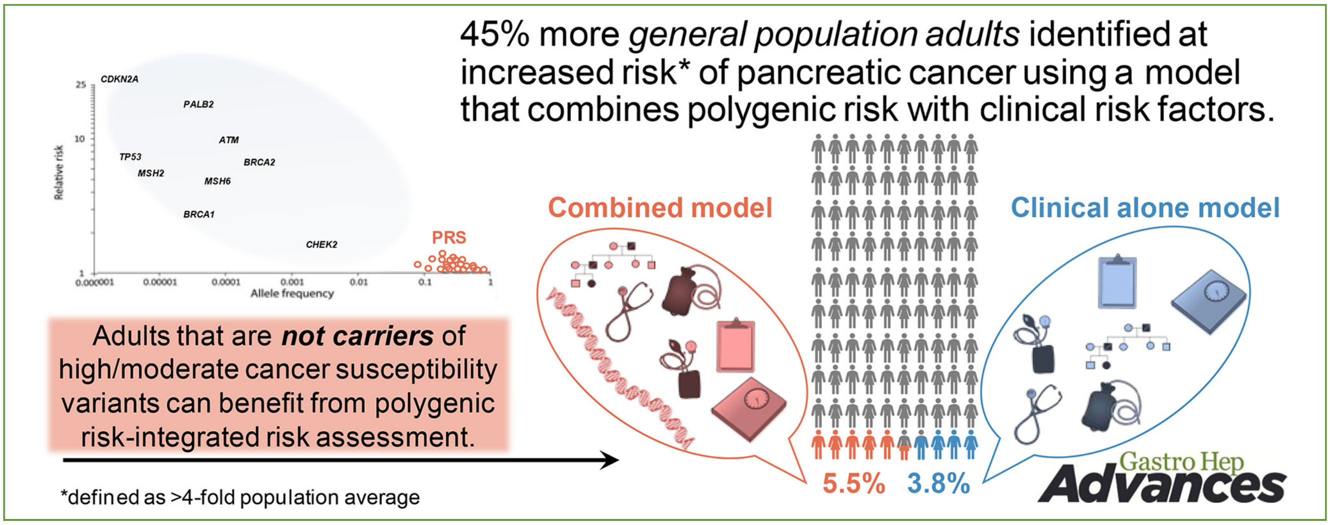
ORIGINAL RESEARCH—CLINICAL

Predicting 10-Year Risk of Pancreatic Cancer Using a Combined Genetic and Clinical Model

Gillian S. Dite,[1] Erika Spaeth,[2] Chi Kuen Wong,[1] Nicholas M. Murphy,[1] and Richard Allman[1]

[1]Genetic Technologies Limited, Fitzroy, Victoria, Australia; and [2]Phenogen Sciences Inc, Charlotte, North Carolina



45% more *general population adults* identified at increased risk* of pancreatic cancer using a model that combines polygenic risk with clinical risk factors.

Adults that are ***not carriers*** of high/moderate cancer susceptibility variants can benefit from polygenic risk-integrated risk assessment.

*defined as >4-fold population average

BACKGROUND AND AIMS: Pancreatic cancer has the poorest 5-year survival rate of any major solid tumor, but when diagnosed at an early stage, survival rates improve. Population screening is impractical because pancreatic cancer is rare with a lifetime risk of 1.7%, but accurate risk stratification in the general population could enable health care providers to focus early detection strategies to at-risk individuals. Here, we validate a combined risk prediction model that integrates a polygenic risk score and a clinical risk model. **METHODS:** Using the UK Biobank, we conducted a prospective cohort study assessing 10-year pancreatic cancer risks based on a polygenic risk score, a clinical risk score, and a combined risk score. We assessed the association, discrimination, calibration, cumulative hazards, and standardized incidence ratios compared to population incidence rates for the risk scores. We also conducted net reclassification analyses. **RESULTS:** While all of the risk scores discriminated well between affected and unaffected participants, the combined risk score – with a Harrell's C-index of 0.714 (95% confidence interval [CI] = 0.698, 0.730) – discriminated better than both the polygenic risk score ($P = .001$) and the clinical risk score ($P = .02$). In terms of calibration, there was no problem with dispersion for the combined risk score ($\beta = 0.952$, 95% CI = 0.865–1.039, $P = .3$) and overall there was a small overestimation of risk ($\alpha = -0.089$, 95% CI = -0.156 to -0.021, $P = .009$). Participants in the top decile of 10-year risk were at 1.413 (95% CI = 1.242–1.607) times population risk. **CONCLUSION:** The combined risk score was able to identify individuals at substantially increased risk of pancreatic cancer and to whom targeted screening could be useful.

Keywords: Pancreatic Cancer; Risk Prediction; Clinical Risk; Polygenic Risk Score

Introduction

Pancreatic cancer has the poorest 5-year survival rate of any major solid tumor, and incidence rates have been on the rise over the last 20 years.[1] When diagnosed at an early stage, survival rates improve, highlighting the importance of screening and early detection. Population screening is not practical due to the low lifetime risk of pancreatic cancer, which is about 1.7% for adults in the Western world. Despite this, accurate risk stratification in the general population could enable health care providers to focus early detection strategies to at-risk individuals.

Abbreviations used in this paper: **AUC, area under the receiver operating characteristic curve; BMI, body mass index; CI, confidence interval; E, expected; NRI, net reclassification improvement; O, observed; OR, odds ratio; PRS, polygenic risk score; SD, standard deviation; SIR, standardized incidence ratio.**

Most current article

2772-5723
https://doi.org/10.1016/j.gastha.2023.05.008

Like many other solid tumor types, some pancreatic cancers are associated with known highly and moderately penetrant germline (likely) pathogenic variants; these may represent between 4%–20% of pancreatic cancer cases.[2,3] However, based on population-based studies of genome-wide loci, the overall heritability of pancreatic cancer lies somewhere between 21%–36%.[3,4] Polygenic risk has been identified as a component of this heritability and has been investigated across multiple ethnic populations.[5–9] While the heritability of pancreatic cancer is important, the majority of pancreatic cancer cases are sporadic and the risk factors contributing to somatic events are largely associated with environmental and clinical risk factors. Several diseases have been found to be associated with an increased risk of developing pancreatic cancer; several forms of pancreatitis, primary sclerosing cholangitis, type 2 diabetes, and hepatitis B viral infection are all associated with or considered a preclinical manifestation of pancreatic cancer.[10–13]

Several clinical risk models have been created in an attempt to stratify the general population of individuals with European ancestry to strengthen the case for screening at-risk individuals. These vary in discriminatory ability with the area under the receiver operating characteritic curve (AUC) ranging from 0.61 for the model by Klein et al,[14] 0.68 for the model by Appelbaum et al,[15] 0.72 for the model by Kachuri et al,[6] and 0.86 for the model by Hippisley-Cox et al.[16] Assessment of the improvement in discrimination from adding a polygenic risk score (PRS) was undertaken by Kachuri et al, with the AUC increasing to 0.74.[6]

In a disease where rapid progression from early to advanced stage occurs within about 14 months,[17] early detection of pancreatic cancer can significantly affect treatment options and survival rates. The application of a risk stratification tool in the general population will allow for selective screening of the most at-risk individuals. Herein, we validate a combined risk prediction model that integrates a PRS developed by Jia et al[5] and a clinical risk model developed by Klein et al[14] to identify asymptomatic at-risk individuals who could benefit from screening for pancreatic cancer.

Materials and Methods

UK Biobank

The UK Biobank is a cohort of over 500,000 participants who were recruited from England, Wales, and Scotland from 2006 to 2009.[18–20] Phenotypic data was collected from participants at a baseline assessment that involved completing a touchscreen questionnaire followed by a face-to-face interview, physical measurements, and biological sample collection with a nurse. The blood, urine, and saliva samples have been analyzed to provide extensive genomic and biomarker data.[18–20] All participants have been linked to cancer registries, death registries, and hospital records, while around 45% of participants have been linked to primary care data.[21] Despite the presence of a healthy volunteer selection bias in the UK Biobank, the size of the resource and the variation in exposure measures means that it is possible to generalize the results of analyses of exposures and disease outcomes to other populations.[22]

Eligibility

UK Biobank participants who had not withdrawn their consent on 22 February 2022, whose gender identity was the same as their sex at birth, who were aged 40–69 years at their baseline assessment date, and who were of European ancestry were initially eligible for this study. Participants were then excluded from all analyses if they had pancreatic cancer diagnosed before their baseline assessment date, had no single-nucleotide polymorphism (SNP) data available, or had died, or been diagnosed with pancreatic cancer within the first 6 weeks of follow-up. To ensure that the dataset did not have any pairs of participants with closer than third-degree relatedness, we used the ukb_gen_samples_to_remove function of the R package ukbtools[23,24] to identify related pairs and randomly chose one to be removed. Details of the eligibility criteria and number of participants eligible and dropped at each step are in Figure A1.

Data extraction

We used UK Biobank data field 21003 to determine age at baseline assessment, UK Biobank data field 22006 to determine European ancestry, and UK Biobank data fields 31 and 22001 to determine sex. A diagnosis of pancreatic cancer was based on self-reported data (UK Biobank data field 20001 equal to 1026) and linked cancer registry data (the first 3 characters of UK Biobank data field 40006 [International Classification of Diseases, Tenth Revision equal to C25 or the first 3 digits of UK Biobank data field 40006 [International Classification of Diseases, Ninth Revision equal to 157). Incident pancreatic cancers were those for which the date of diagnosis (UK Biobank data field 40005) was after the baseline assessment date (UK Biobank data field 53), while prevalent pancreatic cancers were those diagnosed on or before the baseline assessment date (UK Biobank data fields 20007 or 40005 less than or equal to baseline assessment date). We used linked death registry data (UK Biobank data fields 40000 [date of death] and 40007 [age at death]) to identify participants who had died during the follow-up period.

We used Plink version 1.9.[25,26] to extract the Jia et al[5] panel of 22 SNPs for pancreatic cancer from the UK Biobank's SNP imputation dataset. For the variables in the Klein et al[14] clinical model, body mass index (BMI) was taken from UK Biobank data field 21001 (or 23104 if 21001 was not available), smoking status was taken from UK Biobank data field 20116 and ABO genotype was taken from UK Biobank data field 23165. Diabetes of over 3 years duration was identified if one or more of the UK Biobank data fields for diabetes outcomes (130706, 130708, 130710, 130712, and 130714) had a valid date 3 or more years before the baseline assessment date. For heavy alcohol use (>3 drinks per day), UK Biobank data field 1558 was used to identify participants who reported less than daily drinking (and were therefore, not heavy alcohol users), and UK Biobank data fields 1568, 1578, 1588, 1598, and 1608 were used for weekly red wine, champagne, beer/cider, spirits, and fortified wine intake, respectively. Participants with a total over the 5 categories of alcohol of more than 21 drinks per week were heavy alcohol users, while those with a total of 21 drinks or fewer per week were not.

Relative PRS

We used estimates of the odds ratio (OR) per effect allele and effect allele frequency from Jia et al[5] (see Table A1) to calculate a population-standardized PRS (as a relative risk) for each participant.[27,28] For each SNP in the panel of 22 SNPs for pancreatic cancer in Jia et al[5] (see Table A1), the unscaled population average risk was calculated as $\mu = (1-p)^2 + 2p(1-p)OR + p^2OR^2$, where OR was the odds ratio per effect allele and p was the effect allele frequency. For each participant, the adjusted risk (which has a population mean equal to 1) for each of the 22 SNPs was calculated as $\frac{OR^N}{\mu}$, where N was the participant's number of effect alleles for the SNP. Any participants who were missing genotype data for 1 or more SNPs were given an adjusted risk of 1 for each missing SNP. The final PRS (*snprisk*) for each participant was the product of their adjusted risks for the 22 SNPs.

Clinical relative risk score

The clinical relative risk score from Klein et al[14] (*clinrisk*) is calculated as:

$$\begin{aligned} clin\_xb = {} & 0.199 \times smoke\_2 + 0.788 \times smoke\_3 + 0.482 \\ & \times diabetes + 0.470 \times fh + 0.372 \times alcohol \\ & - 0.094 \times bmi\_1 + 0.077 \times bmi\_3 + 0.231 \times bmi\_4 \\ & + 0.207 \times abo\_2 + 0.399 \times abo\_3 + 0.300 \times abo\_4 \\ & + 0.457 \times abo\_5 + 0.365 \times abo\_6 + 0.255 \times rs3790844 \\ & + 0.166 \times rs401681 + 0.239 \times rs9543325 \end{aligned}$$

$$clinrisk = e^{clin\_xb}$$

The indicator variables (0 = no; 1 = yes) were as follows: *smoke_2* for being a former smoker, *smoke_3* for being a current smoker, *diabetes* for having had a diagnosis of diabetes for longer than 3 years, *fh* for having a first-degree family history of pancreatic cancer, *alcohol* for heavy alcohol use (>3 drinks per day), *bmi_1* for having a BMI < 18.5 kg/m^2, *bmi_3* for having a BMI ≥ 25 and < 30 kg/m^2, *bmi_4* for having a BMI ≥ 30 kg/m^2, *abo_2* for having AO as the ABO genotype, *abo_3* for having AA, *abo_4* for having BO, *abo_5* for having BB, and *abo_6* for having AB. The SNPs were coded as the number of effect alleles (0, 1, or 2) as follows: A was the effect allele for rs3790844, T was the effect allele for rs401681, and C was the effect allele for rs9543325.

Family history of pancreatic cancer was not available in the UK Biobank and, therefore, was omitted from the clinical relative risk calculation in this study. If a participant was missing information on a clinical risk factor, we omitted that risk factor from their relative risk calculation. The clinical relative risk score in this study was centered by dividing by its mean in the analysis dataset:

$$clinrisk\_c = \frac{clinrisk}{3.420848}$$

Combined relative risk score

For the calculation of the combined relative risk score in our study, 3 SNPs (rs3790844, rs401681, and rs9543325) were not included in the calculation of the clinical relative risk score because they also appear in the relative PRS from Jia et al.[5] The abridged clinical risk score was centered by dividing by its mean:

$$abr\_clinrisk\_c = \frac{abr\_clinrisk}{1.628917}$$

One SNP, rs505922, is used in both the relative PRS and the determination of the ABO genotype for the clinical relative risk score; therefore, rs505922 was omitted from the calculation of the relative PRS (*abr_snprisk*) for the combined relative risk score. The centered and abridged clinical relative risk score was then multiplied by the abridged relative PRS to obtain the combined relative risk score:

$$combined\_risk = abr\_snprisk \times abr\_clinrisk\_c$$

Absolute 10-year risk scores

We calculated absolute 10-year risk for each of the risk scores (PRS, clinical risk score, and combined risk score) using sex-specific, age-specific (in 5-year groups), and calendar year-specific pancreatic cancer incidence rates for England[29] as population reference rates. We also used a competing mortality adjustment using sex-specific, age-specific (in 5-year groups), and calendar year-specific mortality rates for causes of death other than pancreatic cancer.[30,31]

For the calculation of 10-year risk for each participant (aged b years at their baseline assessment date), we first determined their population incidence from birth to age b years (*popincid*) and to age $b + 10$ years (*popincid10*). For each relative risk score (*riskscore*), we then calculated each participant's cumulative risk to age b years and age $b + 10$ years as $cumul = 1 - e^{-riskscore \times popincid}$ and $cumul10 = 1 - e^{-riskscore \times popincid10}$, respectively. For each participant, their expected survival in the next 10 years was $surv10 = e^{-mort10}$, where *mort10* was their expected mortality in the next 10 years. The 10-year absolute risk of pancreatic cancer was then calculated as $absrisk10yr = \frac{(cumul10 - cumul) \times surv10}{(1 - cumul)}$.

Statistical analysis

In our prospective cohort study, follow-up of each participant began at their baseline assessment and ended at the earliest of their date of completing 10 years of follow-up, date of diagnosis of pancreatic cancer, date of death, or 31 July 2019 (to which linkage to cancer registries is complete). We first assessed the standardized incidence ratio (SIR) of the number of pancreatic cancers expected using sex-specific, age-specific, and calendar year-specific population incidence rates for England[29] compared to the number observed during the 10 years of follow-up, overall, by 10-year age group and by sex.

We used Cox regression with age as the time axis to estimate the hazard ratio per standard deviation (SD) of 10-year risk for the PRS, the clinical risk score, and the combined risk score. Harrell's C-index was used to assess the ability of the risk scores to distinguish between affected and unaffected participants (ie the discrimination of the risk scores). We then plotted Nelson–Aalen cumulative hazard curves for the 3 risk scores stratified by quintile of 10-year risk and extracted the cumulative hazards at 5-year intervals from age 45–75 years.

We evaluated calibration using logistic regression to estimate coefficients for the log odds of the predicted 10-year risk for each of the risk scores and tested whether the coefficients were equal to 1.[32,33] The estimated coefficient is a measure of dispersion, where values < 1 indicate overdispersion, values > 1 indicate underdispersion, and values close to 1 indicate no problem with dispersion. We then constrained the logistic regression models to have a slope of 1 and used the intercept term to assess overall calibration.[33] To illustrate the calibration of the models, we drew calibration plots for deciles of the 10-year risks using the pmcalplot module in Stata.[34,35]

For each of the 10-year risk scores, we used cut points at 0.36% and 0.64% to define risk categories (<0.36% risk was categorized as average-risk, ≥0.36% to <0.64% was categorized as increased-risk and ≥0.64% was categorized as high-risk). The cut-point at 0.36% (which is 1.8 times the median 10-year population risk of 0.2%) was based on the increased risk threshold suggested by Permuth-Wey and Egan,[36] while the cut-point at 0.64% (3.2 times the median 10-year risk) was based on National Comprehensive Cancer Network recommendations[37] for identifying individuals at high-risk. We calculated the change in classification for cases and controls separately and the net reclassification improvement[38] (NRI) for the combined risk score compared to both the PRS and the clinical risk score. We used McNemar's asymptotic test for correlated proportions to test the null hypothesis that the NRI or the classification change for cases and controls separately were equal to 0.

To illustrate the ability of the combined model to stratify pancreatic cancer risk, we calculated the SIR of the number of cases expected using sex- and age-specific population incidence rates for England[29] and the number observed during the 10 years of follow-up for each decile of 10-year risk and using the cut-offs used in the NRI analysis. We used Stata (version 17.0)[34] for the analyses; all statistical tests were 2 sided and *P* values < .05 were considered nominally statistically significant.

Ethics approval

The UK Biobank has Research Tissue Bank approval (REC #11/NW/0382) that covers analysis of data by approved researchers. All participants provided written informed consent to the UK Biobank before data collection began. This research has been conducted using the UK Biobank resource under application number 47401.

Results

There were 376,462 participants (202,215 women and 174,247 men) in the final analysis dataset, 851 (399 women and 452 men) of whom were diagnosed with incident pancreatic cancer during the 10-year follow-up period. Mean age at baseline assessment date was 62.0 years (SD = 6.2 years) for affected participants and 57.4 years (SD = 7.9 years) for unaffected participants. Affected participants had a mean age of 68.0 years (SD = 6.6 years) at diagnosis and a mean of 5.9 years (SD = 2.9 years) of follow-up time until their diagnosis. Unaffected participants had a mean of 9.7 years (SD = 1.0 years) of follow-up time.

Overall, 343,216 (91.2%) participants had all 22 SNPs genotyped and 31,795 (8.5%) were missing one SNP. For the clinical risk score, 371,998 (98.8%) participants had no missing variables and 4385 (1.2%) were missing one variable (see Table A2). The mean relative risks were 0.98 (SD = 0.51) for the PRS, 3.42 (SD = 1.67) for the clinical risk score before centering, 1.0 (SD = 0.49) for the clinical risk score after centering, and 1.00 (SD = 0.66) for the combined risk score. The characteristics of the unaffected and affected participants for the risk factors in the clinical risk score are given in Table A3.

When we compared the number of pancreatic cancers seen during the 10 years of follow-up to the number expected using population incidence rates, there was evidence that there were fewer pancreatic cancers seen in the UK Biobank participants than expected (observed [O] = 851, expected [E] = 1060.8, SIR = 0.802, 95% CI = 0.750–0.858, *P* < .001). This was seen in the 50–59 years age group (O = 195, E = 263.6, SIR = 0.740, 95% CI = 0.643–0.851, *P* < .001 and the 60–69 years age group (O = 604, E = 740.2, SIR = 0.816, 95% CI = 0.754–0.884, *P* < .001), but not in the 40–49 years age group (O = 52, E = 57.0, SIR = 0.913, 95% CI = 0.696–1.198, *P* = .5). When stratified by sex, there were fewer pancreatic cancers than expected for both women (O = 399, E = 493.9, SIR = 0.808, 95% CI = 0.732–0.891) and men (O = 425, E = 566.9, SIR = 0.797, 95% CI = 0.727–0.874).

In affected participants, the mean 10-year risk was 0.383% (SD = 0.250%) for the PRS, 0.390% (SD = 0.244%) for the clinical risk score, and 0.433% (SD = 0.357%) for the combined risk score. In unaffected participants the mean 10-year risk was 0.241% (SD = 0.188%) for the PRS, 0.248% (SD = 0.187%) for the clinical risk score, and 0.247% (SD = 0.224%) for the combined risk score. Figure A2A shows the distribution of the 10-year risk of pancreatic cancer for the combined risk score for unaffected and affected participants.

Table 1 shows the hazard ratio per SD and Harrell's C-index for the three 10-year risk scores. All of the risk scores were

Table 1. Hazard Ratio Per SD of Risk and Harrell's C-Index for the 10-y Risk of the PRS, the Clinical Risk Score, and the Combined Risk Score

Association	Hazard ratio per SD of risk	95% CI	*P* value
Polygenic risk score	1.328	1.263, 1.397	<.001
Clinical risk score	1.328	1.265, 1.394	<.001
Combined risk score	1.310	1.264, 1.357	<.001

Discrimination	Harrell's C-index	95% CI	*P* value[a]
Polygenic risk score	0.702	0.686, 0.717	<.001
Clinical risk score	0.703	0.687, 0.720	<.001
Combined risk score	0.714	0.698, 0.730	<.001

[a] *P* value for test that Harrell's C-index = 0.5.

strongly associated with pancreatic cancer (all $P < .001$). In terms of discrimination of affected and unaffected participants, all of the risk scores discriminated well (all $P < .001$). The combined risk score discriminated better than both the PRS ($z = 3.18$, $P = .001$) and the clinical risk score ($z = 2.27$, $P = .02$).

The calibration plots in Figure 1 show good calibration across almost all the deciles of risk for each of the three 10-year risk scores. The logistic regression estimates for the log odds of the predicted 10-year risks showed that there were no problems with dispersion for any of the PRS (0.988, 95% CI = 0.889–1.087, $P = .8$), clinical risk score (1.035, 95% CI = 0.925–1.145, $P = .5$), or combined risk score (0.952, 95% CI = 0.865–1.039, $P = .3$). For the overall calibration (with the slope constrained to 1), there was marginal evidence of miscalibration for the PRS (−0.065, 95% CI = −0.132 to 0.002, $P = .06$) and evidence of slight overestimation of risks for the clinical risk score (−0.094, 95% CI = −0.161 to −0.026, $P = .006$), and the combined risk score (−0.089, 95% CI = −0.156 to −0.021, $P = .009$).

Figure 2 and Table A4 show the Nelson–Aalen cumulative hazard functions for the three 10-year risk scores stratified by quintile of risk. The cumulative hazard functions for the quintiles of the PRS (Figure 2A) have similar trajectories to age 65 years, after which the top and bottom quintiles diverge. For the clinical risk score (Figure 2B), the top quintile begins to diverge from the bottom quintile from 70 years of age, while for the combined risk score (Figure 2C), the top quintile begins to diverge from the bottom quintile from 60 years of age. By the age of 60 years, the cumulative hazard for the top quintile of the combined risk score (0.0023, 95% CI = 0.0013–0.0041) was 3.8 times that of the bottom quintile (0.0006, 95% CI = 0.0040 to 0.0009). By the age of 75 years, the cumulative hazard for the top quintile of the combined risk score (0.0092, 95% CI = 0.0082–0.0114) was 11.5 times that of the bottom quintile (0.0008, 95% CI = 0.0005–0.0014) and 1.8 times that of the middle quintile (0.0051, 95% CI = 0.0039–0.0066).

Classification tables for the 10-year risks for the PRS vs the combined risk score and the clinical risk score vs the

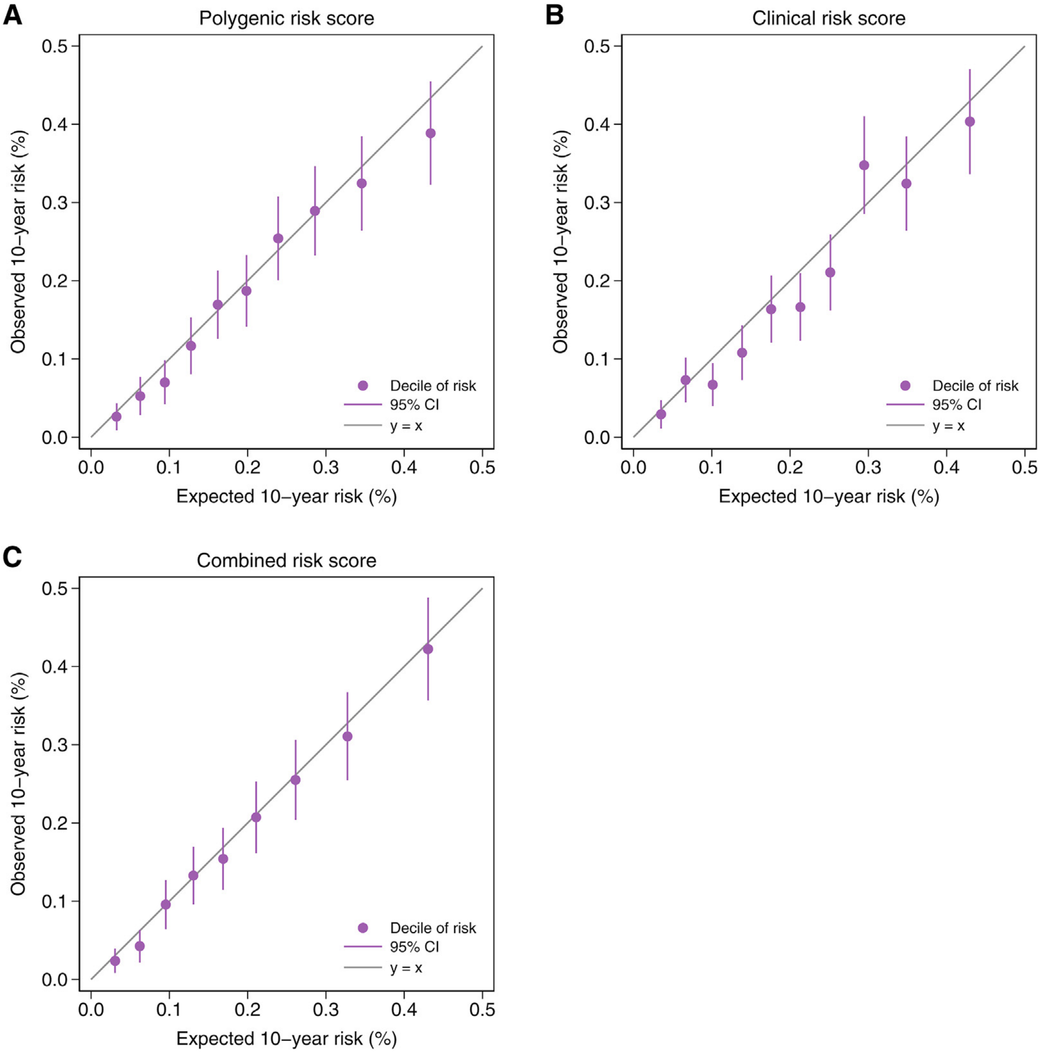


Figure 1. Calibration plots for the 10-year risks of the (A) PRS, (B) clinical risk score, and (C) combined risk score.

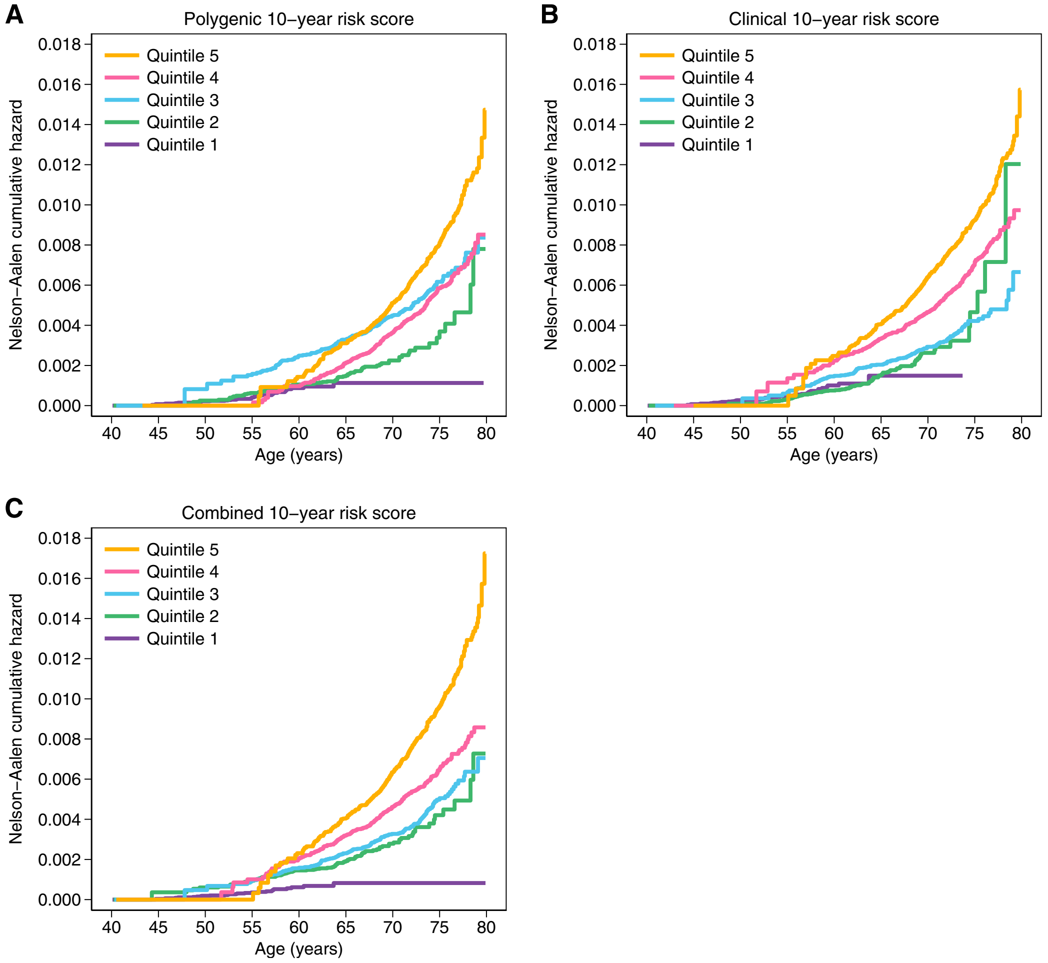


Figure 2. Nelson–Aalen cumulative hazard rates for quintiles of 10-year risk for the (A) PRS, (B) clinical risk score, and (C) combined risk score.

combined risk score are shown in Table 2. The combined risk score improved the overall classification performance compared to the PRS (NRI = 0.066, 95% CI = 0.035–0.098. $P < .001$). This improvement was driven by a net improvement in the classification of affected participants of 0.080 (95% CI = 0.049–0.111, $P < .001$). In unaffected participants, there was a small net change to higher risk categories of 0.014 (95% CI = 0.012–0.015, $P < .001$). Similarly, Table 2 shows that the 10-year combined risk score improved overall classification performance compared to the 10-year clinical risk score (NRI = 0.060, 95% CI = 0.027–0.094. $P < .001$), which was driven by an improvement in classification performance for affected participants of 0.075 (95% CI = 0.042–0.109, $P < .001$). There was a net change to higher risk categories for unaffected participants of 0.015 (95% CI = 0.014–0.016, $P < .001$).

Table 3 and Figure 3 show the SIRs and 95% CIs for deciles of 10-year risk for the 3 risk scores compared to population incidence rates. Participants in the top decile of 10-year risk for the combined risk score had a median risk of 0.66%, were at 1.413 (95% CI = 1.242–1.607) times population risk and were at 4.5 times the risk of the participants in the bottom decile. Participants in the first 8 deciles of 10-year combined risk score were at less than population risk, while the ninth decile was at population risk. Participants in the top decile of 10-year risk for the PRS and the clinical risk score both had a median risk of 0.60% and were at 1.273 (95% CI = 1.114–1.455) and 1.303 (95% CI = 1.142–1.488) times population risk, respectively. When we looked at the SIRs stratified by the risk thresholds used in the NRI analysis in Table 2, the SIRs for the participants in the high-risk category (≥0.64%) were 1.511 (95% CI = 1.247–1.831) for the PRS, 1.493 (95% CI = 1.225–1.820) for the clinical risk score and 1.551 (95% CI = 1.316–1.827) for the combined risk score (Table 4 and Figure A3).

Discussion

Current pancreatic screening recommendations focus on familial and genetic risk of disease,[37] but only 4% of the population have a family history of pancreatic cancer.[39] There needs to be an improvement in clinical screening criteria because the majority of adults ultimately diagnosed with pancreatic cancer do not meet those selective criteria.

Table 2. Classification Tables for the 10-y Combined Risk Score Compared to the 10-y PRS and the 10-y Clinical Risk Score

		Combined risk score category					Combined risk score category			
	Affected	Average	Increased	High	Total	Unaffected	Average	Increased	High	Total
Combined vs polygenic risk score										
Polygenic risk score category	Average	401	72	11	484	Average	278,584	17,415	1827	297,826
	Increased	43	162	58	263	Increased	18,305	35,593	9195	63,093
	High	0	30	74	104	High	1	5058	9633	14,692
	Total	444	264	143	851	Total	296,890	58,066	20,655	375,611
Combined vs clinical risk score										
Clinical risk score category	Average	376	86	7	469	Average	274,258	21,494	1270	297,022
	Increased	67	154	63	284	Increased	22,181	32,216	9861	64,258
	High	1	24	73	98	High	451	4356	9524	14,331
	Total	444	264	143	851	Total	296,890	58,066	20,655	375,611

PRS, polygenic risk score; for each risk score, participants with a 10-y risk of <0.36% were categorized as being at average risk, ≥0.36% and <0.64% were at increased risk and ≥0.64% were at high risk.

Table 3. Number of Cases Observed and Number of Cases Expected Using Population Incidence Rates, SIRs and 95% CIs for Deciles of the 10-y Risk for the PRS, Clinical Risk Score, and Combined Risk Score

Risk score	Observed	Expected	SIR	95% confidence interval
Polygenic 10-y risk score				
Decile 1 (median = 0.04%)	13	23.46	0.554	0.322, 0.954
Decile 2 (median = 0.07%)	19	40.80	0.466	0.297, 0.730
Decile 3 (median = 0.10%)	29	64.82	0.447	0.311, 0.644
Decile 4 (median = 0.14%)	48	88.57	0.542	0.408, 0.719
Decile 5 (median = 0.18%)	66	108.69	0.607	0.477, 0.773
Decile 6 (median = 0.22%)	92	124.31	0.740	0.603, 0.908
Decile 7 (median = 0.27%)	95	136.67	0.695	0.569, 0.850
Decile 8 (median = 0.33%)	128	146.78	0.872	0.733, 1.037
Decile 9 (median = 0.41%)	145	157.02	0.924	0.785, 1.087
Decile 10 (median = 0.60%)	216	169.65	1.273	1.114, 1.455
Clinical 10-y risk score				
Decile 1 (median = 0.04%)	12	21.29	0.564	0.320, 0.993
Decile 2 (median = 0.07%)	26	36.79	0.707	0.481, 1.038
Decile 3 (median = 0.11%)	24	57.67	0.416	0.279, 0.621
Decile 4 (median = 0.15%)	61	83.59	0.730	0.568, 0.938
Decile 5 (median = 0.19%)	50	108.29	0.462	0.350, 0.609
Decile 6 (median = 0.23%)	70	127.58	0.549	0.434, 0.694
Decile 7 (median = 0.28%)	112	141.56	0.791	0.657, 0.952
Decile 8 (median = 0.33%)	123	153.21	0.803	0.673, 0.958
Decile 9 (median = 0.41%)	153	162.01	0.944	0.806, 1.107
Decile 10 (median = 0.59%)	220	168.78	1.303	1.142, 1.488
Combined 10-y risk score				
Decile 1 (median = 0.03%)	9	24.45	0.368	0.192, 0.707
Decile 2 (median = 0.06%)	16	43.31	0.369	0.226, 0.603
Decile 3 (median = 0.10%)	36	67.69	0.532	0.384, 0.737
Decile 4 (median = 0.13%)	50	91.13	0.549	0.416, 0.724
Decile 5 (median = 0.17%)	58	109.93	0.528	0.408, 0.683
Decile 6 (median = 0.21%)	78	124.61	0.626	0.501, 0.782
Decile 7 (median = 0.26%)	96	135.44	0.709	0.580, 0.866
Decile 8 (median = 0.33%)	117	145.41	0.805	0.671, 0.965
Decile 9 (median = 0.43%)	159	154.59	1.029	0.881, 1.202
Decile 10 (median = 0.66%)	232	164.22	1.413	1.242, 1.607

The SIR is the observed number of pancreatic cancer cases divided by the number expected using population incidence rates.
SIR, standardized incidence ratio.

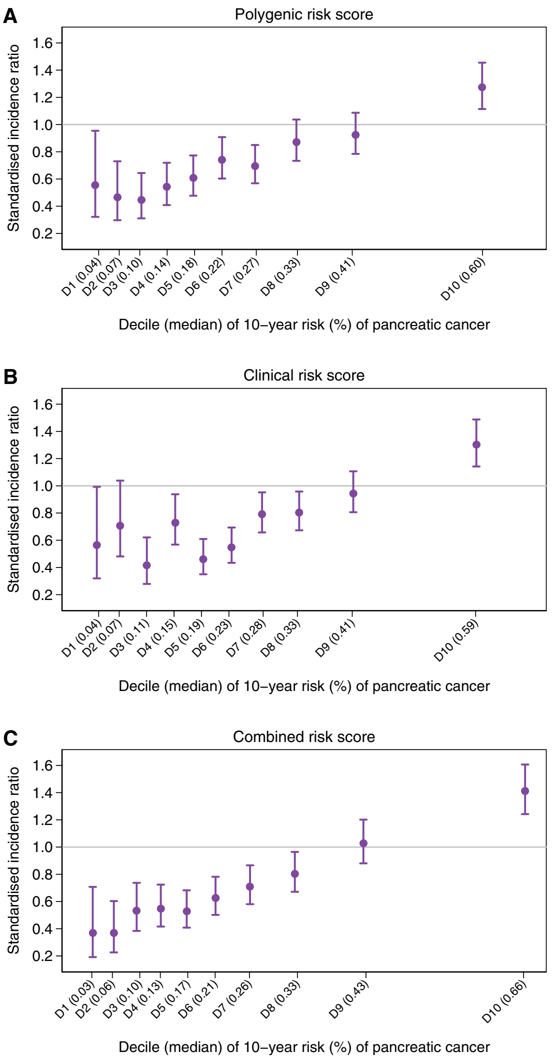


Figure 3. SIRs and 95% CIs for deciles of 10-year risk for the (A) PRS, (B) clinical risk score, and (C) combined risk score, with the deciles placed on the *x*-axis at their median values.

Historically, because the incidence of pancreatic cancer is so low, creating, and validating population-based risk models has been challenging. However, using almost 380,000 adults aged 40–69 years in the UK Biobank, we identified 851 incident cases of pancreatic cancer over a 10-year period. This has enabled us to rigorously validate a risk prediction model that comprises a PRS developed by Jia et al[5] and a clinical risk model developed by Klein et al.[14] We focused on 10-year risk in this manuscript because we had 10 years of follow-up data available for analysis in the UK Biobank, but the relative risk from the combined model can be used with appropriate population incidence rates to calculate absolute risk over any period (Figure A2A–D and Supplementary Methods).

We first looked at the discrimination and calibration of the clinical risk score, the PRS, and the combined risk score. We found a small but statistically significant improvement (due to the large sample size) in discrimination for the combined risk score compared to the PRS and the clinical risk score. The small overestimation of overall risk seen for the 10-year risks for the clinical risk score and combined risk score (and marginal evidence for the PRS) was statistically significant and is consistent with the finding that the UK Biobank had fewer cancers than expected using population incidence rates. Therefore, there might not be an overestimation of risk in a real-world clinical setting, but it is better to overestimate risk than to underestimate risk when it comes to preventive care. Interestingly, the discriminatory performance of the component risk scores was better in our study than in the original papers. For the PRS, the original AUC was 0.639,[5] whereas the Harrell's C-index in this study was 0.702. Similarly, for the clinical risk score, the original AUC was 0.61[14] and the Harrell's C-index in this study was 0.703.

Next, the advantage of using the combined risk score became evident when we looked at the ability of the 3 risk scores to stratify the population. Figure 2 (and Table A4) highlights this stratification potential. Focusing on the top and bottom quintiles of risk, the clinical model was able to statistically significantly differentiate the cumulative hazards by the age of 70 years, and the PRS model was able to differentiate between the top and bottom quintiles of risk by the age of 65. However, the combined model was able to differentiate the cumulative hazards between the top and bottom quintiles of risk by the age of 60. These data highlight the benefit of a combined model over either model independently.

We next looked at the ability of the 3 risk scores to classify participants as being at average, increased, or high risk. Because of clinical guidelines for screening that are based on having a first-degree family history of pancreatic cancer, we used a relative risk of 1.8 (which is the relative risk associated with having a first-degree family history[36]) to establish a clinically equivalent indicator of increased risk. In the general population, the median 10-year risk of pancreatic cancer is 0.2%; therefore the 10-year risk cut-off to identify participants at increased risk was 0.36%. We used a relative risk of 3.2 (which is equivalent to having a moderate penetrant [likely] pathogenic variant[37]) to identify participants at high risk; therefore, the 10-year risk cut-off to identify participants at high risk was 0.64%.

Using these thresholds, we conducted a reclassification analysis comparing the combined risk score with both the PRS and the clinical risk score. In both cases, the combined model improved net reclassification. While the combined and clinical models performed similarly when reclassifying participants at the increased risk threshold, notable reclassification occurs at the high-risk threshold. The combined model identified 17% of affected participants over the high-risk threshold compared with 11% of affected participants identified using the clinical risk score, representing a 50% improvement. This same change was observed in unaffected participants, where 5.5% were identified above the high-risk threshold with the combined risk score compared with 3.8% for the clinical risk score.

Table 4. Number of Cases Observed and Number of Cases Expected Using Population Incidence Rates, SIRs and 95% CIs for Participants at Average, Increased, and High Risk 10-y Risk for the PRS, Clinical Risk Score, and Combined Risk Score

Risk score	Observed	Expected	SIR	95% confidence interval
Polygenic 10-y risk score				
<0.36%	484	722.7	0.670	0.613, 0.732
≥0.36% to <0.64%	263	269.24	0.977	0.866, 1.102
≥0.64%	104	68.83	1.511	1.247, 1.831
Clinical 10-y risk score				
<0.36%	469	714.52	0.656	0.560, 0.719
≥0.36% to <0.64%	284	280.61	1.012	0.901, 1.137
≥0.64%	98	65.63	1.493	1.225, 1.820
Combined 10-y risk score				
<0.36%	444	726.62	0.611	0.557, 0.671
≥0.36% to <0.64%	264	241.93	1.091	0.967, 1.231
≥0.64%	143	92.22	1.551	1.316, 1.827

The SIR is the observed number of pancreatic cancer cases divided by the number expected using population incidence rates.
SIR, standardized incidence ratio.

In the UK Biobank, 1 in 442 adults aged 40–69 years would need to be screened to identify one affected adult over a 10-year period. Using the high-risk categories of the combined risk score or the clinical risk score, 1 in 145 and 1 in 147, respectively, adults would need to be screened. In these high-risk individuals (≥0.64% 10-year risk), the SIRs compared to population incidence rates (Table 4) are all around 1.5, but as shown in Table 2, the combined risk score was able to identify many more at-risk individuals than either the PRS or the clinical risk score. Of note, the comparison of the number needed to screen makes the assumption that pancreatic risk assessment is routinely done in primary care, which it is not. The number needed to screen for the combined risk model could be a significant improvement if the combined model were applied and assessed in a real-world clinical scenario.

Finally, when we looked at the SIRs compared to population incidence rates by decile of risk (Table 3 and Figure 3), the combined risk score was better able to identify participants at high risk than the PRS or the clinical risk score alone. This was evident in terms of the SIR for the top decile, which was 1.4 for the combined risk score and 1.3 for both of the other risk scores. In addition, the median 10-year risk for the top decile of the combined risk score was 0.66%, which was higher than both the PRS (0.60%) and the clinical risk score (0.59%).

A potential limitation of our study is the lack of pancreatic cancer family history data in the UK Biobank. While the inclusion of family history into the risk score may have provided some additional benefit, less than 4% of the general population have a family history of pancreatic cancer (and the population attributable fraction is only 3%),[39] so the contribution to overall risk at a population level is minor. Another limitation is that, while we observed an increase in unaffected participants being reclassified to higher categories in the NRI analysis rather than being reclassified to lower categories, this might have been because there are fewer pancreatic cancers observed in the UK Biobank than expected using population incidence rates. In turn, this might have contributed to the combined models' overestimation of risk and misclassification of unaffected participants.

The vast size of the UK Biobank has allowed us to undertake a comprehensive analysis without being hindered by small numbers when we stratified the analyses by quintile of risk. Additionally, there was very little missing data (Tables A2 and A3), and, therefore, we simply omitted any missing risk factors from the participant's relative risk calculation. This meant that we were able to calculate risk scores for all eligible UK Biobank participants. Future studies will cross validate this model in separate datasets, including those including participants of ethnically diverse genetic ancestries.

Conclusion

While pancreatic cancer is rare, it is important to identify at-risk adults as early as possible to be able to screen efficiently. Although the clinical and PRSs performed well independently, the combined risk score had an advantage in the number of high-risk individuals identified and the net reclassification of affected and unaffected individuals compared with both the PRS and the clinical risk score. The combined risk score also had a clear advantage in the stratification of cumulative hazards, differentiating the top and bottom quintiles at an earlier age than the other risk scores. Based on the analyses presented herein, the combined genetic and clinical risk score was able to identify a greater number of individuals at substantially increased risk of pancreatic cancer. The combined model identifies high-risk individuals who could benefit from existing and emerging targeted screening techniques that could be used more efficiently based on the number needed to screen, thereby providing a net clinical benefit.

We suggest that the use of an improved risk stratification model at the general population level will increase the number of at-risk adults who qualify for high-risk screening programs, representing advancement for clinical practice. Using the clinical risk model alone will miss about 45% of the at-risk population (defined as being at least four-times average risk) identified using the combined model. This means that the opportunity to screen-detect approximately 45% more early stage (presymptomatic) pancreatic cancers will be missed.

A risk-stratification tool paired, in a step-wise manner, with existing or emerging pancreatic screening techniques could lead to clinically significant downstaging of pancreatic cancer diagnoses. This is important because identifying at-risk adults at early stage diagnosis where surgical resection is still possible is associated with increased survival rates.[40,41]

Supplementary Materials

Material associated with this article can be found in the online version at https://doi.org/10.1016/j.gastha.2023.05.008.

References

1. Surveillance, Epidemiology, and End Results Program. SEER cancer stat facts: pancreatic cancer. 2022. https://seer.cancer.gov/statfacts/html/pancreas.html. Accessed September 20, 2022.
2. Hu C, LaDuca H, Shimelis H, et al. Multigene hereditary cancer panels reveal high-risk pancreatic cancer susceptibility genes. JCO Precis Oncol 2018;2:PO.17.00291.
3. Chen F, Childs EJ, Mocci E, et al. Analysis of heritability and genetic architecture of pancreatic cancer: a PanC4 study. Cancer Epidemiol Biomarkers Prev 2019;28(7):1238–1245.
4. Lichtenstein P, Holm NV, Verkasalo PK, et al. Environmental and heritable factors in the causation of cancer — analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med 2000;343(2):78–85.
5. 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;4(3):pkaa021.
6. Kachuri L, Graff RE, Smith-Byrne K, et al. Pan-cancer analysis demonstrates that integrating polygenic risk scores with modifiable risk factors improves risk prediction. Nat Commun 2020;11(1):6084.
7. Nakatochi M, Lin Y, Ito H, et al. Prediction model for pancreatic cancer risk in the general Japanese population. PLoS One 2018;13(9):e0203386.
8. Bogumil D, Conti DV, Sheng X, et al. Replication and genetic risk score analysis for pancreatic cancer in a diverse multiethnic population. Cancer Epidemiol Biomarkers Prev 2020;29(12):2686–2692.
9. Wang J, Conti DV, Bogumil D, et al. Association of genetic pisk score with NAFLD in an ethnically diverse cohort. Hepatol Commun 2021;5(10):1689–1703.
10. 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):1433–1437.
11. Krejs GJ. Pancreatic cancer: epidemiology and risk factors. Dig Dis 2010;28(2):355–358.
12. Risch HA, Yu H, Lu L, et al. Detectable symptomatology preceding the diagnosis of pancreatic cancer and absolute risk of pancreatic cancer diagnosis. Am J Epidemiol 2015;182(1):26–34.
13. Wei MY, Shi S, Liang C, et al. The microbiota and microbiome in pancreatic cancer: more influential than expected. Mol Cancer 2019;18(1):97.
14. Klein AP, Lindstrom 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.
15. Appelbaum L, Cambronero JP, Stevens JP, et al. Development and validation of a pancreatic cancer risk model for the general population using electronic health records: an observational study. Eur J Cancer 2021;143:19–30.
16. Hippisley-Cox J, Coupland C. Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: prospective cohort study. BMJ Open 2015;5(3):e007825.
17. Yu J, Blackford AL, Dal Molin M, et al. Time to progression of pancreatic ductal adenocarcinoma from low-to-high tumour stages. Gut 2015;64(11):1783–1789.
18. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature 2018;562(7726):203–209.
19. 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;12(3):e1001779.
20. Conroy MC, Lacey B, Besevic J, et al. UK Biobank: a globally important resource for cancer research. Br J Cancer 2022;128(4):519–527.
21. UK Biobank. Data providers and dates of data availability. 2021. https://biobank.ctsu.ox.ac.uk/showcase/exinfo.cgi?src=Data_providers_and_dates. Accessed June 14, 2022.
22. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. Am J Epidemiol 2017;186(9):1026–1034.
23. Hanscombe KB, Coleman JRI, Traylor M, et al. ukbtools: an R package to manage and query UK Biobank data. PLoS One 2019;14(5):e0214311.
24. R Core Team. R: a language and environment for statistical computing [computer program]. Vienna, Austria: R Foundation for Statistical Computing, 2020.
25. Purcell S, Chang C. Plink 1.9. 2020. https://www.cog-genomics.org/plink/1.9/. Accessed March 30, 2021.
26. Chang CC, Chow CC, Tellier LCAM, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 2015;4(1):7.
27. Mealiffe ME, Stokowski RP, Rhees BK, et al. Assessment of clinical validity of a breast cancer risk model combining genetic and clinical information. J Natl Cancer Inst 2010;102(21):1618–1627.

28. Conran CA, Na R, Chen H, et al. Population-standardized genetic risk score: the SNP-based method of choice for inherited risk assessment of prostate cancer. Asian J Androl 2016;18(4):520–524.
29. Office for National Statistics. Cancer registration statistics. 2019. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/datasets/cancerregistrationstatisticscancerregistrationstatisticsengland. Accessed July 20, 2021.
30. Office of National Statistics. Mortality statistics - underlying cause, sex and age. 2019. https://www.nomisweb.co.uk/query/construct/summary.asp?mode=construct&version=0&dataset=161. Accessed July 20, 2021.
31. Office of National Statistics. Mortality statistics. 2016. https://www.nomisweb.co.uk/query/construct/summary.asp?mode=construct&version=0&dataset=161. Accessed July 20, 2021.
32. Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. BMC Med 2019;17(1):230.
33. Huang Y, Li W, Macheret F, et al. A tutorial on calibration measurements and calibration models for clinical prediction models. J Am Med Inform Assoc 2020;27(4):621–633.
34. StataCorp. Stata statistical software: release 17. College Station, TX: StataCorp LLC, 2021.
35. Ensor J, Snell KIE, Martin EC. PMCALPLOT: Stata module to produce calibration plot of prediction model performance. 2020. https://ideas.repec.org/c/boc/bocode/s458486.html. Accessed August 18, 2022.
36. Permuth-Wey J, Egan KM. Family history is a significant risk factor for pancreatic cancer: results from a systematic review and meta-analysis. Fam Cancer 2009;8(2):109–117.
37. National Comprehensive Cancer Network. NCCN clinical practice guidelines in oncology. Genetic/familial high-risk assessment: breast, ovarian and pancreatic. Version 1. 2023. 2022. https://www.nccn.org/guidelines/guidelines-detail?category=2&id=1460. Accessed September 20, 2022.
38. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008;27(2):157–172, discussion 207-112.
39. Jacobs EJ, Chanock SJ, Fuchs CS, et al. Family history of cancer and risk of pancreatic cancer: a pooled analysis from the Pancreatic Cancer Cohort Consortium (PanScan). Int J Cancer 2010;127(6):1421–1428.
40. Canto MI, Almario JA, Schulick RD, et al. Risk of neoplastic progression in individuals at high risk for pancreatic cancer undergoing long-term surveillance. Gastroenterology 2018;155(3):740–751.e2.
41. Vasen H, Ibrahim I, Ponce CG, et al. Benefit of surveillance for pancreatic cancer in high-risk individuals: outcome of long-term prospective follow-up studies from three European expert centers. J Clin Oncol 2016;34(17):2010–2019.

Received December 9, 2022. Accepted May 12, 2023.

Correspondence:
Address correspondence to: Gillian S. Dite, PhD, Genetic Technologies Limited, 60–66 Hanover St, Fitzroy, Victoria 3065, Australia. e-mail: gillian.dite@gtglabs.com.

Acknowledgments
We wish to thank Mr Lawrence Whiting for his invaluable expertise in the management of the large data files from the UK Biobank. This research has been conducted using the UK Biobank resource under Application Number 47401. This research used data provided by patients and collected by the NHS as part of their care and support, copyright © (2023), NHS England, and is reused with the permission of the UK Biobank. All rights reserved. This research used data assets made available by National Safe Haven as part of the Data and Connectivity National Core Study, led by Health Data Research UK in partnership with the Office for National Statistics and funded by UK Research and Innovation (grant reference MC_PC_20058).

Authors' Contributions:
Gillian S. Dite: Project administration, conceptualization, methodology, data curation, formal analysis, writing – original draft, writing – review and editing. Erika Spaeth: Supervision, conceptualization, methodology, writing – original draft, writing – review and editing. Chi Kuen Wong: Formal analysis, validation, investigation, writing – review and editing. Nicholas M. Murphy: Resources, investigation, writing – review and editing. Richard Allman: Supervision, conceptualization, methodology, writing – review and editing.

Conflicts of Interest:
The authors disclose the following: Gillian S. Dite, Chi Kuen Wong, and Nicholas M. Murphy are employees of Genetic Technologies Limited. Erika Spaeth is an employee of Phenogen Sciences Incorporated (a subsidiary of Genetic Technologies Limited). Richard Allman was previously an employee of Genetic Technologies Limited and now acts in a consultancy role. Aspects of this manuscript are covered by Australian Provisional Patent Application No: 2022903204, Methods of assessing risk of developing pancreatic cancer.

Funding:
This study was fully funded by Genetic Technologies Limited, which had no role in the conceptualization, design, data analysis, decision to publish, or preparation of the manuscript.

Ethical Statement:
The corresponding author, on behalf of all authors, jointly and severally, certifies that their institution has approved the protocol for any investigation involving humans or animals and that all experimentation was conducted in conformity with ethical and humane principles of research.

Data Transparency Statement:
The data for this study was provided by the UK Biobank and we do not have permission to share the data. Researchers wishing to access the data used in this study can apply directly to the UK Biobank at https://www.ukbiobank.ac.uk/register-apply. Stata 17.0 code for the data management and analysis is available from the corresponding author for noncommercial purposes.

Reporting Guidelines:
STROBE.