scientific reports

OPEN

Identification of patients at risk for pancreatic cancer in a 3-year timeframe based on machine learning algorithms

Weicheng Zhu[1], Long Chen[1], Yindalon Aphinyanaphongs[2], Fay Kastrinos[3], Diane M. Simeone[4], Mark Pochapin[5], Cody Stender[6], Narges Razavian[2,7] & Tamas A. Gonda[5,7]

Early detection of pancreatic cancer (PC) remains challenging largely due to the low population incidence and few known risk factors. However, screening in at-risk populations and detection of early cancer has the potential to significantly alter survival. In this study, we aim to develop a predictive model to identify patients at risk for developing new-onset PC at two and a half to three year time frame. We used the Electronic Health Records (EHR) of a large medical system from 2000 to 2021 (N = 537,410). The EHR data analyzed in this work consists of patients' demographic information, diagnosis records, and lab values, which are used to identify patients who were diagnosed with pancreatic cancer and the risk factors used in the machine learning algorithm for prediction. We identified 73 risk factors of pancreatic cancer with the Phenome-wide Association Study (PheWAS) on a matched case–control cohort. Based on them, we built a large-scale machine learning algorithm based on EHR. A temporally stratified validation based on patients not included in any stage of the training of the model was performed. This model showed an AUROC at 0.742 [0.727, 0.757] which was similar in both the general population and in a subset of the population who has had prior cross-sectional imaging. The rate of diagnosis of pancreatic cancer in those in the top 1 percentile of the risk score was 6 folds higher than the general population. Our model leverages data extracted from a 6-month window of time in the electronic health record to identify patients at nearly sixfold higher than baseline risk of developing pancreatic cancer 2.5–3 years from evaluation. This approach offers an opportunity to define an enriched population entirely based on static data, where current screening may be recommended.

Keywords Pancreas, Pancreatic cancer, Machine learning, Risk factor, Predictive model

Despite the relatively low incidence of pancreatic adenocarcinoma (PC) of 13.2 per 100,000, it now ranks as the third leading cause of cancer death[1]. While advances continue to be made in therapy most patients are diagnosed with advanced and incurable disease. Early detection offers a critical path toward improved overall survival. Pancreatic cancer diagnosed at an early stage has been associated with significantly greater survival and when PC is detected in a screening program, it is associated with significantly better outcomes[2]. However, due to the low prevalence and current lack of effective non-invasive screening modalities, population-level screening has not been endorsed by the US Preventative Services Task Force. Currently, most patients who are enrolled in a screening program are identified based on family history or known germline mutation status (referred to as high-risk individuals; HRI[3]) or after an incidental diagnosis of a cystic neoplasm. However, only 10–20% of patients diagnosed with PC would have been eligible for screening based on family history

[1]Center for Data Science, New York University, New York, NY, USA. [2]Department of Population Health, New York University Grossman School of Medicine, 227 East 30th Street, 6th Floor, New York, NY 10016, USA. [3]Department of Medicine, Division of Digestive and Liver Diseases, Columbia University Irving Medical Center, New York, NY, USA. [4]Moores Cancer Center, UC San Diego Health, San Diego, CA, USA. [5]Division of Gastroenterology and Hepatology, Department of Medicine, New York University, 240 East 38th Street, 23rd Floor, New York, NY 10016, USA. [6]Department of Surgery, New York University, New York, NY, USA. [7]Narges Razavian and Tamas A. Gonda contributed equally to this work. email: Narges.Razavian@nyulangone.org; Tamas.Gonda@nyulangone.org

or germline mutations and no more than 15% of cancers arise from cystic lesions. Therefore, the majority of patients ultimately diagnosed with PC would not have been offered screening according to current practices. In addition, significant disparities exist in the recognition and testing of HRI, which further decreases the efficacy of current screening recommendations[4]. Our goal was to develop a model that can help identify individuals at risk for PC in addition to HRIs or those with known cystic lesions.

Previous studies have identified additional risk factors that are associated with PC, some of which using machine learning models such as logistic regression and random forest[29,31], and some utilizing deep learning based approaches, including deep neural network and gated recurrent unit models[28,30]. However, none of these have sufficiently high odds to help identify a population at high-risk, while deep learning based models often require substantially large training datasets without significant improvement in performance.

Risk associations were developed based on the combination of several of these risk factors and their dynamic change during a period preceding the diagnosis of cancer[23,24]. In addition, some of these risk factors were specifically applied to cohorts with existing pancreatic conditions that are associated with PDAC such as chronic pancreatitis or cystic neoplasms[25–27]. Recent studies have highlighted the opportunities offered by machine learning approaches to mining of the EHR data and the unselected general population (summarized in Supplemental Table 1). A recent transformer model-based study, trained on a Danish cohort and validated in the US Veterans Affair Health System[28], used a prediction time of 0–5 years and another US-based study built a prediction model with a more defined but relatively short time frame between evaluation and diagnosis (180–365 days)[29]. Another recent study evaluated an even shorter lead time of 3 months[30]. Although these models yield promising results, their varying time spans and exclusion criteria may not yet align perfectly with the practical needs of clinical screening timelines.

In our study, we chose a longer time frame then most prior studies between the evaluation period and the predicted diagnosis of pancreatic cancer. This 2.5–3 year time frame was chosen both based on biological data (which has suggested a 3–4 year time between progression from high grade premalignant lesions to cancer[5–7]) and also to increase the potential for clinical impact. A shorter interval may limit the ability to identify early stage lesions and a much longer interval would have been difficult given the timeframe of use of the EHR.

We used the Electronic Health Record (EHR) of our large medical system at [redacted for anonymization] that serves the [redacted for anonymization] Area and vicinity to identify patients who were diagnosed with pancreatic cancer and had a presence for three or more years in the health record preceding the event. We based our model on the presence of diagnostic codes, abnormal laboratory values and demographic variables, specifically to enhance the general applicability and transferability of this model. We also established a framework of selecting significant features from the population data using PheWAS analysis by identifying case and control cohorts and discover top relevant features that can be utilized towards predictive modeling.

Our model identifies individuals at risk of pancreatic cancer based on readily available diagnostic codes, lab values and demographic information. If implemented in any similar EHR it may help identify a broader group of individual who may benefit from screening, beyond those with family history, genetic syndromes or known cysts. The main contribution of this work lies on the establishing a predictive model using much common population data as a way of early screening of high-risk pancreatic cancer patients, instead of merely relying on family histories or incidental diagnoses of a cystic neoplasm. Compared with previous works that utilize deep learning models and EHR data for prediction, our logistic regression based approach is more data efficient and interpretable while achieving similar to better performances. In addition, our approach is more robust to the imbalanced dataset of pancreatic cancer, making it more suitable to deep learning based approaches in practice. We also established a framework of selecting significant features from the population data using PheWAS analysis by identifying case and control cohorts and discover top relevant features that can be utilized towards predictive modeling. This prediction model can bring the capability of early detection of pancreatic cancer to improve the survival of this disease.

Methods

Study design

This predictive modeling study was performed based on EHR at New York University (NYU) Langone Health. This study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guideline for cohort studies. The overview of our approach is described in Fig. 1. We first identify patients with pancreatic cancer and at least three years of records prior to their pancreatic cancer disease onset (N = 1923). We then matched these patients to a set of control patients (N = 7728) based on demographic features and based on presence in the database for at least three years. We used phenome-wide association methodology on this matched case–control set, to identify factors that are statistically significantly associated with pancreatic cancer (73 diagnosis codes and 5 lab values were selected). Our training or discovery dataset was based on a patient population who last visited NYU before or in 2015 (N = 527,027). Our performance results are then reported on a temporally heldout validation cohort, of a new set of patients who visited NYU after 2015 (N = 469,357). This temporally-stratified validation mimics real-world scenarios of training and deploying the model.

Instead of a black box approach, we first perform a statistically rigorous PheWAS feature selection step on a carefully matched PC case–control sub-cohort. We match each pancreatic cancer (PC) patient with non-PC patients based on age, sex, length of history in the EHR, and the number of ICD codes, using k-nearest neighbor algorithm with a 1:4 case-to-control ratio. We then train the model based on the selected features on the full dataset. Furthermore, we validate our model on a heldout cohort that is also temporally different compared to the training and development cohort. This further strengthens the generalizability estimate of the model, given that the practice of medicine and documentation changes, and major world events such as COVID-19 pandemic create a large distribution shift in the data distribution over time.

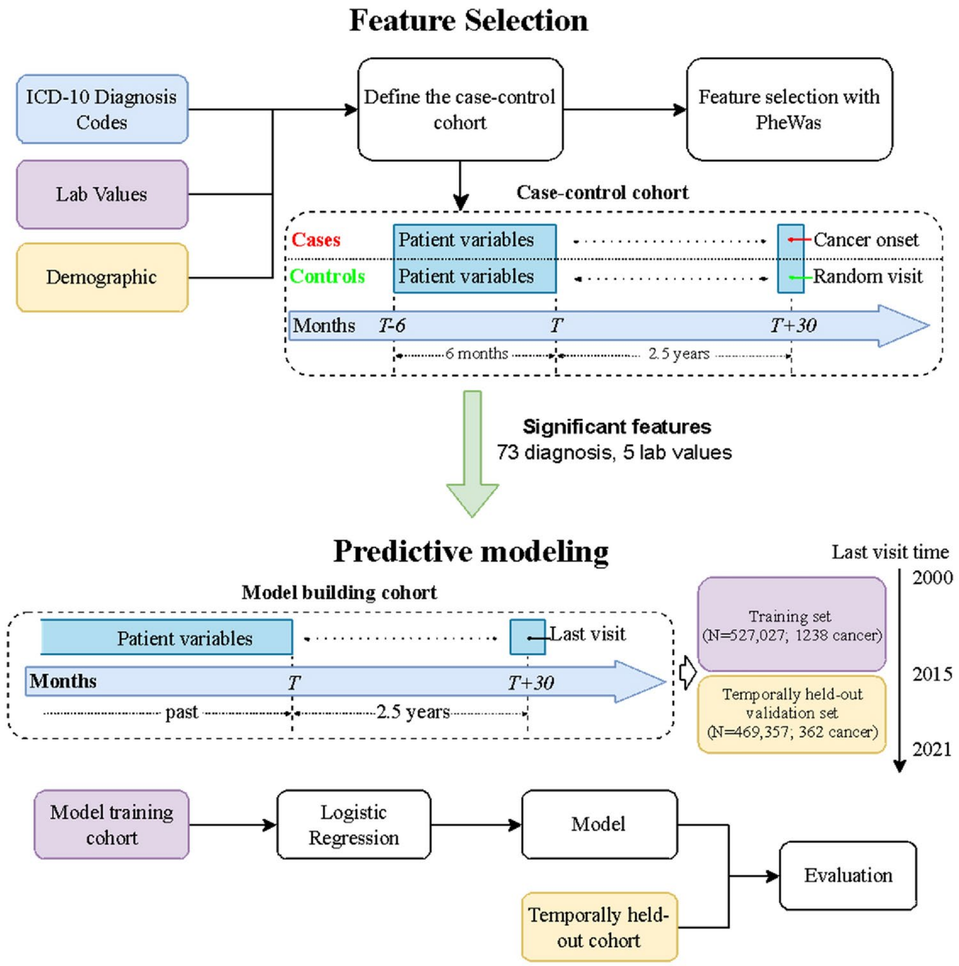

Fig. 1. The pipeline of our model and the timeline of designing case–control and model building cohort.

Overall population, pancreatic cancer definition, and definition of included variables

We identified individuals with at least 3 years of continuous presence in the EHR between January 2000 and January 2021 in the NYU Langone Health system. The EHR data analyzed in this work consists of patients' demographic information, diagnosis records, and lab values. Pancreatic adenocarcinoma diagnoses were determined through specific diagnostic codes (ICD-10 codes: C25.0, C25.1, C25.2, C25.3, C25.7, C25.8, C25.9), detailed in Supplemental Table 2. Any presence of these ICD-10 codes indicated the presence of pancreatic cancer. The onset of pancreatic cancer was marked by the encounter where these diagnostic codes first appeared in the patient's records.

Feature selection with phenome-wide association (PheWAS) analysis

In this section, we introduce feature selection with PheWAS. We first defined the inclusion criteria of positive and negative samples for pancreatic cancer. From them, we matched a case–control cohort under the control of other confounding factors. Then a PheWAS analysis was conducted on the case cohort against the control cohort to select features.

Inclusion criteria and prediction time for pancreatic cancer patients

The timeline of developing pancreatic cancer (PC) is critical in this study. The date of PC diagnosis was defined as the first date at which the PC criteria was observed in the patient records. To predict PC within a 3-year timeframe, we only included PC patients that had at least 3 years' worth of data prior to the first PC diagnosis time. We then defined a date for each patient, called *prediction time*, assuming that we made predictions on that date based on historical data prior to this date. Patients without any records prior to the prediction time were excluded.

Inclusion criteria and prediction time for pancreatic cancer-free patients

For each patient without Pancreatic Cancer, a random encounter date was selected as the "prediction time". Inclusion criteria were defined as having 3 years follow-up at NYU Langone after that "prediction time". Patients without any history prior to the prediction time were excluded.

Included variables for PheWAS analysis

Disease records from a 6-month window prior to the "prediction time", in addition to the *latest* lab and physiological data (unrestricted by time, any time prior to the "prediction time") were used to construct variables for the PheWAS stage (see Fig. 1). Previous diseases of patients prior to "prediction time" were represented as 19,304 ICD-10 codes derived from the EHR. The following 10 lab results or physiological data of patients were included, based on the clinical potential to be a risk factor of PC[25,32,33]. These lab values include A1C, Glucose, Hemoglobin, Albumin, Lipase, Amylase, AST, CA19-9, and BMI. We binarized lab values referring to the standard normal ranges (Supplemental Table 3), and bucketed BMI into underweight (BMI < 18.5), normal (BMI 18.5–25), overweight (BMI 25–30), and obese (BMI > 30).

Phenome-wide association analysis

We introduce Phenome-wide association study (PheWAS)[8,9] to conduct feature selection among thousands of diagnosis variables and lab values. PheWAS is designed to measure the association between a set of phenotypes and target outcomes, with the aim of identifying links between known phenotypic traits (including molecular, biochemical, cellular, and clinical diagnoses) and the target outcomes of interest. However, PheWAS can be influenced by various confounding variables, such as age and sex. To address this, we conducted a case–control subcohort analysis to match each pancreatic cancer (PC) patient with a non-PC patient based on age, sex, length of history in the EHR, and the number of ICD codes. We matched each included PC patient with a non-PC patient using the k-nearest neighbor algorithm, and we used a 1:4 case-to-control ratio. We used the PheWAS approach to test the association between the new onset PC and the included 19,304 diagnosis variables and 10 lab values and physiological measures. Statistically significant variables were used in the subsequent predictive modeling.

Statistical analysis

In the PheWAS analysis, we reported the odds ratios (unadjusted), the corresponding p-values and confidence intervals of the association between the condition and new onset PC in a three-year follow-up. For binary disease indicators (documented at ICD-10), the unadjusted odds ratios (UOR) and p-values of the association were computed based on the Chi-square test; for bucketed numerical values, the UORs were computed on the sub-cohort of patients who had undergone the lab test. The p-values and confidence intervals of the UORs were computed using the proportion z-test. Correction for multiple hypothesis testing was performed using Bonferroni correction, with the level of significance (*) adjusted after Bonferroni correction. Statistically significant diagnosis variables were selected based on the p-values with the threshold of 0.01 after correction for multiple hypothesis testing.

Predictive modeling

Inclusion criteria for predictive modeling with temporal cohort stratification

Different inclusion criteria of patients and included variables are introduced to simulate the real-world predictive operation. We used temporal stratification to separate the development and validation set because we can only use the past data for model development in practice. The development cohort for predictive model development included patients older than 40 years (at their last visit), with or without PC whose last record in EHR was in 2015, who matched the following inclusion criteria (see Fig. 1): For PC patients, we required at least 3 years of data prior to the first date where any record of PC was observed. For PC-free patients, any patient with at least 3 years of records were included. For PC patients, 2.5 years prior to the first PC record was selected as the prediction time. For PC-free patients, "prediction time" was selected as 3 years prior to their last visit. An independent cohort of new patients, not used during the development stage, was similarly constructed, from patients whose last visit happened between 2015 and 2022.

Included variables for predictive modeling

During the predictive modeling, variables selected as significant during the PheWAS stage were constructed for all included patients from all records prior to the "prediction time".

Model development

We built two predictive models, based on regularized logistic regression models with elastic net based regularization[10]. Note that the case–control cohorts defined in PheWAS for identifying significant features are not used for predictive modeling. Instead, the broader cohort which matched inclusion criteria (n = 996,384) is used in predictive modeling. The first model was trained on the full population that is older than 40 years old (N = 527,027). The second model was trained and validated only on patients with no known or documented pancreatic conditions or cross-sectional imaging. We trained with predictive model with regularized Logistic Regression to estimates the probability of pancreatic cancer by $P\left(y=1\right)=1/(1+e^{-(\beta_0+\sum_{i=1}^{m}\beta_i x_i)})$, where $x_1,\cdots,x_m$ are feature variables from EHR, β_0 is the intercept, and $\beta_1,\cdots,\beta_m$ are the coefficients. We tuned the regularization penalty by elastic net (the combination of L1 and L2 regularization). To mitigate the class imbalance shown in Table 1, we applied balanced class weight in logistic regression model. During the development stage, we split the development data set into a training cohort and held-out validation cohort and performed fivefold cross-validation on the training cohort with a 20% validation set. We then reported results on the temporally separate (new patient) held-out cohort described in the previous section.

Patient characteristics	Eligible overall cohort (% of n = 537,410)	Pancreatic cancer free (% of n = 535,478)	New-onset pancreatic cancer (% of n = 1932)	Matched controls (% of n = 7728)
Age, mean (sd)	49.67 (20.10)	49.59 (20.09)	70.25(12.81)	70.01 (12.57)
Sex, n (%)				
Female	335,960 (62.51%)	334,875 (62.54%)	1085 (56.16%)	4358 (56.39%)
Male	201,377 (37.47%)	200,530 (37.45%)	847 (43.84%)	3370 (43.61%)
Unknown	73 (0.01%)	73 (0.01%)	0 (0.00%)	0 (0.00%)
Number of ICD-10 Codes, mean (sd)	3.38 (6.79)	3.36 (6.77)	6.26 (10.57)	6.01 (10.32)
Length of medical history (days), mean (sd)	969.57 (1279.87)	968.3 (1276.89)	1298.43 (1874.38)	1298.30 (1873.54)

Table 1. Demographic and overall characteristics of all eligible individuals, pancreatic cancer patients, all cancer-free individuals and matched cohorts, used during the PheWAS analysis.

Statistical analysis

In predictive modeling, we evaluated the model performance on the held-out validation set with Areas Under the Receiver Operating Characteristic Curve (AUROC). This binary classification metrics is used in similar studies[28,30]. Additional metrics, including Positive Predictive Value (PPV), Negative Predictive Value (NPV), sensitivity, and specificity were computed at different probability thresholds on the predictions, to facilitate the identification of the optimal threshold for screening. These classification metrics are also used in previous work[30,31]. We also reported the fully adjusted odds ratios (OR) and p-values of the variables associated with PC onset based on the logistic regression analysis, as similarly reported by a related work[34].

Sensitivity analysis

The accuracy of our data relies heavily on the accuracy of the use of diagnosis codes, which is a manual chart review of randomly selected cases and controls we found to be approximately 80% accurate. Therefore, to examine the stability of our model on data noise, we conducted a sensitivity analysis by retraining the model based on simulated data where we randomly flipped the labels of 20% cancer samples and the same number of cancer-free samples. We reported the mean and standard deviation 10 times of such experiments to analyze incorrect EHR labels' impact.

Results

Identification of cohorts for PheWAS and predictive modeling

537,410 patients were eligible to be included, where 1932 patients with PC were matched to 7728 PC-free patients based on demographics and length of presence in the database. The characteristics of the case and control cohort are summarized in Table 1.

PheWAS analysis results

After performing PheWAS analysis on the matched case–control, based on 19,304 disease records and 10 laboratory tests or physiological signs (BMI, A1C, Hemoglobin, fasting glucose, CA19-9, Lipase, Amylase, AST, ALT, Albumin), we identified 73 diagnosis codes and 5 lab values that showed significant association with PC onset. Table 2 shows the top 20 variables and S4 Table shows the comprehensive list. Figure 2 provides visualizations of diagnosis and lab value features within the top 20 variables. Most of the features have been associated with pancreatic cancer in the past, however, we included all of these features in predictive modeling. We noted that several of the highly significant features could only have been recognized in patients with either a past medical history of pancreatic diseases or cross-sectional imaging (e.g. pancreatic cyst). As this represented a minority but a potentially distinct and higher-risk population, we developed a model for all eligible individuals (Model 1) and a separate model for those without pancreatic disease or cross-sectional imaging in the time period of feature selection (Model 2).

Performance of the prediction model

For the first model, we trained and evaluated the regularized logistic regression in Model 1 between 2000 and 2021 (N = 996,384). As described in the Methods section, this cohort was temporally stratified into training (patients before 2015) and held-out validation (temporally distinct, new patients after 2015) set. The characteristics of the training set and the held-out validation set are summarized in Table 3. We observe that the dataset is imbalanced and is heavily skew towards negatives for both training cohort and heldout-validation cohort.

Table 4 and Supplemental Table 5 show the coefficients of logistic regression in Model 1 in the temporarily held-out validation cohort. Figure 3 shows the Receiver Operating Characteristic (ROC) curve of Model 1 (the performance of the second model with exclusion is reported in Supplemental Fig. 1). The Areas Under the ROC curve (AUROC) was 0.742 [0.727, 0.757] and this was not significantly different than the performance in Model 2 (Supplemental Fig. 1). As the prevalence of pancreatic cancer is low, the Positive Predictive Value (PPV) is crucial for selecting a reasonable threshold for screening. The PPV of the model at different predicted risk levels is shown in Fig. 4 and Supplemental Fig. 1. Screening the top-1 and top-5 percent of the patients with the highest risks would achieve around 6 and 4 times higher PPV, respectively, than the general patient population in the EHR. We also compared this to the PPV in individuals with Type II Diabetes (T2D; defined as either having any

Diagnosis features	Control (%)	Case (%)	P-value	Unadjusted odds ratio [95% confidence interval]
Disease of pancreas, unspecified (ICD10 K86.9)	2 (0.02%)	11 (0.53%)	5.61E−09 **	22.11 [4.90, 99.83]
Malignant neoplasm of transverse colon (ICD10 C18.4)	1 (0.01%)	5 (0.24%)	1.05E−04	20.05 [2.34, 171.68]
Pseudocyst of pancreas (ICD10 K86.3)	6 (0.07%)	16 (0.77%)	6.04E−10 **	10.74 [4.20, 27.48]
Hypertrophy of breast (ICD10 N62)	4 (0.05%)	9 (0.43%)	8.98E−06 **	9.03 [2.78, 29.37]
Neoplasm of unspecified behavior of digestive system (ICD10 D49.0)	8 (0.10%)	16 (0.77%)	1.05E−08 **	8.05 [3.44, 18.84]
Other specified diseases of pancreas (ICD10 K86.89)	11 (0.13%)	15 (0.72%)	1.50E−06 **	5.49 [2.52, 11.96]
Genetic susceptibility to other malignant neoplasm (ICD10 Z15.09)	10 (0.12%)	13 (0.62%)	1.17E−05 **	5.23 [2.29, 11.94]
Genetic susceptibility to malignant neoplasm of breast (ICD10 Z15.01)	9 (0.11%)	11 (0.53%)	8.97E−05	4.91 [2.03, 11.86]
Malignant neoplasm of unspecified ovary (ICD10 C56.9/57.00)	6 (0.07%)	7 (0.33%)	2.27E−03	4.68 [1.57, 13.94]
Acute pancreatitis without necrosis or infection, unspecified (ICD10 K85.90)	7 (0.08%)	7 (0.33%)	4.98E−03	4.01 [1.41, 11.45]
Family history of malignant neoplasm, unspecified (ICD10 Z80.9)	18 (0.22%)	15 (0.72%)	2.51E−04	3.35 [1.69, 6.66]
Primary osteoarthritis, unspecified hand (ICD10 M19.049)	10 (0.12%)	8 (0.38%)	9.46E−03	3.21 [1.26, 8.14]
Malignant neoplasm of unspecified kidney, except renal pelvis (ICD10 C64.9)	14 (0.17%)	10 (0.48%)	7.89E−03	2.87 [1.27, 6.46]
Cerebral infarction due to unspecified occlusion or stenosis of unspecified cerebral artery (ICD10 I63.50)	24 (0.29%)	15 (0.72%)	3.88E−03	2.51 [1.31, 4.79]
Malignant neoplasm of unspecified site of unspecified female breast (ICD10 C50.919)	101 (1.21%)	43 (2.06%)	2.89E−03	1.72 [1.20, 2.46]
Lab value features	**Control (%)**	**Case (%)**	**P-value**	**Unadjusted odds ratio [95% confidence interval]**
Lipase > 234.0	29 (12.61%)	20 (25.64%)	6.54E−03	2.39 [1.26, 4.53]
Glucose > 126.0	208 (29.80%)	106 (46.70%)	3.01E−06**	2.06 [1.52, 2.81]
Obesity (BMI > 30)	1136 (28.32%)	291 (38.70%)	1.17E−08**	1.60 [1.36, 1.88]
Type II diabetes (E11.9 + A1C)	459 (5.49%)	164 (7.84%)	4.71E−05*	1.47 [1.22, 1.76]
Overweight (BMI 24 ~ 30)	1294 (32.25%)	280 (37.23%)	7.70E−03	1.25 [1.06, 1.47]

Table 2. The top 20 variables that are associated with the development of pancreatic cancer. The unadjusted odds ratios on the case–control cohort and the unadjusted odds ratios of the two models on the overall cohort are shown. *Note that the statistically significant features after Bonferroni correction are marked (*** means corrected p-value < 0.001; ** means corrected p-value < 0.05; * means corrected p-value < 0.1).

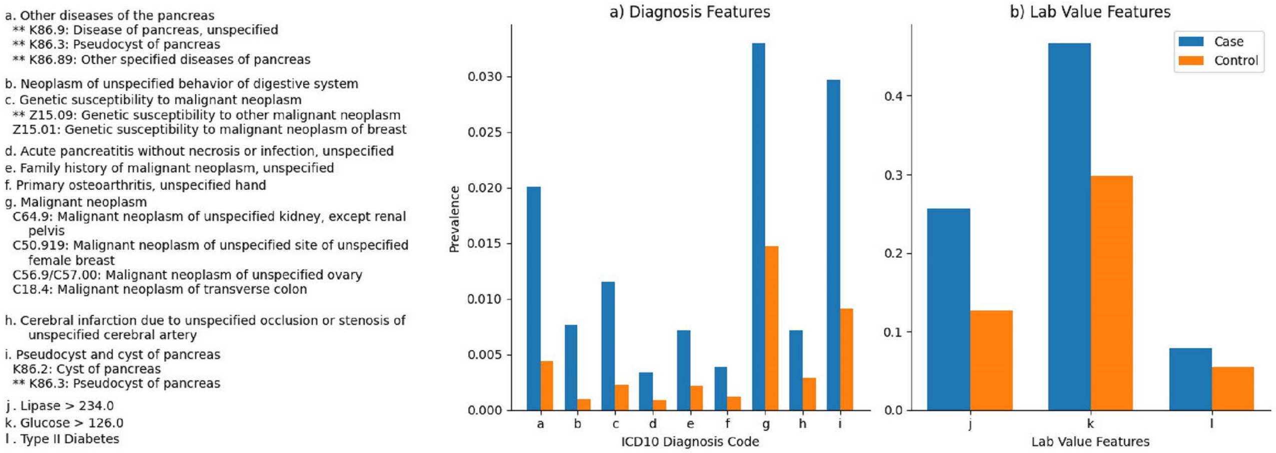

Fig. 2. The prevalence of the top (**a**) diagnosis features, and (**b**) lab features that are associated with the development of pancreatic cancer, compared between control and case groups. *The statistically significant features after Bonferroni correction are marked.

Patient characteristics	Overall cohort (% of n = 996,384)	Training cohort (% of n = 527,027)	Heldout-validation cohort (% of n = 469,357)
Age, mean (sd)	61.54 (12.49)	62.11 (12.76)	60.91 (12.16)
Sex, n (%)			
Female	612,953 (61.51%)	324,481 (61.56%)	288,472 (61.46%)
Male	383,282 (38.46%)	202,436 (38.41%)	180,846 (38.53%)
Unknown	149 (0.03%)	110 (0.03%)	39 (0.02%)
Pancreatic cancer, n (%)			
Positive	1600 (0.16%)	1238 (0.23%)	362 (0.07%)
Negative	994,784 (99.84%)	525,789 (99.77%)	468,995 (99.93%)

Table 3. The characteristic table of the overall cohort for predictive modeling, as well as training and test set separately.

Description	P-value	Adjusted odds ratio [95% confidence interval]
Intercept	3.495E−01	0.811 [0.417, 1.206]
Other specified diseases of pancreas (ICD K86.89)	0.00E + 00***	6.429 [4.734, 8.124]
Neoplasm of unspecified behavior of digestive system (ICD D49.0)	6.99E−08***	4.106 [2.243, 5.970]
Acute pancreatitis without necrosis or infection, unspecified (ICD K85.90)	8.44E−14***	4.019 [1.568, 6.470]
Malignant neoplasm of unspecified kidney, except renal pelvis (ICD C64.9)	2.00E−15***	3.947 [2.634, 5.259]
Primary osteoarthritis, unspecified hand (ICD M19.049)	1.39E−10***	2.868 [1.984, 3.752]
Family history of malignant neoplasm, unspecified (ICD Z80.9)	2.22E−16***	2.768 [2.136, 3.399]
Genetic susceptibility to other malignant neoplasm (ICD Z15.09)	0.00E + 00***	2.159 [0.597, 3.720]
Cyst of pancreas (ICD K86.2)	2.44E−15***	1.924 [1.127, 2.720]
Glucose > 126.0	1.45E−04*	1.910 [1.223, 2.596]
Malignant neoplasm of unspecified site of unspecified female breast (ICD C50.919)	0.00E + 00***	1.880 [1.660, 2.100]
Type 2 diabetes mellitus without complications (ICD E11.9)	0.00E + 00***	1.876 [1.749, 2.002]
Personal history of poliomyelitis (ICD Z86.12)	5.34E−02	1.797 [0.996, 2.597]
Cerebral infarction due to unspecified occlusion or stenosis of unspecified cerebral artery (ICD I63.50)	5.50E−03	1.780 [1.119, 2.442]
Genetic susceptibility to malignant neoplasm of breast (ICD Z15.01)	1.40E−05**	1.758 [0.714, 2.801]
Pseudocyst of pancreas (ICD K86.3)	4.41E−02	1.511 [0.871, 2.151]
Malignant neoplasm of unspecified fallopian tube (ICD C57.00)	4.87E−08***	1.427 [1.241, 1.613]
Disease of pancreas, unspecified (ICD K86.9)	7.97E−04*	1.415 [1.135, 1.696]
Obesity (BMI > 30)	9.39E−02	1.238 [0.945, 1.531]
Lipase > 234.0	5.28E−01	1.111 [0.790, 1.431]
Overweight (BMI 24 ~ 30)	6.10E−01	1.087 [0.779, 1.395]
A1C > 6.5	4.76E−01	1.058 [0.905, 1.211]

Table 4. The intercept, coefficients, and p-values of the positive features in the logistic regression model. *Note that the statistically significant features after Bonferroni correction are marked (***means corrected p-value < 0.001; ** means corrected p-value < 0.05; *means corrected p-value < 0.1).

A1c > 6.5, or an E11.9 diagnosis code) based on published risk assessment and found that this model would have around 3 and 2 times higher PPV than T2D among the top-1 and 5 percent of the patients with highest risk, respectively (Table 5).

Sensitivity analysis

We conducted the sensitivity analysis with 10 trials that repeated training and validation with simulated datasets by randomly flipping outcomes. The AUC of simulated experiments was 0.718 [0.713, 0.722], which is not significantly different from the performance of the original dataset. The sensitivity analysis demonstrated that the model is robust under some noises in EHR.

Discussion

Although screening and early detection of pancreatic cancer hold a realistic promise to impact the survival of this disease, most patients diagnosed with cancer would not have been recommended to undergo screening. An important reason for this is the absence of strong risk factors that have precluded the identification of populations at risk[11,12]. Although high-risk populations have been identified based on familial or genetic risk, it is estimated that only a minority of patients diagnosed with pancreatic cancer have a currently identifiable predisposition[13,14]. Due to the absence of an effective screening tool or modality and the low prevalence (which

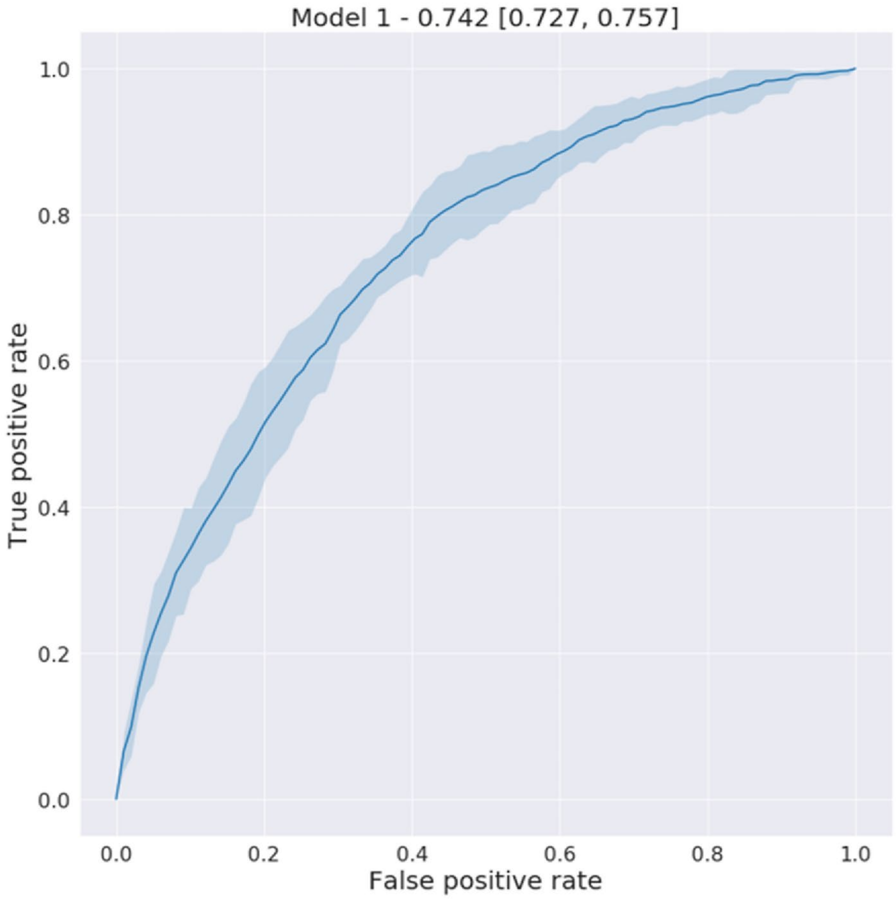

Fig. 3. The receiver operating characteristics curve (AUROC 0.742 [0.727, 0.757]) on the held-out validation set to predict new-onset pancreatic cancer in a 3-year follow-up.

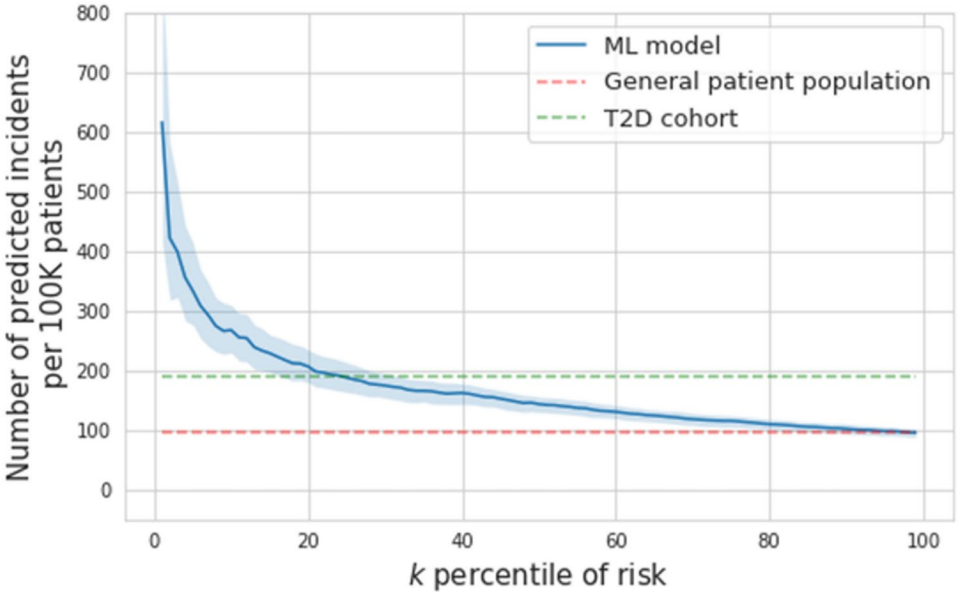

Fig. 4. The positive predictive value (scaled to depict rate of identification in 100,000 patients), of risk of new-onset pancreatic cancer in 3-year follow-up, at different predicted risk levels.

Risk percentiles	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
1%	6.78 [6.13, 7.43]	99.01 [99.01, 99.01]	0.65 [0.58, 0.72]	99.93 [99.92, 99.93]
5%	18.36 [16.82, 19.89]	95.01 [95.01, 95.01]	0.34 [0.31, 0.37]	99.92 [99.92, 99.92]
10%	29.14 [27.70, 30.58]	90.02 [90.02, 90.02]	0.28 [0.26, 0.29]	99.93 [99.92, 99.93]
20%	42.71 [40.66, 44.76]	80.02 [80.02, 80.03]	0.19 [0.19, 0.20]	99.93 [99.92, 99.94]

Table 5. Details on sensitivity, specificity, positive predictive values (NPV), and negative predictive values (NPV) at different risk percentiles based on model 1.

is likely associated with many false positive results), pancreatic cancer screening in the general population is not recommended[15].

Despite the rapidly progressing and aggressive nature of pancreatic adenocarcinoma, there are multiple lines of evidence showing the time frame of progression from advanced precursors or early-stage cancer to advanced cancer around 1.5–3 years[5,7]. In addition, several biomarkers (i.e. CA19-9, glycemic indices, and others) show a significant change 12–36 months prior to the ultimate diagnosis of cancer[16–19]. Perhaps equally important is the observation that survival following the diagnosis and treatment of these late precursors or early cancers is greater than would be expected from lead time bias alone. We therefore hypothesized that recognizing patients at risk for pancreatic cancer approximately three years prior to a clinical diagnosis would be a biologically reasonable time frame that may allow diagnosis of high-grade precursors or early cancers and ultimately impact survival.

We leveraged the large electronic health record of our institution to identify a model that can select patients at risk for pancreatic cancer based on diagnostic variables, laboratory results, and demographic information. We evaluated the presence of these variables in a six-month time period 30 months prior to the cancer diagnosis and matched this cohort at a 1:4 ratio with a control group.

The strength of our approach is that it did not rely on any known association with conditions or diseases although we did limit the inclusion of laboratory tests to those that would have potential clinical relevance to the diagnosis of pancreatic diseases. In our variable selection, we ultimately identified 78 variables; 73 of which were diagnostic codes. The majority of these diagnostic codes had a known or plausible association with an individual risk factor for pancreatic cancer although only a few of them alone could potentially prompt screening for pancreatic cancer. One set of such variables was the diagnosis of certain pancreatic conditions such as pancreatic cysts and pancreatitis. We therefore developed a second model where we excluded patients with a known history of pancreatic disease or a presumed evaluation that would've either identified pancreatic disease or confirmed the presence of a normal pancreas. However, ultimately this represented a small cohort and it had no major impact on model performance.

The identification of other malignancies and pancreatitis further supports the notion that certain genetic syndromes and pancreatic inflammation may predispose to pancreatic cancer. However, it is important to note that none of the variables identified alone would prompt consideration of PC screening in current practice guidelines.

The area under the curve of our model was 0.742. Our precision-recall curve shows that using our model, the top 1% risk percentile is associated with a nearly sixfold increased risk. As the accuracy of the screening instruments is limited (especially lacking the ability to detect precursor lesions), it is somewhat difficult to compare these numbers to other cancer screening. However, the rate of cancer diagnosis in colorectal cancer screening studies per patient in the general population ranges from 1:154–208 depending on the screening method[20,21] and is 1:250 in lung cancer screening by CT scan[22]. This would suggest that the highest risk individuals in the model would constitute a sufficiently enriched cohort to offer screening.

Our study has several limitations. The setting and health record we studied is that of an academic institution and its affiliates and our data is entirely limited to patients who continue to get their care in this system. In any health care setting without a comprehensive EHR, implementation or evaluation of the model would be limited, although the reliance on static diagnostic codes and lab values may permit use of our model. Although ours is a diverse urban population, an analysis on a much greater scale or external validation in a different demographic will be essential. Additionally, our variables are static and changes in metabolic parameters such as weight or glycemic control are not considered. Although this makes the model easier to implement, future approaches will likely need to consider the impact of change. Also, although using diagnostic codes as labels is efficient on modeling large-scaled patient cohort, missingness and mistakes sometimes occur in EHR. Besides simulated sensitivity analysis, more controlled trials should be conducted in the future to validate it impact of the model. In addition, different time frames maybe considered in the future. Current evidence would not support either a longer or shorter time interval but with expansion of molecular and epidemiologic profiling of progression in PC a more exact time frame maybe possible to define.

Conclusion

In conclusion, our study shows that it is feasible to identify a patient population at sixfold higher risk for developing pancreatic cancer than the average population using static variables in a machine learning model. This approach offers an easily implementable approach toward finding an enriched population at sufficiently high baseline risk to warrant a discussion about screening. We believe that an EHR-based approach to identifying individuals coupled with a notification system [SD1] has the chance to reduce disparities in screening and recognize novel high-risk populations. It is however critical to recognize that the current screening tests for pancreatic cancer have significant weakness, which will undoubtedly affect the benefit of finding a high-risk population. Yields of screening in current high-risk cohorts have been relatively modest and this may be due to selection based on only a few variables. Our future efforts will focus on measuring the impact of the implementation of this machine learning based model and this will hopefully occur parallel to continued improvements in the use of screening modalities.

Data availability

The datasets generated and/or analyzed during the current study are not publicly available due to specific institutional requirements governing privacy protection. The EHR data used for the pretraining, fine-tuning, validation and test sets were collected from the NYU Langone Health System EHR maintained by the NYULH Datacore team. This data consists of the production medical records of NYU Langone and cannot be made publicly available. Researchers may obtain a limited de-identified dataset (or a test subset) from NYU Langone Health System by reasonable request after acquiring institutional agreement, subject to local and national ethical approvals. To request a limited de-identified dataset please reach out to corresponding author Dr. Tamas Gonda via email at Tamas.Gonda@nyulangone.org.

Received: 29 August 2024; Accepted: 6 February 2025
Published online: 05 April 2025

References

1. Rahib, L. et al. Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. *Cancer Res.* **74**, 2913–2921 (2014).
2. Canto, M. I. et al. Surgical outcomes after pancreatic resection of screening-detected lesions in individuals at high risk for developing pancreatic cancer. *J. Gastrointest. Surg.* **24**, 1101–1110 (2020).
3. Goggins, M. et al. Management of patients with increased risk for familial pancreatic cancer: updated recommendations from the International Cancer of the Pancreas Screening (CAPS) Consortium. *Gut.* **69**, 7–17 (2020).
4. Gonda, T. A., Everett, J. N., Wallace, M., Simeone, D. M., PRECEDE Consortium. Recommendations for a more organized and effective approach to the early detection of pancreatic cancer from the PRECEDE (Pancreatic Cancer Early Detection) consortium. *Gastroenterology.* **161**, 1751–1757 (2021).
5. Yachida, S. et al. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature.* **467**, 1114–1117 (2010).
6. Yu, J., Blackford, A. L., Dal Molin, M., Wolfgang, C. L. & Goggins, M. Time to progression of pancreatic ductal adenocarcinoma from low-to-high tumour stages. *Gut.* **64**, 1783–1789 (2015).
7. Noë, M. et al. Genomic characterization of malignant progression in neoplastic pancreatic cysts. *Nat. Commun.* **11**, 4085 (2020).
8. Denny, J. C. et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics.* **26**, 1205–1210 (2010).
9. Pendergrass, S. A. et al. The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. *Genet. Epidemiol.* **35**, 410–422 (2011).
10. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 301–320 (2005).
11. Klein, A. P. Pancreatic cancer epidemiology: understanding the role of lifestyle and inherited risk factors. *Nat. Rev. Gastroenterol. Hepatol.* **18**, 493–502 (2021).
12. Midha, S., Chawla, S. & Garg, P. K. Modifiable and non-modifiable risk factors for pancreatic cancer: A review. *Cancer Lett.* **381**, 269–277 (2016).
13. Hu, C. et al. Association between inherited germline mutations in cancer predisposition genes and risk of pancreatic cancer. *JAMA.* **319**, 2401–2409 (2018).
14. Chaffee, K. G. et al. Prevalence of germ-line mutations in cancer genes among pancreatic cancer patients with a positive family history. *Genet. Med.* **20**, 119–127 (2018).
15. US Preventive Services Task Force et al. Screening for pancreatic cancer: US preventive services task force reaffirmation recommendation statement. *JAMA.* **322**, 438–444 (2019).
16. Peters, M. L. B. et al. Progression to pancreatic ductal adenocarcinoma from pancreatic intraepithelial neoplasia: Results of a simulation model. *Pancreatology.* **18**, 928–934 (2018).
17. Shah, N. D. et al. Alcohol-related liver disease is rarely detected at early stages compared with liver diseases of other etiologies worldwide. *Clin. Gastroenterol. Hepatol.* **17**, 2320-2329.e12 (2019).
18. Sharma, A. et al. Model to determine risk of pancreatic cancer in patients with new-onset diabetes. *Gastroenterology.* **155**, 730-739.e3 (2018).
19. Fahrmann, J. F. et al. Lead-time trajectory of CA19-9 as an anchor marker for pancreatic cancer early detection. *Gastroenterology.* **160**, 1373-1383.e6 (2021).
20. Imperiale, T. F. et al. Multitarget stool DNA testing for colorectal-cancer screening. *N. Engl. J. Med.* **370**, 1287–1297 (2014).
21. Pinsky, P. F. & Schoen, R. E. Colorectal cancer incidence by age among patients undergoing surveillance colonoscopy. *JAMA Intern. Med.* **175**, 858–860 (2015).
22. Hoffman, R. M., Atallah, R. P., Struble, R. D. & Badgett, R. G. Lung cancer screening with low-dose CT: a meta-analysis. *J. Gen. Intern. Med.* **35**, 3015–3025 (2020).
23. Sharma, A. et al. Model to determine risk of pancreatic cancer in patients with new-onset diabetes. *Gastroenterology.* **155**(3), 730-739.e3. https://doi.org/10.1053/j.gastro.2018.05.023 (2018).
24. Yuan, C. et al. Diabetes, weight change, and pancreatic cancer risk. *JAMA Oncol.* **6**(10), e202948. https://doi.org/10.1001/jamaoncol.2020.2948 (2020).
25. Klein, A. P. Pancreatic cancer epidemiology: understanding the role of lifestyle and inherited risk factors. *Nat. Rev. Gastroenterol. Hepatol.* **18**(7), 493–502. https://doi.org/10.1038/s41575-021-00457-x (2021).
26. Kirkegård, J., Mortensen, F. V. & Cronin-Fenton, D. Chronic pancreatitis and pancreatic cancer risk: a systematic review and meta-analysis. *Am. J. Gastroenterol.* **112**(9), 1366–1372. https://doi.org/10.1038/ajg.2017.218 (2017).
27. Midha, S., Chawla, S. & Garg, P. K. Modifiable and non-modifiable risk factors for pancreatic cancer: A review. *Cancer Lett.* **381**(1), 269–277. https://doi.org/10.1016/j.canlet.2016.07.022 (2016).
28. Placido, D. et al. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. *Nat. Med.* **29**(5), 1113–1122. https://doi.org/10.1038/s41591-023-02332-5 (2023).
29. Appelbaum, L. 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.* **143**, 19–30. https://doi.org/10.1016/j.ejca.2020.10.019 (2021).
30. Li, X. et al. A deep-learning based prediction of pancreatic adenocarcinoma with electronic health records from the state of Maine. *Int. J. Med. Health Sci.* **14**, 358–365 (2020).

31. Malhotra, A., Rachet, B., Bonaventure, A., Pereira, S. P. & Woods, L. M. Can we screen for pancreatic cancer? Identifying a sub-population of patients at high risk of subsequent diagnosis using machine learning techniques applied to primary care data. *PLoS One.* **16**(6), e0251876. https://doi.org/10.1371/journal.pone.0251876 (2021).
32. Cai, J. et al. Advances in the epidemiology of pancreatic cancer: Trends, risk factors, screening, and prognosis. *Cancer Lett.* **1**(520), 1–11. https://doi.org/10.1016/j.canlet.2021.06.027 (2021).
33. Arslan, A. A. et al. Anthropometric measures, body mass index, and pancreatic cancer: a pooled analysis from the Pancreatic Cancer Cohort Consortium (PanScan). *Arch. Intern. Med.* **170**(9), 791–802. https://doi.org/10.1001/archinternmed.2010.63 (2010).
34. Klatte, D. C. F. et al. Surveillance for pancreatic cancer in high-risk individuals leads to improved outcomes: a propensity score-matched analysis. *Gastroenterology.* **164**(7), 1223–12314. https://doi.org/10.1053/j.gastro.2023.02.032 (2023).

Author contributions

W.Z., Y.A., N.R. and T.A.G. conceptualized and defined the study methodology. The investigation was performed by W.Z., N.R. and T.A.G. Resources provided by Y.A. and N.R. Data curation was performed by W.Z. Formal analysis was performed by W.Z. and N.R. The results were also critically reviewed also by D.S., F.K., C.S. and M.P. Writing of the original manuscript draft was conducted by W.Z., Y.A., N.R., T.A.G. Review and editing of the manuscript was performed by all.

Declarations

Competing interests

The authors declare no competing interests.

Ethics

IRB exempt.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-89607-8.

Correspondence and requests for materials should be addressed to N.R. or T.A.G.