Quantifying surprise in clinical care: Detecting highly informative events in electronic health records with foundation models

Michael C. Burkhart, Bashar Ramadan, Luke Solo, William F. Parker, and Brett K. Beaulieu-Jones

Department of Medicine, University of Chicago,
Chicago, Illinois, USA
E-mail: {burkh4rt,basharramadan,lsolo,wparker,beaulieujones}@uchicago.edu

We present a foundation model-derived method to identify highly informative tokens and events in electronic health records. Our approach considers incoming data for the entire context of a patient's hospitalization to find surprising events. Context enables flagging anomalous events that rule-based approaches would consider within a normal range. We demonstrate that the events our model flags are significantly more useful than average events for predicting downstream patient outcomes and show that a fraction of events we identify as unsurprising can be safely dropped without an adverse impact on performance. Finally, we show how informativeness can help interpret the predictions of prognostic models trained on foundation model-derived representations.

Keywords: Foundation models; Electronic health records; Information quantification; Model explainability.

1. Introduction

Healthcare generates a stream of data, including vitals, labs, medications, and respiratory support. Clinical decision making requires parsing and understanding this information and its importance in the context of each patient's medical history. Oftentimes, event summaries like automatically-collected vitals provide little additional knowledge about a patient.[2,58] Clinicians are commonly notified regardless, resulting in increased cognitive burden and alarm fatigue.[51] For over a decade now, the Joint Commission has included "reduc[ing] patient harm associated with clinical alarm systems" as a National Patient Safety Goal [52, NPSG.06.01.01, since 2014]. In this paper, we investigate the extent to which foundation model (FM)-derived estimates of event information can be used to highlight the most important events in a patient's record. In essence, we explore which events are surprising to the FM based on a comparison between what the model expects to happen next and what is observed. Identifying important or surprising events has the potential to substantially improve our understanding of healthcare delivery and better inform clinicians about the status of their patients.

Divergence between the model's expectation and the actual observation or informativeness broadly indicates one of three things: (1) practice variation, when a clinician makes a decision which deviates from what is typically done in similar contexts in the training data (e.g., prescribing a medication off-label); (2) an unexpected change in patient condition that would generally be observed in clinical measurements (e.g., a lab result which indicates a change in patient state that could not be predicted using observed covariates); or (3) issues of data quality which could be present in either orders or patient measurements (e.g., a typo when entering a value). In all three cases, there is the potential to learn substantially from the model's "surprise," to potentially reduce clinical errors, and to rapidly and succinctly surface the most important information for a clinician. This could be used to both better inform the time-sensitive decision-making that often occurs in the hospital setting as well as to provide a summary of the most important information about a patient for downstream analyses like phenotyping or sub-population identification.

While this work focuses on patients receiving critical care in the inpatient setting, the approach aims to be one which can be generalized to many healthcare settings with longitudinal data and outcomes. This work operates at the level of Electronic Health Records (EHRs) corresponding to individual hospitalization stays. For each hospitalization, we form a sequence of tokens that describe the admitted patient, along with their admission type, and then chronicle vitals, administered medications, lab results, assessments, and a few other categories of data as they become available, ending with a token for discharge.[9,33,49] We perform self-supervised training of a foundation model (FM) to predict the next token in one of these sequences given all previous tokens. Such models have proven remarkably effective across a number of fields[4,41,50] but most importantly in our case for predicting a variety of downstream clinical outcomes.[56,57] Furthermore, these models are generative[34] in the sense that they estimate the joint probability distribution on these sequences. Given a trained model and a novel sequence, we can estimate the context-aware (conditional) information of each token. We call a series of tokens that become available at the same time "an event" and calculate context-aware information for each event. We show that highly informative tokens and events are more

predictive of downstream outcomes and tend to result in greater changes to the model-derived understanding of a patient's current condition.

In this paper, we present the first comprehensive study of FM-derived information quantification for tokens and events in EHRs. Our main contributions are as follows:

(1) We propose a principled FM-derived method to identify highly informative tokens and events in a patient's EHR. As opposed to classical rules-based methods, our context-informed approach identifies anomalous labs and assessments even when a patient has values within what would be considered a normal range. As opposed to the variable importance methods applied to classifiers trained for specific outcomes, our method defines informativeness in terms of the sequences themselves.
(2) We illustrate how the occurrence of highly informative events impacts a patient's prognosis and alters the FM-derived representation that is commonly used for making downstream predictions. In terms of interpretability, this allows us to provide a list of events deemed most informative to the FM. We show that dropping these events from a patient's timeline impacts the performance of downstream prognostic models. Conversely, we show that events carrying the least information can be dropped without sacrificing predictive performance.

2. Related work

Early approaches to modeling sequential data derived from EHRs focused on recurrent neural networks (RNNs) including Long Short-Term Memory [22, LSTM] networks.[5,10,30,39] Approaches shifted from RNNs to transformers[53] beginning with variations on BERT,[11] including BEHRT[29] and Med-BERT.[40] Subsequently, Foresight[27] and ETHOS[42,43] both used generative pretrained transformer [38, GPT] architectures. Wornow, et al. provided a detailed review of FMs for EHRs up to 2023.[57] More recently, Mamba,[19] a selective state-space model, has found applications in ClinicalMamba[59] and EHRMamba.[13]

Some efforts have been made to better understand these types of models. Beaulieu-Jones, et al.[6] noted that sequential EHR models can learn both from the patient's actual state (e.g. the result of a particular lab) and from clinicians' actions (e.g. the fact that a particular lab was ordered). They found that models trained on demographics, admissions data, and charges from the first day of admission (clinician-initiated actions) performed competitively against models trained on full sequences of EHR data. In doing so, they raised an important point about understanding which tokens and events in a patient's sequence drive a model's understanding of that sequence.

Wornow et. al[55] studied, among other things, how sequences derived from EHRs differ from natural language (written English). They showed how EHRs exhibit copy-forwarding of chronic diagnoses, irregular spacing between tokens, and increased perplexity of tokens over time due to disease progression. Their definition of perplexity relates closely to our definition of informativeness, but they did not investigate which types of tokens tend to carry more information, nor did they consider subsequences. In contrast to this work, they focused on much longer-term time horizons consisting of multiple clinical encounters, whereas we focus on single hospitalizations.

3. Methods

3.1. *Data*

We considered 422,765 hospitalizations for adults (age 18 or older) from the Beth Israel Deaconess Medical Center between 2008–2019 (MIMIC-IV-3.1[25]) and 50,440 hospitalizations from the UCMC health system between March 2020–March 2022. We restricted our analysis to patients with stays of at least 24 hours. We formatted EHR data from each health system into the CLIF standard.[44] The MIMIC patients were partitioned into training, validation, and test sets at a 70%-10%-20% rate. We then collected each hospitalization for patients in a given set. In this way, hospitalization records in the test set corresponded to patients with no hospitalizations in the training or validation sets to avoid any potential information leakage. The UCMC data was primarily used as a held-out test set, and so was partitioned at a 5%-5%-90% rate according to the time of each patient's first hospitalization, with training patients coming first, followed by validation and then test.

3.2. *Tokenization*

We converted each hospitalization event from the CLIF standard into a sequence of tokens (represented computationally as non-negative integers) as follows. For a given sequence, the first token always corresponds to a timeline start token. The next three tokens contain patient-level demographic information on race, ethnicity, and sex. The following two tokens correspond to admission-specific information, namely patient age converted to a decile and admission type. Taken together, we refer to the 5 tokens occurring immediately after the timelines start token as the *admission prefix*. Tokens corresponding to a variety of events for a hospitalization are then inserted in the same order in which these events occurred. Transfers are encoded with their standardized location category. Labs are encoded with two tokens and inserted at the time results become available: one for lab category, and a second for the deciled lab value in the training set within that category. We call this strategy, of tokenizing categories and binning their corresponding values according to the training value of the deciles, category-value tokenization. See Figure 1 for an illustration. A handful of other tables receive this type of tokenization: vitals and results according to vital category, medication and dosage by medication category, assessment and results by assessment category. Respiratory information is recorded at the beginning of respiratory support; the encoded information is mode category and device category. We include a token indicating if a patient is placed into a prone position. All hospitalization-related data is encoded this way and inserted in chronological order. Tokens that arrive synchronously correspond to an event and always appear coterminously in a sequence. Timelines then end with a token for discharge category and a dedicated timeline end token. We did not use time-spacing or artificial time tokens[36] as recent studies suggest they do not improve performance.[55]

3.3. *Context-aware information*

Consider the set V^T of length-T sequences of tokens drawn from some vocabulary V. Such sequences correspond directly to tokenized EHR data as described in the previous subsection.

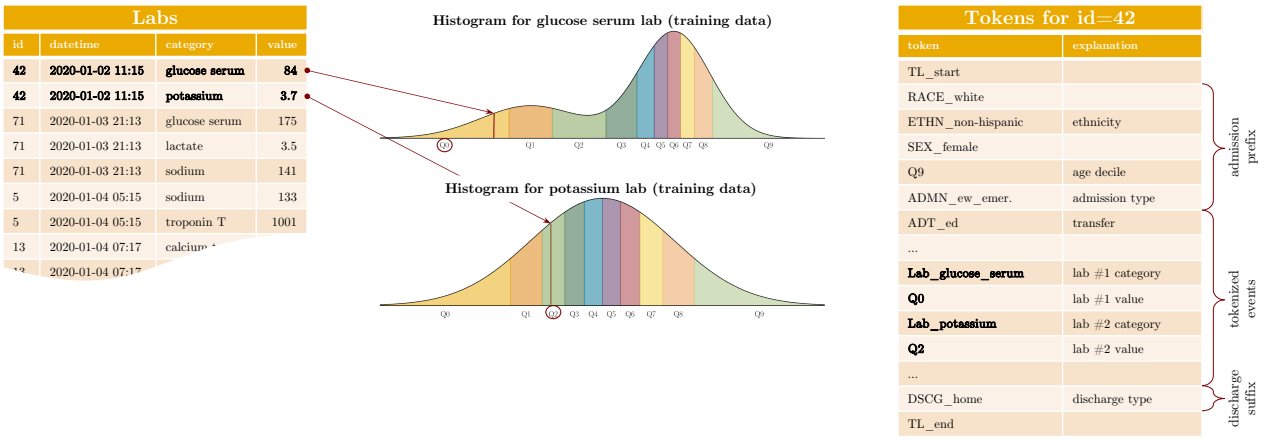

Figure 1. *Category-value tokenization.* We convert lab results into tokens as follows. For each lab category, we determine decile cutoffs (center) using all results corresponding to that lab category available in the training dataset. Each lab value is then encoded as a decile (with `Q0` corresponding to the lowest decile, `Q1` to the next, and so on up to `Q9`) and inserted into the corresponding hospitalization in temporal order.

For a given sequence $x = (x_1, \ldots, x_T)$ and indices $1 \leq u \leq v \leq T$, we let $x_{u:v} = (x_u, x_{u+1}, \ldots, x_v)$ correspond to the subsequence and $x_{<u} = x_{1:u-1}$ to the context at u for $u > 1$. If p is a probability distribution on V^T, we let $p(x_{u:v}) = \mathbb{P}_{X\sim p}(X_{u:v} = x_{u:v})$ denote the marginal distribution and $p(x_{u:v}|x_{y:z}) = \mathbb{P}_{X\sim p}(X_{u:v} = x_{u:v}|X_{y:z} = x_{y:z})$ denote the conditional for indices u, v, y, z. We adopt the convention that $p(x_{u:v}|x_{<1}) = p(x_{u:v})$. With these definitions, the Shannon self-information[46] of a certain realized subsequence $x_{u:v}$ under p is given by $I_p(x_{u:v}) = -\log_2 p(x_{u:v})$. The *context-aware information* associated to a realized subsequence $x_{u:v} \in V$ and context $x_{<u} \in V^{u-1}$ is defined analogously, by

$$\boxed{I_p(x_{u:v}|x_{<u}) = -\log_2 p(x_{u:v}|x_{<u}).} \tag{1}$$

In the case of a single token x_t, we have $t = u = v$ and refer to $I_p(x_t|x_{<t}) = -\log_2 p(x_t|x_{<t})$ as tokenwise context-aware information. As $p(x_{u:v}|x_{<t}) = \prod_{t=u}^{v} p(x_t|x_{<t})$, it follows that

$$I_p(x_{u:v}|x_{<u}) = \sum_{t=u}^{v} I_p(x_t|x_{<t}). \tag{2}$$

Thus, context-aware information is additive.

This quantity plays a pivotal role in the training of standard models. A model is a parameterized distribution p_θ on V^T. Training attempts to find parameters θ that minimize the relative entropy (or Kullback–Leibler divergence[28]) between the empirical distribution $\hat{p}$ given by the training set and p_θ,

$$D(\hat{p}||p_\theta) = \underbrace{\mathbb{E}_{X_{1:T}\sim\hat{p}}[I_{p_\theta}(X_{1:T})]}_{H(\hat{p},p_\theta)} - \underbrace{\mathbb{E}_{X_{1:T}\sim\hat{p}}[I_{\hat{p}}(X_{1:T})]}_{H(\hat{p})}. \tag{3}$$

Here, $H(\hat{p}, p_\theta)$ is the cross-entropy between $\hat{p}$ and p_θ and $H(\hat{p})$ is the entropy of $\hat{p}$. As this latter term is independent of θ, it may be disregarded during training / optimization. By (2), we have the simplification $H(\hat{p}, p_\theta) = \sum_{t=1}^{T} \mathbb{E}_{X_{1:T}\sim\hat{p}}[I_{p_\theta}(X_t|X_{<t})]$. We see that the training process optimizes θ to minimize the expected tokenwise context-aware information over the training set.

In more general terms, training finds the model p_θ that makes the training set least surprising. This is equivalent to maximum likelihood estimation [17, §5.5]. Upon completion of training, p_{θ_*} with optimized parameters θ_* serves as our best approximation to p and can be used to calculate the context-aware information (1) in new timelines for tokens and subsequences.

3.4. *Model architecture and training*

For our parameterized distribution p_θ on sequences of tokens/integers, we train a model from scratch based on the Llama-3.2 model architecture[18] with a hidden size of 1024, intermediate size of 2048, 8 hidden layers, and 8 attention heads, for a total of 67.3 million parameters. Wornow, et al.'s [55, Fig. 1B] architecture comparison indicates that the Llama architecture performs favorably to GPT,[38] Hyena,[37] and Mamba[19] architectures for context lengths of 1000-2000 tokens, such as we use here.

As our vocabulary is created during the tokenization process, we train models from scratch, as opposed to fine-tuning models that have been pre-trained on a tokenized natural language vocabulary. We train weights to minimize (3) with AdamW,[32] a variant of Adam[26] with decoupled weight decay.[21] Training batches were formed by packing tokenized sequences into a $b \times 1024$-dimensional array in row-major order where b is the batch size.[a] We used tree-structured Parzen estimators[1] to tune the learning rate (between $5 \cdot 10^{-5}$ and $5 \cdot 10^{-4}$, inclusive) and effective batch size (between 32 and 96, inclusive). Models were trained on a single compute node with 8×A100 (40GB PCIe) GPUs, connected with 2×16-core 3.0-GHz AMD Milan processors. The model having best-performing loss on the MIMIC evaluation set was selected and provides the p used to calculate context-aware information for the remainder of the paper.

3.5. *Representation-based prognostic models*

As a causal language model or state space-based model processes a sequence $x_{1:T}$ of tokens, it forms a representation $R(x_{1:t}) \in \mathbb{R}^d$ of the subsequence encountered up to the tth token for each $1 \leq t \leq T$, where d tends to be at least a few hundred dimensions. In Llama models, we take the last hidden state to be our representation, with d equal to the "hidden size" parameter, in our case set to 1024. In many FM-based works,[13,48,55] these representations or a function of them provide the basis for all subsequent prognostic predictions. For example, the representation $R(x_{1:t_0})$ of a patient's timeline that contains all tokens occurring prior to some cutoff time will then be used as features in a logistic regression model to predict outcomes for that patient occurring after the cutoff time. All patients start with the same representation, i.e. $R(x_1)$ corresponds to the representation associated to the "timeline start" token. As tokens are added to each timeline, these representations diverge. We are generally interested in the relationship between informativeness and corresponding changes in representation space at both the token

[a]Note, because of this packing strategy, the model does not learn that our sequences always start with the timeline start token. By convention, the true $p(x_1)$ is an indicator function on the timeline start token, so that $I_p(x_1)$ should be $-\log_2 1 = 0$. (Deterministic tokens do not carry information.) The model should learn that the first token after the start token should be a race token, and then an ethnicity token, and so on, because for these predictions, context is supplied.

and event levels of granularity. Establishing a strong relationship between information content and changes in representation could help to explain the predictions of representation-based prognostic models. To this end, we define the magnitude of the change in representation space when token x_t is added as

$$\Delta_t = \|R(x_{1:t}) - R(x_{1:t-1})\| \tag{4}$$

where the norm is taken to be the standard Euclidean norm and define the path length in representation space corresponding to a subsequence $x_{u:v}$ as

$$\Delta_{u:v} = \sum_{t=u}^{v} \Delta_t. \tag{5}$$

3.6. *Redaction experiment*

We restrict our cohort to patients who are admitted to the ICU within the first 24 hours of their admission. We consider two outcomes: *inpatient mortality*, defined as patient death prior to discharge from the hospital, and *long length-of-stay*, defined as discharge occurring ≥ 7 days after admission.

For each timeline truncated at the 24 hour mark, we calculate context-aware information for each event. For each of 10%, 20%, 30%, & 40%, we drop that percentage of either the most or the least informative events, or that percentage of events chosen at random. We do this for each combination of percentage and method (most, least, random), creating 12 partially redacted versions of the original 24-hour timelines.

For each data version, we extract 24-hour representations $R(x_{1:t_0})$ using our model, where $t_0 \leq 1024$ corresponds to the last token to arrive within 24 hours of admission. Note that, in an abuse of notation, t_0 depends on the hospitalization sequence x. Much more information is collected for some patients in the first 24 hours than for others. We then train a logistic regression classifier to predict each outcome (inpatient mortality and long length-of-stay) given the 24-hour representations on the training portion of the MIMIC dataset. We apply each model to the respective versions of both the MIMIC and UCMC test sets.

We perform bootstrap resampling to estimate 95% confidence intervals for test set-based variability in the ROC-AUC [12, *cf.* §13.3]. This method takes a fixed classification model and forms an empirical distribution of performance metrics by resampling test data 10,000 times.

We also use bootstrap sampling to estimate p-values for the hypothesis test of $H_0 : \mathrm{AUC}_0 = \mathrm{AUC}_1$ against the one-sided alternative $H_a : \mathrm{AUC}_0 > \mathrm{AUC}_1$, where AUC_0 corresponds to the original AUC and AUC_1 to the AUC from a fixed classifier built and tested on redacted timelines [12, Algorithm 16.1]. This method compares the observed difference in AUC performance against differences obtained from 10,000 resamplings under the hypothesis that predictions from the two classifiers are exchangeable. This bootstrapping approach also only simulates variability in the test set given fixed classifiers.

4. Results

4.1. *Highlighted timelines*

We present the first 210 tokens from three timelines in row-major order along with comments as Figures 2-4. Starting from the upper left-hand corner, each row can be read off from left to right.

TL_START	RACE_white	ETHN_non-hispanic	SEX_male	Q4	ADMN_observation_admi
ADT_ed	ADT_ward	LAB_hemoglobin	Q5	LAB_platelet_count	Q1
LAB_bicarbonate	Q8	LAB_chloride	Q5	LAB_creatinine	Q4
LAB_glucose_serum	Q2	LAB_magnesium	Q8	LAB_potassium	Q3
LAB_sodium	Q7	LAB_bun	Q2	LAB_calcium_total	Q8
LAB_phosphate	Q5	LAB_ldh	Q9	LAB_troponin_t	Q9
LAB_basophils_percent	Q3	LAB_eosinophils_percent	Q2	LAB_lymphocytes_percent	Q1
LAB_monocytes_percent	Q0	LAB_neutrophils_percent	Q0	LAB_basophils_absolute	Q3
LAB_hemoglobin	Q5	LAB_platelet_count	Q1	LAB_inr	Q4
LAB_pt	Q3	LAB_troponin_t	Q9	LAB_bicarbonate	Q8
LAB_chloride	Q5	LAB_creatinine	Q5	LAB_glucose_serum	Q3
LAB_magnesium	Q7	LAB_potassium	Q1	LAB_sodium	Q7
LAB_bun	Q3	LAB_calcium_total	Q7	LAB_phosphate	Q7
LAB_ldh	Q9	LAB_basophils_percent	Q3	LAB_eosinophils_percent	Q2
_AB_lymphocytes_percent	Q1	LAB_monocytes_percent	Q0	LAB_neutrophils_percent	Q0
LAB_basophils_absolute	Q3	LAB_alt	Q7	LAB_alkaline_..phatase	Q0
LAB_ast	Q7	LAB_bilirubin_total	Q2	LAB_bicarbonate	Q6
LAB_chloride	Q8	LAB_creatinine	Q5	LAB_glucose_serum	Q3
LAB_magnesium	Q7	LAB_potassium	Q2	LAB_sodium	Q7
LAB_bun	Q3	LAB_calcium_total	Q3	LAB_phosphate	Q6
LAB_ldh	Q9	LAB_troponin_t	Q9	LAB_hemoglobin	Q2
LAB_platelet_count	Q1	LAB_basophils_percent	Q3	LAB_eosinophils_percent	Q2
_AB_lymphocytes_percent	Q0	LAB_monocytes_percent	Q0	LAB_neutrophils_percent	Q0
LAB_basophils_absolute	Q3	LAB_hemoglobin	Q2	LAB_bicarbonate	Q5
LAB_chloride	Q7	LAB_creatinine	Q5	LAB_glucose_serum	Q5
LAB_magnesium	Q7	LAB_potassium	Q4	LAB_sodium	Q6
LAB_bun	Q3	LAB_calcium_total	Q3	LAB_phosphate	Q5
LAB_alt	Q7	LAB_alkaline_..phatase	Q0	LAB_ast	Q6
LAB_bilirubin_total	Q1	LAB_ldh	Q9	LAB_inr	Q5
LAB_pt	Q5	LAB_ptt	Q3	LAB_basophils_percent	Q3
LAB_eosinophils_percent	Q2	LAB_lymphocytes_percent	Q1	LAB_monocytes_percent	Q1
LAB_neutrophils_percent	Q0	LAB_basophils_absolute	Q3	LAB_platelet_count	Q1
LAB_troponin_t	Q9	LAB_hemoglobin	Q1	LAB_platelet_count	Q1
LAB_inr	Q5	LAB_pt	Q5	LAB_ptt	Q5
LAB_troponin_t	Q9	LAB_basophils_percent	Q3	LAB_eosinophils_percent	Q2

Figure 2. Timeline highlighted by tokenwise context-aware information for MIMIC hospitalization 24640534 (first 210 tokens). This white, non-Hispanic male of age ~ 60 was admitted to the ED for observation. He had no previous admission history within MIMIC. His stay was 37 days 22 hours in duration. After the stay, he subsequently received ICD-10 diagnoses C9200 for 'Acute myeloblastic leukemia, not having achieved remission' and I214 for 'Non-ST elevation (NSTEMI) myocardial infarction', among others. (Full diagnostic list in appendix.) The model successfully identified low lymphocytes as being potentially clinically relevant (`Lab_lymphocytes_percent`, `Q1`), but overlooked the low neutrophils (`Lab_neutrophils_percent`, `Q0`) and high troponin T (`Lab_troponin_t`, `Q9`).

We see that informative tokens sometimes correspond to lab events or vitals readings that have a direct bearing on the patient's current state. Examples include the lymphocytes percentage lab in Figure 2, arterial PCO_2 in Figure 3, and blood pressure readings in Figure 4. Informative tokens can also correspond to clinician-initiated events, such as the CAM assessment[24] in Figure 3 following a low RASS[45] score. (If the RASS score were -4 as opposed to -3, typically the CAM assessment would not be made until later.) Finally, informative tokens can correspond to measurements that seem implausible and may be worth further investigation, such as the Braden scores[7] in Figure 4.

4.2. *Counts of highly informative tokens and events anticipate negative outcomes*

We consider $T_{\geq 95}$, the number of tokens exceeding the 95th percentile for informativeness, $E_{\geq 95,<99}$, the number of events in the 95th to the 99th percentile, and $E_{\geq 99}$, the number of events exceeding the 99th percentile for informativeness. For these definitions, we restrict to tokens and events that occur within the first 24 hours of admission.

In a logistic regression model for inpatient mortality in the MIMIC test set, we find that $T_{\geq 95}$ $(\hat{\beta} = 0.0269, p < 0.001)$, $E_{\geq 95,<99}$ $(\hat{\beta} = 0.3015, p < 0.001)$, and $E_{\geq 99}$ $(\hat{\beta} = 0.2480, p < 0.001)$ all have positive coefficients and are highly significant. Similarly for long length of stay, we find that

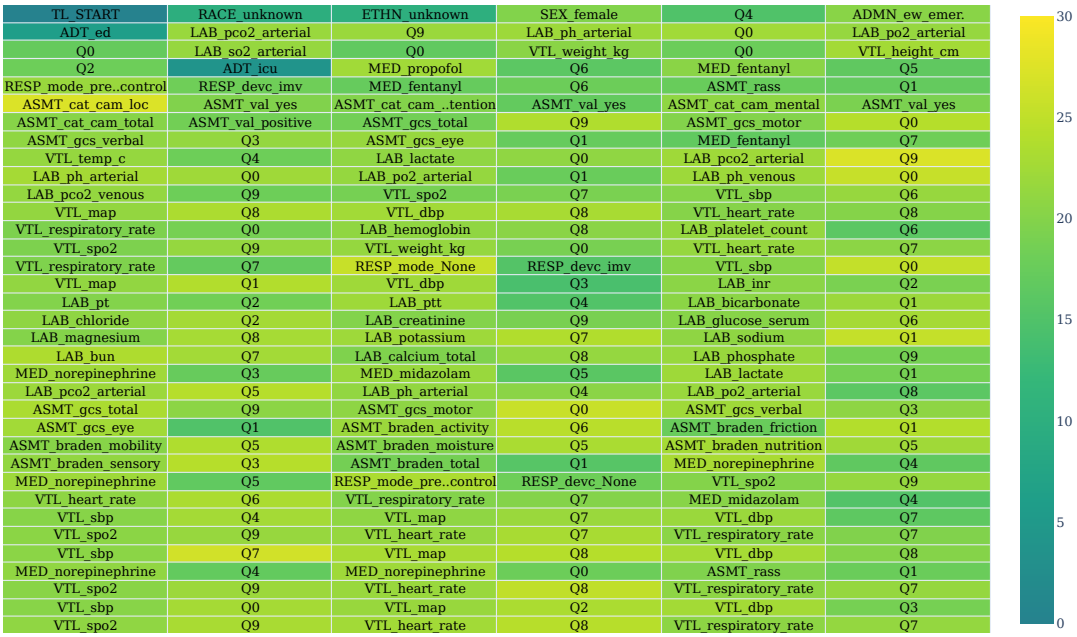

Figure 3. Timeline highlighted by tokenwise context-aware information for MIMIC hospitalization 26886976 (first 210 tokens). This female of unknown race and ethnicity was admitted to the ED at age $\sim$ 73. Her 12 day 2 hour hospital stay ended in death. Afterwards, she subsequently received ICD-10 diagnoses A4189 for 'Other specified sepsis', R6521 for 'Severe sepsis with septic shock', and N179 for 'Acute kidney failure, unspecified', in addition to other diagnoses. After the patient received a low Richmond Agitation-Sedation Scale score [45, RASS] with (`ASMT_rass`, `Q1`) indicating a high likelihood of coma, the model finds the administration of the Confusion Assessment Method [24, CAM] (`ASMT_cat_cam_loc`) to evaluate delirium to be surprising/ informative. The model also finds the high arterial PCO_2 (`LAB_pco2_arterial`, `Q9`) to be of interest.

$T_{\geq 95}$ ($\hat{\beta} = 0.0163, p < 0.001$), $E_{\geq 95,<99}$ ($\hat{\beta} = 0.2872, p < 0.001$), and $E_{\geq 99}$ ($\hat{\beta} = 0.1236, p < 0.001$) are positively and significantly associated. In the UCMC test dataset (where percentiles are based on statistics from the UCMC data), $T_{\geq 95}$ ($\hat{\beta} = 0.0148, p < 0.001$), $E_{\geq 95,<99}$ ($\hat{\beta} = 0.1684, p < 0.001$), and $E_{\geq 99}$ ($\hat{\beta} = 0.4798, p < 0.001$) all associate with inpatient mortality. Similarly, $T_{\geq 95}$ ($\hat{\beta} = 0.0165, p < 0.001$), $E_{\geq 95,<99}$ ($\hat{\beta} = 0.1292, p < 0.001$), and $E_{\geq 99}$ ($\hat{\beta} = 0.4727, p < 0.001$) associate positively with long length of stay.

4.3. *Informative tokens tend to result in larger changes to a patient's latent representation*

In our MIMIC test set, a simple linear regression for Δ_t from (4) given informativeness $I_p(x_t|x_{<t})$ yields $\hat{\beta} = 0.548, p < 0.001$ with $R^2 = 0.212$. For a breakdown of average Δ_t vs. average informativeness by token type, see Figure 5. Positioning and transfer tokens tend to carry less information, while assessment, lab, and quantile `Q` tokens tend to carry more.

At the level of events $x_{u:v}$, a regression for path length $\Delta_{u:v}$ from (5) given event-level informativeness $I_p(x_{u:v}|x_{<u})$ yields $\hat{\beta} = 2.081, p < 0.001$ with $R^2 = 0.997$ (see also Figure 6). More informative events trace out longer paths in representation space. Perhaps surprisingly, there does not appear to be a strong linear relationship between event informativeness and

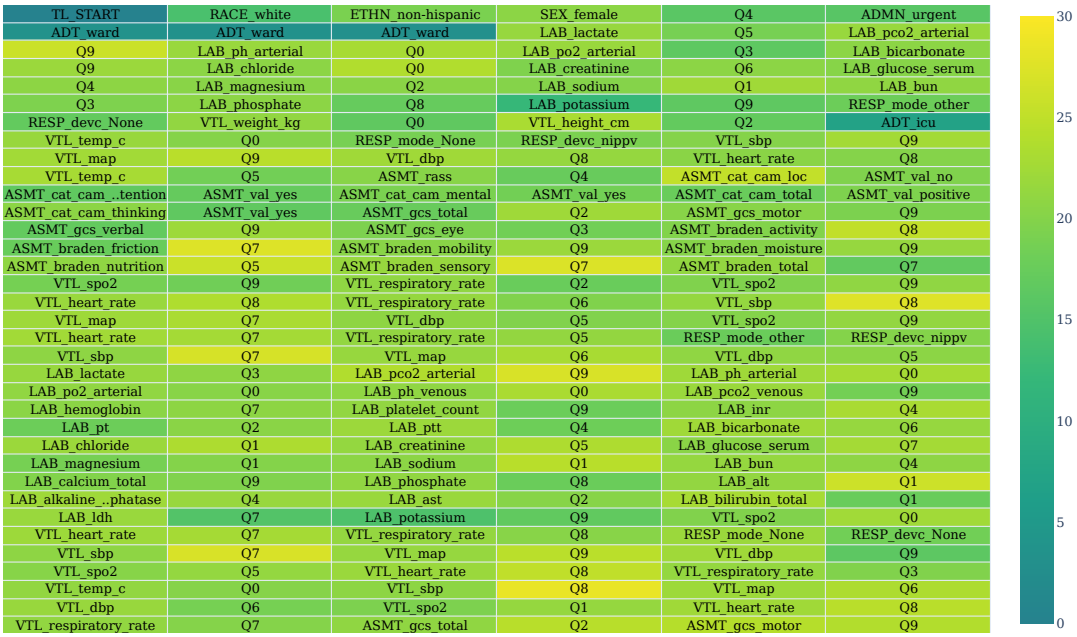

Figure 4. Timeline highlighted by tokenwise context-aware information for MIMIC hospitalization 29022625 (first 210 tokens). This $\sim$ 55 year old white female had previously been seen for a myriad of conditions (see appendix). After a a 30 day 20 hour stay, she received new diagnoses of K2211 for 'Ulcer of esophagus with bleeding', J9601 for 'Acute respiratory failure with hypoxia', J9602 for 'Acute respiratory failure with hypercapnia', A419 for 'Sepsis, unspecified organism', J90 for 'Pleural effusion, not elsewhere classified', E872 for 'Acidosis', and J95851 for 'Ventilator associated pneumonia', among others. The model notices that the Braden scores seem implausibly high, and highlights the hypercarbia (`LAB_pco2_arterial`, `Q9`). It seems to miss the hypoxia (`VTL_spo2`, `Q0`); however, SPO_2 readings up to 93.0 are placed in decile `Q0` so this may be due to the tokenization strategy. The model also emphasizes the patient's rising blood pressure (`VTL_sbp`) over time.

total distance moved in representation space during the course of an event $x_{u:v}$, i.e. $\|R(x_{1:v}) - R(x_{1:u-1})\|$.

4.4. *Redacting informative events significantly reduces prognostic ability*

Results from our redaction experiment (described in §3.6) indicate that dropping highly informative events from a patient's timeline significantly impairs representation-based classifier performance in the MIMIC test set. For ROC-AUC, we find statistically significant performance disparities when dropping as few as 20% of the most informative events. Conversely, events carrying little information can readily be dropped from a timeline without significantly impacting predictive performance. For representation-based logistic regression models trained on the MIMIC training set, predictive performance on the MIMIC and UCMC test sets are available in Table 1. In addition to ROC-AUC, we report PR-AUC, the area under the precision-recall curve, and the Brier score,[8] which corresponds to the mean squared error between predicted probabilities and boolean realizations in online Appendix E. Higher ROC-AUC and PR-AUC values and lower Brier scores correspond to better performing models.

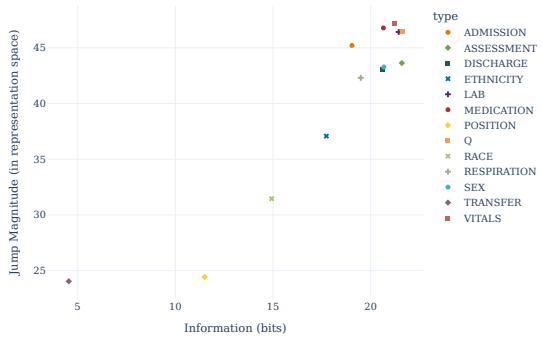

Figure 5. Average Δ_t versus average token-wise informativeness by token type for the 24.7 million tokens x_t in the MIMIC test set.

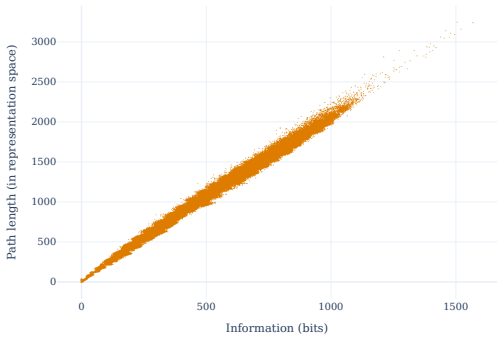

Figure 6. Path length $\Delta_{u:v}$ versus eventwise informativeness for the 2.8 million events $x_{u:v}$ in the MIMIC test set.

5. Discussion

In this work, we developed a method to quantify the informativeness of clinical events as observed in EHRs based on their tokenized representation.[b] We found that highly informative tokens can correspond to measurements of clinical significance, to clinician-initiated events or lab orders, and in some cases to records that seem *prima facie* to be data-entry errors.[35] Tokens that carry more information tend to precipitate larger changes in a patient's latent representation and events that carry more information tend to have longer paths in representation space. Counts of highly informative / surprising tokens and events in the first 24 hours of a patient's stay relate to an increased risk of future negative outcomes like death or long length-of-stay. Redacting highly informative events reduces the predictive performance of representation-based classifiers, while redacting a fraction of relatively uninformative events tends to not result in significant performance drops.

5.1. *Broad applicability beyond clinical prediction*

The FM-derived informativeness measure extends well beyond traditional clinical prediction tasks and opens new avenues for AI applications to healthcare. Context-aware information quantification provides a principled framework for addressing downstream challenges that have historically relied on heuristic approaches.

Our informativeness metric could provide a data-driven solution to clinical alarm fatigue through dynamic alerting systems that prioritize notifications based on contextual surprise rather than static thresholds. For example, a blood pressure of 118/86 mmHg might be routine in most contexts, but could be highly informative if it represents a rapid drop in a patient

[b]Shannon famously estimated the average information content of words in written English to be around 11.82 bits [47, Eq. 7]. Under the assumption that our model p trained on the MIMIC training set adequately approximates the empirical distribution $\hat{p}$, we can average $I_p(x_t|x_{<t})$ over x_t in the MIMIC test set to roughly approximate that tokens in our timelines each carry around 21.23 bits of information on average.

Table 1. ROC-AUC for the two classification tasks on ICU patients in the MIMIC and UCMC test sets. Stars correspond to the significance level of a hypothesis test against the one-sided alternative that the model trained on the original data performs better. A single asterisk * corresponds to $p < 0.05$, two ** to $p < 0.01$ and three *** to $p < 0.001$.

version		Inpatient mortality		Long length-of-stay	
method	pct.	MIMIC	UCMC	MIMIC	UCMC
original	—	0.869 ± 0.009	0.839 ± 0.013	0.740 ± 0.008	0.661 ± 0.011
top	10	0.860 ± 0.010	0.830 ± 0.013	0.735 ± 0.009	0.671 ± 0.011
	20	0.848 ± 0.011 **	0.814 ± 0.014 **	0.726 ± 0.009 **	0.653 ± 0.012
	30	0.833 ± 0.012 ***	0.812 ± 0.013 **	0.720 ± 0.009 **	0.642 ± 0.011 *
	40	0.823 ± 0.011 ***	0.818 ± 0.012 *	0.714 ± 0.009 ***	0.649 ± 0.012
bottom	10	0.867 ± 0.010	0.834 ± 0.011	0.736 ± 0.012	0.659 ± 0.013
	20	0.866 ± 0.009	0.834 ± 0.012	0.732 ± 0.010	0.667 ± 0.012
	30	0.862 ± 0.011	0.829 ± 0.012	0.726 ± 0.009 *	0.667 ± 0.013
	40	0.859 ± 0.012	0.829 ± 0.011	0.724 ± 0.008 **	0.664 ± 0.011
random	10	0.866 ± 0.008	0.838 ± 0.012	0.737 ± 0.006	0.664 ± 0.012
	20	0.863 ± 0.008	0.835 ± 0.011	0.733 ± 0.009	0.667 ± 0.012
	30	0.865 ± 0.010	0.835 ± 0.014	0.728 ± 0.007	0.674 ± 0.009
	40	0.861 ± 0.011	0.835 ± 0.011	0.727 ± 0.008 *	0.674 ± 0.011

being treated for an ischemic stroke or hypertensive emergency. In these conditions, a slower reduction in blood pressure is preferred to avoid complications, making a rapid drop worthy of alerting a clinician. Traditional rule-based clinical decision support alerts would likely consider this a normal reading and therefore fail to identify a potentially concerning change. The ability to use context to differentiate this alarming event from a similar, but clinically appropriate, change in another patient distinguishes FM-based approaches.[23]

Detecting data entry errors represents another potential application. Traditional validation of data entry quality is dedicated to verification of abnormal values, but our approach could identify contextually implausible entries that fall within normal ranges. Overall, surprise quantification could help automate much of the manual chart review process currently required for quality assurance. For clinical research, informativeness patterns could enable novel patient phenotyping approaches. Rather than relying on pre-specified diagnostic codes, researchers could identify patient subgroups characterized by similar patterns of surprising events, potentially revealing previously unrecognized disease subtypes or complex pathologies difficult to capture with traditional algorithms. So-called events-based models[14,61] have already proven remarkably effective at both subtyping and staging neurodegenerative disease,[3,54,60] but could in the future find broader applications.

At the health system level, patterns of informativeness could inform resource allocation decisions. Units characterized by high rates of surprising events might require additional staffing or monitoring capabilities. Our finding that surprising events correlate with negative outcomes suggests informativeness patterns could serve as early warning indicators for periods of increased clinical risk. Furthermore, cases highlighted by our informativeness measure could

serve as valuable educational resources. Clinical scenarios with highly informative events represent situations where standard protocols might be insufficient, making them ideal for training clinicians to recognize complex presentations.

Quantifying the information in EHRs could help clinicians quickly detect anomalous events and identify data entry errors. FMs trained on tokenized EHRs have already demonstrated remarkably good performance on prognostic tasks. Finding ways to better understand and interpret predictions made from these models through informativeness quantification represents a significant step toward more transparent and actionable clinical AI systems. The broad applicability of this informativeness framework across diverse healthcare challenges suggests that context-aware information quantification could become a foundational tool for healthcare analytics, complementing traditional approaches with a more nuanced understanding of clinical surprise and significance.

5.2. *Limitations and directions for future work*

We are actively working to improve our approach to tokenization. Our decision to encode labs and medications at a broad categorical level significantly reduces the overall vocabulary,[c] allowing more efficient training on a relatively small dataset. However, it can merge clinically distinct items into the same token. For instance, potassium measured from whole blood versus serum is treated in the same manner, obscuring an important factor for distinguishing pseudo-hyperkalemia from life-threatening hyperkalemia. Our approach to tokenizing values by decile within each category is standard,[43,55] but sometimes results in deciles that straddle clinically relevant thresholds. Future work may explore tokenization strategies that incorporate lab reference values. We also intend to develop methods to distinguish between the various types of surprising events identified in this work (e.g. to differentiate changes in a patient's condition from data entry errors). Manual chart review would be required to support claims of data entry errors. Finally, we also intend to compare our approach to attention-based mechanisms for identifying important tokens.[15,20,31]

Data and code availability. The MIMIC-IV-3.1 dataset[25] is available to credentialed users on Physionet.[16] The UCMC dataset is available in the CLIF format for federated, privacy-preserving analyses. Reasonable requests may be directed to WFP. Github hosts code to reproduce the results found here.[d] The arXiv version[e] of this work contains an appendix with a listing of the vocabulary, all learned decile cutoffs, additional examples, and detailed results.

Acknowledgments. This work was funded in part by the National Institutes of Health, specifically the National Institute of Neurological Disorders and Stroke grant R00NS114850 to BKB and National Library of Medicine grant R01LM014263 to WFP. This project would not have been possible without the support of the Center for Research Informatics at the University of Chicago and particularly the High-Performance Computing team.

[c]We used 208 unique tokens while Renc et al.[42] used 4,495 and Wornow et al.[55] used 39,818.
[d]See: `https://github.com/bbj-lab/Quantifying-Surprise-EHRs`
[e]See: `https://arxiv.org/abs/2507.22798`

Bibliography

1. T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. In *KDD*, 2019.
2. H. R. Anderson, A. C. Borgen, R. Christnacht, J. Ng, J. G. Weller, H. N. Davison, et al. Stats on the desats: Alarm fatigue and the implications for patient safety. *BMJ Open Qual.*, 12(3), 2023.
3. D. Archetti, A. L. Young, N. P. Oxtoby, D. Ferreira, G. Mårtensson, E. Westman, D. C. Alexander, G. B. Frisoni, and A. Redolfi. Inter-cohort validation of Sustain model for Alzheimer's disease. *Front. Big Data*, 2021.
4. M. Awais, M. Naseer, S. Khan, R. M. Anwer, H. Cholakkal, M. Shah, M.-H. Yang, and F. S. Khan. Foundation models defining a new era in vision: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(4):2245–2264, 2025.
5. B. K. Beaulieu-Jones, P. Orzechowski, and J. H. Moore. Mapping patient trajectories using longitudinal extraction and deep learning in the MIMIC-III critical care database. In *PSB*, pages 123–132, 2018.
6. B. K. Beaulieu-Jones, W. Yuan, G. A. Brat, A. L. Beam, G. Weber, M. Ruffin, and I. S. Kohane. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? *npj Digit. Med.*, 4(1):62, 2021.
7. N. Bergstrom, B. J. Braden, A. Laguzza, and V. Holman. The Braden scale for predicting pressure sore risk. *Nurs. Res.*, 36(4):205–210, 1987.
8. G. W. Brier. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.*, 78(1):1–3, 1950.
9. M. C. Burkhart, B. Ramadan, Z. Liao, K. Chhikara, J. C. Rojas, W. F. Parker, and B. K. Beaulieu-Jones. Foundation models for electronic health records: representation dynamics and transferability. arXiv:2504.10422, 2025.
10. E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun. Doctor AI: Predicting clinical events via recurrent neural networks. arXiv:1511.05942, 2016.
11. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019.
12. B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, 1993.
13. A. Fallahpour, M. Alinoori, W. Ye, X. Cao, A. Afkanpour, and A. Krishnan. EHRMamba: Towards generalizable and scalable foundation models for electronic health records. In *ML4H*, volume PMLR 259, pages 291–307, 2025.
14. H. M. Fonteijn, M. Modat, M. J. Clarkson, J. Barnes, M. Lehmann, N. Z. Hobbs, et al. An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. *NeuroImage*, 60(3):1880–1889, 2012.
15. S. Ge, Y. Zhang, L. Liu, M. Zhang, J. Han, and J. Gao. Model tells you what to discard: Adaptive KV cache compression for LLMs. In *ICLR*, 2024.
16. A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, et al. Physiobank, physiotoolkit, and physionet. *Circulation*, 101(23), 2000.
17. I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2016.
18. A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, et al. The Llama 3 herd of models. arXiv 2407.21783, 2024.
19. A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *COLM*, 2024.
20. Z. Guo, H. Kamigaito, and T. Watanabe. Attention score is not all you need for token importance indicator in KV cache reduction: Value also matters. In *Conference on Empirical Methods in*

Natural Language Processing, 2024.
21. S. Hanson and L. Pratt. Comparing biases for minimal network construction with back-propagation. In *Adv. Neur. Inf. Proc. Sys.*, 1988.
22. S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.
23. M. D. Howell, G. S. Corrado, and K. B. DeSalvo. Three epochs of artificial intelligence in health care. *JAMA*, 331(3):242–244, 2024.
24. S. K. Inouye, C. H. van Dyck, C. A. Alessi, S. Balkin, A. P. Siegal, and R. I. Horwitz. Clarifying confusion: The confusion assessment method. *Ann. Intern. Med.*, 113(12):941–948, 1990.
25. A. E. W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data*, 10, 2023.
26. D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.
27. Z. Kraljevic, D. Bean, A. Shek, R. Bendayan, H. Hemingway, J. A. Yeung, et al. Foresight-a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *Lancet Digit. Health*, 6(4), 2024.
28. S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statistics*, 22:79–86, 1951.
29. Y. Li, S. Rao, J. A. Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, K. Rahimi, and G. Salimi-Khorshidi. BEHRT: Transformer for electronic health records. *Sci. Rep.*, 10, 2020.
30. Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel. Learning to diagnose with LSTM recurrent neural networks. arXiv:1511.03677, 2017.
31. Z. Liu, A. Desai, F. Liao, W. Wang, V. Xie, Z. Xu, A. Kyrillidis, and A. Shrivastava. Scissorhands: Exploiting the persistence of importance hypothesis for LLM KV cache compression at test time. In *Adv. Neur. Inf. Proc. Sys.*, volume 36, pages 52342–52364, 2023.
32. I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.
33. M. McDermott, B. Nestor, P. Argaw, and I. S. Kohane. Event stream GPT: A data pre-processing and modeling library for generative, pre-trained transformers over continuous-time sequences of complex events. In *Adv. Neur. Inf. Proc. Sys.*, volume 36, pages 24322–24334, 2023.
34. A. Ng and M. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *Adv. Neur. Inf. Proc. Sys.*, volume 14, pages 841–848, 2002.
35. H. Niu, O. A. Omitaomu, M. A. Langston, M. Olama, O. Ozmen, H. B. Klasky, A. Laurio, M. Ward, and J. Nebeker. EHR-BERT: A BERT-based model for effective anomaly detection in electronic health records. *J. Biomed. Inf.*, 150:104605, 2024.
36. C. Pang, X. Jiang, K. S. Kalluri, M. Spotnitz, R. Chen, A. Perotte, and K. Natarajan. CEHR-BERT: Incorporating temporal information from structured EHR data to improve prediction tasks. In *Proceedings of Machine Learning for Health*, volume PMLR 158, pages 239–260, 2021.
37. M. Poli, S. Massaroli, E. Nguyen, D. Y. Fu, T. Dao, S. Baccus, Y. Bengio, S. Ermon, and C. Ré. Hyena hierarchy: towards larger convolutional language models. In *ICML*, 2023.
38. A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. OpenAI, 2018.
39. A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, et al. Scalable and accurate deep learning with electronic health records. *npj Digit. Med.*, 1(1):18, 2018.
40. L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digit. Med.*, 4(1):86, 2021.
41. M. Raza, Z. Jahangir, M. B. Riaz, M. J. Saeed, and M. A. Sattar. Industrial applications of large language models. *Sci. Rep.*, 15, 2025.
42. P. Renc, M. K. Grzeszczyk, N. Oufattole, D. Goode, Y. Jia, S. Bieganski, M. B. A. McDermott,

J. Was, A. E. Samir, J. W. Cunningham, D. W. Bates, and A. Sitek. Foundation model of electronic medical records for adaptive risk estimation. arXiv:2502.06124, 2025.
43. P. Renc, Y. Jia, A. E. Samir, J. Was, Q. Li, D. W. Bates, and A. Sitek. Zero shot health trajectory prediction using transformer. *npj Digit. Med.*, 7, 2024.
44. J. C. Rojas, P. G. Lyons, K. Chhikara, V. Chaudhari, S. V. Bhavani, M. Nour, K. G. Buell, K. D. Smith, C. A. Gao, S. Amagai, et al. A common longitudinal intensive care unit data format (CLIF) for critical illness research. *Intensive Care Med.*, 2025.
45. C. N. Sessler, M. S. Gosnell, M. J. Grap, G. M. Brophy, P. V. O'Neal, K. A. Keane, E. P. Tesoro, and R. K. Elswick. The Richmond agitation–sedation scale. *Am. J. Respir. Crit. Care Med.*, 166(10):1338–1344, 2002.
46. C. E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27:379–423,623–656, 1948.
47. C. E. Shannon. Prediction and entropy of printed English. *Bell System Tech. J.*, 30(1):50–64, 1951.
48. E. Steinberg, J. A. Fries, Y. Xu, and N. Shah. MOTOR: A time-to-event foundation model for structured medical records. In *ICLR*, 2024.
49. E. Steinberg, K. Jung, J. A. Fries, C. K. Corbin, S. R. Pfohl, and N. H. Shah. Language models are an effective representation learning technique for electronic health record data. *J. Biomed. Inf.*, 113, 2021.
50. J. Sun, C. Zheng, E. Xie, Z. Liu, R. Chu, J. Qiu, J. Xu, M. Ding, H. Li, M. Geng, et al. A survey of reasoning with foundation models: Concepts, methodologies, and outlook. *ACM Comput. Surv.*, 57(11), 2025.
51. The Joint Commission. Medical device alarm safety in hospitals. *Sentinel Event Alert*, (50), 2013.
52. The Joint Commission. *Hospital: 2025 National Patient Safety Goals*. 2024.
53. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Adv. Neur. Inf. Proc. Sys.*, volume 30, 2017.
54. J. W. Vogel, A. L. Young, N. P. Oxtoby, R. Smith, R. Ossenkoppele, O. T. Strandberg, et al. Four distinct trajectories of tau deposition identified in Alzheimer's disease. *Nat. Med.*, 27(5):871–881, 2021.
55. M. Wornow, S. Bedi, M. A. F. Hernandez, E. Steinberg, J. Fries, C. Ré, O. Koyejo, and N. H. Shah. Context clues: Evaluating long context models for clinical prediction tasks on EHRs. In *ICLR*, 2025.
56. M. Wornow, R. Thapa, E. Steinberg, J. A. Fries, and N. Shah. EHRSHOT: An EHR benchmark for few-shot evaluation of foundation models. In *Neurips Datasets and Benchmarks Track*, volume 36, pages 67125–67137, 2023.
57. M. Wornow, Y. Xu, R. Thapa, B. Patel, E. Steinberg, S. Fleming, M. A. Pfeffer, J. Fries, and N. H. Shah. The shaky foundations of large language models and foundation models for electronic health records. *npj Digit. Med.*, 6(1):135, 2023.
58. D. Xu, F. Liu, X. Ding, J. Ma, Y. Suo, Y.-Y. Peng, J. Li, and X. Fu. Exploring ICU nurses' response to alarm management and strategies for alleviating alarm fatigue: a meta-synthesis and systematic review. *BMC Nursing*, 24, 2025.
59. Z. Yang, A. Mitra, S. Kwon, and H. Yu. ClinicalMamba: A generative clinical language model on longitudinal clinical notes. In *Proceedings of the 6th Clinical NLP Workshop*, pages 54–63, 2024.
60. A. L. Young, R. V. Marinescu, N. P. Oxtoby, M. Bocchetta, K. Yong, N. C. Firth, et al. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference. *Nat. Commun.*, 9(1), 2018.
61. A. L. Young, N. P. Oxtoby, P. Daga, D. M. Cash, N. C. Fox, S. Ourselin, J. M. Schott, and D. C. Alexander. A data-driven model of biomarker changes in sporadic Alzheimer's disease. *Brain*, 137(9):2564–2577, 2014.