Published in final edited form as:

Med Decis Making. 2014 May ; 34(4): 503–512. doi:10.1177/0272989X13514777.

Automatically Annotating Topics in Transcripts of Patient-Provider Interactions via Machine Learning

Byron C. Wallace, PhD¹, M. Barton Laws, PhD¹, Kevin Small, PhD², Ira B. Wilson, MD, MSc¹, and Thomas A. Trikalinos, MD, PhD¹

¹Dept of Health Services, Policy and Practice, 121 South Main Street, Providence, RI, 02903, USA

²National Institutes of Health, 2 Center Drive, Bethesda, MD 20892, USA

Abstract

Background—Annotated patient-provider encounters can provide important insights into clinical communication, ultimately suggesting how it might be improved to effect better health outcomes. But annotating outpatient transcripts with Roter or General Medical Interaction Analysis System (GMIA) codes is expensive, limiting the scope of such analyses. We propose automatically annotating transcripts of patient-provider interactions with topic codes via machine learning.

Methods—We use a conditional random field (CRF) to model utterance topic probabilities. The model accounts for the sequential structure of conversations and the words comprising utterances. We assess predictive performance via 10- fold cross-validation over GMIA-annotated transcripts of 360 outpatient visits (over 230,000 utterances). We then used automated in place of manual annotations to reproduce an analysis of 116 additional visits from a randomized trial that used GMIA to assess the efficacy of an intervention aimed at improving communication around antiretroviral (ARV) adherence.

Results—With respect to six topic codes, the CRF achieved a mean pairwise kappa compared with human annotators of 0.49 (range: 0.47, 0.53) and a mean overall accuracy of 0.64 (range: 0.62, 0.66). With respect to the RCT re-analysis, results using automated annotations agreed with those obtained using manual ones. According to the manual annotations, the median number of ARV-related utterances without and with the intervention was 49.5 versus 76, respectively (paired sign test $p=0.07$). Using automated annotations, the respective numbers were 39 versus 55 ($p=0.04$).

Limitations—While moderately accurate, the predicted annotations are far from perfect. Conversational topics are intermediate outcomes; their utility is still being researched.

Corresponding author: Byron C. Wallace, Dept of Health Services, Policy and Practice, 121 South Main Street, Brown University, Providence, RI, 02903, byron_wallace@brown.edu, Phone: 401-863-6421.

Competing Interests Statement

We declare no competing interests.

Contributorship Statement

BCW, MBL, IBW and TT conceived of project. BCW and KS wrote code to process transcripts and train Conditional Random Fields (CRFs). BCW wrote code to evaluate classifier results. BCW wrote first draft of manuscript; all authors (BCW, KS, MBL, IBW, TT) participated in editing subsequent drafts.

Conclusions—This foray into automated topic inference suggests that machine learning methods can classify utterances comprising patient-provider interactions into clinically relevant topics with reasonable accuracy.

Introduction

Patient-provider communication is a critical component of health-care.¹ Evidence suggests that the patient-provider relationship, and specifically the degree of patient-centeredness in communication, affects patient “enablement, satisfaction, and burden of symptoms”.² Several studies have reported an association between physician-patient communication and health outcomes,^{3–5} and a systematic review of studies investigating patient-provider communication concluded that several verbal behaviors are associated with health outcomes.⁶

The many extant systems for coding and analyzing patient-provider communication have produced a considerable body of literature.^{7,8} These systems are generally based on defining various provider and patient verbal behaviors and counting their frequencies. Analyses using this simple approach have produced substantial insight into provider and patient role relationships, and have described associations between attributes of the relationship and a variety of patient-relevant outcomes.

We focus on patient-provider interactions annotated using the General Medical Interaction Analysis System (GMIAS).⁹ The GMIAS analyzes all of the utterances comprising a patient-provider interaction. It draws on Speech Act Theory^{10–12} to characterize the social acts embodied in each utterance and also classifies their content into condition-specific topic typologies consistent with the widely used Roter Interactional Analysis (RIAS) framework^{13,14} but with much greater specificity (we provide further description in the *GMIAS Topic Coding* subsection of *Methods* and in the Appendix). GMIAS has been used to: characterize interaction processes in physician-patient communication regarding antiretroviral adherence in the context of an intervention trial¹⁵; analyze communication about sexual risk behavior¹⁶; assess the association of visit length with constructs of patient-centeredness¹⁷; describe provider-patient communication regarding ARV adherence compared with communication about other issues¹⁸; and to measure the effectiveness of interventions for improving communication around patient adherence to antiretrovirals.¹⁹

Analysis of outpatient visits coded with salient clinical topics can provide valuable insights into patient-provider communication, but it is a tedious and costly exercise. Although transcribing recorded communications and manually segmenting them into utterances is relatively inexpensive, annotating the utterances is time consuming, and requires highly trained personnel. Because of the cost, large-scale analyses of physician-patient interactions are nontrivial and often impractical. Tools and methods that reduce annotation costs are therefore needed. This work represents an effort to realize this aim: specifically, we use machine learning methods to automatically annotate transcribed and segmented transcripts with GMIAS topic codes. Using an automated statistical approach to label interactions has the potential to drastically reduce annotation costs. Even if less accurate than human annotations, large-scale automated annotation of patient-provider interactions would provide

data to explore potential associations between measureable aspects of patient-provider communication and patient-relevant outcomes. Furthermore, this technology might be used as a screening tool to quickly identify patient-provider encounters that are of interest for a particular research question; or even in quality improvement interventions, to provide feedback to physicians about how they allocate time in visits.

We present a model that uses machine learning to automatically annotate utterances with the topics to which they belong. The input to the model is a patient-provider interaction (recorded, transcribed and segmented into utterances), and its output is GMIAS clinical-topic code predictions for each utterance in the interaction. We describe the model's classification performance as compared with human annotators, and explore a "proof-of-principle" application of the semi-automated system by re-analyzing data from a randomized crossover trial that used sociolinguistic analyses of patient-provider discussions as a secondary outcome.

Methods

We first introduce the GMIAS annotation schema. We then describe the machine learning technologies for automating GMIAS annotations, and the evaluation of their performance. The online Appendix includes a more technical description of the model (also accessible at <http://www.cebm.brown.edu/static/mdm-appendix.pdf>; last accessed on 09/23/2013).

GMIAS Topic Coding

The unit of analysis in the GMIAS is the utterance, which is defined as a completed speech act (see Laws et al.⁹ for details and Searle¹⁰, Austin¹² and Habermas¹¹ for a substantive introduction to Speech Act Theory). The GMIAS annotates each utterance using two types of codes, corresponding to speech acts and topics. Speech acts capture the social act embodied in an utterance: an example is a *commissive*, which refers to a speech act in which the speaker commits to some future action. We do not further consider speech acts in this work; instead we focus on topics. Each utterance is assigned to one from a set of *a priori* defined clinically relevant topics (one of which is specific to ARV). We focus on topic codes because they are more immediately interpretable than speech act codes. Table 1 lists illustrative examples of utterances coded with GMIAS topics. We use the GMIAS for this work, but the proposed approach is equally suitable to similar medical topic coding systems, e.g., the Roter system.¹⁴

The GMIAS defines both general clinical topics (e.g., *biomedical* and *socializing*) and topics specific to ARV therapy and medication adherence. These codes are hierarchically structured, with increasing specificity at deeper nested levels. The complete topic hierarchy is provided in the Appendix. GMIAS covers such topics as *biomedical*, *psychosocial*, *logistics*, and *socializing*. GMIAS instantiations for the study of other conditions can add or modify branches of the topic code tree as needed. Table 2 provides an excerpt from an annotated transcript for illustrative purposes.

In this work we restrict ourselves to predicting high-level codes, that is, topic categories at the first level in the hierarchy defined by GMIAS (each of which subsumes more specific

topic codes). For the purposes of interpretation, we have merged some of the original GMIAS codes (see Appendix Table 1 for details on this mapping). Ultimately, each utterance is predicted to belong to one of the following 6 topics: *biomedical*, *psychosocial*, *logistics*, *socializing*, *ARV*, and *missing/other*.

These topics are described as follows. Clinical observations and diagnostic conclusions are coded as *biomedical* utterances. The *psychosocial* topic includes such issues as substance abuse, recovery, employment, relationships, and criminal justice involvement. Utterances dealing with *logistics* concern the business of providing medical care such as appointments, referrals, record retrieval, prescription refills unrelated to adherence, and studies and trials. The business of conducting the physical examination is a sub-division of *logistics*, and consists largely of physician directives such as “take a deep breath.” *Socializing* refers to casual conversation unrelated to the business of the medical visit, and to social rituals such as greetings. *ARV* applies to utterances that have to do with ARV treatment and adherence. *Missing/other* is a catchall category for utterances that have no other clear designation (e.g., inaudible utterances).

While less granular than using the entire GMIAS topic hierarchy, high-level codes can nonetheless provide insight into clinical interactions. Specifically, they allow one to: quantify which topics occupy patient-provider conversations; describe patterns of topic frequency (e.g., during which point in a visit are treatment logistics usually addressed?); and to describe prototypical sequences of topic transitions (“motifs” of interactions). We also note that we conducted a sensitivity analysis using the original, unmerged high-level topic categories, and the results (not shown) were very similar to those reported below.

The context of the patient-provider interaction and the nearby utterances are generally helpful (and sometimes necessary) for categorizing utterances by clinical topic code. The examples in Table 1 are emblematic of their topics, in that they are easy to classify even without knowledge of the preceding or following utterances, or the context of the interaction in general. Thus automated methods may be able to correctly annotate such utterances by topic simply on the basis of the words they include. But an automated system will need also to handle less clear-cut utterances that can be classified only after considering the context of the interaction. For example, the topic of the utterance “*because you take the -uh-*” is only discernable from the context (in this real example the topic was *ARV*: the provider was interrupted by the patient in the midst of asking about the time of day the patient typically takes the medication). The annotation model will also need to handle very brief utterances (e.g., “ok”, “yes”) whose topic will similarly be determined by context, and other varieties of complex utterances that naturally arise in conversational exchanges.

We have reported evidence for the reliability and validity of the GMIAS elsewhere.²⁰ Briefly, inter-rater kappa for agreement of GMIAS annotations between the developer of the GMIAS (MBL) and 3 other coders was 0.80.²⁰ Agreement was even higher at the high level codes used in this work. (We elaborate on limitations of restricting annotation to this top-level in the Discussion.)

Machine Learning Methods

In brief, we leverage a statistical machine learning model known as a (linear chain) conditional random field (CRF).²¹ CRFs are particularly useful for classifying each instance in a sequence of instances (here, utterances) into mutually exclusive and exhaustive categories (here, the high level topic codes), based on their features (here, the words they include). Successive utterances are more likely to belong to the same topic than utterances that are further away in the patient-provider interaction, as people tend to stay on topic. CRFs capitalize on this autocorrelation structure, i.e., exploit correlations between words and topics, and common patterns of topic sequences. CRFs have previously been used for many sequence prediction tasks,²² and specifically for dialogue act tagging.²³ As far as we aware, this is the first application of such models to patient-provider interactions.

A CRF assumes that the sequence of topics comprising each patient-provider interaction is generated by an underlying latent process. This process governs the succession of topics, rendering certain sequences more likely (e.g., adjacent utterances are likely to share a topic), given the observed utterances. Intuitively, the linear chain first order CRFs that we use here operate over pairs of adjacent utterances, and specifically the associations between the (predicted) topic of the current utterance, the words it comprises, and the (predicted) topics of the immediately preceding and following utterances. The parameters that describe these associations are calculated taking the whole sequence of utterances into account.

As an example, suppose that the current utterance is “hello doc” and belongs to the topic *socializing*. Further assume that it is preceded by another utterance whose topic was also *socializing*. The model calculates the (pseudo) likelihood of the co-occurrence of $\langle \text{current}=\textit{socializing}, \text{previous}=\textit{socializing} \mid \text{word}=\textit{hello} \rangle$ and of $\langle \text{current}=\textit{socializing}, \text{previous}=\textit{socializing} \mid \text{word}=\textit{doc} \rangle$, together with the likelihood of all other observed similar co-occurrences, and encodes it in its parameters. This occurs during the fitting of the parameters of the CRF, or, equivalently, the training of the model. Given a new patient-provider interaction, the CRF estimates for each utterance the probability that it belongs to each topic based on the fitted parameters. We used the CRF implementation in the Mallet toolkit, an open-source natural language processing library.²⁴

Datasets and Their Annotation

We used transcribed and annotated patient-provider interactions from the Enhancing Communication and HIV Outcomes (ECHO) study,¹⁶ and from a randomized cross-over trial of an intervention aimed at improving physicians’ knowledge of patients’ ARV adherence.¹⁹ We used 360 randomly selected patient-provider interactions from the ECHO dataset for developing and validating the machine learning algorithms in the current work (the remaining 55 eligible interactions from the ECHO dataset are kept aside for independent future validation of currently ongoing research). In addition, we demonstrate a proof-of-concept re-analysis of the cross-over trial (n=116 datasets) based on automated topic predictions rather than human annotations, as described in a later section.

ECHO was designed to assess the role of the patient-provider relationship in explaining racial/ethnic disparities in HIV care.¹⁶ Study subjects were HIV care providers and their

patients at four HIV outpatient care sites in separate regions of the United States. Patient provider interactions were annotated with GMIAS. In the cross-over trial the intervention was a report given to the physician before a routine office visit that contained information regarding the patients' ARV usage and their beliefs about ARV therapy. To explore the efficacy of this intervention, 58 paired (116 total) audio recorded visits were annotated with GMIAS codes.¹⁹ In both studies a professional transcription service or a research assistant transcribed audio recordings of visits, and a research assistant reviewed the resulting transcripts for accuracy. Research assistants then coded the transcripts using the GMIAS.

Text Encoding

Utterances (specifically, the words in them) have to be encoded in mathematical form to be analyzed with CRFs. We used standard “bag-of-words”²⁵ text encoding: we first created a list of all the words (the “bag-of-words”) that were observed in the 360 transcribed interactions of the development dataset (the list's length, L , runs in the many thousands). Then any word in any utterance can be represented by a vector that has $L-1$ zeros and a single one: the position of the one indexes the word in the “bag-of-words”. A group of words such as an utterance can be similarly represented as a vector of ones (corresponding to the words in the utterance) and zeros (in all other positions).

More specifically, we removed all words from the corpus that were not observed at least 3 times. We did not remove so-called “stop words” (such as “the” and “a”) or interjections and disfluencies (“uh”, “um”) because they were deemed to be useful discourse markers. We mapped all words to lowercase (e.g., “Hi” to “hi”). We did not perform lemmatization (lemmatization involves mapping terms to canonical forms, e.g., “walking” to “walk”). We assumed all words were demarcated by surrounding whitespace, and did not attempt to further split terms (e.g., we did not split hyphenated words). We experimented with encoding the presence not only of single words (unigrams), but also of adjacent word pairs (bigrams). However, adding bigrams increased model fitting (training) time and did not improve performance (results not shown). Finally, we added an extra feature to each utterance to indicate the speaker identity (patient or provider).

Experimental Setup for Evaluation

We evaluated our model via cross-fold validation, a standard method of evaluating classifier performance. Specifically, we randomly split the 360 visits into 10 folds comprising 36 visits each, without stratifying per provider. We then ran 10 experiments, holding each of the 10 folds out in turn and estimating the CRF parameters using the other 324 annotated visits (9 training folds). We evaluated performance over the held-out set in each fold. (Using hold-out test sets avoids inflating results due to over-fitting.) We report averages and ranges of performance metrics across these folds.

Global and Topic-Specific Evaluation Metrics

For each of the ten test folds, we measured global CRF prediction performance using the human annotations as a reference standard. For each utterance the CRF returned the probability that it belongs to each of the 6 topics. We took the topic with the highest probability to be the CRF prediction. We calculated the model's overall accuracy (the

proportion of utterances that were correctly predicted), and the kappa for agreement between the model predictions and the human annotations (reference standard). Overall accuracy is conceptually simple, but depends on the prevalence of each topic, and can be difficult to interpret. The kappa for agreement quantifies agreement beyond what is expected by chance.

Both overall accuracy and kappa are global metrics that aggregate performance over all topics. We also consider performance with respect to each topic as quantified by three metrics commonly used in the evaluation of classifiers in the natural language processing and information retrieval communities²⁶: recall (sensitivity, the proportion of all utterances belonging a given topic that are correctly predicted as such); precision (the positive predictive value, i.e., the proportion of topic predictions that are correct); and the F-measure, which is the harmonic mean of recall and precision.

Congruence of Observed and Predicted Frequencies of Topics Over the Course of Interactions

It may be difficult to interpret the above metrics in terms of their implications for analyzing provider-patient transcripts. One question, for example, is whether the model is sufficiently accurate to reveal meaningful high-level patterns in interactions. To see if the model indeed captures the global structure of visits, we can compare the predicted versus the true (average) frequencies of topics across time. This average interaction can be interpreted as a “prototypical” visit, and as such we might expect to observe certain trends. For example, one would expect most socializing to take place at the start of visits, and also that logistics probably dominates the end of visits.

To make interactions of varying length (in terms of utterances) comparable, we broke each interaction up into ten equal parts (i.e., we split them into deciles). This discretization “aligns” the earlier, middle and later sections of interactions that have variable length. A crude quantification of a patient-provider interaction is then to count up the relative frequencies of the topics comprising each decile, and to average these across all visits (spanning all 10 test folds). For each decile, we plot these empirical average frequencies together with the corresponding average frequencies based on model predictions. If the predictions track the empirical topic frequencies, then it might suggest that, while imperfect, the automated annotations capture high-level structures in the interactions.

Re-analysis of The Results of a Cross-over Trial Based on Model Predictions

We also evaluated performance by considering the model predictions on 116 independent cases that were collected for the randomized cross-over trial.¹⁹

From the cross-over trial we have one annotated visit for each provider in which they were given the report beforehand (intervention) and one in which they were not (control). The order in which the intervention was administered was randomized; half received the intervention before the first visit, and half received it before a third visit (only the first and third visits were audio recorded and annotated). Those that received it in the third visit had not been provided the report before any prior visits (i.e., the first or second).¹⁹

The intervention increased adherence-related dialogue but did not ultimately improve patient ARV adherence. We re-analyzed the trial using automated rather than manual topic annotations. Specifically, we trained a model over the aforementioned 360 annotated visits from the ECHO study and then used it to generate topic code predictions for the utterances comprising the 116 visits used for the analysis of ARV dialogue (the two datasets are distinct). We then assessed the direction and magnitude of the change in the number of ARV utterances in the paired control versus intervention cases using a Wilcoxon signed rank test, following the original analysis. We compared the results for this analysis calculated using the true (manually assigned) topic codes to the results calculated using the predicted topic codes.

The Appendix includes further technical details of the machine learning models, the data preparation, and fitting of the models.

Results

Cross-Fold Validation

In the ECHO dataset the average kappa, averaged over the 10-fold cross-validation, was indicative of moderate to strong agreement at 0.491 (ranging from 0.470 to 0.526). Average overall accuracy was 63.8% (ranging from 61.8% to 65.9% across the 10 folds).

As shown in Table 3, the most frequent topics were *biomedical* and *logistics* (42% and 21%), followed by *psychosocial* and *ARV* (11% and 10%), and by *missing/other* and *socializing* (9% and 7%). Averaged over the 10 cross-validation folds, all three metrics are generally higher in the more frequent topics, and lower in the less frequent ones.

Figure 1 shows that the average empirical (solid black line) and predicted (dotted grey line) proportions over interaction deciles for each topic seem to agree.

Re-analysis of The Results of a Cross-over Trial Based on Model Predictions

According to the human annotation of topic codes, the median numbers of ARV utterances in the control and intervention visits were 49.5 and 76, respectively (Wilcoxon signed rank test $p=0.067$). According to the model-predicted topic codes, the corresponding medians were 39 and 55 (Wilcoxon signed rank test $p=0.036$). Both the manually assigned and the predicted topic codes support the notion that the intervention increases the number of ARV-related utterances. Using the predicted labels gives rise to a slightly lower p-value compared to that achieved with the manually assigned topic codes, but the difference is small and the qualitative conclusions are identical. The p-value is likely smaller due to the inherent complexity and variability of real-life interactions, which is not fully captured by the model.

Discussion and Limitations

The results are imperfect but promising. An unweighted kappa of 0.5 represents moderate agreement,²⁷ and 64% accuracy is better than the 42% accuracy trivially achievable by assigning each utterance to the most frequent topic code (*biomedical*).

We observed that the model performed better on common topics. For example, the model had trouble identifying utterances that belong to the *socializing* topic; this is perhaps because it is rare. However, correctly identifying rare topics might be important. We thus hope to improve performance with respect to these topics in future work. Collecting more manually annotated data and adding this to the current set with which we estimate model parameters may improve performance with respect to relatively rare topics.

It may be difficult to appreciate how well the model performs in terms of capturing conversational patterns from these scalar metrics. Figure 1 looks to capture this visually by plotting the average empirical proportion of utterances spent on each topic against the average predicted proportion over outpatient visits. One can see that the latter indeed tracks the former. For example, both exhibit the upward tick in logistics at the end of the visit. And when the model does make mistakes, it still seems to capture the larger trends over time. This can be seen in the *biomedical* topic; the first “spike” is overestimated, but the predicted shape agrees with the empirical frequencies.

Our re-analysis of the cross-over trial demonstrated that, in this case, counts of the machine’s predicted annotations picked up the same trend (an increase in the number of ARV utterances) as when using the manually assigned utterances. This suggests that useful trends can be gleaned from transcripts using automated annotation methods.

A limitation of this work is that we have restricted ourselves to predicting “highlevel” topic codes. While such high-level annotations may be useful for certain tasks (e.g., the above example of assessing the effect of an intervention in terms of increasing ARV-related dialogue), information is also lost, which may preclude more nuanced analyses of interactions. For example, without considering *speech act* codes it is impossible to infer who (the patient or the provider) is asking the majority of questions, and in general one cannot glean *how* issues are being discussed from topic codes alone (e.g., is the physician given *biomedical directives* or *asking biomedical questions*?). And without lower-level topic codes, broadly *biomedical* utterances regarding disparate issues such as treatment options and symptoms, will be indistinguishable from one another. A more sophisticated approach (compared to what we have proposed here) would be to model both topics and speech acts *jointly*. Further, to answer more exploratory questions surrounding provider communication (rather than just automating annotation), a more interpretable *generative* model may be preferable to the discriminative Conditional Random Field (CRF) used here.

Another limitation of the present work is that we have focused so far only on transcripts involving HIV patients. The model induced over these specific transcripts would probably suffer in performance if applied to, e.g., transcripts representing a general population of patients. While the approach proposed here is general, if it is to be used on an entirely different population then the model may need to be re-trained (i.e., its parameters re-estimated) using manually labeled data involving patients from the target population. Finally, non-technical issues related to applications that leverage automated coding will need to be explored: for example, physicians might not respond well to an automated system functioning as a screening tool to target interventions.

In sum, while the predictions of the proposed model are obviously imperfect, we believe that they are reliable enough to identify useful patterns of clinical communication. However, this work represents only a first step toward automating the annotation of patient-provider transcripts using computational methods, and these approaches need further evaluation and refinement.

Conclusions

As data becomes bigger and the questions that we ask become more complex (tailored to individual patients), clinical researchers will be overwhelmed, necessitating use of new computational tools. Already, for example, natural language processing has been used to identify patient smoking status from medical discharge records,²⁸ to discern tumor status from unstructured imaging reports,²⁹ and to recognize obesity and comorbidities.^{27,30} The present work extends this line of research to tackle the more complex problem of automating the annotation of transcripts of patient-provider interactions with clinically salient codes (such as those defined by the GMIAS).

We have demonstrated that machine learning methods can automate this annotation with reasonable accuracy. This sort of approach could dramatically reduce the cost of acquiring GMIAS-annotated transcripts, allowing for larger-scale analyses of patient-provider interactions. Such a system might serve as a component in a semi-automated screening tool that monitors patient-provider interactions to ensure that providers are, e.g., spending sufficient time on particular topics.

Acknowledgments

We thank Eugene Charniak and other members of the Brown Laboratory for Linguistic Information Processing (BLLIP) for helpful discussions. We also thank Yoojin Lee for discussions regarding the original RCT analysis.

Financial support for this study was provided in part by the National Institute of Mental Health (2K24MH092242). The annotation of the data used in this work was supported by grant R01MH083595. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report.

References

1. Ong LML, De Haes JCJM, Hoos AM, Lammes FB. Doctor-patient communication: a review of the literature. *Social science & medicine*. 1995; 40(7):903–918. [PubMed: 7792630]
2. Little P, Everitt H, Williamson I, et al. Observational study of effect of patient centredness and positive approach on outcomes of general practice consultations. *Bmj*. 2001; 323(7318):908–911. [PubMed: 11668137]
3. Epstein R, Street RL. Patient-centered communication in cancer care: promoting healing and reducing suffering. National Cancer Institute, US Department of Health and Human Services, National Institutes of Health. 2007
4. Kaplan SH, Greenfield S, Ware JE Jr. Assessing the effects of physician-patient interactions on the outcomes of chronic disease. *Medical care*. 1989; 110–127.
5. Oates J, Weston WW, Jordan J. The impact of patient-centered care on outcomes. *Family practice*. 2000; 49:796–804.
6. Beck RS, Daughtridge R, Sloane PD. Physician-patient communication in the primary care office: a systematic review. *The Journal of the American Board of Family Practice*. 2002; 15(1):25–38. [PubMed: 11841136]

7. Beck RS, Daughtridge R, Sloane PD. Physician-patient communication in the primary care office: a systematic review. *J.Am.Board Fam.Pract.* 2002; 15(1):25–38. [PubMed: 11841136]
8. Fine E, Reid MC, Shengelia R, Adelman RD. Directly observed patient-physician discussions in palliative and end-of-life care: a systematic review of the literature. *J Palliat Med.* 2010; 13(5):595–603. [PubMed: 20491550]
9. Laws MB, Epstein L, Lee Y, Rogers W, Beach MC, Wilson IB. The association of visit length and measures of patient-centered communication in HIV care: A mixed methods study. *Patient education and counseling.* 2011; 85(3):e183–e188. [PubMed: 21592716]
10. Searle, JR. *Speech acts: An essay in the philosophy of language.* Cambridge: university press; 1969.
11. Habermas, J. *The Theory of Communicative Action.* Vol. Volume I. Boston: Beacon; 1984.
12. Austin JL. *How to do things with words.* Oxford UP. 1961
13. Roter, D. *The Roter method of interaction process analysis.* Baltimore: Johns Hopkins University; 1989.
14. Roter D, Larson S. The Roter interaction analysis system (RIAS): utility and flexibility for analysis of medical interactions. *Patient education and counseling.* 2002; 46(4):243–251. [PubMed: 11932123]
15. Wilson IB, Laws MB, Safren SA, et al. Provider focused intervention increases HIV antiretroviral adherence related dialogue, but does not improve antiretroviral therapy adherence in persons with HIV. *Journal of Acquired Immune Deficiency Syndromes & Human Retrovirology.* 2010; 53(3): 338–347.
16. Laws MB, Bradshaw YS, Safren SA, et al. Discussion of sexual risk behavior in HIV care is infrequent and appears ineffectual: a mixed methods study. *AIDS and behavior.* 2011 May; 15(4): 812–822. [PubMed: 20981480]
17. Laws MB, Epstein L, Lee Y, Rogers W, Beach MC, Wilson I. Visit length and constructs of patient-centeredness in HIV care. 2010
18. Laws MB, Beach MC, Lee Y, et al. Provider-patient Adherence Dialogue in HIV Care: Results of a Multisite Study. *AIDS and behavior.* 2012 Jan 31.
19. Wilson IB, Laws MB, Safren SA, et al. Provider-focused intervention increases adherence-related dialogue, but does not improve antiretroviral therapy adherence in persons with HIV. *Journal of acquired immune deficiency syndromes (1999).* 2010; 53(3):338. [PubMed: 20048680]
20. Laws MB. GMIAS Coding Manual. 2013 <https://sites.google.com/a/brown.edu/m-barton-laws/home/gmias>.
21. Lafferty J, McCallum A, Pereira FCN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001
22. Settles, B. Biomedical named entity recognition using conditional random fields and rich feature sets; Paper presented at: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications; 2004.
23. Ji, G.; Bilmes, J. Dialog act tagging using graphical models; Paper presented at: Proc. ICASSP; 2005.
24. McCallum AK. Mallet: A machine learning for language toolkit. 2002
25. Srivastava, A.; Sahami, M. *Text mining: Classification, clustering, and applications.* CRC Press; 2010.
26. Jurafsky, D.; Martin, JH.; Kehler, A.; Vander Linden, K.; Ward, N. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* Vol. Vol 2. MIT Press; 2000.
27. Uzuner O. Recognizing obesity and comorbidities in sparse data. *Journal of the American Medical Informatics Association : JAMIA.* 2009 Jul-Aug;16(4):561–570. [PubMed: 19390096]
28. Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association : JAMIA.* 2008 Jan-Feb;15(1):14–24. [PubMed: 17947624]

29. Cheng LT, Zheng J, Savova GK, Erickson BJ. Discerning tumor status from unstructured MRI reports--completeness of information in existing reports and utility of automated natural language processing. *Journal of digital imaging*. 2010 Apr; 23(2):119–132. [PubMed: 19484309]
30. Yang H, Spasic I, Keane JA, Nenadic G. A text mining approach to the prediction of disease status from clinical discharge summaries. *Journal of the American Medical Informatics Association : JAMIA*. 2009 Jul-Aug;16(4):596–600. [PubMed: 19390098]

References

1. Lafferty J, McCallum A, Pereira FCN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001
2. Koller, D.; Friedman, N. Probabilistic graphical models: principles and techniques. MIT press; 2009.
3. Loeliger HA. An introduction to factor graphs. *Signal Processing Magazine, IEEE*. 2004; 21(1):28–41.
4. McCallum AK. Mallet: A machine learning for language toolkit. 2002

Appendix to Automatically Annotating Topics in Transcripts of Patient-Provider Interactions via Machine Learning

Methods Details

We use a Conditional Random Field (CRF),¹ a machine learning model amenable to many natural language processing tasks, to automatically predict topic codes for the utterances comprising patient-provider transcripts. Inducing a model to automatically tag utterances with topic codes can be viewed as a *structured classification task*. In structured classification, one assumes that the class labels are ordered in a way that reflects the underlying generative process. The most common example of this is sequential classification, in which it is assumed that instances are generated one after another, and that this ordering is somehow reflected by their labels. In our case, for example, it is very likely that adjacent utterances will belong to the same topic, due to the nature of conversations: people tend to stay on topic.

We will denote (representations of) utterances by $\mathbf{X}$ and their labels (topic codes) by $\mathbf{Y}$. We will use t to index time. The vector $\mathbf{X}_t$ thus represents the utterance at time t . We represent utterances via a standard vector representation of text, i.e., using indicator variables that encode whether a word does or does not appear in an utterance (this is called *bag-of-words* encoding and is described further below).

Conditional Random Fields are discriminative models, meaning they model only the conditional likelihood of $\mathbf{Y}$ given $\mathbf{X}$ (they do not model $\mathbf{X}$). The graphical model² corresponding to CRFs is shown in Figure 1 (adapted from Lafferty et al.¹). Darkened circles represent variables not explicitly modeled. CRFs use the formalism of *factor graphs*³ to model the relationship between adjacent observations (utterances). Specifically, *linear-chain* CRFs – the type of interest to us here – assume that the probability of a sequence of labels $\mathbf{Y}$ (topics) corresponding to a given set of utterances $\mathbf{X}$ is:

$$P_{\lambda}(Y|X) = \frac{1}{Z_{\lambda}(X)} \exp\left\{\sum_t \sum_j \lambda_j f_j(y_t, y_{t-1}, X_t)\right\}$$

where λ are the model parameters and Z_{λ} is a normalizing term to ensure a valid probability. The indexes t and j run over time and features (usually words in our case), respectively. The f_j function is a ‘feature function’ that extracts features from utterances. For example, $f_j(y_t, y_{t-1}, X_t)$ might be **1** if and only if state y_t is *biomedical* and state y_{t-1} is *socializing* and the word “pain” occurs in X_t . In this way, every possible label sequence of length two and feature combination gets its own weight λ_j . For example, there would be some weight corresponding to $\langle \text{socializing}, \text{socializing}, \text{hello} \rangle$; the value of this scalar would encode, roughly, the likelihood of a *socializing* utterance containing the word “hello” and following an utterance that also belongs to the *socializing* topic.

We used the CRF implementation in the Mallet toolkit, an open-source natural language processing library.⁴ Due to observed non-convergence during parameter estimation, we tightened the convergence threshold (from the default of 10^{-4} to 10^{-6}), and set the maximum number of iterations to 2000. We set the Gaussian variance around the 0 prior on parameters to 5, making it relatively uninformative. The model converged under these settings in fewer than 2000 iterations for all ten folds.

Figure 1. Average relative topic frequencies over deciles in patient-provider conversations. The solid line is the empirical (true) proportion; the dotted line corresponds to average model predictions (over test sets).

Appendix Figure 1.

Graphical model representation of the conditional random field (CRF).

Appendix Figure 2.
GMIAS topic codes (original schema).

Table 1

Some illustrative GMIAS topic code examples. To demonstrate the hierarchical structure of codes, we show a sub-topic under *Logistics*; recall, however, that in this work we focus only on the highest level of codings.

Topic Codes	Description/Example
<i>Biomedical</i>	Patient health and treatment: "what medication do you take?"
<i>ARV</i>	Adherence barriers; "so you're taking your meds"
<i>Psychosocial</i>	Substance abuse, jobs, housing, etc.; "My job is really stressful right now."
<i>Logistics</i>	
Non-physical	Appointments; "I need to get that script refilled"
<i>Physical examination</i>	"Take a deep breath"
<i>Socializing</i>	"Did you see the ball game?"

Table 2

A sample excerpt from an annotated transcript. The topic codes here include *logistics*, *biomedical* and *psychosocial*.

Topic	Speaker	Utterance
<i>Logistics</i>	Doctor	So it sounds like he's gonna help you out with some of those problems.
<i>Logistics</i>	Patient	I hope so.
<i>Biomedical</i>	Patient	I know blood pressure can be, as they say, the silent killer, but
<i>Biomedical</i>	Doctor	Mhm
<i>Biomedical</i>	Patient	I've had some headaches,
<i>Biomedical</i>	Patient	but I, he just told me take Tylenol
<i>Biomedical</i>	Patient	so that's all I take is the Tylenol.
<i>Biomedical</i>	Patient	Um, My this shoulder here?
<i>Biomedical</i>	Doctor	Mhm?
<i>Biomedical</i>	Patient	I've been having, all of a sudden it's been aching so bad [+3 inaudible words]
<i>Biomedical</i>	Patient	I mean just to move it
<i>Biomedical</i>	Patient	not to rotate it.
<i>Biomedical</i>	Doctor	Mhm.
<i>Biomedical</i>	Patient	But I haven't done anything
<i>Biomedical</i>	Patient	oh, it just aches
<i>Biomedical</i>	Patient	just to move it.
<i>Biomedical</i>	Patient	Like in the joint or in the bone or something
<i>Psychosocial</i>	Patient	and I have to work until the 23rd of September,
<i>Psychosocial</i>	Patient	cuz it's a seasonal job
<i>Psychosocial</i>	Patient	and it ended.

Table 3

Means and ranges (over ten folds) of per-topic metrics. These include recall (sensitivity), precision (positive predictive value) and F-measure (the harmonic mean of precision and recall).

Topic	Occurrences (%)	Recall (range)	Precision (range)	F-measure (range)
Biomedical	97489 (42%)	0.814 (0.784, 0.835)	0.665 (0.619, 0.695)	0.732 (0.692, 0.749)
Logistics	49683 (21%)	0.628 (0.577, 0.674)	0.616 (0.565, 0.663)	0.621 (0.585, 0.652)
Psychosocial	25606 (11%)	0.513 (0.456, 0.543)	0.579 (0.491, 0.683)	0.542 (0.495, 0.605)
ARV	23405 (10%)	0.508 (0.407, 0.576)	0.677 (0.616, 0.738)	0.579 (0.505, 0.630)
Missing/other	20432 (9%)	0.334 (0.293, 0.376)	0.652 (0.573, 0.765)	0.440 (0.410, 0.440)
Socializing	15729 (7%)	0.374 (0.295, 0.465)	0.496 (0.413, 0.593)	0.424 (0.346, 0.525)

Topics are ordered by frequency

Appendix Table 1

The topic codes used in this work. The original topic codes (left-column) are shown in Appendix Figure 2.

GMIAS code	Recoded Topic in This Work
1.1.x, 6.x, 7, 8	1 (<i>biomedical</i>)
2.x	2 (<i>psychosocial</i>)
3.1–3.3	3 (<i>logistics</i>)
4.x	4 (<i>socializing</i>)
5.1x, 5.2–5.33, 5.4	5 (<i>ARV</i>)
80–83, 99, 11	6 (<i>missing value/other</i>)

For more details on the original GMIAS coding schema, see <https://sites.google.com/a/brown.edu/m-barton-laws/home/gmias>.