Structured Evidence Routing for Incident Risk Prediction from Multimodal Longitudinal EHRs

Animesh Agarwal [1] **Meysam Ghaffari** [1] **Nina Fatehi** [1] **Carlos Morato** [1]

Abstract

Incident risk prediction from longitudinal electronic health records (EHRs) is challenging because relevant signals are multimodal, weak in isolation, and distributed across irregular patient histories. We propose structured evidence routing, a router–predictor–reviewer workflow that separates full-record access from disease-specific assessment. The router organizes the complete pre-index EHR into a compact summary and targeted evidence slices; the predictor uses this evidence to form an evidence-linked risk assessment, which the reviewer critiques. For comparison with supervised EHRSHOT baselines, we pair the routed evidence summaries with a supervised classifier readout. Across five 1-year incident diagnosis tasks, our method reaches the AUROC range of established supervised EHRSHOT baselines and remains competitive on AUPRC, while exposing a patient-specific evidence trail. Internal pre-readout ablations further suggest that routing, laboratory evidence, task guidance, and review each contribute to performance.

1. Introduction

Many clinically important conditions emerge gradually, with early warning signals distributed across years of care. The practical challenge is therefore not only to diagnose disease once it is obvious, but to identify patients who are on track to develop a new condition while intervention is still possible. In structured longitudinal electronic health records (EHRs), the relevant signal is rarely concentrated in one modality: clinicians integrate diagnoses, medications, laboratory trajectories, physiologic measurements, procedures, and utilization patterns over time when assessing future risk. The challenge is thus not only prediction from tabular or sequential data, but prediction from *heterogeneous, irregular, multimodal structured health records*.

A large literature has improved prediction from longitudinal structured EHRs using recurrent, transformer-based, and pretrained foundation-style models, including methods for future diagnosis prediction and related forecasting tasks (Choi et al., 2016a;b; Ma et al., 2017; Li et al., 2020; 2023; Rasmy et al., 2021; Yang et al., 2023; Steinberg et al., 2024). These approaches have substantially advanced representation learning for structured patient histories, but incident risk prediction remains difficult when signals are weak, temporally diffuse, and distributed across modalities. In practice, many pipelines still rely primarily on coded diagnoses, procedures, and medications, while laboratory and physiologic observations are not always incorporated in an equally rich or explicit way. More recent work has explored retrieval augmentation and prompting for clinical prediction from structured histories (Xu et al., 2024; Ben Shoham and Rappoport, 2024; Renc et al., 2024). EHRSHOT provides a standardized benchmark for structured EHR prediction (Wornow et al., 2023), and recent work uses it to compare count-based, pretrained sequential, and inference-time prompting pipelines under shared protocols (Gao et al., 2025).

This paper studies evidence organization for incident prediction from multimodal longitudinal EHRs. Passing a full chart directly to a predictor can bury weak disease-specific signals, while narrow feature views can miss repeated abnormal measurements, treatment changes, or cross-visit trends. We propose a router–predictor–reviewer workflow that keeps full-record access separate from disease-specific assessment: the router summarizes the pre-index record and returns targeted evidence slices, the predictor uses task guidance to score risk, and the reviewer evaluates the intermediate assessment for unsupported claims, missing evidence, or temporal inconsistencies.

We evaluate on five 1-year incident diagnosis tasks from EHRSHOT: hypertension, hyperlipidemia, lupus, acute myocardial infarction, and pancreatic cancer. For the main benchmark comparison, we pair routed evidence summaries with a supervised classifier readout, matching the task-label access used by count-based models, CLMBR, and

[1]Optum AI, UnitedHealth Group, Minneapolis, Minnesota, USA. Correspondence to: Animesh Agarwal <animesh.agarwal@optum.com>.

Proceedings of the Workshop on Structured Data for Health at the 43^{rd} International Conference on Machine Learning, Seoul, South Korea.

MoA+Cls. We also report within-framework pre-readout ablations to assess the contribution of routing, laboratory inputs, task guidance, and review.

Our contributions are threefold: (i) we use the full pre-index structured EHR for incident risk prediction, including diagnoses, procedures, medications, laboratory results, measurements, observations, and utilization history; (ii) we introduce structured evidence routing to separate full-record access from disease-specific assessment while producing patient-specific evidence traces; and (iii) we evaluate routed summaries as supervised prediction representations on five EHRSHOT tasks, with comparisons to count-based, CLMBR, and MoA+Cls baselines and pre-readout ablations of routing, laboratory evidence, task guidance, and review.

2. Method

We now define the incident prediction setting and describe the three components shown in Figure 1: the Record Router, Task Predictor, and Reviewer. The design is related to gatekeeper-style diagnostic orchestration, but targets structured longitudinal EHR prediction rather than sequential diagnosis (Nori et al., 2025).

2.1. Problem Setup

For each patient, let $X^{(\leq t^*)} = \{e_t \mid t \leq t^*\}$ denote the structured longitudinal EHR history observed up to prediction time t^*, where each event may include diagnoses, medications, procedures, laboratory results, measurements, and other structured observations. Given a target disease y and prediction horizon h, the task is to estimate

$$\hat{p}_{y,h} \approx P\Big(Y_y(t^* + h) = 1 \mid X^{(\leq t^*)}\Big), \tag{1}$$

that is, the probability that the patient will develop an incident outcome within the future horizon using only information available before the prediction time.

The framework operates directly on the full pre-index longitudinal EHR rather than on a fixed hand-engineered feature vector. For each benchmark instance, the available record includes structured events observed before the task-specific prediction time across diagnoses, procedures, medications, laboratory results, measurements, and observations. All three inference modules are implemented as prompted LLM components, but they have different access privileges and roles.

The **Record Router** is the only module allowed to access the full structured record. It is constrained to organize evidence rather than make the final disease-specific prediction. It serves two roles. First, it constructs a compact patient summary containing salient active conditions, major medications, recent utilization, and notable laboratory or physiologic trends. Second, it returns *targeted evidence slices* in response to disease-specific follow-up requests, such as selected diagnosis history, laboratory trajectories, medication changes, or time-localized observations.

The **Task Predictor** is a disease-specific LLM assessment module, with task guidance derived from established clinical guidelines for the target condition. It does not access the raw chart directly; instead, it receives the router summary, identifies what additional evidence is needed under task-specific clinical criteria, and requests targeted slices from the router. Using the summary, retrieved evidence, and task guidance, the predictor produces an initial incident-risk estimate with a short evidence-linked rationale.

The **Reviewer** is an LLM critique module that evaluates the intermediate assessment for unsupported claims, missing evidence, or temporal inconsistencies. When needed, it triggers iterative refinement: the predictor requests additional evidence from the router, updates its assessment, and may repeat this process until the reviewer finds the prediction sufficiently supported or a preset iteration limit is reached. Figure 1 illustrates the workflow. Prompt templates for the Record Router, Task Predictor, and Reviewer are provided in Appendix B.

3. Experimental Setup

We evaluate on EHRSHOT in the 1-year incident diagnosis setting, following the official task definitions and the latest-label protocol used by Gao et al. (2025). In this setting, each patient contributes at most one prediction time per task: when multiple eligible labels are available, we retain the most recent one. We use the official train/validation/test partitioning from EHRSHOT (Wornow et al., 2023) and evaluate five tasks: hypertension, hyperlipidemia, lupus, acute myocardial infarction, and pancreatic cancer. For each benchmark instance, only structured EHR events observed on or before the prediction time are included.

For the main cross-method benchmark comparison, we report a supervised-readout setting because the strongest structured-EHR baselines in this benchmark, including count-based models and CLMBR, are trained on task labels. We first run the structured evidence-routing framework to construct disease-specific routed evidence summaries. These summaries are embedded with PubMedBERT (Gu et al., 2021) and passed to an XGBoost classifier (Chen and Guestrin, 2016) trained on benchmark training labels. This places our method in the same task-label-access regime as count-based, CLMBR, and MoA+Cls baselines. This comparison treats each method as a supervised prediction pipeline built on a different patient representation. Count-based models use coded event features, CLMBR uses learned longitudinal EHR embeddings, MoA+Cls uses

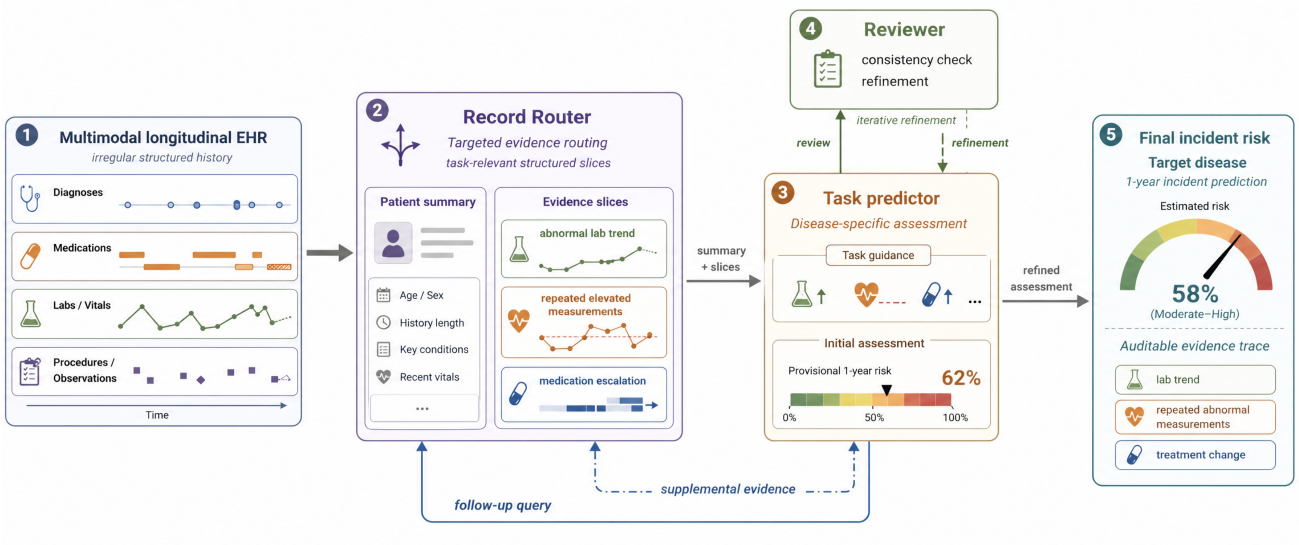

Figure 1. Structured evidence routing for incident risk prediction. A record router organizes long, irregular EHR histories into a compact summary and targeted evidence slices. A review step can trigger refinement before final disease-specific prediction.

LLM-generated summaries with a classifier head, and our method uses routed evidence summaries with a classifier head. The comparison therefore asks whether the evidence surfaced by the routing workflow can support risk scores in the range of established supervised EHR baselines while preserving an auditable evidence trail.

We additionally report internal ablations on the pre-readout framework. These ablations are not intended as cross-method comparisons to EHRSHOT baselines. Instead, they isolate the effect of removing components within the same evidence-routing pipeline. We therefore do not train a separate supervised classifier for each ablated variant; doing so would test whether a downstream classifier can compensate for missing evidence-routing components, rather than whether those components contribute to the framework's own prediction signal. Cohort construction, preprocessing details, and supervised-readout details are provided in Appendix C and Appendix D. The clinical guideline sources used for task guidance are summarized in Appendix E.

4. Results

4.1. Supervised Readout Comparison to Structured-EHR Baselines

Tables 1 and 2 compare our supervised-readout variant with established EHRSHOT baselines. The comparison places routed evidence summaries alongside other supervised patient representations: coded event features, learned longitudinal embeddings, and LLM-generated summaries.

On AUROC, our supervised-readout variant is competitive across all five tasks (Table 1). It improves over CLMBR on every task and matches or exceeds MoA+Cls on all five tasks, including gains on hyperlipidemia, pancreatic cancer, and acute myocardial infarction. Compared with the count-based baseline, our method matches performance on hypertension and hyperlipidemia, exceeds it on lupus and acute myocardial infarction, and is within 0.01 AUROC on pancreatic cancer. These results indicate that routed evidence summaries preserve enough longitudinal risk signal for effective case ranking under a supervised readout.

AUPRC gives a complementary view of the results (Table 2), especially because incident diagnosis tasks can be class-imbalanced. Under this metric, our method improves over CLMBR on all five tasks, suggesting that routed evidence summaries provide a stronger supervised prediction representation than learned longitudinal embeddings in this setting. The method is also competitive with MoA+Cls: it matches MoA+Cls on hypertension, exceeds it on hyperlipidemia and lupus, and remains close on pancreatic cancer and acute myocardial infarction. Against count-based models, which remain a strong supervised reference point, our method reaches a similar range on multiple tasks and obtains the highest AUPRC on lupus. Thus, the precision–recall results support the same overall conclusion: structured evidence routing yields useful prediction features while also exposing the disease-specific evidence behind the score.

4.2. Internal Ablations of the Evidence-Routing Framework

These ablations are within-framework diagnostics: they evaluate whether removing individual components weakens the framework's direct prediction signal before the supervised readout.

Table 1. **Supervised-readout AUROC on EHRSHOT** for 1-year incident diagnosis prediction. Our method uses routed evidence summaries with a trained classifier readout, placing it in the same supervised task-label setting as count-based, CLMBR, and MoA+Cls baselines. For Ours+Cls, ± denotes the bootstrap standard error estimated from 1,000 resamples; baseline uncertainties are reported as provided by prior work.

Task	Count-based	CLMBR	MoA+Cls	Ours+Cls
Hypertension	0.73 ± 0.003	0.70 ± 0.003	0.73 ± 0.01	0.73 ± 0.03
Hyperlipidemia	0.75	0.70	0.72 ± 0.01	0.75 ± 0.02
Lupus	0.76	0.77 ± 0.04	0.82 ± 0.03	0.82 ± 0.04
Pancreatic cancer	0.89	0.82	0.84 ± 0.005	0.88 ± 0.03
Acute myocardial infarction	0.76 ± 0.01	0.74 ± 0.01	0.76 ± 0.01	0.77 ± 0.02

MoA+Cls: Qwen + BGE classifier + trained classifier head.
Ours+Cls: structured evidence-routing summaries + PubMedBERT embeddings + trained classifier head.

Table 2. **Supervised-readout AUPRC on EHRSHOT** for 1-year incident diagnosis prediction. For Ours+Cls, ± denotes the bootstrap standard error estimated from 1,000 resamples; baseline uncertainties are reported as provided by prior work.

Task	Count-based	CLMBR	MoA+Cls	Ours+Cls
Hypertension	0.40 ± 0.01	0.32 ± 0.01	0.36 ± 0.02	0.36 ± 0.03
Hyperlipidemia	0.36 ± 0.005	0.29	0.30 ± 0.01	0.34 ± 0.04
Lupus	0.14 ± 0.04	0.10 ± 0.01	0.16 ± 0.03	0.18 ± 0.07
Pancreatic cancer	0.39 ± 0.04	0.20	0.36 ± 0.01	0.34 ± 0.07
Acute myocardial infarction	0.27 ± 0.01	0.23 ± 0.01	0.28 ± 0.03	0.27 ± 0.04

MoA+Cls: Qwen + BGE classifier + trained classifier head.
Ours+Cls: structured evidence-routing summaries + PubMedBERT embeddings + trained classifier head.

We run targeted ablations on the EHRSHOT pancreatic cancer task, a low-prevalence and clinically heterogeneous endpoint where relevant evidence may be distributed across laboratory results, observations, procedures, and utilization history. We remove the Reviewer, laboratory and observation inputs, the Record Router, and disease-specific task guidance, while keeping the rest of the framework fixed. Table 3 reports pre-readout AUROC for these within-framework comparisons.

The full framework achieves an AUROC of 0.86 ± 0.03. Removing laboratory measurements or bypassing the Record Router produces the largest drops, reducing AUROC to 0.83. Removing review-based refinement or task-specific guidance also lowers AUROC to 0.84. These results suggest that the framework's direct prediction signal is not driven by a single module alone, but by the combination of multimodal structured inputs, mediated access to the longitudinal record, task-specific assessment criteria, and iterative review. Additional details on the ablation setup are provided in Appendix A.

5. Discussion

Across five EHRSHOT tasks, the supervised-readout results show that routed evidence summaries are effective patient representations for incident risk prediction. They improve over CLMBR across tasks, remain competitive with MoA+Cls, and reach the range of strong supervised baselines under both AUROC and AUPRC. These results suggest that explicitly organizing multimodal longitudinal evidence before readout can preserve clinically useful risk signal.

Table 3. **Internal pre-readout ablation** on pancreatic cancer.

Configuration	Pre-readout AUROC
Full framework	0.86 ± 0.03
−Reviewer	0.84 ± 0.033
−Lab measurements	0.83 ± 0.034
−Record Router	0.83 ± 0.037
−Task guidance	0.84 ± 0.030

The value of the framework is not only predictive performance, but the form of the intermediate representation. Unlike count-based features or pretrained longitudinal embeddings, the proposed workflow surfaces a compact patient summary, targeted disease-relevant evidence slices, and an evidence-linked representation before scoring. This makes the path from longitudinal record to risk score more explicit: the readout operates on inspectable evidence rather than an opaque feature vector alone. The ablations further show that routing, laboratory evidence, task guidance, and review each contribute to the direct prediction signal, supporting the claim that evidence selection and organization matter for multimodal longitudinal risk prediction.

6. Limitations

This study has clinical-evaluation limitations. Evaluation is retrospective and based on benchmark labels, so the results do not yet measure how clinicians would interpret or use the surfaced evidence in practice. In particular, we do not evaluate whether the evidence traces improve clinician trust, reduce review burden, or support earlier recognition of incident disease. A natural next step is clinician-facing evaluation of the routed evidence, rationales, and risk estimates in realistic review workflows.

References

Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016a). Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. In *Proceedings of the 1st Machine Learning for Healthcare Conference*. *Proceedings of Machine Learning Research*, **56**, 301–318.

Choi, E., Bahadori, M. T., Kulas, J. A., Schuetz, A., Stewart, W. F., & Sun, J. (2016b). RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems 29*, 3504–3512.

Ben Shoham, O., & Rappoport, N. (2024). CPLLM: Clinical prediction with large language models. *PLOS Digital Health*, **3**(12), e0000680. doi: 10.1371/journal.pdig.0000680.

Renc, P., Jia, Y., Samir, A. E., Was, J., Li, Q., Bates, D. W., & Sitek, A. (2024). Zero shot health trajectory prediction using transformer. *npj Digital Medicine*, 7, 256. https://doi.org/10.1038/s41746-024-01235-0

Ma, F., Chitta, R., Zhou, J., You, Q., Sun, T., & Gao, J. (2017). Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional Recurrent Neural Networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1903–1911. doi: 10.1145/3097983.3098088.

Li, Y., Rao, S., Solares, J. R. A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., & Salimi-Khorshidi, G. (2020). BEHRT: Transformer for electronic health records. *Scientific Reports*, **10**, 7155. doi: 10.1038/s41598-020-62922-y.

Li, Y., Mamouei, M., Salimi-Khorshidi, G., Rao, S., Hassaine, A., Canoy, D., Lukasiewicz, T., & Rahimi, K. (2023). Hi-BEHRT: Hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. *IEEE Journal of Biomedical and Health Informatics*, **27**(2), 1106–1117. doi: 10.1109/JBHI.2022.3224727.

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., Zhi, D., & Zozus, M. N. (2021). Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, **4**, 86. doi: 10.1038/s41746-021-00455-y.

Wornow, M., Thapa, R., Steinberg, E., Fries, J. A., & Shah, N. H. (2023). EHRSHOT: An EHR Benchmark for Few-Shot Evaluation of Foundation Models. In *Advances in Neural Information Processing Systems*.

Yang, Z., Mitra, A., Liu, W., et al. (2023). TransformEHR: Transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nature Communications*, **14**, 7857. doi: 10.1038/s41467-023-43715-z.

Xu, R., Shi, W., Yu, Y., Zhuang, Y., Jin, B., Wang, M. D., Ho, J. C., & Yang, C. (2024). RAM-EHR: Retrieval Augmentation Meets Clinical Predictions on Electronic Health Records. In L.-W. Ku, A. Martins, & V. Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 754–765. Bangkok, Thailand: Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-short.68.

Steinberg, E., Fries, J. A., Xu, Y., & Shah, N. H. (2024). MOTOR: A Time-to-Event Foundation Model for Structured Medical Records. In *International Conference on Learning Representations*. https://openreview.net/forum?id=NialiwI2V6

Gao, J., Rosenthal, M., Wolpin, B., & Cristea, S. (2025). Count-Based Approaches Remain Strong: A Benchmark Against Transformer and LLM Pipelines on Structured EHR. In *The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance at NeurIPS 2025*. https://openreview.net/forum?id=fKJtKew2YQ

Nori, H., Daswani, M., Kelly, C., Lundberg, S., Ribeiro, M. T., Wilson, M., Liu, X., Sounderajah, V., Carlson, J., Lungren, M. P., Gross, B., Hames, P., Suleyman, M., King, D., & Horvitz, E. (2025). Sequential Diagnosis with Language Models. *arXiv preprint arXiv:2506.22405*. https://arxiv.org/abs/2506.22405

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, **3**(1), 1–23. doi: 10.1145/3458754.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. doi: 10.1145/2939672.2939785.

Jones, D. W., Ferdinand, K. C., Taler, S. J., Johnson, H. M., Shimbo, D., Abdalla, M., et al. (2025). 2025 AHA/ACC/AANP/AAPA/ABC/ACCP/ACPM/AGS/AMA/ASPC/NMA/PCNA/SGIM guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Circulation*, **152**(11), e114–e218. doi: 10.1161/CIR.0000000000001356.

Blumenthal, R. S., Morris, P. B., Gaudino, M., Johnson, H. M., Anderson, T. S., Bittner, V. A., Blankstein, R., Brewer, L. C., Cho, L., de Ferranti, S. D., Gianos, E., Gluckman, T. J., Gradney, K. F., Isiadinso, I., Lloyd-Jones, D. M., Marrs, J. C., Martin, S. S., McLain, K. H., Mehta, L. S., Mora, S., Mulugeta, W. M., Natarajan, P., Navar, A. M., Orringer, C. E., Polonsky, T. S., Reynolds, H. R., Saseen, J. J., Shapiro, M. D., Soffer, D. E., Tynes, D. E., Villavaso, C. D., Virani, S. S., & Wilkins, J. T. (2026). 2026 ACC/AHA/AACVPR/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA guideline on the

management of dyslipidemia: A report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Circulation*. Published online March 13, 2026. doi: 10.1161/CIR.0000000000001423.

Aringer, M., Costenbader, K., Daikh, D., Brinks, R., Mosca, M., Ramsey-Goldman, R., Smolen, J. S., Wofsy, D., Boumpas, D. T., Kamen, D. L., Jayne, D., Cervera, R., Costedoat-Chalumeau, N., Diamond, B., Gladman, D. D., Hahn, B., Hiepe, F., Jacobsen, S., Khanna, D., Lerstrøm, K., Massarotti, E., McCune, J., Ruiz-Irastorza, G., Sanchez-Guerrero, J., Schneider, M., Urowitz, M., Bertsias, G., Hoyer, B. F., Leuchten, N., Tani, C., Tedeschi, S. K., Touma, Z., Schmajuk, G., et al. (2019). 2019 European League Against Rheumatism/American College of Rheumatology classification criteria for systemic lupus erythematosus. *Annals of the Rheumatic Diseases*, **78**(9), 1151–1159. doi: 10.1136/annrheumdis-2018-214819.

Thygesen, K., Alpert, J. S., Jaffe, A. S., Chaitman, B. R., Bax, J. J., Morrow, D. A., & White, H. D.; The Executive Group on behalf of the Joint European Society of Cardiology (ESC)/American College of Cardiology (ACC)/American Heart Association (AHA)/World Heart Federation (WHF) Task Force for the Universal Definition of Myocardial Infarction. (2018). Fourth universal definition of myocardial infarction (2018). *Circulation*, **138**(20), e618–e651. doi: 10.1161/CIR.0000000000000617.

Conroy, T., Pfeiffer, P., Vilgrain, V., Lamarca, A., Seufferlein, T., O'Reilly, E. M., Hackert, T., Golan, T., Prager, G., Haustermans, K., Vogel, A., Ducreux, M., & ESMO Guidelines Committee. (2023). Pancreatic cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. *Annals of Oncology*, **34**(11), 987–1002. doi: 10.1016/j.annonc.2023.08.009.

A. EHRSHOT Ablation Details

The ablation study in Table 3 is performed on the EHRSHOT pancreatic cancer task before the supervised classifier readout. Each variant removes one component while keeping the rest of the framework fixed: removing the Reviewer stops the pipeline after the initial prediction round; removing laboratory inputs excludes lab-derived measurements and observations from the patient record; removing the Record Router exposes the raw structured EHR directly to the predictor; and removing task guidance omits guideline-derived disease-specific instructions from the predictor prompt. These ablations isolate how routing, multimodal evidence, task guidance, and review affect the framework's direct prediction signal without training a new readout for each variant.

B. Prompt Templates for EHRSHOT

This appendix summarizes the prompt templates used by the three inference modules in the EHRSHOT experiments: the **Record Router**, the **Task Predictor**, and the **Reviewer**. Variable placeholders are shown as {`variable_name`}. Disease-specific task guidance is injected at runtime as {`task_guidance`}.

B.1. System Overview

The system uses three modules in a structured evidence-routing loop. The **Record Router** is the only module with access to the longitudinal EHR and is responsible for summarization and targeted evidence retrieval. The **Task Predictor** is a disease-specific module that requests evidence, forms an initial risk assessment, and updates that assessment after review. The **Reviewer** evaluates whether the intermediate assessment is evidence-grounded and task-aligned, and can trigger an additional evidence request.

B.2. Record Router Prompts

PROMPT R-1: PATIENT SUMMARIZATION

System: You are a clinical record router with access to a patient's longitudinal structured EHR.

Task: Produce a concise, neutral summary of the patient record for downstream prediction.

Instructions:

- Summarize demographics and major recent or recurrent clinical history.
- Highlight high-signal diagnoses, medications, laboratory findings, and observations relevant to future disease risk.
- Use plain clinical language and avoid unsupported inference.
- Do not diagnose, speculate, or recommend treatment.

Input: {`patient_metadata`}, {`clinical_data`}, {`family_history`}

Output: A brief neutral paragraph summarizing the patient context.

PROMPT R-2: TARGETED EVIDENCE RETRIEVAL

System: You are a clinical record router responding to predictor questions using the structured patient record.

Task: For each question, retrieve only directly relevant evidence from the EHR.

Instructions:

- Search diagnoses, medications, procedures, observations, laboratory results, and family history.
- Return matching evidence with dates and values when available.
- Prefer exact evidence over interpretation.
- If no relevant evidence is present, state that no relevant data were found.

Input: {`patient_metadata`}, {`clinical_data`}, {`family_history`}, {`predictor_questions`}

Output: Structured responses pairing each question with the retrieved evidence.

B.3. Task Predictor Prompts

PROMPT P-1: QUESTION GENERATION

System: You are a {`disease`} predictor. Your assessment should follow the supplied task guidance.

Task: Generate a small number of targeted questions for the Record Router to retrieve the evidence needed to assess 12-month incident risk of {disease}.

Instructions:

- Prioritize primary diagnostic measurements, key symptoms, and major risk factors.
- Ask only questions that can be answered from the structured EHR.
- Do not score risk or interpret evidence at this stage.
- Avoid repeating questions from prior rounds.

Input: {patient_summary}, {prior_questions}, [{reviewer_feedback}], {task_guidance}

Output: A short list of targeted evidence requests.

PROMPT P-2: RISK ASSESSMENT

System: You are a {disease} predictor. Use only the supplied task guidance and the Record Router-provided evidence.

Task: Assess the probability that the patient will receive a *new* {disease} diagnosis within 1 year after the index date.

Instructions:

- Ground the assessment in task-relevant thresholds, recency, and trends.
- Distinguish incident risk from pre-existing disease.
- Treat missing tests as missing evidence, not negative findings.
- When evidence is limited, state that uncertainty explicitly.

Input: {patient_summary}, {router_responses}, {task_guidance}

Output: A structured risk profile containing a risk score, key supporting evidence, countervailing evidence, and a short rationale.

PROMPT P-3: POST-REVIEW UPDATE

Task: Revise the predictor assessment after reviewing the Reviewer feedback.

Instructions:

- Address hard errors and important soft concerns raised by the Reviewer.
- Ask additional questions only if new evidence is needed.
- Do not repeat previously asked questions.
- If no additional information is needed, update the assessment directly.

Input: {patient_summary}, {reviewer_feedback}, {prior_questions}, {router_responses}

Output: Either a short list of new questions or an updated structured risk profile.

B.4. Reviewer Prompts

PROMPT V-1: REVIEW

System: You are the Reviewer in a structured evidence-routing {disease} risk prediction system.

Task: Evaluate whether the current assessment is evidence-grounded, temporally appropriate, and aligned with the incident prediction task.

Instructions:

- Identify unsupported claims, temporal mismatches, and incident-versus-prevalent errors.
- Recommend only small score adjustments unless a clear hard error is present.
- Prefer no change when the assessment is coherent and evidence-grounded.

Input: {current_risk_profile}, {qa_history}

Output: A short structured review containing a summary, hard errors, soft concerns, and an optional score adjustment recommendation.

PROMPT V-2: FINAL PREDICTION SUMMARY

Task: Summarize the final disease-specific prediction after refinement has concluded.

Instructions:

- Base the summary only on the final predictor risk profile.
- Report the final score for the active disease.
- Keep the rationale concise and evidence-based.

Input: {predictor_profile}

Output: A short rationale summarizing the final prediction.

C. EHRSHOT Cohort Construction and Preprocessing

This appendix summarizes the preprocessing steps used for the EHRSHOT evaluation. The goal was to reproduce the benchmark setting as closely as possible while converting the structured EHR stream into the longitudinal input format required by the structured evidence-routing pipeline.

C.1. Benchmark Protocol and Test-Set Selection

We follow the official EHRSHOT task definitions and patient-ID train/validation/test splits (Wornow et al., 2023). We use the latest-label evaluation setting of Gao et al. (2025): because EHRSHOT labels are assigned at the visit level, a patient may have multiple eligible prediction times for the same task, and we retain only the most recent eligible label per patient.

For held-out test evaluation, we load the official test patient IDs, filter the task-specific labeled-patients file to positive and negative examples from that split, and select the row with the latest prediction time for each patient. This produces one benchmark test instance per patient per task. Training patients are used only for supervised-readout training and model selection, as described in Appendix D; held-out test labels are not used during readout training or selection.

C.2. Temporal Cutoff and Longitudinal Input Construction

For every retained patient, only EHR events observed on or before the benchmark prediction time were included in the model input.

The patient record was converted into a chronological encounter-style representation. Events were grouped by calendar date and organized into clinically interpretable modalities, including diagnoses, medications, procedures, measurements, observations, notes, visit details, and device exposures. Within each calendar date, duplicate same-day concept occurrences were collapsed so that the resulting record preserved temporal structure without redundant repetition.

The final output of preprocessing was a patient-level JSON record containing: (i) demographics, (ii) task metadata including prediction time and label, and (iii) a time-ordered list of pre-prediction encounters. This representation preserves the full structured longitudinal history available before the indexed benchmark time while remaining compatible with router-mediated summarization and targeted evidence retrieval.

C.3. Code-to-Description Enrichment

EHRSHOT represents clinical events using structured <SYSTEM>/<CODE> identifiers spanning SNOMED CT, LOINC, RxNorm, and CPT4. Before ingestion into the inference pipeline, these coded fields were enriched with human-readable descriptions using static terminology lookup tables, so that diagnoses, medications, procedures, laboratory tests, and observations were presented with both the original code and an interpretable descriptor.

C.4. Preprocessing Choices Relevant to Interpretation

Three preprocessing choices are especially important for interpreting the EHRSHOT results. First, we used the benchmark's official held-out test patients rather than constructing a new split, so the reported results are directly comparable to prior work following the same protocol. Second, we adopted the latest-label setting, meaning that each patient contributes at most one prediction time per task. Third, unlike count-based benchmark pipelines that operate on ontology roll-ups or restricted modality sets, the proposed method consumes the full pre-prediction structured event stream and relies on the Record Router to organize that evidence for downstream prediction.

D. Supervised Readout over Routed Evidence

The supervised-readout variant uses a classifier over the routed evidence representation produced by the structured evidence-routing framework. This readout is included because the primary EHRSHOT baselines used for comparison, including count-based models and CLMBR, are also trained on task labels. The goal is to compare patient representations under a shared supervised task-adaptation setting.

For each patient in the EHRSHOT training split, we collected the final framework output, including the disease-specific prediction summary, the final round of task-specific reasoning, and the top evidence drivers identified by the system. This information was condensed into a short clinical evidence summary using a GPT-4.1-mini-based extractor. The summaries were embedded into dense vector representations using PubMedBERT embeddings. A supervised classifier head was then trained over these features using XGBoost, with model selection performed by 5-fold stratified cross-validation on the training split.

The selected classifier was applied to the held-out benchmark test set to produce prediction scores for AUROC and AUPRC evaluation. Test labels were not used during readout training or model selection.

E. Clinical Guideline Sources for Task Guidance

For each incident diagnosis task, the disease-specific predictor receives a short task guidance prompt derived from established clinical guidelines summarized in Table E. These references are used to define relevant risk factors, supporting evidence, and temporal patterns for evidence retrieval and risk assessment.

Table 4. Clinical guideline and reference sources used for disease-specific task guidance.

Condition	Guideline / reference
Hypertension	2025 AHA/ACC Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults (Jones et al., 2025)
Hyperlipidemia	2026 AHA/ACC/Multisociety Guideline on the Management of Dyslipidemia (Blumenthal et al., 2026)
Lupus	2019 EULAR/ACR Classification Criteria for Systemic Lupus Erythematosus (Aringer et al., 2019)
Acute MI	Fourth Universal Definition of Myocardial Infarction (2018) (Thygesen et al., 2018)
Pancreatic cancer	ESMO Clinical Practice Guideline for Diagnosis, Treatment and Follow-up (Conroy et al., 2023)