Multi–Cancer Risk Prediction Using Transformers Trained on Large-Scale Longitudinal EHR Data

Asif Khan[1,8]* **Daniel Ritter**[1,3]* **Chunlei Zheng**[2]

Duncan Forster[1] **Moshir Harsh**[1,5,9] **Artem Gazizov**[1]

Debora S. Marks[1,5] **Nathanael R. Fillmore**[2,4]† **Chris Sander**[1,5,6,7,8] †

[1]Harvard Medical School, Boston, MA, USA
[2]VA Boston Healthcare System, Boston, MA, USA
[3]Department of Computer Science, Cornell University, NY, USA
[4]Boston University School of Medicine, Boston, MA, USA
[5]Broad Institute of MIT and Harvard, Boston, MA, USA
[6]CEDAR Center, Knight Cancer Institute, OHSU, Portland, OR, USA
[7]DF/HCC Cancer Center, Boston, MA, USA
[8]Ludwig Center at Harvard, Boston, MA, USA
[9]Max Planck Institute of Molecular Physiology, Dortmund, Germany

Abstract

Early detection of cancers would be of substantial benefit as many cancers are diagnosed too late. Risk assessment from electronic health records (EHRs) can be used to implement efficient surveillance programs, focusing follow-up care on those patients most likely to benefit from screening and timely intervention. We present an end-to-end, multi-task transformer that predicts risk in discrete-time intervals for multiple cancer types from longitudinal EHR trajectories. The model learns latent representations that reflect shared and cancer-specific features to improve performance across different cancer types. Training on a large EHR dataset from the US Department of Veterans Affairs (US-VA), we evaluate model performance for five cancers using positive predictive value (PPV@N) and standardized incidence ratio (SIR@N) for N high-risk patients, as well as AUPRC, under both no-exclusion and 3-month data exclusion windows. The results show improved performance in high-risk cohorts indicating the model's potential utility as a clinical decision support tool for targeted surveillance of patients.

1 Introduction

Cancer is among the most challenging diseases to treat, owing to its complex biology and the difficulties in early detection. Approximately half of cancers are diagnosed at an advanced stage [1], which significantly diminishes the chances of successful treatment and long-term survival. While national screening programs have improved outcomes, they are limited in scope due to a narrow set of risk factors and often fail to identify patients until it is too late for effective intervention.

Many cancers share common risk factors, either inherited (germline) or acquired (environment, life style, etc). For example, inherited loss-of-function BRCA1/2 mutations predispose primarily

*Equal contribution.
†Co-corresponding authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Learning from Time Series for Health.

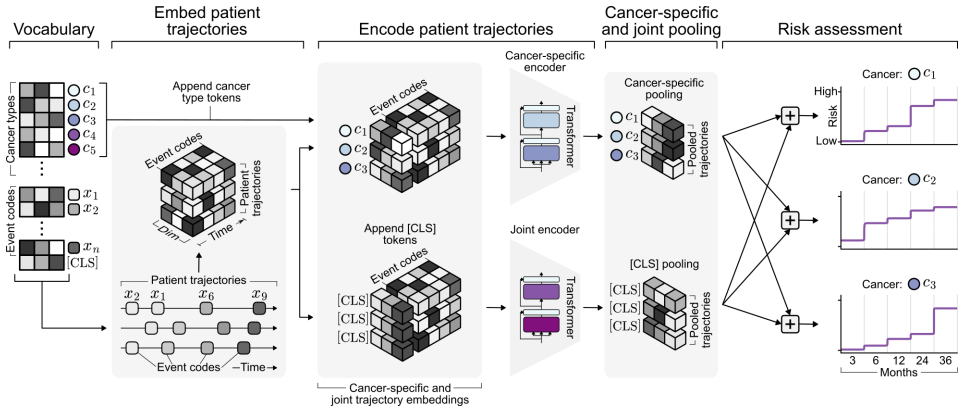

Figure 1: **Multi cancer risk stratification model**. The vocabulary contains cancer type, event code and [CLS] (classification) token embeddings. Patient trajectories go through a cancer-specific encoder to encode cancer-specific latent factors and a joint encoder to capture information from factors common across all cancers. Cancer type and [CLS] tokens are used as the pooled trajectory representations for the cancer-specific and shared encodings, respectively. These representations are combined and passed to a risk assessment head. c_i, cancer type tokens; x_i, event tokens; [CLS], shared pooling token; Dim, feature dimension.

to ovarian and breast cancers, while lifestyle factors such as inflammation, obesity, and smoking, increase risk across multiple cancer types [2, 3]. This interconnected nature of cancer biology highlights the potential for systematic, multi-cancer approaches that jointly learn from shared and cancer-specific features for multi-cancer risk assessment and stratification.

In this work, we present an end-to-end multi-task framework for predicting risk of multiple cancers from longitudinal EHR data. The model uses transformers to extract information from temporal dependencies in diagnosis trajectories to learn both cancer-specific and shared latent representations. We trained the model on the large-scale US Veterans Affairs (US-VA) database, results on five cancer types show the benefit of the multi-cancer risk prediction approach and its clinical utility.

2 Related Work

EHR sequence models. Early sequence models for EHRs used recurrent neural networks (RNNs) and attention mechanisms to model patient visit sequences for next-event [4] and heart failure [5] prediction. More recently, transformer-based architectures such as BEHRT, Med-BERT, G-BERT have demonstrated the benefit of representation learning from EHRs on various downstream clinical outcome tasks [6–8]. Generative style training using large language models (LLMs) also shows the value of pretraining for downstream clinical tasks [9, 10].

Deep time-to-event models. Neural survival models include either a generalization of the continuous time Cox model (DeepSurv [11]) or a discrete time formulation such as DeepHit [12] and Nnet-survival [13]. Transformer based architectures has been explored in these works for a single outcome as well as competing risks, but are not designed for jointly predicting risk of multiple cancer types. The Motor method [14] introduced self-supervised time-to-event pre-training for improved transfer on a broad range of downstream tasks.

Cancer risk models from EHR data. For pancreatic cancer, the Prism method trained logistic regression and MLP models on curated EHR covariates in a multi-source aggregated dataset [15]. Placido et al. [16] used trajectory based deep learning to predict pancreatic cancer risk from diagnosis trajectories, while a further extension used a combination of diagnosis codes and medications [17]. Beyond single cancers, a population-scale registry study used a Bayesian time-dependent Cox model on selected covariates for multi-cancer risk stratification [18].

Our work differs from existing approaches by (a) casting multi-cancer risk assessment as a multi-task learning problem on longitudinal EHR trajectories, (b) explicitly learning cancer-specific and shared

representations via *cancer* task tokens in a transformer encoder, and (c) producing discrete-time risk curves per cancer type for clinically relevant horizons.

3 Method

We propose an end-to-end transformer model for risk prediction of multiple cancer types that learns both shared and cancer-specific latent representations. This formulation improve generalization over single-task models, particularly for cancer types with limited training data. The architecture has three key components: embedding of diagnosis codes, encoding temporal information, and multi-cancer risk prediction. We introduce unique class tokens for each cancer type, which are used in the pooling operation after the transformer encoder to provide cancer specific features. For a shared encoding, a single class token across all cancer types is used to generate the shared representation. These two representations are combined in a multi-head risk prediction model, where each cancer risk is treated as a distinct task with its own time-to-event model, combining both specific and shared features to predict the risk probability curve (Figure 1).

Training Procedure. Each patient's longitudinal trajectory is represented as a sequence of diagnosis codes with associated timestamps. To train the model, we sample multiple sub-trajectories per patient, corresponding to different censoring dates. This provides snapshots of the health state of patients at different time points and serves as a form of data augmentation. We use the difference between the timestamp of each event and the final timestamp to construct the transformer's positional encodings, which capture the irregular spacing of longitudinal events. We also include the patient's age at each event as an additional input due to its strong correlation with cancer incidence. Because the dataset consists of more controls than positive cases, we apply weighted sampling to balance mini-batches during training.

The multi-cancer learning objective is a sum of binary cross-entropy losses computed within each interval, combined across cancer types. Specifically, for each cancer task c, we have a discrete set of follow-up intervals (from the time of assessment, t_a to t), $t \in \{3, 6, 12, 36, 60\}$ months. Let $y^c_{i,t} \in \{0, 1\}$ be a binary indicator if patient i developed cancer c within interval t, and $p^c_{i,t} \in [0, 1]$ be a predicted probability (risk) of cancer diagnosis for patient i, cancer c, at interval t. The training objective is,

$$\mathcal{L}(\hat{p}) = -\sum_{c=1}^{C} \frac{1}{N_c} \sum_{i=1}^{N_c} \sum_{t \in \mathcal{T}} \Big[y^{(c)}_{i,t} \log \hat{p}^{(c)}_{i,t} + \big(1 - y^{(c)}_{i,t}\big) \log \big(1 - \hat{p}^{(c)}_{i,t}\big) \Big], \tag{1}$$

Trajectory Sampling. As there are far more controls (non-cancer) than cancer cases, we use class-ratio weights and balanced sampling to draw an equal number of cancer and control patients in each mini-batch during training. For each patient, we sample k random sub-trajectories with different end-points as a form of data augmentation, allowing the model to learn from patient health states at different time points. For control cases, we ensure patients have at least two years of follow-up after the timestamp of the last event in the sampled trajectory, to rule out false non-cases: patients who may have dropped out of the health care system but later developed cancer. If a patient is diagnosed for more than one cancer type we only predict risk for the first occurrence and exclude the patient from the control set of other cancer types.

4 Dataset and Computational Setup

Dataset. We use longitudinal diagnosis data from the VA Corporate Data Warehouse (CDW) system, which integrates data from all VA medical facilities across the US [19, 20]. The CDW contains inpatient and outpatient encounters; diagnoses are recorded using both ICD9 and ICD10 coding systems. To standardize terminology across coding systems, the data was harmonized to the OMOP common data model [21]. **Cancer labels.** A key challenge in cancer risk assessment is the accuracy of diagnosis labels. In the U.S. healthcare system, ICD diagnosis codes may be assigned for billing or reimbursement purposes, rather than to reflect a patient's actual health state. To handle this problem, we cross-reference all ICD-based cancer diagnoses with the VA central cancer registry, which provides curated, high-quality labels for cancer cases based on pathology reports. We only retain cancer cases recorded in the registry to reduce false positives and provide a reliable signal for

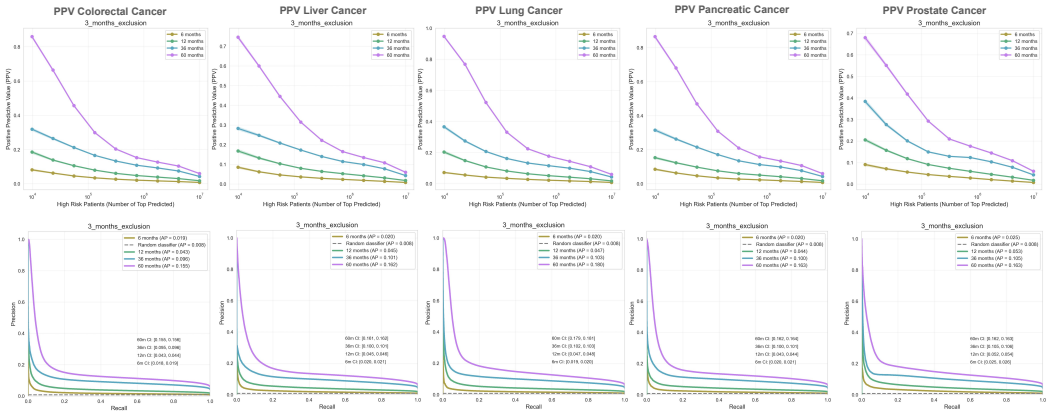

Figure 2: **Prediction performance evaluated on withheld test set**. Top: PPV@N (per 1M screened) for the N highest-risk patients. Bottom: precision–recall curves and area under the curve (AP, average precision) for five cancer types. For clinical implementation one can choose a conservative operating point capped at N to limit false positives; in practice, N is set by clinical capacity and cost.

model training. **Control labels.** For each cancer type, cases were paired with controls defined as patients without a diagnosis of the target cancer at or before the censoring date. Importantly, controls for one cancer type may include patients who develop a different cancer, introducing "hard negatives" and forcing the model to learn cancer-specific risk factors rather than general cancer risk (Table S1).

Evaluation setup. To assess the clinical utility of the model, we compute the standardized incidence ratio (SIR) and positive predicted value (PPV) at different operating points. SIR compares observed incidence in model-identified high-risk subgroups to the expected incidence based on population age-specific rates. PPV at threshold N (PPV@N) is the proportion of true cases among the top N risk-ranked patients; increasing N improves sensitivity but typically lowers PPV, so choosing N balances PPV against false positives.

5 Results and Discussion

We train and evaluate the model under two settings: (i) *no exclusion*, which uses all pre-diagnosis data before cancer diagnosis, and (ii) a *3-month exclusion* that masks events in the 3 months preceding cancer to remove predictive signal from any quasi-symptoms from trajectories. Across all five cancers, the PPV@N curves decrease as N increases (Fig 2). Results without exclusion are reported in Supplementary S1. As expected, performance without the exclusion window is consistently higher (PPV@N, SIR@N, AP).

Among the five cancers, lung and prostate have the highest area under the precision-recall curve (AUPRC) followed by liver and pancreatic, with colorectal being the lowest. These differences may reflect variation in prevalence and number of samples available for training; cancer types with more confirmed cases provide stronger supervision signal. We also report SIR and AUROC for different time intervals in supplementary Figs S2-S3 along with time-to-cancer for 1000 high-risk patients Figs S4-S5. Together, the PPV@N, PR, SIR, and time-to-cancer results show that by choosing an operation point the model can be used to prioritize small high-risk cohorts for clinical management program, focusing resource-intensive early-detection tools such as blood-based CTC, DNA or protein biomarker profiles or targeted imaging where they are most likely to be beneficial. As a clinical decision support tool, the operating point can be set by choosing N (per million computationally analyzed) maximizing true positives while keeping false positives to program capacity. Across five cancers, our results, subject to validation in larger datasets, indicate improved performance relative to prior EHR-based models [15, 16], supporting its potential for deployment.

6 Conclusion, Limitations, and Future Scope

We introduced a transformer-based multi-cancer risk prediction model trained on longitudinal EHR data that makes use of shared and cancer-specific latent representations and outputs discrete-time risk estimates. Trained on the US-VA database, our preliminary results across five cancer types suggest improved PPV and SIR in high-risk cohorts, indicative of potential clinical utility. Clinical benefit can come from decision support for differential diagnoses, for patient consultation suggesting participation in surveillance programs and from increases in the efficiency of intervention programs.

We have not yet completed comprehensive comparisons against other published time-to-event methods; results are limited to five cancer types, and generalization beyond the VA population need to be tested. Next, we will extend to additional cancer types, include other real-world data modalities in the patient trajectories (e.g., medications, lab test values, procedures), and conduct external validation on non-VA populations to assess cross-system transferability. Early cancer detection can improve patient outcomes, help allocate scarce healthcare resources more effectively, and ultimately extend healthy life span. Our proposed risk assessment method is one possible approach to shifting the landscape of cancer diagnosis to earlier time points and cancer therapy to earlier, less aggressive cancer stages.

References

[1] Rebecca L Siegel, Angela N Giaquinto, and Ahmedin Jemal. Cancer statistics, 2024. *CA: a cancer journal for clinicians*, 74(1), 2024.

[2] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77, 2015.

[3] Michael R Stratton, Peter J Campbell, and P Andrew Futreal. The cancer genome. *Nature*, 458 (7239):719–724, 2009.

[4] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR, 2016.

[5] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/231141b34c82aa95e48810a9d1b33a79-Paper.pdf.

[6] 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*, 10:7155, 2020. doi: 10.1038/s41598-020-62922-y. URL https://www.nature.com/articles/s41598-020-62922-y.

[7] 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*, 4:86, 2021. doi: 10.1038/s41746-021-00455-y. URL https://www.nature.com/articles/s41746-021-00455-y.

[8] Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. Pre-training of graph augmented transformers for medication recommendation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-19)*, pages 5953–5959, 2019. URL https://www.ijcai.org/proceedings/2019/0825.pdf.

[9] Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfred Baston, Jack Ross, Esther Idowu, James T. Teo, and Richard J. B. Dobson. Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *The Lancet Digital Health*, 6(4):e281–e290, 2024. doi: 10.1016/S2589-7500(24)00025-6. URL https://doi.org/10.1016/S2589-7500(24)00025-6.

[10] Lin Lawrence Guo, Jason Fries, Ethan Steinberg, Scott Lanyon Fleming, Keith Morse, Catherine Aftandilian, Jose Posada, Nigam Shah, et al. A multi-center study on the adaptability of a shared foundation model for electronic health records. *npj Digital Medicine*, 7(171):1–12, 2024. doi: 10.1038/s41746-024-01166-w. URL https://www.nature.com/articles/s41746-024-01166-w.

[11] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):24, 2018.

[12] Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[13] Michael F. Gensheimer and Balasubramanian Narasimhan. A scalable discrete-time survival model for neural networks. *PeerJ*, 7:e6257, 2019. doi: 10.7717/peerj.6257.

[14] Ethan Steinberg, Jason Alan Fries, Yizhe Xu, and Nigam Shah. Motor: A time-to-event foundation model for structured medical records. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=NialiwI2V6. Spotlight.

[15] Limor Appelbaum, José P Cambronero, Jennifer P Stevens, Steven Horng, Karla Pollick, George Silva, Sebastien Haneuse, Gail Piatkowski, Nordine Benhaga, Stacey Duey, et al. Development and validation of a pancreatic cancer risk model for the general population using electronic health records: An observational study. *European Journal of Cancer*, 143:19–30, 2021.

[16] Davide Placido, Bo Yuan, Jessica X Hjaltelin, Chunlei Zheng, Amalie D Haue, Piotr J Chmura, Chen Yuan, Jihye Kim, Renato Umeton, Gregory Antell, et al. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. *Nature medicine*, 29(5):1113–1122, 2023.

[17] Chunlei Zheng, Asif Khan, Daniel Ritter, and others. Pancreatic cancer risk prediction using deep sequential modeling of longitudinal diagnostic and medication records. https://www.medrxiv.org/content/10.1101/2025.03.03.25323240v1, 2025. medRxiv preprint.

[18] Alexander Wolfgang Jung and Moritz Gerstung. Bayesian cox regression for large-scale inference with applications to electronic health records. *The Annals of Applied Statistics*, 17(2): 1064–1085, 2023.

[19] Stephan D Fihn, Joseph Francis, Carolyn Clancy, Christopher Nielson, Karin Nelson, John Rumsfeld, Theresa Cullen, Jack Bates, and Gail L Graham. Insights from advanced analytics at the veterans health administration. *Health affairs*, 33(7):1203–1211, 2014.

[20] Lauren E Price, Kimberly Shea, and Sheila Gephart. The veterans affairs's corporate data warehouse: uses and implications for nursing research and practice. *Nursing administration quarterly*, 39(4):311–318, 2015.

[21] Paul E Stang, Patrick B Ryan, Judith A Racoosin, J Marc Overhage, Abraham G Hartzema, Christian Reich, Emily Welebob, Thomas Scarnecchia, and Janet Woodcock. Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership. *Annals of internal medicine*, 153(9):600–606, 2010.

Supplementary Material

A Model Architecture and Hyperparameters

The model consists of a transformer encoder, which produces the shared and cancer-specific representations of an input patient history, and a risk-prediction head that takes in the concatenated representations and produces 5 cumulative risk estimates for each time interval in 0-3,0-6,0-12,0-36, or 0-60 months. The transformer encoder consists of two transformer layers (consisting of multi-head self attention follow by a two linear layers, with ReLU applied in between) for the shared representation, and one transformer layer for the cancer-specific one. Each transformer layer has 8 attention heads, and a hidden dimension size of 128. The shared and cancer-specific representations are pooled separately using a learned softmax attention pooling, and then concatenated before being passed to the risk prediction head. The risk prediction head is a linear model mapping the concatenated embeddings to a cumulative probability distribution over the risk prediction intervals. This ensures that, for example, the predicted risk of cancer within 0-6 months is always greater than the predicted risk of cancer within 0-3 months, 0-12 months is greater than 0-6 months, etc. All models were trained with the Adam optimizer for 20 epochs, with a learning rate of 0.001 and a batch size of 128. The final model used in validation was the one that achieved the lowest validation loss during training, evaluated after each epoch.

B Metrics

B.1 Positive Predictive Value at threshold N (PPV@N)

Positive predictive value is defined for a classification task as the fraction of true positives in the total set of predicted positives:

$$PPV = \frac{TP}{TP + FP}$$

In our setting, we determine predicted positives by setting a threshold N. After sorting patients according to their predicted risk, the predicted positives are the N highest risk patients, and the PPV is the fraction of the patients in that set who are true positives.

B.2 Standardized Incidence Ratio at threshold N (SIR@N)

Standardized incidence ratio is defined as the ratio between the number of patients in a high risk cohort who are true positives and the number of true positive patients we would expect to see based on the demographics of the high risk cohort. The high risk cohort is defined by the N highest risk patients, according to the risk model being assessed. SIR can be computed in different ways, depending on what demographic data are used. In our case, we stratify by age, as age is a highly predictive factor for cancer risk. SIR for our setting can then be defined as

$$SIR = \frac{TP}{\sum_i inc(age(i))}$$

where $age(i)$ is the age of the i-th patient in our high risk cohort, and $inc(age)$ is the expected incidence of cancer at the input age. The denomiator, by summing the individual likelihood of cancer across high-risk patients, conditioned on their demographics, tells us how many positive cases we would expect to see in this population without our risk model. A high SIR can then be interpreted as the extent to which our model identifies high-risk patients beyond a simple baseline model using demographic statistics. SIR can be generalized beyond age by stratifying along additional attributes (the inc function in that case would take in more attributes about an individual than just age). We compute SIR separately for multiples values of N, and for all timepoints and cancer types.

Table S1: Cohort statistics.

	Colorectal		Liver		Lung		Pancreatic		Prostate		Control	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
# Male	3600	3421	1919	1845	10366	10155	1149	1025	18123	1025	479515	465958
# Female	88	84	21	30	285	271	36	29	-	-	54845	54970
Age (min/max/mean/median)	[27/95/67/68]	[26/99/67/68]	[32/93/63/62]	[39/93/63/62]	[28/97/67/67]	[28/97/67/67]	[33/98/66/66]	[33/95/66/67]	[30/98/65/65]	[32/104/65/65]	[0/113/62/59]	[0/119/64/65]

C Extended results

C.1 Model trained with no exclusion

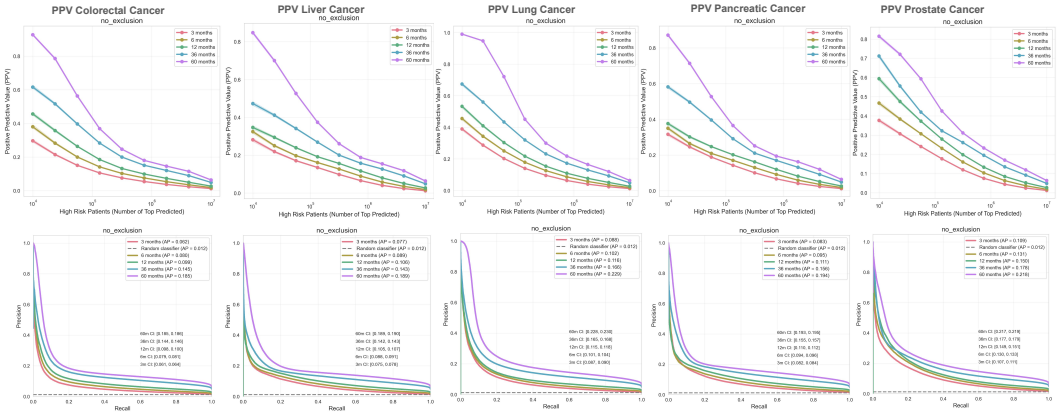

Figure S1: Top: PPV@N (per 1M screened) for the N highest-risk patients. Bottom: precision–recall curves for five cancer types. For clinical implementation one can choose a conservative operating point capped at N to limit false positives; in practice, N is set by clinical capacity and cost.

C.2 Standardized incidence ratio

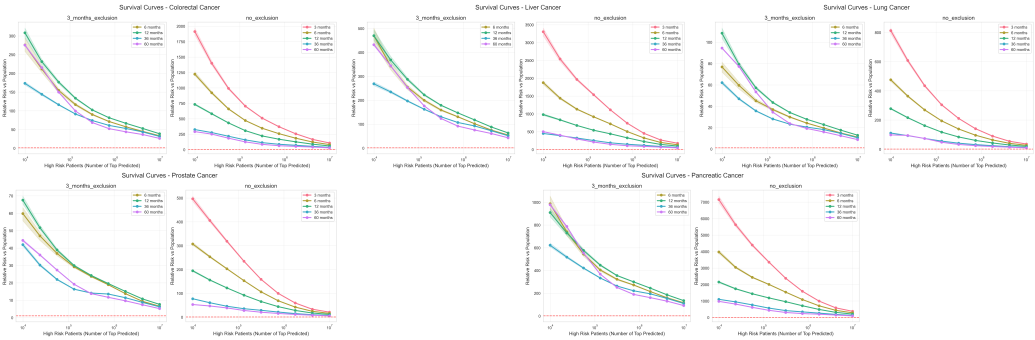

Figure S2: SIR@N (per 1M screened) for the N highest-risk patients. For clinical implementation one can choose a conservative operating point capped at N to limit false positives; in practice, N is set by clinical capacity and cost.

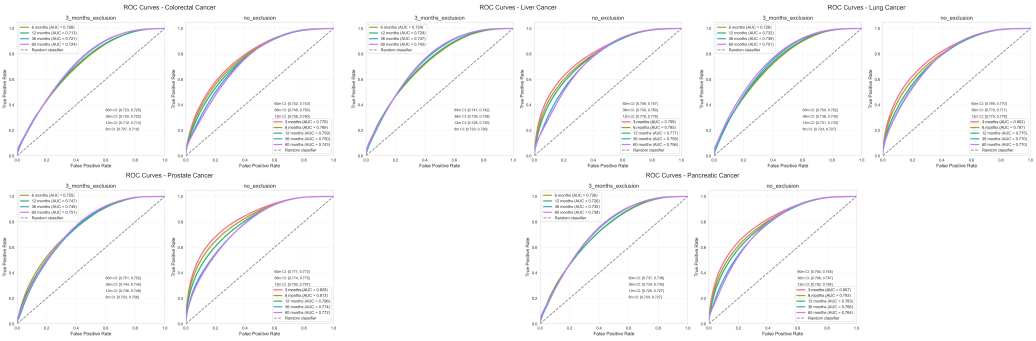

Figure S3: AUROC Scores

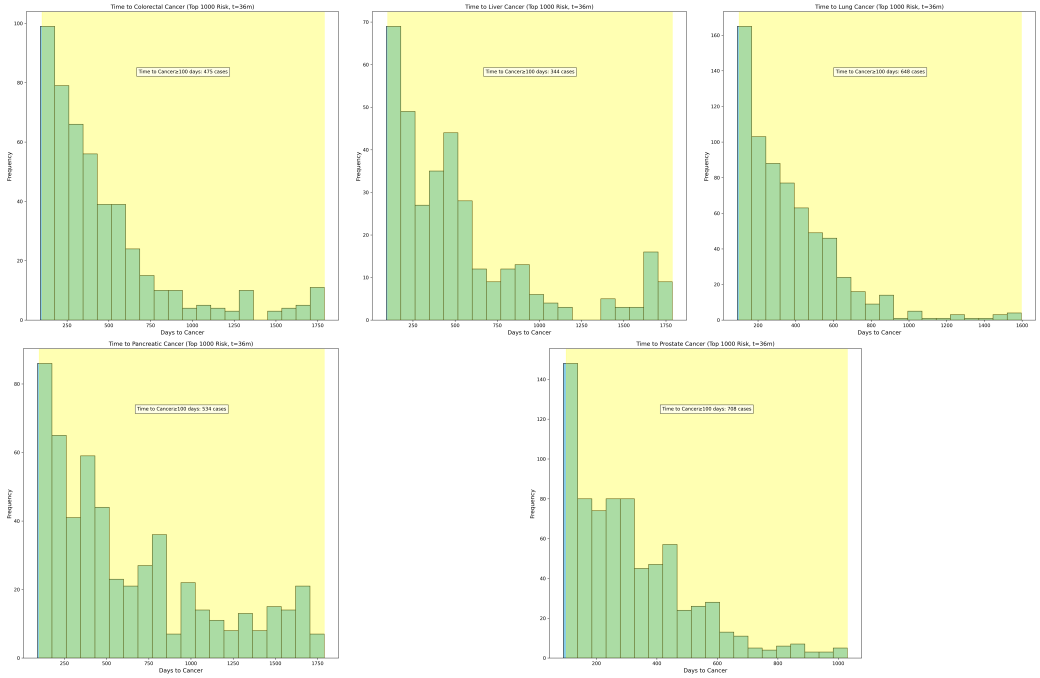

Figure S4: Time to Cancer (3 Months Exclusion).Among patients predicted as high risk who were subsequently diagnosed with cancer, the x-axis represents the interval between the prediction time point and diagnosis, while the y-axis indicates the frequency of patients.

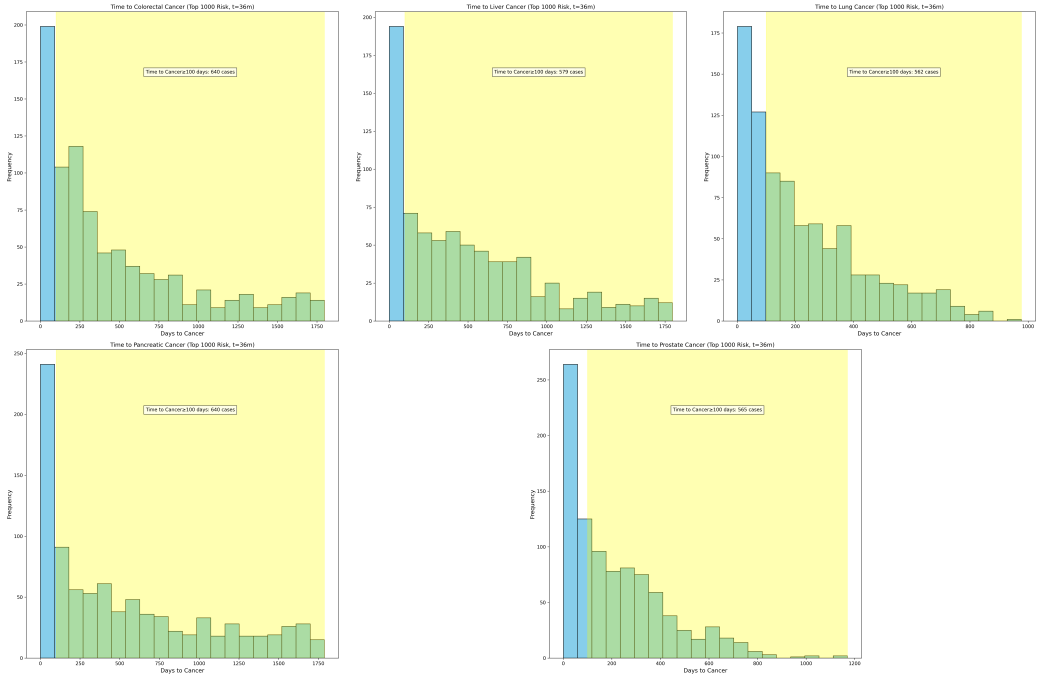

Figure S5: Time to Cancer (No Exclusion).Among patients predicted as high risk who were subsequently diagnosed with cancer, the x-axis represents the interval between the prediction time point and diagnosis, while the y-axis indicates the frequency of patients.