Detecting conversation topics in primary care office visits from transcripts of patient-provider interactions

Jihyun Park,1 Dimitrios Kotzias,1 Patty Kuo,2 Robert L Logan IV,1 Kritzia Merced,2 Sameer Singh,1 Michael Tanana,3 Efi Karra Taniskidou,1 Jennifer Elston Lafata,4,5 David C Atkins,6 Ming Tai-Seale,7 Zac E Imel,2 and Padhraic Smyth1

1Department of Computer Science, University of California, Irvine, Irvine, California, USA, 2Department of Educational Psychology, University of Utah, Salt Lake City, Utah, USA, 3Social Research Institute, University of Utah, Salt Lake City, Utah, USA, 4Division of Pharmaceutical Outcomes and Policy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA, 5Center for Health Policy and Health Services Research, Henry Ford Health System, Detroit, Michigan, USA, 6Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, Washington, USA, and 7Department of Family Medicine and Public Health, University of California San Diego, La Jolla, California, USA

Corresponding Author: Jihyun Park, PhD Student, Donald Bren Hall, University of California, Irvine, CA 92697, USA; jihyupn@ics.uci.edu

ABSTRACT

Objective: Amid electronic health records, laboratory tests, and other technology, office-based patient and provider communication is still the heart of primary medical care. Patients typically present multiple complaints, requiring physicians to decide how to balance competing demands. How this time is allocated has implications for patient satisfaction, payments, and quality of care. We investigate the effectiveness of machine learning methods for automated annotation of medical topics in patient-provider dialog transcripts.

Materials and Methods: We used dialog transcripts from 279 primary care visits to predict talk-turn topic labels. Different machine learning models were trained to operate on single or multiple local talk-turns (logistic classifiers, support vector machines, gated recurrent units) as well as sequential models that integrate information across talk-turn sequences (conditional random fields, hidden Markov models, and hierarchical gated recurrent units).

Results: Evaluation was performed using cross-validation to measure 1) classification accuracy for talk-turns and 2) precision, recall, and F1 scores at the visit level. Experimental results showed that sequential models had higher classification accuracy at the talk-turn level and higher precision at the visit level. Independent models had higher recall scores at the visit level compared with sequential models.

Conclusions: Incorporating sequential information across talk-turns improves the accuracy of topic prediction in patient-provider dialog by smoothing out noisy information from talk-turns. Although the results are promising, more advanced prediction techniques and larger labeled datasets will likely be required to achieve prediction performance appropriate for real-world clinical applications.

Key words: classification, supervised machine learning, patient care, communication
INTRODUCTION

Background

Appropriate documentation of the clinical visit is critical for communication among medical professionals, enabling quality assurance, and accurate billing and reimbursement. The traditional way of documenting a clinical visit in the electronic health record, namely physicians’ notes, provide a source of valuable information on what occurred during the interaction and what physicians consider to be important. Electronic health records have improved the accessibility of medical information, but patients are demanding access to information at a greater scale. Pressure to quickly document the medical visit may lead providers to type into the record during the medical visit, which can negatively impact patient-provider communication. Primary care providers spend about half of their time working on computers, which appears partly responsible for growing concerns of physician burnout across a wide range of physician specialties. In addition, physician generated notes do not always provide an accurate representation of what occurred during the visits.

Technologies that could reduce the burden of documentation on providers, and increase convergence between documentation of a visit and the content of the clinical interaction are greatly needed. Natural language processing (NLP) technologies combined with advances in automatic speech recognition offer potentially promising solutions. Information extraction and summarization technologies built on top of resulting transcripts will be needed to take the next step in reducing the burden of documentation on physicians and in providing clinical decision support to both patients and physicians. If successful, machine learning–enabled automatic speech recognition and charting could free up valuable time for physicians to talk to their patients, rather than typing extensively during clinical encounters.

Patient-provider conversations are complex, multidimensional, and multifunctional. Patients present multiple issues during an office visit requiring clinicians to divide time and effort during a visit to address competing demands such as a patient could be concerned about blood pressure, knee pain, and blurry vision in a single appointment. Moreover, visit content does not solely focus on biomedical issues, but also on psychosocial matters, personal habits, mental health, patient-physician relationship, and small talk. Health communications researchers analyze the content of patient-provider communication by directly labeling the interaction using trained raters to label the topical content of clinical interactions. For example, during periodic health examinations with their primary care physicians, only one-third of patients with mental health needs had a discussion about mental health. These findings suggest that quality improvement efforts that evaluate the content of clinical interactions might help address gaps in service delivery.

However, labeling each talk-turn using coding systems designed to capture the content of medical visits, such as the Multi-Dimensional Interaction Analysis system, is labor intensive and costly. It requires training a team to label the text and establish and maintain reliability. Depending on the extensiveness of the labeling system, it can take several hours to label one patient-provider interaction. This level of effort means that even in research settings with resources for detailed evaluation, studies of patient-provider interaction are often limited in scale. As a result, the direct evaluation of clinical interactions is not feasible in routine care settings for quality improvement purposes. Automated methods capable of extracting the topical content of clinical encounters could support providers who overlook asking patients about critical issues (e.g., suicide, blood pressure medication) when it is clinically indicated, reduce the burden of documentation currently placed on providers, and facilitate large-scale research on the quality of patient-provider communication.

Related work

The past decade has seen an explosion of interest in machine learning and NLP in medical contexts, targeting problems such as automated extraction of information from clinical notes and electronic health records. Of direct relevance to the present article is a growing body of work dedicated to applying methods from machine learning and NLP to the automatic annotation of conversations between providers and patients. The typical approach in such studies begins with labeling a corpus of transcript data (e.g., providing human-generated labels for each utterance or talk-turn in each visit in the corpus). Machine learning techniques are then used to learn a classification model from a subset of the corpus, the training data, and the model’s predictions are evaluated by comparing its predictions with the known human labels on unseen test transcripts. For example, Mayfield et al analyzed patient-provider transcripts from the ECHO (Enhancing Communication and HIV Outcomes) study and employed a logistic regression model to classify utterances into the categories “information-giving” or “information-requesting.” More recently, Kovot et al analyzed transcriptions from motivational interviewing related to pediatric obesity and developed probabilistic machine learning models for classifying patient utterances into classes of behavioral responses. In later work on the same dataset, Hasan et al compared probabilistic models with more recent recurrent neural network approaches and found the latter to be generally more accurate on that dataset.

There has been less prior work on the problem of automated classification of topical content in patient-provider dialog. Wallace et al developed machine learning models to classify dialog utterances into 1 of 6 high-level discussion topics: biomedical, logistics, psychosocial, antiretroviral (ARV), missing/other, and socializing. Using the same ECHO dataset as used in the Mayfield et al study, they evaluated the performance of conditional random fields (CRFs) for this prediction task and concluded that the results showed promise for automated classification of patient-provider interactions into clinically relevant topics. Gaut et al proposed the use of labeled topic models for classifying psychotherapy sessions with 161 possible topic labels, using a dataset published by Alexander Street Press, and again finding that these models showed promise in terms of predictive ability.

These earlier studies demonstrate that machine learning systems can generate plausible annotations of medical dialogues—our work in this article pursues this line of research further. We differ from earlier work on topic classification in a number of aspects. For example, we compare both probabilistic and neural network classification methods for topic classification of talk-turns, whereas Wallace et al and Gaut et al only focused on probabilistic approaches. We also evaluate performance for a more detailed set of 27 topics compared with the 6 high-level topics used in the Wallace et al study. The Gaut et al study also differs from our work in that it primarily focused on session-level labels and did not investigate the use of sequential information across talk-turns for talk-turn–level predictions as we do here. To our knowledge, this is the first study that systematically compares sequential and nonsequential classification methods, for both probabilistic and neural network models, on the
MATERIALS AND METHODS

Dataset

The source data include transcripts of audio-recordings of primary care office visits from the MHD (Mental Health Discussion) study. Each transcript corresponds to a visit between a patient and a provider—a small fraction of the dialog corresponds to other participants in the conversation (such as a nurse and family member). Data collection occurred from 2007 to 2009 in a health system in Michigan with 26 ambulatory care clinics. Patients were 50-80 years of age, all had insurance, and were due for a colorectal cancer screening at the time of appointment. All aspects of the research protocol were approved by relevant organizations’ institutional review boards.

Each visit is comprised of a series of talk-turns, with 122,083 talk-turns in total across 279 visits (median and mean of 408 and 438 talk-turns per visit, respectively, with upper and lower quartiles of 312 and 522) from 59 providers. Each talk-turn was manually assigned by a human coder (labeler) to 1 of 39 different topic labels that were modified from the Multi-Dimensional Interaction Analysis coding system. A topic is defined as an issue raised in a conversation that required a response from the other member of the dyad and had at least two exchanges between the dyad. A small number of talk-turns were split into two if the turn straddled two topics. This resulted in a few of the original talk-turns being represented as two separate talk-turns after coding. Figure 1 illustrates how different topics were assigned to talk-turns during a short portion of a particular visit. Topic labels that occurred in less than 20 visits were merged into a single topic denoted as Other, resulting in a total of 27 unique topics in the corpus.

Table 1 provides the names of the topic labels, a brief description of each, as well as the percentage of talk-turns assigned to each by the labelers across the corpus. The topic label distribution is skewed toward topics relevant to periodic health exams—the 3 most frequent topics (BiomedHistory, PreventiveCare, and MusSkePain) account for more than half of all talk-turns.

Text preprocessing

We applied a number of preprocessing steps to convert the dialog text into a set of tokenized words. We first replaced the patient names and numbers with -NAME- and -NUMBER- tokens to remove potentially identifiable information. After removing symbols, other than a set of punctuation symbols, such as “.”, “?”, “,”, the sentences in each talk-turn were tokenized into a list of words using the standard Python NLTK tokenizer.

For the models that used bag-of-words encoding, the vocabulary included all unigrams and bigram noun phrases that occurred at least 5 times in the corpus, except for 2 sets of stopwords (see Supplementary Appendix A for more information), resulting in a vocabulary of size $V = 14,800$. For our neural network models, the vocabulary consisted of all unigrams, with neither set of stopwords removed (as is customary in neural network models) with a final vocabulary size of 5073 including an unknown token.

Representing talk-turns for classification models

The data for each visit $i$, $1 \leq i \leq 279$, is represented as a sequence of labeled talk-turns $j$, with $L_i$ talk-turns in the $i$th visit, $1 \leq j \leq L_i$. Let $W_{ij}$ and $y_{ij}$ represent the list of word tokens and the topic label, respectively, for the $j$th talk-turn in the $i$th visit. As mentioned earlier, $y_{ij}$ can take values from 1 to 27, corresponding to each of the 27 unique topics. Each word in $W_{ij}$ is encoded as a binary vector (“one-hot encoding”) of length $V$, where $V$ is the vocabulary size. For example, if a word occurs in a talk-turn and the ID in the vocabulary for the word is 10, then the binary/one-hot-encoded vector
becomes a vector of length \( V \), where the 10th entry of the binary word vector is set to 1 and all other entries are set to zero.

For the majority of the classification models we evaluated, the binary word vectors in each talk-turn were aggregated into a single talk-turn vector \( \mathbf{e}_i \) by adding the individual binary word vectors (also known as a bag-of-words encoding) and reweighting using tf-idf weights, a common text preprocessing step that downweights common and uninformative words. The vector \( \mathbf{e}_i \) represents each talk-turn and has dimensionality equal to the size of the vocabulary \( V = 14800 \).

For our neural network models, we used a different representation as follows. We generated one vector \( \mathbf{e}_i \) per talk-turn using a network composed of an embedding layer and a bidirectional set of gated recurrent units (GRUs). The embedding layer, initialized with pretrained GloVe vectors, takes each binary word vector in the talk-turn and maps it to a dense embedded vector representa-
Fully sequential models

To model the sequential dependencies between the topic labels $y_{i,j}$, we used linear-chain CRFs and hidden Markov models (HMMs). The linear-chain CRF is widely used for sequence labeling tasks such as named-entity recognition or part-of-speech tagging—here, we applied it to the problem of predicting the topic label of a talk-turn, given a sequence of talk-turns. The HMMs are constructed by using the output class label probabilities from each of the independent models discussed above. We converted the class probabilities from the classifiers, $p(y_{i,j} = k \mid W_{i,j})$, to emission probabilities, $p(W_{i,j} \mid y_{i,j} = k)$ (needed by the HMM), by using the fact that $p(W_{i,j} \mid y_{i,j} = k) \propto p(y_{i,j} = k \mid W_{i,j})/p(y_{i,j} = k)$, given
that the talk-turn probabilities \( p(W_t j) \) do not depend on a particular topic \( k \). The emission probabilities are combined with the Markov transition probabilities (which can be directly estimated from label sequences in the training data) via the Viterbi algorithm to compute the sequence of topic labels that has the highest joint probability across talk-turns, conditioned on the observed talk-turn words.

We also found that using speaker-specific transition matrices improved the accuracy of our sequential models. Not surprisingly, providers tend to start new topics during a conversation more than the patients do. The figure in Supplementary Appendix B shows the percentage of time that a particular speaker (provider, patient, or other) starts each topic. To incorporate speaker information in the HMM approach, we augmented the standard HMM to use 2 topic transition matrices, 1 for the provider and 1 for the patient or other speakers. Each transition matrix corresponds to the speaker of the state that the HMM is transitioning to (eg, one transition matrix for transitioning to provider and the other for transitioning to all others). The decoding process in the Viterbi algorithm is modified so that it uses the appropriate matrix depending on the speaker.

Another type of sequential model that is entirely neural-network-based does not have direct dependencies between the topic labels, but has bidirectional connections between the hidden states of the talk-turn-level GRUs. The model, which we also refer to as Hier-GRU, has a hierarchical structure having 2 different GRUs, one at the word level to generate talk-turn-level representation, and the other which takes the talk-turn vectors as inputs to predict the output label \( y_j \) for each talk-turn \( j \). Similar to independent GRU, the talk-turn–level GRU output is connected to a fully-connected layer and then a softmax to make prediction.

RESULTS

Experimental methods

We evaluated all models using 10-fold cross-validation and computed evaluation metrics both at the talk-turn level and at the visit level. At the talk-turn level, we computed the classification accuracy by comparing 1) the predicted topic from a model with 2) the human-generated topic for the talk-turn. To obtain results at the visit level, we aggregated the predictions from the individual talk-turns within each visit to generate a visit-level binary-valued prediction vector \( z \) of dimension 27 (the number of topic labels) with 1 for topic label \( k \) if the model predicted topic \( k \) for 1 or more talk-turns in the visit, and 0 otherwise. Using such a vector for each visit, we calculated the accuracy, precision, recall, and F1 scores. The metrics were micro-averaged by globally counting the true positives, false positives, and so on, for all the topic labels.

To evaluate the performance of the classifiers, we compared the results with those of simple baseline models that predict the most common topics. The baseline at the talk-turn level just predicts the most common topic in the corpus, BiomedHistory (as shown in Table 1). At the visit level, the baseline always predicts the set of topic labels that occur in 50% or more of all visits, irrespective of the words within each visit.

Summary of experimental results

Table 2 shows the classification accuracies at the talk-turn level for independent models, windowed models, and sequential models. The most accurate independent model is the GRU and the most accurate windowed model is the windowed LR. The models with sequential information clearly outperform independent models with the Hier-GRU yielding the highest accuracy of 61.77% over all the models. The improvement in accuracy of Hier-GRU and HMM-GRU are both statistically significant, with \( P < .01 \), relative to each of the independent GRU and windowed LR models (using dependent \( t \) tests for paired samples across the 10 folds of cross-validation).

The visit-level evaluation scores are shown in Table 3. Interestingly, the gap in performance (as measured by the micro-averaged F1 score) between independent, windowed, and sequential models is much smaller in the visit-level scores. The primary reason for this is that the independent models tend to have relatively high recall scores, whereas sequential models have relatively high precision scores.

The prediction models also can be evaluated at the level of individual topics to understand variability in prediction accuracy across topics. Precision, recall, and F1 scores were calculated by

| Model            | Talk-turn level accuracy (%) |
|------------------|------------------------------|
| Baseline         | 29.85                        |
| Independent models |                             |
| LR               | 37.00                        |
| SVM              | 36.64                        |
| GRU              | 38.85a                       |
| Window-based models |                             |
| Windowed LR      | 51.12a                       |
| Windowed SVM     | 50.46                        |
| Fully sequential models |                       |
| CRF              | 48.37                        |
| HMM-LR           | 56.89                        |
| HMM-SVM          | 51.52                        |
| HMM-GRU          | 57.60b                       |
| Hier-GRU         | 61.77b                       |

Micro-averaged precision and recall scores are the same accuracy.

CRF: conditional random field; Hier-GRU: hierarchical gated recurrent units; HMM-GRU: hidden Markov model gated recurrent units; LR: logistic regression; SVM: support vector machine;

\( a \)Highest talk-turn level accuracy in each model type.

\( b \)Scores from the two best models.

Table 3. Micro-averaged accuracy, precision, recall, and F1 scores, at the visit level, for different prediction models

| Model            | Visit level (%) |
|------------------|-----------------|
| Model            | Accuracy | Precision | Recall | F1    |
| Baseline         | 72.29     | 73.79      | 62.22  | 67.42 |
| Independent models |         |            |        |       |
| LR               | 75.19     | 67.91      | 84.25  | 75.15a|
| SVM              | 72.45     | 63.40      | 90.40  | 74.50 |
| GRU              | 72.47     | 64.19      | 86.59  | 73.68 |
| Window-based models |        |            |        |       |
| Windowed LR      | 77.28     | 69.82      | 86.47  | 77.21 |
| Windowed SVM     | 79.81     | 75.06      | 82.13  | 78.37a|
| Fully sequential models |     |            |        |       |
| CRF              | 74.19     | 80.43      | 58.42  | 67.64 |
| HMM-LR           | 80.00     | 80.16      | 73.31  | 76.55 |
| HMM-SVM          | 75.21     | 78.70      | 60.90  | 68.63 |
| HMM-GRU          | 79.00     | 74.98      | 79.35  | 77.06 |
| Hier-GRU         | 78.96     | 73.69      | 82.43  | 77.78a|

CRF: conditional random field; Hier-GRU: hierarchical gated recurrent units; HMM-GRU: hidden Markov model gated recurrent units; LR: logistic regression; SVM: support vector machine;

\( a \)Highest F1 score in each model type.
treat each topic label separately at the talk-turn level. Scores from the 2 best-performing models (Hier-GRU and HMM-GRU) are shown in Table 4, sorted by the percentage of each topic. The more common topics (that occur for example in at least 5% of the talk-turns) generally have higher F1 scores. However, there are some less common but highly specific topics, such as the cigarette topic (in 2.43% of talk-turns), that also have relatively high F1 scores.

**DISCUSSION**

Using sequential information across talk-turns in predictive models systematically leads to more accurate predictions, particularly when predicting topic labels for talk-turns. To illustrate this point in more detail, Figure 3 shows sequences of predicted and human-assigned topic labels, for one particular visit, where different colors represent different topic labels. The top plot is from the independent GRU model. The lack of sequential information in the model leads to predictions that are noisy and lack the sequential smoothness of the human-assigned labels (bottom plot). The second plot is from the Hier-GRU model, and the third plot shows the predicted sequence of topics from the Viterbi parse of the HMM-GRU model (ie, the probabilistic predictions from the same GRU model in the top plot, but which are now sequentially smoothed by the HMM transition matrices). It is visually apparent (not only for this visit but for all visits) that the sequential models (second and third) are much more similar to the human labeling (bottom) than the independent model (top).

Figures 4 and 5 provide a more detailed look at portions of the transcript corresponding to Figure 3. We see for example that there are quite a few short talk-turns that have no words with topic-relevant information, such as “Yeah” (talk-turns 5 and 7 in Figure 4 and 222 in Figure 5) and “Okay” (talk-turn 224 in Figure 5). The sequential models are able to use the context information to assign these talk-turns to the same topic as the human labeler. The independent GRU model, however, does not have any context and assigns these talk-turns by default to the topic with the highest marginal probability (BiomedHistory).

While the smoothing in sequential models helps to improve prediction accuracy, it can also produce errors due to oversmoothing. In particular, we found that the HMM-GRU model tends to predict longer topic segments relative to the human-labeled results, as can be seen in talk-turns 100 to 200 of the visit in Figure 3. The human labels contain short bursts of topics GynGenitoUrinary and BiomedHistory that are not detected by the HMM-GRU model. This is further quantified by the visit-level results in Table 3, where the recall scores of the fully sequential models are systematically lower than the independent models, and the reverse for the precision scores.

We also observed that some topics are semantically similar and easily confusable. For example, in Figure 5, from talk-turn 233 to 242, the 2 sequential models predict the topic PhysicalExam, while

| Label             | Precision | Recall | F1     | Precision | Recall | F1     |
|-------------------|-----------|--------|--------|-----------|--------|--------|
| BiomedHistory     | 65.80     | 76.61  | 70.79a | 74.24     | 56.26  | 64.01  |
| PreventiveCare    | 73.34     | 83.41  | 78.05a | 77.98     | 72.49  | 75.13a |
| MusSkePain        | 67.48     | 69.14  | 68.30a | 67.68     | 64.27  | 65.93a |
| VisitFlowMgmt     | 63.64     | 63.58  | 63.61a | 64.97     | 64.54  | 64.76a |
| GynGenitoUrinary  | 67.62     | 54.76  | 60.51  | 72.15     | 57.66  | 64.09a |
| PhysicalExam      | 48.91     | 52.33  | 50.56  | 50.21     | 63.20  | 55.96  |
| Family            | 49.31     | 47.56  | 48.42  | 42.79     | 54.21  | 47.83  |
| HealthCareSystem  | 37.81     | 28.85  | 32.73  | 46.10     | 39.49  | 42.54  |
| WorkLeisureActivity| 47.26    | 46.20  | 46.72  | 52.90     | 53.14  | 53.02  |
| TestDiagnostics   | 49.67     | 32.43  | 39.24  | 37.24     | 52.88  | 43.70  |
| Cigarette         | 71.38     | 85.84  | 77.94a | 81.38     | 72.73  | 76.81a |
| Weight            | 49.54     | 52.60  | 51.02  | 47.20     | 46.77  | 46.98  |
| Other             | 21.42     | 9.59   | 13.25  | 14.45     | 15.20  | 14.81  |
| Depression        | 50.82     | 64.27  | 56.76  | 45.71     | 44.24  | 44.96  |
| DizzyDentHearVision| 23.08    | 14.01  | 17.44  | 46.96     | 50.83  | 48.82  |
| Medication        | 38.97     | 27.09  | 31.96  | 22.46     | 64.29  | 33.29  |
| Exercise          | 56.94     | 59.23  | 58.06  | 45.42     | 67.01  | 54.14  |
| SmallTalk         | 20.61     | 18.70  | 19.61  | 32.79     | 31.09  | 31.92  |
| GeneralAnxieties  | 32.51     | 18.63  | 23.68  | 16.36     | 32.86  | 21.84  |
| MDLife            | 53.81     | 14.19  | 19.99  | 12.79     | 11.69  | 12.22  |
| Diet              | 27.15     | 19.62  | 22.78  | 27.52     | 57.10  | 37.14  |
| Alcohol           | 56.08     | 36.12  | 43.94  | 52.19     | 64.91  | 57.86  |
| Sleep             | 13.40     | 2.33   | 3.97   | 29.80     | 40.32  | 34.27  |
| TherapeuticIntervention| 0.00 | 0.00   | 0.00   | 2.04      | 5.30   | 2.95   |
| RiskyBehavior     | 33.90     | 5.87   | 10.00  | 44.87     | 41.06  | 42.88  |
| OtherAddictions   | 0.00      | 0.00   | 0.00   | 23.55     | 41.40  | 30.02  |
| Age               | 0.00      | 0.00   | 0.00   | 13.23     | 20.00  | 15.93  |

The rows are sorted by the percentage of talk-turns of each topic. In general, the more frequently discussed topics have higher F1 scores. Hier-GRU: hierarchical gated recurrent units; HMM-GRU: hidden Markov model gated recurrent units.

*Highest F1 scores.
Figure 3. Sequences of color-coded topic labels for one of the visits in our dataset. The upper plot shows the predicted topic labels from an independent model, and the center 2 plots show those from fully sequential models. The lower plot corresponds to the human-coded labels. The segments for the MusSkePain topic (1) had lengths of 23 talk-turns for hierarchical gated recurrent units (Hier-GRU), 27 for hidden Markov model gated recurrent units (HMM-GRU), and 26 for human labeled. Similarly, (2) the PreventiveCare segments had lengths 30, 28, and 27, and (3) the TestDiagnostics segment had lengths 10, 31, and 22 in talk-turns, respectively, for Hier-GRU, HMM-GRU, and human labeled. See Supplementary Appendix E for the boxplots of topic segment lengths for the 4 sequences of labels.

| Talk Turn ID | GRU Label       | Hier-GRU Label  | HMM-GRU Label    | Human Label     | Transcript Text                                                                 |
|--------------|-----------------|-----------------|------------------|-----------------|---------------------------------------------------------------------------------|
| 2            | VisitFlowMgmt   | VisitFlowMgmt   | VisitFlowMgmt    | VisitFlowMgmt   | MD: Hello, -name-. How are you today?                                            |
| 3            | VisitFlowMgmt   | VisitFlowMgmt   | VisitFlowMgmt    | VisitFlowMgmt   | PT: Good, thanks. Good.                                                          |
| 4            | VisitFlowMgmt   | VisitFlowMgmt   | VisitFlowMgmt    | VisitFlowMgmt   | MD: Good, good, good.                                                            |
| 5            | BiomedHistory   | VisitFlowMgmt   | VisitFlowMgmt    | VisitFlowMgmt   | PT: Yeah.                                                                        |
| 6            | VisitFlowMgmt   | VisitFlowMgmt   | VisitFlowMgmt    | VisitFlowMgmt   | MD: Doing your, uh, your physical today?                                         |
| 7            | BiomedHistory   | VisitFlowMgmt   | VisitFlowMgmt    | VisitFlowMgmt   | PT: yeah.                                                                       |
| 8            | VisitFlowMgmt   | VisitFlowMgmt   | VisitFlowMgmt    | VisitFlowMgmt   | MD: Okay. Very good. Um, well, uh, let's go over things, then, um, have some specific things you wanted to go over today. |
| 9            | BiomedHistory   | BiomedHistory   | MusSkePain       | VisitFlowMgmt   | PT: Uh-uh.                                                                      |
| 10           | BiomedHistory   | MusSkePain      | MusSkePain       | MusSkePain      | MD: And well, so a problem with your foot there?                                |
| 11           | BiomedHistory   | MusSkePain      | MusSkePain       | MusSkePain      | PT: Right, the left one.                                                         |
| 12           | BiomedHistory   | BiomedHistory   | MusSkePain       | MusSkePain      | MD: Okay. What's been happening?                                                |
| 13           | BiomedHistory   | MusSkePain      | MusSkePain       | MusSkePain      | PT: Not much. I think I had a little fracture in it, and then, but the, uh, little toe and the one next to it still feel a little, little numbness in it. |
| 14           | BiomedHistory   | MusSkePain      | MusSkePain       | MusSkePain      | MD: Really?                                                                     |
| 15           | BiomedHistory   | MusSkePain      | MusSkePain       | MusSkePain      | PT: Right. But that's been, like, over a couple months ago.                      |
| 16           | MusSkePain      | MusSkePain      | MusSkePain       | MusSkePain      | MD: How did you injure your foot?                                               |

Figure 4. The beginning part of the visit shown in Figure 3. Each talk-turn is presented with predicted labels from 3 different models (independent gated recurrent units [GRU], hierarchical gated recurrent units [Hier-GRU], and hidden Markov model gated recurrent units [HMM-GRU]) and the human-coded labels. For the short talk-turns the BiomedHistory topic label is predicted quite often by the independent GRU, while the 2 other models produce label sequences that are more similar to human-coded labels. MD: medical doctor or provider; PT: patient.
the human labeled MusSkePain—from the corresponding transcript text either prediction seems reasonable. Similarly, from talk-turn 249 to 254 the HMM-GRU predicts SmallTalk, while the human labeled WorkLeisure—from the text the corresponding talk-turns appear to be a mixture of both. We also found other examples across the corpus where the model frequently gets confused among small groups of related topics (eg, GeneralAnxieties and Depression; Weight and Diet). The full confusion matrices are shown in Figure 6. There is inevitably a subjective aspect to the human labeling, suggesting that there is likely to be a performance ceiling in terms of the accuracy of any algorithm relative to human labels on this data.

From the topic-specific results in Table 4 we can see that while the predictions are relatively accurate for some topics, for others (eg, Age, TherapeuticIntervention, OtherAddictions, Other, MDLife), the scores are quite low. The broad nature of these topics is a likely contributor to the low accuracies, but the relative lack of training data per topic may also be another contributing factor. These topics account for roughly 1% (or less) of talk-turns in the corpus and many of these talk-turns are relatively short with little topical content, leading to relatively less signal, particularly for training neural network models. One possible approach to improve accuracy would be to incorporate additional external information relevant to these topics, such as incorporating lists of relevant words from ontological sources such as Unified Medical Language System into the training of prediction models, or adding relevant information from sources such as physician or specialist notes.
Patient-provider communication is an essential component of health care. In this context, prediction models that annotate patient-provider transcripts can in principle provide both useful information about the nature of topics discussed in specific conversations as well as contribute to a broader understanding of patient-provider communication. We have demonstrated that machine learning methods show promise for building models that can automatically predict discussion topics in dialog at the talk-turn and visit level. In particular, using a large real-world patient-provider dialog corpus, we investigated the performance of a variety of classification models including LR, SVMs, feedforward neural network model with GRUs, hierarchical GRUs, CRFs, and HMMs. We found that sequential models (eg, Hier-GRU and HMM-GRU) are more accurate compared with nonsequential models for predicting topic labels for talk-turns. In addition, we found that semantic similarity of discussion topics can be a significant contributor to prediction error.

While additional research and model improvement is needed, our results show promise for a number of medical topics that are critical quality indicators in primary care (eg, cigarette smoking, pain). Potential applications might include exploring systems that incorporate prior information from a list of problem areas or prior diagnoses found in the medical record. For example, the presence or absence of smoking cessation counseling during primary care encounters may inform population health management programs aimed at helping patients quit smoking. Deployment of systems such as this in real-world primary care may also be useful for obtaining the scale of data needed to improve model performance.

**FUNDING**
This work was supported by Patient-Centered Outcomes Research Institute number ME-1602-34167. Zac E Imel is the one who received the PCORI grant as an overall PI of the project.

**AUTHOR CONTRIBUTIONS**
All authors assisted in the writing of the manuscript. JP carried out the analyses and was responsible for writing the Materials and Methods and Results sections, with assistance from DK, RLL, SS, and EK-T. PS supervised the analyses. MT and PK were responsible for data management and cleaning along with JP. JEL and MT-S were responsible for collection of source data and description of data and human labeling. ZI, MT-S, PS, and DCA jointly conceived of the project, and ZEI, KM, and PK wrote the introduction.

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

**ACKNOWLEDGMENTS**
We gratefully acknowledge Mary Ann Cook for helpful comments and feedback, Keith Gunnesson and Taylor Shuman for assistance in formatting of the data, and Abhishek Jindal for useful discussions on sequential models.

**CONFLICT OF INTEREST STATEMENT**
ZEI, DCA, and MT are co-founders with equity stakes in a technology company, Lyssn.io, focused on tools to support training, supervision, and quality assurance of psychotherapy and counseling. Also, within the past five years, PS has received research funding from SAP, Qualcomm, eBay, Google, Adobe, and Cylance, and has served as a consultant for Toshiba, Samsung, General Motors, and First American.

**REFERENCES**
1. Simon HB. The write stuff: how good writing can enhance patient care and professional growth. Am J Med 2013; 126 (6): 467–71.
2. Hewett DG, Watson BM, Gallois C, et al. Communication in medical records: intergroup. Language and patient care. J Lang Soc Psychol 2009; 28 (2): 119–38.
3. Curtis JR, Sathiratanaachewin S, Starks H, et al. Using electronic health records for quality measurement and accountability in care of the seriously ill: opportunities and challenges. J Palliat Med 2018; 21: 552–60.

4. Hsiao WC, Yntema DB, Braun P, et al. Measurement and analysis of in-service work. JAMA 1988; 260 (16): 2361–70.

5. White A, Danis M. Enhancing patient-centered communication and collaboration by using the electronic health record in the examination room. JAMA 2013; 309 (22): 2327–8.

6. Singh K, Meyer SR, Westfall JM. Consumer-facing data, information, and tools: self-management of health in the digital age. Health Aff (Millwood) 2019; 38 (3): 352–8.

7. Shachak A, Reis S. The impact of electronic medical records on patient-doctor communication during consultation: a narrative literature review. J Eval Clin Pract 2009; 15 (4): 641–9.

8. Ventres W, Kooienga S, Vuckovic N, et al. Physicians, patients, and the electronic health record: an ethnographic analysis. Ann Fam Med 2006; 4 (2): 124–31.

9. Smiley CA. On presence: a tale of two visits. December 29, 2016. https://catalyst.nejm.org/electronic-health-record-tale-two-visits/ Accessed June 27, 2019.

10. Tai-Seale M, Olson CW, Li J, et al. Electronic health record logs indicate that physicians split time evenly between seeing patients and desktop medicine. Health Aff (Millwood) 2017; 36 (4): 655–62.

11. Arndt BG, Beasley JW, Watkinson MD, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. Ann Fam Med 2017; 15 (5): 419–26.

12. Shanafelt TD, Hasan O, Dyrbye LN, et al. Changes in burnout and satisfaction with work-life balance in physicians and the general US working population between 2011 and 2014. Mayo Clin Proc 2015; 90 (12): 1600–13.

13. Tai-Seale M, Dillon E, Yang Y, et al. Physicians’ well-being linked to in-basket messages generated by algorithms in electronic health records. Health Aff (Millwood) 2019; 38: 1075–78.

14. Friedberg MW, Chen PG, Van Buum KR, et al. Factors affecting physician professional satisfaction and their implications for patient care, health systems, and health policy. Rand Health Q 2014; 3: 1.

15. Thielke S, Hammond K, Helbig S. Copying and pasting of examinations within the electronic medical record. Int J Med Inform 2007; 76 Suppl 1: S122–8.

16. Hammond KW, Helbig ST, Benson CC, et al. Are electronic medical records trustworthy? Observations on copying, pasting and duplication. AMIA Annu Symp Proc 2003; 2003: 269–73.

17. Chiu C-C, Tripathi A, Chou K, et al. Speech recognition for medical conversations. Proc Interspeech 2018; 2018: 2972–6.

18. Rajkomar A, Kannan A, Chen K, et al. Automatically charting symptoms from patient-physician conversations using machine learning. JAMA Intern Med 2019; 179 (6): 836–8.

19. Elwyn G, Barr PJ, Grande SW. Patients recording clinical encounters: a collaborative approach to the study of medical discourse. Soc Sci Med 1994; 39 (7): 955–65.

20. Tai-Seale M, Hatfield LA, Wilson C, et al. Periodic health examinations and missed opportunities among patients likely needing medical care. Am J Manag Care 2016; 22: e350–7.

21. Caperton DD, Atkins DC, Imel ZE. Rating motivational interviewing fidelity from thin slices. Psychol Addict Behav 2018; 32 (4): 434–41.

22. Levinson W, Lesser CS, Epstein RM. Developing physician communication skills for patient-centered care. Health Aff (Millwood) 2010; 29 (7): 1310–8.

23. Hoerger M, Epstein RM, Winters PG, 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 2013; 13 (1): 188.

24. Mork JG, Bodenreider O, Demner-Fushman D, et al. Deriving assertions from clinical note de-identification and its impact on information extraction. J Am Med Inform Assoc 2011; 18 (5): 544–51.

25. Mays VM, Kohlmann LE, Fingerhut L, et al. A study of the effectiveness of machine learning methods for classification of clinical interview fragments into a large number of categories. J Biomed Inform 2016; 62: 21–31.

26. van Osch M, van Dulmen S, van Vliet L, et al. Specifying the effects of physician’s communication on patients’ outcomes: a randomised controlled trial. Patient Educ Couns 2017; 100 (8): 1482–9.

27. Hojat M. The interpersonal dynamics in clinician–patient relationships. In: Hojat M, ed. Empathy in Health Professions Education and Patient Care. Cham, Switzerland: Springer International Publishing; 2016: 129–50.

28. Tai-Seale M, McGuire TG, Zhang W. Time allocation in primary care office visits. Health Serv Res 2007; 42 (5): 1871–94.

29. Foo PK, Frankel RM, McGuire TG, et al. Patient and physician race and the allocation of time and patient engagement efforts to mental health discussions in primary care. J Ambul Care Manage 2017; 40 (3): 246–56.

30. Tai-Seale M, McGuire T, Colenda C, et al. Two-minute mental health care for elderly patients: inside primary care visits. J Am Geriatr Soc 2007; 55 (12): 1903–11.

31. Etom DT, Ridgeway JL, Linzer M, et al. Healthcare provider relational quality is associated with better self-management and less treatment burden in people with multiple chronic conditions. Patient Prefer Adherence 2017; 11: 1635–46.

32. Charon R, Greene MG, Adelman RD. Multi-dimensional interaction analysis: a collaborative approach to the study of medical discourse. Soc Sci Med 1988; 260 (16): 2361–70.

33. Tai-Seale M, McGuire TG, Zhang W. Time allocation in primary care office visits. Health Serv Res 2007; 42 (5): 1871–94.

34. Foo PK, Frankel RM, McGuire TG, et al. Patient and physician race and the allocation of time and patient engagement efforts to mental health discussions in primary care. J Ambul Care Manage 2017; 40 (3): 246–56.

35. Tai-Seale M, McGuire T, Colenda C, et al. Two-minute mental health care for elderly patients: inside primary care visits. J Am Geriatr Soc 2007; 55 (12): 1903–11.

36. Hoerger M, Epstein RM, Winters PG, 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 2013; 13 (1): 188.

37. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc 2011; 18 (5): 544–51.

38. Mayfield E, Laws MB, Small K, et al. Automatically annotating topics in clinical note de-identification and its impact on information extraction. J Am Med Inform Assoc 2013; 20 (1): 84–94.

39. Roberts K, Harabagiu SM. A flexible framework for deriving assertions from electronic medical records. J Am Med Inform Assoc 2011; 18 (5): 568–73.

40. Delegere L, Molnar K, Savova G, et al. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. J Am Med Inform Assoc 2013; 20 (1): 84–94.

41. Mayfield E, Laws MB, Wilson IB, et al. Automating annotation of information-giving for analysis of clinical conversation. J Am Med Inform Assoc 2014; 21 (1): e122–8.

42. Beach MC, Saha S, Korthuis PT, et al. Patient-provider communication differs for black compared with white HIV-infected patients. AIDS Behav 2011; 15 (4): 805–11.

43. Laws MB, Beach MC, Lee Y, et al. Provider-patient adherence dialogue in HIV care: results of a multisite study. AIDS Behav 2013; 17 (1): 148–59.

44. Mayfield E, Laws MB, Wilson IB, et al. Automating annotation of information-giving for analysis of clinical conversation. J Am Med Inform Assoc 2014; 21 (1): e122–8.

45. Beach MC, Saha S, Korthuis PT, et al. Patient-provider communication differs for black compared with white HIV-infected patients. AIDS Behav 2011; 15 (4): 805–11.

46. Laws MB, Beach MC, Lee Y, et al. Provider-patient adherence dialogue in HIV care: results of a multisite study. AIDS Behav 2013; 17 (1): 148–59.

47. Kotov A, Hasan M, Carcone A, et al. Interpretable probabilistic latent variable models for automatic annotation of clinical text. AMIA Annu Symp Proc 2015; 2015: 785–94.

48. Beach MC, Saha S, Korthuis PT, et al. Patient-provider communication differs for black compared with white HIV-infected patients. AIDS Behav 2011; 15 (4): 805–11.

49. Wallace BC, Laws MB, Small K, et al. Automatically annotating topics in transcripts of patient-provider interactions via machine learning. Med Decis Mak 2014; 34 (4): 503–12.
49. Gaut G, Steyvers M, Imel ZE, et al. Content coding of psychotherapy transcripts using labeled topic models. IEEE J Biomed Health Inform 2017; 21 (2): 476–87.

50. Bird S, Klein E, Loper E. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. Newton, MA: O’Reilly Media, Inc.; 2009.

51. Pennington J, Socher R, Manning CD. GloVe: global vectors for word representation. In: Proceedings of Empirical Methods in Natural Language Processing (EMNLP); 2014: 1532–43.

52. Goldberg Y. Neural Network Methods in Natural Language Processing (Synthesis Lectures on Human Language Technologies). San Rafael, CA: Morgan & Claypool; 2017.

53. Lafferty JD, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. Burlington, MA: Morgan Kaufmann; 2001: 282–9.