# Predicting the Outcome of Patient-Provider Communication Sequences using Recurrent Neural Networks and Probabilistic Models

**Mehedi Hasan, BS[1*], Alexander Kotov, PhD[1*], April Idalski Carcone, PhD[2], Ming Dong, PhD[1], Sylvie Naar, PhD[2]**
**[1]Department of Computer Science, Wayne State University, Detroit, Michigan**
**[2]Department of Family Medicine and Public Health Sciences, School of Medicine, Wayne State University, Detroit, Michigan**

**Abstract** *The problem of analyzing temporally ordered sequences of observations generated by a molecular, physiological or psychological process to make predictions about the outcome of that process arises in many domains of clinical informatics. In this paper, we focus on predicting the outcome of patient-provider communication sequences in the context of clinical dialog. Specifically, we consider prediction of the motivational interview success (i.e. eliciting a particular type of patient behavioral response) based on an observed sequence of coded patient-provider communication exchanges as a sequence classification problem. We propose two solutions to this problem, one that is based on Recurrent Neural Networks (RNNs) and another that is based on Markov Chain (MC) and Hidden Markov Model (HMM), and compare the accuracy of these solutions using communication sequences annotated with behavior codes from the real-life motivational interviews. Our experiments indicate that the deep learning-based approach is significantly more accurate than the approach based on probabilistic models in predicting the success of motivational interviews (0.8677 versus 0.7038 and 0.6067 F1-score by RNN, MC and HMM, respectively, when using under-sampling to correct for class imbalance, and 0.8381 versus 0.7775 and 0.7520 F1-score by RNN, MC and HMM, respectively, when using over-sampling). These results indicate that the proposed method can be used for real-time monitoring of clinical interviews progression and more efficient identification of effective provider communication strategies, which can significantly decrease the effort required to develop behavioral interventions and increases their effectiveness.*

## Introduction

Temporally ordered sequences of discrete or continuous observations generated by molecular, psychological or psychological process(es) arise in many different areas of biology and medicine (e.g., DNA base-pairs, protein sequences, ECG measurements, laboratory results, diagnostic codes, utterances in clinical dialog). Classification (or categorization) is a type of analysis of those sequences that has a broad range of important practical applications, from protein function[1] or structure[2] prediction to detecting individuals with a heart disease[3]. Taking into account both the entire set of observations in a sequence, as well as temporal order and potential dependencies between observations, makes sequence classification a more challenging task than classification of independent observations. Predicting the outcome of those sequences (e.g. physiological or behavioral response) can also be viewed as a sequence classification problem.

In general, sequence classification methods fall into one of the three major classes: feature-based, distance-based and model-based. Feature-based methods transform a sequence into a feature vector and apply a standard supervised machine learning method, such as Support Vector Machine[4] or Decision Tree[5] to arrive at a classification decision. The methods in this class have had limited success, since traditional feature representation methods cannot easily account for the order of and dependencies between observations in a sequence.

Distance-based methods classify a sequence by finding the most similar sequences with known classes based on a distance metric. The most commonly used distance metric is Euclidean distance with Dynamic Time Wrapping[6]. However, distance metrics are primarily designed for time series data, in which the observations are discretized by timestamps. The third type of sequence classification methods first creates a probabilistic model, such as the Markov Chain (MC) or Hidden Markov Model[7] (HMM), for sequences in each class based on the training data and then classifies new sequences by applying the created models. While MCs and HMMs can capture first- and second-order dependencies between adjacent observations in a sequence, learning higher-order dependencies with these models requires prohibitively large amounts of data. By encoding sequences into low-dimensional representations, Recurrent

[*]Authors provided equal contribution.

Neural Networks (RNNs) are able to capture both short- and long-term dependencies and were shown to be effective at modeling different types of sequential data[8]. Long Short-Term Memory (LSTM)[9] is a variant of RNNs, which successfully addressed the vanishing gradient problem[10] of traditional RNN. LSTM demonstrated excellent performance in different domains, from speech[11] and handwriting recognition[12] to health informatics[13,14]. Specifically, LSTM was used as part of a multi-label classification method to recognize patterns in multivariate time series of clinical measurements, such as body temperature, heart rate and blood pressure[13]. LSTM was also effectively used for predicting the diagnosis and medication codes, given a sequence of codes from previous patient visits[14]. A further simplification and improvement of LSTM model, called the Gated Recurrent Unit (GRU)[15], was later proposed. LSTM and GRU demonstrated markedly better performance among all other RNN variants for a variety of tasks in different domains.

In this paper, we address the problem of predicting the outcome of coded patient-provider communication sequences in the context of clinical dialog. Specifically, we focus on predicting the success (i.e. eliciting a particular type of patient behavioral response) of motivational interviews with obese adolescents and their caregivers based on an observed sequence of coded patient-provider communication exchanges during those interviews. Childhood obesity is a serious public health concern in the United States. Recent estimates indicate that approximately one-third (31.8%) of U.S. children 2-19 years of age are overweight and 16.9% are obese[16]. Adolescents, who are obese, are likely to be obese in adulthood and have a greater risk of heart disease, type 2 diabetes, stroke, cancer, and osteoarthritis[17]. One approach to effective obesity intervention is Motivational Interviewing (MI), an evidence-based counseling technique to increase intrinsic motivation and self-efficacy for health-related behavior change. The goal of MI is to encourage patients to explore their own desires, ability, reasons, need for and commitment to the targeted behavior change. These statements, collectively referred to as "change talk" (CHT), consistently predict the actual behavior change[18] that can be sustained for as long as 34 months[19] after an interview. However, the ability of providers to consistently elicit this type of patient communication requires knowledge of effective communication strategies for a variety of patients, which can only be obtained through analysis of a large number of annotated interviews. Since manual examination and analysis of MI interview transcripts is a very time-consuming process, designing effective MI interventions and tailoring them to particular populations can take years. Therefore, there is a need for informatics-based methods to facilitate the development of effective behavioral interventions, in general, and theoretically-grounded computational models to explore the mechanisms of MI's efficacy, in particular.

Our goal is to compare the accuracy of probabilistic models, such as MC and HMM, and deep learning methods, such as LSTM and GRU, for the task of predicting the success of clinical interviews (i.e. eliciting a particular type of patient behavioral response, such as CHT) at any point during a clinical interview based on a sequence of coded previous patient-provider communication exchanges in the same interview, which we consider as a sequence classification problem. This study is a continuation of our previous work[20,21], in which we explored several machine learning methods for automatic annotation of clinical interview fragments with a large number of patient and provider behavior codes from a specialized codebook[22]. While there have been some previous qualitative studies of patient-provider dialog in a clinical setting[23], no previous work explored applicability of state-of-the-art methods for sequence modeling to the analysis of patient-provider communication (PPC) exchanges, in general, and predicting the desired patient behavioral response in the context of motivational interviews, in particular.

## Methods

### *Data collection*

The experimental dataset for this work was constructed from the transcripts of 129 motivational interviews, which consist of a total of 50,239 segmented and annotated utterances. Each transcript consists of an MI interview session, which typically involves a counselor, an adolescent and a caregiver. The utterances were annotated based on the MYSCOPE codebook[22], in which the behavior codes are grouped into the patient (adolescent and caregiver) codes and the counselor codes. Annotated utterances were divided into successful and unsuccessful communication sequences. Successful communication sequences are the ones, which resulted in positive change talk (CHT+) or commitment language (CML+) statements by an adolescent or a caregiver, while unsuccessful sequences are the ones, which resulted in negative change talk (CHT-) or commitment language (CML-), or the ones, in which no change talk or commitment language statements were made.

A fragment of an adolescent session transcript is presented in Table 1. In this example, $SS \rightarrow OQO \rightarrow HUPO \rightarrow OQTBN \rightarrow CHT+$ is a successful sequence, in which a counselor starts with an open-ended question and ultimately is able to elicit a positive change talk statement. As follows from this example, similar utterances, such as "Yeah" and "Yes", can be assigned different behavior codes (CHT+ and HUPW, respectively), depending on the context.

**Table 1:** Fragment of the annotated transcript of a dialogue between a counselor and an adolescent. MYSCOPE codes assigned to the utterances and their meaning are shown in the first two columns.

| Code | Behavior | Speaker | Utterance |
|---|---|---|---|
| SS | Structure Session | Counselor | Okay. Can I meet with Xxxx alone for a few minutes? |
| OQO | Open-ended question, other | Counselor | So, Xxxx, how you doing? |
| HUPO | High uptake, other | Adolescent | Fine |
| OQTBN | Open-ended question, target behavior neutral | Counselor | That's good. So, tell me how do you feel about your weight? |
| CHT+ | Change talk positive | Adolescent | It's not the best. |
| CQECHT+ | Closed question, elicit change talk positive | Counselor | It's not the best? |
| CHT+ | Change talk positive | Adolescent | Yeah |
| CQTBN | Closed question, target behavior neutral | Counselor | Okay, so have you tried to lose weight before? |
| HUPW | High uptake, weight | Adolescent | Yes |

The resulting experimental dataset was highly imbalanced. Out of 5143 observed sequences, 4225 or 82.15% were positive and only 918 or 17.85% were negative. No major differences were observed in the average length of successful (9.79 utterances) and unsuccessful (9.65 utterances) sequences.

Since severely imbalanced datasets often distort the true performance of a classification method relative to a simple "majority vote" baseline (e.g. simply classifying every communication sequence as successful would result in 82.15% accuracy on our dataset). Therefore, it is important to properly address class imbalance. We evaluated the performance of probabilistic and deep learning methods using both under-sampling and over-sampling for balancing the number of samples in different classes. Synthetic Minority Over Sampling Technique (SMOTE)[24] is a widely used oversampling method for imbalanced datasets, in which new synthetic examples are generated for minority classes. Specifically, we generated synthetic examples at the borderline between the majority and minority classes[25]. On the other hand, the under-sampling method reduces the number of samples in majority class by replacing the clusters of samples identified by the $k$-means clustering algorithm with the cluster centroids.

Two MC and HMM models were trained, one model was trained on successful sequences, while another was trained on unsuccessful sequences.

***Prediction method***

Generally, a sequence can be viewed as a temporally ordered set of events. In this study, an event is a behavior code that also has a symbolic representation, such as $LUP+$ (low uptake, positive), $OQECHT+$ (open-ended question, elicit change talk positive), etc. Given a sequence of behavior codes $S_i = \{c_1, c_2, ..., c_n\}$ representing patient-provider communication exchanges during some part of a motivational interview, the task of predicting interview success can be viewed as sequence classification. Given a set of class labels $L = \{l_1, l_2, ..., l_m\}$ (in our case, the labels are "successful" and "unsuccessful" motivational interview), a sequence classifier $C$ learns a function $S_i \rightarrow l_i, l_i \in L$ that maps a sequence $S_i$ into a class label $l_i \in L$.

Our proposed baseline prediction method consists of two steps. In the first step, we model successful and unsuccessful patient-provider interactions using first and second-order Markov Chain and Hidden Markov Model, which are popular probabilistic models for discrete observation sequences with finite vocabulary. In the second step, we classify each

test sequence based on the maximum likelihood of generating that sequence from each model. Although HMM was originally developed for speech recognition[7], it is one of the most widely used methods for sequence modeling[26,27]. However, the latest advances in deep learning indicate that RNNs provide better results than conventional machine learning methods for the task of sequence classification. Therefore, we employed two special variants of RNN in our experiments: Long Short-Term Memory (LSTM)[9] and Gated Recurrent Unit (GRU)[15].

**Markov Chain (MC)** is a probabilistic model that predicts the probability of next state based on its current state but not on its past states (Markov property). For the sequential analysis, we built two Markov models $M$ and $\overline{M}$, summarizing counselor strategies and patient responses, in the cases of successful ($M$) and unsuccessful ($\overline{M}$) motivational interviews. A Markov model $M$ can be represented as a weighted directed graph $G = (V, E, p)$, in which:

- $V = \{CML+, CHT+, CHT-, AMB-, LUP+, LUP-, HUPW, OQO, CQTBN, CQECHT+, ...\}$ is a set of vertices, consisting of adolescent, caregiver and counselor MI behavior codes;
- $E \subseteq V \times V$ is a set of edges corresponding to possible transitions from one MI behavior code to the other in a sequence;
- $p_M : E \rightarrow [0...1]$ is a function that assigns probability $p(c_i|c_j)$ to an edge between the MI behavior codes $c_i$ and $c_j$ based on maximum likelihood estimator:

$$P_M(c_j|c_i) = \frac{n_{c_i,c_j}}{n_{c_i}} \tag{1}$$

where $n_{c_i,c_j}$ and $n_{c_i}$ are the number of times a transition between the MI behavior codes $c_i$ and $c_j$ and the code $c_i$ has been observed in the training data. Given a Markov model $M$ (such that $S \subseteq V$), the probability that a sequence of MI behavior codes $S = \{C_1, ..., C_N\}$ has been generated from a Markov model $M$ is:

$$P_M(S) = \prod_{i=2}^{N} p_M(c_i|c_1, \ldots, c_{i-1}) = \prod_{i=2}^{N} p_M(c_i|c_{i-1}) \tag{2}$$

In the second step, we quantify the likelihood of success of a given motivational interview at a certain time point given a sequence of MI behavior codes $S$ observed prior to that point as:

$$p(S \rightarrow successful) = \log\left(\frac{P_M(S)}{P_{\overline{M}}(S)}\right) = \sum_{i=2}^{N} \log p_M(c_i|c_{i-1}) - \sum_{i=2}^{N} \log p_{\overline{M}}(c_i|c_{i-1}) \tag{3}$$

If $p(S \rightarrow successful) > 0$, a communication sequence is predicted to result in positive change talk or commitment language. Otherwise, it would be classified as a negative change talk or commitment language.

The above model also referred as first-order MC, since it only considers immediately preceding behavior code, when computing the state transition probabilities. In our experiment, we also considered second-order Markov model that computes transition probabilities based on previous two states.

**Hidden Markov Model (HMM)** is another probabilistic model used for modeling processes varying in time. HMMs are widely used for sequence analysis because of their ability to incorporate dependencies among elements in sequence. Mathematically, HMM can be defined as the following: $\lambda = (A, B, \pi)$.

- A is an $N \times N$ state transition probability distribution matrix $A = \{a_{ij}\}$
- B is an $N \times M$ matrix $B = \{b_j(k)\}$ with observation symbol probability distribution for each state
- $\pi$ is the initial state distribution vector $\pi = \{\pi_i\}$

Hence, $N$ is a number of hidden states in the model and $M$ is a number of distinct observations symbols per state, i.e. the discrete vocabulary size. The key difference between HMM and MC is that HMM requires specifying the number of hidden states as a model parameter and then the model deduces a sequence of hidden states that best explains the observations along with state transition probabilities and distributions of observation symbols (emission probabilities) for each state. The Baum-Welch algorithm was used to estimate the parameters of HMMs for successful and unsuccessful interviews using the corresponding training set, while the Viterbi algorithm was used to determine the most likely sequence of hidden states for a given sequence of observations. After assignment of hidden states, the log-likelihood of successful outcome can be estimated using Eq. 3.

**Embeddings**: We took the inspiration for the representation of behavior codes from the idea of word embeddings[28]. Word embedding is a representation of words in low-dimensional space by vectors, which contain the features of the words. In our study, we employed embedding in place of one-hot vectors for representation of behavior codes as input to LSTM and GRU, since one-hot vectors are high-dimensional and sparse. Moreover, code embeddings have the ability to represent semantically similar codes with similar vectors in a low-dimensional space by retaining their internal relations. The embedding was learned during the training of the classification model. Figure 1 illustrates the MYSCOPE code embedding in 44-dimensional vector space visualized by t-SNE[29]. It can be seen that positive behavior codes such as OQECHT+, OQECML+, AF, AFL, SUP, RCML+S, CQECML+, etc. formed a cluster in the left part of the figure. The nearest neighbors of CQECML+ are highlighted by their color intensity (i.e. OQECML+ being more purple means closer to CQECML+). The right part of the figure demonstrates another cluster formed with negative behavior codes including CQECML-, AMB-, RCHT-C, OQECHT-, GINFO-, RBAC, LUP-, RCHT-S, RPTBC, RAMBC, AMB-, RCML-S, etc. It is interesting that the behaviors intended to elicit CHT+/CML+ group together whereas the ones intended to elicit CHT-/CML- also group together and are located on opposite ends of the space.


**Figure 1: T-SNE diagram of behavior codes drawn after 4000 steps with perplexity 30**.

**Recurrent Neural Networks (RNN)** are a type of Neural Networks, which can send feedback from the current hidden state to the hidden state of the next time step. The ability of RNN to capture long-term dependencies for predicting the future events is its main advantage over MC and HMM. The capability of remembering past event is very useful in motivational interviews where a behavior observed at some point in the interview is very informative for the future behaviors, that will be observed later, but not immediately. However, it was observed that RNN fails to capture long-term dependencies due to vanishing gradient problem[10]. In order to mitigate this problem, Hochreiter et al.[9] proposed a special kind of RNN called Long Short Term Memory networks or simply LSTM. There are several variants of LSTM

model, among which the most notable one is the Gated Recurrent Unit[30] (GRU). GRUs are simpler than LSTM units and have been shown to experimentally outperform other models[30]. GRU is formally defined as follows (Eq. 4 to 7):

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{4}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{5}$$

$$\tilde{h}_t = tanh(W_h x_t + r_t \odot U_h h_{t-1} + b_h) \tag{6}$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t \tag{7}$$

As these equations, $\sigma$ stands for an element-wise application of the sigmoid function and $\odot$ is an element-wise product. The update gate $z_t$ and reset gate $r_t$ at time step $t$ are computed by the Eq. (4) and (5), where $W$ and $U$ represent weights and $b$ represents biases. The activation $h_t$ of the GRU at time $t$ is a linear combination of previous activation $h_{t-1}$ and the candidate activation $\tilde{h}_t$, which is represented by Eq. (7) and (6).

$$total\ loss = \alpha \cdot \frac{1}{T} \sum_{t=1}^{T} loss(\bar{y}^{(t)}, y^{(t)}) + (1 - \alpha) \cdot loss(\bar{y}^{(T)}, y^{(T)}) \tag{8}$$

The architecture of the proposed Recurrent Neural Network is shown in Figure 2. We used one hidden layer of 15

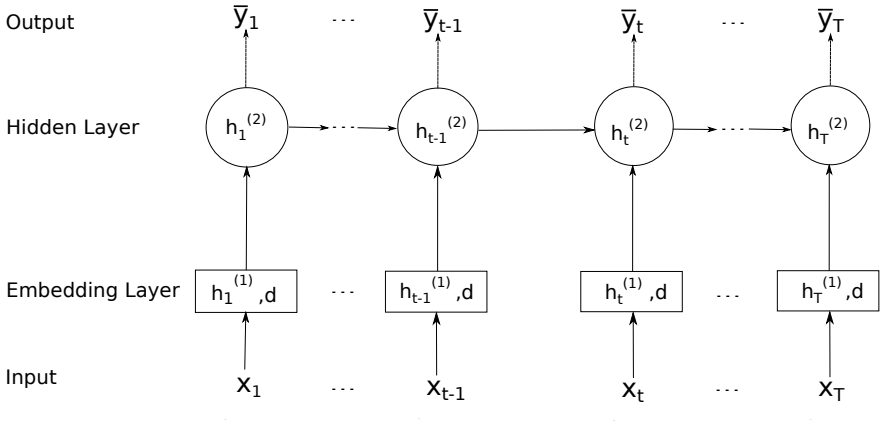


**Figure 2: Proposed RNN model with target replication (TR)**.

LSTM or GRU nodes and embedding of MYSCOPE codes in an observed sequence of communication behaviors as input. Then, we applied a softmax layer at each time-step for predicting the label of the sequence. Since the sequence label was predicted at each time step, the proposed model is also known as Recurrent Neural Network with Target Replication (RNN-TR). As can be seen from Eq. 8, the total loss of our prediction models is the combination of final loss and the mean of the losses over all sequence steps, where $T$ is the total number of elements in a sequence, $\bar{y}^{(t)}$ is the output at step $t$, and $\alpha\ \epsilon\ [0, 1]$ is a hyperparameter that determines the relative importance of intermediate outputs. Experimentally, it was determined that best performance achieved when we use $\alpha = 0.5$. Our model also contains several other hyperparameters such as embedding dimension, number of hidden units, learning rate, batch size, etc., which were determined by the validation set. We implemented our models in Tensorflow with Adam optimizer and early stopping based on the validation loss. Experimentally, we observed that our model converges to optimal after 100 epochs and have no effect of dropout and regularization. The code for all models is publicly available at https://github.com/teanalab/myscope-sequential-analysis.

### *Evaluation metrics*

Performance of the proposed method was evaluated in terms of precision, recall, and F-measure using 10 folds cross-validation and weighted macro-averaging of these metrics over the folds. However, LSTM and GRU are trained on 80% of the data and validated on 10%. The remaining 10% of the data is used as a test set for reporting the performance of the model.

**Results**

Experimental evaluation of the proposed method is conducted on both under and over-sampled sequences. Predictive performance summary of the proposed methods on under and over-sampled sequences is presented in Table 2.

**Table 2:** Performance of MC, HMM, and RNN for predicting the success of under and over-sampled patient-provider communication sequences. The highest value for each performance metric is highlighted in bold.

| Method | Undersampling | | | Oversampling | | |
|---|---|---|---|---|---|---|
| | **Precision** | **Recall** | **F1-Score** | **Precision** | **Recall** | **F1-Score** |
| Markov Chain $1^{st}$ Order | 0.7060 | 0.7044 | 0.7038 | 0.7932 | 0.7799 | 0.7775 |
| Markov Chain $2^{nd}$ Order | 0.6395 | 0.6385 | 0.6379 | 0.7111 | 0.7029 | 0.7000 |
| Hidden Markov Model | 0.6244 | 0.6143 | 0.6067 | 0.7775 | 0.7567 | 0.7520 |
| LSTM RNN | 0.8672 | 0.8626 | 0.8622 | 0.8411 | 0.8372 | 0.8368 |
| LSTM RNN - TR | **0.8733** | **0.8681** | **0.8677** | **0.8424** | **0.8385** | **0.8381** |
| GRU RNN | 0.8674 | 0.8648 | 0.8646 | 0.8379 | 0.8342 | 0.8337 |
| GRU RNN - TR | 0.8705 | 0.8676 | 0.8673 | 0.8412 | 0.8377 | 0.8373 |

***Predictive performance for under-sampled Patient-Provider Communication code sequences***

We applied small learning rate (0.00005) and batch size (8) with early stopping strategy for training deep learning models when the under-sampled dataset is used. Five major conclusions can be drawn from the results in Table 2. First, recurrent neural networks outperform probabilistic models and achieve 16.39%-26.1% higher F1-score. Second, LSTM with target replication has the best performance over all other RNN methods, and achieved F1-score 0.8677 with precision 0.8733 and recall 0.8681. Third, target replication strategy improves the performance of GRU and LSTM, while conventional GRU shows better performance than traditional LSTM. Fourth, among probabilistic models, the MC-based method generally outperforms HMM across all metrics for under-sampled sequences. Fifth, second-order MC has lower precision, recall, and F-measure than first-order MC. In particular, precision, recall and F-measure decrease by 9.42%, 9.36% and 9.36%, when going from first to second-order MC model.

***Predictive performance for over-sampled Patient-Provider Communication code sequences***

Similar to the under-sampled dataset, early stopping strategy was also employed for the over-sampled dataset. For over-sampled data, RNN models were trained with learning rate 0.00010 and batch size 55. Experimental results indicate that HMM had better performance than second-order MC, achieving 9.34%, 7.65%, and 7.43% higher precision, recall, and F-measure, while HMM still has 1.98%, 2.97%, and 3.28% lower precision, recall, and F-measure than first-order Markov chain. Similar to under-sampled sequences, target replication improves the performance of RNN models, and LSTM with target replication has the highest F1-score among all models. However, the predictive performance of recurrent neural networks decreases in over-sampled sequences, while the performance of probabilistic models increases. We also observed that all models have the largest value in precision compared to other performance metrics.

***Most likely communication sequences***

Table 3 provides examples of typical patient-provider communication sequences that frequently appear in successful and unsuccessful motivational interviews. In successful sequences, we observed that information is frequently provided using patient-centered communication (GINFO+) and structure session, when the counselor either explains the therapeutic agenda or attempts to transition to a new topic or session content (SS). Sometimes, counselor acknowledges the clients' communication or an off topic comment (SO). We also observed that the affirm and OQECML+ are consistent with MI theory, which claims their strong effect for eliciting positive change talk or commitment language. For unsuccessful sequences, it can be seen that providing advice using non-patient centered strategies (ADV-) leads to the ambivalence that is weighted against change (AMB-), which is heading in the wrong direction therapeutically.

**Table 3:** Most likely communication sequences in successful and unsuccessful motivational interviews.

| Type | Most likely communication sequences |
|---|---|
| successful | GINFO+: General information, positive → LUP+: Low uptake, positive → OQTBN: Open-ended question, target behavior neutral |
| successful | SS: Structure session → GINFO+: General information, positive → CQECHT+: Closed-ended question, elicit change talk positive |
| successful | SO: Statement, other → LUP+: Low uptake, positive → AF: Affirm → HUPW: High uptake, weight → OQECML+: Open-ended question, elicit commitment language positive. |
| unsuccessful | ADV+: Advise, positive → AMB-: Ambivalence negative → OQECHT-: Open-ended question, elicit change talk negative |
| unsuccessful | CQECHT+: Open-ended question, elicit change talk positive → RCHT-S: Reflect, change talk negative → OQECHT-: Open-ended question, elicit change talk negative |
| unsuccessful | SUP: Support → AF: Affirm → CQTBN: Closed-ended question, target behavior neutral → OQECHT-: Open-ended question, elicit change talk negative → AMB-: Ambivalence negative |

Questions phased to elicit negative change talk or commitment language leads to CHT- or CML-, which is consistent with the manual coding approach or analysis, or AMB-.

**Discussion**

By analyzing the experimental results of different communication sequence outcome prediction methods proposed in this paper, we arrived at the following conclusions. First, the overall predictive performance of RNN models is substantially better than probabilistic models. In particular, the RNN-based method achieves near-human accuracy for predicting the label of motivational interviews. This indicates that RNN is able to capture the structure of discourse in motivational interviews by preserving long-term dependencies among the behavior codes, which reflect the overall progression of the interviews. This provides evidence that the RNN is able to successfully replicate human cognitive processes to integrate previous information when formulating higher level thinking. In addition to that, embeddings allow to reduce the dimensionality of codes in PPC sequences and consequently improve both precision and recall of the prediction method.

Second, using target replication to compute the loss at each time step results in better performance for all configurations of the proposed RNN-based methods. This indicates that mean of the losses over all steps emphasize the dependencies between the pairs of patient and provider codes, which results in more accurate estimates of the model parameters. Better estimates of parameters in RNN models of motivational interviews are propagated to the next step based on the relative importance of intermediate output, where they are aggregated into predictions for the entire sequence. This allows achieving an improvement in the prediction accuracy of the method.

Third, using first-order Markov model results in better prediction accuracy compared to higher-order Markov models, which we attribute to the fact that the number of states in higher-order Markov models grows exponentially with their order. As a result, accurate estimation of transition probabilities requires much larger training set. Using smaller datasets such as under-sampled dataset will result in a sparsity problem, when many transitions are either not observed in the training set at all or observed only a few times, leading to missing or potentially inaccurate probability estimates. Obtaining large training sets cannot be easily accomplished in many domains, including motivational interviewing. In this study, we found out that using first-order Markov models is a reasonable trade-off between efficiency and accuracy.

Fourth, similar to traditional Markov model, HMM achieves a dramatic improvement in the prediction accuracy when larger training set is used. This indicates that sufficient training data is required to estimate optimal hyperparameters such as a number of hidden states, initial state distribution, transition probabilities, and emission probabilities.

Fifth, the proposed method can be used to identify the most effective communication strategies for eliciting a particular type of behavioral response. Awareness of these strategies by researchers can significantly decrease the time and effort required to develop effective interventions to address many public health conditions, such as childhood obesity, and tailor these interventions to particular patient cohorts. Awareness of these strategies by the counselors can lead to a

greater success rate of motivational interviews.

## Conclusion

In this paper, we compared the accuracy of Recurrent Neural Networks with Markov Chain and Hidden Markov Model for the task of predicting the success of motivational interviews. We found out that individual patient-provider communication exchanges are highly indicative of the overall progression and future trajectory of clinical interviews and can be used to predict their overall success. Our proposed methods can facilitate motivational interviewing researchers in establishing causal relationships between different communication strategies and the desired behavioral outcomes during the interviews without resource-intensive manual qualitative analysis of interview transcripts, which can significantly decrease the time and effort required to develop behavioral interventions. Our proposed methods can also help to identify most likely sequences that are common to successful and unsuccessful motivational interviews, which can directly inform clinical practice and increase the effectiveness of behavioral interventions. Our experimental results indicate that the proposed methods can also be used for real-time monitoring of the progression of clinical interviews. This work also has broad implications for public health research by providing a formal theoretically-grounded computational mechanism for qualitative data analysis.

## Acknowledgments

This study was supported by a grant from the National Institutes of Health, NIDDK R21DK108071, Carcone and Kotov, MPIs. We would like to thank the student assistants in the Department of Family Medicine and Public Health Sciences at Wayne State University School of Medicine for their help in developing the training dataset by manually annotating the dataset using the MYSCOPE codebook.

## References

[1] Yakhnenko O, Silvescu A, Honavar V. Discriminatively trained markov model for sequence classification. In: Data Mining, Fifth IEEE International Conference on. IEEE; 2005. p. 8–pp.

[2] Deshpande M, Karypis G. Evaluation of techniques for classifying biological sequences. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer; 2002. p. 417–431.

[3] Wei L, Keogh E. Semi-supervised time series classification. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2006. p. 748–753.

[4] Leslie C, Kuang R. Fast string kernels using inexact matching for protein sequences. Journal of Machine Learning Research. 2004;5(Nov):1435–1455.

[5] Chuzhanova NA, Jones AJ, Margetts S. Feature selection for genetic sequence classification. Bioinformatics. 1998;14(2):139–143.

[6] Keogh EJ, Pazzani MJ. Scaling up dynamic time warping for datamining applications. In: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2000. p. 285–289.

[7] Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE. 1989;77(2):257–286.

[8] Lipton ZC, Berkowitz J, Elkan C. A critical review of recurrent neural networks for sequence learning. arXiv preprint arXiv:150600019. 2015;.

[9] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997;9(8):1735–1780.

[10] Bengio Y, Frasconi P, Simard P. The problem of learning long-term dependencies in recurrent networks. In: Neural Networks, 1993., IEEE International Conference on. IEEE; 1993. p. 1183–1188.

[11] Graves A, Mohamed Ar, Hinton G. Speech recognition with deep recurrent neural networks. In: Acoustics, speech and signal processing (icassp), 2013 ieee international conference on. IEEE; 2013. p. 6645–6649.

[12] Nion T, Menasri F, Louradour J, Sibade C, Retornaz T, Métaireau PY, et al. Handwritten information extraction from historical census documents. In: Document Analysis and Recognition (ICDAR), 2013 12th International Conference on. IEEE; 2013. p. 822–826.

[13] Lipton ZC, Kale DC, Elkan C, Wetzell R. Learning to diagnose with LSTM recurrent neural networks. arXiv preprint arXiv:151103677. 2015;.

[14] Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor ai: Predicting clinical events via recurrent neural networks. In: Machine Learning for Healthcare Conference; 2016. p. 301–318.

[15] Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:14123555. 2014;.

[16] Ogden CL, Carroll MD, Kit BK, Flegal KM. Prevalence of obesity and trends in body mass index among US children and adolescents, 1999-2010. Jama. 2012;307(5):483–490.

[17] General US. Surgeon Generals vision for a healthy and fit nation. Washington, DC: HHS. 2010;.

[18] Apodaca TR, Longabaugh R. Mechanisms of change in motivational interviewing: a review and preliminary evaluation of the evidence. Addiction. 2009;104(5):705–715.

[19] Walker D, Stephens R, Rowland J, Roffman R. The influence of client behavior during motivational interviewing on marijuana treatment outcome. Addictive Behaviors. 2011;36(6):669–673.

[20] Kotov A, Hasan M, Carcone A, Dong M, Naar-King S, BroganHartlieb K. Interpretable probabilistic latent variable models for automatic annotation of clinical text. In: AMIA Annual Symposium Proceedings. vol. 2015. American Medical Informatics Association; 2015. p. 785.

[21] Hasan M, Kotov A, Carcone AI, Dong M, Naar S, Hartlieb KB. A study of the effectiveness of machine learning methods for classification of clinical interview fragments into a large number of categories. Journal of biomedical informatics. 2016;62:21–31.

[22] Carcone AI, Naar-King S, Brogan K, Albrecht T, Barton E, Foster T, et al. Provider communication behaviors that predict motivation to change in black adolescents with obesity. Journal of developmental and behavioral pediatrics: JDBP. 2013;34(8):599.

[23] Eide H, Quera V, Graugaard P, Finset A. Physician–patient dialogue surrounding patients expression of concern: applying sequence analysis to RIAS. Social Science & Medicine. 2004;59(1):145–155.

[24] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002;16:321–357.

[25] Nguyen HM, Cooper EW, Kamei K. Borderline over-sampling for imbalanced data classification. International Journal of Knowledge Engineering and Soft Data Paradigms. 2011;3(1):4–21.

[26] Mutsam N, Pernkopf F. Maximum margin hidden Markov models for sequence classification. Pattern Recognition Letters. 2016;77:14–20.

[27] Won KJ, Prügel-Bennett A, Krogh A. Training HMM structure with genetic algorithm for biological sequence analysis. Bioinformatics. 2004;20(18):3613–3619.

[28] Bengio Y, Ducharme R, Vincent P, Jauvin C. A neural probabilistic language model. Journal of machine learning research. 2003;3(Feb):1137–1155.

[29] Maaten Lvd, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research. 2008;9(Nov):2579–2605.

[30] Cho K, Van Merriënboer B, Bahdanau D, Bengio Y. On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:14091259. 2014;.