A Dataset of Psychiatric Hospital Notes with Temporal Information Annotations

Timothy Miller[1,2] , Gaby Dinh[1], David Harris[1], WonJin Yoon[1,2], Spencer Thomas[1]
Boyu Ren[2,3] , Mei-Hua Hall[2,3*], Guergana Savova[1,2*]
[1] Boston Children's Hospital, MA, USA
[2] Harvard Medical School, MA, USA
[3] McLean Hospital, MA, USA
{firstname.lastname}@childrens.harvard.edu
bren@mgb.org, mhall@mclean.harvard.edu

Abstract

Temporal information extraction is the task of identifying temporal entities in a text and relating them to each other. In medicine, electronic health records (EHRs) contain text that documents the sequence of events during an encounter with a patient, and sometimes the events prior to the encounter (e.g., psychosocial environment and history). Temporality is especially important for the specialty of psychiatry. In this work, we describe the updates to the guidelines that allowed us to create a corpus of temporally-annotated psychiatric discharge summaries and progress notes in English. These updated guidelines were used to create a corpus of over 18,000 events, 2,200 time expressions, and 13,000 temporal relations. Temporal information extraction performance with a baseline system trained on non-psychiatric data obtains an F1 score of 0.152 on relation extraction, indicating the importance of this new dataset for making progress on temporal information extraction in the psychiatric domain.

Keywords: temporality, psychiatry, information extraction

1. Introduction

Temporal information extraction is the task of identifying temporal entities in a text and relating them to each other. Temporal entities include *time expressions* — phrases that refer to spans of time — and *events* — things that have happened whose temporal spans can be placed on a timeline. Temporal relations relate these entities, describing how they are positioned relative to each other on a timeline. Extracting these entities and relations accurately can help create an abstract representation of the temporality of a sequence of events that can be used for downstream purposes.

In medicine, understanding the temporality of events associated with a patient is important for modeling things like diagnosis, prognosis, and health trajectory. Electronic health records (EHRs) contain text that documents the sequence of events during an encounter with a patient, and sometimes the events prior to the encounter (e.g., psychosocial history). Temporality is especially important for the specialty of psychiatry, because many psychiatric conditions are chronic and unfold over a lifespan. A person's psychosocial environment (e.g., recurrent housing instability, chronic unemployment), personal events (e.g., historical traumas) play an outsized role and influence on patient clinical status, inpatient timelines can be rich with event sequencing, and medication adjustment over time are also very common and critical for understanding changes about patient status.

Previous work has created guidelines and datasets for extracting temporal information from EHR text, but primarily focused on cancer (Styler IV et al., 2014; Bethard et al., 2015, 2016, 2017; Wright-Bettner et al., 2020; Yao et al., 2024), with one dataset for general medicine (Sun et al., 2013). These datasets have been used to organize shared tasks and evaluate and advance the state of the art for temporal information extraction.

Work in psychiatry has created modified guidelines for time expression annotation (Viani et al., 2018), but full guidelines have not been created, and no publicly available datasets exist for psychiatric data. Psychiatric EHR data differs from other medical domains in being more richly narrative, having an extensive psychosocial context, and describing longer timelines of events, and thus temporal annotation is likely to require specialized annotation guidelines. Temporal Information Extraction (IE) methods are also likely to benefit from in-domain training data.

Indeed, we found that when we created a gold standard corpus to evaluate existing methods trained on previous temporal IE corpora, annotators needed updated guidelines to accurately annotate this new domain. In this work, we describe the updates to the guidelines that allowed us to create a corpus of temporally-annotated psychiatric discharge summaries and progress notes, report the size of the dataset and agreement, and report performance results for baseline machine learning methods for extracting temporal information. The

[*] These authors contributed equally as senior authors to this work.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7479–7484
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

annotation guidelines will be made publicly available upon publication. The annotated dataset is available to researchers whose institutions must agree to a data use agreement with the data sponsoring institution.

2. Background

Temporal information extraction typically is defined as the extraction of mentions of time expressions, events, and relation links between them. TimeML (Pustejovsky et al., 2011) defines a markup language for annotation of temporal information, which names these information types as TIMEX3, EVENT, and TLINK, respectively. TLINKs consist of a set of temporal relation category labels which fully describe the ways in which two time spans can be related on a timeline.

Subsequent work introduced the concept of *narrative containers* (Pustejovsky and Stubbs, 2011), which reduce the cognitive complexity of temporal annotations by focusing TLINKs to entities that are central. For example, a surgery event might contain multiple sub-events like hand washing, anesthesia, incision, implantation, suturing, and post-operative recovery, all of which have the same relation to more distant temporal entities. The amount of linking can be reduced by linking sub-events to the narrative container (*surgery*), and then linking only surgery to more distant entities, with the sub-events inheriting the same relation.

The THYME (Temporal History of Your Medical Events) (Styler IV et al., 2014; Wright-Bettner et al., 2020) corpus applied ideas from TimeML and narrative containers to the domain of clinical text. THYME also annotated a special relation type called DocTimeRel, which linked every annotated event to the document creation time with a coarser label set (Before, Overlap, After, Before/Overlap), allowing for all events to be placed on a timeline even if not explicitly linked to a time expression. That dataset applied the modified schema to de-identified clinical, pathology, and radiology notes for patients with colorectal and brain cancer from Mayo Clinic. While their reported agreement statistics suggested that clinical temporal information annotation is challenging for humans, the Contains relation, core to the narrative container-based approach, is both highly prevalent and obtains higher agreement.

In the psychiatric domain, there is little work studying temporal information extraction. One exception studies the adaptation of annotation schema and methods for time expressions to psychiatric data (Viani et al., 2018). Specifically, that study explicitly annotated expressions like *past* and *current*, and expressions related to the patient's age (*28-year-old* man).

3. Methods

3.1. Data

Our data consist of de-identified discharge summaries and progress notes from inpatient stays at a large, academic-affiliated psychiatric institution in the United States, written in English. We sampled 120 notes, evenly balanced between discharge and progress notes, with no patient represented in more than one note. One half of the notes in the corpus are sampled from patients whose primary Axis I diagnosis is a mood disorder (predominantly major depressive disorder), and the other half is sampled from patients whose primary diagnosis is psychosis-related (predominantly schizophrenia).[1] Patients with co-morbid psychiatric diagnoses are common and are included.

In order to avoid leaving out patients from underrepresented groups, we sampled our notes with a Monte Carlo-inspired algorithm. We computed demographic statistics across categories present in our structured dataset. We then collected 1000 samples of 120 notes, and computed the demographic statistics over those 120 patients for each sample. Cosine similarity between the demographic statistics in each sample and the overall cohort was computed. Finally, we selected the sample that had the closest similarity in demographic distribution to the overall cohort for data annotation.

3.2. De-identification

All the notes in our sample were de-identified to remove protected health information (PHI) in a multistep process. First, we ran a modified version of the Philter (Norgeot et al., 2020) de-identification tool across our dataset. Specifically, we disabled de-identification of dates, since they are central to our project, and added a specialized date-of-birth regular expression to remove patient birthdays, as clearly identifying information. Next, a human manually reviewed each note to remove non-narrative content (e.g., structured data like lists of medications that are automatically pulled in by electronic health record software), and was instructed to look for stray PHI and report it. In the next phase, a psychiatry domain expert at the data-providing institution manually reviewed each note, specifically looking for leaked PHI, while also restoring false positives found by Philter that would be clinically important. Finally, during the temporal annotation process, we instructed the annotators to report any incidences of PHI found.

[1] Specifically, we filtered for ICD10 billing codes F2* for psychosis and F3* for mood disorders.

3.3. Annotation Pipeline

During the annotation process, we had three annotators working on the project. The primary annotator did most of the single annotations, and has several decades of experience both with medical coding and with annotating EHRs for NLP tasks. The subject matter annotator was a domain expert, a researcher at the data-generating institution who is familiar with psychiatric data, and was primarily involved during guidelines development to provide domain-specifc expertise. A secondary annotator did a smaller number of annotations to complete the project after the primary annotator retired. The secondary annotator had a background in computational linguistics, and experience annotating temporal data for other projects using EHR data. The primary annotator trained the secondary annotator on the modified schema.

Temporal information annotation has not been done in the psychiatry domain before, so we could not use existing temporal annotation guidelines. Instead, we used the publicly available THYME guidelines as a starting point, and modified them during our annotation process. We used the open-source Anafora tool for performing annotation,(Chen and Styler, 2013) modifying the XML schema from the THYME project which also used that software. Anafora annotations are saved in XML format, and can be processed with the anaforatools software. (Bethard, 2025)

First, we performed **pilot annotations**: A sample of manually de-identified discharge notes were pair-annotated by the primary annotator (with experience in temporal annotations) and the subject matter annotator. During this phase, the primary annotator noted differences from other domains, and the subject matter annotator noted important psychiatry-specific phenomenon that would be important to capture. Findings from these sessions were discussed during group meetings, where tricky questions were resolved. During this phase, we iteratively modified the annotation guidelines specific to this dataset. Major changes included:

- Sections to ignore: We opted to ignore a large number of sections that never contain time expressions or narrative text. While these sections may contain things that could be considered events, they would not be linked to time expressions and would have DocTimeRel label OVERLAP (e.g., *Allergies*, *Test results*). These sections could easily be found in text by human annotators.

- Multi-word expressions: While THYME, by convention, annotated head words of phrases (e.g., in the phrase *colon cancer*, only the word *cancer* would be marked as the event), we decided that multi-word terms that are common and specific in psychiatry should be annotated as events.(Holderness et al., 2019) For example, terms like *self harm* or *suicidal ideation*. A large set of multi-word expressions was enumerated in the guidelines.

After updating the guidelines and training the primary annotator, our annotation pipeline took the following steps:

- **Independent annotations**: Our primary annotator and subject matter annotator annotated another pilot batch independently.

- **Agreement measurement**: We measured annotator agreement and compared results to previously published agreement statistics such as THYME (Styler IV et al., 2014) and i2b2 (Sun et al., 2013). We iteratively improved the guidelines and reverted to *Independent annotations* to improve agreement.

- **Solo annotations**: After agreement was satisfactory, the primary annotator proceeded to finish most of the notes. When the secondary annotator was trained, they completed the remaining annotations, including an overlapping subset with the primary annotator for final agreement measurement.

- ***Do not annotate* annotations**: After completing temporal annotations, we went back and added a layer called *Do not annotate* that covers any sections where there might be nominal temporal information, but in a non-narrative context, to simplify the task for downstream models. The idea was that a pre-processing classifier could be trained to label these sections and pass only the unlabeled sections to a temporal information extraction system.

3.3.1. Inter-annotator agreement

We computed inter-annotator agreement between our primary and secondary annotators on a set of nine notes.The metric we use is F1 score, calculating recall and precision by alternating which annotator is considered the reference and which is considered the predictions. Closure is computed over the set of reference relations when computing precision, and to the predicted relations when computing recall, as recommended by (UzZaman and Allen, 2011). We use the anaforatools scoring tool (Bethard, 2025), as used in previous TempEval shared tasks (Bethard et al., 2017).

3.3.2. Baseline Information Extraction Methods

We tested off-the-shelf models trained on the THYME data to see how much knowledge from

THYME training would apply to our corpus of psychiatric notes. The model we evaluated is a BERT-based model (Devlin et al., 2019) jointly trained to do EVENT extraction, TIMEX3 extraction, and TLINK extraction (Miller et al., 2023).

4. Results

Our de-identification process was found to be successful. The Philter tool is known to have high recall (Norgeot et al. (2020) report recall over 99% on two datasets), and our human review steps in our pipeline did not find any instances of missed PHI that needed to be removed. Many instances of false positives were detected, and these were restored to improve the usability of the data for downstream tasks.

Interannotator agreement was found to be 0.68 on EVENT spans, 0.85 on TIMEX3 spans, and 0.661 on TLINKs (correct argument spans and category). While TLINK agreement is substantially lower, this is in accordance with other temporal annotation efforts. The most common relation, CONTAINS, obtained an agreement F1 of 0.663.

Table 1 shows how many annotations of the primary temporal types were found in our dataset across splits. This represents an EVENT/TIMEX3 ratio of 8:1, compared to the 11:1 reported by Styler IV et al. (2014) in THYME, and an average of 0.74 TLINKs per event, compared to 0.5 reported in THYME. We consider these differences to be reasonable given the large variability in clinical notes across institutions and specialties. We also annotated 561 Aspectual links and marked 461 sections as "Do not annotate."

Table 2 shows the results of the baseline system described in Section 3.3.2 when evaluated off-the-shelf, using the same Anaforatools scoring tool used for inter-annotator agreement. We evaluate both for span correctness (finding EVENT, TIMEX3 text spans, TLINK argument spans) and label correctness (correctly labeling DocTimeRel for EVENTs, time class for TIMEX3s, and relation category for TLINKs). The model trained on non-psychiatric data struggles, with EVENTs and TIMEX3s below 0.7 F1, and TLINKs at 0.152. The model especially struggles to label DocTimeRel without in-domain training data. This illustrates the importance of creating this labeled resource — we now have a resource to both quantify the scale of the challenge and to develop future models for improved performance.

Split	EVENT	TIMEX3	TLINK
Train	10,274	1,317	7,809
Dev	3,033	386	2,171
Test	4,759	525	3,431
Total	18,066	2,228	13,411

Table 1: Statistics of the dataset across training, development, and test splits.

		Precision	Recall	F1
Span	EVENT	0.478	0.727	0.577
	TIMEX3	0.581	0.731	0.647
	TLINK	0.313	0.101	0.152
Label	EVENT	0.056	0.085	0.067
	TIMEX3	0.275	0.346	0.306
	TLINK	0.298	0.096	0.145

Table 2: Performance of off-the-shelf model trained on THYME cancer annotations, evaluated on the development set of the psychiatric data. We allowed for overlapping spans, the more relaxed evaluation setting.

5. Discussion

Annotating temporal information in psychiatry notes was a major challenge. We noted that, unlike the THYME colon cancer notes, psychiatric notes often contained large blocks of non-narrative texts. While we tried to remove many of these with the manual pre-processing, this process was imperfect, and the boundaries of what might be important for downstream use were often blurry. As a result, there was some confusion during early rounds of annotation over whether remaining time expressions should be annotated, even if they did not seem to be in narrative sections. To remedy this, we added the layer of "Do not annotate" annotations as a secondary annotation task. The guidance that we gave annotators was, any text that has time expressions that you do not think are important to annotate should be labeled as "Do not annotate."

This dataset also contained very long notes (primarily discharge summaries). This is another contrast with the THYME data – our data represents in-patient stays, which might last weeks or longer, so discharge summaries may stretch very long and contain extensive narratives This represents a cognitively demanding task for annotators, which may be more likely to lead to missed annotations.

Psychiatry notes are also unique in the depth, importance, and length of psychosocial and social environment history and descriptions. Patient social-environment histories often contain important pre-admission events like illicit drug use, history of trauma, and employment history, interpersonal relationship among other factors that clinicians may find relevant to the patient's mental health. While we believe these events are likely to be important, they do not always have the temporal granularity or specificity of ordering that we might like if we want to piece together a complete social history of the patient (i.e., pre-admission social history events

may have no explicit TLINKs but all get marked as DocTimeRel BEFORE).

6. Conclusion

In conclusion, we found that psychiatric discharge summaries and progress notes are challenging to annotate but rich with temporal information. We developed new guidelines that will be made publicly available upon publication, and may help others create similar datasets. In addition, the de-identified corpus has been approved for sharing with the research community with appropriate data use agreements. Annotation offsets (i.e., with no text data) will be shared publicly via a public GitHub repository. We are confident that this is the first such dataset, let alone that can be made available to the research community. The corpus can be obtained at the following github url: https://github.com/Machine-Learning-for-Medical-Language/thyme-mh.

7. Acknowledgements

Research reported in this publication was supported by the National Institute Of Mental Health of the National Institutes of Health under Award Number R01MH126977. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

8. Bibliographical References

S Bethard, L Derczynski, and G Savova. 2015. Semeval-2015 task 6: Clinical tempeval. *Proc. SemEval*.

Steven Bethard. 2025. Anafora Tools. https://github.com/bethard/anaforatools. Accessed: 2025-09-18.

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. SemEval-2016 Task 12: Clinical TempEval. In *Proceedings of the 10th International Conference on Semantic Evaluation (SemEval 2016)*.

Steven Bethard, Guergana Savova, Martha Palmer, and James Pustejovsky. 2017. SemEval-2017 Task 12: Clinical TempEval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 565–572, Vancouver, Canada. Association for Computational Linguistics.

WT Chen and W Styler. 2013. Anafora: A Web-based General Purpose Annotation Tool. *: The 2013 Annual Conference of the \ldots*.

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. Association for Computational Linguistics.

Eben Holderness, Nicholas Miller, Philip Cawkwell, Kirsten Bolton, Marie Meteer, James Pustejovsky, and Mei-Hua Hall. 2019. Analysis of risk factor domains in psychosis patient health records. *Journal of biomedical semantics*, 10(1):19.

Timothy Miller, Steven Bethard, Dmitriy Dligach, and Guergana Savova. 2023. End-to-end clinical temporal information extraction with multi-head attention. *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2023:313–319.

Beau Norgeot, Kathleen Muenzen, Thomas A. Peterson, Xuancheng Fan, Benjamin S. Glicksberg, Gundolf Schenk, Eugenia Rutenberg, Boris Oskotsky, Marina Sirota, Jinoos Yazdany, Gabriela Schmajuk, Dana Ludwig, Theodore Goldstein, and Atul J. Butte. 2020. Protected Health Information filter (Philter): accurately and securely de-identifying free-text clinical notes. *npj Digital Medicine*, 3(1):57. Publisher: Nature Publishing Group.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Roser Saur. 2011. ISO-TimeML Primer. pages 1–45.

James Pustejovsky and Amber Stubbs. 2011. Increasing informativeness in temporal annotation. *Proceedings of the 5th Law Workshop (LAW V)*, (June):152–160.

William F. Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. 2014. Temporal Annotation in the Clinical Domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association : JAMIA*, 20(5):806–813. Place: England.

Naushad UzZaman and James F Allen. 2011. Temporal Evaluation. *The 49th Annual Meeting of the Association for Computational Linguistics Human Language Technologies*, 271(5):351–356.

Natalia Viani, Lucia Yin, Joyce Kam, Ayunni Alawi, André Bittar, Rina Dutta, Rashmi Patel, Robert Stewart, and Sumithra Velupillai. 2018. Time Expressions in Mental Health Records for Symptom Onset Extraction. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 183–192, Brussels, Belgium. Association for Computational Linguistics.

Kristin Wright-Bettner, Chen Lin, Timothy Miller, Steven Bethard, Dmitriy Dligach, Martha Palmer, James H. Martin, and Guergana Savova. 2020. Defining and Learning Refined Temporal Relations in the Clinical Narrative. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 104–114, Online. Association for Computational Linguistics.

Jiarui Yao, Harry Hochheiser, WonJin Yoon, Eli Goldner, and Guergana Savova. 2024. Overview of the 2024 Shared Task on Chemotherapy Treatment Timeline Extraction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 557–569, Mexico City, Mexico. Association for Computational Linguistics.

Appendix

We include a section of a discharge summary from the dataset that is particularly rich in temporal events in Figure 1.

Episode of care 12/24/20XX - 1/20/20XX : 12/24/XX : ******** returns from acute care at *** ******** Hospital after an unwitnessed fall , for which he was admitted for evaluation . Diagnosis was likely due to bradycardia from beta blocker, which he had been taking prior to his initial visit to ***. He had a negative ECHOcardiogram and is now on a 30 day event monitor to look for any arrhythmias . 12/28 : Per nursing report the patient has been perseverative , eating minimally under staff observation and encouragement, had 3 bites of chicken for bites of pineapple last night . He took the increased dose of escitalopram this morning , however the rest of his morning medications fell into his lap, and no longer wanted to take dose, staff will continue to encourage taking all medications .

Figure 1: Sample from a discharge summary included in the corpus. Timex spans are highlighted in yellow and events are highlighted in cyan.