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], Piotr J Chmura[1], Chen Yuan[2,3], Jihye Kim[6], Renato Umeton[3,6,7,8], Gregory Antell[3], Alexander Chowdhury[3], Alexandra Franz[2,3,4], Lauren Brais[3], Elizabeth Andrews[3], Debora S. Marks[2], Aviv Regev[4,9], Peter Kraft[6], Brian M. Wolpin[2,3,10], Michael Rosenthal[2,3,10], Søren Brunak[1,5,*], 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] Massachusetts Institute of Technology, Cambridge, MA, USA
[8] Weill Cornell Medicine, New York City, NY, USA
[9] Currently: Genentech, South San Francisco, CA, USA
[10] Brigham and Women's Hospital, Boston, USA

[‡] joint first authors: davide.placido@cpr.ku.dk, boyuan@g.harvard.edu, jessica.hu@cpr.ku.dk
[*] joint senior authors: soren.brunak@cpr.ku.dk, chris_sander@hms.harvard.edu

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 addressed by identifying high-risk populations. Here we apply artificial intelligence (AI) methods to a dataset of more than 6 million patient records with 24,000 pancreatic cancer cases in the Danish National Patient Registry (Denmark) and, for comparison, a dataset of one million records with 4,000 pancreatic cancer cases in the Mass General Brigham Healthcare System (Boston, US). 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 and test the ability to predict cancer occurrence in time intervals of 3 to 60 months after risk assessment. We extract from the AI machine an estimate of the contribution to prediction of individual disease features. For cancer occurrence within 36 months, the performance of the best model (AUROC=0.88), trained and tested on disease trajectories in the Danish dataset, substantially exceeds that of a model without time information, even when disease events within a 3 month window before cancer diagnosis are excluded from training (AUROC[3m]=0.84). Independent training and testing on the Boston dataset reaches comparable performance (AUROC=0.87, AUROC[3m]=0.80), while cross-application of the Danish deep learning model on the Boston dataset has lower accuracy (AUROC=0.78, AUROC[3m]=0.70), indicating a requirement of independent training in health systems with different coding practices. 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 future screening trials for high-risk patients, e.g., to serial imaging or blood-based biomarkers to facilitate earlier cancer detection. AI on real-world clinical records has the potential to shift focus from treatment of late-stage 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). Early diagnosis of pancreatic cancer is a key challenge, as the disease is typically detected at a late stage. Approximately 80% of pancreatic cancer patients are diagnosed with locally advanced or distant metastatic disease, when long-term survival is extremely uncommon (2-9% of patients at 5-years) (McGuigan et al. 2018). 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) achieve 5-year overall survival [National Cancer Institute, USA, (Blackford et al. 2020)]. Thus, a better understanding of the risk factors for pancreatic cancer and detection at early stages has great potential to improve patient survival and reduce overall mortality from this aggressive malignancy.

[[Known risk factors of limited use]]

The incidence rate of pancreatic cancer is substantially lower compared with 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, few high-penetrance risk factors are known for pancreatic cancer impeding early diagnosis of this disease. Risk of pancreatic cancer has been assessed for many years based on family history, behavioral and clinical risk factors and, more recently, circulating biomarkers and genetic predisposition (Amundadottir et al. 2009; Petersen et al. 2010; D. Li et al. 2012; Wolpin et al. 2014; Klein et al. 2018; Kim et al. 2020). 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 less than 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 identify high-risk patients, which will 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), predicting the risk of potentially acute disease progression, such as in kidney injury (Tomašev et al. 2019) and the likelihood of a next diagnosis based on past EHR sequences, in analogy to natural language processing (Y. Li et al. 2020).

[[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; X. Li et al. 2020), and from an electronic health record (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 codes, not the time sequence of disease states in a patient trajectory - in analogy to the 'bag-of-words' models in natural language processing that ignore the actual sequence of words. Previous studies had used the Danish health registries to generate population-wide disease trajectories, but in a non-predictive manner (Hu et al. 2019; Jensen et al. 2014).

[[Advance here - better data and better ML]]

Here we exploit the power of 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 initially carried out using the Danish National Patient Registry (DNPR) and data which covers 41 years (1977 to 2018) of clinical records for 8.6 million patients, 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 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 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 of cancer risk ideally should take into account the probability of the disease occurring within a shorter or longer time frame. For this reason, we designed the prediction method to predict not only whether cancer is likely to occur, but also to provide risk assessment in incremental time intervals following the assessment, where time of assessment is defined as the day on which the risk prediction is performed based on the history of clinical records of the particular patient. We also analyzed which diagnoses in a patient's history of disease codes are most informative of cancer risk - not as isolated factors but always in the context of the person's complete history of disease codes. 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 and success rate of clinical screening methods and the overall potential benefit of early treatment (**Supplementary Text, Figure S5**).

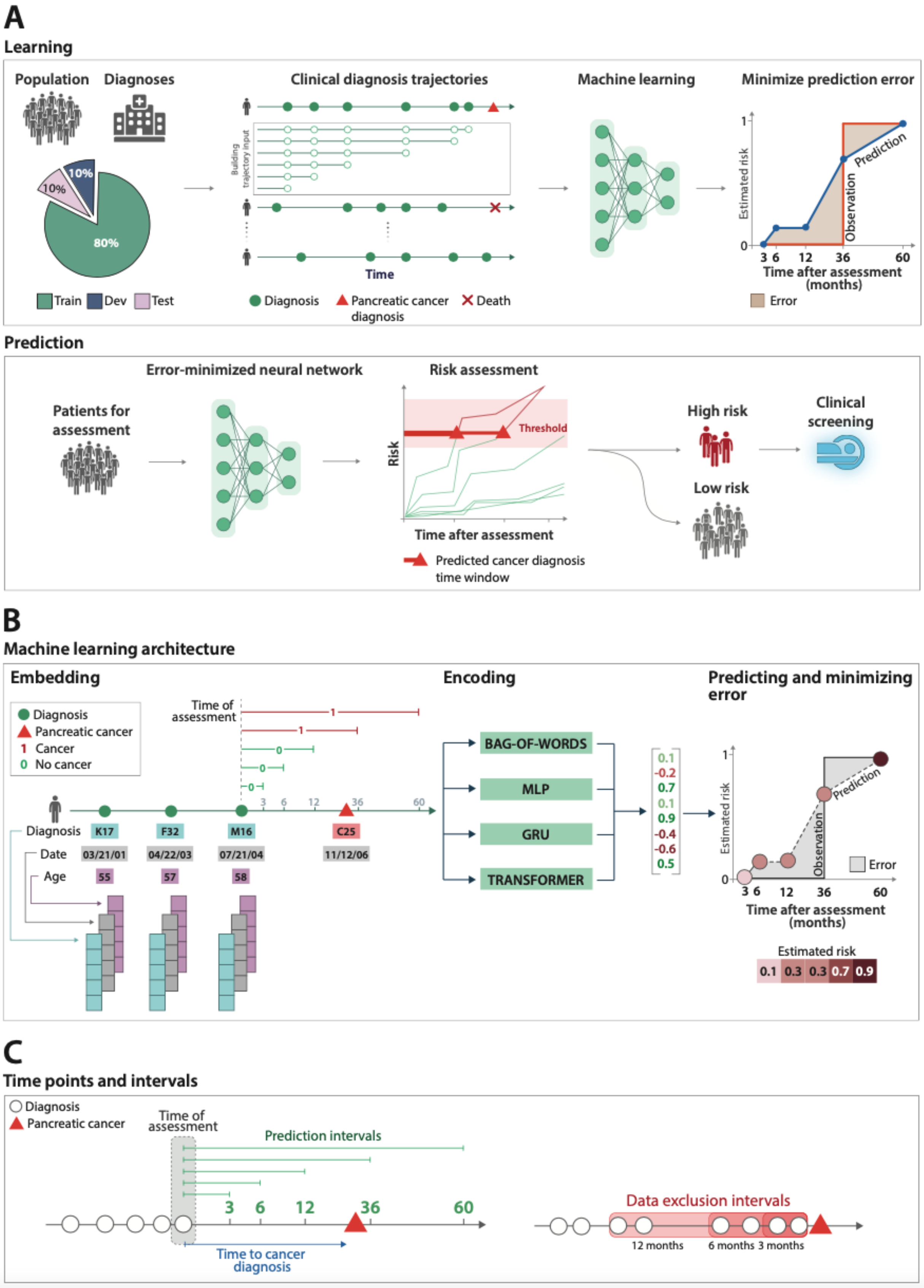

A
Learning
Population
Diagnoses
10%
10%
80%
Train
Dev
Test
Clinical diagnosis trajectories
Building trajectory input
Time
Diagnosis
Pancreatic cancer diagnosis
Death
Machine learning
Minimize prediction error
1
0
Estimated risk
Prediction
Observation
3 6 12 36 60
Time after assessment (months)
Error
Prediction
Patients for assessment
Error-minimized neural network
Risk assessment
Risk
Threshold
Time after assessment
Predicted cancer diagnosis time window
High risk
Low risk
Clinical screening
B
Machine learning architecture
Embedding
Diagnosis
Pancreatic cancer
1 Cancer
0 No cancer
Time of assessment
1
1
0
0
0
3 6 12 36 60
Diagnosis
K17
F32
M16
C25
Date
03/21/01
04/22/03
07/21/04
11/12/06
Age
55
57
58
Encoding
BAG-OF-WORDS
MLP
GRU
TRANSFORMER
0.1
-0.2
0.7
0.1
0.9
-0.4
-0.6
0.5
Predicting and minimizing error
1
0
Estimated risk
Prediction
Observation
Error
3 6 12 36 60
Time after assessment (months)
Estimated risk
0.1 0.3 0.3 0.7 0.9
C
Time points and intervals
Diagnosis
Pancreatic cancer
Time of assessment
Prediction intervals
3 6 12 36 60
Time to cancer diagnosis
Data exclusion intervals
12 months
6 months
3 months

Figure 1. Training and prediction of pancreatic cancer risk from disease trajectories. (**A**) Learning: The general machine learning workflow starts with partitioning the data into training set (Train), development set (Dev) and test set (Test). The trajectories for training input are generated by sampling continuous subsequences of diagnoses for each patient's diagnosis history, each starting with the first record but with different end points. The training and development sets are used for training machine learning models 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. Prediction: A model's ability to generalize is evaluated using the withheld 'test' set. The prediction model, depending on the prediction threshold selected from among possible operational points, discriminates between patients at higher and lower risk of pancreatic cancer. The risk model can guide the development of clinical screening initiatives. (**B**) The model trained with real-world clinical data has three steps: embedding, encoding and prediction. The embedding machine transforms categorical disease codes and time stamps of these disease codes into a latent space. The encoding machine extracts information from a disease history and summarizes each sequence in a characteristic fingerprint (vertical vector). 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. (**C**) Terminology for time points and intervals. The end point of a disease trajectory is the time of assessment. From the time of assessment, cancer risk is assessed within 3, 6, 12, 36 and 60 months. To test the influence of close-to-cancer ICD codes on the prediction of cancer occurrence, exclusion intervals are used to remove diagnoses in the last 3, 6 and 12 months before cancer diagnosis.

Results

Datasets

[[Dataset of disease trajectories: Denmark]]

We used data from the DNPR, where all inpatient admissions to Danish hospitals have been recorded since 1977, and outpatients and emergency visits have been included since 1994. 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 on average 26.7 diagnosis codes per patient. For training we used trajectories of ICD (International Classification of Diseases) codes 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 with discontinuous or very short trajectories (<5 events in total), ending up with 6.2 million patients (**Figure S1A**). The case cohort includes 23,985 pancreatic cancer (PC) cases with cancer occurring at a median age of 70 years (mean age of 65±11 years [men] and 67±12 years [women]) (**Figure 2**, **Table S1**).

[[Dataset of disease trajectories: Boston, US]]

For external validation, we used clinical records from the Mass General Brigham (MGB) hospital system in the US via their Research Patient Data Registry (RPDR), a centralized, access-controlled clinical data warehouse for use in research. As in the Danish dataset, we also used explicit longitudinal records from MGB, i.e., trajectories of ICD codes with explicit time stamps. We used data from 1982 to 2020 and filtered out patients with less than six months of contact or less than five recorded diagnosis codes (**Figure S1b**). The selected dataset (**Figure 2**) has 1.0 million patients with 3,904 pancreatic cancer patients (Methods). The median length of disease trajectories is 13 years and the median number of disease codes per patient is 168; the latter is much higher than in the Danish dataset (**Figure 2C**). The median age of pancreatic cancer diagnosis is 60 years, lower than in the Danish dataset (**Figure 2C**). These statistics might reflect the differences between the health care systems in the two locations, such as referral, billing and documentation practices.

Model architecture

[[Network architecture/layers]]

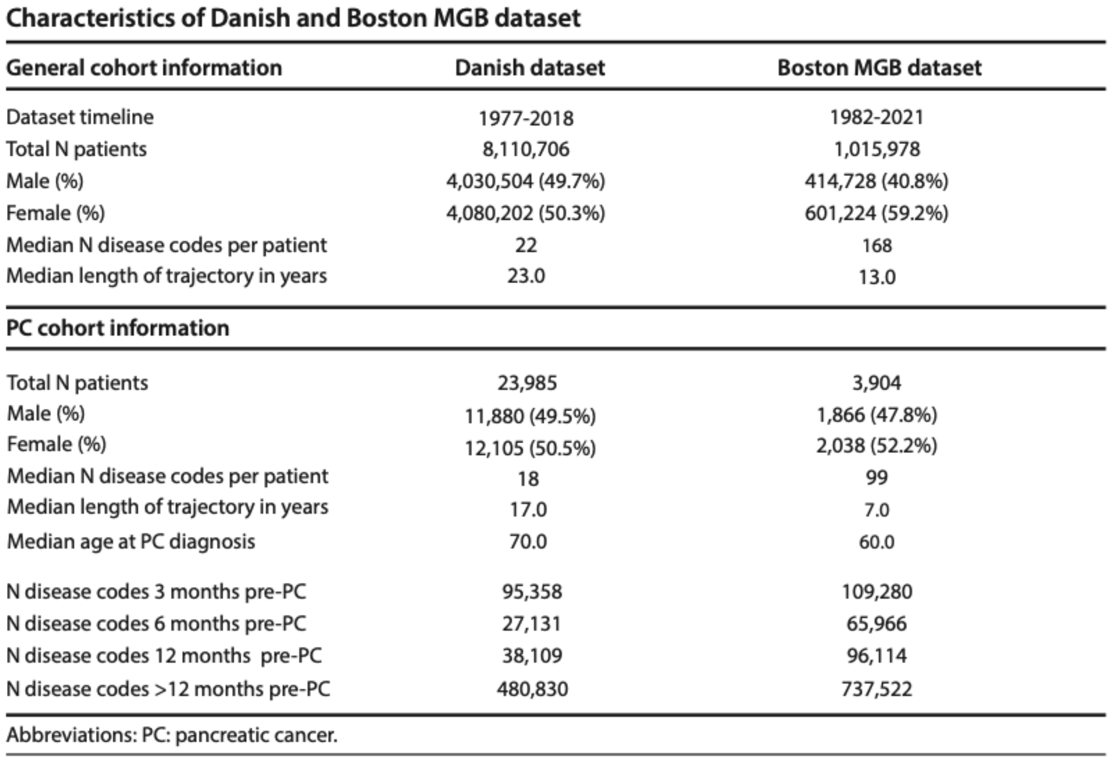
The machine learning model for predicting cancer risk from disease trajectories consists of four parts: (1) **input** data for each event in a trajectory (disease code and time stamps), (2) **embedding** the event features onto real number vectors, (3) **encoding** the trajectories in a lower-dimensional latent space, and (4) **predicting** time-dependent cancer risk. (1) **Input**: In order to best exploit the longitudinality of the EHR data and provide an opportunity to discover early indicators of cancer risk, all contiguous subsequences of diagnoses from a patient's history were sampled, starting with the earliest record and increasing gaps between the end of the trajectory and cancer occurrence for positive cases (Methods). The partial trajectories provide information in support of prediction for different time spans between risk assessment and cancer occurrence, rather than just binary prediction that cancer will occur at any time after assessment. (2) **Embedding**: Each item in a disease trajectory is an event denoted with one of the >2,000 ICD disease codes. To extract informative features from such high-dimensional input, the ML process is designed to embed the categorical input vectors into a continuous, lower-dimensional space. Temporal information, i.e. diagnosis dates and age at diagnosis are also embedded (see Methods). The mapping of the input to the embedding layer is 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). 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 elements of the event vectors. (4) **Predicting:** The embedding and encoding network 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 estimate the risk of cancer within distinct prediction intervals ending a few months or 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, for a given disease trajectory. If and when the risk score exceeds a prediction 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, 120 months after the time t_a of risk assessment; or not to occur at all in 120 months.

[[Scanning hyperparameters for each model type]]

To comprehensively test the performance of different types of ML models, we first conducted an extensive search over hyperparameters and selected the best set of hyperparameters for each model, and then selected the best model type. The model types included transformer, GRU, a multilayer perceptron and bag-of-words. Each model was tested on specific hyperparameter configurations (**Table S2**). To avoid overfitting and to test generalizability of model predictions, we partitioned patient records 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. Subsequently, 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 records.



Characteristics of Danish and Boston MGB dataset

General cohort information	Danish dataset	Boston MGB dataset
Dataset timeline	1977-2018	1982-2021
Total N patients	8,110,706	1,015,978
Male (%)	4,030,504 (49.7%)	414,728 (40.8%)
Female (%)	4,080,202 (50.3%)	601,224 (59.2%)
Median N disease codes per patient	22	168
Median length of trajectory in years	23.0	13.0
PC cohort information		
Total N patients	23,985	3,904
Male (%)	11,880 (49.5%)	1,866 (47.8%)
Female (%)	12,105 (50.5%)	2,038 (52.2%)
Median N disease codes per patient	18	99
Median length of trajectory in years	17.0	7.0
Median age at PC diagnosis	70.0	60.0
N disease codes 3 months pre-PC	95,358	109,280
N disease codes 6 months pre-PC	27,131	65,966
N disease codes 12 months pre-PC	38,109	96,114
N disease codes >12 months pre-PC	480,830	737,522

Abbreviations: PC: pancreatic cancer.

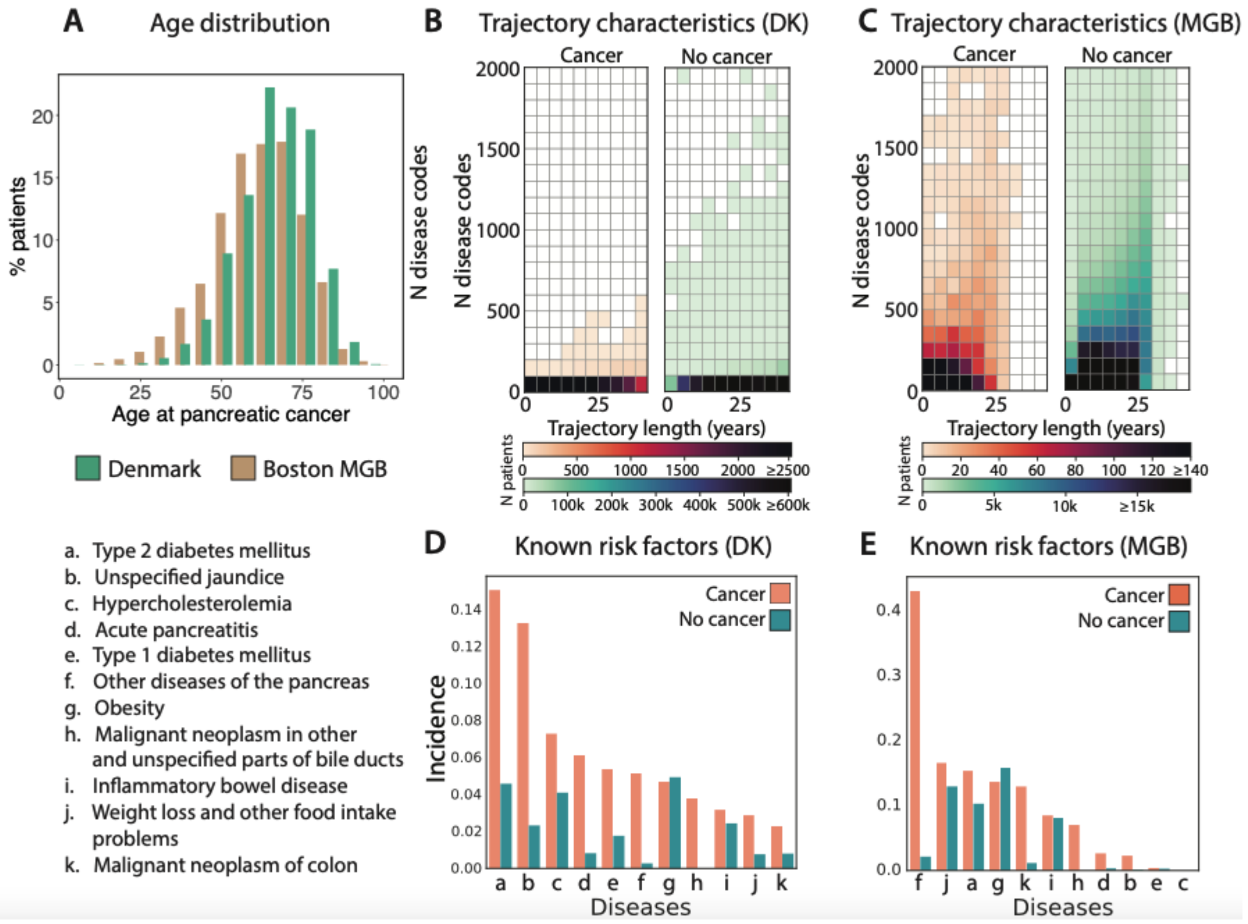


Figure 2. Danish (DK) and Boston (MGB) patient registries used for machine learning of cancer risk. The Danish DNPR database (DK) of clinical records covers over 8 million people for up to 41 years. The Boston MGB (RPDR) database covers only 1 million people with long term data, but has a higher density of disease codes per time interval (**Figure S4**). (**A**) The incidence of pancreatic cancer peaks past the age of 50 years in both datasets. (**B,C**) The machine learning process has to cope with very different distributions of disease trajectories in terms of length of trajectories and density of the number of disease codes. The Danish (DK) dataset has a longer median length of disease trajectories, but lower median number of disease codes per patient compared to the MGB dataset. (**D,E**) An intuitive indication of the association of individual disease codes with subsequent diagnosis of pancreatic cancer is given by the relative incidence of known risk factors in cancer vs. non-cancer patients in the DK (**D**) and MGB (**E**) datasets, counting whether a disease code occurred at least once in a patient's history and excluding events at or after cancer diagnosis.

Evaluation of model performance

[[Picking a best model - DK]]

We evaluated the different models using the precision-recall curve (PRC) and then report performance numbers at the operating point on the receiver-operating curve (ROC) that maximizes the F1 score (**Figure 3**), which strikes a balance between precision (positive predictive value) and recall (sensitivity). In the final performance evaluation of different types of ML models on the test set, the models which explicitly use and encode the time 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 (AUROC=0.879 [0.877-0.880]), followed by GRU (AUROC=0.852 [0.850-0.854]). The bag-of-words model that ignores the time information along disease trajectories performed significantly less well (AUROC=0.807 [0.805-0.809]). At the chosen operating point that maximizes the F1 score (Methods), the model has a precision of 18.1% (17.1-19.9), a recall of 12.3% (11.7-12.9) and a specificity of 99.88% (99.87-99.90). In order to gain a better intuition regarding the impact of applying the model in a real case scenario, we also report the odds ratio (OR) of cancer patients in the high-risk group for the deep learning models. The OR is defined as the odds of getting pancreatic cancer for a high risk patient divided by the odds of getting pancreatic cancer for a low risk patient (**Table S6**). The odds ratio for the Transformer model is 47.5 for 20% recall and 159.0 for 10% recall.

[[Comparison with previous models]]

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; X. Li et al. 2020). These previous studies had encouraging results, but neither used the time sequence of disease histories nor memory or attention mechanisms in the neural network to extract time-sequential longitudinal features. For comparison we implemented analogous approaches, a bag-of-words model and a multilayer perceptron (MLP) model. We evaluated the non-time-sequential models on the DNPR dataset, and the performance for predicting cancer occurrence within 36 months was AUROC=0.807 (0.805-0.809) for the bag-of-words model and

0.845 (0.843-0.847) for the MLP model. Compared to the time-sequential models, e.g., Transformer, which has an odds ratio of 159.0 at 10% recall, the bag-of-words/MLP models have a much lower odds ratio of 4.0/21.0, respectively, also at 10% recall. In other words, when taking time series into account, the odds ratio increases by nearly a factor of 40 (**Table S3**).

[[Prediction for prediction time intervals]]

It is also of clinical interest to consider risk of cancer over different time intervals. The ML models in this work yield 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 within longer rather than shorter time intervals (**Figure 4A&C**). Indeed, prediction performance for the best model decreases from an AUROC of 0.908 (0.906-0.911) for cancer occurrence within 12 months to an AUROC of 0.879 (0.877-0.880) 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).

[[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 1C**). As expected, e.g. when training with data exclusion, the performance decreased to AUROC of 0.862 (0.857-0.866) for 3 months exclusion and a AUROC of 0.834 (0.830-0.838) for 6 months exclusion - both for prediction of cancer occurence within 12 months (**Table S3A**).

[[Information contribution as a function of time gap between of assessment and cancer occurrence]]

The exclusion of trajectories ending very close to pancreatic cancer removes the influence of disease codes that represent symptoms of pancreatic cancer or are otherwise easily attributable to pancreatic cancer. However, data exclusion of such late events alone does not quantify the influence of longer term risk factors on prediction. In an attempt to estimate the performance of the model when possible peri-diagnostic codes are excluded, we report the recall rate of prediction as a function of the time-to-cancer, defined as the time between the end of disease trajectory and the occurrence of cancer (**Figure 4A, C**). As expected, recall levels decrease with time-to-cancer, from 8% for cancer occurring about 1 year after assessment to a recall of 4% for cancer occurring about 3 years after assessment - for both the models trained with and without 3 months data exclusion. This suggests that the model not only learns from symptoms very close to pancreatic cancer but also from longer disease history, albeit at lower accuracy.

[[Performance by cross-application of a trained model to a different dataset]]

In order to assess the predictive performance of the model in other health care systems, we applied the best machine learning model trained on the Danish DNPR to disease trajectories of patients in the Boston MGB dataset, without any adaptation except for mapping the ICD codes from one system to the other. Prediction performance for cancer occurrence within 3 years declined significantly from an AUROC of 0.879 (0.877-0.880), for a Denmark-trained transformer model applied to Danish DNPR patient data (test set), to an AUROC of 0.776 (0.773-0.778), for the same model applied to Boston MGB patient data (**Figure 3H**). Cross-application required mapping the ICD codes used in Denmark (ICD-10 and ICD-8 codes from The Danish Medical Classification System; Sundhedsvæsenets Klassifikations System (SKS)) to the ICD-10-CM and ICD-9-CM codes used in the Boston MGB system. The most striking difference between the two systems is the shorter and more dense disease history in the Boston MGB trajectories compared to the Danish ones (**Figure 2B-C**). These differences plausibly contribute to the lower performance when cross-applying the machine learning model trained in one health system to another. We conclude that independent training is indicated to achieve good performance in a very different dataset.

[[Model performance by independent training on a different dataset]]

Motivated by the decrease in performance when testing the Denmark-derived model on the Boston MGB dataset, we trained and evaluated the model on the Boston MGB dataset from scratch. For the independently trained model, the performance is much higher than in cross-application, with a test-set AUROC of 0.869 (0.867-0.870) for cancer occurring within 36 months. At the operating point maximizing F1 score, the model has a precision of 19.4% (19.1%-19.7%), a recall of 31.0% (30.4%-31.5%) and a specificity of 99.51% (99.50%-99.52%). At 20% recall, the odds ratio for the high risk group for independent training is 112 compared to 7.6 for cross-application. Similar to the models trained independently on the Danish DNPR, the GRU and transformer models performed much better than the model without temporal information (bag-of-words).

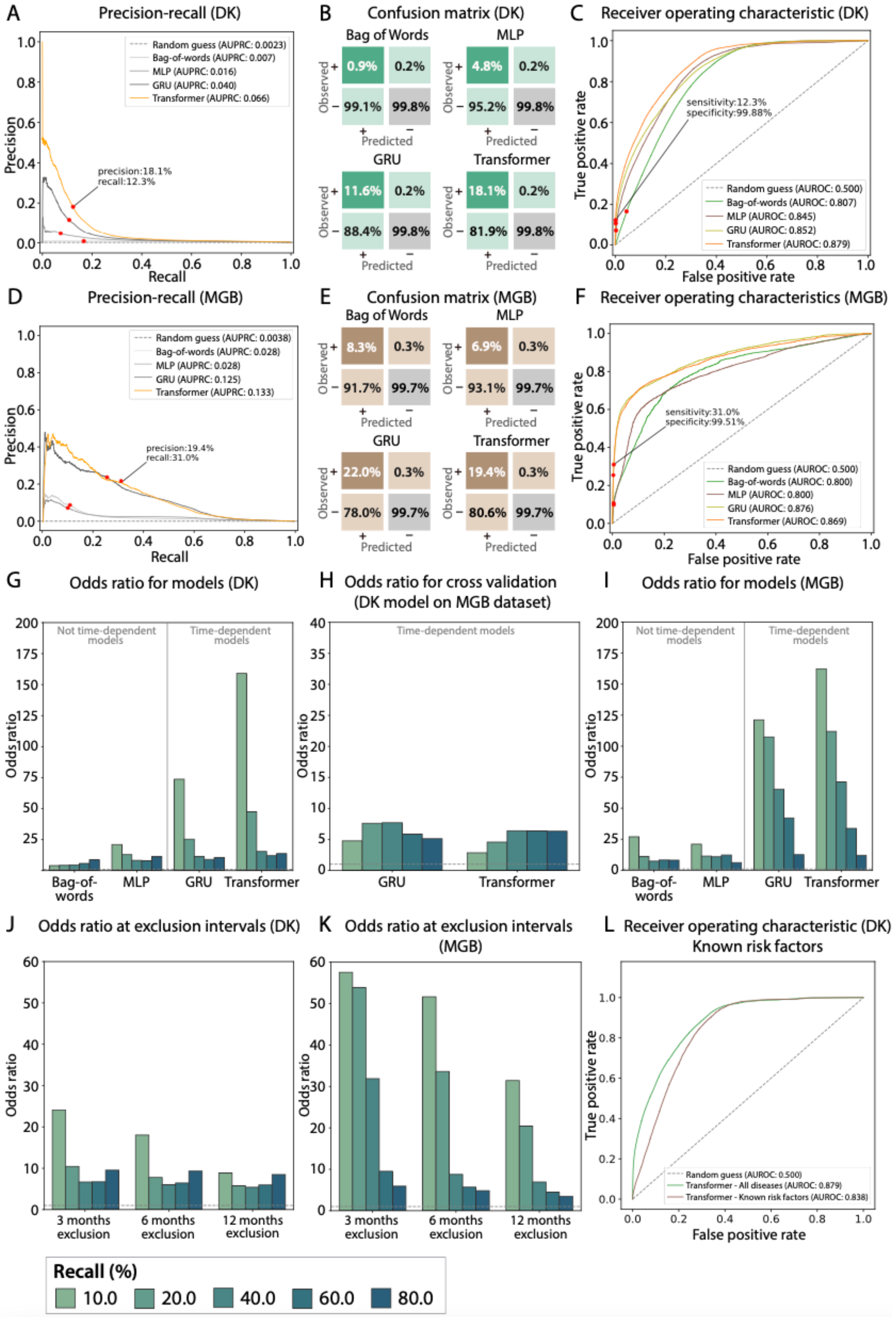

A
Precision-recall (DK)
Random guess (AUPRC: 0.0023)
Bag-of-words (AUPRC: 0.007)
MLP (AUPRC: 0.016)
GRU (AUPRC: 0.040)
Transformer (AUPRC: 0.066)
precision:18.1%
recall:12.3%
Precision
Recall
B
Confusion matrix (DK)
Bag of Words
MLP
GRU
Transformer
Observed
Predicted
0.9% 0.2% 99.1% 99.8%
4.8% 0.2% 95.2% 99.8%
11.6% 0.2% 88.4% 99.8%
18.1% 0.2% 81.9% 99.8%
C
Receiver operating characteristic (DK)
sensitivity:12.3%
specificity:99.88%
Random guess (AUROC: 0.500)
Bag-of-words (AUROC: 0.807)
MLP (AUROC: 0.845)
GRU (AUROC: 0.852)
Transformer (AUROC: 0.879)
True positive rate
False positive rate
D
Precision-recall (MGB)
Random guess (AUPRC: 0.0038)
Bag-of-words (AUPRC: 0.028)
MLP (AUPRC: 0.028)
GRU (AUPRC: 0.125)
Transformer (AUPRC: 0.133)
precision:19.4%
recall:31.0%
Precision
Recall
E
Confusion matrix (MGB)
Bag of Words
MLP
GRU
Transformer
Observed
Predicted
8.3% 0.3% 91.7% 99.7%
6.9% 0.3% 93.1% 99.7%
22.0% 0.3% 78.0% 99.7%
19.4% 0.3% 80.6% 99.7%
F
Receiver operating characteristics (MGB)
sensitivity:31.0%
specificity:99.51%
Random guess (AUROC: 0.500)
Bag-of-words (AUROC: 0.800)
MLP (AUROC: 0.800)
GRU (AUROC: 0.876)
Transformer (AUROC: 0.869)
True positive rate
False positive rate
G
Odds ratio for models (DK)
Not time-dependent models
Time-dependent models
Odds ratio
Bag-of-words
MLP
GRU
Transformer
H
Odds ratio for cross validation (DK model on MGB dataset)
Time-dependent models
Odds ratio
GRU
Transformer
I
Odds ratio for models (MGB)
Not time-dependent models
Time-dependent models
Odds ratio
Bag-of-words
MLP
GRU
Transformer
J
Odds ratio at exclusion intervals (DK)
Odds ratio
3 months exclusion
6 months exclusion
12 months exclusion
K
Odds ratio at exclusion intervals (MGB)
Odds ratio
3 months exclusion
6 months exclusion
12 months exclusion
L
Receiver operating characteristic (DK)
Known risk factors
Random guess (AUROC: 0.500)
Transformer - All diseases (AUROC: 0.879)
Transformer - Known risk factors (AUROC: 0.838)
True positive rate
False positive rate
Recall (%)
10.0
20.0
40.0
60.0
80.0

Figure 3. Performance of the machine learning (ML) models in predicting pancreatic cancer occurrence. Performance of different ML models for prediction of cancer occurrence within 36 months for the Danish DNPR (DK) dataset **(A,B,C)** and Boston Mass General Brigham (MGB) dataset **(D,E,F)**. **(A,D)** Precision-recall curves (PRC): precision (true positives as a fraction of predicted positives) against recall (true positives as a fraction of observed positives) for different models, at different prediction thresholds along the curve. One way to choose an operational point (F1 point) is to balance precision and recall by optimizing the F1 score (red point; Methods). **(B,E)** Confusion matrix for each model, at the F1 point, with the fraction of true positives, true negatives, false positives and false negatives, normalized by column. (**C,F**) Receiver operating characteristic curves (ROC): true positive rate TPR (recall, sensitivity) against false positive rate FPR (false negatives as a fraction of observed negatives = (1-specificity)), at different prediction thresholds along the curve. A random prediction has very low precision for all values of recall (horizontal dotted line in **A** and **D**; AUPRC=incidence=0.004) and equal TPR and FPR (diagonal line in **C** and **F**; AUROC=0.5). The Transformer is the best performing model for 36-month prediction of cancer occurrence for nearly all operational points (**A,C,D,F**). Odds ratios are defined as the odds of getting pancreatic cancer for a high risk patient divided by the odds of getting pancreatic cancer for a low risk patient (**G-K**). Odds ratios for the different ML models for (**G**) the Danish models applied to the Danish dataset, (**H**) the Danish models applied to the Boston dataset and (**I**) the Boston models applied to the Boston dataset. (**J-K**) The odds ratios decrease at higher data exclusion intervals and higher recall thresholds. (**J**) Odds ratios for the Danish GRU model, trained with data exclusion intervals, applied to the Danish dataset. (**K**) Odds ratios for the Boston GRU model, trained with data exclusion intervals, applied to the Boston dataset. (**L**) Prediction performance (by AUROC) was significantly lower when using only 23 known risk factors, rather than 2000 disease codes (no data exclusion).

Predictive Features

[[Interpretation: contribution of known risk factors]]

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 diagnoses are most informative of cancer risk and at what time point? We used two methods for the identification of factors that contribute most to positive prediction. One method uses prior knowledge and limits the input for training and testing to disease types, which have been reported to be indicative of the likely occurrence of pancreatic cancer (Yuan et al. 2020; Klein 2021). The result is that these 23 known risk factors are moderately predictive of cancer but are much less informative compared to the more than 2,000 available diseases (**Table S4**, **Figure 3L**). The relationship between age, number of disease codes and pancreatic cancer occurrence is also consistent with the fact that increasing age has been reported as a major risk factor of pancreatic cancer (**Figure 2A, S6**).

[[Interpretation: contributing factors by gradient method]]

A second, explicitly computational method infers the contribution of a particular input variable to the prediction by the machine learning engine using the integrated gradients (IG) algorithm

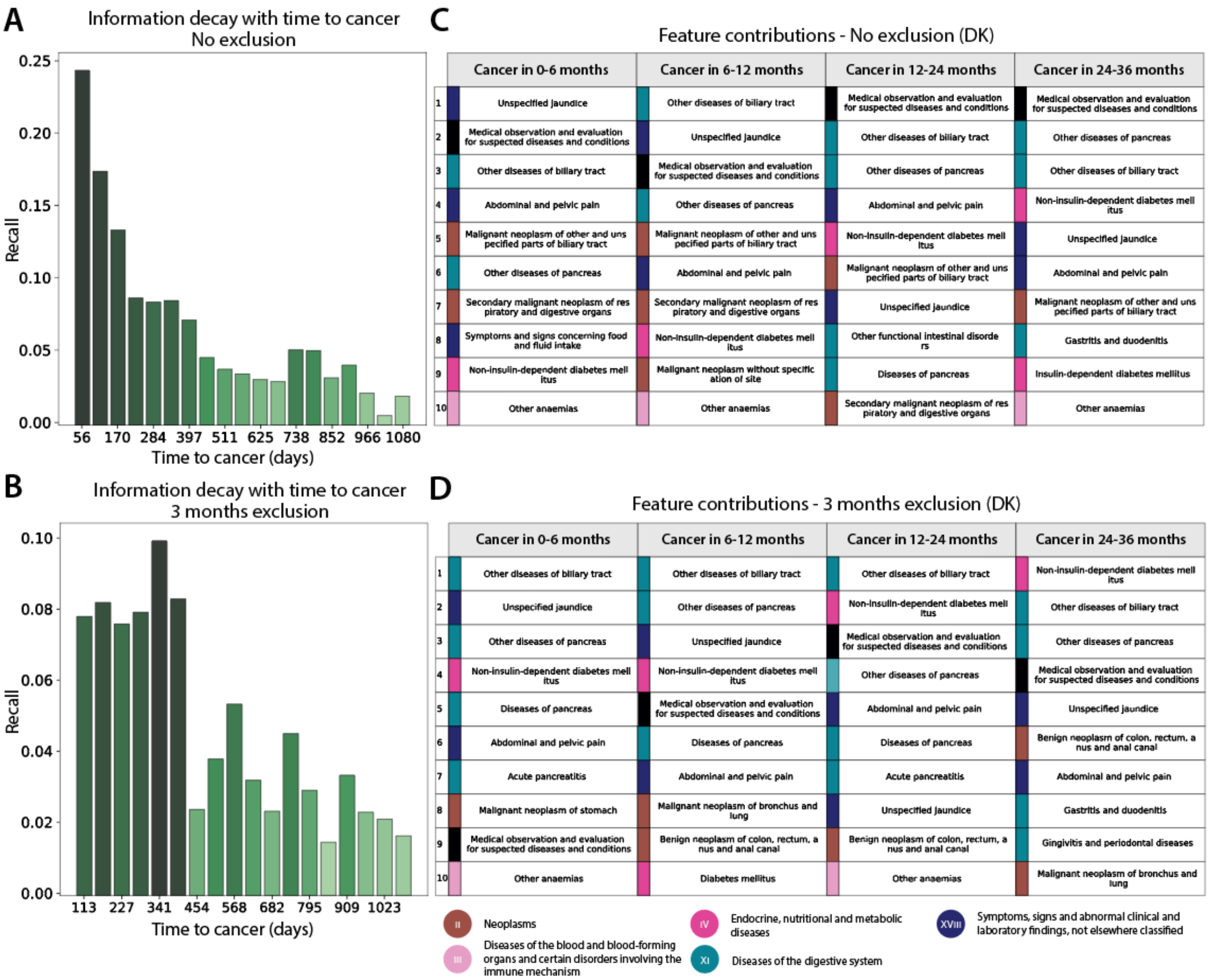
(Sundararajan, Taly, and Yan 2017) (**Figure 4B,D**). The IG contribution was calculated separately for different times to cancer diagnosis, in particular at 0-6 months, 6-12 months, 12-24 months and 24-36 months after assessment, for all patients developing cancer. The aim was to explore how diseases contribute differently to the risk of pancreatic cancer, depending on how close to pancreatic cancer they occurred. There is a tendency for diseases, which in normal clinical practice are known to indicate the potential presence of pancreatic cancer, to have a higher contribution to prediction for trajectories that end closer to cancer diagnosis. On the other hand, putative early risk factors plausibly have a higher IG score for the trajectories that end many months before cancer diagnosis. As an additional check, we computed the contribution for the model trained also with 3 months data exclusion.

[[Interpretation: contributing early factors]]

The top contributing features extracted from the trajectories with time to cancer diagnosis in 0-6 months - such as unspecified jaundice, diseases of biliary tract, abdominal-pelvic pain, weight loss and neoplasms of digestive organs - may be symptoms of or otherwise closely related to pancreatic cancer (**Table S5**). It is also of interest to identify early risk factors for pancreatic cancer**.** For trajectories with longer time between assessment and cancer diagnosis, other disease codes - such as type 2 diabetes and insulin-independent diabetes - make an increasingly large contribution, consistent with epidemiological studies (Yuan et al. 2020; Klein et al. 2013; Kim et al. 2020) and the observed disease distribution in the DNPR dataset (**Figure 4, 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 prior to cancer as a result of cholelithiasis or a hypothetical link between medication by proton pump inhibitors such as omeprazole in reflux disease and the effect of increased levels of gastrin on the state of the pancreas (Alkhushaym et al. 2020).

[[Interpretation for 3 month exclusion interval]]

Overall the contribution of the diseases calculated for the model trained with 3 months data exclusion is similar to the one calculated for the model without data exclusion. The main difference is in the order of the disease contribution, as the diseases that more frequently are diagnosed as a consequence of subclinical pancreatic cancer - which are not included in the training of the 3 months data exclusion model - have lower contribution than the longer term risk factors. The interpretation of individual risk factors from the ML feature list as causative may be subject to misinterpretation as their contribution here is only evaluated in the context of complete disease histories. However, our main goal in this report is to achieve robust predictive power from disease trajectories, rather than mechanistic interpretations.



Feature contributions - No exclusion (DK)

	Cancer in 0-6 months	Cancer in 6-12 months	Cancer in 12-24 months	Cancer in 24-36 months
1	Unspecified jaundice	Other diseases of biliary tract	Medical observation and evaluation for suspected diseases and conditions	Medical observation and evaluation for suspected diseases and conditions
2	Medical observation and evaluation for suspected diseases and conditions	Unspecified jaundice	Other diseases of biliary tract	Other diseases of pancreas
3	Other diseases of biliary tract	Medical observation and evaluation for suspected diseases and conditions	Other diseases of pancreas	Other diseases of biliary tract
4	Abdominal and pelvic pain	Other diseases of pancreas	Abdominal and pelvic pain	Non-insulin-dependent diabetes mellitus
5	Malignant neoplasm of other and unspecified parts of biliary tract	Malignant neoplasm of other and unspecified parts of biliary tract	Non-insulin-dependent diabetes mellitus	Unspecified jaundice
6	Other diseases of pancreas	Abdominal and pelvic pain	Malignant neoplasm of other and unspecified parts of biliary tract	Abdominal and pelvic pain
7	Secondary malignant neoplasm of respiratory and digestive organs	Secondary malignant neoplasm of respiratory and digestive organs	Unspecified jaundice	Malignant neoplasm of other and unspecified parts of biliary tract
8	Symptoms and signs concerning food and fluid intake	Non-insulin-dependent diabetes mellitus	Other functional intestinal disorders	Gastritis and duodenitis
9	Non-insulin-dependent diabetes mellitus	Malignant neoplasm without specification of site	Diseases of pancreas	Insulin-dependent diabetes mellitus
10	Other anaemias	Other anaemias	Secondary malignant neoplasm of respiratory and digestive organs	Other anaemias

Feature contributions - 3 months exclusion (DK)

	Cancer in 0-6 months	Cancer in 6-12 months	Cancer in 12-24 months	Cancer in 24-36 months
1	Other diseases of biliary tract	Other diseases of biliary tract	Other diseases of biliary tract	Non-insulin-dependent diabetes mellitus
2	Unspecified jaundice	Other diseases of pancreas	Non-insulin-dependent diabetes mellitus	Other diseases of biliary tract
3	Other diseases of pancreas	Unspecified jaundice	Medical observation and evaluation for suspected diseases and conditions	Other diseases of pancreas
4	Non-insulin-dependent diabetes mellitus	Non-insulin-dependent diabetes mellitus	Other diseases of pancreas	Medical observation and evaluation for suspected diseases and conditions
5	Diseases of pancreas	Medical observation and evaluation for suspected diseases and conditions	Abdominal and pelvic pain	Unspecified jaundice
6	Abdominal and pelvic pain	Diseases of pancreas	Diseases of pancreas	Benign neoplasm of colon, rectum, anus and anal canal
7	Acute pancreatitis	Abdominal and pelvic pain	Acute pancreatitis	Abdominal and pelvic pain
8	Malignant neoplasm of stomach	Malignant neoplasm of bronchus and lung	Unspecified jaundice	Gastritis and duodenitis
9	Medical observation and evaluation for suspected diseases and conditions	Benign neoplasm of colon, rectum, anus and anal canal	Benign neoplasm of colon, rectum, anus and anal canal	Gingivitis and periodontal diseases
10	Other anaemias	Diabetes mellitus	Other anaemias	Malignant neoplasm of bronchus and lung

II	Neoplasms
III	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
IV	Endocrine, nutritional and metabolic diseases
XI	Diseases of the digestive system
XVIII	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified

Figure 4. Predictive capacity and feature contributions of disease trajectories
(**A-B**) Distribution of recall (sensitivity) values at the F1 operational point as a function of time-to-cancer (time between the end of a disease trajectory and cancer diagnosis). The recall values drop significantly with time-to-cancer. (**A**) For models trained on all data. (**B**) For models trained with 3 months data exclusion. (**C-D**) Top 10 features that contribute to the cancer prediction in time-to-cancer intervals of 0-6, 6-12, 12-24 and 24-36 months. The features are sorted by the contribution score (**Supplementary Tables S5**). We used an integrated gradient (IG) method to calculate the contribution score for each input feature for each trajectory, then summed over all trajectories with cancer diagnosis within the indicated time interval. All data in the figure for the Danish DNPR dataset, 36 months prediction interval.

Discussion

[[Advances in this work]]

Here we present a new framework for applying deep learning methods using comprehensive datasets of disease trajectories to predict cancer risk. Our study was designed to make explicit use of the time sequence of disease events; and, to assess the ability to predict cancer risk for increasing intervals between the time of assessment (the end of the disease trajectory) and cancer occurrence. Earlier work has demonstrated the potential of applying AI methods to assess pancreatic cancer risk but did not exploit the information in the temporal sequence of diseases (Appelbaum, Cambronero, et al. 2021; Chen et al. 2021). Our results indicate that using the time ordering in disease histories as input significantly improves the predictive power of AI methods in anticipating pancreatic cancer occurrence.

[[Comparison of performance in a different healthcare system]]

A single, globally robust model that predicts cancer risk for patients in different countries and different healthcare systems remains elusive. Cross-application of the Danish model to the Boston MGB database had significantly lower performance (Fig. 3H), in spite of common use of ICD disease codes. One of the reasons for this mismatch could be the differences in clinical practice, such as frequency of reporting disease codes in the clinical records, the typical threshold for seeking medical attention, potential influence of billing constraints on what is recorded, as well as referral practice to the local Boston MGB hospitals from other locations, in contrast to the more uniform and comprehensive national nature of the Danish DNPR disease registry. However, the AI methods used are sufficiently robust to achieve a similarly high level of performance in the Boston MGB system when independently trained. With significant differences in healthcare systems, independent model training in different geographical locations may be necessary to achieve desired model performance.

[[Clinical trials and application in clinical practice]]

Successful implementation of early diagnosis and treatment of pancreatic cancer in clinical practice will likely require three essential steps: identification of high-risk patients, detection of early cancer or precancerous states by detailed screening of high-risk patients, and effective treatment after early detection (Singhi et al. 2019; Kenner et al. 2021). The overall impact in clinical practice depends on the success rates in each of these stages. This work only addresses the first stage. With a reasonably accurate method for predicting cancer risk one can direct appropriate high-risk patients into clinical screening trials. A sufficiently enriched pool of high-risk patients would make detailed screening tests more affordable, as such tests are likely to be prohibitively expensive at a population level and enhance the positive predictive value of such tests.

Although the level of performance reported here exceeds that of previous prediction models, implementation in clinical practice requires additional considerations. A careful choice of operational point is required, which is not necessarily the one maximizing F1, which balances precision and recall and was used above as a point of reference. The criteria for initiating clinical screening trials should take into account the cost / benefit balance of screening and intervention (Pandharipande et al. 2016) (example estimate in **Results S1**) as well as the expectations and concerns of patients enrolled in a trial and of those identified as high risk and offered advanced clinical test. The specific design of such trials will require close collaboration between data scientists and practicing clinicians to determine appropriate evaluation and follow-up once high-risk patients are identified by risk assessment tools. Nevertheless, the current late-stage presentation of about 80% of pancreatic cancer patients with incurable disease suggests that innovative approaches will be required to improve patient outcomes for this highly lethal malignancy.

For example, based on the prediction accuracy reported here, one can realistically design clinical screening trials, with software applied to health records of, e.g., 1 million patients, followed by identification of those at highest risk and recruitment into a clinical trial with detailed screening tests for, e.g., 200 high-risk patients. Implementation requires choosing an operational point along the PRC curve with an achievable high positive predictive value, which is important to reduce false positives and therefore minimize unnecessary effort and anxiety. Exploiting the trade-off between precision and recall, one can in this scenario accept lower recall as a clinical trial with limited enrollment cannot in any case detect cancer in a large number of patients. The particular advantage of this ‘predict-select-screen’ process is that computational screening of a *large* population in the first step is inexpensive while the costly second step of sophisticated clinical screening and therapeutic intervention programs is limited to a much *smaller* number of patients, those at highest risk.

[[Challenges for future improvements]]

We expect further increases in prediction accuracy with the availability of data beyond disease codes, such as prescriptions, laboratory values, observations in clinical notes, diagnosis and treatment records from general practitioners (Malhotra et al. 2021) and abdominal imaging (computed tomography, magnetic resonance imaging), as well as inherited genetic profiles. To achieve a globally useful set of prediction rules, access to large data sets of disease histories aggregated nationally or internationally will be extremely valuable. 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 results.

[[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 produce a scalable workflow for early detection of pancreatic cancer in the community, to shift focus from treatment of late- to early-stage cancer, improve the quality of life of patients, and increase the benefit/cost ratio of cancer care.

Methods

Processing of the population-level dataset

[[Danish DNPR dataset]]

The first part of the project was conducted using a dataset of disease histories from the Danish National Patient Registry (DNPR), covering all 229 million hospital diagnoses of 8.6 million patients between 1977-2018. This includes inpatient contacts since 1977 and outpatient and emergency department contacts since 1995, but not data from the general practitioners' records (Schmidt et al. 2015). DNPR access was approved by the Danish Health Data Authority (FSEID-00003092 and FSEID-00004491.) Each entry of the database 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.

The most up-to-date 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*. 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 have since 1968 been assigned a unique lifetime Central Person Registration (CPR) Number, 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 excluded from consideration residents of Greenland, patients who lack a stable place of residence in Denmark, as these would potentially have discontinuous disease trajectories. 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 pancreatic 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 (train/dev/test), to avoid any information leakage between training and development/test datasets.

[[Boston MGB dataset]]

The MGB dataset is from the Mass General Brigham Research Patient Data Registry (RPDR), including data items from the Dana-Farber/ Brigham and Women's Cancer Center, and contains ICD-9-CM and ICD-10-CM codes for disease administration, both are more detailed modifications to the ICD-9/10 international version. Data access for the study was granted under the Institutional Review Board (IRB) Protocol 2019P000993 (*Computational Approaches to Identifying High-Risk Pancreatic Cancer Populations: High Risk Cohorts Through Real World Data*). Analogously to DNPR, we used the three-character category ICD codes for identifying pancreatic cancer, respectively *C25* for ICD-10 and *157* for ICD-9. The end date was similarly defined as the date of death for the patients, the date of the last hospital visit, or, if the patient on file is still alive, the end date used to select from the MGB dataset (2020), whichever is earlier.

Training

The following processing steps were carried out identically for DNPR and MGC 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 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, 120 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.

For patients **without** a pancreatic cancer diagnosis we only include trajectories that end earlier than 2 years before the end of their disease records (due to death or the freeze date of the DNPR data used here). This avoids the uncertainty of cases in which undiagnosed cancer might have existed before the end of the records. The datasets were sampled in small batches for efficient computation, as is customary in ML. Due to the small number of cases of pancreatic cancer compared to controls, we used balanced sampling from the trajectories of the patients in the training set such that each batch has an approximately equal number of positive (cancer) and negative (non-cancer) trajectories.

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. After embedding, each sequence of diagnoses, 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 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 model only uses a single linear layer).

The model output consists of a risk score that monotonically increases for each time interval in the follow-up period after risk assessment. 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 $t = t_a + 3, 6, 12, 36, 60, 120$ months. For a given prediction threshold, scores that exceed such threshold at time *t* are considered to indicate cancer occurrence prior to *t*. 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 1A**).

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

where $t \in \{3,6,12,36,60,120\}$ months; $N_T = 6$ for non-cancer patient and $N_T \leq 6$ 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 and $y_{i,t} = 0$ for no cancer within 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 (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.

Evaluation

The evaluation was carried out separately for each prediction interval of 0-3, 0-6, 0-12, 0-36, and 0-60 months. For example, consider the prediction score for a particular trajectory at the end of the 3 year prediction interval (Fig.1C). If the score is above the threshold, one has a correct positive prediction, if cancer has occurred at any time within 3 years; and a false positive prediction, if cancer has not occurred within 3 years. If the score is below the threshold, one has a false negative prediction if cancer has occurred at any time within 3 years; and a true negative prediction, if cancer has not occurred within 3 years. As both training and evaluation make use of multiple trajectories per patient, with different end-of-trajectory points, the performance numbers are computed over trajectories.

The odds ratio (OR) was calculated as the odds of getting pancreatic cancer when classified at high risk divided by the odds of getting pancreatic cancer when classified at low risk, after picking a specific recall level.

$$OR = \frac{TP/FP}{FN/TN}$$

where TP = True Positives, FP = False Positives, FN = False Negatives, TN = True negatives. For the 0-36 months prediction interval, the observation is diagnosis of pancreatic cancer within 36 months of assessment, yes/no; and the prediction is high risk / low risk at a given operational threshold (e.g., by choosing a specific level of recall).

Cross-application

Few adaptations were necessary in order to test the model trained on the Danish DNPR data on the Boston MGB dataset. In particular, the ICD-9-CM codes were first converted to ICD-10-CM codes using the mapping available on the National Center for Health Statistic (NHCS, www.cdc.gov/nchs) and then, once truncated at the three-characters level, were matched to the respective ICD-10 codes from the DNPR. In this way we created a joint 'vocabulary' where disease codes from the MGB dataset were mapped to the same embedded representation of the matching disease code in DNPR-trained models. In spite of overall semantic agreement of the internationally standardized ICD codes (50,656 out of 53,552 can be matched), the translation from one coding system to the other caused missing values in the input. Indeed, some ICD-9-CM codes (n=969) could not be matched to a single ICD-10-CM code and some ICD-10-CM codes (n=1,927) had no match with the ICD-10 codes in DNPR. We compared the performance results from cross-application to those of the independently trained models by evaluating them against the same test data (subset of Boston MGB data).

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, cumulating the gradients calculated along all the points in the path from the input to the baseline. We chose 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 summing across all the patient trajectories. 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.

References

Alkhushaym, Nasser, Abdulaali R. Almutairi, Abdulhamid Althagafi, Saad B. Fallatah, Mok Oh, Jennifer R. Martin, Hani M. Babiker, Ali McBride, and Ivo Abraham. 2020. "Exposure to Proton Pump Inhibitors and Risk of Pancreatic Cancer: A Meta-Analysis." *Expert Opinion on Drug Safety* 19 (3): 327–34.

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.

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.

Hu, Jessica X., Marie Helleberg, Anders B. Jensen, Søren Brunak, and Jens Lundgren. 2019. "A Large-Cohort, Longitudinal Study Determines Precancer Disease Routes across Different Cancer Types." *Cancer Research* 79 (4): 864–72.

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.

Jensen, Anders Boeck, Pope L. Moseley, Tudor I. Oprea, Sabrina Gade Ellesøe, Robert Eriksson, Henriette Schmock, Peter Bjødstrup Jensen, Lars Juhl Jensen, and Søren Brunak. 2014. "Temporal Disease Trajectories Condensed from Population-Wide Registry Data Covering 6.2 Million Patients." *Nature Communications* 5 (June): 4022.

Kenner, Barbara, Suresh T. Chari, David Kelsen, David S. Klimstra, Stephen J. Pandol, Michael Rosenthal, Anil K. Rustgi, et al. 2021. "Artificial Intelligence and Early Detection of Pancreatic Cancer: 2020 Summative Review." *Pancreas* 50 (3): 251–79.

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.

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.

Li, Xiaodong, Peng Gao, Chao-Jung Huang, Shiying Hao, Xuefeng B. Ling, Yongxia Han, Yaqi Zhang, et al. 2020. “A Deep-Learning Based Prediction of Pancreatic Adenocarcinoma with Electronic Health Records from the State of Maine.” *International Journal of Medical and Health Sciences* 14 (11): 358–65.

Li, Yikuan, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. 2020. “BEHRT: Transformer for Electronic Health Records.” *Scientific Reports* 10 (1): 7155.

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.

McGuigan, Andrew, Paul Kelly, Richard C. Turkington, Claire Jones, Helen G. Coleman, and R. Stephen McCain. 2018. “Pancreatic Cancer: A Review of Clinical Diagnosis, Epidemiology, Treatment and Outcomes.” *World Journal of Gastroenterology: WJG* 24 (43): 4846–61.

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.

Pandharipande, Pari V., Curtis Heberle, Emily C. Dowling, Chung Yin Kong, Angela Tramontano, Katherine E. Perzan, William Brugge, and Chin Hur. 2016. “Targeted Screening of Individuals at

High Risk for Pancreatic Cancer: Results of a Simulation Model.” *Radiology* 278 (1): 306.

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.

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.

Acknowledgements

We thank Adam Yala for expert advice and contributions to methods and software, Regina Barzilay for discussions and guidance, Barbara Kenner for advice, and Julia Sidenius and Inna Chen for advice on clinical matters and known risk factor diseases. We thank Shawn Murphy and Henry Chueh and the Mass General Brigham Health Care Research Patient Data Registry group for facilitating use of their database.

Funding: DP, JXH, ADH and SB acknowledge support from the Novo Nordisk Foundation (grants NNF17OC0027594 and NNF14CC0001). BMW acknowledges support from the Hale Family Center for Pancreatic Cancer Research, Lustgarten Foundation Dedicated Laboratory program, Stand Up to Cancer, NIH grant U01 CA210171, NIH grant P50 CA127003, Pancreatic Cancer Action Network, Noble Effort Fund, Wexler Family Fund, Promises for Purple, and Bob Parsons Fund. MHR acknowledges support from the Hale Family Center for Pancreatic Cancer Research, Lustgarten Foundation Dedicated Laboratory program, and NIH grant U01 CA210171. We thank Stand Up to Cancer, the Lustgarten Foundation and their donors for financial and community support.

Author contributions: DP, BY, JH, AH, SB, CS conceptualization. DP, BY, JH, SB, CS methodology. DP, BY software. DP, BY, JH validation. DP, BY investigation. PC, RU, GA and AC resources. DP, BY, JH, RU, GA, AC data curation. DP, BY, JH, CS original draft. ALL writing review and editing. EA, LB project administration. DM, AR, PK, BW, MR, SB, CS supervision.

Competing Interests: S.B. has ownership 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.

Data and material availability: The software can be made available upon request to cancerriskprediction@gmail.com. The Danish National Patient Registry (DNPR) can only be accessed by researchers authorized by the Danish health authorities. Similarly, Mass General Brigham (MGB) dataset is part of the Research Patient Data Registry (RPDR) and it is only accessible by internal researchers with institutional review board (IRB) approval. All approvals are stated in the manuscript.

Supplementary Materials

Figure S1. Preprocessing and filtering of the DK and MGB disease trajectory datasets.

Filtering of the Danish (DK) and Boston MGB patient registries prior to training: in the Danish dataset, patient status codes were used to remove discontinuous disease histories such as patients living in Greenland, patients with alterations in their patient ID or patients who lack a stable residence in Denmark. We also removed referral and temporary diagnosis codes which are not the final diagnosis codes and can be misleading to use for training. Patients with short trajectories (<5 diagnosis codes) were removed. The final set of patients were split into Training (80 %), Validation (10%) and Testing set (10%).

For the Boston MGB dataset, the first step was an extra layer of patient ID de-identification, which was done by adding a unique random small time shift per patient. Similar to the Danish dataset filtering, short trajectories (<5 diagnosis codes) were removed and patients split into Training (80 %), Validation (10%) and Test set (10%).

Figure S1A - Denmark (DK) DNPR

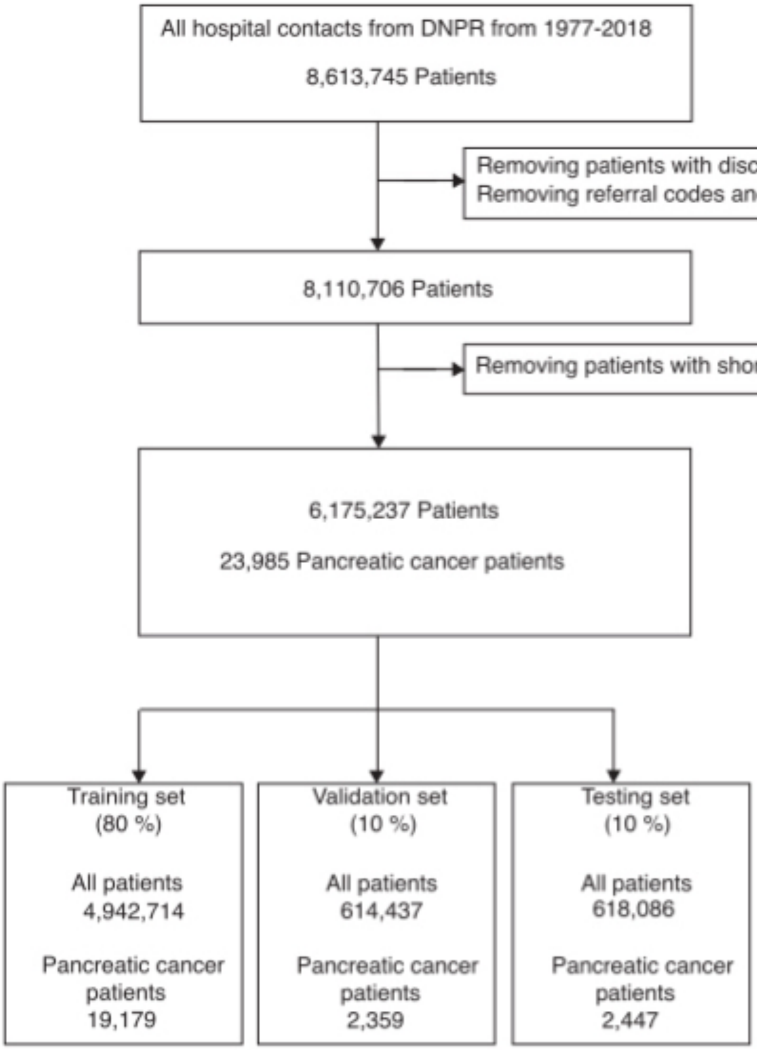

All hospital contacts from DNPR from 1977-2018
8,613,745 Patients
Removing patients with discontinuous disease histories
Removing referral codes and temporary codes
8,110,706 Patients
Removing patients with short (<5 events) disease trajectories
6,175,237 Patients
23,985 Pancreatic cancer patients
Training set
(80 %)
All patients
4,942,714
Pancreatic cancer patients
19,179
Validation set
(10 %)
All patients
614,437
Pancreatic cancer patients
2,359
Testing set
(10 %)
All patients
618,086
Pancreatic cancer patients
2,447

Figure S1B - Boston MGB (RPDR)

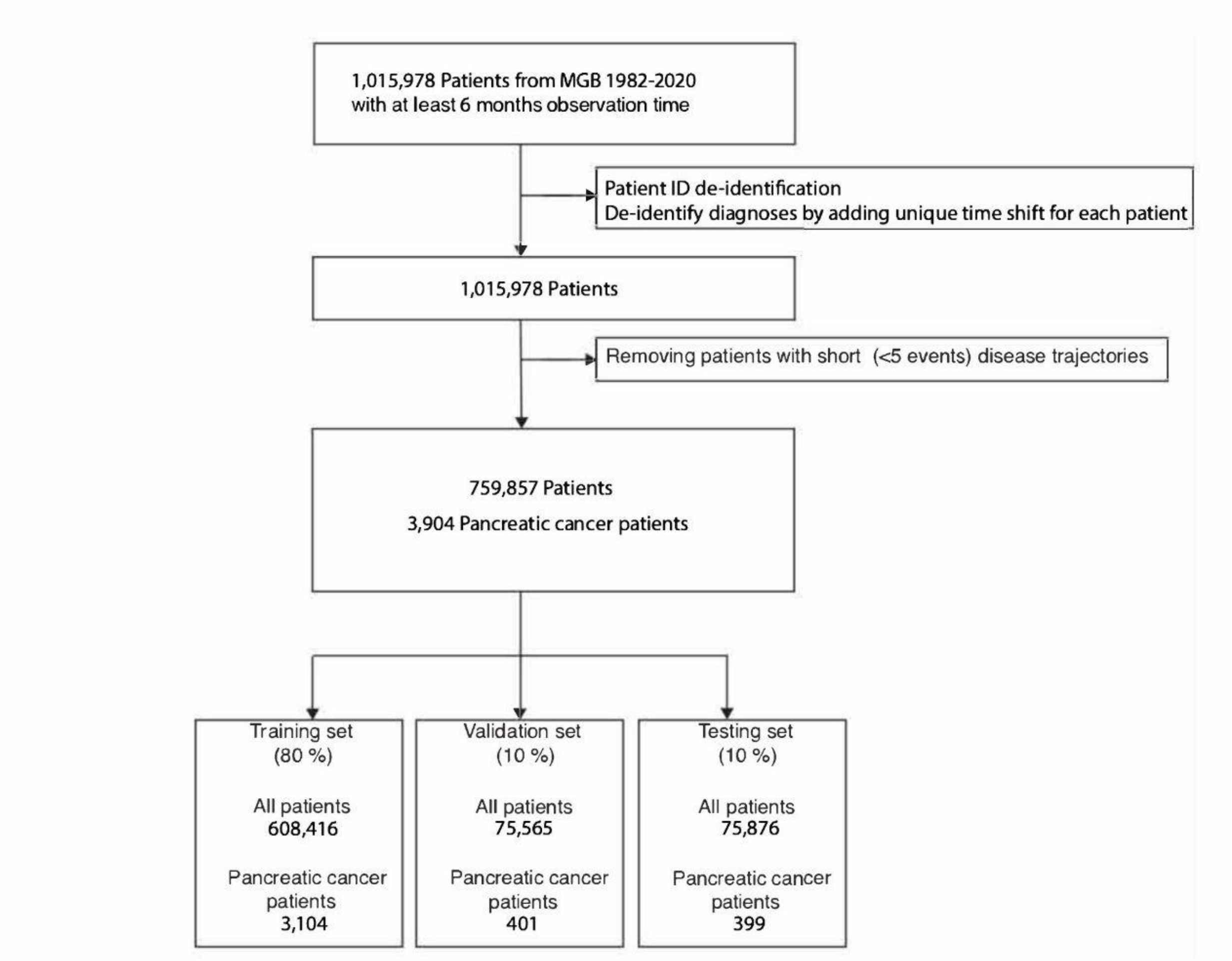

1,015,978 Patients from MGB 1982-2020
with at least 6 months observation time
Patient ID de-identification
De-identify diagnoses by adding unique time shift for each patient
1,015,978 Patients
Removing patients with short (<5 events) disease trajectories
759,857 Patients
3,904 Pancreatic cancer patients
Training set
(80 %)
All patients
608,416
Pancreatic cancer
patients
3,104
Validation set
(10 %)
All patients
75,565
Pancreatic cancer
patients
401
Testing set
(10 %)
All patients
75,876
Pancreatic cancer
patients
399

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

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=23,985)		
	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%)

Table S2. Hyperparameter search for machine learning models.

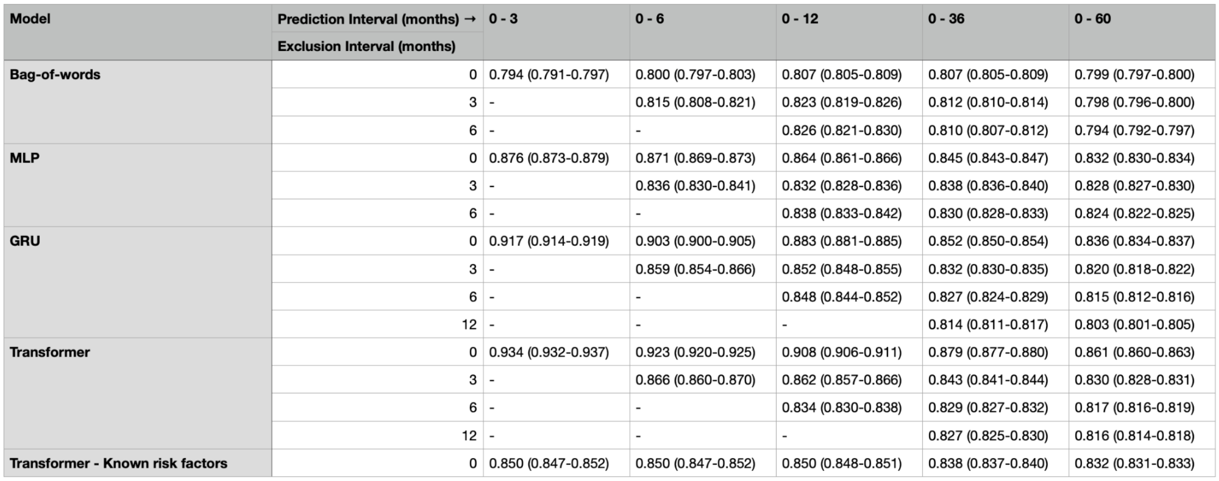
To comprehensively test different types of neural networks and the corresponding hyperparameters, we conducted a large parameter search for each of the network types. The different types of models include simple feed-forward models (LR, MLP) and more complex models that can take the sequential information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

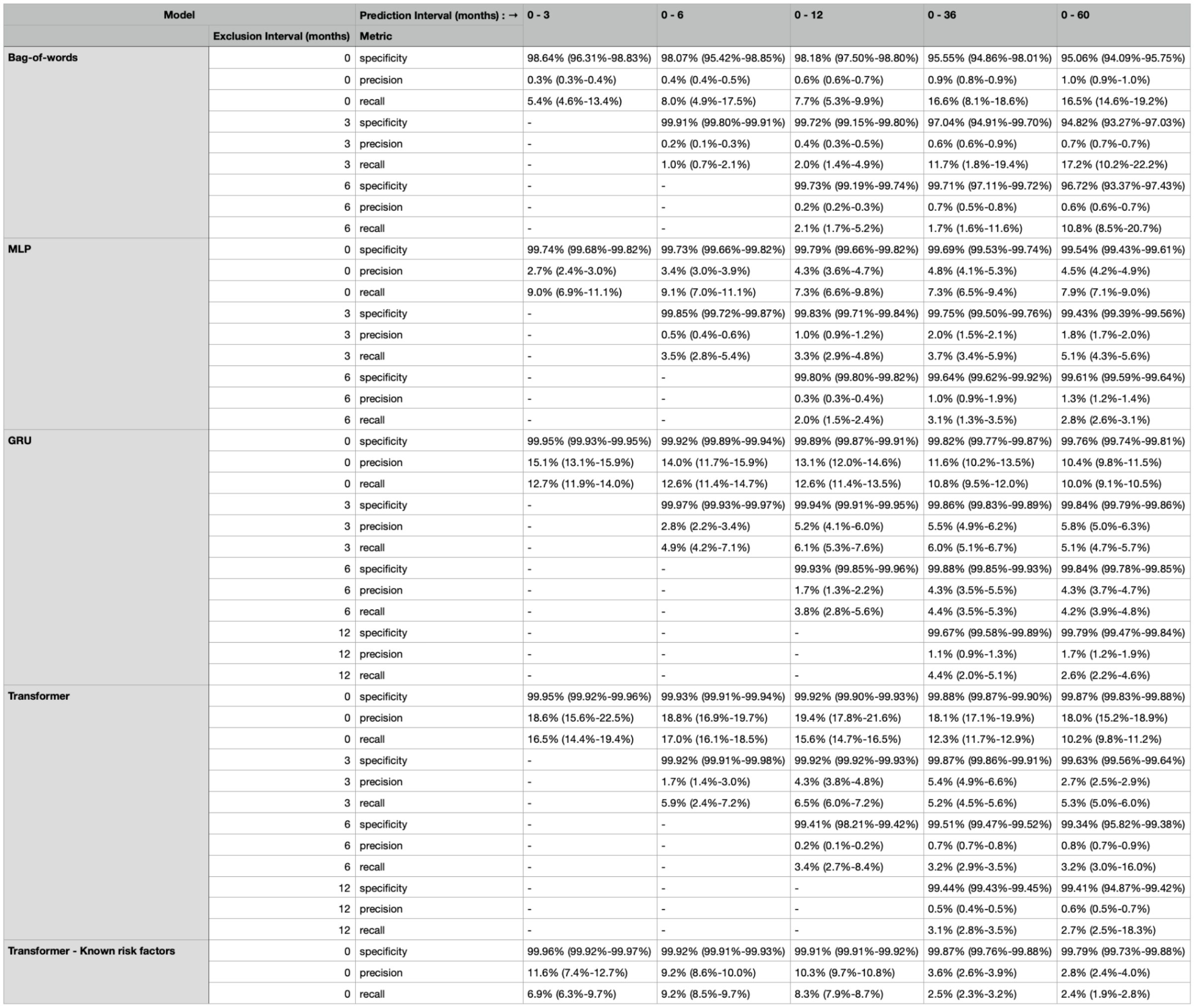
A summary of performance of different models trained with different data exclusion intervals for different prediction intervals. In order to estimate the uncertainty of the performance metrics, 95% confidence interval (CI) were computed using 200 resamples (bootstrapping with replacement); these time intervals may be slightly too narrow due to the estimated small number of trajectories from a single patient in a particular sample, but provide a reasonable guide. Specificity, precision, and recall are for the F1-optimal operational point.

Table S3A. Performance summary DNPR (AUROC)



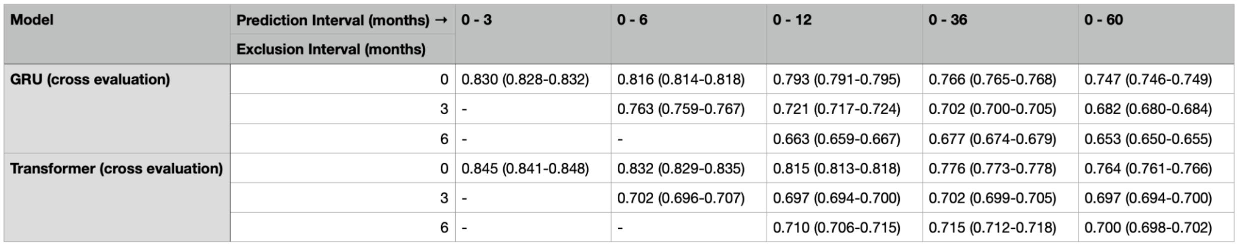
Model	Prediction Interval (months) →	0 - 3	0 - 6	0 - 12	0 - 36	0 - 60
	Exclusion Interval (months)					
Bag-of-words	0	0.794 (0.791-0.797)	0.800 (0.797-0.803)	0.807 (0.805-0.809)	0.807 (0.805-0.809)	0.799 (0.797-0.800)
	3	-	0.815 (0.808-0.821)	0.823 (0.819-0.826)	0.812 (0.810-0.814)	0.798 (0.796-0.800)
	6	-	-	0.826 (0.821-0.830)	0.810 (0.807-0.812)	0.794 (0.792-0.797)
MLP	0	0.876 (0.873-0.879)	0.871 (0.869-0.873)	0.864 (0.861-0.866)	0.845 (0.843-0.847)	0.832 (0.830-0.834)
	3	-	0.836 (0.830-0.841)	0.832 (0.828-0.836)	0.838 (0.836-0.840)	0.828 (0.827-0.830)
	6	-	-	0.838 (0.833-0.842)	0.830 (0.828-0.833)	0.824 (0.822-0.825)
GRU	0	0.917 (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	-	0.859 (0.854-0.866)	0.852 (0.848-0.855)	0.832 (0.830-0.835)	0.820 (0.818-0.822)
	6	-	-	0.848 (0.844-0.852)	0.827 (0.824-0.829)	0.815 (0.812-0.816)
	12	-	-	-	0.814 (0.811-0.817)	0.803 (0.801-0.805)
Transformer	0	0.934 (0.932-0.937)	0.923 (0.920-0.925)	0.908 (0.906-0.911)	0.879 (0.877-0.880)	0.861 (0.860-0.863)
	3	-	0.866 (0.860-0.870)	0.862 (0.857-0.866)	0.843 (0.841-0.844)	0.830 (0.828-0.831)
	6	-	-	0.834 (0.830-0.838)	0.829 (0.827-0.832)	0.817 (0.816-0.819)
	12	-	-	-	0.827 (0.825-0.830)	0.816 (0.814-0.818)
Transformer - Known risk factors	0	0.850 (0.847-0.852)	0.850 (0.847-0.852)	0.850 (0.848-0.851)	0.838 (0.837-0.840)	0.832 (0.831-0.833)

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



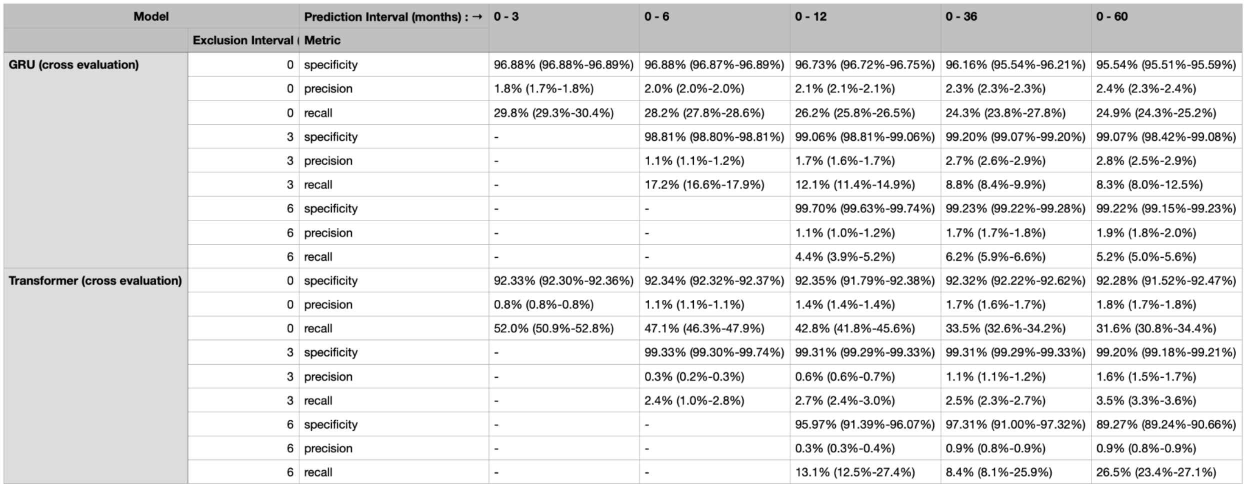
Model		Prediction Interval (months) : →	0 - 3	0 - 6	0 - 12	0 - 36	0 - 60
	Exclusion Interval (months)	Metric					
Bag-of-words	0	specificity	98.64% (96.31%-98.83%)	98.07% (95.42%-98.85%)	98.18% (97.50%-98.80%)	95.55% (94.86%-98.01%)	95.06% (94.09%-95.75%)
	0	precision	0.3% (0.3%-0.4%)	0.4% (0.4%-0.5%)	0.6% (0.6%-0.7%)	0.9% (0.8%-0.9%)	1.0% (0.9%-1.0%)
	0	recall	5.4% (4.6%-13.4%)	8.0% (4.9%-17.5%)	7.7% (5.3%-9.9%)	16.6% (8.1%-18.6%)	16.5% (14.6%-19.2%)
	3	specificity	-	99.91% (99.80%-99.91%)	99.72% (99.15%-99.80%)	97.04% (94.91%-99.70%)	94.82% (93.27%-97.03%)
	3	precision	-	0.2% (0.1%-0.3%)	0.4% (0.3%-0.5%)	0.6% (0.6%-0.9%)	0.7% (0.7%-0.7%)
	3	recall	-	1.0% (0.7%-2.1%)	2.0% (1.4%-4.9%)	11.7% (1.8%-19.4%)	17.2% (10.2%-22.2%)
	6	specificity	-	-	99.73% (99.19%-99.74%)	99.71% (97.11%-99.72%)	96.72% (93.37%-97.43%)
	6	precision	-	-	0.2% (0.2%-0.3%)	0.7% (0.5%-0.8%)	0.6% (0.6%-0.7%)
	6	recall	-	-	2.1% (1.7%-5.2%)	1.7% (1.6%-11.6%)	10.8% (8.5%-20.7%)
MLP	0	specificity	99.74% (99.68%-99.82%)	99.73% (99.66%-99.82%)	99.79% (99.66%-99.82%)	99.69% (99.53%-99.74%)	99.54% (99.43%-99.61%)
	0	precision	2.7% (2.4%-3.0%)	3.4% (3.0%-3.9%)	4.3% (3.6%-4.7%)	4.8% (4.1%-5.3%)	4.5% (4.2%-4.9%)
	0	recall	9.0% (6.9%-11.1%)	9.1% (7.0%-11.1%)	7.3% (6.6%-9.8%)	7.3% (6.5%-9.4%)	7.9% (7.1%-9.0%)
	3	specificity	-	99.85% (99.72%-99.87%)	99.83% (99.71%-99.84%)	99.75% (99.50%-99.76%)	99.43% (99.39%-99.56%)
	3	precision	-	0.5% (0.4%-0.6%)	1.0% (0.9%-1.2%)	2.0% (1.5%-2.1%)	1.8% (1.7%-2.0%)
	3	recall	-	3.5% (2.8%-5.4%)	3.3% (2.9%-4.8%)	3.7% (3.4%-5.9%)	5.1% (4.3%-5.6%)
	6	specificity	-	-	99.80% (99.80%-99.82%)	99.64% (99.62%-99.92%)	99.61% (99.59%-99.64%)
	6	precision	-	-	0.3% (0.3%-0.4%)	1.0% (0.9%-1.9%)	1.3% (1.2%-1.4%)
	6	recall	-	-	2.0% (1.5%-2.4%)	3.1% (1.3%-3.5%)	2.8% (2.6%-3.1%)
GRU	0	specificity	99.95% (99.93%-99.95%)	99.92% (99.89%-99.94%)	99.89% (99.87%-99.91%)	99.82% (99.77%-99.87%)	99.76% (99.74%-99.81%)
	0	precision	15.1% (13.1%-15.9%)	14.0% (11.7%-15.9%)	13.1% (12.0%-14.6%)	11.6% (10.2%-13.5%)	10.4% (9.8%-11.5%)
	0	recall	12.7% (11.9%-14.0%)	12.6% (11.4%-14.7%)	12.6% (11.4%-13.5%)	10.8% (9.5%-12.0%)	10.0% (9.1%-10.5%)
	3	specificity	-	99.97% (99.93%-99.97%)	99.94% (99.91%-99.95%)	99.86% (99.83%-99.89%)	99.84% (99.79%-99.86%)
	3	precision	-	2.8% (2.2%-3.4%)	5.2% (4.1%-6.0%)	5.5% (4.9%-6.2%)	5.8% (5.0%-6.3%)
	3	recall	-	4.9% (4.2%-7.1%)	6.1% (5.3%-7.6%)	6.0% (5.1%-6.7%)	5.1% (4.7%-5.7%)
	6	specificity	-	-	99.93% (99.85%-99.96%)	99.88% (99.85%-99.93%)	99.84% (99.78%-99.85%)
	6	precision	-	-	1.7% (1.3%-2.2%)	4.3% (3.5%-5.5%)	4.3% (3.7%-4.7%)
	6	recall	-	-	3.8% (2.8%-5.6%)	4.4% (3.5%-5.3%)	4.2% (3.9%-4.8%)
	12	specificity	-	-	-	99.67% (99.58%-99.89%)	99.79% (99.47%-99.84%)
	12	precision	-	-	-	1.1% (0.9%-1.3%)	1.7% (1.2%-1.9%)
	12	recall	-	-	-	4.4% (2.0%-5.1%)	2.6% (2.2%-4.6%)
Transformer	0	specificity	99.95% (99.92%-99.96%)	99.93% (99.91%-99.94%)	99.92% (99.90%-99.93%)	99.88% (99.87%-99.90%)	99.87% (99.83%-99.88%)
	0	precision	18.6% (15.6%-22.5%)	18.8% (16.9%-19.7%)	19.4% (17.8%-21.6%)	18.1% (17.1%-19.9%)	18.0% (15.2%-18.9%)
	0	recall	16.5% (14.4%-19.4%)	17.0% (16.1%-18.5%)	15.6% (14.7%-16.5%)	12.3% (11.7%-12.9%)	10.2% (9.8%-11.2%)
	3	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%)
	3	precision	-	1.7% (1.4%-3.0%)	4.3% (3.8%-4.8%)	5.4% (4.9%-6.6%)	2.7% (2.5%-2.9%)
	3	recall	-	5.9% (2.4%-7.2%)	6.5% (6.0%-7.2%)	5.2% (4.5%-5.6%)	5.3% (5.0%-6.0%)
	6	specificity	-	-	99.41% (98.21%-99.42%)	99.51% (99.47%-99.52%)	99.34% (95.82%-99.38%)
	6	precision	-	-	0.2% (0.1%-0.2%)	0.7% (0.7%-0.8%)	0.8% (0.7%-0.9%)
	6	recall	-	-	3.4% (2.7%-8.4%)	3.2% (2.9%-3.5%)	3.2% (3.0%-16.0%)
	12	specificity	-	-	-	99.44% (99.43%-99.45%)	99.41% (94.87%-99.42%)
	12	precision	-	-	-	0.5% (0.4%-0.5%)	0.6% (0.5%-0.7%)
	12	recall	-	-	-	3.1% (2.8%-3.5%)	2.7% (2.5%-18.3%)
Transformer - Known risk factors	0	specificity	99.96% (99.92%-99.97%)	99.92% (99.91%-99.93%)	99.91% (99.91%-99.92%)	99.87% (99.76%-99.88%)	99.79% (99.73%-99.88%)
	0	precision	11.6% (7.4%-12.7%)	9.2% (8.6%-10.0%)	10.3% (9.7%-10.8%)	3.6% (2.6%-3.9%)	2.8% (2.4%-4.0%)
	0	recall	6.9% (6.3%-9.7%)	9.2% (8.5%-9.7%)	8.3% (7.9%-8.7%)	2.5% (2.3%-3.2%)	2.4% (1.9%-2.8%)

Table S3C. Performance summary DNPR model validated on RPDR (AUROC).



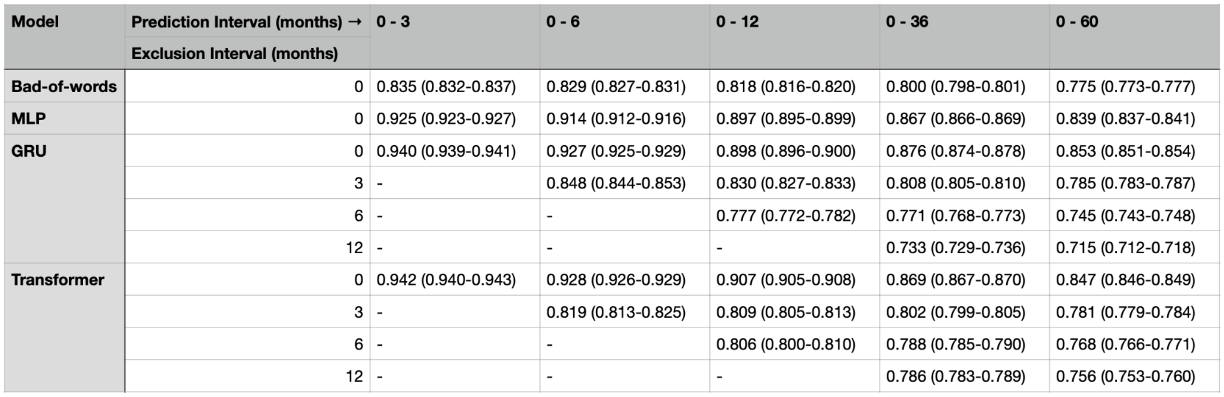
Model	Prediction Interval (months) →	0 - 3	0 - 6	0 - 12	0 - 36	0 - 60
	Exclusion Interval (months)					
GRU (cross evaluation)	0	0.830 (0.828-0.832)	0.816 (0.814-0.818)	0.793 (0.791-0.795)	0.766 (0.765-0.768)	0.747 (0.746-0.749)
	3	-	0.763 (0.759-0.767)	0.721 (0.717-0.724)	0.702 (0.700-0.705)	0.682 (0.680-0.684)
	6	-	-	0.663 (0.659-0.667)	0.677 (0.674-0.679)	0.653 (0.650-0.655)
Transformer (cross evaluation)	0	0.845 (0.841-0.848)	0.832 (0.829-0.835)	0.815 (0.813-0.818)	0.776 (0.773-0.778)	0.764 (0.761-0.766)
	3	-	0.702 (0.696-0.707)	0.697 (0.694-0.700)	0.702 (0.699-0.705)	0.697 (0.694-0.700)
	6	-	-	0.710 (0.706-0.715)	0.715 (0.712-0.718)	0.700 (0.698-0.702)

Table S3D. Performance summary DNPR model validated on RPDR (specificity/precision/recall).



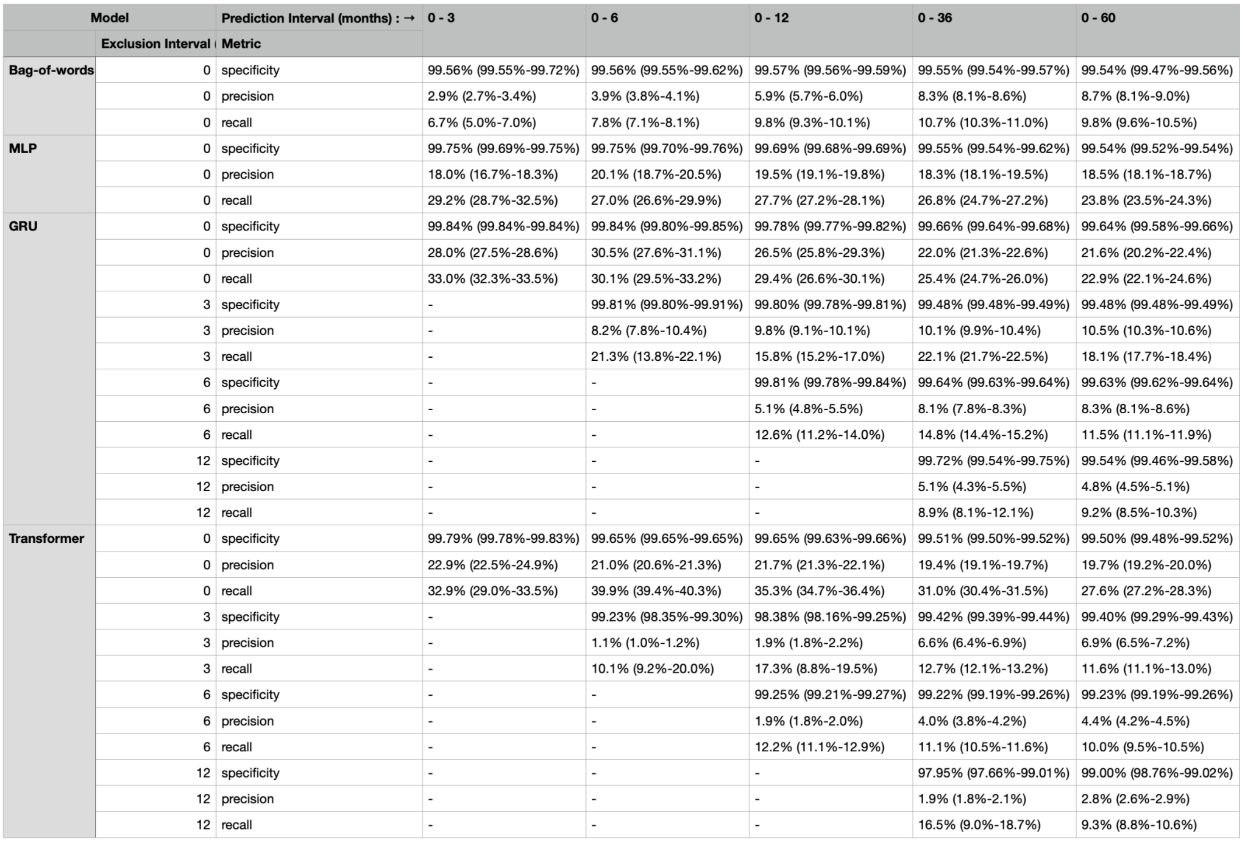
Model		Prediction Interval (months) : →	0 - 3	0 - 6	0 - 12	0 - 36	0 - 60
	Exclusion Interval (	Metric					
GRU (cross evaluation)	0	specificity	96.88% (96.88%-96.89%)	96.88% (96.87%-96.89%)	96.73% (96.72%-96.75%)	96.16% (95.54%-96.21%)	95.54% (95.51%-95.59%)
	0	precision	1.8% (1.7%-1.8%)	2.0% (2.0%-2.0%)	2.1% (2.1%-2.1%)	2.3% (2.3%-2.3%)	2.4% (2.3%-2.4%)
	0	recall	29.8% (29.3%-30.4%)	28.2% (27.8%-28.6%)	26.2% (25.8%-26.5%)	24.3% (23.8%-27.8%)	24.9% (24.3%-25.2%)
	3	specificity	-	98.81% (98.80%-98.81%)	99.06% (98.81%-99.06%)	99.20% (99.07%-99.20%)	99.07% (98.42%-99.08%)
	3	precision	-	1.1% (1.1%-1.2%)	1.7% (1.6%-1.7%)	2.7% (2.6%-2.9%)	2.8% (2.5%-2.9%)
	3	recall	-	17.2% (16.6%-17.9%)	12.1% (11.4%-14.9%)	8.8% (8.4%-9.9%)	8.3% (8.0%-12.5%)
	6	specificity	-	-	99.70% (99.63%-99.74%)	99.23% (99.22%-99.28%)	99.22% (99.15%-99.23%)
	6	precision	-	-	1.1% (1.0%-1.2%)	1.7% (1.7%-1.8%)	1.9% (1.8%-2.0%)
	6	recall	-	-	4.4% (3.9%-5.2%)	6.2% (5.9%-6.6%)	5.2% (5.0%-5.6%)
Transformer (cross evaluation)	0	specificity	92.33% (92.30%-92.36%)	92.34% (92.32%-92.37%)	92.35% (91.79%-92.38%)	92.32% (92.22%-92.62%)	92.28% (91.52%-92.47%)
	0	precision	0.8% (0.8%-0.8%)	1.1% (1.1%-1.1%)	1.4% (1.4%-1.4%)	1.7% (1.6%-1.7%)	1.8% (1.7%-1.8%)
	0	recall	52.0% (50.9%-52.8%)	47.1% (46.3%-47.9%)	42.8% (41.8%-45.6%)	33.5% (32.6%-34.2%)	31.6% (30.8%-34.4%)
	3	specificity	-	99.33% (99.30%-99.74%)	99.31% (99.29%-99.33%)	99.31% (99.29%-99.33%)	99.20% (99.18%-99.21%)
	3	precision	-	0.3% (0.2%-0.3%)	0.6% (0.6%-0.7%)	1.1% (1.1%-1.2%)	1.6% (1.5%-1.7%)
	3	recall	-	2.4% (1.0%-2.8%)	2.7% (2.4%-3.0%)	2.5% (2.3%-2.7%)	3.5% (3.3%-3.6%)
	6	specificity	-	-	95.97% (91.39%-96.07%)	97.31% (91.00%-97.32%)	89.27% (89.24%-90.66%)
	6	precision	-	-	0.3% (0.3%-0.4%)	0.9% (0.8%-0.9%)	0.9% (0.8%-0.9%)
	6	recall	-	-	13.1% (12.5%-27.4%)	8.4% (8.1%-25.9%)	26.5% (23.4%-27.1%)

Table S3E. Performance summary RPDR (AUROC)



Model	Prediction Interval (months) →	0 - 3	0 - 6	0 - 12	0 - 36	0 - 60
	Exclusion Interval (months)					
Bad-of-words	0	0.835 (0.832-0.837)	0.829 (0.827-0.831)	0.818 (0.816-0.820)	0.800 (0.798-0.801)	0.775 (0.773-0.777)
MLP	0	0.925 (0.923-0.927)	0.914 (0.912-0.916)	0.897 (0.895-0.899)	0.867 (0.866-0.869)	0.839 (0.837-0.841)
GRU	0	0.940 (0.939-0.941)	0.927 (0.925-0.929)	0.898 (0.896-0.900)	0.876 (0.874-0.878)	0.853 (0.851-0.854)
	3	-	0.848 (0.844-0.853)	0.830 (0.827-0.833)	0.808 (0.805-0.810)	0.785 (0.783-0.787)
	6	-	-	0.777 (0.772-0.782)	0.771 (0.768-0.773)	0.745 (0.743-0.748)
	12	-	-	-	0.733 (0.729-0.736)	0.715 (0.712-0.718)
Transformer	0	0.942 (0.940-0.943)	0.928 (0.926-0.929)	0.907 (0.905-0.908)	0.869 (0.867-0.870)	0.847 (0.846-0.849)
	3	-	0.819 (0.813-0.825)	0.809 (0.805-0.813)	0.802 (0.799-0.805)	0.781 (0.779-0.784)
	6	-	-	0.806 (0.800-0.810)	0.788 (0.785-0.790)	0.768 (0.766-0.771)
	12	-	-	-	0.786 (0.783-0.789)	0.756 (0.753-0.760)

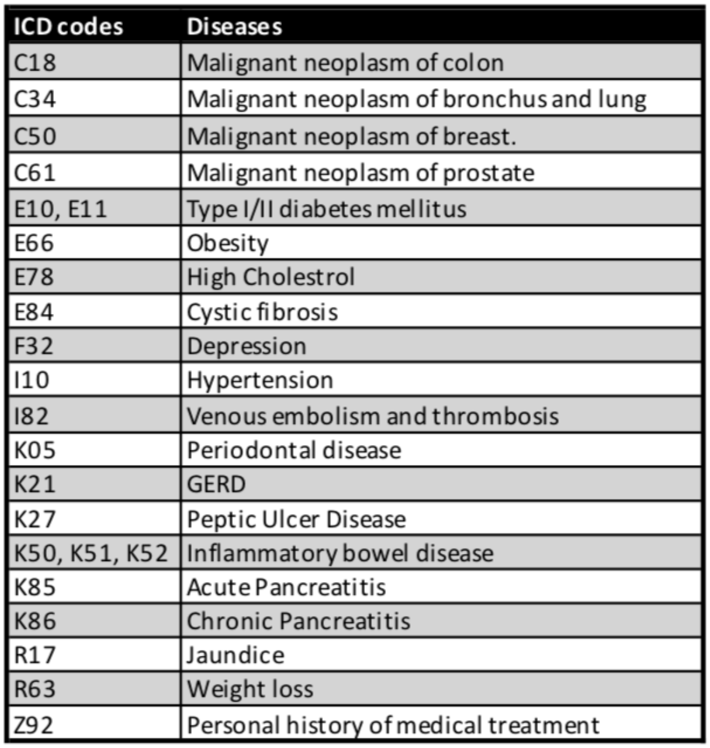
Table S3F. Performance summary RPDR (specificity/precision/recall)



Model		Prediction Interval (months) : →	0 - 3	0 - 6	0 - 12	0 - 36	0 - 60
	Exclusion Interval (	Metric					
Bag-of-words	0	specificity	99.56% (99.55%-99.72%)	99.56% (99.55%-99.62%)	99.57% (99.56%-99.59%)	99.55% (99.54%-99.57%)	99.54% (99.47%-99.56%)
	0	precision	2.9% (2.7%-3.4%)	3.9% (3.8%-4.1%)	5.9% (5.7%-6.0%)	8.3% (8.1%-8.6%)	8.7% (8.1%-9.0%)
	0	recall	6.7% (5.0%-7.0%)	7.8% (7.1%-8.1%)	9.8% (9.3%-10.1%)	10.7% (10.3%-11.0%)	9.8% (9.6%-10.5%)
MLP	0	specificity	99.75% (99.69%-99.75%)	99.75% (99.70%-99.76%)	99.69% (99.68%-99.69%)	99.55% (99.54%-99.62%)	99.54% (99.52%-99.54%)
	0	precision	18.0% (16.7%-18.3%)	20.1% (18.7%-20.5%)	19.5% (19.1%-19.8%)	18.3% (18.1%-19.5%)	18.5% (18.1%-18.7%)
	0	recall	29.2% (28.7%-32.5%)	27.0% (26.6%-29.9%)	27.7% (27.2%-28.1%)	26.8% (24.7%-27.2%)	23.8% (23.5%-24.3%)
GRU	0	specificity	99.84% (99.84%-99.84%)	99.84% (99.80%-99.85%)	99.78% (99.77%-99.82%)	99.66% (99.64%-99.68%)	99.64% (99.58%-99.66%)
	0	precision	28.0% (27.5%-28.6%)	30.5% (27.6%-31.1%)	26.5% (25.8%-29.3%)	22.0% (21.3%-22.6%)	21.6% (20.2%-22.4%)
	0	recall	33.0% (32.3%-33.5%)	30.1% (29.5%-33.2%)	29.4% (26.6%-30.1%)	25.4% (24.7%-26.0%)	22.9% (22.1%-24.6%)
	3	specificity	-	99.81% (99.80%-99.91%)	99.80% (99.78%-99.81%)	99.48% (99.48%-99.49%)	99.48% (99.48%-99.49%)
	3	precision	-	8.2% (7.8%-10.4%)	9.8% (9.1%-10.1%)	10.1% (9.9%-10.4%)	10.5% (10.3%-10.6%)
	3	recall	-	21.3% (13.8%-22.1%)	15.8% (15.2%-17.0%)	22.1% (21.7%-22.5%)	18.1% (17.7%-18.4%)
	6	specificity	-	-	99.81% (99.78%-99.84%)	99.64% (99.63%-99.64%)	99.63% (99.62%-99.64%)
	6	precision	-	-	5.1% (4.8%-5.5%)	8.1% (7.8%-8.3%)	8.3% (8.1%-8.6%)
	6	recall	-	-	12.6% (11.2%-14.0%)	14.8% (14.4%-15.2%)	11.5% (11.1%-11.9%)
	12	specificity	-	-	-	99.72% (99.54%-99.75%)	99.54% (99.46%-99.58%)
	12	precision	-	-	-	5.1% (4.3%-5.5%)	4.8% (4.5%-5.1%)
	12	recall	-	-	-	8.9% (8.1%-12.1%)	9.2% (8.5%-10.3%)
Transformer	0	specificity	99.79% (99.78%-99.83%)	99.65% (99.65%-99.65%)	99.65% (99.63%-99.66%)	99.51% (99.50%-99.52%)	99.50% (99.48%-99.52%)
	0	precision	22.9% (22.5%-24.9%)	21.0% (20.6%-21.3%)	21.7% (21.3%-22.1%)	19.4% (19.1%-19.7%)	19.7% (19.2%-20.0%)
	0	recall	32.9% (29.0%-33.5%)	39.9% (39.4%-40.3%)	35.3% (34.7%-36.4%)	31.0% (30.4%-31.5%)	27.6% (27.2%-28.3%)
	3	specificity	-	99.23% (98.35%-99.30%)	98.38% (98.16%-99.25%)	99.42% (99.39%-99.44%)	99.40% (99.29%-99.43%)
	3	precision	-	1.1% (1.0%-1.2%)	1.9% (1.8%-2.2%)	6.6% (6.4%-6.9%)	6.9% (6.5%-7.2%)
	3	recall	-	10.1% (9.2%-20.0%)	17.3% (8.8%-19.5%)	12.7% (12.1%-13.2%)	11.6% (11.1%-13.0%)
	6	specificity	-	-	99.25% (99.21%-99.27%)	99.22% (99.19%-99.26%)	99.23% (99.19%-99.26%)
	6	precision	-	-	1.9% (1.8%-2.0%)	4.0% (3.8%-4.2%)	4.4% (4.2%-4.5%)
	6	recall	-	-	12.2% (11.1%-12.9%)	11.1% (10.5%-11.6%)	10.0% (9.5%-10.5%)
	12	specificity	-	-	-	97.95% (97.66%-99.01%)	99.00% (98.76%-99.02%)
	12	precision	-	-	-	1.9% (1.8%-2.1%)	2.8% (2.6%-2.9%)
	12	recall	-	-	-	16.5% (9.0%-18.7%)	9.3% (8.8%-10.6%)

Table S4. Known risk factor disease codes.

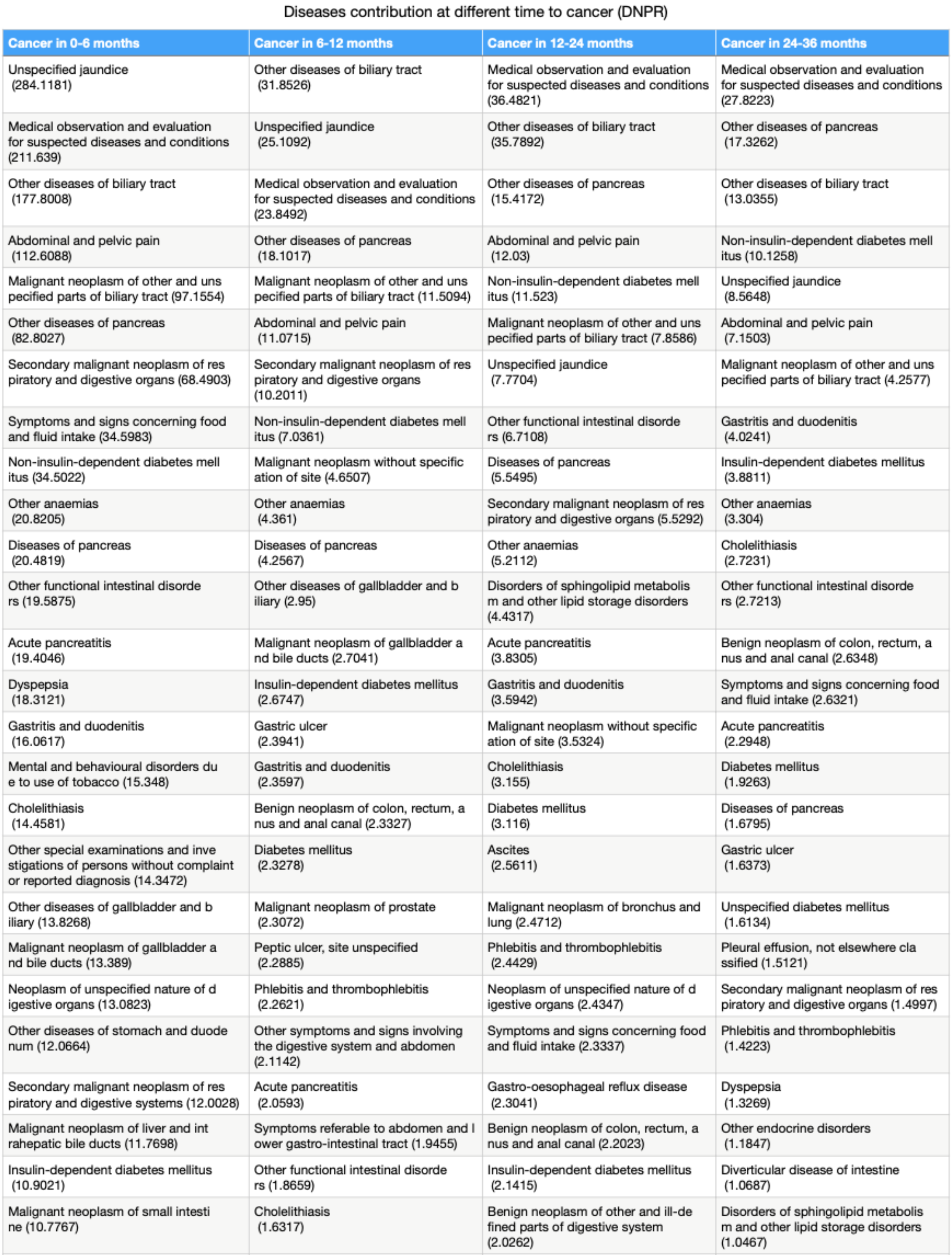
A subset of 23 diseases that have been considered as risk factors for pancreatic cancer (Yuan et al. 2020; Klein 2021) were chosen for the “known risk factor” analysis. Indeed, most of these are flagged by the IG feature extraction method to make a significant contribution to the ML prediction of cancer occurrence (**Figure 4**). These risk factors were used to train a separate time-series model ‘Transformer - known risk factors’ for comparison to the model trained on all ICD codes (Figure 3).



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

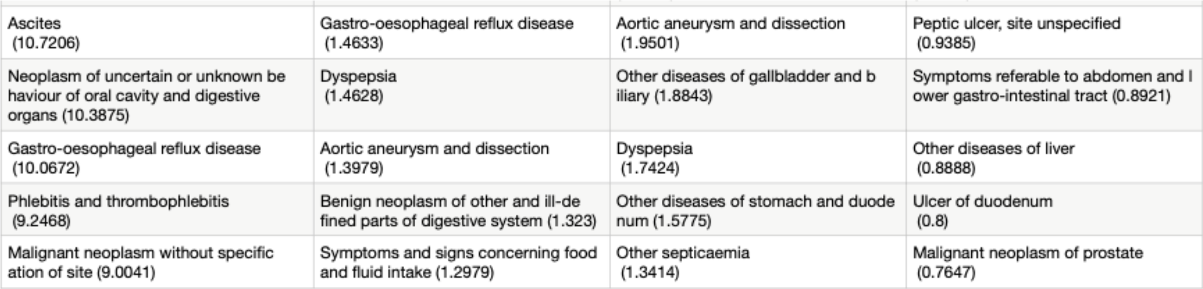
Table S5. Disease attribution without and with 3 months data exclusion

In order to discover the top diseases that contribute to our model’s risk prediction, we calculated the contribution score for all input features using integrated gradients (IG), an attribution method for neural networks. The IG contribution score (arbitrary units) was calculated for trajectories with cancer occurrence in the time windows 0-6 months, 6-12 months, 12-24 months and 24-36 months both without data exclusion (**A**) and with 3 months data exclusion (**B**).

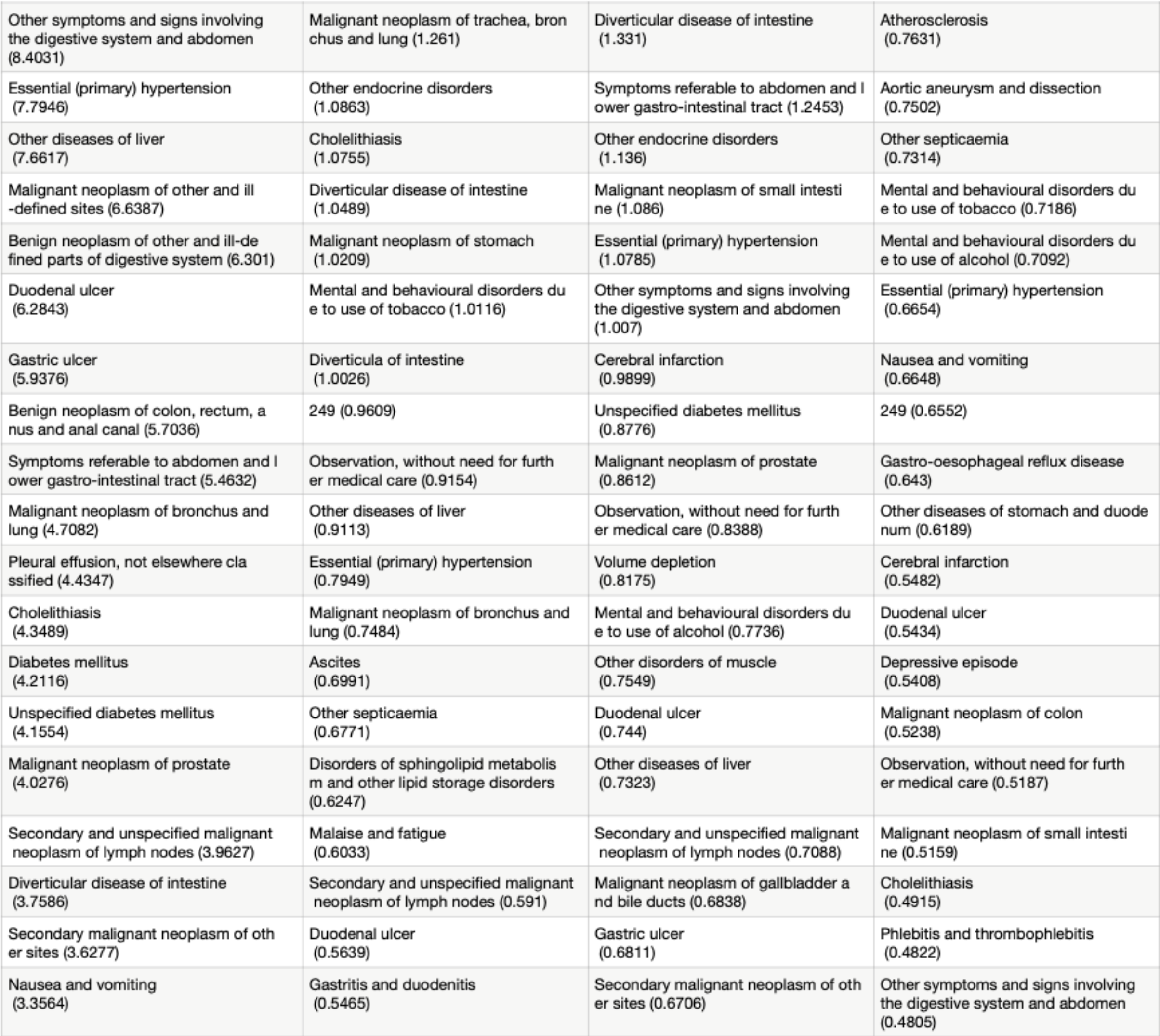


Diseases contribution at different time to cancer (DNPR)

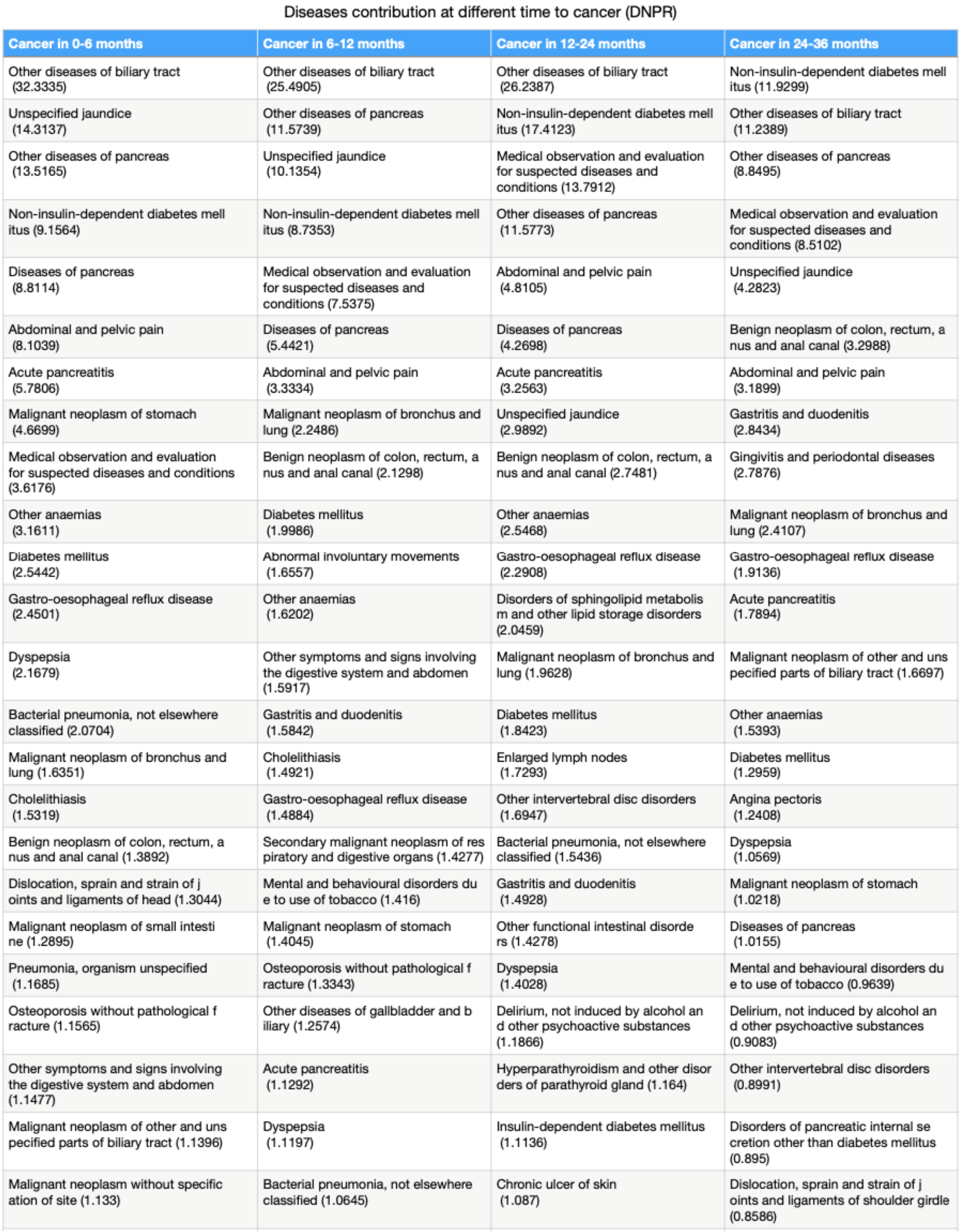
Cancer in 0-6 months	Cancer in 6-12 months	Cancer in 12-24 months	Cancer in 24-36 months
Unspecified jaundice (284.1181)	Other diseases of biliary tract (31.8526)	Medical observation and evaluation for suspected diseases and conditions (36.4821)	Medical observation and evaluation for suspected diseases and conditions (27.8223)
Medical observation and evaluation for suspected diseases and conditions (211.639)	Unspecified jaundice (25.1092)	Other diseases of biliary tract (35.7892)	Other diseases of pancreas (17.3262)
Other diseases of biliary tract (177.8008)	Medical observation and evaluation for suspected diseases and conditions (23.8492)	Other diseases of pancreas (15.4172)	Other diseases of biliary tract (13.0355)
Abdominal and pelvic pain (112.6088)	Other diseases of pancreas (18.1017)	Abdominal and pelvic pain (12.03)	Non-insulin-dependent diabetes mellitus (10.1258)
Malignant neoplasm of other and unspecified parts of biliary tract (97.1554)	Malignant neoplasm of other and unspecified parts of biliary tract (11.5094)	Non-insulin-dependent diabetes mellitus (11.523)	Unspecified jaundice (8.5648)
Other diseases of pancreas (82.8027)	Abdominal and pelvic pain (11.0715)	Malignant neoplasm of other and unspecified parts of biliary tract (7.8586)	Abdominal and pelvic pain (7.1503)
Secondary malignant neoplasm of respiratory and digestive organs (68.4903)	Secondary malignant neoplasm of respiratory and digestive organs (10.2011)	Unspecified jaundice (7.7704)	Malignant neoplasm of other and unspecified parts of biliary tract (4.2577)
Symptoms and signs concerning food and fluid intake (34.5983)	Non-insulin-dependent diabetes mellitus (7.0361)	Other functional intestinal disorders (6.7108)	Gastritis and duodenitis (4.0241)
Non-insulin-dependent diabetes mellitus (34.5022)	Malignant neoplasm without specification of site (4.6507)	Diseases of pancreas (5.5495)	Insulin-dependent diabetes mellitus (3.8811)
Other anaemias (20.8205)	Other anaemias (4.361)	Secondary malignant neoplasm of respiratory and digestive organs (5.5292)	Other anaemias (3.304)
Diseases of pancreas (20.4819)	Diseases of pancreas (4.2567)	Other anaemias (5.2112)	Cholelithiasis (2.7231)
Other functional intestinal disorders (19.5875)	Other diseases of gallbladder and biliary (2.95)	Disorders of sphingolipid metabolism and other lipid storage disorders (4.4317)	Other functional intestinal disorders (2.7213)
Acute pancreatitis (19.4046)	Malignant neoplasm of gallbladder and bile ducts (2.7041)	Acute pancreatitis (3.8305)	Benign neoplasm of colon, rectum, anus and anal canal (2.6348)
Dyspepsia (18.3121)	Insulin-dependent diabetes mellitus (2.6747)	Gastritis and duodenitis (3.5942)	Symptoms and signs concerning food and fluid intake (2.6321)
Gastritis and duodenitis (16.0617)	Gastric ulcer (2.3941)	Malignant neoplasm without specification of site (3.5324)	Acute pancreatitis (2.2948)
Mental and behavioural disorders due to use of tobacco (15.348)	Gastritis and duodenitis (2.3597)	Cholelithiasis (3.155)	Diabetes mellitus (1.9263)
Cholelithiasis (14.4581)	Benign neoplasm of colon, rectum, anus and anal canal (2.3327)	Diabetes mellitus (3.116)	Diseases of pancreas (1.6795)
Other special examinations and investigations of persons without complaint or reported diagnosis (14.3472)	Diabetes mellitus (2.3278)	Ascites (2.5611)	Gastric ulcer (1.6373)
Other diseases of gallbladder and biliary (13.8268)	Malignant neoplasm of prostate (2.3072)	Malignant neoplasm of bronchus and lung (2.4712)	Unspecified diabetes mellitus (1.6134)
Malignant neoplasm of gallbladder and bile ducts (13.389)	Peptic ulcer, site unspecified (2.2885)	Phlebitis and thrombophlebitis (2.4429)	Pleural effusion, not elsewhere classified (1.5121)
Neoplasm of unspecified nature of digestive organs (13.0823)	Phlebitis and thrombophlebitis (2.2621)	Neoplasm of unspecified nature of digestive organs (2.4347)	Secondary malignant neoplasm of respiratory and digestive organs (1.4997)
Other diseases of stomach and duodenum (12.0664)	Other symptoms and signs involving the digestive system and abdomen (2.1142)	Symptoms and signs concerning food and fluid intake (2.3337)	Phlebitis and thrombophlebitis (1.4223)
Secondary malignant neoplasm of respiratory and digestive systems (12.0028)	Acute pancreatitis (2.0593)	Gastro-oesophageal reflux disease (2.3041)	Dyspepsia (1.3269)
Malignant neoplasm of liver and intrahepatic bile ducts (11.7698)	Symptoms referable to abdomen and lower gastro-intestinal tract (1.9455)	Benign neoplasm of colon, rectum, anus and anal canal (2.2023)	Other endocrine disorders (1.1847)
Insulin-dependent diabetes mellitus (10.9021)	Other functional intestinal disorders (1.8659)	Insulin-dependent diabetes mellitus (2.1415)	Diverticular disease of intestine (1.0687)
Malignant neoplasm of small intestine (10.7767)	Cholelithiasis (1.6317)	Benign neoplasm of other and ill-defined parts of digestive system (2.0262)	Disorders of sphingolipid metabolism and other lipid storage disorders (1.0467)



Ascites (10.7206)	Gastro-oesophageal reflux disease (1.4633)	Aortic aneurysm and dissection (1.9501)	Peptic ulcer, site unspecified (0.9385)
Neoplasm of uncertain or unknown be haviour of oral cavity and digestive organs (10.3875)	Dyspepsia (1.4628)	Other diseases of gallbladder and b iliary (1.8843)	Symptoms referable to abdomen and l ower gastro-intestinal tract (0.8921)
Gastro-oesophageal reflux disease (10.0672)	Aortic aneurysm and dissection (1.3979)	Dyspepsia (1.7424)	Other diseases of liver (0.8888)
Phlebitis and thrombophlebitis (9.2468)	Benign neoplasm of other and ill-de fined parts of digestive system (1.323)	Other diseases of stomach and duode num (1.5775)	Ulcer of duodenum (0.8)
Malignant neoplasm without specific ation of site (9.0041)	Symptoms and signs concerning food and fluid intake (1.2979)	Other septicaemia (1.3414)	Malignant neoplasm of prostate (0.7647)

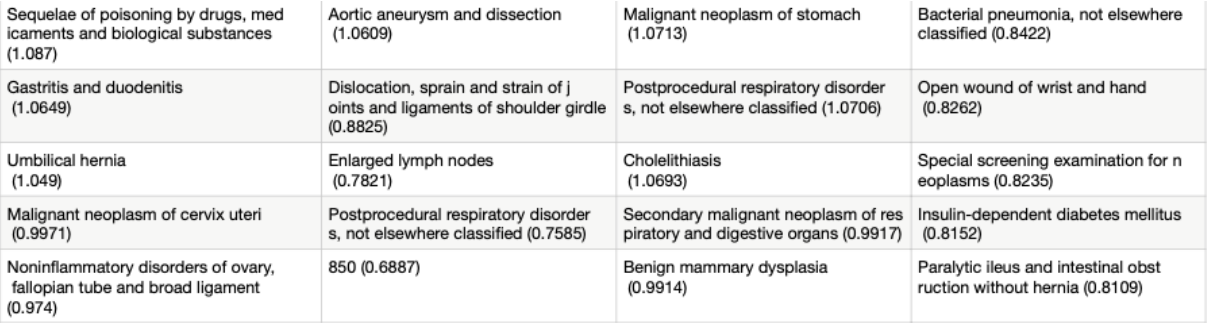


Other symptoms and signs involving the digestive system and abdomen (8.4031)	Malignant neoplasm of trachea, bron chus and lung (1.261)	Diverticular disease of intestine (1.331)	Atherosclerosis (0.7631)
Essential (primary) hypertension (7.7946)	Other endocrine disorders (1.0863)	Symptoms referable to abdomen and l ower gastro-intestinal tract (1.2453)	Aortic aneurysm and dissection (0.7502)
Other diseases of liver (7.6617)	Cholelithiasis (1.0755)	Other endocrine disorders (1.136)	Other septicaemia (0.7314)
Malignant neoplasm of other and ill -defined sites (6.6387)	Diverticular disease of intestine (1.0489)	Malignant neoplasm of small intesti ne (1.086)	Mental and behavioural disorders du e to use of tobacco (0.7186)
Benign neoplasm of other and ill-de fined parts of digestive system (6.301)	Malignant neoplasm of stomach (1.0209)	Essential (primary) hypertension (1.0785)	Mental and behavioural disorders du e to use of alcohol (0.7092)
Duodenal ulcer (6.2843)	Mental and behavioural disorders du e to use of tobacco (1.0116)	Other symptoms and signs involving the digestive system and abdomen (1.007)	Essential (primary) hypertension (0.6654)
Gastric ulcer (5.9376)	Diverticula of intestine (1.0026)	Cerebral infarction (0.9899)	Nausea and vomiting (0.6648)
Benign neoplasm of colon, rectum, a nus and anal canal (5.7036)	249 (0.9609)	Unspecified diabetes mellitus (0.8776)	249 (0.6552)
Symptoms referable to abdomen and l ower gastro-intestinal tract (5.4632)	Observation, without need for furth er medical care (0.9154)	Malignant neoplasm of prostate (0.8612)	Gastro-oesophageal reflux disease (0.643)
Malignant neoplasm of bronchus and lung (4.7082)	Other diseases of liver (0.9113)	Observation, without need for furth er medical care (0.8388)	Other diseases of stomach and duode num (0.6189)
Pleural effusion, not elsewhere cla ssified (4.4347)	Essential (primary) hypertension (0.7949)	Volume depletion (0.8175)	Cerebral infarction (0.5482)
Cholelithiasis (4.3489)	Malignant neoplasm of bronchus and lung (0.7484)	Mental and behavioural disorders du e to use of alcohol (0.7736)	Duodenal ulcer (0.5434)
Diabetes mellitus (4.2116)	Ascites (0.6991)	Other disorders of muscle (0.7549)	Depressive episode (0.5408)
Unspecified diabetes mellitus (4.1554)	Other septicaemia (0.6771)	Duodenal ulcer (0.744)	Malignant neoplasm of colon (0.5238)
Malignant neoplasm of prostate (4.0276)	Disorders of sphingolipid metabolis m and other lipid storage disorders (0.6247)	Other diseases of liver (0.7323)	Observation, without need for furth er medical care (0.5187)
Secondary and unspecified malignant neoplasm of lymph nodes (3.9627)	Malaise and fatigue (0.6033)	Secondary and unspecified malignant neoplasm of lymph nodes (0.7088)	Malignant neoplasm of small intesti ne (0.5159)
Diverticular disease of intestine (3.7586)	Secondary and unspecified malignant neoplasm of lymph nodes (0.591)	Malignant neoplasm of gallbladder a nd bile ducts (0.6838)	Cholelithiasis (0.4915)
Secondary malignant neoplasm of oth er sites (3.6277)	Duodenal ulcer (0.5639)	Gastric ulcer (0.6811)	Phlebitis and thrombophlebitis (0.4822)
Nausea and vomiting (3.3564)	Gastritis and duodenitis (0.5465)	Secondary malignant neoplasm of oth er sites (0.6706)	Other symptoms and signs involving the digestive system and abdomen (0.4805)

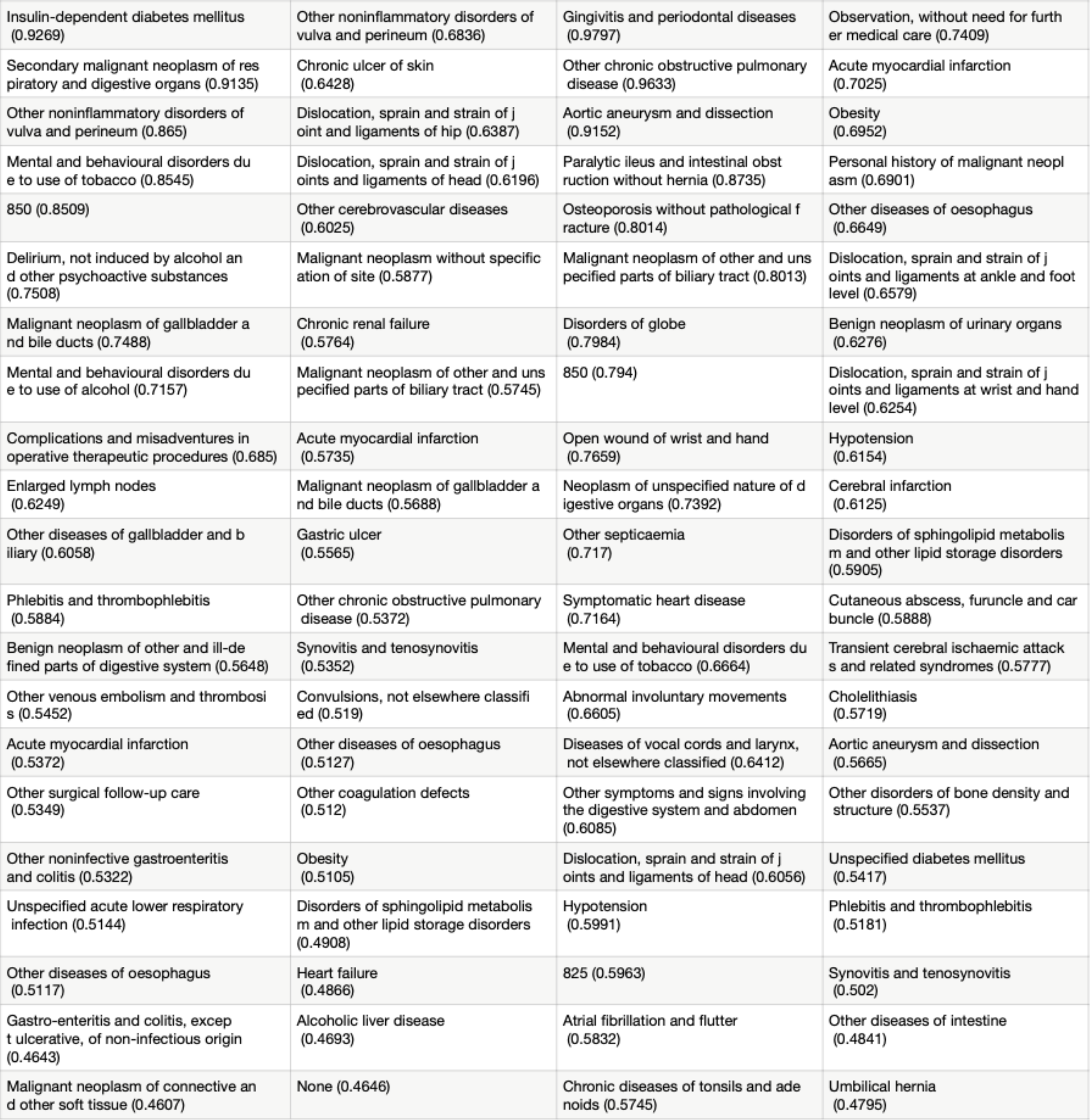


Diseases contribution at different time to cancer (DNPR)

Cancer in 0-6 months	Cancer in 6-12 months	Cancer in 12-24 months	Cancer in 24-36 months
Other diseases of biliary tract (32.3335)	Other diseases of biliary tract (25.4905)	Other diseases of biliary tract (26.2387)	Non-insulin-dependent diabetes mellitus (11.9299)
Unspecified jaundice (14.3137)	Other diseases of pancreas (11.5739)	Non-insulin-dependent diabetes mellitus (17.4123)	Other diseases of biliary tract (11.2389)
Other diseases of pancreas (13.5165)	Unspecified jaundice (10.1354)	Medical observation and evaluation for suspected diseases and conditions (13.7912)	Other diseases of pancreas (8.8495)
Non-insulin-dependent diabetes mellitus (9.1564)	Non-insulin-dependent diabetes mellitus (8.7353)	Other diseases of pancreas (11.5773)	Medical observation and evaluation for suspected diseases and conditions (8.5102)
Diseases of pancreas (8.8114)	Medical observation and evaluation for suspected diseases and conditions (7.5375)	Abdominal and pelvic pain (4.8105)	Unspecified jaundice (4.2823)
Abdominal and pelvic pain (8.1039)	Diseases of pancreas (5.4421)	Diseases of pancreas (4.2698)	Benign neoplasm of colon, rectum, anus and anal canal (3.2988)
Acute pancreatitis (5.7806)	Abdominal and pelvic pain (3.3334)	Acute pancreatitis (3.2563)	Abdominal and pelvic pain (3.1899)
Malignant neoplasm of stomach (4.6699)	Malignant neoplasm of bronchus and lung (2.2486)	Unspecified jaundice (2.9892)	Gastritis and duodenitis (2.8434)
Medical observation and evaluation for suspected diseases and conditions (3.6176)	Benign neoplasm of colon, rectum, anus and anal canal (2.1298)	Benign neoplasm of colon, rectum, anus and anal canal (2.7481)	Gingivitis and periodontal diseases (2.7876)
Other anaemias (3.1611)	Diabetes mellitus (1.9986)	Other anaemias (2.5468)	Malignant neoplasm of bronchus and lung (2.4107)
Diabetes mellitus (2.5442)	Abnormal involuntary movements (1.6557)	Gastro-oesophageal reflux disease (2.2908)	Gastro-oesophageal reflux disease (1.9136)
Gastro-oesophageal reflux disease (2.4501)	Other anaemias (1.6202)	Disorders of sphingolipid metabolism and other lipid storage disorders (2.0459)	Acute pancreatitis (1.7894)
Dyspepsia (2.1679)	Other symptoms and signs involving the digestive system and abdomen (1.5917)	Malignant neoplasm of bronchus and lung (1.9628)	Malignant neoplasm of other and unspecified parts of biliary tract (1.6697)
Bacterial pneumonia, not elsewhere classified (2.0704)	Gastritis and duodenitis (1.5842)	Diabetes mellitus (1.8423)	Other anaemias (1.5393)
Malignant neoplasm of bronchus and lung (1.6351)	Cholelithiasis (1.4921)	Enlarged lymph nodes (1.7293)	Diabetes mellitus (1.2959)
Cholelithiasis (1.5319)	Gastro-oesophageal reflux disease (1.4884)	Other intervertebral disc disorders (1.6947)	Angina pectoris (1.2408)
Benign neoplasm of colon, rectum, anus and anal canal (1.3892)	Secondary malignant neoplasm of respiratory and digestive organs (1.4277)	Bacterial pneumonia, not elsewhere classified (1.5436)	Dyspepsia (1.0569)
Dislocation, sprain and strain of joints and ligaments of head (1.3044)	Mental and behavioural disorders due to use of tobacco (1.416)	Gastritis and duodenitis (1.4928)	Malignant neoplasm of stomach (1.0218)
Malignant neoplasm of small intestine (1.2895)	Malignant neoplasm of stomach (1.4045)	Other functional intestinal disorders (1.4278)	Diseases of pancreas (1.0155)
Pneumonia, organism unspecified (1.1685)	Osteoporosis without pathological fracture (1.3343)	Dyspepsia (1.4028)	Mental and behavioural disorders due to use of tobacco (0.9639)
Osteoporosis without pathological fracture (1.1565)	Other diseases of gallbladder and biliary (1.2574)	Delirium, not induced by alcohol and other psychoactive substances (1.1866)	Delirium, not induced by alcohol and other psychoactive substances (0.9083)
Other symptoms and signs involving the digestive system and abdomen (1.1477)	Acute pancreatitis (1.1292)	Hyperparathyroidism and other disorders of parathyroid gland (1.164)	Other intervertebral disc disorders (0.8991)
Malignant neoplasm of other and unspecified parts of biliary tract (1.1396)	Dyspepsia (1.1197)	Insulin-dependent diabetes mellitus (1.1136)	Disorders of pancreatic internal secretion other than diabetes mellitus (0.895)
Malignant neoplasm without specification of site (1.133)	Bacterial pneumonia, not elsewhere classified (1.0645)	Chronic ulcer of skin (1.087)	Dislocation, sprain and strain of joints and ligaments of shoulder girdle (0.8586)



Sequelae of poisoning by drugs, medicaments and biological substances (1.087)	Aortic aneurysm and dissection (1.0609)	Malignant neoplasm of stomach (1.0713)	Bacterial pneumonia, not elsewhere classified (0.8422)
Gastritis and duodenitis (1.0649)	Dislocation, sprain and strain of joints and ligaments of shoulder girdle (0.8825)	Postprocedural respiratory disorders, not elsewhere classified (1.0706)	Open wound of wrist and hand (0.8262)
Umbilical hernia (1.049)	Enlarged lymph nodes (0.7821)	Cholelithiasis (1.0693)	Special screening examination for neoplasms (0.8235)
Malignant neoplasm of cervix uteri (0.9971)	Postprocedural respiratory disorders, not elsewhere classified (0.7585)	Secondary malignant neoplasm of respiratory and digestive organs (0.9917)	Insulin-dependent diabetes mellitus (0.8152)
Noninflammatory disorders of ovary, fallopian tube and broad ligament (0.974)	850 (0.6887)	Benign mammary dysplasia (0.9914)	Paralytic ileus and intestinal obstruction without hernia (0.8109)



Insulin-dependent diabetes mellitus (0.9269)	Other noninflammatory disorders of vulva and perineum (0.6836)	Gingivitis and periodontal diseases (0.9797)	Observation, without need for further medical care (0.7409)
Secondary malignant neoplasm of respiratory and digestive organs (0.9135)	Chronic ulcer of skin (0.6428)	Other chronic obstructive pulmonary disease (0.9633)	Acute myocardial infarction (0.7025)
Other noninflammatory disorders of vulva and perineum (0.865)	Dislocation, sprain and strain of joint and ligaments of hip (0.6387)	Aortic aneurysm and dissection (0.9152)	Obesity (0.6952)
Mental and behavioural disorders due to use of tobacco (0.8545)	Dislocation, sprain and strain of joints and ligaments of head (0.6196)	Paralytic ileus and intestinal obstruction without hernia (0.8735)	Personal history of malignant neoplasm (0.6901)
850 (0.8509)	Other cerebrovascular diseases (0.6025)	Osteoporosis without pathological fracture (0.8014)	Other diseases of oesophagus (0.6649)
Delirium, not induced by alcohol and other psychoactive substances (0.7508)	Malignant neoplasm without specification of site (0.5877)	Malignant neoplasm of other and unspecified parts of biliary tract (0.8013)	Dislocation, sprain and strain of joints and ligaments at ankle and foot level (0.6579)
Malignant neoplasm of gallbladder and bile ducts (0.7488)	Chronic renal failure (0.5764)	Disorders of globe (0.7984)	Benign neoplasm of urinary organs (0.6276)
Mental and behavioural disorders due to use of alcohol (0.7157)	Malignant neoplasm of other and unspecified parts of biliary tract (0.5745)	850 (0.794)	Dislocation, sprain and strain of joints and ligaments at wrist and hand level (0.6254)
Complications and misadventures in operative therapeutic procedures (0.685)	Acute myocardial infarction (0.5735)	Open wound of wrist and hand (0.7659)	Hypotension (0.6154)
Enlarged lymph nodes (0.6249)	Malignant neoplasm of gallbladder and bile ducts (0.5688)	Neoplasm of unspecified nature of digestive organs (0.7392)	Cerebral infarction (0.6125)
Other diseases of gallbladder and biliary (0.6058)	Gastric ulcer (0.5565)	Other septicaemia (0.717)	Disorders of sphingolipid metabolism and other lipid storage disorders (0.5905)
Phlebitis and thrombophlebitis (0.5884)	Other chronic obstructive pulmonary disease (0.5372)	Symptomatic heart disease (0.7164)	Cutaneous abscess, furuncle and carbuncle (0.5888)
Benign neoplasm of other and ill-defined parts of digestive system (0.5648)	Synovitis and tenosynovitis (0.5352)	Mental and behavioural disorders due to use of tobacco (0.6664)	Transient cerebral ischaemic attacks and related syndromes (0.5777)
Other venous embolism and thrombosis (0.5452)	Convulsions, not elsewhere classified (0.519)	Abnormal involuntary movements (0.6605)	Cholelithiasis (0.5719)
Acute myocardial infarction (0.5372)	Other diseases of oesophagus (0.5127)	Diseases of vocal cords and larynx, not elsewhere classified (0.6412)	Aortic aneurysm and dissection (0.5665)
Other surgical follow-up care (0.5349)	Other coagulation defects (0.512)	Other symptoms and signs involving the digestive system and abdomen (0.6085)	Other disorders of bone density and structure (0.5537)
Other noninfective gastroenteritis and colitis (0.5322)	Obesity (0.5105)	Dislocation, sprain and strain of joints and ligaments of head (0.6056)	Unspecified diabetes mellitus (0.5417)
Unspecified acute lower respiratory infection (0.5144)	Disorders of sphingolipid metabolism and other lipid storage disorders (0.4908)	Hypotension (0.5991)	Phlebitis and thrombophlebitis (0.5181)
Other diseases of oesophagus (0.5117)	Heart failure (0.4866)	825 (0.5963)	Synovitis and tenosynovitis (0.502)
Gastro-enteritis and colitis, except ulcerative, of non-infectious origin (0.4643)	Alcoholic liver disease (0.4693)	Atrial fibrillation and flutter (0.5832)	Other diseases of intestine (0.4841)
Malignant neoplasm of connective and other soft tissue (0.4607)	None (0.4646)	Chronic diseases of tonsils and adenoids (0.5745)	Umbilical hernia (0.4795)

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

Distribution of disease codes for a representative subset of diseases known 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, Gaussian fit to cancer diagnosis data) 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.

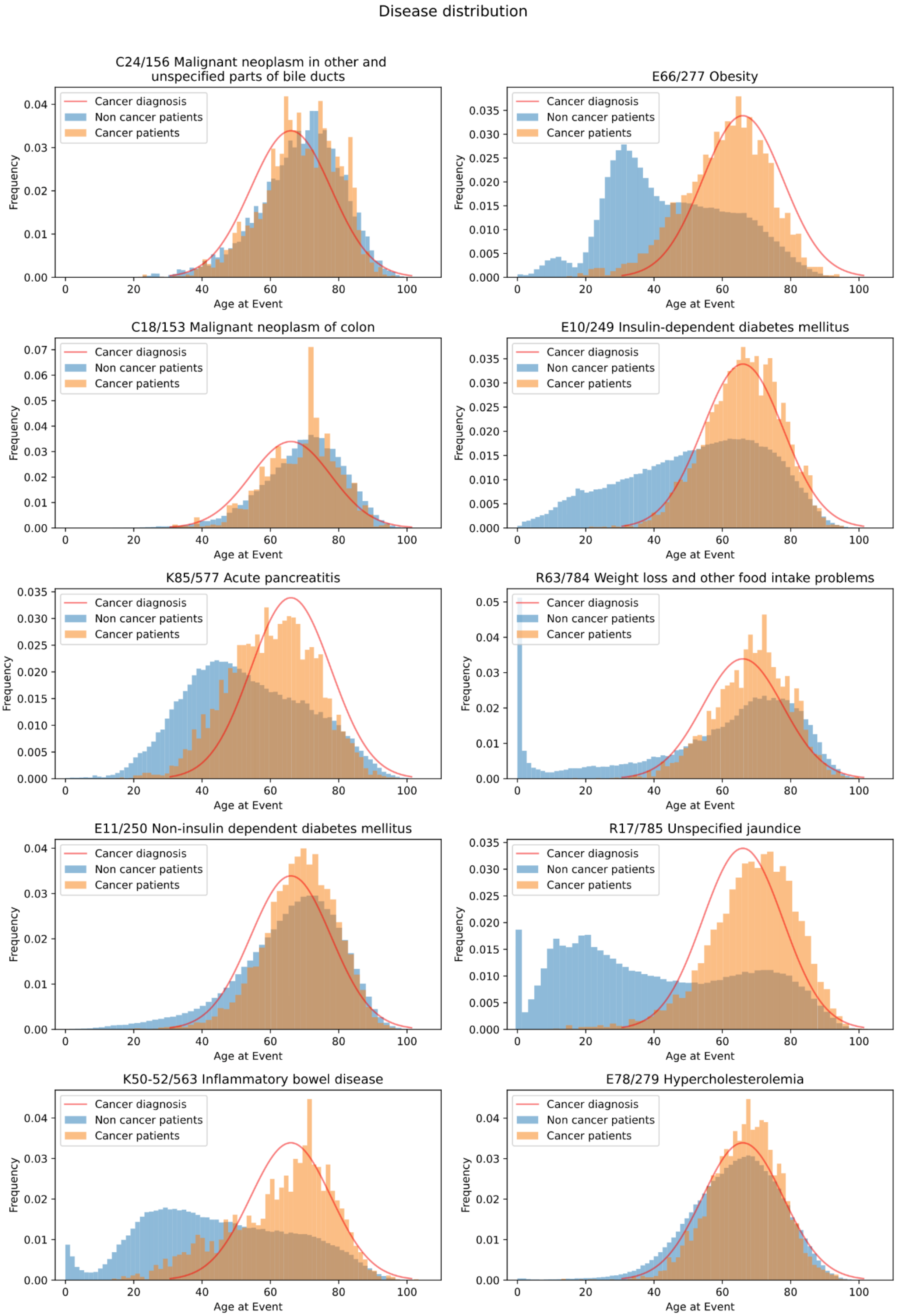

Disease distribution
C24/156 Malignant neoplasm in other and unspecified parts of bile ducts
E66/277 Obesity
C18/153 Malignant neoplasm of colon
E10/249 Insulin-dependent diabetes mellitus
K85/577 Acute pancreatitis
R63/784 Weight loss and other food intake problems
E11/250 Non-insulin dependent diabetes mellitus
R17/785 Unspecified jaundice
K50-52/563 Inflammatory bowel disease
E78/279 Hypercholesterolemia
Cancer diagnosis
Non cancer patients
Cancer patients
Frequency
Age at Event
0
20
40
60
80
100

Figure S4. Distribution of disease codes over years and age in the Danish (DK) and Boston (MGB) datasets.

Distribution of disease codes over time and age for the Danish DNPR (**A,C**) and Boston MGB datasets (**B,D**) for the pancreatic cancer ('cancer') and non-pancreatic-cancer ('non-cancer') cases. The disease code frequency is the total number of disease codes summed over all patients in the selected groups (cancer vs. non-cancer) divided by the total number of disease codes in the entire database.

(**A**) The DNPR dataset has both ICD-8 and ICD-10 disease codes. The transition from ICD-8 to ICD-10 occurred in 1994, after which the disease code frequency increased significantly over the years. This increase could be due to alterations in clinical coding practices or due to higher disease awareness in the population. In this study, we did not perform mapping from ICD-8 to ICD-10 codes. Instead, the model was trained on the non-mapped ICD-8 and ICD-9 codes for it to learn coding patterns independently of a mapping. (**B**) Disease distribution over time for the Boston MGB dataset. The dataset includes both ICD-9 and ICD-10 codes, for which we similarly did not apply any mapping. (**C**) Disease distribution over age for the Danish DNPR dataset showing an interesting increase of disease codes (all diseases) with age for the pancreatic cancer cases. (**D**) Disease distribution over age for the Boston MGB dataset.

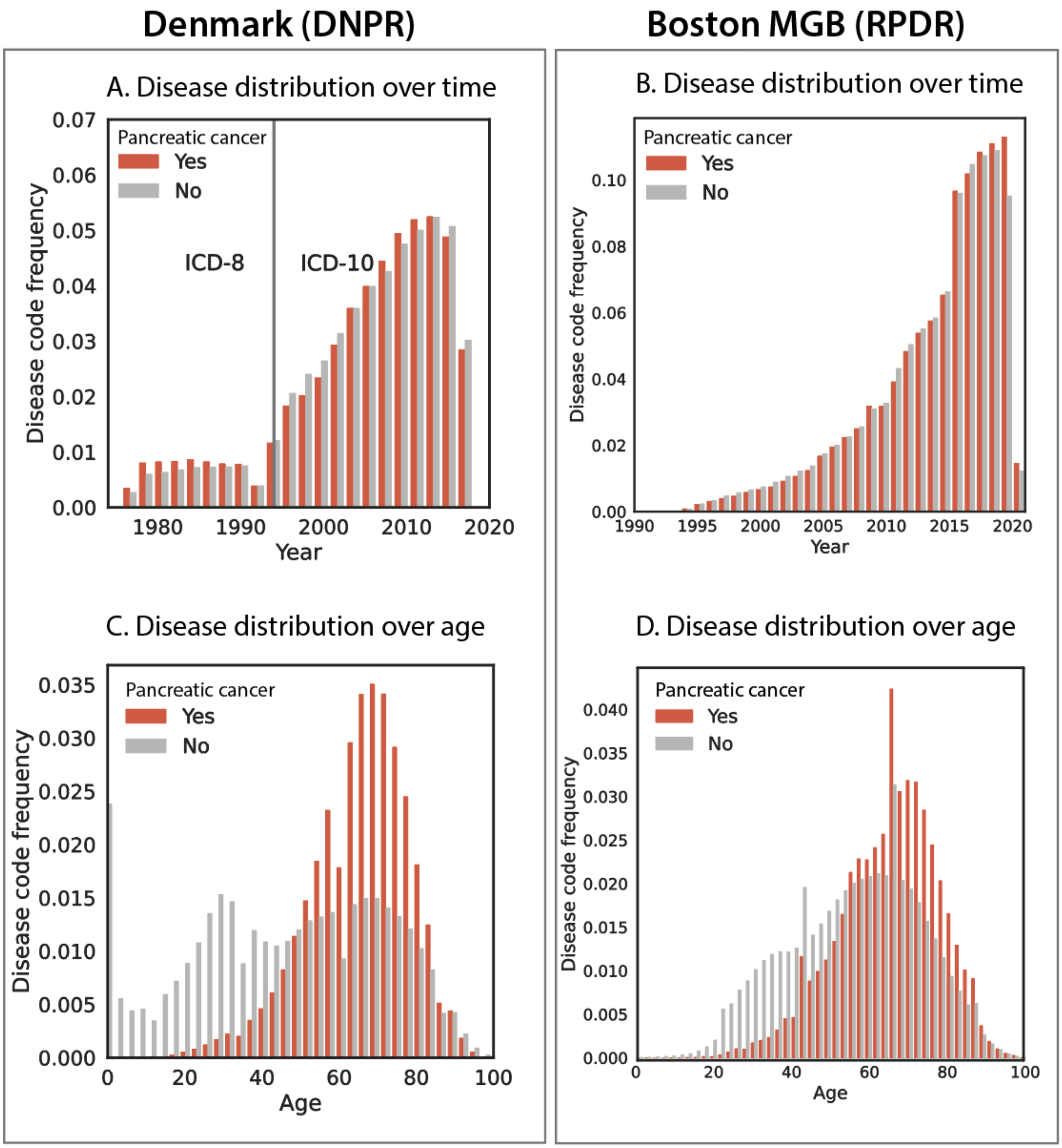
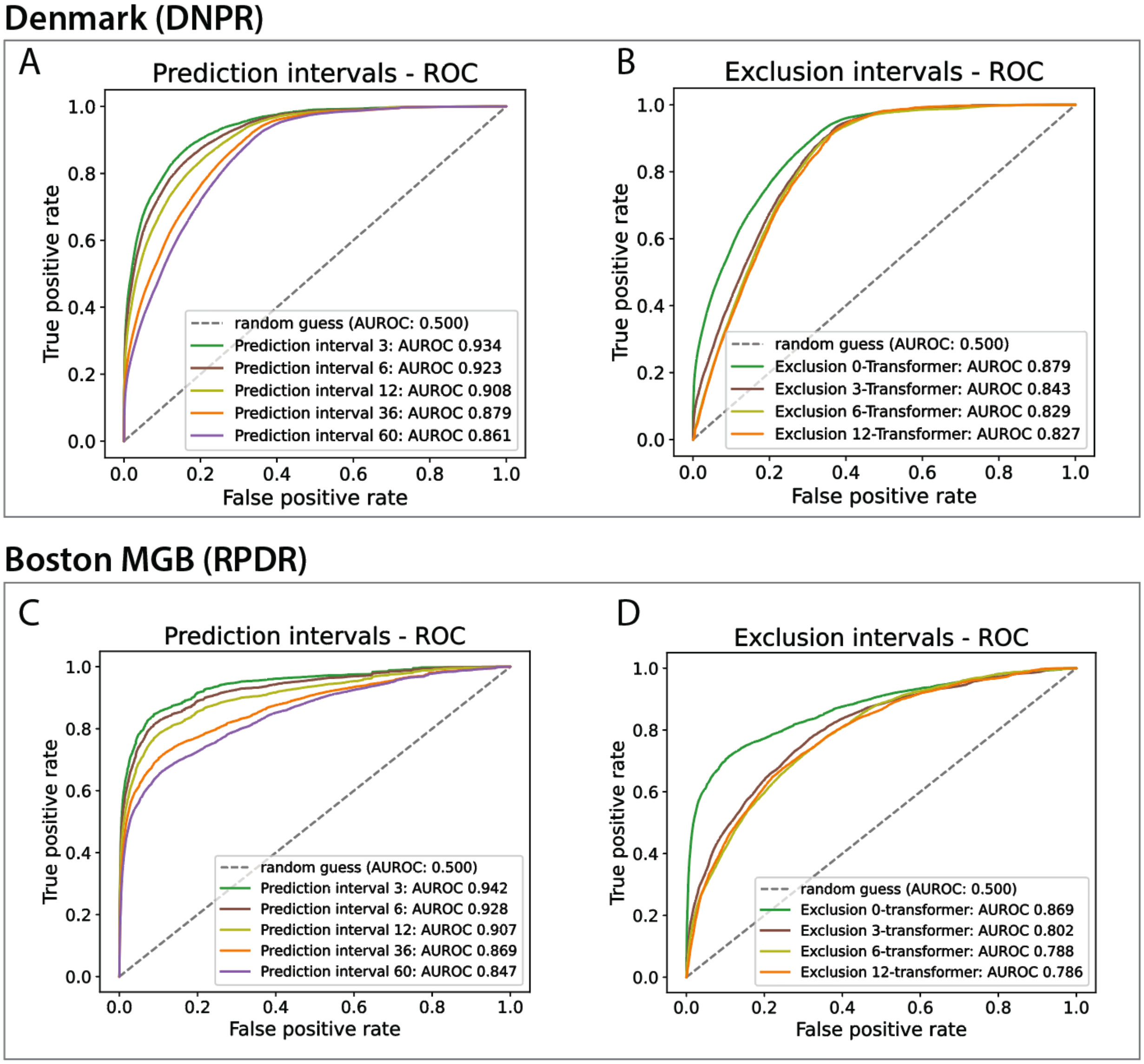

Denmark (DNPR)
Boston MGB (RPDR)
A. Disease distribution over time
Pancreatic cancer
Yes
No
ICD-8
ICD-10
Disease code frequency
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
1980
1990
2000
2010
2020
Year
B. Disease distribution over time
Pancreatic cancer
Yes
No
Disease code frequency
0.00
0.02
0.04
0.06
0.08
0.10
1990
1995
2000
2005
2010
2015
2020
Year
C. Disease distribution over age
Pancreatic cancer
Yes
No
Disease code frequency
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0
20
40
60
80
100
Age
D. Disease distribution over age
Pancreatic cancer
Yes
No
Disease code frequency
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.040
0
20
40
60
80
100
Age

Figure S5. ROC curves for the transformer model for different prediction and exclusion intervals.

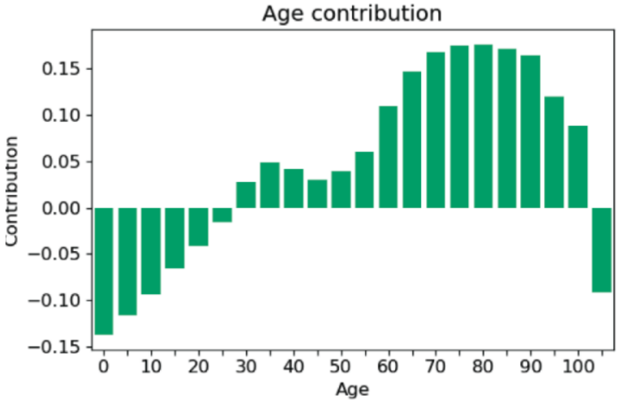
For the transformer model, ROC curves were analysed across different prediction intervals (3, 6, 12, 36 and 60 months) and exclusion intervals (0, 3, 6 and 12 months). As expected, it is more challenging to predict cancer occurrence in longer rather than shorter time intervals. We also see that it becomes more challenging to predict cancer outcomes with higher exclusion intervals.



(**A-B**) The DNPR 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 (diagonal line) would have very low precision for equal TPR and FPR (AUROC=0.5). Exclusion intervals are assessed in 0, 3, 6 or 12 months months. (**A**) 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 by (**B**) excluding from the input diseases diagnoses in the last 0, 3, 6 or 12 months prior to the diagnosis of pancreatic cancer. (**C-D**) The Boston MGB ROC curves for prediction intervals (**C**) and exclusion intervals (**D**).

Figure S6. Age as a contributing factor

The integrated gradient method was used to extract the contribution (arbitrary units) of patient age to the prediction at the time of assessment. This confirmed that the positive contribution to risk rises strongly from age 50. As for the disease contributions, the age contribution was calculated in relation to the 3 year (after the time of assessment/prediction) cancer risk.



Result S1: 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.

Net Benefit = Average benefit for each correctly identified cancer patient * TP
− Monitoring expense for each high-risk patient * P
− Basic cost per enrollee * N

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).

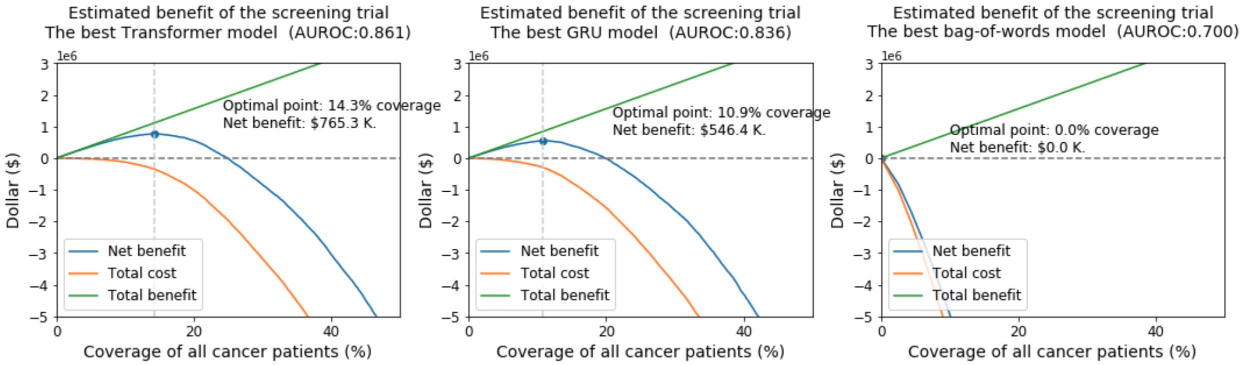


Figure S7. An estimate of financial benefits for different models. We analyzed each possible operational point and calculated the corresponding cost and benefit, using ballpark estimates. We plotted the net benefits as a function of coverage of cancer patients, i.e. recall or sensitivity. Covering more cancer patients plausibly leads to a larger total benefit, but the total cost also increases. The optimal point is picked for 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 but hypothetical design of a screening trial is intended to guide the debate and ultimate decisions regarding implementation with clinicians and healthcare professionals. However, this calculation is based on roughly estimated 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 substantial geographical variability in reimbursement even for the same CPT/billing codes.