npj | precision oncology

Article

Published in partnership with The Hormel Institute, University of Minnesota

https://doi.org/10.1038/s41698-026-01479-x

Polygenic scores for risk of pancreatic ductal adenocarcinoma: evaluation of novel and published models

Check for updates

A list of authors and their affiliations appears at the end of the paper

Polygenic risk scores (PRSs) may enhance risk stratification for pancreatic ductal adenocarcinoma (PDAC), but existing models vary widely in design, predictive performance, and cross-ancestry transferability. We developed genome-wide PRSs using Bayesian methods (LDpred2 and PRS-CS) and *p* value thresholding (PRSice-2) and systematically evaluated these alongside 13 published PRSs to identify models with robust predictive performance across ancestries. Using GWAS summary statistics from 7531 cases and 10,631 controls, we derived the PRSs and tested associations in an independent sample of 4508 PDAC cases and 46,189 controls, with adjustment for well-established PDAC risk factors. Among all models, the genome-wide LDpred2-based PRS showed the strongest association with PDAC (OR = 1.57 per standard deviation increase; 95% CI: 1.51–1.62) and significantly improved discrimination beyond established risk factors alone (AUC = 0.74–0.76; $p < 0.0001$). Importantly, the genome-wide LDpred2 PRS demonstrated consistent associations across African, Admixed American, and European ancestry groups, whereas the best-performing published PRS was associated with PDAC risk only in individuals of European ancestry. These findings support genome-wide PRSs as a promising framework for multi-ancestry risk stratification for PDAC and to inform targeted early detection strategies.

Pancreatic ductal adenocarcinoma (PDAC) is the third leading cause of cancer-related death in the U.S. and is projected to become the second by 2030[1,2]. Globally, PDAC ranks 12th in cancer incidence but 6th in cancer mortality, underscoring its disproportionate lethality[3]. More than 80% of patients present with locally advanced or metastatic disease, contributing to a poor overall 5-year survival rate of 8%[1]. However, the 5-year survival is 3% for metastatic disease compared with 44% for localized tumors, highlighting the critical importance of early detection[1]. Despite this need, the low prevalence of PDAC in the general population makes population-wide screening impractical, as reflected in the 2019 U.S. Preventive Services Task Force recommendation against PDAC screening for average-risk individuals[4]. In contrast, targeted screening of moderate-to-high-risk individuals may offer a more effective approach to early detection[5–7].

Individuals with a first-degree family history of PDAC have an over twofold higher risk of developing PDAC, with the risk rising further as the number of affected relatives increases, underscoring the contribution of hereditary factors[8,9]. However, the genetic basis of PDAC is dominated by somatic driver mutations, most notably in *KRAS*, *TP53*, *CDKN2A*, and *SMAD4*, that play major roles in tumorigenesis[10–14]. In addition, rare pathogenic germline variants in genes such as *ATM*, *BRCA1/2*, *CDKN2A*, *PALB2*, and *TP53* confer moderate-to-high risk and account for approximately 5–10% of cases[15–18]. Common genetic variation also contributes to PDAC susceptibility, but with modest risk magnitudes. Genome-wide association studies (GWASs) have thus far identified 31 PDAC risk loci, including 22 in European ancestry populations (PanScan I-III and PanC4)[19–24], five in a Chinese population[25], and four in Japanese populations[26,27]. An additional locus in the *TERT–CLPTM1L* region was identified through a candidate gene approach by the PANDoRA consortium[28,29]. Notably, there is limited overlap between susceptibility loci identified in the European and East Asian populations, suggesting ancestry-specific genetic architecture underlying risk[22,23,30].

Although GWASs have advanced our understanding of the genetic etiology of PDAC, the individual risk SNPs identified to date have modest effect sizes[19–24]. Aggregating these variants into polygenic risk scores (PRSs) may improve risk stratification and help identify moderate-to-high-risk individuals for targeted surveillance and clinical management[6,31]. At present, at least 13 PRS models have been developed for PDAC[23,32–44], but they vary widely in design, including differences in SNP selection, which limits comparability. Restricting existing PRSs to genome-wide significant SNPs also excludes variants with smaller effects that may nonetheless enhance

e-mail: Antwi.Samuel@mayo.edu; Ann.oberg@mayo.edu

Table 1 | Characteristics of the study participants (*N* = 50,697)

Characteristics	PDAC cases (*n* = 4508)	Controls (*n* = 46,189)
Age at consent		
Mean (SD)	65.7 (10.4)	60.4 (15.2)
Range	21 - 92	18 - 99
BMI at enrollment		
Mean (SD)	28.4 (5.7)	28.8 (6.0)
Range	12.7 - 64.0	11.6 - 73.2
Missing	22	398
Sex		
Male	2568 (57.0%)	19,457 (42.1%)
Female	1940 (43.0%)	26,732 (57.9%)
Family history of pancreatic cancer[a]		
Yes	392 (9.7%)	1768 (4.2%)
No	3660 (90.3%)	40342 (95.8%)
Missing	456	4079
Heavy alcohol use[b]		
Yes	465 (15.2%)	1476 (3.3%)
No	2599 (84.8%)	43231 (96.7%)
Missing	1444	1482
Chronic pancreatitis		
Yes	395 (8.8%)	88 (0.2%)
No	4113 (91.2%)	46101 (99.8%)
ABO blood group		
O	2921 (64.8%)	26647 (57.7%)
Non-O	1585 (35.2%)	19528 (42.3%)
Missing	2	14
Smoking history		
Never	2042 (46.3%)	26468 (57.9%)
Former	1758 (39.8%)	17046 (37.3%)
Current	615 (13.9%)	2210 (4.8%)
Missing	93	465
Type 2 diabetes mellitus		
Yes	1273 (33.3%)	8318 (18.0%)
No	2551 (66.7%)	37871 (82.0%)
Missing	684	0
Genetic ancestry		
African	65 (1.4%)	647 (1.4%)
Admixed American	61 (1.4%)	417 (0.9%)
East Asian	30 (0.7%)	334 (0.7%)
European	4334 (96.1%)	44406 (96.1%)
South Asian	11 (0.2%)	200 (0.4%)
Unassigned	7 (0.2%)	185 (0.4%)

BMI body mass index, *PDAC* pancreatic ductal adenocarcinoma, *SD* standard deviation.
[a]Having two or more first-degree relatives diagnosed with any pancreatic cancer type.
[b]Consuming ≥ 3 drinks of alcohol per day or having a history of alcohol abuse disorder.

predictive performance. Further, because most PDAC GWASs were conducted in European and East Asian populations, PRSs derived from these studies may perform suboptimally in other ancestry groups.

Other PRS modeling approaches, such as LDpred2[45] and PRS-CS[46], apply Bayesian shrinkage methods to adjust SNP effect sizes based on genome-wide linkage disequilibrium (LD) structure and can incorporate large numbers of variants with modest effects to enhance risk prediction. PRSice-2[47], which uses a *p* value thresholding approach, may also enhance risk modeling. These approaches have been used to improve risk modeling for colorectal, breast, and prostate cancers[48,49] and might similarly improve PRS-based risk stratification for PDAC.

The aims of this study were twofold: first, to validate existing published PDAC PRS models in an independent cohort using the same SNPs employed in each study; and second, to compare these models with newly developed genome-wide PRSs generated using LDpred2 and PRS-CS, and the *p* value thresholding PRSice-2 to identify the optimal approach for risk prediction. We hypothesize that genome-wide PRSs will outperform previously published models, providing improved risk stratification and greater potential to inform early detection strategies.

Results

Participant characteristics and risk factor distribution

Distributions of demographic characteristics and PDAC risk factors among the Mayo Clinic cases and controls are shown in Table 1. The PDAC cases were on average six years older than controls and included a higher proportion of men. The cases were also more likely than controls to report a family history of pancreatic cancer, heavy alcohol use, current smoking, and a history of type II diabetes, and they had a higher prevalence of chronic pancreatitis.

Risk factor associations with PDAC

Associations between each covariate and PDAC risk are provided for both the base and full models (Table 2). In the full model, nearly all risk factors were associated with higher PDAC risk, except former smoking, which showed no association, and self-reported BMI, which showed an inverse association. The strongest risk factor was chronic pancreatitis, with affected individuals having over 40-fold higher odds of PDAC (OR = 43.35, 95% CI: 33.92–55.40, $p < 0.0001$).

PRS effect sizes and discriminatory performance

Effect sizes of standardized PDAC PRS models generated using LDpred2, PRS-CS, and PRSice-2 (with *p* value thresholds of 1×10^{-5} and 5×10^{-8}), and previously published PDAC PRSs, were compared (Fig. 1). PRS effect sizes were estimated after adjustment for age, sex, and the top 10 PCs (base model) or for all covariates (full model). Among these, the LDpred2-based PRS showed the largest effect size in the full model (OR = 1.57, 95% CI: 1.51–1.62, $p < 0.0001$). The second most strongly associated PRSs, based on 22 GWAS-identified SNPs (e.g., Klein et al., 2018), showed a slightly smaller effect size (OR = 1.52, 95% CI: 1.47–1.58, $p < 0.0001$). However, the corresponding AUCs for the LDpred2 PRS (AUC = 0.76, 95% CI: 0.75–0.77) and the Klein et al. PRS (AUC = 0.76, 95% CI: 0.75–0.77) did not differ significantly by DeLong's test ($p = 0.12$). Notably, the PRS constructed using only genome-wide significant variants ($p < 5\times10^{-8}$) from our training GWAS had a smaller effect size (OR = 1.43, 95% CI: 1.38–1.48, $p < 0.0001$).

Model discrimination based on AUCs for the base covariate-only model was 0.62 (95% CI: 0.62–0.63), which increased to 0.74 (95% CI: 0.73–0.75) in the full covariate-only model (Fig. 2). Adding the LDpred2 PRS to the base model significantly improved discrimination, increasing the AUC from 0.62 (95% CI: 0.62–0.63) to 0.68 (95% CI: 0.67–0.68); DeLong's test for difference between models $p < 0.0001$. Similarly, adding the LDpred2 PRS to the full model further improved discrimination, increasing the AUC from 0.74 (95% CI: 0.73–0.75) to 0.76 (95% CI: 0.75–0.77); DeLong's test for difference $p < 0.0001$.

We next assessed PRS effect sizes after dichotomizing the PRSs at a potential clinical decision-making cutpoint to define a "high PDAC risk" (Supplementary Table 2). Individuals above the 95th percentile of the LDpred2-based PRS had 2.52-fold higher odds of PDAC (OR = 2.52, 95% CI: 2.26–2.81, $p < 0.0001$) compared with those below the 95th percentile, based on distribution among controls. This threshold identified 12% (n = 545) of PDAC cases in the top 5% of the LDpred2 PRS.

Table 2 | . Associations of variables in the base and full models with risk of PDAC

	Base model[a]		Full model[b]	
Characteristics	OR (95% CI)	*P* value	OR (95% CI)	*P* value
Age, years	1.02 (1.02,1.03)	5.93×10^{-91}	1.03 (1.03,1.03)	3.92×10^{-100}
Sex, male	1.63 (1.54,1.74)	3.69×10^{-53}	1.42 (1.33,1.52)	1.15×10^{-23}
Self-reported BMI, kg/m^2[c]			0.97 (0.97,0.98)	9.42×10^{-16}
Family history of pancreatic cancer[d]			2.18 (1.92,2.48)	6.50×10^{-31}
Heavy alcohol Use[e]			4.41 (3.84,5.06)	6.86×10^{-53}
Chronic pancreatitis			43.35 (33.92,55.40)	2.12×10^{-197}
Non-O ABO blood group			1.34 (1.25,1.43)	6.13×10^{-17}
Smoking, Former			0.95 (0.89,1.03)	0.20
Smoking, Current			3.00 (2.67,3.37)	1.27×10^{-72}
Type 2 diabetes mellitus			1.95 (1.8,2.11)	3.69×10^{-55}

BMI body mass index, *CI* conidence interval, *PC* principal components, *PDAC* pancreatic ductal adenocarcinoma, *OR* odds ratio.

[a]Based model included age (continuous), sex, and top 10 principal components, assessed simultaneously.

[b]Full model included the base model variables plus BMI (continuous), family history of pancreatic cancer (yes vs. no), heavy alcohol use (yes vs. no), chronic pancreatitis (yes vs. no), ABO blood group (non-O vs. O), smoking history (former, current, vs. never), and type 2 diabetes mellitus (yes vs. no).

[c]BMI was calculated based on self-reported usual adult weight and height for cases, and self-reported weight and height at recruitment for controls.

[d]History of any pancreatic cancer type in at least two first-degree relatives.

[e]Consuming ≥ 3 drinks of alcohol per day or having a history of alcohol abuse disorder.

Cross-ancestry evaluation of SNP PRS models

Finally, we evaluated the cross-ancestry performance of two PRSs: a genome-wide model (LDpred2; based on 1.029 million SNPs) and a candidate SNP model (Klein et al., 2018; 22 SNPs), stratified by genetic ancestry with ancestry-specific normalization (Fig. 3). The Klein et al. PRS showed limited transferability, with non-significant associations in both African (OR = 1.18, 95% CI: 0.88–1.58, $p = 0.27$) and Admixed American (OR = 1.22, 95% CI: 0.91–1.64, $p = 0.18$) samples, but with a significant association in European samples (OR = 1.54, 95% CI: 1.49–1.59, $p < 0.0001$). In contrast, the LDpred2 PRS showed a pattern of consistent associations across ancestries: African (OR = 1.51, 95% CI: 1.15–2.00, $p = 0.003$), Admixed American (OR = 1.46, 95% CI: 1.07–2.01, $p = 0.018$), and European (OR = 1.58, 95% CI: 1.53–1.63, $p < 0.0001$). Tests for heterogeneity supported cross-ancestry consistency for LDpred2 (Cochran's Q = 0.34, $p = 0.84$) and suggested a trend toward heterogeneity for the Klein et al. 2018 PRS (Q = 5.51, $p = 0.06$), indicating that genome-wide PRS models may provide more stable performance across diverse ancestries than candidate SNP-based models.

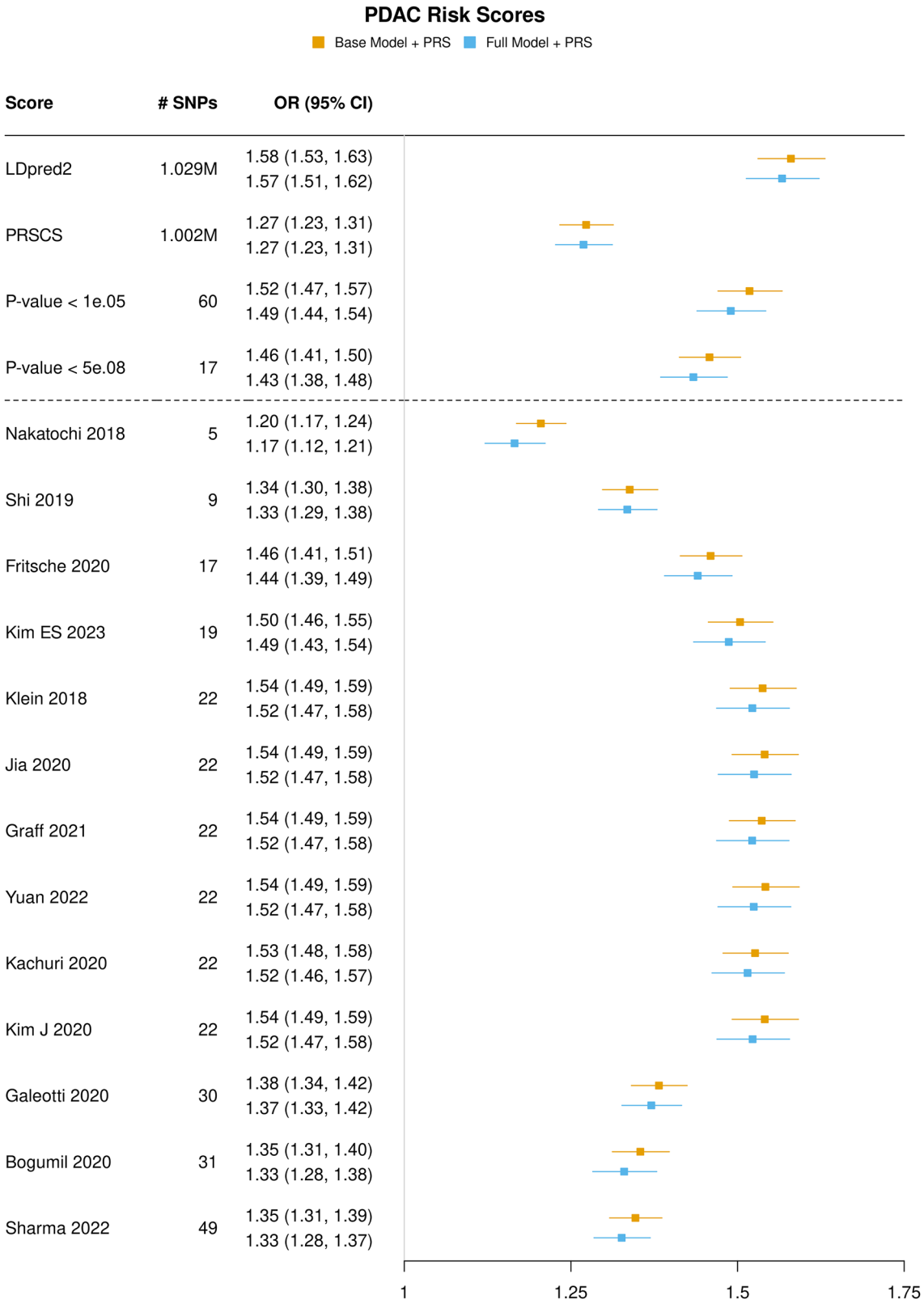
Discussion

We evaluated the performance of previously published PRS models for PDAC in an independent sample using the same SNP sets as in the original studies and applied consistent covariate adjustments across models to enable direct comparison. In parallel, we developed a new genome-wide PRS for PDAC using LDpred2 and PRS-CS and the *p* value thresholding PRSice-2, adjusting for established PDAC risk factors. Under comparable conditions, the genome-wide LDpred2-based PRS model (1.029 million SNPs) had a slightly higher effect size, but the AUC was similar to that of the best performing previously published models that used 22 GWAS-significant SNPs. In the fully adjusted model, each standard deviation increase in the LDpred2 PRS was associated with 1.57-fold higher odds of PDAC, and individuals in the highest 5% of the PRS distribution had 2.52-fold higher odds of PDAC compared with the remaining 95% of participants. AUCs for a full model with non-genetic risk factors only, a model with genetic factors only (LDpred2 PRS and the top 10 PCs), and a model with both genetic and non-genetic factors were 0.74, 0.62, and 0.76, respectively. Notably, the fully adjusted LDpred2 PRS model identified 12% of PDAC cases within the highest 5% of the PRS distribution in controls, considered having a high polygenic risk for PDAC. The LDpred2-based PRS also demonstrated far greater consistency across ancestry groups compared with the best-performing published model[23].

PDAC is a rapidly fatal malignancy, and its early detection remains an intractable challenge[50,51] but can be improved with better risk stratification to identify individuals at elevated risk among whom cancer surveillance and screening can be implemented to improve outcomes[52,53]. Although studies of rare hereditary risk variants for PDAC have translated into clinical genetic testing, a critical barrier is that there is far less known for the ≥90% of patients who do not carry rare germline pathogenic variants in cancer predisposition genes[15,19,54,55]. A viable PRS model for PDAC could be an important adjunct to early detection efforts. However, the potential translation of PRSs to clinically useful tools to identify individuals at elevated PDAC risk has progressed slowly.

Here, we evaluated existing PDAC PRSs alongside newly developed models to identify the best-performing approach for risk stratification. A full covariate-based model demonstrated good discrimination (AUC = 0.74) and could be readily applicable in clinical settings. Adding the PRS yielded a modest but statistically significant improvement (AUC = 0.76). Thus, established risk factors alone provide substantial predictive value, while the incremental benefit of PRS should be interpreted within a multifactorial risk framework. At present, the modest gain in performance does not clearly justify clinical implementation of the genome-wide LDpred2-based PRS beyond the covariate-only model, particularly considering cost and feasibility. However, further refinement through iterative incorporation of additional risk markers and external validation may enhance performance and clinical utility. As genotyping becomes more integrated into clinical practice, PRS-based models may thus have increased relevance in the future. Further, our LDpred2-based PRS showed improved cross-ancestry transferability, particularly among individuals of African and Admixed American ancestries, even after ancestry-specific normalization. However, given the limited sample sizes in these groups, these findings should be interpreted with caution and require validation in larger, more diverse cohorts.

In our study population, the LDpred2-based PRS outperformed the PRS-CS model despite both being genome-wide approaches. It is known that the relative performance of genome-wide PRS approaches depends on factors such as the LD reference panel, variant set, hyperparameter tuning, and the underlying genetic architecture of the trait. LDpred2 and PRS-CS rely on different Bayesian priors and tuning strategies: LDpred2 tunes parameters such as polygenicity and SNP heritability, with an optional sparse formulation, whereas PRS-CS uses a continuous-shrinkage prior with either tuned or automatically learned global shrinkage[45,46]. Accordingly, our findings suggest that, under our analytic conditions, LDpred2 provided a better fit for PDAC risk prediction than PRS-CS.



Previous PDAC PRSs vary widely in design, SNP selection, and covariate adjustment, limiting direct comparison[23,32–44]. Early models based on GWAS-significant SNPs reported about 2–3-fold higher PDAC risk for individuals in the highest PRS strata[23,36,41,43], though many did not adjust for established PDAC risk factors[23,36] or may be affected by sample overlap bias[41]. Further, most were based on summary statistics from the same GWAS (Klein et al. 2018), including its own PRS[23], and they all had similar effect sizes and some had overlapping samples. To address these issues, we evaluated existing PRSs on a unified scale and compared them with genome-wide approaches. The genome-wide LDpred2-based PRS showed slightly higher effect size but similar discrimination to the Klein et al. PRS[23]. Furthermore, despite potential effect size inflation due to

Fig. 1 | Associations of newly constructed Bayesian-based polygenic risk scores (PRS) and previously published models with pancreatic ductal adenocarcinoma (PDAC) risk. The X-axis shows odds ratios (ORs) and 95% confidence intervals (CIs) per one standard deviation increase in the PRS. Bayesian models estimated using the leave-Mayo-out GWAS are displayed above the black horizontal line, and previously published models are shown below. LDpred2 and PRS-CS employ Bayesian frameworks to infer the optimal number of SNPs to include, whereas the other PRSs are based on fixed genome-wide significance thresholds ($p < 5 \times 10^{-08}$) or a less stringent p value ($<1 \times 10^{-05}$). The **base model + PRS** (yellow color) evaluates the association between the PRS and PDAC risk, adjusting for age (continuous), sex, and the top 10 ancestry principal components (continuous). The **full model + PRS** (blue color) includes additional adjustment for body mass index (continuous), first-degree family history of pancreatic cancer (yes/no), heavy alcohol use (yes/no), chronic pancreatitis (yes/no), ABO blood group (O/non-O), smoking history (never/former/current), and diabetes mellitus (yes/no). The PRSs by Nakatochi et al. (2018) and Bogumil et al. (2020) were derived from Japanese and multi-ethnic populations, respectively. The rest were based on European ancestry populations. CI confidence interval, M million, OR odds ratio, PDAC pancreatic ductal adenocarcinoma.

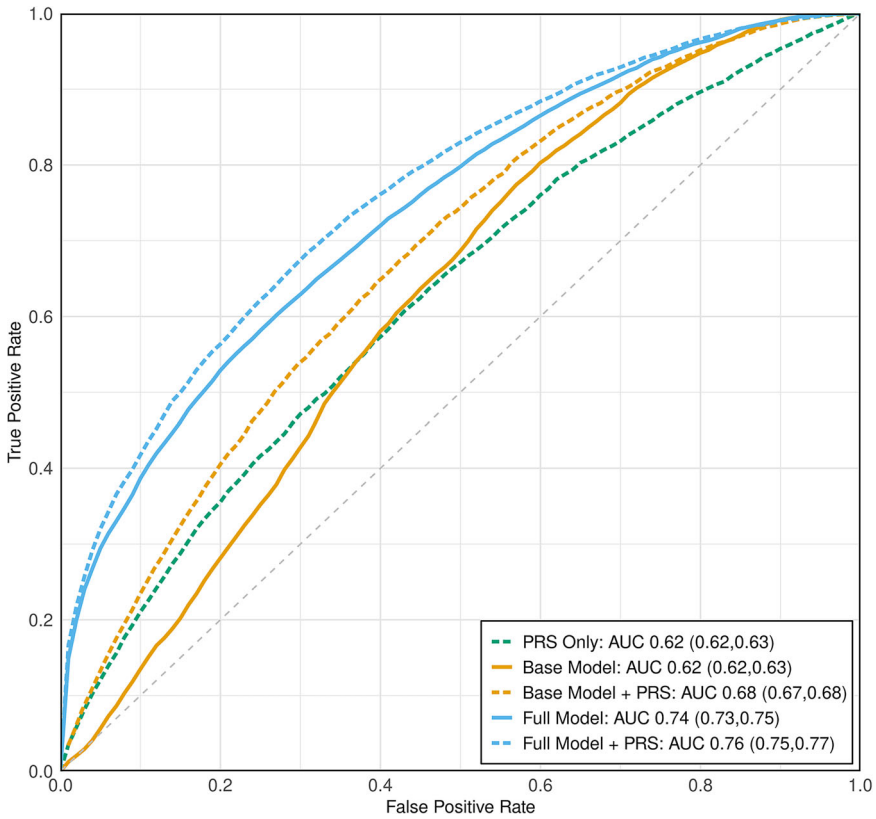


Fig. 2 | Area under the receiver operating characteristic (AUC) curves for prediction of pancreatic adenocarcinoma risk. PRS only (dashed green line) comprised the LDpred2 polygenic risk score (PRS, continuous) and top 10 principal components (continuous). **Base Model** (solid yellow line) comprised age (continuous), sex, and the top 10 principal components of genetic ancestry (continuous). **Base Model + PRS** (dashed yellow line) comprised variables in the base model and the LDpred2 polygenic risk score. **Full Model** (solid blue line) comprised age (continuous), sex, top 10 principal components (continuous), body mass index (continuous), first-degree family history of pancreatic cancer (yes, no), heavy alcohol use (yes, no), chronic pancreatitis (yes, no), ABO blood group (O, non-O), smoking history (never, former, current), and type 2 diabetes mellitus (yes, no). **Full Model + PRS** (dashed blue line) included all variables in the full model and the LDpred2 polygenic risk score.

sample overlap, the Klein et al. PRS still showed a smaller effect than the LDpred2 PRS, which was derived without such overlap. Also, LDpred2 demonstrated more consistent performance across non-European ancestries than the Klein et al. PRS, suggesting potential advantages for cross-ancestry generalization for genome-wide approaches, though this requires verification in other studies.

Key strengths of our study include the large sample size and careful clinical phenotyping of the PDAC cases through individual medical record review by trained abstractors, supervised by subspecialty physicians. Using LDpred2 to derive the PRS allowed us to incorporate latent risk SNPs beyond GWAS-significant variants, expanding on previous PRSs that were largely restricted to genome-wide significant SNPs. We also adjusted for all well-established clinical and lifestyle risk factors for PDAC, reflecting relevant effects under likely clinical risk modeling. The use of an ultra-rapid case-finding process ensured that approximately 86% of our PDAC cases were recruited within 30 days of diagnosis, with an overall high case response rate (68%), minimizing the potential impact of survival bias[56,57]. Additionally, we performed validation analyses of previously published PRS models in an independent sample by recalculating risk estimates for GWAS-significant SNPs identified in the PanScan and PanC4 GWASs after excluding Mayo Clinic samples and using the Mayo samples for independent assessment of the previously published models.

Limitations of our study include our predominantly European ancestry population (96%), although this didn't seem to limit the performance of the LDpred2 model in minority populations. A limitation of the PRS model is that it may reflect multiple underlying pathways or mechanisms that influence PDAC risk that may not be easily detectable for targeting prevention or interception. Further, we did not perform an exhaustive benchmarking of all available genome-wide PRS methods. While no single approach has yet demonstrated consistent superiority across cancers[58–60], it remains possible that alternative methods could have higher predictive performance for PDAC risk. Because of the relative infrequency of PDAC, perhaps more than large numbers of genetic risk variants or few common risk factors might be needed to detect appreciable risk magnitude to enhance risk stratification for this cancer[44,61]. The modifiable risk factors included in our model, such as smoking, BMI, and alcohol use, were measured at a single time point and thus may not reflect cumulative lifetime exposure. Additionally, BMI was self-reported, limiting granularity, and may explain the inverse association observed between BMI and PDAC in this study. The retrospective case-control design also introduces the possibility of recall bias in reporting prior exposures. Our design did not allow for direct estimation of absolute risk, which should be addressed in future prospective cohorts. Also, our study cohort was drawn from a tertiary clinic with high referral rates and might not be a perfect representation of the general population.

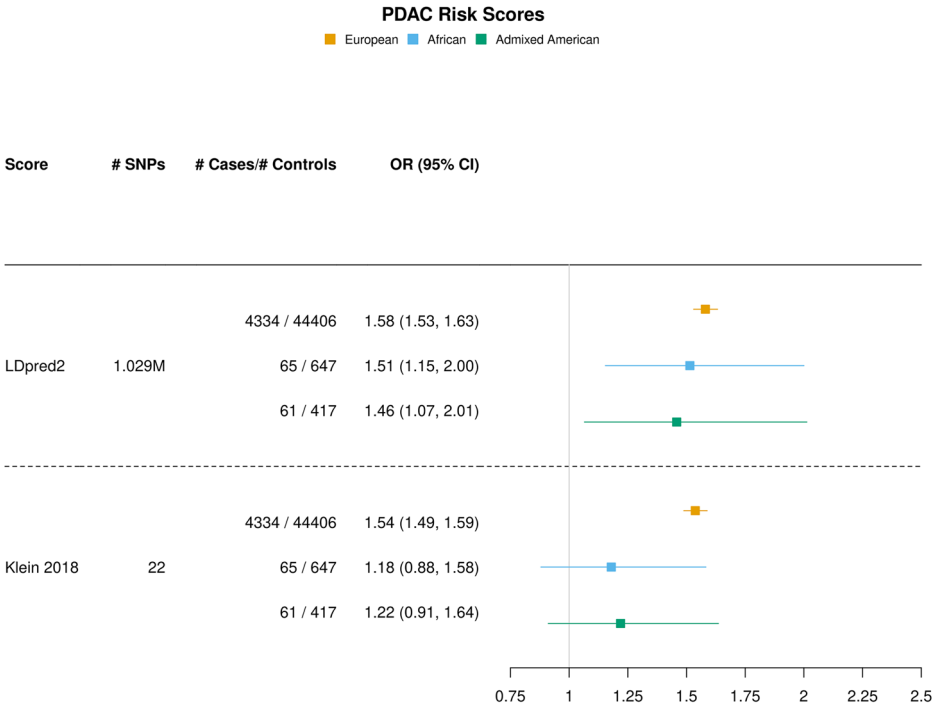


Fig. 3 | Associations of a newly constructed Bayesian-based LDpred2 polygenic risk score (PRS) and a previously published PRS (Klein et al., 2018) with risk of pancreatic ductal adenocarcinoma (PDAC), stratified by genetic ancestry (European, African, and Admixed American). The X-axis displays odds ratios (ORs) and 95% confidence intervals (CIs) per one standard deviation increase in the PRS. Colored bars denote ancestry groups: yellow for European, blue for African, and green for Admixed American participants. Ancestry-specific normalization was performed, adjusting scores using ancestry-specific mean and standard deviations.

In summary, we evaluated previously published PRS models for PDAC and developed a new genome-wide PRS using a Bayesian modeling framework (LDpred2). Under comparable conditions, the LDpred2-based PRS demonstrated a slightly higher effect size, but a similar AUC compared with the second best-performing published model. Notably, the LDpred2 PRS exhibited greater consistency across ancestries. Further refinement of the model could enhance risk stratification, enabling targeted clinical surveillance and screening for moderate-to-high-risk individuals and ultimately supporting earlier detection of this highly lethal cancer across ancestry groups.

Methods

Study population

Genetic and risk factor data were obtained from patients recruited into the Mayo Clinic Biospecimen Resource for Pancreas Research registry between October 2000 and October 2020 who had a pathologically or radiographically confirmed diagnosis of incident PDAC[56,57,62]. Genetic and risk factor data on control patients recruited from primary care clinics at Mayo Clinic between April 2009 and March 2016 and who did not have a history of pancreatic cancer were obtained from the Mayo Clinic Biobank[63]. Participation rates were 68% for the Mayo Clinic Biospecimen Resource for Pancreas Research registry and 21% for the Mayo Clinic Biobank. Details on the study design and recruitment procedures have been published previously[56,57,63]. The Mayo Clinic Biospecimen Resource for Pancreas Research registry and the Mayo Clinic Biobank used similar risk factor questionnaires to solicit information on participant demographics, usual adult weight and height, alcohol use, smoking history, and type II diabetes. We calculated body mass index (BMI, kg/m^2) based on self-reported usual adult weight and height, and abstracted data on chronic pancreatitis from individual medical records. Participants' ABO blood group was genetically determined as non-O versus O using two well-established ABO SNPs (rs505922/rs687289 and rs8176746), as described previously[57]. We also obtained previously published summary statistics and individual-level GWAS data from the PanScan and PanC4 consortia on PDAC cases and controls[19–24,64]. This study was conducted in accordance with the ethical principles of the Declaration of Helsinki and the Declaration of Istanbul. All participants provided written informed consent, including consent for the use of their biospecimens and data in future research studies. The study protocol was approved by the Mayo Clinic Institutional Review Board (IRB #22-001158; approval date June 5, 2023).

Genotyping

Peripheral blood DNA from Mayo Clinic PDAC cases and controls was sequenced by the Regeneron Genetics Center (RGC, Tarrytown, NY) using a custom capture panel as part of the Mayo Clinic Project Generation initiative. The panel incorporated exome regions and common variant coverage through a genotyping-by-sequencing approach[65,66]. Downstream processing and quality control (QC) procedures included variants exclusion if they failed to meet minimum thresholds for call rate (< 95%), minor allele frequency (MAF, <0.5%), or deviation from Hardy-Weinberg equilibrium ($p < 1 \times 10^{-6}$). Sample-level exclusions were made for individuals exhibiting high rates of missing genotypes (> 5%), sex discrepancies, or evidence of aberrant heterozygosity (< 70% across multiple chromosomes)[66]. Relatedness filtering was conducted through iterative analysis with PLINK and PRIMUS, identifying and removing individuals with excess kinship (PI_HAT > 0.1875 with >100 relatives or PI_HAT > 0.08 with >25,000 relatives)[66]. When a second-degree or closer relationship was observed between cases and controls, we retained the case and excluded the control(s). When two cases or two controls had a second-degree or closer relationship, the one with more complete covariate data was retained. Imputation was performed via the TOPMed Imputation Server[67,68], and only variants with high imputation quality (dosage R^2 >0.8) and MAF ≥ 5% were retained. After QC, data from 4508 PDAC cases and 46,189 controls remained for the final analysis (Supplementary Table 1).

Leave-Mayo-out GWAS analysis

Because Mayo Clinic samples contributed to the GWASs used in developing many of the published PRS models (through inclusion in PanScan and PanC4), potential bias could arise from sample overlap influencing SNP selection and effect size estimates. To address this, we leveraged data from a previously published PanScan I-III and PanC4 GWAS, as described by Klein et al.[23], to generate SNP weights after excluding Mayo Clinic participants. Each genotyping array was re-imputed using the TOPMed Imputation Server to ensure consistency across datasets. Mayo samples (1509

cases and 1865 controls) were excluded for use in an independent validation cohort, resulting in a reanalyzed dataset of 7531 cases and 10,631 controls for updating SNP weights. GWAS analysis was then re-run using PLINK2, adjusting for the same covariates as in the original analysis (i.e., study site, geographic region, age, sex, and ancestry principal components [PCs])[23]. Per-array summary statistics were combined using a fixed-effects inverse-variance meta-analysis in METAL[69]. Supplementary Figure 1 shows Manhattan plots comparing the PanScan I-III and PanC4 GWAS results and those obtained after excluding Mayo Clinic participants.

Published PRS models for PDAC

We manually curated published PDAC PRS models from the literature[23,32,41,42,70] and the Polygenic Score (PGS) Catalog[71,72]. Four palindromic SNPs were identified and replaced with proxy SNPs in high linkage disequilibrium ($r^2 > 0.98$). Details of each model, including the number of SNPs used, and the source GWAS from which SNPs were derived, are provided in Supplementary Table 2. The overlap of SNPs across models and information on proxy SNPs used to replace those found to be palindromic are also provided in Supplementary Fig. 2.

PRSs for PDAC

We estimated PRSs for PDAC in the Mayo Clinic sample using GWAS summary statistics derived from the reanalyzed PanScan I-III and PanC4 datasets. Three complementary approaches were applied. First, we implemented LDpred2-auto[73], a Bayesian shrinkage method that adjusts SNP effect sizes based on genome-wide LD structure using the HapMap3+ European LD reference panel, automatically estimating key model parameters without requiring external tuning samples. Second, we used PRS-CS-auto[69], another Bayesian shrinkage method that employs a continuous shrinkage prior to infer SNP weights, optimized with the 1000 Genomes Project European LD reference panel provided in the PRS-CS GitHub. Finally, we applied PRSice-2[47] using the traditional clumping-and-thresholding method, with p value thresholds of 1×10^{-5} and 5×10^{-8} and default clumping parameters.

Imputation of covariate data

To account for missing values in covariates (Supplementary Fig. 3), we performed multiple imputation by chained equations using the *mice* package in the R software[42] incorporating the entire set of covariates (age, BMI, sex, family history of pancreatic cancer, alcohol use, chronic pancreatitis, ABO blood group, smoking status, type II diabetes), PRS, and case-control status in the imputation model. Following iterative chained-equation modeling, 30 imputed datasets were generated.

Statistical analyses

We fit the following logistic regression models treating case-control status as the outcome: (1) a **base model** that included age, sex, and top 10 ancestry PCs; (2) a **base model** + **PRS** that added each PRS of interest separately to the base model; (3) a **full model** that added the following covariates to the base model: BMI, family history of pancreatic cancer, alcohol use, chronic pancreatitis, ABO blood group, smoking status, and type II diabetes; and (4) a **full model** + **PRS** that added each PRS to the full model. This approach allowed comparisons between PRS-only, covariates-only, and integrated models to determine whether the PRS improves prediction beyond what can be predicted with covariates alone. Logistic regression model results were reported as odds ratios (ORs) with 95% confidence intervals (CIs). Each model was fitted to each imputed dataset, and results were pooled using Rubin's rules to account for uncertainty induced by missingness. Model discrimination was assessed using the area under the receiver operating characteristic curve (AUC), and final estimates were pooled as log-transformed values across multiple imputations. We compared the performance of the published PRS models with novel PRS models estimated using LDpred2, PRSCS, and PRSice-2. To assess the potential for clinical decision-making, we assessed risk for individuals in the top five percent of the PRS, determined based on the distribution of the PRS in controls only. To compare association of the PRS across ancestries, we repeated the models above within each genetically determined ancestry (European, African, and Admixed American) having effective sample size $(4 \times p \times [1 - p] \times N$, where p is the sample prevalence) greater than 100. For these analyses, we performed ancestry-specific normalization, adjusting scores using ancestry-specific mean and standard deviation. All statistical analyses were performed using R v.4.4.1 (R CORE Team, Vienna, Austria).

Data availability

The Pancreatic Cancer Cohort Consortium (PanScan) and Pancreatic Cancer Case–Control Consortium (PanC4) GWAS data used to develop the polygenic risk scores, have been deposited in dbGaP (accession numbers phs000206.v5.p3 and phs000648.v1.p1, respectively). Individual-level genetic data from the Project Generation cohort used for validation analyses are subject to contractual restrictions and are not publicly available. These data were generated by the Regeneron Genetics Center as part of a collaboration with Mayo Clinic and may be accessed only by approved investigators within secure Mayo Clinic computing environments, in accordance with institutional and data use agreements. Summary-level results supporting the findings of this study for the validation analyses are available from the corresponding author upon reasonable request, subject to institutional and contractual approvals. Code Availability: The underlying code for this study is not publicly available but may be made available to qualified researchers on reasonable request to the corresponding author.

Code availability

The underlying code for this study is not publicly available but may be made available to qualified researchers on reasonable request to the corresponding author.

Received: 6 January 2026; Accepted: 1 May 2026;
Published online: 27 May 2026

References

1. Siegel, R. L., Kratzer, T. B., Giaquinto, A. N., Sung, H. & Jemal, A. Cancer statistics, 2025. *CA Cancer J. Clin.* **75**, 10–45 (2025).
2. Rahib, L. et al. Projecting cancer incidence and deaths to 2030: The unexpected burden of thyroid, liver, and pancreas cancers in the United States. *Cancer Res.* **74**, 2913–2921 (2014).
3. International Agency for Research on Cancer, World Health Organization. Cancer Today: Cancer Fact Sheets. https://gco.iarc.fr/today/en/fact-sheets-cancers. Accessed November 21, 2025.
4. Owens, D. K. et al. Screening for Pancreatic Cancer: US Preventive Services Task Force Reaffirmation Recommendation Statement. *Jama* **322**, 438–444 (2019).
5. Goggins, M. et al. Management of patients with increased risk for familial pancreatic cancer: updated recommendations from the International Cancer of the Pancreas Screening (CAPS) Consortium. *Gut* **69**, 7–17 (2020).
6. Sawhney, M. S. et al. ASGE guideline on screening for pancreatic cancer in individuals with genetic susceptibility: Summary and recommendations. *Gastrointest. Endosc.* **95**, 817–826 (2022).
7. Biller, L. H., Wolpin, B. M. & Goggins, M. Inherited pancreatic cancer syndromes and high-risk screening. *Surg. Oncol. Clin. N. Am.* **30**, 773–786 (2021).
8. Antwi, S. O. et al. Risk of different cancers among first-degree relatives of pancreatic cancer patients: Influence of probands' susceptibility gene mutation status. *J. Natl. Cancer Inst.* **111**, 264–271 (2019).
9. Klein, A. P. et al. Prospective risk of pancreatic cancer in familial pancreatic cancer kindreds. *Cancer Res* **64**, 2634–2638 (2004).
10. Waddell, N. et al. Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* **518**, 495–501 (2015).

11. Hruban, R. H., Goggins, M. & Kern, S. E. Molecular genetics and related developments in pancreatic cancer. *Curr. Opin. Gastroenterol.* **15**, 404–409 (1999).
12. Hayashi, A., Hong, J. & Iacobuzio-Donahue, C. A. The pancreatic cancer genome revisited. *Nat. Rev. Gastroenterol. Hepatol.* **18**, 469–481 (2021).
13. Raphael, B. J. et al. Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer Cell* **32**, 185–203.e113 (2017).
14. Witkiewicz, A. K. et al. Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. *Nat. Commun.* **6**, 6744 (2015).
15. Hu, C. et al. Association between inherited germline mutations in cancer predisposition genes and risk of pancreatic cancer. *JAMA* **319**, 2401–2409 (2018).
16. Shindo, K. et al. Deleterious germline mutations in patients with apparently sporadic pancreatic adenocarcinoma. *J. Clin. Oncol.* **35**, 3382–3390 (2017).
17. Yu, Y. et al. A whole-exome case-control association study to characterize the contribution of rare coding variation to pancreatic cancer risk. *HGG Adv.* **3**, 100078 (2022).
18. Farinella, R. et al. From classical approaches to artificial intelligence, old and new tools for PDAC risk stratification and prediction. *Semin Cancer Biol.* **112**, 71–92 (2025).
19. Amundadottir, L. et al. Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nat. Genet* **41**, 986–990 (2009).
20. Petersen, G. M. et al. A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. *Nat. Genet* **42**, 224–228 (2010).
21. Wolpin, B. M. et al. Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer. *Nat. Genet* **46**, 994–1000 (2014).
22. Childs, E. J. et al. Common variation at 2p13.3, 3q29, 7p13 and 17q25.1 associated with susceptibility to pancreatic cancer. *Nat. Genet* **47**, 911–916 (2015).
23. Klein, A. P. et al. Genome-wide meta-analysis identifies five new susceptibility loci for pancreatic cancer. *Nat. Commun.* **9**, 556 (2018).
24. Zhang, M. et al. Three new pancreatic cancer susceptibility signals identified on chromosomes 1q32.1, 5p15.33 and 8q24.21. *Oncotarget* **7**, 66328–66343 (2016).
25. Wu, C. et al. Genome-wide association study identifies five loci associated with susceptibility to pancreatic cancer in Chinese populations. *Nat. Genet* **44**, 62–66 (2011).
26. Low, S. K. et al. Genome-wide association study of pancreatic cancer in Japanese population. *PLoS One* **5**, e11824 (2010).
27. Lin, Y. et al. Genome-wide association meta-analysis identifies GP2 gene risk variants for pancreatic cancer. *Nat. Commun.* **11**, 3175 (2020).
28. Campa, D. et al. TERT gene harbors multiple variants associated with pancreatic cancer susceptibility. *Int J. Cancer* **137**, 2175–2183 (2015).
29. Wang, Z. et al. Imputation and subset-based association analysis across different cancer types identifies multiple independent risk loci in the TERT-CLPTM1L region on chromosome 5p15.33. *Hum. Mol. Genet* **23**, 6616–6633 (2014).
30. Campa, D. et al. Lack of replication of seven pancreatic cancer susceptibility loci identified in two Asian populations. *Cancer Epidemiol. Biomark. Prev.* **22**, 320–323 (2013).
31. Chari, S. T. et al. Early detection of sporadic pancreatic cancer: summative review. *Pancreas* **44**, 693–712 (2015).
32. Nakatochi, M. et al. Prediction model for pancreatic cancer risk in the general Japanese population. *PLoS One* **13**, e0203386 (2018).
33. Shi, Z. et al. Systematic evaluation of cancer-specific genetic risk score for 11 types of cancer in The Cancer Genome Atlas and Electronic Medical Records and Genomics cohorts. *Cancer Med.* **8**, 3196–3205 (2019).
34. Fritsche, L. G. 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* **107**, 815–836 (2020).
35. Kim, E. S. et al. Potential utility of risk stratification for multicancer screening with liquid biopsy tests. *NPJ Precis Oncol.* **7**, 39 (2023).
36. Jia, G. et al. Evaluating the utility of polygenic risk scores in identifying high-risk individuals for eight common cancers. *JNCI Cancer Spectr.* **4**, pkaa021 (2020).
37. Graff, R. E. et al. Cross-cancer evaluation of polygenic risk scores for 16 cancer types in two large cohorts. *Nat. Commun.* **12**, 970 (2021).
38. Yuan, C. et al. The age-dependent association of risk factors with pancreatic cancer. *Ann. Oncol.* **33**, 693–701 (2022).
39. Kachuri, L. et al. Pan-cancer analysis demonstrates that integrating polygenic risk scores with modifiable risk factors improves risk prediction. *Nat. Commun.* **11**, 6084 (2020).
40. Kim, J. et al. Genetic and circulating biomarker data improve risk prediction for pancreatic cancer in the general population. *Cancer Epidemiol. Biomark. Prev.* **29**, 999–1008 (2020).
41. Galeotti, A. A. et al. Polygenic and multifactorial scores for pancreatic ductal adenocarcinoma risk prediction. *J. Med. Genet* **58**, 369–377 (2021).
42. Bogumil, D. et al. Replication and genetic risk score analysis for pancreatic cancer in a diverse multiethnic population. *Cancer Epidemiol. Biomark. Prev.* **29**, 2686–2692 (2020).
43. Sharma, S., Tapper, W. J., Collins, A. & Hamady, Z. Z. R. Predicting pancreatic cancer in the UK biobank cohort using polygenic risk scores and diabetes mellitus. *Gastroenterology* **162**, 1665–1674.e1662 (2022).
44. Risch, H. A., Yu, H., Lu, L. & Kidd, M. S. Detectable symptomatology preceding the diagnosis of pancreatic cancer and absolute risk of pancreatic cancer diagnosis. *Am. J. Epidemiol.* **182**, 26–34 (2015).
45. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: Better, faster, stronger. *Bioinformatics* **36**, 5424–5431 (2021).
46. Ge, T., Chen, C. Y., Ni, Y., Feng, Y. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776 (2019).
47. Choi, S. W. & O'Reilly, P. F. PRSice-2: Polygenic risk score software for biobank-scale data. *Gigascience* **8**. https://doi.org/10.1093/gigascience/giz082 (2019).
48. Briggs, S. E. W. et al. Integrating genome-wide polygenic risk scores and non-genetic risk to predict colorectal cancer diagnosis using UK Biobank data: population based cohort study. *BMJ* **379**, e071707 (2022).
49. Kang, J. H. et al. Polygenic risk and rare variant gene clustering enhance cancer risk stratification for breast and prostate cancers. *Commun. Biol.* **7**, 1289 (2024).
50. Singhi, A. D., Koay, E. J., Chari, S. T. & Maitra, A. Early detection of pancreatic cancer: Opportunities and challenges. *Gastroenterology* **156**, 2024–2040 (2019).
51. Stoffel, E. M., Brand, R. E. & Goggins, M. Pancreatic cancer: Changing epidemiology and new approaches to risk assessment, early detection, and prevention. *Gastroenterology* **164**, 752–765 (2023).
52. Canto, M. I. et al. Risk of neoplastic progression in individuals at high risk for pancreatic cancer undergoing long-term surveillance. *Gastroenterology* **155**, 740–751.e742 (2018).
53. Vasen, H. 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.* **34**, 2010–2019 (2016).
54. Petersen, G. M. Familial pancreatic cancer. *Semin Oncol.* **43**, 548–553 (2016).

55. Tempero, M. A. NCCN guidelines updates: Pancreatic cancer. *J. Natl. Compr. Canc Netw.* **17**, 603–605 (2019).
56. Antwi, S. O. et al. Pancreatic cancer: Associations of inflammatory potential of diet, cigarette smoking and long-standing diabetes. *Carcinogenesis* **37**, 481–490 (2016).
57. Antwi, S. O. et al. Influence of cancer susceptibility gene mutations and ABO blood group of pancreatic cancer probands on concomitant risk to first-degree relatives. *Cancer Epidemiol. Biomark. Prev.* **31**, 372–381 (2022).
58. Pain, O. et al. Evaluation of polygenic prediction methodology within a reference-standardized framework. *PLoS Genet* **17**, e1009021 (2021).
59. Zheng, Z. et al. Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex traits within and between ancestries. *Nat. Genet* **56**, 767–777 (2024).
60. Jayasinghe, D., Eshetie, S., Beckmann, K., Benyamin, B. & Lee, S. H. Advancements and limitations in polygenic risk score methods for genomic prediction: a scoping review. *Hum. Genet* **143**, 1401–1431 (2024).
61. Wentzensen, N. & Wacholder, S. From differences in means between cases and controls to risk stratification: a business plan for biomarker development. *Cancer Discov.* **3**, 148–157 (2013).
62. Antwi, S. O. et al. Exposure to environmental chemicals and heavy metals, and risk of pancreatic cancer. *Cancer Causes Control* **26**, 1583–1591 (2015).
63. Olson, J. E. et al. Characteristics and utilisation of the Mayo Clinic Biobank, a clinic-based prospective collection in the USA: Cohort profile. *BMJ Open* **9**, e032707 (2019).
64. King, S. D. et al. Genetic susceptibility to nonalcoholic fatty liver disease and risk for pancreatic cancer: Mendelian randomization. *Cancer Epidemiol. Biomark. Prev.* **32**, 1265–1269 (2023).
65. Gaynor, S. M. et al. Yield of genetic association signals from genomes, exomes and imputation in the UK Biobank. *Nat. Genet* **56**, 2345–2351 (2024).
66. Jazieh, K. et al. Alterations in the CTRB2 gene and response to chemotherapy in pancreatic cancer. *PLoS One* **21**, e0343022 (2026).
67. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
68. Sun, Q. et al. Leveraging TOPMed imputation server and constructing a cohort-specific imputation reference panel to enhance genotype imputation among cystic fibrosis patients. *HGG Adv.* **3**, 100090 (2022).
69. Ni, G. et al. A Comparison of ten polygenic score methods for psychiatric disorders applied across multiple cohorts. *Biol. Psychiatry* **90**, 611–620 (2021).
70. Klein, A. P. et al. An absolute risk model to identify individuals at elevated risk for pancreatic cancer in the general population. *PLoS One* **8**, e72311 (2013).
71. Lambert, S. A. et al. Enhancing the polygenic score catalog with tools for score calculation and ancestry normalization. *Nat. Genet* **56**, 1989–1994 (2024).
72. Lambert, S. A. et al. The polygenic score catalog as an open database for reproducibility and systematic evaluation. *Nat. Genet* **53**, 420–425 (2021).
73. Privé, F., Albiñana, C., Arbel, J., Pasaniuc, B. & Vilhjálmsson, B. J. Inferring disease architecture and predictive ability with LDpred2-auto. *Am. J. Hum. Genet* **110**, 2042–2055 (2023).

Acknowledgements

The late Dr. Gloria M. Petersen played a foundational role in the conception and design of this study, but, regrettably, did not live to witness its completion. We dedicate this work to Dr. Petersen, whose legacy is defined by her intellectual rigor, visionary leadership, and unwavering commitment to scientific discovery. We also extend our sincere gratitude to Dr. Donghui Li, formerly of the MD Anderson Cancer Center in Houston, Texas, whose leadership and scientific vision were instrumental in establishing the PanC4 consortium and ensuring its continued success. We wish her a rewarding and enjoyable retirement. We further extend our deepest gratitude to the pancreatic cancer patients and cancer-free controls whose contributions of biospecimens and data through the various registries and the PanScan and PanC4 consortia made this research possible. We also acknowledge the Mayo Clinic Biobank (MCB) and its dedicated research team, as well as the patient-participants who made the MCB program possible. Support for the MCB was provided by the Mayo Clinic Center for Individualized Medicine. We further acknowledge the Regeneron Genetics Center for generating and providing genetic data for the Mayo Clinic participants included in this study. Funding sources: This work was supported by the National Cancer Institute at the National Institutes of Health [R01CA272668, P30CA15083, P50CA102701, K01CA23787, U01CA210138]; the Lustgarten Foundation for Pancreatic Cancer Research; and the Genetic Epidemiology Risk Assessment Program for Pancreatic Cancer honoring Dr. Gloria Petersen. This study was also supported in part by the Intramural Research Program of the NIH. The contributions of the NIH author(s) are considered Works of the United States Government. The findings and conclusions presented in this paper are those of the authors and do not necessarily reflect the views of the NIH, the U.S. Department of Health and Human Services, or the other funding sources. Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer/World Health Organization. The funders did not play any role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Author contributions

Conceptualization: A.L.O., S.O.A., B.J.C., and N.B.L. Data curation: E.E.C., K.G.R., H.J.A., B.J.C., and N.B.L. Formal analysis: E.E.C., K.G.R., H.J.A., B.J.C., and N.B.L. Funding acquisition: A.L.O., S.O.A., and S.M. Investigation: A.L.O., S.O.A., B.J.C., N.B.L., L.T.A., A.P.K., D.J.S., S.M., E.E.C., K.G.R., and H.J.A. Methodology: A.L.O., S.O.A., B.J.C., N.B.L., E.E.C., K.G.R., and H.J.A. Software: B.J.C., N.B.L., E.E.C., K.G.R., and H.J.A. Project administration: A.L.O. and S.O.A. Resources: A.L.O. Visualization: E.E.C., H.J.A., K.G.R., and B.J.C. Writing – original draft: S.O.A. Writing – review and editing: SOA, BJC, EEC, NBL, KGR, HJA, SM, WRB, DJS, JZ, DM, AAA, LEBF, PMB, FC, TT, JA, MD, SG, PJG, VK, DC, CK, LLM, REN, AVP, JWW, SP, XOS, KV, SKVDE, EW, WZ, DA, GA, PB, SJC, YC, BD, PF, ELG, MG, CH, MH, RJH, MRJ, PK, NM, SCM, KN, UP, MP, NR, MJS, HDS, DTS, MCS, RLM, CYU, HY, CY, MAYO-RGC, RGC, HAR, BMW, RZSS, APK, LTA, ALO. All authors read and approved the final manuscript.

Competing interests

The Authors declare no Competing Non-Financial Interests, but the following competing financial interests exist. Mayo Clinic and Exact Sciences have an intellectual property development agreement under which Dr. Majumder is listed as an inventor and may receive future royalties as a Mayo Clinic employee.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41698-026-01479-x.

Correspondence and requests for materials should be addressed to Samuel O. Antwi or Ann L. Oberg.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2026

Samuel O. Antwi[1] ✉, Brandon J. Coombes[2], Erin E. Carlson[2], Nicholas B. Larson[2], Kari G. Rabe[2], Hunter J. Atkinson[2], Shounak Majumder[3], William R. Bamlet[2], Daniel J. Schaid[2], Jun Zhong[4], David McKean[5], Alan A. Arslan[6,7,8], Laura E. Beane Freeman[9], Paige M. Bracci[10], Federico Canzian[11], Thérèse Truong[12], Josh Atkins[13], Mengmeng Du[14], Steven Gallinger[15], Phyllis J. Goodman[16], Verena Katzke[17], Daniele Campa[18], Charles Kooperberg[19], Loic Le Marchand[20], Rachel E. Neale[21,22], Alpa V. Patel[23], Jean Wactawski-Wende[24], Sandra Perdomo[25], Xiao-Ou Shu[26], Kala Visvanathan[5,27], Stephen K. Van Den Eeden[28], Emily White[19], Wei Zheng[26], Demetrius Albanes[9], Gabriella Andreotti[9], Paul Brennan[25], Stephen J. Chanock[9], Yu Chen[7,8,29], Burcu Darst[19], Pietro Ferrari[30], Edward L. Giovannucci[31,32], Michael Goggins[33], Christopher Haiman[34], Manal Hassan[35], Rayjean J. Hung[15], Miranda R. Jones[27], Peter Kraft[9], Núria Malats[36], Steven C. Moore[9], Kimmie Ng[37], Ulrike Peters[19], Miquel Porta[38,39], Nathaniel Rothman[9], Maria-José Sánchez[40,41,42], Howard D. Sesso[31,43], Debra T. Silverman[9], Melissa C. Southey[44,45,46], Roger L. Milne[44,45,46], Caroline Y. Um[23], Herbert Yu[20], Chen Yuan[37], MAYO-RGC Project Generation*, Regeneron Genetics Center*, Harvey A. Risch[49], Brian M. Wolpin[37], Rachael Z. Stolzenberg-Solomon[9], Alison P. Klein[5,27,33], Laufey T. Amundadottir[4,9] & Ann L. Oberg[2] ✉

[1]Department of Quantitative Health Sciences, Mayo Clinic, Jacksonville, FL, USA. [2]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, USA. [3]Department of Medicine, Mayo Clinic, Rochester, MN, USA. [4]Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. [5]Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins School of Medicine, Baltimore, MD, USA. [6]Department of Obstetrics and Gynecology, New York University Grossman School of Medicine, New York, NY, USA. [7]Department of Population Health, New York University Grossman School of Medicine, New York, NY, USA. [8]Perlmutter Cancer Center, NYU Langone Health, New York, NY, USA. [9]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. [10]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, USA. [11]Genomic Epidemiology Group, German Cancer Research Center (DKFZ), Heidelberg, Germany. [12]Paris-Saclay University, UVSQ, Inserm, Gustave Roussy, CESP, Villejuif, France. [13]Nuffield Department of Population Health, Oxford University, Oxford, UK. [14]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [15]Lunenfeld-Tanenbaum Research Institute, Sinai Health System and University of Toronto, Toronto, ON, Canada. [16]SWOG Statistical Center, Fred Hutchinson Cancer Center, Seattle, WA, USA. [17]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany. [18]Department of Biology, University of Pisa, Pisa, PI, Italy. [19]Division of Public Health Sciences, Fred Hutchinson Cancer Center, Seattle, WA, USA. [20]Cancer Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI, USA. [21]Population Health Program, QIMR Berghofer, Brisbane, QLD, Australia. [22]School of Public Health, University of Queensland, Brisbane, QLD, Australia. [23]Department of Population Science, American Cancer Society, Atlanta, GA, USA. [24]University at Buffalo School of Public Health and Health Professions, Buffalo, NY, USA. [25]Genomic Epidemiology Branch, International Agency for Research on Cancer, Lyon, France. [26]Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN, USA. [27]Department of Epidemiology, Johns Hopkins School of Public Health, Baltimore, MD, USA. [28]Division of Research, Kaiser Permanente Northern California, Pleasanton, CA, USA. [29]New York University Grossman School of Medicine, New York, NY, USA. [30]Nutrition and Metabolism Branch, International Agency for Research on Cancer, World Health Organization, Lyon, France. [31]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [32]Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [33]Department of Pathology, Sol Goldman Pancreatic Cancer Research Center, Johns Hopkins School of Medicine, Baltimore, MD, USA. [34]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. [35]Department of Gastrointestinal Medical Oncology, University of Texas MD Anderson Cancer Center, Houston, TX, USA. [36]Genetic and Molecular Epidemiology Group, Spanish National Cancer Research Centre, Madrid, Spain. [37]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [38]CIBER Epidemiologíay Salud Pública (CIBERESP), Barcelona, Spain. [39]Hospital del Mar Institute of Medical Research (IMIM), Universitat Autònoma de Barcelona, Barcelona, Spain. [40]Escuela Andaluza de Salud Pública (EASP), Granada, Spain. [41]Instituto de Investigación Biosanitaria ibs.GRANADA, Granada, Spain. [42]Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP), Madrid, Spain. [43]Division of Preventive Medicine, Brigham and Women's Hospital, Boston, MA, USA. [44]Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, VIC, Australia. [45]Cancer Epidemiology Division, Cancer Council Victoria, Melbourne, VIC, Australia. [46]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, VIC, Australia. [49]Department of Chronic Disease Epidemiology, Yale School of Public Health, New Haven, CT, USA.
✉e-mail: Antwi.Samuel@mayo.edu; Ann.oberg@mayo.edu

MAYO-RGC Project Generation

James R. Cerhan[47], Fergus J. Couch[47], Nicholas B. Larson[47], Eric W. Klee[47], Zachary S. Fredericksen[47], Mine Cicek[47], Janet E. Olson[47], Lisa K. Colborn[47], Andrew J. Danielsen[47] & Jonathan J. Harrington[47]

[47]Mayo Clinic, Rochester, MN, USA.

Regeneron Genetics Center

Aris Baras[48], Gonçalo Abecasis[48], Adolfo Ferrando[48], Giovanni Coppola[48], Andrew Deubler[48], Luca A. Lotta[48], John D. Overton[48], Jeffrey G. Reid[48], Alan Shuldiner[48], Katherine Siminovitch[48], Jason Portnoy[48], Marcus B. Jones[48], Lyndon Mitnaul[48], Alison Fenney[48], Jonathan Marchini[48], Manuel Allen Revez Ferreira[48], Maya Ghoussaini[48], Mona Nafde[48], William Salerno[48], Cristen Willer[48], Lourdes Crane[48], Christina Beechert[48], Erin Fuller[48], Laura M. Cremona[48], Eugene Kalyuskin[48], Hang Du[48], Caitlin Forsythe[48], Zhenhua Gu[48], Kristy Guevara[48], Michael Lattari[48], Alexander Lopez[48], Kia Manoochehri[48], Prathyusha Challa[48], Manasi Pradhan[48], Raymond Reynoso[48], Ricardo Schiavo[48], Maria Sotiropoulos Padilla[48], Chenggu Wang[48], Sarah E. Wolf[48], Jeffrey G. Reid[48], Mona Nafde[48], Manan Goyal[48], George Mitra[48], Sanjay Sreeram[48], Rouel Lanche[48], Vrushali Mahajan[48], Sai Lakshmi Vasireddy[48], Gisu Eom[48], Krishna Pawan Punuru[48], Sujit Gokhale[48], Benjamin Sultan[48], Pooja Mule[48], Mudasar Sarwar[48], Muhammad Aqeel[48], Xiaodong Bai[48], Lance Zhang[48], Sean O'Keeffe[48], Razvan Panea[48], Evan Edelstein[48], Ayesha Rasool[48], Evan K. Maxwell[48], Boris Boutkov[48], Alexander Gorovits[48], Ju Guan[48], Lukas Habegger[48], Alicia Hawes[48], Olga Krasheninina[48], Samantha Zarate[48], Adam J. Mansfield[48], Joshua Backman[48], Kathy Burch[48], Adrian Campos[48], Liron Ganel[48], Sheila Gaynor[48], Benjamin Geraghty[48], Arkopravo Ghosh[48], Salvador Romero Martinez[48], Christopher Gillies[48], Lauren Gurski[48], Eric Jorgenson[48], Tyler Joseph[48], Michael Kessler[48], Jack Kosmicki[48], Adam Locke[48], Priyanka Nakka[48], Karl Landheer[48], Olivier Delaneau[48], Anthony Marcketta[48], Joelle Mbatchou[48], Arden Moscati[48], Anita Pandit[48], Jonathan Ross[48], Carlo Sidore[48], Eli Stahl[48], Timothy Thornton[48], Sailaja Vedantam[48], Rujin Wang[48], Kuan-Han Wu[48], Bin Ye[48], Blair Zhang[48], Andrey Ziyatdinov[48], Yuxin Zou[48], Jingning Zhang[48], Kyoko Watanabe[48], Mira Tang[48], Frank Wendt[48], Suganthi Balasubramanian[48], Suying Bao[48], Kathie Sun[48], Chuanyi Zhang[48], Sean Yu[48], Aaron Zhang[48], David Corrigan[48], Dhruv Shidhaye[48], Chen Wang[48], Keyrun Adhikari[48], Alexander Lachmann[48], Brian Hobbs[48], Jon Silver[48], William Palmer[48], Rita Guerreiro[48], Amit Joshi[48], Antoine Baldassari[48], Sarah Graham[48], Ernst Mayerhofer[48], Erola Pairo Castineira[48], Mary Haas[48], Niek Verweij[48], George Hindy[48], Jonas Bovijn[48], Tanima De[48], Luanluan Sun[48], Olukayode Sosina[48], Arthur Gilly[48], Peter Dornbos[48], Juan Rodriguez-Flores[48], Moeen Riaz[48], Manav Kapoor[48], Gannie Tzoneva[48], Momodou W. Jallow[48], Anna Alkelai[48], Ariane Ayer[48], Veera Rajagopal[48], Sahar Gelfman[48], Vijay Kumar[48], Jacqueline Otto[48], Jose Bras[48], Silvia Alvarez[48], Jessie Brown[48], Hossein Khiabanian[48], Joana Revez[48], Kimberly Skead[48], Valentina Zavala[48], Jae Soon Sul[48], Lei Chen[48], Sam Choi[48], Amy Damask[48], Nan Lin[48], Charles Paulding[48], Sameer Malhotra[48], Joseph Herman[48], Michelle G. LeBlanc[48], Nadia Rana[48], Jennifer Rico-Varela[48], Jaimee Hernandez[48], Larizbeth Romero[48], Ashley Paynter[48], Randi Schwartz[48], Jody Hankins[48], Anna Han[48], Samuel Hart[48], Ryan Smith[48], Ann Perez-Beals[48], Gina Solari[48], Johannie Rivera-Picart[48], Michelle Pagan[48] & Sunilbe Siceron[48]

[48]Regeneron Genetics Center, Tarrytown, NY, USA.