TimelyMed: AI-Driven Clinical Course Attribution and Temporal Mapping for Psychiatric Medical Records

Chien-Hung Chen[1], **Chi-Shin Wu**[2,3], **Chu-Hsien Su**[2], **Hsin-Hsi Chen**[4]
[1]Graduate Institute of Networking and Multimedia, National Taiwan University, Taiwan
[2]National Center for Geriatrics and Welfare Research, National Health Research Institutes, Taiwan
[3]Department of Psychiatry, National Taiwan University Hospital, Yunlin branch, Taiwan
[4]Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
chchen@nlg.csie.ntu.edu.tw, {chishinwu, 100905}@nhri.edu.tw, hhchen@ntu.edu.tw

Abstract

Timely understanding of a patient's clinical course is crucial for effective treatment. Extracting course-related information, such as temporal and medical events, from unstructured medical records is both challenging and time-consuming, especially when relying on manual identification by physicians. We introduce TimelyMed, a system powered by a locally deployed large language model (LLM) that ensures data security while efficiently organizing key psychiatric events and their corresponding temporal information. Additionally, our system is attributed, allowing clinicians to not only categorize events but also trace them back to their original textual descriptions, ensuring transparency and interpretability in clinical decision-making. By organizing temporal and medical event information into timelines, our system enables physicians to quickly grasp a patient's medical history while effectively reducing their cognitive burden.

1 Introduction

Understanding the longitudinal clinical course, which refers to the progression and development of a psychiatric disorder over time, is crucial for effective decision-making in mental health care. Key factors of the clinical course, including symptom onset, episode recurrence, hospitalization history, and treatment response, are crucial for guiding appropriate interventions [Angst and Sellaro, 2000; Emsley *et al.*, 2013; Treuer and Tohen, 2010]. For example, as illustrated in Figure 1, clinicians often need to determine when psychiatric symptoms first appeared, how many times the patient has experienced psychiatric episodes, and when the most recent psychiatric hospitalization occurred from clinical notes. However, manually extracting and organizing course-related information, such as temporal and medical events, from unstructured medical records is a labor-intensive process, often requiring clinicians to reconstruct patient histories through tedious chart reviews or even hand-drawn timelines [Treuer and Tohen, 2010]. This process is not only time-consuming but also prone to human errors and omissions.

Extracting temporal information from medical records presents a challenging task for large language models (LLMs) [Agrawal *et al.*, 2022; Monajatipoor *et al.*, 2024; Ge *et al.*, 2024]. LLMs may struggle to generate structured outputs that meet specific requirements, hindering downstream tasks. Additionally, they are prone to hallucinations. For example, the sentence "*She was titrated to 200 mg on 8/25.*" might be incorrectly interpreted as referring to "August 25, 2002." Moreover, variations in physicians' writing styles can further impede accurate time recognition. For instance, in the sentence "*27 start watching CKMH Psychiatry, MDD was diagnosed.*" the model may fail to recognize that "*27*" refers to the patient's age.[1]

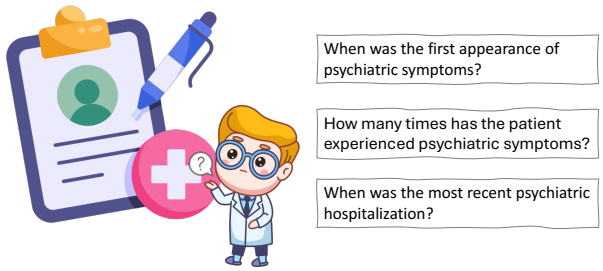

Figure 1: Doctors need to extract key clinical course factors from clinical notes to make decisions.

To address these challenges, we propose TimelyMed,[2] an automated system that extracts temporal and medical event information from psychiatric medical records using a locally deployed LLM. Our system is designed to provide clinicians with an accurate and structured representation of a patient's psychiatric history, reducing the cognitive burden and minimizing errors in timeline reconstruction. TimelyMed leverages a predefined taxonomy of psychiatric events and temporal expressions, allowing it to accurately capture critical courses in a patient's history. Through training on annotated data, we optimize the LLM to extract temporal and medical events from clinical texts. After structuring these elements into a timeline visualization, clinicians can gain a comprehensive view of a patient's mental health trajectory with minimal effort. The attribution feature further enhances the system

[1]We used `Llama-3.1-8B-Instruct` in examples.
[2]http://nlg17.csie.ntu.edu.tw:8501

Category	Example
Symptom/Episode	She began to suffer from low mood, insomnia, and suicidal ideation in 2002.
Remission/Response	Under Solian to 600mg, her symptoms subsided smoothly.
Hospitalization	He was admitted to the acute ward in 2010/02 for 46 days.

Table 1: Examples of Medical Event Categories

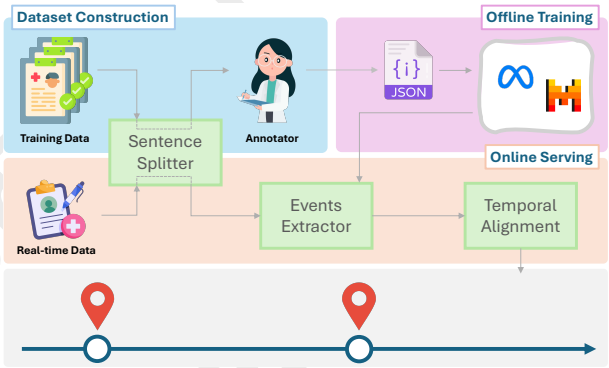

Figure 2: Architecture of the TimelyMed System

by providing direct links between extracted events and their corresponding text, allowing clinicians to verify and contextualize each piece of extracted information.

By converting raw medical records into structured event timelines, TimelyMed enables clinicians to obtain a overview of a patient's psychiatric history, which is critical for both diagnosis and treatment planning. For example, knowing the exact number of psychiatric episodes or the timing of the most recent hospitalization allows for better monitoring of disease progression and focusing on precipitating factors and treatment adherence. Moreover, integrating our system into clinical workflows can support timely decision-making by providing rapid, structured insights from unstructured data sources. This capability is particularly valuable in psychiatric settings, where detailed patient histories are essential for understanding the often chronic and episodic nature of disorders such as schizophrenia, bipolar disorder, and major depressive disorder. The ability to quickly retrieve and synthesize these details from medical records could also improve the accuracy of psychiatric diagnoses and treatment outcomes.

2 Dataset Construction

This study utilized the Integrated Medical Database of National Taiwan University Hospital (NTUH-iMD), approved by the Institutional Review Board of National Taiwan University Hospital (NTUH-201610072RINA). The NTUH-iMD database includes electronic medical records such as admission and discharge notes, and ICD diagnostic codes. This study focused on 500 discharge notes from the psychiatric unit with a principal psychiatric diagnosis (ICD-9-CM: 290–319 or ICD-10-CM: F00-F99) between January 1, 2006, and September 30, 2016. All personal identifying information, including names, addresses, and birthdays, was de-identified, and only sections related to the history of the present illness were included. Furthermore, we apply an additional de-identification tool [Wang *et al.*, 2022; Lee *et al.*, 2024] to replace sensitive details, such as residence, school, workplace, and other elements that could potentially disclose aspects of the patient's private life. Although effective, random replacements can impact data completeness and readability. Therefore, after the tool's application, manual review and modification were conducted to ensure both de-identification and the integrity of the data for subsequent analysis.[3]

All medical chart annotations were performed by a board-certified clinical psychiatrist with 20 years of experience and a researcher specializing in NLP. Annotators were instructed to label each sentence processed by the Sentence Splitter (Stanza toolkit [Qi *et al.*, 2020; Zhang *et al.*, 2021]). Although modern large language models support increasingly long input contexts, allowing most clinical notes to be processed as input, using sentences as the unit of analysis remains a more effective approach. This decision is based on the limited availability of annotated data, computational constraints, and the need for attribution functionality.

Sentences indicating the development or worsening of psychiatric symptoms were categorized as "Symptom/Episode." Those describing treatment remission (fully controlled) or response (slight improvement) were labeled as "Remission/Response," while sentences mentioning admissions to acute psychiatric or rehabilitation wards were classified as "Hospitalization." If a sentence did not contain any of these events, it was categorized as "None". Examples of these medical event categories are provided in Table 1. Notably, a sentence may contain multiple medical events.

Temporal events include "Date", "Age", "Duration", "Ago", and "Persistence". Among them, "Ago" specifies how many years, months, or days before the present admission, while "Persistence" indicates how long a condition continued after a specific event. All temporal events are labeled in their respective standardized formats. A total of 500 discharge summaries were manually annotated and cross-verified by both annotators. In cases of inconsistency, the two reviewers discuss and reach a consensus.

3 TimelyMed System

TimelyMed comprises an offline training phase and an online serving phase (Figure 2). During the offline training phase, a LLM is trained to extract temporal and medical events. In the online serving phase, the system converts clinical notes into course timelines in real time.

[3]The examples in our demo system are written by a psychiatric physician in an anonymous and synthesized manner.

Category	Llama w/o SFT			Llama			Mistral		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Episode/Symptom	0.6974	0.8429	0.7633	0.8768	0.8678	0.8723	0.9233	0.9260	**0.9247**
Hospitalization	0.0000	0.0000	0.0000	0.9385	0.8744	0.9054	0.9817	0.9082	**0.9436**
Remission/Response	0.5342	0.6710	0.5949	0.8093	0.7546	0.7810	0.8805	0.8541	**0.8671**

Table 2: Performance Comparison on Medical Event Extraction Task. "Llama w/o SFT" stands for Llama model without supervised fine-tuning. Since "Mistral w/o SFT" only output "Episode" and "None", we did not report the results here.

Category	Llama w/o SFT			Llama			Mistral		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Age	0.6233	0.7546	0.6827	0.9357	0.9441	0.9399	0.9624	0.9732	**0.9677**
Ago	0.3380	0.3020	0.3190	0.8858	0.8932	0.8895	0.9291	0.9378	**0.9334**
Duration	0.1249	0.4228	0.1928	0.8506	0.8397	0.8452	0.9142	0.8878	**0.9008**
Persistence	0.0376	0.0478	0.0421	0.5714	0.3439	0.4294	0.7371	0.6471	**0.6892**
Time	0.6286	0.6401	0.6343	0.9461	0.9461	0.9461	0.9648	0.9683	**0.9666**

Table 3: Performance Comparison on Temporal Event Extraction Task. "Llama w/o SFT" stands for Llama model without supervised fine-tuning. Since "Mistral w/o SFT" failed to follow our output format correctly, we did not report the results here.

Identify the type of medical events and corresponding temporal events from a given sentence in a clinical note.

 # Steps

1. Analyze the Sentence: Break down the sentence to identify keywords or phrases that indicate a medical event type.
2. Determine Medical Event: Categorize the event as 'Episode/Symptom', 'Hospitalization', 'Remission/Response', or 'None' based on the context and identified keywords.
3. Identify Temporal Event: Recognize any temporal indicators to classify the temporal event as 'Age', 'Ago', 'Duration', 'Persistence', 'Date' or 'None'.

 # Output Format

Provide the output as a JSON object with two fields, "medical event" and "temporal event", indicating the identified types.
{"medical event": "[Medical Event Type]", "temporal event": "[Temporal Event Type]"}
<demonstrations>

Notes

- If the sentence does not clearly indicate any medical or temporal event, classify it as 'None' for one or both categories.
- Pay close attention to temporal cues, as they are crucial for accurately identifying the time event type.
- A sentence may contain multiple medical events; separate them with commas.

Figure 3: System Prompt Design for the TimelyMed System

3.1 Offline Training

Based on annotated data, we leverage supervised fine-tuning to train an Events Extractor to extract temporal and medical events from clinical notes. We employ the LoRA [Hu *et al.*, 2022] to fine-tune LLMs. Besides, we design a system prompt that specifies the task description, event categories, output format, demonstrations, and key constraints, as shown in Figure 3.

3.2 Online Serving

The process begins by segmenting the clinical notes into sentences. A well-trained LLM is then employed to extract both temporal and medical event information. Following extraction, a Temporal Alignment step is applied to format the data according to the requirements of the timeline package.[4] Specifically, all time-related information is converted into a standardized date format to ensure consistency. Additionally, if a medical event is identified without an accompanying temporal event, the system searches the preceding sentence for a temporal event. This process continues iteratively until a relevant temporal reference is found, ensuring that every medical event is appropriately anchored in time. Finally, we will include the source sentences as attribution in the timeline to enhance the system's reliability.

[4]https://github.com/innerdoc/streamlit-timeline

4 Experiments and Discussion

We conducted an evaluation of three models: LLaMA (`Llama-3.1-8B-Instruct`) [Touvron *et al.*, 2023] and Mistral (`Ministral-8B-Instruct-2410`) [Jiang *et al.*, 2023]. The models were assessed using 10-fold cross-validation to ensure robust performance estimation. Model effectiveness was measured using Precision, Recall, and F1-score as evaluation metrics.

Table 2 and Table 3 present the performance results for medical and temporal event extraction, respectively. The results indicate that Mistral achieved the best overall performance after training. In the medical event task, "Remission/Response" exhibited the weakest performance, likely due to the diverse terminology used by different physicians to describe symptom control and alleviation. In the temporal event task, "Persistence" performed poorly, possibly because of its low occurrence in the dataset (approximately 3.5%).

5 Conclusion and Future Work

TimelyMed provides an efficient and privacy-preserving solution for extracting and organizing psychiatric course timelines from medical records. By leveraging a locally deployed LLM, the system ensures data security while enhancing clinical workflows through automated event extraction and attribution. This approach reduces the cognitive burden on clinicians, improves decision-making, and supports more accurate psychiatric diagnoses and treatment planning. Future work will focus on refining event classification and expanding the system's applicability across diverse psychiatric conditions.

Acknowledgments

This work was supported by Ministry of Technology and Science, Taiwan , under grants MOST110-2314-B-400-053-MY3, National Science and Technology Council, Taiwan, under grants NSTC 112-2634-F-002-005-, and Ministry of Education (MOE), Taiwan, under grants NTU-113L900901.

Contribution Statement

Author Chien-Hung Chen and Author Chi-Shin Wu contributed equally to this work.

References

[Agrawal *et al.*, 2022] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are few-shot clinical information extractors. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[Angst and Sellaro, 2000] Jules Angst and Robert Sellaro. Historical perspectives and natural history of bipolar disorder. *Biological psychiatry*, 48(6):445–457, 2000.

[Emsley *et al.*, 2013] Robin Emsley, Bonginkosi Chiliza, Laila Asmal, and Brian H Harvey. The nature of relapse in schizophrenia. *BMC psychiatry*, 13:1–8, 2013.

[Ge *et al.*, 2024] Xueren Ge, Abhishek Satpathy, Ronald Dean Williams, John Stankovic, and Homa Alemzadeh. DKEC: Domain knowledge enhanced multi-label classification for diagnosis prediction. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12798–12813, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[Hu *et al.*, 2022] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[Jiang *et al.*, 2023] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[Lee *et al.*, 2024] You-Qian Lee, Ching-Tai Chen, Chien-Chang Chen, Chung-Hong Lee, Peitsz Chen, Chi-Shin Wu, and Hong-Jie Dai. Unlocking the secrets behind advanced artificial intelligence language models in deidentifying chinese-english mixed clinical text: Development and validation study. *J Med Internet Res*, 26:e48443, Jan 2024.

[Monajatipoor *et al.*, 2024] Masoud Monajatipoor, Jiaxin Yang, Joel Stremmel, Melika Emami, Fazlolah Mohaghegh, Mozhdeh Rouhsedaghat, and Kai-Wei Chang. Llms in biomedicine: A study on clinical named entity recognition. *arXiv preprint arXiv:2404.07376*, 2024.

[Qi *et al.*, 2020] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[Treuer and Tohen, 2010] T Treuer and M Tohen. Predicting the course and outcome of bipolar disorder: a review. *European Psychiatry*, 25(6):328–333, 2010.

[Wang *et al.*, 2022] Chen-Kai Wang, Feng-Duo Wang, You-Qian Lee, Pei-Tsz Chen, Bo-Hong Wang, Chu-Hsien Su, Joseph Chin-Chi Kuo, Chi-Shin Wu, Yi-Ling Chien, Hong-Jie Dai, Vincent S. Tseng, and Wen-Lian Hsu. Principle-based approach for the de-identification of code-mixed electronic health records. *IEEE Access*, 10:22875–22885, 2022.

[Zhang *et al.*, 2021] Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D Manning, and Curtis P Langlotz. Biomedical and clinical english model packages for the stanza python nlp library. *Journal of the American Medical Informatics Association*, 28(9):1892–1899, 06 2021.