Title: Predicting pancreatic cancer in the UK Biobank cohort using polygenic risk scores and diabetes mellitus.

Short title: Polygenic risk scores is associated with diabetogenic pancreatic cancer

Authors:

1- **Shreya Sharma;** MSc, University of Southampton, Human Development and Health.

2- **William J Tapper;** PhD, University of Southampton, Human Development and Health.

3- Andrew Collins; PhD, University of Southampton, Genetic Epidemiology and Bioinformatics Research Group, Human Development and Health.

4- Zaed Z R Hamady; PhD, FRCS, National Institute for Health Research Southampton Biomedical Research Centre, University of Southampton and University Hospital Southampton NHS Foundation Trust.

Corresponding Author:

Zaed Z R Hamady PhD, FRCS.

Consultant Hepato Pancreato Biliary surgeon

National Institute for Health Research Southampton Biomedical Research Centre,

University Hospital Southampton NHS Foundation Trust, Southampton, UK, SO16 6YD.

Tel: +44 238120 5284; E-mail: z.hamady@soton.ac.uk

Abbreviations:

AUC Area Under the Curve

BMI Body Mass Index

DM Diabetes Mellitus

GWAS Genome wide association studies

ICD International Classification of Disease

LD Linkage Disequilibrium

LSDM Long Standing Diabetes Mellitus

NODM New-Onset Diadetes Mellitus

PC Principal Components

PDAC Pancreatic Ductal Adenocarcinoma

PPV Positive Predictive Value

PRS Polygenic Risk Score

QC Quality Control

ROC Receiver Operator Characteristic

SNPs Single nucleotide polymorphisms

T2DM Type 2 Diabetes Mellitus

Disclosures:

Zaed Z R Hamady, partly funded by CRUK grant number C45617/A29908.

Author's contribution

All authors meet the criteria for authorship. They have read and approved the last version of the manuscript.

Design; ZH, SS, WT, AC,

Analysis; SS, WT, ZH,

Writing; WT, ZH, SS, AC

Word count: Abstract, 260 words ; Manuscript, 6517 words

Figures: 5**, Tables:** 2**, Supplementary figures:** 3**, Supplementary tables:** 5

Data Transparency Statement: Data used in this study is available via UK Biobank upon request.

Abstract:

Background and Aims:

Diabetes mellitus (DM) is known to be associated with Pancreatic ductal adenocarcinoma (PDAC), particularly, new-onset DM (NODM). Others have developed polygenic risk scores (PRS) associated with PDAC risk. We aimed to compare the performance of these PRS in an independent cohort to determine if they can discriminate between NODM and long standing DM (LSDM) patients with PDAC.

Methods:

Cases (1,042) and matched cancer free controls (10,420) were drawn from the UK Biobank. Five PRS models were calculated using single nucleotide polymorphisms (SNPs) from previous studies (Nakatochi, Galeotti, Molina, Jia and Rashkin) and a combination of these. Regression models were used to assess the association between PDAC and PRS adjusted for ancestry, smoking, DM, waist circumference, and a family history of digestive cancer. Receiver operator characteristic (ROC) curves and the area under the curve metrics (AUC) were used to assess the performance of each PRS for classifying PDAC risk.

Results:

The combined PRS model achieved the highest AUC (0.605), and significantly improved a clinical risk model in this cohort (AUC=0.83, *P* =0.0002). Individuals within the 5th quintile have a 2.74-fold increased risk of developing PDAC versus those in the 1st quintile (*P* <0.001), and have a 3.05-fold increased risk of developing PDAC if they have DM versus those without DM (*P* <0.001). The positive predictive value (PPV) was 11.9% in participants without DM, 23.9% with LSDM and 86.7% with NODM.

Conclusions:

The PDAC related common genetic variants are more strongly associated with DM. This PRS has the potential for targeting individuals with NODM for PDAC secondary screening measures.

Keywords: Early detection, Pancreatic Adenocarcinoma, Genetic risk score

Introduction:

Pancreatic Ductal Adenocarcinoma (PDAC) is considered to have the worst cancer survival outcome due to the lack of effective strategies directed at early detection [1]. The overwhelming majority of PDAC patients have locally advanced disease or distant metastasis at presentation and only a minority of patients are surgically resectable with curative intent [2]. Improvement of survival is heavily reliant on the development of innovative early detection and novel treatment strategies [3]. As treatment options for patients with resectable cancers continue to improve, including availability of multimodality neoadjuvant therapy [4], and more potent adjuvant regimens [5], a "stage shift" from the current 10-15% resectable proportion to 50% or greater will unequivocally lead to improved survival in this disease. The possibility of a population-wide pancreas screening programme is not viable at this point and unlikely to be available for the foreseeable future due to the relatively low prevalence of PDAC cases in the general population.

There is a list of symptoms recognised as suggestive of PDAC, such as weight loss, new-onset diabetes mellitus (NODM) in people older than 50 years [6] or non-specific gut symptoms [7]. These inconsistent symptoms mean public health awareness may not be as effective in aiding early detection. A strong family history has recently been identified as an indication for surveillance by the National Institute for Clinical Excellence (NICE) [2]. Whereas developments in genetic testing and screening programmes have reduced the burden and death toll of breast cancer, these techniques have not yet been well developed in PDAC [8,9]. Thus, efforts should be focused on developing robust tools for recognising individuals at a high-risk of PDAC, allowing improved potential for monitoring or secondary screening. To date, genetic research on factors associated with PDAC has been limited to rare monogenic, high penetrance genetic syndromes associated with this disease in high-risk families, whilst

more frequently occurring, but low penetrance, single nucleotide polymorphisms (SNPs) have been less well studied.

Genome wide association studies (GWAS) and meta analyses have identified new PDAC associated SNPs [10–13]. A method for translating GWAS results into a clinically useful tool is through the construction of polygenic risk scores (PRS) or risk models. A recent study [14] investigated a PRS for pancreatic cancer within the UK Biobank using 22 PDAC associated SNPs within 432 cases recorded by the national cancer registry. The PRS was predictive but did not consider all of the risk SNPs identified by previous GWAS and was not integrated with the phenotypic or lifestyle factors associated with PDAC. There is a lack of clinical risk model analysis on prospective cohorts based on genetic predisposition. In this study, we generate several PRS models using PDAC risk associated SNPs from three published risk models [14–16] a large meta-analysis [13] and one causal analysis [17] between PDAC and diabetes mellitus (DM) and assess their performance in a prospective cohort of patients, the UK Biobank. The most predictive model is integrated with epidemiological risk factors and its clinical utility is explored in patients with and without DM.

Materials and Methods:

Study Design and Population

The UK Biobank is an ongoing population-based cohort study which aims to improve the prevention, diagnosis and treatment of a wide range of diseases. Extensive genetic and clinical data have been collected for approximately 500K participants from across the UK, aged between 40 and 69 at the time of recruitment in 2006–2010. The design, data collection and processing are described in detail elsewhere [18,19].

Our analysis was restricted to PDAC patients and cancer free controls who were identified using the International Classification of Diseases codes (ICD version 10 and 9) that were recorded by the national cancer registry, hospital admissions and cause of death. Additional PDAC cases were identified using self-reported cancer and operative procedures. The PDAC disease ascertainment and the data fields used are illustrated in supplementary table 1. PDAC cases were considered prevalent if diagnosed before study entry and incident if diagnosed after study entry or without a date of diagnosis if identified by mortality alone. Since we are measuring clinical exposure in addition to genetic profile only incident cases were used for analysis. To prevent bias from analysing heterogeneous molecular subtypes, patients diagnosed with neuroendocrine tumours (8150=Islet cell carcinoma; 8246=Neuroendocrine carcinoma) were excluded.

To select age and sex matched cancer free controls, we used the MatchIt package in R [20] based on a nearest neighbour method with a 1:10 case control ratio in accordance with previous studies [21].

Patients with DM, were identified among cases and controls. Patients with DM were further stratified by type (1 or 2), assuming that participants with unknown type and onset ages greater than 35 were of type 2 diabetes mellitus (T2DM). Finally, particpants with T2DM

were split into cases of new onset DM (NODM) and long-standing DM (LSDM) using the following criteria: NODM if diagnosed within 24 months before or after diagnosis of PDAC cases or within 24 months of last follow up or death for controls; LSDM if diagnosed more than 24 months before diagnosis of PDAC or more than 24 months before last follow up or death for controls. A first-degree family history of either DM or digestive organ cancer was ascertained through a baseline study questionnaire and relevant ICD10 code.

Ethics and Consent

The UK Biobank has an ethical board managing any ethical concerns that may arise and is committed to ensuring high ethical standards throughout the project. All UK Biobank participants provided informed consent and were able to remove their personal data from the study at any point in time. Ethical approval for this study was obtained from the North West Multi-Centre Research Ethics Committee reference number 06/MRE08/65. The current research has been conducted using the UK Biobank Resource under Application Numbers 17749 and 35273.

Genotype Data and Quality Control (QC)

We used the imputed genotype data from the UK Biobank which are described in detail elsewhere[18, 19]. In brief, blood samples were genotyped in batches at the Affymetrix laboratories using either the UK BiLEVE array (807,411 SNPs) or UK Biobank axiom array (825,927 SNPs) which share 95% of SNPs and are therefore highly compatible. Additional SNPs (~9 million) were imputed using the Haplotype Reference Consortium (https://www.ncbi.nlm.nih.gov/pubmed/27548312/), the thousand genomes (https://www.ncbi.nlm.nih.gov/pubmed/23128226/) and the UK 10K (https://www.ncbi.nlm.nih.gov/pubmed/23128226/) projects. Routine quality control (QC) was performed by both Affymetrix and UK Biobank to remove poorly genotyped samples or

SNPs. The Biobank QC tested SNPs for batch effect, plate effect, array effect and gender effect. Departure from Hardy–Weinberg equilibrium (HWE) among controls was assessed through Pearson's chi-squared test. Samples were tested by the Biobank QC for genotype missingness, outlying heterozygosity levels that could not be attributed to admixture or consanguinity and discordance between self-reported sex and genotype inferred sex. Finally, samples for which consent had been withdrawn were also removed.

Construction and assessment of Polygenic Risk Scores

Five PRS were calculated using the SNPs from four previous publications which are hereafter referred to by the first author of each study [14–17]. The Nakatochi, Galeotti, Molina and Jia PRS models consist of 5, 30, 33 and 22 SNPs respectively which predispose to PDAC. The fifth PRS, referred to as combined, was generated by combining all of these SNPs with ten additional SNPs that were associated with pancreatic cancer in a large pan-cancer study [13]. After combining these data, SNPs in strong linkage disequilibrium (LD) were identified using a matrix of pairwise LD and recommended threshold for LD pruning ($r^2 \geq 0.8$) [22]. One SNP from each pair in strong LD with the weakest association with PDAC was removed (Supplementary table 2). More strict pruning, for example $r^2 \geq 0.5$, was not explored as it would remove informative SNPs from regions with multiple correlated signals that were included in the published PRS. A total of 49 out of 54 SNPs remained for construction of the combined PRS. The published effect size, p-value and risk allele for each SNP was sourced through the GWAS catalogue [23]. For SNPs present in two or more studies, the published effect size from the largest and therefore most powerful GWAS was used. The SNPs, risk allele and summary statistics used to create each PRS are shown in supplementary table 3 and the pairwise measures of LD are shown in supplementary table 2

The weighted PRS were generated using Plink [24], which applies the following formula:

$$PRS = \sum_{j=1}^{n} \ln(OR_j) * x_j$$

Where x is the number of risk alleles carried by an individual for SNP j which is weighted by the natural logarithm of the SNP's published effect size ln (ORj) and n is the number of SNPs in the model. PRS were standardised to a mean of zero and standard deviation (SD) of 1 by calculation of Z score (PRS-mean/SD).

Individual SNPs from the PRS were tested for replication in the UK Biobank using logistic regression, with disease status as the dependent variable and additively coded SNPs as the predictor along with covariates for smoking and the first ten genetic principal components (PC) to adjust for population stratification. Association tests were performed using Plink V1.9 [24]. The power to replicate each SNP in the UK Biobank was estimated using the genetic power calculator [25] according to the sample size (1042 cases versus 10420 controls) and the published effect sizes that were used to generate the PRS. The calculation assumed a multiplicative genetic risk model, a type 1 error rate of 0.05 and a disease prevalence of 15.4 per 100,000.

Statistical analysis

To identify relevant phenotypes that differed between the PDAC cases and cancer free controls we compared their baseline characteristics. Chi-squared tests were used to compare categorical features and a student *t*-test or Mann-Whitney test for continuous traits depending on their distribution.

The PRS were assessed in several ways. Density plots were generated using ggplot2 in R (3.6.3) to visualise and compare the distribution of each PRS in cases and controls. Their predictive performance was quantified using receiver operating characteristic curves (ROC) and the area under the curve (AUC) metric. Z-tests were used to compare AUCs between

diagnostic models using a paired design [26]. To investigate how the predicted risk of disease varied with increasing PRS, we performed quintile analyses where samples in the lowest PRS quintile were treated as a reference. Chi-squared tests were used to determine the odds of PDAC risk in each quintile versus the reference. In addition, Cox regression was used to determine if the cumulative hazard of developing PDAC varied between each PRS quintile versus the reference. In Cox regression, follow-up times were taken as the duration between age at study entry and diagnosis of PDAC. Control participants were censored at the date of last follow-up or death. Survival analyses were corrected for ethnicity (PC 1-10) and, in model 1, were also adjusted for smoking (never, current and previous), waist circumference (cm), DM onset (No DM, NODM, LSDM) and first-degree family history of digestive cancer (yes/no). To test for differences in association across DM onset, an interaction term was included for DM onset category (No DM, NODM, LSDM) and PRS (continuous).

For the clinical risk score, age of participants at recruitment, age when DM was diagnosed, DM onset (No DM, NODM, LSDM), waist circumference (cm), and first-degree family history of digestive cancer (yes/no) were included in the model. Age at DM diagnosis and DM onset were tested for collinearity defined as a variance inflation factor above 10. Inclusion of the PRS to the clinical score model was assessed using ROC and by comparing the AUC with those from PRS alone using a Z-test.

To test for differences in association with respect to DM, additional subgroup analyses were performed in participants with and without DM and with either NODM or LSDM, and were restricted to the most predictive combined PRS model. Cox regression analyses were performed as described above with the exception that DM status was not included in the list of covariates. The mean cumulative hazard ratios were plotted along with their 95%

confidence intervals in each subgroup (No DM, DM, NODM and LSDM) and stratified by the combined PRS quintile. Finally, overall associations with PDAC per standard deviation of the combined PRS were determined using logistic regression with adjustment for PC (1-10). Sensitivity, specificity and positive predictive values (PPV) associated with combined PRS at the 5^{th} quintile were calculated by cross tabulation of the PRS with case-control outcome and confidence intervals for proportions obtained.

Two-sided *P*-values less than 0.05 were regarded as statistically significant despite testing five PRS as these model are highly correlated due to using many of the same SNPs (Supplementary table 3). Statistical analyses were conducted using R software version 3.6.3. and SPSS v 27.0.

Results:

Sampling

At the time of analysis, 1,611 PDAC cases were identified with genotypic and phenotypic data which included 1,059 incident cases. Seventeen of the incident PDAC cases were diagnosed with neuroendocrine tumours and therefore excluded leaving 1,042 cases for analysis. A total of 325,379 unrelated individuals without any cancer diagnosis were used to generate the matched control cohort which consisted of 10,420 controls and an average standard mean difference of zero after matching for both age and sex (Supplementary figure 1).

Demographics

The mean age of PDAC cases and cancer free controls at recruitment was 61.3 years and their ethnic backgrounds were similar (*P* =0.816). The mean duration of follow up time from recruitment was 109 months (range 0-166 months). A first-degree family history of digestive organ cancer was reported in 15.4% of cases and 12.6% of controls (*P* =0.01). Smoking history was missing in 0.9% of cases which, as a whole, had a higher proportion of current smokers (15.8%) compared to controls (8.7%) (*P* <0.001). Similarly, DM was more frequently observed in PDAC cases (24.1%) than in controls (9.2%) (*P* <0.001). PDAC cases had a higher average body mass index (BMI) and waist circumference (measured at recruitment) compared to controls (*P* <0.05) although these differences were subtle and therefore unlikely to be clinically significant (Supplementary table 4).

Clinical risk model

As expected, we observed statistically significant associations between smoking, DM and PDAC risk (HR=2.2, 95% CI 1.845-2.64, *P* <0.001 for current smoking, and HR=2.66, 95% CI 2.35-3.06, *P* <0.001 for DM). The risk of PDAC varied depending on the age when DM was

diagnosed, participants diagnosed with DM at 60 years of age and older have a 1.33 times higher risk of developing PDAC compared to participants diagnosed with DM at younger than 60 years (P =0.027), and more than 3 times higher risk compared to participants with no DM (HR=3.32, 95% CI 2.80-3.93 P <0.001). Participants with a waist circumference of more than 100 cm and reported first-degree family history of digestive organs cancer have a higher risk of developing PDAC (HR=1.2, 95% CI 1.04-1.36 and HR=1.19, 95% CI 1.00-1.40 respectively P <0.05). Alcohol drinking status and first-degree family history of DM did not show a correlation with PDAC risk in the UK Biobank cohort (Supplementary Table 5).

The variance inflation factor for age DM diagnosed and DM onset was 1.083 with a tolerance of 0.919, suggesting that there was no collinearity between the two variables. Therefore, the clinical risk model included both age DM diagnosed and DM onset. The resulting clinical risk model achieved an AUC of 79.1% (95% CI 75.4-82.7) (Figure 1).

Polygenic risk scores

Five PRS were constructed to predict the risk of developing PDAC. Density plots of the resulting scores showed that for each PRS model there is a clear shift in the PRS distribution towards higher scores in the PDAC cases compared with the matched cancer free controls. We looked at the predictive accuracy of each PRS model and their association with PDAC risk using ROC and quintile analyses (Supplementary figure 2). The AUC observed in the combined PRS model 60.5% (95% CI 58.7-62.3) was significantly higher than the Nakatochi PRS (P =8.9x10^{-8}) and showed a trend towards superiority compared to the other PRS models (Galeotti, Molina and Jia) (P >0.05). The odds of having PDAC for participants in the 5th quintile in the combined PRS was 2.9 (95% CI 2.34-3.59) compared to the first quintile (P <0.001) and showed a trend of superiority compared to the other PRS models (Nakatochi OR 2.2, 95% CI 1.76-2.75, Galeotti OR 2.59, 95% CI 2.05-3.21, Molina OR 2.54, 95% CI 2.05-3.14

and Jia OR 2.7, 95% CI 2.16-3.37) (Supplementary figure 2). Similarly, when comparing the highest and lowest PRS quintiles using Cox regression, hazard ratios for the Nakotochi model indicated a 2 fold increased risk of PDAC (HR 2.09, 95% CI 1.69-2.60), 2.5 fold higher utilising the Galeotti and Molina models (HR 2.53, 95% CI 2.06-3.11 and HR 2.47, 95% CI 2.02-3.04 respectively), and 2.7 fold higher (HR 2.71 ;95% CI 2.19-3.35) for the Jia PRS model. The combined PRS model was associated with a 2.8 fold increased hazard of developing PDAC in the 5th quintile compared to the 1st (HR 2.83 95%CI 2.31-3.48). Similar results were generated in the multivariable analyses, which is adjusted for smoking, waist circumference, DM and first-degree family history of digestive cancer in addition to the first 10 PC (Table 1). The combined PRS continued to have a trend towards higher association with PDAC risk in the top versus bottom quintile (HR 2.74; 95% CI 2.23-3.37, *P* <0.001). We found that associations with risk for PDAC per SD of PRS were significant among participants with DM and stronger than participants with No DM (*P* for interaction 0.004). Adjusted survival plots, stratified by PRS quintiles are displayed in supplementary figure 3 and are consistent with the hypothesised PRS-related probability gradient across the full age range. Individuals in the 5th quintile expressed a 5% PDAC risk at the age of 65 versus 1.7% PDAC risk for individuals in the 1st quintile.

The addition of the combined PRS to the clinical risk model significantly improves the discrimination ability of the model to an AUC of 83% (95% CI 80-86) (*P* =0.0002) (Figure 1).

Subgroup analysis

Both cases with DM and without DM showed marked skewing towards higher PRS values compared with controls (data not shown). Interestingly, we found that association with risk for PDAC per SD of PRS was significant among participants without DM (OR=1.39 95% CI 1.29-1.48), with DM (OR=1.67 95% CI 1.44-1.94), with LSDM (OR=1.87, 95% CI 1.53-2.29)

and with NODM (OR 1.89 95%CI 1.28-2.78) (Figure 2). However, the highest AUC was observed in participants with DM at 64.5% (95% CI 60.9-68.2) compared to particpants without DM at 59.4% (95% CI 57.3-61.4) (P <0.05) (Figure 3). This suggests that the PRS model is more predictive of PDAC risk in participants with DM compared to participants without DM. Survival plots from Cox-regression, stratified by DM status, are displayed in Figure 4. The curves demonstrate higher PDAC prediction by the combined PRS model at the 5th quintile in participants with DM, compared to those without DM (HR 3.05 95% 2.39-3.91, P <0.001). For individuals in 5th quintile of the combined PRS**,** the positive predictive value (PPV) was 14.4% (95% CI 13-15.9) for the whole cohort, 11.9% (95% CI 10.5-13.4) in participants without DM, 23.9% (95% CI 18.1-30.3) in participants with LSDM and 86.7% (95% CI 73.2-94.9) in participants with NODM. Associated sensitivities and specificities with correspondent subgroups are shown in Table 2.

The specific evaluation of whether these effects differ by age was underpowered due to limited numbers of cases with NODM in this cohort. When stratified by age of diagnosis of DM, there was no substantial difference in the relationship of PRS with PDAC risk (OR 1.63 95% CI 1.292-2.064; OR 1.54 95% CI 1.287-1.837) per 1 SD for participants with DM diagnosed at less than 60 and 60 years or more respectively. A similar association is found when the cohort was divided by gender and waist circumference (Figure 2).

The predicted PDAC risk after adjusting for PC 1-10, smoking, waist circumference and first degree family history of digestive organ cancer, was 29.6% (95% CI 27-32) for participants with DM and 12.4% (95% CI 12-12.7) for participants without DM (P <0.0001) when their combined PRS is within the 5th quintile. Participants with LSDM and NODM were estimated to have similar PDAC risk but the statistical power of this comparison is limited by the small sample size of controls with NODM (Figure 5).

Discussion:

Our study included PDAC cases and matched cancer free controls from the UK Biobank cohort. A previously published PRS model derived from an Asian cohort [15], was less strongly associated with PDAC incidence in this UK based cohort compared with those derived from European cohorts [14,16,17] that were replicated with similar levels of accuracy. The PDAC risk associated SNPs from all of these models were then used to construct a novel combined PRS model for PDAC development with the main focus centred on clinical risk factors. This combined model produced similar or higher discrimination between PDAC cases and controls compared to the PRS from individual studies. We demonstrated that a PRS derived from common genetic variants alone could successfully identify participants at increased risk of PDAC, particularly among individuals with DM. In addition, the inclusion of the combined PRS into a risk model consisting of traditional clincal features improved the discrimination ability of the clinical risk model. The PPV for the combined PRS to predict PDAC was higher in participants with NODM, than with LSDM and without DM. Furthermore, the predictiveness of the combined PRS was related to DM status, with improved performance in participants with DM.

GWAS have emerged as a powerful, hypothesis-free approach to identify common alleles that influence disease risk. In recent years a number of SNPs convincingly associated with PDAC risk have been reported [27,28]. The development of PRS to evaluate the overall predictive power of common risk loci for PDAC has previously been carried out [15–17], however, to date no study has looked at the discriminatory ability of polymorphisms on various types of DM associated with PDAC. Three studies examined the association of susceptibility variants between T2DM and PDAC [29–31], without finding any significantly associated T2DM-related variant with PDAC risk. In the Molina-Montes [17] study, they have

elucidated that LSDM is not causally linked to PDAC, whereas PDAC may cause NODM, if the influencing effects of body weight are ruled out. A caveat to using the same SNPs and effect sizes from these studies is that they were determined from GWAS that were not specific to DM. An adequately powered GWAS analysis specific to PDAC with DM is yet to be performed. Although our PRS positively identifies those at heightened risk of PDAC with DM and less predictive of PDAC without DM, there is still room for improving its discriminatory accuracy. Furthermore, these results suggest possible functional inter-relationships between inherited variation in genes important for pancreatic development, diabetes and PDAC risk.

Our findings on the observational association between DM and PDAC risk are similar to previous studies which demonstrated a doubling of PDAC risk in people with T2DM [32,33]. In concordance with others [17,34,35], NODM was highly associated with PDAC. Although we have noted that the number of NODM cases in the control group is low, this finding was also reported by Molina-Montes [17]. A possible explanation for the low incidence of NODM cases is that individuals in the control group may have undiagnosed diabetes. In UK biobank this may correspond to development of DM after recruitment which passed unnoticed despite linkage to national Hospital Episode Data (HES).

Our study highlights the potential utility of a PRS in PDAC risk stratification in the general population and particularly in people with DM. This may facilitate early cancer detection in this population which currently lacks consistent recommendations for early detection. The benefits of this study lie in utilising PRS to enrich PDAC prevalence in people with DM and make further clinical imaging investigations cost effective. This will need to be proven in a prospective clinical study with a specific aim at PDAC diagnosis.The PRS discrimination between cancer and controls alone is limited and unlikely to be usable clinically on its own.

The strengths of this study are based around the dataset. The UK Biobank is a prospective cohort of patients which reduces the risk of confounding bias. The blood samples and genotyping has been produced prior to incident PDAC. The combined model allowed the largest number of SNPs in a PDAC risk model to be tested. The results at the highest and lowest risk ends were significant which will be the areas of interest most useful for a clinical model.

There are limitations with this study: the UK biobank cannot be considered a representation of the UK population as the participants are a health-conscious, educated cohort of individuals. In addition, the UK Biobank cohort enrolled participants older than 40 years of age, therefore, a younger population is not represented in this cohort. However, PDAC is very rare incident in people under the age of 40 [36]. Overall, the UK Biobank mainly represents European ancestry, meaning further studies into the effect of these SNPs on other ancestries is necessary. Subgroup analysis related to anatomical position of the cancer in relation to pancreas was not possible due to limited number and accuracy of coding in relation to cancer position. Weight loss concurrent with diagnosis of DM has been suggested as a clinical marker for PDAC associated DM, unfortunately this piece of data was not recorded by the UK Biobank cohort.

Further to this study it would be useful to identify and include additional PDAC risk SNPs in the model. Therefore, further studies are needed in order to reach similar numbers of risk associated SNPs as other cancers, where models have been constructed from hundreds of risk associated SNPs [9,37], improving the potential for stronger significance and AUC figures. Additional SNPs could be found by performing a GWAS for PDAC risk related SNPs on UK Biobank data. This can also test whether the results presented in this paper are replicated and provide an opportunity to examine SNPs related to DM to see if they are able to

discriminate between *de novo* and pancreatogenic DM. Further analysis of UK Biobank blood samples to identify other omic factors that may predict PDAC is highly recommended.

Acknowledgments

The authors are thankful to the participants and the team of the UK Biobank study.

Figures Legends:

Figure 1: Receiver operating characteristic (ROC) curves and area under the curve (AUC) metrics for overall accuracy of the clinical risk (CR) model *; solid gray line, compared to CR model after addition of the combined PRS to the model; solid black line (*P* =0.00019). CI; Confidence Interval, PRS; polygenic risk score.

* Model included; age of participants at recruitment, age when DM diagnosed, DM onset (No DM, NODM, LSDM), waist circumference (cm), and first-degree family history of digestive cancer (yes/no).

Figure 2: Subgroup analysis, forest plot shows association between standardised PRS and PDAC risk. Logistic regression for case control status against standardised PRS adjusted for principal components (PC 1-10) within sample subsets defined by age, gender, waist circumference and diabetes mellitus (DM), LSDM; Long-standing DM, NODM; New-Onset DM, DM<60; DM diagnosed at less than 60 years, DM ≥ 60; DM diagnosed at or more than 60 years. Data shown odds ratio (OR) and 95% Confidence Interval (CI).

Figure 3: Receiver operating characteristic (ROC) curves and area under the curve (AUC) metrics for overall accuracy of PDAC prediction by combined PRS in diabetes mellitus (DM) subgroups. Individuals without DM; sold gray line, Individuals with DM; solid black line. CI; Confidence Interval.

Figure 4: Absolute risk estimates for PDAC diagnosis by combined PRS quintile among individuals in the UK Biobank Cohort adjusted for principal components (PC 1-10), smoking, waist circumference and first degree family history of digestive cancer, stratified by diabetes mellitus (DM) status (A) Individuals without DM (n=10251); (B) Individuals with DM (n=1211). Hazard ratio (HR) 3.05 (95% CI 2.394-3.906) comparing participitants in 5^{th} quintile with and without DM (*P* = 3.63 E-19).

Figure 5: Adjusted PDAC risk as obtained by cox regression hazard estimates in UK Biobank participants by combined PRS quintiles, stratified by diabetes mellitus (DM) status (yes/no) (A) and onset of DM (B) prior to censor time. Risk estimates are means with 95% Confidence intervals. The combined PRS is more predictive of PDAC in participants with DM. LSDM; long-standing DM, NODM; New onset DM. Models were adjusted for principle components (PC1-10), smoking, waist circumference and first degree family history of digestive organ cancer.

Supplementary Figure 1. Selection of cases and cancer free controls from the UK Biobank cohort.

Supplementary Figure 2. PRS distribution and performance according to Receiver operating characteristic (ROC) and quintile comparisons. (A) Density plots showing distribution of standardised PRS among PDAC cases and cancer free controls. (B) ROC curves and area under the curve (AUC) metrics for overall accuracy of PDAC prediction by each PRS. (C) Odds ratio (OR) estimates: Quintile plots showing PDAC risk in each quintile versus 1st quintile.

Supplementary figure 3. Risk estimates for PDAC diagnosis by PRS quintile.

Cox regression adjusted for principal components (PC 1-10), smoking, waist circumference, diabetes mellitus status, first degree family history of diabetes, first degree family history of digestive cancer among individuals in quintile 1 to 5 for each PRS (Reference=1). Censored at date of death or last register follow up (01/02/2018). Time = duration between age at study entry to PDAC diagnosis or, for censored participants, date of death or last follow up in register.

Table Legends:

Table 1: Association between genotype scores and risk of PDAC in UK Biobank cohort (case control 1:10 matched for age and sex). Polygenic risk scores (PRS) were calculated using the SNPs from four previous publications which referred to by the first author of each study The (Nakatochi, Galeotti, Molina and Jia) [14–17]. Combined PRS was generated by combining all of the SNPs from aformentioed studies with additional SNPs that were associated with pancreatic cancer in Rashkin meta-analysis [13].

Hazard ratio (HR): PDAC risk among individuals in quintile X versus 1, adjusted for principal components (PC 1-10)

* model 1; adjusted for smoking (Categorical) Waist Circumference (cont.), DM (No/LSDM/NODM), family history of digestive cancer (Yes/No).

** p value for interaction produced by adding interaction term between PRS (continuous) and DM (No DM/LSDM/NODM) in model 1. DM; Diabetes Mellitus; NODM New-onset diabetes mellitus, LSDM long-standing diabetes mellitus.

Table 2: Associated sensitivity/specificity and positive predictive value (PPV) with 95% Confidence Intervals (CI) for the prediction of PDAC in UK biobank cohort with subgroups of DM. the value associated with participants within 5th quintile of combined PRS . DM; Diabetes Mellitus; NODM New-onset diabetes mellitus, LSDM long-standing diabetes mellitus.

Supplementary table 1: Data fields and ICD (International Classification of Disease) Codes used for identification of PDAC and diabetes in the UK Biobank cohort.

Supplementary table 2: SNPs removed due to Linkage Disequilibrium (LD).

~~The~~ threshold for linkage disequilibrium was $r^2 \geq 0.8$. All SNPs on the left were kept.

Supplementary table 3: SNPs used to create PRS, * SNP with risk Allele difference

Supplementary table 4: Baseline general characteristics of the study population, PDAC cases and cancer free controls (Age/Sex matched).

PDAC; Pancreatic ductal adenocarcinoma, DM; Diabetes Mellitus; NODM; New-onset diabetes mellitus, LSDM; long-standing diabetes mellitus

DM status:

Type 2: includes diabetes of unknown type (n=66, 7 cases and 59 controls)

NODM: Defined in cases as type 2 diabetes diagnosed within 24 months before or after diagnosis of PDAC. Defined in controls as type 2 diabetes diagnosed 24 months before death or last follow up.

LSDM: Defined in cases as type 2 diabetes diagnosed more than 24 months before PDAC diagnosis. Defined in controls as type 2 diabetes diagnosed more than 24 months before date of death or date of last follow up.

Supplementary table 5: Univariable cox regression, association between phenotype variables and PDAC risk in UK Biobank cohort with 1:10 Case-control (matched for age and sex). DM; Diabetes Mellitus; NODM, New-onset diabetes mellitus, LSDM, long-standing diabetes mellitus.

Refrences:

1. Rahib L, Smith BD, Aizenberg R, et al. Projecting cancer incidence and deaths to 2030: The unexpected burden of thyroid, liver, and pancreas cancers in the united states. Cancer Res 2014.
2. O'Reilly D, Fou L, Hasler E, et al. Diagnosis and management of pancreatic cancer in adults: A summary of guidelines from the UK National Institute for Health and Care Excellence. Pancreatology 2018.
3. Singhi AD, Koay EJ, Chari ST, et al. Early Detection of Pancreatic Cancer: Opportunities and Challenges. Gastroenterology 2019.
4. Rangarajan K, Pucher PH, Armstrong T, et al. Systemic neoadjuvant chemotherapy in modern pancreatic cancer treatment: A systematic review and meta-analysis. Ann R Coll Surg Engl 2019.
5. Khorana AA, Shapiro M, Mangu PB, et al. Potentially curable pancreatic cancer: American society of clinical oncology clinical practice guideline update. J Clin Oncol 2017.
6. Hart PA, Kamada P, Rabe KG, et al. Weight loss precedes cancer-specific symptoms in pancreatic cancer-associated diabetes mellitus. Pancreas 2011.
7. Mills K, Birt L, Emery JD, et al. Understanding symptom appraisal and help-seeking in people with symptoms suggestive of pancreatic cancer: A qualitative study. BMJ Open 2017;7.
8. Ferlay J, Partensky C, Bray F. More deaths from pancreatic cancer than breast cancer in the EU by 2017. Acta Oncol (Madr) 2016.
9. Lakeman IMM, Rodríguez-Girondo M, Lee A, et al. Validation of the BOADICEA model and a 313-variant polygenic risk score for breast cancer risk prediction in a Dutch prospective cohort. Genet Med 2020;22:1803–1811.
10. Lin Y, Nakatochi M, Hosono Y, et al. Genome-wide association meta-analysis identifies GP2 gene risk variants for pancreatic cancer. Nat Commun 2020.
11. Klein AP, Wolpin BM, Risch HA, et al. Genome-wide meta-analysis identifies five new susceptibility loci for pancreatic cancer. Nat Commun 2018.
12. Campa D, Rizzato C, Capurso G, et al. Genetic susceptibility to pancreatic cancer and its functional characterisation: The PANcreatic Disease ReseArch (PANDoRA) consortium. Dig Liver Dis 2013.
13. Rashkin SR, Graff RE, Kachuri L, et al. Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts. Nat Commun 2020;11:4423.

14. Jia G, Lu Y, Wen W, et al. Evaluating the Utility of Polygenic Risk Scores in Identifying High-Risk Individuals for Eight Common Cancers. JNCI Cancer Spectr 2020.

15. Nakatochi M, Lin Y, Ito H, et al. Prediction model for pancreatic cancer risk in the general Japanese population. PLoS One 2018.

16. Galeotti AA, Gentiluomo M, Rizzato C, et al. Polygenic and multifactorial scores for pancreatic ductal adenocarcinoma risk prediction. J Med Genet 2020.

17. Molina-Montes E, Coscia C, Gómez-Rubio P, et al. Deciphering the complex interplay between pancreatic cancer, diabetes mellitus subtypes and obesity/BMI through causal inference and mediation analyses. Gut 2020.

18. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med 2015.

19. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature 2018;562:203–209.

20. Ho DE, Imai K, King G, et al. MatchIt: Nonparametric preprocessing for parametric causal inference. J Stat Softw 2011;42.

21. Fritsche LG, Patil S, Beesley LJ, et al. Cancer PRSweb: An Online Repository with Polygenic Risk Scores for Major Cancer Traits and Their Evaluation in Two Independent Biobanks. Am J Hum Genet 2020.

22. Euesden J, Lewis CM, O'Reilly PF. PRSice: Polygenic Risk Score software. Bioinformatics 2015;31:1466–1468.

23. Welter D, MacArthur J, Morales J, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 2014.

24. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–575.

25. Purcell S, Cherny SS, Sham PC. Genetic power calculator: Design of linkage and association genetic mapping studies of complex traits. Bioinformatics 2003;19.

26. Zhou X-H, McClish DK, Obuchowski NA. *Statistical methods in diagnostic medicine*. John Wiley & Sons; 2009.

27. Petersen GM, Amundadottir L, Fuchs CS, et al. A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. Nat Genet 2010.

28. Amundadottir L, Kraft P, Stolzenberg-Solomon RZ, et al. Genome-wide association

study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. Nat Genet 2009.

29. Wu L, Rabe KG, Petersen GM. Do variants associated with susceptibility to pancreatic cancer and type 2 diabetes reciprocally affect risk? PLoS One 2015.

30. Pierce BL, Austin MA, Ahsan H. Association study of type 2 diabetes genetic susceptibility variants and risk of pancreatic cancer: An analysis of PanScan-I data. Cancer Causes Control 2011.

31. Tang H, Wei P, Duell EJ, et al. Genes-environment interactions in obesity- and diabetes-associated pancreatic cancer: A GWAS data analysis. Cancer Epidemiol Biomarkers Prev 2014.

32. Batabyal P, Hoorn S Vander, Christophi C, et al. Association of diabetes mellitus and pancreatic adenocarcinoma: A meta-analysis of 88 studies. Ann Surg Oncol 2014.

33. Ben Q, Xu M, Ning X, et al. Diabetes mellitus and risk of pancreatic cancer: A meta-analysis of cohort studies. Eur J Cancer 2011.

34. Sharma A, Kandlakunta H, Nagpal SJS, et al. Model to Determine Risk of Pancreatic Cancer in Patients With New-Onset Diabetes. Gastroenterology 2018.

35. Pannala R, Basu A, Petersen GM, et al. New-onset diabetes: a potential clue to the early diagnosis of pancreatic cancer. Lancet Oncol 2009.

36. Collaborators GBD 2017 PC. The global, regional, and national burden of pancreatic cancer and its attributable risk factors in 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. lancet Gastroenterol Hepatol 2019;4:934–947.

37. Mavaddat N, Michailidou K, Dennis J, et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. Am J Hum Genet 2019;104:21–34.

Author names in bold designate shared co-first authorship