MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190255

Annotating Temporal Relations to Determine the Onset of Psychosis Symptoms

Natalia Viani[a], Joyce Kam[a], Lucia Yin[a], Somain Verma[a], Robert Stewart[a,b], Rashmi Patel[a,b], Sumithra Velupillai[a,c]

[a] Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK,
[b] South London and Maudsley NHS Foundation Trust, London, UK,
[c] School of Electrical Engineering and Computer Science, KTH, Stockholm, Sweden

Abstract

For patients with a diagnosis of schizophrenia, determining symptom onset is crucial for timely and successful intervention. In mental health records, information about early symptoms is often documented only in free text, and thus needs to be extracted to support clinical research. To achieve this, natural language processing (NLP) methods can be used. Development and evaluation of NLP systems requires manually annotated corpora. We present a corpus of mental health records annotated with temporal relations for psychosis symptoms. We propose a methodology for document selection and manual annotation to detect symptom onset information, and develop an annotated corpus. To assess the utility of the created corpus, we propose a pilot NLP system. To the best of our knowledge, this is the first temporally-annotated corpus tailored to a specific clinical use-case.

Keywords:
Schizophrenia; Electronic Health Records; Natural Language Processing

Introduction

For patients with a diagnosis of schizophrenia, duration of untreated psychosis (DUP) is the period of time between the onset of first symptoms and the initiation of adequate treatment [1]. As shown in previous studies, prolonged DUP is associated with poor intervention outcomes, both in the first years of treatment and in the long-term [2,3]. Therefore, to enhance the management of symptoms and improve social functioning, timely treatment is crucial. For determining symptom onset and ultimately reducing DUP, the information collected in clinical practices could be successfully re-used.

With the rapid adoption of electronic health records (EHRs), clinical data are increasingly available in electronic format, allowing for large-scale retrospective research. However, especially in the field of mental health, clinically relevant information (e.g., symptoms, diagnoses, medication) is often documented in unstructured form (free text), for instance through letters and progress notes. To allow analyzing the information enclosed in such clinical text, natural language processing (NLP) techniques are becoming increasingly popular [4]. In the case of determining symptom and treatment onset from clinical notes, NLP methods are needed to both extract clinical concepts (events) and anchor them on a timeline. To this end, two types of temporal information have to be identified: time expressions such as dates and times (TIMEXes), and temporal links (relations) between these and the available events ({hallucinations} in {2002}).

In recent years, a few clinical corpora have been annotated for temporal relations (TLINKs), and used for developing NLP systems tailored to this task. The 2012 i2b2 Temporal Relations Challenge focused on temporal relations in narratives from an intensive care unit [5]: a total of 310 discharge summaries were annotated with events, temporal expressions, and 8 types of temporal relations (e.g., "before", "overlap"). Styler IV et al. developed a corpus (THYME) of 1,254 cancer patient records, which were annotated with clinical and temporal information [6]. This corpus was then used in the 2015 and 2016 Clinical TempEval challenges (440 and 591 documents, respectively), which focused on determining two types of TLINKs [7,8]: relations between events and the document creation time (DCT), and relations between an event or a TIMEX and a narrative container. Successful NLP systems developed on these corpora have mainly relied on supervised machine learning algorithms, using lexical, morphological and syntactical features. A few systems also included heuristics and rule-based components.

Despite the recent advances in temporal relation extraction, developing temporal NLP systems in different clinical domains remains a challenge, due to the inherent complexity of the task – each patient can have several EHRs with clinically relevant information, and in each document every clinically relevant event can in principle be linked to every TIMEX. In this paper, we address the problem of determining temporal relations in mental health records, with a focus on symptom onset identification for schizophrenia patients. To the best of our knowledge, this is the first study on temporal relation extraction that was driven by a specific psychiatric clinical use-case. The NLP task was defined from a clinical perspective, with the final goal to extract relevant information at the patient level. As a crucial initial step to reach this goal, we needed to not only identify the clinically relevant events to be extracted and anchored in time, but also the set of documents that were likely to contain this information.

We have three main aims in this study. First, we propose a methodology for selecting the most relevant documents for the considered use-case. Then, we develop a manual annotation process to temporally anchor all the relevant symptoms, thus enabling the extraction of symptom onset and other information of interest. Finally, we propose a preliminary NLP system to assess the utility of the created corpus.

Methods

Dataset

In this study, we used mental health records from the Clinical Record Interactive Search (CRIS) database [9]. This research

repository contains anonymized patient data (structured and unstructured), derived from the EHR system used at the South London and Maudsley National Health Service (NHS) Foundation Trust (SLaM). Textual documents typically consist of either notes related to specific events or attachments of different types (e.g., assessments, discharge letters). In the system, there are no structured elements indicating whether a document represents a first assessment, which would be helpful to identify relevant content for our use-case.

For selecting the records that would most likely include the information of interest, we focused on documents from early intervention services for people with first episode psychosis (FEP). We considered documents written within a 3-month window by six intervention teams from the team's acceptance date in April 2018, on the assumption that these "early" documents would include the initial assessment and the richest description of a patient's clinical history. The steps followed for document extraction are shown in Figure 1. We focused on longer documents, in which clinicians typically document the presenting history and mental state examination, and excluded questionnaires and forms consisting of short lines. To identify how symptom onset information was typically documented, 70 documents were double-annotated for relevant paragraphs such as: "difficulties were noted for the first time when the patient was 7 years old, as he was displaying aggressive behaviour."

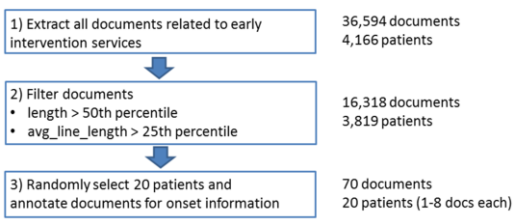

Figure 1 – First Document Extraction.

We analysed the annotated documents in terms of clinical and temporal content, automatically identifying symptoms (i.e., events) and time expressions. For extracting symptoms, we used a previously-developed keyword list [10] containing 598 psychosis-specific symptom terms. To identify TIMEXes, we used SUTime [11], a rule-based tool for temporal expression extraction which we have adapted to the mental health domain [12]. As a result of these analyses, we selected the final set of documents to be used for annotation. In general, documents do not follow a standard format. In some cases, there is a semi-structured format, with section headings (e.g., history of presenting complaint, clinical history, mental state examination), but there is a large variability. It is important to note that, while some events are reported with a specific date (e.g., "presented on 1st Jan. 2014 with hallucinations"), others are not clearly linked to a temporal point (e.g., references to "the past", ongoing symptomology).

Temporal Relation Annotation

Documents were pre-annotated with symptoms and time expressions, using the same tools as for corpus selection. Annotators were asked to try to link each pre-annotated event to a TIMEX, if such relation could be inferred from the text. In addition, they had to assign to each event a polarity value, which could be either "positive" or "negative" (e.g., "denied hallucinations"). This distinction is important, as negated symptoms would not likely indicate onset information, and should be represented differently on a patient's clinical timeline. Figure 2 shows an annotation example.

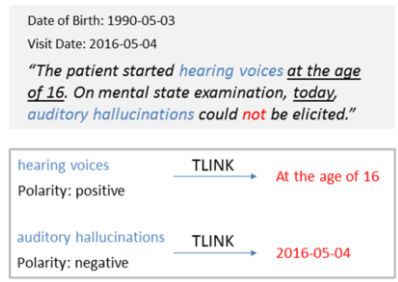

Figure 2 – Annotation Example.

The corpus was divided in batches, each including documents belonging to 9-10 patients (Table 1). Annotations were carried out by three medical students, and all documents were double-annotated. To guide the annotation task, two NLP researchers created specific guidelines, which were enriched with relevant example cases. To create the final version of the corpus, all annotated documents were adjudicated, resolving disagreements and performing corrections when needed.

To investigate if symptoms were linked to time points prior to the clinic visit, we also analysed, for each patient, how far back in time (in terms of days) a symptom referred to. In this analysis, we did not consider symptoms with negative polarity, as these are not likely to represent onset information.

Automated Temporal Relation Extraction

We used the annotated corpus to develop two NLP modules: TLINK extraction and polarity classification. The dataset was randomly split into training, development, and test sets. The training set was used for system development and manual rule engineering, with validation on the development set. The test set was set aside for final evaluation (in future studies).

In the temporal relation module, we addressed TLINK extraction (i.e., determining if the event can be linked to a time expression in the document) and TIMEX assignment (i.e., finding the normalized value of the TIMEX linked to the event) simultaneously. To perform both tasks, we developed a rule-based system relying on a number of features, such as the section in which the event is found and the presence of anchor dates in the text (admission, discharge, clinic dates). More specifically, section labels were identified by using a set of keywords (e.g., "history", "examination"), while anchor dates were extracted with regular expressions. System development was carried out on a subset of the training set, iteratively adding/refining rules. As a result, we developed ten rules to be applied following an order of relevance. For example, if only one TIMEX (representing a date like YYYY-MM-DD) is found in the same sentence including the symptom, a link is created. As another example, if a symptom is mentioned in a section named "mental state examination on admission" and an admission date is available, a link is created.

To assign a polarity value to events, we used ConText [13], a rule-based algorithm which relies on modifiers (e.g., "no", "denies") to determine whether a concept is negated. These modifiers are looked for in a window of words surrounding the event. In this paper, we used 11 "negation" modifiers representing the terms that were found in our corpus.

Evaluation

To evaluate the quality of the developed corpus, we computed inter-annotator (IAA) agreement. For polarity, we calculated accuracy on symptoms: an agreement is obtained when both annotators marked the same value. For the TLINK task, we defined an "adapted" accuracy, considering for each symptom two cases of agreement: i) both annotators identified 0 links, or ii) both annotators identified a link to the same time expressions (in terms of normalized value). All other combinations were regarded as disagreements. Automated extraction systems were evaluated with the same metrics.

Results

Corpus Selection

Figure 3 shows the results of the analysis conducted for corpus selection (on a subset of 70 documents). For each document, we computed the number of automatically extracted symptoms/TIMEXes (orange lines) and of manually annotated onset paragraphs (blue lines). Results (normalized counts) indicate that documents with many clinical/temporal elements are more likely to contain information on symptom onset.

Starting from this observation, we filtered the initial corpus by adding the following criteria: Symptom_count > 0 and Timex_count > 5. The number of symptoms was computed by using a list of 26 keywords developed by two psychiatrists, and the number of TIMEXes was found with the adapted SUTime. This filtering step led to a final set of 9,779 unique documents for 3,433 patients. From this corpus, we extracted 645 documents for 239 randomly selected patients (an average of 2.7 documents per patient) grouped into 24 batches.

Corpus Annotation

Table 1 reports the number of patients, documents, events, and TIMEXes in our corpus.

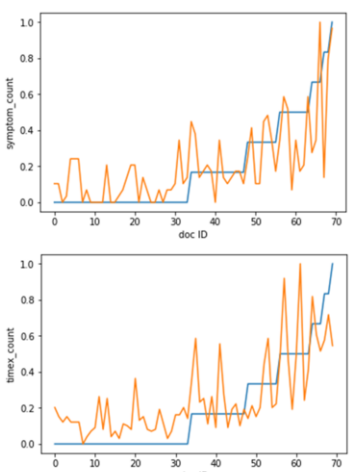

Figure 3 – Relation between Symptom/TIMEX Counts (orange lines) and Manually Annotated Onset Paragraphs (blue lines).

There are 2,590 symptoms (on average, 4 per document). The 5 most frequent (73% of events) are the following (raw counts between brackets): hallucinations (736), delusions (430), delusional (398), paranoia (179), and thought disorder (159). Each symptom was manually annotated for a Polarity value and an optional TLINK. Table 2 reports the IAA, while Table 3 shows the prevalence of Polarity values and temporal relations (for TLINKs, "Yes" represents the existence of a link). These counts were computed on the adjudicated dataset.

Table 1 – Patient, Document, Event and TIMEX Counts

	Total	Train	Dev
Patients (batches)	239 (24)	140 (14)	49 (5)
Documents	645	361	133
Events	2,590	1,465	515
TIMEXes	24,135	13,502	5,061

Table 2 – Inter-Annotator Agreement (IAA) per Annotated Item

Item	IAA (average)	IAA (range/batch)
TLINK	0.73	0.60 - 0.84
Polarity	0.95	0.81 - 1

Table 3 – Annotation Results for TLINK and Polarity

Item	Value	Total	Train	Dev
TLINK	Yes	1,661 (64.1%)	945	302
	No	929 (35.9%)	520	213
Polarity	Pos	1,900 (73.4%)	1,110	368
	Neg	690 (26.6%)	355	147

Starting from the adjudicated annotations, for each patient we computed the difference between the maximum and the minimum dates associated to any "positive" symptom (diff). It is important to note that, for those symptoms that were not explicitely linked to a date, this difference could not be computed. As a result, we were able to compute diff values for 206 patients (Figure 4). Out of these, 41 (20%) had a diff value longer than one year, while 71 (34%) had a zero diff value.

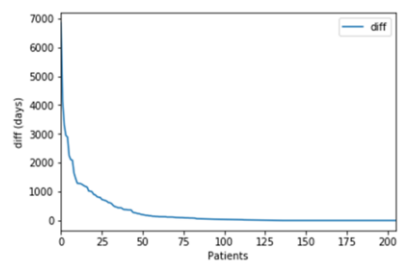

Figure 4 – Difference (in days) between Minimum and Maximum Symptom Dates for 206 Patients

Automated Temporal Relation Extraction

Table 4 shows preliminary results for the two developed NLP modules, using the same metrics as for IAA. For comparison purposes, we also report two baseline results: for the Polarity attribute we classified each event as "positive", while for TLINKs we did not link any event to any specific TIMEX.

Table 4– Performance of NLP modules

Item	Model	Train	Dev
TLINK	baseline	0.47	0.54
	Rule-based	0.67	0.58
Polarity	baseline	0.76	0.72
	ConText	0.93	0.95

Discussion

We have developed a corpus of mental health records for patients with schizophrenia who have been admitted to early onset intervention services, annotated with temporal relations to capture the onset of psychosis symptoms. To the best of our knowledge, this is the first temporally-annotated corpus that was developed for a specific clinical use-case besides clinical timeline reconstruction. In particular, our use-case is related to the analysis of symptom onset and to the calculation of DUP on a large patient cohort. To address this long-term goal, dataset selection was crucial: we applied symptom/TIMEX-based filtering steps to the available CRIS data, and selected multiple documents referred to each patient. Starting from a mention-level annotation task, we aim at proposing a framework that could be also relevant for information extraction on a patient-level. The guidelines and the keywords used in the annotation process, as well as the code for NLP development, are available at: https://github.com/medesto/.

Besides this underlying perspective, our corpus differs from related datasets (i.e., i2b2 2012 and THYME) in two main ways. First, to allow capturing onset information which can be reported across different sentences/paragraphs, we did not require linked entities to be close to each other: each event could be linked to any time expression written in the document. This is a first step towards reconstructing timelines across multiple documents, a problem that remains understudied in the clinical domain. Raghavan et al., for example, proposed a system for cross-document alignment of event sequences [14]. Second, to simplify the annotation task and still obtain useful data, we asked annotators to associate each event to only one time expression (the most relevant one), thus considering only one type of TLINK. Given these differences, we proposed the use of an adapted accuracy to measure the IAA on temporal links, with a final value of 0.73. This represents a promising result, especially considering the inherent difficulty of the task. In some cases, for example, annotators found it difficult to decide whether a symptom should be linked to a given date: even if a temporal link could be reasonably inferred, the relation was not clearly stated in the text. As another cause of disagreements, for a few symptoms that were clearly related to the visit date, this date was not explicitly written in the text: in these cases, the "most likely" date was often chosen. As for the Polarity attribute, we obtained a particularly high IAA, with an accuracy of 0.95. As expected, classifying a symptom as positive or negative was easier than contextualizing it from the temporal point of view.

In the adjudicated corpus, 1661 symptoms (64%) were linked to a specific date. It is interesting to note that 541 (33%) of these symptoms were negated, and therefore do not play a role for symptom onset extraction. As a matter of fact, 197 negated events (36%) were found in "examination" sections, thus representing results of patient visits. We aimed at capturing these temporal links as they could be important for general timeline reconstruction, however they might not be directly relevant to our long-term goal. To assess the utility of our data for symptom onset extraction, we analyzed the temporal gap between the first and the last symptom dates available for each patient. Our assumption was that symptoms going far back in time could represent the actual onset of psychosis. To verify this, we reviewed the documented "early symptoms" for 41 patients having a diff value of more than one year. Out of these 41 instances, 17 corresponded to a clear onset date, while 15 were a close approximation to the onset date (which was specified in other parts of the texts). An example of the first type is given by: "he has been suffering from psychosis since he was 10 years old when he started experiencing hallucinations". The remaining 9 instances resulted either from erroneous dates written in the text (3), or from a long temporal gap between documents associated to the same patient (6). These results indicate that the proposed annotation schema could be useful for correctly capturing information on early symptom onset, as well as for retaining more general temporal information for timeline reconstruction. To further assess this point, we plan to analyze the information annotated for the 165 patients with a diff value lower than one year.

As regards NLP system development, preliminary results indicate that our gold data are consisent enough to allow for automated system development (Table 4). However, the performance of the TLINK module on the development set (0.58) was lower than that on the training set (0.67), showing that more effort should be put into developing a generalizable system. To address this, we plan to both improve the available rules and explore supervised machine learning methods. Given the complexity of our problem, human-in-the-loop approaches could be explored [15]. Moreover, to support real-world usability, it would be important to focus on explainable methods [16]. Once the NLP system is completed, we will run it on a large patient cohort, to quantify the number of patients for which an early symptom onset is documented in free text. In addition, we are interested in assessing which types of TIMEXes are most frequently associated to onset information.

Our study presents two main limitations. First, given the huge amount of textual information available in CRIS, the proposed corpus selection might not be ideal. To investigate this, we are currently annotating different types of documents, in particular those related to first referrals to SLaM (without focusing on early intervention services). Second, the way in which we modeled the problem could be potentially improved. To simplify the annotation task, we only focused on a limited set of symptom keywords; however, these keywords are not suitable to capture more complex linguistic variants. Moreover, addressing the extraction problem at a mention-level is not necessarily the best option. As future work, we will investigate other ways to model our problem, for example by following a question-answering annotation approach. One drawback of this approach could be the low prevalence of symptom onset descriptions in the texts and the need to review even larger sets of documents.

Conclusions

In this study we described a gold standard for temporal relation extraction in the mental health domain, with a focus on symptom onset and DUP extraction. We presented a method for corpus selection and an annotation schema, with promising IAA results. As a proof of concept, we proposed an early rule-based system for TLINK extraction. In the future, this system could be used to temporally anchor symptoms and treatments extracted from mental health records, thus enabling the calculation of DUP and other relevant concepts.

Acknowledgements

NV and SV are supported by the Swedish Research Council (2015-00359), Marie Skłodowska Curie Actions, Cofund,

Project INCA 600398. Ethical approval: Oxford REC C, reference 08/H0606/71+5. RS and RP are part-funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. RP has received

support from a Medical Research Council (MRC) Health Data Research UK Fellowship (MR/S003118/1) and a Starter Grant for Clinical Lecturers (SGL015/1020) supported by the Academy of Medical Sciences, The Wellcome Trust, MRC, British Heart Foundation, Arthritis Research UK, the Royal College of Physicians and Diabetes UK.

References

[1] M. Marshall, S. Lewis, A. Lockwood, R. Drake, P. Jones, and T. Croudace, Association between duration of untreated psychosis and outcome in cohorts of first-episode patients: a systematic review, *Arch. Gen. Psychiatry*. **62** (2005) 975–983.

[2] M. Hill, N. Crumlish, M. Clarke, P. Whitty, E. Owens, L. Renwick, S. Browne, E.A. Macklin, A. Kinsella, C. Larkin, J.L. Waddington, and E. O'Callaghan, Prospective relationship of duration of untreated psychosis to psychopathology and functional outcome over 12 years, *Schizophr. Res.* **141** (2012) 215–221.

[3] M. Penttilä, E. Jääskeläinen, N. Hirvonen, M. Isohanni, and J. Miettunen, Duration of untreated psychosis as predictor of long-term outcome in schizophrenia: systematic review and meta-analysis, *The British Journal of Psychiatry*. **205** (2014) 88–94.

[4] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, and H. Liu, Clinical information extraction applications: A literature review, *Journal of Biomedical Informatics*. **77** (2018) 34–49.

[5] W. Sun, A. Rumshisky, and O. Uzuner, Evaluating temporal relations in clinical text: 2012 i2b2 Challenge, *J Am Med Inform Assoc*. **20** (2013) 806–813.

[6] W.F. Styler IV, S. Bethard, S. Finan, M. Palmer, S. Pradhan, P.C. de Groen, B. Erickson, T. Miller, C. Lin, G. Savova, and J. Pustejovsky, Temporal annotation in the clinical domain, *Transactions of the Association for Computational Linguistics*. **2** (2014) 143–154.

[7] S. Bethard, L. Derczynski, G. Savova, J. Pustejovsky, and M. Verhagen, SemEval-2015 Task 6: Clinical TempEval, *Proceedings of SemEval 2015*. (2015) 806–814.

[8] S. Bethard, G. Savova, W.-T. Chen, L. Derczynski, J. Pustejovsky, and M. Verhagen, Semeval-2016 task 12: Clinical TempEval, *Proceedings of SemEval 2016*. (2016) 1052–1062.

[9] G. Perera, M. Broadbent, F. Callard, C.-K. Chang, J. Downs, R. Dutta, A. Fernandes, R.D. Hayes, M. Henderson, R. Jackson, A. Jewell, G. Kadra, R. Little, M. Pritchard, H. Shetty, A. Tulloch, and R. Stewart, Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource, *BMJ Open*. **6** (2016) e008721.

[10] R. Jackson, R. Patel, S. Velupillai, G. Gkotsis, D. Hoyle, and R. Stewart, Knowledge discovery for Deep Phenotyping serious mental illness from Electronic Mental Health records, *F1000Res*. **7** (2018).

[11] A.X. Chang, and C.D. Manning, SUTIME: A Library for Recognizing and Normalizing Time Expressions, *Proceedings of LREC 2012*. (2012).

[12] N. Viani, L. Yin, J. Kam, A. Alawi, A. Bittar, R. Dutta, R. Patel, R. Stewart, and S. Velupillai, Time Expressions in Mental Health Records for Symptom Onset Extraction, *Proceedings of LOUHI 2018*. (2018) 183-192.

[13] B.E. Chapman, S. Lee, H.P. Kang, and W.W. Chapman, Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm, *Journal of Biomedical Informatics*. **44** (2011) 728–737.

[14] P. Raghavan, E. Fosler-Lussier, N. Elhadad, and A.M. Lai, Cross-narrative Temporal Ordering of Medical Events, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.* **1** (2014) 998–1008.

[15] A. Holzinger, Interactive machine learning for health informatics: when do we need the human-in-the-loop?, *Brain Inform*. **3** (2016) 119–131.

[16] A. Holzinger, From Machine Learning to Explainable AI, *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*. (2018) 55–66.

Address for correspondence

Natalia Viani

Email: natalia.viani@kcl.ac.uk.