Using Multimodal Data to Improve Precision of Inpatient Event Timelines

Gabriel Frattallone-Llado[1*], Juyong Kim[2*], Cheng Cheng[2], Diego Salazar[4], Smitha Edakalavan[3], and Jeremy C. Weiss[4]

[1] Universidad de Puerto Rico, San Juan, PR 00926-1117, Puerto Rico
gabriel.frattallone2@upr.edu
[2] Carnegie Mellon University, Pittsburgh PA 15206, USA
{juyongk,ccheng2}@andrew.cmu.edu
[3] University of Pittsburgh, Pittsburgh, PA 15260, USA SME57@pitt.edu
[4] National Library of Medicine, Bethesda, MD 20894, USA
{jeremy.weiss,diego.salazar}@nih.gov

Abstract. Textual data often describe events in time but frequently contain little information about their specific timing, whereas complementary structured data streams may have precise timestamps but may omit important contextual information. We investigate the problem in healthcare, where we produce clinician annotations of discharge summaries, with access to either unimodal (text) or multimodal (text and tabular) data, (i) to determine event interval timings and (ii) to train multimodal language models to locate those events in time. We find our annotation procedures, dashboard tools, and annotations result in high-quality timestamps. Specifically, the multimodal approach produces more precise timestamping, with uncertainties of the lower bound, upper bounds, and duration reduced by 42% (95% CI 34-51%), 36% (95% CI 28–44%), and 13% (95% CI 10–17%), respectively. In the classification version of our task, we find that, trained on our annotations, our multimodal BERT model outperforms unimodal BERT model and Llama-2 encoder-decoder models with improvements in F1 scores for upper (10% and 61%, respectively) and lower bounds (8% and 56%, respectively). The code for the annotation tool and the BERT model is available (link).

Keywords: Temporal Information · Timeline Construction · Multimodal Data · Absolute Timeline Prediction.

1 Introduction

Temporal data mining involves the extraction of temporal information from different sources and modalities of data, and it has broad application in fields such as law, finance, and healthcare. For instance, in criminal recidivism prediction, the event timeline for a defendant could be extracted from both texts in probation office documents and tables in psychiatric health records [2]; In

Pre-print. G. Frattallone-Llado and J. Kim—Equal contribution.

stock price movement and volatility prediction, financial time series could be extracted from financial news, daily stock market price tables, and verbal and vocal cues in earning calls [1]. In clinical risk prediction, patient timeline could be extracted from electronic health records with both unstructured clinical notes and structured tabular data [10]. This paper focuses on providing a multimodal extraction system and a benchmark dataset for clinical timelines.

Precise clinical event timelines are crucial for prognosis and prediction tasks. These forecasting tasks have been studied with varied prediction times and unstructured data sources [7]. Discharge summaries provide the most complete information. In the 2012 Informatics for Integrating Biology and Bedside (i2b2) Challenge [9], 310 discharge summaries were annotated with temporal information, including clinical events, temporal expressions, and their temporal relations.

This approach yields a relative timeline of clinical events, rather than absolute, leaving many events without the precise timing needed for forecasting tasks. To achieve a more complete event timeline, i2b2 events were annotated with absolute time values [3], by bounding the events with closed intervals in calendar times and temporal uncertainties on the bounds. Their annotation procedure was unimodal, as annotators had access only to the discharge summary text.

Meanwhile, there is a consensus that the integration of structured and unstructured data has a significant impact on constructing models and predicting target variables [7]. In this project, we take advantage of the combination of unstructured and structured data in a multimodal approach, which has proven beneficial in other applications [4,5]. Our work adopts a version of the probabilistic bounds described and applies them to discharge summaries from the i2b2 dataset in order to generate absolute inpatient event timelines. We introduce the following: a visualization and annotation tool, an annotation process with a three-pass system, two types of annotations to better represent the nature of clinical events, and the multimodal annotation approach. By combining the information from unstructured and structured data, this multimodal approach should yield a more precise (*i.e.*, less uncertain) timeline for inpatient events.

Our multimodal approach contributes the following: (i) we introduce absolute timeline intervals without assumption of independence of endpoint uncertainties, (ii) we find that multimodal annotations lead to more precise timelines than the unimodal annotations. (iii) we verify the annotation quality by mapping our annotations to temporal relations, where our relations compare favorably against benchmark annotations (i2b2), and (iv) we demonstrate that a fine-tuned multimodal encoder (BERT) architecture outperforms fine-tuned unimodal encoder and off-the-shelf generative encoder-decoder architectures (Llama-2). Overall, we show the importance of annotation from multimodal data sources, both in the annotation process and for machine learning predictive performance.

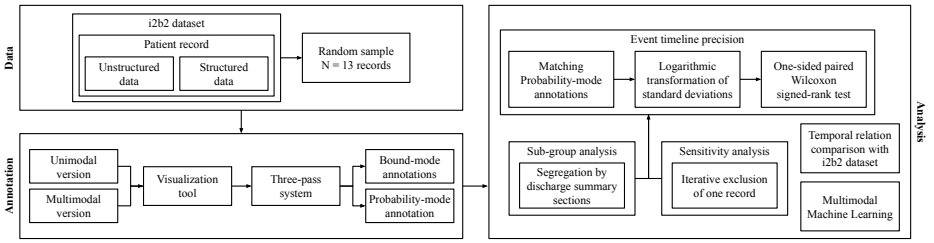

Fig. 1. Flowchart of our method.

2 Methods

We use the i2b2 dataset [12], a compendium of de-identified electronic health records from Partner Healthcare and Beth Israel Deaconess Medical Center, containing discharge summaries with annotated clinical time/events. We located i2b2 patient's records in MIMIC-III with matching discharge summaries, allowing us to collect both structured and unstructured records. Thirteen records were randomly selected from the training set of the 2012 i2b2 temporal relations challenge data [9]. For each of these records, a human annotator with medical experience identified clinical events in the discharge summary and timestamped their endpoints on a timeline.

In our study, an annotation comprises a contiguous, highlighted text, representing a clinical event, and a time interval. Based on the use of the structured and unstructured data, two annotation versions were generated: multimodal and unimodal. In the unimodal version, the annotator only had access to the discharge summary and the admission and discharge times. In the multimodal version, the annotator also had access to the full structured data. Out of the total thirteen records, five were annotated using both unimodal and multimodal versions, while the remaining records were annotated from only multimodal. Upon acceptance, the annotation tool, the annotation files, and the analyses will be made public. A flowchart representation of the process describes the steps of our annotation and analysis (Fig. 1).

2.1 Annotation Tool

The R Shiny annotation tool displays all the unstructured and structured data for any given record (Fig. 2). The structured data is displayed on a graph, where the x-axis contains absolute time values and the y-axis contains event identifiers. The user annotates time intervals by clicking and dragging on the graph (blue overlay). Structured data events contained within this overlay are displayed in a table that the user can search and select to be relevant. On the unstructured data, or discharge summary, section, the user selects the relevant span for annotation by highlighting the text. The polarity of negative events is specified with a checkbox. Both types of annotation (Bounds- and Probability- mode) require the user to generate an overlay on the graphical timeline. In Bounds-mode

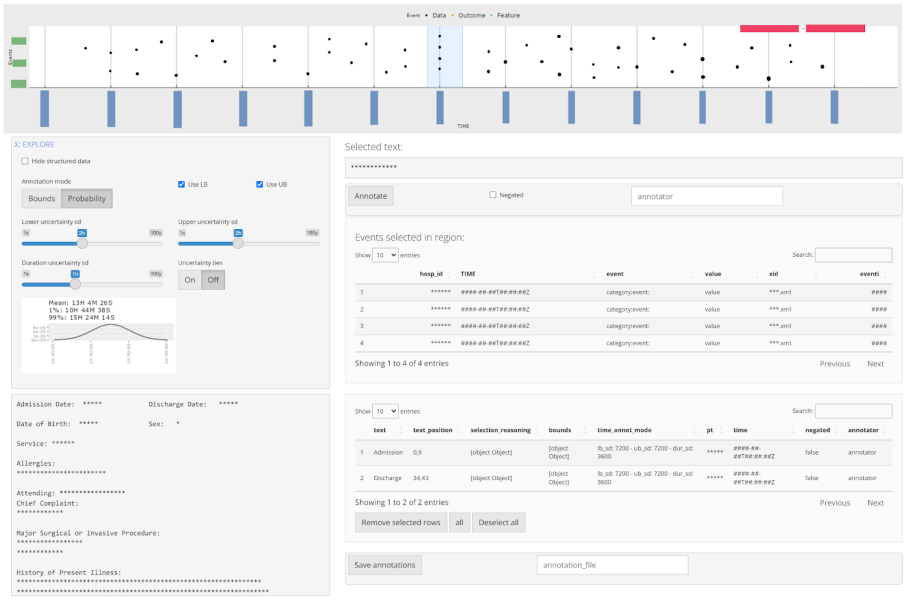

Fig. 2. The web-based annotation tool.

annotation, either the lower or upper bound can be omitted by unchecking a box, which represents *indefiniteness*. In Probability-mode annotation, users also select standard deviations to represent their uncertainty about the lower bound, upper bound, and duration of the event. The choices are pre-defined as follows: 1, 3, or 10 seconds; 1, 3, 10, or 30 minutes; 1, 2, 4, 6, 12, or 24 hours; 2 days; 1 or 2 weeks; 1 month; 1, 10, or 100 years.

2.2 Annotation Process

The annotation process consisted of a three-pass system defined as follows: (1) the annotator peruses the document in its entirety to become familiarized with its content, (2) the annotator makes annotations at the paragraph level, setting lower and upper bounds that apply to all the events in that paragraph, and (3) the annotator makes annotations on particular events, which could represent words or phrases. Events are selected according to the i2b2 annotation guidelines [8]. For simplicity, the one event attribute recorded is the polarity. As stated previously, two types of annotations are available: Bounds-mode and Probability-mode. In Bounds-mode annotation, lower and upper bounds are defined, with the expectation that the event(s) occurs at some point between them, and no information is specified about event duration or timing uncertainty. One can omit either bound when the value is unknowable, as in the lower bound for certain conditions from the past medical history. In this case, the bound defaults to negative or positive infinity, as appropriate. Bounds-mode annotation

is predominantly used in the second pass, because each paragraph may contain various events with different timing attributes. In Probability-mode annotation, the selected lower and upper bounds represent the mean values of two distributions. The mean value of the event's duration is calculated as their difference. All three distributions are assumed to be normal and the annotator must select a standard deviation for each one. These standard deviations represent the annotator's level of uncertainty about the event's timing and will serve as a surrogate of timeline precision for data analysis. Probability-mode annotation is used predominantly in the third pass, where the duration and approximate bounds of particular events may be reasonably determined.

2.3 Statistical Methods

The annotation process above was performed on the thirteen randomly-selected records from the i2b2 training dataset [9]. Five records are annotated from both unimodal and multimodal data by performing the process twice for each of these records, while the rest are annotated solely from the multimodal version of the data. In total, eighteen annotation files were generated, comprising five unimodal and thirteen multimodal versions. These files contain 4884 annotations in total, of which 1156 were in Bounds-mode and 3728 were in Probability-mode.

We performed two descriptive analyses. For the first analysis, we focus on the ten annotation files originated from both unimodal and multimodal data to study and compare the annotation practices across both versions. We defined the main effect as the precision of the event timelines (multimodal vs unimodal), *i.e.*, the difference in uncertainty of events is obtained by comparing the standard deviation for the lower bound, upper bound, and duration for multimodal and unimodal annotations. To compare these event timelines, we selected exact matches from the Probability-mode annotations, and we performed a one-sided paired Wilcoxon signed-rank test on the logarithmic transformation of the standard deviations. The resulting estimated differences between the two versions correspond to a scaling factor, which reflects the degree of increase or decrease in uncertainty, which we call *precision factor*. Confidence intervals were calculated using the bootstrap. To verify the robustness of the results, a sensitivity analysis was performed by re-running the test while iteratively excluding one record pair. Sub-analyses were also performed on different sections of the discharge summaries, termed Medical History (history of present illness, past medical history, home medications, allergies, social and family history), Hospital Stay (hospital course, procedures, service), Examinations and Findings (physical exam, laboratory tests, images, and studies), and Discharge (condition, disposition, instructions, medications, diagnosis).

For the second analyses, we sought to assess the compatibility between our annotations and the i2b2 dataset. We aligned the events in the thirteen multimodal annotation files to the events in the corresponding i2b2 dataset. Since the i2b2 dataset provides only temporal relations between two events – "BEFORE", "AFTER", or "OVERLAP" – while ours provides the absolute time of events, we extracted those i2b2 temporal relations where both text endpoints could be

matched with our annotated events, and compared to the temporal relations computed from our annotations. We used character-level intersection over union to match the text span of the i2b2 events and ours. We restricted our events to be aligned to be Probability-mode and calculated a pair of z-scores: one indicating the likelihood that one event precedes the other and another for the opposite order. If exactly one z-score exceeds a predetermined threshold, we determined that one event precedes the other, otherwise they overlap. After matching the temporal relations between our dataset and the i2b2 dataset, we report the F1 score, inter-annotator agreement, and accuracy of the types of relations.

2.4 Multimodal Learning

To test the utility of multimodal data over single-modal data, we study a reduced version of our absolute timeline prediction. The task involves binary classifications of whether the lower bound (LB) and upper bound (UB) are finite, identifying the type of annotation (Bounds- or Probability-mode), and multiclass classifications of the bounds and the standard deviations of LB, UB, and duration. Each of the bound and standard deviation classifications consists of three classes defined as follows. Bounds are classified relative to admission time, with thresholds at admission and 24 hours post-admission. Standard deviations for LB and UB are grouped into three classes: under 2 hours, 4 hours to 1 day, and over 2 days. Duration standard deviations follow a similar pattern, classified as under 1 hour, 2 hours to 1 day, and over 2 days.

The experiment on the classification version of our dataset employed two BERT-based models. The first model, named Unimodal BERT, is a span classification model [13] where the contextualized BERT embeddings of the annotated text span from clinical notes along with its context are fed to a feedforward network for classification. The second model, named Multimodal BERT, also incorporates the structured event data by applying multi-head attention. In this setup, the BERT embedding of the text span serves as the query, and the keys and values are the contextualized embeddings of the names and the one-hot encodings of the timestamps in the table. The resulting weighted sum of the one-hot encodings is taken logarithm and then added to the logits of the mean predictions. The base BERT of both models is initialized with the BlueBERT-Base [6].

To fit the input length shorter than 128, the maximum sequence length of BlueBERT, we filtered out paragraph-level events that were annotated in the second pass. The fourteen annotation files[5] are split into 5 groups to perform 5-fold cross-validation, and we report the average of test performances across three different random seeds. For all the experiments, we trained the models for 20 epochs with a learning rate of 5e-5. For comparison, we also report the performance of the majority selection baseline, the baseline using Llama-2 (llama-2-13b-chat-hf) on the annotated clinical notes (same as Unimodal BERT) to generate the LB

[5] One additional annotation file that 2012 i2b2 dataset does not include is added.

and UB through few shot prompting [11], and the baseline selecting the LB and UB based on the ground truth selection of the structured events.

3 Results

3.1 Comparison across different modalities

An exploratory data analysis of the 10 paired annotation files (5 multimodal and 5 unimodal) revealed a total of 2718 annotations, of which 693 were in Bounds-mode and 2025 were in Probability-mode. For each record, the number of annotations and their distribution between Bounds and Probability modes were very similar between the unimodal and multimodal versions. Furthermore, 95.7% of the events were annotated as exact matches across versions, meaning that the selected text had the same start and end positions, *i.e.*, span, regardless of annotation mode (Tab. 1A).

Across discharge summary sections, there were relatively more Bounds-mode annotations in the Medical History (67% bounds) and Discharge (35%) sections, and fewer in the larger Hospital Stay (7%) and Examinations and Findings (17%), indicating increased ability to annotate intraencounter events in Probability-mode. Across sections, there was high agreement between unimodal and multimodal annotation, shown by the proportion of exact matches (Tab. 1B).

All matching events annotated in Probability-mode were used to perform a Wilcoxon signed-rank test. We compared the standard deviations annotated for lower bound, upper bound, and duration. The standard deviations reported in the multimodal version were significantly smaller than those reported in the unimodal version. This trend was consistent across the lower bound (p-value < 0.001), upper bound (p-value < 0.001), and duration (p-value < 0.001). In general, we increased the precision on average by a factor of 1.42, 1.36, and 1.13 for the lower bound, upper bound, and duration, respectively. These values suggest that the uncertainty in multimodal annotations was reduced by 42%, 36%, and 13%, compared to unimodal annotations for the corresponding types of bounds and duration (Fig. 3).

Similarly, we conducted a comparison of the standard deviations pertaining to bounds and duration types across various sections present in clinical notes (Fig. 3). Our findings indicate that, in the case of bounds, annotations exhibited an improvement in precision in all the sections. However, for the duration, there was no observed change in uncertainty, except for the Medical History, Exams and Findings, and Discharge sections.

In the sensitivity analysis, the standard deviation effect estimates (lower, upper, and duration) remained close to sample estimate, *i.e.*, the effect of increased precision in the multimodal annotations was maintained. The estimated precision factors were for lower bound, upper bound, and duration: 1.42 (95% CI [1.34-1.51]), 1.36 (95% CI [1.28-1.44]), and 1.13 (95% CI [1.10-1.17]) respectively. The Wilcoxon sign-rank test again showed differences between both versions of annotation (p-value < 0.001).

Comparison of Annotation Practices Across Versions					
			Bounds-Mode	Probability-Mode	Exact Matches
Record[1]	**Version**	**Annotations**	# (%)	# (%)	# (%)
A	Unimodal	161	45 (28.0%)	116 (72.0%)	156 (96.9%)
A	Multimodal	165	46 (27.8%)	119 (72.1%)	156 (94.5%)
B	Unimodal	379	53 (14.0%)	326 (86.0%)	354 (93.4%)
B	Multimodal	380	60 (15.8%)	320 (84.2%)	354 (93.2%)
C	Unimodal	182	58 (31.9%)	124 (68.1%)	167 (91.8%)
C	Multimodal	180	58 (32.2%)	122 (67.8%)	167 (92.8%)
D	Unimodal	352	125 (35.5%)	227 (64.5%)	347 (98.6%)
D	Multimodal	361	115 (31.9%)	246 (68.1%)	347 (96.1%)
E	Unimodal	278	64 (23.0%)	214 (77.0%)	276 (99.3%)
E	Multimodal	280	69 (24.6%)	211 (75.4%)	276 (98.6%)
Total	**-**	**2718**	**693 (25.5%)**	**2025 (74.5%)**	**2600 (95.7%)**

[1]Record numbers have been redacted

(A) Across Versions

Comparison of Annotation Practices Across Sections					
			Bounds-Mode	Probability-Mode	Exact Matches
Section	**Version**	**Annotations**	# (%)	# (%)	# (%)
Medical History	Unimodal	244	161 (66.0%)	83 (34.0%)	234 (95.9%)
Medical History	Multimodal	249	167 (67.1%)	82 (32.9%)	234 (94.0%)
Exams and Findings	Unimodal	484	29 (6.0%)	455 (94.0%)	467 (96.5%)
Exams and Findings	Multimodal	489	32 (6.5%)	457 (93.5%)	467 (95.5%)
Hospital Stay	Unimodal	366	59 (16.1%)	307 (83.9%)	354 (96.7%)
Hospital Stay	Multimodal	369	62 (16.8%)	307 (83.2%)	354 (95.9%)
Discharge	Unimodal	248	96 (38.7%)	152 (61.3%)	234 (94.4%)
Discharge	Multimodal	248	87 (35.1%)	161 (64.9%)	234 (94.4%)

(B) Across Sections

Table 1. Comparison of annotation practices across unimodal and multimodal versions (A) and across discharge summary sections (B), with number of annotations per record/per section, their breakdown into Bounds-mode and Probability-mode, and the exact match between the versions.

3.2 Comparison with i2b2 dataset (2012)

The total number of events in the thirteen annotation files derived from the multimodal data is 3532, with 2721 in Probability-mode annotations. The corresponding i2b2 training data files have 1024 events ("EVENT" tag) and 2140 temporal relations ("TLINK" tag), with 901 comparing events only. Note these i2b2 files do not cover the entire discharge summary. By aligning the text spans with the character-level IOU threshold of 0.5, 761 of the Probability-mode events are matched with the i2b2 events, and 490 of the i2b2 temporal relations can be mapped to our dataset. We use -1.0 for the threshold of z-scores to determine the temporal relations between two Probability-mode events. The confusion matrix between the i2b2 temporal relations and our annotations' temporal relations is shown in Fig. 4. The accuracy, the macro F1-score, and the Cohen's kappa score between the i2b2 and our temporal relations are 0.698, 0.599, and 0.383, respectively. In other words, 70% of the temporal links between the two events agree between our annotations and the i2b2 dataset. The reason for the low kappa score is imbalanced distribution of the categories, leading to a high expected agreement of random assignments.

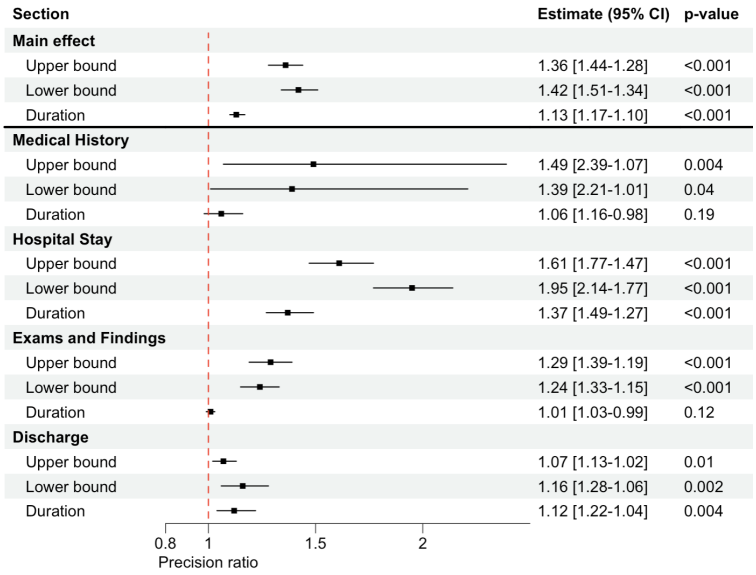

Fig. 3. Precision factor between the two types of annotations. A value greater than 1 indicates a reduction of uncertainty in multimodal annotations (p-values obtained from the Wilcoxon sign-rank test).

To more thoroughly compare the temporal relations between the i2b2 dataset and ours, we conducted human assessment. For each type of mismatched relation, we randomly selected three examples, resulting in a total of 18 temporal relations. Then, a human with medical experience, who was not involved in the annotation process, evaluated the selected examples by examining the input note. This individual was asked to judge which annotation represented temporal relations more accurately. The results of the assessment are shown in Table 2. Out of the 18 examples, 9 were found to be better annotated in our dataset, while 8 were more accurately represented in the i2b2 dataset. This result suggests that the temporal relations solely from textual information are not complete, emphasizing the importance of incorporating structured data, but also suggests potential improvement in our multimodal based annotation process.

3.3 Multimodal Learning

Table 3 shows the results of the classification version of absolute timeline prediction. The F1 score of the lower bound prediction (Mean-LB) and the upper bound prediction (Mean-UB) of Multimodal BERT improved 10% (0.604 vs. 0.551) and 8% (0.680 vs. 0.632) from Unimodal BERT, respectively. This improvements, stemming from the integration of multi-head attention into the final mean prediction logits, demonstrate the benefits of integrating unstructured text with structured patient data. Comparing with Llama-2, our Multimodal BERT

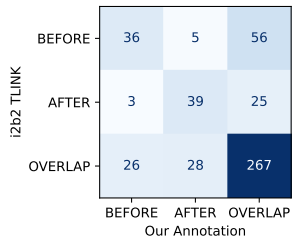

Fig. 4. Confusion matrix between the temporal relations of the 2012 i2b2 dataset and our dataset

Temporal relation (i2b2 / Ours)	Judgment		
	i2b2	Ours	Both
BEFORE / AFTER	2	1	0
BEFORE / OVERLAP	2	1	0
AFTER / BEFORE	1	1	1
AFTER / OVERLAP	0	3	0
OVERLAP / BEFORE	2	1	0
OVERLAP / AFTER	1	2	0
Total	**8**	**9**	**1**

Table 2. Human assessment on the 18 sampled mismatched temporal relations

Classification Version of Absolute Timeline Prediction								
	Type			Mean		Std Dev		
Method	LB inf[†1]	UB inf[†1]	Anno[†1]	LB[2]	UB[2]	LB[2]	UB[2]	Dur[2]
Tabular (Oracle)	-	-	-	0.594	0.789	-	-	-
Majority	0.859	1.000	0.840	0.224	0.238	0.220	0.207	0.267
Llama-2	-	-	-	0.374	0.441	-	-	-
Unimodal BERT	0.935	1.000	0.908	0.551	0.632	**0.446**	**0.447**	0.573
Multimodal BERT	**0.936**	1.000	**0.912**	**0.604**	**0.680**	0.433	0.436	**0.579**

[†]LB/UB inf: definiteness of LB/UB, Anno: Bounds- or Probability-mode
[1]Accuracy, [2]Macro-averaged F1 score

Table 3. Results of the classification version of absolute timeline prediction on our dataset. The best results among the non-oracle methods are highlighted in bold.

shows 61% (0.604 vs. 0.374) and 54% (0.680 vs. 0.441) higher F1 score in bound predictions. While the structured data baseline (Tabular (Oracle)) showed the best results in the upper bound mean prediction, this was under the idealized assumption of perfect attention on the structured data. The inferior lower bound mean prediction of the structured data baseline, compared to its upper bound, stems from the mismatch in labels between events anchored to the admission and the admission time itself.

4 Discussion and Conclusions

Clinical events from unstructured data provide information about the patient's progression, but placing them on an absolute timeline can be challenging. Yet while the timing of events in structured data is more certain, the structured data may miss events or other predictive insights. When both types of data are in alignment following the patient's clinical course, their combination may generate more complete and precise event timelines. Our work demonstrates the benefit of the multimodal approach, both in quality of annotation and prediction.

When comparing Probability-mode annotations, statistical analysis revealed superior precision for the multimodal version of the timeline in all three temporal entities (lower bound, upper bound, and duration). The precision factor for

the bounds was similar at 1.40 and 1.34 (lower and upper, respectively), whereas for the duration it was at 1.13. This follows the intuition that the duration of many clinical events can be estimated based on clinical knowledge. In particular, the specific position of events on a timeline is more uncertain and depends on many factors that cannot be predicted with clinical knowledge alone, and having relevant structured data with precise timestamps greatly reduced timing uncertainty for the lower and upper bounds of events.

A subgroup analysis compared the precision factor of the timelines within different sections of the discharge summary. In the case of the bounds, multimodal annotation increased precision in all sections, but most prominently in the Hospital Stay and Medical History. The improvement in the Hospital Stay section is expected since the vast majority of the structured data parallels the patient's clinical course from admission to discharge. The Medical History section usually describes events that occurred prior to admission. Likewise, some of the structured data is generated during the patient's time in the emergency department, prior to admission. The Exam and Findings section also showed a moderate improvement in precision with the multimodal version. It contains laboratory tests and imaging studies, which could frequently be referenced to structured data with precise timestamps. Thus, the uncertainty of event timing is significantly reduced when the structured data is aligned with the unstructured data,

In the case of event duration, the multimodal version yielded statistically significant improvement in precision only in the Discharge and Hospital Stay section. Events in the Discharge section are very likely to have their upper bound anchored to the time of discharge, a event that was also available in the unimodal annotation. Thus, little improvement is seen in the precision of the upper bound when the rest of the structured data is made available. This aligns with the result that the upper bound experiences less improvement than the lower bound in the Hospital Stay since events here also extend until discharge.

This study has several limitations. First, one annotator was used, which precludes measurement of inter-annotator agreement. Additionally, the uncertainties defined for the bounds and duration could violate the positive semi-definite condition of a multivariate normal distribution, which may be oversimplified for the annotator's belief about the interval. Since a small sample size was used, the selected discharge summaries may not be representative, *e.g.*, due to differences in chief complaints, institutional policies, or note templates. Finally, there was no functionality for adding additional meta-data to events or recording temporal relations without established timelines.

The reported findings support the use of multimodal data to generate more precise event timelines when compared to unstructured data alone. The benefit is especially prominent when a large quantity of structured data aligns with the unstructured data. Further areas of study could include working with a larger sample size and analyzing differences when subjects are measured more frequently, *e.g.*, in critical care units versus the hospital floor.

The compatibility analysis with the i2b2 dataset validated that our dataset provides a complement to the existing text-based annotations. In the multimodal

learning experiments, a BERT model leveraging structured events through multi-head attention improved F1 scores for predicting lower and upper bounds over the unimodal BERT. This demonstrates how temporal localization of clinical events benefits from jointly modeling text and structured data sources. Overall, these strongly support the enhanced utility of our multimodal annotation approach for generating more precise absolute timelines of inpatient events.

Acknowledgments

This research was supported in part by the Intramural Research Program of the National Library of Medicine (NLM), National Institutes of Health.

References

1. Ang, G., Lim, E.P.: Guided attention multimodal multitask financial forecasting with inter-company relationships and global and local news. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (May 2022)
2. Cheng, T.T., Cua, J.L., Tan, M.D., Yao, K.G., Roxas, R.E.: Information extraction from legal documents. In: 2009 eighth international symposium on natural language processing. pp. 157–162. IEEE (2009)
3. Leeuwenberg, A., Moens, M.F.: Towards extracting absolute event timelines from english clinical reports. IEEE/ACM Transactions on Audio, Speech, and Language Processing **28**, 2710–2719 (2020)
4. Liu, S., Wang, X., et al.: Multimodal data matters: language model pre-training over structured and unstructured electronic health records. IEEE Journal of Biomedical and Health Informatics **27**(1), 504–514 (2022)
5. Moldwin, A., Demner-Fushman, D., Goodwin, T.R.: Empirical findings on the role of structured data, unstructured data, and their combination for automatic clinical phenotyping. AMIA Summits on Translational Science Proceedings (2021)
6. Peng, Y., Chen, Q., Lu, Z.: An empirical study of multi-task learning on bert for biomedical text mining. In: Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing. pp. 205–214 (2020)
7. Seinen, T.M., Fridgeirsson, E.A., et al.: Use of unstructured text in prognostic clinical prediction models: a systematic review. Journal of the American Medical Informatics Association **29**(7), 1292–1302 (2022)
8. Sun, W., Rumshisky, A., Uzuner, O.: Annotating temporal information in clinical narratives. Journal of biomedical informatics **46**, S5–S12 (2013)
9. Sun, W., Rumshisky, A., Uzuner, O.: Evaluating temporal relations in clinical text: 2012 i2b2 challenge. Journal of the American Medical Informatics Association **20**(5), 806–813 (2013)
10. Tayefi, M., et al.: Challenges and opportunities beyond structured data in analysis of electronic health records. Computational Statistics (2021)
11. Touvron, H., Martin, L., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint (2023)
12. Uzuner, Ö., South, B.R., Shen, S., DuVall, S.L.: 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association **18**(5), 552–556 (2011)
13. Zhong, Z., Chen, D.: A frustratingly easy approach for entity and relation extraction. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics. pp. 50–61 (2021)