Pancreatic cancer risk predicted from disease trajectories using deep learning

Davide Placido[1,‡,], Bo Yuan[2,3,4,‡,], Jessica X. Hjaltelin[1,‡,], Chunlei Zheng[11,12,‡], 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], Siamack Ayandeh[11], Mary Brophy[11,12], Nhan Do[11,12], Peter Kraft[6], Brian M. Wolpin[2,3,10], Nathanael Fillmore[2,3,11,12], 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] Genentech Inc., South San Francisco, CA, USA
[10] Brigham and Women's Hospital, Boston, USA
[11] VA Boston Healthcare System, Boston, MA, USA
[12] Boston University School of Medicine, Boston, MA, USA

‡ joint first authors: davide.placido@cpr.ku.dk, boyuan@g.harvard.edu, jessica.hu@cpr.ku.dk, chunlei.zheng@va.gov
* joint senior authors: soren.brunak@cpr.ku.dk, chris@sanderlab.org

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 6 million patient records with 24,000 pancreatic cancer cases in the Danish National Patient Registry (DNPR) and, for comparison, a dataset of three million records with 3,900 pancreatic cancer cases in the United States Department of Veterans Affairs (US-VA) healthcare system. 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.

For cancer occurrence within 36 months, the performance of the best model (AUROC=0.88, DNPR), trained and tested on disease trajectories, exceeds that of a model without longitudinal information (AUROC=0.85, DNPR). Performance decreases when disease events within a 3 month window before cancer diagnosis are excluded from training (AUROC[3m]=0.83). Independent training and testing on the US-VA dataset reaches comparable performance (AUROC=0.78, AUROC[3m]=0.76). 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 prediction-surveillance programs based on risk assessment in a large population followed by affordable surveillance of a relatively small number of patients at highest risk. Use of 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.

[[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 high risk due to family history or rare inherited pathogenic variants 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 surveillance programs for early detection. Real-world data are usually noisy and the predictive power can be affected by various confounding factors, therefore, the development of realistic risk prediction methods requires 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 successfully 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; Thorsen-Meyer et al. 2022).

[[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 technology by focusing on the time sequence of clinical events and by predicting the risk of cancer occurrence in specific time intervals after assessment. This investigation was first carried out using the Danish National Patient Registry (DNPR), which contains data from hospital admissions including out-patient visits from 1977 to 2018 (41 years), including diagnosis codes, for a total of 8.6 million patients. About 40,000 of these 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 machine learning 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 diagnosis codes are most informative of cancer risk - not as isolated factors but in the context of the person's complete history of disease codes. Finally, we propose a practical scenario for surveillance programs, taking into consideration the availability of real-world data, the accuracy of prediction on such data, the scope of a surveillance program, the likely cost and success rate of surveillance methods and the overall potential benefit of early treatment (Supplement, **Result S1**).

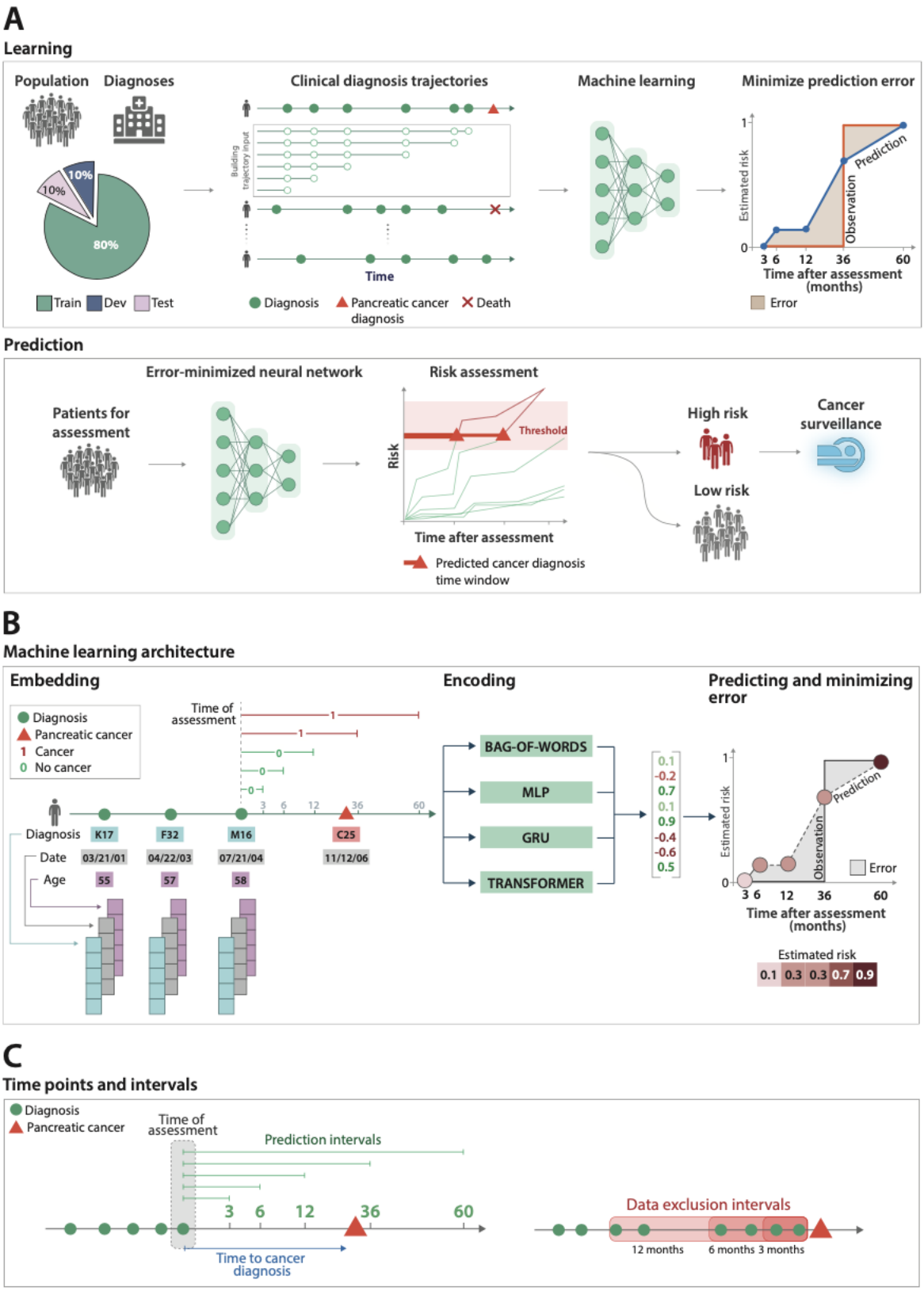


Figure 1. Training and prediction of pancreatic cancer risk from disease trajectories. (**A**) Learning: The general machine learning workflow starts with partitioning the data into training set (Train), development set (Dev) and test set

(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 accurately predict 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 surveillance 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 lower-dimensional real-number continuous space. The encoding machine extracts information from a disease history and summarizes each sequence in a characteristic fingerprint in the latent space (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 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 diagnosis 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-DNPR]]

We used disease trajectory data from the Danish National Patient Registry (DNPR). Demographic information was obtained by linkage to the Central Person Registry, which is possible via the personal identification number introduced in 1968 and uniquely identifies any Danish citizen over the 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 diagnostic codes with explicit time stamps for each hospital contact comprising diagnoses down to the three-character category in the ICD hierarchy. The accuracy of ICD codes for pancreatic cancer, which were used to label cases in training, was estimated to be above 88% (Methods). 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 diagnosed at a median age of 70 years (49.5% man, 50.5 female) (**Figure 2**, **Table S1**).

[[Dataset of disease trajectories: US-VA]]

For validation in another health care system, we used clinical records from the United States Department of Veterans Affairs (US-VA) hospital system via the VA Corporate Data Warehouse, which integrates both electronic health record and cancer registry data from VA facilities nationwide. As in the Danish dataset, we also used explicit longitudinal records from US-VA, i.e., trajectories of ICD codes with explicit time stamps.

Cases of pancreatic cancer were ascertained using VA cancer registry data. The accuracy of the US-VA cancer registry data for pancreatic cancer, which was used to label cases in training, was estimated to be 95-96% (Methods). We used data from 1999 to 2020 and filtered out patients with less than five recorded diagnosis codes (**Figure S1b, Figure 2**). The selected dataset has 3.0 million patients with 3,864 pancreatic cancer patients (**Figure S1b**). On average, the health records in the US-VA dataset have shorter (median 12 years) but substantially denser disease history (median 188 records per patient). For US-VA, the median age of pancreatic cancer diagnosis is 68.0 years (96.7% male, 3.3% female). These numerical differences likely reflect the differences in population (entire population in Denmark vs military veterans in US-VA), in the health care systems in the two datasets, 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: (1) **input:** data for each event in a trajectory (diagnosis 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 health/disease records 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 ending at a series of time points, with 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 a level-3 ICD code. To extract informative features from such high-dimensional input, the machine learning process is designed to embed the categorical input vectors into a continuous, lower-dimensional space. Temporal information, i.e. diagnosis dates and age at each diagnosis are also embedded (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 to capture complex sequential 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 encoded latent space vector is then used as input to a multilayer fully-connected 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 and data exclusion]]

For each of the disease trajectories ending at time t_a, a 5-dimensional risk score is calculated, where each dimension represents the cumulative risk of cancer occurrence within a particular prediction window after t_a, e.g., ending at t_a + 12 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 separately predicts a cancer diagnosis to occur within 3, 6, 12, 36, or 60 months after the time t_a of risk assessment. In addition, to test the influence of close-to-cancer diagnosis 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 during training.

[[Scanning hyperparameters for each model type]]

To comprehensively test the performance of different types of machine learning 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 (MLP) 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 patients from the health system used to train and test the models.

Characteristics of Danish and US-VA dataset

General cohort information	Danish dataset	US-VA dataset
Dataset timeline	1977-2018	1999-2020
Total N patients	8,110,706	2,962,383
Male (%)	4,030,504 (49.7%)	2,538,762 (85.7%)
Female (%)	4,080,202 (50.3%)	423,621 (14.3%)
Median N disease codes per patient	22	188
Median length of trajectory in years	23.0	12.0
PC cohort information		
Total N patients	23,985	3,864
Male (%)	11,880 (49.5%)	3,736 (96.7%)
Female (%)	12,105 (50.5%)	128 (3.3%)
Median N disease codes per patient	18	121
Median length of trajectory in years	17.0	8.0
Median age at PC diagnosis	70.0	68.0
N disease codes 0-3 months pre-PC	95,358	125,305
N disease codes 3-6 months pre-PC	27,131	56,198
N disease codes 6-12 months pre-PC	38,109	97,911
N disease codes >12 months pre-PC	480,830	1,188,199

Abbreviations: PC: pancreatic cancer.

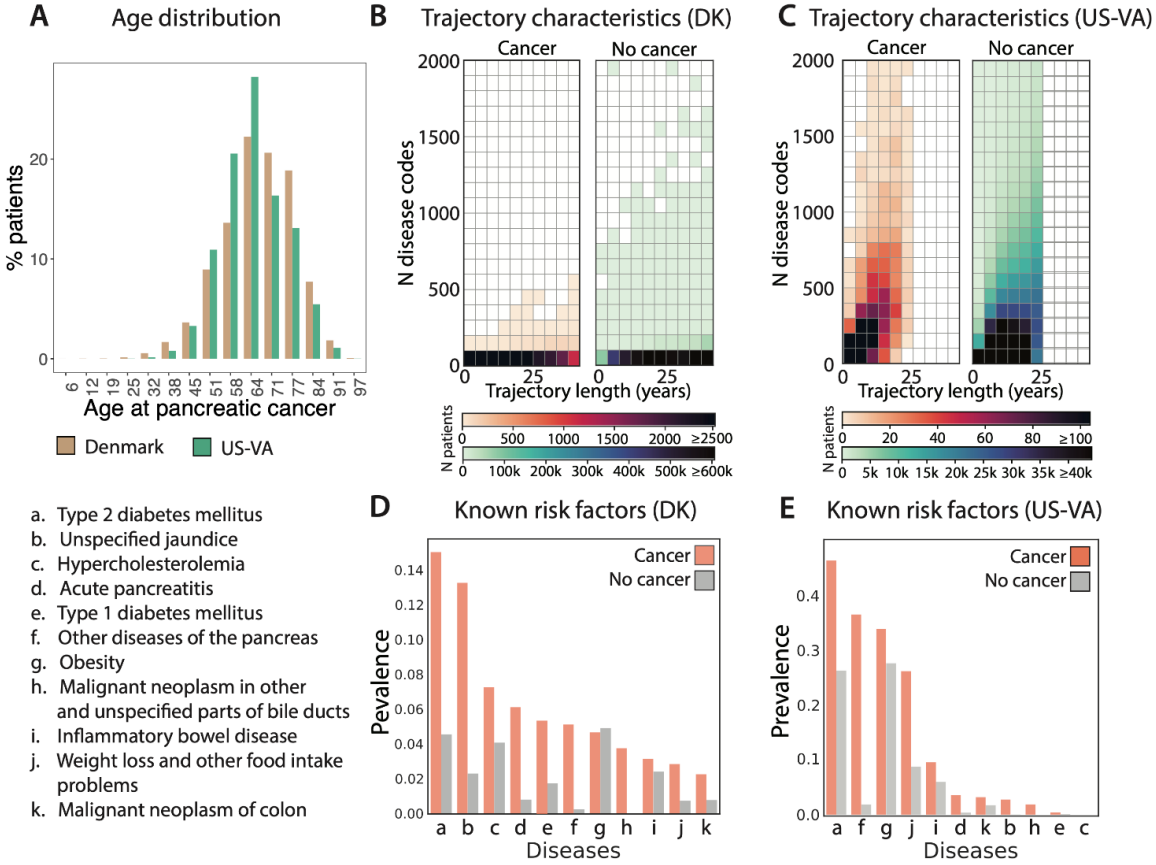


Figure 2. Characteristics of the Danish (DK) and US-Veterans Administration (VA) patient registries.

(**A**) Age distributions for age at pancreatic cancer diagnosis in the two cohorts. (**B,C**) The DK dataset (DNPR) has a longer median length of disease

trajectories, but lower median number of disease codes per patient compared to the VA dataset, so the machine learning process, independently in each dataset, has to cope with very different distributions of disease trajectories in terms of length of trajectories and density of the number of disease codes. Color level indicates the number of patients in a given bin. (**D,E**) Background check on the distribution disease code in the clinical records: prevalence of known risk factors in cancer vs. non-cancer patients in the DK (**D**) and VA (**E**) datasets, counting whether a disease code occurred at least once in a patient's history previous to their pancreatic cancer code (cancer) or two years previous to the end of data (no cancer).

Evaluation of model performance

[[Picking a best model - DK]]

We evaluated the different models trained in the Danish DNPR using the AUROC and relative risk (RR) curves (**Figure 3A-H**). All performance metrics are calculated on the basis of applying each trained risk assessment model to the test set. The test set is strictly withheld during training and hyperparameter search. In the final performance evaluation of different types of machine learning models on the test sets, the models which explicitly use and encode the time sequence of disease codes, i.e., GRU and Transformer, ranked highest by AUROC (**Figure 3A; 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]). In order to gain a better intuition regarding the impact of applying a model in a real case scenario, we also report the relative risk score (RR) of cancer patients in the high-risk group as predicted by the machine learning model.

The relative risk score is defined at a given operational decision point (Methods). This assesses by what factor a prediction method does better than a random model (**Figure 3B**). The relative risk (RR) for the 36 months prediction interval is 104.7 for the Transformer model (with time sequences), at an operational point defined by the N=1000 highest risk patients out of 1 million patients (0.1% at highest risk; notation: N1000).

[[3A Comparison with previous models]]

Earlier work also developed machine learning methods on real-world clinical records to predict 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 to extract

time-sequential longitudinal features. For comparison we implemented analogous approaches, a bag-of-words model and an 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 (**Figure 3A**). The RR is also much lower (2.1 and 26.6, respectively), compared to that of time-sequential models (e.g., 104.7 for Transformer).

[[3C Performance with data exclusion]]

Disease codes within a very short time before diagnosis of pancreatic cancer are most probably directly indicative such that even without any machine learning, well-trained clinicians would include pancreatic cancer in their differential diagnosis. Disease codes just prior to pancreatic cancer occurrence may also encompass it (e.g., neoplasm of the digestive tract) and thus directly reflect the label one wants to infer. To avoid undue influence of these disease codes in training, we therefore separately trained the models excluding input diseases diagnoses from the last 3/6/12 months prior to the diagnosis of pancreatic cancer ('data exclusion'). As expected, when training with data exclusion, the performance decreased from AUROC=0.879 to AUROC=0.843/0.829/0.827 for 3/6/12 months data for the best model - all for prediction of cancer occurrence within 36 months (DNPR dataset, **Figure 3C**).

[[3E Performance for training on known risk factors rather than a full ICD set]]

One can train on a smaller or larger set of disease codes that occur in patient disease trajectories. E.g., one can use prior knowledge and limit the input for training to known risk factors, i.e., diseases that have been reported to be indicative of the likely occurrence of pancreatic cancer (Yuan et al. 2020; Klein 2021). We find that prediction performance with the subset of ICD codes for 23 known risk factors reduces prediction accuracy of the transformer model to AUROC=0.838 compared to 0.879 for all diagnosis codes, and therefore we used the latter throughout the rest of the work (**Figure 3E,F, Table S4**).

[[3G Prediction for time intervals]]

It is of particular clinical interest to consider the risk of cancer over different time intervals. The machine learning models in this work are designed to report 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, as the longer time intervals allow for larger time gaps between the end of the disease trajectory (time of assessment) and the time of cancer diagnosis. Indeed, prediction performance for the transformer 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 (without data exclusion) (**Figure 3G,H**).

[[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 late 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. We therefore computed 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 5A, C**). As expected, recall levels decrease with longer 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 (DNPR, **Figure 5A**). 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.

Generalization for different health care systems

[[4B Performance by cross-application of a model to different datasets]]

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 dataset to disease trajectories of patients in the US-VA dataset, without any adaptation except for mapping the ICD codes from one system to the other. Prediction performance for cancer occurrence within 36 months after assessment declined from an AUROC of 0.879 (0.877-0.880) and RR=104.6, for a Denmark-trained transformer model applied to Danish DNPR patient data (test set), to an AUROC of 0.710 (0.708-0.712) and RR=57.4, for the same model applied to US-VA patient data (**Figure 4B**). The most striking difference in input data between the two systems is the shorter and more dense disease history in the US-VA trajectories compared to the Danish ones (**Figure 2B-C**).

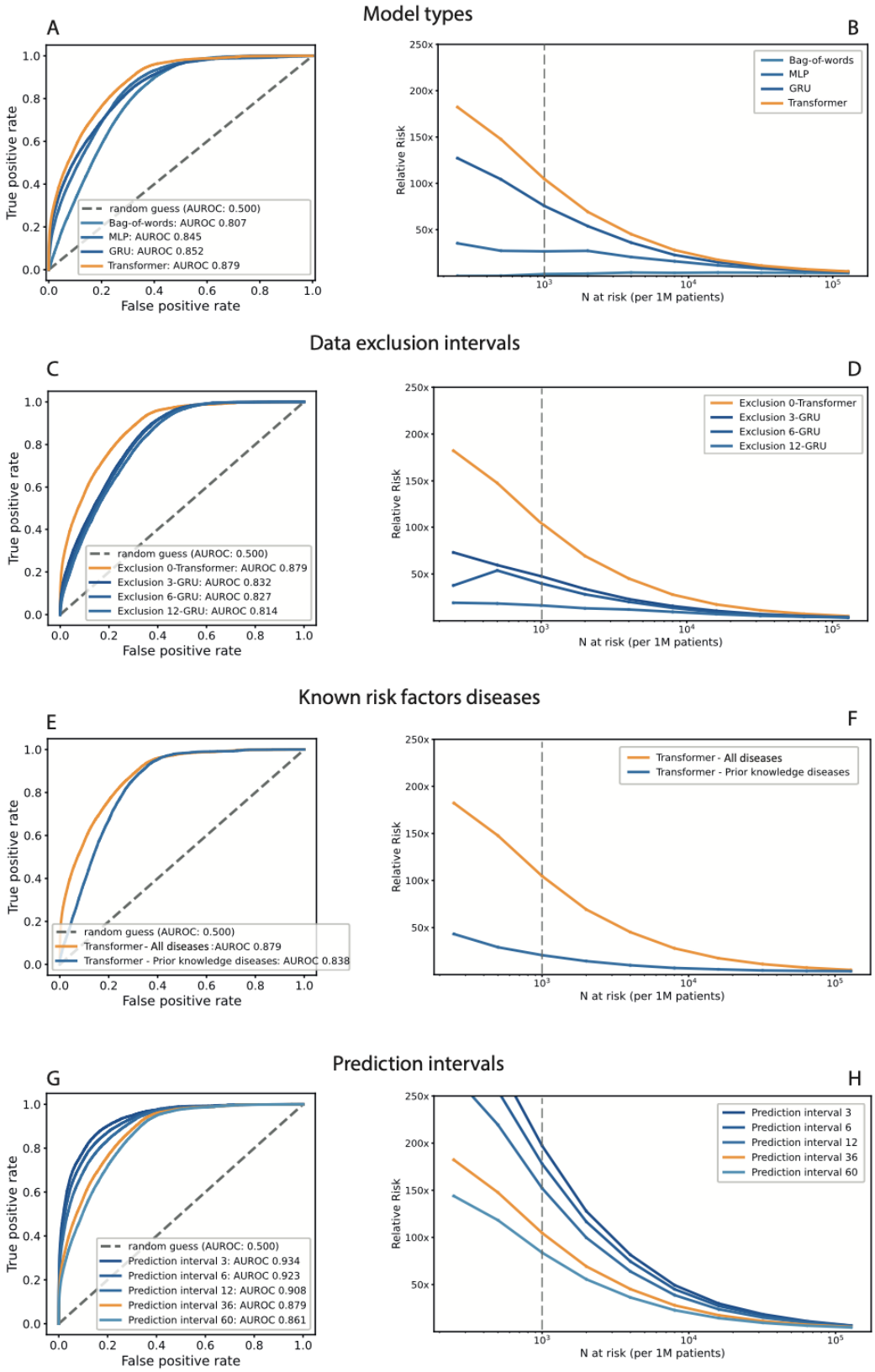
[[4C Performance by independent training on a different dataset]]

Motivated by the decrease in performance when testing the Denmark-derived model on the US-VA dataset, we trained and evaluated the model on the US-VA dataset from scratch. For the independently trained model, the performance is clearly higher than in cross-application, with a test-set AUROC of 0.775 (0.772-0.778) and RR=80.4 at the N1000 operational point (0.1% at highest risk) for cancer occurring within 36 months (**Figure 4C**). The difference in performance of the independently trained models in the two health care systems may be in part due to differences in medical and reporting practices or in demographics, including different age and sex distributions.

[[4D Proposal for a realistic surveillance program]]

A realistic surveillance program has to use an operational point (decision threshold) taken into consideration cost and benefit in real-world clinical practice. While we have no empirical data on cost and benefit in this study, we can evaluate predictive performance, which is a key factor in assessing benefit. An example (**Figure 4D**), for a potentially realistic choice of model and decision threshold, is a surveillance program at an operational point of N=1000 (0.1%) highest risk patients out of 1 million patients, i.e., at a risk fraction of 0.1% of all patients assessed. At this point, the best independently trained models trained with 3 months of data exclusion evaluated for the 12 month prediction interval for patients above age 50 obtain a relative risk of RR=58.7 for DNPR (GRU model) and of RR=33.1 for US-VA (Transformer model). Corresponding models using all the data (no exclusion) have higher values of RR. Considerations for the design of a surveillance program are further discussed below.

447 [[figure 3]]



448

449 **Figure 3. Performance of the machine learning (ML) on clinical record**
450 **trajectories in predicting pancreatic cancer occurrence, in the DK data**
451 **set.** For each model and prediction evaluation, performance is better for larger

AUROC (**A,C,E,G**) and for higher relative risk for the N highest-risk patients (cumulative; **B,D,F,H**). **(A,B)** Choice of algorithm: the transformer algorithm is best with AUROC=0.879 (no data exclusion, 36 months prediction interval). **(C,D)** Choice of input data**:** Prediction performance declines with exclusion, in training, of k=3,6,12 months of data between the end of a disease trajectory and cancer occurrence (best model for each exclusion interval, for 36 month prediction interval). **(E,F)** Choice of input data**:** Prediction is better for all 2000 ICD level-3 disease codes used throughout in training (Methods) compared to only the subset of 23 known risk factors; using transformer, all data, for 36 month prediction interval. **(G,H)** Choice of prediction task**:** Prediction of cancer is more difficult for larger prediction intervals, the time interval within which cancer is predicted to occur after assessment (transformer model, all data). We evaluate prediction performance for the 36-month prediction interval (orange) in the other panels, as this is of practical interest. **(B,D,F,H)** Prediction performance at a **particular operational point**, e.g., N=1000 highest-risk patients out of 1 million, the relative risk (RR) is 104.7 for the 36 month prediction interval using all data and 47.6 with 3 month data exclusion.

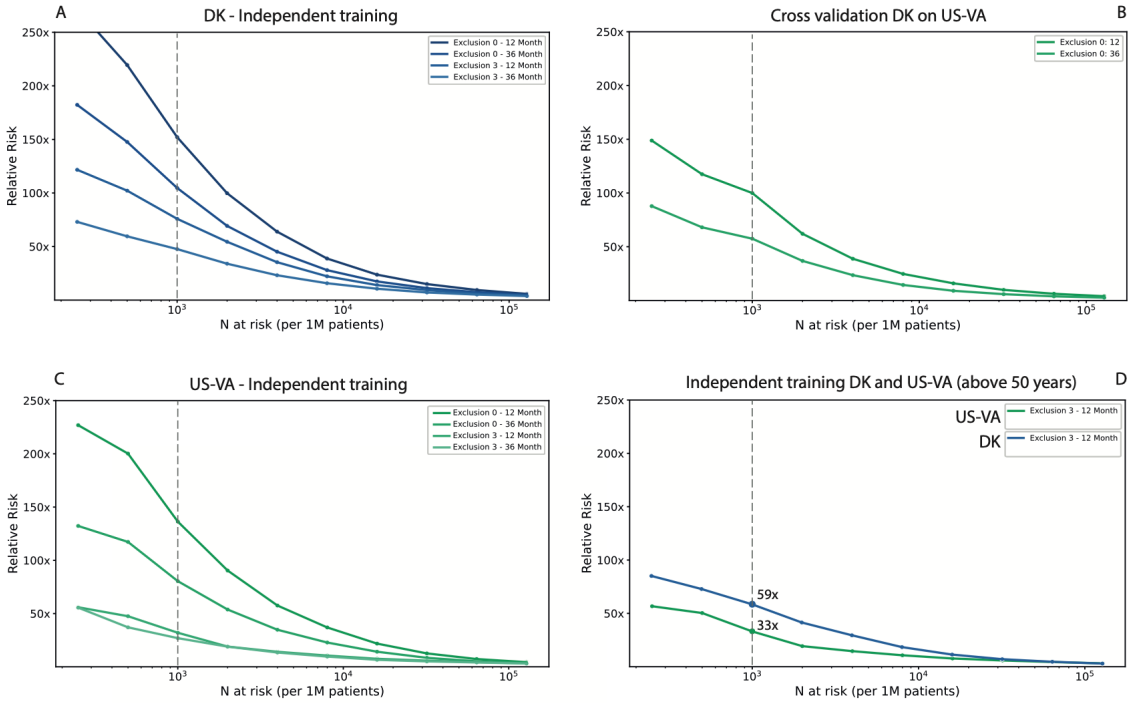


Figure 4. Estimated performance of a surveillance program for high risk patients in different health systems and with different operational choices.

Estimated relative risk for the top N high-risk patients is based on computing the accuracy of prediction on the withheld test set **(A,C,D)** or external dataset **(B)**. In designing surveillance programs, one has a choice between models trained on all data (excl=0) versus models trained excluding data from the last three months before cancer occurrence (excl=3) and between prediction for cancer within 12 months of prediction or within 36 months, with estimates of prospective performance based on independent training on data from the DK **(A)** Performance for the US-VA **(C)** health systems. **(B)** Estimated performance is somewhat lower for cross-application for a model trained on DK data applied to VA patient data, with an AUROC of 0.710 (0.708-0.712), illustrating the challenge of deriving globally valid prediction tools without independent localized or system-specific training. **(D)** A proposed practical choice for a surveillance program with good estimated accuracy of prediction, in either system, would entail application of independently trained models with 3-month data exclusion for a prediction interval of 12 months for patients above age 50.

492 **Predictive Features**

493 **[[Interpretation: contributing factors from gradient method]]**

494 Although the principal criterion for the potential impact of prediction-surveillance
495 programs is robust predictive performance, it is of interest to interpret the features of
496 any predictive method: which diagnoses are most informative of cancer risk?
497 Computational methods can infer the contribution of a particular input variable to the
498 prediction by a machine learning engine, e.g., the integrated gradients (IG) algorithm
499 (Sundararajan, Taly, and Yan 2017) (**Figure 5C-D; Table S5**). The IG contribution
500 was calculated separately for different times between assessment and cancer
501 diagnosis, in particular at 0-6 months, 6-12 months, 12-24 months and 24-36 months
502 after assessment, for all patients who developed cancer. The aim was to explore how
503 diagnosis codes contribute differently to the risk of pancreatic cancer, depending on
504 how close to pancreatic cancer they were administered. As expected, there is a
505 tendency for codes, which in normal clinical practice are known to indicate the potential
506 presence of pancreatic cancer, to have a higher contribution to prediction for
507 trajectories that end closer to cancer diagnosis. On the other hand, putative early risk
508 factors have a higher IG score for trajectories that end many months before cancer
509 diagnosis. With the proviso that the IG scores separately evaluate individual features
510 that actually act collectively, these results provide some level of confidence in the
511 machine learning prediction method. The interpretation of individual risk factors from
512 the machine learning feature list as causative may be subject to misinterpretation as
513 their contribution here is only evaluated in the context of complete disease histories.
514 However, our main goal in this report is to achieve robust predictive power from
515 disease trajectories, rather than mechanistic interpretations.

516 **[[Interpretation: contributing early factors]]**

517 The top contributing features extracted from the trajectories with time to cancer
518 diagnosis in 0-6 months - such as unspecified jaundice, diseases of biliary tract,
519 abdominal-pelvic pain, weight loss and neoplasms of digestive organs - may be
520 symptoms of or otherwise closely related to pancreatic cancer (**Table S5A**). It is also
521 of interest to identify early risk factors for pancreatic cancer**.** For trajectories with longer
522 time between assessment and cancer diagnosis, other disease codes - such as type
523 2 diabetes and insulin-independent diabetes - make an increasingly large contribution,
524 consistent with epidemiological studies (Yuan et al. 2020; Klein et al. 2013; Kim et al.
525 2020) and the observed disease distribution in the DNPR and US-VA datasets (**Figure**
526 **4, S3**). Other factors, such as cholelithiasis (gallstones) and reflux disease, are
527 perhaps of interest in terms of potential mechanistic hypotheses, such as inflammation
528 of the pancreas prior to cancer as a result of cholelithiasis or a hypothetical link
529 between medication by proton pump inhibitors such as omeprazole in reflux disease
530 and the effect of increased levels of gastrin on the state of the pancreas (Alkhushaym
531 et al. 2020).

[[Interpretation: extent of similarity of IG features in the two systems]]

We have additionally performed the disease contribution analysis on the US-VA dataset (**Table S5B**). The disease contribution analysis showed commonly identified disease codes between the two systems such as unspecified jaundice, other disorders of pancreas, other diseases of biliary tract, diabetes mellitus, abdominal and pelvic pain, other diseases of liver, symptoms and signs concerning food and fluid intake. Generally, in addition to disease codes that are rather specific for pancreatic cancer, others are unspecific, indicating the importance of including many variables with limited individual prognostic value to obtain good prognostication.

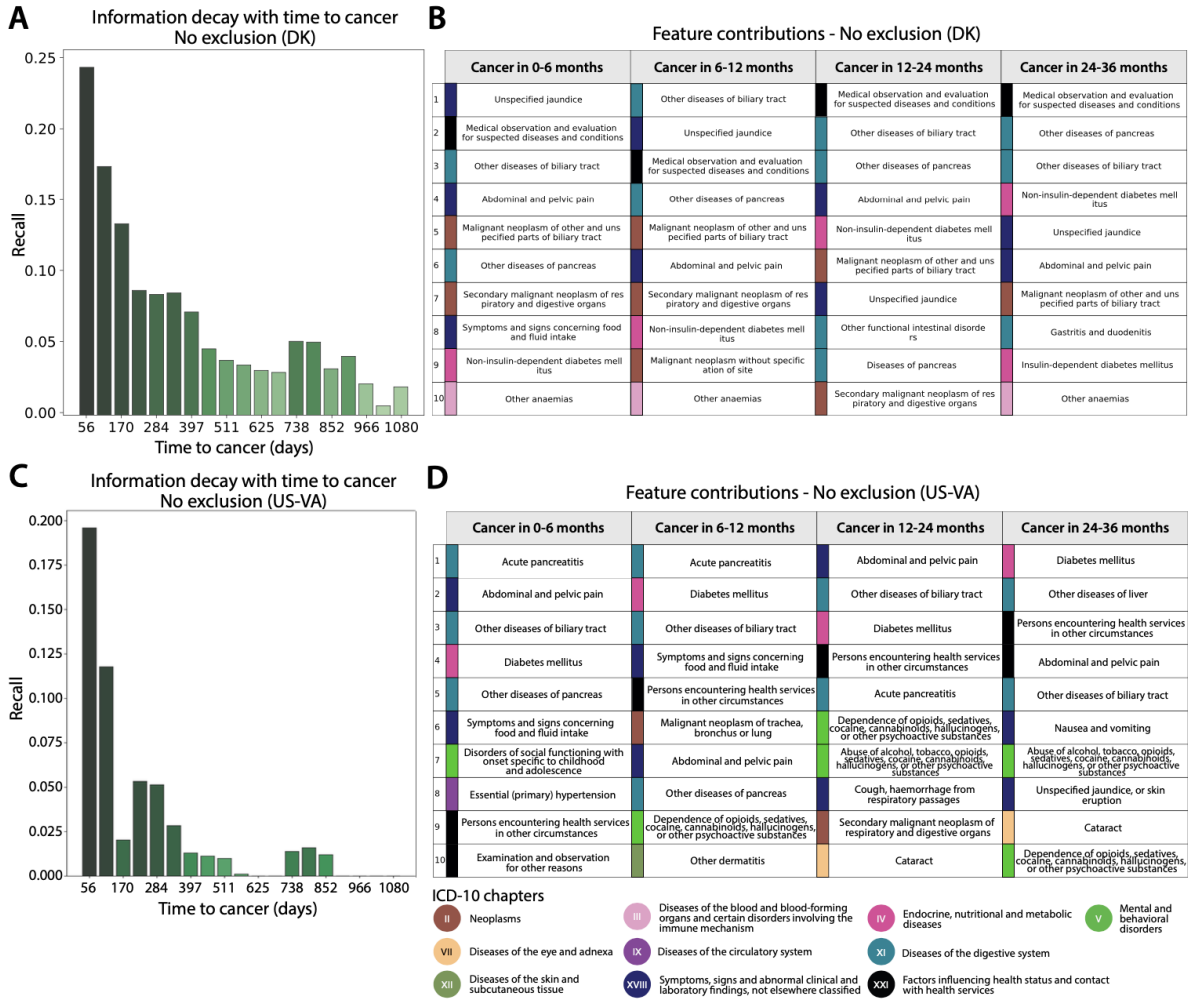


Figure 5. Predictive capacity and feature contributions of disease trajectories

(**A**,**C**) Distribution of recall (sensitivity) values at the F1 operational point (Methods) as a function of time-to-cancer (time between the end of a disease trajectory and cancer diagnosis). The recall values decrease with time-to-cancer. (**A**) Danish system, for models trained on all data (no exclusion). The contribution of age for DNPR is in Figure S3. (**C**) US-VA system, for models

trained on all data. (**B,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 for the **(B)** Danish and **(D)** US-VA systems. The features are sorted by the contribution score (**Suppl. 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.

Discussion

[[Advances in this work]]

We present a new framework for applying deep learning to longitudinal datasets of disease trajectories to predict the risk of a low-incidence but very aggressive cancer. This 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 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 to the learning engine, rather than just disease occurrence at any time, significantly improves the predictive power of AI methods in anticipating pancreatic cancer occurrence.

[[Performance in a different healthcare systems]]

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 US-VA database had lower performance in spite of common use of ICD disease codes **(Figure 4)**. Reasons for this lack of transferability plausibly are differences in clinical practice, such as frequency of reporting disease codes in clinical records, typical thresholds for seeking medical attention, potential influence of billing constraints and billing optimization, as well as entry to and departure from the US-VA health system - all of these in contrast to the more uniform and comprehensive national nature of the Danish disease registries. However, the AI methods used are sufficiently robust to achieve a higher level of performance in the US-VA system via independent training, as opposed to cross-application of an externally trained prediction tool. We conclude that when there are significant differences in healthcare systems, independent model training in different geographical locations is required to achieve locally optimal model performance. However, if independent training is not feasible, e.g. in smaller health care systems, then cross-application may be of value, depending on the actual outcome of cross-performance tests, which can be performed on smaller datasets than those required for training.

[[Application in clinical practice - 3 steps]]

Successful implementation of early diagnosis and treatment of pancreatic cancer in clinical practice will likely require three essential steps: (i) identification of high-risk patients, (ii) detection of early cancer or precancerous states by detailed surveillance of high-risk patients, and (iii) 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 surveillance programs. A sufficiently enriched pool of high-risk patients would make detailed screening tests affordable, as such tests are likely to be prohibitively expensive at a population level, and enhance the positive predictive value of such tests.

[[Design of surveillance programs]]

Based on the prediction accuracy reported here, one can design a potentially realistic prediction-surveillance process, in which software is applied to health records of, e.g., 1 million patients, followed by identification of those at highest risk and recruitment into a surveillance program with detailed screening tests for, e.g., 1000 highest-risk patients (0.1% of all assessed). Implementation requires choosing an operational point along the curve of relative risk against the fraction of patients at high risk (**Figures 3,4**) (or an equivalent point on a precision-recall curve) with an achievable high positive predictive value so as to reduce false positives in the high-risk group and therefore minimize unnecessary effort, potential harm from unnecessary invasive procedures and potential anxiety.

One proposal for the design of a surveillance program is to focus on patients from age 50 (**Figure 4D**). (1) One mode of operation is to supplement existing surveillance practice in a given hospital system, in which patients with, e.g., pancreatitis, abdominal pain and/or unspecified jaundice are routinely under surveillance by clinicians. To supplement these, using a model *trained with 3 months data exclusion* and applied to patients of age 50 or older, our results indicate that out of 1000 patients at highest risk (0.1% of 1 million) 70 patients would be future pancreatic cancer patients (PPV 0.07, 12 months prediction interval, GRU, DNPR), assuming that these typically are not already under surveillance. (2) Another mode of operation is to consider the risk prediction for any patient, including those that are already under surveillance as part of current clinical practice. In this mode one identifies patients at highest risk using a model *trained on all data* (no data exclusion), for which we computed a PPV of 0.32 (12 months prediction interval, Transformer, DNPR, age 50 or older), corresponding to 320 future pancreatic cancer patients out of 1000 at highest risk, (in a population of 1 million assessed by the prediction method). In addition to patients already under surveillance this would plausibly add a non-negligible number of additional candidates for surveillance.

Interestingly, the relative risk as reported here is comparable to that reported for the -surveillance of patients with genetic risk factors such as mutations in the BRCA2, ATM, PALB2 and other genes or those with a high value of a polygenic risk score (Klein 2021; Galeotti et al. 2021). While those with family history of pancreatic cancer are reported to have a 9x higher risk of pancreatic cancer, heritability is estimated to explain only 4-5% of all pancreatic cancers. Here, complementary to established surveillance programs for patients with elevated risk based on germline variants or

family history, the focus is on the application to the entire patient population, for which genetic or detailed family history data is available only in a relatively small number of patients. The real-world AI approach casts a much wider net and is a practical approach in that running the prediction program routinely on, say, 1 million patients - is relatively inexpensive. In addition to prediction performance, cost of screening is then a leading factor in determining the fraction of high-risk patients nominated for a screening program in any particular health care system.

[[Estimated performance of surveillance programs]]

In either scenario we estimate the pool of patients already under surveillance plus additional surveillance candidates would result in approximately 320 (out of 1000) patients (0.1% of 1 million), who actually will develop pancreatic cancer. Whether or not cancer is actually detected early in these patients of course depends on the clinical tests performed. Our study cannot provide estimates on the positive early detection rate of such a surveillance program, only on the number of patients estimated to eventually develop pancreatic cancer. Detailed deliberations with clinicians will be required to design the details of such a program, e.g., to choose the frequency of follow-up visits and tests to be performed, optimizing success rates of tests, minimizing cost and minimizing potential harm from invasive procedures. For cost considerations alone, we provide a hypothetical tradeoff algorithm (**Results S1**). A surveillance program will also have to be supplemented with counseling, in analogy to genetic counseling, to provide patients with an informed choice and to minimize potential anxiety. The numerical estimates of performance of the risk assessment tool presented here are simply one of several considerations needed for the design of a surveillance program.

In our view, the performance reported here, in particular in the Danish system, is an advance relative to earlier methods and is approaching a level in which the design of a surveillance program is a practical proposition. Even a relatively moderate level of additional early detection may be considered of value, provided that cost and implementation issues can be successfully addressed in a real-world implementation.

[[Challenges for future improvements]]

We expect further increases in prediction accuracy with the real-world availability of data beyond disease codes, such as medication, laboratory values, observations in clinical notes, abdominal imaging (computerized tomography, magnetic resonance imaging), treatment records from general practitioners (Malhotra et al. 2021; Lemanska et al. 2022) as well as germline 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, with careful assessment of the accuracy of clinical records. 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. The challenges to achieve federated learning are, however, not only technical but also social and organizational, especially in a competitive health care landscape.

[[Impact on patients]]

The particular advantage of a real-world high-risk prediction-surveillance 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 intervention programs is limited to a much smaller number of patients, those at highest risk. Prediction performance at the level shown here may be sufficient for an initial design of real world clinical prediction-surveillance programs and future improvements are likely. AI on real-world clinical records has the potential to produce a scalable workflow for early detection of 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 disease code trajectory datasets

[[The population-level 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 Institute of Clinical Medicine, Aarhus University Hospital, has been reported to be 98% (accuracy for 19 Charlson conditions in 950 reviewed records) (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.

[[Use of ICD codes]]

The 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*. 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 labeled as ‘temporary’ or ‘referral’ (8.3% removed, **Figure S1A**), as these can be misinterpreted when mixed with the main diagnoses and are not valuable for the purposes of this study.

[[Filtering the Danish dataset]]

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.

[[Subset of Danish dataset used for training]]

The Danish DNPR 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.

[[Quality control using the Danish Cancer Registry dataset]]

To perform a quality check of the pancreatic cancer cases in DNPR, we compared pancreatic cancer cases from both DNPR with the pancreatic cancer cases in the Danish Cancer Registry (DCR). DCR is a population-wide cancer cohort and one of the largest and most comprehensive in the world, which has recorded nation-wide cancer incidences from 1943 (Gjerstorff 2011; Sundhedsstyrelsen 2009). The DCR contains information on diagnosis, cancer type, tumor topography, TNM staging classification and morphology. The purpose of DCR is to keep track of cancer incidences and mortality in Denmark and to study causes, courses and statistics for treatment improvements in the Danish health system. The cancer registry administration was moved from The Danish Cancer Society to the National Board of Health in 1997.

We compared the overlap between the pancreatic cancer cases in DNPR versus the cases in DCR to estimate the quality of our case population used for training. 88% of the pancreatic cancer cases used for training overlapped with the cancer registry. Due to the differing nature of the DCR and DNPR, for which the latter is a more administrative registry with multiple coding purposes such as hospital reimbursement, monitoring hospital services and patient trajectories, quality assurance etc., we did not expect these registries to overlap 100%. As there may be correctly labeled cases in DNPR that did not get captured in DCR, the 88% concordance is a lower bound on accuracy of the pancreatic cancer ICD-10 codes in DNPR, which we used as case labels in training.

784 **[[The military veterans US-VA dataset]]**

785 Analysis in the US-VA system was conducted with approval from the VA Boston
786 Healthcare System Institutional Review Board. For the US-VA system, we used data
787 from the VA Corporate Data Warehouse (CDW), which collates electronic health
788 record and cancer registry information on veterans treated at VA facilities nationwide
789 (Price, Shea, and Gephart 2015), using methods similar to prior work (Elbers et al.
790 2020; Chang et al. 2022)(Wu et al. 2022). The CDW includes electronic health record
791 data from 1999 to the present originating from the comprehensive range of primary
792 care and specialized services the VA provides at its inpatient and outpatient facilities,
793 as well as claims data for care received at outside facilities and paid for by the VA.
794 Each outpatient visit in the database is associated with the date it occurred, and each
795 inpatient visit is associated with an admission and discharge date. Both inpatient and
796 outpatient visits are also associated with International Classification of Diseases codes
797 (ICD-9 before October 1, 2015 and ICD-10 on and after that date) pertaining to the
798 visit.
799

800 **[[The US-VA cancer registry and data quality]]**

801 In addition to electronic health record data, the CDW also includes cancer registry
802 data. VA cancer registry activities were initiated pursuant to a national directive in
803 1998, with incident cases annotated retrospectively from 1995 and prospectively until
804 the present (Zullig et al. 2012). Cases are abstracted manually by trained cancer
805 registrars in accordance with standards of the North American Association of Central
806 Cancer Registries (M. L. Thornton 2022). Potential cases are flagged for review by
807 custom software (OncoTraX) which identifies potential cases automatically based on
808 occurrence of structured data such as ICD codes in the electronic health record. This
809 approach, where semi-automated screening is followed by manual review by trained
810 cancer registrars, results in highly accurate case ascertainment (Zullig et al. 2012,
811 2019). For example, one study found that VA cancer registry data in the CDW had
812 near perfect accuracy in colorectal cancer case ascertainment by all evaluated
813 measures (positive predictive value, negative predictive value, sensitivity, and
814 specificity) as compared to de novo manual review of 200 potential cases (Earles et
815 al. 2018). In contrast, case ascertainment using ICD codes from the electronic health
816 record had only 58% positive predictive value. Regarding pancreatic cancer case
817 ascertainment specifically, we compared pancreatic cancer cases in the VA cancer
818 registry to cases identified based on ICD codes in the VA electronic health record. We
819 found that 94.9% of pancreatic cases in the VA cancer registry between 1999 and
820 2020 had at least one ICD-10 code for pancreatic cancer in the electronic health
821 record, while only 32.4% of patients with a pancreatic cancer ICD code in the electronic
822 health record were found in the cancer registry. Spurious ICD codes for pancreatic
823 cancer can occur in the EHR due to multiple reasons, including outright error, use of
824 the cancer code for a screening or evaluation visit, use of the cancer code for medical
825 history, and use of the cancer code in dual-use scenarios where cancer care is
826 provided outside the VA.
827

828 **[[Filtering the US-VA dataset and subset used for training]]**

829 VA patients were included in the study as described in the flow chart (**Figure S1B**).
830 Out of all 15,933,326 patients with ≥1 ICD code in US-VA Corporate Data Warehouse
831 between 1999 and 2020 and available in our research study database, we randomly
832 sampled a subset of approximately 3 million patients (2,975,110) due to limitations of
833 the computational resources available. Based on the considerations above, we
834 identified patients with pancreatic cancer as the overlap of those with a diagnosis of
835 pancreatic cancer in the VA cancer registry and those with an ICD code in the VA
836 electronic health record system. We excluded patients with an ICD code for pancreatic
837 cancer in the VA electronic health record data who did not have an entry for pancreatic
838 cancer in the VA cancer registry or vice versa, since the status of these patients was
839 unclear. We further excluded patients with short trajectories (<5 visits), as in the
840 Danish dataset. This resulted in a final VA dataset of 1,948,209 patients total, including
841 3,418 patients with pancreatic cancer. We randomly allocated patients in the final VA
842 dataset into training (80%), development (10%), and test (10%) sets, with the condition
843 that all trajectories from a patient were only included in one split group (train/dev/test),
844 to avoid any information leakage between training and development/test datasets, as
845 in the Danish dataset.
846

847 **[[Similarity of survival curves in the two health care systems]]**

848 As a check on the quality of pancreatic cancer case ascertainment in the Danish and
849 US-VA datasets, we plotted overall survival in each dataset, stratified by cancer stage
850 (**Figure S1C)**. Cancer stage was obtained from the respective dataset's cancer
851 registry. Stage was only available on a subset of patients. The overall similarity of the
852 survival curves in the two datasets adds confidence to the quality of the data, as
853 selected for the comparative study.
854

855 **<u>Training</u>**

856

857 The following processing steps were carried out identically for DNPR and US-VA
858 datasets. For each patient, whether or not they ever had pancreatic cancer, the data
859 was augmented by using all continuous **partial trajectories** of (minimal length >=5
860 diagnoses) from the beginning of their disease history and ending at different time
861 points, which we call the time of assessment. For cancer patients, we used only
862 trajectories that end before cancer diagnoses, i.e. $t_a<t_{cancer}<t_{death}$. We used a **step**
863 **function annotation** indicating cancer occurrence at different time points (3, 6, 12,
864 36, 60 months) after the end of each partial trajectory. For the positive ('PC') cases
865 this provides the opportunity to learn from disease histories with a significant time gap
866 between the time of assessment and the time of cancer occurrence. For example, for
867 a patient, who had pancreatitis a month or two just before the cancer diagnosis, it is
868 of interest to learn which earlier disease codes might have been predictive of cancer
869 occurrence going back at least several months or perhaps years. The latter is also
870 explored by separately re-training of the machine learning model excluding data from
871 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 machine learning. 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 machine learning 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 $t = t_a + 3, 6, 12, 36, 60$ 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**).

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

where $t \in \{3,6,12,36,60\}$ months; $N_T = 5$ for non-cancer patient and $N_T \leq 5$ for cancer patients where we only use the time points before the cancer diagnosis; $i = 1,2,3,\ldots,N$ labels samples; Θ is the set of model parameters; λ_2 is the regularization strength; $\hat{p}$ is the estimated risk output by the model; 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 (**Figure1C**). If the score is above the threshold, one has a correct positive prediction, if cancer has occurred at any time

957 within 3 years; and a false positive prediction, if cancer has not occurred within 3 years.
958 If the score is below the threshold, one has a false negative prediction if cancer has
959 occurred at any time within 3 years; and a true negative prediction, if cancer has not
960 occurred within 3 years. As both training and evaluation make use of multiple
961 trajectories per patient, with different end-of-trajectory points, the performance
962 numbers are computed over trajectories.
963
964 The relative risk ratio (RR) is calculated as the odds of getting pancreatic cancer when
965 classified at high risk compared to a random method that just uses the disease
966 incidence in the population. RR is defined as
967

968 $$RR = \frac{precision}{incidence} = \frac{TP/(TP+FP)}{(TP+FN)/(TP+FP+TN+FN)}$$

969
970 where TP = True Positives, FP = False Positives, FN = False Negatives, TN = True
971 negatives. The relative risk score is defined at a given operational decision point along
972 the RR curve as a function of the number of patients predicted to be at high risk
973 **(Figure 3 right; Figure 4)**. The relative risk ratio assesses by what factor a prediction
974 method does better than a random pick based just on the population disease
975 incidence.
976
977 For the purpose of comparing different algorithms, input modes and DK-DNPR vs. US-
978 VA, we report exhaustive performance tables (**Supplement Tables S3A-I**) with
979 precision and recall at the F1 operational point, which maximizes the harmonic mean
980 of recall and precision (Sasaki 2007). However, for consideration of clinical
981 implementation, which requires severe limits on the number of patients that can be
982 advanced to a surveillance program, we use an operational point for the top 1000 high
983 risk patients out of 1 million patients (0.1% of 1 million, see relative risk plots **Figures**
984 **3&4**).
985
986
987 **<u>Cross-application to the US-VA dataset</u>**
988
989 In order to assess the predictive performance of the model in other health care
990 systems, we applied the best machine learning model trained on the Danish DNPR to
991 disease trajectories of patients in a US Veterans Administration dataset. For the US-
992 VA dataset, we directly applied models without any adaptation except for mapping the
993 ICD code from one system to the other. In brief, ICD-9 codes in the US-VA dataset
994 were first mapped to ICD-10 codes, followed by adding a prefix “D” to obtain Danmark
995 compatible ICD-10 codes.
996

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.

Software

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

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. 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 (grant SU2C#6180), the Lustgarten Foundation and their donors for financial and community support. The authors thank Shawn Murphy and Henry Chueh and the Mass General Brigham Health Care Research Patient Data Registry group for facilitating use of their database.

Conflict of Interest Statements

S.B. has ownership in Intomics A/S, Hoba Therapeutics Aps, Novo Nordisk A/S, Lundbeck A/S, ALK Abello 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. C.S. is on the science advisory board of Cytoreason LTD. From August 1, 2020, A.R. is an employee of Genentech.

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.

Chang, Michael S., Jennifer La, Nicole Trepanowski, David Cheng, John R. Bihn, Nhan Do, Mary Brophy, Nathanael R. Fillmore, and Rebecca I. Hartman. 2022. "Increased Relative Proportions of Advanced Melanoma among Veterans: A Comparative Analysis with the Surveillance, Epidemiology, and End Results Registry." *Journal of the American Academy of Dermatology* 87 (1): 72–79.

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.

Earles, Ashley, Lin Liu, Ranier Bustamante, Pat Coke, Julie Lynch, Karen Messer, María Elena Martínez, et al. 2018. "Structured Approach for Evaluating Strategies for Cancer Ascertainment Using Large-Scale Electronic Health Record Data." *JCO Clinical Cancer Informatics* 2 (December): 1–12.

Elbers, Danne C., Nathanael R. Fillmore, Feng-Chi Sung, Spyridon S. Ganas, Andrew Prokhorenkov, Christopher Meyer, Robert B. Hall, et al. 2020. "The Veterans Affairs Precision Oncology Data Repository, a Clinical, Genomic, and Imaging Research Database." *Patterns (New York, N.Y.)* 1 (6): 100083.

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.

Galeotti, Alice Alessandra, Manuel Gentiluomo, Cosmeri Rizzato, Ofure Obazee, John P. Neoptolemos, Claudio Pasquali, Michael Nentwich, et al. 2021. "Polygenic and Multifactorial Scores for Pancreatic Ductal Adenocarcinoma Risk Prediction." *Journal of Medical Genetics* 58 (6): 369–77.

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.

Gjerstorff, Marianne Lundkjær. 2011. "The Danish Cancer Registry." *Scandinavian Journal of Public Health* 39 (7 Suppl): 42–45.

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.

Lemanska, Agnieszka, Claire A. Price, Nathan Jeffreys, Rachel Byford, Hajira Dambha-Miller, Xuejuan Fan, William Hinton, et al. 2022. "BMI and HbA1c Are Metabolic Markers for Pancreatic Cancer: Matched Case-Control Study Using a UK Primary Care Database." *PloS One* 17 (10): e0275369.

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.

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.

Price, Lauren E., Kimberly Shea, and Sheila Gephart. 2015. "The Veterans Affairs's Corporate Data Warehouse: Uses and Implications for Nursing Research and Practice." *Nursing Administration Quarterly* 39 (4): 311–18.

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.

Sundhedsstyrelsen. 2009. *Det moderniserede Cancerregister - metode og kvalitet*.

Tealab, Ahmed. 2018. “Time Series Forecasting Using Artificial Neural Networks Methodologies: A Systematic Review.” *Future Computing and Informatics Journal* 3 (2): 334–40.

Thornton, M. L. 2022. *Standards for Cancer Registries Volume II: Data Standards and Data Dictionary*. Edited by Monica Thornton. Vol. Version 23. Twenty Fourth Edition. North American Association of Central Cancer Registries, Inc. (NAACCR).

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.

Thorsen-Meyer, Hans-Christian, Davide Placido, Benjamin Skov Kaas-Hansen, Anna P. Nielsen, Theis Lange, Annelaura B. Nielsen, Palle Toft, et al. 2022. “Discrete-Time Survival Analysis in the Critically Ill: A Deep Learning Approach Using Heterogeneous Data.” *NPJ Digital Medicine* 5 (1): 142.

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.

Wu, Julie Tsu-Yu, Jennifer La, Westyn Branch-Elliman, Linden B. Huhmann, Summer S. Han, Giovanni Parmigiani, David P. Tuck, et al. 2022. “Association of COVID-19 Vaccination With SARS-CoV-2 Infection in Patients With Cancer: A US Nationwide Veterans Affairs Study.” *JAMA Oncology* 8 (2): 281–86.

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.

Zullig, Leah L., George L. Jackson, Raye Anne Dorn, Dawn T. Provenzale, Rebecca McNeil, Catherine M. Thomas, and Michael J. Kelley. 2012. “Cancer Incidence among Patients

1261 of the U.S. Veterans Affairs Health Care System." *Military Medicine* 177 (6): 693–701.
1262 Zullig, Leah L., Shelley A. Jazowski, Neetu Chawla, Christina D. Williams, David Winski,
1263 Christopher G. Slatore, Alecia Clary, Kelli M. Rasmussen, Leann M. Ticknor, and
1264 Michael J. Kelley. 2019. "Summary of Veterans Health Administration Cancer Data
1265 Sources." *Journal of Registry Management* 46 (3): 76–83.

Supplementary Materials

Figure S1. Preprocessing and filtering of the DK and US-VA disease trajectory datasets.

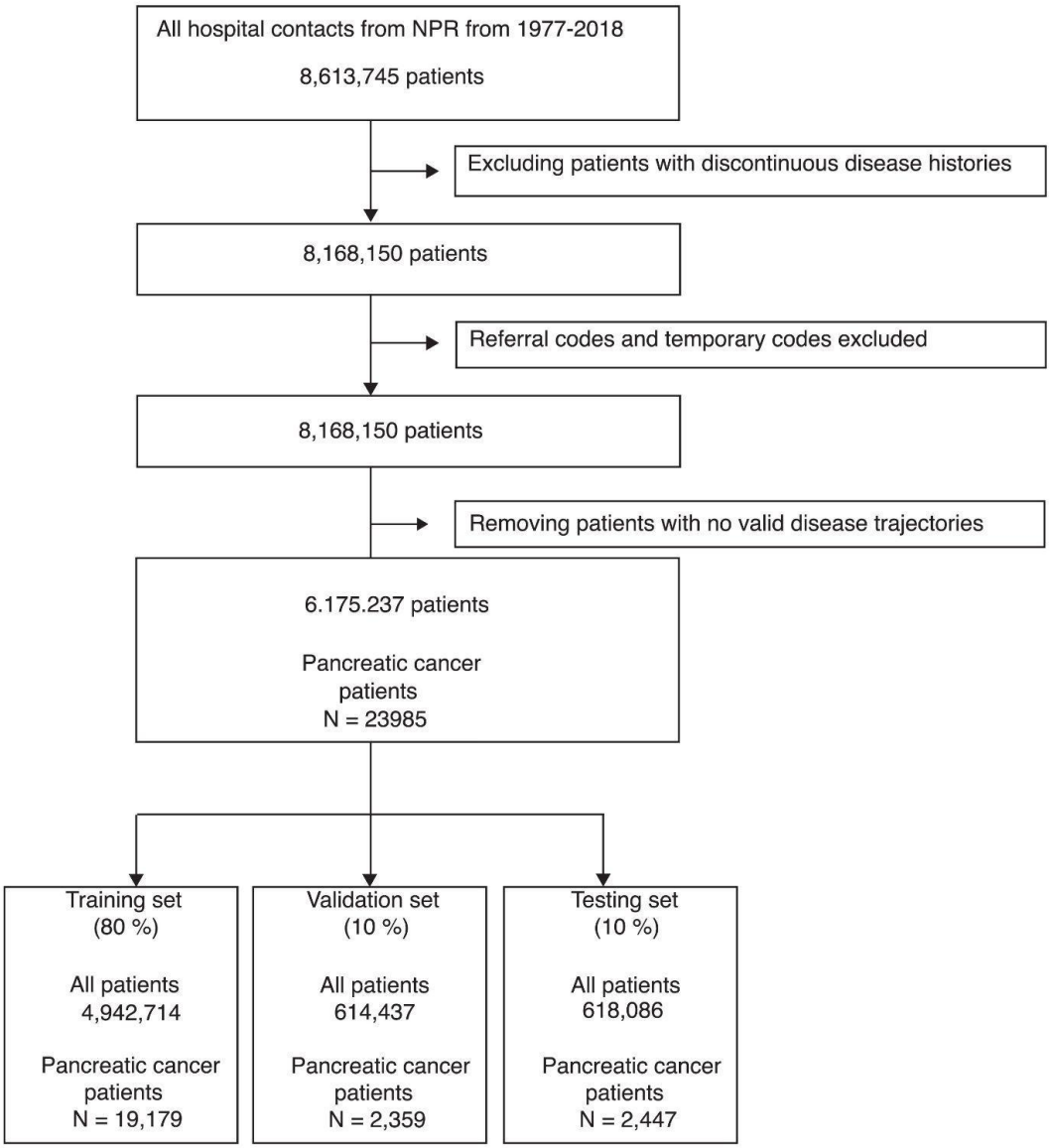
Filtering of the Danish (DK) and US-VA patient registries prior to training.

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 US-VA dataset, around 3 million of patients were randomly sampled due to computational limitations and patients with ICD-9/10 for pancreatic cancer, but without entries in US-VA cancer registry were excluded. Similar to the Danish dataset filtering, short trajectories (<5 diagnosis codes) were removed and patients were split into Training (80 %), Validation (10%) and Test set (10%).

1287 **Figure S1A - Denmark (DK) DNPR**

1288

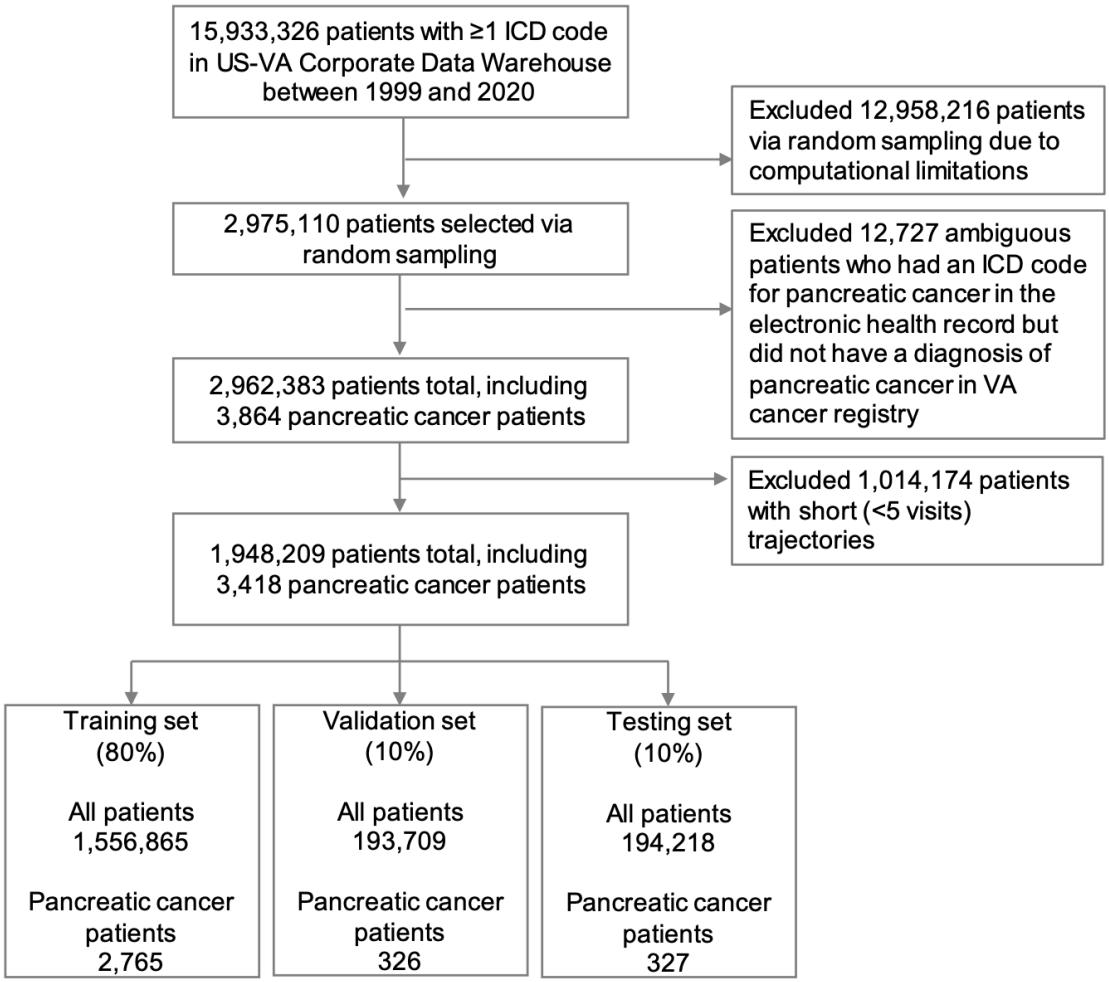


1289

1290

1291 **Figure S1B - US-VA**

1292



1293
1294
1295
1296
1297
1298
1299
1300

Figure S1C - Survival curves

Overall survival in each dataset, stratified by cancer stage. Cases were ascertained using the methods described in Methods, and cancer stage was obtained from the respective dataset's cancer registry. Stage was only available on a subset of patients.

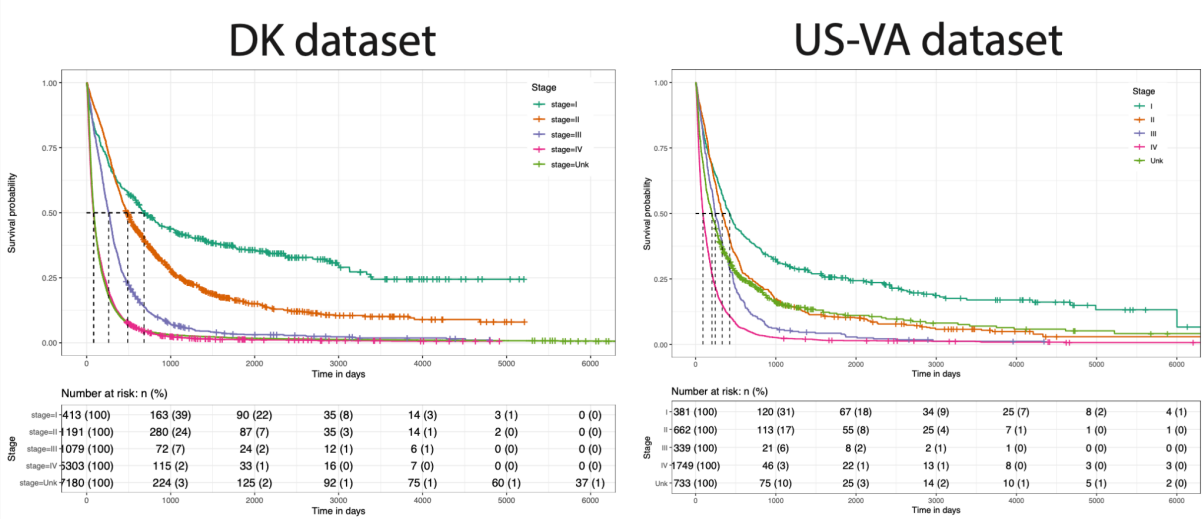
For the Danish dataset, five-year survival was 23% for Stage I, 8.3% for Stage II, 2.3% for Stage III, and 0.8 for Stage IV. Median survival was 645 days for Stage I, 483 days for Stage II, 262 days for Stage III, and 81 days for Stage IV.

For the US-VA dataset, five-year survival was 19.4% for Stage I, 8.8% for Stage II, 3.2% for Stage III, and 1.3% for Stage IV. Median survival was 424 days for Stage I, 330 days for Stage II, 243 days for Stage III, and 91 days for Stage IV.

Ordered for comparison:

Median survival [days]	*DNPR*	*US-VA*
Stage I	*645*	*424*
Stage II	*483*	*330*
Stage III	*262*	*243*
Stage IV	*81*	*91*

The uneven distribution of sex in the VA database may contribute to the differences in survival. The overall similarity of the survival curves in the two datasets adds some confidence to the quality of the data, as selected for the study.



1326 **Table S1. Description of the patient cohorts used in this study (DK).**

1327

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, ∞)	1,422 (0.02%)	7,398 (0.09%)

1328

1329

Pancreatic Cancer Patients (n=23,985)		
	Male	Female
Total Count	11,880 (49.53%)	12,105 50.47%
Age at pancreatic cancer diagnosis [0,30)	13 (0.05%)	19 (0.08%)
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, ∞)	148 (0.62%)	495 (2.06%)

1330
1331
1332

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 - DK dataset and US-VA dataset

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 that maximizes the F1 score, which is the harmonic mean of recall and precision (Sasaki 2007).

Summary of Tables S3

Table S3A	Performance summary DNPR (AUROC)
Table S3B	Performance summary DNPR (Specificity/Precision/Recall)
Table S3C	Performance summary DNPR models applied to US-VA (AUROC)
Table S3D	Performance summary DNPR models applied to US-VA (Specificity/Precision/Recall)
Table S3E	Performance summary US-VA, independent training (AUROC)
Table S3F	Performance summary US-VA, independent training (Specificity/Precision/Recall)
Table S3G	Performance summary US-VA, independent training above 50 (AUROC)
Table S3H	Performance summary US-VA, independent training above 50 (Specificity/Precision/Recall)
Table S3I	Performance summary DNPR above 50 (AUROC)
Table S3J	Performance summary DNPR above 50 (Specificity/Precision/Recall)

Table S3A. Performance summary DNPR (AUROC)

Model	Prediction Interval (months) → Exclusion Interval (months)	0 - 3	0 - 6	0 - 12	0 - 36	0 - 60
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 US-VA (AUROC)

Model	Prediction Interval (months) →	0 - 3	0 - 6	0 - 12	0 - 36	0 - 60
	Exclusion Interval (months)					
Transformer	0	0.881 (0.878-0.883)	0.833 (0.830-0.835)	0.790 (0.788-0.792)	0.710 (0.708-0.712)	0.685 (0.683-0.687)
	3	-	0.665 (0.661-0.670)	0.641 (0.638-0.643)	0.626 (0.624-0.628)	0.627 (0.625-0.629)
	6	-	-	0.593 (0.590-0.598)	0.615 (0.612-0.616)	0.609 (0.607-0.610)

Table S3D. Performance summary DNPR model validated on US_VA (scecificity/precision/recall)

Model		Prediction Interval (months): →	0 - 3	0 - 6	0 - 12	0 - 36	0 - 60
	Exclusion Interval (months)	Metric					
Transformer (cross evaluation)	0	specificity	99.94% (99.92%-99.97%)	99.92% (99.91%-99.92%)	99.91% (99.91%-99.92%)	99.90% (99.90%-99.91%)	99.90% (99.90%-99.90%)
	0	precision	6.8% (5.9%-9.3%)	6.1% (5.8%-6.4%)	6.4% (6.2%-6.7%)	6.2% (6.0%-6.4%)	6.3% (6.1%-6.4%)
	0	recall	13.3% (9.1%-17.7%)	12.9% (12.4%-13.4%)	9.7% (9.2%-10.0%)	5.9% (5.7%-6.1%)	4.9% (4.7%-5.0%)
	3	specificity	-	99.90% (99.90%-99.90%)	99.90% (99.90%-99.90%)	99.90% (99.90%-99.90%)	99.86% (99.72%-99.86%)
	3	precision	-	0.4% (0.4%-0.5%)	0.8% (0.7%-0.9%)	1.3% (1.2%-1.4%)	1.4% (1.1%-1.5%)
	3	recall	-	2.8% (2.5%-3.0%)	1.9% (1.8%-2.2%)	1.4% (1.3%-1.5%)	1.6% (1.6%-2.4%)
	6	specificity	-	-	99.94% (99.94%-99.94%)	81.89% (67.42%-81.94%)	81.89% (79.61%-81.95%)
	6	precision	-	-	0.1% (0.1%-0.1%)	0.1% (0.1%-0.1%)	0.2% (0.2%-0.2%)
	6	recall	-	-	0.2% (0.2%-0.3%)	26.1% (25.8%-45.7%)	26.4% (25.9%-29.5%)

Table S3E. Performance summary US-VA (AUROC)

Model	Prediction Interval (months) →	0 - 3	0 - 6	0 - 12	0 - 36	0 - 60
	Exclusion Interval (months)					
Transformer	0	0.899 (0.895-0.903)	0.857 (0.854-0.862)	0.825 (0.822-0.828)	0.775 (0.772-0.778)	0.762 (0.760-0.765)
	3	-	0.783 (0.779-0.786)	0.761 (0.758-0.764)	0.758 (0.756-0.761)	0.745 (0.741-0.747)
	6	-	-	0.627 (0.622-0.630)	0.678 (0.676-0.680)	0.678 (0.676-0.680)

Table S3F. Performance summary US-VA (specificity/precision/recall)

Model	Prediction Interval (months) →	0 - 3	0 - 6	0 - 12	0 - 36	0 - 60
Exclusion Interval (months)	Metric					
Transformer	0 specificity	99.96% (99.96%-99.97%)	99.96% (99.95%-99.96%)	99.95% (99.94%-99.95%)	99.91% (99.88%-99.95%)	99.89% (99.88%-99.91%)
	0 precision	11.5% (10.5%-13.2%)	11.1% (10.4%-12.1%)	12.0% (10.9%-12.7%)	8.9% (7.8%-12.2%)	8.0% (7.6%-9.1%)
	0 recall	15.4% (13.4%-16.8%)	13.0% (11.6%-13.7%)	10.6% (10.0%-11.3%)	7.9% (6.6%-9.3%)	7.3% (6.7%-7.5%)
	3 specificity	-	99.95% (99.93%-99.96%)	99.94% (99.93%-99.97%)	99.90% (99.89%-99.90%)	99.90% (99.77%-99.90%)
	3 precision	-	2.9% (2.4%-3.7%)	2.8% (2.4%-4.2%)	2.9% (2.7%-3.2%)	2.9% (2.0%-3.2%)
	3 recall	-	3.6% (2.8%-4.2%)	2.8% (2.1%-3.4%)	2.9% (2.6%-3.1%)	2.4% (2.2%-3.5%)
	6 specificity	-	-	97.18% (96.95%-99.60%)	96.59% (96.52%-96.99%)	96.60% (96.52%-97.07%)
	6 precision	-	-	0.2% (0.2%-0.2%)	0.3% (0.3%-0.3%)	0.3% (0.3%-0.3%)
	6 recall	-	-	7.8% (1.2%-8.6%)	8.3% (7.4%-8.7%)	7.8% (6.8%-8.3%)

Table S3G. Performance table US-VA (AUROC) above 50 years

Model	Prediction Interval (months) →	0 - 3	0 - 6	0 - 12	0 - 36	0 - 60
	Exclusion Interval (months)					
Transformer	0	0.900 (0.895-0.904)	0.857 (0.853-0.862)	0.825 (0.821-0.829)	0.775 (0.772-0.778)	0.762 (0.760-0.765)
	3	-	0.782 (0.778-0.787)	0.760 (0.758-0.764)	0.756 (0.754-0.759)	0.744 (0.741-0.746)

Table S3H. Performance summary US-VA (specificity/precision/recall) above 50 years

Model	Prediction Interval (months) →	0 - 3	0 - 6	0 - 12	0 - 36	0 - 60
Exclusion Interval (months)	Metric					
Transformer	0 specificity	99.96% (99.96%-99.97%)	99.95% (99.95%-99.97%)	99.95% (99.94%-99.95%)	99.89% (99.88%-99.95%)	99.89% (99.88%-99.90%)
	0 precision	11.8% (11.0%-12.5%)	11.2% (10.4%-12.7%)	12.0% (11.2%-12.6%)	8.2% (7.8%-12.5%)	8.2% (7.7%-8.6%)
	0 recall	15.7% (14.7%-16.9%)	13.1% (11.5%-13.9%)	10.8% (10.1%-11.6%)	8.9% (6.8%-9.4%)	7.5% (6.9%-7.7%)
	3 specificity	-	99.96% (99.95%-99.96%)	99.96% (99.93%-99.97%)	99.90% (99.85%-99.90%)	99.82% (99.81%-99.90%)
	3 precision	-	3.6% (2.8%-4.1%)	3.7% (2.7%-4.6%)	2.9% (2.5%-3.1%)	2.3% (2.1%-3.1%)
	3 recall	-	3.4% (3.1%-4.0%)	2.5% (2.2%-3.2%)	2.9% (2.7%-3.4%)	2.9% (2.3%-3.2%)

Table S3I: Performance summary DNPR (AUROC) above 50 years

Model	Prediction Interval (months): ->	0-3	0-6	0-12	0-36	0-60
	Exclusion Interval (months)					
Bag-of-words	0	0.600 (0.594-0.605)	0.611 (0.606-0.615)	0.622 (0.618-0.627)	0.621 (0.618-0.624)	0.605 (0.603-0.608)
	3	-	0.646 (0.636-0.656)	0.653 (0.646-0.660)	0.633 (0.629-0.637)	0.609 (0.606-0.612)
	6	-	-	0.655 (0.648-0.662)	0.631 (0.626-0.635)	0.605 (0.601-0.609)
MLP	0	0.758 (0.752-0.763)	0.747 (0.742-0.751)	0.727 (0.723-0.731)	0.690 (0.687-0.694)	0.665 (0.662-0.668)
	3	-	0.686 (0.676-0.696)	0.672 (0.666-0.680)	0.662 (0.659-0.666)	0.644 (0.641-0.648)
	6	-	-	0.654 (0.645-0.662)	0.642 (0.639-0.646)	0.631 (0.628-0.635)
GRU	0	0.854 (0.850-0.858)	0.827 (0.823-0.831)	0.777 (0.774-0.781)	0.717 (0.714-0.720)	0.686 (0.683-0.690)
	3	-	0.733 (0.725-0.744)	0.710 (0.703-0.715)	0.667 (0.663-0.671)	0.645 (0.641-0.648)
	6	-	-	0.694 (0.686-0.703)	0.652 (0.648-0.656)	0.629 (0.625-0.633)
	12	-	-	-	0.629 (0.622-0.634)	0.610 (0.606-0.614)
Transformer	0	0.876 (0.872-0.880)	0.853 (0.850-0.857)	0.821 (0.819-0.825)	0.753 (0.750-0.756)	0.717 (0.714-0.720)
	3	-	0.739 (0.728-0.746)	0.722 (0.715-0.727)	0.674 (0.671-0.678)	0.651 (0.648-0.654)
	6	-	-	0.661 (0.655-0.668)	0.646 (0.642-0.650)	0.623 (0.620-0.627)
	12	-	-	-	0.638 (0.632-0.643)	0.615 (0.611-0.620)
Transformer - Prior knowledge disease	0	0.664 (0.659-0.671)	0.669 (0.665-0.675)	0.669 (0.664-0.673)	0.638 (0.635-0.641)	0.626 (0.623-0.628)

Table S3J: Performance summary DNPR (specificity/precision/recall) above 50 years

Model		Prediction Interval (months): ->	0-3	0-6	0-12	0-36	0-60
	Exclusion Interval (months)	Metric					
Bag-of-words	0	specificity	96.05% (90.42%-97.10%)	95.10% (92.06%-96.09%)	95.82% (92.13%-96.33%)	90.42% (88.67%-94.49%)	89.09% (87.19%-90.67%)
	0	precision	0.4% (0.3%-0.4%)	0.5% (0.4%-0.5%)	0.7% (0.6%-0.7%)	0.9% (0.9%-1.0%)	1.0% (1.0%-1.1%)
	0	recall	7.3% (5.5%-16.3%)	9.6% (7.6%-15.1%)	8.5% (7.4%-14.9%)	17.1% (10.2%-19.7%)	17.4% (15.0%-20.2%)
	3	specificity	-	99.79% (99.54%-99.80%)	99.32% (95.53%-99.52%)	94.59% (88.20%-95.38%)	87.88% (85.36%-93.80%)
	3	precision	-	0.2% (0.2%-0.3%)	0.4% (0.3%-0.5%)	0.6% (0.6%-0.7%)	0.8% (0.7%-0.8%)
	3	recall	-	1.0% (0.7%-1.8%)	2.2% (1.5%-10.5%)	10.5% (9.0%-21.7%)	19.6% (10.6%-23.5%)
	6	specificity	-	-	99.36% (96.63%-99.39%)	94.91% (89.44%-99.33%)	89.80% (85.63%-94.72%)
	6	precision	-	-	0.3% (0.2%-0.3%)	0.5% (0.5%-0.8%)	0.7% (0.6%-0.7%)
	6	recall	-	-	2.2% (1.7%-8.9%)	9.7% (1.7%-19.5%)	16.1% (8.8%-22.2%)
MLP	0	specificity	99.65% (99.42%-99.68%)	99.65% (99.42%-99.68%)	99.64% (99.43%-99.67%)	99.43% (99.08%-99.52%)	99.10% (98.88%-99.26%)
	0	precision	3.6% (3.1%-4.0%)	4.6% (3.9%-4.9%)	5.5% (4.6%-5.9%)	5.6% (4.6%-6.3%)	5.0% (4.6%-5.5%)
	0	recall	7.9% (7.1%-10.9%)	7.9% (7.1%-10.1%)	7.3% (6.8%-8.9%)	7.2% (6.4%-9.5%)	7.8% (6.8%-8.8%)
	3	specificity	-	99.76% (99.74%-99.80%)	99.72% (99.47%-99.73%)	99.08% (98.81%-99.53%)	98.81% (98.43%-99.01%)
	3	precision	-	0.7% (0.5%-0.8%)	1.3% (1.0%-1.5%)	1.7% (1.5%-2.2%)	1.8% (1.6%-1.9%)
	3	recall	-	3.4% (2.7%-4.6%)	3.2% (2.8%-4.9%)	5.2% (3.2%-6.2%)	5.0% (4.4%-6.1%)
	6	specificity	-	-	99.63% (99.62%-99.66%)	99.62% (99.25%-99.86%)	99.55% (98.89%-99.63%)
	6	precision	-	-	0.4% (0.3%-0.5%)	1.3% (1.0%-2.5%)	1.5% (1.1%-1.9%)
	6	recall	-	-	2.0% (1.6%-2.5%)	2.0% (1.2%-3.2%)	1.9% (1.5%-3.1%)
GRU	0	specificity	99.90% (99.83%-99.91%)	99.82% (99.78%-99.87%)	99.77% (99.73%-99.81%)	99.58% (99.55%-99.68%)	99.55% (99.53%-99.58%)
	0	precision	17.9% (14.0%-19.2%)	14.8% (13.4%-17.8%)	14.7% (13.2%-16.5%)	12.2% (11.6%-13.6%)	12.1% (11.6%-12.9%)
	0	recall	13.0% (12.3%-15.9%)	14.2% (12.2%-15.9%)	13.4% (12.4%-14.5%)	12.1% (11.0%-12.8%)	10.2% (9.7%-10.6%)
	3	specificity	-	99.94% (99.87%-99.95%)	99.88% (99.84%-99.91%)	99.71% (99.70%-99.76%)	99.70% (99.56%-99.71%)
	3	precision	-	3.9% (2.7%-4.6%)	6.4% (5.4%-7.5%)	6.4% (6.0%-7.1%)	6.9% (5.6%-7.4%)
	3	recall	-	5.4% (4.4%-8.5%)	6.9% (5.9%-8.0%)	6.4% (5.8%-6.9%)	5.1% (4.8%-5.9%)
	6	specificity	-	-	99.88% (99.76%-99.91%)	99.75% (99.69%-99.87%)	99.55% (99.39%-99.72%)
	6	precision	-	-	2.3% (1.8%-2.9%)	4.3% (3.8%-6.3%)	3.9% (3.3%-5.1%)
	6	recall	-	-	4.4% (3.4%-6.4%)	4.6% (3.5%-5.1%)	4.5% (3.6%-5.3%)
	12	specificity	-	-	-	99.29% (98.96%-99.47%)	99.23% (98.60%-99.63%)
	12	precision	-	-	-	1.0% (0.9%-1.2%)	1.3% (1.1%-1.7%)
	12	recall	-	-	-	4.0% (3.2%-5.3%)	3.2% (2.0%-4.8%)
Transformer	0	specificity	99.85% (99.83%-99.91%)	99.85% (99.81%-99.86%)	99.82% (99.80%-99.86%)	99.75% (99.72%-99.78%)	99.72% (99.62%-99.75%)
	0	precision	18.1% (16.5%-23.4%)	20.3% (18.1%-21.5%)	20.9% (19.7%-23.6%)	19.1% (17.9%-20.9%)	18.7% (15.9%-20.0%)
	0	recall	18.6% (15.6%-20.0%)	17.5% (16.4%-19.1%)	15.9% (14.8%-16.6%)	12.5% (11.7%-13.1%)	10.5% (10.0%-11.7%)
	3	specificity	-	99.82% (99.79%-99.96%)	99.82% (99.81%-99.84%)	99.72% (99.69%-99.80%)	99.13% (98.98%-99.22%)
	3	precision	-	1.8% (1.4%-3.1%)	4.5% (4.0%-4.9%)	5.6% (5.2%-6.7%)	2.7% (2.5%-2.9%)
	3	recall	-	6.2% (2.5%-7.7%)	6.9% (6.2%-7.5%)	5.5% (4.8%-5.8%)	5.6% (5.3%-6.2%)
	6	specificity	-	-	98.98% (98.96%-99.03%)	98.83% (98.74%-98.86%)	98.46% (90.54%-98.63%)
	6	precision	-	-	0.2% (0.2%-0.3%)	0.7% (0.7%-0.8%)	0.8% (0.7%-0.9%)
	6	recall	-	-	3.1% (2.4%-3.7%)	3.3% (3.0%-3.6%)	3.4% (3.0%-16.5%)
	12	specificity	-	-	-	98.83% (98.82%-98.84%)	98.74% (98.73%-98.80%)
	12	precision	-	-	-	0.5% (0.5%-0.6%)	0.7% (0.7%-0.8%)
	12	recall	-	-	-	3.3% (3.0%-3.6%)	2.9% (2.6%-3.2%)
Transformer - Prior knowledge disease	0	specificity	99.91% (99.79%-99.92%)	99.80% (99.78%-99.82%)	99.79% (99.78%-99.80%)	99.68% (99.65%-99.70%)	99.67% (99.33%-99.69%)
	0	precision	12.0% (7.5%-13.0%)	9.5% (8.9%-10.4%)	10.6% (10.1%-11.1%)	3.6% (3.3%-4.0%)	3.6% (2.4%-4.0%)
	0	recall	7.3% (6.7%-10.5%)	9.8% (8.9%-10.3%)	8.8% (8.4%-9.3%)	2.6% (2.4%-2.8%)	2.2% (2.0%-2.9%)

Table S4. Known risk factor disease codes.

A subset of 23 diseases (subset of the 2000 level 3 ICD codes) 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 machine learning 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 with 3 months data exclusion - Denmark DNPR & US-VA

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 for DNPR (**A**) and US-VA (**B**), both with 3 months data exclusion. - A list with features without data exclusion is in **Figure 5**. - We mapped ICD9 diagnosis codes to ICD10 for the US-VA dataset to keep the description comparable, and therefore multiple code descriptions might be listed for a given ICD9 code.

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 mell itus (11.9299)
Unspecified jaundice (14.3137)	Other diseases of pancreas (11.5739)	Non-insulin-dependent diabetes mell itus (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 mell itus (9.1564)	Non-insulin-dependent diabetes mell itus (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, a nus 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, a nus and anal canal (2.1298)	Benign neoplasm of colon, rectum, a nus 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 metabolis m 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 uns pecified 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, a nus and anal canal (1.3892)	Secondary malignant neoplasm of res piratory and digestive organs (1.4277)	Bacterial pneumonia, not elsewhere classified (1.5436)	Dyspepsia (1.0569)
Dislocation, sprain and strain of j oints and ligaments of head (1.3044)	Mental and behavioural disorders du e to use of tobacco (1.416)	Gastritis and duodenitis (1.4928)	Malignant neoplasm of stomach (1.0218)
Malignant neoplasm of small intesti ne (1.2895)	Malignant neoplasm of stomach (1.4045)	Other functional intestinal disorde rs (1.4278)	Diseases of pancreas (1.0155)
Pneumonia, organism unspecified (1.1685)	Osteoporosis without pathological f racture (1.3343)	Dyspepsia (1.4028)	Mental and behavioural disorders du e to use of tobacco (0.9639)
Osteoporosis without pathological f racture (1.1565)	Other diseases of gallbladder and b iliary (1.2574)	Delirium, not induced by alcohol an d other psychoactive substances (1.1866)	Delirium, not induced by alcohol an d other psychoactive substances (0.9083)
Other symptoms and signs involving the digestive system and abdomen (1.1477)	Acute pancreatitis (1.1292)	Hyperparathyroidism and other disor ders of parathyroid gland (1.164)	Other intervertebral disc disorders (0.8991)
Malignant neoplasm of other and uns pecified parts of biliary tract (1.1396)	Dyspepsia (1.1197)	Insulin-dependent diabetes mellitus (1.1136)	Disorders of pancreatic internal se cretion other than diabetes mellitus (0.895)
Malignant neoplasm without specific ation of site (1.133)	Bacterial pneumonia, not elsewhere classified (1.0645)	Chronic ulcer of skin (1.087)	Dislocation, sprain and strain of j oints and ligaments of shoulder girdle (0.8586)

Sequelae of poisoning by drugs, med icaments 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 j oints and ligaments of shoulder girdle (0.8825)	Postprocedural respiratory disorder s, 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 n eoplasms (0.8235)
Malignant neoplasm of cervix uteri (0.9971)	Postprocedural respiratory disorder s, not elsewhere classified (0.7585)	Secondary malignant neoplasm of res piratory 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 obst ruction 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 furth er medical care (0.7409)
Secondary malignant neoplasm of res piratory 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 j oint and ligaments of hip (0.6387)	Aortic aneurysm and dissection (0.9152)	Obesity (0.6952)
Mental and behavioural disorders du e to use of tobacco (0.8545)	Dislocation, sprain and strain of j oints and ligaments of head (0.6196)	Paralytic ileus and intestinal obst ruction without hernia (0.8735)	Personal history of malignant neopl asm (0.6901)
850 (0.8509)	Other cerebrovascular diseases (0.6025)	Osteoporosis without pathological f racture (0.8014)	Other diseases of oesophagus (0.6649)
Delirium, not induced by alcohol an d other psychoactive substances (0.7508)	Malignant neoplasm without specific ation of site (0.5877)	Malignant neoplasm of other and uns pecified parts of biliary tract (0.8013)	Dislocation, sprain and strain of j oints and ligaments at ankle and foot level (0.6579)
Malignant neoplasm of gallbladder a nd 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 du e to use of alcohol (0.7157)	Malignant neoplasm of other and uns pecified parts of biliary tract (0.5745)	850 (0.794)	Dislocation, sprain and strain of j oints 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 a nd bile ducts (0.5688)	Neoplasm of unspecified nature of d igestive organs (0.7392)	Cerebral infarction (0.6125)
Other diseases of gallbladder and b iliary (0.6058)	Gastric ulcer (0.5565)	Other septicaemia (0.717)	Disorders of sphingolipid metabolis m 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 car buncle (0.5888)
Benign neoplasm of other and ill-de fined parts of digestive system (0.5648)	Synovitis and tenosynovitis (0.5352)	Mental and behavioural disorders du e to use of tobacco (0.6664)	Transient cerebral ischaemic attack s and related syndromes (0.5777)
Other venous embolism and thrombosi s (0.5452)	Convulsions, not elsewhere classifi ed (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 j oints and ligaments of head (0.6056)	Unspecified diabetes mellitus (0.5417)
Unspecified acute lower respiratory infection (0.5144)	Disorders of sphingolipid metabolis m 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, excep t 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 an d other soft tissue (0.4607)	None (0.4646)	Chronic diseases of tonsils and ade noids (0.5745)	Umbilical hernia (0.4795)

US-VA

Disease contribution at different time to cancer (US-VA)

Cancer in 0-6 months	Cancer in 6-12 months	Cancer in 12-24 months	Cancer in 24-36 months

Acute pancreatitis Other diseases of pancreas	Non-insulin-dependent diabetes mellitus Insulin-dependent diabetes mellitus Other specified diabetes mellitus	Non-insulin-dependent diabetes mellitus Insulin-dependent diabetes mellitus Other specified diabetes mellitus	Mental and behavioral disorders due to psychoactive substances use
Non-insulin-dependent diabetes mellitus Insulin-dependent diabetes mellitus Other specified diabetes mellitus	Mental and behavioral disorders due to psychoactive substances use	Mental and behavioral disorders due to psychoactive substances use	Non-insulin-dependent diabetes mellitus Insulin-dependent diabetes mellitus Other specified diabetes mellitus
Other symptoms and signs involving the digestive system and abdomen	Persons encountering health services in other circumstances	Persons encountering health services in other circumstances	Persons encountering health services in other circumstances
Persons encountering health services in other circumstances	Other symptoms and signs involving the digestive system and abdomen	Other arthritis	Benign neoplasm of digestive system
Mental and behavioral disorders due to psychoactive substances use	Other arthritis	Other symptoms and signs involving the digestive system and abdomen	Symptoms and signs involving the digestive system and abdomen
Other diseases of biliary tract	Mental and behavioral disorders due to psychoactive substances use	Symptoms and signs involving the digestive system and abdomen	Malignant neoplasm of prostate
Symptoms and signs involving the digestive system and abdomen	Benign neoplasm of digestive system	Benign neoplasm of digestive system	Symptoms and signs involving the circulatory and respiratory systems
Benign neoplasm of digestive system	Other diseases of biliary tract	Malignant neoplasm of prostate	Other arthritis
Other arthritis	Malignant neoplasm of prostate	Mental and behavioral disorders due to psychoactive substances use	Other symptoms and signs involving the digestive system and abdomen
Symptoms and signs involving the circulatory and respiratory systems	Symptoms and signs involving the circulatory and respiratory systems	Abnormal findings on examination of blood, without diagnosis Other disorders of thyroid	Abnormal findings on examination of blood, without diagnosis Other disorders of thyroid
Other diseases of pancreas	Symptoms and signs involving the digestive system and abdomen	Symptoms and signs involving the circulatory and respiratory systems	Age-related cataract and Other cataract
Mental and behavioral disorders due to psychoactive substances use	Acute pancreatitis Other diseases of pancreas	Other diseases of biliary tract	Glaucoma
Age-related cataract and Other cataract	Viral hepatitis	Acute pancreatitis Other diseases of pancreas	Viral hepatitis
Viral hepatitis	Essential (primary) hypertension I169	Radiation-related disorders of the skin and subcutaneous tissue	Mental and behavioral disorders due to psychoactive substances use
Malignant neoplasm of prostate	Abnormal findings on examination of blood, without diagnosis Other disorders of thyroid	Age-related cataract and Other cataract	Radiation-related disorders of the skin and subcutaneous tissue
Long term (current) drug therapy Encounters for other specific health care	Cellulitis	Viral hepatitis	Other disorders of urinary system
Special screening examination for other diseases and disorders	papulosquamous disorders	Other disorders of the skin and subcutaneous tissue	Other diseases of biliary tract

Symptoms and signs involving the skin and subcutaneous tissue	Non-insulin-dependent diabetes mellitus	Other disorders of urinary system	Cellulitis
Abnormal findings on examination of blood, without diagnosis Other disorders of thyroid	Age-related cataract and Other cataract	Follow-up examination after treatment for malignant neoplasms and others	Acute pancreatitis Other diseases of pancreas
Cellulitis	Other diseases of pancreas	Symptoms and signs involving the skin and subcutaneous tissue Oedema, not elsewhere classified Unspecified jaundice	Other disorders of the skin and subcutaneous tissue
Special screening examination for other diseases and disorders	Radiation-related disorders of the skin and subcutaneous tissue	Glaucoma	Other spondylopathies Dorsalgia Other dorsopathies, not elsewhere classified
Essential (primary) hypertension Hypertension Crisis	Symptoms and signs involving the skin and subcutaneous tissue Oedema, not elsewhere classified Unspecified jaundice	Dermatitis and eczema	Dermatitis and eczema
Overweight and obesity	Dermatophytosis	Special screening examination for other diseases and disorders	Cervical disc disorders Other intervertebral disc disorders Postprocedural musculoskeletal disorders, not elsewhere classified
Symptoms and signs concerning food and fluid intake	Special screening examination for other diseases and disorders	Other spondylopathies Dorsalgia Other dorsopathies, not elsewhere classified	Disorders of skin appendages
Other spondylopathies Dorsalgia Other dorsopathies, not elsewhere classified	Other spondylopathies Dorsalgia Other dorsopathies, not elsewhere classified	Heart disease	Symptoms and signs involving the skin and subcutaneous tissue Oedema, not elsewhere classified Unspecified jaundice
Paroxysmal tachycardia Atrial fibrillation and flutter Other cardiac arrhythmias Cardiac arrest Abnormalities of heart beat	Follow-up examination after treatment for malignant neoplasms and others	Cellulitis	Special screening examination for other diseases and disorders
Presence of other functional implants	Glaucoma	Diverticular disease of intestine	Heart disease
Other medical care	Other medical care	Encounters for other specific health care Long term (current) drug therapy	Follow-up examination after treatment for malignant neoplasms and others
Glaucoma	Other disorders of urinary system	papulosquamous disorders	Presence of other devices And other functional implants
Dermatophytosis	Other disorders of the skin and subcutaneous tissue	Other benign neoplasms of skin	Persons with potential health hazards related to socioeconomic and psychosocial circumstances
Follow-up examination after treatment for malignant neoplasms and others	Dermatitis and eczema	Cervical disc disorders Other intervertebral disc disorders Postprocedural musculoskeletal disorders, not elsewhere classified	Acquired deformities of fingers and toes

Other diseases of the urinary system	Cervical disc disorders Other intervertebral disc disorders Postprocedural musculoskeletal disorders, not elsewhere classified	Persons with potential health hazards related to socioeconomic and psychosocial circumstances	Other forms of heart disease
Symptoms and signs involving the skin and subcutaneous tissue Edema, not elsewhere classified Unspecified jaundice	Heart disease	Radiation-related disorders of the skin and subcutaneous tissue	Mental and behavioural disorders due to use of alcohol
Papulosquamous disorders	Persons with potential health hazards related to socioeconomic and psychosocial circumstances	Other malignant neoplasms of skin	Other malignant neoplasms of skin
Other disorders of the skin and subcutaneous tissue	Cholelithiasis	Acquired deformities of fingers and toes	Other benign neoplasms of skin
Other forms of heart disease	Symptoms and signs concerning food and fluid intake	Dermatophytosis	Diverticular disease of intestine
Diverticular disease of intestine	Other forms of heart disease	Other forms of heart disease	Nail disorders
Persons with potential health hazards related to socioeconomic and psychosocial circumstances	Corns and callosities	Pulmonary disease Chronic bronchitis	Varicose veins of lower extremities
Dermatitis and eczema	Diverticular disease of intestine	Mental and behavioral disorders due to use of alcohol	Symptoms and signs concerning food and fluid intake
Disorders of refraction and accommodation	Presence of other devices And other functional implants	Other diseases of digestive system	Other disorders of male genital organs
Cholecystitis Other diseases of gallbladder	Other diseases of digestive system	Varicose veins of lower extremities	Asthma Other chronic obstructive pulmonary disease
Other benign neoplasms of skin	Overweight and obesity	Disorders of arteries, arterioles and capillaries in diseases classified elsewhere	Calculus of kidney and ureter
Varicose veins of lower extremities	Other disorders of male genital organs	Personal history of other diseases and conditions	Disorders of choroid and retina
Malignant neoplasm of trachea Malignant neoplasm of bronchus and lung	Acquired deformities of fingers and toes	Presence of other devices And other functional implants	Herpesviral [herpes simplex] infections
Cholelithiasis	Disorders of arteries, arterioles and capillaries in diseases classified elsewhere	Other cerebrovascular diseases	Other venous embolism and thrombosis
Disorders of skin appendages	Other benign neoplasms of skin	Nail disorders	Dermatophytosis
Non-insulin-dependent diabetes mellitus	Encounters for other specific health care Long term (current) drug therapy	Other disorders of male genital organs	Angina pectoris
Mental and behavioral disorders due to use of alcohol	Nonrheumatic aortic valve disorders	Heart failure	Disorders of arteries, arterioles and capillaries in diseases classified elsewhere

Disorders of arteries, arterioles and capillaries in diseases classified elsewhere	Disorders of skin appendages	Non-insulin-dependent diabetes mellitus	Disorders of external ear in diseases classified elsewhere
Other disorders of male genital organs	Mononeuropathies of upper limb Other mononeuropathies	Disorders of external ear in diseases classified elsewhere	Inguinal hernia

Figure S2. Distribution of disease codes as a function of age in the DNPR (Denmark) 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

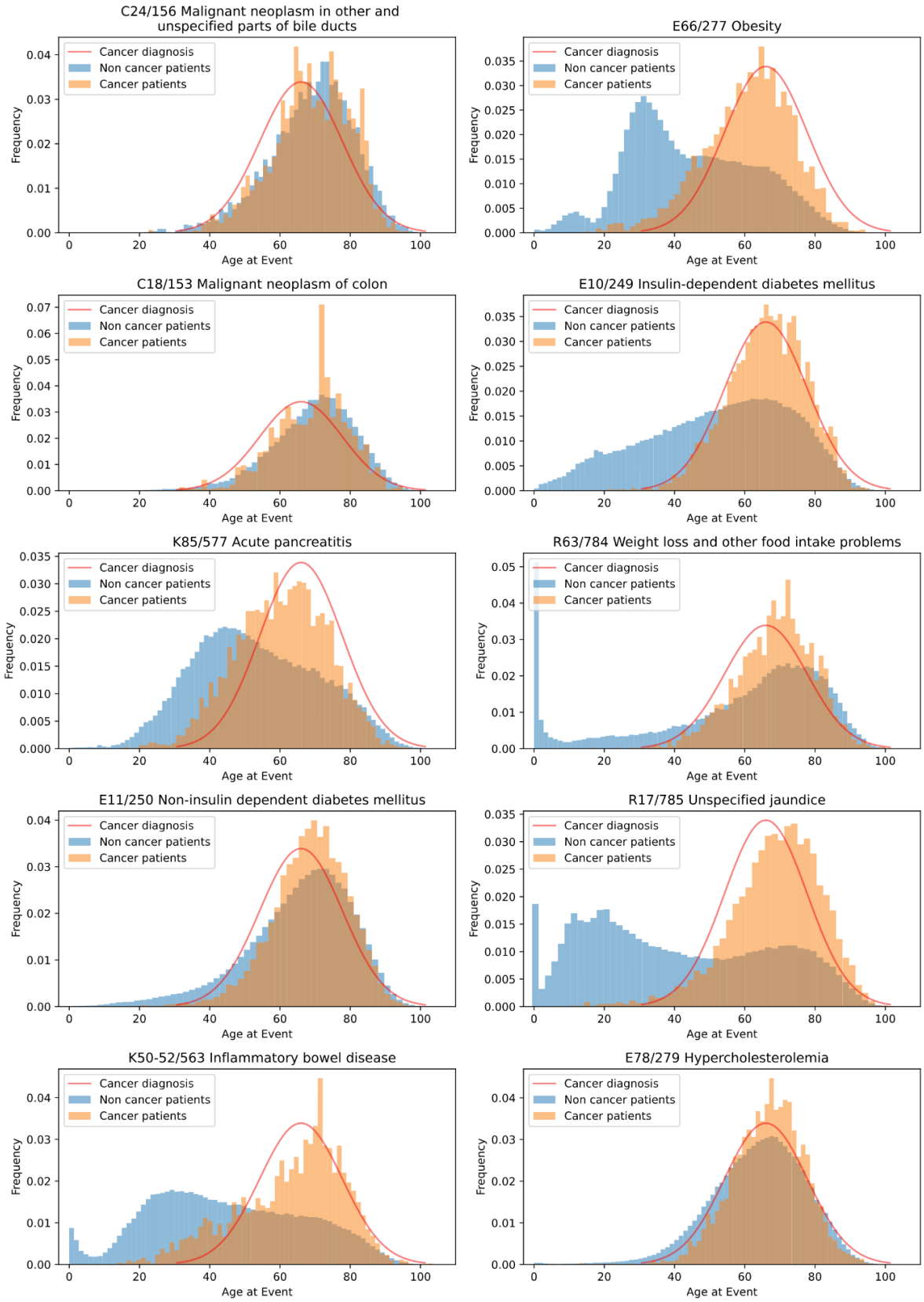
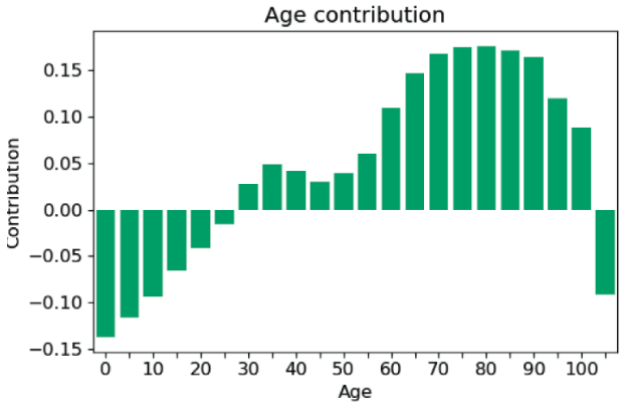


Figure S3. Age as a contributing factor

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 prediction-surveillance program, to be explored in a clinical setting, 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 on a yearly basis and, if identified as high-risk, offered a series of follow-up visits with 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 a diversity of scenarios, it is reasonable to propose a clinical prediction-surveillance program tailored to the health system in a particular country.

To illustrate the economic benefits of such a screening and to stimulate discussion regarding the optimal design of a prediction-surveillance program, we have made a first-order-estimate for a clinical screening trial of 10,000 people using the best model

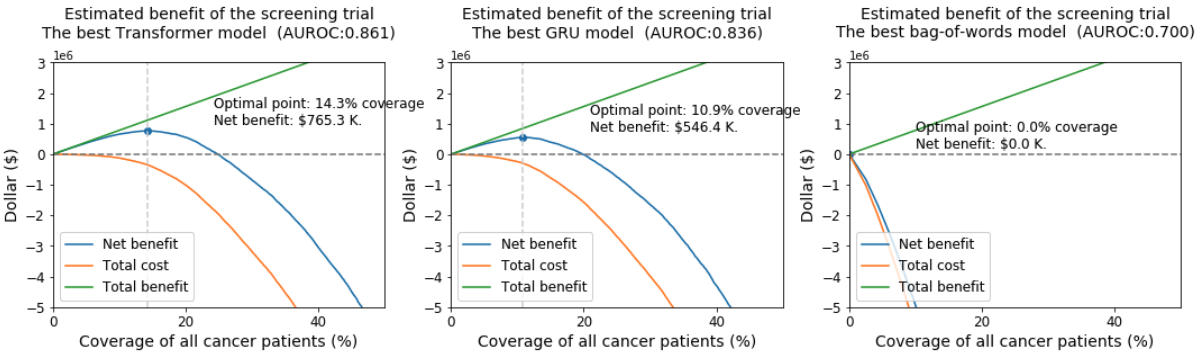
(the transformer model). For simplicity, we have made no assumptions regarding age distribution. Here is a simple economic model.

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).



<u>Estimate of financial benefits for different models.</u> 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 prediction-surveillance 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 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.