



## City Research Online

### City, University of London Institutional Repository

---

**Citation:** Howes, C., Purver, M. & McCabe, R. (2013). Using Conversation Topics for Predicting Therapy Outcomes in Schizophrenia. *Biomedical Informatics Insights*, 6s1(Suppl 1), pp. 39-50. doi: 10.4137/bii.s11661

This is the accepted version of the paper.

This version of the publication may differ from the final published version.

---

**Permanent repository link:** <https://openaccess.city.ac.uk/id/eprint/21767/>

**Link to published version:** <https://doi.org/10.4137/bii.s11661>

**Copyright:** City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

**Reuse:** Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

---

---



**OPEN ACCESS**

Full open access to this and thousands of other papers at <http://www.la-press.com>.

## Using Conversation Topics for Predicting Therapy Outcomes in Schizophrenia

Christine Howes<sup>1</sup>, Matthew Purver<sup>1</sup> and Rose McCabe<sup>2</sup>

<sup>1</sup>Cognitive Science Research Group, School of Electronic Engineering and Computer Science. <sup>2</sup>Barts and the London School of Medicine Queen Mary University of London, London, UK.

Corresponding authors email: [c.howes@qmul.ac.uk](mailto:c.howes@qmul.ac.uk); [m.purver@qmul.ac.uk](mailto:m.purver@qmul.ac.uk); [r.mccabe@qmul.ac.uk](mailto:r.mccabe@qmul.ac.uk)

**Abstract:** Previous research shows that aspects of doctor-patient communication in therapy can predict patient symptoms, satisfaction and future adherence to treatment (a significant problem with conditions such as schizophrenia). However, automatic prediction has so far shown success only when based on low-level lexical features, and it is unclear how well these can generalize to new data, or whether their effectiveness is due to their capturing aspects of style, structure or content. Here, we examine the use of topic as a higher-level measure of content, more likely to generalize and to have more explanatory power. Investigations show that while topics predict some important factors such as patient satisfaction and ratings of therapy quality, they lack the full predictive power of lower-level features. For some factors, unsupervised methods produce models comparable to manual annotation.

**Keywords:** topic modelling, LDA, doctor-patient communication

*Biomedical Informatics Insights* 2013:6 (Suppl. 1) 39–50

doi: [10.4137/BII.S11661](https://doi.org/10.4137/BII.S11661)

This article is available from <http://www.la-press.com>.

© the author(s), publisher and licensee Libertas Academica Ltd.

This is an open access article published under the Creative Commons CC-BY-NC 3.0 license.



## Introduction and Background

### Therapy communication and outcomes

Aspects of doctor-patient communication have been shown to be associated with patient outcomes, in particular patient satisfaction, treatment adherence and health status.<sup>1</sup> For patients with schizophrenia, non-adherence to treatment is a significant problem, with non-adherent patients having an average risk of relapse that is 3.7 times higher than adherent patients.<sup>2</sup> Some recent work suggests that a critical factor is conversation structure, namely how the communication proceeds. In consultations between out-patients with schizophrenia and their psychiatrists, McCabe et al<sup>3</sup> showed that patients who used more other repair i.e., clarified what the doctor was saying, were more likely to adhere to their treatment six months later. However, outcomes are also affected by the content of the conversation i.e., what is talked about. Using conversation analytic techniques, McCabe et al<sup>4</sup> show that doctors and patients have different agendas, which manifests itself in the topics that they talk about; on the same data, with topics annotated by hand, Hermann et al<sup>5</sup> showed that patients attempt to talk about psychotic symptoms, while doctors focus more on medication issues. Importantly, more talk about medication from the patient increases the patient's chances of relapse in the six months following the consultation.<sup>5</sup>

### Automatic prediction

Our previous research has used machine learning techniques to investigate whether outcomes such as adherence, evaluations of the consultation, and symptoms can be predicted from therapy transcripts using features which can be extracted automatically.<sup>6,7</sup> Findings indicate that high-level features of the dialogue structure (backchannels, overlap etc.) do not predict these outcomes to any degree of accuracy. However, by using all words spoken by patients as unigram lexical features, and selecting a subset based on correlation with outcomes over the training set, we are able to predict outcomes to reasonable degrees of accuracy (c. 70% for future adherence to treatment—see Howes et al<sup>6</sup> for details).

These studies show that some aspects of therapy consultations, which can be extracted automatically (thus removing the need for expert annotation), can enable accurate prediction of outcomes. However, as

the successful features encode specific words spoken by the patient, it is unclear whether they relate to dialogue structure or content, or some combination of the two, and thus help little in explaining the results or providing feedback to help improve therapy effectiveness. It is also unclear how generalizable such results are to larger datasets or different settings, given such specific features with a small dataset. More general models or features may therefore be required.

In this paper, we examine the role and extraction of topic. Topic provides a measure of content more general than lexical word features; by examining its predictive power, we hope to provide generalizable models while also shedding more light on the role of content vs. structure. As content is known to be predictive of outcomes to some extent, identification and tracking of topics covered can provide useful information for clinicians, enabling them to better direct their discussions in time restricted consultations, and aid the identification of patients who may subsequently be at risk of relapse or non-adherence to treatment. However, annotating for topic by hand is a time-consuming and subjective process (topics must first be agreed on by researchers, and annotators subsequently trained on this annotation scheme); we therefore examine the use of automatic topic modelling.

### Topic modelling

Probabilistic topic modelling using Latent Dirichlet Allocation (LDA)<sup>19</sup> has been previously used to extract topics from large corpora of texts, e.g., web documents and scientific articles. A “topic” consists of a cluster of words that frequently occur together. Using contextual clues, topic models can connect words with similar meanings and distinguish between uses of words with multiple meanings (for examples, see Steyvers and Griffiths).<sup>8</sup> LDA uses unsupervised learning methods, and learns the topic distributions from the data itself, by iteratively adjusting priors (see Blei<sup>9</sup> for an outline of the algorithms used in LDA). Such techniques have been applied to structured dialogue, such as meetings<sup>10</sup> and tutoring dialogues<sup>11</sup> with encouraging results.

In the clinical domain, probabilistic topic modelling has been applied to patients' notes to discover relevant clinical concepts and connections between patients.<sup>12</sup> In terms of clinical dialogue, there are few studies which apply unsupervised methods

to learning topic models, though recently this has become an active field of exploration. Angus et al<sup>13</sup> apply unsupervised methods to primary care clinical dialogues, to visualise shared content in communication in this domain. However, their data relies on only six dialogues, with the three training dialogues being produced in a role play situation. It is unclear whether using constructed dialogues as the baseline measure maps reliably to genuine dialogues. Additionally, though they did find differences in the patterns of communication based on how the patient had rated the encounter, their task was a descriptive one, not a predictive one and it is unclear if or how their methodology would scale up, especially given that they selected their testing dialogues on the basis of the patient evaluations.

Cretchley et al<sup>14</sup> applied unsupervised techniques to dialogues between patients with schizophrenia and their carers (either professional carers or family members). Patients and carers were instructed to talk informally and given a list of general interest topics such as sport and entertainment. They split their sample into two pre-defined communication styles (“low- or high-activity communicators”) and described differences in the most common words spoken by each type depending on both the type of carer and the type of communicator. Once again, however, this was a descriptive exercise, on a very small number of dyads, and in choosing to predefine the participants by the amount of communicative activity they undertook they may have missed ways to differentiate between groups of patients that can be extracted from the data, rather than being pre-theoretic.

## Research questions

The preliminary studies outlined above demonstrate some of the issues arising from using unsupervised topic modelling techniques to look at clinical dialogues. One of the main issues is in the interpretation of results. Studies described above used visualizations of the data to find patterns; one question that therefore arises is whether we can usefully interpret “topics” without these—for example, just by examining the most common words in a topic. Another question concerns the limited evidence that different styles of communication can be demonstrated using unsupervised topic modelling, and that these differences have a bearing on, for example, the patient’s

evaluations of the communication or their symptoms. Our main questions here are therefore:

- Does identification of topic allow prediction of symptoms and/or therapy outcomes?
- If so, can automatic topic modelling be used instead of manual annotation?
- Does automatic modelling produce topics that are interpretable and/or comparable to human judgements?

## Data

This study used data from a larger study investigating clinical encounters in psychosis,<sup>3</sup> collected between March 2006 and January 2008. Thirty one psychiatrists agreed to participate, and IRB approval for the study was obtained. Patients meeting Diagnostic and Statistical Manual-IV (APA) criteria for a diagnosis of schizophrenia or schizoaffective disorder attending psychiatric outpatient and assertive outreach clinics in three centers (one urban, one semi-urban and one rural) were asked to participate in the study. After complete description of the study to the subjects, written informed consent was obtained from 138 (40%) of those approached. Psychiatrist-patient consultations were then audio-visually recorded using digital video. The dialogues were transcribed, and these transcriptions, consisting only of the words spoken, form our dataset here. The consultations ranged in length, with the shortest consisting of only 617 words (lasting approximately 5 minutes), and the longest 13816 (lasting nearly an hour). The mean length of consultation was 3751 words.

## Outcomes

Patients were interviewed at baseline, immediately after the consultation, by researchers not involved in the patients’ care, to assess their symptoms. Both patients and psychiatrists filled in questionnaires evaluating their experience of the consultation at baseline, and psychiatrists were asked to assess each patient’s adherence to treatment in a follow-up interview six months after the consultation. The measures obtained are described in more detail below.

## Symptoms

Independent researchers assessed patients’ symptoms at baseline on the 30-item Positive and Negative Syndrome Scale (PANSS).<sup>15</sup> The scale assesses positive, negative and general symptoms and is rated on a scale



of 1–7 (with higher scores indicating more severe symptoms). Positive symptoms represent a change in the patients' behavior or thoughts and include sensory hallucinations and delusional beliefs. Negative symptoms represent a withdrawal or reduction in functioning, including blunted affect, and emotional withdrawal and alogia (poverty of speech). Positive and negative sub-scale scores ranged from 7 (absent)—49 (extreme), general symptoms (such as anxiety) scores ranged from 16 (absent)—112 (extreme). Inter-rater reliability using videotaped interviews for PANSS was good (Cohen's kappa = 0.75).

### Patient satisfaction

Patient satisfaction with the communication was assessed using the Patient Experience Questionnaire (PEQ).<sup>16</sup> Three of the five subscales (12 questions) were used as the others were not relevant, having been developed for primary care. The three subscales were communication experiences communication barriers, and emotions immediately after the visit. For the communication subscales, items were measured on a 5-point Likert scale, with 1 = disagree completely and 5 = agree completely. The four items for the emotion scale were measured on a 7-point visual analogue scale, with opposing emotions at either end. A higher score indicates a better experience.

### Therapeutic relationship

The Helping Alliance Scale (HAS)<sup>17</sup> was used after the consultation to assess both patients' and doctors' experience of the therapeutic relationship. The HAS has 5 items in the clinician version and 6 items in the patient version, with questions rated on a scale of 1–10. Items cover aspects of interpersonal relationships between patients and clinician and aspects of their judgement as to the degree of common understanding and the capability to provide or receive the necessary help, respectively. The scores from the individual items were averaged to provide a single value, with lower scores indicating a worse therapeutic relationship.

### Adherence to treatment

Adherence to treatment was rated by the clinicians as good (>75%), average (25%–75%) or poor (<25%) six months after the consultation. Due to the low incidence of poor ratings (only 8 dialogues), this was converted to a binary score of 1 for good adherence

(89 patients), and 0 otherwise (37). Ratings were not available for the remaining 12 dialogues.

### Hand-coded topics

Hermann et al<sup>5</sup> annotated all 138 consultations for topics. First, an initial list of categories was developed by watching a subset of the consultations. The dialogues were then manually segmented and topics assigned to each segment, with the list of topic categories amended iteratively to ensure best fit and coverage of all relevant topics. A subset of 12 consultations was coded independently by two annotators, such that every utterance (and hence every word) was assigned to a single topic; inter-rater reliability was found to be good using Cohen's kappa ( $\kappa = 0.71$ ). The final list of topics used, with descriptions, is outlined in Table 1.

### Topic Modelling

The transcripts from the same 138 consultations were analysed using an unsupervised probabilistic topic model. The model was generated using the Machine Learning for Language Toolkit (MALLET),<sup>18</sup> using standard Latent Dirichlet Allocation<sup>19</sup> with the notion of document corresponding to the transcribed sequence of words spoken (by any speaker) in one consultation. As is conventional,<sup>20</sup> stop words (common words which do not contribute to the content of the talk, such as 'the' and 'to') were removed. The number of topics was specified as 20 to match the number of topics used by the human annotators (see above), and the default setting of 1000 Gibbs sampling iterations was used. The optimal number of LDA topics for outcome prediction may not be equivalent to the number agreed on by human annotators, however, and future work should investigate different numbers of topics. As an uneven distribution of topics was observed in the hand-coded topic data (see below), automatic hyperparameter optimisation was enabled to allow the prominence of topics and the skewedness of their associated word distributions to vary to best fit the data.

### Interpretation

The resulting topics (probability distributions over words) were then assessed by experts for their interpretability in the context of consultations between psychiatrists and out-patients with schizophrenia. The top 20 most probable words in each topic



**Table 1.** Hand-coded topic names and descriptions.

Topic	Name	Description
01	Medication	Any discussion of medication, excluding side effects
02	Medication side effects	Side effects of medication
03	Daily activities	Includes activities such as education, employment, household chores, daily structure etc.
04	Living situation	The life situation of the patient, including housing, finances, benefits, plans with life etc.
05	Psychotic symptoms	Discussion on symptoms of psychosis such as hallucinations and delusional beliefs
06	Physical health	Any discussion on general physical health, physical illnesses, operations, etc.
07	Non-psychotic symptoms	Discussion of mood symptoms, anxiety, obsessions, compulsions, phobias etc.
08	Suicide and self harm	Intent, attempts or thoughts of self harm or suicide (past and present)
09	Alcohol, drugs and smoking	Current or past use of alcohol, drugs or cigarettes and their harmful effects
10	Past illness	Discussion of past history of psychiatric illnesses, including previous admissions and relapses
11	Mental health services	Care coordinator, community psychiatric nurse, social worker or home treatment team etc.
12	Other services	Primary care services, social services, DVLA, employment agencies, police, housing etc.
13	General chat	Includes introductions; general topics; weather; holidays; end of appointment courtesies
14	Explanation about illness	Patients diagnosis, including doctor explanations and patients questions about their illness
15	Coping strategies	Discussions around coping strategies that the patient is using or the doctor is advising
16	Relapse indicators	Relapse indicators and relapse prevention, including early warning signs
17	Treatment	General and psychological treatments, advice on managing anxiety, building confidence etc.
18	Healthy lifestyle	Any advice on healthy lifestyle such as dietary advice, exercise, sleep hygiene etc.
19	Relationships	Family members, friends, girlfriends, neighbours, colleagues and relationships etc.
20	Other	Anything else. Includes e.g., humour, positive comments and non-specific complaints

were presented to two groups independently—one group of experts in the area of psychiatric research (of whom some members were also involved in developing the hand-coded topics), and one group of experts in the area of communication and dialogue (without specific expertise in the context of psychiatry)—and each group produced text descriptions of the topics they felt they corresponded to. The two groups’ interpretations strongly agreed in 13 of the 20 topic assignments (65%) and partially agreed (i.e., there was some overlap in the interpretations) in a further 3 topic assignments (i.e., in total, 80%).

Having assigned a tentative interpretation to the top word lists for each topic, the two groups

reconvened to examine the occurrences of the topics in the raw transcripts, in order to validate these interpretations within the context of the discussion. Excerpts from the dialogues were chosen on the basis of the proportion of words assigned to each topic in the final iteration of the LDA sampling algorithm. Four excerpts were examined for each of the 20 topics, and a final interpretation for each was agreed upon.

The ease of giving the topic lists of most common words a coherent “interpretation” varied greatly. Some topics were easily given compact descriptions, for example topics 6, 12 and 18, whilst other word lists appeared more disparate. The list of topics and interpretations can be seen in Table 2. Example excerpts

**Table 2.** Interpretations of LDA topics.

Interpretation and top words per topic	
0	<b>Sectioning/crisis</b> remember doctor hospital reason police people memory ring shaking headaches door christmas injection weekly mental murder fit girlfriend locked
1	<b>Physical health</b> —side-effects of medication and other medical issues gp side effects injection panic effect dose pain operation body discuss depot related effective attacks move surgery worse legs
2	<b>Non-medical services</b> —liaising with other services letter health advice letters cpn send dla social called weeks number housing copy gp suppose november living leave green
3	<b>Complaining</b> —negative descriptions of lifestyle etc. people care life cope drug friends dry camera live person rang ring bloody living poor mind mental thing sound
4	<b>Meaningful activities</b> —social functioning beyond the illness setting (e.g., work, study) half find people pills friends lost mental voices talking cancer thinking illness years made progress combination work stone light
5	<b>Making sense of psychosis</b> people things give god talking talk ten reason person depressed understand anti mind family cut watching medicine doctor drugs
6	<b>Sleep patterns</b> sleep day time feel bed bit things hours morning sleeping night mind oclock today years drink bad wake till
7	<b>Social stressors</b> —other people who are stressors or helpful under stress things back place years thought bit ago home put day coming hospital told house felt weeks appointment today week
8	<b>Physical symptoms</b> —e.g., pain, hyperventilating absolutely breathing excellent tea completely music nemesis eating fun oxygen big burning black uncomfortable fast broadcasting breathe movie bottle
9	<b>Physical tests</b> —Anxiety/stress arising from physical tests blood drug thing stress tremor mentioned increase dose car seventy tests test met anxiety keen relaxed felt salt red
10	<b>Psychotic symptoms</b> —e.g., voices, etc. voices hear church voice bad people hearing sister doctor sisters telling felt god heard news evil schizophrenia speaking spirit
11	<b>Reassurance/positive feedback</b> —also possibly progress sort kind things day back thing team remember part stuff work point suppose view lot idea bit sense found
12	<b>Substance use</b> —alcohol/drugs drink drinking alcohol contact support work team money beer cut forms cannabis living cmht sounds craving disability mate missed
13	<b>Family/lifestyle</b> mum money dad brother shopping died enjoy tablets blood bad daughter sister meet living involved bus checking school bother
14	<b>Non-psychotic symptoms</b> —incl. mood, paranoia, negative feelings feel medication feeling thoughts time mood low head past illness control treatment helped increase depression coming paranoid tired happening
15	<b>Medication issues</b> medication issues drugs discuss alcohol raise clear memory level longer problems decision reduce term meet today made injection reason
16	<b>External support</b> —positive social support (e.g., work, family, people) good time people feel bit work lot thing week moment back fine life place happy medication find sort problems
17	<b>Weight management</b> —weight issues in the context of drug side-effects weight stone eat medication gain hospital twelve weigh exercise cut gym putting diet lose appetite reduce reducing walking stuff
18	<b>Medication regimen</b> —dose, timings etc. drug time taking doctor night milligrams hundred months morning tablets stop voices moment bit day dose write long problem
19	<b>Leisure</b> —social relationships/social life etc. thing mates pretty thirty lets high world afternoon seventy front pub speak eleventh lock pm sister weird birthday pension



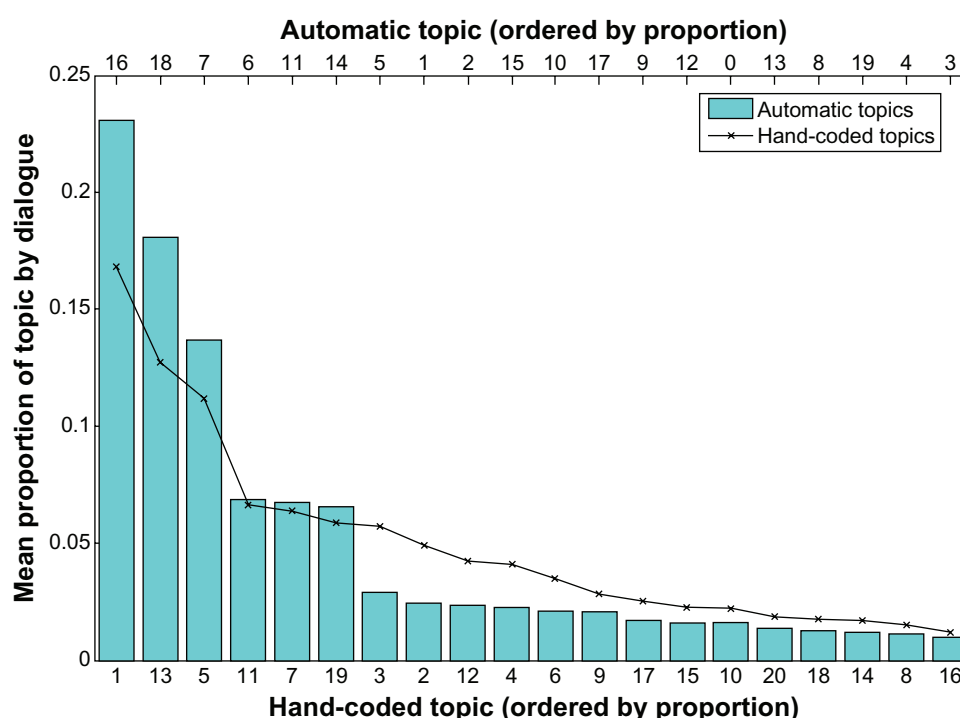
used in interpreting topic 17 (a more coherent topic) and topic 3 (a more disparate topic) can be seen in Table 3 (words assigned to the relevant topic in the final iteration of the LDA sampling algorithm are shown in bold).

## Distribution

Figure 1 shows the distribution of the different topics across the whole corpus; for the automatic LDA version, this is determined from the most likely assignment of observed words to topics. The distribution is highly

**Table 3.** Sample excerpts of LDA topics 17 and 3.

<b>Topic 17: Weight management</b>	
<b>Speaker text</b>	
P	well not really because I put more <b>weight</b> on when I was there
Dr	hehe oh that's not the idea is it
P	no because I was <b>losing</b> it before
Dr	oh
P	because they didn't advocate doing what I was doing they've
Dr	right
P	just said to eat normally but <b>eat smaller portions</b>
Dr	yeah
P	so ah I was <b>putting weight</b> on
Dr	so you actually changed your <b>diet</b> have you
P	yeah
Dr	in order to <b>lose weight</b>
P	yeah
P	when the medication goes down my neck and in my <b>stomach</b> it feeds off the food what I have not the the <b>vegetables</b> but the <b>burger</b> before and it <b>reacts</b> with that really bad
Dr	so you feel there is an <b>interaction</b> between the medication and the food
P	yeah yeah I feel like I'm dying like er like er I going to be dead like I took what I said to the <b>ambulance</b> man the paramedic he said to me the <b>medication</b> you're on you got to watch what you eat that's what he said to me
Dr	uh humm
P	and he said to me as well your <b>blood</b> pressure and your heart is all right but yeah you're <b>putting</b> on you're <b>putting</b> on too much <b>weight</b> like it's <b>expanding</b> outwards
Dr	uh mmm
P	look at my <b>stomach</b> look at that look at this
Dr	mmm have you put on <b>weight</b> recently
P	yeah I <b>weigh</b> fifteen stone now sixteen <b>stone</b>
Dr	right right
<b>Topic 3: Complaining</b>	
<b>Speaker text</b>	
Dr	no NAME it's something we discussed before it's about how you know the the <b>mirroring</b> between how you <b>treat</b> other people and whether that's a <b>model</b> for how I and others should treat you or <b>vice versa</b>
P	no far from it because no that's
Dr	mmm
P	<b>completely garbage</b> because <b>people</b> got haven't got much right to <b>criticise</b> me because I've done a lot to try and <b>make living</b> on my own work the <b>fact</b> that I have put on <b>fourteen</b> stone <b>developed</b> a s serious <b>amount</b> of <b>health</b> problems which are now a <b>life</b> and <b>death</b> risk as NAME said on Friday
Dr	mmm
P	I really do want to <b>die</b> though anybody listening to this I really do but I'm not <b>strong</b> and I'm not <b>strong</b> you know if anybody is videoing this still yeah so a very <b>frightening world</b> we <b>live</b> in I believe the world should be <b>friendly</b> and <b>caring</b> I really do I just I just said to the <b>camera</b> I think the <b>world</b> should be <b>friendly</b> and <b>caring</b>
Dr	mmm mmm
P	there's despair and <b>despondency</b>
Dr	yeah
P	I would go up to ward five this afternoon in the <b>psychiatric</b> hospital and and be <b>assessed</b> but I don't think <b>admittance</b> would
Dr	mmm
P	I mean I just want to be it's just driving me <b>insane</b> having no one to <b>ring</b> and I'm attached to the phone
Dr	mmm
P	it's a <b>dump</b> here



**Figure 1.** Distribution of topics.

skewed, with the largest, topic 16, accounting for about a fifth of all the data, and the smallest, topic 3, only 1.4%. Once stop words had been removed, the corpus consisted of 78,723 tokens. Nearly 18,000 of these (17,957) were therefore most likely to be assigned to topic 16, with just over 1000 (1063) in the smallest topic.

As can be seen from Figure 1, the distribution of automatic topics is consistent with the distribution from the hand-coded topics (Kolmogorov-Smirnov  $D = 0.300$ ;  $p = 0.275$ ). However, it is not clear that the topics themselves correspond so well. For the hand-coded topics, the topic with the highest probability is medication, followed by general chat and then psychotic symptoms; for the LDA topics, the most likely is external support, followed by medication regimen and social stressors—with psychotic symptoms only appearing much further down the list.

### Cross-correlations between hand-coded and automatic topics

We next examined the correspondence between automatic and hand-coded topics directly. Of course, because of the differences in methods, we do not expect these to be equivalent; but examining similarities and differences helps validate or otherwise the interpretations given to the LDA topics, and determine

whether the topics in fact pick out different aspects of the dialogues in each case.

Table 4 shows correlations with coefficients greater than 0.3. Note that this is an arbitrary cut-off point; other smaller significant correlations also exist in the data. These correlations are calculated on the basis of the proportions of each topic in each dialogue; as such, these are overview figures across dialogues and do not tell us about topic assignment at a finer-grained level. For example, we know that highly correlated topics occur in the same dialogues, but not whether they occur in the same sequential sections of those dialogues.

From the data in Table 4 we can see that some of the topics match up well, suggesting that in certain cases the LDA topic model is picking out similar aspects of the content. Examples are the high correlations between the hand and automatically coded substance misuse and physical health topics. Given the relative prominence of the two topics, the high correlation between medication and medication regimen suggests that the LDA topic model is picking out a subset of the talk on medication. This could be linked to the fact that though there may be many different ways of talking about medication (potentially depending on the type of drug, the patient's history etc.) that are understandable to human annotators,

**Table 4.** Correlations between hand-coded and automatic topic distributions.

Hand-coded topic	Automatic topic	r	p
Medication	Medication regimen	0.643	<0.001
Psychotic symptoms	Making sense of psychosis	0.357	<0.001
Psychotic symptoms	Psychotic symptoms	0.503	<0.001
Physical health	Physical health	0.603	<0.001
Non-psychotic symptoms	Sleep patterns	0.376	<0.001
Suicide and self-harm	Weight management	0.386	<0.001
Alcohol, drugs and smoking	Substance use	0.651	<0.001
Mental health services	Non-medical services	0.396	<0.001
General chat	Sectioning/crisis	0.364	<0.001
Treatment	Medication issues	0.394	<0.001
Healthy lifestyle	Weight management	0.517	<0.001
Relationships	Complaining	0.391	<0.001
Relationships	Social stressors	0.418	<0.001
Relationships	Leisure	0.341	<0.001

there is a smaller set of talk about medication which refers to, for example, dosages which is being discovered by LDA topic modelling. Similar considerations may be at play with the link between healthy lifestyle and weight management, and non-psychotic symptoms and sleep issues.

More interestingly the hand-coded psychotic symptoms topic is highly correlated with two automatic topics about psychotic symptoms. Looking at the contexts of these topics, it appears that there may be differences in the ways people talk about their psychotic symptoms depending on whether they are describing the symptoms per se, or looking to make sense of their psychotic symptoms in a wider context.

Interesting differences in the two codings can also be seen in the correlations with relationships, which could illustrate different ways in which they are discussed, both negative (complaining), and positive (leisure). This suggests that the LDA topics are picking up other factors of the communication in addition to the content.

## Prediction of Target Variables

We now turn to examining the association between topics and the target variables we would like to predict: symptoms, doctor and patient evaluations of the therapy, and patient outcomes (specifically, adherence to treatment).

## Correlations with symptoms

Patterns of symptoms are known to affect communication, and we therefore assessed whether

there were correlations between what was talked about, as indexed by hand coded or automatically coded topic, and the three PANSS symptom scales (positive, negative, general).

As can be seen from Table 5 (Table 5 shows correlations above 0.2 only), for the hand coded topics, all three symptom scales were negatively correlated with daily activities (consultations with more ill patients contained less talk about daily activities) and positively correlated with talk about psychotic symptoms. Higher general symptoms were also associated with less talk about healthy lifestyle, and more about suicide and self-harm.

For the automatically extracted topics, consultations with patients with more positive symptoms had more talk in the categories of complaining, making sense of psychosis, physical tests and psychotic symptoms. Consultations with patients with worse negative symptoms had less talk about weight management. As with the hand-coded topics there was some overlap between positive and general symptoms, with general symptoms positively correlated with complaining and making sense of psychosis. These correlations also served as a validation measure of some of the topics, and their interpretations.

## Classification experiments

We performed a series of classification experiments, to investigate whether the probability distributions of topics could enable automatic detection of patient and doctor evaluations of the consultation, symptoms and adherence. In each case, we

**Table 5.** Correlations between symptoms and topics.

Symptom scale	Topic	r	p
<b>Hand-coded</b>			
Positive	Daily activities	−0.249	0.004
	Psychotic symptoms	0.487	<0.001
Negative	Daily activities	−0.211	0.015
	Psychotic symptoms	0.206	0.018
General	Daily activities	−0.254	0.003
	Psychotic symptoms	0.383	<0.001
	Healthy lifestyle	−0.235	0.007
	Suicide and self harm	0.230	0.008
<b>Automatic</b>			
Positive	Complaining	0.265	0.002
	Making sense of psychosis	0.378	<0.001
	Physical tests	0.233	0.007
	Psychotic symptoms	0.316	<0.001
Negative	Weight management	−0.202	0.019
General	Complaining	0.234	0.007
	Making sense of psychosis	0.316	<0.001

used the Weka machine learning toolkit<sup>21</sup> to preprocess data, and a decision tree classifier (J48) and the support vector machine implementation Weka LibSVM<sup>22</sup> as classifiers. Variables to be predicted were binarized into groups of equal size prior to analysis, and for the adherence measure a balanced subset of 74 cases was used. All experiments used 5-fold cross-validation, and the experiments using an SVM classifier used a radial bias function with the best values for cost and gamma determined by a grid search in each case.

Tables 6 and 7 show the accuracy figures for each predicted variable, using a variety of different feature subsets. Doctor factors are the gender and identity of the doctor. Patient factors are the gender and age of the patient, and also the total number of words spoken by both patient and doctor. Topic factors are the total number of words in that topic for the hand-coded topics; and an equivalent value for the automatic topics calculated by multiplying the topic's posterior probability for a dialogue by the total number of words.

From Tables 6 and 7 we can see that there are different patterns of results for the different measures. For the therapeutic relationship (HAS) measures, including doctor factors gives an accuracy of over 70% in all cases, with the identity of the psychiatrist the

most important factor in the decision trees. However, although allowing us a reasonably good fit to the data, the inclusion of the doctor's identity as a feature means that this is not a generalizable result; we would not be able to utilise the information from this factor in predicting the HAS score of a consultation with a new doctor. In this respect, the 65% accuracy when using only the 20 coarse-grained automatic topics is encouraging. In the decision tree, the highest node is social stressors, with a high amount of talk in this category indicating a low rating of the therapeutic relationship from the doctor (66 low/21 high). If there was less talk about social stressors, the next highest node is sleep patterns, with more talk in this area indicating a greater likelihood of a good therapeutic relationship rating (29 high/3 low). Next, more talk about non-psychotic symptoms leads to low ratings (11 low/3 high), and more reassurance leads to a better therapeutic relationship. Interestingly, automatic topics give better accuracy than manual topics when used alone.

For adherence, the best accuracy is achieved by a model which includes doctor features as well as hand-coded topics. Good physician communication is known to increase adherence<sup>23</sup> and in this sample, adherence was also related to the doctor's evaluation of the therapeutic relationship, with 29 of the 37 non-adherent patients rated as having a poor therapeutic relationship by the doctor ( $\chi^2 = 13.364$ ;  $p < 0.001$ ).

Given this, it is surprising that we can predict the therapeutic relationship reasonably well using only automatic topics, but not adherence. Topics also do not appear to give useful performance when predicting patient ratings of the therapeutic relationship (HAS P), or patient evaluations of the consultation (PEQ), although doctor/patient factors seem to have some predictive power. Note that low-level lexical features have shown success in predicting both adherence and patient ratings (Howes et al<sup>6</sup> for example, achieved f-scores of around 70%).

The best predictors for the different types of symptoms are also low, but here the hand-coded topics do better than the automatic topics, with accuracies of 61% for both positive and negative symptoms. For positive symptoms, perhaps unsurprisingly, the decision tree only has one node; if there is more talk on the topic of psychotic symptoms, then the patient is likely to have higher positive symptoms (or vice versa). However, in this respect, especially

**Table 6.** Classification accuracy based on hand-coded topics with different feature groups.

Measure	Topics and Dr/P factors		Topics and P factors		Topics only		Dr/P factors only	
	J48	SVM	J48	SVM	J48	SVM	J48	SVM
HAS Dr	<b>75.8</b>	<b>71.2</b>	47.0	56.8	50.8	56.8	<b>72.0</b>	<b>71.2</b>
HAS P	46.3	49.3	59.0	53.7	50.7	47.0	51.5	52.2
PANSS pos	58.0	59.5	58.8	49.6	<b>61.1</b>	58.0	45.8	59.5
PANSS neg	58.3	59.1	57.6	<b>62.1</b>	<b>61.4</b>	57.6	54.5	52.3
PANSS gen	51.9	55.0	55.0	57.3	55.7	59.5	51.9	53.4
PEQ comm	50.0	56.0	53.7	59.7	55.2	55.2	57.5	<b>61.2</b>
PEQ comm barr	50.7	<b>61.9</b>	56.0	50.7	52.2	52.2	49.3	<b>60.4</b>
PEQ emo	51.2	45.7	47.2	48.0	51.2	49.6	57.5	50.0
Adherence (balanced)	51.4	<b>66.2</b>	47.3	50.0	51.4	44.6	47.3	56.8

**Note:** Accuracy values of over 60% are shown in bold.

given the cross-correlations discussed above, it is surprising that the automatic topics do not allow any prediction of symptoms at above chance levels. For negative symptoms, patients are likely to have more negative symptoms in consultations with little talk on either healthy lifestyle or daily activities.

## Discussion

While both LDA and hand-coded topics seem to have some predictive power, they have different effects for different target variables. Automatic topics do not allow prediction of symptoms, where manual topics do—even though there is a correlation between their corresponding topics relating to psychotic symptoms. This may suggest that LDA used in this way is discovering topics which are a subset of the

manual topics: discussion of symptoms may be wider and include more different conversational phenomena than suggested purely by symptom-related lexical items. On the other hand, LDA topics appear to be better at predicting evaluations of the therapeutic relationship; here, one possible explanation may be that LDA is producing “topics” which capture aspects of style or structure rather than purely content. Further investigation might reveal whether examination of the relevant LDA topics can reveal important aspects of communication style—particularly that of the doctor, given that doctor identity factors also improve prediction of this measure, and are related to patients subsequent adherence.

Although the results from this exploratory study are limited, they are encouraging. We have used only very coarse-grained notions of topics, and a simplistic document-style LDA model, so there is much potential for further research. Using a more dialogue-related model that takes account of topic sequential structure, for example Purver et al<sup>10</sup> or one that can incorporate stylistic material separately to content, as done for function vs. content words by Griffiths and Steyvers<sup>24</sup> should allow us to produce models that better describe the data and can be used to discover more directly what aspects of the communication between doctors and patients with schizophrenia are associated with their symptoms, therapeutic relationship and adherence behaviour.

## Author Contributions

Conceived and designed the experiments: CH, MP, RM. Analyzed the data: CH, MP, RM. Wrote the first

**Table 7.** Classification accuracy based on automatically extracted topics with different feature groups.

Measure	Topics and Dr/P factors		Topics and P factors		Topics only	
	J48	SVM	J48	SVM	J48	SVM
HAS Dr	<b>75.0</b>	<b>75.0</b>	<b>62.9</b>	50.8	<b>65.2</b>	<b>62.9</b>
HAS P	49.3	48.5	50.7	50.7	53.7	47.0
PANSS pos	45.0	58.8	47.3	44.3	51.1	50.4
PANSS neg	50.8	52.3	56.1	56.1	48.5	50.8
PANSS gen	47.3	50.4	52.7	48.9	53.4	48.9
PEQ comm	51.5	56.0	54.5	50.7	56.7	53.7
PEQ comm barr	56.7	<b>60.4</b>	53.7	47.8	51.5	56.0
PEQ emo	57.5	49.6	48.8	51.2	52.8	53.5
Adherence (balanced)	47.3	54.1	47.3	44.6	47.3	51.4

**Note:** Accuracy values of over 60% are shown in bold.





draft of the manuscript: CH. Contributed to the writing of the manuscript: CH, MP, RM. Agree with manuscript results and conclusions: CH, MP, RM. Jointly developed the structure and arguments for the paper: CH, MP, RM. Made critical revisions and approved final version: CH, MP, RM. All authors reviewed and approved of the final manuscript.

## Funding

CH, RM and MP received grants from the Engineering and Physical Sciences Research Council. RM received grants from the Medical Research Council.

## Competing Interests

MP discloses his position as a director of Chatterbox Analytics, a company that supplies a social media analysis software service using language processing techniques with some similarities to those described in this paper. The techniques and results described here are not relevant to the company's area of business, which lies outside the biomedical domain.

## Disclosures and Ethics

As a requirement of publication the authors have provided signed confirmation of their compliance with ethical and legal obligations including but not limited to compliance with ICMJE authorship and competing interests guidelines, that the article is neither under consideration for publication nor published elsewhere, of their compliance with legal and ethical guidelines concerning human and animal research participants (if applicable), and that permission has been obtained for reproduction of any copyrighted material. This article was subject to blind, independent, expert peer review. The reviewers reported no competing interests.

## References

- Ong LM, de Haes JC, Hoos AM, Lammes FB. Doctor-patient communication: a review of the literature. *Soc Sci Med*. Apr 1995;40(7):903–18.
- Fenton WS, Blyler CR, Heinssen RK. Determinants of medication compliance in schizophrenia: Empirical and clinical findings. *Schizophr Bull*. 1997;23(4):637–51.
- McCabe R, Healey PGT, Priebe S, et al. Shared Understanding in Psychiatrist-Patient Communication: Association with Treatment Adherence in Schizophrenia. *Patient Education and Counseling*. 2013.
- McCabe R, Heath C, Burns T, Priebe S. Engagement of patients with psychosis in the consultation: conversation analytic study. *BMJ*. Nov 16, 2002;325(7373):1148–51.
- Hermann P, Lavelle M, Mehnaz S, McCabe R. What do psychiatrists and patients with schizophrenia talk about in psychiatric encounters? In preparation.
- Howes C, Purver M, McCabe R, Healey PGT, Lavelle M. Helping the medicine go down: Repair and adherence in patient-clinician dialogues. In proceeding of: *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2012)*. Paris; Sep 2012.
- Howes C, Purver M, McCabe R, Healey PGT, Lavelle M. Predicting adherence to treatment for schizophrenia from dialogue transcripts. In proceeding of: *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2012 Conference)*. Seoul, South Korea; Jul 2012:79–83.
- Steyvers M, Griffiths T. Probabilistic topic models. In: Landauer T, McNamara D, Dennis S, Kintsch W, editors. *Handbook of Latent Semantic Analysis*. 2007;427(7):424–40.
- Blei D. Probabilistic topic models. *Communications of the ACM*. 2012; 55(4):77–84.
- Purver M, Körding K, Griffiths T, Tenenbaum J. Unsupervised topic modelling for multi-party spoken discourse. In proceedings of: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*. Sydney, Australia; Jul 2006:17–24.
- Arguello J, Rose C. Topic segmentation of dialogue. In *Proceedings of the HLT-NAACL Workshop on Analyzing Conversations in Text and Speech*. New York, NY; Jul 2006.
- Arnold C, El-Saden S, Bui A, Taira R. Clinical case-based retrieval using latent topic analysis. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association; 2010:26.
- Angus D, Watson B, Smith A, Gallois C, Wiles J. Visualising conversation structure across time: Insights into effective doctor-patient consultations. *PLoS One*. 2012;7(6):e38014.
- Cretchley J, Gallois C, Chenery H, Smith A. Conversations between carers and people with schizophrenia: a qualitative analysis using leximancer. *Qual Health Res*. Dec 2010;20(12):1611–28.
- Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophr Bull*. 1987;13(2):261–76.
- Steine S, Finset A, Laerum E. A new, brief questionnaire (PEQ) developed in primary health care for measuring patients' experience of interaction, emotion and consultation outcome. *Fam Pract*. Aug 2001;18(4):410–8.
- Priebe S, Gruyters T. The role of the helping alliance in psychiatric community care: A prospective study. *J Nerv Ment Dis*. Sep 1993;181(9):552–7.
- McCallum AK. MALLET: A machine learning for language toolkit. Available at: <http://mallet.cs.umass.edu>. Accessibility verified May 2013.
- Blei D, Ng A, Jordan M. Latent Dirichlet allocation. *J Machine Learn Res*. 2003;3:993–1022.
- Salton G, McGill M. *Introduction to Modern Information Retrieval*. McGraw-Hill Inc.; 1986.
- Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. *The WEKA Data Mining Software: An update*. *ACM SIGKDD Explorations*. 2009;11(1):10–8.
- EL-Manzalawy Y, Honavar V. WLSVM: Integrating LibSVM into Weka Environment. 2005. Software available at <http://www.cs.iastate.edu/~yasser/wlsvm>.
- Zolnier KB, Dimatteo MR. Physician communication and patient adherence to treatment: a meta-analysis. *Med Care*. Aug 2009;47(8):826–34.
- Griffiths T, Steyvers M. Finding scientific topics. *Proc Natl Acad Sci U S A*. 2004;101(Suppl 1):5228–35.