Forecasting Clinical Risk from Textual Time Series: Structuring Narratives for Temporal AI in Healthcare

Shahriar Noroozizadeh[1, 2*], **Sayantan Kumar**[3*], **Jeremy C. Weiss**[3]

[1]Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
[2] Heinz College of Information Systems and Public Policy, Carnegie Mellon University, Pittsburgh, PA, USA
[3] National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
snoroozi@cs.cmu.edu, {sayantan.kumar, jeremy.weiss}@nih.gov

Abstract

Clinical case reports encode temporal patient trajectories that are often underexploited by traditional machine learning methods relying on structured data. In this work, we introduce the forecasting problem from textual time series, where timestamped clinical findings—extracted via an LLM-assisted annotation pipeline—serve as the primary input for prediction. We systematically evaluate a diverse suite of models, including fine-tuned decoder-based large language models and encoder-based transformers, on tasks of event occurrence prediction, temporal ordering, and survival analysis. Our experiments reveal that encoder-based models consistently achieve higher F1 scores and superior temporal concordance for short- and long-horizon event forecasting, while fine-tuned masking approaches enhance ranking performance. In contrast, instruction-tuned decoder models demonstrate a relative advantage in survival analysis, especially in early prognosis settings. Our sensitivity analyses further demonstrate the importance of time ordering, which requires clinical time series construction, as compared to text ordering, the format of the text inputs that LLMs are classically trained on. This highlights the additional benefit that can be ascertained from time-ordered corpora, with implications for temporal tasks in the era of widespread LLM use.

Introduction

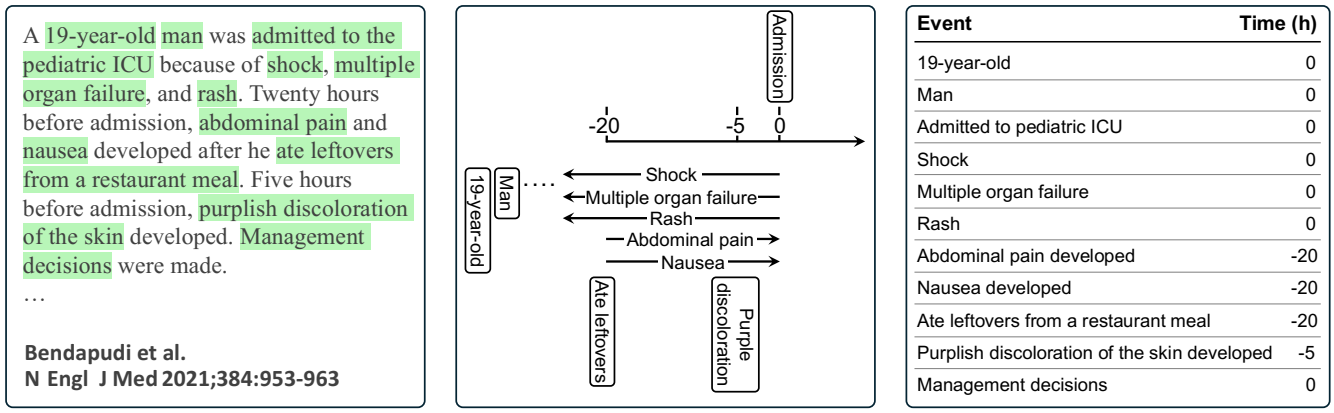
Healthcare disparities persist globally, with preventable deaths from conditions like sepsis disproportionately affecting underserved populations who often present to under-resourced facilities with limited access to specialized expertise. In these settings, much of the critical diagnostic and prognostic information exists only in unstructured clinical narratives—case reports, discharge summaries, and progress notes—as comprehensive structured data infrastructure may be unavailable (Anzalone et al. 2025; Seinen et al. 2025). While machine learning approaches have shown promise on structured data, with recent work demonstrating that incorporating text alongside structured inputs significantly improves predictive performance (Kline et al. 2022), large language models struggle with clinical risk estimation, especially in patient-facing scenarios (Wong et al. 2025). This gap highlights the need for automated risk forecasting from textual clinical records that could democratize access to expert-level clinical reasoning, particularly in resource-constrained environments where timely specialist consultation is not available. The challenge of LLM risk estimation motivates a structured treatment of risk forecasting and survival modeling from unstructured narrative sources.

Among textual sources, retrospective case reports serve as a rich training ground for developing temporally-aware AI systems that can eventually be deployed on real-time clinical notes. Case reports provide holistic accounts of clinical trajectories with explicit temporal reasoning—exactly the type of structured thinking needed for real-time clinical decision support. However, their narrative structure interleaves temporally unordered information, making them difficult to apply in forecasting tasks. Without identifying the time of occurrence for each event, models are prone to causal leakage—using information not available at the time of prediction. Solving this temporal reasoning challenge in case reports establishes the foundation for deployment in live clinical environments where such capabilities could improve patient outcomes.

One might ask whether large pretrained language models can already perform forecasting from clinical narratives, given their exposure to biomedical texts (Peng, Chen, and Lu 2020; Weber et al. 2024). However, these models have architectural limitations for temporal reasoning tasks. Encoder-based models apply random masking that rarely yields temporally coherent representations, while decoder-style models mask text in linear sequence, modeling events in text order rather than time order. Our experiments make these limitations explicit. For instance, on the timeline in Figure 1, the chance of recovering a valid time-ordered mask is below 0.002%, and decoder models show low temporal concordance (e.g., $c = 0.23$). Without a task-specific restructuring, such factors lead to suboptimal forecasting performance.

To address these temporal reasoning challenges, we develop a forecasting framework that transforms free-text narratives into textual time series—sequences of (event, time) tuples—and makes explicit the temporal structure of patient trajectories. Building on recent annotation work (Wang and Weiss 2025; Noroozizadeh et al. 2025), we propose a hybrid pipeline combining rule-based heuristics and LLM-assisted extraction to recover temporally localized clinical findings,

*These authors contributed equally.



Event	Time (h)
19-year-old	0
Man	0
Admitted to pediatric ICU	0
Shock	0
Multiple organ failure	0
Rash	0
Abdominal pain developed	-20
Nausea developed	-20
Ate leftovers from a restaurant meal	-20
Purplish discoloration of the skin developed	-5
Management decisions	0

Figure 1: Example case report (left) with timeline representations (middle) and text-ordered, (event, time) tuples (right).

validated on real clinical notes with expert verification.

Our work builds on extensive research in clinical NLP and temporal reasoning (Kohane 1987; Leeuwenberg and Moens 2020; Cheng and Weiss 2023), but shifts the emphasis from extraction to downstream clinical forecasting. While prior efforts have focused on entity recognition and temporal relation extraction (Zhou and Hripcsak 2007; Uzuner et al. 2011; Sun, Rumshisky, and Uzuner 2013), few have systematically assessed how such representations influence downstream tasks such as risk prediction and survival modeling. We bridge this gap by formalizing textual time series representations and evaluating their effectiveness for temporally grounded clinical prediction.

We empirically assess this framework across multiple tasks and modeling strategies, benchmarking encoder and decoder models—both fine-tuned and prompted versions—on three forecasting tasks: event prediction, survival analysis, and temporal ordering of future clinical events. Our results show no single model dominates across all tasks, highlighting the importance of aligning model architecture with forecasting objectives. We further examine how input ordering (time vs. narrative) impacts generalization and analyze the sensitivity of models to missing historical context through systematic masking experiments.

In summary, this paper makes the following contributions:

(i) **Annotation and Extraction Pipeline:** Building on prior methods for extracting information from the PubMed Open Access corpus (Noroozizadeh and Weiss 2025), this work introduces a pipeline that transforms case reports into textual time series via regular expression filtering and LLM-assisted extraction, yielding temporally anchored (event, time) tuples while minimizing causal leakage by restricting input to past events; (ii) **Comprehensive Model Comparison:** We conduct a comprehensive evaluation of encoder- and instruction-tuned decoder-based models—using both fine-tuned MLP heads and prompting—across event prediction, temporal ordering, and survival analysis tasks. Results show no single model excels universally, underscoring the need to align model choice with the specific forecasting objective; (iii) **Temporal- versus Text- Ordering:** We investigate the impact of annotation order by comparing training on time-ordered versus text-ordered data. Results show that preserving the narrative's natural order can improve generalization to external datasets, while time-based ordering can enhance ranking performance; (iv) **Sensitivity Analysis of Temporal Masking:** We conduct systematic dropout experiments by randomly masking parts of the clinical history to assess their impact on forecasting and event ordering. While higher masking levels reduce classification performance (F1), the concordance index remains largely stable, highlighting differing sensitivities of binary prediction and ranking tasks to historical context in textual time series; (v) **Methodological Framework for Temporal Clinical Forecasting:** Our approach demonstrates how narrative clinical texts can be systematically converted into structured temporal representations for forecasting tasks, providing a replicable methodology that could be adapted to other clinical and non-clinical text sources.

Related Work

Temporal Information Extraction from Clinical Text: Extracting timelines from clinical narratives is a challenging biomedical NLP task. The i2b2 2012 challenge introduced annotated datasets for temporal relation extraction from discharge summaries (Uzuner et al. 2011). Subsequent methods linked clinical events to timestamps or temporal expressions (Leeuwenberg and Moens 2020; Frattallone-Llado et al. 2024), typically assuming pre-defined event spans. We adopt our prior approach (Noroozizadeh and Weiss 2025; Noroozizadeh et al. 2025) to assign timepoints to findings from full-length case reports, enabling finer temporal resolution. Unlike methods using structured EHR data, we focus solely on text—crucial for sources like PubMed that lack structured metadata. By directly supervising event-time alignment, we overcome limitations of span-based annotations (Rosenbloom et al. 2011). This aligns with growing emphasis on temporality in sepsis and critical care phenotyping (Johnson et al. 2018; Kamran et al. 2024), particularly given high missingness rates in structured data (Johnson et al. 2023; Seinen et al. 2025).

Predictive Modeling with Clinical Text: Clinical text has been used for outcome prediction tasks like mortality and readmission, with models like *ClinicalBERT* achieving high levels of performance on EHR notes (Huang, Altosaar, and Ranganath 2019; Gu et al. 2021). Text captures complementary information—symptoms and social factors—not found in structured codes. However, traditional approaches using full-text or bag-of-words representations obscure temporal dynamics. Our approach preserves event order within narratives, treating data as time series. This aligns with recent efforts adapting LLMs for time series tasks, such as Time-LLM mapping numeric sequences to tokens (Jin et al. 2023). To our knowledge, this is the first work to forecast from LLM-derived time series in clinical domains. While prior sepsis prediction models rely on vitals or scores like SOFA and SAPS (Hou et al. 2020; Noroozizadeh, Weiss, and Chen 2023), our method is complementary—leveraging narrative descriptions that include clinician interpretations and context. This is valuable when case reports or patient histories are available but structured data is sparse or unavailable.

Large Language Models in Healthcare: Recent LLMs such as `GPT-3` and `GPT-4` have enabled applications from medical QA to note summarization. These models encode substantial medical knowledge and reasoning capacity—`GPT-4` demonstrates strong performance on board exams and can support patient record abstraction for quality reporting (Boussina et al. 2024). However, their use in clinical forecasting remains largely unexplored. Our benchmark tests LLMs via zero- and few-shot prompting to assess clinical event forecasting from patient narratives, comparing results to fine-tuned models to quantify the gap between general-purpose and task-specific modeling. This contributes evidence that while LLMs hold broad medical knowledge, fine-tuning and structured inputs are often required for clinical reliability (Huang, Altosaar, and Ranganath 2019; Gu et al. 2021). Some studies suggest domain-specific tuning does not always improve over foundation models (Jeong et al. 2024), reinforcing the need for task-specific evaluations like ours.

Methods[1]

We next describe our dataset construction and annotation pipeline, define the forecasting and survival prediction tasks, detail the modeling approaches evaluated, and present the sensitivity analyses used to assess robustness (Figure 2).

Dataset Extraction

We used the PubMed Open Access (PMOA) Subset (as of April 25, 2024) for our analysis. Following prior work (Wang and Weiss 2025), we extracted the body of each case by selecting the text between the `==== Body` and `==== Ref` delimiters in the PMOA corpus. To maintain relevance, we included only documents with case-insensitive matches

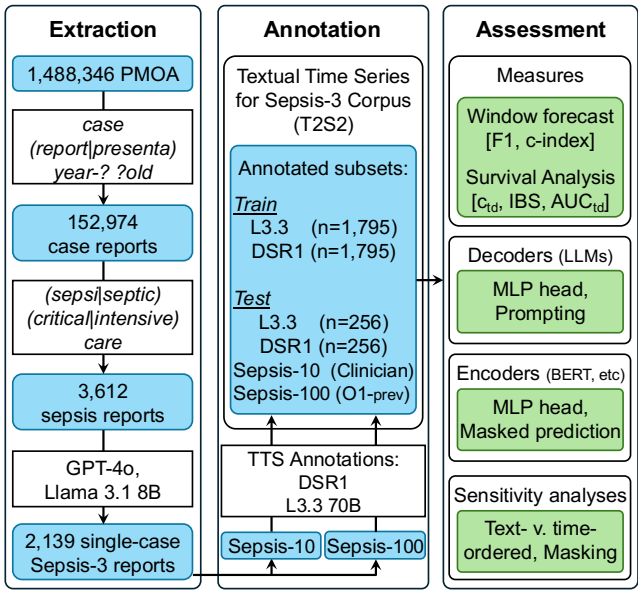


Figure 2: *T2S2* forecasting analysis pipeline

to `(case report|case presenta)` and `year-? ?old`, an approach shown to be more specific than relying on PubMed metadata (Noroozizadeh and Weiss 2025).

Additional filtering was applied to identify potential sepsis-related case reports using the regular expressions `sepsi|septic` and `critical|intensive` care. We then used LLM-based queries to extract key attributes, including sepsis diagnosis, patient count, age, and gender. For this step, both `GPT-4o` and `Llama-3.1-8B-Instruct` were employed. Reports describing more than one case were excluded. A case report was retained if either model classified it as involving a Sepsis-3 diagnosis.

This multi-step process yielded a total of 2319 Sepsis-3 case reports. From this set, we randomly selected two subsets for evaluation: a group of 10 reports (*sepsis-10*), which underwent expert clinical annotation to serve as a gold standard, and a larger group of 100 reports (*sepsis-100*) as a bronze standard for broader testing based on an alternative annotator (`O1-preview`; Figure 2).

Textual Time Series Annotation

Following Noroozizadeh and Weiss (2025), a textual time series case report corpus of Sepsis-3 patients was constructed from the PMOA sepsis reports. A textual time series refers to a sequence of clinical findings, each paired with a timestamp (either absolute or relative to the case presentation time), corresponding to an individual patient. Here, a clinical finding is defined as a free-text expression describing an entity that pertains to or may impact a person's health.

A textual time series differs from conventional approaches to clinical concept annotation (e.g., (Sun, Rumshisky, and Uzuner 2013; Uzuner et al. 2011)) in its extension of the text span to better capture the specificity and contextual meaning of each clinical finding. In our work, the interpretation of a clinical finding differs from the i2b2 con-

[1]Our code can be found at: https://github.com/Shahriarnz14/Textual-Time-Series-Forecasting.

The extended version of this paper, including the full technical appendix, is available in Noroozizadeh, Kumar, and Weiss (2025).

cept guidelines in the following respects: (1) Clinical findings are not restricted to single prepositional phrases following a markable entity. For instance, instead of splitting "pain in chest that radiates substernally" into less informative parts, we retain the full phrase as a single, complete finding to preserve its meaning and context. (2) To improve clarity, compound phrases are broken into individual findings. For example, "metastases in the liver and pancreas" becomes "metastasis in the liver" and "metastasis in the pancreas."

We used `DeepSeek-R1-UD-IQ1` and `Llama-3.3-70B-Instruct` models to generate clinical textual time series from the 2319 sepsis case reports. We refer to this as **T**extual **T**ime-**S**eries for **S**epsis (*T2S2*). Details on the quality evaluation of LLM-extracted temporal annotations and the prompts used for extraction are provided in Appendix A and B, respectively. Our dataset was split into training (n = 1,795; 80% train, 20% validation), testing *T2S2*-test (n = 244), and two external validation sets: *sepsis-10* (n = 10) and *sepsis-100* (n = 90).

Forecasting Tasks

Using the timelines extracted from the annotations, we defined two primary forecasting tasks and a survival analysis task as follows:

Event Occurrence Prediction. Given a prefix of the clinical timeline (all events up to a certain time t), the model is tasked with predicting whether each of the immediate next k events occurs within a specified time horizon. This setup is repeated across multiple time cut-offs to simulate an "online" forecasting scenario, where the model must output a binary label for each of the next k events: does this event occur within h hours after t? Time horizons used include 1 hour, 24 hours (1 day), and 168 hours (1 week). The task is framed as a series of binary classification problems, with one binary decision per event. Evaluation is based on precision, recall, and F1 score, averaged across event positions for each time horizon.

Temporal Ordering Prediction. This task assesses the model's ability to reconstruct the correct sequence of future events. At each time cut-off t, we extract the next k events from the timeline and remove their timestamps. The model must output a permutation of these events that matches their true chronological order. This is framed as a ranking task, evaluated by computing the pairwise concordance between the predicted and true orderings (e.g., proportion of correctly ordered pairs). This tests whether the model can infer temporal progression from unordered event content alone.

Survival Analysis for Mortality Time. We include a classical survival analysis task to model time until death. Many case reports specify whether the patient died and, if so, when (e.g., "the patient died on hospital day 10"), which enables us to define a time-to-event outcome. For this task, we evaluate models at predefined cut-off times— specifically at 0 hours (admission), 24 hours (1 day), and 168 hours (1 week)— and use the extracted event history up to each cut-off as input. A survival model is trained to predict the probability of survival over time beyond each cut-off. We evaluate using the time-dependent concordance to measure alignment of the predicted survival times with actual outcomes.

Modeling Approaches

To evaluate performance on the event forecasting and survival prediction tasks, we implement five modeling paradigms: (i) decoder-only large language models (LLMs) with fine-tuned heads, (ii) prompted LLMs without gradient updates (zero- or few-shot), (iii) encoder-only models with task-specific fine-tuned heads, (iv) encoder-masking models with fine-tuning, and (v) encoder-masking models in zero-shot settings. In Appendix D and E, we provide technical details of encoder-based forecasting methods and the survival modeling framework respectively.

(i) Fine-tuned LLMs. We apply instruction-tuned decoder-only models from the *Llama* and *DeepSeek* families: *Llama-3.3-70B-Instruct*, *Llama-3.1-8B-Instruct*, and their distilled variants (Grattafiori et al. 2024; Liu et al. 2024a). We also include open-source models with documented training corpora (*OLMO-32B-Instruct* (OLMo et al. 2024), *RedPajama-INCITE-7B-Instruct* (Weber et al. 2024)) to address potential data leakage concerns—notably, RedPajama explicitly excludes PubMed from its training data. Additionally, we evaluate a medically fine-tuned decoder model (*MediPhi-PubMed* (Corbeil et al. 2025)) to assess domain-specific benefits.

Each model is paired with a lightweight multilayer perceptron (MLP) head trained for classification or ranking, depending on the task. The input consists of a text-formatted prefix of events (e.g., a clinical timeline up to time t), optionally accompanied by an instruction template. The output layer produces task-specific predictions: binary labels for event occurrence or a permutation over k events for ordering. Training is conducted using cross-entropy loss for classification and pairwise ranking loss for ordering.

(ii) Prompted LLMs. In the zero- or few-shot setting, we use the same LLM architectures as above, but without any fine-tuning. Instead, we supply structured prompts at inference time to guide the model toward task-specific outputs. Each prompt includes: (1) a system instruction establishing the model's role (e.g., "You are an expert physician."), (2) a user instruction describing the prediction task, and (3) one or more input-output examples in a few-shot format. Prompt designs are customized for each task and are detailed in Appendix C. Model outputs are parsed to extract binary labels or ordered lists, and errors in generation or ambiguity are excluded from evaluation.

(iii) Fine-tuned Encoder-Only Models. We also evaluate a range of encoder-based models trained end-to-end on each task. Architectures include general-purpose models (*BERT-base-uncased* (Devlin et al. 2019), *RoBERTa-base* (Liu et al. 2019), *DeBERTa-v3-small* (He, Gao, and Chen 2021), *ModernBERT-base/large* (Warner et al. 2024)) and biomedically-pretrained variants (*BioClinical-ModernBERT-base/large* (Sounack et al. 2025)). For each model, we append a task-specific MLP head for event occurrence or ordering prediction. The model input is the same

flattened prefix of the event sequence used in other settings, tokenized and formatted according to the respective architecture. These models are trained using standard supervised learning objectives and evaluated on the same metrics as other methods: F1 score for event occurrence, and pairwise concordance for temporal ordering. For survival analysis, the final hidden states are passed into a time-to-event prediction head and evaluated using the time-dependent concordance index (Antolini, Boracchi, and Biganzoli 2005).

(iv, v) Encoder-Masking Models. We adapt masked language modeling (MLM) for our temporal reasoning tasks. For event occurrence, models predict masked tokens ("`yes`"/"`no`") in prompts such as "Will [event] happen within 24 hours? `[MASK]`". For temporal ordering, models predict "`before`"/"`after`" tokens in prompts comparing event pairs. We evaluate both fine-tuned variants (trained on task-specific objectives) and zero-shot variants (using pretrained MLM capabilities without additional training). This approach leverages the bidirectional context of encoder models while maintaining interpretable predictions through constrained vocabulary.

Sensitivity Analyses

We conduct sensitivity analyses to examine how the temporal structure and completeness of input sequences affect forecasting performance. These analyses probe the robustness of models to changes in event ordering and to varying levels of missing historical information.

Time-ordering Strategies. We evaluate two strategies for ordering clinical events within textual time series inputs. In the *text-ordered* setting, events are presented in the order they appear in the original case report narratives, preserving the narrative structure produced during extraction. In the *time-ordered* setting, events are sorted chronologically by their documented occurrence time, enforcing strict temporal alignment. This comparison isolates the effect of narrative sequencing versus explicit temporal ordering on model performance. For text-ordered inputs, to prevent causal leakage, events occurring after the forecast cutoff time t are masked for encoder-based models and omitted for decoder models.

Timestep Dropout. To assess robustness to incomplete patient histories, we introduce a *timestep dropout rate* (TDR), defined as the proportion of input timesteps randomly removed prior to inference. This procedure simulates varying degrees of missing clinical documentation that may arise in deployment settings. We vary TDR from 0% (full history) to 90% (severely truncated history) in increments, masking events independently and uniformly at random. This setup enables controlled evaluation of model performance under partial information, probing the extent to which predictive accuracy and event ordering depend on the completeness of historical context.

Results

In this section, we present results for our three main evaluation settings: event forecast within a subsequent time window, temporal ordering of forecasted events, and survival analysis. We provide sensitivity analyses on the forecasting tasks, examining the effects of temporal ordering and historical context availability in Appendix H and I.

Forecasting Tasks

Event forecast within next 24 hours: F1 performance. Our results show that encoder-based models outperform decoder-based LLMs in event forecasting across both `DeepSeek-R1` and `Llama-3.3-70B` annotations (Table 1). Forecasting performance across all models is shown in Tables A3 and A4 (Appendix F). Encoder models, especially those with a fine-tuned MLP head, achieve substantially higher F1 scores than decoder LLMs, reinforcing the effectiveness of encoder-based representations for forecasting.

Among decoders, we observe varied performance patterns. *RedPajama-INCITE-7B-Instruct*, whose training excludes PubMed, achieves 0.352 F1 at 24h—better than some decoder models like *MediPhi-PubMed* (0.268) and within the typical decoder performance range, though below top performers like *DeepSeek-Llama-3.3-70B* (0.482). *OLMO-32B-Instruct* performs competitively (0.411), suggesting that open models without biomedical pretraining can still achieve strong temporal forecasting results.

Among encoder models, both general-purpose and biomedically-pretrained variants show strong performance. While *ModernBERT-base* achieves slightly higher F1 scores on internal test sets (0.576 vs 0.559), *BioClinical-ModernBERT-base* demonstrates superior concordance (0.677 vs 0.646 for *DeepSeek-R1*) and notably better generalization to external datasets, with F1 scores of 0.635 and 0.662 on *sepsis-100* compared to 0.607 and 0.626 for the general-purpose variant. This suggests that while both *ModernBERT* architectures are effective, biomedical pretraining particularly enhances temporal reasoning and cross-dataset generalization.

Different training strategies yield distinct performance patterns. The MLP head fine-tuning approach consistently outperforms both fine-tuned masking and zero-shot masking models, particularly in long-horizon forecasting. While *ModernBERT-base* achieves strong F1 scores, *BioClinical-ModernBERT-base* demonstrates the best overall performance when considering both accuracy and generalization. Zero-shot masking models, particularly standard *BERT* and *RoBERTa*, fail to make meaningful predictions at the 1-hour mark. However, *ModernBERT* variants exhibit relatively better zero-shot performance, with *BioClinical-ModernBERT-base* achieving notably higher scores in zero-shot settings (0.246 F1 at 1h), indicating that biomedical pretraining improves out-of-the-box temporal reasoning.

Performance trends remain consistent across forecasting windows. F1 scores improve as the window increases, with 1-hour predictions being most challenging and 168-hour predictions yielding highest scores. Models perform worse on external validation sets than internal *T2S2* test sets, though *BioClinical-ModernBERT* models show the smallest performance degradation, maintaining their advantage in cross-dataset generalization.

	T2S2 (DeepSeek-R1)								*T2S2* (Llama-3.3-70B)							
	F1(1h)	F1(1d)	F1(1w)	c-index	F1(1d)-10	c10	F1(1d)-100	c100	F1(1h)	F1(1d)	F1(1w)	c-index	F1(1d)-10	c10	F1(1d)-100	c100
LLM-MLP head																
DS-L3.3 70B	**0.075**	0.482	**0.796**	**0.632**	**0.397**	**0.578**	**0.613**	0.595	0.140	0.563	0.811	**0.624**	**0.397**	**0.585**	**0.629**	**0.598**
OLMO 32B Instruct	0.000	0.411	0.688	0.621	0.273	0.561	0.432	0.593	0.190	0.315	0.764	0.604	0.238	0.551	0.315	0.596
RedPajama-INCITE 7B Instruct	0.000	0.352	0.651	0.618	0.282	0.572	0.471	0.594								
MediPhi-PubMed	0.000	0.268	0.751	0.626	0.169	0.561	0.210	0.597	0.000	0.507	0.801	0.623	0.277	0.564	0.534	0.600
Prompting																
L3.3 70B few-shot	**0.095**	**0.380**	0.652	–	**0.360**	–	0.452	–	0.064	**0.460**	0.729	–	0.313	–	0.480	–
OLMO 32B few-shot ordinal	0.013	0.315	0.651	–	0.247	–	0.422	–	0.046	0.413	0.647	–	0.300	–	0.424	–
Encoder-MLP head																
ModernBERT-base	0.306	**0.576**	0.877	0.646	0.428	0.562	0.607	0.558	0.257	0.645	**0.903**	0.594	0.431	0.590	0.626	0.565
Bioclinical ModernBERT-base	0.264	0.559	**0.879**	**0.677**	0.395	**0.598**	**0.635**	**0.610**	0.290	**0.653**	0.902	**0.650**	**0.449**	**0.618**	**0.662**	**0.627**
Encoder-masking-fine-tuned																
ModernBERT-base	**0.186**	0.449	**0.697**	0.672	0.325	0.576	**0.503**	0.613	**0.139**	**0.489**	0.700	0.632	**0.334**	0.595	**0.478**	0.612
Bioclinical ModernBERT-base	0.166	0.450	0.692	**0.676**	0.336	0.604	0.499	**0.655**	0.129	0.481	0.733	0.653	0.331	**0.615**	0.475	**0.658**
Encoder-masking-zeroshot																
ModernBERT-base	0.169	**0.468**	**0.804**	0.496	**0.358**	0.501	**0.511**	0.499	0.132	**0.525**	**0.840**	0.556	**0.352**	0.503	0.511	0.501
Bioclinical ModernBERT-base	**0.246**	0.458	0.614	**0.498**	0.301	0.509	0.457	0.498	**0.197**	0.460	0.533	0.499	0.299	0.520	0.465	0.499

Table 1: Forecasting performance (event occurrence: F1 and correct event ordering: concordance-index) of the ensuing $k = 8$ events. All models are trained/fine-tuned on time-ordered annotations from either `DeepSeek-R1` or `Llama 3.3-70B`. **Bold** values indicate best in category within each column group (refer to Tables A3 and A4 in Appendix F for detailed performance statistics on all models). Performance statistics for all models are presented in A3 and A4, respectively. Abbreviations: F1(1d)-10/100: F1 (1 day) for sepsis-10 and sepsis-100 respectively. c10/100: c-index for sepsis-10 and sepsis-100 respectively.

Temporal ordering of forecasted events: concordance. Encoder models consistently outperform decoders in correctly ranking the order of upcoming events, as measured by concordance (c-index; Table 1; complete results in Tables A3 and A4 for `DeepSeek-R1` and `Llama-3.3-70B` annotations, respectively, with full details in Appendix F. Decoder models with potential biomedical exposure achieve modest concordance (0.618-0.632), while *RedPajama-INCITE-7B-Instruct* without PubMed data in its training corpus performs similarly, suggesting domain-specific knowledge is less critical here than for event prediction. The comparable performance of *RedPajama-INCITE-7B-Instruct* to other general models (with access to PubMed in their pretraining) also provides evidence that the *T2S2* prediction task is distinct from raw PubMed text and assuages causal leakage concerns.

Biomedical pretraining in encoders demonstrates clear advantages, with *BioClinical-ModernBERT-base* achieving the highest concordance (0.677 with an MLP head, 0.676 with fine-tuned masking) and strong performance across datasets. MLP-head training yields superior F1 scores, while fine-tuned masking excels in concordance. *BioClinical-ModernBERT-base* also generalizes well to external datasets (c-index $>$0.60), suggesting that biomedical pretraining provides stability in temporal reasoning under dataset shifts.

Survival analysis: model performance across timepoints and cohorts. Tables 2 and A5 (Appendix G) show time-dependent concordance results where instruction-tuned decoder models clearly outperform encoder-based models. This pattern contrasts with the forecasting tasks (Table 1), where encoders excelled. Across both the *T2S2* and *sepsis-100* cohorts, decoder models consistently achieve the strongest results. Notably, *RedPajama-INCITE-7B-Instruct* attains the highest concordance overall (0.76 at 168h, *sepsis-100*) despite lacking any biomedical text in pretraining, while larger models such as *DeepSeek-R1-Llama-70B* and *OLMO-32B-Instruct* also perform robustly. Encoder models—including biomedically fine-tuned variants like *BioClinical-ModernBERT-base*, which is competitive on *sepsis-100* but not on *T2S2*—are generally surpassed by decoder models.

Model	0h	24h	168h
bert-base-uncased	0.60	0.61	0.53
roberta-base	0.55	0.60	0.60
deberta-v3-small	0.56	0.57	0.57
ModernBERT-base	0.52	0.58	0.58
ModernBERT-large	0.53	0.59	0.58
Bioclinical ModernBERT-base	0.57	0.53	0.53
Bioclinical ModernBERT-large	0.54	0.55	0.54
DeepSeek-R1-Llama-70B	**0.64**	**0.63**	0.59
Llama-3.3-70B-Instruct	0.62	**0.63**	0.58
DeepSeek-R1-Llama-8B	0.60	0.58	0.58
Llama-3.1-8B-Instruct	0.62	0.61	**0.61**
OLMO 32B Instruct	0.60	0.62	0.59
RedPajama-INCITE 7B Instruct	0.56	0.60	**0.61**
MediPhi-PubMed	0.54	0.62	0.55

Table 2: Time-dependent concordance index for survival analysis on the `DeepSeek-R1` *T2S2* annotations evaluated at 0h, 24h, and 168h. Extended *sepsis-10* and *sepsis-100* results are provided in Appendix G.

Predictive performance does not universally improve with more observation time. Models evaluated on the *T2S2* cohort, often peak at earlier windows (0h or 24h) rather than at 168h. Meanwhile, the *sepsis-10* cohort experiences clear ceiling effects, with many models achieving perfect concordance (1.00). These results suggest that, unlike the forecasting tasks, survival prediction in our dataset benefits more from the capabilities inherent in large decoder models than from domain-specific pretrained encoders.

Discussion and Conclusion

Our findings highlight the superiority of encoder-based models over decoder-based LLMs for event forecasting, em-

phasizing the limitations of autoregressive models in structured prediction tasks. Among encoder models, *BioClinical-ModernBERT-base* consistently performs best, achieving superior concordance (0.677 vs 0.646) and better external generalization despite slightly lower internal F1 scores than general-purpose *ModernBERT-base*. MLP-head fine-tuning excels in F1 scores while fine-tuned masking models achieve higher concordance, particularly in long-term forecasting. This suggests that classification and ranking tasks benefit from distinct optimization strategies.

The consistent advantage of biomedically-pretrained models across tasks deserves special attention. While general-purpose models achieve strong internal performance, biomedical pretraining provides crucial benefits for real-world deployment: *BioClinical-ModernBERT* variants show the smallest performance degradation on external datasets and maintain higher zero-shot capabilities (0.246 F1 at 1h vs near-zero for standard *BERT*/*RoBERTa*). This suggests that domain-specific pretraining enhances not just accuracy but also robustness and generalization—critical factors when considering clinical deployment where distribution shifts are common.

The gradual improvement of F1 scores over longer time horizons indicates that event patterns become more predictable over time, whereas short-term forecasting remains challenging due to higher variability. Performance drops on external validation datasets highlight generalization challenges, though *ModernBERT* models exhibit relative robustness. For survival analysis, instruction-tuned LLMs like *Llama-3.3-70B-Instruct* outperform traditional transformer baselines, with several models achieving peak concordance at early timepoints rather than with extended observation.

Our sensitivity analysis reveals several trade-offs: time-ordered training generally improves concordance while text-ordering can yield better F1 scores on externally-annotated datasets. Robustness experiments show F1 scores degrade significantly beyond 60% timestep dropout, while concordance remains stable, indicating event ranking is less sensitive to partial history than event classification.

Additional insights into the forecasting results are presented in Appendix J. Specifically, we summarize two observations that clarify model behavior and apparent anomalies: (i) why decoders fare better on survival but not short-horizon forecasting, and (ii) interpreting very low F_1 scores and external validation drops.

Methodological Contributions and Potential Impact. Our framework demonstrates how narrative clinical texts can be systematically converted into structured temporal representations for forecasting tasks. The ability to extract temporally structured insights from unstructured clinical text could potentially support clinical decision-making, particularly in settings with limited structured data or specialized expertise.

Importantly, while consumers frequently consult LLMs about health risk via prompting in chats, this study demonstrates that the prompting approach performs substantially worse in risk prediction than alternative approaches, at least with respect to precision/recall/F1, with prompted LLMs achieving at best 0.460 F1 at 24h compared to 0.653 for fine-tuned encoders. Our study highlights several alternative approaches and characterizes their relative performance strengths and weaknesses. In addition, our finding that concordance remains stable with 60% missing context suggests robustness to incomplete documentation scenarios, which highlights the degradation pattern in performance as it relates to context availability.

Our approach also demonstrates that our language model systems can reliably capture the temporal reasoning that clinicians use in practice based on their clinical documentation. However, additional work would be required to validate performance on real-time clinical data and demonstrate measurable impact on patient outcomes in deployment. Further discussion on societal impact of our work is in Appendix M.

Limitations and Future Directions. Our study has important limitations to consider. First, our pipeline relies on case reports from the PubMed Open Access (PMOA) corpus. Because these reports often highlight rare or atypical presentations and differ from routine clinical notes (e.g., progress notes, discharge summaries), generalizability to real-world health-system text may be limited. However, case reports provide rich, temporally structured narratives with explicit clinical rationale, making them well suited for evaluating model performance under sparse documentation. Accordingly, we present a general framework for textual time series forecasting that uses case reports as an interpretable, temporally rich foundation for method development, with extensions to real-world corpora such as MIMIC discharge summaries. See Appendix L for a detailed discussion on the generalizability of our pipeline beyond case report details, as well as comparisons with other clinical sources.

Second, while we focused on sepsis due to its clinical relevance and prevalence in case reports, our framework is fundamentally disease-agnostic. Preliminary results from over 125K PMOA case reports across various diagnostic conditions show promising generalization, with future work planned on broader diagnostic categories and real-world corpora such as MIMIC-IV discharge summaries.

Finally, despite mitigation steps (timestamp extraction, temporal masking, external test sets), pretrained language models may still have been exposed to PMOA content. Although they were not trained on our derived *(event, time)* sequences or temporally reordered data, the original narratives could appear in pretraining. Our pipeline applies temporal reordering and masking of future events beyond the prediction cutoff, enforcing a causal structure absent from the originals and yielding a distinct setup that requires reasoning over partial, temporally grounded inputs. While we included *RedPajama-INCITE-7B-Instruct*, an open-source model with a documented PubMed-free corpus for pretraining, and saw its performance (0.352 F1 at 24h; concordance 0.618) was comparable to decoders potentially exposed to PubMed, other strategies supporting the argument that the *T2S2* task is sufficiently distinct from raw PMOA text and mitigates causal leakage concerns could be examined. See Appendix K for a detailed discussion of how open-source models mitigate potential data leakage, as well as additional forecasting results on 115 case reports published after the pretraining cutoffs of encoder-based models (post-2020).

Acknowledgments

This research was supported in part by the Division of Intramural Research (DIR) of the National Library of Medicine (NLM), National Institutes of Health. This work utilized the computational resources of the NIH HPC Biowulf cluster. S.N. was supported by Carnegie Mellon University TCS Presidential Fellowship, and Natural Sciences and Engineering Research Council of Canada (NSERC) PGS-D award. S.N. was also supported in part by an appointment to the National Library of Medicine Research Participation Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and the National Library of Medicine, National Institutes of Health. ORISE is managed by ORAU under DOE contract number DE-SC0014664. All opinions expressed in this paper are the authors' and do not necessarily reflect the policies and views of NIH, NLM, DOE, or ORAU/ORISE.

References

Antolini, L.; Boracchi, P.; and Biganzoli, E. 2005. A time-dependent discrimination index for survival data. *Statistics in medicine*, 24(24): 3927–3944.

Anzalone, A. J.; Geary, C. R.; Dai, R.; Watanabe-Galloway, S.; McClay, J. C.; and Campbell, J. R. 2025. Lower electronic health record adoption and interoperability in rural versus urban physician participants: a cross-sectional analysis from the CMS quality payment program. *BMC Health Services Research*, 25(1): 128.

Bai, S.; Kolter, J. Z.; and Koltun, V. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Boussina, A.; Krishnamoorthy, R.; Quintero, K.; Joshi, S.; Wardi, G.; Pour, H.; Hilbert, N.; Malhotra, A.; Hogarth, M.; Sitapati, A. M.; et al. 2024. Large language models for more efficient reporting of hospital quality measures. *NEJM AI*, 1(11): AIcs2400420.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Cheng, C.; and Weiss, J. C. 2023. Typed markers and context for clinical temporal relation extraction. In *Machine Learning for Healthcare Conference*, 94–109. PMLR.

Corbeil, J.-P.; Dada, A.; Attendu, J.-M.; Abacha, A. B.; Sordoni, A.; Caccia, L.; Beaulieu, F.; Lin, T.; Kleesiek, J.; and Vozila, P. 2025. A Modular Approach for Clinical SLMs Driven by Synthetic Data with Pre-Instruction Tuning, Model Merging, and Clinical-Tasks Alignment. *arXiv preprint arXiv:2505.10717*.

Dai, N.; Liang, J.; Qiu, X.; and Huang, X.-J. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 5997–6007.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186.

Frattallone-Llado, G.; Kim, J.; Cheng, C.; Salazar, D.; Edakalavan, S.; and Weiss, J. C. 2024. Using multimodal data to improve precision of inpatient event timelines. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 322–334.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; and Poon, H. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1): 1–23.

He, P.; Gao, J.; and Chen, W. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Hou, N.; Li, M.; He, L.; Xie, B.; Wang, L.; Zhang, R.; Yu, Y.; Sun, X.; Pan, Z.; and Wang, K. 2020. Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost. *Journal of translational medicine*, 18: 1–14.

Huang, K.; Altosaar, J.; and Ranganath, R. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Jeong, D. P.; Garg, S.; Lipton, Z. C.; and Oberst, M. 2024. Medical Adaptation of Large Language and Vision-Language Models: Are We Making Progress? *arXiv preprint arXiv:2411.04118*.

Jin, M.; Wang, S.; Ma, L.; Chu, Z.; Zhang, J. Y.; Shi, X.; Chen, P.-Y.; Liang, Y.; Li, Y.-F.; Pan, S.; et al. 2023. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*.

Johnson, A.; Pollard, T.; Horng, S.; Celi, L. A.; and Mark, R. 2023. MIMIC-IV-Note: Deidentified free-text clinical notes (version 2.2).

Johnson, A. E.; Aboab, J.; Raffa, J. D.; Pollard, T. J.; Deliberato, R. O.; Celi, L. A.; and Stone, D. J. 2018. A comparative analysis of sepsis identification methods in an electronic database. *Critical care medicine*, 46(4): 494–499.

Kamran, F.; Tjandra, D.; Heiler, A.; Virzi, J.; Singh, K.; King, J. E.; Valley, T. S.; and Wiens, J. 2024. Evaluation of sepsis prediction models before onset of treatment. *NEJM AI*, 1(3).

Kline, A.; Wang, H.; Li, Y.; Dennis, S.; Hutch, M.; Xu, Z.; Wang, F.; Cheng, F.; and Luo, Y. 2022. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5(1): 171.

Kohane, I. S. 1987. *Temporal reasoning in medical expert systems*. Ph.D. thesis, Boston University.

Leeuwenberg, A.; and Moens, M.-F. 2020. Towards extracting absolute event timelines from english clinical reports. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 2710–2719.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; and Liang, P. 2024b. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12: 157–173.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Noroozizadeh, S.; Kumar, S.; Chen, G. H.; and Weiss, J. C. 2025. PMOA-TTS: Introducing the PubMed Open Access Textual Time Series Corpus. *arXiv preprint arXiv:2505.20323*.

Noroozizadeh, S.; Kumar, S.; and Weiss, J. C. 2025. Forecasting Clinical Risk from Textual Time Series: Structuring Narratives for Temporal AI in Healthcare. *arXiv preprint arXiv:2504.10340*.

Noroozizadeh, S.; and Weiss, J. C. 2025. Reconstructing Sepsis Trajectories from Clinical Case Reports using LLMs: the Textual Time Series Corpus for Sepsis. *arXiv preprint arXiv:2504.12326*.

Noroozizadeh, S.; Weiss, J. C.; and Chen, G. H. 2023. Temporal supervised contrastive learning for modeling patient risk progression. In *Machine Learning for Health (ML4H)*, 403–427. PMLR.

OLMo, T.; Walsh, P.; Soldaini, L.; Groeneveld, D.; Lo, K.; Arora, S.; Bhagia, A.; Gu, Y.; Huang, S.; Jordan, M.; et al. 2024. 2 OLMo 2 Furious. *arXiv preprint arXiv:2501.00656*.

Peng, Y.; Chen, Q.; and Lu, Z. 2020. An Empirical Study of Multi-Task Learning on BERT for Biomedical Text Mining. *BioNLP 2020*, 205.

Rosenbloom, S. T.; Denny, J. C.; Xu, H.; Lorenzi, N.; Stead, W. W.; and Johnson, K. B. 2011. Data from clinical notes: a perspective on the tension between structure and flexible documentation. *Journal of the American Medical Informatics Association*, 18(2): 181–186.

Seinen, T. M.; Kors, J. A.; van Mulligen, E. M.; and Rijnbeek, P. R. 2025. Using Structured Codes and Free-Text Notes to Measure Information Complementarity in Electronic Health Records: Feasibility and Validation Study. *Journal of Medical Internet Research*, 27: e66910.

Sounack, T.; Davis, J.; Durieux, B.; Chaffin, A.; Pollard, T. J.; Lehman, E.; Johnson, A. E.; McDermott, M.; Naumann, T.; and Lindvall, C. 2025. BioClinical ModernBERT: A State-of-the-Art Long-Context Encoder for Biomedical and Clinical NLP. *arXiv preprint arXiv:2506.10896*.

Sun, W.; Rumshisky, A.; and Uzuner, O. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5): 806–813.

Uzuner, Ö.; South, B. R.; Shen, S.; and DuVall, S. L. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5): 552–556.

Wang, J.; and Weiss, J. 2025. A Large-Language Model Framework for Relative Timeline Extraction from PubMed Case Reports. In *Proceedings of the AMIA Informatics Summit*. American Medical Informatics Association.

Warner, B.; Chaffin, A.; Clavié, B.; Weller, O.; Hallström, O.; Taghadouini, S.; Gallagher, A.; Biswas, R.; Ladhak, F.; Aarsen, T.; et al. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Weber, M.; Fu, D.; Anthony, Q.; Oren, Y.; Adams, S.; Alexandrov, A.; Lyu, X.; Nguyen, H.; Yao, X.; Adams, V.; et al. 2024. Redpajama: an open dataset for training large language models. *Advances in Neural Information Processing Systems*, 37: 116462–116492.

Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Wong, L.; Ali, A.; Xiong, R.; Shen, S. Z.; Kim, Y.; and Agrawal, M. 2025. Retrieval-augmented systems can be dangerous medical communicators. *CoRR*, abs/2502.14898.

Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; and Zhang, W. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 11106–11115.

Zhou, L.; and Hripcsak, G. 2007. Temporal reasoning with medical data—a review with emphasis on medical natural language processing. *Journal of Biomedical Informatics*, 40(2): 183–202.