Original Paper

Automated Extraction and Coding of Qualitative Data Within Physician-Patient Communication in Cancer Care Using Natural Language Processing: Development and Validation Study

Jay Sureshbhai Patel[1], BDS, MS, PhD; Laura A Siminoff[2], BA, MA, PhD; Zasim Azhar Siddiqui[3], BDS, MS, PhD; Maria D Thomson[4], PhD; Ritwik Katiyar[1], BS, MS; Oqba Memar[5], BDS, MSc, DMD

[1]Center for Artificial Intelligence, Data Science, and Informatics, Department of Oral Health Sciences, Kornberg School of Dentistry, Temple University, Philadelphia, PA, United States

[2]Department of Social and Behavioral Sciences, College of Public Health, Temple University, Philadelphia, PA, United States

[3]Department of Pharmaceutical Systems and Policy, West Virginia University, Morgantown, WV, United States

[4]Department of Social and Behavioral Sciences, School of Public Health, Virginia Commonwealth University, Richmond, VA, United States

[5]Department of Restorative Dentistry, Kornberg School of Dentistry, Temple University, Philadelphia, PA, United States

Corresponding Author:

Jay Sureshbhai Patel, BDS, MS, PhD
Center for Artificial Intelligence, Data Science, and Informatics, Department of Oral Health Sciences
Kornberg School of Dentistry, Temple University
400-C, 3223 N Broad St
Philadelphia, PA 19140
United States
Phone: 1 (215) 204-7000
Email: patel.jay@temple.edu

Abstract

Background: Large-scale analysis of qualitative clinical communication data typically requires substantial manual coding. Natural language processing (NLP) may reduce this burden by converting unstructured physician-patient communication transcripts into structured qualitative data.

Objective: This study aimed to develop, test, and validate an NLP application for identifying quality-of-life factors in cancer care communication. We also compared traditional machine learning (ML), rule-based, transformer-based, and out-of-the-box (OoB) large language model (LLM) approaches.

Methods: We analyzed 434 standardized patient encounter transcripts of physician-patient interactions collected from a prior study. A rule-based SpaCy model, traditional ML models, bidirectional encoder representations from transformer (BERT), Robustly Optimized BERT Pretraining Approach (RoBERTa), and the OoB Qwen3 LLM were evaluated using F_1-scores. The highest-performing model was incorporated into a web-based NLP application and further assessed through manual validation.

Results: Transformer-based models demonstrated the strongest performance. RoBERTa achieved the highest initial F_1-score of 83%, outperforming BERT at 80%, the OoB Qwen3 LLM model at 39%, and the rule-based SpaCy baseline at 34%. Manual validation of the RoBERTa model produced an F_1-score of 91%. The comparatively low performance of Qwen3 indicated that OoB LLM use alone may not produce reliable results for qualitative, domain-specific clinical communication tasks.

Conclusions: A fine-tuned transformer model (RoBERTa) was better suited to capturing the complexity of patient encounter transcripts than traditional ML, rule-based approaches, or an OoB LLM. The findings demonstrate the feasibility of using a manually validated, web-based NLP application trained on gold-standard annotations to transform large-scale physician-patient communication transcripts into structured qualitative data. Hybrid systems combining rule-based precision with the contextual capabilities of transformer models may further improve the identification of longer conversational sequences, including emotional expressions, question-answer exchanges, and multitopic utterances.

JMIR Form Res 2026;10:e88404; doi: 10.2196/88404

Keywords: cancer care; qualitative research; natural language processing; physician-patient relations; large language models

Introduction

Cancer is a major global health challenge, affecting millions of people each year. Treatments such as chemotherapy, immunotherapy, and radiation often impose a heavy burden, leading to reduced quality of life (QoL) and, in some cases, posttraumatic stress disorder [1-4]. Research consistently shows that open and effective communication between patients, caregivers, and health care providers empowers patients to make informed decisions, reduces the emotional toll of diagnosis and treatment, and improves overall well-being [5-11].

In the social and behavioral sciences, qualitative research methods have played an important role in understanding how patients make treatment decisions and weigh the risks, benefits, and preferences of different options [12-17]. By capturing patients' thoughts, emotions, behaviors, and interactions, qualitative approaches can reveal patterns and experiences that might otherwise be overlooked [18-20]. In medicine, these methods, such as interviews, focus groups, and ethnographic observation, have been used to explore health care services in depth, providing a humanistic, person-centered perspective [21-24]. Their flexibility allows researchers to examine complex social and cultural factors that quantitative methods may not fully capture, informing policy, improving practice, and guiding the development of interventions that better match the needs of patients.

Despite these strengths, traditional qualitative research has notable limitations. Data collection and analysis are often constrained by small sample sizes. Processing recorded interviews through transcription, coding, and review is labor-intensive, time-consuming, and susceptible to human bias [25-27]. In health care, qualitative data frequently exist in unstructured formats, such as transcripts from clinical encounters, interviews, focus groups, or written narratives. Large-scale manual review of these materials is costly, difficult to scale, and rarely feasible outside of well-funded studies [28].

Natural language processing (NLP) offers a promising solution. NLP can automatically process large volumes of unstructured text, converting rich, narrative data into structured formats suitable for qualitative and statistical analysis [29]. This approach could enable rapid, consistent, and scalable analysis of clinical communications, including triadic assessments involving clinicians, patients, and caregivers that often occur in cancer care. Cancer care, with its complex and emotionally charged interactions, presents an opportunity for advancing these methods. By automating the coding process, NLP can improve efficiency, reduce resource demands, and uncover insights from large datasets [25]. Early NLP systems, such as cTAKES (Clinical Text Analysis and Knowledge Extraction System) and CLAMP (Clinical Language Annotation, Modeling, and Processing), have been developed to extract structured clinical information from unstructured health records [30,31]. In more recent studies, an AI-based tool to analyze open-ended patient feedback with high accuracy has been developed [32]. Furthermore, pretrained transformer models have been applied to several clinical NLP tasks, from processing clinical notes to the sentiment analysis of clinical conversations. These studies show the growing value of transformer-based models for clinical text analysis [33,34].

Transformer-based models and large language models (LLMs) have shown promise in health care; however, their application to qualitative patient data remains limited [35,36]. Qualitative studies examining patient feedback data commonly use unsupervised NLP methods such as word-frequency analysis, latent Dirichlet allocation (LDA), and BERTopic rather than supervised models because creating gold standard annotations for transcript text is time-consuming. Since transformer models require large amounts of data, qualitative analysis of patient transcripts is rarely performed using supervised methods [36,37]. While several studies compare LLMs with human annotations or LDA [38-40], few or no studies directly compare the use of traditional machine learning (ML) and NLP embeddings against transformers and LLMs in qualitative studies. Finally, although these models have been developed, studies suggest that their translation into real-world applications remains limited [36]. This creates a gap in understanding of how transformers and LLMs compare with traditional ML models in identifying QoL factors in physician-patient communication transcripts and in supporting manually validated and real-world applications.

The objective of this study is to develop, test, and validate an application to process large-scale qualitative clinical communication data into structured formats, reducing the need for manual coding while comparing classic supervised ML models with the latest transformer and LLMs. A successful application could process transcribed conversational data quickly, optimize resources, and support generalized research findings. However, the novelty of this study is not in showing NLP's potential to process clinical texts, which has already been established. The novelty of this study is in its comparison of ML with transformer-based and out-of-the-box (OoB) LLMs to extract qualitative variables from large-scale physician-patient communication transcripts. The study also uses large-scale, gold-standard manual annotations, develops a web-based application, and conducts a manual evaluation of the model for real-world use.

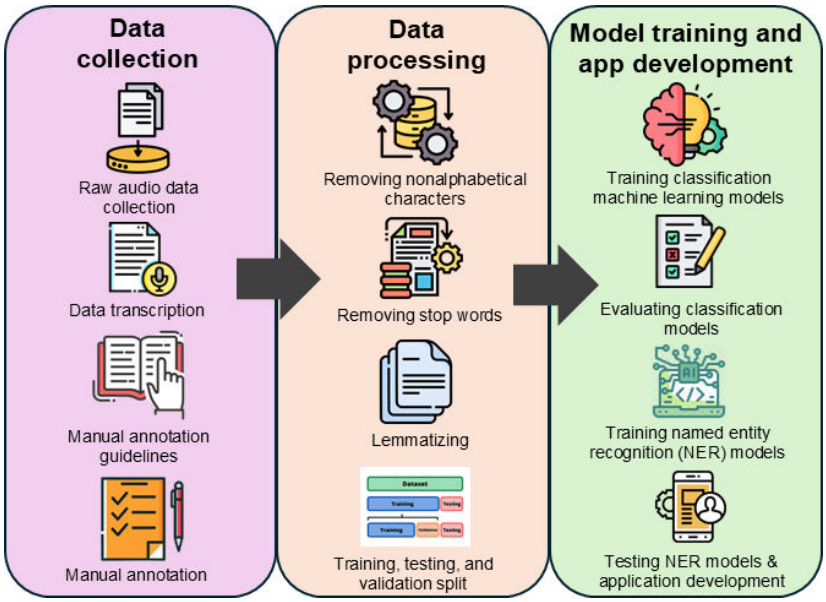
Methods

Dataset

We used data from a prior study led by author-researcher LAS that examined how physician-patient communication, patient gender, and race influence diagnostic decisions for colorectal cancer in primary care [41]. Researchers contacted 500 primary care physician (PCP) practices, of which

265 agreed to participate. From these, 220 PCPs were randomly selected to receive at least one visit from a trained unannounced standardized patient (USP), resulting in 414 total visits. USPs, representing 4 race-gender combinations, presented as 47-year-old new patients with symptoms suggestive of colorectal cancer. Each PCP saw 2 USPs, with assignments counterbalanced to ensure equal representation across patient cases. Consultations were covertly audio-recorded and transcribed, and medical records were later reviewed for diagnostic documentation and follow-up. In this study, we used 434 transcripts of the audio recordings to develop NLP applications, with each transcript, on average, containing approximately 2928 words and 311 lines. The overall workflow of this study is displayed in Figure 1.

Figure 1. Overall workflow diagram for this study, outlining the process from data collection to model training, development, and testing. NER: named entity recognition.



Ethical Considerations

The study was approved by the Temple University Institutional Review Board (IRB; #HM10930). Participating physicians consented to audio recording of clinical encounters. The patients were trained as standardized patients and therefore did not need consent. The transcripts were handled in accordance with institutional privacy and confidentiality requirements, and the LLM was evaluated locally to minimize the risk of data exposure. Participating practices were reimbursed for standardized patient visits, and physicians received a US $50 honorarium for completing the poststudy survey; no additional compensation was provided for the present analysis [41]. Finally, the transcripts were deidentified to protect the privacy of the standardized patients and participating physicians.

Data Preprocessing

We applied our previously developed data cleaning pipeline to remove punctuation, numbers, and special characters [41,42]. We then manually cleaned the dataset by reviewing the 300 most frequent features. Irrelevant segments were then removed. For example, filler words, such as "Humm," "Umm," and "Aaa," that appeared throughout the transcripts were excluded. To address class imbalance, we applied random oversampling and undersampling techniques for model development. Classes with fewer examples were randomly oversampled to increase representation, while classes with more examples were randomly undersampled to reduce prevalence during training [43,44].

Developing Annotation Guidelines

We developed manual annotation guidelines based on observations and a review of detailed transcribed datasets. Three trained reviewers examined a subset of transcripts to identify relevant qualitative topics, including symptoms, emotions, financial concerns, health behaviors, medical history, psychosocial information, treatment, testing, and follow-up planning. The categories reflect domains commonly examined in qualitative research on patient experiences, QoL, and patient-clinician communication among patients with cancer [45]. Clear and detailed annotation guidelines are crucial for ensuring consistency and reliability, as they reduce variability between annotators and enhance the reliability of the dataset. This consistency is essential for training accurate and robust ML models. Additionally, comprehensive guidelines promote reproducibility by allowing other researchers to understand the criteria used for annotations. Table 1 presents the detailed annotation guidelines used in the manual annotation process.

Table 1. Guidelines for manual annotations and to create a gold standard dataset.

Transcript classification	Description	Examples
Symptoms	Description of pain, duration of symptoms, disease, fatigue, and location of pain	"Stomach pain," "How long has it been going on?"
Emotions	Annoyed/frustrated about symptoms, concerned/worried about symptoms, and general anxiety during the visit	"Um, it's just like a maddening," "That is why we worry."
Financial	Test costs, insurance, and medication costs	"How much does it cost?," "Around five hundred dollars"
Follow-up and closing	Evaluate medical trial for decision when to return, general goodbye, and logistics for testing	"See me after 1st Jan," "Take care."
Health behaviors	Alcohol use, exercise, sleep, and tobacco use	"Do you use alcohol?" "Do you exercise?"
Medical history	Surgeries, pregnancies, previous PCP[a], and screenings	"Have you ever had surgeries?," "tonsils"
Medications	OTC GI[b] offered, rationale for medication, and prescription offered	"If you ever take anything like Prilosec," "can be treated with"
Physical examination	Vitals, rectal examination, and hernia examination	"Be sure you are not losing any blood." "Take a deep breath."
Psychosocial information	Age, marital status, children, and mental health	"How old are you?" "Are you single?," "stressed"
Test offered	Test offered, rationale for tests, and sequential diagnostic process discussion	"Diagnosis of that is an x-ray," "If you are just not getting better"

[a]PCP: primary care physician.
[b]OTC GI: over-the-counter gastrointestinal medication.

Manual Annotation

To create a gold standard dataset, 3 domain experts manually reviewed 434 transcripts using the guidelines. We used "the Extensible Human Oracle Suite of Tools" (eHOST) to annotate clinical notes. Each concept was tagged in detail, as shown in Table 1. We performed the manual annotation process with 4 iterations. After every iteration, we calculated the interrater agreement using the Cohen κ test. The final Cohen κ value was 0.89, indicating excellent agreement. Disagreed concepts between the annotators were discussed and resolved through consensus. By the end of this step, we had created a gold standard dataset of 434 manually annotated clinical transcripts. For more information on the final data and the classification of various words and phrases based on the guidelines, refer to the word clouds shown in Figure S1 in Multimedia Appendix 1.

ML Models

First, we used ML classification models to determine the overall category of each sentence in the transcripts. This sentence-level approach enabled us to evaluate the broader context of each sentence within the transcripts, allowing for a structured understanding of how information was conveyed throughout the conversations.

To evaluate the performance of different models, we split the dataset into an 80:20 train-test split and tested 4 vectorization techniques: term frequency and inverse document frequency (TF-IDF), bag-of-words (BoW), Word2Vec, and bidirectional encoder representations from transformer (BERT) embeddings. These embeddings transformed the text into numerical formats suitable for model training and comparison. Additionally, we created 2 versions of our dataset, one retaining common stop words (eg, "the," "a," and "an") and one with these words removed. This was done to determine whether excluding nonessential terms improves classification performance. We then trained and tested multiple ML classifiers, including support vector machines (SVMs), naive Bayes, random forests, feed-forward neural networks, and BERT models [46].

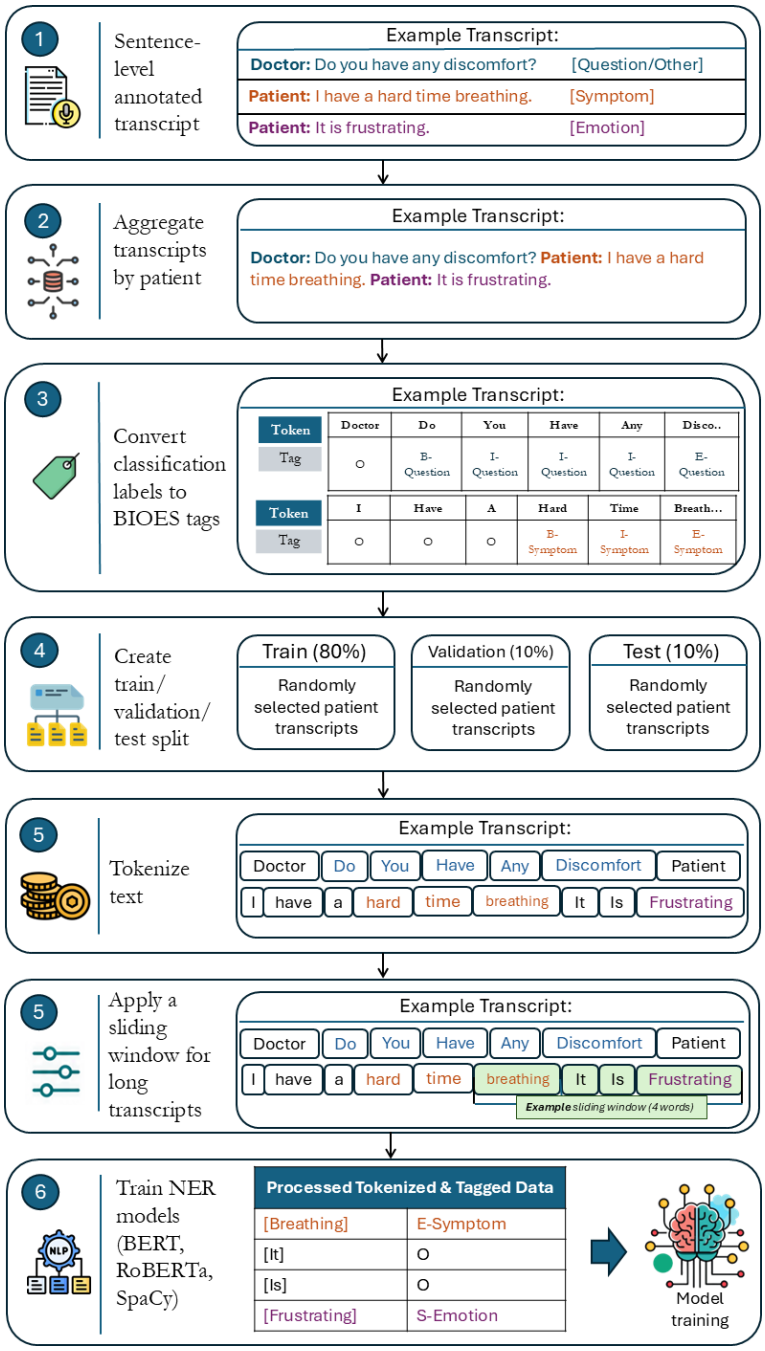
Named Entity Recognition Development

While sentence-level classification provided a high-level understanding of the transcript content by assigning a single category to each sentence, it lacked the granularity needed to identify multiple specific concepts that might coexist within a single sentence. To capture finer semantic details, we extended our analysis by developing a named entity recognition (NER) system.

To prepare our data for NER, we first converted all classification labels into a sequence-labeling format. We aggregated transcripts at the patient level rather than by individual sentence, then applied the BIOES (Beginning, Inside, Outside, End, and Single) tagging scheme to label the data for model training. BIOES assigns each utterance a tag indicating whether it marks the beginning, inside, end, or a single-utterance span of a given topic, and tags all nontopical utterances as "O" or background. The dataset was then split into an 80:10:10—80% for training, 10% for validation, and 10% for testing. Once the tagging and splitting of the data were completed, 3 separate NER models were selected for training: SpaCy, standard BERT, and the Robustly Optimized BERT Pretraining Approach (RoBERTa). SpaCy is a fast, rule-based NLP library that processes text with pretrained models, making it suitable for real-time applications. We particularly used a medically aware SpaCy configuration that incorporated a biomedical scispaCy model and rule-based medical phrase patterns. BERT is a deep learning model that captures contextual meaning by analyzing entire sentences, making it particularly effective for recognizing complex patterns and relationships in medical text. Finally, RoBERTa is a transformer-based model built on BERT, optimized through strategies such as dynamic masking, removal of the next-sentence prediction objective, and larger-scale pretraining to enhance token representation [47,48].

BERT-based models have a maximum input token limit of 512 tokens. Therefore, we implemented a sliding-window approach during training and inference to handle our longer transcripts. This allowed us to process lengthy input sequences by breaking them into overlapping segments, ensuring that contextual information was preserved across boundaries. (See Figure 2 for details of the data processing performed for NER model training.)

Figure 2. Natural language processing (NLP) workflow for preparing and training named entity recognition (NER) models on physician-patient transcripts. Transcripts were first aggregated by patient, converted from classification labels into BIOES (Beginning, Inside, Outside, End, and Single) token-level tags, and split into training, validation, and test sets. Text was then tokenized, and sliding windows were applied to handle longer transcripts before training. BERT: bidirectional encoder representations from transformer; RoBERTa: Robustly Optimized BERT Pretraining Approach.



LLM: Zero-Shot Prompting

We used zero-shot prompting with a locally hosted, open-source LLM. Due to the sensitive nature of the dataset, which contains patient-provider transcripts, data security was a primary concern. Therefore, we evaluated the LLM locally on our machines to minimize the risk of data exposure. Given our available computational resources, we selected the Qwen3 model for the analysis. This model was selected because, according to the Qwen3 technical report, the Qwen3 model family demonstrates strong benchmark performance across various NLP tasks, comparable to, or in some cases better than, GPT-4o and Gemini 2.5 [49].

The Qwen3 model was provided with the following prompt to generate its results:

> *Classify individual utterances from these clinical transcripts. Choose exactly one label from the allowed list. Classify only what is explicitly stated in the supplied utterance. Do not infer missing information. When no category applies, choose O. Allowed labels and definitions are as follows: [see Table 1] Return only a JSON object with one key named "label." Do not explain your answer.*

Model Evaluation

Posttraining, the model evaluation began with the construction of a confusion matrix, a tabular representation that summarizes the performance of a classification model by comparing predicted labels with actual labels. We report the counts of true positives (TPs), false positives (FPs), true negatives (TNs), and false negatives (FNs) across all classes using the confusion matrix. From these values, we compute several performance metrics. Precision measures the proportion of correct positive predictions. Sensitivity (also known as recall) evaluates the model's ability to identify actual positive instances. Specificity assesses the model's ability to correctly identify negative cases and avoid FPs. The F_1-score combines precision and recall into a single metric, which is especially useful when dealing with imbalanced classes. Accuracy provides an overall measure of how often the model's predictions are correct. Additionally, we calculated the area under the curve (AUC), which quantifies the model's ability to discriminate between classes across all classification thresholds. These metrics were used for both the NER model and the ML classification model to compare their performance. Since the dataset contained imbalanced categories, overall model performance was calculated using support-weighted averages. In this approach, each class contributes to the overall metric in proportion to the number of labeled instances in that class. This prevents rare classes from being overrepresented while also allowing class-based performance to be reported separately. Weighted precision, weighted recall, weighted specificity, weighted F_1-score, and AUC were used as the primary overall performance metrics for both the NER and ML classification models [50,51].

Manual Validation

Manual validation was conducted to assess the accuracy and verify the practical usefulness of the model-labeled transcripts using an out-of-fold test set. Based on the power analysis, a minimum of 166 transcript instances was required for manual review. To exceed this minimum threshold, 200 transcript instances were manually evaluated. Each instance was reviewed to determine whether the model-generated label was correct, using the same scoring criteria for TP, FP, TN, and FN as described above for model evaluation.

Application Development

Once the most effective model was established, we built a web application using tools such as Flask to enable users to interact with the model efficiently and analyze transcripts in real time. The application was then hosted on PythonAnywhere to provide easy and secure access online. This application ensures that transcribed data are quickly classified automatically.

Results

Physician Participants' Demographics

This study included 220 PCPs; 70% (n=154) of whom were family medicine practitioners. About 56% (n=123) of the PCPs were male participants, and 73.9% (n=163) identified as White. The average age of physician participants was 51.1 years, with an average of 22 years of practice experience.

Patient Sample Demographics

The final sample consisted of 367 USP visits. These visits were evenly distributed across 4 demographic groups: 25.6% (n=94) Caucasian men, 25.1% (n=92) Caucasian women, 23.7% (n=87) African American men, and 25.6% (n=94) African American women.

ML Models

We evaluated the classification models using 4 different vectorization methods: TF-IDF, BoW, Word2Vec, and BERT embeddings, both with and without stop words. The results are presented in Table 2. Among our classification models, the BERT-based classification model with stop words achieved the highest sensitivity of 66%, an F_1-score of 70%, and an accuracy of 70% when using BERT embeddings. These results highlight the effectiveness of transformer-based architectures in handling phrase-classification tasks, outperforming traditional ML algorithms across all vectorization methods. Further analysis of the results reveals that the BERT-based classification model performed well in classification, but frequently misclassified filler words like "okay" and "yeah" as medical history. Despite this performance, the classification models failed to discern key nuances within each sentence and failed to handle overlapping entity types within a single sentence. Because this approach required splitting transcripts into separate sentences, it could assign only one label per sentence. As a result, it misclassified sentences that contained multiple entity types. For example, in the sentence "I am really worried about my bowel movements with constant diarrhea and constipation," the model classified it as an emotion instead of a symptom simply because the sentence included an emotional expression.

Table 2. Average weighted (W) performance metrics of all trained models across different vectorization methods and classifications.

Classifications	Accuracy (W, %)	F_1-score (W, %)	Precision (W, %)	Recall/sensitivity (W, %)	Specificity (W, %)
Including stop words (a, an, the, etc)					
TF-IDF[a] embedding					
SVM[b]	64	64	65	64	96
Naive Bayes	60	59	66	60	95
Random forest	62	61	64	62	95
Neural network	66	65	66	66	96
BERT[c,d]	—	—	—	—	—
Bag-of-words embedding					
SVM	63	63	65	63	95
Naive Bayes	63	63	65	63	95
Random forest	62	61	63	62	95
Neural network	66	66	67	66	96
BERT[c]	—	—	—	—	—
Word2vec embedding					
SVM	52	50	54	52	94
Naive Bayes	31	33	43	31	92
Random forest	56	54	57	56	94
Neural network	57	56	58	57	95
BERT[c]	—	—	—	—	—
BERT embedding					
SVM[e]	—	—	—	—	—
Naive Bayes	11	11	26	11	91
Random forest	47	43	51	47	93
Neural network	49	47	52	49	94
BERT	70	70	73	66	96
Excluding stop words (a, an, the, etc)					
TF-IDF embedding					
SVM	66	65	66	66	96
Naive Bayes	61	60	65	61	95
Random forest	62	61	63	62	95
Neural network	65	65	66	65	96
BERT[c]	—	—	—	—	—
Bag-of-words embedding					
SVM	62	62	65	62	58
Naive Bayes	63	62	64	63	95
Random forest	61	60	61	61	95
Neural network	65	65	66	65	96
BERT[c]	—	—	—	—	—
Word2vec embedding					
SVM	58	57	59	58	95
Naive Bayes	36	38	48	36	93
Random forest	58	57	58	58	95
Neural network	67	56	58	57	95
BERT[c]	—	—	—	—	—
BERT embedding					
SVM[e]	—	—	—	—	—
Naive Bayes	10	10	26	10	90
Random forest	47	44	52	47	93
Neural network	47	45	51	47	93
BERT	67	66	70	63	93

[a]TF-IDF: term frequency and inverse document frequency.
[b]SVM: support vector machine.
[c]The BERT model does not accept any embedding other than its own BERT embeddings.
[d]BERT: bidirectional encoder representations from transformer.
[e]SVM is a computationally intensive algorithm; since our dataset is large, the algorithm was unable to complete training with BERT embeddings.

NER Model Results

Among our NER models, the RoBERTa-based model achieved the best overall performance, with precision, sensitivity (recall), and an F_1-score of 83%. By comparison, the SpaCy model achieved an F_1-score of 34%, precision of 40%, sensitivity of 35%, and specificity of 91%. The RoBERTa model demonstrated strong performance in identifying entities such as family medical history, symptoms, medical history, and physical examination, achieving F_1-scores of 95%, 89%, 85%, and 85%, respectively. However, the model showed limitations in classifying more ambiguous categories, such as possible diagnoses and tests offered, with lower F_1-scores of 72% and 74%, respectively (refer to Table 3 and Figure 3 for more details).

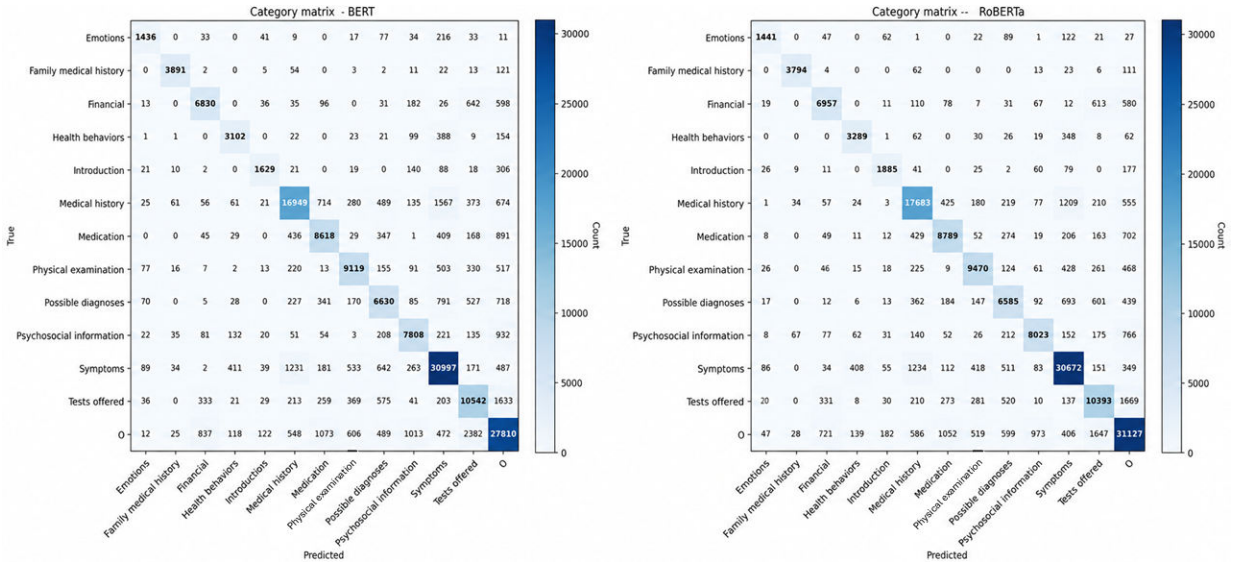
Both BERT models outperformed SpaCy, particularly in the medications, introduction, health behaviors, and emotions categories. In the medication category, SpaCy identified only generic terms like "vitamin," while the BERT models recognized full phrases with context, such as "I can

Table 3. The per-class named entity recognition metrics of the BERT[a] and RoBERTa[b] models.

Categories	F_1-score (W[c], %)	Precision (W, %)	Sensitivity (W, %)	Specificity (W, %)	AUC[d], %	Support[e]
BERT						
Emotions	77	79	75	100	99	1907
Family medical history	95	95	94	100	100	4124
Financial	82	83	80	99	99	8489
Health behaviors	80	80	81	100	99	3820
Introduction	77	83	72	100	99	2254
Medical history	82	85	79	99	99	21,405
Medication	77	76	78	99	99	10,973
Physical examination	82	82	82	99	99	11,063
Possible diagnoses	69	69	69	99	98	9592
Psychosocial information	80	79	80	99	99	9702
Symptoms	87	86	88	98	99	35,080
Tests offered	71	69	74	98	98	14,304
Background "O"	79	80	78	95	95	35,507
Overall performance (mean)	80	81	80	98	98	168,220
RoBERTa						
Emotions	81	84	79	100	99	1833
Family medical history	95	96	94	100	100	4013
Financial	83	83	82	99	99	8485
Health behaviors	84	83	86	100	99	3845
Introduction	81	81	81	100	99	2315
Medical history	85	84	85	99	99	20,677
Medication	81	80	82	99	99	10,714
Physical examination	85	85	85	99	99	11,142
Possible diagnoses	72	72	72	99	98	9151
Psychosocial information	83	84	82	100	99	9791
Symptoms	89	89	90	98	99	34,113
Tests offered	74	73	75	98	98	13,888
Background "O"	83	84	82	95	96	38,026
Overall performance (mean)	83	83	83	98	98	167,993

[a]BERT: bidirectional encoder representations from transformer.
[b]RoBERTa: Robustly Optimized BERT Pretraining Approach.
[c]W: average weighted.
[d]AUC: area under the curve.
[e]Support refers to the total number of labeled token instances in the test set for each category. BIOES (Beginning, Inside, Outside, End, and Single) tags belonging to the same category were aggregated for a simpler view, with decimal places rounded for clarity.

Figure 3. Confusion matrices for BERT and RoBERTa named entity recognition models. BERT: bidirectional encoder representations from transformer; RoBERTa: Robustly Optimized BERT Pretraining Approach.



recommend some vitamins and stuff to take." Similarly, for the introduction, SpaCy detected only basic greetings like "hey" and mislabeled other greeting words like "good morning," whereas the BERT models accurately extracted a wider range of greetings, including complete sentences like "Hi, good morning, how are you, hi doctor." Moreover, in the health behaviors category, SpaCy picked up individual words like "smoking," "walking," "cigarettes," and "glass of wine," but once again failed to capture the full contextual meaning. In contrast, the BERT models recognized complete sentences, such as "I am getting five to six hours of sleep at night." Finally, for the emotions category, SpaCy was once again only able to identify keywords such as "annoying," "concerned," whereas the BERT models were able to identify full contextual information like, "yeah, I am really concerned about my health lately."

Furthermore, as an overall evaluation of the RoBERTa model and SpaCy, Figure S2 in Multimedia Appendix 1 presents the length distribution of annotated text, showing that RoBERTa produced longer spans that better preserved contextual information. Similarly, Figure S3 in Multimedia Appendix 1 compares the average word count per annotation, indicating that RoBERTa's annotations were more detailed, while SpaCy's were shorter and lacked full contextual understanding. These findings further confirm RoBERTa's superior performance in extracting entities with greater accuracy and depth.

LLM: Zero-Shot Prompting

The OoB Qwen 3 LLM achieved comparatively poor performance compared to our RoBERTa model, with an F_1-score of 39%, a precision of 61%, a sensitivity (recall) of 46%, and a specificity of 94%. This indicates poor model performance without finetuning.

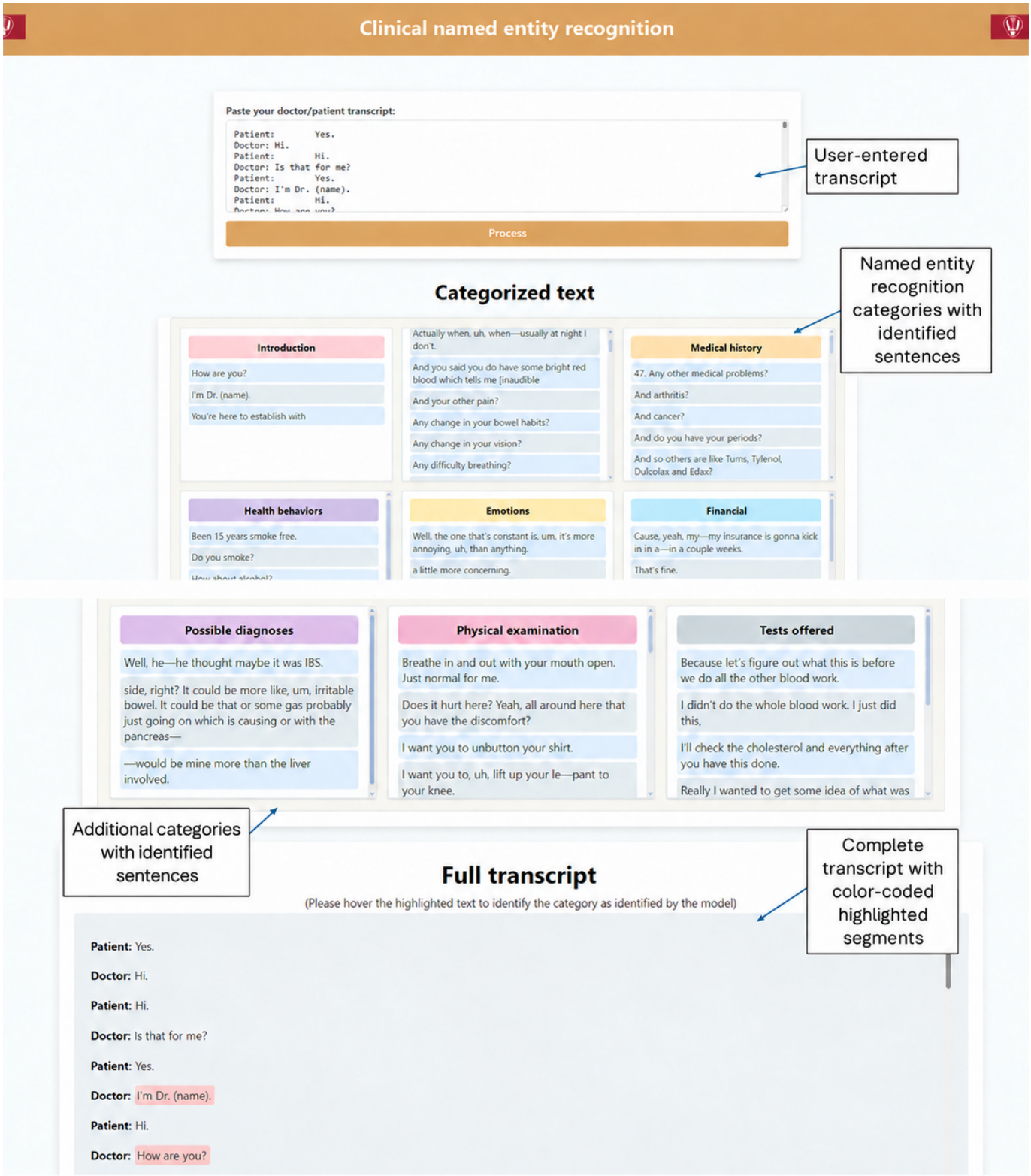
Manual Evaluation

Manual evaluation revealed an F_1-score of 91%, sensitivity of 89%, and specificity of 74%, indicating strong model performance. The scores suggest that the model achieved a strong balance between precision and recall, and correctly identified relevant instances within the transcript.

Application

Since the RoBERTa model performed the best among our classification and NER models, we used it for the NER application. The application was built and deployed on PythonAnywhere. The details of the application are outlined in Figure 4. Furthermore, minor rule-based algorithms were developed to ensure and adjust the model's output in the application for clarity. If the model fails to highlight the full text, "please take two every day" as medication, and only highlights "two every day," the rule-based algorithm adjusts the highlighted section to include the full sentence for clarity. The application was developed with usability and human-computer interaction principles in mind. This allowed us to provide access to our model through a web device, enabling quick transcript analysis on the go.

Figure 4. Web-based clinical named entity recognition (NER) application using the fine-tuned RoBERTa model. Users can enter physician-patient encounter transcripts into the application, which then processes the text and categorizes the identified sentences. The full transcript is also displayed, with color-coded, highlighted segments to show where the model identified each entry. RoBERTa: Robustly Optimized BERT Pretraining Approach.



Discussion

Principal Findings

This study demonstrates that transformer-based models can effectively extract and classify patient information from cancer care transcripts. In sentence-level classification tasks, traditional models such as SVM, random forest, feedforward neural networks, and naive Bayes achieved moderate performance with TF-IDF and BoW embeddings. The BERT-based classification model achieved the highest sentence-level accuracy of 70% using BERT

embeddings. However, sentence-level classification was limited to a single-sentence annotation; the classification models struggled with sentences containing multiple entities. Because each sentence could receive only one label, multitopic statements, such as those combining symptoms with emotional expressions, were often misclassified. The NER approach helped address this limitation by evaluating text at a more granular level, allowing multiple labels to be identified within the same sentence.

Among the NER models tested, the RoBERTa model achieved the strongest performance, with an overall F_1-score of 83%, sensitivity of 83%, and specificity of 98%. These results suggest that the model had a strong balance between identifying information correctly and avoiding FPs. The high sensitivity indicates that the model captured most of the labels correctly, while the high specificity shows that it was effective at not labeling unrelated text. RoBERTa also outperformed the standard BERT NER model (F_1-score=80%, sensitivity=80%, and specificity=98%), the rule-based SpaCy baseline (F_1-score=34%, sensitivity=35%, and specificity=91%), and the zero-shot prompted OoB Qwen-3 LLM model (F_1-score=39%, sensitivity=46%, and specificity=94%). These results suggest that domain-specific fine-tuning is important for this task and that general-purpose LLMs may not be sufficient. Finally, manual validation of the RoBERTa model showed similar strong performance (F_1-score=91%, sensitivity=89%, and specificity=74%), further showcasing the model's capability.

Our results also compare favorably with prior work. For example, our RoBERTa model outperformed the classifier from the "One in a Million" primary care consultation study (F_1-score=0.83 vs 0.45-0.55) [35] and exceeded the receiver operating characteristic AUC reported in a recent study on NLP for cancer interview data (0.98 vs 0.94). Our study also builds on previous studies by extracting a broader set of entities, including, but not limited to, patient-reported symptoms, emotions, financial concerns, health behaviors, and diagnostic testing [52]. We also achieved comparable performance to studies using topic modeling approaches, such as LDA, and standard ML algorithms [36]. However, our LLM significantly underperformed by comparison. We believe that this is likely due to a lack of fine-tuning, thereby showcasing that OoB LLMs are not suitable for NER tasks in qualitative studies. Importantly, we manually validated our model and developed a web-based application that delivers near-instant annotations, making this work more directly translatable into research and clinical workflows.

This study's findings have practical value for qualitative research. By rapidly coding large collections of physician-patient transcripts, the developed NLP application/tool could reduce the time and resources required for manual review and support larger, multicenter studies that examine communication patterns across different health care settings. Researchers could use the application to qualitatively examine how patients describe and interpret symptoms, express emotions and financial concerns, discuss health behaviors and treatment experiences, communicate preferences, and participate in decision-making. In clinical quality improvement settings, the system could also help identify patterns in how clinicians discuss diagnostic testing, treatment concerns, follow-up plans, and patient questions.

The results indicate the value of future hybrid approaches. Clinical communication is relational and context-rich, often requiring models to capture longer sequences of text, such as emotional expressions, question-answer exchanges, and multitopic utterances. While transformer-based models provided the strongest performance in this study, a hybrid system that combines the precision of rule-based methods with the contextual depth of transformer models may better handle this variability. This concept parallels manual qualitative coding strategies such as "lumping," where longer text segments are grouped to identify overarching themes, and "splitting," where text is divided into smaller codes to capture finer details [53]. Similar considerations exist in quantitative research, where broad psychometric measures may fail to capture the full dimensionality of a construct [54]. Extending hybrid approaches could make NLP pipelines more adaptable to domains where conversation topics vary widely.

Despite its strengths, this study has several limitations. While the RoBERTa model performed well, it can still generate incorrect or incomplete annotations. Therefore, the application is best suited as a human-guided tool rather than a replacement for qualitative coding. Researchers could use model-generated annotations as a starting point and then review and edit them as needed, rather than manually annotating each transcript from scratch. Although this dataset was larger than those used in previous studies, future work should scale evaluation across more diverse health care contexts to improve robustness. We also acknowledge that larger and more recent proprietary LLMs, such as GPT 5.2 and Gemini 3.5, may achieve stronger zero-shot prompt performance. However, we could not evaluate these models in our local environment due to computational limitations and privacy concerns. Additionally, LLMs are known to be affected by different prompts and wording, which was not explored in this study; a single standard prompt was used to generate the results. Moreover, further clinical validation through direct application of our developed tool is needed to assess its performance in a qualitative study. This validation will be essential to determine its clinical usefulness further and support its future implementation in practice. Future research should also explore fine-tuning LLMs and further validate the NER tool in additional studies, including its ability to handle longer transcripts without a sliding window for stronger contextual understanding.

Conclusions

This study demonstrates the feasibility and effectiveness of transformer-based models to process large-scale qualitative clinical communication data in cancer care. The RoBERTa NER model achieved the highest overall performance, outperforming traditional ML, rule-based approaches, and the OoB Qwen3 model. The real promise of this approach lies in its ability to enable large-scale, multi-center studies of communication and decision-making that can model systems-level factors. At present, small, single-center

studies are limited by nesting effects, which undermine statistical power and prevent the large sample sizes necessary to answer high-level questions about structural and policy factors that shape communication. By accurately identifying and classifying clinically relevant information from complex medical dialogues, our approach addresses a key bottleneck in qualitative health research: the time-intensive and labor-intensive nature of manual coding. The development of a real-time, web-based application further enhances the practical use of this work, enabling rapid transcript analysis that can support qualitative research. These findings highlight the potential for integrating NLP into qualitative research workflows to improve efficiency, consistency, and scalability.

Acknowledgments

The authors would like to acknowledge Mr Manan Patel for programming assistance with this study.

Generative AI Use

The authors attest that there was no use of generative AI technology in the generation of text, figures, or other informational content in this manuscript.

Funding

This study was funded by the National Cancer Institute (R01 CA124607), with LAS serving as the principal investigator. This study was funded for a total of 5 years starting from 2008 to 2012. The study was funded for a total of US $3,840,853.

Data Availability

The dataset used in this paper was generated from the NIH-funded grant R01CA124607. Interview transcripts collected as part of this project are used for secondary analysis in this study. This dataset cannot be made publicly available to protect the privacy and security of patients and participants.

Authors' Contributions

Conceptualization: JSP, LAS, MDT
Formal analysis: JSP, ZAS, RK, OM
Writing – review & editing: JSP, LAS, MDT, RK, OM

Conflicts of Interest

None declared.

Multimedia Appendix 1

Word cloud, the distribution of annotated entity text lengths, and the comparison of average word counts.
[DOCX File (Microsoft Word File), 1444 KB-Multimedia Appendix 1]

References

1. Granek L, Nakash O, Ariad S, Shapira S, Ben-David M. Mental health distress: oncology nurses' strategies and barriers in identifying distress in patients with cancer. Clin J Oncol Nurs. Feb 1, 2019;23(1):43-51. [doi: 10.1188/19.CJON.43-51] [Medline: 30681995]
2. Kaul S, Avila JC, Mutambudzi M, Russell H, Kirchhoff AC, Schwartz CL. Mental distress and health care use among survivors of adolescent and young adult cancer: a cross-sectional analysis of the National Health Interview Survey. Cancer. Mar 1, 2017;123(5):869-878. [doi: 10.1002/cncr.30417] [Medline: 27859009]
3. Kwekkeboom KL, Seng JS. Recognizing and responding to post-traumatic stress disorder in people with cancer. Oncol Nurs Forum. May 2002;29(4):643-650. [doi: 10.1188/02.ONF.643-650] [Medline: 12011911]
4. Zebrack B, Kayser K, Sundstrom L, et al. Psychosocial distress screening implementation in cancer care: an analysis of adherence, responsiveness, and acceptability. J Clin Oncol. Apr 1, 2015;33(10):1165-1170. [doi: 10.1200/JCO.2014.57.4020] [Medline: 25713427]
5. Cusatis R, Holt JM, Williams J, et al. The impact of patient-generated contextual data on communication in clinical practice: a qualitative assessment of patient and clinician perspectives. Patient Educ Couns. Apr 2020;103(4):734-740. [doi: 10.1016/j.pec.2019.10.020] [Medline: 31744702]
6. de Haes JC, van Knippenberg FC. The quality of life of cancer patients: a review of the literature. Soc Sci Med. 1985;20(8):809-817. [doi: 10.1016/0277-9536(85)90335-1] [Medline: 3890194]
7. Ekberg S, Parry R, Land V, et al. Communicating with patients and families about illness progression and end of life: a review of studies using direct observation of clinical practice. BMC Palliat Care. Dec 8, 2021;20(1):186. [doi: 10.1186/s12904-021-00876-2] [Medline: 34876096]
8. Fenton ATHR, Fletcher KM, Kizildag D, et al. Cancer caregivers' prognostic and end-of-life communication needs and experiences and their impact. J Pain Symptom Manage. Jan 2023;65(1):16-25. [doi: 10.1016/j.jpainsymman.2022.09.012] [Medline: 36198337]

9. Kwame A, Petrucka PM. A literature-based study of patient-centered care and communication in nurse-patient interactions: barriers, facilitators, and the way forward. BMC Nurs. Sep 3, 2021;20(1):158. [doi: 10.1186/s12912-021-00684-2] [Medline: 34479560]
10. Michael N, Gorelik A, Georgousopoulou E, et al. Patient-caregiver communication concordance in cancer—refinement of the cancer communication assessment tool in an Australian sample. Support Care Cancer. Sep 2022;30(9):7387-7396. [doi: 10.1007/s00520-022-07163-7] [Medline: 35612665]
11. Sharkiya SH. Quality communication can improve patient-centred health outcomes among older patients: a rapid review. BMC Health Serv Res. Aug 22, 2023;23(1):886. [doi: 10.1186/s12913-023-09869-8] [Medline: 37608376]
12. Bourque MA, Loiselle CG. Patients' cancer care perceptions conceptualized through the Cancer Experience Measurement Framework. BMC Health Serv Res. May 23, 2022;22(1):693. [doi: 10.1186/s12913-022-07946-y] [Medline: 35606765]
13. Låver J, McAleavey A, Valaker I, Castonguay LG, Moltu C. Therapists' and patients' experiences of using patients' self-reported data in ongoing psychotherapy processes—a systematic review and meta-analysis of qualitative studies. Psychother Res. Mar 2024;34(3):293-310. [doi: 10.1080/10503307.2023.2222896] [Medline: 37322037]
14. Mulley AG, Trimble C, Elwyn G. Stop the silent misdiagnosis: patients' preferences matter. BMJ. Nov 8, 2012;345:e6572. [doi: 10.1136/bmj.e6572] [Medline: 23137819]
15. Sanda MG, Dunn RL, Michalski J, et al. Quality of life and satisfaction with outcome among prostate-cancer survivors. N Engl J Med. Mar 20, 2008;358(12):1250-1261. [doi: 10.1056/NEJMoa074311] [Medline: 18354103]
16. Sitzia J, Wood N. Patient satisfaction: a review of issues and concepts. Soc Sci Med. Dec 1997;45(12):1829-1843. [doi: 10.1016/s0277-9536(97)00128-7] [Medline: 9447632]
17. Truog RD. Patients and doctors—evolution of a relationship. N Engl J Med. Feb 16, 2012;366(7):581-585. [doi: 10.1056/NEJMp1110848] [Medline: 22335734]
18. Davidson E, Edwards R, Jamieson L, Weller S. Big data, qualitative style: a breadth-and-depth method for working with large amounts of secondary qualitative data. Qual Quant. 2019;53(1):363-376. [doi: 10.1007/s11135-018-0757-y] [Medline: 30930496]
19. Massie MJ. Prevalence of depression in patients with cancer. J Natl Cancer Inst Monographs. Jul 1, 2004;2004(32):57-71. [doi: 10.1093/jncimonographs/lgh014]
20. Söllner W, DeVries A, Steixner E, et al. How successful are oncologists in identifying patient distress, perceived social support, and need for psychosocial counselling? Br J Cancer. Jan 2001;84(2):179-185. [doi: 10.1054/bjoc.2000.1545] [Medline: 11161373]
21. Greenfield G, Ignatowicz AM, Belsi A, et al. Wake up, wake up! It's me! It's my life! Patient narratives on person-centeredness in the integrated care context: a qualitative study. BMC Health Serv Res. Nov 29, 2014;14(1):619. [doi: 10.1186/s12913-014-0619-9] [Medline: 25471663]
22. Kusurkar RA, Mak-van der Vossen M, Kors J, et al. "One size does not fit all": the value of person-centred analysis in health professions education research. Perspect Med Educ. Aug 2021;10(4):245-251. [doi: 10.1007/s40037-020-00633-w] [Medline: 33284407]
23. Liang M, Liang J, Xu J, Chen Q, Lu Q. Experience of multidisciplinary medical teams on humanistic palliative care in oncology wards: a descriptive qualitative study in Southern China. BMJ Open. Feb 26, 2024;14(2):e074628. [doi: 10.1136/bmjopen-2023-074628] [Medline: 38413159]
24. Ullrich C, Stürmlinger A, Wensing M, Krug K. Qualitative research methods in medical dissertations: an observational methodological study on prevalence and reporting quality of dissertation abstracts in a German university. BMC Med Res Methodol. Dec 10, 2020;20(1):301. [doi: 10.1186/s12874-020-01186-6] [Medline: 33302886]
25. Abram MD, Mancini KT, Parker RD. Methods to integrate natural language processing into qualitative research. Int J Qual Methods. Jan 1, 2020;19:1609406920984608. [doi: 10.1177/1609406920984608]
26. Crowston K, Allen EE, Heckman R. Using natural language processing technology for qualitative data analysis. Int J Soc Res Methodol. Nov 2012;15(6):523-543. [doi: 10.1080/13645579.2011.625764]
27. McMullin C. Transcription and qualitative methods: implications for third sector research. Voluntas. 2023;34(1):140-153. [doi: 10.1007/s11266-021-00400-3] [Medline: 34522070]
28. Carminati L. Generalizability in qualitative research: a tale of two traditions. Qual Health Res. Nov 2018;28(13):2094-2101. [doi: 10.1177/1049732318788379] [Medline: 30043686]
29. Allen KS, Hood DR, Cummins J, Kasturi S, Mendonca EA, Vest JR. Natural language processing-driven state machines to extract social factors from unstructured clinical documentation. JAMIA Open. Jul 2023;6(2):ooad024. [doi: 10.1093/jamiaopen/ooad024] [Medline: 37081945]

30. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010;17(5):507-513. [doi: 10.1136/jamia.2009.001560] [Medline: 20819853]
31. Soysal E, Wang J, Jiang M, et al. CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. J Am Med Inform Assoc. Mar 1, 2018;25(3):331-336. [doi: 10.1093/jamia/ocx132] [Medline: 29186491]
32. Buchalter RB, Kamath SD, Nair KG, Liska D, Khorana AA, Schmit SL. Novel hot and cold spots of young-onset colorectal cancer mortality in United States counties. Gastroenterology. Jun 18, 2022;163(4):1101. [doi: 10.1053/j.gastro.2022.06.044] [Medline: 35728692]
33. Cheligeer C, Wu G, Lee S, et al. BERT-based neural network for inpatient fall detection from electronic medical records: retrospective cohort study. JMIR Med Inform. Jan 30, 2024;12:e48995. [doi: 10.2196/48995] [Medline: 38289643]
34. Arnaud É, Elbattah M, Gignon M, Dequen G. Learning embeddings from free-text triage notes using pretrained transformer models. Presented at: 15th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2022) — Scale-IT-up / HEALTHINF track; Feb 9-11, 2022; Vienna, Austria. [doi: 10.5220/0011012800003123]
35. Pyne Y, Wong YM, Fang H, Simpson E. Analysis of “One in a Million” primary care consultation conversations using natural language processing. BMJ Health Care Inform. Apr 2023;30(1):e100659. [doi: 10.1136/bmjhci-2022-100659] [Medline: 37116948]
36. Feizollah A, Lin CY, O’Malley L, Thompson W, Listl S, Byrne M. The use of natural language processing to interpret unstructured patient feedback on health services: scoping review. J Med Internet Res. Aug 14, 2025;27:e72853. [doi: 10.2196/72853] [Medline: 40811803]
37. Lázaro E, Moscardó V. Qualitative health-related quality of life and natural language processing: characteristics, implications, and challenges. Health Care (Don Mills). 2024;12(19):2008. [doi: 10.3390/healthcare12192008]
38. Castellanos A, Jiang H, Gomes P, Vander Meer D, Castillo A. Large language models for thematic summarization in qualitative health care research: comparative analysis of model and human performance. JMIR AI. Apr 4, 2025;4:e64447. [doi: 10.2196/64447] [Medline: 40611510]
39. Parker RD, Mancini K, Abram MD. Natural language processing enhanced qualitative methods: an opportunity to improve health outcomes. Int J Qual Methods. Oct 2023;22:16094069231214144. [doi: 10.1177/16094069231214144]
40. Mathis WS, Zhao S, Pratt N, Weleff J, De Paoli S. Inductive thematic analysis of healthcare qualitative interviews using open-source large language models: how does it compare to traditional methods? Comput Methods Programs Biomed. Oct 2024;255:108356. [doi: 10.1016/j.cmpb.2024.108356] [Medline: 39067136]
41. Rogers HL, Dumenci L, Epstein RM, Siminoff LA. Impact of patient gender and race and physician communication on colorectal cancer diagnostic visits in primary care. J Womens Health (Larchmt). May 2019;28(5):612-620. [doi: 10.1089/jwh.2018.6961] [Medline: 30489201]
42. Sureshbhai Patel J, Rao R, Brandon R, et al. Develop a natural language processing pipeline to automate extraction of periodontal disease information from electronic dental clinical notes. Presented at: ICMHI ’22: Proceedings of the 6th International Conference on Medical and Health Informatics; May 13-15, 2022; Virtual Event Japan. [doi: 10.1145/3545729.3545744]
43. Patel JS, Brandon R, Tellez M, et al. Developing automated computer algorithms to phenotype periodontal disease diagnoses in electronic dental records. Methods Inf Med. Dec 2022;61(S 02):e125-e133. [doi: 10.1055/s-0042-1757880] [Medline: 36413995]
44. Wang X, Wang Y. Sentence-level resampling for named entity recognition. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics; 2022:2151-2165. [doi: 10.18653/v1/2022.naacl-main.156]
45. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321-357. [doi: 10.1613/jair.953]
46. Epstein RM, Street RL Jr. Patient-centered communication in cancer care: promoting healing and reducing suffering. National Cancer Institute, National Institutes of Health, US Department of Health and Human Services; 2007. URL: https://cancercontrol.cancer.gov/sites/default/files/2020-06/pcc_monograph.pdf [Accessed 2026-10-12]
47. Petridis C. Text classification: neural networks vs machine learning models vs pre-trained models. arXiv. Preprint posted online on Dec 30, 2024. [doi: 10.48550/arXiv.2412.21022]
48. Liu Y, Ott M, Goyal N, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv. Preprint posted online on Jul 26, 2019. [doi: 10.48550/arXiv.1907.11692]
49. Neumann M, King D, Beltagy I, Ammar W. scispaCy: fast and robust models for biomedical natural language processing. In: Proceedings of the 18th BioNLP Workshop and Shared Task. Association for Computational Linguistics; 2019:319-327. [doi: 10.18653/v1/W19-5034]

50. Yang A, Li A, Yang B, et al. Qwen3 technical report. arXiv. Preprint posted online on May 14, 2025. [doi: 10.48550/arXiv.2505.09388]
51. Grandini M, Bagli E, Visani G. Metrics for multi-class classification: an overview. arXiv. Preprint posted online on Aug 13, 2020. [doi: 10.48550/arXiv.2008.05756]
52. Opitz J. A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. Trans Assoc Comput Linguist. Jun 25, 2024;12:820-836. [doi: 10.1162/tacl_a_00675]
53. Fang C, Markuzon N, Patel N, Rueda JD. Natural language processing for automated classification of qualitative data from interviews of patients with cancer. Value Health. Dec 2022;25(12):1995-2002. [doi: 10.1016/j.jval.2022.06.004] [Medline: 35840523]
54. Saldaña J. The Coding Manual for Qualitative Researchers. SAGE Publications Ltd; 2025. URL: https://collegepublishing.sagepub.com/products/the-coding-manual-for-qualitative-researchers-5-287917 [Accessed 2026-10-12]

Abbreviations

AUC: area under the curve
BERT: bidirectional encoder representations from transformer
BIOES: Beginning, Inside, Outside, End, and Single
BoW: bag-of-words
CLAMP: Clinical Language Annotation, Modeling, and Processing
cTAKES: Clinical Text Analysis and Knowledge Extraction System
eHOST: Extensible Human Oracle Suite of Tools
FN: false negative
FP: false positive
IRB: institutional review board
LDA: Latent Dirichlet Allocation
LLM: large language model
ML: machine learning
NER: named entity recognition
NLP: natural language processing
OoB: out-of-the-box
PCP: primary care physician
QoL: quality of life
RoBERTa: Robustly Optimized BERT Pretraining Approach
SVM: support vector machine
TF-IDF: term frequency and inverse document frequency
TN: true negative
TP: true positive
USP: unannounced standardized patient

Edited by Amaryllis Mavragani; peer-reviewed by Mahmoud Elbattah, Maria Chatzimina, Md Zakir Hossain, Yannan Lin; submitted 24.Nov.2025; final revised version received 04.Aug.2026; accepted 04.Aug.2026; published 26.Nov.2026

Please cite as:
Patel JS, Siminoff LA, Siddiqui ZA, Thomson MD, Katiyar R, Memar O
Automated Extraction and Coding of Qualitative Data Within Physician-Patient Communication in Cancer Care Using Natural Language Processing: Development and Validation Study
JMIR Form Res 2026;10:e88404
URL: https://formative.jmir.org/2026/1/e88404
doi: 10.2196/88404

© Jay Sureshbhai Patel, Laura A Siminoff, Zasim Azhar Siddiqui, Maria D Thomson, Ritwik Katiyar, Oqba Memar. Originally published in JMIR Formative Research (https://formative.jmir.org), 26.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.