



OPEN ACCESS

EDITED BY

Phoey Lee Teh,
Glyndŵr University, United Kingdom

REVIEWED BY

Daisy Monika Lal,
Lancaster University, United Kingdom
Chukwudi Uwasomba,
The Open University, United Kingdom

*CORRESPONDENCE

Sarah B. Floyd
✉ sbf@clermson.edu

RECEIVED 04 December 2024

ACCEPTED 26 March 2025

PUBLISHED 24 April 2025

CITATION

Floyd SB, Almeldien AG, Smith DH, Judkins B, Krohn CE, Reynolds ZC, Jeray K and Obeid JS (2025) Using artificial intelligence to develop a measure of orthopaedic treatment success from clinical notes.

Front. Digit. Health 7:1523953.

doi: 10.3389/fdgth.2025.1523953

COPYRIGHT

© 2025 Floyd, Almeldien, Smith, Judkins, Krohn, Reynolds, Jeray and Obeid. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Using artificial intelligence to develop a measure of orthopaedic treatment success from clinical notes

Sarah B. Floyd^{1*}, Ahmed G. Almeldien², D. Hudson Smith³, Benjamin Judkins⁴, Claire E. Krohn⁴, Zachary Cole Reynolds⁴, Kyle Jeray⁴ and Jihad S. Obeid⁵

¹Department of Public Health Sciences, Clemson University, Clemson, SC, United States, ²Biomedical Data Science and Informatics Program, Clemson University, Clemson, SC, United States, ³Department of Mathematical and Statistical Sciences, Clemson University, Clemson, SC, United States, ⁴Department of Orthopaedic Surgery, Prisma Health, Greenville, SC, United States, ⁵Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC, United States

Introduction: A readily available outcome measure that reflects the success of a patient's treatment is needed to demonstrate the value of orthopaedic interventions. Patient-reported outcome measures (PROMs) are survey-based instruments that collect joint-specific and general health perceptions on symptoms, functioning, and health-related quality of life. PROMs are considered the gold standard outcome measure in orthopaedic medicine, but their use is limited in real-world practice due to challenges with technology integration, the pace of clinic workflows, and patient compliance. Clinical notes generated during each encounter patients have with their physician contain rich information on current disease symptoms, rehabilitation progress, and unexpected complications. Artificial intelligence (AI) methods can be used to identify phrases of treatment success or failure captured in clinical notes and discern an indicator of treatment success for orthopaedic patients.

Methods: This was a cross-sectional analysis of clinical notes from a sample of patients with an acute shoulder injury. The study included adult patients presenting to a Level-1 Trauma Center and regional health system for an acute Proximal Humerus Fracture (PHF) between January 1, 2019 and December 31, 2021. We used the progress note from the office visit for PHF-related care (ICD10: S42.2XXX) or shoulder pain (ICD10: M45.2XXX) closest to 1-year after the injury date. Clinical notes were reviewed by an orthopaedic resident and labeled as treatment success or failure. A structured comparative analysis of classifiers including both machine and deep learning algorithms was performed.

Results: The final sample included 868 clinical notes from patients treated by 123 physicians across 35 departments within one regional health system. The study sample was stratified into 465 notes labeled as treatment success and 403 labeled as treatment failure. The Bio-ClinicalBERT model had the highest performance of 87% accuracy (AUC = 0.87 ± 0.04) in correctly distinguishing between treatment success and failure notes.

Discussion: Our results suggest that text classifiers applied to clinical notes are capable of differentiating patients with successful treatment outcomes with high levels of accuracy. This finding is encouraging, signaling that routinely collected clinical note content may serve as a data source to develop an outcome measure for orthopaedic patients.

KEYWORDS

proximal humerus fractures, artificial intelligence, treatment outcomes, clinical text, natural language processing, electronic health record system

Introduction

A readily available outcome measure that reflects the success of a patients' treatment is needed to demonstrate the value of orthopaedic interventions (1–4). PROMs are viewed as the gold standard outcome measure in orthopaedics (5, 6), and could be used to track improvements in patient status over time to assess whether a patient achieved their treatment goals (1, 7, 8). However, PROM collection in real-world practice has proven difficult due to challenges with technology integration, data collection disruptions to clinic workflows, and low patient compliance completing PROM questionnaires (9, 10). Most orthopaedic practices in the United States view the collection of PROMs as time consuming, cumbersome, and costly (4, 11), and it is estimated that only half of orthopaedic practices collect PROMs (9, 12). Without widespread availability of PROMs, evaluating the success of orthopaedic treatments has been limited to process-based measures and end points such as survival and surgical complications, which can readily be found in structured fields in administrative data (4). Unfortunately, these data points paint an incomplete picture of a patient's medical experience and whether treatment was successful in achieving a patient's prioritized outcome goals (e.g., reduced pain, improved range of motion, return to sports or work, etc.) (13–16). A crucial step to demonstrating if orthopaedic care achieves meaningful patient-specific improvement is to have outcome data available that measure if treatment was successful in achieving the outcome goals prioritized by patients.

Unstructured clinical notes represent untapped potential for evaluating the success of orthopaedic treatments for individual patients (17–24). Clinical notes are generated for each encounter patients have with their physician. These notes contain rich information on current disease symptoms, rehabilitation progress, and unexpected complications, and captures an essence of the patient state that is not stored elsewhere in the Electronic Health Record (EHR) (19, 25–30). However, because these notes are in an unstructured format, they have not historically been thought of as a useful data source to develop consistent measures of patient treatment success. Applications of artificial intelligence (AI) and natural language processing (NLP) methods in healthcare are rapidly expanding, and have been used to identify similar patient groups with a defined set of clinical markers and symptoms, automate patient cohort selection, and develop predictive models (17, 25, 31–33). In orthopaedic medicine, NLP applications are being deployed to solve a range of orthopaedic-related problems (34–36). We proposed that NLP methods can be employed to identify phrases denoting treatment success or failure recorded in clinical notes, thereby discerning indicators of treatment success for orthopaedic patients.

Clinical notes are typically documented using a subjective, objective, assessment and plan (SOAP) format, where the subjective section of the progress note reflects the patient's story (the interim history since the last visit), told to and interpreted by the physician (28). The rest of the unstructured SOAP notes include physical findings (objective findings), medical reasoning (assessment), and patient care (the plan) and (37–39) reveal distinct trajectories of patient outcomes after treatment (38, 40). In

successful cases the progress note documents the degree of improvement or relief experienced and reported by patients to their clinicians (30). Conversely when symptoms have not resolved, or when subsequent complications arise, these ongoing patient complaints and persistent treatment utilization are documented in the notes. Note content highlights symptoms and outcome dimensions, which are valued by patients and contribute to a patient's individual definition of success (41).

A patient-specific outcome measure is needed in orthopaedic medicine to assess if a patient's treatment goals were met and help demonstrate the value of orthopaedic treatments. The objective of this study was to develop a clinical text classifier, using different AI approaches, including traditional machine learning and deep learning, capable of identifying whether a patient achieved treatment success for a sample of patients receiving follow-up care for an acute proximal humerus fracture. Proximal humerus fracture (PHF) was chosen as a use case for testing this method of outcome generation because many different treatment options can be used for this injury and better outcome data are needed to guide future clinical decision-making for this condition (42–45). To that end, we began by exploring the feasibility of generating an outcome measure from clinical notes by evaluating various AI-based text classifiers for patients treated for a PHF.

Materials and methods

Study sample

This was a cross-sectional analysis of clinical notes from a sample of patients with an acute PHF. The study included adult patients presenting to a Level-1 Trauma Center regional health system for an acute PHF between January 1, 2019 and December 31, 2021. The first visit for a patient for PHF during the study period was defined as the index visit. We then identified all health system encounters (hospital encounters, office visits, etc.) with a diagnosis of PHF or shoulder pain from the index PHF visit to 365 days after the index PHF visit and extracted the clinical note from the office visit for PHF-related care (ICD10: S42.2XXX) or shoulder pain (ICD10: M45.2XXX) closest to 1-year after the injury date. This resulted in one representative note per person. Patients were excluded from the study if they were less than 18 years of age, did not have at least one office visit with a diagnosis of PHF or shoulder pain that occurred 45 days or more days after the index visit, or if their last office visit was less than 500 characters, since smaller notes did not have the necessary sections for outcome information. We found that less than 1% of the eligible sample had a note length less than 500 characters. A minimum of 45 days after index was used as this is the minimal time needed for healing of a PHF, prior to which, treatment success cannot be assessed. We randomly sampled 1,000 patients meeting these inclusion criteria, and after physician review of the sample, a total of 868 notes were used for this study. Additional exclusions after physician review included notes from patients receiving care for bilateral PHFs or notes that did not reflect care for PHF. This study was approved by the Prisma Health Institutional Review Board.

Data labeling process

The University of South Carolina Patient Engagement Studio brings together patients and caregivers, community groups, health system innovators, clinicians, and academic researchers to produce meaningful research that advances health research outcomes. The PES membership includes over 100 patients with diverse backgrounds and clinical experiences from across the United States trained to provide feedback and collaborate with research teams (46, 47). The first author (S.F.) led three Patient Engagement Studio (PES) sessions to involve patients in the development of the definitions of treatment success. There is no one standard acceptable definition of treatment success, as it varies based upon patient lifestyle and desired goals (4, 48). However, generally, the goal of orthopaedic treatment is to restore the function of the joint, minimize pain and maximize quality of life for patients after the injury. Together, the PES members and research team defined four outcome states and associated labels that described the range of positive and negative outcomes that could follow care for PHF. The outcome states, definitions, and note text indicators can be found in Table 1.

Four orthopaedic residents, each with a minimum of 2 years of experience in orthopaedic medicine participated in the labeling of the clinical notes. Each orthopaedic resident received a 1-h training session on the note labels and definitions. Residents were instructed to label each note based on the outcome state reflected in the current visit and documented in the note. In addition to the outcome labels, orthopaedic residents had the option to select “Indeterminant—not enough content to label” or flag a note for expert review. We assigned overlap in note review subsets, so that agreement could be assessed across residents. An attending orthopaedic surgeon, the Chair of Department of Orthopaedic Surgery, served as the final arbitrator when discordance occurred between residents’ labels or when expert review was requested. Each label provided by the Chair was viewed as the final label for that note.

REDCap (49) was used to organize and store notes and labels. Notes were truncated to 10,000 words and HTML tags were added for formatting to aid in the review process. The final labeled dataset produced by the orthopaedic residents represented the gold standard dataset used for model development and validation. The four outcome labels were aggregated into a binary outcome representing treatment success or failure. Treatment success was represented by notes labeled as success. Failure was comprised of all remaining labels (improvement, deterioration and failure) as each of those notes had documentation of lingering, symptomatic problems requiring ongoing care. This was done to avoid class imbalance between the treatment success and failure classes, since our data set had many more cases in the treatment success category compared to the remaining three unfavorable outcome categories.

Stratified data segmentation protocol

We maintained a consistent data division strategy to ensure a fair and unbiased comparison across all models, so 90% of the data was allocated for model training, while the remaining unseen 10% was reserved for testing. Within these partitions, we preserved the proportion of outcome labels observed in the full labeled data set. We used a five-fold cross-validation strategy to robustly measure the generalizability of all models. Moreover, due to the relatively small number of cases, we repeated the testing cycle 10 times through train and test data partitioning (Figure 1). A critical aspect of our experimental design was ensuring fair model comparison by using identical data splits across all models within each run. For each of the ten experimental iterations, we created a single stratified train-test split that was subsequently used for all models being evaluated. This approach eliminates potential confounding effects from data

TABLE 1 Outcome labels and corresponding definition and indicators.

Outcome state	Definition	Indicators of outcome state found in clinical notes
Treatment success	Treatment success occurs when a patient is able to resume desired activities, has a sufficient range of motion, and is in minimal/mild or no pain. After PHF it is possible for there to be some lingering motion limitations (patient may never return to 100%) or minimal pain, but these issues should not require ongoing treatment or be prohibitive to their desired lifestyle or daily activities.	• Radiographic healing • Making good progress/improvements with current treatment or stopping treatment • Patient returns to work or play. • No major complaints documented. • Only follow-up as needed
Improvement of condition	Improvement occurs when there is a record of some levels of pain or functional problems that are somewhat prohibitive to the patient’s desired activities, but improvement is occurring. In these situations, physicians may continue to monitor patients, but do not alter care or treatment courses.	• Radiographic healing or signs of healing occurring. • Moderate loss of function or pain which interferes with desired activities, but no change in treatment. • Ongoing treatment and monitoring progress • Return in 2–6 weeks for repeat x-rays and re-check
Deterioration of condition	Deterioration occurs when there is a record of some levels of pain or functional problems that are becoming more prohibitive to the patient’s desired activities. No real improvements are occurring, and physicians may escalate or alter care or treatment courses.	• Negative radiographic changes observed. • Moderate loss of function or pain which interferes with desired activities requiring a change in treatment. • Initiating or continuing treatment and monitoring progress. • Return in 2–6 weeks for repeat x-rays and re-check.
Treatment Failure	Treatment failure occurs when the patient is experiencing significant pain or limitations and requires subsequent fracture-related care. Failing is occurring when patients are unable to resume desired activities and may include fracture sequelae, complications, or nonunion.	• Ongoing, persistent treatment (injections, surgeries) for symptoms related to PHF. • Unrelenting pain • Surgical complications • Loss of significant motion • Extreme pain • Fracture related sequelae (e.g., avascular necrosis)

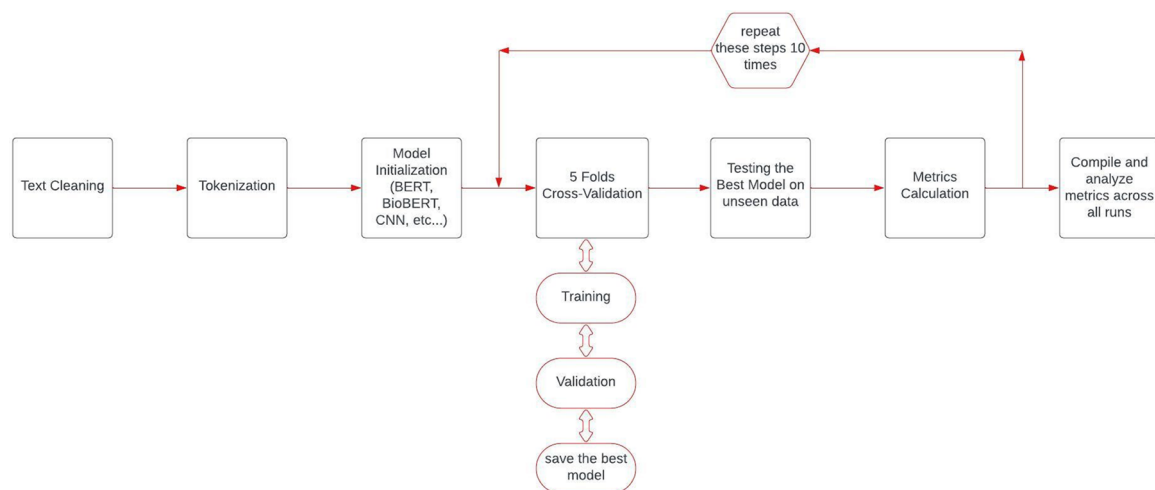


FIGURE 1

AI model training and evaluation pipeline for treatment success measure assessment.

partitioning and ensures that any observed performance differences can be attributed to the models themselves rather than to variations in training or testing data.

Bag-of-words based models

We tested the performance of several baseline classifiers that utilize a bag-of-words approach, primarily to provide a reference point for comparing more advanced modeling techniques. For these traditional machine learning models, we employed GridSearchCV with 5-fold cross-validation to systematically search for optimal hyperparameters. The text data were vectorized using term frequency-inverse document frequency (TF-IDF) (50) with parameters set to ignore terms appearing in more than 90% of documents ($\text{max_df} = 0.9$), include terms appearing in at least 5 documents ($\text{min_df} = 5$), and incorporate both unigrams and bigrams [$\text{ngram_range} = (1,2)$]. Text preprocessing was consistent across all models, involving lowercase conversion, special character removal (except for periods, numbers, and alphabets), whitespace normalization, and lemmatization for words longer than three characters. This approach preserved medical abbreviations that might be clinically significant while reducing vocabulary size and standardizing word forms.

After applying TF-IDF vectorization, we tested five distinct machine learning classifiers: Logistic Regression, with an inverse regularization strength (C) equal to 0.1; Support Vector Machine with Linear Kernel (51); Random Forest model with 200 trees, with each tree having a maximum depth of 10 and using “entropy” as the criterion for measuring the quality of splits (52); Gradient Boosting Classifier with 200 boosting stages and a learning rate of 0.01, implying effectiveness in gradual learning over numerous iterations (53); and XGBClassifier with 100 estimators, a learning rate of 0.01, a maximum tree depth of 4, and a subsample ratio of 0.9, indicating a preference for moderate complexity and high sample coverage in training (54).

Transformer-based classifiers

In this study, we adopted a family of transformer-based language models to analyze clinical note content, primarily because transformers utilize a multi-head self-attention mechanism that pinpoints how different words in a sequence relate to each other, even when they are far apart. This mechanism effectively captures non-linear contextual dependencies by allowing multiple “heads” to attend to different segments of the input in parallel, thereby highlighting important clinical cues that might not be locally adjacent. Of the transformer-based models (55), we used BERT (56), BioBERT (57), and Bio-ClinicalBERT (58), and to analyze note content. We selected BERT as our base architecture model due to its bidirectional training paradigm, which uses both preceding and succeeding words for context. Building upon BERT, BioBERT further refines the model with additional pre-training on domain-specific biomedical corpora, effectively incorporating specialized terminology. Similarly, Bio-ClinicalBERT further refines BioBERT by fine-tuning on the Medical Information Mart for Intensive Care (MIMIC-III) (59) clinical notes dataset to augment its capability in deciphering medical terminology and understanding complex clinical semantics. Each of our transformer-based models accepted a maximum of 512 tokens. We selected the last 512 tokens instead of the first 512 tokens based on our observation that later parts of the note typically contained the most-clinically relevant information about ongoing symptoms and necessary treatment or discharge from care.

For these transformer-based models, we employed a consistent training approach. Each model was trained for 8 epochs with a batch size of 8, using the AdamW optimizer (71) with a learning rate of 3×10^{-5} and epsilon of 1×10^{-8} . To optimize the learning rate schedule, we implemented a cosine annealing learning rate scheduler that gradually reduced the learning rate from the initial value to zero throughout the training process. We applied gradient clipping with a maximum norm of 1.0 to prevent exploding gradients during backpropagation. This technique

helped stabilize the training process, particularly important for the fine-tuning of pre-trained language models on domain-specific data. The transformer-based models processed our common data splits through their respective tokenization pipelines, ensuring consistent evaluation across experimental runs.

Convolutional neural network classifier

We also tested the performance of Convolutional Neural Network (CNN)-based classifiers for textual data, primarily because CNNs excel at extracting localized patterns from sequences—such as *n*-gram features—by sliding convolutional filters over the embedded text (60). We first converted the text into integer sequences and right padded these to a fixed sequence length (61, 62). Our CNN architecture was specifically designed for text classification, featuring an embedding layer followed by multiple parallel convolutional layers with varying filter sizes (3–12) to capture different *n*-gram patterns in the text. Each convolutional layer utilized 200 filters and was connected to an adaptive max pooling layer. The architecture incorporated a three-layer fully connected network (200 units → 100 units → 1 unit) with ReLU activations and dropout (rate = 0.5) for regularization. The CNN model was trained for up to 50 epochs with early stopping (patience = 5) using the Adam optimizer with an initial learning rate of 0.001 and a step learning rate scheduler (step size = 5, gamma = 0.1). Unlike the transformer models, the CNN implementation used a custom tokenizer and a batch size of 16, which resulted in a different processing pipeline for the same underlying data splits.

Statistical analysis and model evaluation

Fleiss' kappa statistics were used to assess the degree of agreement in note labels between orthopaedic residents. We used the benchmarks for agreement measures for categorical data as described by Landis and Koch, where 0.00–0.20, 0.21–0.40, 0.41–0.60, 0.61–0.80, and 0.81–1.00 indicate poor, fair, moderate, substantial, and almost perfect agreement, respectively (63). Patient characteristics including patient age, sex, race, insurance status, fracture diagnosis, and visit characteristics were extracted from the EHR and used to assess differences in patient and visit characteristics associated with note labels. Analyses were performed with SAS (Cary, NC) version 15.2, R studio, and excel.

To evaluate the performance of each model on the binary outcome measure we evaluated the F1 score, sensitivity, specificity, positive and negative predictive value, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) metric. The AUC-ROC metric provides a single value that encapsulates the trade-off between the true positive rate and the false positive rate at various decision threshold levels. Model comparisons are displayed using ordered Box and Whisker plots where each box represents the interquartile range (IQR) for a specific model's AUC scores and data points represent outliers. To assess statistical significance between model performance, we conducted paired

t-tests between models that shared identical data splits and unpaired *t*-tests (Welch's *t*-test) for comparisons involving the CNN model. Since the transformer-based and traditional machine learning models used identical data splits in each run, paired *t*-tests were appropriate for these comparisons. In contrast, due to the CNN model's different tokenization and processing pipeline, unpaired *t*-tests were necessary for comparing CNN with other models. A significance level of $p < 0.05$ was set.

Results

Progress note characteristics

The sample of 868 clinical notes came from patients treated by 123 physicians across 35 departments within one regional health system. The average age of the patient was 67.5 years of age and 80% of the sample were female patients. Patients were primarily white (91%) and enrolled in the Medicare system (64%). The notes came from fracture-related encounters ranging from the 2nd to the 22nd encounter after the index PHF visit, with a mean of 4, representing the 4th PHF-related encounter. The note lengths ranged from 981 to 15,297 characters with a mean length of 4,961 characters.

The sample was stratified into 465 notes labeled as treatment success and 403 notes labeled as treatment failure. Patients experiencing treatment success did not differ from those labeled as treatment failure in age, sex, or race. Additionally, treatment success notes did not differ in time since index visit, or the number of PHF-related encounters. However, notes labeled as treatment success had a higher proportion of Medicare (65.8% vs. 62.3%) and commercially insured patients (23.4% vs. 22.1%), compared to notes labeled as treatment failure. Additionally, treatment success notes had a lower proportion of surgically treated patients (16.3% vs. 24.3%) than notes labeled as treatment failure. Treatment success notes were also significantly shorter in length (4,673 vs. 5,291 characters). Table 2 contains sample characteristics by note label status.

Orthopaedic resident agreement in note labeling

Each orthopaedic resident was assigned and labeled 268 notes with overlap in note assignment across orthopaedic residents. Interrater agreement for classification of clinical notes was moderate [pairwise percent agreement = 75.31%, Fleiss' $\kappa = 0.49$ (95% CI: 0.30–0.68)].

Model performance metrics

Table 3 shows results for both machine and deep learning models explored in our analysis. The XGB Classifier model achieved the highest AUC score (0.84 ± 0.03) among the bag-of-words models and the Bio-ClinicalBERT model had the highest AUC (0.87 ± 0.04) among the deep learning models. Overall, the

TABLE 2 Sample characteristics by note label Status.

Patient and note characteristics	Total sample	Treatment success notes	Treatment failure notes	P-value
	N (%)	N (%)	N (%)	
	N = 868 (100)	N = 465 (53.4)	N = 403 (46.4)	
Patient Age, Mean (SD)	67.5 (14.9)	68.1 (15.1)	66.8 (14.7)	0.19
Patient Sex, N (%)				0.18
Male	175 (20.2)	86 (18.5)	89 (22.1)	
Female	693 (79.8)	379 (81.5)	314 (77.9)	
Patient Race, N (%)				0.68
White	791 (91.1)	427 (91.8)	364 (90.3)	
Black	46 (5.3)	23 (4.9)	23 (5.7)	
American Indian, Alaskan, or Hawaiian	4 (1.0)	3 (0.6)	1 (0.2)	
Hispanic	13 (1.5)	6 (1.3)	7 (1.7)	
Asian	4 (0.5)	2 (0.4)	2 (0.5)	
Other	10 (1.0)	4 (1.0)	6 (1.5)	
Insurance Provider				0.04
Medicare	557 (64)	306 (65.8)	251 (62.3)	
Medicaid	37 (4.2)	16 (3.4)	21 (5.2)	
Private	198 (23)	109 (23.4)	89 (22.1)	
Other	70 (8.1)	34 (7.3)	36 (8.9)	
Note Characteristics				
Days from Index, Mean (SD)	135 (79.9)	131.2 (72.0)	140.8 (87.8)	0.08
PHF-related encounter, Mean (SD)	4.4 (1.8)	4.4 (1.6)	4.5 (2.01)	0.26
Patient treated surgically, N (%)	174 (20)	76 (16.3)	98 (24.3)	<0.01
Note character length, Mean (SD)	4,961 (2,725)	4,673.4 (2,511.9)	5,294.1 (2,921.3)	<0.001

TABLE 3 Performance metrics for Bag-of-word and deep learning models.

Model name	F1 score	Specificity	Sensitivity	Precision	NPV	AUC
Bag-of-words models						
Logistic regression	0.70 (± 0.03)	0.64 (± 0.03)	0.82 (± 0.04)	0.72 (± 0.02)	0.73 (± 0.04)	0.80 (± 0.03)
SVC	0.72 (± 0.03)	0.60 (± 0.03)	0.81 (± 0.03)	0.71 (± 0.02)	0.74 (± 0.03)	0.79 (± 0.03)
Random forest	0.72 (± 0.03)	0.68 (± 0.03)	0.80 (± 0.02)	0.73 (± 0.02)	0.73 (± 0.03)	0.80 (± 0.01)
Gradient boosting classifier	0.75 (± 0.02)	0.70 (± 0.02)	0.81 (± 0.02)	0.75 (± 0.02)	0.75 (± 0.02)	0.82 (± 0.01)
XGBClassifier	0.75 (± 0.02)	0.70 (± 0.06)	0.78 (± 0.03)	0.75 (± 0.03)	0.74 (± 0.03)	0.84 (± 0.03)
Deep learning models						
BERT	0.76 (± 0.08)	0.76 (± 0.04)	0.75 (± 0.06)	0.79 (± 0.05)	0.72 (± 0.08)	0.83 (± 0.05)
BioBERT	0.75 (± 0.03)	0.71 (± 0.07)	0.77 (± 0.04)	0.77 (± 0.05)	0.73 (± 0.03)	0.81 (± 0.03)
Bio-ClinicalBERT	0.77 (± 0.04)	0.74 (± 0.09)	0.80 (± 0.04)	0.78 (± 0.07)	0.76 (± 0.02)	0.87 (± 0.04)
CNN	0.73 (± 0.09)	0.65 (± 0.14)	0.77 (± 0.07)	0.71 (± 0.10)	0.69 (± 0.11)	0.80 (± 0.05)

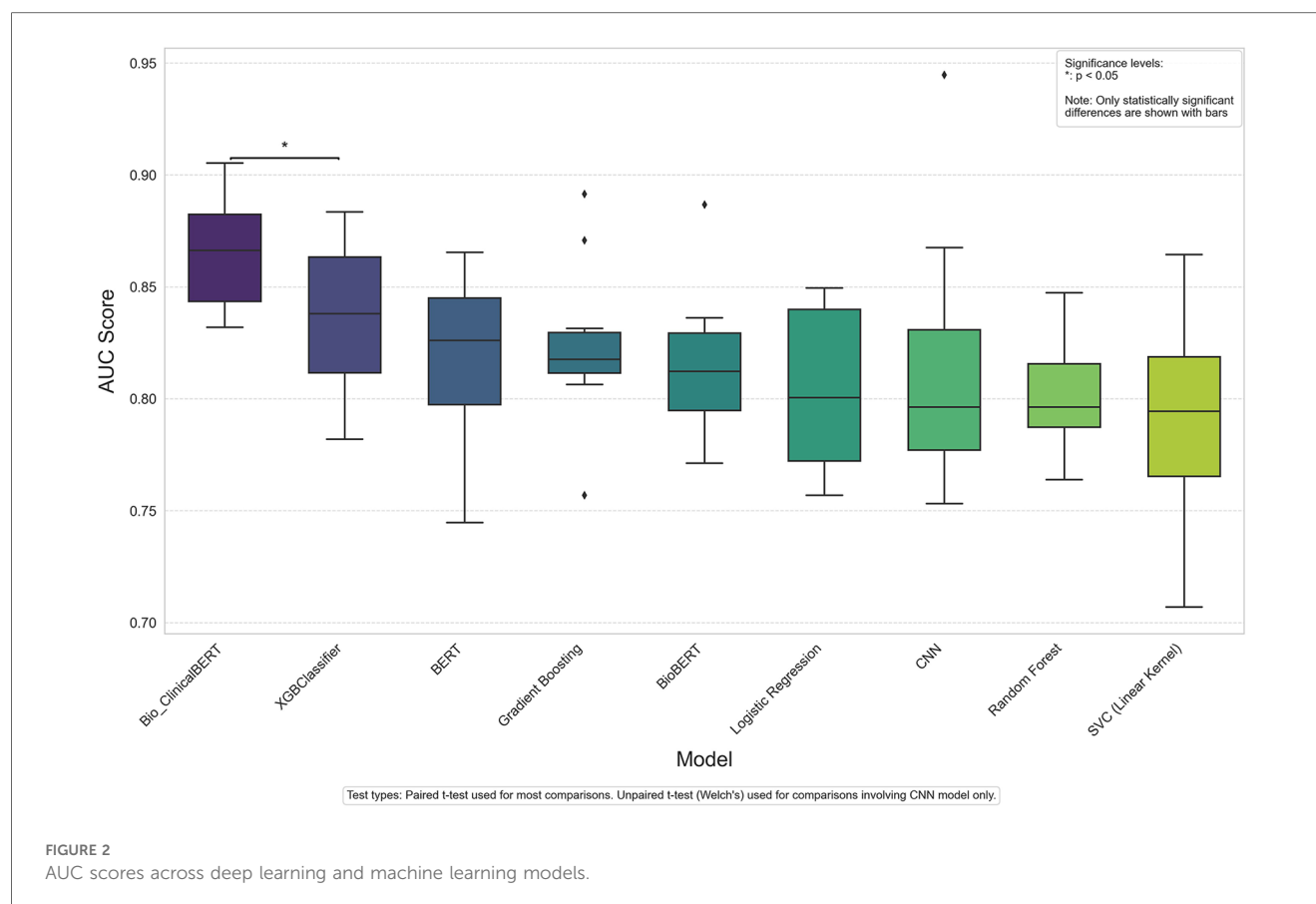
Bio-ClinicalBERT model had the highest performance (AUC = 0.87 ± 0.04) in correctly distinguishing between treatment success and failure notes. Table 3 shows complete performance metrics for machine and deep learning models and Figure 2 displays comparative model performance for all models.

Discussion

This study is the first to develop a clinical text classifier capable of identifying whether a patient achieved treatment success following care for an acute PHF. Our results suggest that AI, more specifically, text classification models in combination with clinical note text can be used to differentiate patients with good

and bad treatment outcomes. This finding is encouraging, and it suggests that routine clinical note content may offer a solution to the much-needed problem of outcome data availability in orthopaedic medicine. This AI approach could translate to producing a widely available treatment outcome measure, which could be used for comparative effectiveness research studies and for the development of clinical quality metrics (64) to improve the quality of orthopaedic care.

A crucial step in demonstrating that orthopaedic care achieves meaningful patient-specific improvement is to have outcome data available that closely measures if treatment was successful in achieving the outcome goals prioritized by patients. However, a leading challenge in outcome measurement remains that orthopaedic outcomes are highly individualized and patient



specific. This fact undermines the value of even the industry-wide standard PROMs, as PROM overall scores often assume equal weighting of outcome dimensions for all patients. Yet, as an example, patients may more heavily value pain reduction over range of motion, which would not be accurately reflected in summary PROM scores. In contrast, our approach and resulting measure is highly flexible and responsive to each individual patient's treatment goals. Orthopaedic-related encounters are highly focused on each patient and how musculoskeletal conditions and treatment affect outcome dimensions and a patient's quality of life. Each encounter is naturally tailored to the patient and their unique musculoskeletal needs, and physicians record goal achievements or challenges that are revealed through the encounter discussion, exam and treatment planning.

The applications of AI and NLP are rapidly expanding in medicine (17, 32, 65, 66) and orthopaedic medicine (23, 34). Yet, our work is the first to use a text classifier to develop a binary measure of treatment success following care for an acute shoulder fracture. Recent studies have used similar methods to achieve related tasks. For example, Humbert-Droz et al., developed a NLP pipeline to extract documented PROMs from the EHR for patients with rheumatoid arthritis (24), and Zanotto et al., used NLP models to identify patient characteristics and health status PROMs (21). However, these studies differ from our approach in that they are still dependent upon the collection of PROM data during clinical care. Our approach offers advantages in that it does not

require any additional data collection efforts on behalf of physicians, staff, or practice managers, as we know there are a myriad of challenges in collecting PROMs that may never be eliminated (10). Our approach is flexible and responsive with respect to the natural language typically recorded during the clinical encounter.

Our sample contained clinical notes from office visits for PHF-related care ranging from over 123 physicians across 35 departments within a regional health system. This finding signals that patients are receiving follow-up care for their PHFs from many different providers. While many of these visits were with an orthopaedic provider, this is an elderly patient population that regularly seeks care from primary care. Therefore, the diversity in providers and departments in our sample spans clinical notes authored by both orthopaedic and primary care physicians. This might signal that our approach is robust to notes beyond those authored just by orthopaedic physicians, to also include a variety of physicians that may discuss healing following orthopaedic treatment. Although there were not many differences observed between the characteristics of success and failure notes, we did find that treatment success notes were significantly shorter in length (4,673 vs. 5,291 characters) compared to treatment failure notes. This is not surprising, as good outcomes can be discussed and recorded more quickly and typically don't require ongoing treatment and related discussion and documentation. However, clinical documentation can be highly varied and more robust patient and visit characteristics should be assessed in future work

to explore their relationship with clinical documentation and patient outcomes.

We found that of the models we tested, the Bio-ClinicalBERT model had the highest performance ($AUC = 0.87 \pm 0.04$) in correctly distinguishing between treatment success and failure notes. This finding was consistent with prior research, which has illustrated Bio-ClinicalBERT's enhanced capability in comprehending intricate medical terminologies and thematic elements inherent to clinical documentation compared to other models. In a similar study, Lee et al. also found that Bio-ClinicalBERT models performed well when identifying documented goals-of-care discussions within the EHR (22).

This work was the first study to develop a clinical text classifier capable of identifying whether a patient achieved treatment success following care for an acute PHF. Our ongoing work will continue to refine our approach to how to use clinical progress note content to form assessments of patient outcomes following orthopaedic treatment. Beyond validation and moving toward use in clinical practice, we envision the next step would be to consider integrating this measure with the EHR to support clinical decision support (CDS) use and quality reporting at the health system level. Development of this measure provides science at scale which enables unlimited comparative effectiveness studies across orthopaedic treatments and future tailored treatment decision making through CDS use to improve individual and population health.

This work is not without limitations. This was a relatively small feasibility study to assess if text classifiers could be applied to clinical notes to generate a measure of orthopaedic treatment success. Although we were successful in achieving an outcome measure with a high level of accuracy, we used a relatively small sample of patients with one shoulder condition. The patient sample we used lacked diversity, being primary female and white, and model performance should be assessed in more diverse patient samples to ensure unbiased model performance among patient subsets. Additionally, clinical note structure can vary significantly across physicians, medical specialties, and health systems. To inform the generalizability of our work, we must test our approach using multi-institutional data across a range of orthopaedic providers with variable documentation practices. Although we expect our approach would be generalizable to other orthopaedic conditions, we have not explored our methods in orthopaedic conditions beyond PHF. Future work will focus on applying our approach to acute and chronic orthopaedic conditions across adult and pediatric populations. Furthermore, the transformer-based models we adopted including BERT, BioBERT, and ClinicalBioBERT, are relatively limited in their parameter scope and the breadth of training data compared to more recent advancements in large language models (LLMs). These newer models, which boast significantly more parameters and have been trained on substantially larger datasets, may offer improved accuracy and better generalizability across diverse medical texts and clinical conditions. Exploring the application of LLMs in future studies could address some of the current limitations and provide a more robust tool for clinical text analysis.

As a next step for this work, we plan to assess the validity of our newly developed measure using PROMs or prospective patient follow-up. There are mixed reports of the accuracy of clinical note content, therefore we must evaluate if our measure accurately reflects a state of patient improvement or decline our new measure is reporting. However, multiple studies have found that healthcare professionals produce accurate documentation for concrete and overt symptoms, such as range of motion and impaired physical functioning (67–70). Therefore, because of the highly focused nature of orthopaedic encounters and symptoms, we believe that clinical note content remains a valuable and valid source of data from which to develop an indicator of treatment success for patients (17, 19, 38). Lastly, although we engaged patients in the development of outcome labels and the definition of treatment success, we ultimately used the physician perspective to develop the gold standard dataset labels of treatment success or failure. It is recognized that patients and physicians may differ in their definition of treatment success, therefore future work will explore incorporating patient perspectives into the definition of treatment success.

Conclusions

We believe the development of an AI-based measure of treatment success using clinical notes may represent a paradigm shift in how outcomes are collected in orthopaedic medicine. Orthopaedic-related healthcare encounters are highly focused on the ways in which musculoskeletal treatments affect outcome priorities such as pain, range of motion, and patient's quality of life. This method, if widely adopted and implemented on clinical note content stored in the EHR could translate to widely available success measures which could be used to for comparative effectiveness research studies and for the development of clinical quality metrics (64) to improve the quality of orthopaedic care.

Data availability statement

The data used in this analysis is not readily available because patient identifiers are contained in the dataset and they must remain stored in a HIPAA compliant location. Requests to access the datasets should be directed to sbf@clemson.edu.

Ethics statement

The studies involving humans were approved by Institutional Review Board at Prisma Health. The studies were conducted in accordance with the local legislation and institutional requirements. The ethics committee/institutional review board waived the requirement of written informed consent for participation from the participants or the participants' legal guardians/next of kin because the research poses minimal risk, is non-interventional, and obtaining consent would be impractical.

Author contributions

SF: Writing – original draft, Writing – review & editing, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision. AA: Writing – review & editing, Data curation, Formal analysis, Methodology, Writing – original draft. DS: Writing – review & editing, Data curation, Formal analysis, Methodology, Supervision, Writing – original draft. BJ: Writing – review & editing, Data curation, Formal analysis, Writing – original draft. CK: Writing – review & editing, Data curation, Formal analysis, Writing – original draft. ZR: Writing – review & editing, Data curation, Methodology, Writing – original draft. KJ: Writing – review & editing, Conceptualization, Data curation, Methodology, Supervision, Writing – original draft. JO: Writing – review & editing, Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Supervision, Writing – original draft.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was

supported by grant R03HS029060 from the Agency for Healthcare Research and Quality.

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.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

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.

References

- Jayakumar P, Bozic KJ. Advanced decision-making using patient-reported outcome measures in total joint replacement. *J Orthop Res.* (2020) 38(7):1414–22. doi: 10.1002/jor.24614
- Briffa N. The employment of patient-reported outcome measures to communicate the likely benefits of surgery. *Patient Relat Outcome Meas.* (2018) 9:263–6. doi: 10.2147/PROM.S132746
- Makhni EC. Meaningful clinical applications of patient-reported outcome measures in orthopaedics. *J Bone Joint Surg Am.* (2021) 103(1):84–91. doi: 10.2106/jbjs.20.00624 (Accessed June 1, 2024).
- Ayers D, Bozic K. The importance of outcome measurement in orthopaedics. *Clin Orthop Relat Res.* (2013) 471(11):3409–11. doi: 10.1007/s11999-013-3224-z
- Gagnier J. Patient reported outcomes in orthopaedics. *J Orthop Res.* (2017) 35(10):2098–108. doi: 10.1002/jor.23604
- American Academy of Orthopaedic Surgeons. Patient reported outcome measures. (2023). Available at: <https://www.aaos.org/quality/research-resources/patient-reported-outcome-measures/performance-measures-by-orthopaedic-subspecialty/> (Accessed June 1, 2024).
- Power JD, Perruccio AV, Canizares M, McIntosh G, Abraham E, Attabib N, et al. Determining minimal clinically important difference estimates following surgery for degenerative conditions of the lumbar spine: analysis of the Canadian spine outcomes and research network (CSORN) registry. *Spine J.* (2023) 23(9):1323–33. doi: 10.1016/j.spinee.2023.05.001
- Antosh IJ, Svoboda SJ, Peck KY, Garcia EJ, Cameron KL. Change in KOOS and WOMAC scores in a young athletic population with and without anterior cruciate ligament injury. *Am J Sports Med.* (2018) 46(7):1606–16. doi: 10.1177/0363546518768753
- DelPrete A. According to AAOE survey, only 35% of orthopaedic practices are collecting PROMs (2017). Available at: <https://www.aaos.org/news/354456/According-to-AAOE-Survey-Only-35-of-Orthopaedic-Practices-are-Collecting-PROMs.htm> (Accessed June 1, 2024).
- Nguyen H, Butow P, Dhillon H, Sundaresan P. A review of the barriers to using patient-reported outcomes (PROs) and patient-reported outcome measures (PROMs) in routine cancer care. *J Med Radiat Sci.* (2021) 68(2):186–95. doi: 10.1002/jmrs.421
- Sisodia RC, Dankers C, Orav J, Joseph B, Meyers P, Wright P, et al. Factors associated with increased collection of patient-reported outcomes within a large health care system. *JAMA Network Open.* (2020) 3(4):e202764. doi: 10.1001/jamanetworkopen.2020.2764
- Shapiro LM, Spindler K, Cunningham B, Koh J. Patient-reported outcome measure collection and utilization: a survey of American academy of orthopaedic surgeons members. *J Am Acad Orthop Surg.* (2024) 32(3):114–22. doi: 10.5435/JAAOS-D-23-00879
- Miller JM, Velanovich V. The natural language of the surgeon's clinical note in outcomes assessment: a qualitative analysis of the medical record. *Am J Surg.* (2010) 199(6):817–22. doi: 10.1016/j.amjsurg.2009.06.037
- Joarder I, Taniguchi S, Mendoza A, Snow ME. Defining “successful” treatment outcomes in adolescent idiopathic scoliosis: a scoping review. *Eur Spine J.* (2023) 32(4):1204–44. doi: 10.1007/s00586-023-07592-w
- Turcotte J, Kelly M, Lynch K, Pipkin K, Patton C. Patient goal-directed care in an orthopaedic spine specialty clinic. *J Am Acad Orthop Surg.* (2021) 29(17):e880–e7. doi: 10.5435/JAAOS-D-20-01105
- Paradis BE, Klein BJ, Bartlett LE, Cohn RM, Bitterman AD. Goals of care discussions in orthopaedic surgery: geriatric hip fractures. *J Arthroplasty.* (2024) 39(5):1144–8. doi: 10.1016/j.arth.2024.03.001
- 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 Suppl B):99–104. doi: 10.1302/0301-620X.102B7.BJJ-2019-1574.R1
- Martin-Sanchez F, Verspoor K. Big data in medicine is driving big changes. *Yearb Med Inform.* (2014) 23(01):14–20. doi: 10.15265/IY-2014-0020
- Hashir M, Sawhney R. Towards unstructured mortality prediction with free-text clinical notes. *J Biomed Inform.* (2020) 108:103489. doi: 10.1016/j.jbi.2020.103489
- Murdoch TB, Detsky AS. The inevitable application of big data to health care. *Jama.* (2013) 309(13):1351–2. doi: 10.1001/jama.2013.393
- Zanotto BS, Beck da Silva Etges AP, Dal Bosco A, Cortes EG, Ruschel R, De Souza AC, et al. Stroke outcome measurements from electronic medical records: cross-sectional study on the effectiveness of neural and nonneural classifiers. *JMIR Med Inform.* (2021) 9(11):e29120. doi: 10.2196/29120
- Lee RY, Kross EK, Torrence J, Li KS, Sibley J, Cohen T, et al. Assessment of natural language processing of electronic health records to measure goals-of-care discussions as a clinical trial outcome. *JAMA Network Open.* (2023) 6(3):e231204. doi: 10.1001/jamanetworkopen.2023.1204
- 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):927–34. doi: 10.1016/j.arth.2020.09.055

24. Humbert-Droz M, Izadi Z, Schmajuk G, Gianfrancesco M, Baker MC, Yazdany J, et al. Development of a natural language processing system for extracting rheumatoid arthritis outcomes from clinical notes using the national rheumatology informatics system for effectiveness registry. *Arthritis Care Res (Hoboken)*. (2023) 75(3):608–15. doi: 10.1002/acr.24869
25. Assale M, Dui LG, Cina A, Seveso A, Cabitza F. The revival of the notes field: leveraging the unstructured content in electronic health records. *Front Med (Lausanne)*. (2019) 6:66. doi: 10.3389/fmed.2019.00066
26. Hanson JL, Stephens MB, Pangaro LN, Gimbel RW. Quality of outpatient clinical notes: a stakeholder definition derived through qualitative research. *BMC Health Serv Res*. (2012) 12:407. doi: 10.1186/1472-6963-12-407
27. Podder V, Lew V, Ghassemzadeh S. *SOAP Notes*. Treasure Island, FL: StatPearls Publishing (2021).
28. Hunter KM. *Doctors' Stories: The Narrative Structure of Medical Knowledge*. Princeton, N.J.: Princeton Univ. Press (1993).
29. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst*. (2014) 2:3. doi: 10.1186/2047-2501-2-3
30. Resnik P, Niv M, Nossal M, Kapir A, Toren R. Communication of clinically relevant information in electronic health records: a comparison between structured data and unrestricted physician language. *Perspect Health Inform Manag*. (2008) 2008:1–10.
31. Obaid J, Weeda E, Matuskowitz A, Gagnon K, Crawford T, Carr C, et al. Automated detection of altered mental status in emergency department clinical notes: a deep learning approach. *BMC Med Inform Decis Mak*. (2019) 19(1):164. doi: 10.1186/s12911-019-0894-9
32. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *J Biomed Inform*. (2017) 73:14–29. doi: 10.1016/j.jbi.2017.07.012
33. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. *J Biomed Inform*. (2018) 77:34–49. doi: 10.1016/j.jbi.2017.11.011
34. AlShehri Y, Sidhu A, Lakshmanan LVS, Lefavre KA. Applications of natural language processing for automated clinical data analysis in orthopaedics. *J Am Acad Orthop Surg*. (2024) 32(10):439–46. doi: 10.5435/JAAOS-D-23-00839
35. Sasanelli F, Le KDR, Tay SBP, Tran P, Verjans JW. Applications of natural language processing tools in orthopaedic surgery: a scoping review. *Appl Sci*. (2023) 13(20):11586. doi: 10.3390/app132011586
36. Thirukumaran CP, Zaman A, Rubery PT, Calabria C, Li Y, Ricciardi BF, et al. Natural language processing for the identification of surgical site infections in orthopaedics. *J Bone Joint Surg Am*. (2019) 101(24):2167–74. doi: 10.2106/JBJS.19.00661
37. Varpio L, Rashotte J, Day K, King J, Kuziemsky C, Parush A. The EHR and building the patient's story: a qualitative investigation of how EHR use obstructs a vital clinical activity. *Int J Med Inf*. (2015) 84(12):1019–28. doi: 10.1016/j.ijmedinf.2015.09.004
38. Boag W, Doss D, Naumann T, Szolovits P. What's in a note? Unpacking predictive value in clinical note representations. *AMIA Jt Summits Transl Sci Proc*. (2018) 2018:26–34.
39. Vest JR, Grannis SJ, Haut DP, Halverson PK, Menachemi N. Using structured and unstructured data to identify patients' need for services that address the social determinants of health. *Int J Med Inform*. (2017) 107:101–6. doi: 10.1016/j.ijmedinf.2017.09.008
40. Carrell DS, Schoen RE, Leffler DA, Morris M, Rose S, Baer A, et al. Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. *J Am Med Inform Assoc*. (2017) 24(5):986–91. doi: 10.1093/jamia/ocx039
41. Bayley KB, London MR, Grunkemeier GL, Lansky DJ. Measuring the success of treatment in patient terms. *Med Care*. (1995) 33(4 Suppl):As226–35.
42. Davey MS, Hurley ET, Anil U, Condren S, Kearney J, O'Tuile C, et al. Management options for proximal humerus fractures—a systematic review & network meta-analysis of randomized control trials. *Injury*. (2022) 53(2):244–9. doi: 10.1016/j.injury.2021.12.022
43. Patel AH, Wilder JH, Ofa SA, Lee OC, Savoie FH, Brien O, et al. Trending a decade of proximal humerus fracture management in older adults. *JSES International*. (2022) 6(1):137–43. doi: 10.1016/j.jseint.2021.08.006
44. Baker HP, Gutbrod J, Strelzow JA, Maassen NH, Shi L. Management of proximal humerus fractures in adults—a scoping review. *J Clin Med*. (2022) 11(20):6140. doi: 10.3390/jcm11206140
45. Hao KA, Patch DA, Reed LA, Spitler CA, Horneff JG, Ahn J, et al. Factors influencing surgical management of proximal humerus fractures: do shoulder and trauma surgeons differ? *J Shoulder Elbow Surg*. (2022) 31(6):e259–69. doi: 10.1016/j.jse.2021.11.016
46. Kennedy AB, Mitcham A, Parris K, Albertson F, Sanchez Ferrer L, O'Boyle C, et al. Wonderings to research questions: engaging patients in long COVID research prioritization within a learning health system. *Learn Health Syst*. (2024) 8(S1):e10410. doi: 10.1002/lrh2.10410
47. Natafagi N, Parris K, Walker E, Gartner T, Coffin J, Mitcham A, et al. Through their eyes: using photovoice to capture the capacity-building journey of long COVID patient experts. *Health Expect*. (2024) 27(6):e70094. doi: 10.1111/hex.70094
48. Wright AA, Hensley CP, Gilbertson J, Leland JM III, Jackson S. Defining patient acceptable symptom state thresholds for commonly used patient reported outcomes measures in general orthopedic practice. *Man Ther*. (2015) 20(6):814–9. doi: 10.1016/j.math.2015.03.011
49. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. (2009) 42(2):377–81. doi: 10.1016/j.jbi.2008.08.010
50. Manning CD, Schütze H, Raghavan P. *Introduction to Information Retrieval*. Cambridge, USA: Cambridge University Press (2008).
51. Ma Y, Guo G, editors. *Support Vector Machines Applications*. Cham: Springer (2014).
52. Breiman L. Random forests. *Mach Learn*. (2001) 45(1):5–32. doi: 10.1023/A:1010933404324
53. Klug M, Barash Y, Bechler S, Resheff YS, Tron T, Ironi A, et al. A gradient boosting machine learning model for predicting early mortality in the emergency department triage: devising a nine-point triage score. *J Gen Intern Med*. (2020) 35(1):220–7. doi: 10.1007/s11606-019-05512-7
54. Inoue T, Ichikawa D, Ueno T, Cheong M, Inoue T, Whetstone WD, et al. XGBoost, a machine learning method, predicts neurological recovery in patients with cervical spinal cord injury. *Neurotrauma Rep*. (2020) 1(1):8–16. doi: 10.1089/neur.2020.0009
55. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY: Curran Associates Inc. (2017). p. 6000–10.
56. Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, editors. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, MN: Association for Computational Linguistics (2019). p. 4171–86.
57. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. (2019) 36(4):1234–40. doi: 10.1093/bioinformatics/btz682
58. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. *arXiv [Preprint]*. *arXiv:1904.03323* (2019). Available at: <https://arxiv.org/abs/1904.03323> (Accessed December 3, 2023).
59. Johnson AEW, Pollard TJ, Shen L, Lehman L-w, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. (2016) 3(1):160035. doi: 10.1038/sdata.2016.35
60. Zhang Y, Wallace B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv [Preprint]*. *arXiv:1510.03820* (2015). Available at: <https://arxiv.org/abs/1510.03820> (Accessed December 3, 2023).
61. Jacovi A, Shalom OS, Goldberg Y. Understanding convolutional neural networks for text classification. *arXiv [Preprint]*. *arXiv:1809.08037* (2018). Available at: <https://arxiv.org/abs/1809.08037> (Accessed December 3, 2023).
62. Rios A, Kavuluru R. Convolutional neural networks for biomedical text classification: application in indexing biomedical articles. *ACM BCB*. (2015) 2015:258–67. doi: 10.1145/2808719.2808746
63. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. (1977) 33(1):159–74.
64. Tonner C, Schmajuk G, Yazdany J. A new era of quality measurement in rheumatology: electronic clinical quality measures and national registries. *Curr Opin Rheumatol*. (2017) 29(2):131–7. doi: 10.1097/bor.0000000000000364
65. Allada AK, Wang Y, Jindal V, Babee M, Tizhoosh HR, Crowley M. Analysis of language embeddings for classification of unstructured pathology reports. *Annu Int Conf IEEE Eng Med Biol Soc*. (2021) 2021:2378–81. doi: 10.1109/EMBC46164.2021.9630347
66. Velupillai S, Suominen H, Liakata M, Roberts A, Shah AD, Morley K, et al. Using clinical natural language processing for health outcomes research: overview and actionable suggestions for future advances. *J Biomed Inform*. (2018) 88:11–9. doi: 10.1016/j.jbi.2018.10.005
67. Sneeuw KC, Sprangers MA, Aaronson NK. The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease. *J Clin Epidemiol*. (2002) 55(11):1130–43. doi: 10.1016/S0895-4356(02)00479-1
68. Field D, Douglas C, Jagger C, Dand P. Terminal illness: views of patients and their lay carers. *Palliat Med*. (1995) 9(1):45–54. doi: 10.1177/026921639500900107
69. Jones JM, McPherson CJ, Zimmermann C, Rodin G, Le LW, Cohen SR. Assessing agreement between terminally ill cancer patients' reports of their quality of life and family caregiver and palliative care physician proxy ratings. *J Pain Symptom Manage*. (2011) 42(3):354–65. doi: 10.1016/j.jpainsymman.2010.11.018
70. Yang JS, Keener JD, Yamaguchi K, Chen J, Stobbs-Cucchi G, Patton R, et al. Reliability of patient self-assessment of shoulder range of motion and strength after shoulder arthroplasty. *J Shoulder Elbow Surg*. (2015) 24(1):1–10. doi: 10.1016/j.jse.2014.08.025
71. Loshchilov I, Hutter F. Decoupled weight decay regularization. *arXiv [Preprint]*. *arXiv:1711.05101* (2017). Available at: <https://arxiv.org/abs/1711.05101> (Accessed January 8, 2024).