HHS Public Access

Author manuscript

Gastroenterology. Author manuscript; available in PMC 2018 March 01.

Published in final edited form as:

Gastroenterology. 2017 March ; 152(4): 840–850.e3. doi:10.1053/j.gastro.2016.11.046.

A Clinical Prediction Model to Assess Risk for Pancreatic Cancer Among Patients With New-onset Diabetes

Ben Boursi, M.D.[1,2,3], **Brian Finkelman, Ph.D.**[1], **Bruce J. Giantonio, M.D.**[1,2], **Kevin Haynes, PharmD, M.S.C.E.**[1], **Anil K. Rustgi, M.D.**[1], **Andrew D. Rhim, M.D.**[4], **Ronac Mamtani, M.D., M.S.C.E.**[1,2,*], and **Yu-Xiao Yang, M.D., M.S.C.E.**[1,*]

[1]Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania, USA;

[2]Abramson Cancer Center, University of Pennsylvania, Philadelphia, PA, USA;

[3]Tel-Aviv University, Tel-Aviv, Israel;

[4]Sheikh Ahmed Bin Zayed Al Nahyan Center for Pancreatic Cancer Research and Department of Gastroenterology, Hepatology and Nutrition, University of Texas M.D. Anderson Cancer Center

Abstract

Background and aims—Approximately 50% of all patients with pancreatic ductal adenocarcinoma (PDA) develop diabetes mellitus before their cancer diagnosis. Screening individuals with new-onset diabetes might therefore allow earlier diagnosis of PDA. We sought to develop and validate a PDA risk prediction model to identify high-risk individuals among those with new-onset diabetes.

Methods—We conducted a retrospective cohort study in a population representative database from the UK. Individuals with incident diabetes after the age of 35 and 3 or more years of follow

Correspondence to: Yu-Xiao Yang, MD, MSCE, yangy@mail.med.upenn.edu, Fax: (215)349-5915; Telephone: (215)573-5027.

*Drs. Mamtani and Yang served as co-senior investigators

Department of Medicine and Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania, and Sackler School of Medicine, Tel-Aviv University, Ben Boursi postdoctoral fellow | Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania, Brian Finkelman postdoctoral researcher | Division of Hematology/Oncology, Department of Medicine, University of Pennsylvania, Bruce J Giantonio associate professor of Medicine | Healthcore Inc., Wilmington, Kevin Haynes clinical epidemiologist | Division of Gastroenterology, Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Anil K Rustgi T. Grier Miller professor of Medicine and Genetics | Sheikh Ahmed Bin Zayed Al Nahyan Center for Pancreatic Cancer Research and Department of Gastroenterology, Hepatology and Nutrition, University of Texas M.D. Anderson Cancer Center, Andrew D. Rhim assistant professor of Medicine | Division of Hematology/Oncology, Department of Medicine, University of Pennsylvania, Ronac Mamtani assistant professor of Medicine | Division of Gastroenterology, Department of Medicine and Center for Clinical Epidemiology and Biostatistics, Perelman School of Medicine, University of Pennsylvania, 733 Blockley Hall, 423 Guardian Drive, Philadelphia, PA, USA 19104, Yu-Xiao Yang associate professor of Medicine and Epidemiology

Disclosures: The authors have nothing to disclose.

Authors contribution: Boursi B, Finkelman B, Giantonio BJ, Haynes K, Rustgi AK, Mamtani R, and Yang YX contributed to conception and design of the study; Boursi B and Yang YX acquired the data; Boursi B, Finkelman B, Haynes K, Mamtani R, and Yang YX contributed to analysis of data; Boursi B, Finkelman B, Giantonio BJ, Haynes K, Rhim AD, Rustgi AK, Mamtani R, and Yang YX contributed to drafting the article or revising it critically for important intellectual content; and final approval of the version to be published.

Author names in bold designate shared co-first authorship

up after diagnosis of diabetes were eligible for inclusion. Candidate predictors consisted of epidemiologic and clinical characteristics available at the time of diabetes diagnosis. Variables with P values below .25 in the univariable analyses were further evaluated using backward stepwise approach. Model discrimination was assessed using receiver operating characteristic curve analysis. Calibration was evaluated using the Hosmer–Lemeshow test. Results were internally validated using a bootstrapping procedure.

Results—We analyzed data from 109,385 patients with new-onset diabetes. Among them, 390 (0.4%) were diagnosed with PDA within 3 years. The final model (area under the curve, 0.82; 95% CI, 0.75–0.89) included age, body mass index, change in body mass index, smoking, use of proton pump inhibitors and anti-diabetic medications, as well as levels of HbA1C, cholesterol, hemoglobin, creatinine, and alkaline phosphatase. Bootstrapping validation showed negligible optimism. If the predicted risk threshold for definitive PDA screening was set at 1% over 3 years, only 6.19% of the new-onset diabetes population would undergo definitive screening, which would identify patients with PDA with 44.7% sensitivity, 94.0% specificity, and a positive predictive value of 2.6%.

Conclusion—We developed a risk model based on widely available clinical parameters to help identify patients with new-onset diabetes who might benefit from PDA screening.

Keywords

BMI; pancreatic cancer; insulin; glucose

Introduction

Despite comprising only 3% of all new cancer diagnoses in the United States, pancreatic ductal adenocarcinoma (PDA) remains the fourth most common cause of cancer death, and it is expected to rise to the second most common cause by 2030[1]. According to the US Surveillance, Epidemiology and End Results (SEER) program, there will be an estimated 53,070 new cases and 41,780 deaths from the disease in 2016.[2] The 5-year survival rate for pancreatic cancer is only 7.7% overall.[2] The reason for this abysmal prognosis is that the vast majority (>80%) of patients with PDA are diagnosed at advanced stages.

During the last decade, the median survival of patients with metastatic disease remained around 6-11 months despite recent therapeutic advances.[3-5] According to data from cancer research UK between 7-25% of patients with resectable pancreatic cancer survive for 5 years or more.[6] A similar stage-specific survival trend is also observed in the US. According to SEER data 2006-2012, the 5-year survival for localized pancreatic cancer is 29%, compared to 11% for those with regional lymph node spread and 2.6% for those with distant metastasis.[2] Thus, the only means by which to significantly improve the prognosis of PDA is to detect the cancer at early stages, when the tumor is contained in the pancreas, and prior to the development of overt symptoms. While imaging tests such as CT and MRI and endoscopic ultrasound can identify contained pancreatic tumors as small as 0.5cm, it is not feasible to screen the general population for PDA given the relatively low incidence of the disease. Thus, central to the efforts to improve early diagnosis is the need to enhance our ability to identify high-risk individuals for this disease. The feasibility and cost-effectiveness

of this approach has already been shown in recent screening protocols applied to members of families with familial pancreatic cancer (FPC) and patients with select germline mutations.[7-11] However, only 10-20% of all cases of PDA can be attributed to FPC; the vast majority of PDA arise sporadically with limited family history of this disease.[12]

The epidemiological association between diabetes mellitus and PDA has been reported in numerous studies. It has been estimated that approximately 50% of newly diagnosed PDA patients have diabetes at diagnosis[13, 14]. The risk of PDA is the highest among those with recent onset diabetes[1, 15-18] and/or those with recent initiation of insulin therapy,[15, 19] suggesting that PDA may induce the onset of diabetes mellitus or worsen existing diabetes mellitus. A previous population-based cohort study by Chari et al. indicated that the 3-year cumulative incidence of PDA among patients with new-onset diabetes may reach 0.85%, which is nearly eight times higher than expected.[18] Such PDA-associated diabetes may be a paraneoplastic phenomenon caused by the cancer rather than a result of direct destruction by PDA and loss of normal pancreatic tissue.[20, 21] Frequent resolution of diabetes mellitus after resection of the tumor provides further evidence for such reverse causality.[13] These data and computational models featuring phylogenetic analysis of pancreatic tumors in the same host suggest that the onset of diabetes mellitus may precede the clinical identification of metastatic spread, in at least some cases.[22, 23] Furthermore, it is possible that this association is bi-directional with higher glucose exposure promoting tumor growth. This hypothesis was supported by a meta-analysis demonstrating a linear increase of 14% in pancreatic cancer risk with every 0.56mmol/L increase in fasting blood glucose levels.[24]

Because of the high incidence of diabetes mellitus in the general population and the lack of a low-risk and low-cost PDA screening test, conducting mass PDA screening in all patients with new-onset diabetes mellitus is not feasible. Nevertheless, targeted screening for PDA (e.g., using endoscopic ultrasound [EUS]) may be feasible if diabetic individuals at the highest risk for PDA-associated diabetes could be identified.

Based on existing knowledge, it is highly plausible that a number of epidemiological and/or clinical characteristics (e.g., anthropometric variables such as weight or weight changes[25, 26], lifestyle factors such as smoking[25], medical comorbidities such as pancreatitis[27], medications such as metformin, insulin[28, 29], and laboratory studies such as glucose and cholesterol levels[30]) easily ascertainable around the time of diabets diagnosis may predict PDA risk. Similar risk prediction tools exist in other malignancies, such as the Gail model for breast cancer. To date, epidemiological investigation of predictors of PDA-associated diabetes mellitus has been limited to assessment of single predictors in small single-center samples.

The aim of the current study was to develop and validate a PDA risk prediction model among patients with new-onset diabetes mellitus in a large population-representative electronic medical records database. Such a model would be based on information readily available to clinicians and can be applied to all patients with newly diagnosed diabetes to efficiently identify patients at high risk for PDA.

Methods

Data Source

The study used The Health Improvement Network (THIN), a large primary care electronic research database from the United Kingdom (UK). THIN contains comprehensive medical records on approximately 11 million patients (more than 5% of the UK population) treated by general practitioners in 570 practices throughout the UK. The demographic and geographic distributions of the THIN population are broadly representative of those of the general UK population.[31] The UK National Health System (NHS) provides universal health care coverage through general practitioners. Ninety-eight percent of the UK population is registered with general practitioners who are affiliated with the NHS.[31] In the UK, general practitioners act as the main means of access to all forms of health care provision in the NHS, including specialty referrals and routine hospital admissions. As such, the medical records kept by general practitioners contain complete medical histories of their patients. The THIN database is fundamentally different from claims-based medical databases; it is essentially an electronic version of the actual patient medical record.

To ensure completeness and accuracy, participating practices follow predefined protocols for the recording of computerized clinical data and transfer anonymized patient-based clinical records on a regular basis to the research database. Each medical diagnosis is defined using Read diagnostic codes, which is the standard coding system used by general practices in the UK.[32, 33] Each prescription issued must be entered, including date, dosage, quantity dispensed and duration of therapy. The computerized information includes demographics, diagnosis resulting from general practitioner's consultation, a summary of specialists' clinical notes, hospital discharge letters, and a free text section. Data from each practice are routinely examined to determine whether the research protocol has been followed and to perform quality assessment checks. Data from practices that fail to meet research criteria are not entered as valid data onto the database.[34] Numerous epidemiological studies have been performed using THIN, showing excellent quality of information on prescriptions and medical diagnosis.[31, 34, 35] The study was approved by the Institutional Review Board at the University of Pennsylvania and by the Scientific Review Committee of THIN.

Study Population

The target study population consisted of patients with new-onset diabetes mellitus during follow-up in THIN. All people receiving medical care from a THIN practitioner between 1995-2013 with at least one Read code for diabetes mellitus during the follow-up period were potentially eligible for inclusion. The exclusion criteria were: patients without acceptable medical records (i.e., patients with incomplete documentation or out of sequence date of birth, registration date, date of death, or date of exit from the database); subjects who were diagnosed with diabetes mellitus within the first year after initiation of follow-up in order to avoid prevalent diabetes cases[36]; subjects younger than 35 years of age at the time of diabetes diagnosis that have exceedingly low risk for pancreatic cancer and might have type 1 diabetes mellitus; subjects with a diagnosis of PDA prior to the initial diagnosis of diabetes; subjects with a diagnosis of PDA >3 years after the diagnosis of diabetes since PDA-associated diabetes mellitus is generally diagnosed within 3 years after diabetes

diagnosis.[13,18] In the current source cohort, 169 individuals developed pancreatic cancer beyond the first 3 years after the diabetes diagnosis, with a mean time from diabetes diagnosis to cancer diagnosis of 5.9 (±2.2 SD) years. This group constituted 30.2% (169/559) of all pancreatic cancer patients in the source cohort. Among subjects who did not develop PDA following the initial diagnosis of diabetes mellitus, we excluded those who had <3 years of follow-up in THIN following diabetes diagnosis.

Primary outcome

The outcome of interest was an incident diagnosis of PDA (defined according to diagnostic Read codes) within 3 years following the diagnosis of diabetes mellitus. Based on previous works, PDA diagnosed within a short time (e.g., < 6 months) following a diagnosis of diabetes had potentially a slightly better prognosis compared to PDA cases without previous history of new onset diabetes due to earlier stage at diagnosis[21], suggesting that there is clinical utility in predicting and screening for these cancer cases as well. Therefore, these cases were included in the analysis.

Predictors

As candidate predictors, we included a comprehensive list of PDA risk factors as well as variables related to glucose metabolism (54 candidate variables in total). These predictors included anthropometric variables, lifestyle factors, medical comorbidities, medications, and laboratory studies (supplementary index 1) and were selected based on literature review, biological plausibility, and clinical sense. All variables were available in the practice medical record at the time of initial diabetes diagnosis, the period in which the prediction model is intended to be used. The only exception was anti-diabetic medications for which the definition included prescriptions at or within 6 months after diabetes diagnosis. For laboratory studies, last values at the time or up to 1 year before diabetes diagnosis were used.

Sample size considerations

All available data in the THIN database were used to maximize the power and generalizability of the results. Similar to previous works in THIN[37], we identified 109,385 eligible incident diabetes mellitus patients. Among this cohort of new-onset diabetes mellitus we had 390 cases of PDA with a 3 year cumulative incidence of 0.4%. Thus, our sample size was substantially larger than the recommended 10 events per candidate variable for the derivation of a model and at least 100 events[38] for validation studies.

Missing data

For candidate predictors with less than 60% missingness, we performed multiple imputation using multivariate normal regression (MVN), imputing a total of 20 datasets.[39-42] The MVN method has been shown to be valid whether or not all imputed variables follow a normal distribution.[43] Like all multiple imputation methods, the MVN method assumes that the data are missing at random, which is a less restrictive assumption than that required by complete case analysis.[43]

Statistical Analysis

Given the binary outcome, we developed our prediction model using logistic regression. All analyses were performed using Stata 13 (Stata Corp, College Station, TX).

Model building procedures

Initial variable selection was based on univariable analysis adjusted for the duration of follow-up from registration date to diabetes diagnosis date. All variables associated with a p-value<0.25 in univariable analyses were further assessed by multivariable logistic regression. For all continuous predictors we assessed normality and linearity. Second degree fractional polynomials were used in the presence of nonlinear relationships of the continuous predictors and the outcome.[44, 45] In each of the imputed datasets we used a backward stepwise approach for the multivariable logistic regression with p-values of <0.001 and >0.05 as the inclusion and exclusion thresholds, respectively. Predictors that were selected in ≥50% of the imputation models were included in the final multivariable model. Backward elimination is generally preferred over forward selection as an automated predictor selection procedure because it takes into consideration the correlations among predictors.[46] In a sensitivity analysis we repeated the analysis using a forward stepwise approach. The imputed data sets were then combined (using Rubin's rule) to produce an overall estimate of model coefficients, while taking into account uncertainty in the imputed values.[40-42] The final multivariable model was tested for collinearity defined as variance inflation factor (VIF)>10. Clinically meaningful interactions were tested in the regression model. Specifically, interactions between obesity and cholesterol levels, history of coronary artery disease (CAD), and prescriptions of metformin or proton pump inhibitors as well as the interaction between creatinine levels and metformin prescription were tested.

Measuring prediction model performance

The performance of the prediction model in the derivation cohort was evaluated by examining measures of discrimination and calibration. Discrimination is the ability of the risk score to differentiate between patients who do and do not experience an event (in this case, the diagnosis of PDA) during the study period. This measure is quantified by calculating the area under the receiver operating characteristic curve statistic.[47] Calibration reflects the agreement between predicted probabilities from the model and observed outcomes. We used the Hosmer–Lemeshow test[48] to statistically determine the extent of agreement between the predicted and the observed probabilities.

Internal Validation

We performed an internal validation using a bootstrapping procedure.[49, 50] This approach uses the entire data in order to develop the prediction model and in addition accounts for model overfitting or uncertainty quantifying any optimism in the final prediction model. Moreover, it provides a shrinkage factor that can be used to adjust the regression coefficients and apparent performance for optimism. The bootstrapping in the current study was performed using 100 bootstrap resamples of 44,000 individuals each, each time selecting variables and developing a model within the sample. The discrimination for each new model

was calculated both within the sample as well as in the original cohort allowing us to calculate optimism according to Harrell's algorithm.[51]

Sensitivity analysis

As a sensitivity analysis, we repeated the model building and validation procedures only among patients with PDA that were diagnosed more than 6 months and up to 3 years following the diagnosis of diabetes mellitus. This analysis aimed to evaluate the model after excluding patients with immediate diagnosis of PDA after diabetes mellitus.

Patient involvement—No patients were involved in setting the research question or the outcome measures, nor were they involved in developing plans for recruitment, design, or implementation of the study. No patients were asked to advice on interpretation or writing up of results. There are no plans to disseminate the results of the research to the relevant patient community.

Results

From 179,264 individuals with incident diabetes mellitus in THIN (1995-2013) we identified 109,385 eligible patients with new-onset diabetes (Figure 1). Among this cohort, 390 individuals (0.4%) were diagnosed with PDA within 3 years of diabetes diagnosis (Figure 1). The median follow-up from diabetes diagnosis to PDA diagnosis date was 0.55 years (IQR 0.21 to 1.38).

We evaluated 54 candidate predictors. The median number of missing variables per person was 9 with IQR 5-13, and the mean was 8.8±4.5. Characteristics of individuals with any missing data were comparable to those with completely observed data, in terms of age (62.4±12.4 vs. 62.2±11.7, respectively), sex (53.7% vs. 52.5% males, respectively), duration of follow-up before diabetes diagnosis (5.3±3.1 vs. 6.6±3.2 years, respectively) and percent of individuals that developed PDA within 3 years of diabetes diagnosis (0.4±0.6 vs. 0.3±0.5, respectively) between the groups. Six variables were excluded from the analyses due to missingness of more than 60% (i.e., AST, amylase, ESR, CRP, uric acid, and urinary microalbumin). Among the forty-eight remaining candidate variables 28 had complete data and 20 (mostly lab values) contained <60% missingness and were subjected to the multiple imputation procedure. Of the 48 variables that were analyzed in the univariate logistic regression, 30 variables had a p-value<0.25 (3 out of 5 anthropometry variables, 1 out of 3 behavioral/lifestyle characteristics, 3 out of 8 medical comorbidities, 9 out of 15 medications, and 14 out of 17 laboratory studies). Neutrophils, NLR, alkaline phosphatase, and triglycerides were associated with pancreatic cancer risk in a nonlinear fashion, and second degree fractional polynomials were used in the multivariable regression model (Table 1).

The full multivariable prediction model based on the backward stepwise approach is presented in Table 2. This model included age, BMI, change in BMI per year, smoking, anti-diabetic medications and PPIs, as well as HbA1C, hemoglobin, total cholesterol, creatinine and alkaline phosphatase. Forward selection approach yielded a similar model among imputed data sets (supplementary index 2). The area under the curve of the model was 0.82

(95%CI 0.75 to 0.89) (Figure 2) and the p-value for the Hosmer and Lemeshow goodness of fit test ranged from 0.10 to 0.78 in 18 of the 20 imputed data sets (in two remaining imputed sets the p-value was <0.05). Internal validation of the model using bootstrapping procedure revealed minimal optimism of 0.0003 (95%CI -0.00574 to 0.00571). Figure 3 presents the predictiveness curve for PDA according to the prediction model. If the risk threshold for further PDA screening was set at 10% over 3 years, 0.08% of the new-onset diabetes population would undergo screening, and the sensitivity, specificity and positive predictive value of the model would be 5.5%, 99.9%, and 25.0%, respectively. For a risk threshold of 1% over 3 years, 6.19% of the new-onset diabetes population would undergo screening, and the corresponding sensitivity, specificity and positive predictive value would be 44.7%, 94.0%, and 2.6%. Table 3 presents the sensitivity, specificity and positive predictive value for three different probability cut offs (1%, 5% and 10%).

The sensitivity analysis evaluating PDA risk 6 months to 3 years following new onset diabetes included 109,203 patients with new-onset diabetes, of them 208 individuals (0.2%) were diagnosed with PDA. Results of the backward stepwise approach are presented in supplementary index 3. The full multivariable prediction model included age, change in BMI per year, smoking, anti-diabetic medications, hemoglobin, triglycerides, creatinine and alkaline phosphatase. The area under the curve of the model was 0.77 (95%CI 0.68 to 0.87), similar to the primary model of 0 to 3 years.

Discussion

We developed and internally validated a novel statistical model for the prediction of PDA-associated diabetes mellitus (i.e., PDA diagnosed within 3 years of diabetes onset) among patients with new-onset diabetes mellitus. The model was developed among a large cohort of patients with new-onset diabetes in a population-representative database with validated clinical information. The model was based on demographic, behavioral and clinical variables for which information would be routinely available at the time of diabetes diagnosis. The final model was shown to have excellent discrimination (0.82, 95%CI 0.75 to 0.89) with negligible optimism (0.0003, 95%CI: -0.00574 to 0.00571, based on bootstrapping internal validation) and adequate goodness-of-fit.

The incidence of PDA has been rising in the past ten years by approximately 0.6% per year. Less than 10% of newly diagnosed patients have a localized, potentially resectable disease at the time of diagnosis, and only 7.7% are expected to survive 5 years.[1] Although the morbidity and mortality secondary to PDA are high, universal screening using potential biomarkers such as CA 19-9 or clinical tests (e.g., endoscopic ultrasound, CT or MRI) is not feasible. The existing biomarkers have limited diagnostic accuracy. While highly sensitive and specific, the clinical diagnostic modalities entail substantial cost and/or risk. Therefore, there is an urgent clinical need for novel prediction models and screening methods for the detection of asymptomatic early stage disease that would be both efficient and cost-effective.[52]

Approximately 50% of PDA patients have a diagnosis of new-onset diabetes mellitus within the 3 years prior to cancer diagnosis.[13,14] Similarly, the risk of PDA is markedly increased

up to eight times more than expected within the first 3 years following a new diagnosis of diabetes mellitus. Thus, new-onset diabetes mellitus may define a population which harbors a substantial burden of PDA.[13] However, conducting mass PDA screening using costly and/or invasive tests among all patients with newly diagnosed diabetes mellitus would not be an efficient approach because the vast majority of these patients do not have PDA-associated diabetes mellitus. Our prediction model provides a low-cost and low-risk solution to this problem by identifying high-risk individuals for definitive diagnostic testing. In real-world application, the optimal probability cut-off for definitive screening can be easily adjusted according to resources availability and the performance characteristics of the definitive screening modality. For instance, if EUS with fine needle aspiration (FNA) is used with >90% sensitivity and specificity, one would likely choose a higher cut-off which provides a higher PPV due to the morbidity associated with this invasive test (perforation, infection, pancreatitis, and hemorrhage in 2-3% of patients[53]). However if a non-invasive test, such as imaging with CT or MRI (with one-time sensitivity and specificity of ~80%), is used as the next step in screening before a definitive diagnostic test, we might prefer a lower cut-off which provides higher sensitivity for the screening test. For example, using a 1% predicted risk of PDA as the threshold for proceeding with definitive testing, only 6.19% of the entire new-onset diabetes mellitus population would need to undergo the definitive testing, and yet nearly half of all PDA-associated diabetes mellitus cases in this population would be captured with a number needed to screen of 38. It is important to note that this strategy would allow these cancers to be diagnosed months to as much as 3 years earlier than when they would have been diagnosed under current practice, thus potentially improving prognosis. Our model could also be used to efficiently identify a high-risk subpopulation among those with the new-onset diabetes mellitus in which additional humoral or genetic biomarker evaluation could be applied before proceeding with the invasive definitive testing; indeed the need for an inexpensive, rapid pre-screening “sieve” for patients with new-onset diabetes mellitus was emphasized recently.[12]

There were several strengths to the study. The study cohort included approximately 180,000 patients with new-onset diabetes mellitus and at least 3 years of follow-up. The large sample size and number of events minimized the risk of model overfitting (i.e., selecting spurious predictors) or underfitting (i.e., failing to include important predictors). The quality of medical diagnosis and prescription information in THIN was previously shown in numerous pharmacoepidemiology studies.[31, 34, 35] The incidence of cancer in THIN was shown to be valid compared to cancer registry data in the entire population of the UK.[54, 55] In an additional study the PPV for incidence PDA identified using Read codes in THIN was 97% based on manual chart review, further supporting the validity of PDA diagnosis in THIN.[56]

While our main analysis focused on PDA within 3 years following the diagnosis of diabetes mellitus, the sensitivity analysis focusing on PDA diagnosed 6 month to 3 years following diabetes diagnosis identified a model containing a similar set of predictors and having similar performance characteristics as our primary analysis, indicating that our model was robust in predicting PDA risk for the entire 3 year period following diabetes diagnosis.

In addition, the associations observed in the multivariable prediction model provided potential mechanistic insights regarding the pathogenesis involved in PDA-induced diabetes

mellitus (a paraneoplastic phenomenon) versus type 2 diabetes mellitus. Variables conventionally associated with increased risk of type 2 diabetes (such as BMI, hypercholesterolemia and hypertriglyceridemia) were associated with lower PDA risk. Renal failure and elevated creatinine levels, both known complications of long lasting type 2 diabetes mellitus, were associated with decreased PDA risk; and higher HbA1C levels and initial treatment with insulin, not commonly used as first line treatment in patients with type 2 diabetes mellitus were associated with higher cancer risk, similar to an additional work that was recently published.[56] Our results regarding BMI were also consistent with a previous study demonstrating a link between weight loss preceding the onset of diabetes mellitus and the subsequent risk of PDA diagnosis.[26]

The current study had several potential limitations. Several known risk factors for PDA, such as family history and dietary pattern, are not included in the THIN database. In addition we were not able to include genetic polymorphisms, epigenetic changes and information regarding circulating tumor cells that are emerging as novel tumor markers, although this is similar to other risk assessment tools in other malignancies. However, the current model was shown to have excellent predictive power. We also had substantial missing data among a number of laboratory-based variables; we implemented multiple imputation procedures to account for the missing data. Furthermore, the model was not intended to be a definitive diagnostic test but rather part of a sequential approach to identify individuals at high-risk for PDA. Indeed, its sole reliance on information routinely available as a part of a general practice medical records means that this model can be easily applied in practice to virtually every patient with new-onset diabetes mellitus with negligible cost or risk. A recent cost analysis further estimated that screening for pancreatic cancer using MRI/MRCP for 3 years among individuals with new onset diabetes over 50 with either weight loss or smoking is affordable with a cost per added year ranging from $356.42 based on Medicare costs to $1418.92 based on national average.[57] However this analysis did not consider the incidence of new-onset diabetes and the cost-benefit balance at the population level. A dedicated cost-utility analysis, incorporating all relevant risk estimates, performance characteristics of diagnostic tests and costs, is necessary to adequately evaluate the cost-effectiveness of the various screening strategies among new-onset diabetes patients.

Since THIN lacks detailed cancer stage data, we were not able to assess whether the model is better at detecting early versus late stage disease and thus to evaluate the full impact on PDA survival. By implementing a screening strategy among new-onset diabetes patients which would allow detection of a substantial burden of pancreatic cancers up to 3 years earlier than the time of clinical diagnosis under the current practice, we hope it would allow us to detect a higher proportion of early-stage pancreatic cancers. Future prospective studies are needed to demonstrate this effect. Further research is also needed to evaluate whether using this prediction model results in a higher percent of patients that are eligible for potentially curative surgery as well as an improvement in mortality.

The current study was conducted among a new-onset diabetes mellitus cohort from the UK general practice. The 3-year cumulative incidence of pancreatic cancer observed in our cohort is somewhat lower than that reported in a regional US population-based cohort[18] but slightly higher than the incidence reported from the US nation-wide VA cohort.[16, 17] The

exact reason for these differences is not clear, but they highlight the need for external validation of our prediction model.

In summary, we developed and internally validated an easily automatable and inexpensive clinical prediction tool to identify individuals at high-risk for PDA among the enormous population of patients with new-onset diabetes mellitus. Because approximately 50% of all PDA cases are associated with recent onset diabetes mellitus, this novel prediction model could potentially lead to improved prognosis for a substantial proportion of all PDA cases. External validation of this prediction model, preferably within a prospective cohort of new-onset diabetes patients, would be imperative to confirm and potentially improve the performance characteristics of our model before it can be considered for clinical use. Additionally, before use in clinical practice, it will be necessary to evaluate the impact of the suggested screening approach on important clinical outcomes as well as on the sensitivity of the follow-up tests. There is also a need to define the appropriate follow-up screening modalities and schedules. Further risk factors such as CA 19-9 levels and specific genetic alterations might be combined in the future as additional steps before a definitive diagnostic test is performed in order to decrease the number of people exposed to the potential morbidity of invasive diagnostic procedures.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank James D. Lewis, M.D., M.S.C.E. for his guidance and critical review of the manuscript.

Grant support: This study was supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant UL1TR000003 and the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) P30-DK050306 Center for Molecular Studies in Digestive and Liver Diseases. Dr. Mamtani was supported by NIH K23 grant CA187185. Dr. Rhim was supported by NIH grants DK088945 and CA177857 and a Rising Stars Award from the Cancer Prevention Research Institute of Texas. Dr. Rustgi was supported by the Lustgarten Family Fund and by NIH R01 grant DK060694. There was no support from any other organization for the submitted work.

References

1. Siegel R, Miller KD, Jemal A. Cancer statistics, 2015. CA Cancer J Clin. 2015; 65(1):5–29. [PubMed: 25559415]
2. http://seer.cancer.gov/statfacts/html/pancreas.html, accessed 10/12/2016.
3. Burris HA 3rd, Moore MJ, Andersen J, et al. Improvements in survival and clinical benefit with gemcitabine as first-line therapy for patients with advanced pancreas cancer: a randomized trial. J Clin Oncol. 1997; 15(6):2403–13. [PubMed: 9196156]
4. Von Hoff DD, Ervin T, Arena FP, et al. Increased survival in pancreatic cancer with nab-paclitaxel plus gemcitabine. N Engl J Med. 2013; 369(18):1691–703. [PubMed: 24131140]
5. Conroy T, Desseigne F, Ychou M, et al. FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer. N Engl J Med. 2011; 364(19):1817–25. [PubMed: 21561347]
6. http://www.cancerresearchuk.org/about-cancer/type/pancreatic-cancer/treatment/statistics-and-outlook-for-pancreatic-cancer, accessed 10/12/2016.
7. Canto MI, Hruban RH, Fishman EK, et al. Frequent detection of pancreatic lesions in asymptomatic high-risk individuals. Gastroenterology. 2012; 142(4):796–804. [PubMed: 22245846]

8. Bartsch DK, Slater EP, Carrato A, et al. Refinement of screening for familial pancreatic cancer. Gut. 2016; 65(8):1314–21. [PubMed: 27222532]
9. Harinck F, Konings IC, Kluijt I, et al. A multicentre comparative prospective blinded analysis of EUS and MRI for screening of pancreatic cancer in high-risk individuals. Gut. 2015; doi: 10.1136/gutjnl-2014-308008
10. Shin EJ, Topazian M, Goggins MG, et al. Linear-array EUS improves detection of pancreatic lesions in high-risk individuals: a randomized tandem study. Gastrointest Endosc. 2015; 82(5): 812–8. [PubMed: 25930097]
11. Rustgi AK. Familial pancreatic cancer: genetic advances. Genes Dev. 2014; 28(1):1–7. [PubMed: 24395243]
12. Chari ST, Kelly K, Hollingsworth MA, et al. Early detection of sporadic pancreatic cancer: summative review. Pancreas. 2015; 44(5):693–712. [PubMed: 25931254]
13. Pannala R, Leirness JB, Bamlet WR, et al. Prevalence and clinical profile of pancreatic cancer-associated diabetes mellitus. Gastroenterology. 2008; 134(4):981–7. [PubMed: 18395079]
14. Aggarwal G, Rabe KG, Petersen, et al. New-onset diabetes in pancreatic cancer: a study in the primary care setting. Pancreatology. 2012; 12(2):156–61. [PubMed: 22487526]
15. Wang F, Gupta S, Holly EA. Diabetes mellitus and pancreatic cancer in a population-based case-control study in the San Francisco Bay Area, California. Cancer Epidemiology, Biomarkers & Prevention. 2006; 15(8):1458–63.
16. Gupta S, Vittinghoff E, Bertenthal D, et al. New-onset diabetes and pancreatic cancer. Clin Gastroenterol Hepatol. 2006; 4(11):1366–72. [PubMed: 16945591]
17. Munigala S, Singh A, Gelrud A, et al. Predictors for Pancreatic Cancer Diagnosis Following New-Onset Diabetes Mellitus. Clin Transl Gastroenterol. 2015; 6:e118.doi: 10.1038/ctg.2015.44 [PubMed: 26492440]
18. Chari ST, Leibson CL, Rabe KG, et al. Probability of pancreatic cancer following diabetes: a population-based study. Gastroenterology. 2005; 129(2):504–11. [PubMed: 16083707]
19. Bonelli L, Aste H, Bovo P, et al. Exocrine pancreatic cancer, cigarette smoking, and diabetes mellitus: a case-control study in northern Italy. Pancreas. 2003; 27(2):143–9. [PubMed: 12883263]
20. Sah RP, Nagpal SJ, Mukhopadhyay D, et al. New insights into pancreatic cancer-induced paraneoplastic diabetes. Nat Rev Gastroenterol Hepatol. 2013; 10(7):423–33. [PubMed: 23528347]
21. Pelaez-Luna M, Takahashi N, Fletcher JG, et al. Resectability of presymptomatic pancreatic cancer and its relationship to onset of diabetes: a retrospective review of CT scans and fasting glucose values prior to diagnosis. Am J Gastroenterol. 2007; 102(10):2157–63. [PubMed: 17897335]
22. Yachida S, Jones S, Bozic I, et al. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature. 2010; 467(7319):1114–7. [PubMed: 20981102]
23. Campbell PJ, Yachida S, Mudie LJ, et al. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature. 2010; 467(7319):1109–13. [PubMed: 20981101]
24. Liao WC, Tu YK, Wu MS, et al. Blood glucose concentration and risk of pancreatic cancer: systematic review and dose-response meta-analysis. BMJ. 2015; 349:g7371.doi: 10.1136/bmj.g7371 [PubMed: 25556126]
25. Klein AP, Lindström S, Mendelsohn JB, 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. [PubMed: 24058443]
26. Hart PA, Kamada P, Rabe KG, et al. Weight loss precedes cancer-specific symptoms in pancreatic cancer-associated diabetes mellitus. Pancreas. 2011; 40(5):768–72. [PubMed: 21654538]
27. Lowenfels AB, Maisonneuve P, Cavallini G, et al. Pancreatitis and the risk of pancreatic cancer. International Pancreatitis Study Group. N Engl J Med. 1993; 328(20):1433–7. [PubMed: 8479461]
28. Mohammed A, Janakiram NB, Brewer M, et al. Antidiabetic Drug Metformin Prevents Progression of Pancreatic Cancer by Targeting in Part Cancer Stem Cells and mTOR Signaling. Transl Oncol. 2013; 6(6):649–59. [PubMed: 24466367]
29. Lonardo E, Cioffi M, Sancho P, et al. Metformin targets the metabolic Achilles heel of human pancreatic cancer stem cells. PLoS One. 2013; 8(10):e76518. [PubMed: 24204632]

30. Stocks T, Bjørge T, Ulmer H, et al. Metabolic risk score and cancer risk: pooled analysis ofseven cohorts. Int J Epidemiol. 2015; 44(4):1353–63. [PubMed: 25652574]
31. Blak BT, Thompson M, Dattani H, et al. Generalisability of The Health Improvement Network (THIN) database: demographics, chronic disease prevalence and mortality rates. Inform Prim Care. 2011; 19(4):251–5. [PubMed: 22828580]
32. Chisholm J. The Read clinical classification. BMJ. 1990; 300(6732):1092. [PubMed: 2344534]
33. Benson T. The history of the Read Codes: the inaugural James Read Memorial Lecture 2011. Inform Prim Care. 19(3):173–82. [PubMed: 22688227]
34. Lewis JD, Schinnar R, Bilker WB, et al. Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research. Pharmacoepidemiol Drug Saf. 2007; 16(4): 393–401. [PubMed: 17066486]
35. Mamtani R, Haynes K, Boursi B, et al. Validation of a coding algorithm to identify bladder cancer and distinguish stage in an electronic medical records database. Cancer Epidemiol Biomarkers Prev. 2015; 24(1):303–7. [PubMed: 25389114]
36. Lewis JD, Bilker WB, Weinstein RB, et al. The relationship between time since registration and measured incidence rates in the general practice research database. Pharmacoepidemiol Drug Saf. 2005; 14(7):443–451. [PubMed: 15898131]
37. Boursi B, Mamtani R, Haynes K, et al. The effect of past antibiotic exposure on diabetes risk. Eur J Endocrinol. 2015; c172(6):639–48.
38. Vergouwe Y, Steyerberg EW, Eijkemans MJ, et al. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol. 2005; 58(5):475–83. [PubMed: 15845334]
39. Schafer JL. Multiple imputation: a primer. Stat Methods Med Res. 1999; 8(1):3–15. [PubMed: 10347857]
40. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011; 30(4):377–99. [PubMed: 21225900]
41. Marshall A, Altman DG, Holder RL, et al. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Med Res Methodol. 2009; 9:57. [PubMed: 19638200]
42. Wood AM, White IR, Royston P. How should variable selection be performed with multiply imputed data? Stat Med. 2008; 27(17):3227–46. [PubMed: 18203127]
43. Lee KJ, Carlin JB. Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. Am J Epidemiol. 2010; 171(5):624–32. [PubMed: 20106935]
44. Royston P, Sauerbrei W. Building multivariable regression models with continuous covariates in clinical epidemiology--with an emphasis on fractional polynomials. Methods Inf Med. 2005; 44(4):561–71. [PubMed: 16342923]
45. Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Stat Med. 2007; 26(30):5512–28. [PubMed: 18058845]
46. Mantel N. Why stepdown procedures in variable selection? Technometrics. 1970; 12:621–5.
47. Metz CE. Basic principles of ROC analysis. Semin Nucl Med. 1978; 8(4):283–98. [PubMed: 112681]
48. Hosmer, DW., Lemeshow, S. Applied Logistic Regression. Wiley; 2000.
49. Efron B. Bootstrap Methods: Another Look at the Jackknife. Ann Statist. 1979; 7:1–26.
50. Steyerberg, EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York: Springer; 2009.
51. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine. 1996; 15(4):361–87. [PubMed: 8668867]
52. Kemmer BJ, Chari ST, Cleeter DF, et al. Early detection of sporadic pancreatic cancer, strategic map for innovation - a white paper. Pancreas. 2015; 44(5):686–92. [PubMed: 25938853]

53. Eloubeidi MA, Tamhane A, Varadarajulu S, et al. Frequency of major complications after EUS-guided FNA of solid pancreatic masses: a prospective evaluation. Gastrointest Endosc. 2006; 63(4):622–9. [PubMed: 16564863]

54. Haynes K, Forde KA, Schinnar R, et al. Cancer incidence in The Health Improvement Network. Pharmacoepidemiol Drug Saf. 2009; 18(8):730–6. [PubMed: 19479713]

55. Cea Soriano L, Soriano-Gabarro M, Garcia Rodriguez LA. Validity and completeness of colorectal cancer diagnoses in a primary care database in the United Kingdom. Pharmacoepidemiol Drug Saf. 2016; 25(4):385–91. [PubMed: 26436320]

56. Lu Y, Garcia Rodriguez LA, Malgerud L, et al. New-onset type 2 diabetes, elevated HbA1c, anti-diabetic medications, and risk of pancreatic cancer. Br J Cancer. 2015; 113(11):1607–14. [PubMed: 26575601]

57. Bruenderman E, Martin RC 2nd. A cost analysis of a pancreatic cancer screening protocol in high-risk populations. Am J Surg. 2015; 210(3):409–16. [PubMed: 26003200]

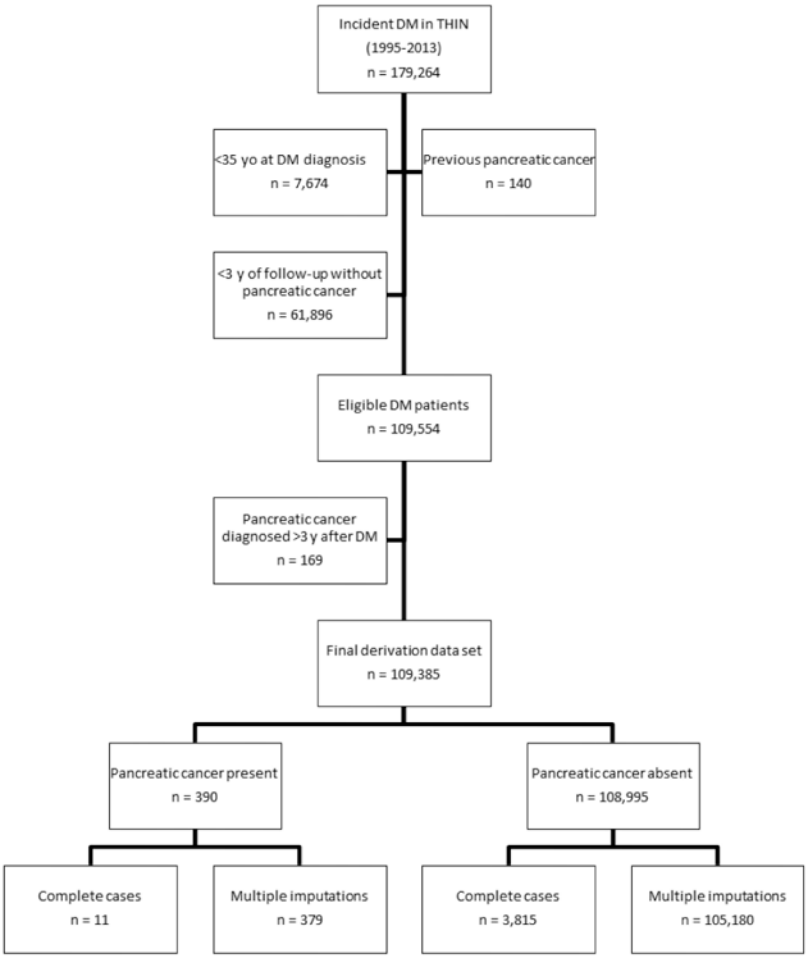

Figure 1. Participants flow diagram

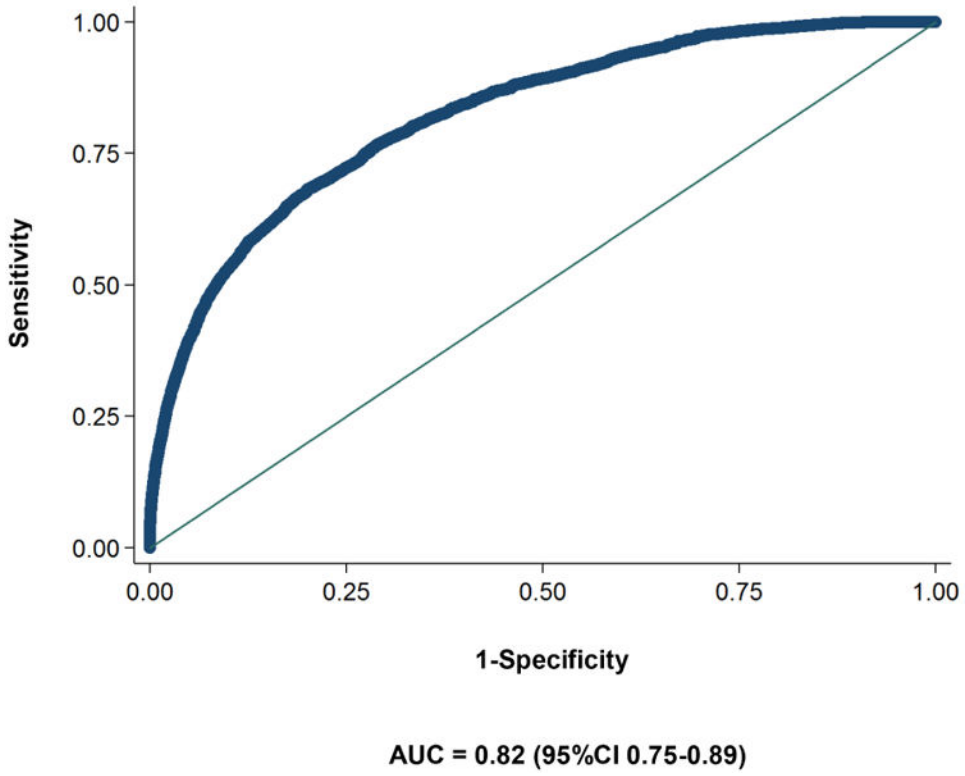

Figure 2. Reciever operator curve of the final model

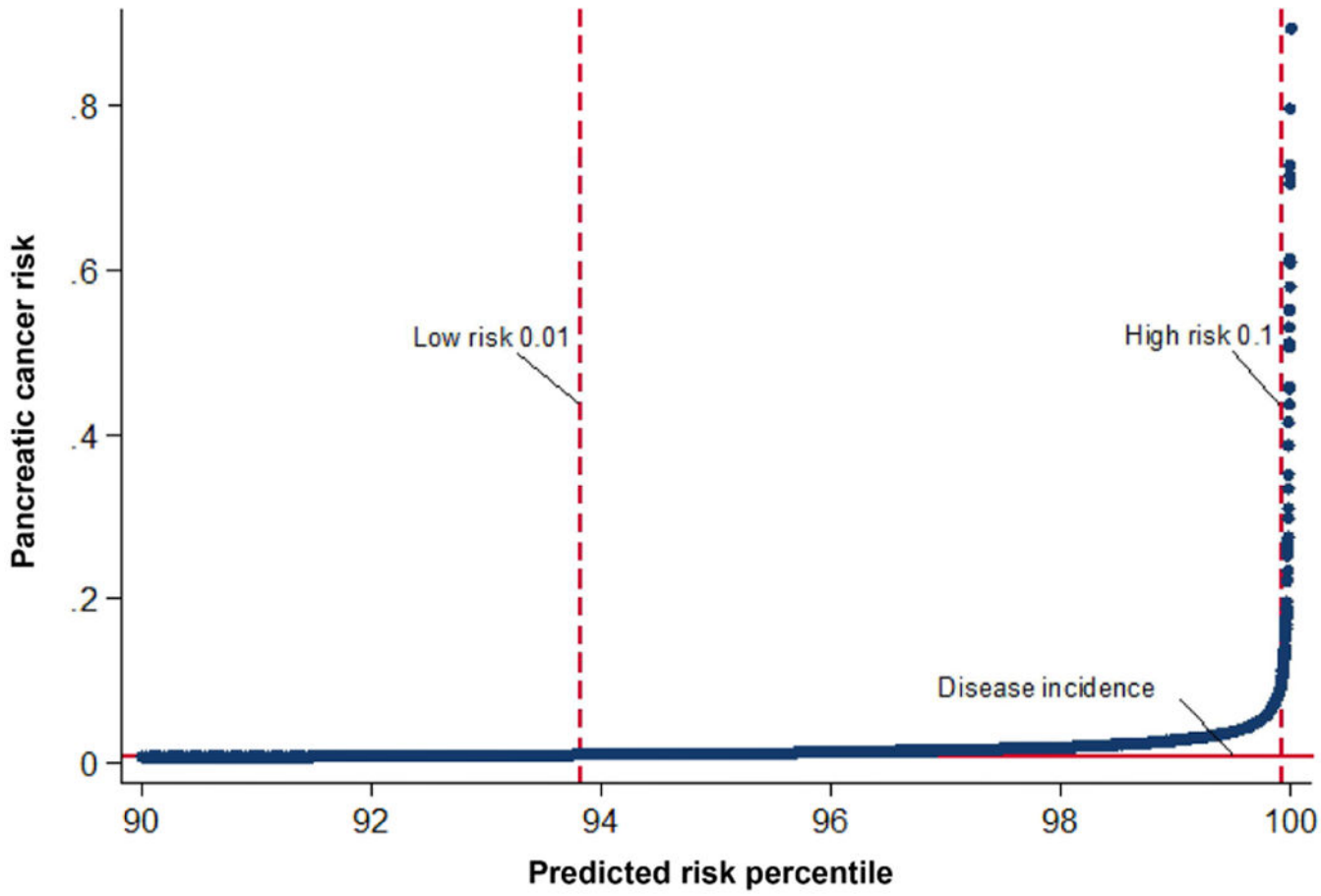

Figure 3. Predictiveness curve for the pancreatic cancer risk model in table 2. Shown are the risk thresholds for <0.01 for low risk >0.1 for high risk

Gastroenterology. Author manuscript; available in PMC 2018 March 01.

Table 1

Participants' characteristics and effect of pancreatic cancer risk in the univariable logistic regression analysis

Predictor[1,2]	Entire study cohort N (%) for dichotomous and median (IQR) for continous variables	Individuals with pancreatic cancer during follow-up N (%) for dichotomous and median (IQR) for continous variables	Individuals without pancreatic cancer during follow-up	Missingness N (%)	Crude OR (95% CI)[3]	P-value
Overall	**109,385**	**390**	**108,995**	-	-	-
Anthropometry						
Age, Median (IQR)	62.7 (53.4-71.6)	70.7 (63.0-78.2)	62.7 (53.4-71.6)	0	1.06 (1.05-1.07)	<0.001
Sex (Males), N(%)[4]	58,678 (53.6)	209 (53.6)	58,469 (53.6)	0	1.01 (0.83-1.23)	0.92
BMI, Median (IQR)	29.9 (26.5-34.1)	27.9 (24.9-31.6)	29.9 (26.5-34.1)	11,228 (10.3)	0.94 (0.92-0.95)	<0.001
ΔBMI/year, Median (IQR)[5]	-0.18 (-0.43-0.04)	0 (-0.20-0.21)	-0.18 (-0.43-0.04)	58,332 (53.3)	1.24 (1.15-1.35)	<0.001
Height (m), Median (IQR)[4]	1.68 (1.60-1.76)	1.68 (1.61-1.75)	1.68 (1.60-1.76)	53,298 (48.7)	0.69 (0.19-2.56)	0.58
Behavioral/lifestyle characteristics						
Smoking (ever), N(%)	52,800 (48.3)	216 (55.4)	52,584 (48.2)	0	1.18 (0.97-1.45)	0.1
Alcohol use, N(%)[4]	58,248 (53.3)	225 (57.7)	58,023 (53.2)	0	0.99 (0.81-1.22)	0.94
Alcohol use disorder, N(%)[4]	770 (0.7)	3 (0.8)	767 (0.7)	0	0.94 (0.30-2.93)	0.91
Medical comorbidities						
Personal history of cancer,N(%)[4]	4,349 (4.0)	20 (5.1)	4,329 (4.0)	0	1.25 (0.80-1.97)	0.98
Coronary artery disease	15,337 (14.0)	72 (18.5)	15,265 (14.0)	0	1.41 (1.09-1.83)	0.008
Cerbrovascular disease	4,510 (4.1)	32 (8.2)	4,478 (4.1)	0	2.28 (1.59-3.28)	<0.001
Hypertension[4]	41,947 (38.4)	159 (40.8)	41,788 (38.3)	0	1.05 (0.86-1.28)	0.65
Neuropathy[4]	514 (0.5)	3 (0.8)	511 (0.5)	0	1.71 (0.55-5.33)	0.36
Retinopathy	266 (0.2)	2 (0.5)	264 (0.2)	0	2.30 (0.57-9.29)	0.24
Auto-immune disease[4]	2,965 (2.7)	12 (3.1)	2,953 (2.7)	0	1.11 (0.62-1.97)	0.72
Chronic pancreatitis[4]	211 (0.2)	0 (0)	211 (0.2)	0	NA	NA
Medication use						
1). Diabetes treatment						
Insulin	794 (0.7)	11 (2.8)	783 (0.7)	0	4.28 (2.34-7.84)	<0.001

Predictor[1,2]	Entire study cohort N (%) for dichotomous and median (IQR) for continous variables	Individuals with pancreatic cancer during follow-up N (%) for dichotomous and median (IQR) for continous variables	Individuals without pancreatic cancer during follow-up	Missingness N (%)	Crude OR (95% CI)[3]	P-value
Oral anti-diabetics (not metformin)	11,804 (10.8)	149 (38.2)	11,655 (10.7)	0	5.49 (4.46-6.74)	<0.001
Metformin	32,103 (29.4)	160 (41.0)	31,943 (29.3)	0	1.55 (1.26-1.90)	<0.001
First treatment lifestyle modification	71,167 (65.1)	164 (42.1)	71,003 (65.1)	0	0.41 (0.33-0.50)	<0.001
2). Other medications						
Aspirin	26,387 (24.1)	123 (31.5)	26,264 (24.1)	0	1.41 (1.13-1.74)	0.002
Statins	34,066 (31.1)	143 (36.7)	33,923 (31.1)	0	1.15 (0.93-1.42)	0.19
Fibrates	2,807 (2.6)	15 (3.9)	2,792 (2.6)	0	1.46 (0.87-2.45)	0.15
Proton pump inhibitors (PPIs)	19,146 (17.5)	116 (29.7)	19,030 (17.5)	0	1.89 (1.52-2.36)	<0.001
ACE inhibitors[4]	26,394 (24.1)	106 (27.2)	26,288 (24.1)	0	1.10 (0.88-1.38)	0.41
Angiotensin receptor blockers (ARBs) [4]	8,289 (7.6)	34 (8.7)	8,255 (7.6)	0	1.09 (0.76-1.55)	0.64
Bisphosphanates[4]	849 (0.8)	2 (0.5)	847 (0.8)	0	0.67 (0.17-2.68)	0.57
Hydroxychloraquine[4]	111 (0.1)	1 (0.3)	110 (0.1)	0	2.51 (0.35-18.06)	0.36
Lithium[4]	380 90.4)	1 (0.3)	379 (0.4)	0	0.73 (0.10-5.21)	0.75
Digoxin	3,710 (3.4)	20 (5.1)	3,690 (3.4)	0	1.54 (0.98-2.42)	0.06
Penicillin (>5 courses)[4]	3,996 (3.7)	24 (6.2)	3,972 (3.6)	0	1.11 (0.71-1.73)	0.64
Laboratory studies						
HbA1C	6.4 (5.7-7.9)	7.5 (6.4-10.8)	6.4 (5.7-7.8)	62,581 (57.2)	1.22 (1.16-1.29)	<0.001
Glucose (mmol/L)	8.4 (6.6-12.2)	9.6 (7.4-14.6)	8.4 (6.6-12.2)	38,219 (34.9)	1.08 (1.04-1.11)	<0.001
Hemoglobin (g/dL)	14.4 (13.5-15.4)	14.0 (12.9-15.1)	14.4 (13.5-15.4)	58,511 (53.5)	0.83 (0.76-0.89)	<0.001
WBC ($\times 10^9$/L)	7.2 (6.1-8.7)	7.4 (6.4-8.6)	7.2 (6.1-8.7)	51,761 (47.3)	1.03 (0.99-1.06)	0.08
Platelets ($\times 10^9$/L)[4]	248 (208-296)	245 (202-300)	248 (208-296)	51,576 (47.2)	1.0 (0.99-1.0)	0.62
Neutrophils	4.3 (3.4-5.6)	4.7 (3.8-6.0)	4.3 (3.4-5.6)	54,982 (50.3)	1.11 (1.04-1.18)	0.001
Neutrophils^2[6]	-	-	-	-	0.99 (0.99-1.0)	0.001
Lymphocytes	2.2 (1.7-2.8)	1.9 (1.5-2.5)	2.2 (1.7-2.8)	55,004 (50.3)	0.98 (0.96-1.0)	0.12
Neutrophil to lymphocyte ratio (NLR)	2.0 (1.5-2.7)	2.4 (1.8-3.4)	2.0 (1.5-2.7)	55,473 (50.7)	1.18 (1.08-1.29)	<0.001
NLR^2[6]	-	-	-	-	0.99 (0.99-1.0)	0.004

Predictor[1,2]	Entire study cohort N (%) for dichotomous and median (IQR) for continous variables	Individuals with pancreatic cancer during follow-up N (%) for dichotomous and median (IQR) for continous variables	Individuals without pancreatic cancer during follow-up	Missingness N (%)	Crude OR (95% CI)[3]	P-value
Creatinine (μmol/L)	87 (76-100)	88 (75-100)	87 (76-100)	28,144 (25.7)	1.00 (0.99-1.01)	0.16
Bilirubin (μmol/L)[4]	10 (8-14)	10 (8-14)	10 (8-14)	45,175 (41.3)	1.01 (0.99-1.03)	0.48
ALT (U/L)	30 (21-44)	25 (18-44)	30 (21-44)	57,127 (52.2)	1.01 (1.0-1.01)	<0.001
Alkaline phosphatase (U/L)	84 (68-106)	98 (76-131)	84 (68-106)	44,332 (40.5)	1.01 (1.0-1.01)	<0.001
Alkaline phosphatase^2[6]	-	-	-	-	0.99 (0.99-1.00)	<0.001
Albumin (g/L)	42 (40-45)	42 (39-44)	42 (40-45)	46,603 (42.6)	0.93 (0.91-0.96)	<0.001
Total Cholesterol (mmol/L)	5.3 (4.5-6.2)	4.9 (4.3-5.7)	5.3 (4.5-6.2)	32,027 (29.3)	0.80 (0.72-0.89)	<0.001
LDL Cholesterol (mmol/L)	3.1 (2.4-3.8)	2.8 (2.1-3.3)	3.1 (2.4-3.8)	62,718 (57.3)	0.76 (0.65-0.89)	0.001
HDL Cholesterol (mmol/L)[4]	1.2 (1.0-1.4)	1.2 (1.0-1.4)	1.2 (1.0-1.4)	48,883 (44.7)	0.91 (0.63-1.31)	0.61
Triglycerids (mmol/L)	1.9 (1.3-2.7)	1.6 (1.2-2.4)	1.9 (1.3-2.7)	47,436 (43.4)	0.80 (0.70-0.91)	0.001
Triglycerids^2[6]	-	-	-	-	1.0 (0.99-1.0)	0.09

[1] All variables were defined prior to diabetes diagnosis except anti-diabetic medications that included prescriptions up to 6 months after diabetes diagnosis.

[2] The following predictors were excluded due to missingness of more than 60%: AST, amylase, ESR, CRP, uric acid, and urinary microlbumin.

[3] Adjusted to duration of follow-up from registration date to diabetes index date.

[4] P-value<0.25, the variable was not included in the backward stepwise procedure.

[5] Calculated by subtracting the last BMI measurement before diabetes diagnosis from the first measured BMI during follow-up, and dividing the difference by the time between the tests. This calculation was performed only for individuals with ≥ 1year between the two BMI measurements.

[6] Quadratic term for a non-linear parameter.

Table 2

Final multivariable prediction model* and a case example

Predictor	β Coefficient	SE	Odds ratio (OR)	95%CI	P-value
Anthropometry					
Age	0.0500853	0.005061 8	1.05	1.04-1.06	<0.001
BMI	-0.0245365	0.010285 5	0.98	0.96-1.00	0.017
ΔBM I/year	0.1421699	0.054726 7	1.15	1.04-1.28	0.01
Behavioral/lifestyle characteristics					
Smoking	0.4150802	0.107410 2	1.51	1.23-1.87	<0.001
Medication use					
Insulin	0.8695054	0.341289 9	2.39	1.22-4.66	0.011
Oral hypoglycemics (not metformin)	1.129888	0.145582 6	3.10	2.33-4.12	<0.001
Metformin	0.3105286	0.131609 6	1.36	1.05-1.77	0.018
PPIs	0.4138479	0.117458 3	1.51	1.20-1.90	<0.001
Laboratory studies					
HbA1C	0.0989848	0.037920 3	1.10	1.02-1.19	0.011
Hb (g/dL)	-0.1419052	0.040488 8	0.87	0.80-0.94	0.001
Total cholesterol (mmol/L)	-0.1902255	0.055398 5	0.83	0.74-0.92	0.001
Creatinine (μmol/L) Creatinine^2	-0.0190353 0.0000495	0.005534 40.000018 4	0.98 1.00005	0.97-0.99 1.00001-1.00009	0.001 0.007
Alkaline phosphatase(U/L)Alkaline phosphatase^2	0.0084294 -2.97e-06	0.000764 15.67e-07	1.01 0.99	1.00-1.010.999996-0.999998	<0.001 <0.001
Intercept	-6.35211	0.894360 1			

*The formula of the resulting logistic model is:

P_3 year probability for pancreatic cancer following diabetes diagnosis = $e(X\beta)/1+ e(X\beta)$

$X\beta$ = 0.0500853 × age – 0.0245365 × BMI + 0.1421699 × ΔBMI/t + 0.4150802 × ever smoking + 0.8695054 × insulin + 1.129888 × oral hypoglycemic drugs that are not metformin + 0.3105286 × metformin + 0.4138479 × PPIs + 0.0989848 × HbA1C – 0.1419052 × Hb – 0.1902255 × total cholesterol – 0.0190353 creatinine + 0.0000495 × creatinine^2 + 0.0084294 × alkaline phosphatase – 2.97e-06 alkaline phosphatase^2 – 6.35211

Case example: a 70-year-old male with new-onset diabetes mellitus who is a smoker with a BMI of 20 and a drop of 2 BMI units over the past year. His medications include insulin, an oral hypoglycemic agent other than metformin and a PPI. His lab values include HbA1C of 7.5, total cholesterol of 4mmol/L (155mg/dl), Hb 10g/dl, creatinine 90μmol/L (1.02mg/dl) and alkaline phosphatase of 120U/L. According to the model, his predicted 3-year risk for PDA would be 11.9%.

Table 3

Model diagnostic performance at different predicted probability cut-offs

Probability cut-off	Diagnostic performance
1%:	
Sensitivity	44.74%
Specificity	93.95%
PPV	2.6%
5%:	
Sensitivity	10.53%
Specificity	99.74%
PPV	12.94%
10%:	
Sensitivity	5.53%
Specificity	99.94%
PPV	25.0%

PPV= positive predictive value