Pancreatic cancer risk prediction using deep sequential modeling of longitudinal diagnostic and medication records

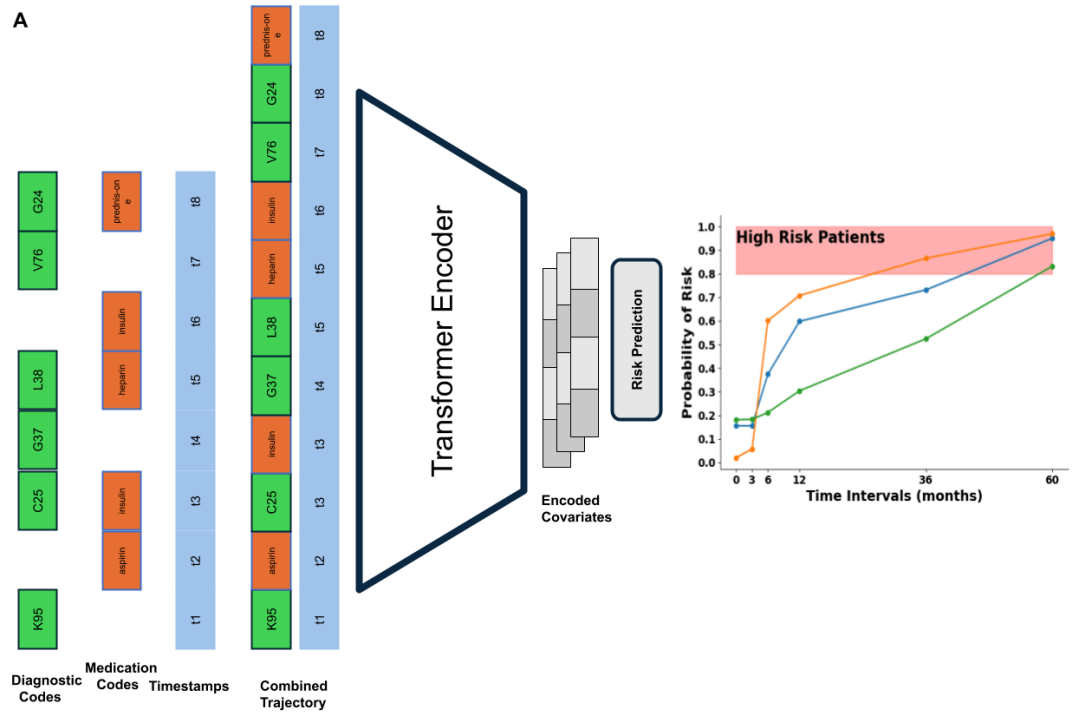
Chunlei Zheng[1,], Asif Khan[2,6,*], Daniel Ritter[2,3,*], Debora S. Marks[2,5], Nhan V. Do[1,4], Nathanael R. Fillmore[1,&], Chris Sander[2,5,6,&]*

[1] VA Boston Healthcare System, Boston, MA, USA.
[2] Harvard Medical School, 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] Ludwig Center at Harvard, Boston, MA, USA.

***Joint first authors**
&Joint senior authors

Abstract

Background: Pancreatic ductal adenocarcinoma (PDAC) is a rare, aggressive cancer often diagnosed late with low survival rates, due to the lack of population-wide screening programs and the high cost of currently available early detection methods.
Methods: To facilitate earlier treatment, we developed an AI-based tool that predicts the risk of pancreatic cancer diagnosis within 6, 12 and 36 months of assessment, using time sequences of diagnostic and medication events from real-world electronic health records (EHRs). Trained on a large US Veterans Affairs dataset with 19,000 PDAC cases and millions of controls, the tool employs a Transformer-based model that can capture and benefit from information synergy between diagnoses and medications.
Findings: Risk prediction is improved when incorporating medication data alongside diagnostic codes. For N patients predicted to be at highest risk out of 1 million, risk of cancer within 3 years is substantially higher than using a reference estimate based on age and gender alone (standard incidence ratio SIR=115 to 70 for N=1000 to 5000). Detection of the most predictive features generates clinical hypotheses such as X and Y. We quantify prediction bias between different socioeconomic subpopulations.
Interpretation: The risk prediction tool is intended to be the first step in a three-step clinical program: identification of high-risk individuals using AI tools, followed by a stratified surveillance program for early detection and intervention, aiming to benefit patients and lower health-care costs.
Funding: US CDMRP Pancreatic Cancer Risk Using Artificial Intelligence.



Patient's combined diagnosis and medication trajectory.

Trajectory as an input to a model and output is a probability of risk over time.

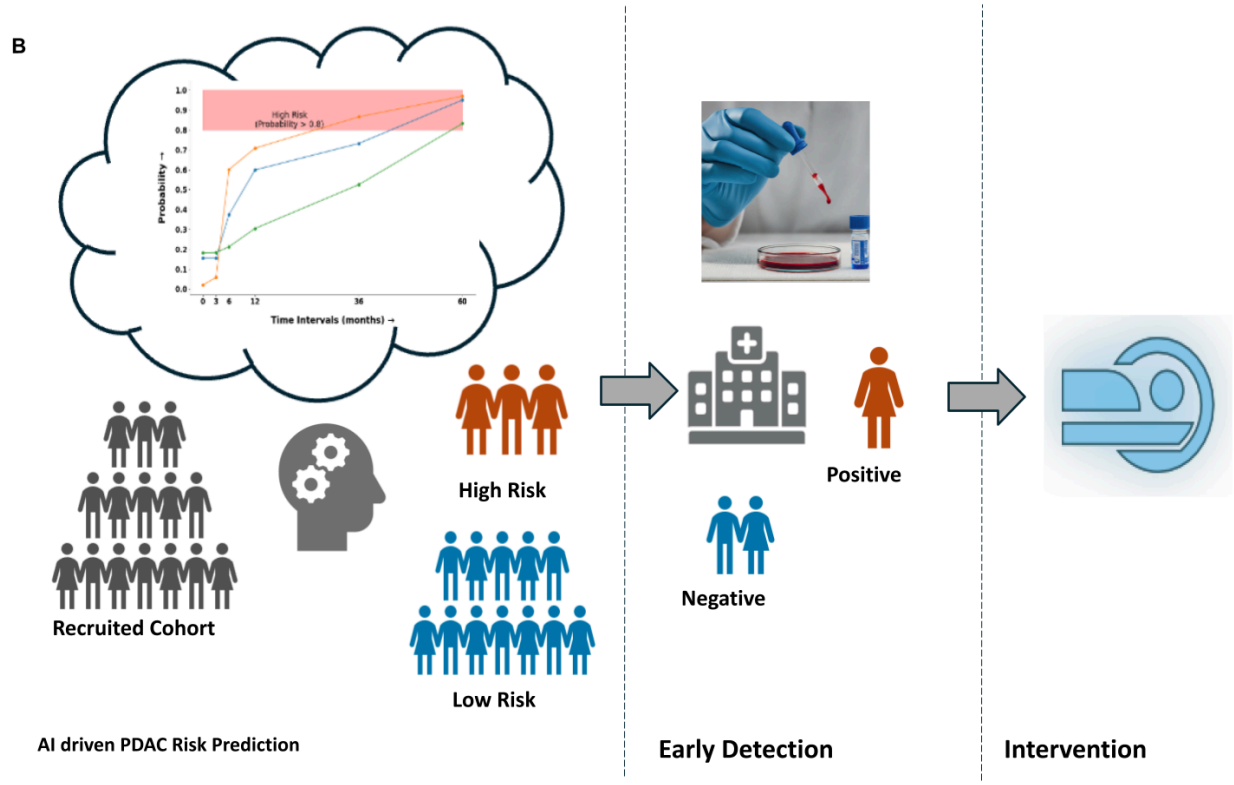


Figure 1: AI Tool for Risk Prediction. A: *Example of input trajectory based on the combination of diagnostic (ICD) and medication (RxNorm) codes. A trajectory represents the patient's medical history, encompassing diagnoses and prescribed medications along with respective timestamps. The combined trajectory is given as input to the AI prediction tool, which uses a Transformer encoder to extract latent risk factors. These factors are then fed to a risk prediction model that calculates the probability of cancer risk over different time intervals.* ***B:*** *Surveillance program that uses the AI tool to identify high-risk patients. Identified high risk cohorts can benefit from early detection technologies and possible healthcare interventions.*

Introduction

PDAC is one of the most challenging cancers to diagnose with a poor prognosis, often detected at advanced stages when curative treatment options are limited. Its increasing incidence makes it one of the leading causes of cancer-related deaths worldwide (1)[,(2)]. Approximately 80% of pancreatic cancer patients are diagnosed with locally advanced or distant metastatic disease and have poor survival rates (only ~9% survive for 5 years) (3). In contrast, for the ~20% of patients with early stage disease (stage IA) survival rates are much higher (~80% of these survive for 5 years, according to the US Veterans Affairs (US-VA) cancer registry (4)). These patients can be effectively treated by a combination of surgery, chemotherapy and radiotherapy (5). Therefore, identifying risk factors and improved detection at early stages could significantly improve outcomes and reduce mortality from this aggressive malignancy.

Early risk models on selected cohorts. Early detection programs can improve patient outcomes, but universal implementation is currently impractical due to the risk of false positives and high cost of screening for some cancer types, such as PDAC. An affordable selective screening approach targets high-risk individuals with an acceptable low rate of false positives. Previous works on selected (but relatively small) cohorts have developed risk models using factors such as family history, lifestyle, environmental, and genetic variants (6–11), and identified individuals at increased risk of PDAC when such data was available (12).

Using real-world EHR data in predictive AI models. Electronic health records (EHRs) offer a rich and generally available source of patient data, including diagnoses, medications, and laboratory results, that can support population-wide risk prediction. Prior works have demonstrated the application of machine learning (ML) in EHR systems, such as BEHRT, a Transformer model specifically designed for general prediction of future diagnoses of patients (13). Similar ML approaches have been explored in other recent works (14–17). Training specific ML prediction models on routinely collected large-scale real-world EHR data could enable targeted screening and

preventive interventions for high risk PDAC patients. Recent works have used EHR data to develop several PDAC risk prediction models (3);(18–22), which can be used to identify high risk cohorts.

Medication data for improved risk prediction. Most cancer risk prediction models based on EHR data predominantly rely on diagnosis codes alone and only a few use other forms of data types, such as medication records (3,19,21). Diagnosis codes, especially in the US healthcare system, are influenced by billing incentives and reimbursement practices, which can lead to “billing pollution”, where differential or secondary diagnosis codes are upcoded to justify costly procedures or maximize reimbursement (23,24). This data noise can result in a somewhat inaccurate representation of a patient’s state of health. In contrast, medication data plausibly provide more objective information as prescriptions are based on well-trained physicians’ decisions and are more likely to represent objective clinical necessity rather than financial motivation. In addition, use of medications can provide direct information on disease risk due to known preventive effects or negative side effects.

Drug effects on cancer risk. The usage of certain medications has been linked to altered cancer risk,with either positive or negative effects. E.g., consider omeprazole, a proton pump inhibitor commonly used to treat gastroesophageal reflux disease. Recent epidemiological studies have raised concerns based on potential association between long-term use of omeprazole and an increased risk of developing PDAC (25–27). This association is attributed to hypergastrinemia, a condition characterized by elevated levels of the hormone gastrin, which has been shown to stimulate the growth of pancreatic cancer cells via interaction with the cholecystokinin B receptor (CCKBR). An opposite example is metformin, commonly used to treat type 2 diabetes, which has been associated with a decreased risk of PDAC (28). Metformin works by lowering blood glucose levels (29,30) and also reduces levels of inflammatory markers (31,32), both of which are factors that can contribute to cancer development. While such associations with altered risk may be influenced by various confounding factors, medications may provide non-redundant information about cancer risk.

Non-sequential versus sequential risk models

Previous risk prediction efforts have primarily relied on a pre-identified set of diagnosis codes as input features, overlooking valuable longitudinal information within EHR and the sequential nature of clinical events. Recently, a PDAC RISk Model (‘PRism’) (3) trained on a large-scale TriNetX (33) database, incorporated composite features, including diagnosis records, medication history, and laboratory results to encode each patient’s history as a fixed 87-dimensional vector, which was then used to train multilayer perceptron (MLP) and logistic regression (LR) models. The choice of input

features in PRism does not model the sequential ordering and timing of events explicitly. Unlike PRism, our AI tool uses deep sequential learning, explicitly capturing the sequential order and timing of clinical events for more nuanced PDAC risk prediction.

Combination model trained on medication and diagnostic data

Previously, CancerRiskNet (34) introduced a Transformer-based sequential model for PDAC risk prediction, using only diagnosis codes. Here, we extend this architecture to longitudinal sequences of both diagnosis and medication codes, allowing the model to learn a shared encoded space that captures both clinical diagnostic histories and prescribed drugs over time. Our model utilizes self-attention to capture interactions among disease and medication events, while accounting for temporal spacing using positional encoding based on relative event timings. Developed and tested on a large-scale US-VA dataset, this model is a novel approach to integrate diagnostic and medication data for PDAC risk prediction. Our work facilitates identification of high risk cohorts who could benefit from a three step clinical program – prediction, detection, and intervention (Figure 1B), to improve patient outcomes.

(A)	Overall	Non-cancer	Cancer
n (number of patients)	210848	192722	18126
By gender, n (%)			
Female	28071 (13.3)	27566 (14.3)	505 (2.8)
Male	182774 (86.7)	165153 (85.7)	17621 (97.2)
By race, n (%)			
Asia/Pacific Islander	3066 (2.1)	2917 (2.2)	149 (1.0)
Black	25708 (17.3)	22670 (16.9)	3038 (20.6)
Native	1159 (0.8)	1062 (0.8)	97 (0.7)
White	119061 (79.9)	107582 (80.1)	11479 (77.8)
Trajectories-ICD diagnoses			
Lengths (# events), median [Q1,Q3]	91.0 [15.0,298.0]	82.0 [13.0,283.0]	195.0 [73.0,433.0]
Time (years), median [Q1,Q3]	7.0 [2.0,14.0]	8.0 [2.0,14.0]	7.0 [3.0,12.0]
Trajectories-medications			
Number of prescriptions median (# events) [Q1,Q3]	50.0 [14.0,129.0]	48.0 [13.0,126.0]	73.0 [29.0,152.0]
Time (years), median [Q1,Q3]	8.0 [3.0,14.0]	8.0 [3.0,15.0]	7.0 [2.0,12.0]

(B)	Overall	Non-cancer	Cancer
n (number of patients)	987693	986393	1300
By gender, n (%)			
Female	141224 (14.3)	141179 (14.3)	45 (3.5)
Male	846468 (85.7)	845213 (85.7)	1255 (96.5)
By race, n (%)			
Asia/Pacific Islander	15544 (2.3)	15534 (2.3)	10 (0.9)
Black	114733 (16.6)	114512 (16.6)	221 (20.6)
Native	5449 (0.8)	5445 (0.8)	4 (0.4)
White	553383 (80.3)	552544 (80.3)	839 (78.1)
Trajectories-ICD diagnoses			
Lengths (# events), median [Q1,Q3]	83.0 [13.0,282.0]	82.0 [13.0,282.0]	207.0 [78.0,458.5]
Time (years), median [Q1,Q3]	8.0 [2.0,14.0]	8.0 [2.0,14.0]	7.0 [3.0,12.0]
Trajectories-medications			
Number of prescriptions median (# events) [Q1,Q3]	48.0 [13.0,126.0]	48.0 [13.0,126.0]	78.0 [31.0,167.0]
Time (years), median [Q1,Q3]	8.0 [3.0,15.0]	8.0 [3.0,15.0]	7.0 [2.0,12.0]

Table 1. Patient records in the US-VA dataset for model training and evaluation. *US-VA data includes patient visits in both inpatient and outpatient settings across the nation. (A) The dataset for model development has 18K cancer cases and a randomly selected set of 193K controls. (B) The dataset for model evaluation (also called held-out dataset in Methods) has a subset of 1300 cancer cases and, to reflect the balance in real-world data, 986K controls . Q1: first quantile and Q3: third quantile.*

Results

Performance of combination model trained with data exclusion

The ICD codes appearing shortly before the diagnosis of cancer can represent quasi-symptoms that would be easily interpreted by a clinician. To determine whether the model is primarily based on these quasi-symptoms or learns non-trivial covariates predictive of cancer, we trained models with and without a three-month data exclusion window. The data exclusion involves removing all ICD codes and medications assigned within three months prior to the diagnosis of pancreatic cancer. With this data exclusion, compared to models from either ICD or medication alone, the combination model performed best (for each of the 'cancer within k months' intervals). For instance, for the combined model the AUROC is 0.857 (95% CI: 0.852-0.860) for the 36-month prediction interval, while the diagnosis and medication models have lower AUROCs of 0.787 (95% CI: 0.776-0.791) and 0.746 (95% CI: 0.741-0.751) respectively (Figure 2A). As expected, the scores of models trained with data exclusion are lower when compared to the model without exclusion (see Supplementary), which can be attributed to the absence of quasi-symptoms.

With our focus on highest risk patients to be nominated for moderate-size affordable surveillance programs, we evaluate model performance (computed on the withheld test set) as a function of N highest risk patients out of 1 million real-world all-comers. The standard incidence ratio (SIR), which quantifies model performance relative to a simple model based only on incidence population statistics for a given gender and age bracket, decreases as one goes down the list of N patients ranked by predicted risk (Figure 2C). The curve reflects the tendency of generally higher accuracy at smaller N (except at very small N where small sample size results in higher uncertainty intervals).
The curve can be used to make a particular choice of decision threshold such that N represents a conservative choice that minimizes false positives and optimizes performance by restricting the number N of high-risk patients nominated for surveillance. In a real-world application, the N-threshold should be selected where the confidence interval is low and stable, ensuring a balance between the cost and practical requirements of a clinical surveillance program targeting the highest-risk individuals.

To address the question of when in a prediction interval cancer is likely to occur, we evaluate, as an example, the actual time to cancer (in the test set) for N=1000 highest risk patients for the (Figure 3). In this example, 37% of the highest risk patients have cancer occurrence after more than 100 days. This kind of evaluation can be a guide to choice of tests and follow-up repeat visit schedules in the design of surveillance programs.

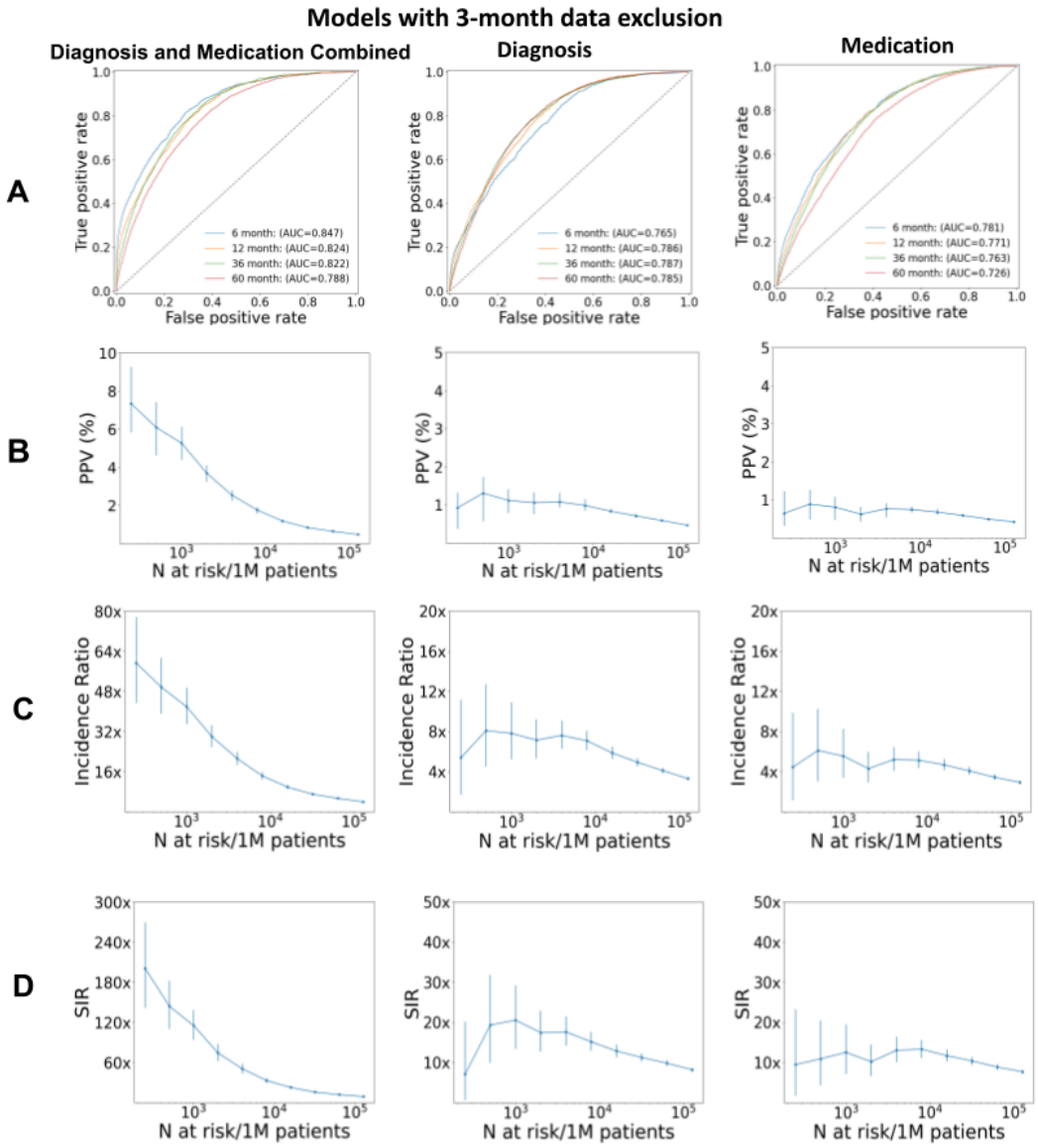


Figure 2. Comparison of the performance of model trained with a three-month data exclusion window for the 36-month prediction interval: diagnostic data only, medication data only, and combination of both. *Left: model trained on the combination of diagnosis and medication data. Middle: trained on diagnosis data only. Right: model trained on medication data only. (A) ROC performance curve and area under the curve (AUC) for different prediction time intervals. (B) Positive Predictive Value (PPV) and (C) Incidence ratio and (D) Standardized incidence ratio (SIR) for the N highest-risk patients out of 1 million patients. The curve can be used to choose a decision threshold for realistic implementations.*

Interpretation of most predictive features

To highlight the most relevant features driving the prediction of the risk model (attribution analysis), we applied the Integrated Gradients (IG)(35) method to a set of

high-risk patient trajectories. Features consisting of ICD codes and medications were ranked by their IG scores, with the top 10% identified as most predictive (details in supplementary; list in Table 2).

Essential hypertension and diabetes have widely been associated with cancer risk (36), while lipoprotein metabolism disorders link to disrupted lipid metabolism (37). Conditions such as acute pancreatitis (38) and chronic inflammatory conditions such as ischemic heart disease as well as chronic obstructive pulmonary disease are linked to sustained inflammation (39). These diagnoses include several metabolic, cardiovascular, and inflammatory conditions which are known to be associated with high risk of pancreatic cancer.

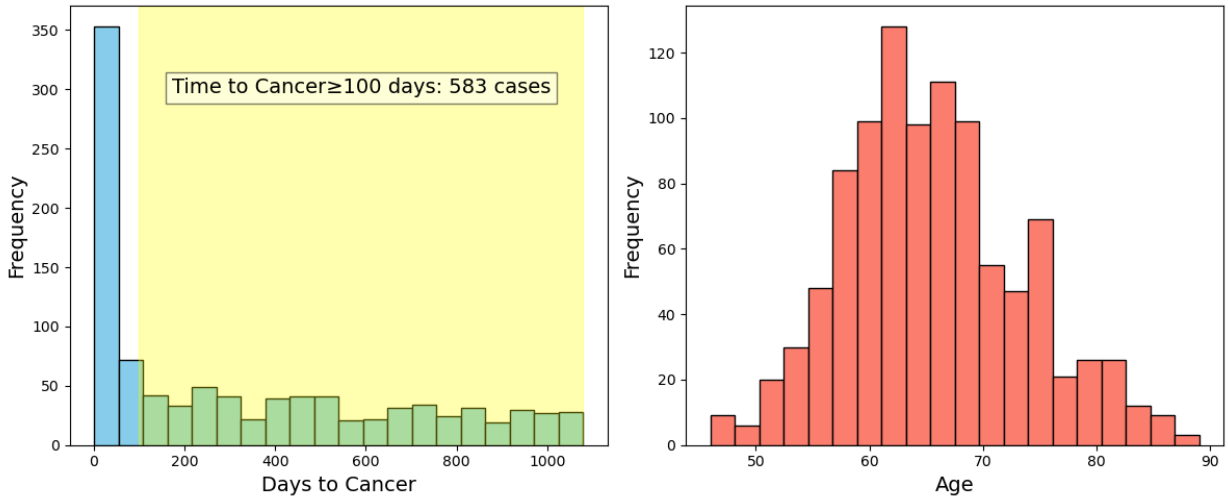


Figure 3. Characteristics of highest-risk subpopulation using a model trained with a three-month data exclusion window for the 36-month prediction interval (N=1000) (Left) *Patients who were predicted to be at* ***high risk*** *and eventually* ***got cancer*** *are sampled to determine the time difference between the end of the disease trajectory (time point of prediction) and the actual cancer diagnosis (x-axis). y-axis: frequency of patients in the respective time gap. The frequency at larger time differences (yellow) indicates to what extent the model can correctly predict cancer diagnosis at larger time intervals before the actual diagnosis.* ***(Right)*** *Age distribution in the sampled high-risk cohort indicates good coverage.*

	Diagnosis Codes
1	Essential (primary) hypertension
2	Non-insulin-dependent diabetes mellitus, Insulin-dependent diabetes mellitus, Other specified diabetes mellitus
3	Disorders of lipoprotein metabolism and other lipidaemias, Other metabolic disorders, Disorders of glycoprotein metabolism
4	Acute pancreatitis
5	Chronic ischaemic heart disease
6	Psoriasis, Parapsoriasis, Pityriasis rosea Other papulosquamous disorders, Other dermatitis
7	Benign neoplasm of other and ill-defined parts of digestive system
8	Other chronic obstructive pulmonary disease
9	Polyarthrosis, Other arthrosis, Coxarthrosis
10	Glaucoma, Other congenital malformations of eye

	Medication Codes
1	Lisinopril
2	Simvastatin
3	Hydrochlorothiazide
4	Aspirin
5	Metformin
6	Hydrocortisone
7	Nitroglycerin
8	Docusate
9	Insulin
10	Vitamin B12

Table 2. ***Top ten most predictive diagnosis and medication codes using a model trained with a three-month data exclusion window for the 36-month prediction interval.*** *We conducted an attribution analysis on high-risk patients and list the most predictive diagnosis (left) and medication (right) codes for high-risk patients as identified by our model.*

Metformin, prescribed medication for type 2 diabetes, and insulin, used for blood glucose regulation, both are associated with elevated glucose levels known to increase risk of cancer (40,41). An association between the initiation of antidiabetic and anticoagulant medications and the diagnosis of pancreatic cancer in a two year pre-diagnosis time window was also reported in a study of controlled cohorts of nurses and health professionals (42). Aspirin and simvastatin, prescribed in cardiovascular or chronic inflammation conditions, may influence cancer risk through their anti-inflammatory effects (43,44). These medications likely indicate underlying metabolic or inflammatory conditions that may increase the likelihood of future cancer development, suggesting their role as markers of disease progression rather than direct causative agents.

These interpretations are to be considered with caution, as the IG method does not provide isolated effects of individual features; instead, it measures the influence of each feature change within the context of other features in the input vector. This implies that

the predictive power of features emerges from their interdependencies rather than their independent contributions.

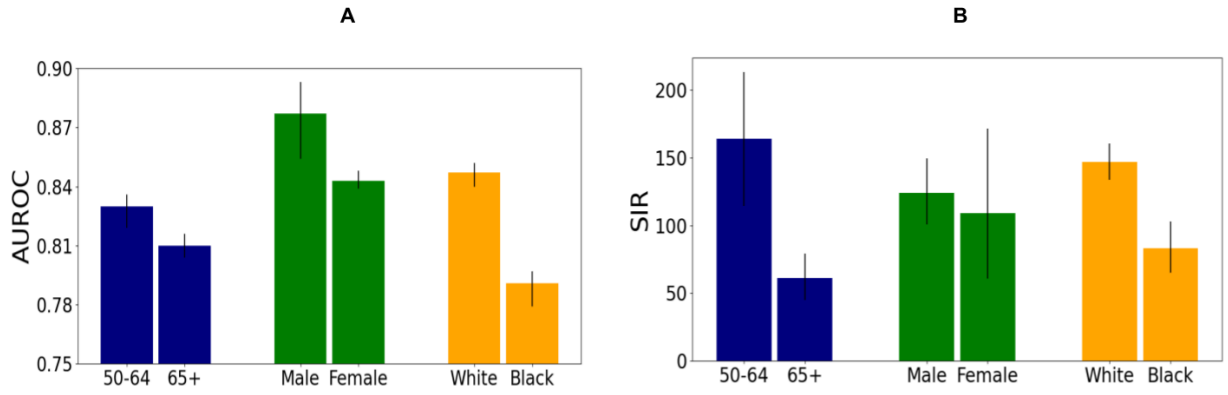


Figure 4. Evaluation of prediction performance in subpopulations using a model trained with a three-month data exclusion window for the 36-month prediction interval (A) *AUROC and* ***(B)*** *SIR, stratified by race, gender, and age. Notably, the SIR score is higher for the white subpopulation, as well as for male patients, and individuals in the 50-64 age group. The observed differences in performance across subpopulation groups can be attributed to biases in the dataset, as it consists of a higher proportion of white patients compared to black patients, more male patients compared to female patients, and a greater number of individuals in the 50-64 age group compared to 65+ age group, as detailed in Table 1.*

Evaluation in subpopulations

AI models can be biased towards specific subpopulations based on the distribution of samples from different groups within the training dataset (45). For instance, if the training data predominantly includes samples from one demographic group, the model is likely to perform better when evaluated on that specific group. Investigating such biases is important for assessing whether a model performs comparably across all subgroups.

We, therefore, stratified the evaluation dataset and validated the model on subpopulations defined by race and sex. The AUROC for white patients in a 12-month prediction window is 0.901 (95%CI: 0.897-0.904), compared to 0.870 (95%CI: 0.860-0.880) for black patients. For male patients it is 0.892 (95%CI: 0.888-0.896), compared to 0.900 (95%CI: 0.883-0.913) for female patients. Moreover, when examining the 1,000 high-risk cohorts, we observed that the SIR for white patients is higher than for black patients, and the SIR for male patients is higher than for female patients. There are more white patients and more male patients in the database (Table 1), contributing to the observed performance disparities for the models trained on all

patients. These gaps highlight the model's bias plausibly resulting from the uneven representation of subpopulations in the dataset (Figure 4).

Methods

Datasets for model development and evaluation

We use the US-VA database to construct training and evaluation datasets, based on a total of 15,926,415 patients, including 19,426 cases of pancreatic cancer. Overall data statistics including data splits used for model development and evaluation are in Table 1. We use each patient's time-based trajectory of International Classification of Diseases (ICD) codes for diagnoses and active ingredients from prescription medications. Consistent with the approach in earlier work (34), we restrict ICD codes to three-character categories within the ICD hierarchy to focus on fairly broad diagnostic groups. Timestamps associated with the diagnosis or medication event codes are incorporated in the model as positional encodings to account for the temporal aspect of patients' clinical histories.

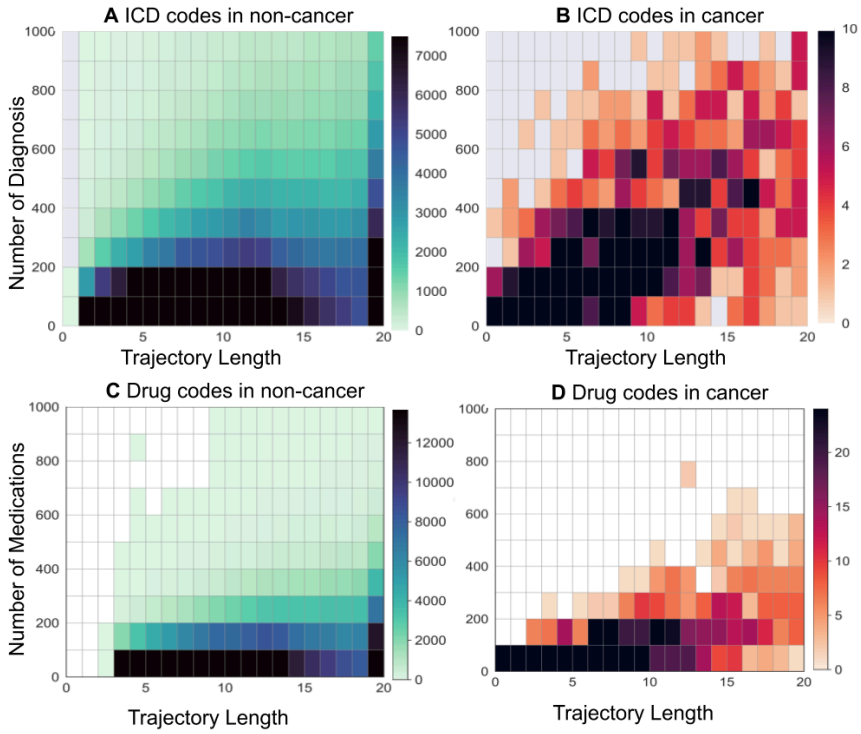
To train the AI model, we subsample trajectories of fixed lengths for each patient, starting at the first recorded event and ending at various endpoints. This subsampling approach ensures a more comprehensive modeling of patient risk, as it captures periods when specific symptoms develop. As a result, the same patient can have trajectories representing variations in cancer risk over time. These variations can occur for a number of reasons such as changes in lifestyle, disease states or the use of specific medications. This approach allows us to interpret each trajectory as a realistic and dynamic representation of a patient's health over time.

Evaluation of prediction performance

Using different data types, including patient ICD codes, medication usage and their combination, we trained three models and then evaluated their performance using four metrics: AUROC, positive predictive value (PPV), incidence ratio (IR) and standardized incidence ratio (SIR). Further details on each metric are in the Supplementary. The SIR is particularly relevant for developing a realistic surveillance program, as it compares the model's predictions to standardized incidence rates from population-wide studies conditioned on age and gender. A SIR of 1 indicates a model prediction is equivalent to frequency estimates based on population-wide studies, while higher values indicate better predictive performance. Considering the high cost of surveillance programs, these are currently most realistic for a relatively small number of patients at a very high risk of pancreatic cancer. We, therefore, select the top 1,000 high-risk patients from a cohort of 1 million and report the SIR scores for different models.

Risk prediction model

The architecture of the risk prediction model is based on the framework described earlier (34). The earlier implementation only uses diagnostic codes. The current model takes as input combined trajectories of medication and diagnostic codes, along with their respective timestamps, ending at the time of prediction. Initially, all events in a given trajectory of clinical data items are encoded into a fixed-length representation using an embedding layer. The timestamp information is incorporated in form of relative positional encoding of events, which, when combined with the embedding, provides a time-dependent encoding of trajectories. These encodings are then input into a Transformer model which uses a self-attention mechanism to learn interdependencies between all codes within a trajectory, both diagnostic and medication. The Transformer's output is aggregated along the time dimension to produce a fixed-length latent representation of a given input trajectory. This representation is subsequently fed into a supervised risk regression model, which predicts the probability of cancer occurrence within five time intervals: 0-3, 0-6, 0-12, 0-36, and 0-60 months. Further details are in the Supplementary.



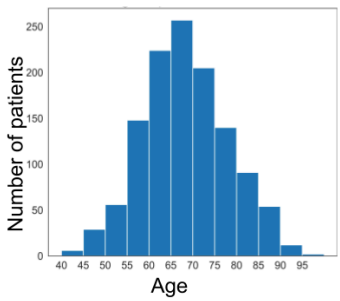


Figure 3 Age distribution of pancreatic cancer patients in US-VA dataset.

Figure 5. The density of medications and diagnosis codes for different data trajectory lengths*. Cancer patients (A,C), and non-cancer patients (B, D). The intensity*

of color is a number of patients in a bin. For non-cancer patients, patients with less than two years of follow-up were removed.

Discussion

Combination model trained on diagnostic and medication trajectories.

This work highlights the information value of combining diagnosis and medication data in patients' health histories using a Transformer neural network for risk prediction of pancreatic cancer. Medication data carries information complementary to diagnosis codes, and helps address the hard-to-detect and hard-to-correct upcoding of diagnosis codes, as medication codes are less likely to be influenced by finance-motivated billing practices (23). In future, application of natural language processing of clinical notes to filter diagnosis codes and ML-learned feature selection may be useful to mitigate billing related noise in EHR datasets. In addition, additional patient data such as blood test results and, in future, widely available genetic information will probably further enrich the model's ability to learn from the most informative combinations of risk factors in a real-world population.

Sequential modeling of EHR trajectories.

By using a sequential model, the AI tool makes use of temporal dependencies between health events. Our results indicate that a sequential model, as opposed to a 'bag-of-events' model, can capture temporal contributions of and informational synergy between disease and medication events, leading to improved performance in predicting cancer risk compared to non-sequential models. (Supplementary Figure 2S).

Focus on large real-world patient data and on the highest risk bracket.

There is an informational trade-off between large real-world EHR datasets and specialized data from small specifically recruited cohorts. This study focuses on large real-world patient datasets in contrast to pioneering earlier prediction models in smaller controlled cohort studies, which have developed risk prediction models using carefully acquired well-defined risk factors such as family history, genetic markers, unintended weight loss, and systematic changes in fasting glucose levels (6–11). However, such valuable information is not generally available in the EHRs of all-comers in large populations. To cast a wider net, this study used EHRs from real-world populations based only on generally available information in health records, such as diagnoses and prescriptions.

In a trade-off that balances cost and benefit of surveillance programs, we focus on the selection of high-risk cohorts, which is only a small fraction of the large total population:

patients at very high risk for the cancer with the highest probability of accurate prediction (lowest false positive prediction rate), for whom expensive screening tests are justified and affordable. For example, from all patients in a hospital system aged 50 years or older, out of 1 million patients on whose EHRs the prediction tool is applied, we propose to identify a subpopulation of a few thousand highest-risk patients and focus on a 12-month prediction interval. E.g., for 1000 patients at highest risk, the result of SIR=115 for a model trained with a 3-month exclusion window (not using data from a 3 month interval before cancer occurrence) quantifies how much better the AI tool is in identifying high-risk patients than usage of a background model that uses population wide incidence rates, only using age and gender information. The tool thus may have concrete utility as a decision support tool aiding clinicians in an early detection program in a large hospital system.

Prediction performance in subpopulations.

We have also evaluated the prediction performance of the inclusively trained model in different race groups and highlight potential biases to be addressed in future work. Our retrospective (on a withheld test set) evaluation of the PDAC risk prediction model across different race groups highlights the potential impact of inherent biases in data distribution. Specifically, the model trained on the entire population had a lower AUROC score of 0.82 for the black ethnic group compared to 0.85 for the White ethnic group; a lower score of 0.86 for male subpopulation compared to 0.88 for female subpopulation; and higher score of 0.86 for patients aged 50-64 compared to 0.82 for age 65+ patients (Fig. 4). These disparities in performance can plausibly be attributed to inherent differences in the training data distribution, which leads to biases in prediction performance. For instance, there are more white than black patients in the EHR database (Table 2) so that disease and medication data items corresponding to white are more likely to be reflected in the parameters of the trained model. In addition, differences in the type of EHR data items in each subpopulation often reflect existing disparities within society, especially socioeconomic status (SES) (46). This is because social determinants of health (SDOH), such as access to health insurance and healthcare services, are typically influenced by financial income (47). Factors such as historical demographics, data collection procedures, and access to healthcare services can directly influence the composition of population-specific data in a healthcare system and lead to underrepresentation or overrepresentation of specific populations in the dataset.

Addressing disparities in performance across demographics groups.

While differences in the performance of cancer risk prediction methods in subpopulations have been noted earlier (3), addressing these as part of an improved ML process is an open problem, which we have not addressed in this study. A relevant

future task would be to evaluate models of this type on groups with diverse socio-economic status (SES), perhaps by the use of residential information, such as postal zip codes, if this information is available. With sufficient population structure data, one can train a separate model for each ethnic or socioeconomic group in a particular health system. Access to such data could be obtained within a unified national healthcare system, which is not currently available in the US, or by federated learning procedures with software access to data behind institutional firewalls, which are technically feasible but difficult to negotiate. The key aim is to remove bias in the trained model(s) and achieve equal accuracy in spite of unequal risk. We anticipate that future work will achieve this aim in support of more equitable access to early detection and prevention programs.

Stratified surveillance programs.

The proposed PDAC risk model, which predicts cancer risk probabilities over various time intervals after risk assessment, offers a promising approach to differentially screen individuals based on their risk profiles. Such stratification can be either binary (at risk or not), or more granular using the real-number probabilities to stratify patients into, e.g., very high, high, medium, and low-risk groups. One can also determine thresholds in the probability values as decision parameters for various screening actions. Differential action can imply different time intervals for repeat screening visits, different procedures such as CT scans, MRI scans, endoscopic ultrasound or different advanced blood tests, e.g., for sequence variants, methylation or fragments in cell free DNA or for serum proteins, once these are generally accepted. Cost of the program and value of the kind of information desired would be decision parameters. Suggesting details of stratified screening programs is beyond the scope of this study. These are likely to be worked out by clinicians in consultation with developers of detection technologies, machine learning experts and hospital administrators. Careful definition and prospective evaluation of stratified efficient surveillance programs is likely to lower the barrier for general acceptance in clinical practice.

Outlook

The proposed AI-based risk prediction tool can be an important component of a clinical cancer management program, identifying only high-risk individuals drawn from a large, real-world population for an initial moderate-size lower-cost targeted surveillance program. By focusing on such high-risk cohorts, one can achieve a reduction in predicted false positives and a reasonably precise identification of pre-cancerous and early-stage cases. Lowering screening costs helps ease the healthcare system's burden, while early identification facilitates preventive measures, such as lifestyle modifications, that may delay or prevent pancreatic cancer onset. An AI-driven

surveillance program—combining prediction, detection, and treatment within a clinical collaboration—provides a promising framework for equitable, effective cancer surveillance across diverse healthcare environments.

Contributors

Author contributions: concepts C.Z., D.R., A.K., D.S.M, N.R.F, and C.S.; computation C.Z., D.R., and A.K.; supervision D.S.M, N.R.F, and C.S.; manuscript C.Z., D.R., A.K., D.S.M, N.R.F, and C.S. By mutual agreement, joint first authors can list the article as Name et al.

Declaration of Interests

Competing interest: C.S. is on the SAB of Cytoreason Ltd. D.S.M. is an advisor for Dyno Therapeutics, Octant, Jura Bio, Tectonic Therapeutic and is a cofounder of Seismic Therapeutic.

Acknowledgements

NONE

References

1. Rahib L, Smith BD, Aizenberg R, Rosenzweig AB, Fleshman JM, Matrisian LM. Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. Cancer Res [Internet]. 2014 Jun 1 [cited 2024 Jun 10];74(11). Available from: https://pubmed.ncbi.nlm.nih.gov/24840647/

2. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. CA: A Cancer Journal for Clinicians. 2024 Jan 1;74(1):12–49.

3. Jia K, Kundrot S, Palchuk MB, Warnick J, Haapala K, Kaplan ID, et al. A pancreatic cancer risk prediction model (Prism) developed and validated on large-scale US clinical data. EBioMedicine [Internet]. 2023 Dec [cited 2024 Jun 6];98. Available from: https://pubmed.ncbi.nlm.nih.gov/38007948/

4. Price LE, Shea K, Gephart S. The Veterans Affairs's Corporate Data Warehouse: Uses and Implications for Nursing Research and Practice. Nurs Adm Q. 2015;39(4):311.

5. Wolfgang CL, Herman JM, Laheru DA, Klein AP, Erdek MA, Fishman EK, et al. Recent progress in pancreatic cancer. CA Cancer J Clin. 2013 Sep;63(5):318–48.

6. Klein AP, Wolpin BM, Risch HA, Stolzenberg-Solomon RZ, Mocci E, Zhang M, et al. Genome-wide meta-analysis identifies five new susceptibility loci for pancreatic cancer. Nat Commun. 2018 Feb 8;9(1):1–11.

7. Kim J, Yuan C, Babic A, Bao Y, Clish CB, Pollak MN, et al. Genetic and Circulating Biomarker Data Improve Risk Prediction for Pancreatic Cancer in the General Population. Cancer Epidemiol Biomarkers Prev [Internet]. 2020 May [cited 2024 Jun 18];29(5). Available from: https://pubmed.ncbi.nlm.nih.gov/32321713/

8. Li D, Duell EJ, Yu K, Risch HA, Olson SH, Kooperberg C, et al. Pathway analysis of genome-wide association study data highlights pancreatic development genes as susceptibility factors for pancreatic cancer. Carcinogenesis. 2012 Apr 20;33(7):1384–90.

9. Petersen GM, Amundadottir L, Fuchs CS, Kraft P, Stolzenberg-Solomon RZ, Jacobs KB, et al. A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. Nat Genet. 2010 Jan 24;42(3):224–8.

10. Amundadottir L, Kraft P, Stolzenberg-Solomon RZ, Fuchs CS, Petersen GM, Arslan AA, et al. Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. Nat Genet [Internet]. 2009 Sep [cited 2024 Jun 13];41(9). Available from: https://pubmed.ncbi.nlm.nih.gov/19648918/

11. Pancreatic Cancer Risk Factors [Internet]. 2024 [cited 2024 Oct 17]. Available from: https://www.cancer.org/cancer/types/pancreatic-cancer/causes-risks-prevention/risk-factors.html

12. Ratner B, Bonislawski A. Early Detection: Catching Cancer When It's Curable. OR Books; 2024.

13. Li Y, Rao S, Solares JRA, Hassaine A, Ramakrishnan R, Canoy D, et al. BEHRT: Transformer for Electronic Health Records. Sci Rep. 2020 Apr 28;10(1):1–12.

14. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. npj Digital Medicine. 2021 May 20;4(1):1–13.

15. Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission [Internet]. 2019 [cited 2024 Jun 7]. Available from: http://arxiv.org/abs/1904.05342

16. Shang J, Ma T, Xiao C, Sun J. Pre-training of graph augmented transformers for medication recommendation. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI 2019. International Joint Conferences on Artificial Intelligence; 2019. p. 5953–9.

17. Rupp M, Peter O, Pattipaka T. ExBEHRT: Extended Transformer for Electronic Health Records. In: ICLR 2023 Workshop on Trustworthy Machine Learning for Healthcare [Internet]. 2023 [cited 2024 Jun 11]. Available from: https://openreview.net/pdf?id=C09sAJ960u

18. Klein AP, Lindström S, Mendelsohn JB, Steplowski E, Arslan AA, Bas Bueno-de-Mesquita H, et al. An Absolute Risk Model to Identify Individuals at Elevated Risk for Pancreatic Cancer in the General Population. PLoS One. 2013 Sep 13;8(9):e72311.

19. McGuigan A, Kelly P, Turkington RC, Jones C, Coleman HG, McCain RS. Pancreatic cancer: A review of clinical diagnosis, epidemiology, treatment and outcomes. World J Gastroenterol [Internet]. 2018 Nov 21 [cited 2024 Jun 6];24(43). Available from: https://pubmed.ncbi.nlm.nih.gov/30487695/

20. Baecker A, Kim S, Risch HA, Nuckols TK, Wu BU, Hendifar AE, et al. Do changes in health reveal the possibility of undiagnosed pancreatic cancer? Development of a risk-prediction model based on healthcare claims data. PLoS One. 2019 Jun 25;14(6):e0218580.

21. Appelbaum L, Cambronero JP, Stevens JP, Horng S, Pollick K, Silva G, et al. Development and validation of a pancreatic cancer risk model for the general population using electronic health records: An observational study. Eur J Cancer [Internet]. 2021 Jan [cited 2024 Jun 6];143. Available from: https://pubmed.ncbi.nlm.nih.gov/33278770/

22. Chen W, Zhou Y, Xie F, Butler RK, Jeon CY, Luong TQ, et al. Derivation and External Validation of Machine Learning-Based Model for Detection of Pancreatic Cancer. Official journal of the American College of Gastroenterology | ACG. 2023 Jan;118(1):157.

23. Coustasse A. Upcoding Medicare: is healthcare fraud and abuse increasing? Perspect Health Inf Manag [Internet]. 2021 [cited 2024 Jun 13];18(4). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8649706/

24. Joiner KA, Lin J, Pantano J. Upcoding in medicare: where does it matter most? Health Econ Rev. 2024 Jan 2;14(1):1.

25. Lassalle M, Le Tri T, Afchain P, Camus M, Kirchgesner J, Zureik M, et al. Use of Proton Pump Inhibitors and Risk of Pancreatic Cancer: A Nationwide Case–Control Study Based on the French National Health Data System (SNDS). Cancer Epidemiol Biomarkers Prev. 2022 Mar 3;31(3):662.

26. Hong HE, Kim AS, Kim MR, Ko HJ, Jung MK. Does the Use of Proton Pump Inhibitors Increase the Risk of Pancreatic Cancer? A Systematic Review and Meta-Analysis of Epidemiologic Studies. Cancers [Internet]. 2020 Aug [cited 2024 Jun 10];12(8). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7463819/

27. Kearns MD, Boursi B, Yang YX. Proton pump inhibitors on pancreatic cancer risk and survival. Cancer Epidemiol. 2017 Feb;46:80.

28. Aljofan M, Riethmacher D. Anticancer activity of metformin: a systematic review of the literature. Future Science OA [Internet]. 2019 Sep [cited 2024 Jun 12];5(8). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6745597/

29. Harada N. Effects of metformin on blood glucose levels and bodyweight mediated through intestinal effects. J Diabetes Investig. 2020 Nov;11(6):1420.

30. Diabetes Prevention Program Research Group. Reduction in the Incidence of Type 2 Diabetes with Lifestyle Intervention or Metformin. 2002 Feb 7 [cited 2024 Jun 26]; Available from: https://www.nejm.org/doi/full/10.1056/NEJMoa012512

31. Cheng X, Liu YM, Li H, Zhang X, Lei F, Qin JJ, et al. Metformin Is Associated with Higher Incidence of Acidosis, but Not Mortality, in Individuals with COVID-19 and Pre-existing Type 2 Diabetes. Cell Metab [Internet]. 2020 Oct 6 [cited 2024 Jun 12];32(4). Available from: https://pubmed.ncbi.nlm.nih.gov/32861268/

32. Xian H, Liu Y, Nilsson AR, Gatchalian R, Crother TR, Tourtellotte WG, et al. Metformin inhibition of mitochondrial ATP and DNA synthesis abrogates NLRP3 inflammasome activation and pulmonary inflammation. Immunity. 2021 Jul 13;54(7):1463–77.e11.

33. Topaloglu U, Palchuk MB. Using a Federated Network of Real-World Data to Optimize Clinical Trials Operations. JCO clinical cancer informatics [Internet]. 2018 Dec [cited 2024 Jun 6];2. Available from: https://pubmed.ncbi.nlm.nih.gov/30652541/

34. Placido D, Yuan B, Hjaltelin JX, Zheng C, Haue AD, Chmura PJ, et al. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. Nat Med. 2023 May;29(5):1113–22.

35. Sundararajan M, Taly A, Yan Q. Axiomatic Attribution for Deep Networks [Internet]. 2017 [cited 2024 Jun 11]. Available from: http://arxiv.org/abs/1703.01365

36. Ben Q, Xu M, Ning X, Liu J, Hong S, Huang W, et al. Diabetes mellitus and risk of pancreatic cancer: A meta-analysis of cohort studies. Eur J Cancer [Internet]. 2011 Sep [cited 2024 Oct 18];47(13). Available from: https://pubmed.ncbi.nlm.nih.gov/21458985/

37. Swierczynski J, Hebanowska A, Sledzinski T. Role of abnormal lipid metabolism in development, progression, diagnosis and therapy of pancreatic cancer. World J Gastroenterol. 2014 Mar 7;20(9):2279.

38. Bansal P, Sonnenberg A. Pancreatitis is a risk factor for pancreatic cancer. Gastroenterology. 1995 Jul;109(1):247–51.

39. Grivennikov SI, Greten FR, Karin M. Immunity, inflammation, and cancer. Cell [Internet]. 2010 Mar 19 [cited 2024 Oct 18];140(6). Available from: https://pubmed.ncbi.nlm.nih.gov/20303878/

40. Roy A, Sahoo J, Kamalanathan S, Naik D, Mohan P, Kalayarasan R. Diabetes and pancreatic cancer: Exploring the two-way traffic. World J Gastroenterol. 2021 Aug 14;27(30):4939.

41. Chan MT, Lim GE, Skovsø S, Yang YHC, Albrecht T, Alejandro EU, et al. Effects of insulin on

human pancreatic cancer progression modeled in vitro. BMC Cancer. 2014 Nov 6;14:814.

42. Zhang Y, Wang QL, Yuan C, Lee AA, Babic A, Ng K, et al. Pancreatic cancer is associated with medication changes prior to clinical diagnosis. Nat Commun. 2023 Apr 28;14(1):2437.

43. Boudreau DM, Yu O, Johnson J. Statin Use and Cancer Risk: A Comprehensive Review. Expert Opin Drug Saf. 2010 Jul;9(4):603.

44. Bosetti C, Rosato V, Gallus S, Cuzick J, La Vecchia C. Aspirin and cancer risk: a quantitative review to 2011. Ann Oncol. 2012 Jun 1;23(6):1403–15.

45. Usc-Isi NM, Usc-Isi FM, Usc-Isi NS, Usc-Isi KL, Usc-Isi AG. A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys (CSUR) [Internet]. 2021 Jul 13 [cited 2024 Jun 6]; Available from: https://dl.acm.org/doi/10.1145/3457607

46. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science [Internet]. 2019 Oct 25 [cited 2024 Jun 10]; Available from: https://www.science.org/doi/10.1126/science.aax2342

47. Huang J, Gretton A, Borgwardt K, Schölkopf B, Smola A. Correcting Sample Selection Bias by Unlabeled Data. Adv Neural Inf Process Syst [Internet]. 2006 [cited 2024 Jun 10];19. Available from: https://proceedings.neurips.cc/paper_files/paper/2006/file/a2186aa7c086b46ad4e8bf81e2a3a19b-Paper.pdf

48. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need [Internet]. 2017 [cited 2024 Jun 13]. Available from: http://arxiv.org/abs/1706.03762

49. Merrill RM. Environmental Epidemiology: Principles and Methods. Jones & Bartlett Learning; 2008. 495 p.

1. Rahib L, Smith BD, Aizenberg R, Rosenzweig AB, Fleshman JM, Matrisian LM. Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. Cancer Res [Internet]. 2014 Jun 1 [cited 2024 Jun 10];74(11). Available from: https://pubmed.ncbi.nlm.nih.gov/24840647/

2. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. CA: A Cancer Journal for Clinicians. 2024 Jan 1;74(1):12–49.

3. Jia K, Kundrot S, Palchuk MB, Warnick J, Haapala K, Kaplan ID, et al. A pancreatic cancer risk prediction model (Prism) developed and validated on large-scale US clinical data. EBioMedicine [Internet]. 2023 Dec [cited 2024 Jun 6];98. Available from: https://pubmed.ncbi.nlm.nih.gov/38007948/

4. Price LE, Shea K, Gephart S. The Veterans Affairs's Corporate Data Warehouse: Uses and Implications for Nursing Research and Practice. Nurs Adm Q. 2015;39(4):311.

5. Wolfgang CL, Herman JM, Laheru DA, Klein AP, Erdek MA, Fishman EK, et al. Recent progress in pancreatic cancer. CA Cancer J Clin. 2013 Sep;63(5):318–48.

6. Klein AP, Wolpin BM, Risch HA, Stolzenberg-Solomon RZ, Mocci E, Zhang M, et al. Genome-wide meta-analysis identifies five new susceptibility loci for pancreatic cancer. Nat Commun. 2018 Feb 8;9(1):1–11.

7. Kim J, Yuan C, Babic A, Bao Y, Clish CB, Pollak MN, et al. Genetic and Circulating Biomarker Data Improve Risk Prediction for Pancreatic Cancer in the General Population. Cancer Epidemiol Biomarkers Prev [Internet]. 2020 May [cited 2024 Jun 18];29(5). Available from: https://pubmed.ncbi.nlm.nih.gov/32321713/

8. Li D, Duell EJ, Yu K, Risch HA, Olson SH, Kooperberg C, et al. Pathway analysis of genome-wide association study data highlights pancreatic development genes as susceptibility factors for pancreatic cancer. Carcinogenesis. 2012 Apr 20;33(7):1384–90.

9. Petersen GM, Amundadottir L, Fuchs CS, Kraft P, Stolzenberg-Solomon RZ, Jacobs KB, et al. A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. Nat Genet. 2010 Jan 24;42(3):224–8.

10. Amundadottir L, Kraft P, Stolzenberg-Solomon RZ, Fuchs CS, Petersen GM, Arslan AA, et al. Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. Nat Genet [Internet]. 2009 Sep [cited 2024 Jun 13];41(9). Available from: https://pubmed.ncbi.nlm.nih.gov/19648918/

11. Pancreatic Cancer Risk Factors [Internet]. 2024 [cited 2024 Oct 17]. Available from: https://www.cancer.org/cancer/types/pancreatic-cancer/causes-risks-prevention/risk-factors.html

12. Ratner B, Bonislawski A. Early Detection: Catching Cancer When It’s Curable. OR Books; 2024.

13. Li Y, Rao S, Solares JRA, Hassaine A, Ramakrishnan R, Canoy D, et al. BEHRT: Transformer for Electronic Health Records. Sci Rep. 2020 Apr 28;10(1):1–12.

14. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. npj Digital Medicine. 2021 May 20;4(1):1–13.

15. Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission [Internet]. 2019 [cited 2024 Jun 7]. Available from: http://arxiv.org/abs/1904.05342

16. Shang J, Ma T, Xiao C, Sun J. Pre-training of graph augmented transformers for medication recommendation. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI 2019. International Joint Conferences on Artificial Intelligence; 2019. p.

5953–9.

17. Rupp M, Peter O, Pattipaka T. ExBEHRT: Extended Transformer for Electronic Health Records. In: ICLR 2023 Workshop on Trustworthy Machine Learning for Healthcare [Internet]. 2023 [cited 2024 Jun 11]. Available from: https://openreview.net/pdf?id=C09sAJ960u

18. Klein AP, Lindström S, Mendelsohn JB, Steplowski E, Arslan AA, Bas Bueno-de-Mesquita H, et al. An Absolute Risk Model to Identify Individuals at Elevated Risk for Pancreatic Cancer in the General Population. PLoS One. 2013 Sep 13;8(9):e72311.

19. McGuigan A, Kelly P, Turkington RC, Jones C, Coleman HG, McCain RS. Pancreatic cancer: A review of clinical diagnosis, epidemiology, treatment and outcomes. World J Gastroenterol [Internet]. 2018 Nov 21 [cited 2024 Jun 6];24(43). Available from: https://pubmed.ncbi.nlm.nih.gov/30487695/

20. Baecker A, Kim S, Risch HA, Nuckols TK, Wu BU, Hendifar AE, et al. Do changes in health reveal the possibility of undiagnosed pancreatic cancer? Development of a risk-prediction model based on healthcare claims data. PLoS One. 2019 Jun 25;14(6):e0218580.

21. Appelbaum L, Cambronero JP, Stevens JP, Horng S, Pollick K, Silva G, et al. Development and validation of a pancreatic cancer risk model for the general population using electronic health records: An observational study. Eur J Cancer [Internet]. 2021 Jan [cited 2024 Jun 6];143. Available from: https://pubmed.ncbi.nlm.nih.gov/33278770/

22. Chen W, Zhou Y, Xie F, Butler RK, Jeon CY, Luong TQ, et al. Derivation and External Validation of Machine Learning-Based Model for Detection of Pancreatic Cancer. Official journal of the American College of Gastroenterology | ACG. 2023 Jan;118(1):157.

23. Coustasse A. Upcoding Medicare: is healthcare fraud and abuse increasing? Perspect Health Inf Manag [Internet]. 2021 [cited 2024 Jun 13];18(4). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8649706/

24. Joiner KA, Lin J, Pantano J. Upcoding in medicare: where does it matter most? Health Econ Rev. 2024 Jan 2;14(1):1.

25. Lassalle M, Le Tri T, Afchain P, Camus M, Kirchgesner J, Zureik M, et al. Use of Proton Pump Inhibitors and Risk of Pancreatic Cancer: A Nationwide Case–Control Study Based on the French National Health Data System (SNDS). Cancer Epidemiol Biomarkers Prev. 2022 Mar 3;31(3):662.

26. Hong HE, Kim AS, Kim MR, Ko HJ, Jung MK. Does the Use of Proton Pump Inhibitors Increase the Risk of Pancreatic Cancer? A Systematic Review and Meta-Analysis of Epidemiologic Studies. Cancers [Internet]. 2020 Aug [cited 2024 Jun 10];12(8). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7463819/

27. Kearns MD, Boursi B, Yang YX. Proton pump inhibitors on pancreatic cancer risk and survival. Cancer Epidemiol. 2017 Feb;46:80.

28. Aljofan M, Riethmacher D. Anticancer activity of metformin: a systematic review of the literature. Future Science OA [Internet]. 2019 Sep [cited 2024 Jun 12];5(8). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6745597/

29. Harada N. Effects of metformin on blood glucose levels and bodyweight mediated through intestinal effects. J Diabetes Investig. 2020 Nov;11(6):1420.

30. Diabetes Prevention Program Research Group. Reduction in the Incidence of Type 2 Diabetes with Lifestyle Intervention or Metformin. 2002 Feb 7 [cited 2024 Jun 26]; Available from: https://www.nejm.org/doi/full/10.1056/NEJMoa012512

31. Cheng X, Liu YM, Li H, Zhang X, Lei F, Qin JJ, et al. Metformin Is Associated with Higher Incidence of Acidosis, but Not Mortality, in Individuals with COVID-19 and Pre-existing Type 2 Diabetes. Cell Metab [Internet]. 2020 Oct 6 [cited 2024 Jun 12];32(4). Available from: https://pubmed.ncbi.nlm.nih.gov/32861268/

32. Xian H, Liu Y, Nilsson AR, Gatchalian R, Crother TR, Tourtellotte WG, et al. Metformin inhibition of mitochondrial ATP and DNA synthesis abrogates NLRP3 inflammasome activation and pulmonary inflammation. Immunity. 2021 Jul 13;54(7):1463–77.e11.

33. Topaloglu U, Palchuk MB. Using a Federated Network of Real-World Data to Optimize Clinical Trials Operations. JCO clinical cancer informatics [Internet]. 2018 Dec [cited 2024 Jun 6];2. Available from: https://pubmed.ncbi.nlm.nih.gov/30652541/

34. Placido D, Yuan B, Hjaltelin JX, Zheng C, Haue AD, Chmura PJ, et al. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. Nat Med. 2023 May;29(5):1113–22.

35. Sundararajan M, Taly A, Yan Q. Axiomatic Attribution for Deep Networks [Internet]. 2017 [cited 2024 Jun 11]. Available from: http://arxiv.org/abs/1703.01365

36. Ben Q, Xu M, Ning X, Liu J, Hong S, Huang W, et al. Diabetes mellitus and risk of pancreatic cancer: A meta-analysis of cohort studies. Eur J Cancer [Internet]. 2011 Sep [cited 2024 Oct 18];47(13). Available from: https://pubmed.ncbi.nlm.nih.gov/21458985/

37. Swierczynski J, Hebanowska A, Sledzinski T. Role of abnormal lipid metabolism in development, progression, diagnosis and therapy of pancreatic cancer. World J Gastroenterol. 2014 Mar 7;20(9):2279.

38. Bansal P, Sonnenberg A. Pancreatitis is a risk factor for pancreatic cancer. Gastroenterology. 1995 Jul;109(1):247–51.

39. Grivennikov SI, Greten FR, Karin M. Immunity, inflammation, and cancer. Cell [Internet].

2010 Mar 19 [cited 2024 Oct 18];140(6). Available from: https://pubmed.ncbi.nlm.nih.gov/20303878/

40. Roy A, Sahoo J, Kamalanathan S, Naik D, Mohan P, Kalayarasan R. Diabetes and pancreatic cancer: Exploring the two-way traffic. World J Gastroenterol. 2021 Aug 14;27(30):4939.

41. Chan MT, Lim GE, Skovsø S, Yang YHC, Albrecht T, Alejandro EU, et al. Effects of insulin on human pancreatic cancer progression modeled in vitro. BMC Cancer. 2014 Nov 6;14:814.

42. Zhang Y, Wang QL, Yuan C, Lee AA, Babic A, Ng K, et al. Pancreatic cancer is associated with medication changes prior to clinical diagnosis. Nat Commun. 2023 Apr 28;14(1):2437.

43. Boudreau DM, Yu O, Johnson J. Statin Use and Cancer Risk: A Comprehensive Review. Expert Opin Drug Saf. 2010 Jul;9(4):603.

44. Bosetti C, Rosato V, Gallus S, Cuzick J, La Vecchia C. Aspirin and cancer risk: a quantitative review to 2011. Ann Oncol. 2012 Jun 1;23(6):1403–15.

45. Usc-Isi NM, Usc-Isi FM, Usc-Isi NS, Usc-Isi KL, Usc-Isi AG. A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys (CSUR) [Internet]. 2021 Jul 13 [cited 2024 Jun 6]; Available from: https://dl.acm.org/doi/10.1145/3457607

46. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science [Internet]. 2019 Oct 25 [cited 2024 Jun 10]; Available from: https://www.science.org/doi/10.1126/science.aax2342

47. Huang J, Gretton A, Borgwardt K, Schölkopf B, Smola A. Correcting Sample Selection Bias by Unlabeled Data. Adv Neural Inf Process Syst [Internet]. 2006 [cited 2024 Jun 10];19. Available from: https://proceedings.neurips.cc/paper_files/paper/2006/file/a2186aa7c086b46ad4e8bf81e2a3a19b-Paper.pdf

48. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need [Internet]. 2017 [cited 2024 Jun 13]. Available from: http://arxiv.org/abs/1706.03762

49. Merrill RM. Environmental Epidemiology: Principles and Methods. Jones & Bartlett Learning; 2008. 495 p.

Supplementary Material

Detailed Methodology

Data source - US VA Corporate Data Warehouse

We use longitudinal ICD codes and medication data from the Department of Veterans Affairs Corporate Data Warehouse (CDW) derived in The Observational Medical Outcomes Partnership (OMOP5) format. The CDW is one of the largest databases of health data collected from VA institutions across the United States, including both inpatient and outpatient settings. Patient diagnosis is recorded in both ICD9 and ICD10 codes. We obtained diagnoses for 15,933,326 patients from 1999 to 2022 directly from the CDW as described previously[37].

The medication data in CDW is challenging to use directly because it is dispersed across multiple tables, which requires data extraction and integration efforts to obtain medication histories of patients. To overcome this complexity, we obtained medication usage information by leveraging the advantage of an OMOP (version 5) database, which is derived from CDW and has data processed for research use.

In brief, the *drug_exposure* table in OMOP was constructed by mapping and aggregating multiple raw medication tables from CDW. Subsequently, a standardized algorithm developed by the observational health data sciences and informatics (OHDSI) community was used to process medication information from the *drug_exposure* table to form the *drug_era* table[1]. The *drug_era* table contains continuous intervals for each drug, representing periods during which a patient is assumed to be exposed to the drug, categorized under the concept class "*Ingredient*" in the OMOP common data model. We extracted all medication usage information for each patient from the *drug_era* table.

To ensure relevance and accuracy of the medication data used for training the model, we included only active medications. Inactive ingredients were filtered out using the FDA's inactive ingredient information[2]. The following sections describe in detail how we constructed the training and evaluation datasets (Figure 1S).

---

[1] http://ohdsi.github.io/CommonDataModel/sqlScripts.html#Drug_Eras

[2] https://www.fda.gov/drugs/drug-approvals-and-databases/inactive-ingredients-database-download

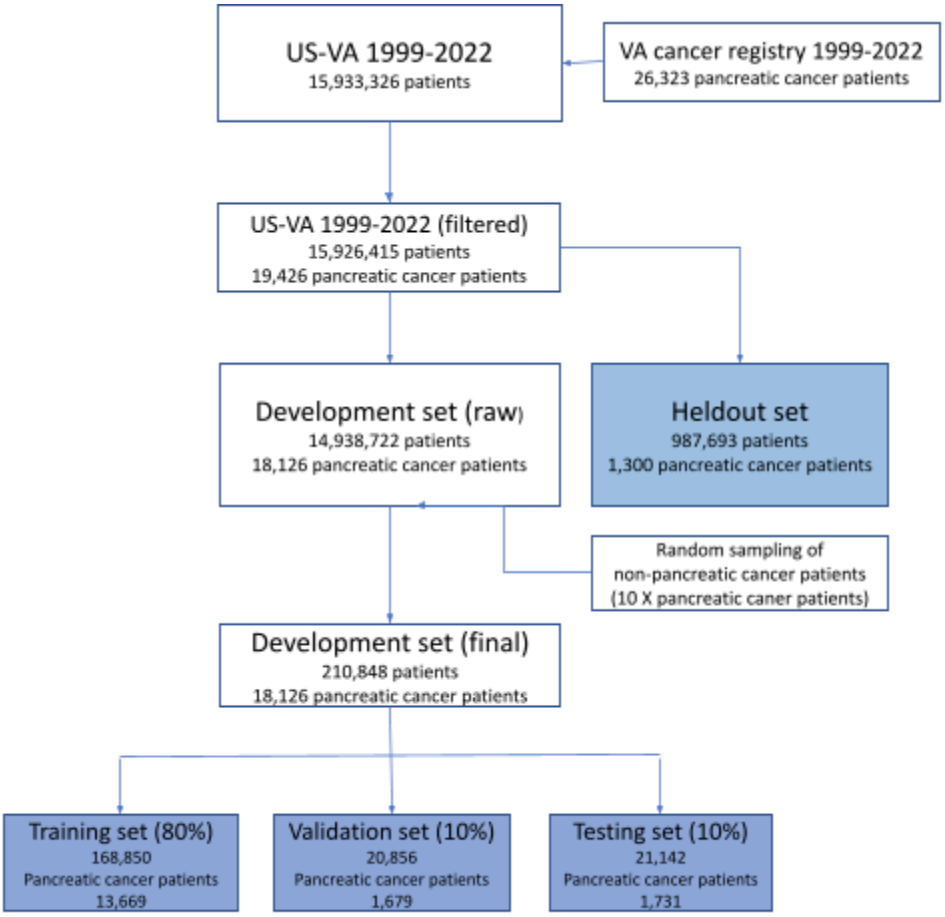


Figure 1S. Flowchart of data extraction and preparation for training ML models.

Filtering Data Against US VA Cancer Registry

Due to the issue of upcoding of diagnosis codes discussed in the introduction, the ICD codes for patients often do not accurately reflect their actual health conditions. This discrepancy complicates the identification of true pancreatic cancer cases; simply relying on the presence of ICD codes denoting pancreatic cancer would result in a high number of false positive labels (nearly 50%, data not shown). To remedy this issue, we retrieve a list of verified pancreatic cancer patients from the VA Cancer Registry, a centralized repository of cancer patients data within the VA system. Inclusion of patients in the registry is governed by stringent criteria and verification steps that extend beyond individual ICD codes.

Using a set of verified positive cases from the cancer registry, we filter out our dataset by excluding any patients who had ICD codes indicating pancreatic cancer but did not appear in the cancer registry. Conversely, we also excluded patients listed in the cancer registry who lacked corresponding ICD codes in the CDW EHR, although such

instances were rare. The filtering step enhances the reliability and accuracy of pancreatic cancer labels in our dataset.

Construction of a development and a retrospective evaluation set

To construct our training and evaluation datasets, we first extracted ~10 million patient records from the CDW. We hold-out ~1 million of these patients to form a retrospective evaluation set. From the remainder, we include all verified positive pancreatic cancer patients and a subset of negative patients, resulting in a development set totaling 210,848 patients. This approach reduces the class imbalance between positive and negative cases in the training process, which improves the model's ability to learn effectively from both classes – particularly from a diverse set of positive cancer histories, while managing computational resources. This set is used during model development, and a retrospective evaluation set is used to assess performance after model development is completed.

Data preprocessing

We use trajectory sampling to increase the number and diversity of patient history sequences used during training. Trajectory sampling involves sampling multiple subsequences from patients' histories, providing a larger and more varied set of inputs for the model. Specifically, for each patient, we randomly select an event within their sequence of ICD codes and medication records. We then use the patient's history up to that event as an input data and predict the risk from that event forward. We refer to these sampled subsequences as patient 'trajectories'. This approach allows the model to learn from different stages of a patient's medical history, capturing a wider range of health conditions.

For patients with pancreatic cancer, trajectories are sampled from time intervals before the cancer diagnosis. For patients without pancreatic cancer, trajectories are sampled from at least two years before the end of their record to minimize cases where a patient developed pancreatic cancer shortly after the prediction timepoint, which might not be recorded in their EHR history, e.g., if they leave the health care system. This sampling procedure means that positive cancer patients can produce trajectories with negative labels. For e.g., if a patient's history is taken up to an event in the fifth year of their record, but they do not develop cancer until the fifteenth recorded year (past the largest predicted time interval of our model). Such trajectories provide challenging negative samples, as ideally the model should learn not only that someone may get cancer, but also predict a specific time interval in which this could occur.The sampling process further addresses the imbalance caused by low incidence of pancreatic cancer. The purpose of the test set is to evaluate the model's performance on data with a data

distribution similar to the training set. Additionally, a separate hold-out set is used to assess the model's performance on data that more closely reflects a realistic setting.

During training, four random trajectories are sampled for each patient in each epoch. At evaluation time, rather than random sampling, we take the last four trajectories before either the time of cancer diagnosis or two years before the end of the record (depending on whether the patient has cancer). This ensures that different models are evaluated on a consistent set of trajectories, using the most information available in the patient records.

Risk-prediction model development

This section describes the ML framework that uses the longitudinal nature of EHRs for PDAC risk prediction. The input to the model consists of contiguous subsequences of diagnosis and prescription codes, along with their corresponding timestamps, extracted from a patient's medical record. These subsequences enable the model to assess a patient's evolving cancer risk over a series of time points leading up to a potential cancer diagnosis.

Transformer encoder

The first step in the model's architecture is to transform each event (i.e. one diagnostic or medication code) in a trajectory into a continuous feature representation, achieved using a two-step process. First, each event is mapped to a low-dimensional continuous feature vector using an embedding lookup table (learned during training). Next, positional encodings are added to the embeddings to account for the temporal order of events within a trajectory. These positional encodings are derived from the timestamps following (Placido et al., 2023) and added to the embeddings of respective events. Additionally, patient age, a known risk factor for pancreatic cancer, is also encoded in a similar manner. This encoding allows the model to handle irregularly spaced events within trajectories. In this way, a trajectory of length N is mapped to a feature matrix $X \in R^{N \times d_{in}}$, where d is the dimensionality of the latent space.

Positional encodings. To incorporate temporal information, we apply a time encoding method based on the timestamps associated with the sequence events. This approach is particularly useful for capturing the temporal dynamics of medical events and aids in handling trajectories with irregular intervals. The first step involves generating positional time embeddings P, using a set of multipliers, defined as follows,

$$M = \frac{2\pi}{linspace(start, end, d_{in})} \in R^{d_{in}}$$

where $start$ and end denote the minimum and maximum values for the embedding range, $linspace$ generates linearly spaced d_{in} points between $start$ and end. Given the time deltas Δt, for each event in the sequence, the positional embeddings are computed using the cosine function as, $P = \cos(\Delta t \cdot M) \in R^{N \times d_{in}}$. The embeddings P are then integrated with the original feature matrix X by a combination of two transformations: E_{scale} that scales the original features, and E_{add} that analogous to a bias is added to features. Both transformations are derived from P using linear transformations $W_{scale}, W_{bias} \in R^{d_{in} \times d_{in}}$ and ,

$$E_{scale} = PW_{scale}, \quad E_{add} = PW_{add}$$

The final time-informed feature embedding is computed as:

$$X = E_{scale} \odot X + E_{add}$$

where $\odot$ denotes the element-wise multiplication.

Transformer neural network. Next, we use a Transformer architecture to capture sequential interdependencies between events within a trajectory. The Transformer uses a self-attention mechanism, enabling it to learn pairwise relationships between all events in the input trajectory (48). The Transformer's output is then aggregated over the entire trajectory length, resulting in a fixed-size latent representation that conceptually encapsulates the patient's health state at a specific time interval. Formally, consider Q, K, and V be the query, key, and value matrices, respectively, obtained by linear projections of the input X,

$$Q = XW_Q^T, \; K = XW_K^T, \; V = XW_V^T$$

where $W_Q, W_K, W_V \in R^{d \times d_{in}}$, and d is the dimensionality of a projection space. The attention mechanism is then defined as,

$$A = \frac{QK^T}{\sqrt{d}}, \qquad Z = \sigma(A)V$$

here A is a self-attention matrix, σ is softmax along the column, and Z is the intermediate representation of an input trajectory. It is typical to use a Transformer with multiple heads that increase its capacity to model intricate relationships. The multi-headed attention mechanism combines multiple attention heads to capture different types of dependencies:

$$Z_{agg} = [Z_1 \cdots Z_H] \, W_{agg}$$

where, $[\cdots]$ is a concatenation along the columns, $W_{agg} \in R^{Hd\times d}$ is a learnable weight matrix that combines the outputs of the attention heads, and Z_{agg} is the latent representation aggregated across all heads. The output of the multi-headed attention mechanism is passed through a feed-forward layer, layer normalization, a linear layer, and a pooling layer to obtain the fixed length encoding $z \in R^{d}$ of an input trajectory.

Risk regression model. In the final step, the fixed-length latent representation z serves as covariates in a risk prediction model. Specifically, a regression model includes parameters $\beta \in R^{d\times 5}$, bias $b \in R^{d\times 1}$that predicts the risk within five pre-defined time intervals (e.g., 0-3, 3-6, 6-12, 12-36, 36-60 months),

$$\lambda(t|z) = ReLU(z^{T}\beta + b)$$

where $ReLU(x) = max(0, x)$ stands for rectified-linear unit. The cumulative risk of cancer occurrence for time intervals, such as, 0-3, 0-6, 0-12, 0-36, 0-60 months, can then be obtained as,

$$s_{T}(z) = s(T \geq t \mid z) = \lambda(z) + \sum_{t}^{T} \lambda(t|z)$$

where $\lambda(z) = z^{T}c$ is a baseline risk of an input trajectory parametrized through $c \in R^{d\times 1}$; the cumulative probability of cancer in the respective time interval is obtained as,

$$p_{T}(z) = sigmoid(s_{T}(z)).$$

For the five selected intervals, the respective probabilities $\{p_1, p_2, p_3, p_4, p_5\}$ are optimized using binary cross entropy loss against a target sequence $\{y_1, y_2, y_3, y_4, y_5\}$. The parameters of the risk model, embedding layer, and Transformer layer together referred as θ are learned jointly by optimizing the combined binary cross-entropy loss over the five chosen time intervals.

$$loss = \frac{1}{N}\frac{1}{N_T} \sum_{i,t}^{N\,N_T} [y_{i,t} \log(p_t(z_i)) + (1 - y_{i,t}) \log(1 - p_t(z_i))] + \lambda_2 ||\theta||_2$$

Training and evaluation setup

The development dataset of VA subjects described in Table 1 was split into training, development, and test datasets according to an 80:10:10 split. This test set simply gives us an estimate of performance on data with an inflated number of positive cases. This is a useful sanity check after model training, but not strictly necessary as our final reported performance is based on a held-out evaluation set. Transformer models were trained using the training dataset, the best performing model was chosen using development

datasets. Due to the low incidence of pancreatic cancer patients, balanced sampling of the trajectories of patients in the training set was used to obtain approximately equal numbers of trajectories from cancer and non-cancer patients. The binary cross-entropy loss function was used for model training with an initial learning rate of 0.001 and trained for 20 epochs. During training, if the validation AUROC failed to improve for more than four epochs, the model with the best validation AUROC up to that point was reloaded, the learning rate was halved, and training then continued. The final model saved after training was the one with the highest validation AUROC across all training epochs. All models were trained on an A100 GPU with 40GB of GPU memory.

Model architecture and hyperparameter setup.

We use a Transformer architecture as an encoder neural network in our model. Specifically, the encoder consists of *'L'* Transformer layers, each using a multi-head self-attention (MHSA) mechanism with '*H'* heads. Each has a dimensionality of *'d'* for the key, query and value transformations. The output of the MHSA module is combined with the original input through a residual connection and then normalized using a layer normalization to improve training stability. This is followed by a position-wise feed-forward network consisting of two fully connected layers of dimensionality 'd' with ReLU activation functions. The output of the feed-forward network is combined with its input via a second residual connection and layer normalization.

For risk regression, we implement a time-to-event probability transformation. This transformation includes two fully connected linear layers: one maps the input features (dimensionality *'d'*) to hazard rates across the five intervals, and the other computes a baseline hazard term. The hazard outputs are passed through ReLU activation to ensure they are non-negative and are combined with the baseline hazard to output the risk probability for five intervals (as discussed above).

We conducted a grid search over the following configurations of hyperparameters in the encoder network L={1,2,4}, H={4,8,16}, d={64, 128}. Additionally, we also included different values of learning rate lr={0.0005, 0.001, 0.005, 0.01}, weight decay={0, 0.01, 0.1, 0.5} in the Adam optimizer and dropout probability={0, 0.1, 0.2, 0.5}. The optimal hyperparameters were determined to be L=1, H=16, d=64, lr=0.001, and without dropout. We trained all models for a maximum of 100 epochs, monitoring convergence every five epochs.

Evaluation

Evaluation of each model was done using a retrospective set described in the second part of Table 1. The AUROC score was used for evaluation of overall model performance and the 95% confidence interval was computed using 100 resampling of bootstrapping with replacement. We further report the incidence ratio (IR), positive

predictive value (PPV) and standardized incidence ratio (SIR) for sets of highest-risk patients to quantify the benefit of our model in a realistic application of surveillance programs.

Standardized Incidence Ratio (SIR) is a score used to assess how well a risk prediction model can identify high-risk patients for a disease, such as PDAC. It compares the observed positive cases in a high-risk cohort to the expected baseline incidence. The expected baseline incidence in high-risk patients is calculated by summing the age-group-specific incidence rates from the standardized nationwide database, weighted by the number of patients in each age group, race, and sex combination present in the database.

Denote $AgeGroups = \{50 - 54,\ 55 - 59,\ \cdots,\ 85\ +\}$ as the set of age intervals used for risk stratification, $Race = \{Black,\ Whites,\ Others\ (Asian,\ Hispanic)\}$ as the set of racial groups, and $Gender = \{Male,\ Female\}$. Let H be the set of high-risk patients (age >= 50), $f^{x,y,s}$ be the number of patients with age x, race y, sex s in the database, $TP_{x,y,s}$ be a count of true positives within that group, and $Incidence(x,\ y,\ s)$ be the baseline PDAC incidence rate (per year) that can be obtained from the US-VA or the SEER database for the corresponding age, race, and sex stratum. SIR is then defined as,

$$SIR = \frac{Observed\ Positive\ Cases\ in\ H}{Expected\ baseline\ Incidence\ in\ H} = \frac{\sum_{x \in AgeGroup} \sum_{y \in Race} \sum_{s \in Sex} TP_{x,y,s}}{\sum_{x \in AgeGroup} \sum_{y \in Race} \sum_{s \in Sex} \sum_{t \in T} Incidence\ (x+t, y, s) \times f^{x,y,s}}$$

where T = {0, 1, 2, 3} is the time-window of cancer-occurence in units of years. For values of t less than 1 year, we round down to the previous year when querying the database. For example, when computing SIR for a patient at the 6 or 12 month timepoints there is no summation over T.

The 95% confidence interval of SIR is computed using the formula below (49),

$$95\%\ CI = \left(\sqrt{N_{Observed}} \pm 0.98\right)^2 / N_{Expected}$$

$N_{Observed}$ = Observed positive cases in high risk cohort H.

$N_{Expected}$ = Expected baseline incidence in high risk cohort H.

A SIR value greater than 1 indicates that the risk prediction model has identified a cohort with a higher-than-expected incidence rate than the baseline population, suggesting that the model effectively identifies high-risk individuals who are more likely to develop PDAC. Conversely, a SIR value less than 1 implies that the model is not accurately identifying the high-risk cohort. The SIR metric provides a standardized way

to evaluate the performance of a risk prediction model in identifying high-risk patients, accounting for differences in age, race, and sex distributions between the study population and the baseline population. It is particularly useful in PDAC, where early detection and risk stratification are crucial for improving patient outcomes.

Incidence Ratio (IR) is defined as the ratio of the precision of the prediction model to the baseline incidence rate obtained from the population,

$$IR = \frac{PPV}{Incidence}$$

where $Incidence$ refers to the baseline PDAC incidence rate also determined from the US-VA dataset and PPV stands for positive predictive value. The key difference between IR and SIR is that SIR adjusts for demographic factors such as age, race, and gender, while the IR does not account for these variables. Placido et al., 2023 refers to IR as relative risk (RR).

Positive Predictive Value (PPV) measures the proportion of true positive predictions among all positive predictions made by the model. Formally,

$$PPV = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)}$$

where TP stands for true positives and FP for false positives. A higher value means the model has a better rate of TPs relative to FPs. If the model predicts a positive outcome, PPV measures how probable it is that the prediction is correct.

Identifying clinically relevant codes using attribution analysis.

To highlight the most predictive features to our model, we use the Integrated Gradients (IG) (Sundarajan et al., 2017) method to compute feature attributions. IG provides estimates of the contributions of each individual feature in an input to the model output. To do so, IG computes a path integral over the gradients of the model output with respect to inputs. The integration path starts from an uninformative baseline input and ends at the input for which we are computing the attributions. Formally, this can be written as

$$IG_i(F, x, x^{base}) = (x_i - x_i^{base}) \times \int_{\alpha=0}^{1} \frac{\delta F(x^{base} + \alpha(x - x^{base}))}{\delta x_i} d\alpha$$

where $x,\ x^{base}$ and F are the input, baseline and model, respectively. IG_i is a local attribution method, and only computes the feature attributions for a single input.

To estimate the model's reliance on particular features across many examples, we compute the fraction of the time an ICD code or medication occurs in the top 10% highest attributed features in an input sequence. This approach allows us to identify diagnoses or medications that are important to model predictions. Relevant features should consistently rank highly across many inputs, while less important features should appear in the top rankings less frequently. By computing the fraction of the time a code is highly ranked and normalizing this by its overall occurrence, we ensure that frequently occurring codes are not deemed important solely due to their high frequency. In summary, this attribution approach highlights the global importance of codes while accounting for contextual relevance across trajectories.

We apply IG with the above aggregation scheme to the 1000 highest and lowest predicted risk trajectories in our training data. This approach targets confident model predictions, avoiding less informative low or mid risk features. True positive cases that have only a moderate level of predicted risk would also likely have uninformative attributions, since those attributions are a function of the model. Taking 1000 of those with the highest predicted risk means that the true positives are also predicted to be very likely positive, and therefore we would expect the attributions to look reasonable.

Baseline models

We compare our proposed model with two baseline methods: a MLP and LR. The LR model uses a 'bag-of-events' representation of a patient's ICD and medication history as input and produces five monotonically increasing risk scores using the risk regression model as discussed above. This representation is obtained by transforming the sequence of ICD and medication codes into a vector, where each dimension corresponds to the frequency of a specific code normalized by the total number of codes in the trajectory. Although this encoding captures information about the occurrence of codes, it ignores their temporal ordering. In contrast, the MLP uses the same input embeddings as our Transformer model but processes each embedding independently through two fully connected layers, instead of using self-attention to learn interdependencies among the codes. The output of the MLP is obtained by averaging over tokens and passed through the same risk regression head used in the Transformer model to produce the five output risk scores.

Extended Results

Comparison with baseline models

Here, we compare the sequential model to non-sequential LR and MLP, both trained on a bag-of-events (no time stamps) representation of patient trajectories. The sequential

model, which uses a Transformer architecture, achieves the highest AUROC scores (Figure 2S), supporting the hypothesis that a temporal model of diagnostic and medication data is a more effective approach for cancer risk prediction.

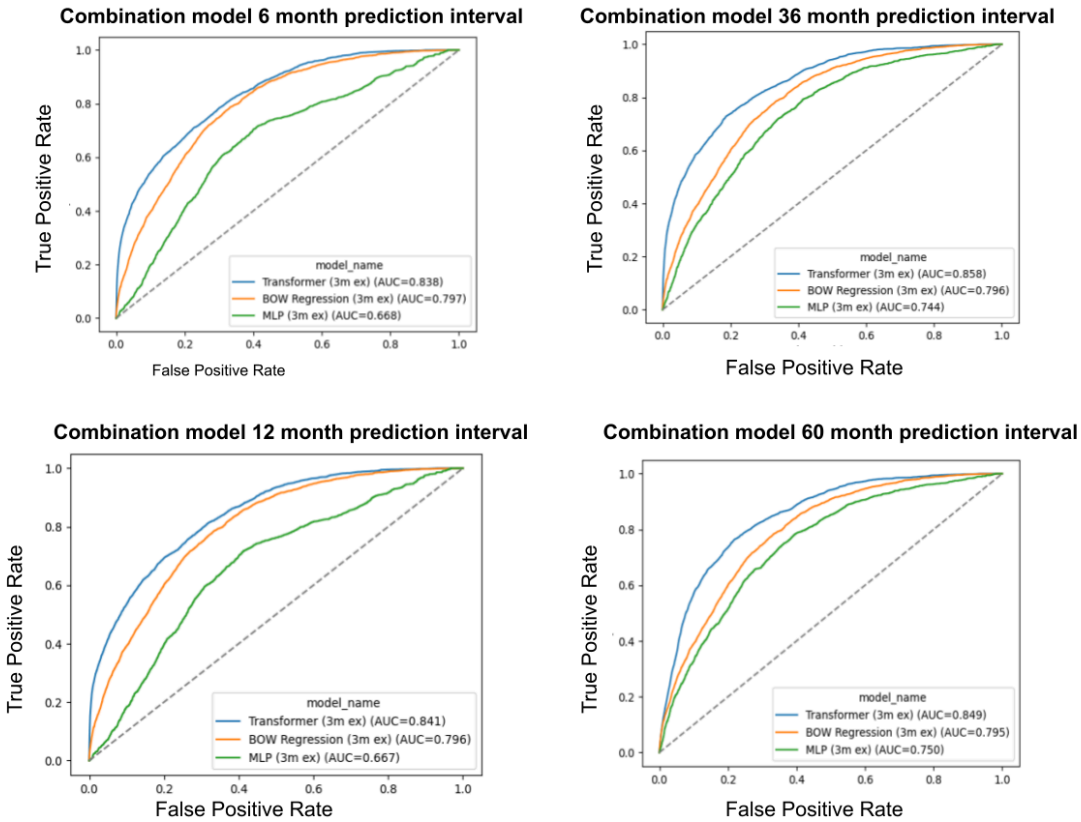


Figure 2S. Comparison of sequential models with baselines on a test dataset. *All baselines are trained on combined (medication and diagnosis) dataset. Each plot compares AUROC scores of the Transformer (sequential) model with two non-sequential baselines, the logistic regression (BOW = bag of words) and MLP for the 6, 12, 36, and 60 months prediction windows.*

Model trained without three month data exclusion

Following (Placido et al., 2023) we also run our model without data exclusion, results reported in Figure 3S. The prediction performance improved compared to the three month exclusion, likely due to the inclusion of quasi-symptoms that appear before an actual diagnosis. These results demonstrate the AI tool can capture features that a well-trained clinician might identify.

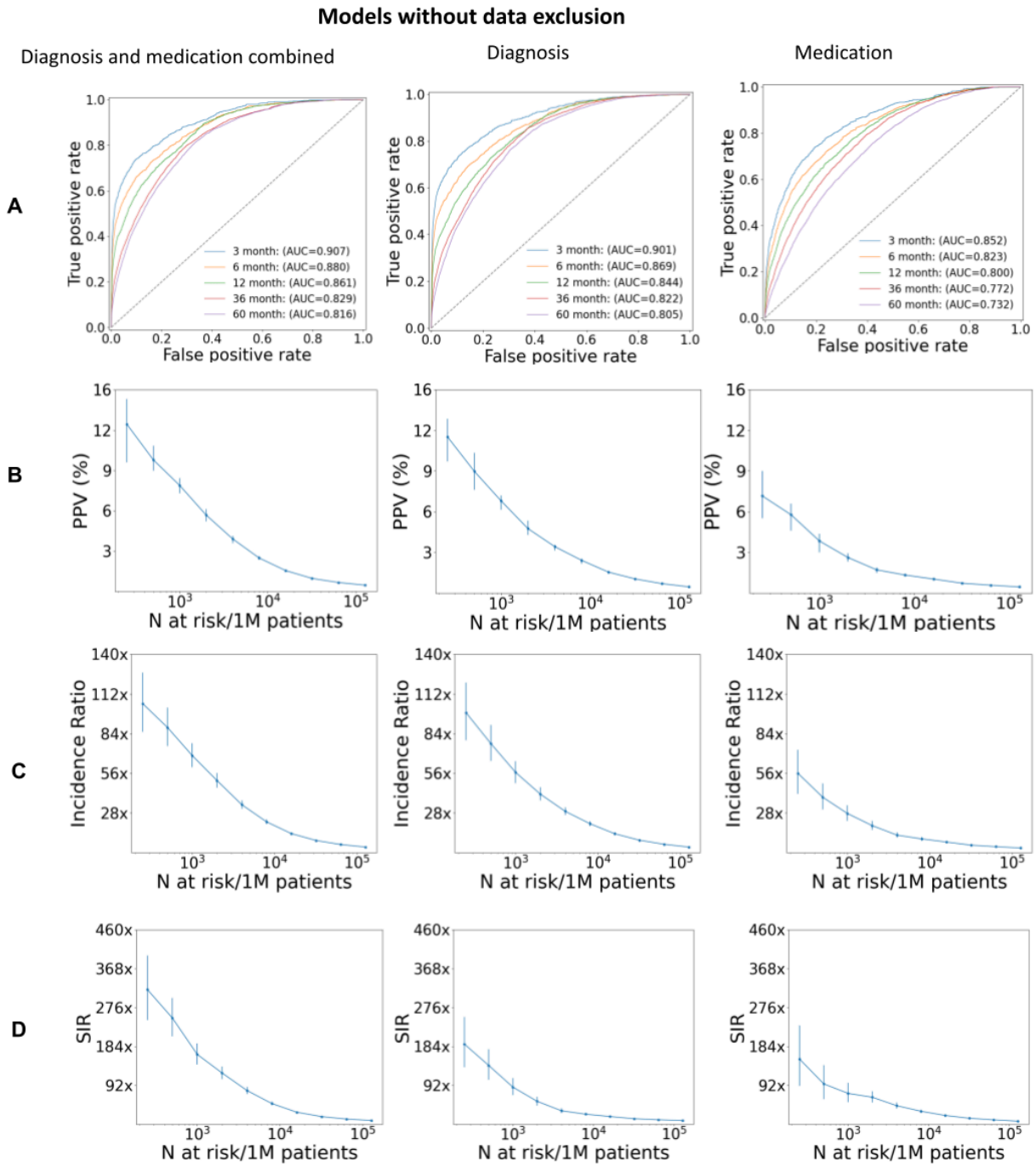


Figure 3S. Comparison of the performance of models trained without a data exclusion window: diagnostic data only, medication data only, and combination of both. *Left panels: combination of diagnosis and medication data. Middle panels: diagnosis data only. Right panel: medication data only. (A) AUROCs for different prediction time intervals, (B) Positive predictive value (PPV) for high-risk groups per 1 million patients, (C) Incidence ratio (IR) of high-risk groups per 1 million patients, and (D) Standardized incidence ratio (SIR) of high-risk groups per 1 million patients.*

Using SEER incidence rate in SIR calculation

The reported SIR and IR plots utilize baseline incidence data from the US-VA database. Next, we employ the Surveillance, Epidemiology, and End Results (SEER) database, which offers estimates of nationwide incidence rates. Using SEER incidence rates allows us to estimate model performance across a broader national population, extending beyond the veteran cohort. Results reported in Figure 4S.

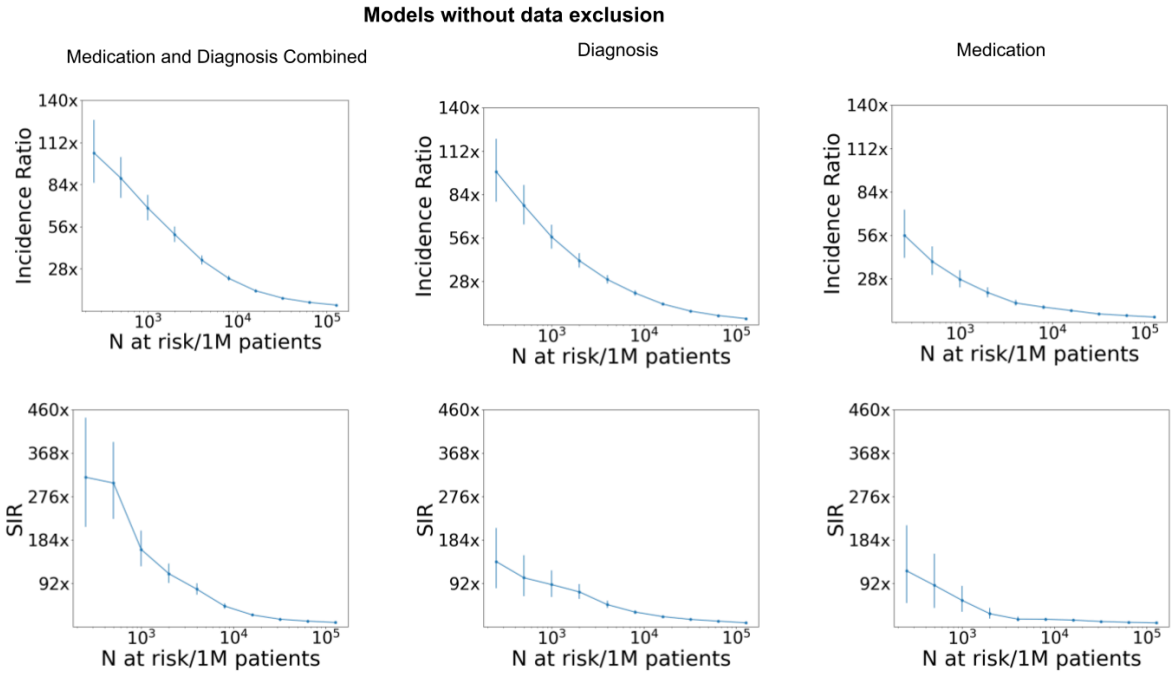


Figure 4S. Comparison of Incidence Ratio and SIR of models trained without a data exclusion window across three data types, using SEER incidence rate: diagnostic data only, medication data only, and combination of both. *Left panels: combination of diagnosis and medication data. Middle panels: diagnosis data only. Right panel: medication data only.*

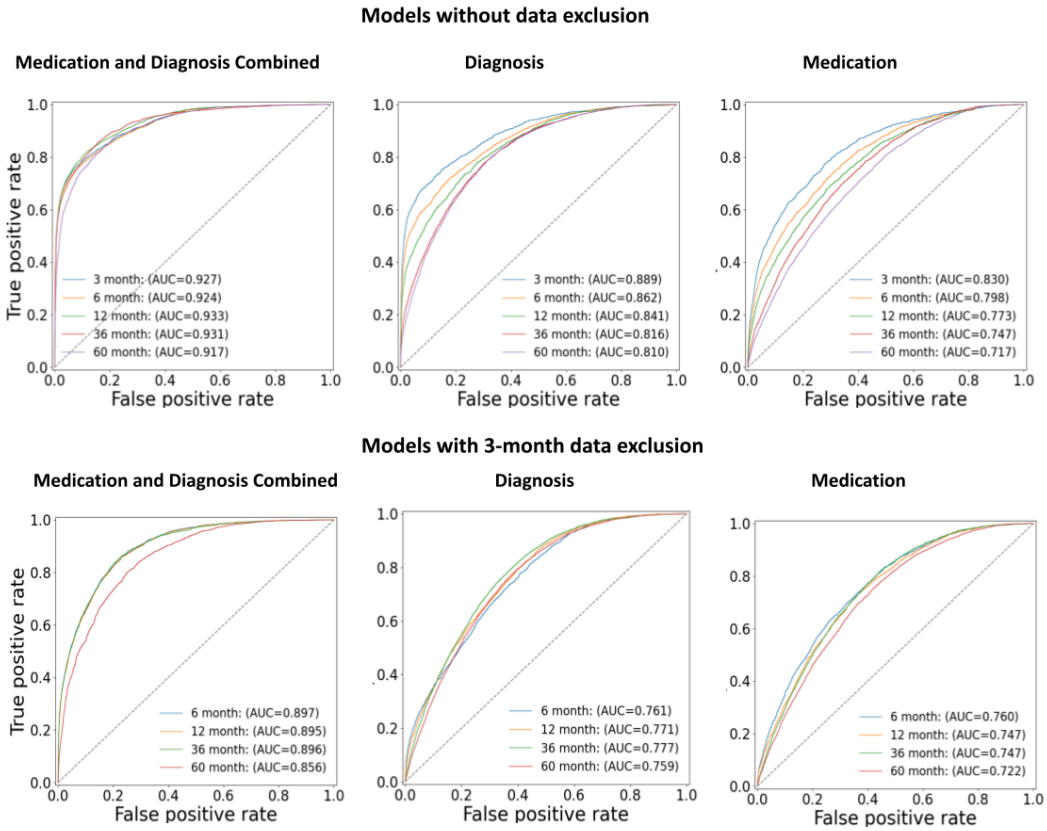


Figure 5S. *Evaluation on a held-out test set comprising 10% of the data reported in Table 1A, with a higher proportion of PDAC cases compared to the test data reported in Table 1B. These results reflect the model's performance on a test set that mirrors the cancer-to-non-cancer ratio present in the training set.*

Results on 10% test set

Figure 5S reports performance on the held-out test set that reflects the ratio of positives and negatives in the training set.

Baseline Incidence in US-VA and SEER database

Table 1S. Baseline incidence rate stratified by age, and race in US-VA and SEER database. *The SEER incidence were obtained using SEER*Stat software version 8.4.3*[3]*. The pancreatic cancer incidences in US-VA were computed using our* ***Incidence rate = (New cancers / Population)*** *with subsampling described in Data source.* For each sampled trajectory, the age, gender, sex and its pancreatic cancer status within 12 months were computed, the incidence of pancreatic cancer is defined by the ratio of the number of trajectories with pancreatic cancer to the total number of trajectories.

---

[3] Surveillance Research Program, National Cancer Institute SEER*Stat software (seer.cancer.gov/seerstat) version 8.4.3.

Age Group	Race (White, Black)	Sex	SEER Incidence	US-VA Incidence
50-54	black	Female	0.00013338	0.00036411
50-54	black	Male	0.00019127	8.512E-05
50-54	white	Female	7.779E-05	0.00020017
50-54	White	Male	0.00012217	0.00028827
55-59	black	Female	0.00023298	6.122E-05
55-59	black	Male	0.00033398	0.00081471
55-59	white	Female	0.00014521	0.00011076
55-59	White	Male	0.00022726	0.00049353
60-64	black	Female	0.00036045	0.00018398
60-64	black	Male	0.00051581	0.00065415
60-64	white	Female	0.00022726	0.00026773
60-64	White	Male	0.00033968	0.00069301
65+	black	Female	0.00087209	0.00026856
65+	black	Male	0.00099674	0.00049979
65	white	Female	0.00064624	0.00022857
65	White	Male	0.00082106	0.00030998

Limitations

Medication effects

We acknowledge several limitations associated with the use of medication data in the risk prediction models. Patient adherence to prescribed medication regimens, including issues of under- or over-dosing, can complicate the informational contribution of medication effects. Additionally, there is a distinction between curative medications, such as insulin for diabetes management and preventive medications like statins for cholesterol reduction or GLP-1 receptor agonists like ozempic for weight reduction. Both may affect cancer risk by different mechanisms. Understanding these diverse impacts on disease progression is important for accurately interpreting risk profiles. This study

does not look into potentially causative contributions of drugs to increase or decrease cancer risk. Investigating such relationships to establish clearer associations between medications and risk factors of cancer is of interest for future research, but prediction performance does not depend on mechanistic explanation of risk factors.

Comparison of the AI tool with standard clinical practice

Diagnosis trajectories can contain quasi-symptoms just prior to cancer diagnosis that are obvious risk factors and do not require ML for prediction. However, their occurrence overlaps with non-obvious risk factors, so there are no sharp boundaries. We therefore recommend using models that are trained on all or most disease states even within a few months or days prior to cancer diagnosis. This would result in the benefit of comprehensive prediction, even if some predictions of high risk based on observing quasi-symptoms (e.g., unspecified jaundice) would trivially be made by experienced clinicians. Evaluation of innovative added benefit from the AI tool should of course carefully distinguish between trivial and non-trivial contributions in a patient trajectory. In reporting prediction performance we therefore use models trained excluding the last three months of disease codes before cancer diagnosis (Figure 2). Larger data exclusion intervals just before cancer may also be useful. Direct comparison of prediction performance with that from standard clinical practice, however, is out of scope and would require a carefully designed controlled clinical trial.

Time-to-cancer diagnosis

Ideally, one would be able to predict the initiation of cancer and its growth to a detectable level. However, the AI prediction model is trained on the time point of cancer diagnosis as captured in EHRs or clinical notes, which is much later - especially for pancreatic cancer. One can perhaps improve this type of model by incorporating background knowledge of how long it typically takes for cancer to reach the stages or grades reported at first diagnosis, but this is challenging. A related question on the evaluation of prediction performance is the time interval between the time of assessment (prediction) and the time of cancer diagnosis. In the current model, we predict cancer diagnosis to occur within, e.g., 36 months, and if cancer occurs at any time point within that time frame, the prediction is scored as accurate. However, the time point of cancer diagnosis in this high-risk group in general is distributed over the 36 months and the variation in time-to-cancer after the time point of prediction is not modeled. It may be more precisely captured - if the data are sufficiently informative - by alternative formulations of the probability model. Nonetheless, the current formulation does provide reasonable sufficient information for the design of stratified surveillance problems.