Count-Based Approaches Remain Strong: A Benchmark Against Transformer and LLM Pipelines on Structured EHR

Jifan Gao
Department of Data Science
Dana-Farber Cancer Institute
Boston, MA

Michael Rosenthal
Department of Radiology
Dana-Farber Cancer Institute
Boston, MA

Brian Wolpin
Department of Medical Oncology
Dana-Farber Cancer Institute
Boston, MA

Simona Cristea *
Department of Data Science
Dana-Farber Cancer Institute
Boston, MA

Abstract

Structured electronic health records (EHR) are essential for clinical prediction. While count-based learners continue to perform strongly on such data, no benchmarking has directly compared them against more recent mixture-of-agents LLM pipelines, which have been reported to outperform single LLMs in various NLP tasks. In this study, we evaluated three categories of methodologies for EHR prediction using the EHRSHOT dataset: count-based models built from ontology roll-ups with two time bins, based on LightGBM and the tabular foundation model TabPFN; a pretrained sequential transformer (CLMBR); and a mixture-of-agents pipeline that converts tabular histories to natural-language summaries followed by a text classifier. We assessed eight outcomes using the EHRSHOT dataset. Across the eight evaluation tasks, head-to-head wins were largely split between the count-based and the mixture-of-agents methods. Given their simplicity and interpretability, count-based models remain a strong candidate for structured EHR benchmarking. The source code is available at: https://github.com/cristea-lab/Structured_EHR_Benchmark.

1 Introduction

Structured electronic health record (EHR) data consists of clinical information from various domains such as medical conditions and drug exposures. These records are stored with standardized medical concepts and organized into relational tables, using standardized frameworks like the Observational Medical Outcomes Partnership (OMOP) Common Data Model [Hripcsak et al., 2015]. Compared to other data modalities, such as clinical free-text notes and medical imaging, structured EHR data offer unique advantages that make them valuable in clinical outcome prediction. Standardized vocabularies and schemas enable cross-site analyses, reproducibility, and fair benchmarking [Arora et al., 2023, Wang et al., 2025]; the structured data can scale to million-patient cohorts for efficient cohort construction and population-level training [Gamal et al., 2021]; and the organized nature of structured data significantly simplifies adherence to regulatory and ethical guidelines [Tayefi et al.,

*Correspondence to: scristea@ds.dfci.harvard.edu

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

2021], such as those mandated by the Health Insurance Portability and Accountability Act (HIPAA) [Edemekong et al., 2018].

Various modeling strategies have been developed for structured EHR data. Early approaches often relied on count-based representations [Reps et al., 2018, Khalid et al., 2021], where the occurrence or frequency of medical concepts (*e.g.*, diagnoses, procedures, medications) is aggregated into feature vectors. These representations are straightforward and compatible with common machine-learning models such as logistic regression, random forests, or gradient boosting machines, and have demonstrated strong performance on various clinical prediction tasks [Gao et al., 2024b, Bergquist et al., 2024, 2023]. Despite their simplicity, count-based models largely ignore the temporal order and contextual relationships between events.

Motivated by advances in natural language processing, more recent methods treat EHR data as sequential tokens, applying transformer architectures designed to capture contextual dependencies. Transformer-based models such as Delphi [Shmatko et al., 2024], CLMBR [Steinberg et al., 2021, Guo et al., 2024] and MOTOR [Steinberg et al., 2023] convert time-stamped medical codes into sequences and employ attention mechanisms to model longitudinal patterns in patient histories. These pretrained frameworks have shown promising performance across tasks. In addition, researchers are exploring the use of large language models (LLMs) for structured EHR prediction by converting tabular records into textual narratives [Wornow et al., 2024, Kim et al., 2025, Kirchler et al., 2025]. This paradigm allows multimodal integration beyond free text and leverages the text understanding capabilities of LLMs. In the general text domain, Wang et al. [2024] proposed a mixture-of-agents (MoA) approach and demonstrated that sequentially passing the output of one LLM to another leads to improvements in performance. Based on this collaborative potential of LLMs, Gao et al. [2025] proposed the Mixture-of-Multimodal-Agents (MoMA) for multimodal EHR data, including structured EHR. MoMA first generates concise and clinically meaningful summaries of a patient's tabular record and then passes these summaries to a specialized text classifier for downstream prediction.

Despite rapid advances in employing LLMs for EHR data, recent evidence suggests that strong count-based tabular models remain competitive for structured EHR prediction [Brown et al., 2025], particularly when data exhibit skewed or heavy-tailed distributions, where gradient boosting machines (GBMs) often have an advantage over neural networks [McElfresh et al., 2023]. At the same time, LLM evaluations for clinical prediction in structured EHR data indicate that prompt-engineered single-LLM baselines do not reliably surpass classical count-based models on structured EHR tasks [Chen et al., 2024]. In addition to labeled data, the EHRSHOT initiative [Wornow et al., 2023a] also introduced the pretrained sequence model CLMBR [Steinberg et al., 2021], which ranks at or near the top on many EHRSHOT task groups, as well as a public leaderboard with the recorded performance of various methods on EHRSHOT tasks. However, the EHRSHOT leaderboard does not include MoA pipelines that convert tabular records to intermediate clinical summaries before classification.

Motivated by this knowledge gap in understanding how different classes of models perform on predicting various medical outcomes from EHR data, we hereby established a benchmark that directly compares state-of-the-art count-based learners (including GBM and the prior-data-fitted TabPFN [Hollmann et al., 2022], a strong small-data tabular foundation model) with a MoA LLM architecture for structured EHR prediction on EHRSHOT. Our paper contributes through the following:

- Head-to-head comparisons across three methodology categories: (i) count-based features with GBM and with the TabPFN foundation model, (ii) CLMBR (transformer-based), and (iii) our MoA approach;
- Introduction of a MoA baseline and its contribution to the EHRSHOT leaderboard;
- Quantitative analyses to help understand the behavior of the introduced MoA approach, demonstrating that it produces intermediate summaries adapted to the task at hand.

Figure 1 presents the overall study design and the model schema.

2 Methods

In this section, we first introduce the construction of cohorts. We then describe in detail the three modeling categories used in our benchmark: (i) count-based models, which transform OMOP concepts into aggregated feature arrays for traditional machine-learning classifiers; (ii) a pretrained

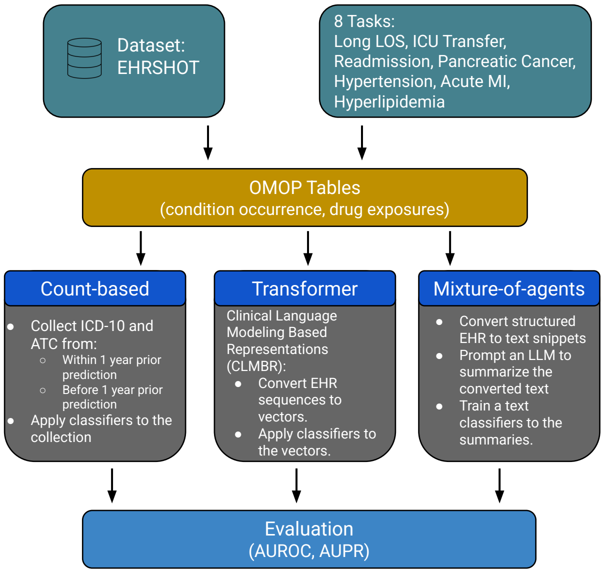

Dataset:
EHRSHOT
8 Tasks:
Long LOS, ICU Transfer,
Readmission, Pancreatic Cancer,
Hypertension, Acute MI,
Hyperlipidemia
OMOP Tables
(condition occurrence, drug exposures)
Count-based
Collect ICD-10 and ATC from:
Within 1 year prior prediction
Before 1 year prior prediction
Apply classifiers to the collection
Transformer
Clinical Language Modeling Based Representations (CLMBR):
Convert EHR sequences to vectors.
Apply classifiers to the vectors.
Mixture-of-agents
Convert structured EHR to text snippets
Prompt an LLM to summarize the converted text
Train a text classifiers to the summaries.
Evaluation
(AUROC, AUPR)

Figure 1: **Study design and benchmarking pipeline**. We evaluated three modeling categories on the EHRSHOT benchmark dataset, with OMOP-standardized cohorts and eight tasks: Long Length of Stay (LOS), ICU Transfer, Readmission, Pancreatic Cancer, Hypertension, Acute Myocardial Infarction (MI), and Hyperlipidemia. From the OMOP tables, we construct: (i) a count-based tabular pipeline that aggregates ICD-10 and ATC codes over a 1-year look-back window and trains strong tabular models (LightGBM and TabPFN); (ii) a pretrained sequential model (CLMBR) that tokenizes time-stamped OMOP concepts and learns vector representations for downstream prediction, introduced with the EHRSHOT dataset; and (iii) a MoA LLM pipeline that converts each patient's longitudinal record into a concise natural-language summary before classification. All models use the same cohorts, prediction windows, and splits; performance is compared on held-out data.

transformer model (CLMBR) that learns sequential representations of longitudinal patient records, introduced together with the EHRSHOT dataset; and (iii) a MoA LLM pipeline that converts tabular histories into natural-language summaries before classification.

2.1 Cohort construction

All analyses were conducted on the publicly available EHRSHOT benchmark, using the predefined training, validation, and test splits provided with the dataset (see Supplementary Table 1 for cohort characteristics). Labels and prediction windows followed the definitions established in Wornow et al. [2023a]. We benchmarked the three described methodology categories for eight different tasks, including three operational outcomes and five new diagnostic prediction outcomes. The operational outcomes include: Long Length of Stay (predict whether a patient's total length of stay during a hospital visit will be at least 7 days), Readmission (predict whether a patient will be re-admitted to the hospital within 30 days after being discharged from a visit), and ICU Transfer (predict whether a patient will be transferred to the ICU during a hospital visit). The new diagnosis tasks are to predict the first diagnosis of one of these diseases within the following year after a hospital visit: pancreatic cancer, hypertension, acute myocardial infarction, hyperlipidemia, and lupus. As EHRSHOT assigns labels per hospital visit, a single patient with multiple hospital visits may contribute multiple labels. To limit bias from repeated measurements, we report the performance for predicting both the earliest assigned label per patient, as well as the latest label (earlier labels are not used to predict future labels from the same patient, following OHDSI's cohort-defining guidelines [OHDSI, 2019]).

2.2 Problem setup

We consider supervised prediction from longitudinal EHR data. For each patient $i \in \{1, \ldots, N\}$ in a dataset of N total patients, we fix a prediction time τ_i and collect all structured events prior to τ_i. $S_i = \{(c_{ij}, t_{ij})\}_{j=1}^{T_i}$ therefore represents the event stream of patient i [McDermott et al., 2023], where $c_{ij} \in \mathcal{C}$ are OMOP concept IDs and $t_{ij} < \tau_i$ are the corresponding time-stamps of these events. The binary task-specific label is $y_i \in \{1, \ldots, K\} \in \{0, 1\}$.

2.3 Count-based models

To construct tabular representations, we apply a map $m : \mathcal{C} \to \mathcal{G}$ to aggregate fine-grained concepts into ICD-10/ATC section categories. For conditions, concept identifiers were first mapped to their corresponding ICD-10 codes and then rolled up into higher-level diagnostic categories (*e.g.*, "I20-I25: Ischemic heart diseases"). For drug exposures, drug identifiers were first mapped to ATC codes and subsequently aggregated at the pharmacological section level (*e.g.*, "C07: Beta blocking agents"). This ontology-based roll-up strategy reduces feature sparsity and improves performance, as has been shown in prior EHR-based data competitions [Bergquist et al., 2023, 2024].

We define two look-back windows relative to τ_i: $\mathcal{W}_{\text{recent}} = [\tau_i - 365 \text{ days}, \tau_i)$ and $\mathcal{W}_{\text{history}} = (-\infty, \tau_i - 365 \text{ days})$. This time-stratified approach allows us to capture both the recent, as well as the long-term history of diagnoses and medication exposure.

For category $g \in \mathcal{G}$ and window $w \in \{\text{recent}, \text{history}\}$, the count-based feature is

$$n_{ig}^{(w)} = \sum_{j=1}^{T_i} \mathbf{1}[m(c_{ij}) = g \ \wedge \ t_{ij} \in \mathcal{W}_w] .$$

The count-based feature representation for patient i is

$$\mathbf{x}_{i,count} = \left[n_{ig}^{(\text{recent})}, n_{ig}^{(\text{history})} \right]_{g \in \mathcal{G}} \in \mathbb{R}^{2|\mathcal{G}|}$$

In our experiment, $|\mathcal{G}| =$ Total number of ICD-10 and ATC sections $= 388$.

Given the count-based feature representation $\mathbf{x}_{i,count}$ for each patient i, we pass them to two representative tabular models: LightGBM [Ke et al., 2017] and TabPFN [Hollmann et al., 2022]. LightGBM is a gradient boosting framework that builds an ensemble of shallow decision trees, where each tree focuses on correcting the mistakes of the previous ones. It is widely used in structured data competitions and clinical prediction tasks, where it achieved outstanding performance [Bergquist et al., 2023, 2024]. TabPFN is a recently developed foundation model for tabular data. Instead of training a new model for each dataset, TabPFN has been pretrained on millions of synthetic classification problems and produces a learned Bayes predictive distribution in a single forward pass conditioned on the entire training data (features and labels) and the test features, yielding the posterior predictive label distribution without task-specific retraining. TabPFN has been reported to outperform boosting models on small- to medium-sized tabular datasets [Hollmann et al., 2022].

2.4 Pretrained transformers

We also evaluated CLMBR (Clinical Language Model-Based Representations) [Steinberg et al., 2021, Guo et al., 2024], a pretrained transformer model designed for structured EHR data and introduced together with the EHRSHOT dataset. Patient histories were organized as longitudinal sequences of timestamped medical events, with each medical concept (*e.g.*, diagnoses and medications) represented as a token. CLMBR learns contextual embeddings of these sequences using a masked language modeling objective, allowing the model to capture temporal dependencies and co-occurrence patterns among medical concepts. The learned patient representations can then be used as input to lightweight classifiers for downstream prediction tasks.

Formally, given the original event stream $S_i = \{(c_{ij}, t_{ij})\}_{j=1}^{T_i}$ with $c_{ij} \in \mathcal{C}$ (no roll-up), the CLMBR yields contextual states $\{\mathbf{h}_{ij}\}$; CLMBR uses a transformer f_θ to form a fixed-length patient representation

$$\mathbf{r}_i = f_\theta(S_i).$$

This vector $\mathbf{r}_i$ is then passed to a downstream classifier. We used LightGBM and TabPFN as classifier, trained on $\{(\mathbf{r}_i, y_i)\}$ in the training set, and evaluated on the test set.

2.5 Mixture-of-agents pipelines

The MoA pipeline adapts LLMs to structured EHR by introducing a collaborative multi-agent workflow. Let M_i denote the plain-text serialization of S_i, formatted as a time-ordered sequence of *(medical concept name, age at event)* pairs. In this setting, a summarizer agent $\mathcal{A}$ first converts M_i into readable and concise textual summaries.

$$U_i \;=\; \mathcal{A}(M_i),$$

which is then consumed by a predictor $\mathcal{P}$ to produce class predictions per patient:

$$\hat{p}_i \;=\; \mathcal{P}\big(U_i\big).$$

In our implementation, we adopted Llama-3-8b [AI@Meta, 2024], reported to perform well with structured EHR data [Chen et al., 2024, Gao et al., 2025], and Qwen2.5-14B-Instruct, which has demonstrated stronger reasoning ability in long-context settings [QwenTeam, 2024]. Both models were prompted to summarize raw tabular data into clinically meaningful structured text (prompts in Supplementary Section P1 to P8), and the resulting summaries served as input for the downstream classifiers. We used BGE-large-en-v1.5 [Xiao et al., 2023] and ClinicalBERT [Alsentzer et al., 2019] as classifiers due to their outstanding performance on clinical text representation tasks [Myers et al., 2025, Gao et al., 2024a]. In particular, we extracted the final hidden state corresponding to the [CLS] token and passed it through a single feed-forward layer, yielding the logit predictions for classification.

3 Results

In this section, we report metrics for the assessed methods across eight tasks, highlight the MoA's intermediate summary quality, and examine medical concepts contributing to prediction.

3.1 Discrimination performance

Figure 2 shows the results of our evaluations, and Supplementary Table 2 and Supplementary Table 3 report the corresponding numerical results. Across tasks, wins concentrate in the count-based and MoA methodologies. For clarity, Figure 2 only reports a single winning method within each of the three methodology categories: (i) count-based features with LightGBM *vs.* count-based features with TabPFN, (ii) CLMBR embeddings with LightGBM *vs.* CLMBR embeddings with TabPFN, and (iii) the best LLM choice within the MoA pipeline. The results of other variants for each methodology category are shown in Supplementary Table 4 to Supplementary Table 15. In the earliest-label setting, count-based methods take 5/8 AUROC wins (MoA: 3/8) and 6/8 AUPR wins (MoA: 2/8). In the latest-label setting and excluding readmission here due to its abnormally low prevalence (see Supplementary Table 1), count-based methods lead AUROC with 4/7 wins (MoA: 2/7; CLMBR: 1/7) and split AUPR (3/7) with MoA (3/7) and CLMBR (1/7). In short, most wins are shared by count-based and MoA methods, with count-based finishing slightly ahead overall and consistently outperforming CLMBR. Therefore, count-based methods remain a strong baseline for structured EHR benchmarking, while MoA methods offer gains on specific tasks.

3.2 Quality of the intermediate summary

In the MoA pipeline, an LLM agent is prompted to produce intermediate summaries relevant to the prediction target. To compare the differences in information content between the original text (raw structured EHR) and the intermediate summary, we first used MedCat [Kraljevic et al., 2021] to extract concepts as ICD-10 chapters from both texts, computed the percentages of mentions of each ICD-10 chapter, and plotted the percentage change (summary minus original EHR) in Figure 3. Positive values indicate that the summary amplifies the chapter, while negative values indicate that information from the chapter is down-weighted. Chapters which move in opposite directions depending on the task highlight how the MoA pipeline adapts to the clinical prediction target. For example, for pancreatic cancer prediction, the summary amplifies Endocrine/Metabolic and Mental/Behavioral

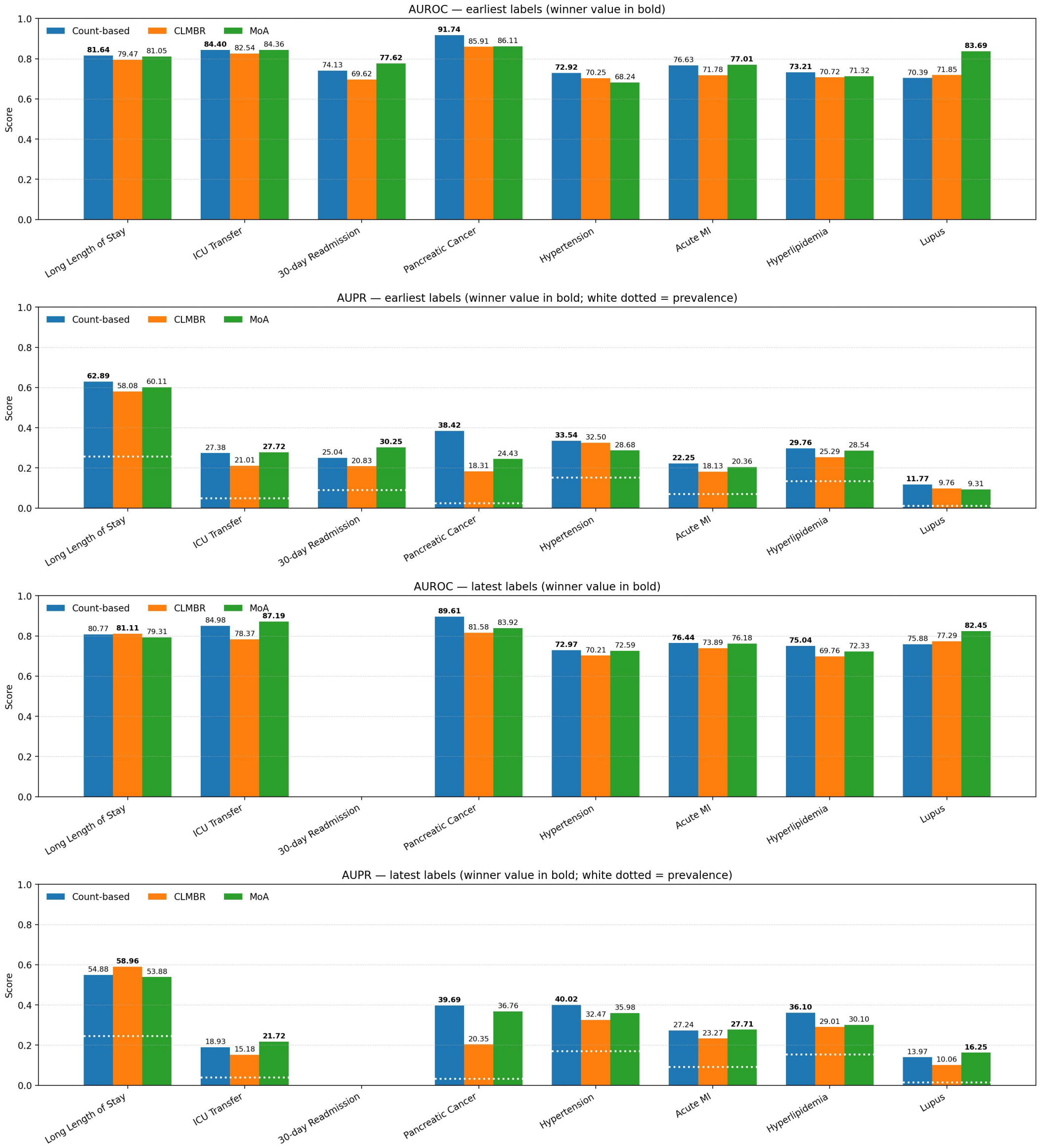


Figure 2: **Performance of the benchmarked models across the eight prediction tasks in EHRSHOT under two label definitions (earliest and latest).** Top panels show AUROC; bottom panels show AUPR. Bolded bars indicate the best-performing model for each task; the white dotted line in each AUPR panel marks the outcome prevalence. Across tasks, wins are shared mainly between the count-based methods and the MoA pipeline, with count-based methods holding a slight overall edge and generally outperforming CLMBR.

content, while down-weighting it for ICU transfer prediction. Conversely, the ICU-transfer summary elevates Nervous system mentions, while de-emphasizing them in the pancreatic cancer summary. This assessment directly shows that intermediate summaries help steer the input information content towards the prediction task.

3.3 A case study for the Mixture-of-agent pipeline

The case study in Figure 4 shows how the MoA pipeline summarizes a patient's recent history for pancreatic cancer prediction. The structured EHR data are first converted into "EVENT at AGE" snippets and passed to the summarizer (Qwen2.5-14B-Instruct). The prompt requires a JSON output

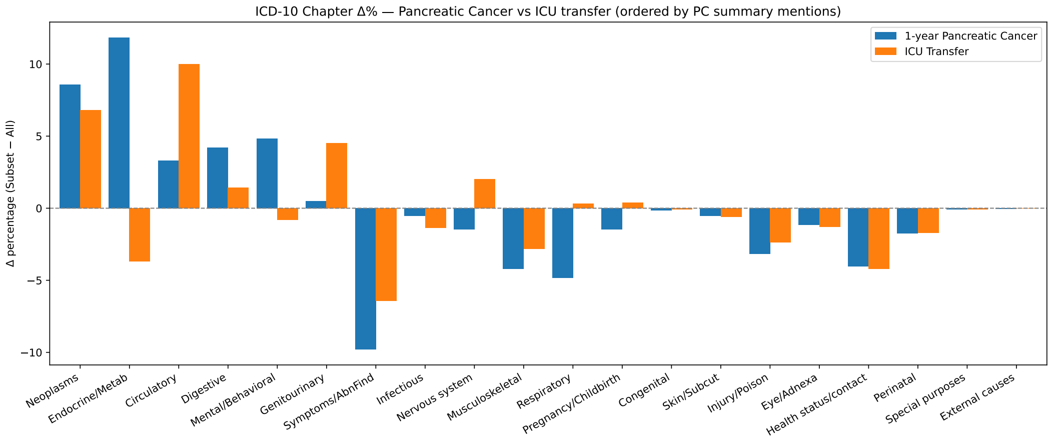


Figure 3: **Task-adaptive emphasis in MoA intermediate summaries.** Bars show the change in ICD-10 chapter mention percentages between the MoA summary and the original structured EHR ($\Delta\% =$ percentage mentions in summary $-$ percentage mentions in original), computed with MedCat concept extraction. Positive values indicate chapters the summary amplifies; negative values indicate down-weighting. Chapters with opposite sign directions illustrate adaptation to the prediction task. For example, *Endocrine/Metabolic* is amplified for pancreatic cancer but reduced for ICU transfer, whereas *Nervous system* is boosted for ICU transfer and de-emphasized for pancreatic cancer.

with a risk category/score, positive and negative drivers, as well as a short justification. In this example, given the input EHR text (*e.g.*, “obstruction of bile duct,” “cholangitis,” “abdominal pain", "multiple analgesics"), Qwen produces an intermediate summary that labels the case as Moderate risk with a score of 0.5, cites biliary obstruction and cholangitis as positive drivers, and briefly explains its underlying rationale. When ingested by the downstream text classifier, this evidence yields towards a positive prediction, which is consistent with the true positive label for this patient.

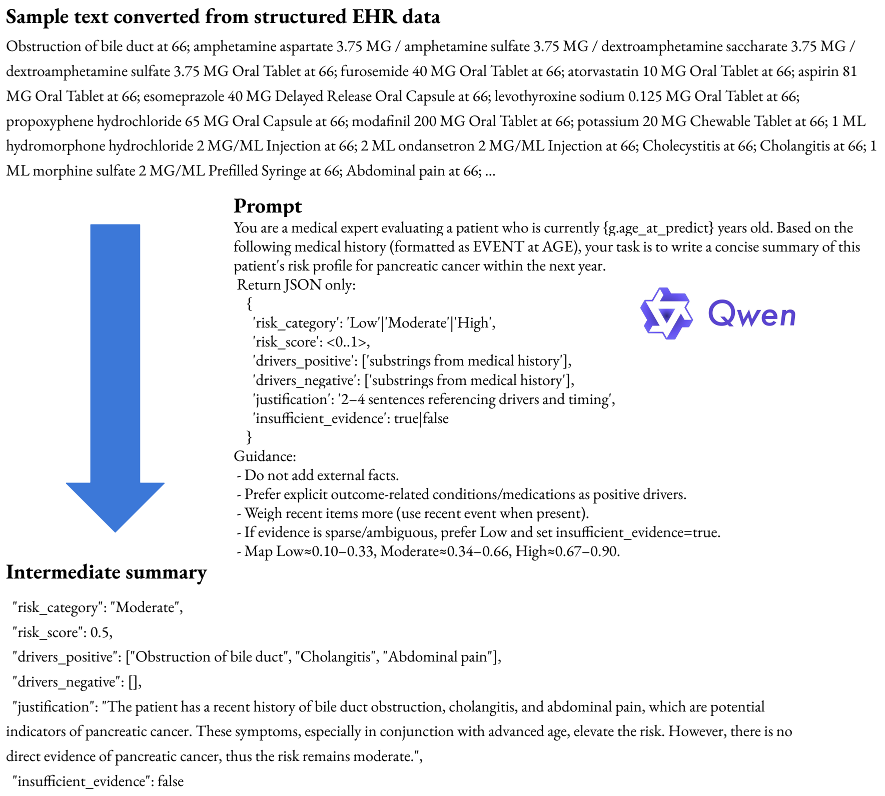
3.4 SHAP values of LightGBM

To better understand the risk factors driving model predictions, we examined the SHAP values [Lundberg and Lee, 2017] from the LightGBM model features (Figure 5). All the top five contributing factors came from the most recent time bin (within one year before prediction time), suggesting that events in the immediate history carry the strongest signal for prediction, while factors from longer look-back windows may not have a substantial influence on the outcomes.

For long length-of-stay, hematologic disorders such as aplastic anemia, complications following transfusion, and frequent health care encounters were among the strongest predictors, consistent with patients requiring prolonged and complex hospital care. In the readmission task, medication use and cancer-related diagnoses showed high importance. The model highlighted anti-inflammatory drugs, systemic antihistamines, and throat preparations, along with malignant neoplasms, suggesting that both chronic disease burden and frequent medication use contribute to higher readmission risk. For pancreatic cancer prediction, diagnoses related to hepatobiliary and digestive disorders were most influential, which is expected given the biological links between gallbladder, biliary tract, and pancreatic diseases. Symptoms such as abdominal complaints and general digestive neoplasms also appeared as strong predictors. In acute myocardial infarction, cardiovascular conditions were the most prominent. Ischemic heart disease and beta-blocker prescriptions were the leading contributing factors, along with kidney disease, lipid-lowering therapy, and muscle relaxants, reflecting established risk factors and common co-treatments in patients at high risk of acute cardiac events.

4 Discussion

In this study, we benchmarked three categories of methods for clinical prediction using structured EHR data: (i) count-based models using ontology roll-ups and time-binned features, (ii) a pretrained sequential transformer (CLMBR), and (iii) a mixture-of-agents LLM pipeline that summarizes



Sample text converted from structured EHR data

Obstruction of bile duct at 66; amphetamine aspartate 3.75 MG / amphetamine sulfate 3.75 MG / dextroamphetamine saccharate 3.75 MG / dextroamphetamine sulfate 3.75 MG Oral Tablet at 66; furosemide 40 MG Oral Tablet at 66; atorvastatin 10 MG Oral Tablet at 66; aspirin 81 MG Oral Tablet at 66; esomeprazole 40 MG Delayed Release Oral Capsule at 66; levothyroxine sodium 0.125 MG Oral Tablet at 66; propoxyphene hydrochloride 65 MG Oral Capsule at 66; modafinil 200 MG Oral Tablet at 66; potassium 20 MG Chewable Tablet at 66; 1 ML hydromorphone hydrochloride 2 MG/ML Injection at 66; 2 ML ondansetron 2 MG/ML Injection at 66; Cholecystitis at 66; Cholangitis at 66; 1 ML morphine sulfate 2 MG/ML Prefilled Syringe at 66; Abdominal pain at 66; ...

Prompt

```
You are a medical expert evaluating a patient who is currently {g.age_at_predict} years old. Based on the
following medical history (formatted as EVENT at AGE), your task is to write a concise summary of this
patient's risk profile for pancreatic cancer within the next year.
 Return JSON only:
   {
     'risk_category': 'Low'|'Moderate'|'High',
     'risk_score': <0..1>,
     'drivers_positive': ['substrings from medical history'],
     'drivers_negative': ['substrings from medical history'],
     'justification': '2–4 sentences referencing drivers and timing',
     'insufficient_evidence': true|false
   }
Guidance:
 - Do not add external facts.
 - Prefer explicit outcome-related conditions/medications as positive drivers.
 - Weigh recent items more (use recent event when present).
 - If evidence is sparse/ambiguous, prefer Low and set insufficient_evidence=true.
 - Map Low≈0.10–0.33, Moderate≈0.34–0.66, High≈0.67–0.90.
```

Intermediate summary

```
 "risk_category": "Moderate",
 "risk_score": 0.5,
 "drivers_positive": ["Obstruction of bile duct", "Cholangitis", "Abdominal pain"],
 "drivers_negative": [],
 "justification": "The patient has a recent history of bile duct obstruction, cholangitis, and abdominal pain, which are potential
indicators of pancreatic cancer. These symptoms, especially in conjunction with advanced age, elevate the risk. However, there is no
direct evidence of pancreatic cancer, thus the risk remains moderate.",
 "insufficient_evidence": false
```

Figure 4: **Case study: LLM-generated intermediate summary from structured EHR in the MoA pipeline.** Structured EHR is first converted into an "EVENT at AGE" text sequence and further passed to Qwen with a task prompt to predict the risk of a new pancreatic cancer diagnosis in the next year. The model is instructed to return a constrained JSON object containing `risk_category` (Low/Moderate/High), `risk_score` $[0, 1]$, `drivers_positive/drivers_negative` from the input, a 2 to 4 sentence `justification` that references evidence and timing, and an `insufficient_evidence` flag. In this example, Qwen highlights biliary obstruction, cholangitis, and abdominal pain, and returns a *Moderate* risk with score 0.5, while also yielding an interpretable intermediate summary used for downstream text classifiers in the MoA pipeline.

tabular histories before classification. We evaluated eight binary prediction tasks on the EHRSHOT dataset: Long Length of Stay, ICU Transfer, Readmission, Pancreatic Cancer, Hypertension, Acute Myocardial Infarction, and Hyperlipidemia. Across tasks, victories are split primarily between the count-based models and the MoA pipeline, with count-based methods holding a slight overall edge and generally outperforming CLMBR. This shows how, despite recent advances in multi-agent LLM pipelines, count-based approaches remain a strong, practical choice for structured EHR prediction.

To our knowledge, this is the first benchmark to compare count-based models head-to-head with a MoA pipeline on structured EHR data. Although MoA pipelines have demonstrated superior performance compared to single LLM pipelines, previous studies have largely only evaluated single LLM baselines on structured EHR data [Chen et al., 2024]. To fill in this knowledge gap, we hereby adapted and evaluated a MoA pipeline designed specifically for tabular data. We also equipped the count-based approach with an up-to-date tabular foundation model (TabPFN) in addition to a strong GBM baseline (LightGBM), offering a fair and contemporary comparison against sequential and LLM-based methods under the same cohorts and splits.

Although predicting the latest label for a patient provides richer contextual information compared to using the earliest label (Supplementary Table 1), overall performance patterns remain largely unchanged. In the latest label setting, count-based models and the MoA pipeline still dominated the evaluation tasks. The only exception is that in the Long Length of Stay task with the latest labels, CLMBR shows an outstanding improvement and takes the lead. This suggests that count-

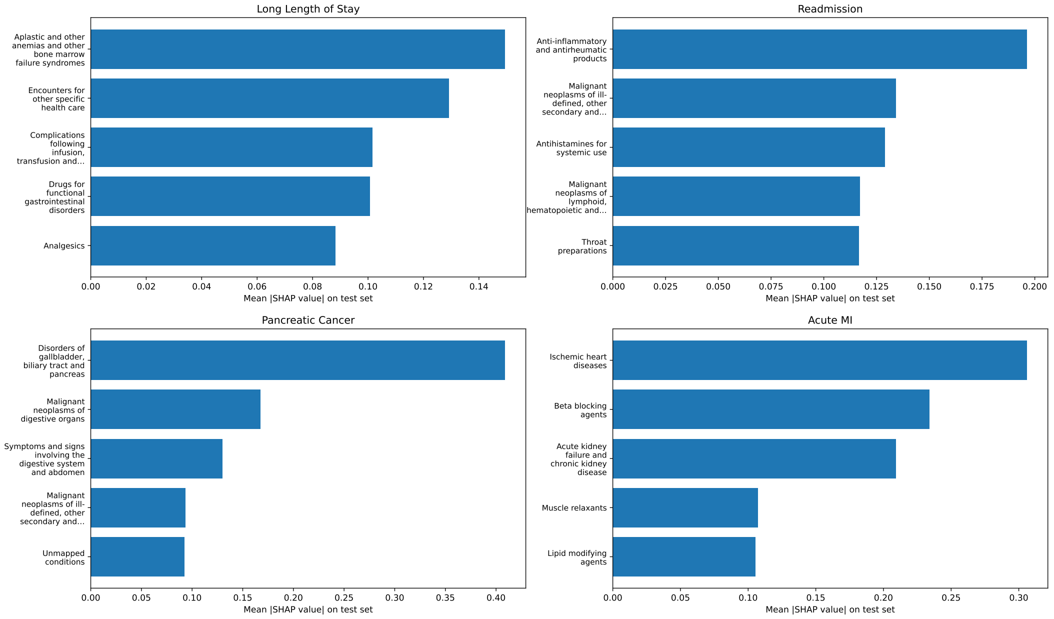


Figure 5: **SHAP analysis of LightGBM models for four prediction tasks**. For each task, we show the mean absolute SHAP values on the test set for the top five most influential features. All of the contributing factors come from the most recent time bin (within one year before the prediction time). This detail is not explicitly labeled in the figure, as the concept names are already lengthy.

based models and the MoA pipeline are comparatively robust to context richness, whereas the transformer-based CLMBR may benefit more from more contextual details.

Our interpretability analysis for the count-based methods shows that the top five SHAP contributors come from the most recent time bin (within one year before prediction time). This suggests that immediate clinical history dominates model decisions, while long-past events contribute with less signal. One implication of these findings is that fine-grained temporal ordering and long-range dependencies may be less impactful for the prediction tasks assessed here in the EHRSHOT dataset.

We performed a series of ablation studies on the MoA pipelines and the count-based baselines. For the MoA pipeline, we varied both the summarizing agent (Llama-3-8b-instruct *vs.* Qwen2.5-14B-Instruct) as well as the classifier (ClinicalBERT *vs.* BGE-large-en-v1.5), and we also evaluated prompts that produce unstructured free-text summaries (Supplementary Section P9 to P16) rather than the constrained JSON. As reported in Supplementary Table 8 to Supplementary Table 15, the combination Qwen2.5-14B-Instruct + BGE-large-en-v1.5 with structured prompts yielded the strongest performance among all MoA variants, underscoring that agent choice and prompt format substantially affect the performance of MoA pipelines. For the count-based models, we compared pipelines with *vs.* without ontology roll-up; in every task, ontology roll-up improved performance over the non-rolled counterparts.

Several limitations warrant mention. Our analysis is based on a single-institution dataset and a limited set of outcomes, which constrains external validity. We also focused only on diagnoses and medications training data. Further integration of labs, vitals, and procedures in model training might improve performance. Future work should expand across multiple institutions, include a broader set of clinical information resources, and test additional LLM agents and summarization strategies.

Overall, our results reaffirm the strength of count-based modeling for structured EHR prediction, even in the era of LLMs. With up-to-date LLM agents, the MoA pipeline achieves superior performance in a subset of the evaluation tasks. These findings highlight that traditional methods still provide strong baselines, but generative and transformer-based approaches open new possibilities under specific clinical scenarios. As clinical AI moves toward foundation-model paradigms [Moor et al., 2023, Wornow et al., 2023b], it will be worthwhile to systematically evaluate how these approaches leverage richer histories, scale across diverse health systems, and integrate into decision-making workflows.

References

AI@Meta. Llama 3 model card. 2024. URL `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

Anmol Arora, Joseph E Alderman, Joanne Palmer, Shaswath Ganapathi, Elinor Laws, Melissa D Mccradden, Lauren Oakden-Rayner, Stephen R Pfohl, Marzyeh Ghassemi, Francis McKay, et al. The value of standards for health datasets in artificial intelligence-based applications. *Nature medicine*, 29(11):2929–2938, 2023.

Timothy Bergquist, Thomas Schaffter, Yao Yan, Thomas Yu, Justin Prosser, Jifan Gao, Guanhua Chen, Łukasz Charzewski, Zofia Nawalany, Ivan Brugere, et al. Evaluation of crowdsourced mortality prediction models as a framework for assessing artificial intelligence in medicine. *Journal of the American Medical Informatics Association*, 31(1):35–44, 2023.

Timothy Bergquist, Johanna Loomba, Emily Pfaff, Fangfang Xia, Zixuan Zhao, Yitan Zhu, Elliot Mitchell, Biplab Bhattacharya, Gaurav Shetty, Tamanna Munia, et al. Crowd-sourced machine learning prediction of long covid using data from the national covid cohort collaborative. *EBioMedicine*, 108, 2024.

Katherine E Brown, Chao Yan, Zhuohang Li, Xinmeng Zhang, Benjamin X Collins, You Chen, Ellen Wright Clayton, Murat Kantarcioglu, Yevgeniy Vorobeychik, and Bradley A Malin. Large language models are less effective at clinical prediction tasks than locally trained machine learning models. *Journal of the American Medical Informatics Association*, 32(5):811–822, 2025.

Canyu Chen, Jian Yu, Shan Chen, Che Liu, Zhongwei Wan, Danielle Bitterman, Fei Wang, and Kai Shu. Clinicalbench: Can llms beat traditional ml models in clinical prediction? *arXiv preprint arXiv:2411.06469*, 2024.

Peter F Edemekong, Pavan Annamaraju, and Micelle J Haydel. Health insurance portability and accountability act. 2018.

Aya Gamal, Sherif Barakat, and Amira Rezk. Standardized electronic health record data modeling and persistence: A comparative review. *Journal of biomedical informatics*, 114:103670, 2021.

Jifan Gao, Guanhua Chen, Ann P O'Rourke, John Caskey, Kyle A Carey, Madeline Oguss, Anne Stey, Dmitriy Dligach, Timothy Miller, Anoop Mayampurath, et al. Automated stratification of trauma injury severity across multiple body regions using multi-modal, multi-class machine learning models. *Journal of the American Medical Informatics Association*, 31(6):1291–1302, 2024a.

Jifan Gao, Philip Mar, Zheng-Zheng Tang, and Guanhua Chen. Fair prediction of 2-year stroke risk in patients with atrial fibrillation. *Journal of the American Medical Informatics Association*, 31 (12):2820–2828, 2024b.

Jifan Gao, Mahmudur Rahman, John Caskey, Madeline Oguss, Ann O'Rourke, Randy Brown, Anne Stey, Anoop Mayampurath, Matthew M Churpek, Guanhua Chen, et al. Moma: A mixture-of-multimodal-agents architecture for enhancing clinical prediction modelling. *arXiv preprint arXiv:2508.05492*, 2025.

Lin Lawrence Guo, Jason Fries, Ethan Steinberg, Scott Lanyon Fleming, Keith Morse, Catherine Aftandilian, Jose Posada, Nigam Shah, and Lillian Sung. A multi-center study on the adaptability of a shared foundation model for electronic health records. *NPJ Digital Medicine*, 7(1):171, 2024.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*, 2022.

George Hripcsak, Jon D Duke, Nigam H Shah, Christian G Reich, Vojtech Huser, Martijn J Schuemie, Marc A Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R Rijnbeek, et al. Observational health data sciences and informatics (ohdsi): opportunities for observational researchers. In *MEDINFO 2015: eHealth-enabled Health*, pages 574–578. IOS Press, 2015.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

Sara Khalid, Cynthia Yang, Clair Blacketer, Talita Duarte-Salles, Sergio Fernández-Bertolín, Chungsoo Kim, Rae Woong Park, Jimyung Park, Martijn J Schuemie, Anthony G Sena, et al. A standardized analytics pipeline for reliable and rapid development and validation of prediction models using observational health data. *Computer Methods and Programs in Biomedicine*, 211: 106394, 2021.

Junmo Kim, Namkyeong Lee, Jiwon Kim, and Kwangsoo Kim. Medrep: Medical concept representation for general electronic health record foundation models. *arXiv preprint arXiv:2504.08329*, 2025.

Matthias Kirchler, Matteo Ferro, Veronica Lorenzini, Christoph Lippert, Andrea Ganna, et al. Large language models improve transferability of electronic health record-based predictions across countries and coding systems. *medRxiv*, 2025.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, et al. Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit. *Artificial intelligence in medicine*, 117:102083, 2021.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Matthew McDermott, Bret Nestor, Peniel Argaw, and Isaac S Kohane. Event stream gpt: a data pre-processing and modeling library for generative, pre-trained transformers over continuous-time sequences of complex events. *Advances in Neural Information Processing Systems*, 36: 24322–24334, 2023.

Duncan McElfresh, Sujay Khandagale, Jonathan Valverde, Vishak Prasad C, Ganesh Ramakrishnan, Micah Goldblum, and Colin White. When do neural nets outperform boosted trees on tabular data? *Advances in Neural Information Processing Systems*, 36:76336–76369, 2023.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

Skatje Myers, Dmitriy Dligach, Timothy A Miller, Samantha Barr, Yanjun Gao, Matthew Churpek, Anoop Mayampurath, and Majid Afshar. Evaluating retrieval-augmented generation vs. long-context input for clinical reasoning over ehrs. *arXiv preprint arXiv:2508.14817*, 2025.

OHDSI. *The Book of OHDSI. Chapter 10: Defining Cohorts*. 2019. URL `https://ohdsi.github.io/TheBookOfOhdsi/TheBookOfOhdsi.pdf`.

QwenTeam. Qwen2.5: A party of foundation models, September 2024. URL `https://qwenlm.github.io/blog/qwen2.5/`.

Jenna M Reps, Martijn J Schuemie, Marc A Suchard, Patrick B Ryan, and Peter R Rijnbeek. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *Journal of the American Medical Informatics Association*, 25(8):969–975, 2018.

Artem Shmatko, Alexander Wolfgang Jung, Kumar Gaurav, Søren Brunak, Laust Mortensen, Ewan Birney, Tom Fitzgerald, and Moritz Gerstung. Learning the natural history of human disease with generative transformers. *MedRxiv*, pages 2024–06, 2024.

Ethan Steinberg, Ken Jung, Jason A Fries, Conor K Corbin, Stephen R Pfohl, and Nigam H Shah. Language models are an effective representation learning technique for electronic health record data. *Journal of biomedical informatics*, 113:103637, 2021.

Ethan Steinberg, Jason Fries, Yizhe Xu, and Nigam Shah. Motor: A time-to-event foundation model for structured medical records. *arXiv preprint arXiv:2301.03150*, 2023.

Maryam Tayefi, Phuong Ngo, Taridzo Chomutare, Hercules Dalianis, Elisa Salvi, Andrius Budrionis, and Fred Godtliebsen. Challenges and opportunities beyond structured data in analysis of electronic health records. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(6):e1549, 2021.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024.

Liwei Wang, Andrew Wen, Sunyang Fu, Xiaoyang Ruan, Ming Huang, Rui Li, Qiuhao Lu, Heather Lyu, Andrew E Williams, and Hongfang Liu. A scoping review of omop cdm adoption for cancer research using real world data. *NPJ Digital Medicine*, 8(1):189, 2025.

Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason Fries, and Nigam Shah. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. *Advances in Neural Information Processing Systems*, 36:67125–67137, 2023a.

Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A Pfeffer, Jason Fries, and Nigam H Shah. The shaky foundations of large language models and foundation models for electronic health records. *npj digital medicine*, 6(1):135, 2023b.

Michael Wornow, Suhana Bedi, Miguel Angel Fuentes Hernandez, Ethan Steinberg, Jason Alan Fries, Christopher Ré, Sanmi Koyejo, and Nigam H Shah. Context clues: Evaluating long context models for clinical prediction tasks on ehrs. *arXiv preprint arXiv:2412.16178*, 2024.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding, 2023.

Supplementary Material

Supplementary Table 1 summarizes the cohorts across the eight prediction tasks.

Table Supplementary Table 1: Cohort characteristic

Task	Long LOS	Readmission	Pancreatic Cancer	Acute MI	ICU transfer	Hypertension	Hyperlipidemia	Lupus
Cohort characteristics								
Encounters, n	3,855	3,718	3,864	3,834	3,617	2,328	2,650	3,864
Age, median (IQR)	58 (40, 69)	58 (43, 69)	58 (41, 69)	58 (41, 69)	58 (41, 69)	50 (35, 64)	51 (35, 65)	58 (41, 69)
Female, n (%)	2,067 (53.6)	1,977 (53.2)	2,071 (53.6)	2,071 (54.0)	1,959 (54.2)	1,416 (60.8)	1,594 (60.2)	2,058 (53.3)
White, n (%)	2,248 (58.3)	2,190 (58.9)	2,251 (58.3)	2,231 (58.2)	2,103 (58.1)	1,355 (58.2)	1,529 (57.7)	2,253 (58.3)
Outcome and context information: earliest (if multiple)								
Prevalence (%)	24.9	8.2	4.6	7.1	4.5	14.3	12.7	2.6
Number of distinct diagnoses, median (IQR)	19 (10, 34)	21 (11, 38)	21 (11, 37)	21 (11, 36)	19 (10, 34)	17 (9, 31)	17 (9, 31)	21 (11, 37)
Number of diagnoses, median (IQR)	31 (15, 67)	44 (18, 99)	43 (18, 97)	43 (18, 95.5)	31 (15, 73)	32 (13, 76)	34 (14, 77)	43 (18, 96)
Number of distinct medications, median (IQR)	21 (10, 35)	37 (22, 56)	36 (21.5, 55)	36 (21, 55)	22 (11, 36)	30 (17, 46.25)	31 (18, 49)	36 (22, 55)
Number of medications, median (IQR)	25 (11, 43)	75 (40, 150)	73 (38, 148)	73 (38, 146)	25 (11, 46.75)	58 (30, 118)	61 (32, 126)	73 (38, 147)
Number of distinct dates, median (IQR)	9 (4, 24)	16 (8, 32)	15 (8, 31)	15 (8, 31)	10 (4, 26)	13 (7, 27)	13 (7, 27)	15 (8, 31)
Outcome and context information: most recent (if multiple)								
Prevalence (%)	24.3	0.5	5.5	9.3	4.6	16.5	15.4	3.2
Number of distinct diagnoses, median (IQR)	28 (14, 55)	32 (15, 60)	30 (15, 57)	29.5 (15, 55)	28 (14, 55)	22 (10, 41.5)	23 (11, 45)	30 (15, 57)
Number of diagnoses, median (IQR)	21 (55, 163)	77 (29, 211)	73 (26, 195)	70 (26, 184.75)	56 (20, 169)	46 (17, 123)	51 (19, 138)	73 (26, 195)
Number of distinct medications, median (IQR)	33 (15, 63)	50 (30, 81)	36 (21.5, 55)	36 (21, 55)	33 (15, 64)	37 (20, 60)	40 (22, 66)	48 (28, 78)
Number of medications, median (IQR)	43 (17, 167)	121 (55, 306)	73 (38, 148)	73 (38, 146)	44 (17, 173)	76 (36, 194)	86.5 (39, 226.25)	113 (51, 283.75)
Number of distinct dates, median (IQR)	19 (6, 58)	27 (11, 69)	23 (10, 59.75)	23 (10, 57)	20 (6, 59)	18 (8, 43)	19 (9, 46)	25 (10, 63)

Supplementary Table 2 shows the numerical benchmark results using the earliest visit as the prediction time.

Table Supplementary Table 2: Earliest labels — best of each group (row-wise winners in **bold**).

	AUROC			AUPR		
Task	Count-based	CLMBR	MoA	Count-based	CLMBR	MoA
Long LOS	**0.8164**	0.7947	0.8105	**0.6289**	0.5808	0.6011
ICU Transfer	**0.8440**	0.8254	0.8436	0.2738	0.2101	**0.2772**
Readmission	0.7413	0.6962	**0.7762**	0.2504	0.2083	**0.3025**
Pancreatic Cancer	**0.9174**	0.8591	0.8611	**0.3842**	0.1831	0.2443
Hypertension	**0.7292**	0.7025	0.6824	**0.3354**	0.3250	0.2868
Acute MI	0.7663	0.7178	**0.7701**	**0.2225**	0.1813	0.2036
Hyperlipidemia	**0.7321**	0.7072	0.7132	**0.2976**	0.2529	0.2854
Lupus	0.7039	0.7185	**0.8369**	**0.1177**	0.0976	0.0931

Supplementary Table 3 shows the numerical benchmark results using the last visit as the prediction time.

Table Supplementary Table 3: Latest labels — best of each group (row-wise winners in **bold**).

	AUROC			AUPR		
Task	Count-based	CLMBR	MoA	Count-based	CLMBR	MoA
Long LOS	0.8077	**0.8111**	0.7931	0.5488	**0.5896**	0.5388
ICU Transfer	0.8498	0.7837	**0.8719**	0.1893	0.1518	**0.2172**
Readmission	N/A	N/A	N/A	N/A	N/A	N/A
Pancreatic Cancer	**0.8961**	0.8158	0.8392	**0.3969**	0.2035	0.3676
Hypertension	**0.7297**	0.7021	0.7259	**0.4002**	0.3247	0.3598
Acute MI	**0.7644**	0.7389	0.7618	0.2724	0.2327	**0.2771**
Hyperlipidemia	**0.7504**	0.6976	0.7233	**0.3610**	0.2901	0.3010
Lupus	0.7588	0.7729	**0.8245**	0.1397	0.1006	**0.1625**

Supplementary Table 4, Supplementary Table 5, Supplementary Table 6, and Supplementary Table 7 show results of count-based and CLMBR variants.

Table Supplementary Table 4: Earliest labels — **AUROC** (mean ± sd) across tasks and methods for count-based and CLMBR variants.

Task	**Count-based + LGBM**	**Count-based + TabPFN**	**CLMBR + LGBM**	**CLMBR + TabPFN**	**MoA**
Long LOS	0.8047 ± 0.0039	**0.8164 ± N/A**	0.7835 ± 0.0072	0.7947 ± N/A	0.8105 ± 0.0257
ICU Transfer	**0.8440 ± 0.0058**	0.8377 ± N/A	0.7456 ± 0.0156	0.8254 ± N/A	0.8436 ± 0.0054
Readmission	0.7413 ± 0.0032	0.7366 ± N/A	0.6962 ± 0.0093	0.6320 ± N/A	**0.7762 ± 0.0176**
Pancreatic Cancer	0.8821 ± 0.0070	**0.9174 ± N/A**	0.8004 ± 0.0087	0.8591 ± N/A	0.8611 ± 0.0150
Hypertension	0.7233 ± 0.0041	**0.7292 ± N/A**	0.6956 ± 0.0105	0.7025 ± N/A	0.6824 ± 0.0256
Acute MI	0.7663 ± 0.0119	0.7590 ± N/A	0.7093 ± 0.0161	0.7178 ± N/A	**0.7701 ± 0.0091**
Hyperlipidemia	0.7289 ± 0.0044	**0.7321 ± N/A**	0.6592 ± 0.0041	0.7072 ± N/A	0.7132 ± 0.0140
Lupus	0.6689 ± 0.0100	0.7039 ± N/A	0.7079 ± 0.0107	0.7185 ± N/A	**0.8369 ± 0.0129**

Table Supplementary Table 5: Earliest labels — **AUPR** (mean ± sd) across tasks and methods for count-based and CLMBR variants..

Task	**Count-based + LGBM**	**Count-based + TabPFN**	**CLMBR + LGBM**	**CLMBR + TabPFN**	**MoA**
Long LOS	0.6172 ± 0.0116	**0.6289 ± N/A**	0.5673 ± 0.0137	0.5808 ± N/A	0.6011 ± 0.0533
ICU Transfer	0.2381 ± 0.0220	0.2738 ± N/A	0.1519 ± 0.0206	0.2101 ± N/A	**0.2772 ± 0.0089**
Readmission	0.2504 ± 0.0203	0.2449 ± N/A	0.2083 ± 0.0027	0.1603 ± N/A	**0.3025 ± 0.0225**
Pancreatic Cancer	0.3144 ± 0.0463	0.3842 ± **N/A**	0.1722 ± 0.0313	0.1831 ± N/A	0.2443 ± 0.0636
Hypertension	0.3241 ± 0.0070	**0.3354 ± N/A**	0.3188 ± 0.0064	0.3250 ± N/A	0.2868 ± 0.0327
Acute MI	0.2113 ± 0.0052	**0.2225 ± N/A**	0.1735 ± 0.0045	0.1813 ± N/A	0.2036 ± 0.0169
Hyperlipidemia	**0.2976 ± 0.0124**	0.2955 ± N/A	0.2029 ± 0.0077	0.2529 ± N/A	0.2854 ± 0.0231
Lupus	0.0866 ± 0.0252	**0.1177 ± N/A**	0.0976 ± 0.0469	0.0608 ± N/A	0.0931 ± 0.0418

Supplementary Table 8, Supplementary Table 9, Supplementary Table 10, Supplementary Table 11 present ablation studies on MoA architectures.

Supplementary Table 12, Supplementary Table 13, Supplementary Table 14, Supplementary Table 15 show the comparison of count-based models with *vs.* without ontology roll-up.

Table Supplementary Table 6: Latest labels — **AUROC** (mean ± sd) for count-based and CLMBR variants.

Task	Count-based + LGBM	Count-based + TabPFN	CLMBR + LGBM	CLMBR + TabPFN	MoA
Long LOS	0.7918 ± 0.0015	0.8077 ± N/A	0.7851 ± 0.0037	**0.8111 ± N/A**	0.7931 ± 0.0030
ICU Transfer	0.8213 ± 0.0036	0.8498 ± N/A	0.7837 ± 0.0090	0.7680 ± N/A	**0.8719 ± 0.0169**
Readmission	N/A ± N/A	N/A ± N/A	N/A ± N/A	N/A ± N/A	N/A ± N/A
Pancreatic Cancer	0.8609 ± 0.0113	**0.8961 ± N/A**	0.7867 ± 0.0039	0.8158 ± N/A	0.8392 ± 0.0058
Hypertension	**0.7297 ± 0.0025**	0.7142 ± N/A	0.6966 ± 0.0033	0.7021 ± N/A	0.7259 ± 0.0066
Acute MI	**0.7644 ± 0.0083**	0.7500 ± N/A	0.7389 ± 0.0056	0.7015 ± N/A	0.7618 ± 0.0088
Hyperlipidemia	0.7261 ± 0.0048	**0.7504 ± N/A**	0.6795 ± 0.0073	0.6976 ± N/A	0.7233 ± 0.0147
Lupus	0.7027 ± 0.0022	0.7588 ± N/A	0.7729 ± 0.0371	0.7660 ± N/A	**0.8245 ± 0.0291**

Table Supplementary Table 7: Latest labels — **AUPR** (mean ± sd) for count-based and CLMBR variants.

Task	Count-based + LGBM	Count-based + TabPFN	CLMBR + LGBM	CLMBR + TabPFN	MoA
Long LOS	0.5342 ± 0.0037	0.5488 ± N/A	0.5458 ± 0.0113	**0.5896 ± N/A**	0.5388 ± 0.0080
ICU Transfer	0.1533 ± 0.0078	0.1893 ± N/A	0.1518 ± 0.0244	0.1057 ± N/A	**0.2172 ± 0.0157**
Readmission	N/A ± N/A	N/A ± N/A	N/A ± N/A	N/A ± N/A	N/A ± N/A
Pancreatic Cancer	0.3418 ± 0.0392	**0.3969 ± N/A**	0.1610 ± 0.0013	0.2035 ± N/A	0.3676 ± 0.0130
Hypertension	**0.4002 ± 0.0084**	0.3805 ± N/A	0.3247 ± 0.0146	0.3125 ± N/A	0.3598 ± 0.0279
Acute MI	0.2724 ± 0.0092	0.2477 ± N/A	0.2327 ± 0.0083	0.2103 ± N/A	**0.2771 ± 0.0305**
Hyperlipidemia	0.3261 ± 0.0056	**0.3610 ± N/A**	0.2543 ± 0.0068	0.2901 ± N/A	0.3010 ± 0.0104
Lupus	0.1395 ± 0.0427	0.1397 ± N/A	0.1006 ± 0.0101	0.0780 ± N/A	**0.1625 ± 0.0323**

Table Supplementary Table 8: MoA variants (earliest labels) — **AUROC** (mean ± sd).

Task	Qwen + BGE	Qwen + ClinicalBERT	Llama-3 + ClinicalBERT	Qwen only	Qwen + BGE (unstructured prompt)
Long LOS	**0.8105 ± 0.0257**	0.8023 ± 0.0102	0.7727 ± 0.0104	0.7943 ± 0.0114	0.7369 ± 0.0123
ICU Transfer	**0.8436 ± 0.0054**	0.8114 ± 0.0123	0.7793 ± 0.0191	0.8027 ± 0.0139	0.7497 ± 0.0170
Readmission	**0.7762 ± 0.0176**	0.6969 ± 0.0174	0.6338 ± 0.0062	0.6879 ± 0.0165	0.6086 ± 0.0130
Pancreatic Cancer	**0.8611 ± 0.0150**	0.8497 ± 0.0118	0.7826 ± 0.0294	0.8416 ± 0.0117	0.7345 ± 0.0233
Hypertension	**0.6824 ± 0.0256**	0.6691 ± 0.0196	0.6392 ± 0.0198	0.6624 ± 0.0190	0.6231 ± 0.0188
Acute MI	**0.7701 ± 0.0091**	0.7516 ± 0.0102	0.7321 ± 0.0103	0.7436 ± 0.0109	0.7069 ± 0.0119
Hyperlipidemia	**0.7132 ± 0.0140**	0.6993 ± 0.0107	0.6795 ± 0.0102	0.6918 ± 0.0122	0.6648 ± 0.0136
Lupus	**0.8369 ± 0.0129**	0.7981 ± 0.0202	0.7686 ± 0.0196	0.7909 ± 0.0214	0.7421 ± 0.0205

Table Supplementary Table 9: MoA variants (earliest labels) — **AUPR** (mean ± sd).

Task	Qwen + BGE	Qwen + ClinicalBERT	Llama-3 + ClinicalBERT	Qwen only	Qwen + BGE (unstructured prompt)
Long LOS	**0.6011 ± 0.0533**	0.5822 ± 0.0105	0.5427 ± 0.0104	0.5683 ± 0.0125	0.5205 ± 0.0113
ICU Transfer	**0.2772 ± 0.0089**	0.2187 ± 0.0151	0.1486 ± 0.0198	0.2109 ± 0.0161	0.1316 ± 0.0184
Readmission	**0.3025 ± 0.0225**	0.1652 ± 0.0153	0.1298 ± 0.0199	0.1577 ± 0.0172	0.1183 ± 0.0171
Pancreatic Cancer	**0.2443 ± 0.0636**	0.2305 ± 0.0198	0.2209 ± 0.0152	0.2258 ± 0.0188	0.2048 ± 0.0168
Hypertension	**0.2868 ± 0.0327**	0.2539 ± 0.0197	0.2292 ± 0.0149	0.2446 ± 0.0193	0.2177 ± 0.0179
Acute MI	**0.2036 ± 0.0169**	0.1926 ± 0.0388	0.1852 ± 0.0101	0.1884 ± 0.0214	0.1696 ± 0.0114
Hyperlipidemia	**0.2854 ± 0.0231**	0.2687 ± 0.0191	0.2489 ± 0.0194	0.2617 ± 0.0189	0.2341 ± 0.0168
Lupus	**0.0931 ± 0.0418**	0.0884 ± 0.0194	0.0785 ± 0.0193	0.0851 ± 0.0196	0.0729 ± 0.0187

Table Supplementary Table 10: MoA variants (latest labels) — **AUROC** (mean ± sd)

Task	Qwen + BGE	Qwen + ClinicalBERT	Llama-3 + ClinicalBERT	Qwen only	Qwen + BGE (unstructured prompt)
Long LOS	**0.7931 ± 0.0030**	0.7867 ± 0.0106	0.7629 ± 0.0103	0.7786 ± 0.0112	0.7321 ± 0.0097
ICU Transfer	**0.8719 ± 0.0169**	0.8443 ± 0.0154	0.7791 ± 0.0195	0.8323 ± 0.0176	0.7396 ± 0.0191
Readmission	N/A ± N/A	N/A ± N/A	N/A ± N/A	N/A ± N/A	N/A ± N/A
Pancreatic Cancer	**0.8392 ± 0.0058**	0.8336 ± 0.0117	0.7392 ± 0.0288	0.8184 ± 0.0107	0.7064 ± 0.0218
Hypertension	**0.7259 ± 0.0066**	0.7112 ± 0.0098	0.6981 ± 0.0106	0.7068 ± 0.0096	0.6729 ± 0.0112
Acute MI	**0.7618 ± 0.0088**	0.7513 ± 0.0095	0.7013 ± 0.0107	0.7445 ± 0.0102	0.6817 ± 0.0124
Hyperlipidemia	**0.7233 ± 0.0147**	0.7087 ± 0.0102	0.6986 ± 0.0102	0.7061 ± 0.0121	0.6811 ± 0.0121
Lupus	**0.8245 ± 0.0291**	0.7991 ± 0.0197	0.7587 ± 0.0194	0.7894 ± 0.0235	0.7324 ± 0.0204

Table Supplementary Table 11: MoA variants (latest labels) — **AUPR** (mean ± sd).

Task	**Qwen + BGE**	**Qwen + ClinicalBERT**	**Llama-3 + ClinicalBERT**	**Qwen only**	**Qwen + BGE (unstructured prompt)**
Long LOS	**0.5388 ± 0.0080**	0.5036 ± 0.0124	0.4752 ± 0.0104	0.4897 ± 0.0116	0.4453 ± 0.0090
ICU Transfer	**0.2172 ± 0.0157**	0.1914 ± 0.0158	0.1087 ± 0.0197	0.1816 ± 0.0179	0.0926 ± 0.0181
Readmission	N/A ± N/A	N/A ± N/A	N/A ± N/A	N/A ± N/A	N/A ± N/A
Pancreatic Cancer	**0.3676 ± 0.0130**	0.2863 ± 0.0191	0.1053 ± 0.0195	0.2681 ± 0.0185	0.0894 ± 0.0170
Hypertension	**0.3598 ± 0.0279**	0.3385 ± 0.0206	0.2982 ± 0.0196	0.3326 ± 0.0214	0.2781 ± 0.0194
Acute MI	**0.2771 ± 0.0305**	0.2698 ± 0.0342	0.1941 ± 0.0104	0.2597 ± 0.0283	0.1697 ± 0.0116
Hyperlipidemia	**0.3010 ± 0.0104**	0.2794 ± 0.0192	0.2491 ± 0.0193	0.2719 ± 0.0187	0.2264 ± 0.0172
Lupus	**0.1625 ± 0.0323**	0.1491 ± 0.0295	0.1186 ± 0.0192	0.1413 ± 0.0264	0.1015 ± 0.0199

Table Supplementary Table 12: Count-based (earliest labels) — **AUROC**

Task	**LGBM (rollup)**	**LGBM (non-rollup)**	**TabPFN (rollup)**	**TabPFN (non-rollup)**
Long LOS	0.8047 ± 0.0039	0.7864 ± 0.0052	0.8164 ± N/A	0.8092 ± N/A
ICU Transfer	0.8440 ± 0.0058	0.8247 ± 0.0064	0.8377 ± N/A	0.8315 ± N/A
Readmission	0.7413 ± 0.0032	0.7215 ± 0.0039	0.7366 ± N/A	0.7281 ± N/A
Pancreatic Cancer	0.8821 ± 0.0070	0.8582 ± 0.0078	0.9174 ± N/A	0.8824 ± N/A
Hypertension	0.7233 ± 0.0041	0.7019 ± 0.0048	0.7292 ± N/A	0.7053 ± N/A
Acute MI	0.7663 ± 0.0119	0.7395 ± 0.0131	0.7590 ± N/A	0.7331 ± N/A
Hyperlipidemia	0.7289 ± 0.0044	0.7062 ± 0.0051	0.7321 ± N/A	0.7078 ± N/A
Lupus	0.6689 ± 0.0100	0.6396 ± 0.0111	0.7039 ± N/A	0.6902 ± N/A

Table Supplementary Table 13: Count-based (earliest labels) — **AUPR**

Task	**LGBM (rollup)**	**LGBM (non-rollup)**	**TabPFN (rollup)**	**TabPFN (non-rollup)**
Long LOS	0.6172 ± 0.0116	0.5774 ± 0.0126	0.6289 ± N/A	0.5862 ± N/A
ICU Transfer	0.2381 ± 0.0220	0.1993 ± 0.0241	0.2738 ± N/A	0.2515 ± N/A
Readmission	0.2504 ± 0.0203	0.2118 ± 0.0217	0.2449 ± N/A	0.2087 ± N/A
Pancreatic Cancer	0.3144 ± 0.0463	0.2684 ± 0.0479	0.3842 ± N/A	0.3332 ± N/A
Hypertension	0.3241 ± 0.0070	0.2897 ± 0.0079	0.3354 ± N/A	0.3055 ± N/A
Acute MI	0.2113 ± 0.0052	0.1766 ± 0.0058	0.2225 ± N/A	0.1864 ± N/A
Hyperlipidemia	0.2976 ± 0.0124	0.2519 ± 0.0136	0.2955 ± N/A	0.2461 ± N/A
Lupus	0.0866 ± 0.0252	0.0528 ± 0.0268	0.1177 ± N/A	0.0827 ± N/A

Table Supplementary Table 14: Count-based (latest labels) — **AUROC**

Task	**LGBM (rollup)**	**LGBM (non-rollup)**	**TabPFN (rollup)**	**TabPFN (non-rollup)**
Long LOS	0.7918 ± 0.0015	0.7726 ± 0.0019	0.8077 ± N/A	0.7813 ± N/A
ICU Transfer	0.8213 ± 0.0036	0.7957 ± 0.0041	0.8498 ± N/A	0.8208 ± N/A
Readmission	N/A ± N/A	N/A ± N/A	N/A ± N/A	N/A ± N/A
Pancreatic Cancer	0.8609 ± 0.0113	0.8359 ± 0.0121	0.8961 ± N/A	0.8617 ± N/A
Hypertension	0.7297 ± 0.0025	0.7035 ± 0.0031	0.7142 ± N/A	0.6886 ± N/A
Acute MI	0.7644 ± 0.0083	0.7344 ± 0.0091	0.7500 ± N/A	0.7222 ± N/A
Hyperlipidemia	0.7261 ± 0.0048	0.6998 ± 0.0053	0.7504 ± N/A	0.7216 ± N/A
Lupus	0.7027 ± 0.0022	0.6791 ± 0.0027	0.7588 ± N/A	0.7442 ± N/A

Table Supplementary Table 15: Count-based (latest labels) — **AUPR** (mean ± sd).

Task	**LGBM (rollup)**	**LGBM (non-rollup)**	**TabPFN (rollup)**	**TabPFN (non-rollup)**
Long LOS	0.5342 ± 0.0037	0.4973 ± 0.0042	0.5488 ± N/A	0.5081 ± N/A
ICU Transfer	0.1533 ± 0.0078	0.1196 ± 0.0086	0.1893 ± N/A	0.1683 ± N/A
Readmission	N/A ± N/A	N/A ± N/A	N/A ± N/A	N/A ± N/A
Pancreatic Cancer	0.3418 ± 0.0392	0.2962 ± 0.0406	0.3969 ± N/A	0.3491 ± N/A
Hypertension	0.4002 ± 0.0084	0.3584 ± 0.0093	0.3805 ± N/A	0.3368 ± N/A
Acute MI	0.2724 ± 0.0092	0.2337 ± 0.0101	0.2477 ± N/A	0.2414 ± N/A
Hyperlipidemia	0.3261 ± 0.0056	0.2865 ± 0.0062	0.3610 ± N/A	0.3186 ± N/A
Lupus	0.1395 ± 0.0427	0.0991 ± 0.0439	0.1397 ± N/A	0.1162 ± N/A

Below are the prompts used for the four clinical tasks. We used a single, manually engineered prompt template that is task-agnostic and requires only minimal edits to change the task name and time horizon.P2

Prompt S1: Long length-of-stay

```
You are a medical expert evaluating a patient who is currently
    {g.age_at_predict} years old. Based on the following medical history
    (formatted as EVENT at AGE), your task is to write a concise summary of
    this patient’s risk profile for long length-of-stay during this admission.
 Return JSON only:
    {
      'risk_category': 'Low'|'Moderate'|'High',
      'risk_score': <0..1>,
      'drivers_positive': ['substrings from medical history'],
      'drivers_negative': ['substrings from medical history'],
      'justification': 2 to 4 sentences referencing drivers and timing',
      'insufficient_evidence': true|false
    }
Guidance:
 - Do not add external facts.
 - Prefer explicit outcome-related conditions/medications as positive drivers.
 - Weigh recent items more (use recent event when present).
 - If evidence is sparse/ambiguous, prefer Low and set
     insufficient_evidence=true.
 - Map Low0.10 to 0.33, Moderate to 0.34 to 0.66, High0.67 to 0.90.
```

Prompt S2: readmission

```
You are a medical expert evaluating a patient who is currently
    {g.age_at_predict} years old. Based on the following medical history
    (formatted as EVENT at AGE), your task is to write a concise summary of
    this patient’s risk profile for readmission within the next 30 days after
    discharge.
 Return JSON only:
    {
      'risk_category': 'Low'|'Moderate'|'High',
      'risk_score': <0..1>,
      'drivers_positive': ['substrings from medical history'],
      'drivers_negative': ['substrings from medical history'],
      'justification': 2 to 4 sentences referencing drivers and timing',
      'insufficient_evidence': true|false
    }
Guidance:
 - Do not add external facts.
 - Prefer explicit outcome-related conditions/medications as positive drivers.
 - Weigh recent items more (use recent event when present).
 - If evidence is sparse/ambiguous, prefer Low and set
     insufficient_evidence=true.
 - Map Low0.10 to 0.33, Moderate to 0.34 to 0.66, High0.67 to 0.90.
```

Prompt S3: ICU transfer

```
You are a medical expert evaluating a patient who is currently
    {g.age_at_predict} years old. Based on the following medical history
    (formatted as EVENT at AGE), your task is to write a concise summary of
    this patient’s risk profile for an ICU transfer during this admission
 Return JSON only:
    {
      ’risk_category’: ’Low’|’Moderate’|’High’,
      ’risk_score’: <0..1>,
      ’drivers_positive’: [’substrings from medical history’],
      ’drivers_negative’: [’substrings from medical history’],
      ’justification’: 2 to 4 sentences referencing drivers and timing’,
      ’insufficient_evidence’: true|false
    }
Guidance:
 - Do not add external facts.
 - Prefer explicit outcome-related conditions/medications as positive drivers.
 - Weigh recent items more (use recent event when present).
 - If evidence is sparse/ambiguous, prefer Low and set
     insufficient_evidence=true.
 - Map Low0.10 to 0.33, Moderate to 0.34 to 0.66, High0.67 to 0.90.
```

Prompt S4: Pancreatic cancer

```
You are a medical expert evaluating a patient who is currently
    {g.age_at_predict} years old. Based on the following medical history
    (formatted as EVENT at AGE), your task is to write a concise summary of
    this patient’s risk profile for pancreatic cancer within the next year.
 Return JSON only:
    {
      ’risk_category’: ’Low’|’Moderate’|’High’,
      ’risk_score’: <0..1>,
      ’drivers_positive’: [’substrings from medical history’],
      ’drivers_negative’: [’substrings from medical history’],
      ’justification’: 2 to 4 sentences referencing drivers and timing’,
      ’insufficient_evidence’: true|false
    }
Guidance:
 - Do not add external facts.
 - Prefer explicit outcome-related conditions/medications as positive drivers.
 - Weigh recent items more (use recent event when present).
 - If evidence is sparse/ambiguous, prefer Low and set
     insufficient_evidence=true.
 - Map Low0.10 to 0.33, Moderate to 0.34 to 0.66, High0.67 to 0.90.
```

Prompt S5: Hypertension

```
You are a medical expert evaluating a patient who is currently
    {g.age_at_predict} years old. Based on the following medical history
    (formatted as EVENT at AGE), your task is to write a concise summary of
    this patient’s risk profile for hypertension within the next year.
 Return JSON only:
    {
      ’risk_category’: ’Low’|’Moderate’|’High’,
      ’risk_score’: <0..1>,
      ’drivers_positive’: [’substrings from medical history’],
      ’drivers_negative’: [’substrings from medical history’],
      ’justification’: 2 to 4 sentences referencing drivers and timing’,
      ’insufficient_evidence’: true|false
    }
Guidance:
 - Do not add external facts.
 - Prefer explicit outcome-related conditions/medications as positive drivers.
 - Weigh recent items more (use recent event when present).
 - If evidence is sparse/ambiguous, prefer Low and set
     insufficient_evidence=true.
 - Map Low0.10 to 0.33, Moderate to 0.34 to 0.66, High0.67 to 0.90.
```

Prompt S6: Acute mycardial infarction

```
You are a medical expert evaluating a patient who is currently
    {g.age_at_predict} years old. Based on the following medical history
    (formatted as EVENT at AGE), your task is to write a concise summary of
    this patient’s risk profile for acute MI within the next year.
 Return JSON only:
    {
      ’risk_category’: ’Low’|’Moderate’|’High’,
      ’risk_score’: <0..1>,
      ’drivers_positive’: [’substrings from medical history’],
      ’drivers_negative’: [’substrings from medical history’],
      ’justification’: 2 to 4 sentences referencing drivers and timing’,
      ’insufficient_evidence’: true|false
    }
Guidance:
 - Do not add external facts.
 - Prefer explicit outcome-related conditions/medications as positive drivers.
 - Weigh recent items more (use recent event when present).
 - If evidence is sparse/ambiguous, prefer Low and set
     insufficient_evidence=true.
 - Map Low0.10 to 0.33, Moderate to 0.34 to 0.66, High0.67 to 0.90.
```

Prompt S7: Hyperlipidemia

```
You are a medical expert evaluating a patient who is currently
    {g.age_at_predict} years old. Based on the following medical history
    (formatted as EVENT at AGE), your task is to write a concise summary of
    this patient’s risk profile for hyperlipidemia within the next year.
 Return JSON only:
   {
     ’risk_category’: ’Low’|’Moderate’|’High’,
     ’risk_score’: <0..1>,
     ’drivers_positive’: [’substrings from medical history’],
     ’drivers_negative’: [’substrings from medical history’],
     ’justification’: 2 to 4 sentences referencing drivers and timing’,
     ’insufficient_evidence’: true|false
   }
Guidance:
 - Do not add external facts.
 - Prefer explicit outcome-related conditions/medications as positive drivers.
 - Weigh recent items more (use recent event when present).
 - If evidence is sparse/ambiguous, prefer Low and set
     insufficient_evidence=true.
 - Map Low0.10 to 0.33, Moderate to 0.34 to 0.66, High0.67 to 0.90.
```

Prompt S8: Lupus

```
You are a medical expert evaluating a patient who is currently
    {g.age_at_predict} years old. Based on the following medical history
    (formatted as EVENT at AGE), your task is to write a concise summary of
    this patient’s risk profile for lupus within the next year.
 Return JSON only:
   {
     ’risk_category’: ’Low’|’Moderate’|’High’,
     ’risk_score’: <0..1>,
     ’drivers_positive’: [’substrings from medical history’],
     ’drivers_negative’: [’substrings from medical history’],
     ’justification’: 2 to 4 sentences referencing drivers and timing’,
     ’insufficient_evidence’: true|false
   }
Guidance:
 - Do not add external facts.
 - Prefer explicit outcome-related conditions/medications as positive drivers.
 - Weigh recent items more (use recent event when present).
 - If evidence is sparse/ambiguous, prefer Low and set
     insufficient_evidence=true.
 - Map Low0.10 to 0.33, Moderate to 0.34 to 0.66, High0.67 to 0.90.
```

Unstructued Prompt S9: Long LOS

```
You are a medical expert evaluating a patient who is currently
    {g.age_at_predict} years old. Based on the following medical history
    (formatted as EVENT at AGE), your task is to write a concise summary of
    this patient’s risk profile for a long length-of-stay during this admission.
```

Unstructued Prompt S10: Readmission

```
You are a medical expert evaluating a patient who is currently
    {g.age_at_predict} years old. Based on the following medical history
    (formatted as EVENT at AGE), your task is to write a concise summary of
    this patient’s risk profile for a readmission within the next 30 days after
    discharge.
```

Unstructued Prompt S11: ICU transfer

```
You are a medical expert evaluating a patient who is currently
    {g.age_at_predict} years old. Based on the following medical history
    (formatted as EVENT at AGE), your task is to write a concise summary of
    this patient’s risk profile for lupus within the next year
```

Unstructued Prompt S12: Pancreatic cancer

```
You are a medical expert evaluating a patient who is currently
    {g.age_at_predict} years old. Based on the following medical history
    (formatted as EVENT at AGE), your task is to write a concise summary of
    this patient’s risk profile for an ICU transfer during this admission.
```

Unstructued Prompt S13: Hypertension

```
You are a medical expert evaluating a patient who is currently
    {g.age_at_predict} years old. Based on the following medical history
    (formatted as EVENT at AGE), your task is to write a concise summary of
    this patient’s risk profile for hypertension within the next year
```

Unstructued Prompt S14: Acute mycardial infarction

```
You are a medical expert evaluating a patient who is currently
    {g.age_at_predict} years old. Based on the following medical history
    (formatted as EVENT at AGE), your task is to write a concise summary of
    this patient’s risk profile for acute mycardial infarction within the next
    year
```

Unstructued Prompt S15: Hyperlipidemia

```
You are a medical expert evaluating a patient who is currently
    {g.age_at_predict} years old. Based on the following medical history
    (formatted as EVENT at AGE), your task is to write a concise summary of
    this patient’s risk profile for hyperlipidemia within the next year
```

Unstructued Prompt S16: Lupus

```
You are a medical expert evaluating a patient who is currently
    {g.age_at_predict} years old. Based on the following medical history
    (formatted as EVENT at AGE), your task is to write a concise summary of
    this patient’s risk profile for lupus within the next year
```