Extracting Adherence Information from Electronic Health Records

Jordan Sanders,[ɵ] **Meghana Gudala**,[ϕ] **Kathleen Hamilton**,[ɵ] **Nishtha Prasad**,[ɵ]
Jordan Stovall,[ϕ] **Eduardo Blanco**,[ɵ] **Jane E. Hamilton**[ɣ] and **Kirk Roberts**[ϕ]
[ɵ]Department of Computer Science and Engineering, University of North Texas
[ϕ]School of Biomedical Informatics, University of Texas Health Science Center at Houston
[ɣ]McGovern Medical School, University of Texas Health Science Center at Houston
{JordanSanders3,KathleenHamilton,NishthaPrasad}@my.unt.edu eduardo.blanco@unt.edu
{Meghana.Gudala,Jordan.GodfreyStovall,Jane.E.Hamilton,kirk.roberts}@uth.tmc.edu

Abstract

Patient adherence is a critical factor in health outcomes. We present a framework to extract adherence information from electronic health records, including both sentence-level information indicating general adherence information (full, partial, none, etc.) and span-level information providing additional information such as adherence type (medication or nonmedication), reasons and outcomes. We annotate and make publicly available a new corpus of 3,000 de-identified sentences, and discuss the language physicians use to document adherence information. We also explore models based on state-of-the-art transformers to automate both tasks.

1 Introduction

Patient adherence, also known as compliance, is the degree to which a patient follows medical advice. Adherence includes not only taking medications as prescribed, but also using medical devices as instructed, following diet and exercise recommendations, etc. As we shall see, adherence is a critical factor in health outcomes, particularly in patients with chronic illnesses. Patients play the central role following medical advice and communicating adherence to health providers: they largely self-administer, self-regulate and self-report. While insurance claims and filling prescriptions on time may be available as structured data, these sources of information only provide a partial view of patient adherence. Indeed, not going to a specialist or not filling a prescription are sufficient to detect some forms of non-adherence, but doing so is insufficient to guarantee adherence. Additionally, many other forms of non-adherence are common (e.g., skipping medications, taking the wrong dosage, following a diet only occasionally).

From a computational perspective—including both annotation and model building efforts—patient adherence has been primarily modeled as a binary decision, with an emphasis on pinpointing non-adherence. Adherence, however, ranges from absolute non-adherence to full adherence. Partial adherence and non-adherence can be due to many factors. Adherence models by medical experts include intentional and unintentional non-adherence (Ng et al., 2014), social support (Simoni et al., 2006) and other patient attributes such as age and time since diagnosis (Weaver et al., 2005).

In this paper, we extract medication adherence information from electronic health records. Natural language processing is a requirement to solve this problem, as physicians record self-reported patient adherence (or non-adherence) in unstructured free text. We identify not only whether a patient has adhered fully or not at all to medical advice, but also partial adherence as well as doctors reviewing or instructing patients about treatments and the importance of future adherence. In addition, we target for the first time information to better understand adherence. This information includes adherence type (medication or non-medication), the source of adherence information (the patient, a care giver, etc.), reason of non-adherence (cost, intolerance, forgetfulness, etc.), speculation (doesn't remember, is not sure, etc.) and negation (forgot, stopped, etc.) and others. The main contributions of this paper are:[1]

[1]The corpus has been approved for public release by an Institutional Review Board and will be made available at https://github.com/irpepper/EHR_adherence along with the full annotation guidelines and code to replicate our experiments.

Proceedings of the 28th International Conference on Computational Linguistics, pages 680–695
Barcelona, Spain (Online), December 8-13, 2020

- a corpus of 3,000 de-identified sentences from electronic health records and annotations indicating which adherence information is discussed (full, partial, none, review) and additional information including 10 attributes (medication, non-medication, source, target, reason, outcomes, etc.);
- analysis discussing the language doctors use to record adherence information; and
- experimental results showing that transformer models can partially automate this task, and error analyses providing insights into when the task is most challenging.

2 Background and Previous Work

Electronic health records have emerged in the last decade as a standard in healthcare (Jha et al., 2009; Henry et al., 2018). Studies have shown that they are beneficial both in small practices and large organizations (Buntin et al., 2011). While electronic health records are notoriously difficult to annotate and mine for information (Chapman et al., 2011; Friedman et al., 2013), modern machine learning and in particular deep learning has seen many successes (Shickel et al., 2018). Some of the remaining challenges include the lack of a universal standard and the inherent difficulties of working in the medical domain (abbreviations, jargon, incomplete sentences, etc.)

The Consequences of Non-Adherence. Among individuals with chronic illnesses, medication non-adherence is an important contributor to (a) poor patient outcomes related to increased morbidity and mortality, and (b) increased healthcare costs related to avoidable hospitalizations and emergency department visits (Ho et al., 2006; Yang et al., 2009; Asche et al., 2011). Non-adherence is associated with \$100–\$300 billion of avoidable health care costs (3-10% of the total) in the US annually (Viswanathan et al., 2012; Iuga and McGuire, 2014). Despite these facts, studies have shown that up to 50% of patients in the US with chronic illnesses stop taking their medications within one year of being prescribed (Lee et al., 2006), and over 31% of prescriptions are not filled within 9 months (Tamblyn et al., 2014).

Adherence and Electronic Health Records. Previous work has measured medication adherence objectively using medication event monitoring systems (Wu et al., 2008). To our knowledge, the work by Turchin et al. (2008) is the first to extract adherence from physician notes. Their system, however, is based on 87 heuristics using pattern matching, and their evaluation considers only 82 notes.

Physician-patient transcripts have been studied as a source of information to predict non-adherence. For example, Howes et al. (2012) conclude that unigrams in psychiatrist-patient transcripts are good predictors of future adherence to treatment for schizophrenia (as determined by the physician), and Howes et al. (2013) investigate the role of topics automatically identified. More recently and still using traditional machine learning, Wallace et al. (2014) automatically identify whether utterances discuss adherence barriers. Similar to our sentence-level annotations, all these previous works consider adherence as a text classification task. That is, they assign one label to a piece of text (e.g., a full dialogue transcript, sentence, utterance). Unlike them, we also consider span-level adherence information including adherence type (medication or non-medication), the reasons and outcomes for adhering (or not adhering), etc.

Social media has also been studied as a source of adherence information, Onishi et al. (2018) find that out of 400 tweets mentioning drugs, 9 tweets indicate non-adherence and 6 include the reason for not adhering (e.g., adverse effects). They do not, however, attempt to automatically extract any adherence information. Moseley et al. (2020) present a corpus (1,102 discharge summaries and 1,000 nursing progress notes) annotated with 13 patient phenotypes, one of which is non-adherence. They also present results with a CNN, but do not report results identifying non-adherence. Their definition of non-adherence is equivalent to our *partial* and *none* adherence (Section 3), and like the other works described above, they reduce the problem to text classification and disregard span-level information. Deep learning approaches have been proven useful to extract information from electronic health records beyond adherence. For example, Jagannatha and Yu (2016) show that BiLSTM networks are useful for medical event detection, and Rosenthal et al. (2019) experiment with transfer learning, GRUs and BERT to identify sections (e.g., allergies, chief complaint, examination, family history, procedures, etc.) in electronic health records.

Sent.-Level	Span-Level
FULL	1: [By report]$^{\text{SOURCE}}$, there is good compliance treatment, [good tolerance of treatment and poor symptom control]$^{\text{OUTCOME}}$.
PARTIAL	2: [Pt]$^{\text{SOURCE}}$ [unable to tolerate]$^{\text{NEGATION}}$ [prolixin]$^{\text{MEDICATION}}$ [due to restlessness]$^{\text{REASON}}$ and preferred [zyprexa]$^{\text{MEDICATION}}$ which he complied with.
NONE	3: [Pt reports]$^{\text{SOURCE}}$ he [hasn't taken]$^{\text{NEGATION}}$ [meds]$^{\text{MEDICATION}}$ [since *phi_date*]$^{\text{TIME}}$ and [is not taking]$^{\text{NEGATION}}$ [meds]$^{\text{MEDICATION}}$ [at this time]$^{\text{TIME}}$.
REVIEW	4: [The patient]$^{\text{TARGET}}$ was counseled regarding risk factor reductions and importance of compliance with treatment.
UNKNOWN	5: Treatment has always been minimal because [she]$^{\text{TARGET}}$ was [not interested in taking]$^{\text{NEGATION}}$ [medications]$^{\text{MEDICATION}}$ [to control her]$^{\text{REASON}}$ [arthritis]$^{\text{PROBLEM}}$.

Table 1: Examples of sentence-level and span-level adherence annotations. Additional examples can be found in the supplementary materials.

3 Annotating Adherence

Large corpora of electronic health records are not publicly available because of privacy concerns thus we work with new sentences. We summarize below the final annotation guidelines. The guidelines were refined after several iterations with pilot annotations and discussions with the annotators.

Source sentences We work with sentences retrieved from real electronic health records of patients suffering from diabetes or mental health disorders. The criterion to select sentences consisted on checking for a set of keywords likely to discuss adherence: *adhere*, *adhered*, *adherence*, *adhering*, *compliance*, *complied*, *taking medications* and *taking meds*. A manual process ensured that all protected health information (e.g., patients' names, hospitals, locations, dates) was replaced with dummy tokens before the annotation process started. We selected and de-identified 3,000 sentences following these steps.

Sentence-Level Annotations The fist annotation task is to determine whether the patient adheres to the treatment prescribed by the doctor generating the electronic health record. We use five labels:

- FULL: patient is completely adherent to the treatment as prescribed by the doctor;
- PARTIAL: patient is following some of the treatment, but not exactly as prescribed by the doctor;
- NONE: patient is not following the treatment at all;
- REVIEW: either (a) patient received guidance about how to adhere to treatment or (b) the doctor reviews the prescribed treatment, but the patient's adherence (or lack thereof) is not discussed; or
- UNKNOWN: the sentence is ambiguous, there is not enough information to choose another label.

FULL adherence requires not only medication adherence, but also adherence to other treatment aspects such as diet, not lifting weights, or resting. PARTIAL adherence includes patient adherence ranging from just above NONE adherence to almost FULL adherence regardless of the reason: forgetting, intolerance, etc. We use REVIEW to indicate sentences describing instructions to patients that do not report on the patient's adherence. Tables 1 and 5 exemplify the five sentence-level adherence labels. Annotators could also discard sentences not discussing medical adherence, e.g., *Umbilical cord is still adhered.* Annotators discarded only 223 sentences (6.9%) because they did not discuss medical adherence.

Span-Level Annotations Regardless of the sentence-level annotations, we further annotate sentences with spans indicating ten attributes related to adherence. These more detailed annotations could be described as slots for an *adherence* event or following FrameNet (Baker et al., 1998), the frame elements of an instance of the *adherence* frame. We use ten span-level labels:

- MEDICATION: medications that are part of a treatment;
- NON-MEDICATION: any form of treatment other than medications (e.g., diet, exercise);
- NEGATION: phrases indicating that the patient is non-adherent to the treatment;
- SPECULATION: phrases indicating uncertainty by the doctor;
- REASON: explanation or justification for non-adherence (e.g., forget, financially unable, unable to travel to pharmacy, adverse effects) or (rarely) adherence (e.g., encouragement by family);

		Annot. Agreement			Adjud. Agreement		
	Freq. (%)	Cohen's κ			Cohen's κ		
Sentence-Level Adherence Information							
FULL	21	0.78			0.75		
PARTIAL	31	0.55			0.77		
NONE	7	0.81			0.85		
REVIEW	30	0.81			0.92		
UNKNOWN	11	0.23			0.56		
All	100	0.73			0.79		
	Freq. (%)	P	R	F1	P	R	F1
Span-Level Adherence Information							
MEDICATION	68	0.93	0.92	0.92	0.97	0.96	0.97
NON-MEDICATION	41	0.81	0.65	0.72	0.89	0.75	0.81
NEGATION	40	0.81	0.90	0.85	0.91	0.89	0.90
SPECULATION	1	0.65	0.37	0.47	1.00	1.00	1.00
REASON	6	0.63	0.67	0.65	0.60	0.60	0.60
OUTCOME	14	0.41	0.64	0.64	0.88	0.58	0.70
SOURCE	23	0.68	0.70	0.69	0.97	0.62	0.76
TARGET	55	0.72	0.83	0.77	0.82	0.76	0.79
TIME	22	0.69	0.57	0.62	0.76	0.80	0.78
PROBLEM	28	0.72	0.65	0.68	0.73	0.73	0.73

Table 2: Label frequencies (percentage of sentences with each sentence-level and span-level label) and agreements. We detail agreements in the annotation and adjudication phases.

- OUTCOME: results of non-adherence (e.g., lack of recovery) or adherence (e.g., full recovery, adverse effect patient had to deal with in order to be adherent);
- SOURCE: person who is reporting adherence (usually the patient, but also relatives and care givers);
- TARGET: individual responsible for adherence (e.g., the patient, a care giver);
- TIME: phrases expressing time and related to adherence;
- PROBLEM: phrases explaining the need for medication of treatment (e.g., illness, symptoms).

Tables 1 and 6 provide examples, and we briefly summarize important criteria from the annotation guidelines. First, multiple smaller spans are preferred to one larger span covering several individual elements (e.g., *[the patient]*$^{\text{SOURCE}}$ *and [patient's daughter]*$^{\text{SOURCE}}$ *[...]*). Second, MEDICATION includes specific medications as well as general medications (e.g., *[Symbicort]*$^{\text{MEDICATION}}$ *and [cough meds]*$^{\text{MEDICATION}}$ *[...]*). Third, if the SOURCE and TARGET are the same entity (e.g., the *patient*, Example 2), we only annotate the SOURCE. Additionally, we only annotate the first occurrence of the TARGET. Fourth, SOURCE can be human (e.g, *she (the patient)* in Example 5) or non-human (e.g., *[medical history indicates [...]]*$^{\text{SOURCE}}$. Fifth, TIME indicates a temporal expression answering *when* adherence (or non-adherence) took place including absolute (e.g., *[since February]*$^{\text{TIME}}$) and relative (e.g., *[since last visit]*$^{\text{TIME}}$) expressions. Sixth, NEGATION includes not only explicit negations such as *[has not taken]*$^{\text{NEGATION}}$ *[meds]*$^{\text{MEDICATION}}$ but also more nuanced negations such as *[limited adherence]*$^{\text{NEGATION}}$ and *has been [missing]*$^{\text{NEGATION}}$ *[simvastatin]*$^{\text{MEDICATION}}$ *[occasionally]*$^{\text{TIME}}$.

Annotation Process The annotation process consisted of two phases and involved four individuals with complementary expertise. In the first phase, two natural language processing practitioners completed both annotation tasks independently. In the second phase, a medical scribe and a doctor adjudicated the annotations resulting from the first phase. In the annotation phase, both annotators annotated each sentence independently. In the adjudication phase, both adjudicators adjudicated 20% of sentences, and the remaining 80% were adjudicated by one adjudicator. We discuss agreements in Section 4.

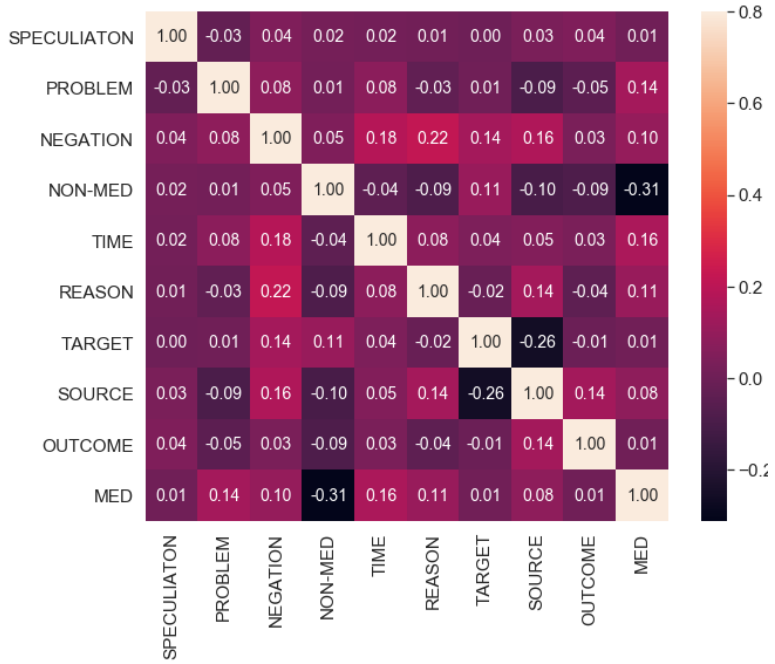

Figure 1: Heatmap of the correlations of observing two span types in the same sentence.

4 Corpus Analysis

4.1 Sentence-Level Annotations

The top block of of Table 2 presents the distribution of sentence-level labels and agreements. Annotators could make a decision with the vast majority of sentences (UNKNOWN: 11%). Surprisingly, we found that doctors often record in electronic health records when they REVIEW treatments and emphasize the importance of adherence (30%). This may possibly signal lack of adherence or that a doctor suspects future non-adherence and wishes to address the (potential) issue. Complete non-adherence is rare (7%), and patients partially adhering is the most common case (31%), followed by FULL adherence (21%).

We calculate agreements with Cohen's κ (Cohen, 1960) in order to discount the probability of agreeing by chance. Inter-annotator and especially inter-adjudicator agreements are high (All, κ: 0.73 and 0.79). Cohen's κ inter-adjudicator coefficients range between 0.56 and 0.92, and are above 0.75 with all labels except UNKNOWN, which is the least frequent label. κ in the 0.6–0.8 range indicate substantial agreements, and over 0.8 (nearly) perfect (Artstein and Poesio, 2008).

4.2 Span-Level Annotations

The bottom block of Table 2 presents the frequency of spans per sentence and agreements. Most sentences discussing adherence mention a MEDICATION (68%), and surprisingly, a substantial amount also mention NON-MEDICATION (41%). Many sentences also include sentences indicating non-adherence (NEGATION, 40%), and around a quarter discuss TIME (22%), the SOURCE reporting adherence (23%) or the PROBLEM requiring treatment (and adherence, 28%). These counts show that our span-level adherence annotations provide additional information to sentence-level annotations.

Calculating inter-annotator agreements of span-level annotations is more involved than sentence-level annotations (a span covering a sequence of tokens vs. one label per sentence). We follow previous work to consider partial matches in the agreement calculations (Chinchor and Sundheim, 1993; Collins et al., 2016). More specifically, we calculate Precision, Recall and F1 of the token level annotations after marking each token as correct, incorrect, partial, spurious or missing. Unlike the aforecited works, we consider a token as a partial match as long as it overlaps with the gold annotations, regardless of whether the last token of the span is the same. We observe again high inter-annotator and inter-adjudicator agreements (F1: 0.6–1.0, all above 0.7 except REASON, which is present in only 6% of sentences).

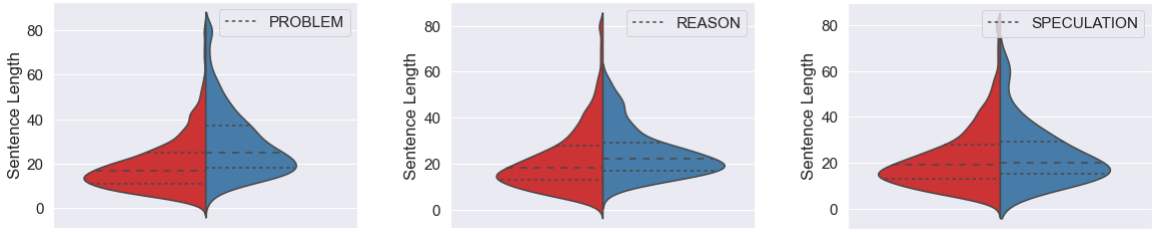

Figure 2: Distribution of sentence lengths depending on whether selected span-level annotations are present. Red (left) and blue (right) indicate that the span-level annotation is not present and is present respectively.

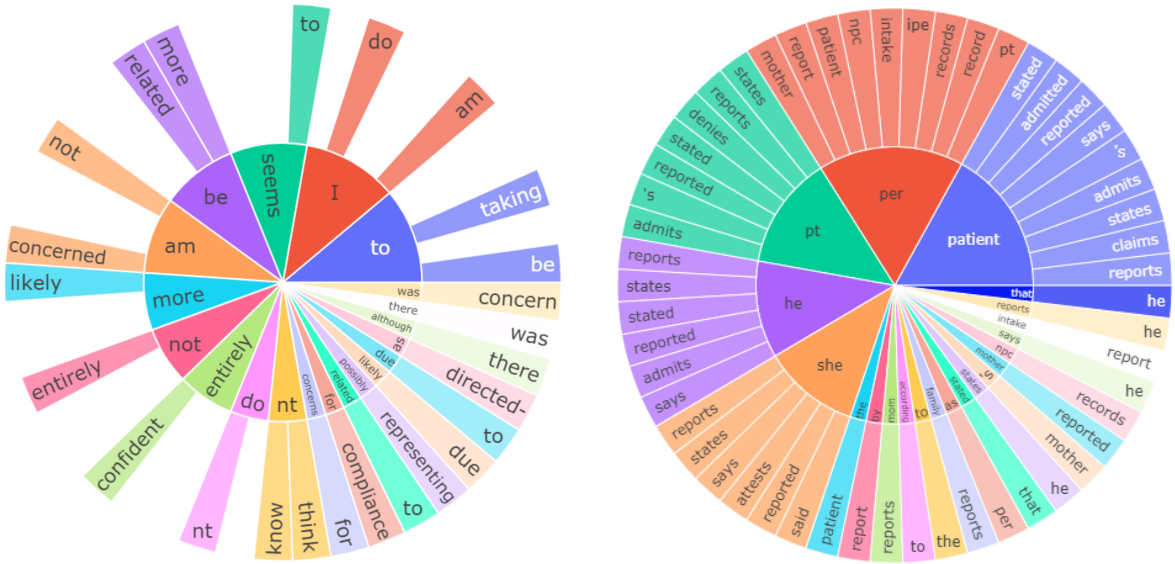

Figure 3: Most common bigrams in SPECULATION (left) and SOURCE (right) spans.

Span Correlations Figure 1 presents a heatmap of the correlation of observing two span-level annotations in the same sentence. Specifically, we generate the heatmap taking into account the presence of span-level annotations and as opposed to the count. We observe that doctors tend to not discuss MEDICATION and NON-MEDICATION adherence (-0.31), or SOURCE and TARGET information in the same sentence (-0.26). On the other hand, they tend to use NEGATION to discuss REASON (0.22), TIME (0.18) and SOURCE (0.16), and MEDICATION correlates with TIME (.16) and PROBLEM (0.14).

Spans and Sentence Length Intuitively, longer sentences have more span-level annotations and certain span-level annotations include more tokens than others. The plots in Figure 2 confirm this intuition. For example, sentences that include PROBLEM, REASON and SPECULATION tend to be longer than sentences that do not contain these spans, as describing this information requires more tokens than, for example, introducing negation. Sentence length is roughly uniform regardless of whether other span-level annotations are present. The supplementary materials provide similar plots for all span-level annotations.

Spans and Common Words Figure 3 plots the most common bigrams (i.e., sequences of two words) for two span-level annotations: SPECULATION and SOURCE. We observe that doctors document SPECULATION in first person (*I do*, *I am*), with verbs bringing up uncertainty (*seems*, *think*), and speculative adverbs (*entirely confident*, *likely*, *possibly*). We also note that negation is quite common to indicate SPECULATION: *not entirely*, *(do)n't think*, *(do)n't know*, *(I) am not*. Regarding SOURCE, we observe that the patient is rarely named (*PHI_PERSON*, de-identified). Rather, they refer to the SOURCE with *patient* or the shorthand *pt*, personal pronouns (*She*, *he*), and a long list of sources that are not the patient: *per record*, *family*, *mother*, *mom*, etc. Additionally, common communication verbs are used to indicate the

	BERT-base			BERT-large			DistillBERT			ClinicalBERT		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
FULL	0.78	0.67	0.72	0.75	0.74	0.74	0.82	0.71	0.76	0.83	0.82	0.83
PARTIAL	0.48	0.52	0.50	0.63	0.55	0.59	0.57	0.56	0.57	0.55	0.58	0.57
NONE	0.83	0.94	0.88	0.86	0.96	0.91	0.82	0.96	0.89	0.86	0.96	0.91
REVIEW	0.93	0.90	0.91	0.85	0.86	0.85	0.92	0.85	0.89	0.93	0.88	0.91
UNKNOWN	0.33	0.28	0.31	0.38	0.19	0.25	0.35	0.34	0.35	0.47	0.25	0.33
All (Micro Avg.)	0.78	0.79	0.78	0.78	0.80	0.79	0.80	0.80	0.80	0.82	0.83	0.82

Table 3: Results obtained classifying sentences into their sentence-level adherence information. The last system is the only one using pretraining specific to the clinical domain.

SOURCE: *says*, *states*, *reports*, *indicates*, *admits*, *attests*, *claims*, etc.

The supplementary materials include sunbursts for the remaining span-level annotations. We note that most NEGATION spans do not include common negation cues such as *not*, *non* or *never*. Instead, doctors commonly document NEGATION with verbs and adjectives that indicate negation in a nuanced manner: *stopped taking*, *unable to (take, follow, etc.)*, *prefers to*, *been skipping*, *minimal compliance*, etc.

5 Experiments and Results

We create stratified train, development, and test splits (75/5/20) and experiment with transformers to automate both annotation tasks. More specifically, we work with four versions of BERT (Devlin et al., 2019). Three versions are pre-trained with general English: the base and large BERT models (110M and 340M parameters) and a smaller version (65M parameters) shown to be as effective, DistillBERT (Sanh et al., 2019). The fourth transformer, ClinicalBERT (Si et al., 2019), is pretrained in the clinical domain. We use the transformers package by Wolf et al. (2019), keras (Chollet and others, 2015), and ktrain (Maiya, 2020) to tune the models with the train and development splits.

Sentence-Level Annotations The models for sentence-level classification are simple: a transformer to obtain a distributed representation, 0.2 dropout, and a fully connected layer of size 5 with softmax activation to make the classification. Table 3 shows the results obtained with each transformer on the test split after training and tuning with the train and development splits. It is difficult for humans to draw non-binary conclusions about adherence given a single sentence. Therefore, as intuition would suggest, PARTIAL and UNKNOWN are the most challenging labels for the model to identify correctly. Across all four transformer models, these labels have F1-measures significantly below that of the less broad FULL, NONE, and REVIEW labels, even with NONE representing only 7% of the dataset. ClinicalBERT yields the best results by a small margin (F1: 0.82 vs. 0.78–0.80), particularly in sentences annotated with FULL label (F1: 0.83 vs. 0.72–0.76). As adherence is related to improved overall health outcomes, it is likely that ClinicalBERT's domain knowledge is able to interpret medical jargon related to outcomes better. This advantage is noticeably lacking in the NONE adherence label, but negation already provides significant indication in this case, and all transformers obtain similar results (F1: 0.88–0.91).

Span-Level Annotations The models for predicting span-level annotations are slightly more complex. We use a transformer to obtain contextualized word embeddings, and then use a BiLSTM (Graves and Schmidhuber, 2005) of size 100, 0.5 dropout and an additional fully connected layer of size 40 with softmax activation to train a sequence-to-sequence model that outputs span-level annotations. We use the BILOU standard (Ratinov and Roth, 2009) to represent spans. Table 6 shows the results obtained with each transformer. All models achieve the same precision and recall (0.87), though there are differences amongst span performance. We note, however, that BERT-large is the only one to detect SPECULATION. As SPECULATION occurs infrequently within the dataset (1% of sentences), the larger model size—featuring more than three times as many parameters as the other transformers—is the only rich enough language model to learn any correlation given so few instances. More importantly, we note that

	BERT-base			BERT-large			DistillBERT			ClinicalBERT		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
MEDICATION	0.91	0.87	0.89	0.92	0.88	0.90	0.91	0.87	0.89	0.90	0.89	0.90
NON-MEDICATION	0.79	0.79	0.79	0.77	0.79	0.78	0.76	0.81	0.79	0.76	0.79	0.78
NEGATION	0.85	0.86	0.86	0.85	0.86	0.85	0.85	0.85	0.85	0.87	0.84	0.85
SPECULATION	0.00	0.00	0.00	0.29	0.29	0.29	0.00	0.00	0.00	0.00	0.00	0.00
REASON	0.60	0.57	0.58	0.68	0.60	0.64	0.64	0.62	0.63	0.70	0.64	0.67
OUTCOME	0.54	0.44	0.49	0.55	0.44	0.49	0.55	0.44	0.49	0.63	0.47	0.54
SOURCE	0.87	0.75	0.81	0.85	0.80	0.82	0.83	0.82	0.83	0.85	0.80	0.82
TARGET	0.86	0.86	0.86	0.83	0.84	0.83	0.84	0.85	0.85	0.84	0.85	0.85
TIME	0.63	0.68	0.65	0.63	0.70	0.66	0.61	0.70	0.65	0.65	0.68	0.67
PROBLEM	0.71	0.70	0.70	0.71	0.70	0.70	0.72	0.73	0.72	0.71	0.74	0.73
All (Micro Avg.)	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87

Table 4: Results obtained identifying span-level adherence information. The last system is the only one using pretraining specific to the clinical domain.

Error Type	%	Gold	Predicted	Example
Uncertainty	25	UNKNOWN	FULL	Taking tylenol gives good relief to episodes, does not like taking meds, he needs 2 tylenol in a week.
Implication	20	PARTIAL	NONE	He admits that he has not returned to his [. . .] and is not always consistent with taking medications.
Contrast	16	PARTIAL	FULL	He has excellent adherence, though he reports missing 1-2 doses a month.
Medical Terms	11	UNKNOWN	PARTIAL	NIHSS:3 , TPA: y, BNS:6, HPHQ :1 Dc meds: 7, no antidepressant pre and [. . .], Pre MRS:0: , MRS 3:,ESS:6, TPHQ 0: Med adherence:8
Short Sents.	9	UNKNOWN	REVIEW	Question adherence to therapy.

Table 5: Most frequent error types predicting sentence-level adherence information.

ClinicalBERT stands out predicting two important adherence spans: REASON (F1: 0.67 vs. 0.57–0.64) and PROBLEM (F1: 0.73 vs. 0.70–0.72).

6 Error Analysis

We conduct a manual error analysis of a sample of 100 sentences from the test set in order to identify the most commons error types made by the best models in each task. We identified well-known linguistic phenomena that are challenging in any domain as well as specific issues present in electronic health records. Our future plans include improving the models based on these insights.

Sentence-Level Adherence Information. We exemplify the most common error types predicting sentence-level annotations in Table 5. We observe ***uncertainty*** in 25% of errors; in the example, it is unclear if taking tylenol was prescribed by a doctor or it is simply describing the patient's background and medical history. Other *uncertainty* errors include lack of information (e.g., *The patient is taking medications, but did not bring list and does not remember name of the medications.*). Solving 20% of errors would require making ***implications***, i.e., inferring adherence information that can be drawn from the sentence even though it is not explicitly stated. In the example, *not always consistent* implicates *sometimes consistent* and thus PARTIAL adherence. Other examples include *helps him adhere to his medications' schedule better* (so he is already partially adhering), *would like to see increased compliance* (so there is already some compliance), and *work together to improve adherence to CPAP* (so there is already some adherence). In 16% of errors, the sentence contains a ***contrast***: the doctor first states FULL adherence (e.g., *excellent adherence*) and then qualifies this statement to indicate PARTIAL adherence (e.g., *reports missing 1-2 doses a month*). Other *contrast* errors include *She reports compliance with medicine but has*

Error Type	%	Example
Long Phrase or Clause (OUTCOME)	62	*Gold:* his [fev1 today is significantly improved]$^{\text{outcome}}$ and he is increasing his medical adherence. *Pred.:* his fev1 today is [significantly]$^{\text{outcome}}$ improved and he is increasing his medical adherence.
Infrequent Therapies (NON-MEDICATION)	68	*Gold:* __phi_location__ eval of cpap machine and trial of [nasal pillows]$^{\text{NON-MEDICATION}}$ to fit for better adherence. *Pred.:* __phi_location__ eval of cpap machine and trial of nasal pillows to fit for better adherence.
Generic MEDICATION	53	*Gold:* [...] start having this pain again, patient state he tried [soma-muscle relexant]$^{\text{MEDICATION}}$ of his wife with no improvement [...] *Pred.:* start having this pain again, patient state he tried [soma-muscle]$^{\text{PROBLEM}}$ relexant of his wife with no improvement [...]
Unspecific TIME	39	*Gold:* taking amitriptyline for fibromyalgia, but feeling sleepy and drowsy when taking meds; only took [occasionally]$^{\text{TIME}}$. *Pred.:* taking amitriptyline for fibromyalgia, but feeling sleepy and drowsy when taking meds; only took occasionally.

Table 6: Most frequent error types depending on gold span-level adherence information. Error types are sorted by decreasing frequency in the test set (e.g, errors involving OUTCOME are the most frequent). We only show gold and predicted labels for the error being exemplified to improve readability.

been missing simvastatin occasionally. A less frequent error (11%) occurs when adherence information is buried in a long list of ***medical terms***. The example in Table 5, additionally, indicates adherence with a numeric score. Finally, our analysis identifies ***short sentences*** as the cause of 9% of errors. In the example, the model is mislead by the communication verb *question*(s), which appears to indicate REVIEW although there is not enough information to make this judgment.

Span-Level Adherence Information. Most sentences discussing adherence have more than one span, and we conduct an error analysis to analyze the most common errors depending on the gold label (Table 6). Most errors mislabeling OUTCOMEs (62%) occur when the gold outcome is a ***long phrase or clause*** and the model misses part of it. In the example, the adverb *significantly* is the only token identified as outcome. The most common error regarding NON-MEDICATIONs is to miss them altogether (68%). The model is quite effective at identifying non-medication treatments related to diet and exercising regardless of the specific wording (e.g., avoid carbs, eliminate fat), but it often misses ***infrequent therapies*** (e.g., nasal pillows). Mentions to ***generic medications***, unlike specific drugs such as methyldopa, hydrocodone and benzopril, are the most common source of errors mislabeling MEDICATIONs. TIME spans are most challenging (39% of errors) when they refer to ***unspecific temporal information*** (e.g., occasionally, current). Another common error with TIME is identifying temporal expressions related to when *medical treatment* was prescribed, not when adherence (or non-adherence) took place.

7 Conclusions

Patient adherence is critical to positive health outcomes, especially for patients with chronic illnesses. Non-adherence results in $100–$300 billion of avoidable healthcare costs annually in the US alone (Viswanathan et al., 2012; Iuga and McGuire, 2014). Doctor-patient communication is critical for adherence, but ultimately patients have the responsibility to follow medical advice: they largely self-administer, self-regulate and self-report. Crucially, patient adherence is documented by physicians in electronic health records using unstructured free text. Thus, extracting adherence information requires natural language processing.

Previous work on adherence worked primarily with physician-patient transcripts. Unlike them, we work with electronic health records as generated by physicians, which are commonplace in small practices and large organizations. Beyond determining whether patients are fully adherent or not (a binary

decision), like previous work does, we differentiate between FULL and PARTIAL adherence, and no adherence at all (NONE). We also identify when doctors document themselves reviewing medical treatments with patients in order to improve adherence. More importantly, in addition to approaching adherence as a text classification task at the sentence level, we also consider span-level information including adherence type (MEDICATION or NON-MEDICATION), REASONs and OUTCOMEs. Our annotation effort (3,000 sentences) shows that sentences often contain more than one span.

The work presented here opens the door to applications that would benefit from extracting adherence information from electronic health records—the largest source of physician-generated patient records. For example, it enables studies to identify the most common reasons and outcomes of non-adherence, and tools to anticipate potential patient non-adherence so that medical staff can apply an intervention.

Acknowledgements

Research reported in this article was partially funded through a Patient-Centered Outcomes Research Institute® (PCORI®) Award (PCORI/ME-2018C1-10963). The views, statements and opinions presented in this article are solely the responsibility of the author and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute® (PCORI®), its Board of Governors or Methodology Committee.

References

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Carl Asche, Joanne LaFleur, and Christopher Conner. 2011. A review of diabetes treatment adherence and the association with clinical and economic outcomes. *Clinical therapeutics*, 33(1):74–109.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada, August. Association for Computational Linguistics.

Melinda Beeuwkes Buntin, Matthew F Burke, Michael C Hoaglin, and David Blumenthal. 2011. The benefits of health information technology: a review of the recent literature shows predominantly positive results. *Health affairs*, 30(3):464–471.

Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D'Avolio, Guergana K Savova, and Ozlem Uzuner. 2011. Overcoming barriers to nlp for clinical text: the role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association: JAMIA*, 18(5):540.

Nancy Chinchor and Beth Sundheim. 1993. MUC-5 evaluation metrics. In *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*.

François Chollet et al. 2015. Keras. `https://keras.io`.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Sarah A Collins, Emily Gesner, Perry L Mar, Doreen M Colburn, and Roberto A Rocha. 2016. Prioritization and refinement of clinical data elements within ehr systems. In *AMIA Annual Symposium Proceedings*, volume 2016, page 421. American Medical Informatics Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Carol Friedman, Thomas C Rindflesch, and Milton Corn. 2013. Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the national library of medicine. *Journal of biomedical informatics*, 46(5):765—773, October.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.

J Henry, Y Pylypchuk, T Searcy, and V Patel. 2018. Adoption of electronic health record systems among us non-federal acute care hospitals: 2008-2015. 2016. *ONC Data Brief*, 35.

P Michael Ho, John S Rumsfeld, Frederick A Masoudi, David L McClure, Mary E Plomondon, John F Steiner, and David J Magid. 2006. Effect of medication nonadherence on hospitalization and mortality among patients with diabetes mellitus. *Archives of internal medicine*, 166(17):1836–1841.

Christine Howes, Matthew Purver, Rose McCabe, Patrick G. T. Healey, and Mary Lavelle. 2012. Predicting adherence to treatment for schizophrenia from dialogue transcripts. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 79–83, Seoul, South Korea, July. Association for Computational Linguistics.

Christine Howes, Matthew Purver, and Rose McCabe. 2013. Investigating topic modelling for therapy dialogue analysis. In *Proceedings of the IWCS 2013 Workshop on Computational Semantics in Clinical Text (CSCT 2013)*, pages 7–16, Potsdam, Germany, March. Association for Computational Linguistics.

Aurel O Iuga and Maura J McGuire. 2014. Adherence and health care costs. *Risk management and healthcare policy*, 7:35.

Abhyuday N Jagannatha and Hong Yu. 2016. Bidirectional RNN for medical event detection in electronic health records. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 473–482, San Diego, California, June. Association for Computational Linguistics.

Ashish K. Jha, Catherine M. DesRoches, Eric G. Campbell, Karen Donelan, Sowmya R. Rao, Timothy G. Ferris, Alexandra Shields, Sara Rosenbaum, and David Blumenthal. 2009. Use of electronic health records in u.s. hospitals. *New England Journal of Medicine*, 360(16):1628–1638. PMID: 19321858.

Jeannie K Lee, Karen A Grace, and Allen J Taylor. 2006. Effect of a pharmacy care program on medication adherence and persistence, blood pressure, and low-density lipoprotein cholesterol: a randomized controlled trial. *Jama*, 296(21):2563–2571.

Arun S. Maiya. 2020. ktrain: A low-code library for augmented machine learning. *arXiv*, arXiv:2004.10703 [cs.LG].

Edward T. Moseley, Joy T. Wu, Jonathan Welt, John Foote, Patrick D. Tyler, David W. Grant, Eric T. Carlson, Sebastian Gehrmann, Franck Dernoncourt, and Leo Anthony Celi. 2020. A corpus for detecting high-context medical conditions in intensive care patient notes focusing on frequently readmitted patients. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1362–1367, Marseille, France, May. European Language Resources Association.

Jamie Ng, Seung Ki Moon, Taezoon Park, and Wah-Pheow Tan. 2014. Intentional and unintentional medication nonadherence – comparing older and younger adults. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 58(1):160–164.

Takeshi Onishi, Davy Weissenbacher, Ari Klein, Karen O'Connor, and Graciela Gonzalez-Hernandez. 2018. Dealing with medication non-adherence expressions in twitter. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 32–33, Brussels, Belgium, October. Association for Computational Linguistics.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155, Boulder, Colorado, June. Association for Computational Linguistics.

Sara Rosenthal, Ken Barker, and Zhicheng Liang. 2019. Leveraging medical literature for section prediction in electronic health records. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4864–4873, Hong Kong, China, November. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi. 2018. Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1589–1604.

Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. 2019. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, page ocz096, July.

Jane M Simoni, Pamela A Frick, and Bu Huang. 2006. A longitudinal evaluation of a social support model of medication adherence among hiv-positive men and women on antiretroviral therapy. *Health Psychology*, 25(1):74.

Robyn Tamblyn, Tewodros Eguale, Allen Huang, Nancy Winslade, and Pamela Doran. 2014. The incidence and determinants of primary nonadherence with prescribed medication in primary care: a cohort study. *Annals of internal medicine*, 160(7):441–450.

Alexander Turchin, Holly I Wheeler, Matthew Labreche, Julia T Chu, Merri L Pendergrass, and Jonathan S Einbinder. 2008. Identification of documented medication non-adherence in physician notes. In *AMIA Annual Symposium Proceedings*, volume 2008, page 732. American Medical Informatics Association.

Meera Viswanathan, Carol E Golin, Christine D Jones, Mahima Ashok, Susan J Blalock, Roberta CM Wines, Emmanuel JL Coker-Schwimmer, David L Rosen, Priyanka Sista, and Kathleen N Lohr. 2012. Interventions to improve adherence to self-administered medications for chronic diseases in the united states: a systematic review. *Annals of internal medicine*, 157(11):785–795.

Byron C Wallace, M Barton Laws, Kevin Small, Ira B Wilson, and Thomas A Trikalinos. 2014. Automatically annotating topics in transcripts of patient-provider interactions via machine learning. *Medical Decision Making*, 34(4):503–512.

Kathryn E Weaver, María M Llabre, Ron E Durán, Michael H Antoni, Gail Ironson, Frank J Penedo, and Neil Schneiderman. 2005. A stress and coping model of medication adherence and viral load in hiv-positive men and women on highly active antiretroviral therapy (haart). *Health psychology*, 24(4):385.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Jia-Rong Wu, Debra K Moser, Misook L Chung, and Terry A Lennie. 2008. Predictors of medication adherence using a multidimensional adherence model in patients with heart failure. *Journal of cardiac failure*, 14(7):603–614.

Yi Yang, Vennela Thumula, Patrick F Pace, BF Banahan, Noel E Wilkin, WB Lobb, and Drug Benefit Trends. 2009. Medication nonadherence and the risks of hospitalization, emergency department visits, and death among medicare part d enrollees with diabetes. *Drug Benefit Trends*, 21(12):9.

8 Appendix A. Additional Corpus Analysis

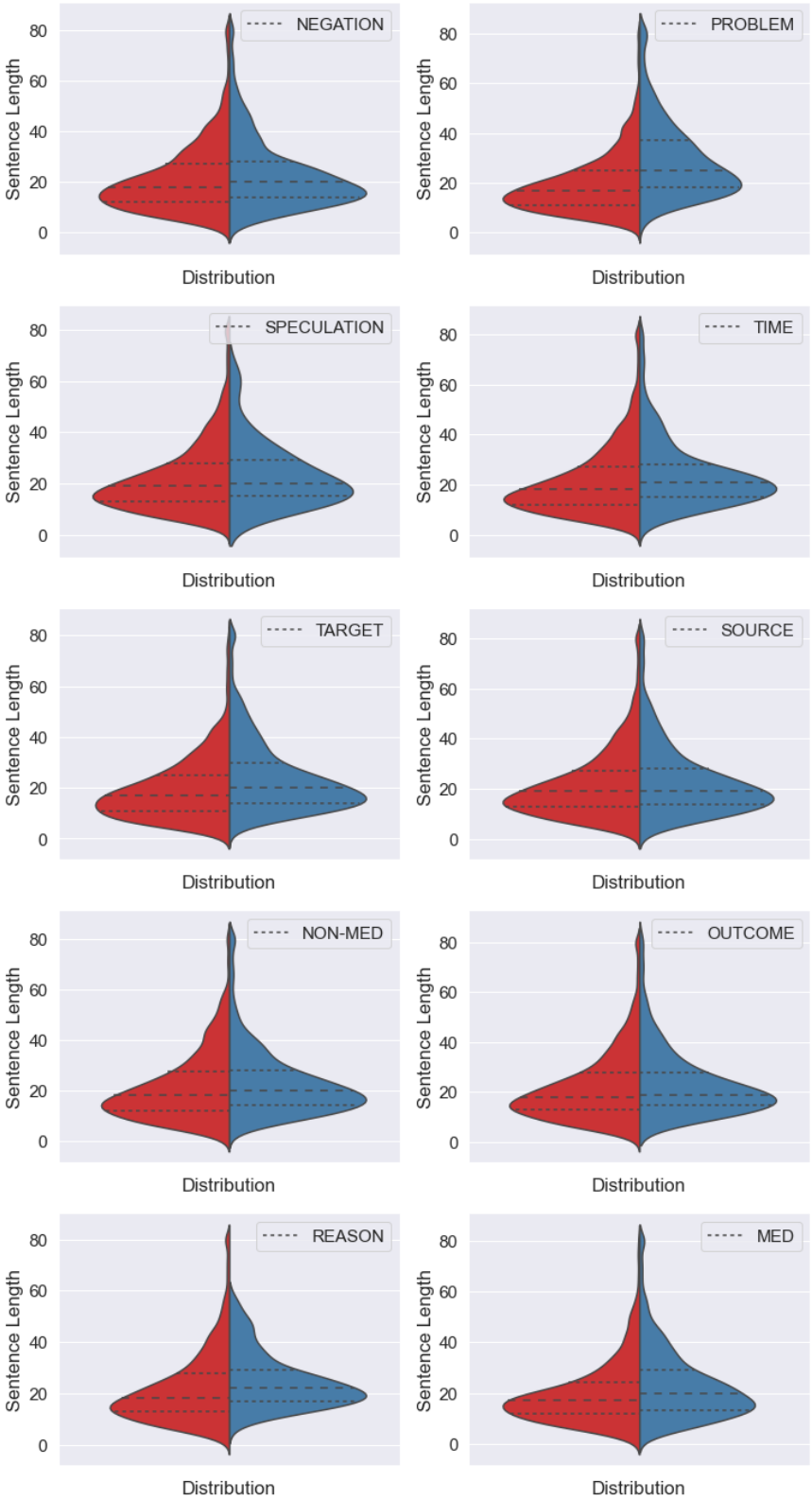

Figure 4: Distribution of sentence lengths depending on whether span-level annotations are present. Red (left) and blue (right) indicate that the span-level annotation is not present and is present respectively. This figure complements Figure 2 in the main paper.

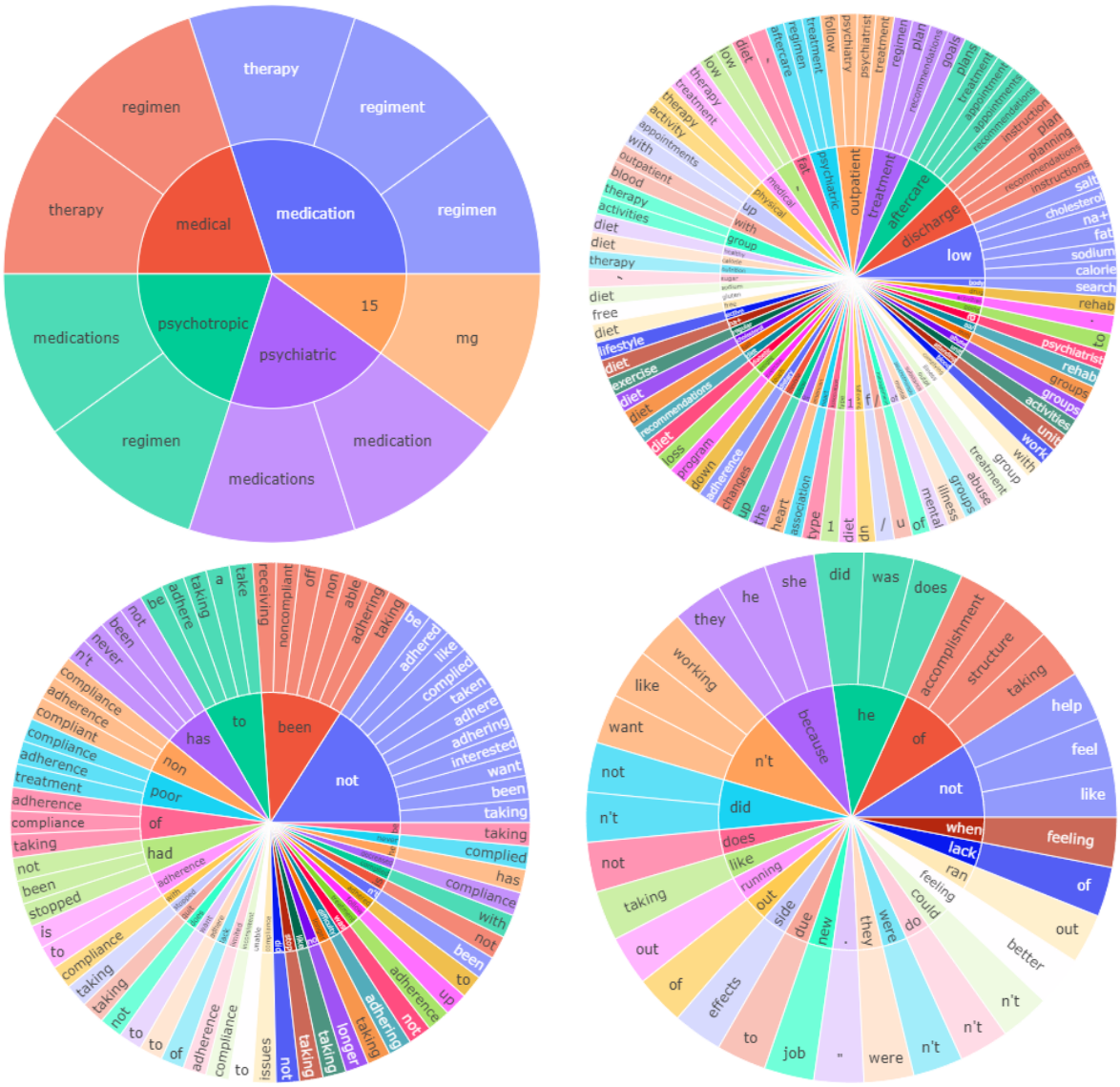

Figure 5: Most common bigrams in the following span-level annotations (bottom-down, left-to-right): MEDICATION, NON-MEDICATION, NEGATION, and REASON. This figure complements Figure 3 in the main paper.

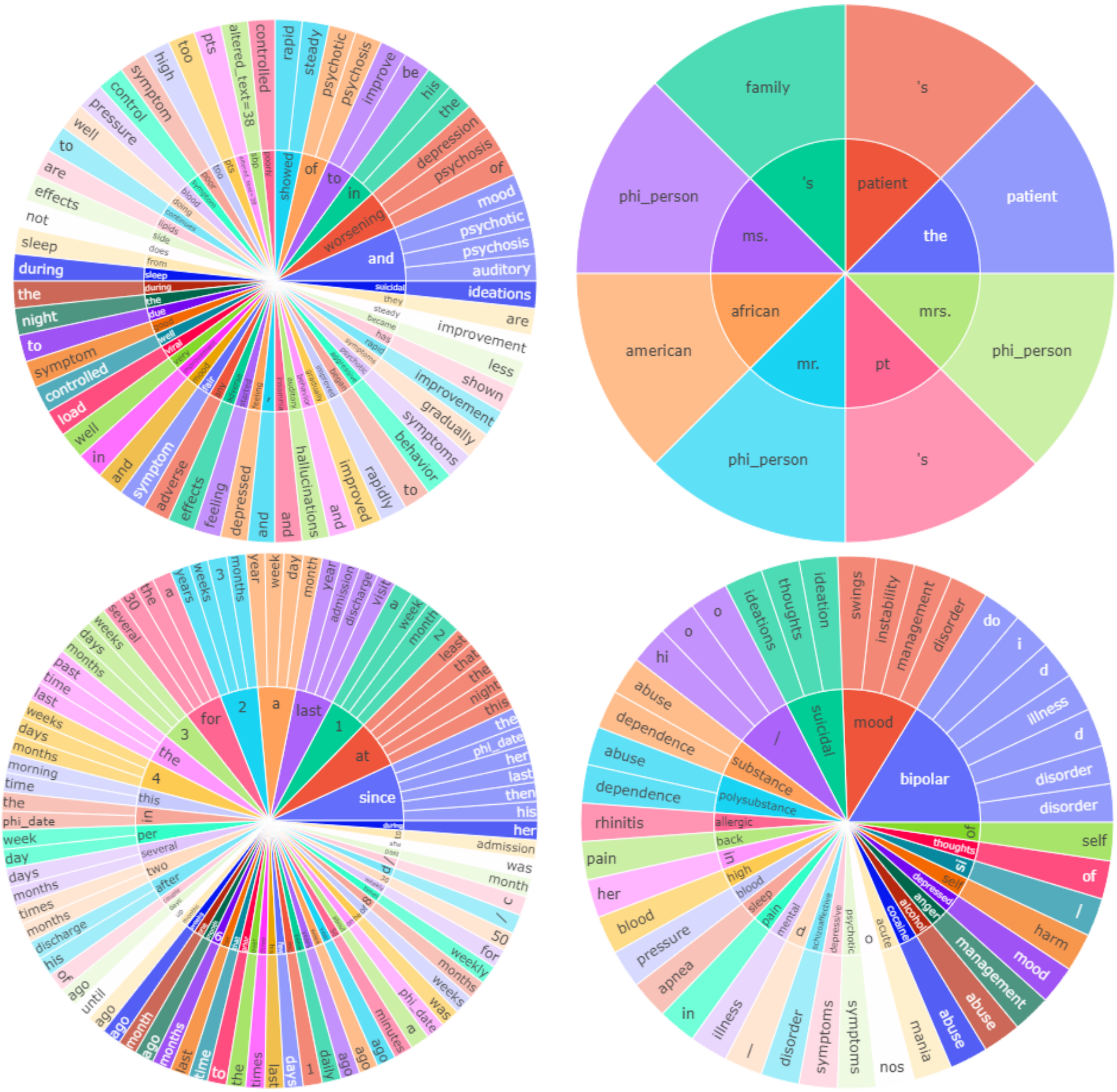

Figure 6: Most common bigrams in the following span-level annotations (bottom-down, left-to-right): OUTCOME, TARGET, TIMEand PROBLEM. This figure complements Figure 3 in the main paper.

Figure 7: Examples of span-level adherence annotations. This figure complements Table 1 in the main paper.