Natural language processing for (mostly population) health

Byron C Wallace
byron.wallace@utexas.edu | byron.ischool.utexas.edu
This talk

Illustrative applications of NLP and Machine Learning methods, aiming to improve healthcare in an era of information overload.

Figure from BMJ; 2014
Talk overview

• A tour of work in NLP + health, including:
  – Evidence-based medicine (EBM)
  – Modeling patient-doctor communication
  – Social media (surveillance)

• Caveat: This is not a general survey! NLP + health is a huge sub-area; this is an extremely biased sampling of work I’ve done or am familiar with.
  – No coverage of, e.g., EHR mining
Evidence-based medicine + NLP/ML
Evidence-Based Medicine *n.*
The conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients
… only 20 percent of medical practices are based on rigorous research evidence … The rest are based on a kind of folklore.
From biomedical articles to actionable evidence

<table>
<thead>
<tr>
<th>Studies</th>
<th>Estimate (95% C.I.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>0.358 (0.152, 0.565)</td>
</tr>
<tr>
<td>- Carroll, 1997</td>
<td>0.411 (0.202, 0.621)</td>
</tr>
<tr>
<td>- Grant, 1981</td>
<td>0.370 (0.126, 0.615)</td>
</tr>
<tr>
<td>- Peck, 1987</td>
<td>0.353 (0.113, 0.593)</td>
</tr>
<tr>
<td>- Donat, 2003</td>
<td>0.254 (0.093, 0.416)</td>
</tr>
<tr>
<td>- Stewart, 1990</td>
<td>0.337 (0.094, 0.580)</td>
</tr>
<tr>
<td>- Young, 1995</td>
<td>0.397 (0.169, 0.626)</td>
</tr>
</tbody>
</table>
An old publication

...
An old publication

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit, z = 300, in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod, ù = 0.5, tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.
An old publication

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit, z = 300, in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod, ā = 0.5, tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Workspace

Chart
The data deluge

On average, 75 articles describing results from clinical trials are published every day. 

The automation of systematic reviews
Would lead to best currently available evidence at the push of a button
Lots of work in this space

- Two recent surveys:
  - More resources at: https://github.com/bwallace/automating-ebm-resources/wiki/Papers

- I’ll present just a specific piece of this work in class today
Semi-automating data extraction

this work supported by NIH grant R01LM012086

Semi-automating Risk of Bias (RoB) assessment


Automating PICO extraction

Randomized Control Trials (RCTs)

Population

Volunteers

Randomize to

Intervention group

Control group

Outcomes

Outcomes
Risk of Bias (RoB)

A key step in evidence synthesis: assessing the reliability of individual trials

– Assess *risks of bias* across several ‘domains’
<table>
<thead>
<tr>
<th></th>
<th>Blinding of participants and personnel</th>
<th>Incomplete outcome assessment</th>
<th>Other sources of bias</th>
<th>Selective reporting</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Random sequence generation</strong></td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td><strong>Welschen 2012</strong></td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td><strong>Soureti 2011</strong></td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td><strong>Powers 2011</strong></td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>+</td>
</tr>
<tr>
<td><strong>Benner 2008</strong></td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td><strong>Grover 2007</strong></td>
<td>?</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td><strong>Maasland 2007</strong></td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>?</td>
</tr>
<tr>
<td><strong>Steenkiste 2007</strong></td>
<td>+</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td><strong>Sheridan 2006</strong></td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td><strong>McAlister 2006</strong></td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>

Key:
- **low risk of bias**
- **high risk of bias**
- **unclear risk of bias**

No risk of bias criteria are specified in the table.
**Bias**  Allocation concealment

**Authors judgement**  Low risk

**Support for judgement**  Quote: "The Family Practice Research Coordinator at the University of British Columbia held this sequence independently and remotely"
drug as monotherapy. If proven to be more effective than single-drug therapy, this therapeutic approach may have important clinical implications for tobacco-dependence treatment. Exploration of combination therapy with existing drugs may provide the best opportunity to advance treatment in the absence of any new pharmacotherapies for tobacco dependence.

To investigate the efficacy of combination pharmacotherapy with varenicline and bupropion SR for smoking cessation, compared with varenicline monotherapy, we conducted a multicenter, randomized, phase 3 clinical trial.

Methods

Study Design
A randomized, blinded, placebo-controlled clinical trial was conducted at Mayo Clinic in Rochester, Minnesota, a Mayo Clinic Health System site in La Crosse, Wisconsin, and the University of Minnesota in Minneapolis between October 2009 and April 2013. The study consisted of a 12-week treatment period with follow-up through week 52. The institutional review boards of Mayo Clinic and the University of Minnesota approved all study procedures. The trial ended when recruitment was achieved and follow-up was completed.

Screening and Eligibility Criteria
Individuals were eligible to participate if they were at least 18 years of age, smoked at least 10 cigarettes per day for at least 6 months, were motivated to become smoking abstinent, completed written informed consent, and were in good health.

Potentially eligible participants were excluded if they were pregnant, lactating, or likely to become pregnant and 30 days) with another tobacco dependence investigational drug; or (17) current (30 days) bupropion or varenicline use.

Study Procedures
The study consisted of a telephone screening call, 11 clinic visits, and 3 follow-up telephone calls (Figure). One follow-up telephone call occurred during the medication phase at the time of the target quit date and 2 calls occurred after the medication phase. Two clinic visits occurred before the medication phase, 6 during the medication phase, and 3 after the medication phase.

For each participant, demographic data, tobacco use history, and self-reported information on race and ethnicity according to National Institutes of Health guidelines and recommendations for federally funded research were collected. Smoking dependence was assessed using the Fagerström Test for Nicotine Dependence (score range, 0-10). Depressive symptomatology was assessed using the Beck Depression Inventory, second edition. The Columbia-Suicide Severity Rating Scale assessed for suicidal ideation or behaviors. Both assessments were completed at baseline and weeks 2, 4, 8, 14, 26, and 52.

A central pharmacy randomly assigned study medication in a 1:1 ratio using a computer-generated randomization sequence with variable-sized blocks ranging from 2 to 8 stratified by study site. Study medication was labeled and dispensed according to participant identification, ensuring that treatment assignment remained concealed from the participant, investigators, and all study personnel having participant contact. Following provision of informed consent, participants received randomly assigned medication at the baseline visit.

https://robot-reviewer.vortext.systems/
https://github.com/ijmarshall/robotreviewer
The machine learning task

Input: a full-text paper

Machine Learning

Output: RoB assessments and supporting quotes
Traditional supervised learning

- labeled data $\mathcal{L}$
- unlabeled data $\mathcal{U}$
- learner
- predictive model
Collecting annotations is expensive and time-consuming.

Instead, **we will use previously conducted reviews** to train ML models.
The Cochrane Database of Systematic Reviews (CDSR)

- We’ve linked 13,000 CDSR entries to published full-text PDFs describing trials
- We derive labels on articles and sentences from the CDSR
Distant supervision
alternatively, supervision by database

Craven & Kumlien, AAAI, 1999
Abstract

We develop a model for automatically extracting the sample size from the free text of clinical trial abstracts. We demonstrate that training this model via distant supervision (by leveraging the Cochrane Database of Systematic Reviews) results in performance comparable to a fully supervised approach, with extremely minimal explicit human annotation. This work demonstrates the promise of distantly supervised models with existing data sources for automating data extraction in the biomedical domain.

1 Introduction & Motivation

Evidence-Based Medicine (EBM) looks to inform clinical practice in light of the best available evidence. Systematic reviews are rigorous syntheses of the studies pertaining to specific clinical questions. They are vital tools of EBM, and are increasingly used to inform all levels of healthcare, from bedside decisions to policy-making (Petticrew, 2001). But producing and maintaining such reviews is becoming increasingly onerous due in part to the exponential expansion of the biomedical literature base. Producing a single review requires between 1000–2000 person hours (Allen and Olkin, 1999). Conducting systematic reviews is therefore laborious and expensive, especially as reviews are typically undertaken by highly trained individuals. The number of systematic reviews is rapidly increasing, thus multiplying costs (over 27,000 reviews were published in 2012). The generation of primary evidence is now outpacing our ability to synthesize it given pragmatic resource constraints (Bastian et al., 2010; Wallace et al., 2013). If we are to keep systematic reviews and related EBM products current, we need to optimize the steps involved in conducting evidence synthesis. Data extraction (extracting variables of interest from the free text of articles describing clinical trials) represents a substantial part of the systematic review workload. In this paper we explore automating the data extraction task. Here we simplify the task by focusing on extracting the trial sample size (number of participants) from article abstracts. Methods to extract such information from free text have the potential to greatly mitigate the workload involved in producing systematic reviews.

Our main contribution is a distantly supervised (Snow et al., 2004; Mintz et al., 2009) approach to sample size extraction that leverages the Cochrane Database of Systematic Reviews (CDSR) (The Cochrane Collaboration, 2014). We show that by exploiting this resource we can achieve extraction performance comparable to a fully supervised approach, with extremely few labels (we use labeled data only for hyper-parameter tuning). This is promising because it implies that existing curated resources can be leveraged to automate data extraction. Such databases already exist and are growing in size, and include the CDSR and the recently established Systematic Review Data Repository (SRDR) (Ip et al., 2012).

We next describe the data we use in this work (including a new corpus of biomedical abstracts annotated by the authors). We review related work in Section 3, describe our methods in Section 4, report results in Section 5 and end with a brief discussion in Section 6.
Machine learning approach overview

• Regularized linear models (parameterized by $w$)

• Very high-dimensional, sparse feature space

• Parameter estimation via stochastic gradient descent
Document-level objective

\[
\arg\min_{\mathbf{w}_d^q} \alpha \|\mathbf{w}_d^q\|^2 + \sum_{i=1}^{n^q} \mathcal{L}\{\mathbf{w}_d^q \cdot \mathbf{x}_i, y_i^q\}
\]

- Regularizer
- Empirical loss

“low” or “high/unknown” risk of bias for domain \( q \)
... and basically the same for sentence model

\[
\arg\min_{w^q_s} \alpha \|w^q_s\|^2 + \sum_{i=1}^{n^q} \sum_{j=1}^{m_i} \mathcal{L}\{w^q_s \cdot s_{ij}, l^q_{ij}\}
\]

\(s\) subscript for sentences

indicates whether sentence \(j\) in article \(i\) supports risk of bias judgement for domain \(q\)
But article level assessments are not independent of supporting sentences.
A simple joint model

\[ y_i^q = \text{sign}\{ w_d^q \cdot x_i + w_{ds}^q \cdot s_{i*}^q \} \]

as before: document level features

features that indicate tokens present in the supporting sentence for this domain
A simple joint model

\[ y^q_i = \text{sign}\{ w^q_d \cdot x_i + w^q_{ds} \cdot s^q_{i*}\} \]

e.g., computer generated indicates low risk for poor randomization; double blind does so for proper blinding
A simple joint model

At test time, we don’t know which sentences support assessments for which domains, so we use the predictions.

$$y^q_i = \text{sign}\{w_d^q \cdot x_i + \hat{l}^q_{i0} \cdot (w_{ds}^q \cdot s_{i0}^q) + \ldots + \hat{l}^q_{im_i} \cdot (w_{ds}^q \cdot s_{im_i}^q)\}$$
This model ignores correlations between domains.

We use a *multi-task* approach to tie weight vectors across domains in a joint model.
Multi-task learning

• Predict multiple outputs from a shared representation

• Allows ‘borrowing of strength’ across tasks
A ‘frustratingly easy’ approach

<table>
<thead>
<tr>
<th>N x K</th>
<th>Q entries for d0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0 1 0 1 0 ... 0</td>
</tr>
<tr>
<td>1</td>
<td>0 1 0 1 0 ... 0</td>
</tr>
<tr>
<td>1</td>
<td>0 1 0 1 0 ... 0</td>
</tr>
<tr>
<td>0</td>
<td>0 0 0 0 0 ... 0</td>
</tr>
<tr>
<td>1</td>
<td>0 0 0 0 0 ... 0</td>
</tr>
</tbody>
</table>

\[ y = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 1 \end{pmatrix} \]

\[ x_i = \begin{pmatrix} 0 1 0 1 0 & 0 1 0 1 0 & 0 0 0 0 0 \end{pmatrix} \]

\[ x_i^q = \begin{pmatrix} 0 0 0 0 0 & 0 0 0 0 0 & 0 1 0 1 0 \end{pmatrix} \]

[Daumé III, 09]
• **Joint** model achieves an average of 3+% absolute improvement in accuracy over baseline (mean 0.70 v 0.73)

• Still 5-10% behind humans (~80% accurate)
Sentence evaluation

- We showed domain experts sentences extracted for different domains by
  (1) random guessing (a baseline approach)
  (2) human reviewers (i.e., from the Cochrane database)
  (3) our model

- They didn’t know where these sentences came from.

- They rated sentences as *highly relevant*, *somewhat relevant*, or *not relevant*. 
## Sentence evaluation

<table>
<thead>
<tr>
<th>Domain</th>
<th>Trials (n)</th>
<th>baseline</th>
<th>cochrane</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>378</td>
<td>0.50%</td>
<td>56.50%</td>
</tr>
<tr>
<td>1. Random sequence generation</td>
<td>81</td>
<td>0.00%</td>
<td>60.50%</td>
</tr>
<tr>
<td>2. Allocation concealment</td>
<td>75</td>
<td>0.00%</td>
<td>60.00%</td>
</tr>
<tr>
<td>3. Blinding of participants and personnel</td>
<td>76</td>
<td>0.00%</td>
<td>68.40%</td>
</tr>
<tr>
<td>4. Blinding of outcome assessment</td>
<td>56</td>
<td>0.00%</td>
<td>57.10%</td>
</tr>
<tr>
<td>5. Incomplete reporting of outcomes</td>
<td>67</td>
<td>3.00%</td>
<td>50.80%</td>
</tr>
<tr>
<td>6. Selective reporting</td>
<td>23</td>
<td>0.00%</td>
<td>4.60%</td>
</tr>
</tbody>
</table>

percent of sentences deemed ‘highly relevant’ by experts
Sentence evaluation

<table>
<thead>
<tr>
<th>Domain</th>
<th>Trials (n)</th>
<th>baseline</th>
<th>top1</th>
<th>top3</th>
<th>cochrane</th>
<th>top1 v cochrane</th>
<th>top3 v cochrane</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>378</td>
<td>0.50%</td>
<td>45.00%</td>
<td>60.40%</td>
<td>56.50%</td>
<td>-11.6% (-18.5% to -4.4%); P&lt;0.001</td>
<td>+3.9%, (-3.2% to +10.9%); P=0.141</td>
</tr>
<tr>
<td>1. Random sequence generation</td>
<td>81</td>
<td>0.00%</td>
<td>55.60%</td>
<td>65.40%</td>
<td>60.50%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Allocation concealment</td>
<td>75</td>
<td>0.00%</td>
<td>44.00%</td>
<td>60.00%</td>
<td>60.00%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Blinding of participants and personnel</td>
<td>76</td>
<td>0.00%</td>
<td>55.30%</td>
<td>72.40%</td>
<td>68.40%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Blinding of outcome assessment</td>
<td>56</td>
<td>0.00%</td>
<td>39.30%</td>
<td>62.50%</td>
<td>57.10%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Incomplete reporting of outcomes</td>
<td>67</td>
<td>3.00%</td>
<td>40.90%</td>
<td>57.60%</td>
<td>50.80%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Selective reporting</td>
<td>23</td>
<td>0.00%</td>
<td>0.00%</td>
<td>4.60%</td>
<td>4.60%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

performance is actually **better**, and at least non-inferior, to human performance if we consider the top-3 sentences extracted by the model.
Statistical models of patient-doctor communication


Patient-doctor communication

• Patient-doctor communication is a critical part of quality care

• Especially for patient-centered care
  – Patients need to understand *what* is wrong with them, *steps* to fix it and *why* those steps will work

• There are significant correlations between verbal behaviors and health outcomes

• But it’s difficult to study
## Patient-doctor communication

<table>
<thead>
<tr>
<th>Role</th>
<th>Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>D</strong></td>
<td>Let me just write down some of these issues here so I get them straight in my mind.</td>
</tr>
<tr>
<td><strong>P</strong></td>
<td>Doctor you ain’t got to tell me nuttin’.</td>
</tr>
<tr>
<td><strong>P</strong></td>
<td>I’m in very good hands when I’m around you.</td>
</tr>
<tr>
<td><strong>P</strong></td>
<td>If push comes to a shove, you open the window and throw me out.</td>
</tr>
<tr>
<td><strong>D</strong></td>
<td>I wanted to ask you, too - you know you had that colonic polyp - is it two years from now that they’re going to be doing the repeat?</td>
</tr>
<tr>
<td><strong>P</strong></td>
<td>Yeah.</td>
</tr>
<tr>
<td><strong>D</strong></td>
<td>We’ll do the repeat coloscopy in about two years.</td>
</tr>
</tbody>
</table>
Patient-doctor communication

<table>
<thead>
<tr>
<th>Role</th>
<th>Utterance</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>Let me just write down some of these issues here so I get them straight in my mind.</td>
<td>Logistics</td>
</tr>
<tr>
<td>P</td>
<td>Doctor you ain’t got to tell me nuttin’.</td>
<td>Socializing</td>
</tr>
<tr>
<td>P</td>
<td>I’m in very good hands when I’m around you.</td>
<td>Socializing</td>
</tr>
<tr>
<td>P</td>
<td>If push comes to a shove, you open the window and throw me out.</td>
<td>Socializing</td>
</tr>
<tr>
<td>D</td>
<td>I wanted to ask you, too - you know you had that colonic polyp -</td>
<td>Biomedical</td>
</tr>
<tr>
<td>D</td>
<td>- is it two years from now that they’re going to be doing the repeat?</td>
<td>Biomedical</td>
</tr>
<tr>
<td>P</td>
<td>Yeah.</td>
<td>Biomedical</td>
</tr>
<tr>
<td>D</td>
<td>We’ll do the repeat coloscopy in about two years.</td>
<td>Biomedical</td>
</tr>
<tr>
<td>Topic Codes</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Biomedical</td>
<td>Patient health and treatment: “what medication do you take?”</td>
<td></td>
</tr>
<tr>
<td>ARV</td>
<td>Adherence barriers; &quot;so you're taking your meds&quot;</td>
<td></td>
</tr>
<tr>
<td>Psychosocial</td>
<td>Substance abuse, jobs, housing, etc.; &quot;My job is really stressful right now.&quot;</td>
<td></td>
</tr>
<tr>
<td>Logistics</td>
<td>Appointments; &quot;I need to get that script refilled&quot;</td>
<td></td>
</tr>
<tr>
<td>Physical examination</td>
<td>“Take a deep breath”</td>
<td></td>
</tr>
<tr>
<td>Socializing</td>
<td>“Did you see the ball game?”</td>
<td></td>
</tr>
</tbody>
</table>
The utility of topic annotations

- Quantitatively address questions about communication

- Consider an intervention intended to alter doctor communication around ARV adherence
  - How do we know if it worked?
Wilson et al., 2010

- Administered an intervention to a bunch of doctors

- Counted ARV adherence utterances in conversations before and after intervention: is there a difference?

- 116 visits manually annotated (58 visits before/58 after)
  - Median ARV utterances in controls (no intervention): 49.5
  - And in cases (intervention): 76
  - \( p \)-value = 0.067

- But annotation is laborious. Can we automate it?
Predicting topics given utterances

### Input

| “How do you feel?” | Biomedical |
| “My stomach hurts” | Biomedical |

- Standard **structured learning** problem
- Standard structured learning approach (that you’re now familiar with) **conditional random field**

\[
p_\theta(y|x) = \frac{1}{Z_\theta(x)} \exp \left\{ \sum_{t=1}^{T} \sum_{k=1}^{K} \theta_k f_k(y_{t-1}, y_t, x_t) \right\}
\]
Topic Prediction Results

Average overall accuracy: about 64% (62% to 66%)

Average Kappa: .49 (.47 to .53)
Topic prediction results

We report averages and ranges of two performance metrics: accuracy and kappa. The former captures overall agreement of the models topic predictions with the humans annotations; the latter adjusts this measure to account for agreement by chance (one could achieve 42% accuracy by just coding each utterance as biomedical, the most frequent topic). We report these in Section 4. (Results using a generative model were largely comparable, though slightly lower.)

The results are imperfect but promising; 50% kappa is moderately strong agreement [82]. To contextualize these results in terms of their implications for interpreting physician-patient interactions, we can compare the predicted versus true (mean) frequencies of topics over the duration of visits to see if they agree. To this end, we broke each visit into 10 equally long time periods (deciles). Figure 2 plots the average empirical (solid) and predicted (dotted) proportions of each topic over these deciles. Encouragingly, the latter indeed tracks the former, and seems to capture broad trends. For example, both exhibit the upward tick in logistics at the end of visits.

We also used the automatic annotations to reproduce the results from the MAP study [85] (we described this dataset in Section 2.2). Wilson et al. [85] demonstrated that this intervention indeed increased adherence-related dialogue, as assessed by manually annotating utterances with topic codes. Here we explore whether the predicted codes tell the same story, demonstrating a potential practical application of automated annotation technology. We tally codes predicted by the model and compare these to counts of manually assigned codes (following the original analysis). According to the manually assigned topic codes, the median numbers of ARV utterances in the control and intervention visits are 49.5 and 76, respectively. According to the predicted topic codes, the corresponding medians are 39 and 55. Using a Wilcoxon signed rank test to assess significance, we calculate a p-value of .067 (corresponding to the null hypothesis that the median numbers of ARV utterances in the case and control visits are equal) with respect to the manually assigned topic codes. The corresponding p-value calculated using the predicted topic codes is .036. Thus 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 using manually assigned topic codes, but the difference is small and the qualitative conclusions are identical.

In preliminary work more closely related to aim 2, some in our group have investigated methods for automatically analyzing information flow in provider-patient visits using machine learning [57]. These methods were able to identify patterns in interactions that correlated with patient survey response data. Specifically, this work introduced a metric designed to capture ‘information flow’; this metric is a function over speech act codes. It was shown that this metric correlates with patient-reported measures of communication quality, and this correlation holds even when predicted codes are used in place of manually assigned ones.
Reproducing the RCT analysis

- From manual codes: 49.5 median ARV utterances for control visits and 76 for cases \( (p\text{-value } .067) \)
- Using *predicted* codes: 39 for control visits; and 55 for cases \( (p\text{-value } .036) \)
- So predicted codes reveal the same trend at a comparable significance level
So we can predict topic codes, but is that enough?

- Tells us *what* is being discussed but not *how* it is

- “Would you please take your ARV meds?” vs. “You need to take your ARV meds!”
  - Both are *ARV adherence* utterances, but the communication styles are very different

- Enter *speech acts*
A bit of sociolinguistics
Speech acts in GMIAS

- GMIAS includes following speech act codes: ask question, commissive, conversation management, directive, empathy, give information, humor/levity, and social-ritual.
### Patient-Doctor communication

<table>
<thead>
<tr>
<th>Role</th>
<th>Utterance</th>
<th>Topic</th>
<th>Speech act</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>Let me just write down some of these issues here so I get them straight in my mind.</td>
<td>Logistics</td>
<td>Commissive</td>
</tr>
<tr>
<td>P</td>
<td>Doctor you ain’t got to tell me nuttin’.</td>
<td>Socializing</td>
<td>Directive</td>
</tr>
<tr>
<td>P</td>
<td>I’m in very good hands when I’m around you.</td>
<td>Socializing</td>
<td>Give Info.</td>
</tr>
<tr>
<td>P</td>
<td>If push comes to a shove, you open the window and throw me out.</td>
<td>Socializing</td>
<td>Humor/Levity</td>
</tr>
<tr>
<td>D</td>
<td>I wanted to ask you, too - you know you had that colonic polyp - is it two years from now that they’re going to be doing the repeat?</td>
<td>Biomedical</td>
<td>Conv. Mgmt.</td>
</tr>
<tr>
<td>D</td>
<td></td>
<td>Biomedical</td>
<td>Ask Q.</td>
</tr>
<tr>
<td>D</td>
<td></td>
<td>Biomedical</td>
<td>Ask Q.</td>
</tr>
<tr>
<td>P</td>
<td>Yeah.</td>
<td>Biomedical</td>
<td>Conv. Mgmt.</td>
</tr>
<tr>
<td>D</td>
<td>We’ll do the repeat coloscopy in about two years.</td>
<td>Biomedical</td>
<td>Give Info.</td>
</tr>
</tbody>
</table>

**Table 1:** An excerpt from a patient-doctor interaction, annotated with topic and speech act codes. The "D" and "P" roles denote doctor and patient, respectively. "Conv. Mgmt." abbreviates conversation management; "Ask Q." abbreviates ask question.
Jointly modeling topics and speech acts

- Want an interpretable generative model to analyze interactions (not just predictions)
- But standard structural generative models only handle univariate case
Markov-Multinomial model
Markov-Multinominal model

- Decompose sequence into transitions and emissions

- Transitions:

\[ P(y_t|y_0, \ldots, y_{t-1}) = P(y_t|y_{t-1}) = \lambda_{y_{t-1},y_t} \]

- Emissions:

\[ P(u_t|y_t) = \prod_{w \in u_t} P(w|y_t) = \prod_{w \in u_t} \tau_{y_t,w} \]
Jointly modeling topics and speech acts

We have outlined a novel, fully generative model of topics and speech acts that, in addition to improving automated annotation (code prediction) alone, a discriminative model may outperform this fully generative approach. As we will discuss in Aim 2, this generative, additive component model will allow for quantitative exploration of factors that affect communication patterns.

We will assume an analogous conditional transition probability for speech acts. Model parameters can be estimated using standard optimization techniques, e.g., gradient descent. To predict annotations for a new utterance, the observed frequency of topic codes by themselves are affected both by the current topic and the current speech act. Denoting the log of the "background" frequency for a given topic as conditional transition probability for speech acts.

$$\text{Jointly modeling topics and speech acts}$$

$$P(\text{Topic}_t, \text{Speech Act}_t | \text{Topic}_{t-1}, \text{Speech Act}_{t-1}) = \frac{\exp \left( \sum_{w \in \text{words}} \beta_w \left( \text{Topic}_t \right) \delta(w) \right) \cdot \sum_{s \in \text{speech acts}} \exp \left( \sum_{w \in \text{words}} \beta_w \left( \text{Speech Act}_t \right) \delta(w) \right)}{\sum_{s \in \text{speech acts}} \sum_{y \in \text{topics}} \exp \left( \sum_{w \in \text{words}} \beta_w \left( \text{Topic}_t \right) \delta(w) \right) \cdot \sum_{s \in \text{speech acts}} \exp \left( \sum_{w \in \text{words}} \beta_w \left( \text{Speech Act}_t \right) \delta(w) \right)}$$

The graphical representation of our proposed model is shown in Figure 3. First, we will model the probability of a word being used in an utterance because it will provide machinery to tackle aim 2 (as we discuss below). By FHMM does not obviously lend itself to discrete observations (typically Gaussian distributions), and the training time (via maximum likelihood) is too intensive to be feasible for our application.

$$P(\text{Utterance}_t, \text{Topic}_t, \text{Speech Act}_t | \text{Utterance}_{t-1}, \text{Topic}_{t-1}, \text{Speech Act}_{t-1})$$

$$= \frac{\exp \left( \sum_{w \in \text{words}} \beta_w \left( \text{Topic}_t \right) \delta(w) \right) \cdot \sum_{s \in \text{speech acts}} \exp \left( \sum_{w \in \text{words}} \beta_w \left( \text{Speech Act}_t \right) \delta(w) \right)}{\sum_{s \in \text{speech acts}} \sum_{y \in \text{topics}} \exp \left( \sum_{w \in \text{words}} \beta_w \left( \text{Topic}_t \right) \delta(w) \right) \cdot \sum_{s \in \text{speech acts}} \exp \left( \sum_{w \in \text{words}} \beta_w \left( \text{Speech Act}_t \right) \delta(w) \right)}$$

An obvious means of jointly modeling topics and speech acts is to treat the Cartesian product of these two sets as a single state space, but this space is too large and sparse for this approach to be practical. Instead, we will propose a model that captures both the topical content and the speech act type associated with each utterance.

$$P(\text{Utterance}_t, \text{Topic}_t, \text{Speech Act}_t | \text{Utterance}_{t-1}, \text{Topic}_{t-1}, \text{Speech Act}_{t-1})$$

$$= \frac{\exp \left( \sum_{w \in \text{words}} \beta_w \left( \text{Topic}_t \right) \delta(w) \right) \cdot \sum_{s \in \text{speech acts}} \exp \left( \sum_{w \in \text{words}} \beta_w \left( \text{Speech Act}_t \right) \delta(w) \right)}{\sum_{s \in \text{speech acts}} \sum_{y \in \text{topics}} \exp \left( \sum_{w \in \text{words}} \beta_w \left( \text{Topic}_t \right) \delta(w) \right) \cdot \sum_{s \in \text{speech acts}} \exp \left( \sum_{w \in \text{words}} \beta_w \left( \text{Speech Act}_t \right) \delta(w) \right)}$$

We hypothesize that this model will provide more accurate automatic annotations than existing act codes. Thus we will explore such a model for our task, and compare its performance to the approach.
An additive component sequential model: transitions

\[ P(s_t | s_{t-1}, y_{t-1}) = \frac{1}{Z_s} \exp\{\pi_s + \sigma_{s_{t-1},s_t} + \sigma_{y_{t-1},s_t} + \sigma(y_{t-1},s_{t-1}),s_t\} \]

- \( \pi_s \) is the baseline probability of \( s_t \).
- \( Z_s \) is a normalizing constant for speech acts.
- \( \pi_s \) and \( \sigma \) are components corresponding to previous topic and speech act, respectively.
- Interaction component for topic/speech act interactions.

With the exception that we do not explicitly model the distribution of topics, we use a corpus of patient-provider visits annotated with GMIAS codes. The GMIAS has been used in context of an intervention trial (Wilson et al., 2010); we assess convergence by calculating predictive performance on a held-out portion (5%) of the training dataset at each step, with the exception that we do not explicitly model the distribution of topics. We use a Newton optimization method similar to the approach outlined by Rabiner and Juang (1986) over a matrix of transitions from topic/speech act pair (\( C \)) to \( C \) transitions out of this pair. The term \( \exp\{\pi_s + \sigma \} \) gives rise to both the topic and speech acts; these, in turn, jointly induce a distribution over words, topics and speech ground frequencies.
Jointly modeling topics \textit{and} speech acts

\[ P(y_t) \text{ is independent of } P(s_t) \text{ given } y_{t-1} \text{ and } s_{t-1} \text{ because time is a blocking agent} \]
An Additive Component Sequential Model: Emissions

\[ P(w|y_t, s_t) = \frac{1}{Z_w} \exp\{\theta_w + \eta_{w}^{y_t} + \eta_{w}^{s_t} + \eta_{w}^{s_t,y_t}\} \]

- component corresponding to current topic
- baseline probability of \( w \)
- component corresponding to speech act
- topic / speech act interaction component
Putting it all Together

\[
P(y_t, s_t|s_{t-1}, y_{t-1}, u_t) = \\
P(u_t|y_t, s_t) \cdot P(y_t|y_{t-1}, s_{t-1}) \cdot P(s_t|s_{t-1}, y_{t-1})
\]

- Optimization via gradient descent
- Prediction via Viterbi decoding
Experimental Results

Our evaluation includes two parts. First, we perform standard cross-validation over the aforementioned 360 annotated interactions, evaluating F-measure for each topic/speech act pair. Second, we look to automatically reproduce an analysis of a randomized control trial that assessed the efficacy of an intervention meant to alter physician-patient communication. We show that JAS outperforms the baseline approach with respect to both tasks. We will make our source code available upon publication.

We emphasize that while we are here comparing predictive performance, we are specifically interested in fully generative models of conversations due to the longer-term applications we have in mind. We would like, e.g., to use this model to assess the variation in communicative approaches across different hospitals, and generative models are more naturally amenable to answering such exploratory questions. Indeed, the additive component based model we have developed here allows us to easily add physician- and center-specific parameters. Further, we may soon have access to many unannotated transcripts, and we would like to learn from these; generative approaches allow straightforward exploitation of unlabeled data. For these reasons, Results (Macro-averaged)
Revisiting the ARV Study

- Median (lower, upper) counts of utterances that have topic ARV and speech act give information over control (no intervention before visit) and intervention visits

<table>
<thead>
<tr>
<th></th>
<th>True</th>
<th>MM</th>
<th>JAS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>control</td>
<td>intervention</td>
<td>control</td>
</tr>
<tr>
<td></td>
<td>10 (4, 28)</td>
<td>23 (11, 39)</td>
<td>13 (5, 33)</td>
</tr>
</tbody>
</table>
Physician-specific parameters

\[ P(s_t | s_{t-1}, d) \propto \exp\{\pi_{s_t} + \sigma_{s_{t-1},s_t} + \sigma^d_{s_{t-1},s_t}\} \]

doctor \(d\)'s specific tendency to transition from speech act \(s_{t-1}\) to \(s_t\)
Physician-specific parameters

\[
\begin{bmatrix}
\hat{\theta}_{0,0} & \hat{\theta}_{0,1} & \cdots & \hat{\theta}_{0,|S|} \\
\hat{\theta}_{1,0} & \hat{\theta}_{1,1} & \cdots & \hat{\theta}_{1,|S|} \\
\vdots & \vdots & \ddots & \vdots \\
\hat{\theta}_{|D|,0} & \hat{\theta}_{|D|,1} & \cdots & \hat{\theta}_{|D|,|S|}
\end{bmatrix}
\]

Speech act to speech act transition parameters

Identify clusters (communication styles)
explained by the clustering. Note that this is a more
doctor characteristics and heterogeneity that is un-
tercept represents the combined effects of omitted
the nesting of patients within doctors. This random
ndal, 2008) with a random intercept to account for
ear mixed effects model (Rabe-Hesketh and Skro-
physician communication, we used a two-level lin-
regression analysis of the cluster assignments.
be illustrative; we also performed a more rigorous
was 1.73. The difference in means is meant only to
cluster, the analogous aggregate summary response
were 1.57; for physicians in the 'blue' (lower right)
doctors assigned to the 'gray' cluster (upper left)
tions in Table 2 provided after sessions involving
we show the identified clusters in the PCA-reduced
linear component model in Figure 2. On the left,
We show clustering results using the proposed log-
Newton-based gradient descent method, similar to
we fit the other model components with a standard
acies across doctors. We fix the
under review. We then take the (doctor-specific)
ing topics
(though excluding doctor specific terms) for model-
posed a variant of this additive sequential model
therein).
Figure 2: discovered physician clusters (and mean patient responses from sessions involving doctors
: coefficients and 95% CI's from a regression of patient survey data on cluster assignments.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|}
\hline
Question & Coefficient & 95% CI \tabularnewline
\hline
Q1 & 1.73 & (1.57, 1.89) \tabularnewline
Q2 & 1.57 & (1.40, 1.74) \tabularnewline
Q3 & 1.00 & (0.83, 1.17) \tabularnewline
Q4 & 2.00 & (1.83, 2.17) \tabularnewline
Q5 & 1.50 & (1.30, 1.70) \tabularnewline
Q6 & 2.00 & (1.80, 2.20) \tabularnewline
Q7 & 1.50 & (1.30, 1.70) \tabularnewline
Q8 & 1.00 & (0.80, 1.20) \tabularnewline
Q9 & 2.00 & (1.80, 2.20) \tabularnewline
Q10 & 1.50 & (1.30, 1.70) \tabularnewline
Q11 & 1.00 & (0.80, 1.20) \tabularnewline
Q12 & 2.00 & (1.80, 2.20) \tabularnewline
\hline
\end{tabular}
\caption{Coefficient estimates and 95% confidence intervals for patient responses to survey questions.}
\end{table}

\section*{5 Results and Discussion}

\par
We acknowledge that we have conducted multi-
Recall that lower is better here.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{clustering_physicians.png}
\caption{Clustering Physicians}
\end{figure}
Are the Clusters Meaningful?

How is the provider who takes care of your HIV at ...

**Overall**
Q1 ... explaining the results of tests in a way that you understand?
Q2 ... giving you facts about the benefits and risks of treatment?
Q3 ... telling you what to do if certain problems or symptoms occur?
Q4 ... demonstrating caring, compassion, and understanding?
Q5 ... understanding your health worries and concerns?

**HIV-specific**
Q6 ... talking with you about your sex life?
Q7 ... asking you about stresses in your life that may affect your health?
Q8 ... asking about problems with alcohol?
Q9 ... asking about problems with street drugs like heroin and cocaine?

**Adherence**
Q10 ... giving you information about the right way to take your antiretroviral medicines?
Q11 ... understanding the problems you have taking your antiretroviral medicines?
Q12 ... helping you solve problems you have taking your antiretroviral medicines the right way?
Clustering Physicians

- Q1: p = 0.091
- Q2: p = 0.404
- Q3: p = 0.103
- Q4: p = 0.030
- Q5: p = 0.200
- Q6: p = 0.489
- Q7: p = 0.388
- Q8: p = 0.616
- Q9: p = 0.637
- Q10: p = 0.121
- Q11: p = 0.121
- Q12: p = 0.009