ScienceDirect

Contents lists available at sciencedirect.com
Journal homepage: www.elsevier.com/locate/jval

Patient-Reported Outcomes

Natural Language Processing for Automated Classification of Qualitative Data From Interviews of Patients With Cancer

Chao Fang, PhD, Natasha Markuzon, PhD, Nikunj Patel, PharmD, Juan-David Rueda, PhD

ABSTRACT

Objectives: This study sought to explore the use of novel natural language processing (NLP) methods for classifying unstructured, qualitative textual data from interviews of patients with cancer to identify patient-reported symptoms and impacts on quality of life.

Methods: We tested the ability of 4 NLP models to accurately classify text from interview transcripts as "symptom," "quality of life impact," and "other." Interview data sets from patients with hepatocellular carcinoma (HCC) (n = 25), biliary tract cancer (BTC) (n = 23), and gastric cancer (n = 24) were used. Models were cross-validated with transcript subsets designated for training, validation, and testing. Multiclass classification performance of the 4 models was evaluated at paragraph and sentence level using the HCC testing data set and analyzed by the one-versus-rest technique quantified by the receiver operating characteristic area under the curve (ROC AUC) score.

Results: NLP models accurately classified multiclass text from patient interviews. The Bidirectional Encoder Representations from Transformers model generally outperformed all other models at paragraph and sentence level. The highest predictive performance of the Bidirectional Encoder Representations from Transformers model was observed using the HCC data set to train and BTC data set to test (mean ROC AUC, 0.940 [SD 0.028]), with similarly high predictive performance using balanced and imbalanced training data sets from BTC and gastric cancer populations.

Conclusions: NLP models were accurate in predicting multiclass classification of text from interviews of patients with cancer, with most surpassing 0.9 ROC AUC at paragraph level. NLP may be a useful tool for scaling up processing of patient interviews in clinical studies and, thus, could serve to facilitate patient input into drug development and improving patient care.

Keywords: Bidirectional Encoder Representations from Transformers, natural language processing, patient interviews, patient-reported outcomes.

VALUE HEALTH. 2022; 25(12):1995–2002

Introduction

The US Food and Drug Administration and European Medicines Agency have encouraged the collection and use of patient experience data to inform clinical practice, medical product development, and regulatory decision making, with the overarching aim of improving patient care.[1,2] In oncology, the value of collecting patient experience data is becoming increasingly recognized given that it may benefit patients' physical health and wellbeing, treatment decision making, delivery of care, clinical research, and policy making.[3,4] Cancer continues to be one of the leading causes of death worldwide.[5] A cancer diagnosis can have a negative impact on an individual's mental health, psychosociological wellbeing, and quality of life (QoL),[6–8] and patients may face a poor prognosis and short survival[9]; thus, incorporating patient experience data into cancer research and practice is important for improving patient care. In turn, there has been a drive by various stakeholders to identify and promote best practice in collecting patient-reported outcomes (PROs) in cancer clinical trials and implementing PROs in cancer care.[10–13]

In clinical trials of patients with cancer, PROs and QoL measures have been shown to be important prognostic tools through associations with tumor response and survival.[14–16] Symptom self-reporting during trials has demonstrated a variety of clinical benefits such as improved patient outcomes, including survival and QoL measures, and better use of healthcare resources.[17,18] Clinical trials have also found that regularly collecting QoL data in routine clinical practice improves patients' QoL and emotional wellbeing, as well as providers' awareness of patients' health.[19,20] These findings demonstrate how involving patients in research and measuring PROs and QoL may improve patient care.

In addition to quantitative methods, qualitative research methods, such as interviews and focus groups, are commonly used in cancer clinical trials to capture patients' feelings, beliefs, and attitudes toward their treatment and disease. This is important, given that research shows patients' priorities may differ from

physicians' priorities when making cancer treatment decisions.[21] Interview-based studies have provided important insights into the patient journey in cancer, from identifying barriers and facilitators to patient screening, to patient preferences in post-treatment follow-up.[22-24] Interviews help gain understanding of patients' perspectives, such as identifying methods to help raise prognostic awareness and the communication needs of patients,[25,26] as well as patients' experiences, for example, with lifestyle interventions,[27] palliative care,[28] and pain management,[29] which may help inform providers and policy makers.

Despite the valuable information that qualitative interviews can capture, processing interviews represents a major barrier to their widespread use in clinical research. Processing of interviews is generally performed manually, representing a time-consuming, labor-intensive, and expensive method.[30] An automated approach to interview processing could increase efficiency and decrease interobserver bias, improving the overall time, cost, and quality of interview processing.[31] Natural language processing (NLP) is an artificial intelligence methodology allowing the automatic processing of unstructured and free-form text that has been gaining popularity and growing in sophistication.[32] NLP has been used in a variety of medically related applications, including summarizing and extracting information from published biomedical texts[33]; evaluating speech features in interviews and other free texts to predict risk or onset of disease, such as psychosis and Alzheimer's disease[34,35]; and automating the clinical trial screening process to identify eligible participants.[36,37] In oncology, NLP has also been used to process event and temporal information in electronic medical records to establish clinical timelines, potentially aiding healthcare practitioners in patient diagnosis and care.[38] It has been used to extract and organize clinical information from free-text pathology reports to identify diagnoses, patient characteristics, and meaningful outcomes of interest and establish a pathologic diagnosis.[39,40] In conjunction with machine learning, NLP has also shown the ability to predict postoperative complications and hospital readmissions among women with ovarian cancer.[41] These examples highlight the potential for NLP to extract information that is of interest to providers to improve patient care.

Symptoms and their impacts on QoL are important information for healthcare providers, and the ability to identify them can help shape patient care and the drug development process. For example, in oncology, identifying symptoms and their impacts may improve the management of adverse events, help tailor information given to patients, and provide greater clarity on the benefit-risk ratio of treatments.[42] Previous interview-based studies in oncology using manual processing identified symptoms of importance to guide the development of standardized PROs that adequately assess outcomes.[43-48] Both within and outside of the field of oncology, studies have investigated the use of NLP to detect and analyze symptom information from free-text formats in electronic health records for capturing symptoms, classifying or characterizing disease, and studying adverse events and clinical outcomes.[49-52] Nevertheless, knowledge of NLP application to improve cancer-related symptoms and their impact on QoL from qualitative patient interview data is limited. No such application exists within hepatocellular carcinoma (HCC), biliary tract cancer (BTC), or gastric cancer (GC)/gastroesophageal junction cancer PRO research. The ability to extract QoL information from patients' interviews is critical to understand and address the limitations of current treatments, particularly in the oncology field where QoL is often impacted by treatment.[53] Additionally, such methods have the potential to expand the use of qualitative patient interview research in cancer clinical trials to evaluate patient-reported insights not captured by traditional surveys. Although potentially applicable to any disease, this may be of particular value in oncology, where interview-based studies are used for understanding patients' perceptions of their symptoms and impact on QoL and for collecting patient experience data to support patient-centered research and care. This proof-of-concept study sought to explore the use of novel NLP methods as a tool for classifying unstructured, qualitative textual data collected from interviews of patients with cancer to identify patient-reported symptoms and impacts on QoL. The aim was to test the predictive performance of different NLP models in the classification of elements of free text, namely "symptoms," "QoL impacts," and "other," from interview transcripts and to evaluate their suitability as a potential tool for identifying the impacts of symptoms on QoL in the oncology setting. In addition, this study aimed to understand how well learning from one model can be transferred to another cancer type without affecting accuracy.

Methods

Data Sets

The study comprised 3 interview data sets of transcripts from patients with liver cancer (HCC, n = 25), BTC (n = 23), and GC (n = 24). Manual processing of HCC and BTC interview data sets has previously been reported.[48,54] Participant demographics and clinical characteristics for each data set are listed in the Appendix in the Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.06.004. Semistructured interviews, roughly 1 hour and 15 minutes in length, were conducted by trained interviewers and covered a variety of topics, such as the patients' demographic background, disease background, signs, symptoms, and impacts on their daily life.

Each interview data set contained 3 files: conversation pattern, quotation manager, and codebook. The conversation pattern was made up of the "question" from trained interviewers and "answer" from patients, obtained from the raw interview transcripts. The quotation manager contained human-assigned classifications from 2 independent coders, the number required to establish intercoder reliability,[55] who were external PhD scientists trained in qualitative research methods and related data analysis from IQVIA™. The coders mapped quotations to "detailed classifications" (ie, detailed descriptions such as "abdominal distress," "back pain," and "cough"), with assigned classifications serving as the ground truth for training, validating, and evaluating the NLP models. The codebook contained "detailed classifications" mapped to a "general grouped classification" (ie, "symptom," "QoL impact," and "other"), which was used when training the symptom-specific and QoL impact-specific classifications. The "other" classification represents text not classified as "symptom" or "QoL impact," thereby containing text that does not provide information on symptoms or impacts on QoL and highlighting information that may not be of interest.

NLP Models

We explored 4 different NLP models for automating the processing of patient interviews: (1) a traditional word-encoding approach, term frequency-inverse document frequency (TF-IDF)[56]; (2) an advanced global vectorization word-embedding approach, Global Vectors for Word Representation (GloVe)[57]; (3) a sequential data processing deep learning approach, recurrent neural networks (RNN)[58]; and (4) a cutting-edge bidirectional encoder language representation model, Bidirectional Encoder Representations from Transformers (BERT).[59,60] These 4 models

were selected as they take into consideration an increasing amount of contextual information, in turn representing increasing levels of sophistication and complexity, in the order they are listed.

Term frequency-inverse document frequency

TF-IDF is a context-free model, given that it does not consider the contextual information in a collection of documents or "corpus," representing the simplest model tested. It is a traditional word-encoding approach that serves as a baseline performance model.[56] TF-IDF acts as a model for determining the importance of words, by reducing the importance of words that occur more frequently and increasing the importance of words that appear less frequently. The model counts how many times a word appears in a single document, for example in an interview, normalized by the count of the most common word, known as the term frequency.[61] Subsequently, the model counts how many times a word appears in a corpus, for example, a data set of interviews, known as the document frequency.[61] The inverse of document frequency, inverse document frequency, when multiplied by term frequency in turn measures the importance of a word in a corpus.[61] For TF-IDF, data preprocessing involved the following steps: (1) each data set was standardized by lowercasing and punctuation stripping; (2) each data set was split into substrings (ie, words); (3) words or "tokens" were assigned a unique integer value to obtain index tokens; and (4) vector features were created by transforming each data set using the index into a dense float vector of embedding representation. Finally, we applied word cleaning steps, namely stop word removal and stemming, to reduce the dimensionality of the vocabulary.

Global Vectors for Word Representation

GloVe is an NLP model that considers the contextual information of a data set by obtaining how frequently a word appears in the context of another word; it represents a count-based model that takes into consideration the linear relationship between words and thereby their semantic relationship. GloVe is used for unsupervised learning of word representation that obtains vector (numerical) representations for words; this allows it to capture both global (ie, word-word cooccurrence counts) and local (ie, word similarity) statistics of a corpus.[57] For GloVe, data preprocessing steps were as follows: (1) pretrained GloVe embeddings were downloaded (http://nlp.stanford.edu/data/glove.6B.zip), which contain text-encoded vectors of different sizes (50-, 100-, 200-, and 300-dimensional vectors); (2) dictionary mapping words (ie, strings) to the vector representations were generated; (3) an embedding matrix was prepared, where the entry at index i is the pretrained vector for the word of index i in our vectorizer's vocabulary; and (4) the pretrained GloVe word embeddings matrix was loaded into the model's embedding layer and the GloVe embedding weights were fixed during model training. We built a neural network with one embedding layer that weighs every word in the sequence with the corresponding vector.

Recurrent neural networks

RNN models are sequential models that are powerful for processing sequence data, such as time series or natural languages.[58] An RNN sequential model iterates over the timesteps of a sequence and "memorizes" the information about the timesteps it has seen so far, thereby taking into consideration contextual information and processing information unidirectionally.[58] In this study, the RNN sequential model used an embedding layer as an encoder to encode patient interview sentences. This embedding layer (input vocabulary of size 1000 and output embedding dimension of size 64) stores one vector per word by converting the sequences of word indices to sequences of vectors. We then applied an RNN layer to process sequence input by iterating through the elements. In this study, we tested 2 types of RNN layers: long-short term memory[62] and gated recurrent unit.[63] In this study, we tested an RNN sequential model with 1 or 2 RNN layers.

Bidirectional Encoder Representations from Transformers

BERT represents the most sophisticated NLP model used in this study, given that it can capture the context of a word given its position within a sentence by considering bidirectional contextual information. In addition, BERT models are pretrained on large unlabeled data sets and then fine-tuned for a specific downstream task; for example, in this study, this model was fine-tuned using the patient interview data sets. This transfer learning process brings robustness to the models. State-of-the-art BERT models learn dynamic word embedding using a self-attention mechanism, which is a mechanism that relates different positions of a sequence to compute its representation.[59,60] The BERT family of models uses the Transformer encoder architecture to process each token of input text in the full context of all tokens before and after, thereby considering the whole contextual information. For BERT models, text inputs were transformed to numeric token IDs and arranged in several vectors before being input to BERT. TensorFlow Hub (https://www.tensorflow.org/hub) provides preprocessing models for BERT models. We applied the built-in preprocessing layers from BERT to preprocess text. We fine-tuned pretrained small version BERT models downloaded from TFHub (https://tfhub.dev/google/collections/bert/1), using transfer learning on the patient interview data set. Five BERT models were evaluated, each with a different number of Transformer blocks (4, 6, 8, 10, or 12).

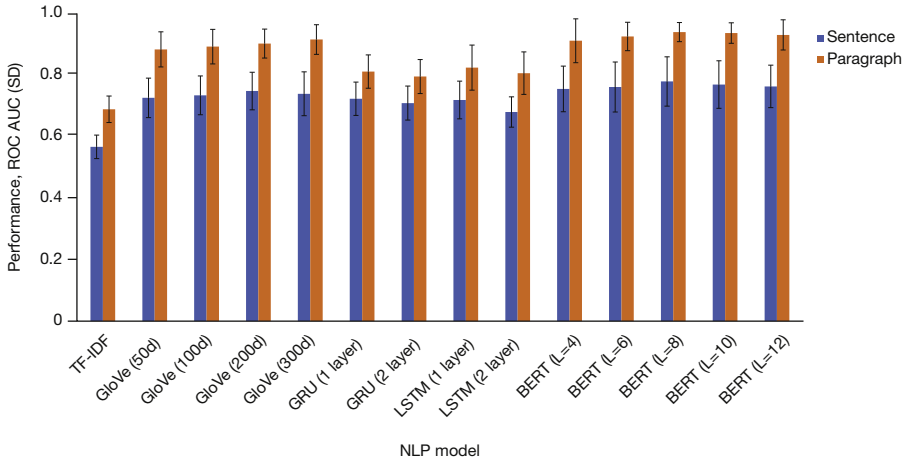
Analyses

Cross-validation was used to evaluate the performance of the NLP models by dividing the data sets into training, validation, and test data sets. The training data set was used to train the models, the validation data set was used to fine-tune the hyperparameters of the models, and the test data set was used to evaluate the performance of the models. The multiclass classification performance of the 4 models in predicting text classification ("symptom," "QoL impact," and "other") was evaluated at paragraph and sentence level across all patient interviews within a type of cancer and analyzed using the mean receiver operating characteristic area under the curve (ROC AUC) score. Due to an imbalance in classifications ("symptom" and "QoL impact" accounted for 70%-80% of classifications), a bootstrap approach was applied to randomly sample an equal number of "symptom," "QoL impact," and "other" classifications to form a balanced data set for training the NLP models. No resampling methods were applied to test data sets.

The BERT model was further analyzed using the LIME visualization tool[64] and confusion matrix. The confusion matrix compares the classification performance of a model on a set of test data with human-assigned classifications, referred to as the ground truth, to determine the predictive ability of a model in accurately classifying text.

The translatability of a model's multiclass classification performance from one type of cancer to a different type of cancer was evaluated on the BERT model. For cross-cancer–type evaluation analysis, the HCC or BTC training data set was used to train the

Figure 1. Multiclass classification performance of NLP models, trained and evaluated using the HCC data set, in predicting text classification ("symptom," "QoL impact," and "other") at the paragraph and sentence level. Error bars represent SD.



BERT indicates Bidirectional Encoder Representations from Transformers; d, dimension; GloVe, Global Vectors for Word Representation; GRU, gated recurrent unit; HCC, hepatocellular carcinoma; L, number of Transformer blocks; LSTM, long-short term memory; NLP, natural language processing; NN, neural network; QoL, quality of life; ROC AUC, receiver operating characteristic area under the curve; TF-IDF, term frequency-inverse document frequency.

model, and HCC, BTC, or GC test data sets were used to evaluate the performance of the BERT model in predicting text classification ("symptom," "QoL impact," and "other"). Multiclass classification performance was analyzed using the confusion matrix and one-versus-rest technique using the ROC AUC score.

Results

For identifying "symptom," "QoL impact," and "other" content in patient interviews, all NLP models performed better at paragraph versus sentence level in multiclass classification using the HCC data set (Fig. 1). Similar performance was achieved using the BTC and GC data sets (data not shown). The BERT models generally outperformed all other NLP models at both paragraph and sentence levels. This was followed by GloVe and RNN models, and TF-IDF was the lowest performing model. Among the various BERT models, improved predictive performance was seen when increasing the Transformer blocks from 4 to 6 and 8, with no observable improvement with larger Transformer blocks. With the GloVe models, the larger the vector size the higher the predictive performance. For the RNN models, long-short term memory outperformed gated recurrent unit networks at both the 1- and 2-layer level.

Using the HCC data sets, the BERT model achieved a mean ROC AUC of 0.94 in predicting the classification of "symptom," "QoL impact," and "other" words from 25 patient interviews (Fig. 2). Further analyses show examples of the prediction probabilities the BERT model predicted for each classification ("symptom," "QoL impact," and "other") and the weighting of words that contributed to this prediction. In Appendix Figure 1A in the Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.06.004, "nausea" and "radiation" had the highest weighting in positively and negatively contributing toward the symptom prediction probability, respectively. In Appendix Figure 1B in the Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.06.004, "activities" had the highest weighting in positively contributing toward QoL impact prediction probability.

Figures 2 to 4 show results from the cross-cancer–type evaluation analysis to examine the translatability of a model's performance from one type of cancer to another. Initially, each BERT model was trained and evaluated using the HCC data set (Fig. 2) and then evaluated using the BTC data set (Fig. 3) and GC data set (Fig. 4). The BERT model was also trained used the BTC data set and evaluated using the HCC data set (Appendix Fig. 2 in the Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.06.004).

For identifying "symptom," "QoL impact," and "other" words in patient interviews, the BERT model showed similar predictive performance in multiclass classification between balanced and imbalanced data sets, regardless of the data set used to train (HCC or BTC) and test (HCC, BTC, or GC) the model (Figs. 3 and 4, Appendix Fig. 2 in the Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.06.004). The highest predictive performance was seen when using the HCC data set to train and the HCC or BTC data set to test the BERT model (Fig. 3).

Discussion

Qualitative patient interviews of individuals affected by a particular disease or condition are an important method for understanding the impact of disease and treatment on QoL, particularly in disease areas such as cancer, which represent a substantial clinical and humanistic burden to patients.[6,65] Of the 4 NLP models we examined, the BERT model generally outperformed all other models in accurately predicting the multiclass classification of text, showing a similar predictive performance in classifying text regardless of which data set was used to train or test the model. These findings demonstrate the ability of the BERT model to accurately classify text from interviews conducted in different patient populations with various types of cancer collected from separate studies, highlighting the potential for NLP to be used as a tool for automating the processing of interviews of patients with cancer.

In this study, the predictive performance of 4 different types of NLP models was tested. Unlike TF-IDF and GloVe, which do not consider the order of words from inputs, the RNN and BERT models consider the unidirectional and bidirectional contextual information from inputs, respectively. BERT models take into

Figure 2. Cross-validation of the BERT model for multiclass classification using the HCC data set.

		Prediction		
	Total N = 1627	**Symptom**	**QoL impact**	**Other**
Ground truth	**Symptom**	**787**	139	44
	QoL impact	54	**299**	20
	Other	17	14	**253**
		Mean ROC AUC (SD): 0.940 ± 0.028		

BERT indicates Bidirectional Encoder Representations from Transformers; HCC, hepatocellular carcinoma; QoL, quality of life; ROC AUC, receiver operating characteristic area under the curve.

consideration the whole contextual information of any given word and are fine-tuned for specific tasks through pretraining on large unlabeled data sets by transfer learning. These features bring complexity, robustness, and scalability to the models, demonstrated by the higher predictive performance seen with the BERT models in this study. These findings show that the ability of the BERT models to consider contextual information underpins the relevance of the BERT models to real-world application, such as clinical trial data sets.

The similar predictive performance of the BERT model across interviews from different patient populations and studies highlights the potential generalizability of the model. Although validation of the model is required in other types of cancers and interviews from various studies, these results suggest that the BERT model may be scalable across a range of patient populations. This is of particular importance given that there are >100 different types of cancer.[66]

The current study suggests that NLP can be applied to the processing of important qualitative data. NLP may provide several advantages over manual processing. First, the BERT model was able to classify text as "symptom," "QoL impact," or "other" at a high accuracy, with predictive probabilities ranging between 75%

Figure 3. Performance of the BERT model using the HCC training data set and BTC test data set, with (A) imbalanced and (B) balanced data sets.

A

		Prediction		
	Total N = 1254	**Symptom**	**QoL impact**	**Other**
Ground truth	**Symptom**	**672**	70	64
	QoL impact	29	**178**	**10**
	Other	20	11	**200**
		ROC AUC (one-vs-rest): 0.944		

B

		Prediction		
	Total N = 651	**Symptom**	**QoL impact**	**Other**
Ground truth	**Symptom**	**184**	11	22
	QoL impact	29	**178**	10
	Other	20	9	**188**
		ROC AUC (one-vs-rest): 0.953		

BERT indicates Bidirectional Encoder Representations from Transformers; BTC, biliary tract cancer; HCC, hepatocellular carcinoma; QoL, quality of life; ROC AUC, receiver operating characteristic area under the curve.

Figure 4. Performance of the BERT model using the HCC training data set and GC test data set, with (A) imbalanced and (B) balanced data sets.

A

Prediction

Total N = 2207	Symptom	QoL impact	Other
Symptom	**1088**	172	140
QoL impact	77	**258**	24
Other	67	33	**348**
	ROC AUC (one -vs-rest): 0.890		

Ground truth

B

Prediction

Total N = 1077	Symptom	QoL impact	Other
Symptom	**265**	50	44
QoL impact	77	**258**	24
Other	57	26	**276**
	ROC AUC (one -vs-rest): 0.889		

Ground truth

BERT indicates Bidirectional Encoder Representations from Transformers; GC, gastric cancer; HCC, hepatocellular carcinoma; QoL, quality of life; ROC AUC, receiver operating characteristic area under the curve.

and 95%. The automatic processing of interviews to identify symptoms and their impacts could scale up this process, highlighting the potential application of NLP in a clinical care and trial setting. Furthermore, using NLP to process patient interview data eliminates the potential for subjectivity introduced by experience and personal biases of researchers[67] and the potential impact of interobserver variability, in the context of manual processing performed using a predefined coding system.

There are several limitations to using NLP, which highlight future work that should be conducted to further evaluate NLP as a potential tool for processing patient experience data. Given the objective rules governing NLP, subtleties in the meaning and intention of the interviewee that may be apparent to a human will not be captured in the NLP of interview transcripts, which may introduce inaccuracies. This brings to the forefront a simultaneous advantage and disadvantage of NLP, namely the generation of simpler and more specific data from rich and complex data, respectively. Although examining the latter may provide more meaningful insights, obtaining key information through NLP is essential to more widespread collection of patient experience data on a larger scale. Hence, simple versus complex data can arguably be complementary to one another. As with other research methods, relatively small data sets may further hinder the generalizability of findings given that sampling may not be well representative. In the current study, data sets from distinct populations of patients with cancer at different stages and across a wide age range were included; nevertheless, validation in larger patient interview data sets is required to address the aforementioned issues. Although NLP would improve time and cost-effectiveness, interviews themselves are a time-consuming and laborious process. Therefore, future research may seek to concentrate on not only automating the interview process but also evaluating patient experience from a variety of unstructured text formats, including online health communities and social media and surveys allowing free-text responses.[68] In addition, it is important to acknowledge the extent of automation of NLP; human input is still required, for example, in analyzing and interpreting the classified elements in their context, by considering other elements that may be of importance in the interview transcripts. In this study, the output of NLP provides interview transcripts highlighted with text classified as "symptom," "QoL impact," or "other." The subsequent step would involve a human analyzing the findings and producing figures or tables that capture frequency counts of symptom and impacts of QoL alongside relevant quotes, given that these are the elements of interest. Moreover, investigations into NLP for classifying elements other than "symptom" and "QoL impact" are needed to further explore the value of NLP as a tool for processing patient experience data. Future NLP modeling may benefit, for example, from classifying specific disease or treatment-related symptoms, disease severity, and disease duration. Further work will identify the minimal amount of information needed to train the model and run the processing of patient experience data and to evaluate whether supervised techniques, with reinforced learning, may provide additional insights into the current unsupervised approach.

The patient's voice and PROs are critical elements of patient-focused drug development in oncology, and their importance in medical product development and regulatory decision making is recognized by health authorities.[1,2] This proof-of-concept study demonstrates the application of NLP to understand the impact of disease and treatment on patient-reported symptoms and QoL. The automated processing of interview transcripts with NLP can be applied to future cancer clinical trials to build on our approach and explore its integration, for example, by comparing data processed from interviews at baseline with later timepoints. Furthermore, NLP for automated processing of patient interviews

could be applied outside of the field of oncology to other disease areas, particularly where patient-reported symptoms and QoL are critical components of patient care. Overall, by scaling up and automating the processing of patient experience data, NLP could aid in identifying information that is important and relevant to patients, thereby improving the design of future clinical trials and quality of care.

Conclusions

Understanding patient-reported symptoms and impacts on QoL is an important element of medical product development. Qualitative patient interview methods are often used to collect patients' experience of a disease or treatment, generating rich, unstructured data, and its processing can be laborious. The BERT NLP model used in our research demonstrated a proof-of-concept approach that might be more accurate in characterizing patient-reported symptoms and QoL using interview transcripts from patients with HCC, BTC, or GC. Automatic processing of interview transcripts to extract relevant patient experience data may scale up processing, ultimately facilitating patient input into drug development and improving patient care within oncology and other therapy areas.

Supplemental Materials

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2022.06.004.

Article and Author Information

Accepted for Publication: June 12, 2022

Published Online: July 12, 2022

doi: https://doi.org/10.1016/j.jval.2022.06.004

Author Affiliations: Oncology Biometrics ML/AI, AstraZeneca, Waltham, MA, USA (Fang, Markuzon); US Medical Affairs, AstraZeneca, Gaithersburg, MD, USA (Patel); Oncology Market Access and Pricing, AstraZeneca, Gaithersburg, MD, USA (Rueda).

Correspondence: Natasha Markuzon, PhD, Oncology Biometrics ML/AI, AstraZeneca, 760 Winter St, Waltham, MA 02451, USA. Email: natasha.markuzon@astrazeneca.com

Author Contributions: *Concept and design:* Fang, Markuzon, Patel, Rueda
Acquisition of data: Fang, Patel, Rueda
Analysis and interpretation of data: Fang, Markuzon, Patel, Rueda
Drafting of the manuscript: Fang, Markuzon, Patel, Rueda
Critical revision of the paper for important intellectual content: Fang, Markuzon, Patel, Rueda
Statistical analyses: Fang, Markuzon, Patel, Rueda
Provision of study materials or patients: Patel
Obtaining funding: Patel
Administrative, technical, or logistic support: Markuzon, Patel, Rueda
Supervision: Markuzon, Patel, Rueda

Conflict of Interest Disclosures: All authors are employed by and reported stock ownership in AstraZeneca. No other disclosures were reported.

Funding: This study was funded by AstraZeneca.

Role of Funder: The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Proprietary Statement: Primary qualitative patient interview data are owned by AstraZeneca. Data from secondary analysis of primary data through natural language processing, models, and methods described in this article are nonproprietary.

Acknowledgment: The authors thank the study participants. The authors acknowledge Jennifer Philippou for her contributions toward the development of the models described in this article. Medical writing support, under the guidance of the authors, was provided by Sonya Frazier, PhD, CMC Connect, McCann Health Medical Communications, with funding from AstraZeneca in accordance with Good Publication Practice (GPP3) guidelines (Ann Intern Med 2015;163:461-464).

REFERENCES

1. FDA patient-focused drug development guidance series for enhancing the incorporation of the patient's voice in medical product development and regulatory decision making. US Food and Drug Administration. https://www.fda.gov/drugs/development-approval-process-drugs/fda-patient-focused-drug-development-guidance-series-enhancing-incorporation-patients-voice-medical. Accessed July 1, 2022.
2. Regulatory guidance for the use of health-related quality of life (HRQL) measures in the evaluation of medicinal products. European Medicines Agency. https://www.ema.europa.eu/en/regulatory-guidance-use-health-related-quality-life-hrql-measures-evaluation-medicinal-products. Accessed July 1, 2022.
3. Bottomley A, Pe M, Sloan J, et al. Analysing data from patient-reported outcome and quality of life endpoints for cancer clinical trials: a start in setting international standards. *Lancet Oncol*. 2016;17(11):e510–e514.
4. Appendix 2 to the guideline on the evaluation of anticancer medicinal products in man - the use of patient-reported outcome (PRO) measures in oncology studies. European Medicines Agency. https://www.ema.europa.eu/en/appendix-2-guideline-evaluation-anticancer-medicinal-products-man-use-patient-reported-outcome-pro. Accessed July 1, 2022.
5. Total cancers. Global Health Metrics. https://www.thelancet.com/pb-assets/Lancet/gbd/summaries/diseases/neoplasms.pdf. Accessed July 1, 2022.
6. Kang D, Shim S, Cho J, Lim HK. Systematic review of studies assessing the health-related quality of life of hepatocellular carcinoma patients from 2009 to 2018. *Korean J Radiol*. 2020;21(6):633–646.
7. Kouhestani M, Ahmadi Gharaei H, Fararouei M, Ghahremanloo HH, Ghaiasvand R, Dianatinasab M. Global and regional geographical prevalence of depression in gastric cancer: a systematic review and meta-analysis [published online May 20, 2020]. BMJ Support Palliat Care. https://doi.org/10.1136/bmjspcare-2019-002050.
8. Kuswanto CN, Stafford L, Sharp J, Schofield P. Psychological distress, role, and identity changes in mothers following a diagnosis of cancer: a systematic review. *Psychooncology*. 2018;27(12):2700–2708.
9. Allemani C, Weir HK, Carreira H, et al. Global surveillance of cancer survival 1995-2009: analysis of individual data for 25,676,887 patients from 279 population-based registries in 67 countries (CONCORD-2). *Lancet*. 2015;385(9972):977–1010.
10. Kluetz PG, Slagle A, Papadopoulos EJ, et al. Focusing on core patient-reported outcomes in cancer clinical trials: symptomatic adverse events, physical function, and disease-related symptoms. *Clin Cancer Res*. 2016;22(7):1553–1558.
11. Stover AM, Tompkins Stricker C, Hammelef K, et al. Using stakeholder engagement to overcome barriers to implementing patient-reported outcomes (PROs) in cancer care delivery: approaches from 3 prospective studies. *Med Care*. 2019;57(suppl 5 suppl 1):S92–S99.
12. Bhatnagar V, Hudgens S, Piault-Louis E, et al. Patient-reported outcomes in oncology clinical trials: stakeholder perspectives from the accelerating anticancer agent development and validation workshop 2019. *Oncologist*. 2020;25(10):819–821.
13. Core patient-reported outcomes in cancer clinical trials. Guidance for industry. US Food and Drug Administration. https://www.fda.gov/media/149994/download. Accessed July 1, 2022.
14. Victorson D, Soni M, Cella D. Metaanalysis of the correlation between radiographic tumor response and patient-reported outcomes. *Cancer*. 2006;106(3):494–504.
15. Quinten C, Coens C, Mauer M, et al. Baseline quality of life as a prognostic indicator of survival: a meta-analysis of individual patient data from EORTC clinical trials. *Lancet Oncol*. 2009;10(9):865–871.
16. Efficace F, Collins GS, Cottone F, et al. Patient-reported outcomes as independent prognostic factors for survival in oncology: systematic review and meta-analysis. *Value Health*. 2021;24(2):250–267.
17. Basch E, Deal AM, Kris MG, et al. Symptom monitoring with patient-reported outcomes during routine cancer treatment: a randomized controlled trial. *J Clin Oncol*. 2016;34(6):557–565.
18. Denis F, Lethrosne C, Pourel N, et al. Randomized trial comparing a web-mediated follow-up with routine surveillance in lung cancer patients [published correction appears in *J Natl Cancer Inst*. 2018;110(4):436]. *J Natl Cancer Inst*. 2017;109(9):djx029.
19. Velikova G, Booth L, Smith AB, et al. Measuring quality of life in routine oncology practice improves communication and patient well-being: a randomized controlled trial. *J Clin Oncol*. 2004;22(4):714–724.

20. Detmar SB, Muller MJ, Schornagel JH, Wever LDV, Aaronson NK. Health-related quality-of-life assessments and patient-physician communication: a randomized controlled trial. *JAMA*. 2002;288(23):3027–3034.
21. Rocque GB, Rasool A, Williams BR, et al. What is important when making treatment decisions in metastatic breast cancer? A qualitative analysis of decision-making in patients and oncologists. *Oncologist*. 2019;24(10):1313–1321.
22. Sutkowi-Hemstreet A, Vu M, Harris R, Brewer NT, Dolor RJ, Sheridan SL. Adult patients' perspectives on the benefits and harms of overused screening tests: a qualitative study. *J Gen Intern Med*. 2015;30(11):1618–1626.
23. Whitaker KL, Macleod U, Winstanley K, Scott SE, Wardle J. Help seeking for cancer 'alarm' symptoms: a qualitative interview study of primary care patients in the UK. *Br J Gen Pract*. 2015;65(631):e96–e105.
24. Roorda C, de Bock GH, Scholing C, et al. Patients' preferences for post-treatment breast cancer follow-up in primary care vs. secondary care: a qualitative study. *Health Expect*. 2015;18(6):2192–2201.
25. Hermann M, Kühne F, Rohrmoser A, Preisler M, Goerling U, Letsch A. Perspectives of patients with multiple myeloma on accepting their prognosis-a qualitative interview study. *Psychooncology*. 2021;30(1):59–66.
26. Farias AJ, Ornelas IJ, Hohl SD, et al. Exploring the role of physician communication about adjuvant endocrine therapy among breast cancer patients on active treatment: a qualitative analysis. *Support Care Cancer*. 2017;25(1):75–83.
27. Chang P-H, Lin C-R, Lee Y-H, et al. Exercise experiences in patients with metastatic lung cancer: a qualitative approach. *PLoS One*. 2020;15(4):e0230188.
28. Pini S, Hackett J, Taylor S, et al. Patient and professional experiences of palliative care referral discussions from cancer services: a qualitative interview study. *Eur J Cancer Care (Engl)*. 2021;30(1):e13340.
29. Rustøen T, Gaardsrud T, Leegaard M, Wahl AK. Nursing pain management–a qualitative interview study of patients with pain, hospitalized for cancer treatment. *Pain Manag Nurs*. 2009;10(1):48–55.
30. Malterud K. Qualitative research: standards, challenges, and guidelines. *Lancet*. 2001;358(9280):483–488.
31. Crowston K, Allen EE, Heckman R. Using natural language processing technology for qualitative data analysis. *Int J Soc Res Methodol*. 2011;15(6):523–543.
32. Kreimeyer K, Foster M, Pandey A, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *J Biomed Inform*. 2017;73:14–29.
33. Moradi M, Dorffner G, Samwald M. Deep contextualized embeddings for quantifying the informative content in biomedical text summarization. *Comput Methods Programs Biomed*. 2020;184:105117.
34. Bedi G, Carrillo F, Cecchi GA, et al. Automated analysis of free speech predicts psychosis onset in high-risk youths. *NPJ Schizophr*. 2015;1:15030.
35. Mahajan P, Baths V. Acoustic and language based deep learning approaches for Alzheimer's dementia detection from spontaneous speech. *Front Aging Neurosci*. 2021;13:623607.
36. Shivade C, Hebert C, Regan K, Fosler-Lussier E, Lai AM. Automatic data source identification for clinical trial eligibility criteria resolution. *AMIA Annu Symp Proc*. 2016;2016:1149–1158.
37. Zhang K, Demner-Fushman D. Automated classification of eligibility criteria in clinical trials to facilitate patient-trial matching for specific patient populations. *J Am Med Inform Assoc*. 2017;24(4):781–787.
38. Denny JC, Peterson JF, Choma NN, et al. Development of a natural language processing system to identify timing and status of colonoscopy testing in electronic medical records. *AMIA Annu Symp Proc*. 2009;2009:141.
39. Kehl KL, Xu W, Lepisto E, et al. Natural language processing to ascertain cancer outcomes from medical oncologist notes. *JCO Clin Cancer Inform*. 2020;4:680–690.
40. Acevedo F, Armengol VD, Deng Z, et al. Pathologic findings in reduction mammoplasty specimens: a surrogate for the population prevalence of breast cancer and high-risk lesions. *Breast Cancer Res Treat*. 2019;173(1):201–207.
41. Barber EL, Garg R, Persenaire C, Simon M. Natural language processing with machine learning to predict outcomes after ovarian cancer surgery. *Gynecol Oncol*. 2021;160(1):182–186.
42. Rydén A, Blackhall F, Kim HR, et al. Patient experience of symptoms and side effects when treated with osimertinib for advanced non-small-cell lung cancer: a qualitative interview substudy. *Patient*. 2017;10(5):593–603.
43. Garcia SF, Rosenbloom SK, Beaumont JL, et al. Priority symptoms in advanced breast cancer: development and initial validation of the National Comprehensive Cancer Network-Functional Assessment of Cancer Therapy-Breast Cancer Symptom Index (NFBSI-16). *Value Health*. 2012;15(1):183–190.
44. Holmstrom S, Naidoo S, Turnbull J, Hawryluk E, Paty J, Morlock R. Symptoms and impacts in metastatic castration-resistant prostate cancer: qualitative findings from patient and physician interviews. *Patient*. 2019;12(1):57–67.
45. Lee GL, Pang GSY, Akhileswaran R, et al. Understanding domains of health-related quality of life concerns of Singapore Chinese patients with advanced cancer: a qualitative analysis. *Support Care Cancer*. 2016;24(3):1107–1118.
46. Niklasson A, Paty J, Rydén A. Talking about breast cancer: which symptoms and treatment side effects are important to patients with advanced disease? *Patient*. 2017;10(6):719–727.
47. Williams LA, Bruera E, Badgwell B. In search of the optimal outcome measure for patients with advanced cancer and gastrointestinal obstruction: a qualitative research study. *Ann Surg Oncol*. 2020;27(8):2646–2652.
48. Patel N, Maher J, Lie X, et al. Understanding the patient experience in hepatocellular carcinoma: a qualitative patient interview study. *Qual Life Res*. 2021;31(2):473–485.
49. Koleck TA, Dreisbach C, Bourne PE, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *J Am Med Inform Assoc*. 2019;26(4):364–379.
50. Dreisbach C, Koleck TA, Bourne PE, Bakken S. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *Int J Med Inform*. 2019;125:37–46.
51. DiMartino L, Miano T, Wessell K, Bohac B, Hanson LC. Identification of uncontrolled symptoms in cancer patients using natural language processing. *J Pain Symptom Manag*. 2022;63(4):610–617.
52. Karmen C, Hsiung RC, Wetter T. Screen Internet forum participants for depression symptoms by assembling and enhancing multiple NLP methods. *Comput Methods Programs Biomed*. 2015;120(1):27–36.
53. Lewandowska A, Rudzki G, Lewandowski T, et al. Quality of life of cancer patients treated with chemotherapy. *Int J Environ Res Public Health*. 2020;17(19):6938.
54. Patel N, Lie X, Gwaltney C, et al. Understanding patient experience in biliary tract cancer: a qualitative patient interview study. *Oncol Ther*. 2021;9(2):557–573.
55. O'Connor C, Joffe H. Intercoder reliability in qualitative research: debates and practical guidelines. *Int J Qual Methods*. 2020;19:1–13.
56. Manning CD, Raghavan P, Schütze H. Tf-idf weighting. In: *Introduction to Information Retrieval*. Cambridge, United Kingdom: Cambridge University Press; 2008:118–120.
57. Pennington J, Socher R, Manning C. GloVe: global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). https://nlp.stanford.edu/pubs/glove.pdf. Accessed July 1, 2022.
58. Schmidhuber J. Deep learning in neural networks: an overview. *Neural Netw*. 2015;61:85–117.
59. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 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). https://aclanthology.org/N19-1423/. Accessed July 1, 2022.
60. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017. https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html. Accessed July 1, 2022.
61. Kim S-W, Gil J-M. Research paper classification systems based on TF-IDF and LDA schemes. *Hum Cent Comput Inf Sci*. 2019;9(30). https://doi.org/10.1186/s13673-019-0192-7.
62. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9(8):1735–1780.
63. Cho K, van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. Proceedings of the 2014 Conference on empirical methods in natural language processing (EMNLP). https://aclanthology.org/D14-1179/; 2014. Accessed July 1, 2022.
64. Ribeiro M, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. https://dl.acm.org/doi/10.1145/2939672.2939778; 2016. Accessed July 1, 2022.
65. Carrato A, Falcone A, Ducreux M, et al. A systematic review of the burden of pancreatic cancer in Europe: real-world impact on survival, quality of life and costs. *J Gastrointest Cancer*. 2015;46(3):201–211.
66. What is cancer? NIH National Cancer Institute. https://www.cancer.gov/about-cancer/understanding/what-is-cancer#types. Accessed July 1, 2022.
67. Anderson C. Presenting and evaluating qualitative research. *Am J Pharm Educ*. 2010;74(8):141.
68. Conway M, Hu M, Chapman WW. Recent advances in using natural language processing to address public health research questions using social media and consumergenerated data. *Yearb Med Inform*. 2019;28(1):208–217.