CANCEREMBED: LEARNING TEMPORAL PATIENT REPRESENTATIONS WITH COMORBIDITY-GUIDED GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Identifying the primary site of cancer from longitudinal electronic health records (EHRs) is challenging due to heterogeneous clinical evidence, irregular trajectories, and nonspecific phenotypes. We introduce *CancerEmbed*, a representation-learning framework that formulates primary-site identification as trajectory-aware patient retrieval over structured and unstructured EHRs. CancerEmbed combines a diagnosis-relative temporal Transformer that integrates summarized clinical notes with structured context, cluster-specific sparse attention (CSA) that extracts cancer-specific evidence directly from full clinical notes, and comorbidity-guided deep metric learning (DML) that structures patient representations using phenotypic similarity beyond cancer labels. Across five cancer types, CancerEmbed demonstrates strong generalization to an independent cohort of diagnostically challenging cases. The temporal Transformer with SupCon achieves 86.8% H@5 and 61.1% P@5, improving over the same representation without DML by 28.3 and 15.1 percentage points, respectively, and reaches 64.2% classification accuracy and 45.4% macro-F1. Complementarily, CSA achieves 81.1% H@5 and 53.6% P@5 without DML using full clinical notes alone, without structured EHR inputs at inference. These results show that jointly modeling temporal, semantic, and phenotypic similarity improves generalization under distribution shift and that metric-learning gains depend on the underlying representation. More broadly, *CancerEmbed* provides a framework for organizing heterogeneous longitudinal clinical data into meaningful patient neighborhoods for retrieval-based reasoning.

1 INTRODUCTION

Cancer is a leading cause of premature death and disability worldwide (Organization, 2026), with new cases and deaths projected to reach 35.3 million and 18.5 million, respectively, in 2050 (Bizuayehu et al., 2024). As this burden grows, shortening the interval between initial presentation and definitive diagnosis has become an important lever for improving care (Neal et al., 2015). This interval is often prolonged because a new clinical or radiological finding carries a broad differential diagnosis, spanning malignancy, infectious, inflammatory, and benign conditions (Mohseni et al., 2014; Dufay et al., 2012), and because many cancers have symptoms that do not point to a specific organ (Jensen et al., 2022). Even once malignancy is established, identifying the primary tumor site remains essential, as it determines staging, prognosis, and the choice of therapy (Liu et al., 2024; Krämer et al., 2024).

This difficulty is not confined to rare presentations, as roughly 15% of cancers arrive in the clinic as metastatic disease (Greco et al., 2012). Nearly 3% of cancer diagnosis in the US (American Cancer Society, 2026) are classified as cancers of unknown primary (CUP) with median survival of a few months (Urban et al., 2013). Patients would typically have to undergo protracted investigation before a diagnosis is confirmed (National Institute for Health and Care Excellence, 2010). To resolve uncertain primary sites, clinicians integrate evidence across longitudinal EHRs and reasoning from similar prior cases. Computationally, this poses three challenges: **heterogeneous evidence** across structured and unstructured data; **irregular trajectories**, where patients are observed at different disease stages and intervals; and **clinically meaningful similarity**, which must capture not only diagnosis but also disease trajectory and comorbidity context.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Existing approaches address these challenges only partially, requiring additional procedures and operating at a single time point, thereby disregarding a patient's longitudinal history (Lu et al., 2021; Moon et al., 2023). Approaches using representation-learning models for EHR, capture rich longitudinal context, however they are optimized for outcome prediction rather than patient similarity (Li et al., 2020; Rasmy et al., 2021; Pang et al., 2021; Steinberg et al., 2024). Deep metric learning (DML) offers a direct route to retrieval, however it disregards temporal or phenotypic structure that clinicians depend on, while earlier patient similarity techniques depend on static features (Sharafoddini et al., 2017).

Therefore, we propose *CancerEmbed*, a framework that recasts primary-site identification as trajectory-aware patient retrieval. By comparing evolving clinical trajectories across longitudinal EHRs, CancerEmbed characterizes diagnostically uncertain patients against prior precedents. It uses a Transformer with diagnosis-relative temporal encoding to align events across varying disease paces, structured via comorbidity-guided deep metric learning to capture phenotypic, semantic, and temporal similarities. The resulting representations group patients with shared progression patterns, providing interpretable precedents to inform site-of-origin prediction.

In summary, our contributions are as follows:

(1) Formulated primary-site identification as **trajectory-aware patient retrieval over longitudinal EHRs**, modeling patient similarity jointly across semantic, temporal, and phenotypic dimensions.

(2) Introduced a **diagnosis-relative temporal Transformer** that represents irregularly sampled encounters according to their temporal proximity to the index diagnosis, enabling disease-stage-aware patient representations.

(3) Proposed a **cluster-specific sparse-attention framework** that distinguishes shared clinical evidence from primary-site-specific evidence, producing interpretable, cancer-conditioned patient representations with source-level traceability.

(4) Developed a **comorbidity-guided deep metric learning framework** that uses phenotypic similarity to organize the patient embedding space beyond conventional single-label supervision.

(5) Validated the proposed framework across five primary cancer sites and an independent **external cohort of diagnostically challenging cases**.

2 Related Work

Cancer Primary-Site Identification. Accurate primary-site identification is critical for treatment selection and prognosis but remains challenging when presentations are metastatic, nonspecific, or lack pathological evidence, often leading to CUP classification. Recent computational methods infer tumor origin using whole-slide histopathology (Lu et al., 2021) or genomic signatures (Jiao et al., 2020; Penson et al., 2020; Moon et al., 2023). However, these specialized data are often unavailable during early diagnostic evaluation. In contrast, longitudinal EHRs provide routinely collected evidence, such as symptoms, comorbidities, and disease progression, whose temporal patterns reveal shared trajectories. Leveraging these longitudinal patterns for patient retrieval and primary-site identification remains underexplored.

Temporal Modeling of Longitudinal EHRs and Clinical Notes. Longitudinal EHRs contain irregular sequences of diagnoses, procedures, medications, laboratory results, and clinical encounters. For longitudinal EHR representation learning, models such as BEHRT represent diagnoses augmented with visit, age, and positional information (Li et al., 2020), while GT-BEHRT capture relationships among medical concepts and visits (Poulain & Beheshti, 2024). On the other hand, clinical notes provide essential context such as symptoms, expert-diagnostic reasoning, evolving clinical impressions, and uncertainty that may help distinguish otherwise similar presentations. However, their length, redundancy, and longitudinal complexity make them difficult to model directly. Recent work has explored large language models (LLM) for generating concise summaries of longitudinal clinical text (Hegselmann et al., 2024). Yet limited work jointly contextualizes narrative notes with structured EHR information at the encounter level and aggregates these representations across an extended pre-diagnosis trajectory. Our study addresses this gap by constructing encounter-level representations from summarized clinical notes and contextualized EHR data and evaluating temporal aggregation strategies for clinically meaningful patient retrieval.

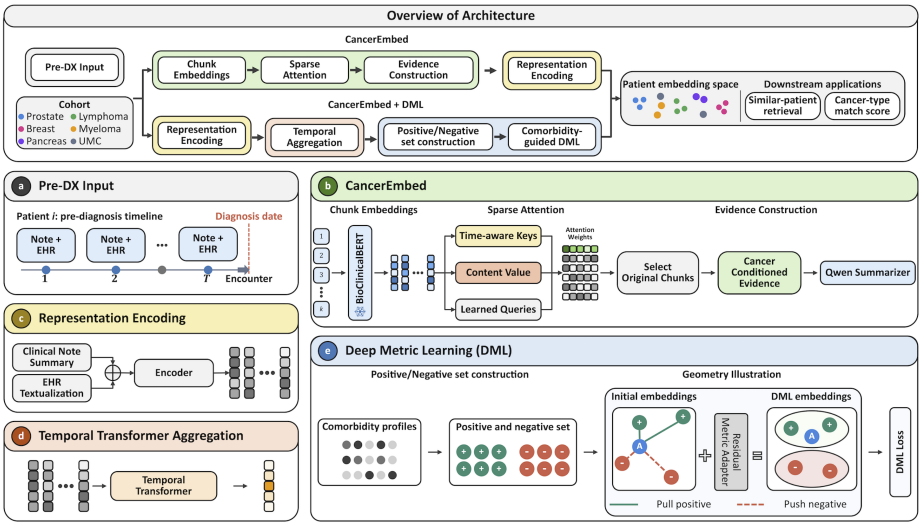


Figure 1: **Overview of CancerEmbed (CSA) and CancerEmbed+DML (Temporal Transformer + SupCon).** (a) Pre-diagnosis clinical notes and EHR data are collected longitudinally. CancerEmbed uses (b) sparse attention to construct cancer-conditioned evidence and (c) encodes the evidence summary with textualized EHR data. CancerEmbed+DML further (d) aggregates encounter representations with a temporal Transformer and (e) applies comorbidity-guided deep metric learning to refine patient embeddings for similar-patient retrieval and cancer-type match scoring.

Sparse and Cluster-Specific Attention. Sparse-attention methods restrict interactions to improve efficiency and reduce irrelevant information; for example, BigBird combines local, global, and random attention patterns (Zaheer et al., 2020), while clustered attention represents groups of related queries using centroids (Vyas et al., 2020). In clinical modeling, GRASP incorporates cohort-level information from similar patients (Zhang et al., 2021). Unlike these approaches, Cluster-Specific Sparse Attention (CSA) learns distinct sparse aggregation patterns across clinical clusters. We evaluate CSA as an alternative to global temporal aggregation and examine its interaction with metric learning.

Deep Metric Learning for Clinical Representation and Retrieval. Retrieval-oriented representation learning has progressed from triplet objectives such as FaceNet (Schroff et al., 2015) to supervised contrastive learning with multiple positives (Khosla et al., 2020) and ranking-aware objectives such as FastAP (Cakir et al., 2019). More recent methods incorporate neighborhood structure and contextual similarity to improve retrieval quality (Liao et al., 2023). However, performance remains sensitive to representations, sampling, and evaluation protocols (Musgrave et al., 2020). Patient-similarity methods have likewise evolved from clinical variables to learned representations over multimodal EHRs, including recent graph-based approaches that model relationships among clinical concepts (Conroy et al., 2017; Huang et al., 2023). Yet similarity is often defined by task-specific labels or outcomes, leaving semantic, temporal, and phenotypic relationships underexplored. Moreover, how metric-learning objectives interact with temporal EHR aggregation for patient retrieval remains poorly understood.

3 METHOD

Figure 1 summarizes CANCEREMBED, which uses clinical information from the 100 days preceding, but excluding, the index diagnosis. The framework contains two complementary pathways. The **temporal pathway** aggregates longitudinal encounters using time-to-diagnosis information, whereas the **cluster-specific sparse-attention (CSA) pathway** separates shared context from evidence specific to five cancer types: breast cancer, lymphoma, prostate cancer, pancreatic cancer, and myeloma. All cancer-specific branches are evaluated without requiring labels at inference. Rep-

resentations from either pathway can be refined through comorbidity-guided metric learning and matched against a reference gallery using cosine similarity for retrieval and primary-site prediction.

3.1 CANCEREMBED–CLUSTER-SPECIFIC SPARSE ATTENTION (CSA)

Temporally contextualized chunks. Eligible pre-cutoff notes are divided into overlapping chunks $\{\chi_{i,m}\}_{m=1}^{M_i}$. A frozen BioClinicalBERT encoder (Alsentzer et al., 2019) produces a mean-pooled embedding $\mathbf{h}_{i,m} = f_{\text{BioClinicalBERT}}(\chi_{i,m})$ for each chunk. Let $\widetilde{\delta}_{i,m}$ denote its normalized time to the patient-specific cutoff. Each embedding is projected into a time-aware key and a clinical-content value: $\mathbf{k}_{i,m} = \text{LN}\Big(\mathbf{W}_k\mathbf{h}_{i,m} + g_{\text{time}}^{\text{chunk}}(\widetilde{\delta}_{i,m})\Big)$, $\mathbf{v}_{i,m} = \mathbf{W}_v\mathbf{h}_{i,m}$. The key determines attention by combining clinical content with temporal proximity, whereas the value provides the content aggregated by the resulting attention weights.

Shared–specific routing. We learn a shared query q^{sh} and cancer-specific queries q_c^{sp} to aggregate representations from chunk keys K_i and values V_i via masked sparse attention (Peters et al., 2019):

$$r_i^{\text{sh}} = 1.5\text{-entmax}\Big(q^{\text{sh}}K_i^\top/\sqrt{d}\Big)V_i, \qquad r_{ic}^{\text{sp}} = 1.5\text{-entmax}\Big(q_c^{\text{sp}}K_i^\top/\sqrt{d}\Big)V_i. \tag{1}$$

The specific attention weights are normalized independently for each candidate cancer, excluding padded positions via a validity mask. Unlike softmax, 1.5-entmax assigns exact zero mass to irrelevant chunks while remaining differentiable, enabling different hypotheses to focus on distinct evidence within the same record. We form the candidate-conditioned representation by blending shared clinical context with hypothesis-specific evidence: $z_{ic} = \text{nrm}\big(\lambda_{\text{mix}}r_i^{\text{sh}} + (1-\lambda_{\text{mix}})r_{ic}^{\text{sp}}\big) = \frac{\lambda_{\text{mix}}r_i^{\text{sh}} + (1-\lambda_{\text{mix}})r_{ic}^{\text{sp}}}{\left\|\lambda_{\text{mix}}r_i^{\text{sh}} + (1-\lambda_{\text{mix}})r_{ic}^{\text{sp}}\right\|_2}$. Finally, classification logits are computed via cosine similarity between z_{ic} and a learned normalized prototype $\bar{p}_c$: $\ell_{ic} = z_{ic}^\top\bar{p}_c/\tau$. Thus, logits are generated by separately routed views of the patient record, each measured against its matching prototype.

Selector learning. The primary objective is class-weighted cross-entropy $\mathcal{L}_{\text{cls}}$, with class weights $w_c \propto n_c^{-1/2}$. An auxiliary specific-only loss $\mathcal{L}_{\text{sp}}$ applies the same objective to $\text{nrm}(r_{ic}^{\text{sp}})^\top\bar{p}_c/\tau$, encouraging conditioned queries to remain predictive without the shared channel. We set $\lambda_{\text{mix}} = 0.5$ to weight the shared and cancer-specific representations equally. To regularize prototype geometry, we compute fixed reference centroids from the frozen embeddings of the internal training patients: $\widetilde{\boldsymbol{\mu}}_c = \frac{1}{|\mathcal{I}_c^{\text{train}}|}\sum_{i\in\mathcal{I}_c^{\text{train}}}\mathbf{z}_i^{\text{ref}}$, $\boldsymbol{\mu}_c = \frac{\widetilde{\boldsymbol{\mu}}_c}{\|\widetilde{\boldsymbol{\mu}}_c\|_2}$. The target relation matrix $R_{cc'} = \boldsymbol{\mu}_c^\top\boldsymbol{\mu}_{c'}$ measures the cosine similarity between cancer types. We encourage the learned prototypes to preserve this geometry through $\mathcal{L}_{\text{rel}} = \frac{1}{C(C-1)}\sum_{c\neq c'}\left(\bar{p}_c^\top\bar{p}_{c'} - R_{cc'}\right)^2$. To control router evidence usage, let $\boldsymbol{\alpha}$ denote an attention vector satisfying $\alpha_t \geq 0$ and $\sum_t \alpha_t = 1$. Its effective support is measured by the participation ratio $N_{\text{eff}}(\boldsymbol{\alpha}) = \left(\sum_{t=1}^T \alpha_t^2\right)^{-1}$. Given the budget $B = 4$, we penalize only attention distributions whose effective support exceeds the budget: $\phi_B(\boldsymbol{\alpha}) = B^{-2}\left[N_{\text{eff}}(\boldsymbol{\alpha}) - B\right]_+^2$, $[x]_+ = \max(x, 0)$. The budget loss averages this penalty over the shared and cancer-specific routers: $\mathcal{L}_{\text{budget}} = \frac{1}{N}\sum_i \phi_B\big(\boldsymbol{\alpha}_i^{\text{sh}}\big) + \frac{1}{NC}\sum_{i=1}^N\sum_{c=1}^C \phi_B(\boldsymbol{\alpha}_{ic}^{\text{sp}})$. The final selector objective is

$$\mathcal{L}_{\text{selector}} = w_{\text{cls}}\mathcal{L}_{\text{cls}} + w_{\text{sp}}\mathcal{L}_{\text{sp}} + w_{\text{rel}}\mathcal{L}_{\text{rel}} + w_{\text{budget}}\mathcal{L}_{\text{budget}}, \tag{2}$$

where the nonnegative weights balance the four components. Attention remains differentiable during training, and hard evidence selection is applied only after optimization.

Auditable evidence construction and encoding. With routing parameters frozen post-training, patient i and hypothesis c yield an evidence set $\mathcal{E}_{ic} = \mathcal{E}_i^{\text{sh}} \cup \mathcal{E}_{ic}^{\text{sp}}$, comprising up to four shared (identical across branches) and four hypothesis-specific chunks. Chunks are ordered chronologically, pruning lowest-ranked specific chunks first if exceeding a 2,048-token limit. A fine-tuned Qwen model jointly summarizes these chunks for each view using the prompt in Appendix A—differing from the temporal pathway's per-encounter summarization. The prompt includes only raw text, dates, and opaque IDs, excluding cancer labels and attention weights; its output preserves supported findings, negation, timing, uncertainty, and source citations without inferring diagnoses. Utilizing

the structured-EHR context $r_{i,t}$, described in further detail in Section 3.2, we form $p_{ic} = s_{ic} \,\|\, r_i$. A frozen MedEmbed encoder maps p_{ic} to $\mathbf{u}_{ic} = f_{\text{enc}}(p_{ic}) \in \mathbb{R}^{768}$, producing five candidate embeddings per patient.

3.2 CancerEmbed - Temporal Comorbidity-Guided pathway.

Pre-Diagnosis Encounter Encoding For patient i, let $t \in \{1, \ldots, T\}$ index the eligible pre-diagnosis encounters in chronological order As shown in Figure 1(a), the note is summarized , while the structured variables are converted into a textual description . The two are concatenated and encoded by a frozen MedEmbed encoder, the resulting sequence $(\mathbf{e}_{i,1}, \ldots, \mathbf{e}_{i,T_i})$ is then passed to the temporal aggregation module.

Clinical-note summarization. Clinical notes contain substantial information unrelated to primary-site identification. We therefore use an open-source, fine-tuned Qwen-based medical summarization model[1] to extract a concise, task-relevant summary ($\widetilde{n}_{i,t}$) for each encounter. Full prompting, decoding, and preprocessing details are provided in Appendix A.

Structured EHR textualization. For each encounter, we construct a structured clinical context from the most recent values within the preceding year for tobacco use, alcohol use, employment status, depression screening, insurance class, and body mass index. Raw values are deterministically mapped to normalized categories, with missing and implausible values represented explicitly. We serialize these features using a fixed template and transform them into a concise natural-language representation ($r_{i,t}$) using a frozen T5-based paraphrasing model (Raffel et al., 2020).

Temporal Transformer and optimization. Figure 1(d) illustrates the temporal transformer module. We normalize pre-diagnosis interval as $\widetilde{\Delta}_{i,t} = \Delta_{i,t}/W$, where $W = 100$ days.

During training, the encounter encoder MedEmbed remains frozen. We optimized temporal Transformer with two auxiliary prediction heads jointly. The first classification head g_{cls} predicts the cancer label y_i, whereas the second regression head g_{fut} predicts a post-diagnosis target representation $\mathbf{z}_i^{\text{post}}$. For a minibatch $\mathcal{B}$, the temporal classification loss is $\mathcal{L}_{\text{temp-cls}} = |\mathcal{B}|^{-1} \sum_{i \in \mathcal{B}} \text{CE}(g_{\text{cls}}(\mathbf{z}_i^{\text{trans}}), y_i)$. Let $\mathcal{B}_{\text{post}} \subseteq \mathcal{B}$ denote the patients with an eligible post-diagnosis target. The future-representation loss is $\mathcal{L}_{\text{fut}} = \max(1, |\mathcal{B}_{\text{post}}|)^{-1} \sum_{i \in \mathcal{B}_{\text{post}}} \left[1 - \cos\big(g_{\text{fut}}(\mathbf{z}_i^{\text{trans}}), \mathbf{z}_i^{\text{post}}\big)\right]$. The temporal-training objective is

$$\mathcal{L}_{\text{temp}} = \mathcal{L}_{\text{temp-cls}} + \lambda_{\text{fut}} \mathcal{L}_{\text{fut}} \tag{3}$$

The post-diagnosis representation is used only as an auxiliary training target and is never provided as input to the temporal Transformer. If a minibatch contains no eligible post-diagnosis target, we set $\mathcal{L}_{\text{fut}} = 0$. After training, both auxiliary heads are discarded, and the temporal Transformer is frozen before fitting the residual metric adapter.

Comorbidity-Guided Deep Metric Learning Cancer-type supervision separates cancer cohorts but does not explicitly preserve local similarity among patients with comparable clinical histories. We therefore use DML to refine representations produced by temporal Transformer. Let $\mathbf{z}_i^{\text{init}}$ denote the initial representation of patient i. The same residual adapter architecture and DML objectives are applied independently to each aggregator, without using cancer-type labels to define patient relations.

Comorbidity-derived relations. For each patient, we collect ICD-10 codes recorded during the year preceding, but excluding, the index diagnosis and map them to an 18-dimensional multi-hot comorbidity vector. After adding demographic variables, this yields a clinical profile $\mathbf{c}_i \in \{0, 1\}^{24}$. Patient similarity is measured by the Jaccard index, $S_{ij} = \frac{\mathbf{c}_i^\top \mathbf{c}_j}{\|\mathbf{c}_i\|_1 + \|\mathbf{c}_j\|_1 - \mathbf{c}_i^\top \mathbf{c}_j}$, we define $\mathcal{P}_i = \{j \in \mathcal{B} \setminus \{i\} : S_{ij} \geq \alpha_+\}$, $\mathcal{N}_i = \{j \in \mathcal{B} \setminus \{i\} : S_{ij} \leq \alpha_-\}$, where $\alpha_+ = 0.45$ and $\alpha_- = 0.25$. Pairs between these thresholds are excluded. Optimization uses valid anchors $\mathcal{V} = \{i \in \mathcal{B} : |\mathcal{P}_i| > 0, |\mathcal{N}_i| > 0\}$; minibatches without valid anchors do not produce a DML update.

[1] Model card: `https://huggingface.co/mradermacher/Qwen2.5-7B-Instruct-medical_summary_latest-GGUF`.

Residual metric adapter. We first normalize the initial representation as $\bar{\mathbf{z}}_i^{\text{init}} = \text{nrm}(\mathbf{z}_i^{\text{init}})$. Rather than fine-tuning the full aggregator, we learn $\Delta\mathbf{z}_i = \gamma f_\theta\big(\text{LN}_\phi(\bar{\mathbf{z}}_i^{\text{init}})\big)$, $\mathbf{z}_i^{\text{dml}} = \text{nrm}\big(\bar{\mathbf{z}}_i^{\text{init}} + \Delta\mathbf{z}_i\big)$. Here, f_θ is a two-layer MLP with GELU activation and dimensions $768 \to 256 \to 768$, and $\gamma = 0.05$ controls the residual magnitude. Its output layer is initialized to zero, allowing the adapter to refine the original embedding geometry gradually. A separate adapter is trained for each aggregator–objective combination. The underlying aggregator remains frozen, and only the adapter and LayerNorm affine parameters are updated.

To limit distortion of the original embedding space, we use an identity penalty $\mathcal{L}_{\text{id}}$ for cosine deviation from the normalized input and a local-geometry penalty $\mathcal{L}_{\text{geom}}$ for changes in pairwise cosine distance among positive pairs. For $\omega \in \{\text{supcon}, \text{triplet}, \text{fastap}\}$, the total loss is $\mathcal{L}_{\text{dml}}^{(\omega)} = \mathcal{L}_\omega + 0.4\mathcal{L}_{\text{id}} + 0.25\mathcal{L}_{\text{geom}}$.

Comorbidity-guided metric learning objective. Our primary DML objective is supervised contrastive learning (SupCon), adapted to the comorbidity-derived relations (Khosla et al., 2020). Let $\mathcal{C}_i = \mathcal{P}_i \cup \mathcal{N}_i$ denote the candidate set for anchor i. We optimize

$$\mathcal{L}_{\text{supcon}} = -\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\frac{1}{|\mathcal{P}_i|}\sum_{p\in\mathcal{P}_i}\log\frac{\exp\big((\mathbf{z}_i^{\text{dml}})^\top\mathbf{z}_p^{\text{dml}}/\tau\big)}{\sum_{a\in\mathcal{C}_i}\exp\big((\mathbf{z}_i^{\text{dml}})^\top\mathbf{z}_a^{\text{dml}}/\tau\big)}, \qquad \tau = 0.1. \tag{4}$$

This objective increases similarity to qualifying positive patients relative to qualifying negatives. Because supervision is derived from pre-diagnosis clinical profiles, patients from different cancer cohorts may form positive pairs, while ambiguous pairs are excluded.

To evaluate sensitivity to the DML objective, we additionally consider triplet loss and FastAP in the ablation study using the same relation sets. For triplet learning, we select the least-similar qualifying positive and the most-similar qualifying negative: $p(i) = \arg\min_{j\in\mathcal{P}_i} S_{ij}$, $n(i) = \arg\max_{j\in\mathcal{N}_i} S_{ij}$. The triplet objective is $\mathcal{L}_{\text{triplet}} = \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\Big[\|\mathbf{z}_i^{\text{dml}} - \mathbf{z}_{p(i)}^{\text{dml}}\|_2 - \|\mathbf{z}_i^{\text{dml}} - \mathbf{z}_{n(i)}^{\text{dml}}\|_2 + \mu\Big]_+$. Thus, triplets are selected near the clinical-relation boundaries rather than according to evolving embedding distances. We also evaluate FastAP (Cakir et al., 2019), which optimizes a differentiable approximation of retrieval average precision: $\mathcal{L}_{\text{fastap}} = 1 - \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\widetilde{\text{AP}}_i$. FastAP uses Euclidean distances between normalized embeddings and ten uniformly spaced histogram intervals over $[0, 2]$. For both alternative objectives, only patients in $\mathcal{C}_i$ are considered, excluding ambiguous pairs.

4 Results

4.1 Dataset Description

Our model was developed and evaluated on a leak-free internal cohort of 11,440 patients spanning five cancer types: breast cancer ($N = 4{,}827$), lymphoma ($N = 881$), prostate cancer ($N = 4{,}548$), pancreatic cancer ($N = 731$), and myeloma ($N = 453$). Patients were partitioned into mutually exclusive training ($N = 7{,}992$), validation ($N = 1{,}162$), and internal test ($N = 2{,}286$) sets. External evaluation was conducted on a locked CUP (Unidentified Mass Clinic - UMC) cohort of 53 diagnostically challenging cases that was not used for model training, hyperparameter tuning, or checkpoint selection. The cohort comprised 3 patients with breast cancer, 40 with lymphoma, 2 with prostate cancer, 3 with pancreatic cancer, and 5 with myeloma with definitive diagnosis. The UMC cohort therefore provides a challenging external evaluation setting, with both diagnostically difficult cases and a substantial shift in cancer-type prevalence relative to the internal cohort, most notably the predominance of lymphoma.

4.2 Quantitative Performance

Classification Performance Table 1 reports five-class classification performance on the internal test and external UMC cohorts. Temporal Transformer achieved strong internal performance without DML, reaching 97.7% accuracy, 95.5% macro-F1, and 99.9% ROC-AUC, but its external accuracy and macro-F1 decreased to 49.1% and 26.9%, respectively. Incorporating comorbidity-guided DML in CancerEmbed (Temporal Transformer + SupCon) substantially improved external performance to

Table 1: Internal and external five-class cancer classification performance. F1, precision, recall, and one-vs-rest ROC-AUC are macro-averaged across the five cancer classes; accuracy is micro-averaged by definition. All entries are percentages.

	Internal reserve set ($N = 2{,}286$)					**External UMC** ($N = 53$)				
Method	Accuracy	F1	ROC-AUC	Precision	Recall	Accuracy	F1	ROC-AUC	Precision	Recall
DeepPhe (Savova et al., 2017)	86.4	75.5	96.9	84.3	70.7	45.3	27.5	**77.9**	31.8	43.5
Qwen3 0.6B (Zhang et al., 2025)	96.5	92.7	99.7	94.4	91.3	60.4	40.0	75.2	41.1	48.5
LLaMa 3.1 8B-Instruct (He et al., 2024)	92.9	82.0	99.3	93.3	77.4	20.8	10.3	54.5	16.1	11.7
Mean	96.0	91.4	99.6	94.6	89.0	54.7	36.0	74.9	34.6	47.0
Time-Decay	96.3	92.4	99.6	95.1	90.3	60.4	38.9	77.3	36.9	48.5
Temporal Transformer	**97.7**	**95.5**	**99.9**	**96.2**	**94.7**	49.1	26.9	74.9	31.6	33.2
CancerEmbed (CSA)	91.0	84.2	99.0	80.8	89.0	30.2	27.3	76.6	34.1	49.3
CancerEmbed (Temporal Transformer + SupCon)	95.5	90.8	99.5	92.2	89.6	**64.2**	**45.4**	75.0	**54.5**	**52.8**
CancerEmbed (CSA + SupCon)	89.7	82.7	98.6	79.3	87.7	20.8	16.4	63.1	27.6	28.2

64.2% accuracy and 45.4% macro-F1—the highest values in the table—while maintaining strong internal performance (95.5% accuracy and 90.8% macro-F1). These results suggest that supervised contrastive learning using comorbidity-derived patient relationships improves the external generalizability of the temporal representations. In contrast, CancerEmbed (CSA), which operates directly on full clinical notes without requiring structured comorbidity features at inference, achieved lower classification accuracy and macro-F1 on both cohorts. Nevertheless, its external ROC-AUC of 76.6% was the highest among the CancerEmbed configurations, suggesting that CSA retained useful cancer-discriminative ranking information even though this information did not translate consistently into correct final class assignments.

Retrieval Performance Table 2 reports similar patient retrieval performance for the internal test set ($N_q = 2{,}286$; $N_g = 9{,}154$) and the external UMC cohort ($N_q = 53$; $N_g = 11{,}440$). While most methods perform strongly on the internal cohort, their differences become pronounced under external evaluation. The temporal Transformer without DML achieves external H@5/H@10 of 58.5%/69.8% and P@5/P@10 of 46.0%/46.8%, substantially outperforming mean pooling and the general-purpose LLM baselines. Notably, CancerEmbed-CSA, which operates directly on full clinical notes without structured EHR inputs, provides the strongest external retrieval without DML, achieving H@5/H@10 of 81.1%/86.8% and P@5/P@10 of 53.6%/52.3%.

Comorbidity-guided DML substantially strengthens the external generalization of trajectory-aware representations. CancerEmbed with the temporal Transformer and SupCon achieves the best overall external performance, reaching 86.8% H@5, 90.6% H@10, 61.1% P@5, and 59.6% P@10. Relative to the Transformer without DML, these represent gains of 28.3, 20.8, 15.1, and 12.8 percentage points, respectively. Importantly, CancerEmbed improves both retrieval coverage and neighborhood purity. For example, DeepPhe attains a relatively high external H@10 of 79.3%, but only 22.6% P@10, indicating that same-cancer patients are often retrieved but constitute a small fraction of the neighborhood. In contrast, CancerEmbed achieves 90.6% H@10 together with 59.6% P@10, retrieving a relevant patient for most queries while maintaining substantially greater same-cancer concentration among the nearest neighbors. The effect of DML is representation-dependent. Adding SupCon to CSA decreases H@5/H@10 from 81.1%/86.8% to 73.6%/83.0%, while modestly increasing P@5/P@10 from 53.6%/52.3% to 55.5%/55.5%. Together, these results highlight complementary strengths within CancerEmbed: CSA extracts transferable cancer-specific evidence directly from full clinical notes, whereas comorbidity-guided DML organizes trajectory-aware representations into neighborhoods with both high coverage and high cancer-specific purity.

4.3 Qualitative Performance

Patient Embedding Space Analysis. Figure 2 visualizes the two CancerEmbed representation pathways before and after comorbidity-guided DML. The temporal Transformer produces strongly cancer-structured representations (Figure 2 (a)), consistent with its auxiliary classification objective. After SupCon (Figure 2 (b)), cancer-specific structure remains visible while increased local mixing suggests incorporation of cross-cancer phenotypic similarity. This reorganization is consistent with the strong external retrieval and classification gains of CancerEmbed (Temporal Transformer + SupCon). CancerEmbed (CSA), which operates directly on full clinical notes without structured

Table 2: Internal and external patient-retrieval performance. $H@K$ is the percentage of queries with at least one same-cancer patient among the top K results, and $P@K$ is the mean percentage of same-cancer patients among the top K results. All entries are percentages.

	Internal Set ($N_q = 2{,}286$)						**External UMC** ($N_q = 53$)					
Method	$H@1$	$P@1$	$H@5$	$P@5$	$H@10$	$P@10$	$H@1$	$P@1$	$H@5$	$P@5$	$H@10$	$P@10$
DeepPhe (Savova et al., 2017)	77.3	77.3	92.8	76.6	95.7	75.8	22.6	22.6	67.9	24.9	79.3	22.6
Qwen3 0.6B (Zhang et al., 2025)	90.6	90.6	96.7	89.3	97.9	88.5	17.0	17.0	35.8	18.1	54.7	17.7
LLaMA 3.1 8B-Instruct (He et al., 2024)	92.7	92.7	97.3	90.1	98.1	88.6	17.0	17.0	35.8	15.8	50.9	15.5
Mean	90.3	90.3	96.2	88.5	97.4	87.7	22.6	22.6	35.8	17.7	45.3	15.8
Time decay	90.7	90.7	96.7	89.3	97.9	88.6	24.5	24.5	37.7	23.0	50.9	19.4
Temporal Transformer	96.9	96.9	98.0	97.0	98.3	97.0	45.3	45.3	58.5	46.0	69.8	46.8
CancerEmbed (CSA)	86.2	86.2	95.4	86.3	96.9	86.1	**56.6**	**56.6**	81.1	53.6	86.8	52.3
Mean + SupCon	83.8	83.8	94.6	80.7	96.8	79.2	47.2	47.2	77.4	42.3	86.8	41.3
Time decay + SupCon	84.7	84.7	95.4	82.3	97.1	81.0	52.8	52.8	71.7	47.9	83.0	47.4
CancerEmbed (Temporal Transformer + SupCon)	95.3	95.3	**98.2**	93.5	**99.0**	92.5	**56.6**	**56.6**	**86.8**	**61.1**	**90.6**	**59.6**
CancerEmbed (CSA + SupCon)	74.2	74.2	86.4	72.6	89.6	71.8	52.8	52.8	73.6	55.5	83.0	55.5

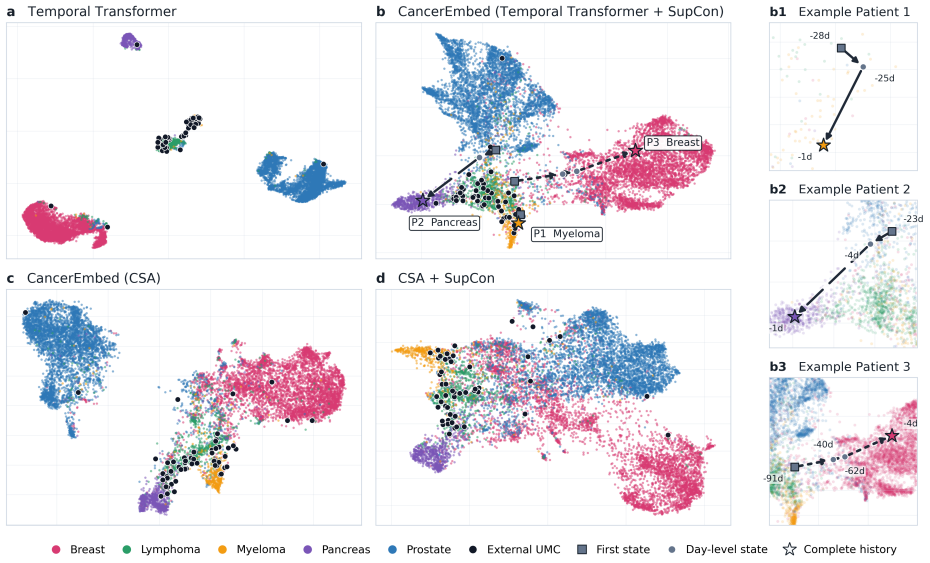


Figure 2: **Patient embedding spaces and patient trajectories.** UMAP projections of internal (colored by cancer type) and external UMC (black) patient embeddings for **(a)** Temporal Transformer, **(b)** CancerEmbed (Temporal Transformer + SupCon), **(c)** CancerEmbed (CSA) and **(d)** CSA + SupCon. **(b1–b3)** Trajectories of three held-out patients (myeloma, pancreas, breast) in the CancerEmbed (Temporal Transformer + SupCon) space, from the first available note (square) to the complete history (star).

comorbidity inputs, exhibits broader and more overlapping cancer-specific regions (Figure 2 (c)) while maintaining strong external retrieval. SupCon further increases cross-cohort mixing (Figure 2(d)), consistent with its modest increase in retrieval precision but reduced hit rate. Overall, these visualizations suggest that the effect of DML depends on the geometry of the underlying representation: SupCon effectively introduces phenotypic structure into the well-separated Transformer space, whereas CSA already captures broader cross-cancer clinical similarity. As UMAP can distort distances and neighborhoods, we interpret these patterns qualitatively alongside Tables 2.

Trajectory of Patient Representations Over Time. Figure 2 illustrates how CancerEmbed (Temporal Transformer + SupCon) representations evolve as clinical evidence accumulates toward diagnosis. Early representations may be influenced by correlated comorbidities and shared clinical characteristics learned via DML, placing patients near regions associated with other cancer types. As additional diagnostic evidence accumulates in clinical documents, the three representative patients progressively move toward their eventual cancer-specific regions: P1 from lymphoma toward myeloma, P2 from the prostate-lymphoma boundary toward pancreatic cancer, and P3 from lym-

Table 3: Ablation of deep metric learning objectives for five-class cancer classification and patient retrieval on the internal reserve set and external UMC cohort. For classification, F1, precision, recall, and one-vs-rest ROC-AUC are macro-averaged across the five cancer classes, while accuracy is micro-averaged. Retrieval is evaluated using hit rate ($H@K$) and precision ($P@K$). All values are percentages. Bold indicates the best result in each column; ties are all highlighted.

(a) Five-class cancer classification

	Internal reserve set ($N = 2{,}286$)					**External UMC** ($N = 53$)				
Method	Accuracy	F1	ROC-AUC	Precision	Recall	Accuracy	F1	ROC-AUC	Precision	Recall
Temporal Transformer + Triplet	**97.7**	**95.5**	**99.9**	**96.0**	**95.1**	47.2	26.6	73.7	31.7	32.7
Cluster-specific sparse attention + Triplet	89.8	82.5	98.6	78.9	88.1	50.9	31.0	71.3	33.2	36.3
Temporal Transformer + FastAP	96.9	93.5	99.8	94.2	92.9	52.8	28.1	67.5	32.6	34.2
Cluster-specific sparse attention + FastAP	91.0	84.5	98.9	81.2	89.0	24.5	27.1	75.2	44.3	41.7
Proposed – CancerEmbed (CSA)	91.0	84.2	99.0	80.8	89.0	30.2	27.3	**76.6**	34.1	49.3
Proposed – CancerEmbed (Temporal Transformer + SupCon)	95.5	90.8	99.5	92.2	89.6	**64.2**	**45.4**	75.0	**54.5**	**52.8**

(b) Retrieval Task

	Internal reserve set ($N_q = 2{,}286$)						**External UMC** ($N_q = 53$)					
Method	$H@1$	$P@1$	$H@5$	$P@5$	$H@10$	$P@10$	$H@1$	$P@1$	$H@5$	$P@5$	$H@10$	$P@10$
Mean + Triplet	90.7	90.7	96.5	88.9	97.4	87.9	24.5	24.5	39.6	18.9	49.1	16.4
Time decay + Triplet	91.1	91.1	97.0	89.5	98.0	88.9	20.8	20.8	39.6	23.0	47.2	20.9
Temporal Transformer + Triplet	**96.9**	**96.9**	**98.2**	96.9	98.6	96.8	39.6	39.6	54.7	43.0	60.4	44.0
Cluster-specific sparse attention + Triplet	77.9	77.9	93.8	77.6	96.7	76.8	45.3	45.3	69.8	40.4	84.9	40.9
Mean + FastAP	89.7	89.7	97.3	88.2	98.4	86.9	45.3	45.3	75.5	37.7	84.9	36.8
Time decay + FastAP	90.5	90.5	97.6	89.2	98.5	88.3	39.6	39.6	75.5	40.8	88.7	43.8
Temporal Transformer + FastAP	**96.9**	**96.9**	98.0	**97.0**	98.4	**97.0**	54.7	54.7	73.6	55.1	81.1	52.6
Cluster-specific sparse attention + FastAP	84.1	84.1	91.7	83.7	93.7	83.2	35.8	35.8	56.6	34.3	64.2	29.4
Proposed – CancerEmbed (CSA)	86.2	86.2	95.4	86.3	96.9	86.1	**56.6**	**56.6**	81.1	53.6	86.8	52.3
Proposed – CancerEmbed (Temporal Transformer + SupCon)	95.3	95.3	**98.2**	93.5	**99.0**	92.5	**56.6**	**56.6**	**86.8**	**61.1**	**90.6**	**59.6**

phoma toward breast cancer. These trajectories illustrate how CancerEmbed balances phenotypic similarity with accumulating disease-specific evidence, dynamically refining patient representations as diagnosis approaches.

4.4 Ablation

As shown in Tables 3, the effects of the DML objectives depend strongly on the underlying patient representation. CancerEmbed (Temporal Transformer + SupCon) achieved the strongest external classification performance (64.2% accuracy and 45.4% macro-F1) and retrieval performance (90.6% H@10 and 59.6% P@10). For the temporal Transformer, Triplet produced no external classification gain and reached only 60.4% H@10, whereas FastAP modestly improved classification accuracy to 52.8% and retrieval to 81.1% H@10 and 52.6% P@10, but remained below SupCon. For CSA, Triplet increased classification accuracy from 30.2% to 50.9% but reduced recall, ROC-AUC, and retrieval quality, while FastAP further reduced accuracy to 24.5% and retrieval to 64.2% H@10 and 29.4% P@10. Because all three objectives use the same residual adapter and comorbidity-derived relations, these results indicate that SupCon most effectively reorganizes the temporal Transformer representation, whereas Triplet and FastAP provide less consistent, representation-dependent benefits across classification and retrieval.

5 Conclusion

We introduced *CancerEmbed*, a framework that reframes primary-site identification as trajectory-aware representation learning and retrieval over longitudinal EHRs. To our knowledge, CancerEmbed is the first framework to learn cancer patient representations that jointly integrate longitudinal structured EHR data and unstructured clinical notes while explicitly modeling temporal and phenotypic similarity. Technically, CancerEmbed combines diagnosis-relative temporal modeling, cancer-conditioned sparse attention, and comorbidity-guided metric learning to structure patient embeddings beyond endpoint labels. External evaluation on diagnostically challenging cases demonstrates the generalizability of these representations, while ablations show that metric-learning gains depend strongly on the underlying representation geometry. CancerEmbed provides a framework for organizing heterogeneous longitudinal clinical data into meaningful patient neighborhoods for retrieval-based reasoning.

AI USE STATEMENT

In this work, we used generative AI tools for improving the manuscript's clarity and grammar, assisting with figure organization, supporting reference formatting, providing editorial feedback on presentation of results. Also, it was used as part of code-implementation. We have not used generative AI tools to generate data, conduct experiments, or generate ideas. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final manuscript and research results.

REFERENCES

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 72–78. Association for Computational Linguistics, 2019. doi: 10.18653/v1/W19-1909.

American Cancer Society. Key statistics for cancers of unknown primary, 2026. URL https://www.cancer.org/cancer/types/cancer-unknown-primary/about/key-statistics.html. Last revised January 14, 2026; accessed September 25, 2026.

Habtamu Mellie Bizuayehu, Kedir Y Ahmed, Getiye Dejenu Kibret, Abel F Dadi, Sewunet Admasu Belachew, Tanmay Bagade, Teketo Kassaw Tegegne, Rebecca L Venchiarutti, Kelemu Tilahun Kibret, Aklilu Habte Hailegebireal, et al. Global disparities of cancer and its projected burden in 2050. *JAMA network open*, 7(11):e2443198, 2024.

Fatih Cakir, Kun He, Xide Xia, Brian Kulis, and Stan Sclaroff. Deep metric learning to rank. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1861–1870. IEEE, 2019.

Bryan Conroy, Minnan Xu-Wilson, and Asif Rahman. Patient similarity using population statistics and multiple kernel learning. In *Machine Learning for Healthcare Conference*, pp. 191–203. PMLR, 2017.

C Dufay, A Abdelli, V Le Pennec, and L Chiche. Mesenteric tumors: diagnosis and treatment. *Journal of visceral surgery*, 149(4):e239–e251, 2012.

FA Greco, K Oien, M Erlander, R Osborne, G Varadhachary, J Bridgewater, D Cohen, and H Wasan. Cancer of unknown primary: progress in the search for improved and rapid diagnosis leading toward superior patient outcomes. *Annals of oncology*, 23(2):298–304, 2012.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, et al. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders. *arXiv preprint arXiv:2410.20526*, 2024.

Stefan Hegselmann, Shannon Zejiang Shen, Florian Gierse, Monica Agrawal, David Sontag, and Xiaoyi Jiang. A data-centric approach to generate faithful and high quality patient summaries with large language models. *arXiv preprint arXiv:2402.15422*, 2024.

Yefan Huang, Feng Luo, Xiaoli Wang, Zhu Di, Bohan Li, and Bin Luo. A one-size-fits-three representation learning framework for patient similarity search. *Data Science and Engineering*, 8(3): 306–317, 2023.

Ellen Jensen, Jette Kolding Kristensen, Rikke Tveden Bjerglund, Søren Paaske Johnsen, and Janus Laust Thomsen. The pathway and characteristics of patients with non-specific symptoms of cancer: a systematic review. *BMC cancer*, 22(1):574, 2022.

Wei Jiao, Gurnit Atwal, Paz Polak, Rosa Karlic, Edwin Cuppen, Alexandra Danyi, Jeroen de Ridder, Carla van Herpen, Martijn P Lolkema, et al. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nature communications*, 11(1):728, 2020.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Alwin Krämer, Tilmann Bochtler, Chantal Pauli, Kai-Keen Shiu, Natalie Cook, Juliana Janoski de Menezes, Roberto A Pazo-Cid, Ferran Losa, Debbie GJ Robbrecht, Jiří Tomášek, et al. Molecularly guided therapy versus chemotherapy after disease control in unfavourable cancer of unknown primary (cupisco): an open-label, randomised, phase 2 study. *The Lancet*, 404(10452): 527–539, 2024.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):7155, 2020.

Christopher Liao, Theodoros Tsiligkaridis, and Brian Kulis. Supervised metric learning to rank for retrieval via contextual similarity optimization. In *International Conference on Machine Learning*, pp. 20906–20938. PMLR, 2023.

Xin Liu, Xiaowei Zhang, Shiyu Jiang, Miao Mo, Qifeng Wang, Yanli Wang, Liangping Zhou, Silong Hu, Huijuan Yang, Yifeng Hou, et al. Site-specific therapy guided by a 90-gene expression assay versus empirical chemotherapy in patients with cancer of unknown primary (fudan cup-001): a randomised controlled trial. *The Lancet Oncology*, 25(8):1092–1102, 2024.

Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, Melissa Zhao, Maha Shady, Jana Lipkova, and Faisal Mahmood. Ai-based pathology predicts origins for cancers of unknown primary. *Nature*, 594(7861):106–110, 2021.

Shahrzad Mohseni, Abolfazl Shojaiefard, Zhamak Khorgami, Shahriar Alinejad, Ali Ghorbani, and Ali Ghafouri. Peripheral lymphadenopathy: approach and diagnostic tools. *Iranian journal of medical sciences*, 39(2 Suppl):158, 2014.

Intae Moon, Jaclyn LoPiccolo, Sylvan C Baca, Lynette M Sholl, Kenneth L Kehl, Michael J Hassett, David Liu, Deborah Schrag, and Alexander Gusev. Machine learning for genetics-based classification and treatment response prediction in cancer of unknown primary. *Nature medicine*, 29(8): 2057–2067, 2023.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *European Conference on Computer Vision*, pp. 681–699. Springer, 2020.

National Institute for Health and Care Excellence. Metastatic malignant disease of unknown primary origin in adults: Diagnosis and management. Clinical Guideline CG104, National Institute for Health and Care Excellence, London, UK, 2010. URL `https://www.nice.org.uk/guidance/cg104`. Updated April 26, 2023; accessed September 25, 2026.

Richard D Neal, Puvanendran Tharmanathan, Barbara France, NU Din, Sian Cotton, Julia Fallon-Ferguson, William Hamilton, Alison Hendry, Martin Hendry, Ruth Lewis, et al. Is increased time to diagnosis and treatment in symptomatic cancer associated with poorer outcomes? systematic review. *British journal of cancer*, 112(1):S92–S107, 2015.

World Health Organization. *Global status report on cancer 2026: the future we choose together*. World Health Organization, 2026.

Chao Pang, Xinzhuo Jiang, Krishna S Kalluri, Matthew Spotnitz, RuiJun Chen, Adler Perotte, and Karthik Natarajan. Cehr-bert: Incorporating temporal information from structured ehr data to improve prediction tasks. In *Machine learning for health*, pp. 239–260. PMLR, 2021.

Alexander Penson, Niedzica Camacho, Youyun Zheng, Anna M Varghese, Hikmat Al-Ahmadie, Pedram Razavi, Sarat Chandarlapaty, Christina E Vallejo, Efsevia Vakiani, Teresa Gilewski, et al. Development of genome-derived tumor type prediction to inform clinical cancer care. *JAMA oncology*, 6(1):84–91, 2020.

Ben Peters, Vlad Niculae, and André F. T. Martins. Sparse sequence-to-sequence models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1504–1519. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1146.

Raphael Poulain and Rahmatollah Beheshti. Graph transformers on ehrs: Better representation improves downstream performance. In *The Twelfth International Conference on Learning Representations*, 2024.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, 2021.

Guergana K Savova, Eugene Tseytlin, Sean Finan, Melissa Castine, Timothy Miller, Olga Medvedeva, David Harris, Harry Hochheiser, Chen Lin, Girish Chavan, et al. Deepphe: a natural language processing system for extracting cancer phenotypes from clinical records. *Cancer research*, 77(21):e115–e118, 2017.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Anis Sharafoddini, Joel A Dubin, and Joon Lee. Patient similarity in prediction models based on health data: a scoping review. *JMIR Medical Informatics*, 5(1):e7, 2017. doi: 10.2196/medinform.6730.

Ethan Steinberg, Jason Fries, Yizhe Xu, and Nigam Shah. Motor: A time-to-event foundation model for structured medical records. In *International Conference on Learning Representations*, volume 2024, pp. 41038–41077, 2024.

Damien Urban, Aparna Rao, Mathias Bressel, Yaacov R Lawrence, and Linda Mileshkin. Cancer of unknown primary: a population-based analysis of temporal change and socioeconomic disparities. *British journal of cancer*, 109(5):1318–1324, 2013.

Apoorv Vyas, Angelos Katharopoulos, and François Fleuret. Fast transformers with clustered attention. *Advances in Neural Information Processing Systems*, 33:21665–21674, 2020.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17283–17297. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c8512d142a2d849725f31a9a7a361ab9-Paper.pdf.

Chaohe Zhang, Xin Gao, Liantao Ma, Yasha Wang, Jiangtao Wang, and Wen Tang. Grasp: generic framework for health status representation learning based on incorporating knowledge from similar patients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 715–723, 2021.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

A CLINICAL-NOTE SUMMARIZATION

Although both representation pathways use the same fine-tuned Qwen medical summarizer, they employ different summarization tasks. In the temporal transformer pathway, Qwen summarizes each encounter note into a short three-to-four-sentence narrative, which is subsequently concatenated with the corresponding textualized EHR variables by the prompt in 3. In contrast, the cluster-specific sparse-attention pathway summarizes a routed evidence view containing shared and cancer-specific chunks selected from multiple longitudinal notes. Its prompt in figure 4 produces four to eight oncology-focused bullet points and requires each statement to end with exact source-chunk citations.

```
SYSTEM:
You are a clinical-note summarizer. Use only information documented in the supplied
↪  clinical note.

USER:
Summarize the following clinical note in three to four concise sentences.

Retain documented findings from pathology, radiology, ancillary studies, and relevant
↪  medical history. Preserve important positive and negative findings, dates, and
↪  uncertainty when they are explicitly stated. Do not introduce unsupported information
↪  or infer a final diagnosis.

If the note contains no relevant oncology or hematology information, return exactly:

"No relevant oncology or hematology information."

Clinical note:
[NOTE_TEXT]
```

Figure 3: **Reconstructed prompt specification for encounter-level clinical-note summarization in the temporal Transformer pathway.** Each eligible encounter note is summarized independently before being concatenated with the corresponding textualized structured EHR variables. The instruction retains documented pathology, radiology, ancillary-study, and relevant medical-history information while discouraging unsupported inferences. Notes without relevant content are mapped to a predefined sentinel sentence. This prompt is reconstructed from the documented summarization protocol because the original executable prompt was not available in the retained codebase.

```
SYSTEM:
You are a careful clinical evidence summarizer.

USER:
Summarize the supplied oncology/hematology clinical evidence in English. Write
4-8 clinically informative bullet points, normally 180-300 words when the
evidence supports that length. Cover all supported information relevant to
oncology/hematology, including presenting symptoms and findings, diagnoses and
pathology, tests or imaging, treatments or procedures, family or genetic history,
risk factors, dates, important negative findings, and uncertainty. Omit categories
that are not supported. Use only facts explicitly stated in the evidence and do
not infer or state a final diagnosis. Each bullet must contain one coherent factual
statement and end with one or more exact source chunk IDs in square brackets. Do
not repeat a chunk ID within the same bullet. If the evidence is genuinely sparse,
use fewer bullets and state that limitation using a relevant citation when
possible. Return bullets only, without a title.

Evidence:
[CHUNK_ID_1] date=NOTE_DATE_1: CHUNK_TEXT_1
[CHUNK_ID_2] date=NOTE_DATE_2: CHUNK_TEXT_2
[CHUNK_ID_3] date=NOTE_DATE_3: CHUNK_TEXT_3
...

Rewrite the complete summary as 4-8 informative cited bullets when evidence
supports it; use fewer only for sparse evidence. Cover supported
symptoms/findings, diagnoses/pathology, tests, treatment, family/genetic history,
risk factors, dates, important negatives, and uncertainty. Omit unsupported
categories. End every bullet with exact allowed chunk IDs. No title, repeated
citation, final diagnosis, or unsupported fact.
```

Figure 4: **Prompt used for cluster-specific sparse-attention evidence summarization.** The same prompt template is used for all five cancer hypotheses; only the routed evidence chunks differ between branches. Cancer labels and attention weights are not included in the prompt.

The same prompt template is used for all five cancer hypotheses; cancer conditioning is introduced solely through the branch-specific evidence supplied to Qwen. Consequently, the cluster-specific prompt provides broader longitudinal coverage and chunk-level traceability, whereas the original prompt produces a general encounter-level summary.

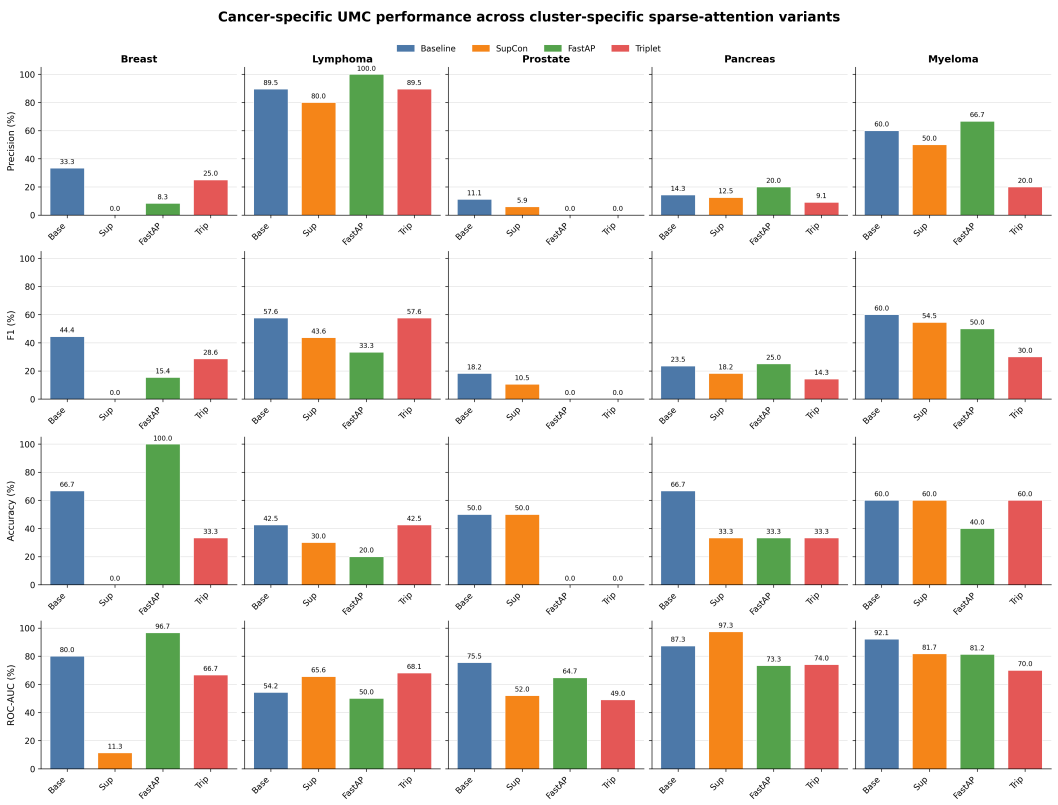


Figure 5: Comparison between different cancer group classification performance

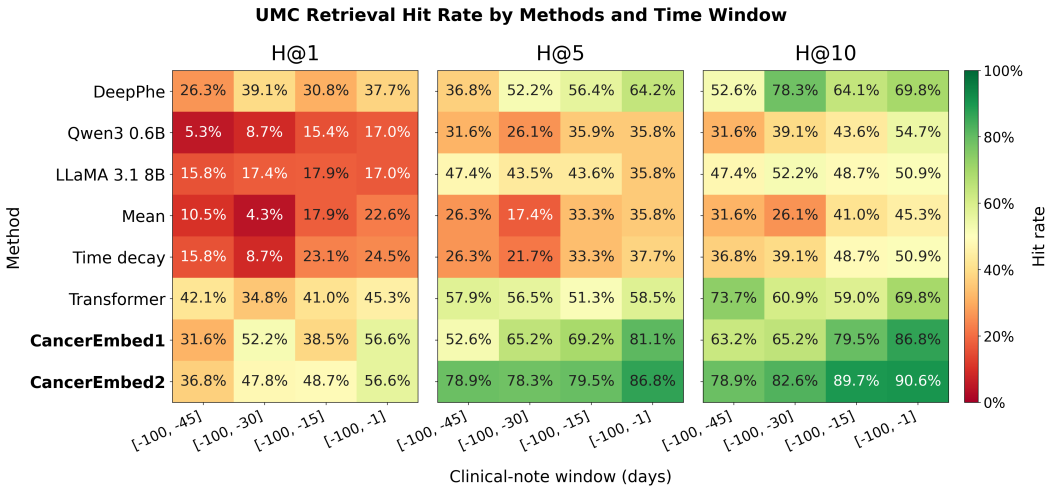
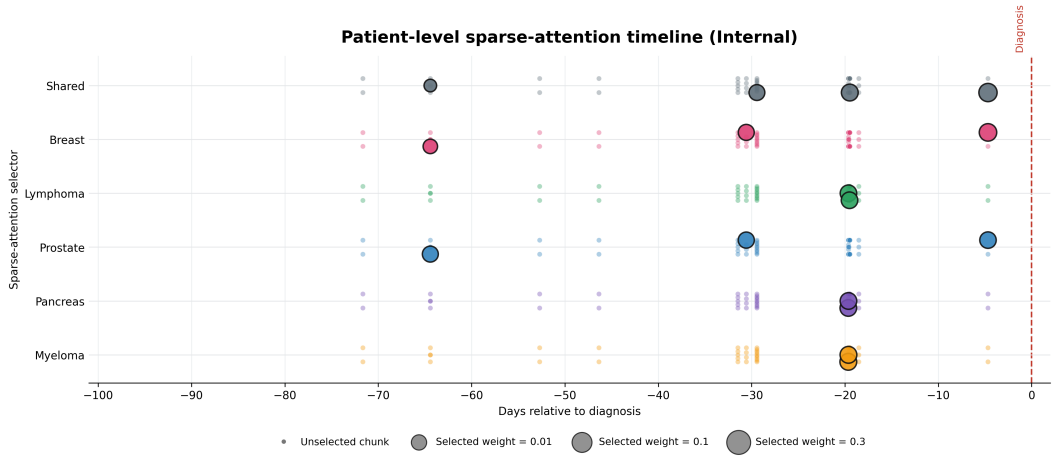


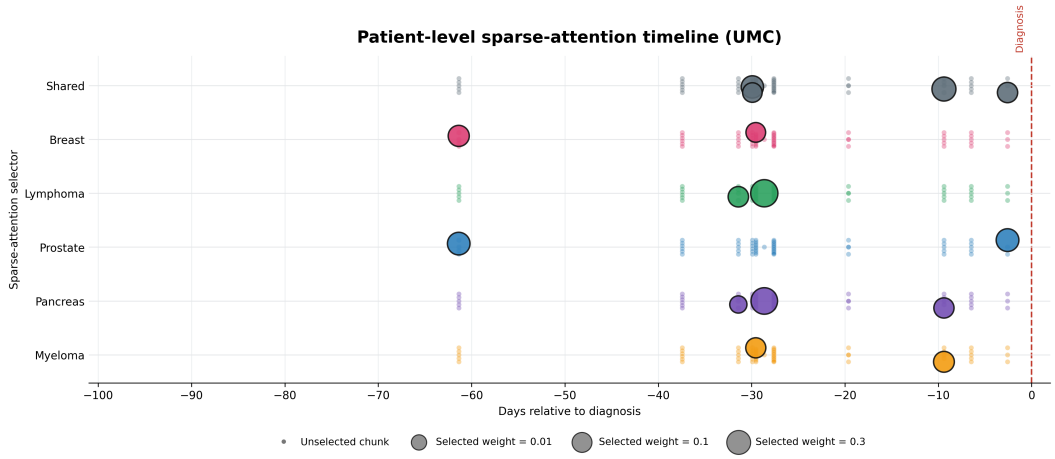
Figure 6: UMC hit rate performance at different time windows [-100, -45], [-100, -30], [-100, -15], [-100, -1]. CancerEmbed1 represents the CancerEmbed (CSA), CancerEmbed2 represents the CancerEmbed (Temporal Transformer + SupCon)

B Quantitative Results Supplementary

B.1 Subgroup Performance for Each Cancer in UMC



(a) Internal test patient with prostate cancer: 54 chunks from 19 clinical notes.



(b) External UMC patient with lymphoma: 71 chunks from 12 clinical notes.

Figure 7: **Patient-level sparse-attention timelines.** Each row represents the shared selector or one of the five cancer-specific selectors. Small points denote eligible chunks that were not retained, whereas large outlined points denote the final evidence passed to Qwen. The size of a selected point is proportional to its Entmax attention weight. The red dashed line marks the diagnosis date.