Pancreatic cancer risk predicted from disease trajectories using deep learning

Davide Placido[1,‡,*], Bo Yuan[2,3,4,‡,*], Jessica X. Hjaltelin[1,‡,*], Amalie D. Haue[1,5], Chen Yuan[2,3], Jihye Kim[6], Renato Umeton[3], Gregory Antell[3], Alexander Chowdhury[3], Alexandra Franz[2,3,4], Lauren Brais[3], Elizabeth Andrews[3], Aviv Regev[4,7], Peter Kraft[6], Brian M. Wolpin[2,3], Michael Rosenthal[2,3,8], Søren Brunak[1,6,#,*], Chris Sander[2,3,4,#,*]

[1] Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark
[2] Harvard Medical School, Boston, USA
[3] Dana-Farber Cancer Institute, Boston, USA
[4] Broad Institute of MIT and Harvard, Boston, USA
[5] Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark.
[6] Harvard T.H. Chan School of Public Health, Boston, USA
[7] Current address: Genentech, 1 DNA Way, South San Francisco, CA, USA
[8] Brigham and Women's Hospital, Boston, USA
‡ joint first authors
# joint supervising authors

*Correspondence: email to cancer.risk.prediction@gmail.com reaches the principal authors: D.P., B.Y., J.X.H., S.B., C.S.

Abstract

Pancreatic cancer is an aggressive disease that typically presents late with poor patient outcomes. There is a pronounced medical need for early detection of pancreatic cancer, which can be facilitated by identifying high-risk populations. Here we apply artificial intelligence (AI) methods to a large corpus of more than 6 million patient records spanning 40 years with 24,000 pancreatic cancer cases in the Danish National Patient Registry. In contrast to existing methods that do not use temporal information, we explicitly train machine learning models on the time sequence of diseases in patient clinical histories. In addition, the models predict the risk of cancer occurrence in time intervals of 3 to 60 months duration after risk assessment. For cancer occurrence within 12 months, the performance of the best model trained on full trajectories (AUROC=0.91) substantially exceeds that of a model without time information (AUROC=0.81). For the best model, lower performance (AUROC=0.86) results when disease events within a 3 month window before cancer diagnosis are excluded from training, reflecting the decreasing information value of earlier disease events. These results raise the state-of-the-art level of performance of cancer risk prediction on real-world data sets and provide support for the design of real-world population-wide clinical screening trials, in which high risk patients are assigned to serial imaging and measurement of blood-based markers to facilitate earlier cancer detection. AI on real-world clinical records has the potential to shift focus from treatment of late- to early-stage cancer, benefiting patients by improving lifespan and quality of life.

Introduction

Clinical need for early detection

Pancreatic cancer is a leading cause of cancer-related deaths worldwide with increasing incidence (Rahib et al. 2014). Approximately 80% of pancreatic cancer patients are diagnosed at a late stage, when long-term survival is extremely uncommon. However, patients who present with early-stage disease can be cured by a combination of surgery, chemotherapy and radiotherapy. Indeed, more than 80% of patients with stage IA pancreatic ductal adenocarcinoma (PDAC) have 5-year overall survival [National Cancer Institute, USA, (Blackford et al. 2020)]. Thus, earlier detection of pancreatic cancer has great potential to prolong patient survival and reduce overall mortality from this difficult malignancy.

Known risk factors of limited use

The incidence rate of pancreatic cancer is substantially less than other high mortality cancers, such as lung, breast and colorectal cancer. Thus, age-based population screening is difficult due to poor positive predictive values for potential screening tests and large numbers of futile evaluations for patients with false-positive results. Moreover, very few high-penetrance risk factors are known for pancreatic cancer. Cancer risk has been assessed for many years based on family history, behavioral and clinical risk factors and, more recently, genetic predisposition using genome-wide association studies (GWAS), including for pancreatic cancer (Amundadottir et al. 2009; Petersen et al. 2010; Li et al. 2012; Wolpin et al. 2014; Klein et al. 2018). Currently, some patients with familial risk due to family history or inherited genetic mutation or cystic lesions of the pancreas undergo serial pancreas-directed imaging to detect early pancreatic cancers, but these patients account for <20% of those who develop pancreatic cancer. To address the challenge of early detection of pancreatic cancer in the general population (Pereira et al. 2020; Singhi et al. 2019), we aim to predict the risk of pancreatic cancer from real-world longitudinal clinical records and to facilitate the design of screening trials for early detection. Development of realistic risk prediction methods requires access to high-quality clinical records and a choice of appropriate machine learning methods, in particular deep learning techniques that work on large and noisy sequential datasets (Dietterich 2002; LeCun, Bengio, and Hinton 2015).

Earlier clinical ML work

We build on earlier work in the field of risk assessment based on clinical data and disease trajectories using machine learning technology (Nielsen et al. 2019; Thorsen-Meyer et al. 2020). AI methods have been applied to a number of clinical decision support problems (Shickel et al. 2018), such as choosing optimal time intervals for actions in intensive care

units (Hyland et al. 2020), assessing cancer risk from images (Esteva et al. 2017; Yala et al. 2019; Yamada et al. 2019) or predicting the risk of potentially catastrophic disease progression, such as in kidney injury (Tomašev et al. 2019). Building clinically applicable prediction tools for pancreatic cancer screening is challenging, in part due to the low incidence in the general population and the consequent difficulty of achieving a low false positive rate (high specificity).

Earlier ML work on PDAC risk

For risk assessment of pancreatic cancer, recently machine learning predictive models using patient records have been built using health interview survey data (Muhammad et al. 2019), general practitioners' health records controlled against patients with other cancer types (Malhotra et al. 2021), real-world hospital system data (Appelbaum, Cambronero, et al. 2021), and from an EHR database provided by TriNetX, LLC.
(Chen et al. 2021; Appelbaum, Berg, et al. 2021). While demonstrating the information value of health records for cancer risk, these previous studies used only the occurrence of disease states for a patient in limited time intervals - not the time sequence of disease states over up to 40 years - in analogy to the ‘bag-of-words’ models in natural language processing that ignore the actual sequence of words.

Advance here - better data and better ML

Here we apply more advanced machine learning (ML) technology by focusing on the time sequence of clinical events and by predicting the risk of cancer occurrence over a multi-year time interval. This investigation was carried out using the Danish National Patient Registry (DNPR) and data which covers 40 years (1977 to 2018) of clinical records for 8.6 million persons, of which about 40,000 had a diagnosis of pancreatic cancer (Schmidt et al. 2015; Siggaard et al. 2020). To maximize predictive information extraction from these records we tested a range of statistical and ML methods. These methods range from regression methods and machine learning without time dependence to time series methods such as Gated Recurrent Units (GRU) and Transformer, adapting AI methods that have been very successful, e.g., in natural language processing and analysis of other time series data (Cho et al. 2014; Tealab 2018; Vaswani et al. 2017).

Advance - prediction time intervals

The likely action resulting from a personalized positive prediction ideally should take into account the probability of cancer occurring in a shorter or longer time frame. For this reason, we designed the prediction method to predict not only whether or not cancer is ever likely to occur but also to provide risk assessment in incremental time intervals following the assessment. We also analyzed which diagnoses from the past are most informative of cancer risk. Finally, we propose a practical scenario for broadly-based screening trials, taking into consideration typically available real-world data, the accuracy

of prediction on such data, the scope of a screening trial, the cost of clinical screening methods and the overall potential benefit of early treatment (**Supplementary Text, Figure S4**).

Results

Figure 1. Training and prediction of pancreatic cancer risk from data in the Danish National Patient Registry. The general machine learning workflow starts with partitioning the data into training set (train), development set (dev) and test set (test). The first two data sets are used for training a machine learning model (multiple models were explored) to fit a risk score function (prediction) to a step function (observation) that represents the occurrence of a pancreatic cancer diagnosis, by minimizing the prediction error over all instances (top). Once the best model is found, the model's ability to generalize is assessed using the test set. The model, depending on the threshold selected from among different operational points, discriminates between patients at higher and lower risk of pancreatic cancer (bottom).

Datasets

We used data from DNPR, where data from all encounters (inpatient admissions, outpatient visits and emergency visits) with Danish hospitals have been recorded since

1977. DNPR access was approved by the Danish Health Data Authority (FSEID-00003092 and FSEID-00004491.) Demographic information was obtained by linkage to the Central Person Registry, which is possible via the personal identification number introduced in 1968, that identifies any Danish citizen uniquely over the entire lifespan (Schmidt, Pedersen, and Sørensen 2014). DNPR covers approximately 8.6 million patients with 229 million hospital diagnoses, with an average of 26.7 entries per patient. Each entry includes data on the start and end date of an admission or visit, as well as diagnosis codes. The diagnoses are coded according to the International Classification of Diseases (ICD-8 until 1994 and ICD-10 since then). The accuracy of cancer diagnosis disease codes, as examined by the Danish Health and Medicines Authority, has been reported to be 98% accurate (89.4% correct identification for inpatients and 99.9% for outpatients) (Thygesen et al. 2011). For cancer diagnoses specifically, the reference evaluation was based on detailed comparisons between randomly sampled discharges from five different hospitals and review of a total of 950 samples (Schmidt et al. 2015). We used both the ICD-8 code 157 and ICD-10 code C25, *malignant neoplasm of pancreas*, to define pancreatic cancer (PC) cases. For training we used patient trajectories with explicit time stamps for each hospital contact comprising diagnoses down to the three-character category in the ICD hierarchy. We used data from January 1977 to April 2018 and filtered out patients either with discontinuous disease history or too short a disease history (<5 events in total), ending up with 6.2 million patients (**Figure S1**). The eventual case cohort includes 23,985 pancreatic cancer (PC) cases with cancer occurring at a mean age of 64.6±11.3 years (men) and 67.1±12.1 years (women) (**Table S1**).

Model architecture

Network architecture

The machine learning model for predicting cancer risk from disease trajectories consists of four parts: (1) **input** data for each event in a trajectory (disease code and time stamps), (2) **embedding** of the event features onto real number vectors, (3) **encoding** the trajectories in a latent space, and (4) **predicting** time-dependent cancer risk. (1) **Input**: In order to best exploit the completeness and longitudinality of the DNPR data, all subsequences of diagnoses, starting with the date of birth, from each patient's history were sampled (see Methods). Such data augmentation not only increases the amount of data used in training, but also enables prediction for different time spans between risk assessment and cancer occurrence rather than just a binary prediction that cancer will occur at any time after assessment. (2) **Embedding**: Each element of the subsequence is one of the >2,000 ICD codes. To extract informative features from such high-dimensional inputs, the ML process was set up to embed the categorical input vectors into a continuous, lower-dimensional space. Temporal information, i.e. diagnosis dates and ages at diagnoses are also embedded (see Methods). The embedding layer is data-driven and trained together with other parts of the model. (3) **Encoding**: The longitudinal

nature of the disease trajectories allows us to construct time-sequence models using sequential neural networks, such as gated recurrent units (GRU) models (Cho et al. 2014) and recurrent neural network (RNN) models (Tealab 2018). We also used the Transformer model (Vaswani et al. 2017) which uses an attention mechanism and therefore can capture time information and complex interdependencies. For comparison, we also tested a bag-of-words (i.e., bag-of-disease-codes) approach that ignores the time and order of disease events by pooling the event vectors. (4) **Predicting:** The embedding and encoding layers map each disease trajectory onto a characteristic fingerprint vector in a low-dimensional latent space. This vector is then used as input to a feedforward network to make a prediction of future cancer occurrence within distinct time windows over a period of several years after the end of a trajectory (the time of risk assessment).

Prediction of occurrence within a time interval

For each of the disease trajectories ending at time t_a, a 5-dimensional risk score is calculated, where each dimension represents the risk of cancer occurrence within a particular prediction window after t_a, e.g., 6-12 months or 12-36 months (Lin et al. 2008; Yala et al. 2021). The risk score is constrained to monotonically increase with time as the risk of cancer occurrence naturally increases over time. If and when the risk score exceeds a decision threshold, cancer diagnosis is predicted to have occurred (**Figure 1**) In this way, the model uses a time-sequence of disease codes for one person as input and predicts a cancer diagnosis to occur within 3, 6, 12, 36, 60 months after the time t_a of risk assessment; or not to occur at all in 60 months.

Scanning hyperparameters for each model type

To comprehensively test the performance of different types of neural network architectures, we first conducted an extensive search over hyperparameters and selected the best set of hyperparameters for each network type, and then selected the best network architecture. The model types included transformer, GRU, a multilayer perceptron and bag-of-words. Each model was tested on specific hyperparameter configurations, as shown in (**Table S2**). To avoid overfitting and to test generalizability of model predictions, data from all of the 6.2M patients (including 23,895 pancreatic cancer patients) in the DNPR were partitioned randomly into 80%/10%/10% training/development/test sets. We conducted training only on the training set and used the development set to examine the performance for different hyperparameter settings, which guides model selection. The performance of the selected models was evaluated on the fully withheld test set and reported as an estimate of performance in prospective applications in health care settings with similar availability of longitudinal disease state records.

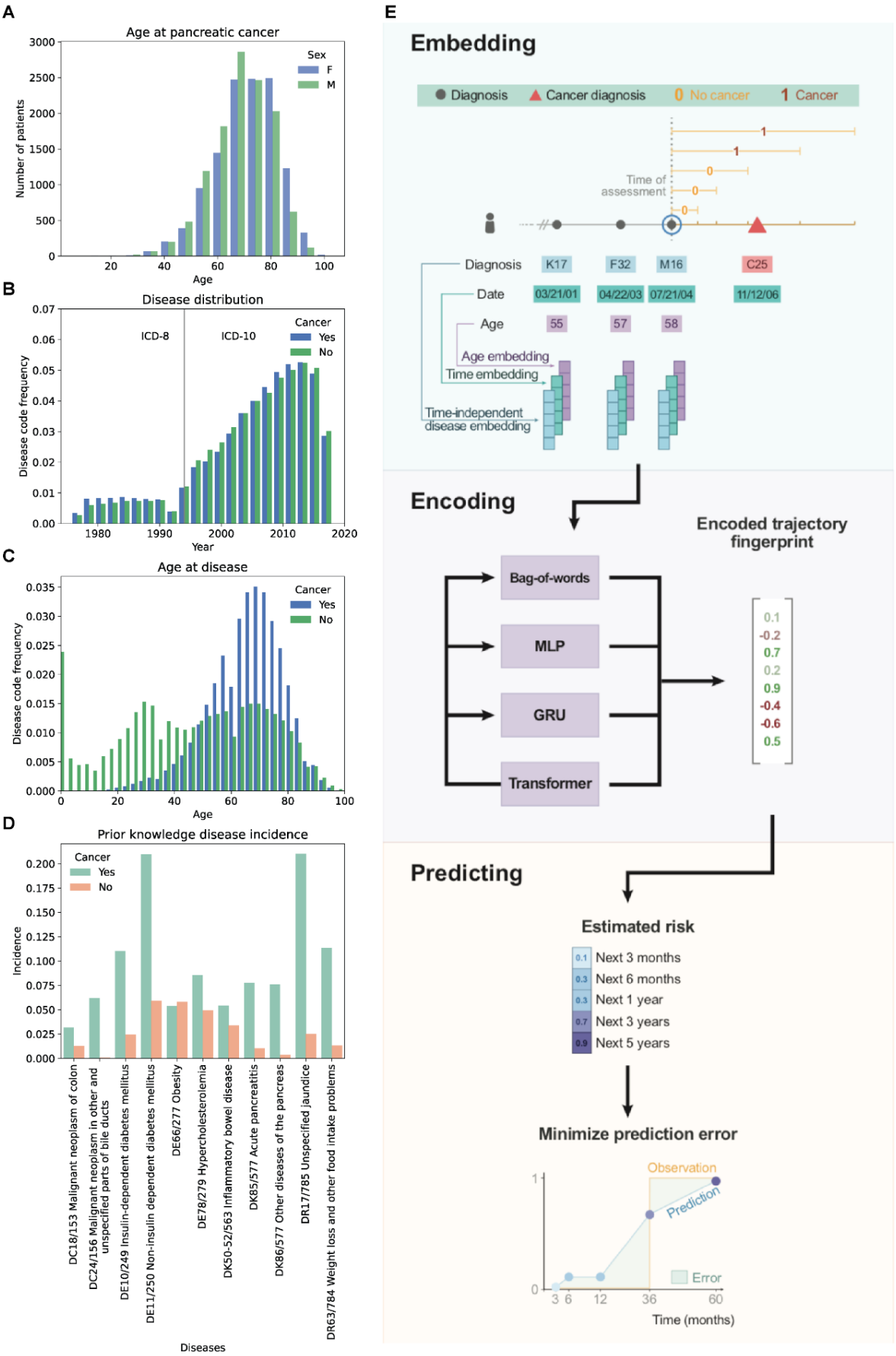


Figure 2. The DNPR database of clinical records covers over 8 million people for up to 40 years, and we used such rich longitudinal information to train a range of deep learning models. (**A-D**) We analyzed the DNPR dataset used for training in terms of the age distribution of pancreatic cancer patients (F female, M male) (**A**); the date distribution (**B**) and age distribution (**C**) of the total number of disease codes in the database; and the incidence of a selected set of previously known risk diseases (**D**). (**E**) The model we developed and trained with real-world clinical data has three steps: embedding, encoding and prediction. The embedding machine transforms categorical disease codes and continuous time stamps into a latent space. The encoding machine extracts information from a disease history and summarizes each sequence in a characteristic fingerprint. The prediction machine then uses the fingerprint to generate predictions for cancer occurrence within different time intervals after the time of assessment (3, 6, 12, 36, 60 months). The model parameters are trained by minimizing the difference between the predicted and the actually observed cancer occurrence.

Evaluation of model performance

Picking a best model - Transformer

We evaluate the different models using the precision-recall curve (PRC) and then highlight the chosen best operational point on the receiver-operating curve (ROC) (**Figure 3**), with the aim to strike a balance between precision and recall taking into account the relatively low incidence of pancreatic cancer in the general population. In the final performance evaluation of different types of ML models on the test set, the models which explicitly use and encode the time information in the sequence of disease codes, i.e., GRU and Transformer, ranked highest (**Figure 3A-C, Table S3**). For the prediction of cancer incidence within 3 years of the assessment date (the date of risk prediction), the Transformer model had the best performance (AUPRC=0.066 [95% confidence interval, 0.063-0.069], AUROC=0.879 (0.877-0.880), followed by GRU (AUPRC=0.040 [0.038-0.041], AUROC=0.852 [0.850-0.854]). The bag-of-words model that ignores the time information along disease trajectories performed significantly less well (AUPRC=0.007 [0.007-0.007], AUROC=0.807 [0.805-0.809]).

Prediction for time intervals

While most risk prediction methods aim to make a binary distinction of whether cancer occurs or not, it is also of interest to consider the time interval within which cancer is likely to occur. The ML models in this work yield explicit risk scores for pancreatic cancer occurrence within 3, 6, 12, 36 and 60 months of the date of risk assessment. As expected, it is more challenging to predict cancer occurrence in longer rather than shorter time intervals. Indeed, prediction performance for the best model decreases from a precision

of 19.4% [18.2-21.9] and recall of 15.6% [14.5-16.5] (99.91% [99.90-99.93] specificity), AUPRC 0.081 [0.076-0.084] for cancer occurrence within 12 months to a precision of 18.2% [17.1-19.7] and recall of 12.4% [11.5-12.9] (99.88% [99.87-99.90] specificity), AUPRC 0.066 [0.063-0.069] for occurrence within 3 years (**Figure 3D-E**). For each ML model and each prediction interval, we picked the operational points that maximize the F1 score, which is the harmonic mean of recall and precision (Sasaki 2007).

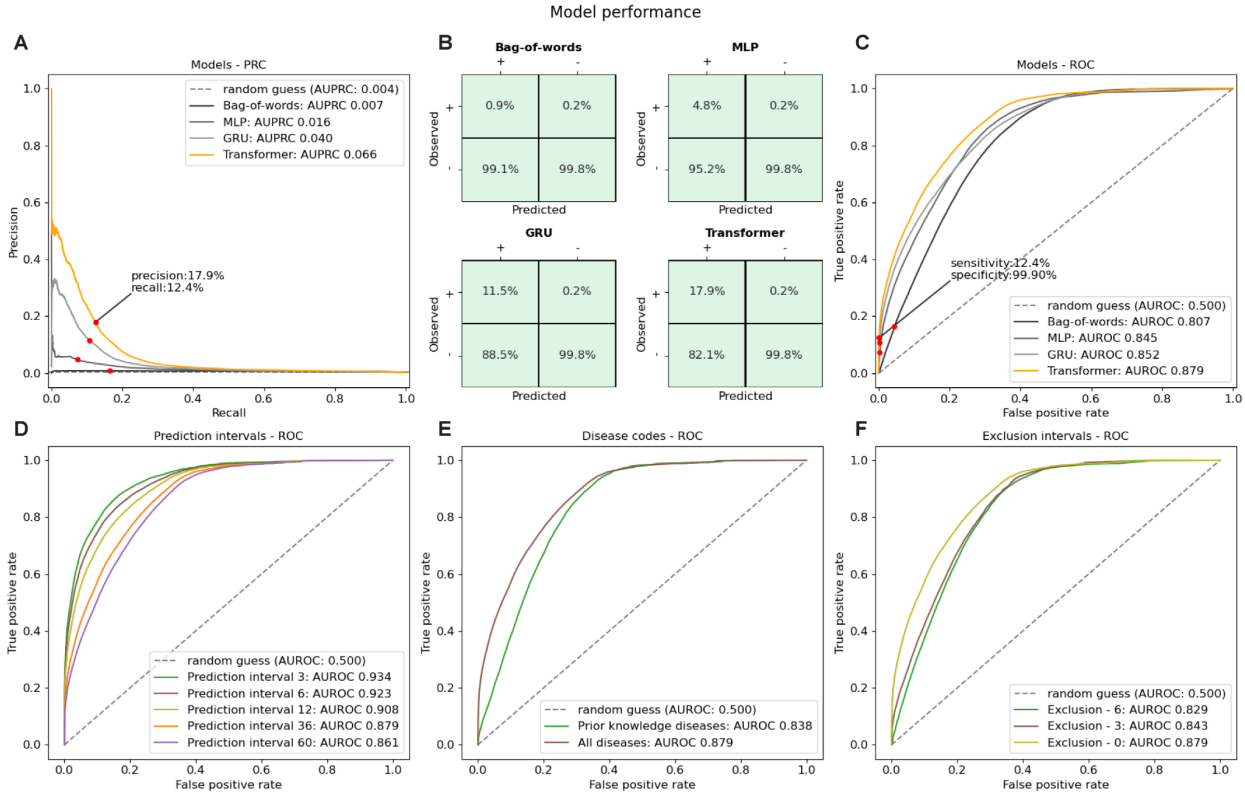


Figure 3. Performance of the deep learning model in predicting pancreatic cancer occurrence. (A-C) Evaluating performance of different ML models for prediction of cancer occurrence within 36 months. (**A**) The precision-recall curve (PRC) plots precision (true positives as a fraction of all predicted positives) against recall (true positives as a fraction of observed positives) for different models at different prediction thresholds along the curve. (**B**) The confusion matrix for each model at the best operational point selected on the PRC curves contains the fraction of true positives, true negatives, false positives and false negatives, normalized by column, to highlight the precision of the models. The operational points were selected to maximize the F1 score. (**C-F**) The ROC curves plot true positive rate (TPR) against false positive rate (FPR) different prediction thresholds, where TPR is the true positives as a fraction of observed positives (recall) and FPR is the false negatives as a fraction of observed negatives (1-specificity). A random prediction (horizontal dotted line in **B**, diagonal line in **C-F**) would have very low precision (AUPRC=incidence=0.004) for all values of recall and equal TPR and FPR (AUROC=0.5). The Transformer is the best performing model for 36-month prediction of cancer occurrence (**B, C**). (**D**) The best-performing Transformer models are evaluated for different prediction intervals starting at the time of assessment and ending at time points up to 60 months. The performance of the transformer model is best for the 0-6 month time interval, but still reasonable up to the 0-60 month prediction interval. Transformer performance (36-month)

compared to the same model trained (**E**) only on a subset of prior knowledge diseases (n=23, **Table S4**) or (**F**) excluding from the input diseases diagnoses in the last 3 or 6 months prior to the diagnosis of pancreatic cancer.

Potential use of prediction time intervals in screening trials

The prediction performance for different time intervals is a crucial parameter in the design of screening trials. In particular, the frequency and timespan of clinical screening tests or the consideration of preventive intervention may differ for different prediction intervals. For example, a patient predicted to be diagnosed with cancer within three months, may be nominated for extensive clinical tests, such as pancreas-directed imaging or sensitive tests of bodily fluids, such as for circulating proteins or cell-free DNA. In contrast, a patient with likely occurrence within a 5-year prediction interval may be nominated for less extensive tests and offered testing at repeat intervals.

Performance with data exclusion

Disease codes within a short time before diagnosis of pancreatic cancer are most probably directly predictive such that even without any machine learning, well-trained clinicians would include pancreatic cancer in their differential diagnosis. Even more so, disease codes just prior to pancreatic cancer occurrence are either semantically similar to it or encompass it (e.g., neoplasm of the digestive tract). To infer earlier detection, we therefore separately trained the models excluding from the input diseases diagnoses in the last 3 or 6 months prior to the diagnosis of pancreatic cancer (**Figure S2**). As expected, e.g., when training with 3-month exclusion of data, the performance for the transformer model decreased to a precision of **4.3%** [3.9-4.7] and recall of 6.5% [5.9-7.0] (99.92% [99.92-99.93] specificity) for cancer occurrence, relative to a precision of **19.4%** [18.2-21.9] and recall of 15.6% [14.5-16.5] (99.91% [99.90-99.93]% specificity) when training with all data - both for cancer occurrence within 12 months (**Figure 3F, Table S3**).

Comparison to previous methods

Earlier work also developed ML methods on real-world data clinical records and predicted pancreatic cancer risk (Appelbaum, Cambronero, et al. 2021; Appelbaum, Berg, et al. 2021; Chen et al. 2021). These previous studies did not use the time sequence of disease histories and for comparison we implemented an analogous approach, a bag-of-words model. We evaluated this model on the DNPR dataset, and the performance for predicting cancer occurrence within 12 months was AUROC=0.807 (0.805-0.809), or **0.62%** (0.58-0.66) precision, 7.7% (5.3-10.0) recall and 98.18% (97.59-98.82) specificity, at an optimal operational point (maximum F1). With time sequence information taken into consideration, the transformer model result is AUROC=0.908 (0.907-0.910), or **19.4%** (18.2-21.9) precision, 15.6% (14.5-16.5) recall and 99.91% (99.90-99.93) specificity. In

other words, when taking time series into account, precision is nearly 20 times higher at double the recall (**Table S3**).

Predictive Features

Interpretation - feature contribution

Although the principal criterion for the potential impact of screening trials is robust predictive performance, it is of interest to interpret the features of any predictive method: which disease diagnoses are most informative of cancer risk? We have therefore used two methods for the identification of factors that contribute most to positive prediction. One method computationally infers the contribution of a particular input variable on the accuracy of prediction. The other method uses prior knowledge and limits the input to disease types, which have been reported to be indicative of the likely occurrence of pancreatic cancer (Yuan et al. 2020; Klein 2021). These prior-knowledge diseases are somewhat predictive of cancer but are less informative compared to the more than 2,000 available diseases (**Table S4**, **Figure 3E**). Our analysis of the contribution of age to the final prediction confirmed the importance of increasing age as a risk factor for pancreatic cancer. (**Figure 4A**).

Interpretation - no data exclusion

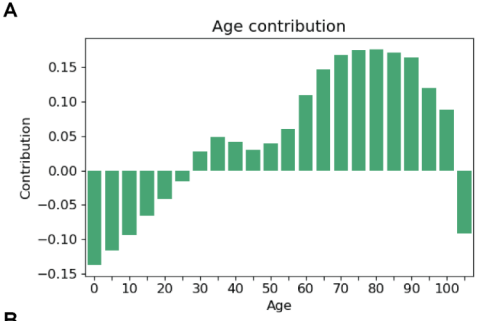
To infer computationally which features contribute most to the assessment of cancer risk, we used the integrated gradients (IG) algorithm (Sundararajan, Taly, and Yan 2017) and ranked the input features by their contribution to the final predictions (**Figure 4B,D**). The top of the IG-extracted list of features may be symptoms of or otherwise closely related to pancreatic cancer, such as '*neoplasms of the digestive tract*' or '*neoplasms without specification of site*', etc. While these certainly can and would be used to predict pancreatic cancer, it is more interesting to identify non-trivial combinations of disease codes that are early risk factors for pancreatic cancer. While one could carefully filter out obviously cancer-related disease codes by expert curation, it is more objective and algorithmically well defined to inspect the IG-extracted predictive features after explicit re-training of the ML model excluding data from the last three or six months before cancer diagnosis (**Figure 4C,E**).

Interpretation - with data exclusion

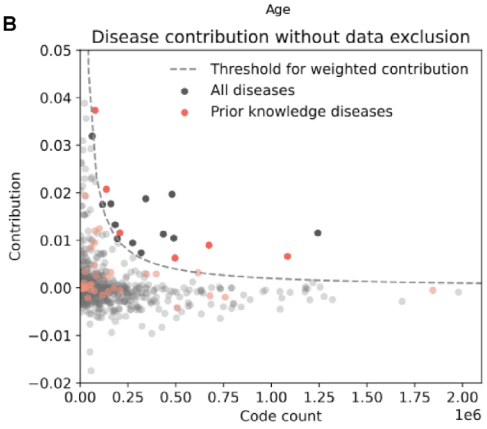
The interpretation of individual risk factors from the ML feature list as mechanistic or causative is typically subject to misinterpretation as their contribution here is only evaluated in the context of complete disease histories. It is, however, interesting to compare the features learned by the neural network system with known risk factors or diagnostic indicators. For example, type II diabetes (non-insulin-dependent), obesity, weight loss or jaundice ranked high in the feature list as important risk factors (**Figure 4**),

consistent with the epidemiological studies (Yuan et al. 2020; Klein et al. 2013; Kim et al. 2020) and the observed disease distribution in the DNPR dataset (**Figure S3**). Other factors, such as cholelithiasis (gallstones) and reflux disease, are perhaps of interest in terms of potential mechanistic hypotheses, such as inflammation of the pancreas as a results of cholelithiasis (gallstones) or a potential link between medication by proton pump inhibitors such as omeprazole in reflux disease (Kearns, Boursi, and Yang 2017; Lai 2021). Alternatively, they may be confounding covariates with other high-scoring factors or may be due to reverse causation by the tumor, i.e., not etiologically related to cancer development, but still useful for risk modeling if the cancer causes them with enough lead time to find the cancer early. Some of the features inferred computationally appear unexpected and may appear as a result of oversampling rare events in the training phase or may be newly identified predictive features. However, our main goal in this report is to achieve robust predictive power from disease histories, rather than mechanistic interpretations.

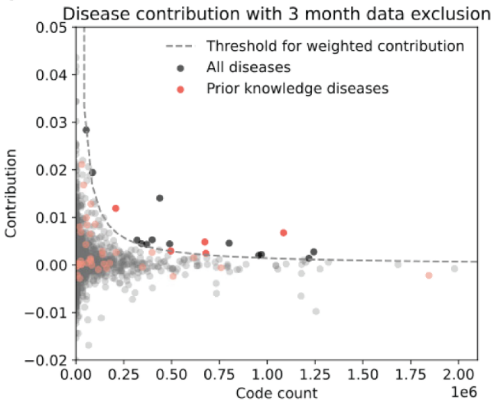
A



B



C



D

Disease contribution without data exclusion

ICD code	Description	Contribution	Frequency weighted contribution
DZ03	Medical observation and evaluation for suspected diseases and conditions	0.003	22578.3
DR10	Abdominal and pelvic pain	0.0115	14341.9
DC78	Secondary malignant neoplasm of respiratory and digestive organs	0.0197	9457.2
DE11	Type 2 diabetes mellitus	0.0066	7165.8
DD64	Other anaemias	0.0187	6424.4
DC34	Malignant neoplasm of bronchus and lung	0.009	6048.4
DK80	Cholelithiasis	0.0104	5115.8
DK59	Other functional intestinal disorders	0.0113	4907.1
197	Secondary malignant neoplasm of respiratory and digestive systems	0.0515	3337.3
DC61	Malignant neoplasm of prostate	0.0062	3099.1
DR63	Symptoms and signs concerning food and fluid intake	0.0373	2917.6
162	Malignant neoplasm of trachea, bronchus and lung	0.0207	2840.4
DK30	Functional dyspepsia	0.0177	2836.2
DC79	Secondary malignant neoplasm of other and unspecified sites	0.0094	2589.1
DK29	Gastritis and duodenitis	0.0133	2428.1
DK21	Gastro-oesophageal reflux disease	0.0116	2402.7
DD12	Benign neoplasm of colon, rectum, anus and anal canal	0.0074	2349.1
DF17	Mental and behavioural disorders due to use of tobacco	0.0175	2048.9
DI80	Phlebitis and thrombophlebitis	0.0104	2013.1

E

Disease contribution with 3-month data exclusion

ICD code	Description	Contribution	Frequency weighted contribution
DZ03	Medical observation and evaluation for suspected diseases and conditions	0.0016	12428.5
DE11	Type 2 diabetes mellitus	0.0068	7342.7
DZ12	Special screening examination for neoplasms	0.014	6152.3
DZ01	Other special examinations and investigations of persons without complaint or reported diagnosis	0.0004	5479.5
DZ50	Care involving use of rehabilitation procedures	0.0046	3664
DR10	Abdominal and pelvic pain	0.0027	3418.9
DC34	Malignant neoplasm of bronchus and lung	0.0048	3255.2
DK21	Gastro-oesophageal reflux disease	0.0119	2475.3
DK80	Cholelithiasis	0.0044	2166.5
DS60	Superficial injury of wrist and hand	0.0022	2106
DR06	Abnormalities of breathing	0.0053	2097.4
DZ13	Special screening examination for other diseases and disorders	0.002	1917.2
DZ38	Liveborn infants according to place of birth	0.0014	1725.3
DK64	Haemorrhoids and perianal venous thrombosis	0.0194	1700.9
DE66	Obesity	0.0025	1672.3
DD12	Benign neoplasm of colon, rectum, anus and anal canal	0.0052	1658.9
DN92	Excessive, frequent and irregular menstruation	0.0043	1599.3
DD64	Other anaemias	0.0045	1527.2
DN46	Male infertility	0.0284	1518
DC61	Malignant neoplasm of prostate	0.0029	1457.8

Figure 4. Interpreting the predictive features of the best-performing model with the integrated gradient method. (**A**) Contribution of age, a known major risk factor. (**B-C**) Distribution of all the disease codes in regard to the contribution and relative frequency in the database for models trained including (**B**) or excluding (**C**) diagnoses that occurred within 3 months before cancer diagnoses. We used an integrated gradients (IG) method to calculate the contribution score for each input feature and weighted them by their overall frequency in the database. The top 20 features (most predictive and frequent) are above the dashed line. (**D**-**E**) Top features that contribute the most to the cancer prediction, without data exclusion (**D**) or with 3-month data exclusion (**E**). The features are sorted by the weighted contribution (intensity of red with increasing value, last column), which is the product of the contribution value (intensity of red with increasing value, next to last column) and the disease frequency (not shown). Prior-knowledge diseases (red text), i.e., those reported as risk factors in the literature (**Table S4)**, are among the top-scoring predictive features.

Discussion

Advances in this work

Here we presented a new framework for applying deep learning methods using EHR disease histories to predict disease risk in support of early detection. The main potential advantages in our study are the explicit use of the time series of disease events in ML of cancer risk and the higher quality and consistency of the clinical records in the Danish health system with nationally organized lifetime patient records. Earlier ML methods applied to large-scale clinical records have demonstrated the potential of applying AI methods to assess pancreatic cancer risk. However, these studies did not exploit the information in the temporal sequence of diseases. Our results indicate that using the time ordering in disease histories as input significantly improves the predictive power of AI methods in anticipating disease occurrence.

Challenge: different health systems

A future challenge for this work, beyond the scope of the current study, is to apply the same ML training methodology to clinical records from other healthcare systems, in which large real-world data sets are available, e.g., by transfer learning or completely independent training. Given the reasonable semantic agreement of the international ICD disease codes, one can apply the ML engine with parameters learned in one system directly to make predictions for data from another system. However, the transfer of rules learned from the Danish data to a system, e.g., in the US, is not guaranteed to lead to a satisfactory level of prediction performance, given the differences in population distribution of diseases, in how events are coded, the number of codes in national implementations of ICD, in diagnostic/screening procedures, and in incidence patterns for comorbidities. Nevertheless, the models derived here offer a critical starting point for training on other datasets, and potentially be combined with transfer learning to broaden the applicability of the proposed prediction tools.

Challenge: more data

In the future, we expect further improvements by augmenting the ICD-10 disease codes with other data items, such as abdominal CT scans, MRIs, laboratory values, genetic data, prescriptions, observations extracted from clinical notes that are not in disease procedure codes as well as diagnosis and treatment records from general practitioners (Malhotra et al. 2021). To achieve a globally useful set of prediction rules, one would like to have access to large data sets of disease histories aggregated nationally or internationally. An ideal scenario for a multi-institutional collaboration would be to employ federated learning across a number of different healthcare systems (Konečný et al. 2016). Federated learning obviates the need for sharing primary data and only requires

permission to run logically identical computer codes at each location and then share and aggregate the results.

Personalized screening design

There is an additional opportunity to tailor the nature of clinical investigations for different sections of the high risk population. For example, a patient predicted to get cancer within the next 3 to 6 months would immediately be assigned to detailed testing with imaging and blood tests, such as CA19-9, cell free DNA for mutations or changes in methylation patterns (Widschwendter et al. 2017; Liu et al. 2020; Fahrmann et al. 2021). On the other hand, someone projected to get cancer in the two to three year time frame might be assigned to a serial surveillance program, with periodic reassessment of risk. Such personalized screening would be facilitated by further improvement in the time aspect of risk prediction and by the availability of more sensitive biomarkers for early cancer.

Retraining for pre-selected cohorts

One attractive scenario is a two-step selection process. For example, a screening trial may be limited to people above a certain age (e.g., > 50 years) who also have type II diabetes (NCT03731637). With these selection criteria, one can in a fairly straightforward manner retrain the machine learning prediction methods to optimize the testing and intervention details of a screening trial in the restricted cohort (“Which aspects shall we test for that are likely risk factors?” “Which are the best tests given the history of the patient?”) and assess the expected prediction accuracy. This in turn would lead to revised cost/benefit estimates and revised cohort selection (“how many and which patients?”) of a screening program.

Decision support for clinicians

A reasonably accurate method for predicting cancer risk affords the opportunity to direct high risk patients into clinical screening trials, which if successful would produce a scalable workflow for early detection of pancreatic cancer in the community. The risk assessment tools described herein could then be made available as decision support tools for clinicians. A sufficiently enriched pool of high-risk patients would make detailed clinical tests affordable, which currently are prohibitively expensive at a population level, and greatly enhance the positive predictive value of these tests.

Impact on patients

Prediction performance at the level shown here may be sufficient for the design of real world clinical screening trials, in which high risk patients are assigned to high specificity screening tests and, if cancer is detected, offered early treatment. AI on real-world clinical records has the potential to shift focus from treatment of late- to early-stage cancer, improving the quality of life of patients and increasing the benefit/cost ratio.

Methods

Processing of the population-level dataset

The project was conducted using a dataset of disease history from the Danish National Patient Registry (DNPR), covering all 229 million hospital diagnoses of 8.6 million patients between 1977-2018 including inpatient, outpatient, and emergency department contacts (Schmidt et al. 2015). DNPR contains disease codes annotated according to the International Classification of Diseases (ICD), as well as corresponding admission dates as well as demographic information for each patient, such as age and sex, via the Civil Registration System.

The most updated ICD classification system has a hierarchical structure, from the most general level, e.g., *C: Neoplasms*, to the most specific four-character subcategories e.g. *C25.1: Malignant neoplasm of body of pancreas*. DNPR contains ICD-10 codes for disease administration after 1994 and ICD-8 codes for the remaining period of the registry. The Danish version of the ICD-10 is more detailed than the international ICD-10 but less detailed than the clinical modification of the ICD-10 (ICD-10-CM). In this study, we used the three-character category ICD codes (n=2,997) in constructing the predictive models and defined “pancreatic cancer (PC) patients” as patients with at least one code under *C25: Malignant neoplasm of pancreas*. The "non-PC patients" are defined as any other patients with at least 2 years free of PC before the end of the medical history. The end date is defined as the date of death, the end date used to select from the DNPR dataset (April 2018), the date of emigration, or the date of the last hospital visit, whichever is earlier.

For the diagnosis codes in the DNPR, we removed disease codes labelled as ‘temporary’ or ‘referral’ (8.3% removed, **Figure S1**), as these can be misinterpreted when mixed with the main diagnoses and are not valuable for the purposes of this study.

Danish citizens are assigned a unique lifetime Civil Registration Number (CPR), which is useful for linking to person-specific demographic data. Using these we retrieved patient status as to whether patients are active or inactive in the CPR system as well as information related to residence status. We applied a demographic continuity filter. For example, we removed from consideration residents of Greenland, patients who lack a stable place of residence in Denmark or moved abroad, as these would potentially have discontinuous observation times. By observation time we mean active use of the healthcare system.

At this point, the dataset comprised a total of 8,110,706 patients, of which 23,601 had the ICD-10 pancreatic cancer code C25 and 14,720 had the ICD-8 pancreatic cancer code 157. We used both ICD-10 and ICD-8 independently, without semantic mapping, while retaining the cancer occurrence label, assuming that machine learning is able to combine information from both. Subsequently, we removed patients that have too few diagnoses (<5 events). The number of positive patients used for training after applying the length filter are 23,985 (82% ICD-10 and 18% ICD-8). Coincidentally, this resulted in a more

strict filtering for ICD-8 events which were used only in 1977-1994. The final dataset was then randomly split into training (80%), development (10%) and test (10%) data, with the condition that all trajectories from a patient were only included in one split group, to avoid any information leakage between training and development/test datasets.

For each patient, whether or not they ever had pancreatic cancer, the data was augmented by using all continuous partial trajectories of (minimal length >=5 diagnoses) from the beginning of their disease history and ending at different time points, which we call the time of assessment. For cancer patients, we used only partial trajectories that end before cancer diagnoses, i.e. $t_a<t_{cancer}<t_{death}$. We used a step function annotation indicating cancer occurrence at different time points (3, 6, 12, 36, 60 months) after the end of each partial trajectory. For the positive ('PC') cases this provides the opportunity to learn from disease histories with a significant time gap between the time of assessment and the time of cancer occurrence. For example, for a patient, who had pancreatitis a month or two just before the cancer diagnosis, it is of interest to learn which earlier disease codes might have been predictive of cancer occurrence going back at least several months or perhaps years. The latter is also explored by separately re-training of the ML model excluding data from the last three or six months before cancer diagnosis.

Model development

A desired model for such diagnosis trajectories consists of three parts: embedding of the categorical disease features, encoding time sequence information, and assessing the risk of cancer. We embed the discrete and high-dimensional disease vectors in a continuous and low-dimensional latent space (Mikolov et al. 2013; Gehring et al. 2017). Such embedding is data-driven and trained together with other parts of the model. For ML models not using embedding, each categorical disease was represented in numeric form as a one-hot encoded vector. The longitudinal records of diagnoses allowed us to construct time-sequence models with sequential neural networks. Each sequence of diagnoses, after embedding, was encoded into a feature vector using different types of sequential layers (recurrent neural network, RNN, and gated recurrent units, GRU), attention layers (transformer), or simple pooling layers (bag-of-words regression model and multilayer perceptron model [MLP]). The encoding layer also included age inputs, which has been demonstrated to have a strong association with pancreatic cancer incidence (Klein 2021). Finally, the embedding and encoding layers were connected to a fully-connected feedforward network (FF) to make predictions of future cancer occurrence following a given disease history (the bag-of-words regression model only uses a single logistic regression layer).

The model output consists of a risk score that monotonically increases for each time interval in the followup period. As cancer by definition occurs before cancer diagnosis, the risk score at a time point t is interpreted as quantifying the risk of cancer occurrence between t_a, the end of the disease trajectory (the time of assessment), and time $t = t_a +$

3, 6, 12, 36, 60 months. For a given risk score threshold, which is an operational parameter in making predictions, only scores that exceed the threshold are considered to indicate cancer occurrence. We currently do not distinguish between different stages of cancer, neither in training from cancer diagnoses nor in the prediction of cancer occurrence.

The model parameters were trained by minimizing the prediction error quantified as the difference between the observed cancer diagnosis in the form of a step function (0 before the occurrence of cancer, 1 from the time of cancer diagnosis) and the predicted risk score in terms of a positive function that monotonically increases from 0, using a cross-entropy loss function, with the sum over the five time points, and L2 regularization on the parameters (**Figure 2**).

$$loss = \frac{1}{N}\frac{1}{N_T}\sum_{i,t}\left[y_{i,t}\ \log\left[\ \hat{p}_{\Theta,t}(x_i)\ \right] + \left(1 - y_{i,t}\right)\log\left[1 - \hat{p}_{\Theta,t}(x_i)\right]\right] + \lambda_2||\Theta||_2$$

where $t \in \{3,6,12,36,60\}$ months; N_T is the number of time points, $N_T = 5$ for non-cancer patients and $N_T \leq 5$ for cancer patients where we only use the time points before the cancer diagnosis; $i = 1,2,3 \ldots N$ samples; Θ is the set of model parameters; λ_2 is the regularization strength; $\hat{p}$ is the model prediction; x_i are the input disease trajectories, $y_{i,t} = 1$ for cancer occurrence or 0 for no cancer within a t-month time window.

The transformer model, unlike the recurrent models, does not process the input as a sequence of time steps but rather uses an attention mechanism to enhance the embedding vectors correlated with the outcome. In order to enable the transformer to digest temporal information such as the order of the exact dates of the diseases inside the sequence, we used positional embedding to encode the temporal information into vectors which were then used as weights for each disease token. Here we adapted the positional embedding from (Vaswani et al. 2017) using the values taken by cosine waveforms at 128 frequencies observed at different times. The times used to extract the wave values were the age at which each diagnosis was administered and the time difference between each diagnosis. In this way the model is enabled to distinguish between the same disease assigned at different times as well as two different disease diagnoses far and close in time. The parameters in the embedding machine, which addresses the issue of data representation suitable for input into a deep learning network, were trained together with the encoding and prediction parts of the model with back propagation (**Figure 2**).

To comprehensively test different types of neural networks and the corresponding hyperparameters, we conducted a large parameter search for each of the network types

(**Table S2**). The different types of models include simple feed-forward models (LR, MLP) and more complex models that can take the sequential information of disease ordering into consideration (RNN, GRU and Transformer). See supplementary table with comparison metrics across different models (**Table S3**). In order to estimate the uncertainty of the performances, the 95% confidence interval was constructed using 200 resamples of bootstrapping with replacement.

For patients without a pancreatic cancer diagnosis, we only include trajectories that end at least 2 years before the end of their disease records (death or the end date of the DNPR). This avoids the uncertainty of cases in which undiagnosed cancer might have existed before the end of the records. For patients with a pancreatic cancer diagnosis code, we used a time window of 5 years prior to cancer onset to define valid positive cases. Trajectories that end no earlier than 5 years before cancer diagnosis are considered positive samples, and the others are counted as negative samples. Due to the small number of cases of pancreatic cancer compared to controls, we used dataset balancing in training. For each training batch/epoch, we oversampled the pancreatic cancer cases to match the count of negative controls.

Interpreting clinically relevant features

In order to find the features that are strongly associated with pancreatic cancer, we have used an attribution method for neural networks called integrated gradients (Sundararajan, Taly, and Yan 2017). This method calculates the contribution of input features, called attribution, aggregating the gradients along the path from the input to the baseline. Here we picked the output of interest to be the 36-month prediction. Positive and negative attribution scores (contribution to prediction) indicate positive correlation with pancreatic cancer patients and non-pancreatic-cancer patients, respectively. Since the gradient cannot be calculated with respect to the indices used as input of the embedding layer, the input used for the attribution was the output of the embedding layer. Then, the attribution at the token level was obtained summing up over each embedding dimension and averaged across all the patient trajectories. We therefore developed a normalized contribution score by weighting the attribution with the overall frequency of each disease respectively. Similarly, for each trajectory, we calculated the age contribution as the sum attribution of the integrated gradients of the input at the age embedding layer.

Software

The software will be made freely available in source code on a github repository.

Acknowledgements

We thank Adam Yala for expert advice and methodological contributions, Regina Barzilay for discussion and guidance. We thank V. Yeung, S. Knemeyer and T. Rogers from SciStories for their help with displayed items. DP, JXH, ADH and SB acknowledge support from the Novo Nordisk Foundation (grants NNF17 OC0027594 and NNF14CC0001). BMW acknowledges support from the Hale Family Center for Pancreatic Cancer Research, Lustgarten Foundation Dedicated Laboratory program, NIH grant U01 CA210171, NIH grant P50 CA127003, Stand Up to Cancer, Pancreatic Cancer Action Network, Noble Effort Fund, Wexler Family Fund, Promises for Purple, and Bob Parsons Fund. We thank Stand Up to Cancer, the Lustgarten Foundation and their donors for financial and community support.

Conflict of Interest Statements

S.B. has ownerships in Intomics A/S, Hoba Therapeutics Aps, Novo Nordisk A/S, Lundbeck A/S and managing board memberships in Proscion A/S and Intomics A/S. B.M.W. notes grant funding from Celgene and Eli Lilly; consulting fees from BioLineRx, Celgene, and GRAIL. A.R. is a co-founder and equity holder of Celsius Therapeutics, an equity holder in Immunitas, and was an SAB member of ThermoFisher Scientific, Syros Pharmaceuticals, Neogene Therapeutics and Asimov until July 31, 2020. From August 1, 2020, A.R. is an employee of Genentech.

REFERENCES

Amundadottir, Laufey, Peter Kraft, Rachael Z. Stolzenberg-Solomon, Charles S. Fuchs, Gloria M. Petersen, Alan A. Arslan, H. Bas Bueno-de-Mesquita, et al. 2009. “Genome-Wide Association Study Identifies Variants in the ABO Locus Associated with Susceptibility to Pancreatic Cancer.” *Nature Genetics* 41 (9): 986–90.

Appelbaum, Limor, Alexandra Berg, Jose Pablo Cambronero, Thurston Hou Yeen Dang, Charles Chuan Jin, Lori Zhang, Steven Kundrot, et al. 2021. “Development of a Pancreatic Cancer Prediction Model Using a Multinational Medical Records Database.” ASCO GI Symposium, January. https://doi.org/10.1200/JCO.2021.39.3_suppl.394.

Appelbaum, Limor, José P. Cambronero, Jennifer P. Stevens, Steven Horng, Karla Pollick, George Silva, Sebastien Haneuse, et al. 2021. “Development and Validation of a Pancreatic Cancer Risk Model for the General Population Using Electronic Health Records: An Observational Study.” *European Journal of Cancer* 143 (January): 19–30.

Blackford, Amanda L., Marcia Irene Canto, Alison P. Klein, Ralph H. Hruban, and Michael Goggins. 2020. “Recent Trends in the Incidence and Survival of Stage 1A Pancreatic Cancer: A Surveillance, Epidemiology, and End Results Analysis.” *Journal of the National Cancer Institute* 112 (11): 1162–69.

Chen, Qinyu, Daniel R. Cherry, Vinit Nalawade, Edmund M. Qiao, Abhishek Kumar, Andrew M. Lowy, Daniel R. Simpson, and James D. Murphy. 2021. “Clinical Data Prediction Model to Identify Patients With Early-Stage Pancreatic Cancer.” *JCO Clinical Cancer Informatics* 5 (March): 279–87.

Cho, Kyunghyun, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. “Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1406.1078.

Dietterich, Thomas G. 2002. “Machine Learning for Sequential Data: A Review.” In *Structural, Syntactic, and Statistical Pattern Recognition*, 15–30. Springer Berlin Heidelberg.

Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. 2017. “Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks.” *Nature* 542 (7639): 115–18.

Fahrmann, Johannes F., C. Max Schmidt, Xiangying Mao, Ehsan Irajizad, Maureen Loftus, Jinming Zhang, Nikul Patel, et al. 2021. “Lead-Time Trajectory of CA19-9 as an Anchor Marker for Pancreatic Cancer Early Detection.” *Gastroenterology* 160 (4): 1373–83.e6.

Gehring, Jonas, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. “Convolutional Sequence to Sequence Learning.” In *Proceedings of the 34th International Conference on Machine Learning*, edited by Doina Precup and Yee Whye Teh, 70:1243–52. Proceedings of Machine Learning Research. PMLR.

Hyland, Stephanie L., Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, et al. 2020. “Early Prediction of Circulatory Failure in the Intensive Care Unit Using Machine Learning.” *Nature Medicine* 26 (3): 364–73.

Kearns, Malcolm D., Ben Boursi, and Yu-Xiao Yang. 2017. “Proton Pump Inhibitors on Pancreatic Cancer Risk and Survival.” *Cancer Epidemiology* 46 (February): 80–84.

Kim, Jihye, Chen Yuan, Ana Babic, Ying Bao, Clary B. Clish, Michael N. Pollak, Laufey T. Amundadottir, et al. 2020. “Genetic and Circulating Biomarker Data Improve Risk Prediction for Pancreatic Cancer in the General Population.” *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* 29 (5): 999–1008.

Klein, Alison P. 2021. “Pancreatic Cancer Epidemiology: Understanding the Role of Lifestyle and Inherited Risk Factors.” *Nature Reviews. Gastroenterology & Hepatology*, May.

https://doi.org/10.1038/s41575-021-00457-x.

Klein, Alison P., Sara Lindström, Julie B. Mendelsohn, Emily Steplowski, Alan A. Arslan, H. Bas Bueno-de-Mesquita, Charles S. Fuchs, et al. 2013. "An Absolute Risk Model to Identify Individuals at Elevated Risk for Pancreatic Cancer in the General Population." *PloS One* 8 (9): e72311.

Klein, Alison P., Brian M. Wolpin, Harvey A. Risch, Rachael Z. Stolzenberg-Solomon, Evelina Mocci, Mingfeng Zhang, Federico Canzian, et al. 2018. "Genome-Wide Meta-Analysis Identifies Five New Susceptibility Loci for Pancreatic Cancer." *Nature Communications* 9 (1): 556.

Konečný, Jakub, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2016. "Federated Learning: Strategies for Improving Communication Efficiency." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1610.05492.

Lai, Shih-Wei. 2021. "Proton Pump Inhibitors and the Risk of Pancreatic Cancer." *Journal of Gastroenterology* 56 (3): 293–94.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521 (7553): 436–44.

Li, Donghui, Eric J. Duell, Kai Yu, Harvey A. Risch, Sara H. Olson, Charles Kooperberg, Brian M. Wolpin, et al. 2012. "Pathway Analysis of Genome-Wide Association Study Data Highlights Pancreatic Development Genes as Susceptibility Factors for Pancreatic Cancer." *Carcinogenesis* 33 (7): 1384–90.

Lin, Ray S., Susan D. Horn, John F. Hurdle, and Alexander S. Goldfarb-Rumyantzev. 2008. "Single and Multiple Time-Point Prediction Models in Kidney Transplant Outcomes." *Journal of Biomedical Informatics* 41 (6): 944–52.

Liu, M. C., G. R. Oxnard, E. A. Klein, C. Swanton, M. V. Seiden, and CCGA Consortium. 2020. "Sensitive and Specific Multi-Cancer Detection and Localization Using Methylation Signatures in Cell-Free DNA." *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO* 31 (6): 745–59.

Malhotra, Ananya, Bernard Rachet, Audrey Bonaventure, Stephen P. Pereira, and Laura M. Woods. 2021. "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.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1301.3781.

Muhammad, Wazir, Gregory R. Hart, Bradley Nartowt, James J. Farrell, Kimberly Johung, Ying Liang, and Jun Deng. 2019. "Pancreatic Cancer Prediction Through an Artificial Neural Network." *Frontiers in Artificial Intelligence* 2 (May): 2.

Nielsen, Annelaura B., Hans-Christian Thorsen-Meyer, Kirstine Belling, Anna P. Nielsen, Cecilia E. Thomas, Piotr J. Chmura, Mette Lademann, et al. 2019. "Survival Prediction in Intensive-Care Units Based on Aggregation of Long-Term Disease History and Acute Physiology: A Retrospective Study of the Danish National Patient Registry and Electronic Patient Records." *The Lancet. Digital Health* 1 (2): e78–89.

Pereira, Stephen P., Lucy Oldfield, Alexander Ney, Phil A. Hart, Margaret G. Keane, Stephen J. Pandol, Debiao Li, et al. 2020. "Early Detection of Pancreatic Cancer." *The Lancet. Gastroenterology & Hepatology* 5 (7): 698–710.

Petersen, Gloria M., Laufey Amundadottir, Charles S. Fuchs, Peter Kraft, Rachael Z. Stolzenberg-Solomon, Kevin B. Jacobs, Alan A. Arslan, et al. 2010. "A Genome-Wide Association Study Identifies Pancreatic Cancer Susceptibility Loci on Chromosomes 13q22.1, 1q32.1 and 5p15.33." *Nature Genetics* 42 (3): 224–28.

Rahib, Lola, Benjamin D. Smith, Rhonda Aizenberg, Allison B. Rosenzweig, Julie M. Fleshman, and Lynn M. Matrisian. 2014. "Projecting Cancer Incidence and Deaths to 2030: The Unexpected Burden of Thyroid, Liver, and Pancreas Cancers in the United States." *Cancer*

Research 74 (11): 2913–21.
Sasaki, Yutaka. 2007. “The Truth Oh the F--Measure.” *Manchester: School of Computer Science, University of Manchester*.
Schmidt, Morten, Lars Pedersen, and Henrik Toft Sørensen. 2014. “The Danish Civil Registration System as a Tool in Epidemiology.” *European Journal of Epidemiology* 29 (8): 541–49.
Schmidt, Morten, Sigrun Alba Johannesdottir Schmidt, Jakob Lynge Sandegaard, Vera Ehrenstein, Lars Pedersen, and Henrik Toft Sørensen. 2015. “The Danish National Patient Registry: A Review of Content, Data Quality, and Research Potential.” *Clinical Epidemiology* 7 (November): 449–90.
Shickel, Benjamin, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. 2018. “Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis.” *IEEE Journal of Biomedical and Health Informatics* 22 (5): 1589–1604.
Siggaard, Troels, Roc Reguant, Isabella F. Jørgensen, Amalie D. Haue, Mette Lademann, Alejandro Aguayo-Orozco, Jessica X. Hjaltelin, Anders Boeck Jensen, Karina Banasik, and Søren Brunak. 2020. “Disease Trajectory Browser for Exploring Temporal, Population-Wide Disease Progression Patterns in 7.2 Million Danish Patients.” *Nature Communications* 11 (1): 4952.
Singhi, Aatur D., Eugene J. Koay, Suresh T. Chari, and Anirban Maitra. 2019. “Early Detection of Pancreatic Cancer: Opportunities and Challenges.” *Gastroenterology* 156 (7): 2024–40.
Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. 2017. “Axiomatic Attribution for Deep Networks.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1703.01365.
Tealab, Ahmed. 2018. “Time Series Forecasting Using Artificial Neural Networks Methodologies: A Systematic Review.” *Future Computing and Informatics Journal* 3 (2): 334–40.
Thorsen-Meyer, Hans-Christian, Annelaura B. Nielsen, Anna P. Nielsen, Benjamin Skov Kaas-Hansen, Palle Toft, Jens Schierbeck, Thomas Strøm, et al. 2020. “Dynamic and Explainable Machine Learning Prediction of Mortality in Patients in the Intensive Care Unit: A Retrospective Study of High-Frequency Data in Electronic Patient Records.” *The Lancet. Digital Health* 2 (4): e179–91.
Thygesen, Sandra K., Christian F. Christiansen, Steffen Christensen, Timothy L. Lash, and Henrik T. Sørensen. 2011. “The Predictive Value of ICD-10 Diagnostic Coding Used to Assess Charlson Comorbidity Index Conditions in the Population-Based Danish National Registry of Patients.” *BMC Medical Research Methodology* 11 (May): 83.
Tomašev, Nenad, Xavier Glorot, Jack W. Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, et al. 2019. “A Clinically Applicable Approach to Continuous Prediction of Future Acute Kidney Injury.” *Nature* 572 (7767): 116–19.
Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is All You Need.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1706.03762.
Widschwendter, Martin, Michal Zikan, Benjamin Wahl, Harri Lempiäinen, Tobias Paprotka, Iona Evans, Allison Jones, et al. 2017. “The Potential of Circulating Tumor DNA Methylation Analysis for the Early Detection and Management of Ovarian Cancer.” *Genome Medicine* 9 (1): 116.
Wolpin, Brian M., Cosmeri Rizzato, Peter Kraft, Charles Kooperberg, Gloria M. Petersen, Zhaoming Wang, Alan A. Arslan, et al. 2014. “Genome-Wide Association Study Identifies Multiple Susceptibility Loci for Pancreatic Cancer.” *Nature Genetics* 46 (9): 994–1000.
Yala, Adam, Constance Lehman, Tal Schuster, Tally Portnoi, and Regina Barzilay. 2019. “A Deep Learning Mammography-Based Model for Improved Breast Cancer Risk Prediction.” *Radiology* 292 (1): 60–66.
Yala, Adam, Peter G. Mikhael, Fredrik Strand, Gigin Lin, Kevin Smith, Yung-Liang Wan, Leslie

Lamb, Kevin Hughes, Constance Lehman, and Regina Barzilay. 2021. “Toward Robust Mammography-Based Models for Breast Cancer Risk.” *Science Translational Medicine* 13 (578). https://doi.org/10.1126/scitranslmed.aba4373.

Yamada, Masayoshi, Yutaka Saito, Hitoshi Imaoka, Masahiro Saiko, Shigemi Yamada, Hiroko Kondo, Hiroyuki Takamaru, et al. 2019. “Development of a Real-Time Endoscopic Image Diagnosis Support System Using Deep Learning Technology in Colonoscopy.” *Scientific Reports* 9 (1): 14465.

Yuan, Chen, Ana Babic, Natalia Khalaf, Jonathan A. Nowak, Lauren K. Brais, Douglas A. Rubinson, Kimmie Ng, et al. 2020. “Diabetes, Weight Change, and Pancreatic Cancer Risk.” *JAMA Oncology* 6 (10): e202948.

Supplementary Materials

Figure S1. Preprocessing and filtering of the DNPR diagnosis data.

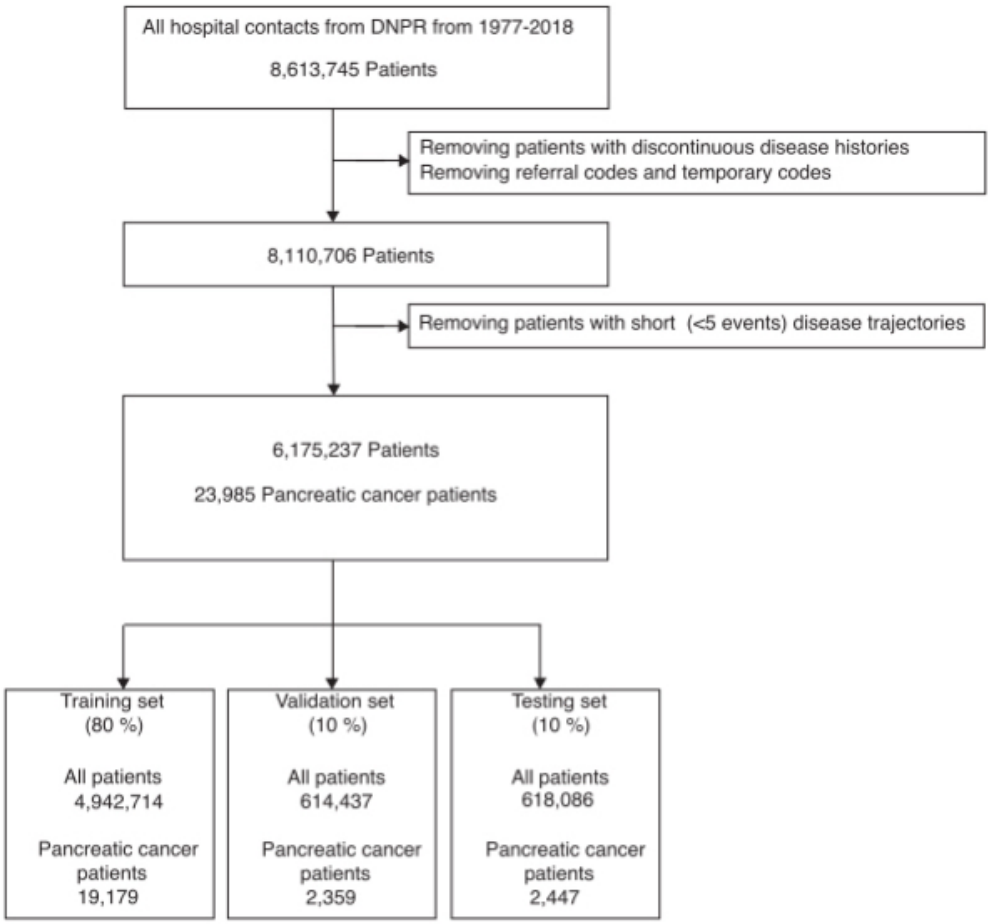


Table S1. Description of the patient cohorts used in this study (DNPR).

Population Metadata (n=8,110,706 persons)		
Gender	Male	Female
Total Count	4,030,504 (49.69%)	4,080,202 (50.31%)
Alive	2,754,152 (33.96%)	2,827,021 (34.86%)
Dead	1,276,352 (15.74%)	1,253,181 (15.45%)
After continuity and length filtering	2,938,248 (36.23%)	3,239,989 (39.95%)
Age at last record (0-10)	216,329 (2.67%)	204,774 (2.52%)
Age at last record (10-20)	332,326 (4.10%)	314,445 (3.88%)
Age at last record (20-30)	322,802 (3.98%)	298,219 (3.68%)
Age at last record (30-40)	283,200 (3.49%)	305,470 (3.77%)
Age at last record (40-50)	323,811 (3.99%)	380,730 (4.69%)
Age at last record (50-60)	368,686 (4.55%)	419,100 (5.17%)
Age at last record (60-70)	373,220 (4.60%)	402,625 (4.96%)
Age at last record (70-80)	394,789 (4.87%)	408,890 (5.04%)
Age at last record (80-90)	258,193 (3.18%)	342,174 (4.22%)
Age at last record (90-100)	63,470 (0.78%)	156,154 (1.93%)
Age at last record (100-110)	1,422 (0.02%)	7,391 (0.09%)
Age at last record (110-120)		7 (0.00%)

Pancreatic Cancer Patients (n=23,895)		
	Male	Female
Total Count	11,880 (49.53%)	12,105 50.47%
Age at pancreatic cancer diagnosis (0-10)	1 (0.00%)	1 (0.00%)
Age at pancreatic cancer diagnosis (10-20)	1 (0.00%)	7 (0.03%)
Age at pancreatic cancer diagnosis (20-30)	11 (0.05%)	11 (0.05%)
Age at pancreatic cancer diagnosis (30-40)	92 (0.38%)	93 (0.39%)
Age at pancreatic cancer diagnosis (40-50)	474 (1.98%)	417 (1.74%)
Age at pancreatic cancer diagnosis (50-60)	1,626 (6.78%)	1,304 (5.44%)
Age at pancreatic cancer diagnosis (60-70)	3,585 (14.95%)	2,950 (12.30%)
Age at pancreatic cancer diagnosis (70-80)	4,017 (16.75%)	4,076 (16.99%)
Age at pancreatic cancer diagnosis (80-90)	1,925 (8.03%)	2,751 (11.47%)
Age at pancreatic cancer diagnosis (90-100)	148 (0.62%)	490 (2.04%)
Age at pancreatic cancer diagnosis (100-110)		5 (0.02%)

Figure S2. Data augmentation and exclusion experiments.

For each patient in the dataset, the data was augmented by using all continuous partial trajectories of (minimal length >=5 diagnoses) from the beginning of their disease history and ending at different time points, which we call the time of assessment. We used a step function annotation indicating cancer occurrence at different time points (3, 6, 12, 36, 60 months) after the end of each partial trajectory. For the positive cases this provides the opportunity to learn from disease histories with a significant time gap between the time of assessment and the time of cancer occurrence. For example, for a patient, who had abdominal pain a month or two just before the cancer diagnosis, it is of interest to learn which earlier disease codes might have been predictive of cancer occurrence going back at least several months or perhaps years. The latter is also explored by separately re-training of the ML model excluding data from the last three or six months before cancer diagnosis.

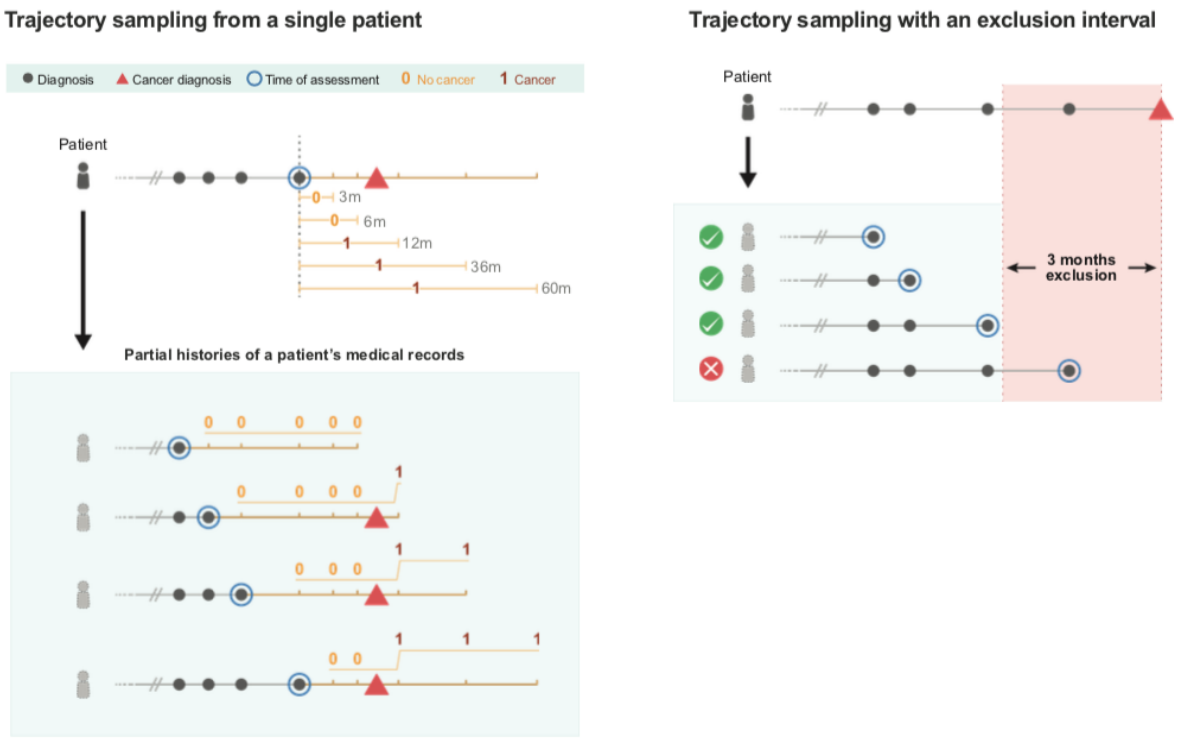


Table S2. Hyperparameter search for machine learning models.

To comprehensively test different types of neural networks and the corresponding hyperparameters, we conducted a large parameter search for each of the network types. The different types of models include simple feed-forward models (LR, MLP) and more complex models that can take the sequential information of disease ordering into consideration (RNN, GRU and Transformer). The hyperparameters of the best performing model are highlighted in bold.

	Type of ML model			
Hyper-parameters	Bag of words	MLP	GRU	**Transformer**
Dropout	0	0,0.1	0,0.1	**0**, 0.1
Weight decay	0.001	0,0.001	0,0.001	0, **0.001**
Only prior knowledge diseases	False, True	False	False	**False**, True
Dimension of hidden layer	-	32, 128, 256	32, 64, 128, 256	32, **256**
Number of hidden layers	-	1, 2	1, 2, 4	**1**, 2, 4
Age input	None	None	None, positional embedding	None, **positional embedding**
Time input	None	None	None, positional embedding	None, **positional embedding**
Number of Heads	-	-	-	8, **16**, 32

Table S3. Performance of exclusion experiments.

A summary of performance of different models trained with different data exclusion intervals for different prediction intervals. In order to estimate the uncertainty of the performance metrics, a 95% confidence interval was computed using 200 resamples (bootstrapping with replacement). The best performance for each scenario is highlighted in bold. Specificity, precision, and recall are for the F1-optimal operational point.

Table S3A. Performance summary (AUROC)

Model		Up to 3 months	Up to 6 months	Up to 12 months	Up to 36 months	Up to 60 months
Bag-of-words	No exclusion	0.794 (0.791-0.797)	0.800 (0.797-0.803)	0.807 (0.805-0.809)	0.807 (0.805-0.809)	0.799 (0.797-0.800)
GRU	No exclusion	0.916 (0.914-0.919)	0.903 (0.900-0.905)	0.883 (0.881-0.885)	0.852 (0.850-0.854)	0.836 (0.834-0.837)
	3 months	-	0.859 (0.854-0.865)	0.852 (0.848-0.856)	0.833 (0.830-0.835)	0.820 (0.818-0.822)
	6 months	-	-	**0.848 (0.845-0.852)**	0.827 (0.825-0.830)	0.815 (0.813-0.816)
MLP	No exclusion	0.876 (0.873-0.878)	0.871 (0.869-0.874)	0.864 (0.862-0.866)	0.845 (0.844-0.848)	0.832 (0.831-0.834)
Transformer	No exclusion	**0.935 (0.932-0.937)**	**0.923 (0.921-0.924)**	**0.908 (0.907-0.910)**	**0.879 (0.877-0.880)**	**0.861 (0.859-0.863)**
	3 months	-	**0.865 (0.860-0.870)**	**0.861 (0.858-0.865)**	**0.842 (0.840-0.845)**	**0.830 (0.828-0.831)**
	6 months	-	-	0.834 (0.829-0.839)	**0.829 (0.827-0.832)**	**0.817 (0.815-0.819)**
Transformer - Prior knowledge disease	No exclusion	0.849 (0.847-0.852)	0.850 (0.847-0.852)	0.850 (0.848-0.852)	0.838 (0.837-0.839)	0.832 (0.830-0.833)

Table S3B. Performance summary (specificity/precision/recall)

Model	Data exclusion	Metric	Up to 3 months	Up to 6 months	Up to 12 months	Up to 36 months	Up to 60 months
Bag-of-words	No exclusion	specificity	98.66% (96.16%-98.83%)	98.05% (95.42%-98.86%)	98.18% (97.59%-98.82%)	95.49% (94.87%-97.81%)	95.09% (94.13%-95.65%)
		precision	0.3% (0.3%-0.4%)	0.4% (0.4%-0.5%)	0.6% (0.6%-0.7%)	0.9% (0.8%-0.9%)	1.0% (0.9%-1.0%)
		recall	5.3% (4.6%-13.8%)	8.0% (4.8%-17.5%)	7.7% (5.3%-10.0%)	16.8% (8.5%-18.8%)	16.5% (14.5%-19.2%)
GRU	No exclusion	specificity	99.95% (99.93%-99.95%)	99.92% (99.89%-99.94%)	99.89% (99.87%-99.92%)	99.82% (99.77%-99.87%)	99.76% (99.74%-99.80%)
		precision	15.1% (13.0%-16.0%)	14.0% (11.8%-15.8%)	13.2% (12.1%-14.8%)	11.5% (10.1%-13.6%)	10.4% (9.9%-11.2%)
		recall	12.7% (12.0%-14.4%)	12.7% (11.3%-14.8%)	12.6% (11.3%-13.6%)	10.8% (9.5%-12.1%)	10.0% (9.2%-10.5%)
	3 months	specificity	-	**99.97% (99.94%-99.97%)**	**99.94% (99.91%-99.95%)**	**99.86% (99.83%-99.89%)**	**99.84% (99.79%-99.86%)**
		precision	-	**2.8% (2.2%-3.6%)**	**5.1% (4.2%-5.9%)**	**5.5% (5.0%-6.3%)**	**5.8% (5.0%-6.4%)**
		recall	-	**4.9% (3.9%-6.7%)**	**6.0% (5.2%-7.5%)**	**6.0% (5.3%-6.7%)**	**5.1% (4.7%-5.8%)**
	6 months	specificity	-	**-**	**99.93% (99.85%-99.95%)**	**99.88% (99.85%-99.93%)**	**99.84% (99.78%-99.85%)**
		precision	-	**-**	**1.7% (1.3%-2.1%)**	**4.3% (3.5%-5.8%)**	**4.4% (3.7%-4.7%)**
		recall	-	**-**	**3.9% (2.9%-6.3%)**	**4.5% (3.6%-5.4%)**	**4.2% (3.9%-4.8%)**
MLP	No exclusion	specificity	99.75% (99.68%-99.82%)	99.73% (99.66%-99.82%)	99.79% (99.66%-99.82%)	99.69% (99.53%-99.73%)	99.54% (99.49%-99.61%)
		precision	2.7% (2.4%-3.0%)	3.3% (3.0%-3.8%)	4.3% (3.6%-4.7%)	4.8% (4.1%-5.3%)	4.5% (4.2%-4.9%)
		recall	8.9% (6.9%-10.8%)	9.0% (7.1%-11.2%)	7.5% (6.6%-9.9%)	7.4% (6.5%-9.4%)	7.8% (7.0%-8.5%)
Transformer	No exclusion	specificity	**99.95% (99.92%-99.96%)**	**99.93% (99.91%-99.93%)**	**99.91% (99.90%-99.93%)**	**99.88% (99.87%-99.90%)**	**99.87% (99.84%-99.88%)**
		precision	**18.6% (15.4%-22.5%)**	**18.8% (16.8%-20.1%)**	**19.4% (18.2%-21.9%)**	**18.2% (17.1%-19.7%)**	**17.8% (16.0%-18.7%)**
		recall	**16.5% (14.2%-19.4%)**	**17.0% (16.1%-18.6%)**	**15.6% (14.5%-16.5%)**	**12.4% (11.5%-12.9%)**	**10.3% (9.8%-10.9%)**
	3 months	specificity	**-**	99.92% (99.91%-99.98%)	99.92% (99.92%-99.93%)	99.87% (99.86%-99.91%)	99.63% (99.56%-99.64%)
		precision	-	1.7% (1.4%-3.2%)	4.3% (3.9%-4.7%)	5.4% (4.9%-6.5%)	2.7% (2.5%-2.9%)
		recall	-	5.9% (2.3%-7.1%)	6.5% (5.9%-7.0%)	5.2% (4.5%-5.6%)	5.3% (4.9%-5.9%)
	6 months	specificity	-	-	99.41% (98.15%-99.41%)	99.51% (99.47%-99.52%)	99.35% (95.83%-99.40%)
		precision	-	-	0.2% (0.1%-0.2%)	0.7% (0.7%-0.8%)	0.8% (0.7%-0.9%)
		recall	-	-	3.4% (2.7%-7.9%)	3.2% (2.9%-3.4%)	3.2% (2.8%-15.8%)
Transformer (n=23 prior knowledge disease only)	No exclusion	specificity	99.96% (99.92%-99.97%)	99.92% (99.91%-99.93%)	99.91% (99.91%-99.92%)	99.87% (99.77%-99.88%)	99.79% (99.73%-99.88%)
		precision	11.5% (7.3%-12.6%)	9.2% (8.6%-9.9%)	10.2% (9.7%-10.8%)	3.6% (2.5%-3.9%)	2.8% (2.4%-4.0%)
		recall	6.9% (6.3%-9.8%)	9.2% (8.6%-9.8%)	8.2% (7.8%-8.7%)	2.5% (2.3%-3.1%)	2.5% (1.9%-2.8%)

Table S4. Prior knowledge disease codes.

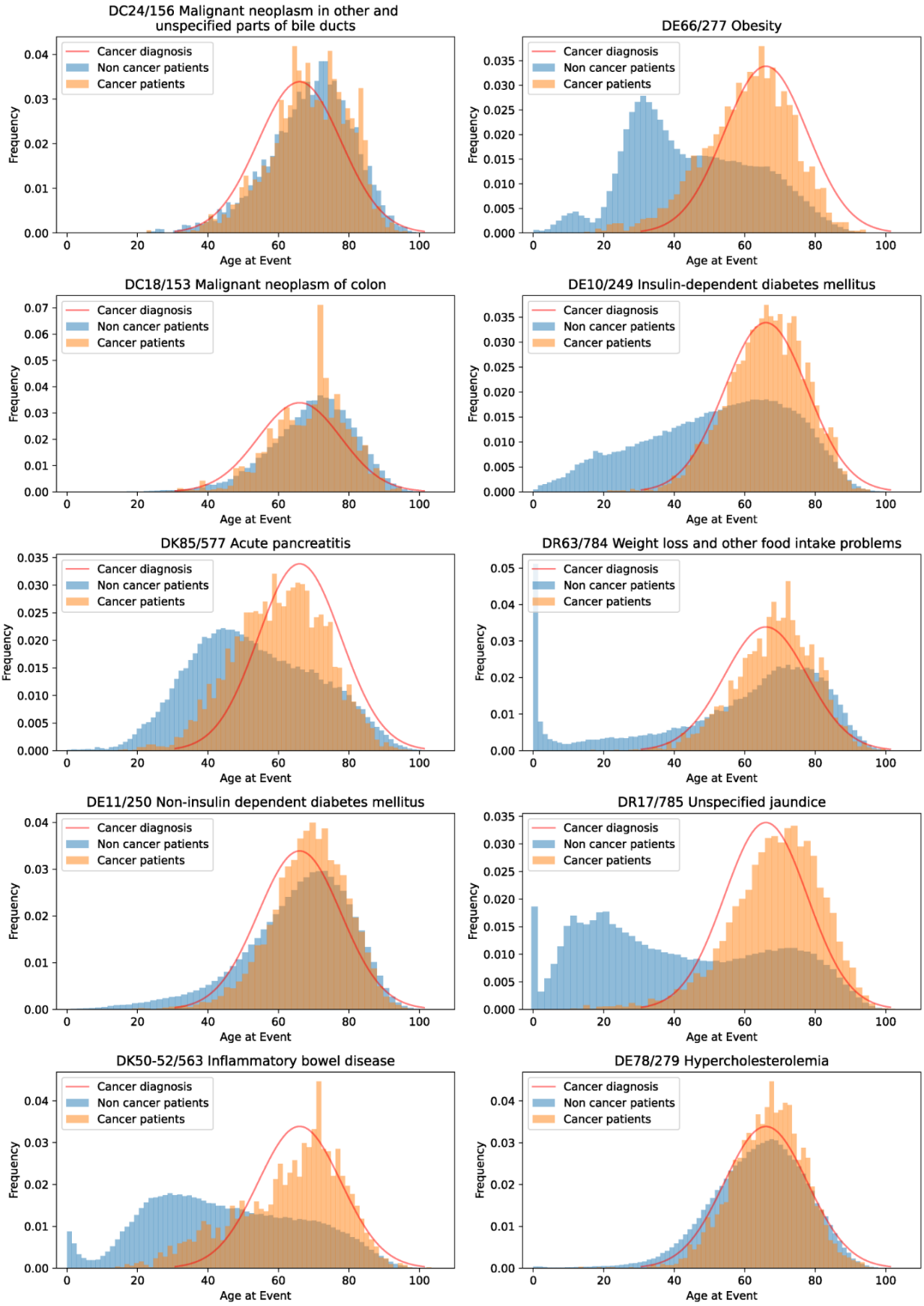
A subset of 23 diseases that have been considered as risk factors for pancreatic cancer (Yuan et al. 2020; Klein 2021). Indeed, most of these make a significant contribution to the ML prediction of cancer occurrence (**Figure 4**). We also used this subset of diseases in training for comparison, called prior-knowledge diseases.

ICD codes	Diseases
C18	Malignant neoplasm of colon
C34	Malignant neoplasm of bronchus and lung
C50	Malignant neoplasm of breast.
C61	Malignant neoplasm of prostate
E10, E11	Type I/II diabetes mellitus
E66	Obesity
E78	High Cholestrol
E84	Cystic fibrosis
F32	Depression
I10	Hypertension
I82	Venous embolism and thrombosis
K05	Periodontal disease
K21	GERD
K27	Peptic Ulcer Disease
K50, K51, K52	Inflammatory bowel disease
K85	Acute Pancreatitis
K86	Chronic Pancreatitis
R17	Jaundice
R63	Weight loss
Z92	Personal history of medical treatment

Figure S3. Distribution of disease codes as a function of age in the database.

Distribution of disease codes for a representative subset of diseases reported to contribute to the risk of pancreatic cancer, as a fraction of all pancreatic cancer patients (orange) and all non-cancer patients (blue). The similarity of the distributions for some of these diseases with the distribution of occurrence of pancreatic cancer (red line) is consistent with either a direct or indirect contribution to cancer risk - but not taken as evidence in this work. The disease codes are ICD-10/ICD-8.

Disease distribution



Supplementary Results

Draft economic considerations for the design of clinical screening trial

We propose a toy estimate of a practical scenario for a screening trial, taking into account typically available real-world data, the accuracy of prediction on such data, the estimated cost of a screening trial, the cost of clinical screening methods and the overall potential benefit of treatment.

The detailed design of a screening program, to be explored in clinical trials, depends on the organization of a particular health care system. In a 'walk in' scenario, in approximate analogy to colonoscopic screening for colorectal cancer, patients older than, e.g., age 50 would be invited for assessment of their risk by the prediction tool every 5 years and, if identified as high-risk, offered extensive clinical testing. In a 'national system' scenario, possible in centralized health systems with location-independent centralized aggregation of electronic health records, risk assessment could be done on an ongoing basis, possibly for each patient whenever a new disease event occurs. If a high-risk prediction is triggered, the responsible physician would receive an alert. With this diversity of scenarios, it is reasonable to propose clinical screening trials in several countries tailored to their particular health system.

To illustrate the economic benefits of such a screening and to stimulate discussion regarding the optimization of trial design, we have made a first-order-estimate for a clinical screening trial of 10,000 people using the best model (the transformer model). For simplicity, we have made no assumptions regarding age distribution. Here is a simple economic model.

$$\begin{aligned}\text{Net Benefit} = {} & \text{Average benefit for each correctly identified cancer patient} * TP \\ & - \text{Monitoring expense for each high-risk patient} * P \\ & - \text{Basic cost per enrollee} * N\end{aligned}$$

where the screening cohort is N=10,000 and TP is the number of true positives, i.e., the number of correctly identified high-risk patients, and P is the number of actual positive patients, which we estimated using cancer incidence of the DNPR dataset. In our cost-benefit estimate, we arbitrarily set the screening trial cost at $200 per enrollee, the additional monitoring expense for a patient predicted at high risk by screening at $10,000 and the extra cost saved for advanced treatment for each monitored patient at $200,000, averaged over those in which cancer is detected (savings in excess of $200,000) and those in which it is not detected (no savings).

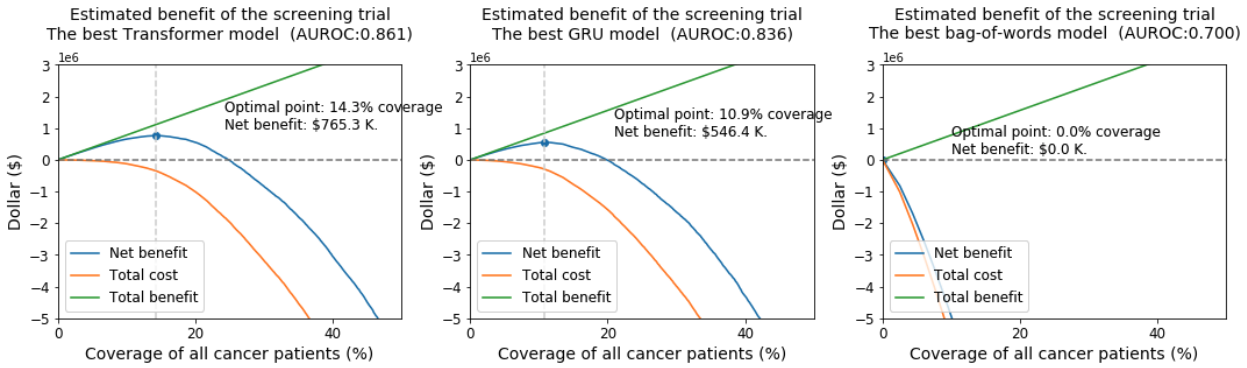


Figure S4. An estimate of financial benefits for different models. We analyzed each possible operational point and calculated the corresponding cost and benefit, respectively. We plotted the net benefits as a function of coverage of cancer patients, i.e. recall or sensitivity. Covering more cancer patients apparently leads to a larger total benefit, but the total cost also scales up non-linearly. The optimal point is picked for the maximal net benefit.

An optimal decision threshold has to balance the cost of assessment and testing against the potential financial benefit for reducing treatment cost. Using this simplified model, we estimated the net benefits of different models with all possible operational points. Such a screening trial for 10,000 people would have $760,000 net benefit by choosing the balance between true and false positives such that the net benefit is optimal. This corresponds to a precision of 14.0% and a specificity of 99.7%. In contrast, a less good model GRU would have $540K net benefits but a bag-of-words model (baseline) would have no net benefits for any operational point because of the low incidence of pancreatic cancer.

The proposed concrete design of a screening trial is intended to guide the debate and ultimate decisions regarding implementation with clinicians and healthcare professionals. However, this hypothetical calculation is based on arbitrary numbers and does not reflect real-world cost analysis. Nor does this economic model reflect the non-monetary benefits to patients' quality of life, which should be the dominant factor in the design of trials and early intervention programs. In a real-world scenario, clinicians and payers in a particular health system have the opportunity to optimize the design of such screening trials with realistic cost-benefit parameters, as well as consideration of communication ethics and the non-financial aspects of patient benefit.

A key challenge for future realistic economic estimates is the mapping between ICD (diagnosis) codes to CPT (billing) codes that are used for expense calculations and reimbursements. In addition, in the US, there is usually a lot of geographical variability in reimbursement even for the same CPT/billing codes.