct application of NLP techniques for data analysis. It should be noted that currently, despite high-impact publications governing sample calculations for other aspects of AI inference,[48] there are currently no peer-reviewed published guidelines regarding calculation of the optimum sample size for NLP development. This is likely to depend significantly on the NLP approach (for example, large language model (LLM) development/fine-tuning vs a rule-based algorithm), and should be a key research priority moving forwards.

Other applications of AI to T&O appear to suffer from similar issues when considering study reporting. Dijkstra et al[49] evaluated 45 machine learning (ML)-based prediction models and identified that the risk of bias (according to the Prediction model Risk of Bias Assessment Tool, (PROBAST) tool)[50] was high across the majority of included studies, with documented issues around small sample sizes, inadequate management of missing data, and lack of appropriate study reporting.

It therefore appears that the key methodological issues around study design and reporting are consistent across AI applications within T&O. The importance of model calibration appears to be particularly underappreciated, which is likely impacted by limited understanding of AI terminology and interpretation by orthopaedic surgeons.[51] There is a need for a unified and collaborative approach encompassing all key stakeholders (clinicians, data scientists, statisticians, patients, providers) to maximize future applicability. Use of a development and deployment structure is integral to this process and to realizing the potential of NLP applications in the field of T&O.

Limitations of our study include the wide spectrum of different NLP approaches ranging from simple rule-based methods to LLMs. This makes a focused assessment challenging due to the heterogeneity of how these methods are typically applied and reported. Given the lack of currently available validated reporting criteria related to NLP, we used a published non-specific checklist that may be limited in some methodological domains and categorization accuracy. This study does, however, provide the first structured assessment of current applications of NLP within the T&O literature, which provides an understanding of some of the current limitations and subsequent lack of progress towards real-world implementation. It also highlights typical key deficiencies in reporting that can guide improvements in future NLP research.

In conclusion, NLP techniques have significant potential to revolutionize current approaches to data analysis, allowing use and assessment of vast quantities of unstructured free-text data that were previously a largely untapped resource. There are, however, several issues with study design and reporting that must be addressed to realize the potential for clinical practice integration. Appreciation of the importance of model calibration remains low. Sharing of code and data (where feasible) should become part of routine practice in order to maximize transparency in keeping with open science principles.

Supplementary material

Table showing the example search strategy, and the PRISMA checklist.

References

1. **Farrow L**, **Zhong M**, **Ashcroft GP**, **Anderson L**, **Meek RMD**. Interpretation and reporting of predictive or diagnostic machine-learning research in Trauma & Orthopaedics. *Bone Joint J*. 2021;103-B(12):1754–1758.
2. **Bitran H**. From free text to FHIR: text analytics for health launches new feature to boost interoperability. 2022. https://techcommunity.microsoft.com/t5/ai-azure-ai-services-blog/from-free-text-to-fhir-text-analytics-for-health-launches-new/ba-p/3257066 (date last accessed 10 February 2025).
3. **Farrow L**, **Wilde K**, **Dymiter J**, **et al**. Use of "hidden in plain sight" de-identification methodology in electronic healthcare data provides

minimal risk of misidentification: results from the icaird safe haven artificial intelligence platform. *Int J Popul Data Sci*. 2022;25(3):2023.

4. **Ford E**, **Oswald M**, **Hassan L**, **Bozentko K**, **Nenadic G**, **Cassell J**. Should free-text data in electronic medical records be shared for research? A citizens' jury study in the UK. *J Med Ethics*. 2020;46(6):367–377.
5. **Young M**, **Holmes NE**, **Kishore K**, **et al**. Natural language processing diagnosed behavioural disturbance phenotypes in the intensive care unit: characteristics, prevalence, trajectory, treatment, and outcomes. *Crit Care*. 2023;27(1):425.
6. **Li J**, **Hu S**, **Shi C**, **et al**. A deep learning and natural language processing-based system for automatic identification and surveillance of high-risk patients undergoing upper endoscopy: a multicenter study. *EClinMed*. 2022;53:101704.
7. **Marafino BJ**, **Park M**, **Davies JM**, **et al**. Validation of prediction models for critical care outcomes using natural language processing of electronic health record data. *JAMA Netw Open*. 2018;1(8):e185097.
8. **Shah RF**, **Bini S**, **Vail T**. Data for registry and quality review can be retrospectively collected using natural language processing from unstructured charts of arthroplasty patients. *Bone Joint J*. 2020;102-B(7_Supple_B):99–104.
9. **Kolanu N**, **Brown AS**, **Beech A**, **Center JR**, **White CP**. Natural language processing of radiology reports for the identification of patients with fracture. *Arch Osteoporos*. 2021;16(1):6.
10. **Moher D**, **Liberati A**, **Tetzlaff J**, **Altman DG**, **PRISMA Group**. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med*. 2009;6(7):e1000097.
11. **No authors listed**. OrthoSearch. https://orthosearch.org.uk (date last accessed 10 February 2025).
12. **Campbell M**, **McKenzie JE**, **Sowden A**, **et al**. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. *BMJ*. 2020; 368:l6890.
13. **Langerhuizen DWG**, **Brown LE**, **Doornberg JN**, **Ring D**, **Kerkhoffs GMMJ**, **Janssen SJ**. Analysis of online reviews of orthopaedic surgeons and orthopaedic practices using natural language processing. *J Am Acad Orthop Surg*. 2021;29(8):337–344.
14. **Mohammadi R**, **Jain S**, **Namin AT**, **et al**. Predicting unplanned readmissions following a hip or knee arthroplasty: retrospective observational study. *JMIR Med Inform*. 2020;8(11):e19761.
15. **Groot OQ**, **Bongers MER**, **Karhade AV**, **et al**. Natural language processing for automated quantification of bone metastases reported in free-text bone scintigraphy reports. *Acta Oncol*. 2020;59(12):1455–1460.
16. **Pinto Dos Santos D**, **Brodehl S**, **Baeßler B**, **et al**. Structured report data can be used to develop deep learning algorithms: a proof of concept in ankle radiographs. *Insights Imaging*. 2019;10(1):93.
17. **Wang Y**, **Mehrabi S**, **Sohn S**, **Atkinson EJ**, **Amin S**, **Liu H**. Natural language processing of radiology reports for identification of skeletal site-specific fractures. *BMC Med Inform Decis Mak*. 2019;19(Suppl 3):73.
18. **Blaker K**, **Wijewardene A**, **White E**, **et al**. Electronic search programs are effective in identifying patients with minimal trauma fractures. *Osteoporos Int*. 2022;33(2):435–441.
19. **Karhade AV**, **Bongers MER**, **Groot OQ**, **et al**. Natural language processing for automated detection of incidental durotomy. *Spine J*. 2020;20(5):695–700.
20. **Karhade AV**, **Bongers MER**, **Groot OQ**, **et al**. Development of machine learning and natural language processing algorithms for preoperative prediction and automated identification of intraoperative vascular injury in anterior lumbar spine surgery. *Spine J*. 2021;21(10):1635–1642.
21. **Wagholikar A**, **Zuccon G**, **Nguyen A**, **et al**. Automated classification of limb fractures from free-text radiology reports using a clinician-informed gazetteer methodology. *Australas Med J*. 2013;6(5):301–307.
22. **Grundmeier RW**, **Masino AJ**, **Casper TC**, **et al**. Identification of Long Bone Fractures in Radiology Reports Using Natural Language Processing to support Healthcare Quality Improvement. *Appl Clin Inform*. 2016;7(4): 1051–1068.
23. **Do BH**, **Wu AS**, **Maley J**, **Biswal S**. Automatic retrieval of bone fracture knowledge using natural language processing. *J Digit Imaging*. 2013; 26(4):709–713.
24. **Sagheb E**, **Ramazanian T**, **Tafti AP**, **et al**. Use of natural language processing algorithms to identify common data elements in operative notes for knee arthroplasty. *J Arthroplasty*. 2021;36(3):922–926.
25. **Wyles CC**, **Tibbo ME**, **Fu S**, **et al**. Use of natural language processing algorithms to identify common data elements in operative notes for total hip arthroplasty. *J Bone Joint Surg Am*. 2019;101-A(21):1931–1938.
26. **Bovonratwet P**, **Shen TS**, **Islam W**, **Ast MP**, **Haas SB**, **Su EP**. Natural language processing of patient-experience comments after primary total knee arthroplasty. *J Arthroplasty*. 2021;36(3):927–934.
27. **Menendez ME**, **Shaker J**, **Lawler SM**, **Ring D**, **Jawa A**. Negative patient-experience comments after total shoulder arthroplasty. *J Bone Joint Surg Am*. 2019;101-A(4):330–337.
28. **Tibbo ME**, **Wyles CC**, **Fu S**, **et al**. Use of natural language processing tools to identify and classify periprosthetic femur fractures. *J Arthroplasty*. 2019;34(10):2216–2219.
29. **Fu S**, **Wyles CC**, **Osmon DR**, **et al**. Automated detection of periprosthetic joint infections and data elements using natural language processing. *J Arthroplasty*. 2021;36(2):688–692.
30. **Olthof AW**, **Shouche P**, **Fennema EM**, **et al**. Machine learning based natural language processing of radiology reports in orthopaedic trauma. *Comput Methods Programs Biomed*. 2021;208:106304.
31. **Karhade AV**, **Lavoie-Gagne O**, **Agaronnik N**, **et al**. Natural language processing for prediction of readmission in posterior lumbar fusion patients: which free-text notes have the most utility? *Spine J*. 2022;22(2): 272–277.
32. **Foufi V**, **Lanteri S**, **Gaudet-Blavignac C**, **Remy P**, **Montet X**, **Lovis C**. Automatic annotation tool to support supervised machine learning for scaphoid fracture detection. *Stud Health Technol Inform*. 2018;255:210–214.
33. **Dominy CL**, **Arvind V**, **Tang JE**, **et al**. Scoliosis surgery in social media: a natural language processing approach to analyzing the online patient perspective. *Spine Deform*. 2022;10(2):239–246.
34. **Galbusera F**, **Cina A**, **Bassani T**, **Panico M**, **Sconfienza LM**. Automatic diagnosis of spinal disorders on radiographic images: leveraging existing unstructured datasets with natural language processing. *Glob Spine J*. 2023;13(5):1257–1266.
35. **Thirukumaran CP**, **Zaman A**, **Rubery PT**, **et al**. Natural language processing for the identification of surgical site infections in orthopaedics. *J Bone Joint Surg Am*. 2019;101-A(24):2167–2174.
36. **Buchlak QD**, **Clair J**, **Esmaili N**, **Barmare A**, **Chandrasekaran S**. Clinical outcomes associated with robotic and computer-navigated total knee arthroplasty: a machine learning-augmented systematic review. *Eur J Orthop Surg Traumatol*. 2022;32(5):915–931.
37. **Jungmann F**, **Kämpgen B**, **Hahn F**, **et al**. Natural language processing of radiology reports to investigate the effects of the COVID-19 pandemic on the incidence and age distribution of fractures. *Skeletal Radiol*. 2022;51(2):375–380.
38. **Borjali A**, **Magnéli M**, **Shin D**, **Malchau H**, **Muratoglu OK**, **Varadarajan KM**. Natural language processing with deep learning for medical adverse event detection from free-text medical narratives: a case study of detecting total hip replacement dislocation. *Comput Biol Med*. 2021;129:104140.
39. **Karhade AV**, **Bongers MER**, **Groot OQ**, **et al**. Can natural language processing provide accurate, automated reporting of wound infection requiring reoperation after lumbar discectomy? *Spine J*. 2020;20(10): 1602–1609.
40. **Tan WK**, **Hassanpour S**, **Heagerty PJ**, **et al**. Comparison of natural language processing rules-based and machine-learning systems to identify lumbar spine imaging findings related to low back pain. *Acad Radiol*. 2018;25(11):1422–1432.
41. **Huhdanpaa HT**, **Tan WK**, **Rundell SD**, **et al**. Using natural language processing of free-text radiology reports to identify type 1 modic endplate changes. *J Digit Imaging*. 2018;31(1):84–90.
42. **Wyles CC**, **Fu S**, **Odum SL**, **et al**. External validation of natural language processing algorithms to extract common data elements in tha operative notes. *J Arthroplasty*. 2023;38(10):2081–2084.
43. **Karhade AV**, **Oosterhoff JHF**, **Groot OQ**, **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):1766–1775.
44. **Flores-Balado Á**, **Castresana Méndez C**, **Herrero González A**, **et al**. Using artificial intelligence to reduce orthopedic surgical site infection surveillance workload: Algorithm design, validation, and implementation in 4 Spanish hospitals. *Am J Infect Control*. 2023;51(11):1225–1229.
45. **Tavabi N**, **Pruneski J**, **Golchin S**, **et al**. Building large-scale registries from unstructured clinical notes using a low-resource natural language processing pipeline. *Health Informatics*. 2022.

46. **Kita K**, **Uemura K**, **Takao M**, **et al**. Use of artificial intelligence to identify data elements for The Japanese Orthopaedic Association National Registry from operative records. *J Orthop Sci*. 2023;28(6):1392–1399.

47. **Davidson EM**, **Poon MTC**, **Casey A**, **et al**. The reporting quality of natural language processing studies: systematic review of studies of radiology reports. *BMC Med Imaging*. 2021;21:142.

48. **Riley RD**, **Ensor J**, **Snell KIE**, **et al**. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441.

49. **Dijkstra H**, **van de Kuit A**, **de Groot T**, **et al**. Systematic review of machine-learning models in orthopaedic trauma. *Bone Jt Open*. 2024;5(1):9–19.

50. **Moons KGM**, **Wolff RF**, **Riley RD**, **et al**. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Ann Intern Med*. 2019;170(1):W1–W33.

51. **Ormond MJ**, **Clement ND**, **Harder BG**, **Farrow L**, **Glester A**. Acceptance and understanding of artificial intelligence in medical research among orthopaedic surgeons. *Bone Jt Open*. 2023;4(9):696–703.

Author information

L. Farrow, MBChB, BSc(Intercalated), MRCS, SCREDS Clinical Lecturer,
Institute of Applied Health Sciences, University of Aberdeen, Aberdeen, UK;
Grampian Orthopaedics, Aberdeen Royal Infirmary, Aberdeen, UK.

A. Raja, Medical Student, School of Medicine, University of Edinburgh, Edinburgh, UK.

M. Zhong, PhD, Lecturer Computer Science
L. Anderson, PhD MPHe BSc(Hons) PGCHET FHEA, Professor of Health Data Science
Institute of Applied Health Sciences, University of Aberdeen, Aberdeen, UK.

Author contributions

L. Farrow: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing, He is the Guarantor..
A. Raja: Conceptualization, Data curation, Investigation, Project administration, Writing – review & editing.
M. Zhong: Conceptualization, Supervision, Writing – review & editing.
L. Anderson: Conceptualization, Methodology, Supervision, Writing – review & editing.

Funding statement

The author(s) disclose receipt of the following financial or material support for the research, authorship, and/or publication of this article: funding for open access publication is provided by the Chief Scientist Office.

ICMJE COI statement

The author(s) disclose receipt of the following financial or material support for the research, authorship, and/or publication of this article: L. Farrow is currently in receipt of a Chief Scientist Office Scotland Clinical Academic Fellowship, which is focused on the use of artificial intelligence techniques (including natural language processing) to improve the clinical care pathway in those referred for hip and knee arthroplasty. L. Farrow is also the guarantor and confirms that all listed authors meet the authorship criteria.

Data sharing

The data that support the findings for this study are available to other researchers from the corresponding author upon reasonable request.

Ethical review statement

Ethical review was not required due to the nature of the study.

Open access funding

The open access funding for this paper was provided by the Chief Scientist Office (ref. CAF 21/06).