Novel Computational Linguistic Measures, Dialogue System and the Development of SOPHIE: Standardized Online Patient for Healthcare Interaction Education

Mohammad Rafayet Ali, Taylan Sen, Benjamin Kane, Shagun Bose, Thomas M Carroll, Ronald Epstein, Lenhart Schubert, Ehsan Hoque, *Member, IEEE*

Abstract—In this paper, we describe the iterative participatory design of SOPHIE, an online virtual patient for feedback-based practice of sensitive patient-physician conversations, and discuss an initial qualitative evaluation of the system by professional end users. The design of SOPHIE was motivated from a computational linguistic analysis of the transcripts of 383 patient-physician conversations from an essential office visit of late stage cancer patients with their oncologists. We developed methods for the automatic detection of two behavioral paradigms, lecturing and positive language usage patterns (sentiment trajectory of conversation), that are shown to be significantly associated with patient prognosis understanding. These automated metrics associated with effective communication were incorporated into SOPHIE, and a pilot user study identified that SOPHIE was favorably reviewed by a user group of practicing physicians.

Index Terms—Virtual agent, Sentiment, Physician education, Communication skills, Physician-patient relations, Oncology, Cancer, Palliative care

1 Introduction

Effective patient-physician communication is fundamental to a patient's right to be fully informed and actively involved in health decision making. Good communication skills facilitate physicians' understanding of patients' symptoms, concerns, and treatment wishes [1]. Effective communication in the clinical setting has further been correlated with better patient health outcomes [2], [3], [4], [5]. Alternatively, a lack of effective communication has been associated with patients underestimating their disease severity [6] and overestimating their prognosis [7]. Together, these findings suggest that training the physicians on the fundamentals of how to communicate with patients, including taking turns, asking questions, showing empathy, and being positive is a very important part of medical education. In addition to in-person training, the state of the art medical education involves using trained actors to play the role of standardized patients who provide medical students feedback [8], [9], [10], [11], [12], [13], [14]. These techniques have the limitations of being expensive in terms of time and money as well as being prone to individual variation. Medical schools in the developing world may not even have the resources to provide such training [15], [16]. There exists a dire need to improve patient-physician communication training that is not only evidence-based and standardized, but is also rapidly customizable, cost-effective, and ready to be deployed online across geographical boundaries.

The 2020 pandemic saw a dramatic increase in online interactions. In-person communications were aggressively replaced with virtual interactions across a wide spectrum of domains from education to healthcare. Even the most technologically inexperienced and averse, from preschool students to senior citizens, were forced to learn and find the means to participate online. The medical education system also felt pressure to accelerate physician training. The loss of in-person interaction, may be likely to exacerbate the current deficiencies in patient-physician communication. This problem is further complicated by the ever decreasing amount of time physicians have to spend with their patients. Additionally, with increasing medical technologies to learn and ever more specialized fields of medical training, physicians have less and less time for training in patient-physician communication.

In this paper, we focus on patient-physician communication in cancer care. Communication between oncologists and patients is especially important due to the complexity and the emotions involved in discussing the patient's life expectancy. In addition, the oncologists need to explain the severity of cancer, the multiple treatment options available, and the correlates of patient involvement in complex decision-making [17], while expressing appropriate emotion and empathy. Despite decades of communication training and research, studies have shown that over 60% of late stage cancer patients do not understand their prognosis [12]. It is thus clear that identifying modifiable correlates of effective communication is essential for developing physician

- *Mohammad Rafayet Ali, Taylan Sen, Benjamin Kane, Shagun Bose, Lenhart Schubert, and Ehsan Hoque are with the Department of Computer Science, University of Rochester, New York.*
- *Thomas M Carroll and Ronald Epstein are with the Department of Family Medicine, University of Rochester Medical Center.*

Manuscript received August 2020.

communication training. We present a multi-stage research project leading to the development of an online virtual patient for physicians' communication training. We begin with the development of automatic detection methods of two behavioral paradigms, *lecturing* and positive language usage patterns (the *sentiment trajectory* of conversation). We have used a data set that consists of 382 transcripts of conversations between late stage (stage 3 or 4) cancer patients (Male=172, Female=210) and their physicians (Male=25, Female=13) and a measure of each patient's prognosis understanding [7]. Computational linguistic analysis of the conversation transcripts enabled us to develop automatic metrics for evaluating the degree of lecturing-like structure in a conversation. In addition, we identify that most physicians tend to use one of three styles of varying their sentiment over time (i.e., there are three styles of sentiment trajectory). We show that these metrics have a significant association with patients' level of prognosis understanding. We then developed an online virtual agent-based communication skills development system, SOPHIE (Standardized Online Patient for Healthcare Interaction Education), which gives users feedback on lecturing and positive language usage. SOPHIE also provides feedback on the user's speech rate and number of questions asked. SOPHIE presents herself as a late stage cancer patient and designed using the physician communication training protocol – SPIKES [18]. Fig. 1 shows a physician practicing communication skills with SOPHIE in his home. SOPHIE allows users to practice in their own private environment.

Prior interventions with physicians and patients have promoted discussions about prognosis but have not improved prognostic understanding. We undertook the first set of analyses to discover patterns of communication that had not been previously described that might affect the outcome of prognosis conversations, with the intention of applying findings from those analyses into the design of SOPHIE. Since the target outcome was to improve the prognosis understanding, the feedback was designed in such a way that has a direct association with the prognosis understanding. This is why we developed SOPHIE utilizing an existing data set. We first identify the affective components in the dataset on which we can give feedback such as sentiment and lecturing style of communication. We then validate it with statistical analysis. Finally, we implement the feedback of SOPHIE using the knowledge we have from our analysis.

Our contributions include:

- The development of an automated metric for measuring the lecturing-like structure of a patient-physician conversation transcript,
- The identification that most doctors use one of three styles of sentiment trajectory (i.e., pattern of modifying their positive language usage over the course of a patient-physician conversation).
- Demonstration that the degree of lecturing structure is significantly associated with the level of prognosis misunderstanding.
- Finding that a certain sentiment trajectory style (one which involves delivering technical information and ending with positive language) is associated with better prognosis understanding.
- Presentation of an iterative participatory design process, and an initial end-user evaluation with eight practicing physicians, of an online virtual patient (i.e., SOPHIE) for feedback-based practice of critical patient-physician conversations.

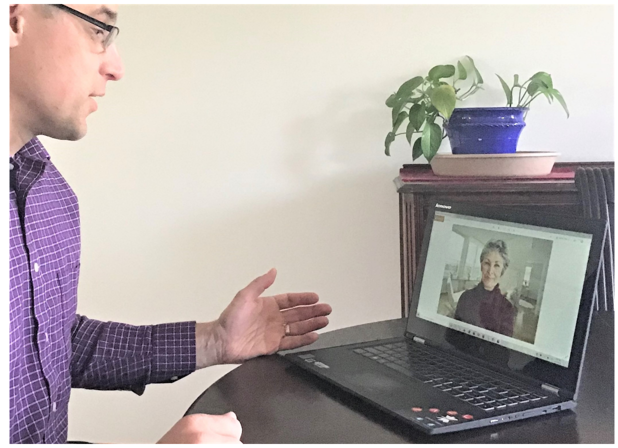


Fig. 1: A physician practicing communication skills with SOPHIE virtual patient.

In this paper, in collaboration with oncologists and medical educators from University of Rochester Medical Center (URMC), we provide early ideas on how inspiration from affective computing and online interactions could potentially transform current medical education.

2 Related Work

This work encompasses on several interconnected areas including, affect and sentiment analysis, prognosis understanding, patient-physician communication, virtual patients, and communication skills development programs [19]. Here we highlight the related research in these intersecting domains.

Affective computing and sentiment analysis has been used by researchers for health-care monitoring and disease symptoms detection. Zucco et al. [20] proposed sentiment and affective computing based architecture for depression detection. In subsequent work the authors [21] analyzed sentiment to detect the dropout of patients in tele-homecare service. The association of positive patient outcomes with physician affect has received limited experimental examination and differing conclusions have been reached in regard to the patient health outcomes. Hall et al. [22] found that the negative affects of the physicians such as showing anger and anxiety are correlated with patients' contentment. In contrast, Verheul et al. [23] in a study with 30 female patients found that warm and empathetic communication helped decrease the state of anxiety among patients. Similarly, Di Blasi, et al. [24] found in their review of 25 randomized controlled trials on affective physician communication, show inconsistency regarding emotional and cognitive care. Sen et al. [25] also found a lack of association of overall conversational positive sentiment with patient ratings of their oncologists' communication skills. Prior studies have also studied the association of physician affect on patient information recall, prognosis understanding, and better health outcomes in general. In a study of women receiving simulated breast cancer-related communications from a videotaped oncologist, van Osch et al. [26] found that affective communication improves information recall. When physicians used positive affect statements participants provided significantly more correct answers on a questionnaire testing the participants' recall of details in the diagnosis, prognosis, and treatment options. A similar study involving participant viewing of videotaped simulated oncologist communications, Shapiro et al. [27] found that participants who received communication from a worried physician as opposed to the standard, recalled significantly less medical information. The use of negative and positive affect does not seem to have a consistent effect on patient-physician communication. This is why we should focus on not only the overall affective state but also the timing of the affect. The importance of sentiment variation over time has long been recognized in storytelling [28] and more recently has been shown to be relevant in natural language analysis and face-to-face communication [29], [30], [31].

In health care communication skills training, virtual agents, and online platforms have been used in an attempt to provide an effective and reproducible experience. In the past, affective computing helped design intelligent virtual agent-based interactions for tele-health [32], [33]. Prendinger et al. [34] presented their initial work on a virtual character that analyzes physiological data in real-time, interprets emotions, and addresses users' negative affective states with empathic feedback. Peddle et al. [35] developed a virtual patient (VP) to develop and practice non-technical knowledge, skills, and attitudes among undergraduate health professionals. In a study with second and third-year nursing students, the authors found that interactions with VPs developed knowledge and skills across all categories of non-technical skills to varying degrees. Third-year students suggested that interactions with VPs helped develop knowledge and skills in a clinical setting. Angus et al. [36] developed a graphical visualization tool to model patient-physician dialogue, to identify patterns of engagement between individuals including communication accommodation, engagement, and repetition. Kleinsmith et al. [37] developed a chat-based interactive virtual patient for early-stage medical students to practice empathetic conversation. During the training, students can gather information regarding the history of the present illness, medical history, family history and social history. Additionally, during each session, the VPs delivered a statement of concern. These statements, termed empathetic opportunities, were designed to elicit an empathetic response from the user. In a study, medical students interacted with the VP and standardized patients. The responses of the participants were then rated by coders, and it turned out that responses were more empathetic with virtual patients than with standardized patients.

In this work, we have focused on improving prognosis understanding among late-stage cancer patients. To this end, we designed a virtual patient to conduct conversations with oncologists. To provide feedback to users on communication skills, we first developed algorithms to detect behavioral cues in patient-physician conversations and then engaged practicing physicians in participatory design to refine the program's feedback module.

TABLE 1: Study Data: Counts and Prognosis Survey Options

Resp. #	Description
0	100%
1	about 90%
2	about 75%
3	about 50-50
4	about 25%
5	about 10%
6	0%
X	don't know

3 MATERIALS

We performed a post-hoc analysis of a study ([38]) involving 382 visits between cancer patients ($N = 382$) and their oncologists ($N = 38$). The data includes a transcript of the conversations, in addition to both patient and physician surveys associated with each visit. The survey included questions to the physician and to the patient regarding the patient's prognosis [39]. Specifically, the prognosis question directed to the physicians was: *"What do you believe are the chances that this patient will live for 2 years or more?"*; the options provided for a response are shown in Table 1.

Patients were separately asked *"What do you believe your doctor thinks are the chances that you will live for 2 years or more?"*, with the same options for a response. When the absolute difference of the responses is greater than 1, the patient-physician prognostic conversation is defined as being misunderstood. Data in which either the physician or patient refused to answer were not used.

4 METHODS

In patient-physician communication there are several behavioral paradigms that help prognosis understanding. Among many behavioral paradigms, we have explored two patterns of behavior – *lecturing*, and the *sentiment trajectory* of conversation. We first present how we set about detecting these phenomena automatically and determining how they are associated with prognosis understanding. Then we explain our feedback design for these two behavioral patterns, applicable in conversation practice with a virtual conversational agent.

4.1 Lecturing

Lecturing generally occurs when the physician delivers a lot of information without giving the patient a chance to ask questions or to respond [1], [40]. We developed an algorithm for calculating the LECT-UR Score (Lecturing Estimation through Counting Turns with an Unbalanced-length Ratio), a measure of lecturing-related conversational structure. The algorithm compares the number of words spoken by the physician to the number of words spoken by the patient across a sliding window of a number of patient-physician turns. When the average number of words spoken by the physician exceeds a given threshold, while the average number of words spoken by the patient is below the threshold, the conversation segment is counted as a lecturing event. Fig. 2 shows the area where a lecturing event can occur in the space of the number of words spoken

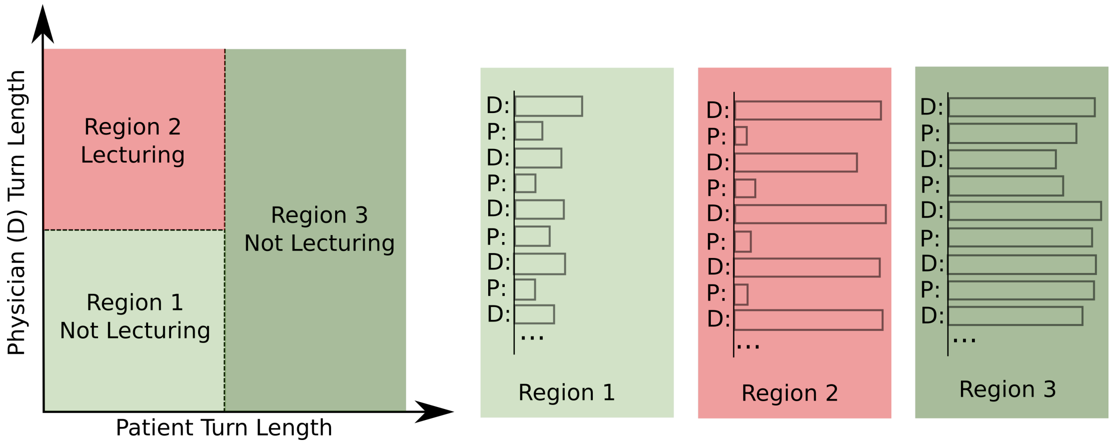


Fig. 2: Classification of Physician (D) and Patient (P) Turn Lengths over Window as Lecturing and Not-Lecturing a) Regions of Lecturing and Not-Lecturing as P length vs. D length b) example transcript turn lengths

by the physician (y-axis) and patient (x-axis). The thresholds are determined by maximizing the entropy of the outcome variables (i.e., prognosis misunderstanding).

As shown in Fig. 2 region 1, when both the physician and patient speak with brief turns it is not counted as an instance of lecturing. Similarly, in region 3, when the patient is speaking with a long turn length is not labeled as lecturing. Only when the physician's average turn length exceeds a threshold, and the patient's average turn length does not exceed the threshold, (i.e., Region 2), is the window labeled as an instance of lecturing.

This algorithm is expressed in the following equations:

$$L = \sum_{\forall_k} I\left(\sum_{i=kw_i \in D}^{k+W} \omega_i - \tau\right) \times I\left(\tau - \sum_{i=kw_i \in P}^{k+W} \omega_i\right) \quad (1)$$

$$I(x) = \{ 0 : x < 01 : x \geq 0$$

where,

L : LECT-UR Score
W : window length in number of turns
τ : turn length disparity threshold
ω : words in the transcript
D : physician utterances
P : Patient utterances

Referring to equation 1, a value for the τ parameter must be determined. As τ approaches zero, the area of region 1 in Fig. 2 will also approach zero. Alternatively, if a very large value is used for τ, region 1 will cover the entire data space. In order to be useful, the LECT-UR score should have variability. Borrowing concepts from information theory, the amount of *information* in a signal can be measured by the signal's *entropy*, where entropy is a measure of the amount of uncertainty [41]. More specifically, for a given data set X, the definition of the entropy, $H(X)$, is:

$$H(X) = \sum_{i=1}^{n} \mathrm{P}(x_i) \log_b \frac{1}{\mathrm{P}(x_i)} \quad (2)$$

where $P(x_i)$ represents the probability of observing the $i^t h$ data point. As the probability of an event x_i approaches certainty (i.e. $P(x_i) \approx 1$), the information content approaches zero. Similarly, as the probability of an event x_i approaches zero, the contribution of such events to the total information content in the data approaches zero. Thus, in order to maximize the information contained in the LECT-UR score, the scores should be well distributed (i.e. maximizing the entropy).

In order to determine the optimal τ and W, we perform a grid search. For a given τ and W we first calculate the LECT-UR score L based on equation 1. We then applied the kernel density estimation method [42] to compute the probability density function $P(x)$. From the probability density function we then obtain the entropy of L using equation 2. In Fig.3, the entropy values for different values of τ and W are shown. The maximal entropy occurs with $\tau = 103$ and $W = 20$. After calculating the LECT-UR score with the optimal parameters for each office visit transcript, we partition the data into high and low LECT-UR groups based on the median value. We then use the Z-score two-tailed population proportion test to see the difference in the percentage of prognosis misunderstanding.

To understand the effects of the confounding variables we performed a logistic regression analysis. We applied logistic regression on gender, age, disease severity, average sentiment of the conversation, study site, study arm, and the LECT-UR to predict the percentage of prognosis misunderstanding. We first normalized the independent variables and fit a logistic regression model predicting the prognosis understanding. In the section 5 we present the regression weights and the expected prognosis misunderstanding percentage for different quantiles of the LECT-UR score.

4.2 Sentiment Trajectory

The sentiment of a text segment, generally represents the emotional tone of the segment. In this work, we focus on positive language usage. We define the sentiment trajectory as the change that occurs in physician positive sentiment over the course of the conversation. Prior research suggests that the change of affective states is more important than the overall affective state. For example, Ali et al. [30] showed being positive at the beginning and at the end of a conversation is more effective than being positive overall. However, the physician sentiment trajectory over a conversation has not been well-studied in the context of patients' prognosis

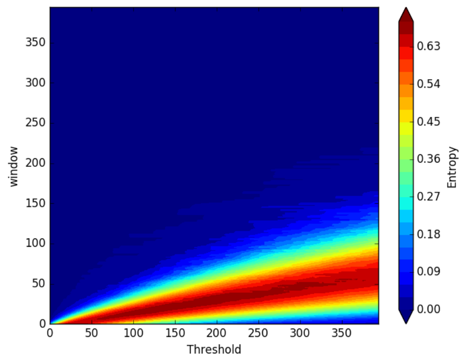


Fig. 3: Finding the Optimal Lecturing Threshold and window size based on the heatmap of Entropy.

understanding. First, we describe how we define sentiment trajectories and identify the number of sentiment trajectory styles. Later, we present the association between the trajectory styles and prognosis understanding.

We utilized the VADER ([43]) automatic text analysis tool. VADER calculates sentiment through the use of a rule-based model that employs a sentiment lexicon (dictionary of words containing an associated valence measure). The VADER positive sentiment feature is the result of a large number of human raters' understanding of positive and negative emotion associated with particular words. The VADER positive sentiment score was evaluated for each turn of the conversation. These physician and patient sentiment scores were used in two ways —- 1) average analysis, and 2) sentiment trajectory. The average sentiment scores for the physician were calculated for each transcript. The transcripts were split into two groups based on the median of the physician average sentiments (i.e. a High Sentiment group and a Low Sentiment group). The outcome measure (Prognosis Misunderstanding%) was then compared between the two groups using the z-score population proportion test.

The sentiment trajectory is defined as the time series of average physician positive sentiment over the segmented conversation. More specifically, we partitioned each conversation transcript into a number of non-overlapping segments and calculated the physician's average positive sentiment within each segment. Each conversation's sentiment trajectory is represented as a multidimensional vector, each dimension corresponding to the average sentiment within a corresponding segment of the conversation. We next determined whether distinct styles of physician sentiment trajectory existed among the conversations and investigated whether any of these physician styles demonstrated significant differences in prognosis understanding. To determine whether distinct styles of sentiment trajectory exist, we applied the k-means clustering algorithm ([44]). The number of clusters k was selected using the widely used Silhouette method ([45]), in which a grid search over a finite space of integer values for the k parameters is performed to maximizes the Silhouette score. Fig. 4 shows the steps of finding the sentiment trajectories. In order to determine whether any of the resulting sentiment trajectory clusters had statistically significant differences in the outcome measures,

TABLE 2: Average Prognosis Misunderstanding scores in High and Low LECT-UR Groups

Group	Prognosis Misunderstanding %	p-value	effect size
High LECT-UR	83.6	0.00058	0.37
Low LECT-UR	72.3		

we applied the inference test for population proportions pairwise between the groups.

We analyzed the effects of confounding variables by performing a logistic regression analysis with the sentiment trajectory styles. Specifically, we applied logistic regression on gender, age, disease severity, average sentiment of the conversation, study site, study arm, and the conversation styles to predict the outcome measures. After fitting data to logistic regression, we again can compare the relative effect that each of the input variables has on predicting whether a given data point (conversation) results in a "Don't understand prognosis" classification. After normalizing the inputs (i.e., scaling and shifting to have mean=0 and variance=1) we fit the model and hence find the model weights. We then investigate the weights of the logistic models and the prognosis misunderstanding percentage for each of the conversation styles.

In addition to the binary prognosis misunderstanding we have looked at the linear score of misunderstanding. We performed a linear regression analysis. The details are in Appendix A and B.

5 Findings

5.1 Association between LECT-UR Score and Prognosis Understanding

As shown in Table 2, the High LECT-UR Score group has a larger percentage of prognosis misunderstanding than the Low LECT-UR Score group (83.6 vs. 72.3) with a corresponding p-value of 0.00058 and an estimated Cliff's d effect size of 0.37 [46].

Fig.5 shows the logistic regression weights when predicting the prognosis misunderstanding %. The (*) marked features had a p-value less than 0.05. Among all the features the disease severity had the highest positive correlation with the prognosis misunderstanding. This shows that the more the disease has progressed the more the patients are likely to misunderstand their prognosis. Although the LECT-UR score has small positive weight than age and severity, it was significant. This model thus suggests that the higher a conversation's LECT-UR score, the more likely a patient will misunderstand their prognosis. Fig. 6 shows the prognosis misunderstanding percentage for the different quantile values of the LECT-UR score. To understand this let's select a quantile value of LECT-UR. For example, the oncologists who are above the 80th percentile based on their LECT-UR score had more than 54% of patients fail to understand their prognosis.

5.2 Association between Sentiment and Prognosis Understanding

The difference in the prognosis misunderstanding % between the high and low average positive sentiment groups did not show a significant difference. Out of the analyzed

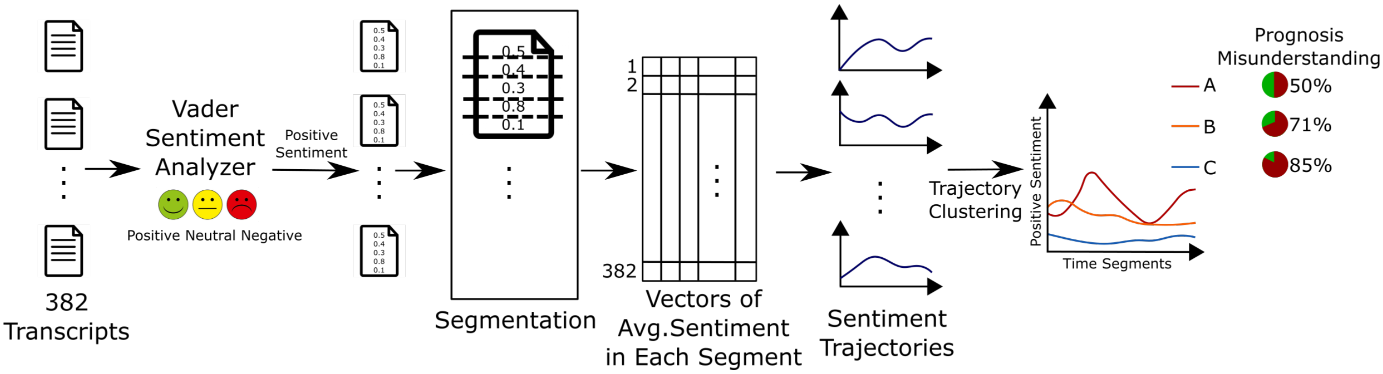


Fig. 4: Sentiment Trajectory Analysis Steps. (From left to right) 382 physician-patient conversations transcribed, VADER tool is used to calculate the positive sentiment of each physician turn, the full conversation is segmented into equal regions, the physician sentiment in each region is averaged into a trajectory, 382 trajectories are clustered in three clusters (i.e. trajectory styles) that best fit the data, statistical comparisons was done of the patient prognosis understanding in each cluster.

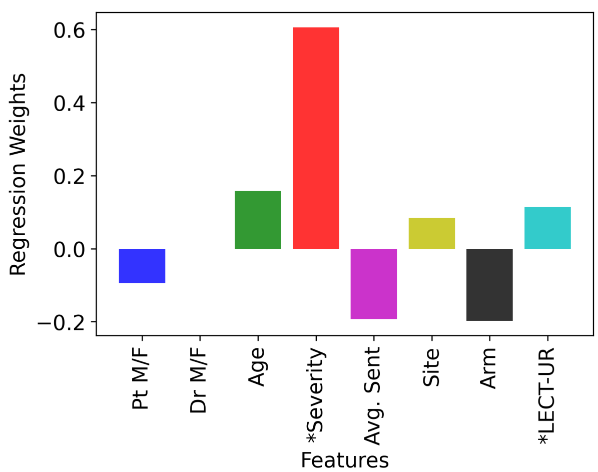


Fig. 5: Logit model weights for predicting whether the prognosis is misunderstood.

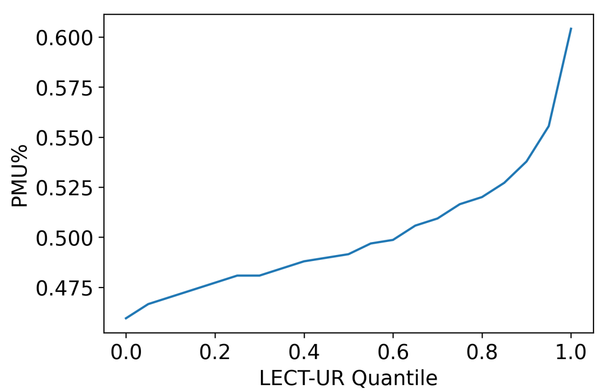


Fig. 6: Prognosis misunderstanding percentage for the different quantile values of the LECT-UR score.

number of clusters (k = 2 through 10), the number of trajectory clusters that had the highest Silhouette score was k=3. In addition, the BIC (Bayesian information criterion [47]) analysis also identified that the optimal value for k is 3. Shown in Fig. 7 are the resulting three trajectory clusters: cluster A (red, n = 15); cluster B (orange, n = 58), and cluster C (blue, n = 191). Cluster A (Dynamic) is characterized by

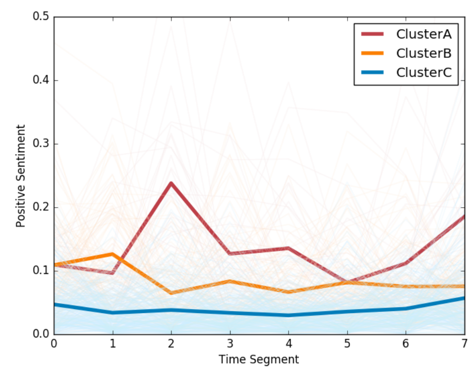


Fig. 7: Resulting Sentiment Trajectory Clusters for best K=3.

TABLE 3: Prognosis misunderstanding % in three sentiment trajectory clusters.

Trajectory cluster (size)			Pairwise statistical comparison		
A (15)	B (58)	C (191)	P_{AB}	P_{BC}	P_{AC}
46.1	52.6	67.9	0.34	0.04	0.06

a more dynamic shape, with increases in positive sentiment at 25% into the conversation (segment 2), as well as at the end of the conversation (segment 7). By contrast, Clusters B (Medium) and C (low) have a mostly flat sentiment level throughout the conversation with approximate average VADER sentiment levels of 0.1 and 0.05 respectively.

Shown in Table 3 are the outcome measures for each of the three trajectory cluster groups along with pairwise population percentage inference test p-values. As shown by the Prognosis Misunderstanding %, the low cluster (cluster C) showed the highest percentage with 67.9 % of the patients having a discordant understanding of their prognosis. The p-values for comparing the percentages between low and dynamic and low with medium clusters were 0.04 and 0.06 respectively.

Fig. 8 shows the logistic regression weights when predicting the Prognosis Misunderstanding %. The variables marked with a (*) had $p < 0.05$. In Fig. 8 the highest positive value was assigned to severity. Although this is not significant, it indicates that patients with a higher severity level of the disease are more likely to misunderstand their

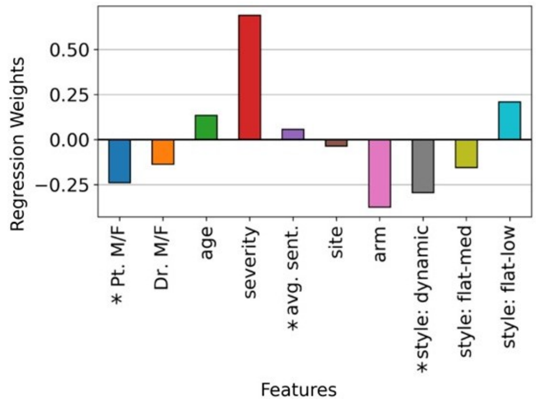


Fig. 8: Logistic regression model weights for predicting whether the prognosis is misunderstood.

TABLE 4: Confounder-Adjusted Logit Model

Trajectory Cluster	PMU %	β	p-val
A (Dynamic)	49.76	-0.294	0.033
B (Medium)	70.74	-0.155	-
C (Low)	84.85	0.209	-

prognosis. Patient gender had negative weight which indicates that female patients were more likely to misunderstand their prognosis. This is also true for physician gender but not significant. Average physician sentiment has low positive weight but significant. This indicates being positive overall is associated with misunderstanding prognosis. This finding is similar to what we have seen in the past where being positive had a negative correlation with how the patients rate their physicians [25]. Among all the clusters, the dynamic cluster has the lowest value. This indicates that when physicians used the dynamic sentiment pattern throughout the conversation, the patients were less likely to misunderstand their prognosis.

Unlike linear regression, with logistic regression there is no simple way to adjust the output (i.e., "correct" the output) for the effect of confounding variables of each data point. This is because the actual outputs are binary, whereas the model output is a probability. Instead, we can compare the predicted model Prognosis Misunderstanding % for each cluster. When all confounding variables are set to have the average value over our data set, we compute the models' predicted Prognosis Misunderstanding % for each cluster (see table 4). The Wald test p-value of the logistic regression is also shown in table 4 (marked * in Fig. 8 when $p < 0.05$). This again indicates that with confounding adjustment, the dynamic style cluster has the lowest Prognosis Misunderstanding % among all clusters.

6 Design of SOPHIE

Our aim is to develop a virtual standardized patient for practicing communication skills. In medical education, students practice with a standardized patient – an actor/actress pretends to have a medical condition. Students interact with the standardized patients and later they receive feedback on their interaction. Our goal is to allow the medical students to practice their communication skills with a virtual agent, allowing multiple repetitions in each student's own environment, which would be difficult to achieve with actual standardized patients.

6.1 Scenario

We have developed a prototype of the SOPHIE program, which allows individuals to have a conversation with a virtual agent concerning prognosis and treatment options. SOPHIE presents herself as a late-stage cancer patient. We have used the SPIKES protocol to guide the conversation [18]. The SPIKES protocol was developed to train physicians deliver bad news. The SPIKES protocol has six steps – 1) setting up the interview, 2) assessing patients' perception, 3) obtaining patients' invitation, 4) giving knowledge and information to the patient, 5) addressing the patient's emotion with empathetic responses, and 6) strategy and summary. With SOPHIE, at the beginning of the conversation (SPIKES step 1) SOPHIE introduces herself and mentions that she has lung cancer. Then SOPHIE raises the topic of her sleep pattern at night and asks if she needs to change her pain medication, allowing the physician to assess her perception (SPIKES step 2). She states that her current pain medication, *Lortab*, is not working anymore. SOPHIE then turns attention to her test results, giving the physician a chance to obtain SOPHIE's invitation to talk about more difficult topics (SPIKES step 3), before asking more specifically about her prognosis if the physician did not already address it, thus allowing the physician to provide information to the patient (SPIKES step 4). SOPHIE then asks about what her options are, allowing the physician to give empathetic responses (SPIKES step 5). Finally, she follows up by discussing whether chemotherapy remains an option, whether she should focus on comfort care, what the side effects of chemotherapy are, and how to break the news to her family, allowing for the physician to provide strategy & summary information (SPIKES step 6).

This type of discussion promotes understanding of the patient, gathering information from the patient, discussing critical information, and responding with empathy.

6.2 Dialogue System

The SOPHIE program is built on top of Eta, a general purpose dialogue management framework representing a further development of the LISSA system [48], [49], [50]. Each dialogue agent built within the Eta framework defines a flexible, modifiable dialogue schema, which specifies a sequence of intended and expected interactions with the user. The body of a dialogue schema consists of a sequence of formal assertions that express either actions intended by the agent, or inputs expected from the user. These events are dynamically instantiated into a dialogue plan over the course of the conversations. As the conversation proceeds, this plan is subject to modification based on the interpretation of each user input in the context of the agent's previous utterance. For instance, if a planned query to the user has already been answered by some part of a user's previous input, the dialogue manager can skip that query. The dialogue manager can also expand steps into subplans by instantiating sub-schemas in the case of more complex interactions.

The dialogue management framework captures the users' response from the audio stream using an automatic speech recognition technique. Both interpretation of the user's replies and generation of the agent's responses are handled using transduction to and from simple context-independent English sentences called *gist-clauses*. The dialogue manager interprets each user's input in the context of SOPHIE's previous question, using this context to select topically relevant pattern transduction hierarchies to use to interpret the user's response. The context of the previous question is useful for resolving anaphora, ellipsis, and other pragmatic phenomena. The rules in the selected hierarchies are then used to derive one or more gist-clauses from the user's input, containing explicit representations of both statements and questions detected in the user's utterance. For example, if SOPHIE asks "Do you think I should take stronger pain medication?" and the user answers "Yes.", the gist-clause extracted would be "I think you should take stronger pain medication." If the user replies "Can you tell me more about how you're feeling?", the gist-clause extracted would be "Can you tell me more about your pain?", having interpreted the question as an inquiry about SOPHIE's pain in the particular context of her question.

As mentioned, the gist-clauses are derived using hierarchical pattern transduction methods. Each transduction hierarchy specifies patterns at its nodes that are to be matched to input, with terminal nodes providing result templates to be used according to various directives (e.g. storing as a gist-clause, outputting the result, specifying a sub-schema to be instantiated, etc.). The pattern templates look for particular words or word features, including "wildcards" matching any word sequence of some length. In the case of a failure to match, the system first tries siblings of the pattern before backtracking to the previous level; the efficiency of the hierarchical pattern matching approach lies in the fact that higher levels can segment utterances into meaningful parts, thus reducing the amount of backtracking necessary to interpret the user's input.

The agent's responses to the user are likewise determined using hierarchical pattern transduction. In the case where the gist-clause from the user's utterance is a simple statement, the agent selects a reaction to the gist-clause and either instantiates a sub-schema to ask a follow-up question, or proceeds to the next topic in the main schema. If the gist-clause from the user's utterance is a question, the agent instantiates a sub-schema to select a reply to the user's question and await either a follow-up question or closure from the user. The system also has the potential to form replies to multiple gist clauses from a single user turn, for instance reacting to the user's statement before responding to a final question by the user.

The transduction hierarchies themselves were designed in a modular fashion, with a "backbone" of transduction trees detecting general questions that SOPHIE might expect a user to ask, with additional transduction trees for detecting questions and replies specific to the current topic of the conversation. In the case of a failure to match a specific response, the dialogue manager can fall back to the current general question, and if this fallback fails, simply output a generic default response.

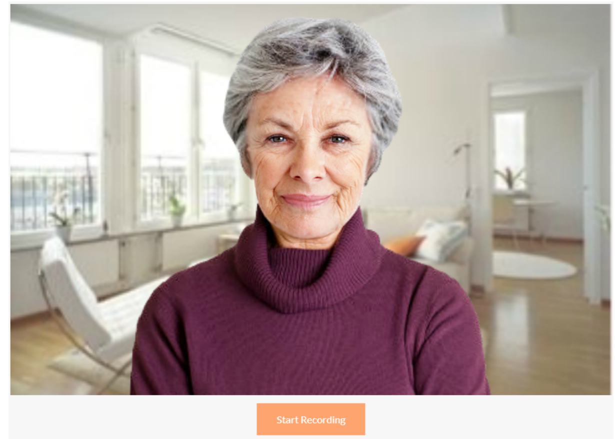


Fig. 9: SOPHIE virtual agent

6.3 Interface

The SOPHIE system features a virtual agent (shown in fig. 9). At the beginning, users start the conversation by pressing the "start recording" button. Users can then proceed to conversing with SOPHIE, and when the conversation is over the program takes the user to the feedback page. The feedback interface is shown in Fig. 10. On the left side of the feedback interface we show the transcript. The red marked speech is considered too long for the patients. On the right side of the interface we show the speech rate of the user, number of questions user asked, turn taking, and the sentiment trajectory. Past literature has established that conversational speech rate is important in enabling patients to understand their prognosis. Also, asking questions of the patient is important for ensuring that the patient understands what is being said ([1]). The turn taking annotation shows the length of each turn by SOPHIE and the user. The example was chosen to illustrate the lecturing style of conversation. The feedback shows the sentiment trajectory of both SOPHIE and the user. Additionally, the feedback shows a suggested sentiment trajectory for the user. The feedback page displays explanations of individual items when users hover their mouse on them.

6.4 Pilot Study

We conducted a pilot study with nine practicing clinicians (fellows, residents, and nurse practitioners) from the University of Rochester Medical Center. Participation was voluntary and we did not offer any payment for their participation. Additionally, we made it clear that not participating or stopping the study in the middle will have no consequences. Among these participants, one participant dropped out due to the bad audio quality of her computer. All participants were white and aged between 30 and 55. Three participants were female and all were native English speakers. Our goal was to gather more information about their experience with SOPHIE, any limitations, and how we could improve the system. The study was performed with one participant at a time on the Zoom communication platform. Each day, we asked the invited participant to have a conversation with SOPHIE and to look at the feedback.

After conversing and receiving the feedback, the participants were interviewed. We have performed a thematic analysis on the interview transcripts; our findings are below.

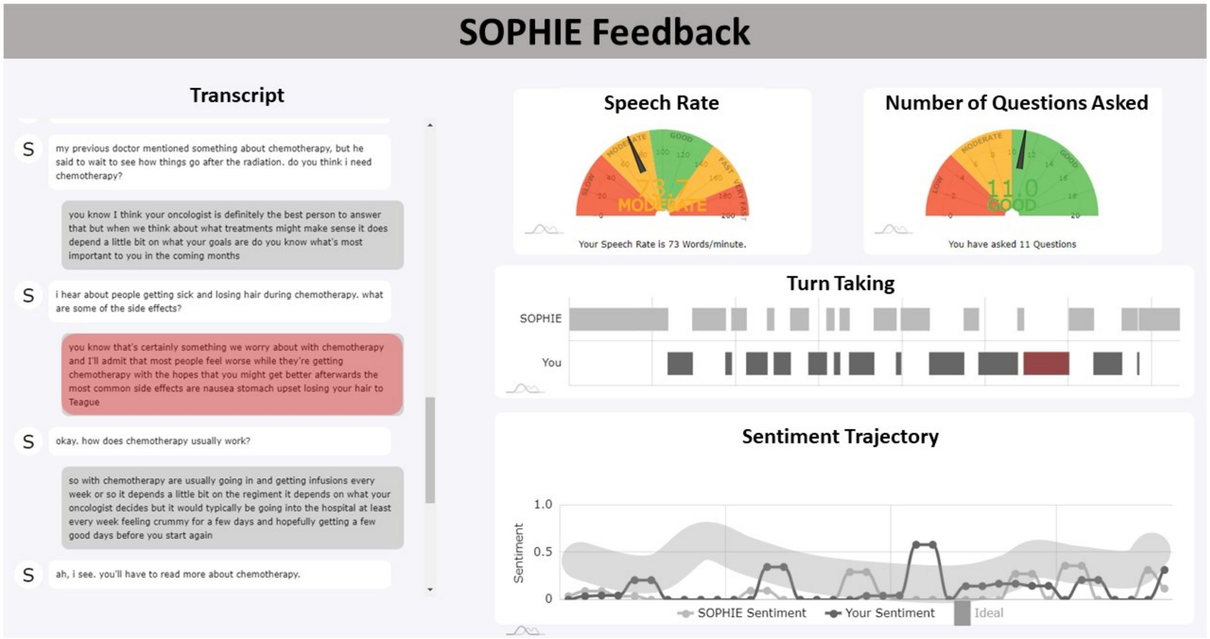


Fig. 10: Feedback interface of SOPHIE. On the left side the conversation transcript is shown. On the right (from top to bottom) speech rate, number of questions are shown. Turn taking shows the turn length and at the bottom the sentiment trajectory of both physician and SOPHIE are shown with the ideal/suggested sentiment trajectory.

6.4.1 Medical History

All the participants mentioned that a brief medical history should be presented before starting the conversation with SOPHIE. One participant said,

> "I think some kind of medical record would be extremely helpful. I thought I don't have any information to say to her."

The participants mentioned that in a regular standardized patient visit, they are given a medical record before they go into the room. They suggested the same scenario should be replicated for SOPHIE. They also mentioned that the way SOPHIE initiated presentation of her symptoms was unusual. In most cases, patients do not actively start the conversation. Rather, the physician looks at the patient's medical record and then starts asking about any new symptoms. In the future, we expect to modify the dialogues so that SOPHIE appears more passive and lets the users ask questions, though completely user-driven conversation remains beyond the state of the art.

6.4.2 Topics of Conversation

Participants (four out of eight) mentioned that SOPHIE jumped between topics and did not allow full coverage of a given topic. For example, SOPHIE begins talking about her pain medication, but the participants often asked questions about the current dosage and about other pain medication she had taken in the past. Since SOPHIE's limited dialogue repertoire falls short of covering those questions, she starts talking about her current medication (i.e., Lortab). One participant said,

> "When she mentioned pain and I was trying to find more about the pain in order to help her with her question. But the answers that I gave her to her questions did not really fit and she just jumped to the next topic so I jumped with it but that was a little bit jarring to me."

Although SOPHIE changed the conversation topics abruptly, the questions she asked were found to be realistic. Five participants felt that SOPHIE was able to express her concerns and make them feel the seriousness of the situation.

6.4.3 Feedback on Speech Rate

Participants (seven out of eight) mentioned that the speech rate feedback was easy to understand and very useful. One participant said,

> "I know I tend to speak very fast, receiving feedback on my speech rate is going to be very useful."

Another participant mentioned that in normal practice there is no way of measuring the speech-rate. However, with SOPHIE we could provide the information about how fast the physicians are speaking, which is useful.

> "I think the feedback (speech-rate) was useful. I never had someone measure my speech rate before. Sometimes I try to be cognizant of speaking a little bit slower with the patients but it was nice to actually get some feedback like you are doing okay."

However, the participants also noted that SOPHIE's speech rate was constant, making it difficult for them to adjust their speech rate depending on whether they are discussing serious issues or a casual topic. In the future, we plan to adjust SOPHIE's speech rate based on the seriousness of the topic being discussed.

6.4.4 Number of Questions Asked

Seven out of eight participants expressed that feedback on the number of questions asked was very useful. One participant said,

> "It was helpful to get the information about how many questions you have asked, because I think a lot of the times we walk away from the conversation thinking that we really invited the patients into the talk, when maybe we didn't and we did a lot of lecturing."

In addition to the number of questions asked, participants suggested that it should be highlighted what type of questions were asked, for example, how many history-taking questions were asked and how many emotional questions were asked.

6.4.5 Explanation of Sentiment

Seven participants mentioned that they did not understand the meaning of the sentiment values. They also said that the sentiment feedback is hard to interpret and they often confused it with empathy. Four participants wanted to see an example sentence of positive and negative sentiment. The participants also mentioned that changing or adjusting sentiment while engaged in the conversation may add to the cognitive load. They suggested that instead of asking the user to be positive at certain moments we should just stress the importance of dynamically adjusting sentiment.

6.4.6 Additional Feedback

The participants also asked for additional feedback that they found useful in practice. Two participants said that there are a few expressions of empathy in the dialogue and they should be highlighted in transcripts so that users could look back and understand how they responded. One participant said that the turn-taking feedback is useful, however, it does not show the total amount of time a person was speaking. The participant said,

> "I tend to speak a lot, but I don't want to make the patients feel that I am not listening. I want to know that I am giving a chance to ask questions."

He suggested adding a bar chart to the feedback page that indicates the total speech length and time.

Three participants suggested giving feedback on nonverbal behaviors, such as eye contact. One of them said,

> "One of the things I think is important, and I have seen it in other clinicians, is eye contact. I think it's super important when we are giving bad news or having difficult conversations."

7 Discussion, Limitations, and Future Work

We have described two novel contributions to communication research; empiric associations between automatically detected behaviors and patient prognostic understanding, and the development of SOPHIE, an automated system for teaching and evaluating patient-physician communication. In addition to the communication training program, the automatic detection of behaviors can be applied in prerecorded standardized patient interactions to evaluate communication skills. We acknowledge some limitations in the development of SOPHIE and the use of such a system as a basis for feedback.

Our finding of associations between trajectory styles and lecturing tendencies with prognosis understanding measures may not be causal. The lecturing analysis was motivated from prior research that suggested that when a physician tends towards lecturing, this results in the patient not retaining as much of the information presented [1]. An alternative explanation could be that when physicians sense that patients do not understand, physicians are motivated to speak more, explain in greater detail, leading to a more lecturing-like structured conversation. Additionally, apparently passive patients may just lack understanding, which can result in poor engagement, and this may result in conversations with a high LECT-UR score. Additionally, the extent to which the LECT-UR score correlates with human annotated instances of "ground truth" lecturing should be investigated. However, it should be noted that despite any difference between the LECT-UR lecturing-like structure measure and human-labelled ground truth instances of lecturing, our results establish that the LECT-UR score serves as a useful metric in its association with patient prognosis misunderstanding.

In explaining the association of higher prognosis understanding with the dynamic sentiment trajectory style, we surmise that being dynamic keeps the patient more engaged, and that ending on a positive note keeps the patient less depressed and more likely to remember the information just presented. However, again, an alternative *anticausal* explanation could be that patients' lack of prognosis understanding, and their physician's perception of this, motivates the physician to speak in a calmer, less dynamic way (e.g., sentiment trajectory styles B or C).

We found that three clusters (k=3) represent the data best according to the Silhouette score. While the Silhouete score is trusted method for finding the optimal number of clusters for k¿=2, the Silhouette method is unable to evaluate when the data is better represented by a single group (i.e. k=1). In order to determine that our finding of three clusters is not an artifact of the techniques used, we used the Bayesian Information Criterion (BIC) [47] as an additional method of validating the optimal k. We used a related clustering technique, the Gaussian Mixture Model (GMM), together with BIC to determine whether the data is better represented by a single cluster. The GMM-BIC analysis also found that the optimal k=3, and importantly showed that k=1 was inferior. While it is possible to use a Gaussian Mixture Model as our primary clustering method instead of k-means, there are multiple reasons why k-means is more appropriate such as skewed distribution and fixed intervals. In addition to considering the number of clusters we have experimented with a range of values for the number of segments. A large segment is not suitable for trajectory analysis since it may contain the bulk of the conversation, and a small segment size is also not suitable since it may not contain representative turns from both physicians and patients. Thus we experimented with five, eight, ten, and fifteen as our number of segments. In this paper, we have shown results for the choice of eight segments, omitting the others as they produced similar results.

Some limitations exist with regard to the bigger picture of SOPHIE-like virtual agents. Past research suggests that while conversing with a virtual agent or AI-driven conversational agent, humans tend to use shorter turns [51]. This could be a limitation of using SOPHIE to train users to avoid lecturing, since users might use shorter turns regardless of feedback. Our LECT-UR scoring method utilizes a window of consecutive turns that also includes the virtual agent's turn. This allows the lecturing feedback to dynamically adapt to the conversation states and to the user's behavior. We think that this can help circumvent the limitation posed by using feedback trained on human-human conversation with a computerized dialogue system, though addressing this concern through a randomized study remains part of our future work.

The current dialogue manager itself also has some limitations, which we aim to address in the future. First, the output of the currently used automatic speech recognition (ASR) software[1] does not include punctuation. This limits the agent's ability to interpret the user correctly. Secondly, as discussed in Section 6.4.2, the dialogue manager tended to abruptly jump to the next topic in the main dialogue schema in cases where it failed to understand the user's input. This will be addressed by further expanding the interpretation patterns on the basis of the dialogues we observed in this study, as well as by allowing for more robust default strategies, such as staying on topic when it appears that the agent misinterpreted the user's input or when the user's input appears irrelevant to the agent's question.

While our study focused on high patient prognosis understanding as a positive goal, it should be acknowledged that patients sometimes don't want to know specifically how much time they have left [52]. In designing a communication training program we should incorporate options as to how much information the physician should deliver. Another limitation of this work may be that our findings are limited to patient-physician relationships involving diseases and conditions as serious and sensitive as advanced cancer care and end-of-life communication.

8 Conclusion

In summary, in this paper, we provide early results of our multi-stage research examining patient-physician conversations, identification of effective traits (not lecturing, asking questions, delivering news on a positive note), development of an automated way of evaluating these traits, and the design of a real-time online standardized patient-physician communication training system where an avatar plays the role of a standardized patient. We structured our exploration in the context of conversations between final stage cancer patients (i.e., terminal patients) and oncologists.

In [53] McGreevey et al. presented a few considerations for implementing AI-driven conversational agents in health care. One important consideration is the level of risk associated with a conversational agent when it makes a mistake. SOPHIE is a low-risk program, and can be augmented with traditional training modules. In addition to being low-risk, SOPHIE allows access by individuals beyond geographical boundaries. This will promote the fair use of the program by reaching the lower socio-economic areas. Indeed, we believe that successful SOPHIE-like systems could have broader global impacts. Two-thirds of cancer deaths happen in low- and mid-income countries such as those in Latin America and sub-Saharan Africa [15], [16]. However, most of the seriously ill patients don't have access to quality palliative care (PC) because of inadequate PC training programs. Current medical training in the countries of these regions focuses on treating diseases. Comfort care in chronic life-threatening diseases such as cancer is still in its infancy. In Africa, some countries–Kenya, Uganda and Botswana–have initiated post-graduate training programs for palliative care [54], [55]; only South Africa has a well-established post-graduate and research program on palliative care [56]. We are hopeful that online programs such as SOPHIE can provide a basis for helping these communities develop training programs for PC physicians.

1. https://www.nuance.com/index.html

9 Acknowledgments

This research was partially funded by the following grants NSF NRT-DESE Award 1449828 and in part by grant W911NF-19-1-0029 with the US Defense Advanced Research Projects Agency and the Army Research Office.

References

[1] A. L. Back, R. M. Arnold, W. F. Baile, J. A. Tulsky, and K. Fryer-Edwards, "Approaching difficult communication tasks in oncology 1," *CA: a cancer journal for clinicians*, vol. 55, no. 3, pp. 164–177, 2005.

[2] S. H. Kaplan, S. Greenfield, and J. E. Ware Jr, "Assessing the effects of physician-patient interactions on the outcomes of chronic disease," *Medical care*, pp. S110–S127, 1989.

[3] E. W. Nawar, R. W. Niska, and J. Xu, "National hospital ambulatory medical care survey: 2005 emergency department summary," 2007.

[4] J. Oates, W. W. Weston, and J. Jordan, "The impact of patient-centered care on outcomes," *Fam Pract*, vol. 49, no. 9, pp. 796–804, 2000.

[5] R. S. Beck, R. Daughtridge, and P. D. Sloane, "Physician-patient communication in the primary care office: a systematic review." *The Journal of the American Board of Family Practice*, vol. 15, no. 1, pp. 25–38, 2002.

[6] J. C. Weeks, P. J. Catalano, A. Cronin, M. D. Finkelman, J. W. Mack, N. L. Keating, and D. Schrag, "Patients' expectations about effects of chemotherapy for advanced cancer," *New England Journal of Medicine*, vol. 367, no. 17, pp. 1616–1625, 2012.

[7] R. Gramling, K. Fiscella, G. Xing, M. Hoerger, P. Duberstein, S. Plumb, S. Mohile, J. J. Fenton, D. J. Tancredi, R. L. Kravitz *et al.*, "Determinants of patient-oncologist prognostic discordance in advanced cancer," *JAMA oncology*, vol. 2, no. 11, pp. 1421–1426, 2016.

[8] A. Eid, M. Petty, L. Hutchins, and R. Thompson, ""breaking bad news": standardized patient intervention improves communication skills for hematology-oncology fellows and advanced practice nurses," *Journal of Cancer Education*, vol. 24, no. 2, pp. 154–159, 2009.

[9] A. K. Sachdeva, P. J. Wolfson, P. G. Blair, D. R. Gillum, E. J. Gracely, and M. Friedman, "Impact of a standardized patient intervention to teach breast and abdominal examination skills to third-year medical students at two institutions," *The American journal of surgery*, vol. 173, no. 4, pp. 320–325, 1997.

[10] M. Ju, A. T. Berman, and N. Vapiwala, "Standardized patient training programs: an efficient solution to the call for quality improvement in oncologist communication skills," *Journal of Cancer Education*, vol. 30, no. 3, pp. 466–470, 2015.

[11] R. M. Epstein, J. C. Levenkron, L. Frarey, J. Thompson, K. Anderson, and P. Franks, "Improving physicians' hiv risk-assessment skills using announced and unannounced standardized patients," *Journal of general internal medicine*, vol. 16, no. 3, pp. 176–180, 2001.

[12] R. M. Epstein, P. R. Duberstein, J. J. Fenton, K. Fiscella, M. Hoerger, D. J. Tancredi, G. Xing, R. Gramling, S. Mohile, P. Franks *et al.*, "Effect of a patient-centered communication intervention on oncologist-patient communication, quality of life, and health care utilization in advanced cancer: the voice randomized clinical trial," *JAMA oncology*, vol. 3, no. 1, pp. 92–100, 2017.

[13] J. J. Fenton, R. L. Kravitz, A. Jerant, D. A. Paterniti, H. Bang, D. Williams, R. M. Epstein, and P. Franks, "Promoting patient-centered counseling to reduce use of low-value diagnostic tests: a randomized clinical trial," *JAMA internal medicine*, vol. 176, no. 2, pp. 191–197, 2016.

[14] G. M. Reger, A. M. Norr, P. Sylvers, J. Peltan, D. Fischer, M. Trimmer, S. Porter, P. Gant, J. S. Baer *et al.*, "Virtual standardized patients vs academic training for learning motivational interviewing skills in the us department of veterans affairs and the us military: A randomized trial," *JAMA network open*, vol. 3, no. 10, pp. e2 017 348–e2 017 348, 2020.

[15] M. Stoltenberg, D. Spence, B.-R. Daubman, N. Greaves, R. Edwards, B. Bromfield, P. E. Perez-Cruz, E. L. Krakauer, M. A. Argentieri, and A. E. Shields, "The central role of provider training in implementing resource-stratified guidelines for palliative care in low-income and middle-income countries: Lessons from the jamaica cancer care and research institute in the caribbean and universidad católica in latin america," *Cancer*, vol. 126, pp. 2448–2457, 2020.

[16] W. Y. van der Plas, S. Benjamens, and S. Kruijff, "The increased need for palliative cancer care in sub-saharan africa," *European Journal of Surgical Oncology*, 2020.

[17] L. Millard, C. Hallett, and K. Luker, "Nurse–patient interaction and decision-making in care: patient involvement in community nursing," *Journal of Advanced Nursing*, vol. 55, no. 2, pp. 142–150, 2006.

[18] W. F. Baile, R. Buckman, R. Lenzi, G. Glober, E. A. Beale, and A. P. Kudelka, "Spikes—a six-step protocol for delivering bad news: application to the patient with cancer," *The oncologist*, vol. 5, no. 4, pp. 302–311, 2000.

[19] M. E. Hoque and R. W. Picard, "Rich nonverbal sensing technology for automated social skills training," *Computer*, vol. 47, no. 4, pp. 28–35, 2014.

[20] C. Zucco, B. Calabrese, and M. Cannataro, "Sentiment analysis and affective computing for depression monitoring," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2017, pp. 1988–1995.

[21] C. Zucco, S. Bella, C. Paglia, P. Tabarini, and M. Cannataro, "Predicting abandonment in telehomecare programs using sentiment analysis: a system proposal," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2018, pp. 1734–1739.

[22] J. A. Hall, D. L. Roter, and C. S. Rand, "Communication of affect between patient and physician," *Journal of Health and Social Behavior*, pp. 18–30, 1981.

[23] W. Verheul, A. Sanders, and J. Bensing, "The effects of physicians' affect-oriented communication style and raising expectations on analogue patients' anxiety, affect and expectancies," *Patient education and counseling*, vol. 80, no. 3, pp. 300–306, 2010.

[24] Z. Di Blasi, E. Harkness, E. Ernst, A. Georgiou, and J. Kleijnen, "Influence of context effects on health outcomes: a systematic review," *The Lancet*, vol. 357, no. 9258, pp. 757–762, 2001.

[25] T. Sen, M. R. Ali, M. E. Hoque, R. Epstein, and P. Duberstein, "Modeling doctor-patient communication with affective text analysis," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 170–177.

[26] M. van Osch, M. Sep, L. M. van Vliet, S. van Dulmen, and J. M. Bensing, "Reducing patients' anxiety and uncertainty, and improving recall in bad news consultations." *Health Psychology*, vol. 33, no. 11, p. 1382, 2014.

[27] D. E. Shapiro, S. R. Boggs, B. G. Melamed, and J. Graham-Pole, "The effect of varied physician affect on recall, anxiety, and perceptions in women at risk for breast cancer: an analogue study." *Health Psychology*, vol. 11, no. 1, p. 61, 1992.

[28] K. Vonnegut, *Palm Sunday: an autobiographical collage*. Dial Press, 1999.

[29] M. I. Tanveer, S. Samrose, R. A. Baten, and M. E. Hoque, "Awe the audience: How the narrative trajectories affect audience perception in public speaking," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–12.

[30] M. R. Ali, T. Sen, D. Crasta, V.-D. Nguyen, R. Rogge, and M. E. Hoque, "The what, when, and why of facial expressions: An objective analysis of conversational skills in speed-dating videos," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 203–209.

[31] M. K. Hasan, W. Rahman, A. Bagher Zadeh, J. Zhong, M. I. Tanveer, L.-P. Morency, and M. E. Hoque, "UR-FUNNY: A multimodal language dataset for understanding humor," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, Nov. 2019, pp. 2046–2056.

[32] M. R. Ali, K. Van Orden, K. Parkhurst, S. Liu, V.-D. Nguyen, P. Duberstein, and M. E. Hoque, "Aging and engaging: A social conversational skills training program for older adults," in *23rd International Conference on Intelligent User Interfaces*, ser. IUI '18, New York, NY, USA, 2018, p. 55–66.

[33] R. Ali, E. Hoque, P. Duberstein, L. Schubert, S. Z. Razavi, B. Kane, C. Silva, J. S. Daks, M. Huang, and K. Van Orden, "Aging and engaging: A pilot randomized controlled trial of an online conversational skills coach for older adults," *The American Journal of Geriatric Psychiatry*, 2020.

[34] H. Prendinger and M. Ishizuka, "What affective computing and life-like character technology can do for tele-home health care," in *Proc. Workshop HCI and Homecare*. Citeseer, 2004.

[35] M. Peddle, L. Mckenna, M. Bearman, and D. Nestel, "Development of non-technical skills through virtual patients for undergraduate nursing students: an exploratory study," *Nurse education today*, vol. 73, pp. 94–101, 2019.

[36] D. Angus, B. Watson, A. Smith, C. Gallois, and J. Wiles, "Visualising conversation structure across time: Insights into effective doctor-patient consultations," *PloS one*, vol. 7, no. 6, 2012.

[37] A. Kleinsmith, D. Rivera-Gutierrez, G. Finney, J. Cendan, and B. Lok, "Understanding empathy training with virtual patients," *Computers in human behavior*, vol. 52, pp. 151–158, 2015.

[38] M. Hoerger, R. M. Epstein, P. C. Winters, K. Fiscella, P. R. Duberstein, R. Gramling, P. N. Butow, S. G. Mohile, P. R. Kaesberg, W. Tang *et al.*, "Values and options in cancer care (voice): study design and rationale for a patient-centered communication and decision-making intervention for physicians, patients with advanced cancer, and their caregivers," *BMC cancer*, vol. 13, no. 1, p. 188, 2013.

[39] J. C. Weeks, E. F. Cook, S. J. O'day, L. M. Peterson, N. Wenger, D. Reding, F. E. Harrell, P. Kussin, N. V. Dawson, A. F. Connors Jr *et al.*, "Relationship between cancer patients' predictions of prognosis and their treatment preferences," *Jama*, vol. 279, no. 21, pp. 1709–1714, 1998.

[40] L. A. Siminoff, P. Ravdin, N. Colabianchi, and C. M. S. Sturm, "Doctor-patient communication patterns in breast cancer adjuvant therapy discussions," *Health expectations*, vol. 3, no. 1, pp. 26–36, 2000.

[41] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE mobile computing and communications review*, vol. 5, no. 1, pp. 3–55, 2001.

[42] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[43] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Eighth international AAAI conference on weblogs and social media*, 2014.

[44] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[45] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[46] N. Cliff, "Dominance statistics: Ordinal analyses to answer ordinal questions." *Psychological bulletin*, vol. 114, no. 3, p. 494, 1993.

[47] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[48] M. R. Ali, D. Crasta, L. Jin, A. Baretto, J. Pachter, R. D. Rogge, and M. E. Hoque, "Lissa — live interactive social skill assistance," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015, pp. 173–179.

[49] S. Z. Razavi, M. R. Ali, T. H. Smith, L. K. Schubert, and M. E. Hoque, "The lissa virtual human and asd teens: An overview of initial experiments," in *International Conference on Intelligent Virtual Agents*. Springer, 2016, pp. 460–463.

[50] S. Z. Razavi, L. K. Schubert, M. R. Ali, and M. E. Hoque, "Managing casual spoken dialogue using flexible schemas," *Pattern Transduction Trees, and Gist Clauses*, 2017.

[51] J. Hill, W. R. Ford, and I. G. Farreras, "Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations," *Computers in human behavior*, vol. 49, pp. 245–250, 2015.

[52] R. A. Rodenbach, K. Brandes, K. Fiscella, R. L. Kravitz, P. N. Butow, A. Walczak, P. R. Duberstein, P. Sullivan, B. Hoh, G. Xing *et al.*, "Promoting end-of-life discussions in advanced cancer: effects of patient coaching and question prompt lists," *Journal of Clinical Oncology*, vol. 35, no. 8, p. 842, 2017.

[53] J. D. McGreevey, C. W. Hanson, and R. Koppel, "Clinical, legal, and ethical aspects of artificial intelligence–assisted conversational agents in health care," *JAMA*.

[54] P. Malloy, J. Boit, A. Tarus, J. Marete, B. Ferrell, and Z. Ali, "Providing palliative care to patients with cancer: Addressing the needs in kenya," *Asia-Pacific journal of oncology nursing*, vol. 4, no. 1, p. 45, 2017.

[55] E. S. Kamonyo, "The palliative care journey in kenya and uganda," *Journal of pain and symptom management*, vol. 55, no. 2, pp. S46–S54, 2018.

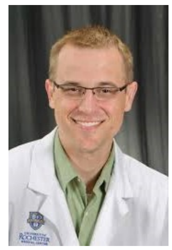
[56] L. Gwyther and F. Rawlinson, "Palliative medicine teaching program at the university of cape town: Integrating palliative care principles into practice," *Journal of pain and symptom management*, vol. 33, no. 5, pp. 558–562, 2007.

Mohammad Rafayet Ali received his Ph.D. in Computer Science from the University of Rochester in 2020. He received MS degree from the same institution in 2016. He earned a B.Sc. degree in Computer Science and Engineering from Bangladesh University of Engineering and Technology in 2013. Currently, he is a postdoctoral fellow at the University of Rochester. His research topics include AI approaches to develop virtual agents for conversational skill coaching and evaluation.

Taylan Sen received his Ph.D. in Computer Science from the University of Rochester in 2021. He received the BS degree in Electrical Engineering and MS degree in Biological Engineering from Cornell University and the JD degree from University at Buffalo. He has seven years industry experience as a software engineer and five years experience as an intellectual property attorney. His research focuses on computational models of nonverbal communication.

Benjamin Kane is a second-year computer science Ph.D. student studying knowledge representation and natural language understanding at the University of Rochester. He received his bachelor's degrees in computer science and economics from the University of Rochester in 2019. His current research involves developing a general-purpose dialogue manager that is capable of enabling realistic and meaningful dialogues between a virtual agent and human.

Shagun Bose Shagun Bose received her Bachelor's degree in Computer Science and Psychology from the University of Rochester in 2020. During her time with the ROC HCI group, her work revolved around creating empathetic and intuitive design. She is currently working as a Software Engineer at Intuit, Inc.



Thomas Carroll is an Associate Professor of Medicine at the University of Rochester. He received his M.D. and Ph.D. from the University of Connecticut and then completed his internal medicine training at the University of Rochester where he subsequently served as Chief Resident and completed the fellowship program in Hospice and Palliative Medicine. His interests include communication training, medical education at all stages of training, and bioethics.

Ronald Epstein MD has conducted groundbreaking research into communication in medical settings and developed innovative educational programs that promote mindfulness, communication and self-awareness. Dr. Epstein co-directs the Center for Communication and Disparities Research and Mindful Practice Programs at the University of Rochester, where he is Professor of Family Medicine, Oncology and Palliative Care. A graduate of Harvard Medical School, he has received numerous humanism awards and fellowships, and the American Cancer Society's highest award, the Clinical Research Professorship.

Lenhart Schubert pursues research in language, dialogue agents, knowledge representation, and reasoning, relevant to agents with common sense, self-motivation, and the ability to acquire knowledge through language. He received a Ph.D. from the University of Toronto in 1970, was a postdoctoral fellow at Johns Hopkins University, a faculty member in Computing Science at the University of Alberta from 1973-1988, and has been a professor of computer science at the University of Rochester, since 1988. He has over 150 publications, was an Alexander von Humboldt Fellow and is a AAAI Fellow.

Ehsan Hoque received his Ph.D. degree from the Massachusetts Institute of Technology in 2013. He is an associate professor of computer science with the University of Rochester where he co-leads the ROC HCI Group. Hoque's research aims to use techniques from artificial intelligence to amplify human ability. His research has been recognized with the MIT TR35 Award, NSF CAREER Award, Emerging Leader by the National Academy of Medicine, among others. He is a member of the ACM, IEEE and AAAI.