Prediction of pancreatic cancer risk in patients with new-onset diabetes using a machine learning approach based on routine biochemical parameters

Cichosz, Simon Lebech; Jensen, Morten Hasselstrøm; Hejlesen, Ole; Henriksen, Stine Dam; Drewes, Asbjørn Mohr; Olesen, Søren Schou

Published in:
Computer Methods and Programs in Biomedicine

DOI (link to publication from Publisher):
10.1016/j.cmpb.2023.107965

Publication date:
2024

Document Version
Publisher's PDF, also known as Version of record

Computer Methods and Programs in Biomedicine 244 (2024) 107965

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Prediction of pancreatic cancer risk in patients with new-onset diabetes using a machine learning approach based on routine biochemical parameters

Simon Lebech Cichosz [a,*], Morten Hasselstrøm Jensen [a], Ole Hejlesen [a], Stine Dam Henriksen [b,c], Asbjørn Mohr Drewes [c,d], Søren Schou Olesen [c,d]

[a] Department of Health Science and Technology, Aalborg University, Aalborg, Denmark
[b] Department of Gastrointestinal Surgery and Clinical Cancer Research Center, Aalborg University Hospital, Aalborg, Denmark
[c] Department of Clinical Medicine, Aalborg University Hospital, Aalborg, Denmark
[d] Centre for Pancreatic Diseases and Mech-Sense, Department of Gastroenterology and Hepatology, Aalborg University Hospital, Aalborg, Denmark

ARTICLE INFO

Keywords:
Pancreatic cancer
Pancreatic ductal adenocarcinoma
New-onset diabetes
Machine learning
Risk prediction
Type 2 diabetes

ABSTRACT

Objective: To develop a machine-learning model that can predict the risk of pancreatic ductal adenocarcinoma (PDAC) in people with new-onset diabetes (NOD).
Methods: From a population-based sample of individuals with NOD aged >50 years, patients with pancreatic cancer-related diabetes (PCRD), defined as NOD followed by a PDAC diagnosis within 3 years, were included (n = 716). These PCRD patients were randomly matched in a 1:1 ratio with individuals having NOD. Data from Danish national health registries were used to develop a random forest model to distinguish PCRD from Type 2 diabetes. The model was based on age, gender, and parameters derived from feature engineering on trajectories of routine biochemical variables. Model performance was evaluated using receiver operating characteristic curves (ROC) and relative risk scores.
Results: The most discriminative model included 20 features and achieved a ROC-AUC of 0.78 (CI:0.75–0.83). Compared to the general NOD population, the relative risk for PCRD was 20-fold increase for the 1 % of patients predicted by the model to have the highest cancer risk (3-year cancer risk of 12 % and sensitivity of 20 %). Age was the most discriminative single feature, followed by the rate of change in haemoglobin A1c and the latest plasma triglyceride level. When the prediction model was restricted to patients with PDAC diagnosed six months after diabetes diagnosis, the ROC-AUC was 0.74 (CI:0.69–0.79).
Conclusion: In a population-based setting, a machine-learning model utilising information on age, sex and trajectories of routine biochemical variables demonstrated good discriminative ability between PCRD and Type 2 diabetes.

1. Introduction

Pancreatic ductal adenocarcinoma (PDAC) is a leading cause of cancer-related death, with a 5-year survival rate of approximately 10 % [1]. Most patients are diagnosed at an advanced stage, which limits treatment possibilities to palliative care [1]. However, patients with resectable PDAC have better treatment options and prognoses, and the 5-year survival rate increases to 39 % when PDAC is diagnosed at stage 1 or 2 [2].

Population-based screening for PDAC is not feasible due to its low incidence [3]. Therefore, attention has been directed towards subgroups of individuals with a higher-than-average risk of PDAC who may benefit from surveillance [4]. One such subgroup is people over 50 years old with new-onset diabetes (NOD), who have a 6–8-fold increased risk of PDAC compared to the background population [5]. However, the absolute incidence of PDAC in this subgroup is still too low for direct

Abbreviations: ATC, Anatomical therapeutic chemical; AUC, Area under curve; CI, Confidence interval; ENDPAC, Enriching new onset diabetes for pancreatic cancer; HbA1c, Glycated haemoglobin; NOD, New-onset diabetes; NPV, Negative predictive values; PPV, positive predictive value; RaoC, rate of change; RF, random forest; ROC, receiver operating characteristic curves; RR, relative risk; PCRD, pancreatic cancer-related diabetes.

* Corresponding author.
E-mail address: simcich@hst.aau.dk (S.L. Cichosz).

https://doi.org/10.1016/j.cmpb.2023.107965
Received 16 October 2023; Received in revised form 16 November 2023; Accepted 30 November 2023
Available online 2 December 2023

implementation of surveillance. Therefore, models that identify people with high PDAC risk have been developed [4]. The most widely employed model is the Enriching New Onset Diabetes for Pancreatic Cancer (ENDPAC) model, which uses information on age, changes in fasting plasma glucose and body weight during the year preceding NOD diagnosis [6,7]. Although this model has acceptable discriminative performance, the performance of this and other currently available models may be limited by the underlying mathematical algorithms, as they are typically developed using classical statistical models focused on linear associations between parameters [6]. Such models may not capture the complex patterns and interactions of predictor variables relevant for the prediction of complex biological phenomena such as PDAC.

Machine learning is a branch of artificial intelligence that can learn from data without explicit rules or equations. The method can handle high-dimensional and heterogeneous data, discover nonlinear and interactive effects, and adapt to changing environments. Machine learning, therefore, has the potential to improve the performance and generalizability of clinical prediction models by exploiting the rich information available in modern health data sources [8,9]. In recent years, there has been growing interest in the application of such models for the early detection of PDAC [6,10–12]. The models have incorporated a range of parameters, including patient-reported symptoms (e.g., weight and pain characteristics), biochemical variables, and disease codes. However, the reliance on parameters that reflect patients' subjective experiences, such as pain, or healthcare provider behaviour, such as the use of diagnostic codes, challenge the broader clinical implementation of these models beyond their original developmental settings [6,10,12]. A notable example of this was evident in a recent study based on trajectories of diagnosis codes where the model required recalibration to achieve acceptable performance in an external validation cohort [10]. These findings may be explained by the limitations associated with subjective parameters and healthcare provider-dependent information, which can hinder the generalizability and reliability of models across different clinical settings. To address this limitation, one solution may be to shift the focus towards entirely objective parameters that are less influenced by human behaviour and closely associated with the underlying biology of PDAC. For example, restricting model parameters to plasma haemoglobin A1c (HbA1c), plasma triglyceride levels and other biochemical variables could provide reliable indicators of metabolic and other disease-related changes associated with the development of PDAC [13–15].

In this study, we hypothesise that profiles of routine biochemical variables and information on age and sex can be used to create a prediction model for the early detection of PDAC in individuals with NOD. The aims of this study were to (i) develop a machine-learning model built on trajectories of biochemical parameters that can determine the risk of PDAC in people with NOD and (ii) explore the diagnostic performance of this model in two NOD populations with different PDAC incidence rates.

2. Methods

2.1. Study design and data sources

This was a retrospective nationwide cohort study in Denmark from 1998 to 2018 containing 8.1 million Danish citizens. Data from the Danish National Patient Registry, Danish National Prescription Registry, and Civil Registration System were used [16–18]. The registries were linked using unique identification numbers assigned to Danish residents [18]. Diagnoses of NOD and PDAC were obtained from the Danish National Patient Registry, which includes outpatient and inpatient hospital contacts [19]. The registry has been based on ICD-10 codes since 1994. Medication prescriptions were extracted from the Danish National Prescription Registry, which covers all prescription medicines since 1996 and includes dispensing dates and product information based on ATC codes [17]. Laboratory data were extracted from the National LABKA database [20], which collects data from Denmark's largest clinical biochemistry and immunology laboratories. The study followed the principles of the TRIPOD statement for transparent reporting of prediction models [21,22].

2.2. Study cohort

People with diabetes were identified based on either an ICD-10 diagnosis code of diabetes (E10–14.x, G63.2, H28.0, H36.0, M14.2, O24.x or R73.x) or an ATC code of glucose-lowering medication (A10.x.) using a previously published algorithm [23,24]. This algorithm identifies people with incident diabetes in both the primary care setting (based on the prescription of glucose-lowering therapies) and hospital-based settings (based on ICD-10 codes and prescription of glucose-lowering therapies). The diabetes onset date was defined as the first occurrence of an ICD-10 or an ATC code.

Patients diagnosed with diabetes before the 1st of January 1998 or under age 50 at NOD diagnosis were excluded. Thus, the final study cohort comprised of people with NOD diagnosed at age 50 or older during the study period.

2.3. Classification of diabetes subtypes

Among all people diagnosed with NOD, patients with PDAC were identified (ICD-10 codes: C25.x or Z850F). People diagnosed with PDAC before or three years after NOD diagnosis were excluded, and the remaining group was defined as pancreatic cancer-related diabetes (PCRD). The threshold of three years for PCRD diagnosis was chosen based on previous pathophysiological investigations showing that glycaemic changes in the context of PDAC can be detected for up to three years before PDAC diagnosis [13–15]. We excluded patients without information on HbA1c at the time of NOD diagnosis. In the group of people with NOD without PDAC, we classified people as having Type 1 diabetes if they had received at least one ICD-10 code of Type 1 diabetes mellitus (E10.x) and no ATC code of "blood glucose-lowering drugs except for insulins" (A10B.x). People with Type 1 diabetes were subsequently excluded, and the remaining people were pragmatically classified as Type 2 diabetes. Finally, patients with PCRD and information on HbA1c at NOD diagnosis were matched 1:1 with a random subsample of people with NOD (without subsequent PDAC diagnosis) who also had information on HbA1c (i.e., random under-sampling). This final cohort was used for model development and validation.

2.4. Predictor variables

We obtained predictor variables from an extensive set of routine biochemical measurements. These included parameters related to glucose and lipid metabolism, liver function tests, nutritional markers, and markers of systemic inflammation. A full list of the parameters and rationale for their inclusion as predictor variables are provided in supplementary Table S1. We included data from up to 3 years prior to NOD diagnosis for all biochemical parameters. In addition, information on age and sex were included as predictor variables.

2.5. Feature engineering

Feature engineering is a process in machine learning where new features (i.e., variables, predictors, or inputs) are created from raw data by transforming, aggregating, or extracting information from existing variables. The goal is to improve the performance of machine learning algorithms by creating new features that better represent the underlying relationships in the data and increase prediction accuracy. [25]

To extract information from underlying patterns in the biochemical parameters, we calculated a battery of 83 features. For each parameter, we calculated the mean, standard deviation, maximum value, minimum value, number of observations, last value, rate of change (RaoC) and

RaoC-condensed. RaoC was defined by the change from the oldest value to the value closest to NOD diagnosis divided by the number of days between the measured values. Likewise, the RaoC-condensed was calculated using the two values closest to NOD diagnosis.

2.6. Missing data

The data analysed in this study originated from clinical practice with the implication that most patients did not have all biochemical parameters measured. The intended usage of the model developed is in a real-world clinical situation, which the data reflect. Hence, we did not exclude cases based on missing data. Additionally, we did not impute missing values, as the measurement (or lack of measurement) of a certain biochemical parameter could hold discrimination information between the classes. Missing data were handled in the modelling using a nonlinear approach and by filling missing values with a 'dummy' value. This allowed the model to treat missing data as independent information.

2.7. Model building procedures

Given the binary outcome (PCRD vs Type 2 diabetes) and the need for a nonlinear approach, we developed our prediction model using random forest (RF) classification. RF is a versatile and robust machine learning method [26] that uses a collection of decision trees to make predictions. RF has several benefits, including low variance, high accuracy, relative robustness to overfitting, ability to handle nonlinear relationships, and interpretability using feature importance. All analyses were performed using Python (v3) and the Scikit-learn package (v0.23.2) for machine learning utilities. The modelling approach is illustrated in supplementary figure S1.

The data were randomly split into 70 % for training and validation and 30 % for testing. This method enhances external validity [27].

The initial 83 extracted features were reduced to a subset of features before model training. Feature selection was conducted using the training data and calculation of feature importance. The model was trained with a subset of the 20 features with the highest feature importance. The reduction procedure was used to simplify the model and to eliminate redundant or indiscriminative information from potential features, leading to improved accuracy [28].

Hyperparameters (i.e., the number of trees, minimum cases per leaf, maximum depth of trees, and the number of features to consider per split) were determined using 3-fold cross-validation. This procedure was implemented to minimise overfitting of the model on the training data [29]. The final model was then tested on the test dataset. We did not recalibrate the model after the training procedure.

2.8. Model evaluation metrics

The area under the receiver operating characteristic curve (ROC-AUC) was used to evaluate the discrimination of the model [30]. Confidence intervals (CIs) for ROC-AUC were estimated using bootstrap replicas ($n = 1000$). To assess the clinical implications of using the model, we calculated the sensitivity, specificity, positive predictive value (PPV), and negative predictive values (NPV) for different cut-off points and assessed relative risk (RR) curves. To address the variations in PDAC incidence among different populations of NOD, we evaluated performance metrics (NPV, PPV, and RR) using two distinct populations estimates with varying 3-year population-based cumulative incidence rates of PDAC among individuals with NOD: Denmark (0.6 %) (current study) and the United States (1.0 %) [5].

2.9. Model interpretability

The random forest classifier uses a subset of "strong variables" for classification, resulting in better performance on high dimensional data [31]. The Gini importance is a measure of feature importance that reflects the outcome of this implicit feature selection and can be used as a general indicator of the individual features' importance in the classification. The feature importance score is a by-product of the random forest training and provides a relative ranking of the features [28]. To interpret the individual features' impact on the prediction model, we used the average feature importance over all trees in the model.

2.10. Sensitivity analysis

In a sensitivity analysis, we restricted the PCRD subgroup to patients diagnosed with PCAD between 6 months and 3 years after being diagnosed with NOD. The objective of this analysis was to assess the model's performance by excluding patients who were immediately diagnosed with PDAC after being diagnosed with NOD.

3. Results

We enrolled 716 patients with PCRD and 716 individuals with pragmatic-defined Type 2 diabetes in the cohorts for model development. A study flowchart illustrating the selection process is presented in supplementary figure S2. Demographic and HbA1c at NOD diagnosis for the two subgroups are shown in Table 1. The mean age at NOD diagnosis was 71.0 years for patients with PCRD vs. 66.7 years for people with Type 2 diabetes. Among patients with PCRD, 48.3 % were women vs. 39.7 % in the Type 2 diabetes group. The HbA1c at baseline was 57 (IQR, 50;77) mmol/mol in the PCRD group vs. 52 (IQR, 48;63) mmol/mol in the Type 2 diabetes group. The median follow-up time from NOD diagnosis to PDAC diagnosis in the PCRD subgroup was 0.45 (IQR, 0.14–1.2) years. The population-based cumulative 3-year incidence of PDAC among people older than 50 years at NOD diagnosis was 0.59 % (95 % CI; 0.57 to 0.62).

3.1. Model performance

The RF model had a ROC-AUC of 0.78 (CI95; 0.75–0.83) on the test dataset (Fig. 1). Diagnostic performance characteristics for the model calibrated for different sensitivity levels are reported in Table 2. In the population-based setting, the NPVs ranged from 99.5 % to 100 % across all sensitivity levels. This reflects the relatively low incidence of PDAC even among individuals with NOD (3-year cumulative PDAC incidence 0.6 %). The PPVs were significantly influenced by the selected sensitivity levels. For example, a sensitivity level of 20 % resulted in a PPV of 8.4 % and a NPV of 99.5 %, while a sensitivity of 50 % resulted in a PPV of 2.3 % and a NPV of 99.7 %.

In Fig. 2, the estimated performance of a surveillance program in population-based settings using the model is presented, with the RR illustrated in relation to the number of patients at-risk in a simulation of 1 million individuals with NOD. For example, the RR for PDAC was 20-fold increased for the 1 % of individuals predicted to have the highest cancer risk compared to the general NOD population in Denmark. This corresponds to a cumulative 3-year cancer risk of 12 %. For the 10 % of individuals predicted to be at highest cancer risk, a 5-fold increase in RR was observed compared to the general NOD population. This corresponds to a cumulative 3-year cancer risk of 3 %. We also simulated RR

Table 1
Baseline characteristics of patients with pancreatic cancer-related diabetes (PCRD) and Type 2 diabetes (T2D).

	PCRD	T2D
Number of subjects	716	716
Age, mean (SD)	71 (9.0)	66.7 (10.1)
Sex, Female, n (%)	346 (48,3)	284 (39.7)
Sex, Male, n (%)	370 (51.7)	432 (60.3)
HbA1c at baseline, mean (IQR)	57 (50;77)	52 (48;63)

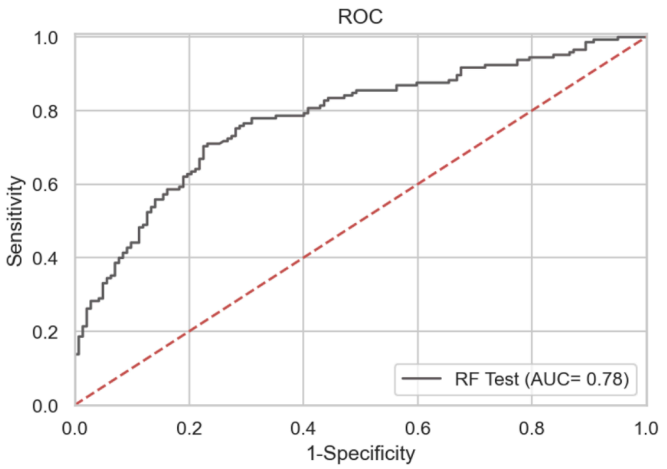

Fig. 1. Receiver operating characteristic (ROC) for the Random Forest (RF) model applied on the test dataset.

Table 2
Diagnostic performance characteristics at different sensitivity levels.

	Model performance					
			PDAC risk 0.6 % *		PDAC risk 1.0 % #	
	Sensitivity	Specificity	PPV	NPV	PPV	NPV
Sensitivity 10	13.8	100	100	99.5	100	99.1
Sensitivity 20	21.4	98.6	8.4	99.5	13.3	99.2
Sensitivity 30	33.1	95.1	3.9	99.6	6.4	99.3
Sensitivity 40	40.0	92.3	3.0	99.6	4.9	99.3
Sensitivity 50	52.4	87.3	2.4	99.7	4.0	99.5
Sensitivity 60	59,3	80.9	1.8	99.7	3.1	99.5
Sensitivity 70	70.3	77.5	1.8	99.8	3.1	99.6
Sensitivity 80	80.1	59.1	1.2	99.8	1.9	99.7
Sensitivity 90	89.6	32.3	0.8	99.8	1.4	99.7
Sensitivity 100	100	4.9	0.6	100	1.1	100

The sensitivity calibration is based on the ROC cut-off closest to a given sensitivity between 10 and 100 %. Population-based cumulative 3-year incidence estimates of PDAC among individuals with NOD are used to calculate PPV and NPV:.

* 3-year cumulative PDAC risk 0.6 % (Danish general population).

3-year cumulative PDAC risk 1.0 % (Unites States general population).

estimates for a NOD population with a 3-year cumulative risk of PDAC of 1.0 %, which corresponds to the population-based risk reported in the United States (Fig. 2).

3.2. Feature importance and model explainability

Feature importance analysis showed that age was the most important single feature to discriminate PCRD from Type 2 diabetes, contributing to approximately 20 % of the discriminative ability (Fig. 3). Also, the rate of change in HbA1c was an important discriminator, contributing approximately 11 %. In addition, information derived from triglycerides, alanine aminotransferase, and alkaline phosphatase also held discriminative information.

Violin plots illustrating the feature distributions with median and interquartile range for selected features are presented in Fig. 4. The rate of change in HbA1c was, on average, increased 3-fold in patients with PCRD vs. Type 2 diabetes. Patients with PCRD were also characterised by lower triglyceride levels but higher alkaline phosphatase levels compared to Type 2 diabetes. However, the distributions indicated substantial overlap between parameters, which makes single parameters unable to differentiate the two groups.

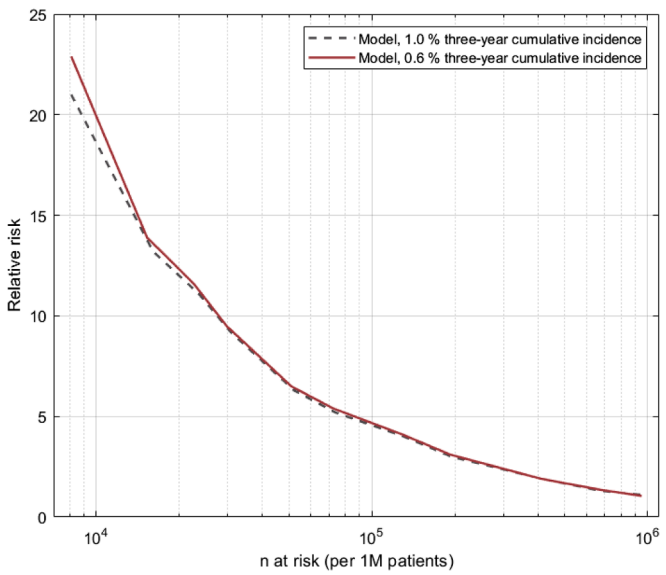

Fig. 2. Estimated performance of a surveillance program for high-risk NOD patients. Estimated relative risk (RR) for the n (horizontal axis) high-risk patients is based on evaluating the accuracy of prediction in NOD patients (>50 years of age) with 3-year cumulative PDAC incidence of either 0.6 % (Danish general population) or 1 % (United States general population).

3.3. Sensitivity analysis

In the sensitivity analysis excluding patients immediately diagnosed with PDAC after NOD diagnosis (<6 months), the ROC-AUC was 0.74 (95 % CI; 0.69–0.79).

4. Discussion

In a population-based cohort of patients with NOD ≥50 years of age, we developed and internally validated a novel machine-learning model to discriminate PCRD from Type 2 diabetes. The model was based on age, sex and trajectories of routine biochemical variables, for which information would be available at the time of diabetes diagnosis. The final model was shown to have good discrimination between Type 2 diabetes and PCRD (AUC-ROC 0.78). The most important discriminators were older age and rapid increase in HbA1c. In addition, biochemical parameters associated with changes in lipid metabolism and bile duct obstruction held discriminative information.

4.1. Model performance

Our model had a diagnostic performance (AUC-ROC 0.78) comparable to that observed in most previous models (AUC-ROC 0.71 to 0.87) designed for PDAC risk determination in people with NOD [6]. For example, the ENDPAC model, which is based on age and one-year changes in body weight and fasting plasma glucose levels prior to NOD onset, had an AUC-ROC of 0.87 in the derivation cohort [7]. In two subsequent independent validation studies, the AUC-ROCs were 0.75 and 0.69, respectively [32,33]. Another model derived from a primary care health database in the United Kingdom was based on 11 parameters and had an AUC-ROC of 0.81 [15]. The specificity of that model was 93 %, at a sensitivity of 44 %, which is comparable to the performance of our model.

We utilized a machine learning-based modelling approach, which, to our knowledge, has only been applied in one prior study concerning the determination of PDAC risk in individuals with NOD [11]. In that study, the most effective model relied on age, changes in weight, and HbA1c (including the rate of change in HbA1c), demonstrating a C-index of 0.82 in distinguishing PCRD from other diabetes subtypes. Our findings

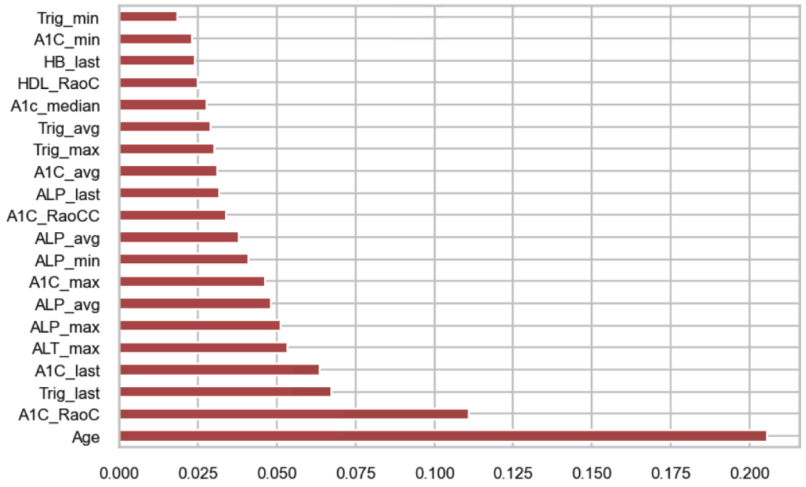

Fig. 3. Feature importance scores for the 20 most important features. Feature importance score shows the relative importance of each feature in the Random Forest (RF) model. Abbreviations: HbA1c (A1C), triglycerides (Trig), high-density lipoproteins (HDL), haemoglobin (HB), alkaline phosphatase (ALP), alanine transaminase (ALT), last measurement (last), minimum (min), maximum (max), average (avg), rate of change (RaoC).

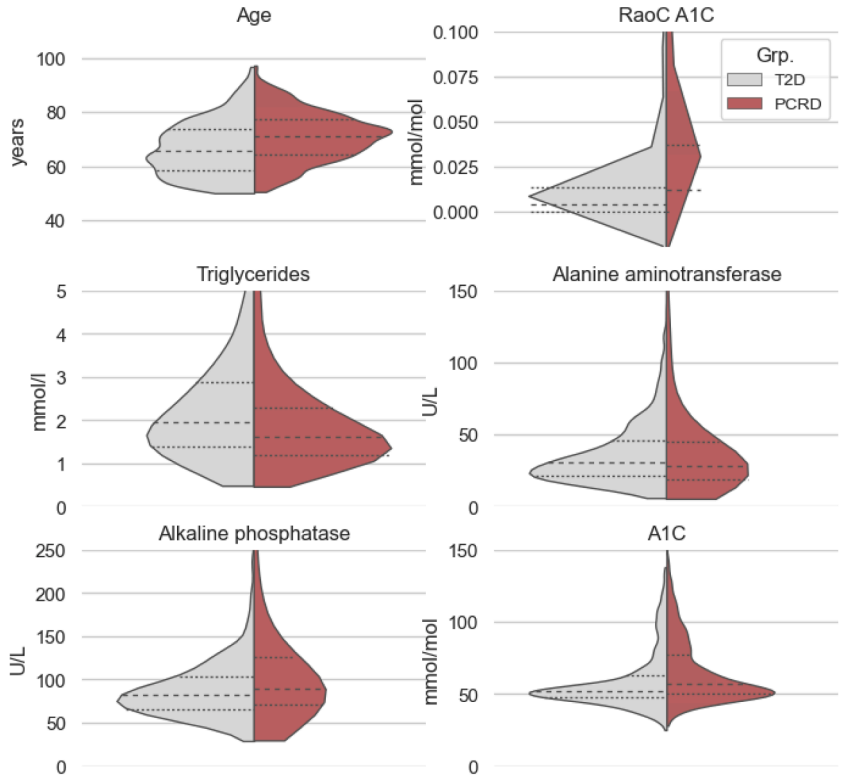

Fig. 4. Violin plot for selected demographic and biochemical features shown for the PCRD and T2D subgroups.

align with this, as age and the rate of change in HbA1c emerged as the most significant discriminators. However, the previous study also incorporated changes in weight and other symptom features, differing from our study's model, which solely relied on routine biochemical data, age, and sex. Standardized serial weight assessments are often not available in a clinical setting, and thus, information on weight may be flawed by recall bias and influenced by a lack of standardization across equipment and settings. This emphasizes the need for a model which do not include this parameter in the risk assessment of NOD patients at the time of diagnosis. Therefore, we consider our model to be more practical and robust for implementation in a clinical setting than previous models, which included information on weight changes, but this will need to be further investigated.

4.2. *Feature importance and model explainability*

A unique aspect of our study was the implementation of feature engineering on routine biochemical variables. This allowed us to maximise the information contained in the biochemical variables. In keeping with this, six of the 20 most discriminative features were derived from HbA1c trajectories underlining the relevance of feature engineering in this context. Our observations align with previous studies that demonstrated exponentially increasing plasma glucose levels in PDAC patients prior to cancer diagnosis [13–15]. The underlying

mechanisms responsible for the glycaemic changes are multifactorial and extend beyond tumour-induced damage to pancreatic islet cells. For instance, pancreatic cancer induces insulin resistance and beta-cell dysfunction, which can be reversed upon tumour resection. This suggests the involvement of paraneoplastic mediators [34].

Another distinctive biochemical difference between PCRD and Type 2 diabetes was reflected in plasma triglyceride trajectories prior to the onset of diabetes, with four out of the 20 most discriminative features based on triglyceride information. This observation also aligns with previous findings and may be linked to the process of exosome-induced browning of subcutaneous adipose tissue during the development and progression of PDAC [14,15,35]. Additionally, we observed variations between PCRD and Type 2 diabetes in features derived from liver function and cholestatic parameters. These may be linked to obstruction of the biliary tract or the presence of liver metastases caused by the tumour.

4.3. Strengths and limitations

A strength of our study is the high degree of model explainability linking predictor variables to biologically plausible mechanisms. Another strength is the use of feature engineering to extract the most possible information from biochemical variables and trajectories of biochemical changes, which offers a robust set of model features that most likely can be implemented across countries and clinical settings.

We intentionally focused our model on well-established and explainable biochemical parameters for PDAC prediction, including HbA1c. This has shown the most discriminative performance in previous models using conventional statistical methods [7,15]. Consequently, we restricted our PCRD cohort to individuals with information on HbA1c at NOD diagnosis. However, this may compromise the generalizability of our model. Therefore, the diagnostic performance of models without this constraint should be explored in future studies. Indeed, in a pilot study, lab test administration per patient (i.e., frequency of lab test administration) was found to be the most discriminative feature [36]. Another important limitation is the lack of an external validation cohort.

Noteworthy, difference in sex distributions between PCRD and Type 2 Diabetes were observed in our study. Variations in the prevalence of Type 2 Diabetes between genders are well-established, especially in some ethnic groups [37]. This difference is also observed in the overall population of NOD>50 years of age (n = 343,938) from the Danish national health registries [38]. Despite this sex difference, sex was not included in the final model, emphasizing that sex only had a moderate discriminative performance that was outweighed by age and biochemical-derived parameters.

The estimates of the cumulative 3-year incidence of PDAC among people with NOD used to assess the model's performance in population-based settings (i.e., Denmark and the United States) must be taken with some caution. Hence, the population-based studies providing these estimates may be influenced by differences in the definition of diabetes, differences in the timing of diagnosis, and ethnic/racial composition of the population. However, until more definite studies become available, we consider these register-based estimates to represent reasonable approximations [39]. Finally, a limitation of this study is the definition of PCRD based on ICD-10 codes from the Danish national health registries, without histological and/or clinical case verification.

5. Conclusions

We developed and internally validated a novel machine-learning model to discriminate PCRD from Type 2 diabetes. The model was based on predictors derived from feature engineering on routine biochemical parameters. If successfully externally validated, we envision the model can be implemented in various clinical settings to identify people with NOD at high risk of PDAC who are amendable to surveillance.

6. Disclosures

Author Jensen is employed and holds shares in Novo Nordisk and has received consultant fees from Abbott Laboratories A/S.

Funding

The study was not funded.

CRediT authorship contribution statement

Simon Lebech Cichosz: Data curation, Writing – original draft. **Morten Hasselstrøm Jensen:** Writing – review & editing. **Ole Hejlesen:** Writing – review & editing. **Stine Dam Henriksen:** Writing – review & editing. **Asbjørn Mohr Drewes:** Data curation, Writing – original draft. **Søren Schou Olesen:** Data curation, Writing – original draft.

Declaration of Competing Interest

No potential conflicts of interest relevant to this article were reported

Data availability

Approval from an ethical committee is not required for epidemiological studies conducted in Denmark using national health registries. The anonymized data (project identifier 708466) has been made accessible by Statistics Denmark. Only research institutions authorized by The Danish Health Data Authority are eligible to apply for this data.

Acknowledgements

None

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2023.107965.

References

[1] J.D. Mizrahi, R. Surana, J.W. Valle, R.T. Shroff, Pancreatic cancer, Lancet 395 (10242) (2020) 2008–2020 [Internet][cited 2023 May 22]Available from, pubmed.ncbi.nlm.nih.gov/32593337/.
[2] R.L. Siegel, K.D. Miller, H.E. Fuchs, A. Jemal, Cancer Statistics, 2021, CA Cancer J. Clin. 71 (1) (2021) 7–33 [Internet][cited 2023 Jun 9]Available from, pubmed.ncbi.nlm.nih.gov/33433946/.
[3] R. Pannala, A. Basu, G.M. Petersen, S.T. Chari, New-onset diabetes: a potential clue to the early diagnosis of pancreatic cancer, Lancet Oncol. 10 (1) (2009) 88–95 [Internet][cited 2023 May 22]Available from, pubmed.ncbi.nlm.nih.gov/19111249/.
[4] A.D. Singhi, E.J. Koay, S.T. Chari, A. Maitra, Early detection of pancreatic cancer: opportunities and challenges, Gastroenterology 156 (7) (2019) 2024–2040 [Internet][cited 2023 May 22]Available from, pubmed.ncbi.nlm.nih.gov/30721664/.
[5] S.T. Chari, C.L. Leibson, K.G. Rabe, J. Ransom, M. De Andrade, G.M. Petersen, Probability of pancreatic cancer following diabetes: a population-based study, Gastroenterology 129 (2) (2005) 504–511 [Internet][cited 2023 May 22]Available from, pubmed.ncbi.nlm.nih.gov/16083707/.
[6] R. Santos, H.G. Coleman, V. Cairnduff, A.T. Kunzmann, Clinical prediction models for pancreatic cancer in general and at-risk populations: a systematic review, Am. J. Gastroenterol. 118 (1) (2023) 26–40 [Internet][cited 2023 May 22]Available from, pubmed.ncbi.nlm.nih.gov/36148840/.
[7] A. Sharma, H. Kandlakunta, S.J.S. Nagpal, et al., Model to determine risk of pancreatic cancer in patients with new-onset diabetes, Gastroenterology 155 (3) (2018) 730–739 [Internet][cited 2023 May 22]e3. Available from, pubmed.ncbi.nlm.nih.gov/29775599/.
[8] J.A.M. Sidey-Gibbons, C.J. Sidey-Gibbons, Machine learning in medicine: a practical introduction, BMC Med. Res. Methodol. 19 (1) (2019) 1–18 [Internet][cited 2023 May 22]Available from, link.springer.com/articles/10.1186/s12874-019-0681-4.
[9] D. Bzdok, N. Altman, M. Krzywinski, Statistics versus machine learning, Nat. Methods 15 (4) (2018).

[10] D. Placido, B. Yuan, J.X. Hjaltelin, et al., A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories, Nat. Med. (2023) [Internet][cited 2023 May 22];Available from, pubmed.ncbi.nlm.nih.gov/37156936/.
[11] W. Chen, R.K. Butler, E. Lustigova, et al., Risk prediction of pancreatic cancer in patients with recent-onset hyperglycemia: a machine-learning approach, J. Clin. Gastroenterol. 57 (1) (2023) 103–110 [Internet][cited 2023 May 26]Available from, pubmed.ncbi.nlm.nih.gov/35470312/.
[12] W. Chen, Y. Zhou, F. Xie, et al., Derivation and external validation of machine learning-based model for detection of pancreatic cancer, Am. J. Gastroenterol. 118 (1) (2023) 157–167 [Internet][cited 2023 May 26]Available from, pubmed.ncbi.nlm.nih.gov/36227806/.
[13] A. Sharma, T.C. Smyrk, M.J. Levy, M.A. Topazian, S.T. Chari, Fasting blood glucose levels provide estimate of duration and progression of pancreatic cancer before diagnosis, Gastroenterology 155 (2) (2018) 490–500 [Internet][cited 2023 May 22]e2. Available from, pubmed.ncbi.nlm.nih.gov/29723506/.
[14] R.P. Sah, A. Sharma, S. Nagpal, et al., Phases of metabolic and soft tissue changes in months preceding a diagnosis of pancreatic ductal adenocarcinoma, Gastroenterology 156 (6) (2019) 1742–1752 [Internet][cited 2023 May 22] Available from, pubmed.ncbi.nlm.nih.gov/30677401/.
[15] P.S. Tan, C. Garriga, A. Clift, et al., Temporality of body mass index, blood tests, comorbidities and medication use as early markers for pancreatic ductal adenocarcinoma (PDAC): a nested case-control study, Gut 72 (3) (2023) 512–521 [Internet][cited 2023 May 22]Available from, pubmed.ncbi.nlm.nih.gov/35760494/.
[16] G. AF, E. R, N. AG, F. T, T RW, Existing data sources for clinical epidemiology: the clinical laboratory information system (LABKA) research database at Aarhus University, Denmark, Clin Epidemiol 3 (2011) 133 [Internet][cited 2023 May 22] Available from, pubmed.ncbi.nlm.nih.gov/21487452/.
[17] A. Pottegård, S.A.J. Schmidt, H. Wallach-Kildemoes, H.T. Sørensen, J. Hallas, M. Schmidt, Data resource profile: the danish national prescription registry, Int. J. Epidemiol. 46 (3) (2017) 798 [Internet][cited 2023 May 22]Available from, pubmed.ncbi.nlm.nih.gov/27789670/.
[18] M. Schmidt, L. Pedersen, H.T. Sørensen, The Danish civil registration system as a tool in epidemiology, Eur. J. Epidemiol. 29 (8) (2014) 541–549 [Internet][cited 2023 May 22]Available from, pubmed.ncbi.nlm.nih.gov/24965263/.
[19] M. Schmidt, S.A.J. Schmidt, J.L. Sandegaard, V. Ehrenstein, L. Pedersen, H. T. Sørensen, The Danish national patient registry: a review of content, data quality, and research potential, Clin Epidemiol 7 (2015) 449–490 [Internet][cited 2023 May 22]Available from, pubmed.ncbi.nlm.nih.gov/26604824/.
[20] A. Grann, R. Erichsen, A. Nielsen, T. Frøslev, R. Thomsen, Existing data sources for clinical epidemiology: the clinical laboratory information system (LABKA) research database at Aarhus University, Denmark, Clin Epidemiol 3 (2011) 133 [Internet][cited 2023 Mar 9]Available from, pubmed.ncbi.nlm.nih.gov/21487452/.
[21] P. Heus, J.A.A.G. Damen, R. Pajouheshnia, et al., Uniformity in measuring adherence to reporting guidelines: the example of TRIPOD for assessing completeness of reporting of prediction model studies, BMJ Open 9 (4) (2019) 1–6.
[22] the TRIPOD Group. Tripod checklist.
[23] S.S. Olesen, R. Viggers, A.M. Drewes, P. Vestergaard, M.H. Jensen, Risk of major adverse cardiovascular events, severe hypoglycemia, and all-cause mortality in postpancreatitis diabetes mellitus versus type 2 diabetes: a nationwide population-based cohort study, Diabetes Care. 45 (6) (2022) 1326–1334 [Internet][cited 2023 May 22]Available from, pubmed.ncbi.nlm.nih.gov/35312752/.
[24] R. Viggers, M.H. Jensen, H.V.B. Laursen, A.M. Drewes, P. Vestergaard, S.S. Olesen, Glucose-lowering therapy in patients with postpancreatitis diabetes mellitus: a nationwide population-based cohort study, Diabetes Care. 44 (9) (2021) 2045–2052 [Internet][cited 2023 May 22]Available from, pubmed.ncbi.nlm.nih.gov/34362812/.
[25] Alice Z., Amanda C. Feature engineering for machine learning: principles and techniques for data – Alice Zheng, Amanda Casari Google Bøger. 1. O'Reilly Media, inc.; 2018.
[26] L. Breiman, Random forests, Mach Learn 45 (1) (2001) 5–32 [Internet][cited 2022 Dec 21]Available from, link.springer.com/article/10.1023/A:1010933404324.
[27] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection [Internet]. 1995 [cited 2021 Mar 9]. Available from: http://robotics.stanford.edu/~ronnyk.
[28] B.H. Menze, B.M. Kelm, R. Masuch, et al., A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data, BMC Bioinf. 10 (1) (2009) 1–16 [Internet][cited 2023 Feb 14]Available from, link.springer.com/articles/10.1186/1471-2105-10-213.
[29] T. Fushiki, Estimation of prediction error by using K-fold cross-validation, Stat Comput 21 (2) (2011) 137–146 [Internet][cited 2023 Feb 14]Available from, link.springer.com/article/10.1007/s11222-009-9153-8.
[30] C.E. Metz, Basic principles of ROC analysis, Semin. Nucl. Med. 8 (4) (1978) 283–298.
[31] Breiman L. Consistency for a simple model of random forests. 2004.
[32] B. Boursi, T. Patalon, M. Webb, et al., Validation of the enriching new-onset diabetes for pancreatic cancer model: a retrospective cohort study using real-world data, Pancreas 51 (2) (2022) 196–199 [Internet][cited 2023 May 22]Available from, pubmed.ncbi.nlm.nih.gov/35404897/.
[33] W. Chen, R.K. Butler, E. Lustigova, S.T. Chari, B.U. Wu, Validation of the enriching new-onset diabetes for pancreatic cancer model in a diverse and integrated healthcare setting, Dig. Dis. Sci. 66 (1) (2021) [Internet][cited 2023 May 22] Available from, pubmed.ncbi.nlm.nih.gov/32112260/.
[34] R.P. Sah, S.J.S. Nagpal, D. Mukhopadhyay, S.T. Chari, New insights into pancreatic cancer-induced paraneoplastic diabetes, Nat. Rev. Gastroenterol. Hepatol. 10 (7) (2013) 423–433 [Internet][cited 2023 May 22]Available from, pubmed.ncbi.nlm.nih.gov/23528347/.
[35] G. Sagar, R.P. Sah, N. Javeed, et al., Pathogenesis of pancreatic cancer exosome-induced lipolysis in adipose tissue, Gut 65 (7) (2016) 1165–1174 [Internet][cited 2023 May 22]Available from, pubmed.ncbi.nlm.nih.gov/26061593/.
[36] L. Appelbaum, A. Berg, J.P. Cambronero, et al., Development of a pancreatic cancer prediction model using a multinational medical records database, Gastrointestinal Cancers Symposium 39 (2021) 394, 3_suppl394.
[37] CDC, National Diabetes Statistics Report 2020. Estimates of Diabetes and Its Burden in the United States, 2020.
[38] M.H. Jensen, S.L. Cichosz, O. Hejlesen, S.D. Henriksen, A.M. Drewes, S.S. Olesen, Risk of pancreatic cancer in people with new-onset diabetes: a Danish nationwide population-based cohort study, Pancreatology 23 (6) (2023) 642–649 [Internet][cited 2023 Nov 15]Available from, pubmed.ncbi.nlm.nih.gov/37422338/.
[39] A. Maitra, A. Sharma, R.E. Brand, et al., A prospective study to establish a new-onset diabetes cohort: from the consortium for the study of chronic pancreatitis, diabetes, and pancreatic cancer, Pancreas 47 (10) (2018) 1244–1248 [Internet][cited 2023 Nov 16]Available from, pubmed.ncbi.nlm.nih.gov/30325864/.