European Journal of Cancer 143 (2021) 19–30

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejcancer.com

Original Research

Development and validation of a pancreatic cancer risk model for the general population using electronic health records: An observational study

Limor Appelbaum [a,*,1], José P. Cambronero [b,1], Jennifer P. Stevens [c], Steven Horng [d], Karla Pollick [c], George Silva [c], Sebastien Haneuse [e], Gail Piatkowski [c], Nordine Benhaga [a], Stacey Duey [f], Mary A. Stevenson [a], Harvey Mamon [g], Irving D. Kaplan [a,2], Martin C. Rinard [b,2]

[a] *Beth Israel Deaconess Medical Center, Department of Radiation Oncology, 330 Brookline Ave, Boston, MA, 02215, USA*
[b] *Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory, 32 Vassar St, Cambridge, MA, 02139, USA*
[c] *Beth Israel Deaconess Medical Center, Center for Healthcare Delivery Science, 330 Brookline Ave, Boston, MA, 02215, USA*
[d] *Beth Israel Deaconess Medical Center, Division of Emergency Medicine Informatics, 330 Brookline Ave, Boston, MA, 02215, USA*
[e] *Harvard University, T.H. Chan School of Public Health, 677 Huntington Ave, Boston, MA, 02115, USA*
[f] *Brigham and Women's Hospital, Partners Research IS and Computing, Information Systems Department, 75 Francis Street, Boston, MA, 02115, USA*
[g] *Dana Farber Cancer Institute/Radiation Oncology, Brigham and Women's Hospital, Harvard Medical School, 75 Francis Street, Boston, MA, 02115, USA*

Received 3 July 2020; received in revised form 15 October 2020; accepted 28 October 2020

KEYWORDS
Pancreatic carcinoma;
Adenocarcinoma;
Electronic health records;

Abstract ***Aim:*** Pancreatic ductal adenocarcinoma (PDAC) is often diagnosed at a late, incurable stage. We sought to determine whether individuals at high risk of developing PDAC could be identified early using routinely collected data.

Methods: Electronic health record (EHR) databases from two independent hospitals in Boston, Massachusetts, providing inpatient, outpatient, and emergency care, from 1979 through

* *Corresponding author*: 330 Brookline Avenue, Boston, MA, 02215, USA.

E-mail address: lappelb1@bidmc.harvard.edu (L. Appelbaum), jcamsan@mit.edu (J.P. Cambronero), jpsteven@bidmc.harvard.edu (J.P. Stevens), shorng@bidmc.harvard.edu (S. Horng), kpollick@bidmc.harvard.edu (K. Pollick), gssilva@bidmc.harvard.edu (G. Silva), shaneuse@hsph.harvard.edu (S. Haneuse), gpiatkow@bidmc.harvard.edu (G. Piatkowski), nbenhaga@bidmc.harvard.edu (N. Benhaga), sduey@partners.org (S. Duey), mstevens@bidmc.harvard.edu (M.A. Stevenson), hmamon@bwh.harvard.edu (H. Mamon), ikaplan@bidmc.harvard.edu (I.D. Kaplan), rinard@csail.mit.edu (M.C. Rinard).

[1] L.A. and J.P.C. contributed equally to this work. [2] I.D.K. and M.C.R. contributed equally to this work.

https://doi.org/10.1016/j.ejca.2020.10.019

Logistic regression models;
AUC

2017, were used with case–control matching. PDAC cases were selected using International Classification of Diseases 9/10 codes and validated with tumour registries.

A data-driven feature selection approach was used to develop neural networks and L2-regularised logistic regression (LR) models on training data (594 cases, 100,787 controls) and compared with a published model based on hand-selected diagnoses ('baseline'). Model performance was validated on an external database (408 cases, 160,185 controls). Three prediction lead times (180, 270 and 365 days) were considered.

Results: The LR model had the best performance, with an area under the curve (AUC) of 0.71 (confidence interval [CI]: 0.67–0.76) for the training set, and AUC 0.68 (CI: 0.65–0.71) for the validation set, 365 days before diagnosis. Data-driven feature selection improved results over 'baseline' (AUC = 0.55; CI: 0.52–0.58).

The LR model flags 2692 (CI 2592–2791) of 156,485 as high risk, 365 days in advance, identifying 25 (CI: 16–36) cancer patients. Risk stratification showed that the high-risk group presented a cancer rate 3 to 5 times the prevalence in our data set.

Conclusion: A simple EHR model, based on diagnoses, can identify high-risk individuals for PDAC up to one year in advance. This inexpensive, systematic approach may serve as the first sieve for selection of individuals for PDAC screening programs.

1. Introduction

Pancreatic ductal adenocarcinoma (PDAC) is the third most lethal cancer in the USA, with annual incidence on the rise since the 1990s [1]. In the United States alone, estimates predict 56,770 new cases of pancreatic cancer and 45,750 deaths in 2019 [2].

Despite significant progress in understanding its biology [3], PDAC is often diagnosed at an advanced stage. However, studies have shown that symptoms exist months to years before diagnosis [4,5]. For example, new-onset diabetes occurs in up to 50% of patients, 24–36 months before diagnosis [6,7], with weight loss beginning as early as 12–15 months before diagnosis [8,9]. In addition, metabolomics studies suggest that tumour-induced biochemical processes such as lipolysis [10,11], low-density lipoprotein consumption [12], and protein breakdown [13] occur during PDAC development.

Detection of early-stage disease has been shown to be effective in individuals at high risk for PDAC due to genetic predisposition. Screening performed in these patients, resulted in a significantly better three-year survival for screened patients, than for patients whose tumours were detected after symptom development (85% versus 25%, respectively). Detection of lesions in their high-grade precursor stages resulted in an overall survival of 100% [14].

The increasing availability of electronic health records (EHRs) – 9.4% of non-federal acute-care hospitals in the US had basic EHRs in 2008 compared with 83.8% in 2015 [15] – and the varied types of indicators they provide have increased their use as a data source for risk prediction models [16].

Our hypothesis is that a simplistic model based on EHR diagnostic codes can be effective in early identification of individuals at increased risk of developing pancreatic cancer. Prior PDAC risk prediction models have often focused on specific high-risk subgroups [17–19], such as diabetics, and use a relatively parsimonious set of handcrafted known PDAC-associated features [17–22]. Only some of these models consider a prediction time window in which high-risk individuals can be detected [5,17,18,20]. Previous work has used diverse data sources such as survey data and case–control or cohort studies to train their models, but few use EHR databases [17,18].

In this study, we develop and validate two novel machine learning PDAC risk models based on patients' prior diagnoses derived from EHR data, to predict pancreatic cancer 6–12 months before an eventual diagnosis date.

2. Methods

This study was exempt from review by the institutional review board (IRB) of Dana-Farber Cancer Institute (DFCI). Beth Israel Deaconess Medical Center (BIDMC) IRB ceded review to DFCI.

2.1. Data source

This study is a retrospective case–control analysis, using two EHRs data sets: BIDMC data and Partners HealthCare (PHC) data. BIDMC data were sourced from the BIDMC Clinical Data Repository (CDR) that contains data from the hospital emergency department, inpatient and outpatient medical records (OMR), PHC data were sourced from the Partners' Research Patient Data Registry (RPDR), which includes data from two large hospitals, Brigham and Women's Hospital (BWH) and Massachusetts General Hospital, as well as community and speciality hospitals in the Boston area.

2.2. Cohort selection

PDAC patients of all ages were identified by International Classification of Diseases (ICD) 9 157.x and ICD10 C25.x codes, from July 1997 to December 2017 (BIDMC) and 1979–2017 (PHC). Patients with pancreatic neuroendocrine tumours (ICD9 code 157.4 and ICD10 code C25.4) were excluded.

To ensure inclusion of patients for whom pathology was lacking/inconclusive and diagnosis was clinical (based on tumour blood markers, imaging and/or multidisciplinary tumour board agreement) rather than histologic, we chose to utilise ICD codes for case selection and not pathology reports.

2.3. Cohort validation

To ensure model training on confirmed cases, cancer patients were cross-checked against the BIDMC tumour registry: patients with no corresponding registry record were removed. Based on our experience with the BIDMC data set, we used the BWH tumour registry upfront to query the RPDR. Patients without a corresponding PDAC ICD code (essential for index date determination) were excluded.

2.4. Controls

Patients who interacted with BIDMC between 2005 and 2017, and were never diagnosed with PDAC, were included in the BIDMC data control group. A PDAC diagnosis at other sites of care is possible; however, we assume it is unlikely given the long track history at BIDMC for the majority of controls.

2.5. Case–control matching

Controls were matched to cases based on sex and age to limit potential confounding by these factors on other comorbidities (i.e. hypertension, diabetes, heart disease). No other criteria were used for matching to maximise the ability of machine learning to identify other discriminating features.

In an attempt to account, at least in part, for the relative rarity of this disease, approximately 170 age- and sex- matched controls were sampled for each cancer patient in the training set. For validation, control patients in PHC data were sampled by querying RDPR for individuals who never had a PDAC diagnosis in a ~1:150 ratio to cases.

The first encounter immediately after diagnosis was defined as the index date, and, hence, age was determined at the date of diagnosis, with all dates computed relative to the diagnosis date.

We considered controls' age as of their latest clinical interaction and computed all dates relative to this index date.

2.6. Missing data

Patients from outside practices are often referred after PDAC diagnosis to BIDMC/BWH as these are large tertiary-care centres. In line with this, we observed a significantly higher rate of missing historic data in the cancer group before PDAC diagnosis. To minimise potential bias caused by missing information for referred cases, we excluded patients in both data sets that did not have an entry in their EHR at least 6 months (or earlier) before the diagnosis date (if cancer) or the last encounter (if control). This is referred to as 'heuristic filter'. We assume that the six-month period immediately adjacent to PDAC diagnosis, likely represents the cancer workup period (with a safety margin).

2.7. Premature death

Collecting observation data for PDAC patients poses the challenge of premature death (i.e. the patient dying early on and providing limited or no observational data in our data set). To mitigate this risk, we focused our data collection on a major hospital system which increases the volume of PDAC patients observed and by the use of the heuristic filter to ensure our patients have an observable history.

2.8. Prediction time cut-offs

Cut-offs of 180, 270 and 365 days before the index date were chosen based on the assumption that the tumour would most likely still be in its precursor or very early stages (T1a/T1b) in this time frame. We consider patients' history earlier than each cut-off date (Supplementary Figure 1).

2.9. Training and testing

BIDMC data were split into training (80%) and test (20%) splits. Hyperparameters were tuned over the training split (see also 'algorithms'). After choosing the final hyperparameters, we trained on the full training split and evaluated on the test split (reported as BIDMC test). The entire BIDMC data were trained to produce fully trained models.

To reduce the risk of measurement error which results from lack of a gold standard tool for comparison, we evaluated model performance on an external validation data set: PHC data (reported as PHC test).

Ten-fold cross-validation (CV) was performed on PHC data, and the training folds were used to re-learn model weights for the same diagnosis codes used in the BIDMC model. The test folds were used for evaluation (reported as PHC-retrained CV).

2.10. Performance metrics

Similar to previous work [20–23], models were compared using rea under the receiver operating characteristic curve (AUROC). Model sensitivity and positive predictive values (PPVs) at different specificity cut-offs are presented, as well as an analysis of our risk score, for our external validation.

2.11. Feature selection

In contrast to prior work, we did not predefine the set of diagnoses used. Any diagnosis observed less than 100 times in the training split of BIDMC data was removed, resulting in a final set of 4150 diagnosis codes. L2 regularisation and dropout were used to address possible overfitting.

2.12. Statistical analysis

95% confidence intervals computed using the empirical bootstrap are reported, with test observations resampled 1000 times.

AUROCs were pairwise compared using a two-sided DeLong test, a conservative non-parametric test with null hypothesis of no difference. A p-value below 0.05 was considered statistically significant. To correct for multiple comparisons, the conservative Bonferroni correction was used.

2.13. Modelling

2.13.1. Baseline model

Because we did not predefine the set of diagnoses used, we assessed if this approach would outperform a manual feature selection approach, by comparing our model to a recently published one in which a logistic regression (LR) was used over expert-identified diagnoses from Medicare claims [20]. To build our baseline, we used the ICD9 codes published in the study by Baecker et al. [20] and automatically extended these with corresponding ICD10 codes where possible. Demographic factors outlined in the original paper were also included: the patient gender, five-year age group, and race. This model was trained on our data sets to perform comparisons, and is referred to as clinical LR (baseline).

2.13.2. Risk score

Multiplicative PDAC risk score were computed using the weights learnt by LR. A patient's risk score is defined as the odds implied by his diagnoses and model coefficients.

Risk groups are based on the distribution of scores in the test split of BIDMC data, where a low risk score is below the 75th percentile, an intermediate score is between the 75th and 99th percentile, and high score is above the 99th percentile. Thresholds were computed separately for each censoring period.

For 'features', 'algorithms' and 'software, see supplementary material.

3. Results

From the records of 1,099,321 patients that were accessible to us from the BIDMC CDR (1997–2017), after applying the aforementioned exclusion criteria, we identified 594 eligible cases, and 100,787 eligible age- and sex- matched controls (Figure 1).

PHC data include diagnoses data for 408 eligible cases and 160,185 eligible age- and sex- matched controls, from BWH inpatient and outpatient records, used to externally validate our results (Figure 2).

Demographic and clinical indicators across BIDMC data (594 cases, 100,787 controls) and PHC data (408 cases, 160,185 controls) are shown in Table 1. Figures 1 and 2 show flow diagrams for BIDMC data and PHC data, respectively.

3.1. Model performance

The LR model had an area under the curve (AUC) of 0.71 (CI: 0.67–0.76) for the training set, and AUC of 0.68 (CI: 0.65–0.71) for the validation set, 365 days before diagnosis (Figure 3 and Table 2). LR had a superior performance to both the feed-forward neural network (NN) and baseline models in AUC, as well as in sensitivities at different specificities (Supplementary Table 1) and in PPVs (Supplementary Table 2).

NNs and LR are not significantly different in the BIDMC test, but the same comparison in PHC test yields significant differences for the 180 and 365 day cut-offs (Figure 3 and Table 2).

At the 365 day cut-off, the 'baseline' model had an AUC of 0.55; CI: 0.52–0.58. The average AUC for LR outperforms the baseline (clinical LR [baseline] in all experiments. This difference in AUC is statistically significant (p-value <0.05) across all cut-offs in the BIDMC test and PHC test (Figure 3 and Table 2).

3.2. Risk score distribution

The LR model flags 2692 (CI 2592–2791) out of 156,485 individuals as high risk, 365 days in advance, identifying 25 (CI 16–36) cancer patients (Figure 4 and Supplementary Table 3). Cancer patients' empirical cumulative distribution function (ECDF) dominates controls'. Risk stratification applied to our external

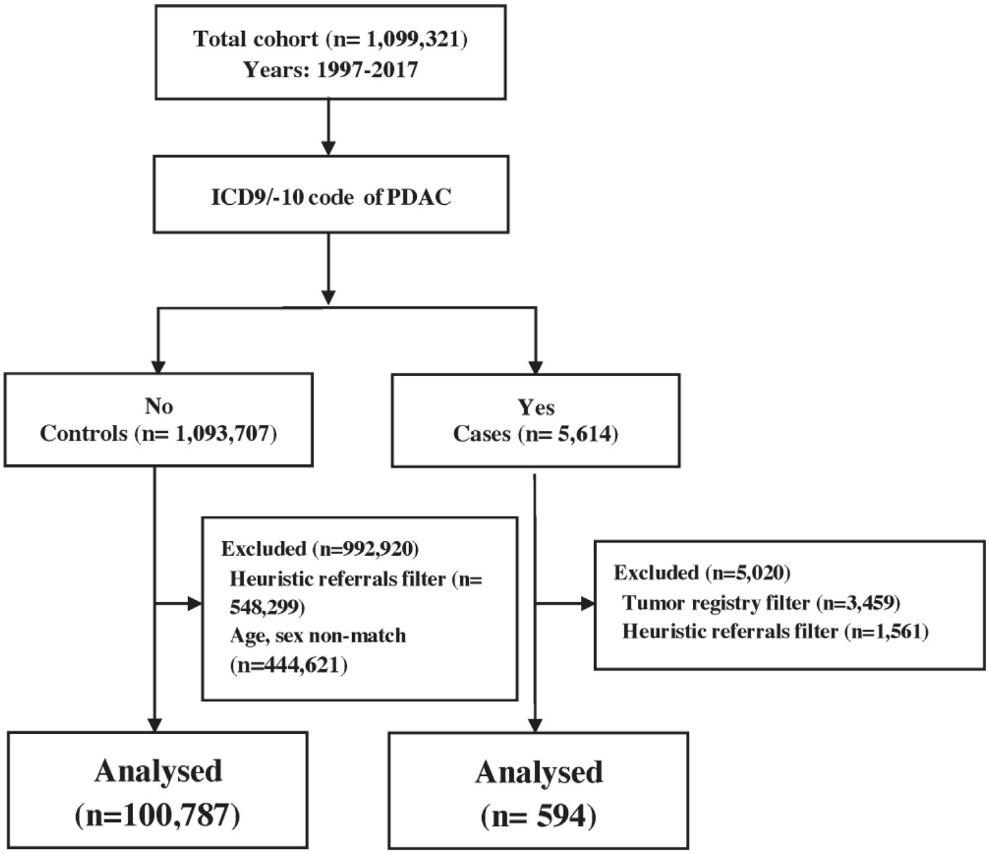
Total cohort (n= 1,099,321)
Years: 1997-2017
ICD9/-10 code of PDAC
No
Controls (n= 1,093,707)
Yes
Cases (n= 5,614)
Excluded (n=992,920)
Heuristic referrals filter (n= 548,299)
Age, sex non-match (n=444,621)
Excluded (n=5,020)
Tumor registry filter (n=3,459)
Heuristic referrals filter (n=1,561)
Analysed (n=100,787)
Analysed (n= 594)

Fig. 1. The flow diagram for BIDMC data. Cases were initially identified using ICD codes for PDAC. Cases and remaining controls that did not have data going back further than 6 months from the date of PDAC diagnosis were then excluded ('heuristic filter'). Only those cases that were also in the tumour registry were left for analysis. Controls were sex and age matched to those cases. BIDMC, Beth Israel Deaconess Medical Center; ICD, International Classification of Diseases; PDAC, pancreatic adenocarcinoma.

validation data set (PHC data) showed that the high-risk group presented a rate of cancer between three and five times (e.g. 1.18% at 6 months cut-off), the prevalence in our data set (0.25%), With censoring of 180, 270, and 365 days, LR identifies as high risk: 2548 (CI 2452–2651) with 30 cases (CI 19–41) out of 160,593 patients; 2634 (CI 2535–2745) with 22 cases (CI 13–32) out of 158,518 patients; and 2692 (CI 2592–2791) with 25 cases (CI 16–36) out of 156,485 patients, respectively (Figure 5 and Supplementary Table 3).

3.3. PDAC correlates

Supplementary Table 4 shows the top 20 (of 4150) diagnoses based on LR weights when trained on BIDMC data using 365-day censoring. Diagnoses include known correlates, such as diabetes, personal history of cancer, and tobacco use, and novel correlates such as hypertension. Complete model weights for LR can be downloaded from https://github.com/ josepablocam/pdac-diag-model-matched/tree/master/model-weights.

4. Discussion

In this study, we developed a PDAC risk model using EHR diagnoses and evaluated it on an external data set.

Importantly, we show that high-risk individuals can be identified 6–12 months before PDAC diagnosis, a time 'window' in which tumour detection would likely allow for curative resection [24,14]. A time period closer to the actual diagnosis date would likely be less clinically meaningful.

Training on a broad set of diagnoses, similar to the 'data-driven' approach taken by Barak-Corren et al. [25] and Razavian et al. [26], rather than on a handcrafted feature set, as used in other models [20,22], improved discrimination. Internal validation on the test split of BIDMC data showed that LR obtained a higher AUROC (0.71) than clinical LR (0.60) a year before diagnosis, and this discrimination differential generalised to a separate patient population (PHC data).

Top diagnoses associated with PDAC incidence included known correlates such as diabetes mellitus, obesity, and smoking, but interestingly also previously unknown correlates, such as certain skin conditions and hypertension. We believe these factors could serve

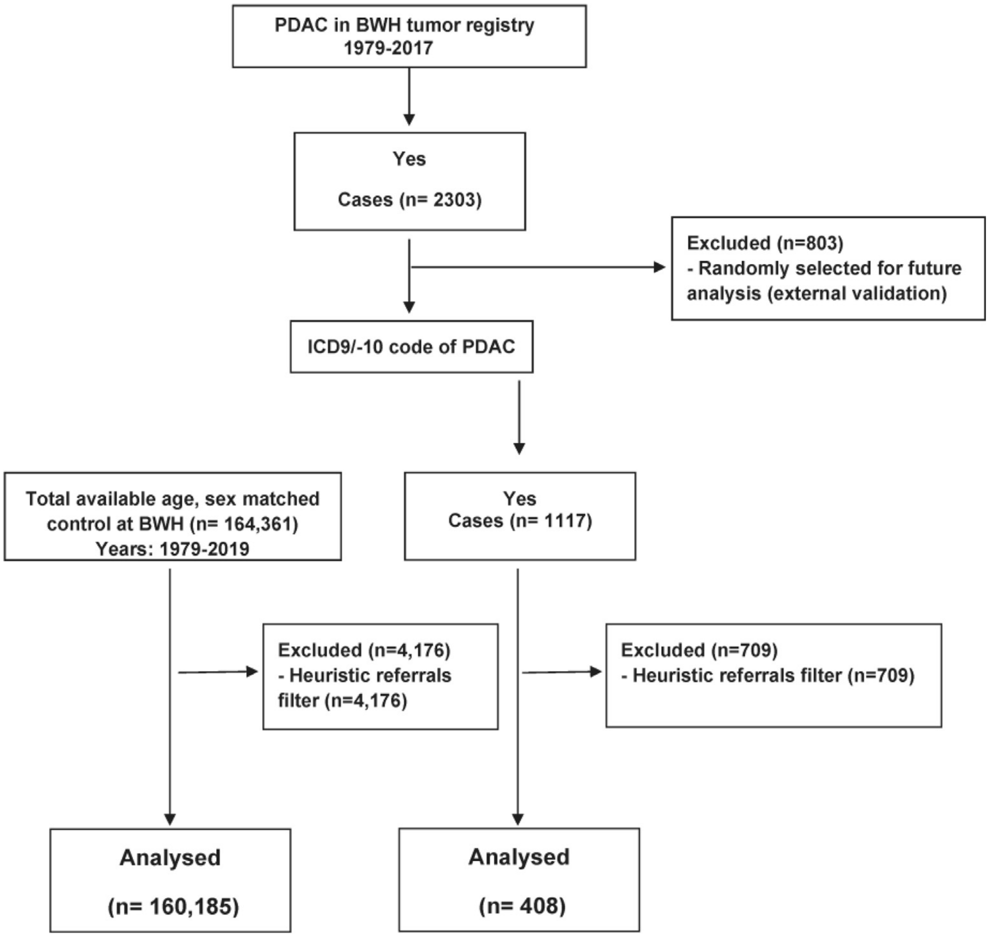
PDAC in BWH tumor registry
1979-2017
Yes
Cases (n= 2303)
Excluded (n=803)
- Randomly selected for future analysis (external validation)
ICD9/-10 code of PDAC
Total available age, sex matched control at BWH (n= 164,361)
Years: 1979-2019
Yes
Cases (n= 1117)
Excluded (n=4,176)
- Heuristic referrals filter (n=4,176)
Excluded (n=709)
- Heuristic referrals filter (n=709)
Analysed
(n= 160,185)
Analysed
(n= 408)

Fig. 2. The flow diagram for PHC data. Cases were initially identified using the BWH tumour registry and cross-checked with PDAC ICD codes. These were matched by age and sex to controls. The 'heuristic filter' was applied to cases and controls to exclude referrals. BWH, Brigham and Women's Hospital; ICD, International Classification of Diseases; PDAC, pancreatic adenocarcinoma; PHC, Partners HealthCare.

as a valuable candidate list for future clinical investigation.

ICD codes are a systematic list of medical classifications, compiled by the World Health Organization, and used in healthcare systems worldwide. These codes represent a simplification of the patient's clinical status and include information on diseases, signs, symptoms, complaints, abnormal laboratory findings, and social circumstances (http://www.who.int/classifications/icd/factsheet/en/).

Models based on ICD codes, such as ours, enable reproduction and can facilitate inexpensive implementation [28]. In contrast, previous PDAC models have focused on data that are less readily available, such as survey data or previous case–control studies [21,22].

ICD codes are naturally limited: they are influenced by institutional/country practices, may conflate disease phenotypes, and are only available for patients that interact with an institution with appropriate record keeping. However, numerous risk models have proven their usefulness [29–31].

To mitigate the impact of ICD codes on our ability to identify true PDAC cases during data collection, we validated ICD-coded cases with tumour registries. This step addressed potential false positives in our training data. We emphasise that this manual validation is only required to ensure valid training, but is not required for end users.

Although our validation on an independent data set demonstrated generalisation, it was applied only to the American population. Therefore, additional prospective validation across a larger range of institutions and diverse populations is required. In spite of this, we believe that it may be applicable to at least some of the European population, as demonstrated by prior work (18) utilising a similar approach to ours, performed on a UK population database.

We would also like to point out that the use of observational healthcare data (e.g. EHR) to learn predictive models in inevitably affected by the systemic issues prevalent in our healthcare system: in particular the varying level and quality of access to health care in vulnerable populations. For our study, we sought to mitigate this issue by collecting as large a data set as possible from two different major hospital systems in a major urban centre, where we had a better chance observing data for patients in non-White population (a key group in the USA). However, because published literature [32] has shown significant differences in PDAC

Table 1
Patients characteristics.

Stat	Value	BIDMC cases	BIDMC controls	PHC cases	PHC controls
Age	<50	26 (4.4)	4418 (4.4)	18 (4.4)	12,642 (7.9)
Age	50–60	91 (15.3)	15,443 (15.3)	65 (15.9)	25,377 (15.8)
Age	60–65	63 (10.6)	10,689 (10.6)	63 (15.4)	20,002 (12.5)
Age	65–70	93 (15.7)	15,778 (15.7)	67 (16.4)	22,345 (13.9)
Age	70–75	113 (19.0)	19,167 (19.0)	53 (13.0)	24,173 (15.1)
Age	75–80	80 (13.5)	13,572 (13.5)	71 (17.4)	17,736 (11.1)
Age	>80	128 (21.5)	21,720 (21.6)	71 (17.4)	37,910 (23.7)
Gender	Female	307 (51.7)	52,089 (51.7)	219 (53.7)	83,997 (52.4)
Gender	Male	287 (48.3)	48,698 (48.3)	189 (46.3)	76,188 (47.6)
Race	American Indian	1 (0.2)	93 (0.1)	2 (0.5)	220 (0.1)
Race	Asian	19 (3.2)	8461 (8.4)	5 (1.2)	3036 (1.9)
Race	Black	51 (8.6)	7977 (7.9)	31 (7.6)	8379 (5.2)
Race	Latino	14 (2.4)	3485 (3.5)	12 (2.9)	4817 (3.0)
Race	Middle Eastern	0 (0.0)	26 (0.0)	0 (0.0)	2 (0.0)
Race	Native American	18 (3.0)	6551 (6.5)	3 (0.7)	2421 (1.5)
Race	Pacific Islander	2 (0.3)	52 (0.1)	0 (0.0)	77 (0.0)
Race	Unknown	30 (5.1)	4176 (4.1)	23 (5.6)	30,756 (19.2)
Race	White	459 (77.3)	69,966 (69.4)	332 (81.4)	110,477 (69.0)
Diagnoses	Abdominal pain	248 (41.8)	20,477 (20.3)	172 (42.2)	18,029 (11.3)
Diagnoses	Angina pectoris	30 (5.1)	6359 (6.3)	17 (4.2)	5140 (3.2)
Diagnoses	Asthma	69 (11.6)	7326 (7.3)	33 (8.1)	6086 (3.8)
Diagnoses	Atherosclerotic heart disease	116 (19.5)	12,551 (12.5)	66 (16.2)	12,227 (7.6)
Diagnoses	Calculus gallbladder	55 (9.3)	3197 (3.2)	30 (7.4)	2686 (1.7)
Diagnoses	Chest pain	114 (19.2)	15,517 (15.4)	130 (31.9)	24,394 (15.2)
Diagnoses	Chronic pancreatitis	54 (9.1)	694 (0.7)	27 (6.6)	451 (0.3)
Diagnoses	Coronary heart disease	1 (0.2)	119 (0.1)	0 (0.0)	48 (0.0)
Diagnoses	Diabetes mellitus	222 (37.4)	22,474 (22.3)	108 (26.5)	13,518 (8.4)
Diagnoses	Emphysema	12 (2.0)	2464 (2.4)	9 (2.2)	2276 (1.4)
Diagnoses	Essential hypertension	382 (64.3)	49,673 (49.3)	196 (48.0)	35,454 (22.1)
Diagnoses	Family history pancreatic cancer	35 (5.9)	5607 (5.6)	16 (3.9)	1842 (1.1)
Diagnoses	Jaundice	71 (12.0)	494 (0.5)	75 (18.4)	502 (0.3)
Diagnoses	Stroke	42 (7.1)	6842 (6.8)	40 (9.8)	8226 (5.1)
Diagnoses	Ulcer	24 (4.0)	2075 (2.1)	19 (4.7)	1287 (0.8)

Counts are performed with the latest values for each patient, without censoring based on the prediction cut-off date. Percentages are calculated within each case/control group, as appropriate. Population characteristics reflect hospital population distribution in the Boston metropolitan area. BIDMC, Beth Israel Deaconess Medical Center; PHC, Partners HealthCare.

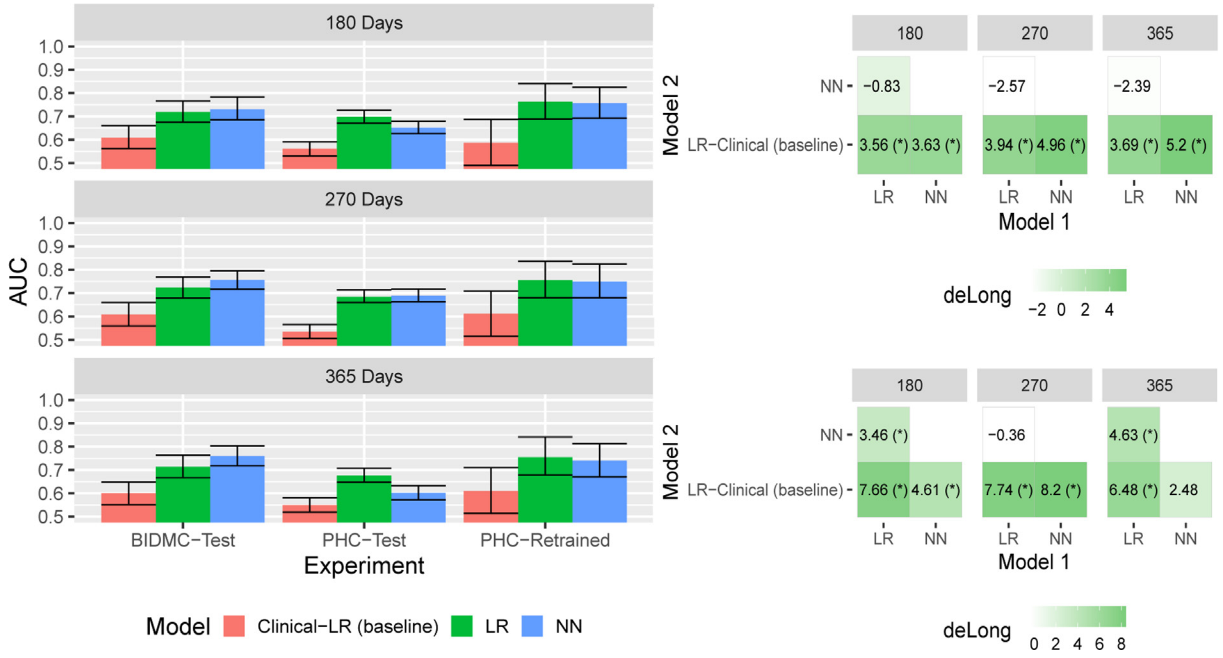

Fig. 3. Comparison of performance across models. (a) Comparison of AUC across the three models: clinical LR, LR, and NNs. Clinical LR refers to a published model using pre-selected features, which we refit to matching diagnostic codes from our data set, with the purpose of establishing a baseline to which we could compare our models developed using a data-driven approach to manual feature selection. LR refers to the models developed, validated and retrained using logistic regression. The NN refers to the models using feed-forward neural networks. The BIDMC test presents performance on the test split of BIDMC data for models trained on the training split of BIDMC data. The PHC test presents performance when models are trained on all of BIDMC data and evaluated on PHC data. PHC-retrained CV presents 10-fold CV performance where the training folds are used to relearn weights. Note that LR performs similarly in the BIDMC test and PHC test, demonstrating that the model can be applied to new data. (b) Statistically significant (p-value <0.05) DeLong tests are noted with an asterisk. The Delong test is used, given that AUC is computed over the same population. Bonferroni adjustment was performed. The BIDMC test AUCs for LR and NNs, when compared with those of clinical LR (baseline), are significantly different for all cut-offs, while they are not significantly different from each other. (c) PHC test AUCs for LR, when compared with those of clinical LR (baseline), are statistically different for all cut-offs. LR and NN results are significantly different from each other for 180 day and 365 day cut-offs. AUC, Area under the curve; BIDMC, Beth Israel Deaconess Medical Center; LR, logistic regression; NN, neural network; PHC, Partners HealthCare.

Table 2
Summary of AUCs across BIDMC data and PHC data.

Model	Cut-off	BIDMC test	PHC test	PHC-retrained
Clinical LR (baseline)	180	0.61 (0.56 −0.66)	0.56 (0.53 −0.59)	0.59 (0.49 −0.69)
LR	180	0.72 (0.67 −0.77)	0.70 (0.67 −0.73)	0.76 (0.69 −0.84)
NN	180	0.73 (0.69 −0.78)	0.65 (0.63 −0.68)	0.76 (0.69 −0.83)
Clinical LR (baseline)	270	0.61 (0.56 −0.66)	0.54 (0.51 −0.57)	0.61 (0.52 −0.71)
LR	270	0.72 (0.68 −0.77)	0.68 (0.66 −0.71)	0.76 (0.68 −0.84)
NN	270	0.76 (0.72 −0.80)	0.69 (0.66 −0.72)	0.75 (0.68 −0.82)
Clinical LR (baseline)	365	0.60 (0.55 −0.65)	0.55 (0.52 −0.58)	0.61 (0.51 −0.71)
LR	365	0.71 (0.67 −0.76)	0.68 (0.65 −0.71)	0.75 (0.68 −0.84)
NN	365	0.76 (0.72 −0.80)	0.60 (0.57 −0.63)	0.74 (0.67 −0.81)

Values in parentheses correspond to 95% confidence intervals estimated by bootstrapping.
BIDMC, Beth Israel Deaconess Medical Center; LR, logistic regression; NN, neural network; PHC, Partners HealthCare.

risk for specific groups, this topic should be further explored in future work.

Furthermore, our observations are limited to a population receiving care in either the hospital outpatient clinics or as inpatients, where diagnoses' distributions may differ from those interacting with community-based services only.

PDAC incidence in the general population is low [1]. We attempt to reflect this in our population by using an imbalance of cases to controls. Although our resulting prevalence is still higher, it represents a methodological improvement over most existing PDAC modelling literature [20–22,27].

Our risk groups are based on distribution percentiles and are intended exclusively to analyse properties of our score. Risk stratification for clinical practice can start with our continuous score, but thresholds should be context-dependent and appropriately reflect clinical and cost implications [33].

We did not exploit the longitudinal nature of EHRs, as our features are summarised by aggregating over time. Additional clinical indicators, such as lab tests, may improve performance further but this hypothesis remains to be tested. We highlight that despite these opportunities, granular use of EHR has well-known challenges [16], including data quality [34–36], which must be addressed.

4.1. Implications for future research and clinical practice

The only existing guidelines for PDAC screening are limited to patients with an inherited predisposition. The screening inclusion criteria are set at a fivefold increased risk relative to the general population (or a 5% absolute risk) [37]. Overall, prediction of PDAC in the general

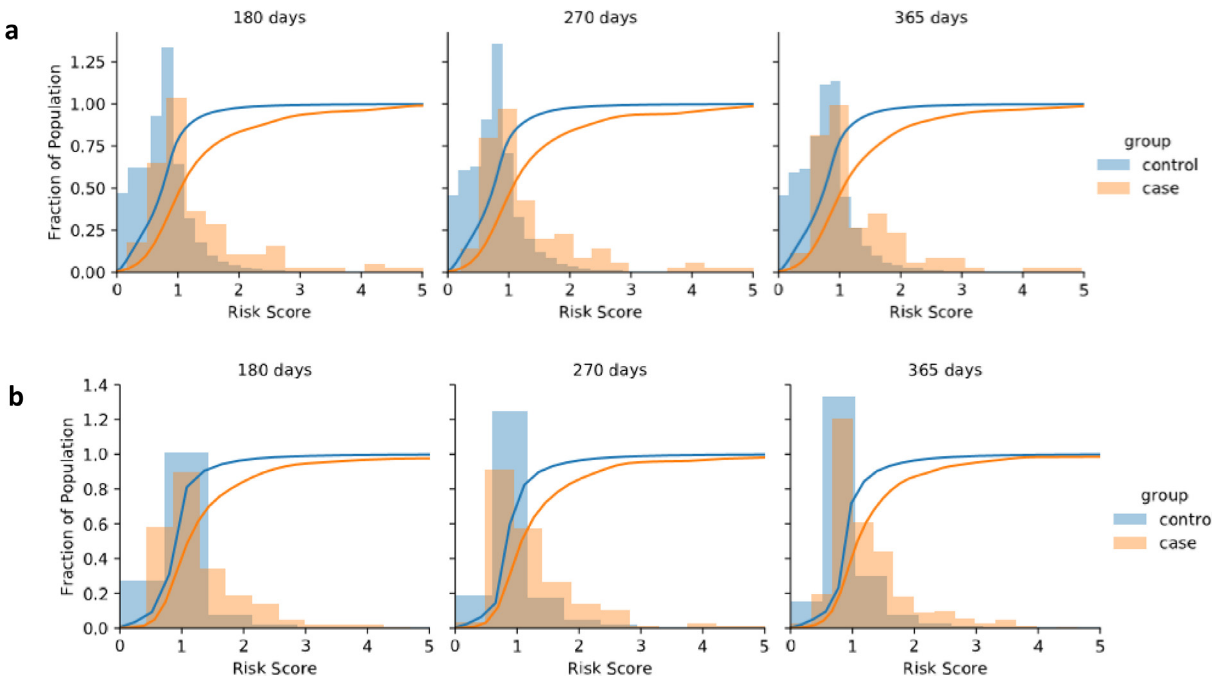

Fig. 4. Distribution of risk scores produced by the LR model. **(a)** Distribution of risk scores produced by LR model for BIDMC data (test split) and **(b)** PHC data. As expected, cancer patients' risk scores dominate those of control patients'. This difference is more pronounced in BIDMC data (test split), as LR was trained on BIDMC data. Virtually all patients' risk score are in (0, 5]. BIDMC, Beth Israel Deaconess Medical Center; LR, logistic regression; PHC, Partners HealthCare.

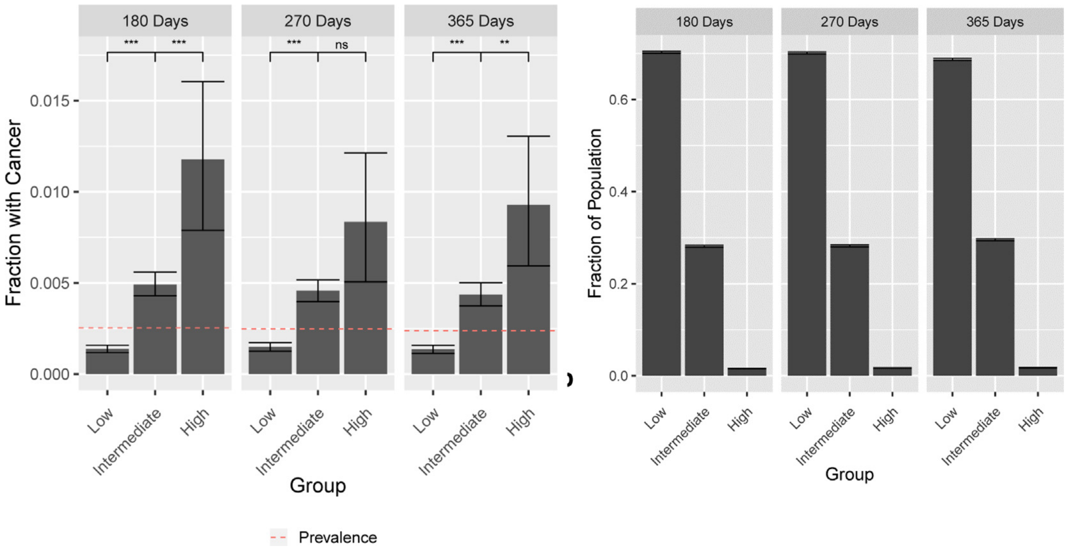

Fig. 5. The rate of cancer and fraction of population, in PHC data based on risk score stratification. (a) The rate of cancer in the risk group and (b) The fraction of population in the risk group. Risk score groups are defined using percentile-based thresholds computed from the test split of BIDMC data. Low is defined as below 75 percentile score, intermediate as between 75th and 99th percentile, and high as above 99th percentile. A total of six comparisons across the three time cut-offs (low versus intermediate, intermediate versus high, and the time cut-offs) were performed. We used Fisher's exact test (The null hypothesis is that the count of true cancer/control patients is independent of the risk group), and adjusted p-values accordingly using Bonferroni adjustment. We use '∗∗∗' to note a significance <0.001, '∗∗' <0.01, '∗' <0.05, and 'ns' when there is no statistical significance. The low/intermediate comparison is always statistically significant, and the intermediate/high is significant for 180 and 365 days. The intermediate/high comparison is not significant for 270 days after carrying out the Bonferroni adjustment. PHC, Partners HealthCare.

population, as opposed to focussing on known risk groups, is challenging and resulted in a relatively low AUC. However, we believe that PDAC risk modelling efforts should aim to provide predictive performance that matches (or outperforms) that provided by these inclusion criteria, delivering an inexpensive and systematic approach to identify patients for further screening.

We present a first step towards this goal. A two-step screening approach can be potentially utilised to identify high-risk patients who would benefit from PDAC diagnostic workup. First, this model can be implemented in clinical practice at the point of care, as a decision support system (DSS), embedded into existing hospital EHR systems. Next, the DSS would alert primary care physicians of individuals at high risk for developing PDAC and trigger physician action, such as additional biomarker or genetic testing. This approach can be prospectively validated in future work.

5. Conclusion

We propose and externally validate a model that uses routinely available diagnostic codes to identify individuals at risk for pancreatic cancer 6–12 months before diagnosis. Our risk score could potentially be used as an initial filter, preceding additional biomarkers or genetic testing, to select patients from the general population for closer surveillance.

Author contributions

Limor Appelbaum: Conceptualization, Data curation, Investigation, Methodology, Validation, Visualization, Writing - original draft; **José Pablo Cambronero**: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing - original draft, Visualization; **Jennifer P. Stevens**: Supervision, Writing - review and editing; **Steven Horng**: Writing - review and editing; **Karla Pollick**: Project administration; **George Silva**: Investigation, Software; **Sebastien Haneuse**: Methodology, Writing - review and editing; **Gail Piatkowski**: Validation; **Nordine Benhaga**: Resources; **Stacey Duey**: Investigation; **Mary Ann Stevenson**: Supervision; **Harvey Mamon**: Investigation, Writing - review and editing; **Irving D. Kaplan**: Conceptualization, Supervision, Writing - review and editing; **Martin C. Rinard**: Conceptualization, Supervision, Validation

Role of the funding source

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank the following individuals for their assistance and feedback during the development of this project: Stephen Berry, Michael Passe, Matthew Ghofrani, Laurie Bogosian, Dr. Vikas P. Sukhatme, Dr. Rebecca A. Miksad, Dr. Joseph Feuerstein, Stephanie Li, and Adam Yala.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejca.2020.10.019.

References

[1] Gordon-Dseagu Vanessa L, Devesa Susan S, Goggins Michael, et al. Pancreatic cancer incidence trends: evi- dence from the surveillance, epidemiology and end results (seer) population-based data. Int J Epidemiol 2017;47(2):427–39.

[2] N Howlader, AM Noone, M Krapcho, et al. Seer cancer statistics review, 1975–2016, national cancer institute. Bethesda, MD. https://seer.cancer.gov/csr/1975_2016/2019.

[3] Vincent Audrey, Herman Joseph, Schulick Rich, et al. Pancreatic cancer. Lancet 2011;378(9791):607–20.

[4] Paparrizos John, White Ryen W, Horvitz Eric. Screening for pancreatic adenocarcinoma using signals from web search logs: feasibility study and results. J Oncol Pract 2016;12(8):737–44.

[5] Evans Julie, Chapple Alison, Salisbury Helen, et al. "it can't be very important because it comes and goes"—patients' accounts of intermittent symptoms preceding a pancreatic cancer diagnosis: a qualitative study. BMJ Open 2014;4(2):e004215.

[6] Pannala Rahul, Leibson Cynthia L, Rabe Kari G, et al. Temporal association of changes in fasting blood glucose and body mass index with diagnosis of pancreatic cancer. Am J Gastroenterol 2009;104(9):2318.

[7] Pannala Rahul, Basu Ananda, Petersen Gloria M, et al. New-onset diabetes: a potential clue to the early diagnosis of pancreatic cancer. Lancet Oncol 2009;10(1):88–95.

[8] Hart Phil A, Kamada Pratima, Rabe Kari G, et al. Weight loss precedes cancer specific symptoms in pancreatic cancer associated diabetes mellitus. Pancreas 2011;40(5):768.

[9] Sah Raghuwansh P, Nagpal Sajan Jiv Singh, Mukhopadhyay Debabrata, et al. New insights into pancreatic cancer-induced paraneoplastic diabetes. Nat Rev Gastroenterol Hepatol 2013;10(7):423.

[10] Macášek Jaroslav, Vecka Marek, Aleš Žák, et al. Plasma fatty acid composition in patients with pancreatic cancer: correlations to clinical parameters. Nutr Canc 2012;64(7):946–55.

[11] Ritchie Shawn A, Akita Hirofumi, Takemasa Ichiro, et al. Metabolic system alterations in pancreatic cancer patient serum: potential for early detection. BMC Canc 2013;13(1):416.

[12] Lin Xianchao, Zhan Bohan, Wen Shi, et al. Metabonomic alterations from pancreatic intraepithelial neoplasia to pancreatic ductal adenocarcinoma facilitate the identification of biomarkers in serum for early diagnosis of pancreatic cancer. Mol Biosyst 2016;12(9):2883–92.

[13] Mayers Jared R, Wu Chen, Clish Clary B, et al. Elevation of circulating branched-chain amino acids is an early event in human pancreatic adenocarcinoma development. Nat Med 2014;20(10): 1193.

[14] Canto Marcia Irene, Almario Jose Alejandro, Schulick Richard D, et al. Risk of neoplastic progression in individuals at high risk for pancreatic cancer undergoing long-term surveillance. Gastroenterology 2018;155(3):740–51.

[15] Henry J, Pylypchuk Yuriy, Searcy Talisha, et al. Adoption of electronic health record systems among us non-federal acute care hospitals: 2008-2015. ONC Data Brief 2016;35:1–9.

[16] Goldstein Benjamin A, Navar Ann Marie, Pencina Michael J, et al. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inf Assoc 2017;24(1):198–208.

[17] Sharma Ayush, Kandlakunta Harika, Nagpal Sajan Jiv Singh, et al. Model to determine risk of pancreatic cancer in patients with new-onset diabetes. Gastroenterology 2018; 155(3):730–9.

[18] Ben Boursi, Finkelman Brian, Giantonio Bruce J, et al. A clinical prediction model to assess risk for pancreatic cancer among patients with new-onset diabetes. Gastroenterology 2017;152(4): 840–50.

[19] Hsieh Meng Hsuen, Sun Li-Min, Lin Cheng-Li, et al. Development of a prediction model for pancreatic cancer in patients with type 2 diabetes using logistic regression and artificial neural network models. Canc Manag Res 2018;10:6317.

[20] Baecker Aileen, Kim Sungjin, Risch Harvey A, et al. Do changes in health reveal the possibility of undiagnosed pancreatic cancer? development of a risk-prediction model based on healthcare claims data. PloS One 2019;14(6):e0218580.

[21] Muhammad Wazir, Hart Gregory R, Bradley Nartowt, et al. Pancreatic cancer prediction through an artificial neural network. Front Artif Intell 2019;2:2.

[22] Klein Alison P, Lindström Sara, Mendelsohn Julie B, et al. An absolute risk model to identify individuals at elevated risk for pancreatic cancer in the general population. PloS One 2013;8(9): e72311.

[23] James A Hanley, McNeil Barbara J. The meaning and use of the area under a receiver operating characteristic (roc) curve. Radiology 1982;143(1):29–36.

[24] Yu Jun, Blackford Amanda L, dal Molin Marco, et al. Time to progression of pancreatic ductal adenocarcinoma from low-to-high tumour stages. Gut 2015;64:1783–9.

[25] Barak-Corren Yuval, Castro Victor M, Solomon Javitt, et al. Predicting suicidal behavior from longitudinal electronic health records. Am J Psychiatr 2016;174(2):154–62.

[26] Razavian Narges, Blecker Saul, Schmidt Ann Marie, et al. Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. Big Data 2015;3(4):277–87.

[27] Nakatochi Masahiro, Lin Yingsong, Ito Hidemi, et al. Prediction model for pancreatic cancer risk in the general Japanese population. PloS One 2018;13(9).

[28] Gilbert Thomas, Neuburger Jenny, Kraindler Joshua, et al. Development and validation of a hospital frailty risk score focusing on older people in acute care settings using electronic hospital records: an observational study. Lancet 2018;391(10132): 1775–82.

[29] Rajkomar Alvin, Oren Eyal, Chen Kai, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018;1(1):18.

[30] Gensheimer Michael F, Solomon Henry A, Wood Douglas J, et al. Automated survival prediction in metastatic cancer patients using high-dimensional electronic medical record data. JNCI (J Natl Cancer Inst): Journal of the National Cancer Institute 2019; 111(6):568–74.

[31] Abelson Sagi, Collord Grace, Ng Stanley WK, et al. Prediction of acute myeloid leukaemia risk in healthy individuals. Nature 2018; 559(7714):400−4.

[32] Huang BZ, Stram DO, Le Marchand L, et al. Interethnic differences in pancreatic cancer incidence and risk factors: The Multiethnic Cohort. Cancer Med 2019;7(8):3592.

[33] Wynants Laure, van Smeden Maarten, McLernon David J, et al. Three myths about risk thresholds for prediction models. BMC Med 2019;17(1):192.

[34] Farid Mina, Roatis Alexandra, Ihab F Ilyas, et al. Clams: bringing quality to data lakes. In: Proceedings of the 2016 International Conference on Management of Data. ACM; 2016. p. 2089−92.

[35] Ouzzani Mourad, Tang Nan, Castro Fernandez Raul. Data civilizer: end-to-end support for data discovery, integration, and cleaning. In: Making Databases Work: the Pragmatic Wisdom of Michael Stonebraker. Association for Computing Machinery; 2018. p. 291−300.

[36] Gray Weiskopf Nicole, Weng Chunhua. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inf Assoc 2013;20(1):144−51.

[37] Canto Marcia Irene, Harinck Femme, Hruban Ralph H, et al. International cancer of the pancreas screening (caps) consortium summit on the management of patients with increased risk for familial pancreatic cancer. Gut 2013;62(3):339−47.