Journal of the American Medical Informatics Association, 2023, **30(10)**, 1673–1683
https://doi.org/10.1093/jamia/ocad139
Advance access publication 21 July 2023
Research and Applications

Research and Applications

Is the patient speaking or the nurse? Automatic speaker type identification in patient–nurse audio recordings

Maryam Zolnoori[1,2,*], Sasha Vergez[2], Sridevi Sridharan[2], Ali Zolnour[3], Kathryn Bowles[2], Zoran Kostic[4], and Maxim Topaz[1,2]

[1]School of Nursing, Columbia University, New York, New York, USA, [2]Center for Home Care Policy & Research, VNS Health, New York, New York, USA, [3]School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran, and [4]Department of Electrical Engineering, Columbia University, New York, New York, USA

***Corresponding Author:** Maryam Zolnoori, PhD, School of Nursing, Columbia University, 390 Fort Washington Avenue, New York, NY 825, USA; mz2825@cumc.columbia.edu, m.zolnoori@gmail.com

ABSTRACT

Objectives: Patient–clinician communication provides valuable explicit and implicit information that may indicate adverse medical conditions and outcomes. However, practical and analytical approaches for audio-recording and analyzing this data stream remain underexplored. This study aimed to 1) analyze patients' and nurses' speech in audio-recorded verbal communication, and 2) develop machine learning (ML) classifiers to effectively differentiate between patient and nurse language.

Materials and Methods: Pilot studies were conducted at VNS Health, the largest not-for-profit home healthcare agency in the United States, to optimize audio-recording patient–nurse interactions. We recorded and transcribed 46 interactions, resulting in 3494 "utterances" that were annotated to identify the speaker. We employed natural language processing techniques to generate linguistic features and built various ML classifiers to distinguish between patient and nurse language at both individual and encounter levels.

Results: A support vector machine classifier trained on selected linguistic features from term frequency-inverse document frequency, Linguistic Inquiry and Word Count, Word2Vec, and Medical Concepts in the Unified Medical Language System achieved the highest performance with an AUC-ROC = 99.01 ± 1.97 and an F1-score = 96.82 ± 4.1. The analysis revealed patients' tendency to use informal language and keywords related to "religion," "home," and "money," while nurses utilized more complex sentences focusing on health-related matters and medical issues and were more likely to ask questions.

Conclusion: The methods and analytical approach we developed to differentiate patient and nurse language is an important precursor for downstream tasks that aim to analyze patient speech to identify patients at risk of disease and negative health outcomes.

Key words: natural language processing; patient-nurse verbal communication; home healthcare; machine learning; audio-recording procedure

INTRODUCTION

Verbal communication between patients and clinicians is a rich and underused data source. This communication includes explicit and implicit information that can aid in identifying social and clinical cues that may indicate communication deficits, signs and symptoms of social or clinical instability, or pathological conditions.[1] For example, our previous study conducted in a home healthcare setting found that almost half of the clinical risk factors and interventions discussed between nurses and patients during home visits were not documented in the electronic health records (EHRs), neither in the clinical notes or structured data.[2] While patients and clinicians often have positive attitudes towards audio recording their verbal communication, little is known about practical pipeline and analytic approaches needed to audio-record and analyze this valuable data stream.[3]

Home healthcare is a setting where skilled clinicians (often registered nurses) provide healthcare services to patients in their homes.[4] Home healthcare patients are generally older adults aged ≥65 years and often are clinically complex and vulnerable patients with multiple chronic conditions.[5] Audio-recording and processing the verbal communication between patients and clinicians during home healthcare visits can potentially help to automatically screen patients for specific diseases (eg, Alzheimer's disease, Anxiety, and Depression)[6,7] and identify risk factors for negative outcomes.[2]

Emerging studies are starting to utilize natural language processing (NLP) methods to measure changes in linguistic parameters of the patient's speech for identifying those with severe health conditions (eg, neurological and mental disorders).[6,7] However, most speech analysis studies thus far were conducted in laboratory settings where patients were instructed to complete some speech production tasks (eg, reading tasks) in a short time (a few minutes). Little is known about our ability to analyze data collected during routine clinical encounters. One of the first steps in the automated analysis of data collected during routine patient–clinician verbal communication is recognizing the speaker type; who is speaking, patient or clinician (and potentially other parties)? We encountered this issue in a study aimed at analyzing patient–nurse communication during routine home

Received: 23 March 2023. **Revised:** 6 June 2023. **Editorial Decision:** 28 June 2023. **Accepted:** 16 July 2023
© The Author(s) 2023. Published by Oxford University Press on behalf of the American Medical Informatics Association. All rights reserved.
For permissions, please email: journals.permissions@oup.com

healthcare visits to improve the identification of patients at risk for hospitalizations.[1,8]

This pioneering study is starting to bridge the gaps in previous literature by building and testing an analytical pipeline (a chain of connected data processing steps) to differentiate the patient and nurse language in audio-recorded data. Specifically, we aimed to: (1) analyze the language used by patients and nurses during audio-recorded patient–nurse verbal communication using NLP feature extraction methods and (2) to create and test machine learning (ML) classifiers that can effectively differentiate between patient and nurse language. The output of this speaker-type identification work will help with downstream tasks that use patient–clinician communication to develop risk identification algorithms to identify patients at risk of a specific disease or negative outcome.

MATERIALS AND METHODS

This study was conducted at Visiting Nurse Service (VNS) Health, the largest not-for-profit home healthcare agency in the United States and approved by VNS Health's IRB (reference no. E20-003). We recruited 5 registered nurses caring for older adults in their homes who were willing to have their patient encounters recorded. The nurse introduced the study to the patient and if interested the research assistant called the patient for informed consent. To be eligible, patients needed to be fluent in English, able to communicate with nurses without caregiver help, and have the cognitive ability to read, understand, and sign informed consent independently. Figure 1 provides a schematic view for the metrology of this study.

Procedure of audio-recording patient–nurse encounters

We conducted a series of pilot studies to identify convenient procedures for audio-recording patient–nurse verbal communication. Specifically, we evaluated the functionality and usability of several audio-recording devices in the home care setting by the participation of nurses who evaluated the usability of the devices using System Usability Scale[9] (SUS) questionnaire and reporting their feedback in a semistructured interview. Details were presented in our previous study.[1] The SUS is a commonly used tool for assessing the ease of use and user satisfaction of a given product or system. Overall, Saramonic Blink658 (further referred to as Saramonic) received the highest usability score (SUS=65%) compared to other devices. This device is both portable and lightweight, equipped with 2 wireless microphones that can be attached to the clothing of both the patient and the nurse. The speech captured by the microphones is transferred to transmitters connected to the device, such as an iPod, where it is stored in 2 separate channels (see Supplementary Appendix SA for description of the device). Nurses found the procedure of audio-recording comfortable. Patients also reported feeling comfortable during the audio recording and reported no impact on their communication with nurses.[1]

Accuracy of automatic speech recognition system and speaker diarization

We used Amazon Web Service (AWS)-General Transcribe (GT) service as an automatic speech recognition (ADR) system. We used AWS-GT because it had the lowest word error rate (WER)=26% in transcribing patient–nurse verbal communication compared to other ASR systems, specifically AWS Medical Transcribe (AWS-MT) and Wave2Vec. Further details are presented in the previous study.[1]

Accuracy of speaker diarization

Additionally, we computed the accuracy of speaker diarization (a task to label audio recordings with classes that correspond to speaker identity) provided by AWS-GT. To do this, the member of the research team (SV) manually reviewed the accuracy of AWS-GT's identified speaker and then computed the overall accuracy (see Overall, we achieved high accuracy of speaker diarization = 96%. Supplementary Appendix SB provides a schematic view of the outcome of AWS-GT Transcribe).

Thanks for raising this concern. We have revised the section previously titled "Developing a pipeline for audio-recording patient-nurse verbal communication in home healthcare" to "Procedure of audio-recording patient-nurse encounters" in which we succinctly detail the procedure of audio-recording and the specific device used for this purpose. We directed readers to our previously published work for further information. We also addressed this concern in the previous comments (comment #1).

Sample of the study

In total, 46 patient–nurse encounters were audio-recorded for 23 eligible patients, each of whom underwent 2 audio-recorded encounters. We audio-recorded 2 encounters per patient to capture potential interaction variability and enhance our data's robustness. This approach diminishes the effects of extraordinary or atypical sessions, for example, when health-related issues, such as pain, lead patients to participate less actively in discussions. All the 46 patient–nurse encounters were transcribed using AWS-GT Transcribe system.

Feature-extraction methods

All audio-recorded encounters (verbal communications) were automatically transcribed to text using AWS-GT. To separate the patient's and nurse's language, we used the following feature-generation methods:

- *Term frequency-inverse document frequency (TF-IDF)*: TF-IDF is a statistical measure that determines the importance of words in a free-text document within a collection of documents. TF-IDF effectively distinguishes patient and nurse language for 2 reasons. The "term frequency" method considers how often certain words are used by either a patient or nurse, giving a higher value to frequently used terms. The "inverse document frequency" method reduces the significance of words common to both parties, highlighting unique terms that characterize patients or nurses. Previous studies have shown that TF-IDF is useful for classifying patient-generated data (patient messages) and characterizing clinicians' language.[10,11]
- *Linguistic inquiry and word count (LIWC) 2015*: LIWC 2015 is a manually curated lexical-based natural language processing tool developed by experts in the psychology of language. It contains a large selection of commonly used words and terms organized into 11 top-level categories, including function words, affective processes, social processes, cognitive processes, perceptual processes, biological processes, drives, relativity, informal language, personal

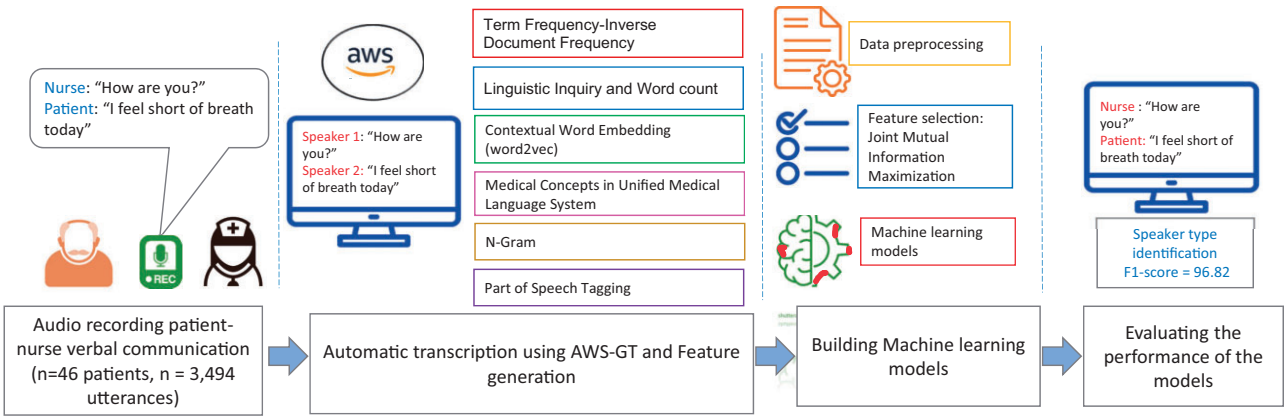

Figure 1. A schematic view of the methodology of the study.

concerns, and time orientation. This comprehensive categorization allows for a nuanced analysis of language use. For instance, patients' words and phrases might be more prevalent in categories related to affective processes (expressing emotions) and personal concerns. In contrast, clinicians might use more terms from categories of cognitive processes (problem-solving, cause-effect relationships). Secondly, LIWC can help identify linguistic patterns and trends unique to each group. For example, specific words or phrases (eg, netspeak and assent) might be more commonly used by nurses and less so by patients, or vice versa. These features helped characterize patients' and clinicians' language in several previous studies.[12,13]

- *Word2Vec*: Word2Vec is a popular word embedding technique that creates dense vector representations of words by learning their context from large, unlabeled free-text documents. By learning the context of words from surrounding words and identifying words with semantic similarities, Word2vec can distinguish how the same terms are used differently by nurses and patients, enhancing the differentiating power of ML algorithm built on the Word2Vec features for differentiating patient and nurse language. Word2Vec has been applied in several clinical studies to characterize clinicians' notes[14,15] and identify informative clues from patients' self-reported messages.[16,17]
- *Medical Concepts in Unified Medical Language System (UMLS)*: The UMLS is a comprehensive database that collects standardized medical terminology from various sources in the biomedical field. It links synonyms for medical concepts across different terminologies using concept unique identifiers (CUIs). Each CUI has a specific name and semantic type, such as "Headache," with a semantic type of "Sign/symptom." The standardized collection of medical concepts provided by the UMLS simplifies comparing and analyzing language patterns between patients and nurses. This is possible because nurses typically use these standardized medical concepts when discussing healthcare concerns and providing instructions. On the other hand, patients usually rely on every day, colloquial language to express their health worries and experiences. Thus, UMLS becomes a crucial tool in discerning the linguistic differences between the communication styles of patients and nurses. The UMLS has been widely used in various studies for extracting features from health-related documents for characterizing patients' and clinicians' reports and classification purposes.[18,19]
- *N-gram*: N-gram is a consecutive set of one or more words found in a text document, such as a patient's or nurse's language. N-grams, by capturing sequences of words that frequently co-occur in specific language groups (eg, patient and nurse), N-gram can unveil patterns in language use. By examining the frequency and distribution of these unique sequences, it is possible to discern those that are more prevalent in either patient or nurse language, thereby aiding in their differentiation. Additionally, previous research has demonstrated that n-grams can offer insights into the mode of communication between clinicians and patients, as well as the patients' health literacy.[20] In this study, we evaluated the effectiveness of n-grams ranging from 1-gram (unigram) to 10-gram for distinguishing between patient and nurse language. Among these, we found that the unigram was the most effective in distinguishing the language of the 2 groups. See "Results" section for more information.
- *Part of speech (POS) tagging*: POS tagging involves identifying and labeling the syntactic role of each word in a sentence, such as verbs, adjectives, adverbs, and nouns. POS tagging helps understand the language composition and syntactic structure of the language used by different individuals. Nurses, being medical professionals, may use more complex sentence structures and certain verbs or verb tenses. On the other hand, patients, especially those with lower health literacy, may use simpler sentence structures and different tenses when describing their symptoms or experiences. POS tagging can help highlight these differences. Previous studies showed that POS tagging can offer insights into the patient's health literacy and everyday language use.[21,22]

Data preparation: manual annotation of speaker types in patient–nurse encounter transcriptions

During patient–nurse encounters in home healthcare, speakers take turns talking, with each uninterrupted block of speech referred to as an utterance. The AWS-GT automated transcription output includes a speaker indicator for each utterance, designating either Speaker #1 or Speaker #2. Two members of the study team independently annotated each

transcribed recording by assigning a speaker type (patient or nurse) to each speaker (Speaker #1 or Speaker #2). Any discrepancies were resolved through discussion between the annotators. The study's sample consisted of 3494 manually annotated utterances for patients and nurses.

Development of ML classifier at the utterance level

To analyze the individual utterances and determine the speaker type (patient or nurse), we constructed ML classifiers using the following steps:

- *Step 1. Data preprocessing and feature generation*: For the TF-IDF method, the data were preprocessed by lowercasing and removing punctuation, symbols, numbers, and stop words. An example of a "stop word" could be commonly used words such as "the," "and," or "is," which often do not carry significant meaning on their own in text analysis. The same preprocessing steps were followed for Word2Vec and N-gram, except that stop words were not removed for Word2Vec as they are crucial for generating accurate word embeddings. For part of speech tagging and LIWC, only lowercasing was performed as punctuations and stop words are important for analyzing the psychology of language and the grammatical structure of sentences.
- *Step 2. Extracting and normalizing medical concepts*: The Quick UMLS tool was used to extract and normalize medical vocabulary mentioned in the patient and nurse's utterances, mapping the vocabulary to UMLS concepts and their corresponding CUIs.
- *Step 3. ML classifier*: To determine each feature set's effectiveness in distinguishing between a patient language and nurse language, we used a support vector machine (SVM) algorithm. SVM is a well-known classifier that is particularly useful when the feature vectors have high dimensionality, and the number of training samples is smaller than the number of features, as is the case in this study. We also evaluated the performance of other classifiers, including Logistic Regression Random Forest, Extra Trees, Adaptive Boosting, and XGBoost.
- *Step 4. Feature selection*: We evaluated the performance of an ML classifier using a combination of the most informative features from each feature set, selected using the Joint Mutual Information Maximization (JMIM)[23] method. JMIM selects a subset of features by maximizing the joint mutual information between the selected features and the outcome class, while minimizing redundancy among the selected features. JMIM has a high generalization ability, especially on small samples with many generated features.[23,24] See Supplementary Appendix SC for more information about the JMIM.
- *Step 5. ML classifier evaluation*: Five-fold cross-validation method with standard performance metrics area under curve-receiver operating characteristic (AUC-ROC) and F1-score (the harmonic mean of sensitivity and precision) were used to evaluate the accuracy of SVM algorithm.

To test whether removing short utterances could improve the performance of SVM algorithm, we conducted an experiment in which we removed utterances with lengths ranging from 1 to 50 tokens, one at a time. In our study, a "token" is defined as an individual unit of language data, which in this case refers to a single word, such as "okay." Our hypothesis was that short utterances (eg, those with one token such as "okay") may introduce noise to the algorithm because they do not contain enough information to distinguish between patient and nurse language. After each removal, we measured the performance of the SVM classifier.

Development of ML classifier at the encounter level

To investigate the potential benefits of analyzing longer speech samples, we evaluated the SVM classifier's performance on aggregated utterances from each patient–nurse encounter. Our hypothesis was that the classifier's performance would improve by considering longer utterances of speech data, as it may contain more informative signals for differentiating between patient and nurse language. We followed the same process as with the utterance-level classifier, including preprocessing, feature generation, feature selection, and evaluation using AUC-ROC and F1-score metrics.

RESULTS

Description of the study sample

Table 1 shows the characteristics of 23 patients who participated in the study. The sample included an equal number of men and women. About half of the patients lived alone, and 96% had no heart issues and did not need assistance with

Table 1. Clinical and demographic characteristics of study patients

Patient characteristics	Entire cohort $N = 23$
Age	59.5 (13.3)
Gender	
Female	12 (52%)
Male	11 (48%)
Race	
Black	20 (87%)
Hispanic	2 (9%)
Other	1 (4%)
Discharged in the last 14 days	
None	10 (43%)
Nursing facility	1 (4%)
Hospital discharge	12 (52%)
Living arrangement	
Home with others	12 (52%)
Alone	11 (48%)
Dyspnea	
Exertion	7 (30%)
Minor exertion and rest	2 (9%)
Never	14 (61%)
Currently reports exhaustion: Yes	11 (48%)
Decline in mental, emotional, behavioral status: Yes	2 (9%)
Cognitive deficit: Yes	4 (17%)
Confused: Yes	8 (35%)
Anxious: Yes	10 (43%)
Depressed: Yes	13 (57%)
Lack of interest: Yes	14 (61%)
The need for assistance with activities of daily living	5 (22%)
Falls risk: Yes	21 (91.0%)
Congestive heart failure: Yes	1 (4%)
Peripheral vascular disease: Yes	1 (4%)
Cerebrovascular disease: Yes	1 (4%)
Chronic obstructive pulmonary disease: Yes	3 (13%)
Renal disease: Yes	2 (9%)
Cancer: Yes	3 (13%)
Metastatic solid tumor: Yes	1 (4%)

Table 2. Recording duration and frequencies of spoken words for both patients and nurses at the encounter and utterance levels

	Average (standard deviation)	25% quartile	50% quartile	75% quartile
Total number of audio-recorded patient–nurse encounters ($N = 46$)				
Duration of audio-recorded patient–nurse encounters (in min)	13 (6)	10	11	15
Count of spoken words (tokens) in the sample	1477 (965)	822	1262	1797
Count of spoken words (tokens) by patients	626 (536)	209	501	801
Count of spoken words (tokens) by nurses	851 (625)	378	709	1149
Total number of utterances in the sample: patients ($N = 1731$), nurses ($N = 1763$)				
Count of utterances in audio-recorded patient–nurse encounters	75 (63)	38	66	87
Count of spoken words (tokens) in the sample at the utterance level	19 (35)	3	9	22
Count of spoken words (tokens) by the patient at the utterance level	16 (26)	2	8	20
Count of spoken words (tokens) by nurses at the utterance level	22 (43)	4	10	23

daily activities. Around 30% of patients showed symptoms of depression and anxiety.

Description of the sample of audio-recorded patient–nurse verbal communication

Table 2 presents descriptive statistics on the duration of audio-recorded patient–nurse encounters and spoken words by patients and nurses at both encounter and utterance levels. To compute information in this table, we employed a series of methods to analyze our audio-recorded data. We utilized Python's Natural Language Processing library to count the number of words in each text, and we used AWS-GT to determine the number of utterances by patients and nurses, with 2 team members independently assigning speaker roles. Lastly, we computed the duration of each recording using Python's wavfile.read function to read and calculate the length of each audio file.

The average duration of an encounter was 13 min, with 25% of encounters being relatively short, lasting <10 min. On average, each encounter included 75 utterances, with the median number of utterances being 66. Overall, nurses' average count of spoken words was higher than patients at the encounter level. This pattern was also observed at the level of individual utterances. As indicated by the 50% quartile, the median number of words per utterance was around 9 words, indicating that in about half of the conversations between patients and nurses, the pace of turn-changing was relatively fast.

Performance of ML classifier at the utterance level

Figure 2A and B shows the AUC-ROC and F1-score of the SVM classifier on linguistic features generated using the feature generation methods after incrementally removing utterances with lengths ranging from 1 to 50 tokens. Table 3 presents the AUC-ROC and F1-score of 4 samples of utterances: the entire sample with a minimum length of 1 token and samples with a minimum length of 5 tokens, 30 tokens, and 50 tokens. The performance of the TF-IDF, LIWC, and Word2Vec methods was similar with AUC-ROC between (70.57, 72.71) for the entire sample and AUC-ROC between (87.58, 91.85) for a sample of utterances with utterance length more than 50 tokens. The performance of Unigram and UMLS was significantly lower than TF-IDF, LIWC, and Word2Vec. We used Unigram because it performed best out of n-grams in the range of 1–10 using the SVM classifier in our initial analysis (see Supplementary Appendix SD for more detailed results). UMLS performed poorly (AUC-ROC between [58.71, 68.15]) compared to other methods, mostly because about 53% (1852/3494) of the utterances did not include any UMLS concepts. POS tagging performed poorly, with an AUC below 0.5 in almost all specified ranges of utterance lengths, indicating that the part of speech tagging is less informative than using the words themselves (Unigram) for differentiating patient and nurse language. The performance of these feature selection methods (except POS tagging) was largely consistent, indicating that utterances with a length of at least 30 tokens contain sufficient informative features to differentiate patient and nurse language. In general, the SVM classifier exhibited better performance than the other ML methods (Logistic Regression, Random Forest, Extra Trees, Adaptive Boosting, and XGBoost). As a result, we opted to exclude their results from this report.

Figure 2C and D displays the results of combining selected features from TF-IDF, LIWC, Word2Vec, Unigram, and UMLS using the JMIM method, following the iterative removal of utterances with lengths ranging from 1 to 50 tokens, one at a time. Overall, in each iteration, JMIM selected a limited number of features for each feature generation method. For example, for the first iteration, including the sample of all utterances, JMIM selected 20 TF-IDF features ($N = 6338$), 66 Word2Vec features ($N = 200$), 66 LIWC features ($N = 93$), and 25 UMLS features ($N = 673$). Overall, the classifier's performance built on TF-IDF+LIWC+Word2Vec was slightly enhanced by incorporating the selected features from the Unigram and UMLS feature sets. This indicates that Unigram and UMLS do not provide additional valuable information for differentiating patient and nurse language, which is expected given their lower performance compared to other feature generation methods.

In summary, we observed that the performance of the SVM classifier was slightly improved on the combination of selected features from TF-IDF+LIWC+Word2Vec+Unigram +UMLS compared to using only TF-IDF with 6338 features (see Table 3), suggesting that including all features from TF-IDF are useful for distinguishing between patient and nurse language. This approach may lower the classifier's generalizability, as reflected in the higher standard deviation of the classifiers.

Performance of ML classifier at the encounter level

Table 4 presents the performance of the SVM classifier on the sample of aggregated utterances for patients ($N = 46$) and nurses ($N = 46$) at the encounter level. Similar to the performance of SVM classifiers at the utterance level, SVM built on TF-IDF generated features (AUC = 97.45 ± 2.36) slightly outperformed Word2Vec and LIWC on the aggregated utterances at the encounter level. Also, like the utterance level, SVM classifiers based on POS-tagging features (AUC =

Figure 2. (A) The AUC-ROC of the SVM classifier on linguistic features generated using the feature generation methods after incrementally removing utterances with lengths ranging from 1 to 50 tokens. (B) F1-score of the SVM classifier on linguistic features generated using the feature generation methods after incrementally removing utterances with lengths ranging from 1 to 50 tokens. (C) AUC-ROC of combining selected features from TF-IDF, LIWC, Word2Vec, Unigram, and UMLS using the JMIM method, following the iterative removal of utterances with lengths ranging from 1 to 50 tokens, one at a time. (D) AUC-ROC and F1-score of combining selected features from TF-IDF, LIWC, Word2Vec, Unigram, and UMLS using the JMIM method, following the iterative removal of utterances with lengths ranging from 1 to 50 tokens, one at a time.

31.48 ± 15.27) performed notably worse compared to the other feature generation methods. Overall, the SVM classifier's standard deviation was higher as opposed to the utterance level, primarily due to the significant decrease in the sample size at the encounter level.

We also observed a slight improvement in the performance of SVM classifiers (AUC-ROC $= 99.01 \pm 1.97$) when combining selected features from TF-IDF and Word2Vec using the JMIM method. The feature selection method did not significantly impact the average of AUC-ROC, but we noticed a significant reduction in standard deviation. Overall, using the JMIM method can enhance the SVM classifier's generalizability, as shown by the decrease in standard deviation. These results imply that samples with a higher number of spoken words contain more informative features for differentiating patient and nurse language, as demonstrated by the encounter-level sample in comparison to the utterance-level sample.

Characterizing patient and nurse language

Figure 3 shows the most informative features selected by JMIM for LIWC, TF-IDF, and UMLS for patients and nurses. The LIWC features (Figure 3A) reveal that patients tend to use keywords related to "religion," "home," and "money" and have a preference for informal language as indicated by features of "nonfluencies," "netspeak," and "assent." Conversely, nurses tend to use more complex sentences indicated by "words>6 letters," "conjunction," and "total function words." Additionally, nurses are more likely to use language that reflects social and affectionate processes and interrogation (ask questions), indicated by "social processes," "affective processes," and "interrogatives."

According to the TF-IDF features selected by JMIM (as depicted in Figure 3B), patients tended to use informal language, such as "telling," "really," and "call"; while the use of health-related terms like "care," "wound," and "health" was more frequent among nurses. Figure 3C displays the semantic

Table 3. SVM classifier performance at utterance level for various utterance lengths

Feature generation method	SVM performance	Number of tokens in the utterance≥1	Number of tokens in the utterance≥5	Number of tokens in the utterance≥30	Number of tokens in the utterance≥50
	Metrics	Patient utterance ($N = 1731$) Nurse utterance ($N = 1763$)	Patient utterance ($N = 1068$) Nurse utterance ($N = 1290$)	Patient utterance ($N = 274$) Nurse utterance ($N = 358$)	Patient utterance ($N = 135$) Nurse utterance ($N = 195$)
Performance of individual feature generation method (no feature selection method was used)					
TF-IDF	**AUC-ROC**	**71.26 ± 1.51**	**79.45 ± 0.63**	**89.7 ± 1.93**	**91.85 ± 1.95**
	F1-score	**67.64 ± 1.59**	**70.88 ± 2.09**	**80.65 ± 2.04**	**81.78 ± 4.28**
LIWC	AUC-ROC	70.57 ± 1.31	76.01 ± 76.01	88.53 ± 3.45	87.67 ± 2.77
	F1-score	65.27 ± 1.16	68.92 ± 1.55	80.37 ± 5.41	78.1 ± 2.53
Word2Vec	AUC-ROC	72.71 ± 1.41	79.1 ± 1.51	88.1 ± 3.21	87.58 ± 2.6
	F1-score	67.16 ± 0.88	70.82 ± 1.9	80.44 ± 3.05	75.56 ± 3.52
Unigram	AUC-ROC	65.99 ± 2.07	70.95 ± 1.83	78.69 ± 3.79	81.88 ± 3.48
	F1-score	67.52 ± 1.49	63.86 ± 1.29	69.61 ± 4.22	65.82 ± 4.54
UMLS	AUC-ROC	58.71 ± 2.22	60.4 ± 2.09	68.29 ± 3.66	68.15 ± 4.96
	F1-score	66.68 ± 0.67	61.84 ± 1.57	60.9 ± 2.81	58.05 ± 9.00
POS-tagging	AUC-ROC	51.68 ± 7.48	46.01 ± 2.81	46.5 ± 4.25	48.43 ± 4.59
	F1-score	65.81 ± 0.84	62.17 ± 0.38	60.16 ± 0.74	58.52 ± 0.93
Performance of a combination of feature generation methods after selecting the most informative features using JMIM method					
TF-IDF (JMIM)+LIWS(JMIM)	AUC-ROC	68.98 ± 1.59	76.29 ± 2.18	87.97 ± 3.09	90.66 ± 1.22
	F1-score	68.24 ± 2.19	70.71 ± 2.03	82.67 ± 3.45	83.4 ± 4.07
TF-IDF (JMIM)+LIWS (JMIM)+Word2Vec (JMIM)	AUC-ROC	71.72 ± 1.9	79.03 ± 2.29	90.06 ± 3.48	91.85 ± 1.59
	F1-score	67.16 ± 1.53	70.19 ± 2.34	82.12 ± 3.47	80.3 ± 3.5
TF-IDF (JMIM)+LIWS (JMIM)+Word2Vec (JMIM)+Unigram (JMIM)	AUC-ROC	71.75 ± 1.87	78.93 ± 2.26	89.89 ± 3.43	91.68 ± 1.82
	F1-score	67.09 ± 1.47	70.22 ± 2.32	81.88 ± 3.94	80.36 ± 3.05
TF-IDF (JMIM)+LIWS (JMIM)+Word2Vec (JMIM)+Unigram (JMIM)+UMLS (JMIM)	**AUC-ROC**	**72.72 ± 1.95**	**79.67 ± 2.07**	**89.94 ± 2.88**	**92.61 ± 2.02**
	F1-score	**68.24 ± 1.47**	**70.71 ± 2.33**	**82.67 ± 3.94**	**83.4 ± 3.05**

Note: The feature generation method that demonstrated the highest performance, as indicated by the AUC-ROC and F1-score values, was highlighted in bold.

Table 4. Performance of SVM classifier at the encounter level[a]

Feature generation methods	AUC-ROC	F1-score
Sample: $N = 46$ patients, $N = 46$ nurses		
Performance of Individual feature generation method (no feature selection method was used)		
TF-IDF	**97.45 ± 2.36**	**93.13 ± 4.87**
POS-tagging	31.48 ± 15.27	64.17 ± 12.91
N-grams	89.45 ± 10.79	85.21 ± 6.48
LIWC	95.67 ± 4.55	82.59 ± 11.15
Word2Vec	92.51 ± 4.05	83.02 ± 8.52
UMLS	77.75 ± 5.48	77.36 ± 3.16
Performance of combination of feature generation methods after selecting the most informative features using JMIM method		
TF-IDF (JMIM)+Word2Vec (JMIM)	99.01 ± 1.97	93.66 ± 5.16
TF-IDF (JMIM)+Word2Vec (JMIM)+LIWC (JMIM)	98.54 ± 1.81	91.07 ± 5.84
TF-IDF (JMIM)+Word2Vec (JMIM)+LIWC (JMIM)+Unigram (JMIM)+UMLS (JMIM)	**99.28 ± 0.98**	**96.82 ± 4.1**

Note: The feature generation method(s) that demonstrated the highest performance, as indicated by the AUC-ROC and F1-score values, was highlighted in bold.

[a] Utterances for each identified speaker (speaker 1 and speaker 2) were aggregated at encounter level.

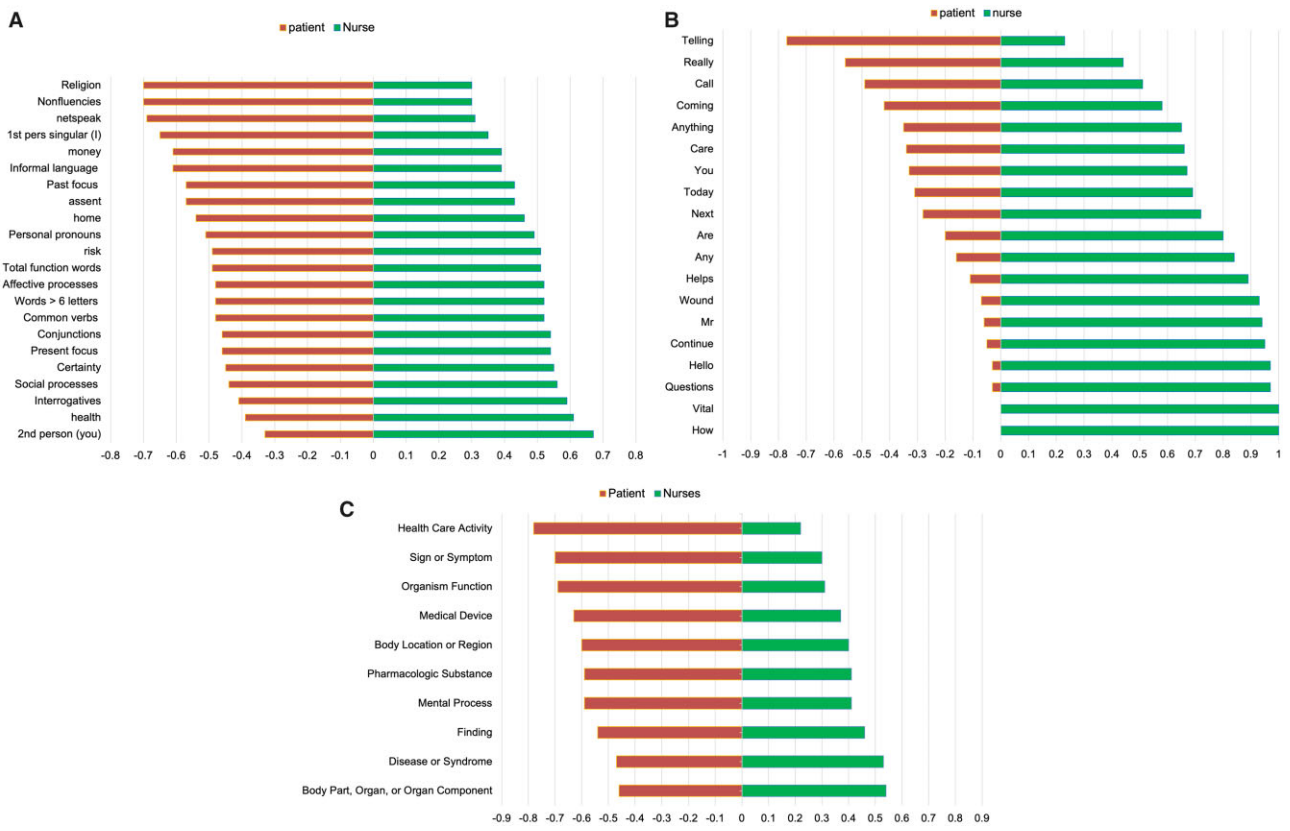

Figure 3. (A) Characterizing patient and nurse language using LIWC. The most informative features in these figures were selected using the JMIM feature selection method. (B) Characterizing patient and nurse language using TF-IDF. The most informative features in these figures were selected using the JMIM feature selection method. (C) Characterizing patient and nurse language using UMLS semantic type. The most informative features in these figures were selected using the JMIM feature selection method.

type of the most informative UMLS CUIs selected by JMIM. Patients primarily discuss their health concerns, such as "signs or symptoms," "body location or region," and "healthcare activities." Meanwhile, nurses tend to focus on clinical findings ("findings"), "diseases or syndromes", and the "mental processes" of their patients.

DISCUSSION

This study is the first to use routinely recorded patient–nurse communication to generate NLP-driven linguistic features for characterizing the language of both patients and nurses. ML classifiers were constructed on these linguistic features to differentiate between patient and nurse speech. The classifiers achieved the highest performance, with an AUC-ROC of 99.28 and an F-score of 96.82. The ability to automatically identify speaker type is crucial for downstream tasks, such as detecting risk factors and communication deficits and developing decision support tools to identify patients at risk of negative outcomes.

To generate the linguistic features, we used a combination of lexical, syntactic, and conceptual word embedding techniques (LIWC, UMLS, TF-IDF, N-gram, POS tagging, Word2Vec) at both individual and aggregated utterances during the

patient–nurse encounter. TF-IDF, LIWC, and Word2Vec achieved high and comparable results (AUC between [0.8, 097] for individual and aggregated utterances), indicating that these techniques were effective in identifying distinct psychological and linguistic features in the language used by patients and nurses.

Several conversational analysis systems, such as Roter's Interaction Analysis System[25] (RIAS) and the Coding Linguistic Elements in Clinical Interactions[25] (CLECI), are available for modeling patient–clinician verbal communication and driving features from their conversations. These sophisticated coding systems needs involvement of annotators to annotate cues within conversations and analyze utterances, focusing on elements like displays of concern, instances of laughter, and information exchange tasks like "asking for understanding" or "bidding for repetition." While manual annotation can shed light on the primary themes of a conversation, it necessitates a significant investment of time and effort due to its labor-intensive and time-consuming nature. The focus of this study was on exploration of automated features generation methods, which can provide more efficient and streamlined way for driving features for automated speaker type identification. In our upcoming research, we intend to investigate how effectively the coding systems can differentiate patient and nurse language.

There are various factors that may contribute to differences in the language used by patients and nurses during communication. Patients and nurses may come from different social, cultural, and educational backgrounds and have varying levels of health literacy, leading to differences in the terms and concepts they use, as was shown by the most informative features of TF-IDF, LIWC, and UMLS. Our results show that patients tended to use informal language more than nurses. In contrast, nurses used more sentences focusing on health-related issues and medical problems. Additionally, the power dynamic between patients and nurses can influence the language used, with patients potentially having less control over the conversation, affecting their communication of needs and concerns. This was demonstrated through selected features from LIWC and TF-IDF, which showed that nurses were more inclined than patients to ask questions during communication.

We tested performance of ML classifiers at both levels—individual utterances and aggregated utterances at an encounter level. The findings indicate that longer utterances contain more informative clues for differentiating the patient and nurse language. Specifically, the SVM classifier achieved the highest performance (AUC-ROC = 0.99) when applied to aggregated utterances of encounters. This is particularly relevant for downstream tasks aimed at analyzing patient language or modeling of patient–nurse verbal communication. At the end of audio-recorded encounters, the transcribed audio data by an ADR system (eg, AWS-GT) can be aggregated by the identified speaker (eg, speaker #1 and speaker #2) to determine the type of speaker (patient or clinician) using ML classifiers. However, this approach is impractical for real-time speech analysis systems that require real-time speaker type identification.

In addition, one particularly promising area of this study is its potential to alleviate the documentation burden that healthcare professionals, especially nurses, often face. By automatically identifying when the nurse is speaking and accurately transcribing crucial aspects of their dialogues, we could potentially automate parts of the documentation process. This automated process could lead to more precise and efficient recording of transcribing health-related concerns expressed during patient–nurse verbal communication into EHRs. Consequently, this could result in considerable time savings for approximately 4 million nurses in the United States, thereby contributing to a significant enhancement in healthcare delivery.

The distinction between the language used by patients and clinicians has been a focal point in various studies investigating patient–clinician interactions. For instance, Drew et al[26] conducted a study that analyzed patient–practitioner interactions to identify practitioners' communication patterns in patient communication. In another study, Mejdahl et al[27] analyzed patient–clinician interaction in epilepsy outpatient clinics to explore the impact of patients' self-reported data on the outcome. Also, Chang et al[28] analyzed communicative behaviors between physicians and patients in rehabilitation centers. In all these studies, the differentiation between patient and clinician language was manually annotated for modeling the interaction between the patient and the clinicians. These studies, along with several other studies published in the field of conversational analysis,[29–31] highlight the necessity of automating speaker type identification in audio-recorded patient–clinician verbal communication. Our study demonstrates the significant potential of natural language processing and ML methods in automatically differentiating between patient and nurse language in verbal communication. This automation is crucial for developing automatic data analysis pipelines for modeling patient–clinician verbal communication to improve patient outcomes.

Audio recording patient–nurse verbal communication is not currently part of the clinical workflow. We conducted a series of pilot studies to identify convenient procedure for audio-recording patient–nurse verbal communication. The findings of the studies showed that both patients and nurses were comfortable with the procedure of audio-recording and patients found it particularly useful for personal use (eg, reviewing the clinician's instruction). To protect patient and nurse confidentiality, the audio recordings were securely stored in a HIPAA-compliant environment on the AWS cloud equipped with speech recognition tools and AI capabilities (eg, GPUs for processing large bodies of text) to develop automated speech processing systems to model the patient speech and their interaction with clinicians during encounters. The involvement of healthcare stakeholders, especially clinicians and managers, plays a crucial role in the effective implementation of the speech processing system. Their active participation is essential for determining the integration of audio recording into clinical workflows and establishing the necessary processing methods for continued use in patient care management.[32,33]

Limitations

The study has several limitations. First, it is limited by a relatively small sample size. While a high-performing classifier was built using 46 audio-recorded patient–nurse verbal communication for 23 patients, the results may not be generalizable to other home healthcare settings. Additionally, the audio data were only collected from VNS Health, the largest nonprofit home healthcare organization in the United States, which could limit the generalizability of the study's findings to other healthcare settings. Furthermore, while the study

explored the performance of various feature-generation methods for differentiating patient and nurse language, the performance of other feature-generation methods and ML classifiers was not investigated.

In future studies, the authors plan to evaluate the performance of deep learning classifiers, such as BiLSTM with an attention layer, for distinguishing between patient and nurse language.

CONCLUSION

This study aimed to build an ML classifier for distinguishing patient and nurse language during home healthcare visits. This analytical approach is crucial for downstream tasks that analyze patient speech to identify patients at risk of diseases and negative outcomes. Further research should use larger samples of audio-recorded patient–nurse verbal communication is required to assess the classifier's universal applicability for speaker type identification.

FUNDING

This study was supported by K99AG076808 and R01AG081928 from National Institute on Agin; Amazon in collaboration with Columbia University Center of AI Technology; VNS Health Doyle Fund for pilot studies; and Columbia University School of Nursing Pilot Award.

AUTHOR CONTRIBUTIONS

MZ: contribution to the conception, study design, data acquisition, data analysis, drafting the manuscript. SV: data acquisition and data annotation. SS: data analysis and interpretation. AZ: data analysis and interpretation of data. KB: data acquisition, and reviewing the manuscript critically for important intellectual content. ZK: reviewing the manuscript critically for important intellectual content. MT: contributions to the conception, study design, data acquisition, and reviewing the manuscript critically for important intellectual content.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

Due to the limitations imposed by the IRB protocol of VNS Health, no data is available for public access.

REFERENCES

1. Zolnoori M, Vergez S, Kostic Z, *et al.* Audio recording patient-nurse verbal communications in home health care settings: pilot feasibility and usability study. *JMIR Hum Factors* 2022; 9 (2): e35325.
2. Song J, Zolnoori M, Scharp D, *et al.* Do nurses document all discussions of patient problems and nursing interventions in the electronic health record? A pilot study in home healthcare. *JAMIA Open* 2022; 5 (2): ooac034.
3. Barr PJ, Bonasia K, Verma K, *et al.* Audio-/videorecording clinic visits for patient's personal use in the United States: cross-sectional survey. *J Med Internet Res* 2018; 20 (9): e11308.
4. Romagnoli KM, Handler SM, Hochheiser H. Home care: more than just a visiting nurse. *BMJ Qual Saf* 2013; 22 (12): 972–74.
5. Shang J, Russell D, Dowding D, *et al.* A predictive risk model for infection-related hospitalization among home healthcare patients. *J Healthc Qual* 2020; 42 (3): 136–47.
6. Petti U, Baker S, Korhonen A. A systematic literature review of automatic Alzheimer's disease detection from speech and language. *J Am Med Inform Assoc* 2020; 27 (11): 1784–97.
7. Low DM, Bentley KH, Ghosh SS. Automated assessment of psychiatric disorders using speech: a systematic review. *Laryngoscope Investig Otolaryngol* 2020; 5 (1): 96–116.
8. Song J, Zolnoori M, Scharp D. Is auto-generated transcript of patient-nurse communication ready to use for identifying the risk for hospitalizations or emergency department visits in home health care? A natural language processing pilot study. In: AMIA Annual Symposium Proceedings; Vol. 2022; American Medical Informatics Association; 2022: 992; Washington, DC.
9. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. *Int J Hum Comput Interact* 2008; 24 (6): 574–94.
10. López Seguí F, Ander Egg Aguilar R, de Maeztu G, *et al.* Teleconsultations between patients and healthcare professionals in primary care in Catalonia: the evaluation of text classification algorithms using supervised machine learning. *Int J Environ Res Public Health* 2020; 17 (3): 1093.
11. Nagamine T, Gillette B, Kahoun J, *et al.* Data-driven identification of heart failure disease states and progression pathways using electronic health records. *Sci Rep* 2022; 12 (1): 17871.
12. Bahgat M, Wilson S, Magdy W. LIWC-UD: classifying online slang terms into LIWC categories. In: 14th ACM Web Science Conference 2022; New York: Association for Computing Machinery; 2022: 422–32.
13. Belz FF, Adair KC, Proulx J, Frankel AS, Sexton JB. The language of healthcare worker emotional exhaustion: a linguistic analysis of longitudinal survey. *Front Psychiatry* 2022; 13: 1044378.
14. Sulieman L, Gilmore D, French C, *et al.* Classifying patient portal messages using convolutional neural networks. *J Biomed Inform* 2017; 74: 59–70.
15. van Buchem MM, Neve OM, Kant IMJ. Analyzing patient experiences using natural language processing: development and validation of the artificial intelligence patient reported experience measure (AI-PREM). *BMC Med Inform Decis Mak* 2022; 22: 1–11.
16. Gogoulou E, Boman M, Abdesslem FB. Predicting treatment outcome from patient texts: the case of internet-based cognitive behavioural therapy. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume; Stroudsburg, PA: Association for Computational Linguistics; 2021: 575–80.
17. Adikari A, Nawaratne R, De Silva D, *et al.* Emotions of COVID-19: content analysis of self-reported information using artificial intelligence. *J Med Internet Res* 2021; 23 (4): e27341.
18. Boyd AD, Dunn Lopez K, Lugaresi C, *et al.* Physician nurse care: a new use of UMLS to measure professional contribution: are we talking about the same patient a new graph matching algorithm? *Int J Med Inform* 2018; 113: 63–71.
19. Lange LL. Representation of everyday clinical nursing language in UMLS and SNOMED. In: Proceedings of the AMIA Annual Fall Symposium; American Medical Informatics Association; 1996: 140; Washington, DC.
20. Lucini FR, Krewulak KD, Fiest KM, *et al.* Natural language processing to measure the frequency and mode of communication between healthcare professionals and family members of critically ill patients. *J Am Med Inform Assoc* 2021; 28 (3): 541–8.

21. Balyan R, Crossley SA, Brown W, *et al*. Using natural language processing and machine learning to classify health literacy from secure messages: the ECLIPPSE study. *PLoS One* 2019; 14 (2): e0212488.
22. Ferrario A, Luo M, Polsinelli AJ, *et al*. Predicting working memory in healthy older adults using real-life language and social context information: a machine learning approach. *JMIR Aging* 2022; 5 (1): e28333.
23. Bennasar M, Hicks Y, Setchi R. Feature selection using Joint Mutual Information Maximisation. *Expert Syst Appl* 2015; 42 (22): 8520–32.
24. Varatharajah Y, Ramanan VK, Iyer R, Vemuri P; Alzheimer's Disease Neuroimaging Initiative. Predicting short-term MCI-to-AD progression using imaging, CSF, genetic factors, cognitive resilience, and demographics. *Sci Rep* 2019; 9 (1): 1–15.
25. Stortenbeker I, Salm L, Olde Hartman T. Coding linguistic elements in clinical interactions: a step-by-step guide for analyzing communication form. *BMC Med Res Methodol* 2022; 22: 191.
26. Drew P, Chatwin J, Collins S. Conversation analysis: a method for research into interactions between patients and health-care professionals. *Health Expect* 2001; 4 (1): 58–70.
27. Mejdahl CT, Schougaard LMV, Hjollund NH, Riiskjær E, Lomborg K. Patient-reported outcome measures in the interaction between patient and clinician–a multi-perspective qualitative study. *J Patient Rep Outcomes* 2020; 4: 1–10.
28. Chang CL, Park BK, Kim SS. Conversational analysis of medical discourse in rehabilitation: a study in Korea. *J Spinal Cord Med* 2013; 36 (1): 24–30.
29. Halpin SN, Konomos M, Roulson K. Using applied conversation analysis in patient education. *Glob Qual Nurs Res* 2021; 8: 23333936211012990.
30. Pino M, Doehring A, Parry R. Practitioners' dilemmas and strategies in decision-making conversations where patients and companions take divergent positions on a healthcare measure: an observational study using conversation analysis. *Health Commun* 2021; 36 (14): 2010–21.
31. Jones A. Nurses talking to patients: exploring conversation analysis as a means of researching nurse–patient communication. *Int J Nurs Stud* 2003; 40 (6): 609–18.
32. Ball SL. Implementation of a patient-collected audio recording audit & feedback quality improvement program to prevent contextual error: stakeholder perspective. *BMC Health Serv Res* 2021; 21: 1–11.
33. Smith SM, Stelmar J, Lee G, Carroll PR, Garcia MM. Use of voice recordings in the consultation of patients seeking genital gender-affirming surgery: an opportunity for broader application throughout surgery? *J Surg Res* 2022; 5: 618–25.