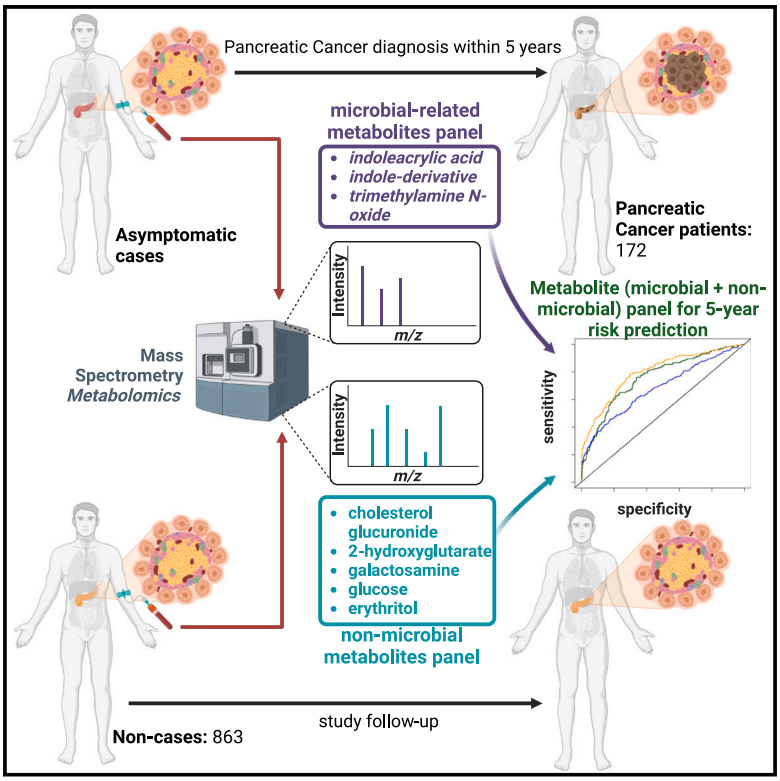
Cell Reports
Medicine

Article

A blood-based metabolomic signature predictive of risk for pancreatic cancer

Graphical abstract



Highlights

- Microbial metabolites in blood inform on risk of developing pancreatic cancer
- A metabolite panel predicts 5-year risk of pancreatic cancer
- The metabolite panel complements CA19-9 for risk prediction of pancreatic cancer

Authors

Ehsan Irajizad, Ana Kenney, Tiffany Tang, ..., Brian M. Wolpin, Sam Hanash, Johannes F. Fahrmann

Correspondence

shanash@mdanderson.org (S.H.), jffahrmann@mdanderson.org (J.F.F.)

In brief

Irajizad et al. report a blood-based metabolite panel that identifies individuals at high risk of developing pancreatic cancer within 5 years of blood draw. The metabolite panel provides a potential tool to identify individuals at high risk of pancreatic cancer who would benefit from surveillance and/or from potential cancer interception strategies.

Irajizad et al., 2023, Cell Reports Medicine *4*, 101194
September 19, 2023
https://doi.org/10.1016/j.xcrm.2023.101194

CellPress

Cell Reports Medicine

CellPress
OPEN ACCESS

Article

A blood-based metabolomic signature predictive of risk for pancreatic cancer

Ehsan Irajizad,[1,2] Ana Kenney,[3] Tiffany Tang,[3] Jody Vykoukal,[2] Ranran Wu,[2] Eunice Murage,[2] Jennifer B. Dennison,[2] Marta Sans,[5] James P. Long,[1] Maureen Loftus,[4] John A. Chabot,[8] Michael D. Kluger,[8] Fay Kastrinos,[8,9] Lauren Brais,[4] Ana Babic,[4] Kunal Jajoo,[5] Linda S. Lee,[5] Thomas E. Clancy,[6] Kimmie Ng,[4] Andrea Bullock,[7] Jeanine M. Genkinger,[9,10] Anirban Maitra,[11] Kim-Anh Do,[1] Bin Yu,[3] Brian M. Wolpin,[4] Sam Hanash,[2,12,*] and Johannes F. Fahrmann[2,12,13,*]

[1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[2]Department of Clinical Cancer Prevention, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[3]Department of Statistics, University of California, Berkeley, Berkeley, CA, USA
[4]Dana-Farber Brigham and Women's Cancer Center, Division of Gastrointestinal Oncology, Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA
[5]Division of Gastroenterology, Hepatology and Endoscopy, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
[6]Dana-Farber Brigham and Women's Cancer Center, Division of Surgical Oncology, Department of Surgery, Brigham and Women's Hospital, Harvard Medical School, Boston, MA USA
[7]Division of Hematology/Oncology, Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA
[8]Division of Digestive and Liver Diseases, Columbia University Irving Medical Cancer and the Vagelos College of Physicians and Surgeons, New York, NY, USA
[9]Herbert Irving Comprehensive Cancer Center, Columbia University Irving Medical Center, New York, NY, USA
[10]Department of Epidemiology, Columbia Mailman School of Public Health, New York, NY, USA
[11]Department of Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[12]These authors contributed equally
[13]Lead contact
*Correspondence: shanash@mdanderson.org (S.H.), jffahrmann@mdanderson.org (J.F.F.)
https://doi.org/10.1016/j.xcrm.2023.101194

SUMMARY

Emerging evidence implicates microbiome involvement in the development of pancreatic cancer (PaCa). Here, we investigate whether increases in circulating microbial-related metabolites associate with PaCa risk by applying metabolomics profiling to 172 sera collected within 5 years prior to PaCa diagnosis and 863 matched non-subject sera from participants in the Prostate, Lung, Colorectal, and Ovarian (PLCO) cohort. We develop a three-marker microbial-related metabolite panel to assess 5-year risk of PaCa. The addition of five non-microbial metabolites further improves 5-year risk prediction of PaCa. The combined metabolite panel complements CA19-9, and individuals with a combined metabolite panel + CA19-9 score in the top 2.5th percentile have absolute 5-year risk estimates of >13%. The risk prediction model based on circulating microbial and non-microbial metabolites provides a potential tool to identify individuals at high risk of PaCa that would benefit from surveillance and/or from potential cancer interception strategies.

INTRODUCTION

Pancreatic cancer is highly lethal and is projected to become the second leading cause of cancer death in the United States by 2040.[1] Surgical resection of localized disease represents the greatest chance for curative therapy. Unfortunately, only a minority (15%–20%) of patients present with surgically resectable disease.[2,3]

The low incidence of pancreatic cancer in the average-risk population (~8–12 per 100,000)[4,5] makes it challenging to implement effective screening programs for pancreatic cancer. The United States Preventative Services Task Force (USPSTF) currently recommends against screening for pancreatic cancer in the general population using any method.[6] Yet, the USPSTF recognizes that screening in persons who are at an increased risk may be warranted.[6] There remains an opportunity to develop blood-based signatures that can identify individuals at increased risk who would benefit from screening and, potentially, from preventive interventions.

The microbiota is a complex ecosystem integral to human health. Microbial diversity is site specific and varies depending on the organ location.[7] Increasing evidence suggests that alterations in the microbiome are associated with risk for certain cancers, including pancreatic cancer.[8] Studies suggest that loss of microbial diversity and community stability coupled with increases in pathogenic microbes increase cancer susceptibility.[9] In the

Table 1. Patient and tumor characteristics for PLCO cohort

	Subject/control subject status			
	Non-subject		Subject	
	N	%	N	%
Total	863	100	173	100
Gender				
Female	357	41.4	72	41.6
Male	506	58.6	101	58.4
Age at randomization				
≤59	183	21.2	37	21.4
60–64	206	23.9	41	23.7
65–69	321	37.2	64	37.0
≥70	153	17.7	31	17.9
Race				
White	783	90.7	157	90.8
Black	30	3.5	6	3.5
Other	50	5.8	10	5.9
Cigarette smoking status				
Never smoked cigarettes	420	48.7	63	36.4
Current cigarette smoker	74	8.6	36	20.8
Former cigarette smoker	369	42.8	74	42.8
BMI at baseline (in kg/m^2)				
Not answered	7	0.8	0	0.0
0–18.5	8	0.9	3	1.7
18.5–25	300	34.8	56	32.4
25–30	365	42.3	71	41.0
30+	183	21.2	43	24.9
Diabetic status				
Unknown	1	0.1	0	0.0
Yes	55	6.4	22	12.7
No	807	93.5	151	87.3
SEER staging (subjects only)				
Unknown	–	–	15	8.7
Localized	–	–	35	20.2
Regional	–	–	33	19.1
Distant	–	–	90	52.0

context of pancreatic cancer, the composition of the microbiome has been linked to alterations in the local microenvironment and to promotion of oncogenesis through immune suppression,[10–12] with implications for response to therapy and survival.[13]

Microbiome colonization has been associated with metabolic changes that can perpetuate inflammation and increase an individual's risk of developing cancer.[7,14–16] Microbiome-related metabolites include short-chain fatty acids, butyrate and acetate, secondary bile acids, indole-derivatives, cadaverine, trimethylamine N-oxide (TMAO), and lipopolysaccharides.[17] A study of serum methionine-related metabolites identified elevated serum levels of TMAO, a gut microbiota-derived metabolite,[18] as associated with pancreatic cancer.[19,20] Other metabolites consisting of indoleacrylic acid and indole-3-acetate have been shown to differentiate subjects with newly diagnosed pancreatic cancer from control subjects.[21]

We designed our study to quantify the extent to which microbiome-related and other metabolites in circulation are elevated among subjects that were subsequently diagnosed with pancreatic cancer using sera collected from participants in the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial. Using a training and testing approach, we established a microbial-related metabolite panel for 5-year risk assessment of pancreatic cancer. The performance of the microbiome metabolite panel for risk prediction of pancreatic cancer was further evaluated in an independent cohort of patients with newly-diagnosed pancreatic cancer compared with non-cancer control subjects. The complementary value of other non-microbial-related metabolites as well as CA19-9 was also determined.

RESULTS

Quantification of microbial-related metabolites

Using untargeted metabolomics, we screened for microbial-derived metabolites in sera from 172 subjects diagnosed within 5 years of blood draw and 863 non-subject participants from the PLCO screening trial (Table 1). A total of 14 microbial-related metabolites were detected and quantified across all specimens, including 9 indole derivatives,[22,23] two secondary bile acids,[24,25] 5-hydroxy-tryptophan,[26] acetylcadaverine,[27] and TMAO.[28,29] Of the 14 metabolites, indoleacrylic acid, TMAO, and indole-derivative_2 had adjusted odds ratios (ORs) per unit standard deviation (SD) increase ≥1.2 for risk of pancreatic cancer (Figures S1 and S2). Elevated levels of TMAO and indoleacrylic acid have been associated with phyla of *Bacillota*, *Bacteroidota*, *Actinomycetota*, and *Pseudomonadota* (species of *Clostridium sporogenes* [*Cs*], *Eubacterium rectale* [*Er*], *Bacteroides thetaiotaomicron* [*Bt*], *Parabacteroides distasonis* [*Pd*], *Collinsella aerofaciens* [*Ca*], and *Edwardsiella tarda* [*Et*]),[30] all of which have relevance to pancreatic cancer (Figures 1A and 1B).[31–34]

Model building and testing of microbial-related metabolite panel

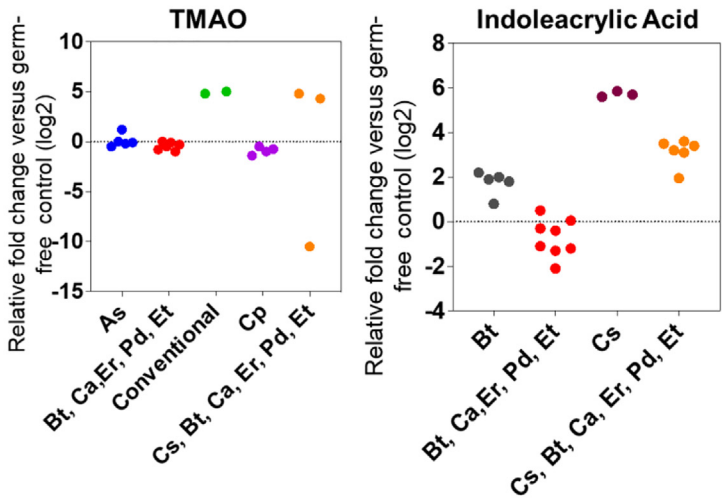
To establish a combination rule, all 14 microbial-related metabolites were considered. Seven different models were trained and optimized in the development set (Figure S3; Table S1). LASSO regression with three selected features achieved the highest prediction performance among all models in the validation set, yielding an area under the curve (AUC) of 0.64 (95% confidence interval [CI]: 0.54–0.73) and an adjusted OR of 1.42 (95% CI 0.94–2.13) per unit SD increase for 5-year probability of pancreatic cancer (Tables 2 and S2). To verify the reproducibility of our finding, we adhered to the predictability, computability, and stability (PCS) framework[35] and stress tested the 3-marker microbial panel to ensure its reliability. Stable performance in terms of AUC and adjusted OR across various data perturbations and stability checks demonstrated the robustness of the 3-marker microbial panel (Table S3).

Performance of the 3-marker microbial panel in the test set

In the test set, the 3-marker microbial panel yielded an AUC of 0.64 (95% CI: 0.53–0.76) and an adjusted OR of 1.72 (95% CI: 1.25–2.37) per unit SD increase for 5-year probability of

A

Annotation	Complete name
As	Anaerostipes sp
Cp	Citrobacter portucalensis
Cs	Clostridium sporogenes
Er	Eubacterium rectale
Bt	Bacteroides thetaiotaomicron
Pd	parabacteroides distasonis
Ca	collinsella aerofaciens
Et	Edwardsiella tarda



B

Species	Phylum	Class	Order	Associated with PDAC	Reference
Clostridium sporogenes (Cs)	Bacillota	Clostridia	Clostridiales	Increased in fecal samples of PDAC patients compared to controls	Zhou et al., 2021
Eubacterium rectale (Er)	Bacillota	Clostridia	Clostridiales	Decreased in fecal samples of PDAC patients compared to controls	Zhou et al., 2021
Bacteroides thetaiotaomicron (Bt)	Bacteroidota	Bacteroidia	Bacteroidales	Bacteriodales species elevated in fecal samples of PDAC patients compared to controls	Half et al., 2019 Matsukawa et al., 2021)
Parabacteroides distasonis (Pd)	Bacteroidota	Bacteroidia	Bacteroidales	Bacteriodales species elevated in fecal samples of PDAC patients compared to controls	Half et al., 2019 Matsukawa et al., 2021
Collinsella aerofaciens (Ca)	Actinomycetota	Coriobacteriia	Coriobacteriales	Relative abundances in fecal samples are associated with poor prognosis among PDAC patients	Matsukawa et al., 2021
Edwardsiella tarda (Et)	Pseudomonadota	Gammaproteobacteria	Enterobacterales	Associated with hepatobiliary disease including pancreatic cancer	Kamiyama et al., 2019

Figure 1. Relationship between TMAO and indoleacrylic acid and microbial species

(A) Association between TMAO and indoleacrlyic acid with different microbial species. Data were derived from the Metabolomics Data Explorer database (see STAR methods).[30] Data were derived from N = 2–8 biological replicates.

(B) Association between referenced microbial species and pancreatic cancer.

pancreatic cancer (Table 3). When considering subjects diagnosed within 2 years of blood draw, the 3-marker microbial panel yielded an AUC of 0.61 (95% CI: 0.48–0.74) and an adjusted OR of 1.43 (95% CI: 0.98–2.03) per unit SD increase for risk prediction of pancreatic cancer (Table 3). Prediction performance of the 3-marker microbial panel for risk assessment of pancreatic cancer was similar among diabetic and non-diabetic individuals (Table S4).

We further assessed the prediction performance of the 3-marker microbial panel in an independent set of samples from 99 subjects with newly diagnosed, resectable pancreatic ductal adenocarcinoma (PDAC), 50 patients with chronic

Table 2. Performance of microbial-related metabolites panels in different learning models in the PLCO validation set

Model	Hyperparameters	AUC (95% CI)	Adj OR[a]
Logistic regression	–	0.57 (0.46–0.67)	1.30 (0.85–2.02)
Logistic regression with ridge (L_2) regularization	penalty weight = 0.22	0.58 (0.48–0.68)	1.32 (0.87–2.05)
Logistic regression with LASSO (L_1) regularization	penalty weight = 0.023, number of selected features = 3	0.64 (0.54–0.73)	1.42 (0.94–2.13)
Iterative random forest	number of iterations = 4	0.52 (0.41–0.62)	1.28 (0.80–1.77)
Deep neural network model	number of cross-validation folds = 4, hidden layers = 2 with 64 nodes in each layer	0.55 (0.45–0.65)	1.17 (0.75–1.80)
GBM	number of trees = 36, max depth = 6	0.53 (0.41–0.65)	1.12 (0.76–1.58)
Auto machine learning (ML)	selected model = randomized trees	0.57 (0.45–0.68)	1.04 (0.64–1.63)

C.I., confidence interval.
[a]Age, gender, BMI, and smoking status were included as covariates in adjusted odds ratios (ORs).

pancreatitis (CP), and 100 healthy control subjects (Table S5). Compared to healthy control subjects, the 3-marker microbial panel had an OR of 2.83 (95% CI: 1.83–4.82) per unit SD increase for probability of CP, an OR of 1.55 (95% CI: 1.13–2.23) for pancreatic cancer, and an OR of 2.07 (95% CI: 1.45–3.18) for pancreatic disease (cancer or CP) (Figure S4). Of note, the 3-marker microbial panel performed best for identifying CP, which may be linked with innate pro-inflammatory properties of the microbial metabolites.[36–39]

Contributions of non-microbial metabolites for improved risk prediction of pancreatic cancer

We assessed the contribution of non-microbial metabolites for pancreatic cancer risk assessment. A total of 1,009 non-microbial metabolites were quantified in the PLCO specimen set (Table S6). Five non-microbial metabolites (cholesterol glucuronide, 2-hydroxyglutarate, galactosamine, glucose, and erythritol) exhibited statistically significant ($p < 0.05$) adjusted ORs in the development set (Table S7). We subsequently applied the PCS framework to develop and stress test a model based on the five non-microbial metabolites. A logistic regression model was selected based on exhibiting the highest predictive performance in the validation set, with a resultant AUC of 0.72 (95% CI: 0.65–0.97) and an adjusted OR of 2.10 (95% CI: 1.04–2.80) for 5-year risk prediction of pancreatic cancer (Table S8). In the set-aside test set, the 5-marker non-microbial panel yielded an AUC of 0.74 (95% CI: 0.65–0.83) and an adjusted (adj) OR of 2.72 (95% CI: 1.83–4.24) for 5-year risk prediction of pancreatic cancer (Table S9).

To assess the contributions of the 3-marker microbial panel and the 5-marker non-microbial panel, we fitted a logistic regression with the 3-marker microbial panel scores and the 5-marker non-microbial panel scores as two separate predictors. The combined metabolite panel yielded an AUC of 0.79 (95% CI: 0.71–0.88) and an adj OR of 3.13 (95% CI: 2.08–4.98) per unit SD increase for 5-year probability of pancreatic cancer in the set-aside test set (Tables 3 and S4). When considering subjects diagnosed within 0–2 years and 2–5 years of blood draw, the combined metabolite panel had respective AUCs of 0.82 (95% CI: 0.72–0.93) and 0.74 (95% CI: 0.60–0.86) (Table 3).

Contribution of the combined metabolite panel with CA19-9 for pancreatic cancer risk assessment

We previously demonstrated that levels of CA19-9 were increased in subjects with PDAC in the PLCO cohort, with an exponential rise starting 2 years prior to diagnosis.[40] We

Table 3. Performance estimates of the 3-marker microbial panel and a combined 3-marker microbial panel + 5-marker non-microbial panel for 5-year risk prediction of pancreatic cancer in the set-aside test set and the entire PLCO specimen set

Time to Dx	Subjects, N	Non-subjects, N	AUC (95% CI)	Adj OR[a] (95% CI)	p value	AUC (95% CI)	Adj OR[a] (95% CI)	p value
			3-marker microbial panel			3-marker microbial panel + 5-marker non-microbial panel		
Set-aside test set								
[0–5)	37	225	0.64 (0.53–0.76)	1.72 (1.25–2.37)	<0.001	0.79 (0.71–0.88)	3.13 (2.08–4.98)	<0.001
[0–2)	24	225	0.61 (0.48–0.74)	1.43 (0.98–2.03)	0.04	0.82 (0.72–0.93)	3.80 (2.33–6.74)	<0.001
[2–5)	13	225	0.70 (0.50–0.90)	2.11 (1.33–3.43)	<0.001	0.74 (0.60–0.86)	1.90 (1.08–3.37)	0.02
Entire set								
[0–5)	172	861	0.62 (0.57–0.67)	1.50 (1.28–1.76)	<0.001	0.76 (0.72–0.80)	2.75 (2.25–3.38)	<0.001
[0–2)	92	861	0.60 (0.54–0.67)	1.43 (1.18–1.74)	<0.001	0.81 (0.76–0.86)	3.66 (2.81–4.84)	<0.001
[2–5)	80	861	0.64 (0.57–0.70)	1.53 (1.28–1.87)	<0.001	0.69 (0.63–0.75)	1.92 (1.51–2.44)	0.02

C.I., confidence interval; Dx, diagnosis.
[a]Age, gender, BMI, and smoking status were included as co-variables in adjusted odd ratios.

Table 4. Performance estimates of the CA19-9 and a combined CA19-9 + 3-marker microbial panel + 5-marker non-microbial panel for 5-year risk prediction of pancreatic cancer in the set-aside test set and the entire PLCO specimen set

Time to Dx	Numbers	AUC (95% CI)	Adj OR (95% CI)[a]	p value	AUC (95% CI)	Adj OR 95% CI)[a]	p value	Diff. of AUCs (95% CI)	p value of difference	Diff. of adj OR (95% CI)	p value of difference
		CA19-9			CA19-9 + 3-marker microbial panel + 5-marker non-microbial panel			Difference			
Set-aside test set											
[0–5)	N0 = 225N1 = 37	0.66 (0.55–0.77)	2.2 (1.53–3.30)	<0.001	0.84 (0.76–0.91)	9.67 (4.56–23.30)	<0.001	0.18 (0.08–0.25)	<0.001	7.47 (2.10–15.97)	0.003
[0–2)	N0 = 225N1 = 24	0.70 (0.57–0.82)	2.55 (1.66–4.19)	<0.001	0.86 (0.77–0.95)	14.99 (5.76–47.66)	<0.001	0.16 (0.05–0.29)	0.006	12.44 (2.30–47.40)	0.01
[2–5)	N0 = 225N1 = 13	0.60 (0.40–0.81)	1.64 (0.94–2.89)	0.01	0.79 (0.67–0.90)	5.10 (1.93–15.88)	0.002	0.19 (0.02–0.37)	0.02	3.46 (−0.06 to 13.20)	0.06
Entire set											
[0–5)	N0 = 861N1 = 172	0.68 (0.63–0.73)	2.27 (1.89–2.76)	<0.001	0.80 (0.75–0.83)	8.44 (5.80–12.20)	<0.001	0.12 (0.07–0.16)	<0.001	6.17 (1.80–8.77)	0.004
[0–2)	N0 = 861N1 = 92	0.75 (0.69–0.81)	3.21 (2.50–4.20)	<0.001	0.87 (0.83–0.91)	20.02 (11.51–36.97)	<0.001	0.12 (0.07–0.16)	<0.001	16.81 (2.10–27.31)	<0.001
[2–5)	N0 = 861N1 = 80	0.60 (0.53–0.67)	1.48 (1.18–1.87)	<0.001	0.71 (0.65–0.77)	3.52 (2.36–5.32)	<0.001	0.11 (0.57–0.70)	0.001	2.04 (1.20–5.86)	0.04

Log transformation of the values were considered for adjusted odds ratio calculation. N0, number of non-subjects; N1, number of subjects.
C.I., confidence interval; Dx, diagnosis.
[a]Age, gender, BMI, and smoking status were included as co-variables in adjusted odd ratios.

therefore assessed whether the combined metabolite panel (3-marker microbial panel + the 5-marker non-microbial panel) would be complementary with CA19-9 for risk prediction of pancreatic cancer. In the set-aside test set, the combined metabolite panel + CA19-9 had an AUC of 0.84 (95% CI: 0.76–0.91) and an adj OR of 9.67 (95% CI: 4.56–23.30) per unit SD increase for 5-year probability of pancreatic cancer (Table 4; Figure 2A). For subjects diagnosed within 2 years after blood draw, the combined metabolite panel + CA19-9 yielded an AUC of 0.86 (95% CI: 0.77–0.95), which was markedly improved compared to CA19-9 alone (AUC: 0.70 [0.57–0.82], comparison of AUCs p value: 0.006) (Table 4).

Performance of the combined metabolite panel + CA19-9 for 5-year risk assessment of pancreatic cancer in the entire PLCO specimen set

In the entire PLCO specimen set, the combined metabolite panels + CA19-9 had an AUC of 0.80 (95% CI: 0.75–0.83) and an adj OR of 8.44 (95% CI: 5.80–12.20) for 5-year probability of pancreatic cancer and an AUC of 0.87 (95% CI: 0.83–0.91) with an adj OR of 20.02 (95% CI: 11.51–36.97) per unit SD increase for 2-year probability of pancreatic cancer (Tables 4 and S10; Figure 2B).

5-year absolute risk estimates adjusted for prevalence of disease based on the entire intervention arm of the PLCO population[41,42] for individuals with combined metabolite panel + CA19-9 model scores in the 80th, 90th, 95th, and 97.5th percentiles were 1.07%, 2.05%, 4.52%, and 13.33%, respectively (Figure 3).

DISCUSSION

Meaningful reductions in pancreatic cancer-related mortality may be realized through effective screening programs for earlier detection of disease. The low incidence of pancreatic cancer necessitates that a screening test for the general population yields adequate sensitivity at exceptionally high specificity. No such tests yet exist that meet performance criteria necessary for implementation for pancreatic cancer screening in the general population. However, the USPSTF has recognized that high-risk individuals, such as those with inherited risk or individuals with a history of CP, may benefit from surveillance and screening.[6] Here, we performed a metabolite screen for reported microbial-related metabolites in the blood and evaluated their association with pancreatic cancer risk. We developed and validated a 3-marker microbial-associated metabolite panel that offers potential utility for identifying individuals at high risk of developing pancreatic cancer within 5 years. A broader metabolite screen resulted in a blood-based metabolite panel consisting of microbial and non-microbial metabolites that yielded further improvements for identifying individuals at high risk of developing pancreatic cancer within 5 years.

Enriching for individuals who are at high risk of pancreatic cancer increases the positive predictive value of pertinent cancer-detection tests while reducing the number of false positive tests. To this end, we showed that the risk prediction model based on circulating microbial + non-microbial metabolites is additive to CA19-9 for identifying individuals who went on to

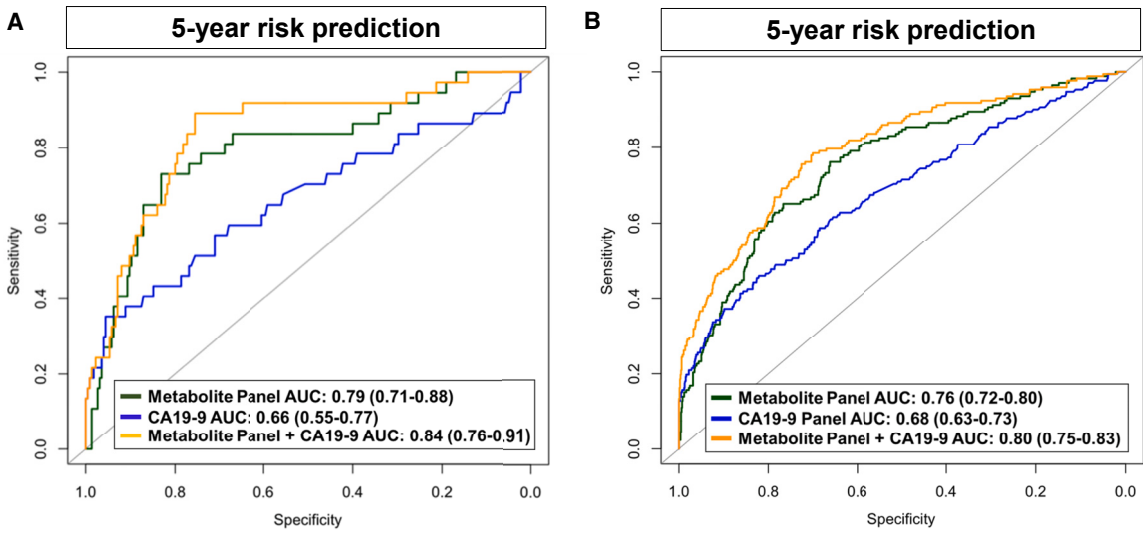


Figure 2. Area under the reciever operating characteristic curves for 5-year risk prediction of pancreatic cancer in the PLCO cohort

Predictive performance estimates for the metabolite (microbial + non-microbial) panel, CA19-9, and the metabolite panel + CA19-9 for 5-year risk prediction of pancreatic cancer in the PLCO set-aside test set (A) and the entire PLCO specimen set (B).

receive a pancreatic cancer diagnosis within 5 years of blood draw. Specifically, our findings demonstrated that individuals with combined metabolite panel + CA19-9 scores in the top 2.5th percentile have estimated 5-year absolute risks of >13%, which would warrant more intensive follow-up and trigger an imaging-based modality such as contrast-enhanced pancreas protocol computed tomography (CT) or MRI/magnetic resonance cholangiopancreatography (MRCP).

The microbial-related metabolite panel includes indoleacyrlic acid, an indole-derivative, and TMAO. TMAO- and indoleacrylic-acid-producing bacteria include those in the phyla of *Bacillota*, *Bacteroidota*, *Actinomycetota*, and *Pseudomonadota*. *Bacillota* species such as *Cs* and *Er* and *Bacteriodota* species including *Bt* and *Pd* have been shown to be increased in fecal samples of patients with PDAC compared with control subjects.[31–33] Relative abundances of fecal *Collinsella aeofaciens*, a species of *Actinomycetota*, is associated with poor prognosis in PDAC.[32]

Indole and associated derivatives are derived through the catabolism of tryptophan via the microbiome that may serve as ligands for the aryl hydrocarbon receptor (AHR) to modulate the immune and inflammatory response.[43–45] Notably, indole and indole derivatives are thought to be largely derived from commensal microbes with reported anti-inflammatory properties.[46]

TMAO is a gut microbiota-derived metabolite of dietary choline, betaine, and L-carnitine that has been reported to be associated with increased risk of several cancer types including pancreatic cancer.[23,47–49] Prior studies have shown that TMAO is elevated in pancreatic cystic fluid of individuals presenting with high-risk intraductal papillary mucinous neoplasms or pancreatic cancer compared with those harboring non-cancerous cysts.[50] Moreover, levels of TMAO in cystic fluid were positively correlated with bacterial clusters corresponding to *Enterobacteriacea*, *Granulicatella*, *Klebsiella*, *Stenotrophomonas*, *Streptococcus*, *Haemophilus*, and *Fusobacterium*,[50] which have previously been reported to be associated with pancreatic cancer.[15,51] Mechanistically, studies have shown that TMAO induces activation of inflammatory pathways, including the nuclear factor κB (NF-κB) pathway and the thioredoxin-interactive protein (TXNIP)-NLRP3 inflammasome, resulting in increased oxidative stress, DNA damage, and release of inflammatory cytokines that may potentiate cancer development.[36–39] We observed TMAO to also be particularly elevated in patients presenting with CP, further suggesting a relationship between TMAO, inflammation of pancreas tissues, and pancreatic cancer risk.[52,53]

Non-microbial metabolites in the metabolite panel included 2-hydroxyglutarate, cholesterol glucuronide, galactosamine, glucose, and erythritol. Production of the oncometabolite 2-hydroxyglutarate is largely associated with mutations in isocitrate dehydrogenase 1 (IDH1) and IDH2, neomorphic enzymes that convert α-ketoglutarate to 2-hydroxyglutarate.[54] 2-Hydroxyglutarate can also be produced through alternative metabolic pathways with pro-tumoral effects. For instance, recent data also suggest that, under hypoxic conditions, lactate dehydrogenase produces 2-hydroxyglutarate to maintain stemness and facilitate immune evasion in pancreatic cancer.[55] Cholesterol glucuronide is a natural metabolite of cholesterol generated in the liver by UDP glucuronyltransferase. Prior studies have shown that elevated levels of cholesterol glucuronide is prognostic for poor survival in patients with pancreatic cancer.[56]

The onset of diabetes is often a manifestation that precedes diagnosis of pancreatic cancer, and new-onset glucose intolerance is a frequent and characteristic feature of pancreatic cancer.[57,58] To this end, in a prior population-based case-control study of 736 pancreatic cancer subjects and 1,875 age- and gender-matched control subjects, 40.2% of subjects with pancreatic cancer had diabetes.[58] In another study, 50% of patients with stage I and II pancreatic cancer had diabetes.[57–59]

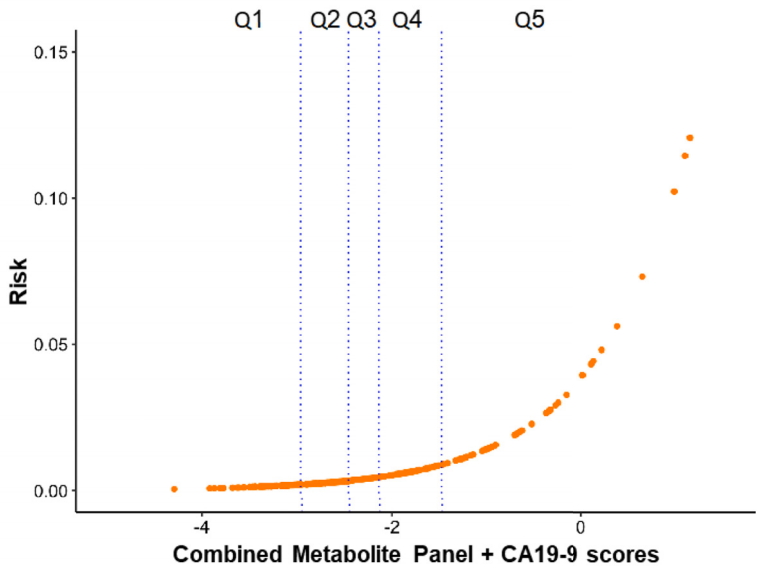


Figure 3. Absolute 5-year risk estimates for individuals with CA19-9 + 3-marker microbial panel + 5-marker non-microbial panel scores

Vertical lines represent 20th, 40th, 60th, and 80th percentile values. Table on the bottom provides absolute 5-year risk estimates for individuals with CA19-9, combined metabolite panel (3-marker microbial metabolite panel + 5-marker non-microbial metabolite panel), and the combined metabolite panel + CA19-9 scores.

	5-year absolute risk (%)		
Percentiles	**CA19-9**	**Combined Metabolite Panel**	**Combined Metabolite Panel + CA19-9**
20.0%	0.450	0.262	0.227
40.0%	0.517	0.425	0.350
60.0%	0.609	0.652	0.528
80.0%	0.870	1.245	1.066
90.0%	1.389	1.890	2.049
95.0%	2.159	2.880	4.521
97.5%	10.060	4.740	13.330

Thus, elevated levels of glucose and galactosamine, a hexosamine derived from galactose,[60] likely reflect an onset of diabetes that temporally occurs with the development of pancreatic cancer. Although prior studies reported that elevations in circulating branched chain amino acids (BCAAs) were associated with increased risk of PDAC,[20,61] we did not observe any statistically significant between BCAA levels and PDAC risk in the PLCO cohort.

There are some considerations to our study. Given the low incidence of PDAC in the general population, procurement of pre-diagnostic specimens for biomarker discovery and testing is challenging. In our study, we leveraged pre-diagnostic sera from the multi-institutional PLCO cancer screening trial to test the merits of microbial-associated and other non-microbial metabolites for risk assessment of PDAC. While we acknowledge the limited sample size of subjects with PDAC in the PLCO specimen set, we emphasize rigor in our statistical approach, adhering to the PCS framework for modeling and evaluation of model stability and robustness,[35] as well as the use of an independent set of plasmas from patients with newly diagnosed PDAC . Information regarding new-onset diabetes versus long-standing diabetes as well as other clinical measurements, such as HbA1C or weight loss, were not available. Moreover, the frequency of diabetes in the PLCO cohort is also likely to be underestimated. Consequently, we were unable to evaluate the complementarity of the metabolite panel together with other risk models based on patient characteristics[62] for risk assessment of pancreatic cancer. 16S sequencing data to assess stool- or tissue-level microbial diversity and composition were not available for analyzed samples, thus preventing direct correlative studies between specific microbial species and the established microbial-related metabolite panel. CP status, fasting status, and food intake for PLCO participants was not available. Fasting status and food and drink uptake were not controlled for in the PLCO cohort, and information was not available. Time-dependent performance estimates were derived based on availability of serum samples at various time points preceding cancer diagnosis from individual patients. Availability of serial samples would allow for the development of longitudinal algorithms for assessment of pancreatic cancer risk. Whether the metabolite panel to inform on risk of other cancer types warrants consideration. Specificity of the metabolite panel for risk of pancreatic cancer can be improved through testing of recognized high-risk populations, including those with inherited risk[63,64] or with mucinous cysts of the pancreas[65] or individuals older than 50 with new-onset diabetes.[57,58]

In conclusion, the metabolite panel has the potential to identify individuals at high risk of pancreatic cancer who may benefit from surveillance and/or potential cancer interception strategies such as vaccines. Integration of the panel with other risk models of pancreatic cancer may yield further improvements for risk assessment.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrm.2023.101194.

ACKNOWLEDGMENTS

J.F.F. was supported by the McKee Early Career Investigator in Pancreatic Cancer Research. B.M.W. was supported by NIH grant U01 CA210171, the Hale Family Center for Pancreatic Cancer Research, the Lustgarten Foundation Dedicated Laboratory program, Stand Up To Cancer, NIH grant P50 CA127003, the Pancreatic Cancer Action Network, the Noble Effort Fund, the Wexler Family Fund, and Promises for Purple. The funding specific to sample collection and management for this cohort was NIH grant U01 CA210171. Work was supported by generous philanthropic contributions to The University of Texas MD Anderson Cancer Center Moon Shots Program TM. A.M. and S.H. are supported by MCL (5U01CA196403-05), EDRN (5U01CA200468-05), and the MD Anderson Cancer Center GI SPORE (5-P50-CA221707-02). J.P.L. is supported by EDRN (5U01CA200468-05). T.T. was supported by the NSF Graduate Research Fellowship Program DGE-2146752. A.K. and B.Y. were supported by a Chan Zuckerberg Biohub Intercampus Research Award. B.Y. was also supported by NSF DMS-1613002, IIS 1741340, and a Weill Neurohub grant.

AUTHOR CONTRIBUTIONS

Conceptualization, E.I., S.H., and J.F.F.; methodology, E.I., B.Y., R.W., E.M., J.B.D., and J.F.F.; formal analysis, E.I., A.K., T.T., B.Y., and J.F.F.; investigation, E.I. and J.F.F.; resources, J.A.C., M.D.K., F.K., L.B., K.J., L.S.L., T.E.C., K.N., A.B., A.M., B.M.W., and S.H.; data curation, E.I., R.W., E.M., and J.F.F.; writing – original draft preparation, E.I. and J.F.F.; writing – review & editing, A.K., T.T., J.V., R.W., E.M., J.B.D., M.S., J.P.L., M.L., J.A.C., M.D.K., F.K., L.B., A.B., K.J., L.S.L., T.E.C., K.N., A.B., J.M.G., A.M., K.-A.D., B.Y., B.M.W., and S.H.; visualization, E.I. and J.F.F.; supervision, S.H. and J.F.F.; project administration, S.H. and J.F.F.; funding acquisition, J.P.L., T.T., A.K., B.Y., A.M., B.M.W., S.H., and J.F.F.

DECLARATION OF INTERESTS

An invention disclosure report related to findings reported herein has been submitted to the University of Texas. B.M.W. receives research funding from Celgene and Eli Lilly and does consulting for BioLineRx, Celgene, and GRAIL.

Received: August 30, 2022
Revised: December 20, 2022
Accepted: August 21, 2023
Published: September 19, 2023

REFERENCES

1. Rahib, L., Wehner, M.R., Matrisian, L.M., and Nead, K.T. (2021). Estimated Projection of US Cancer Incidence and Death to 2040. JAMA Netw. Open *4*, e214708. https://doi.org/10.1001/jamanetworkopen.2021.4708.
2. Ryan, D.P., Hong, T.S., and Bardeesy, N. (2014). Pancreatic adenocarcinoma. N. Engl. J. Med. *371*, 1039–1049. https://doi.org/10.1056/NEJMra1404198.
3. Kleeff, J., Korc, M., Apte, M., La Vecchia, C., Johnson, C.D., Biankin, A.V., Neale, R.E., Tempero, M., Tuveson, D.A., Hruban, R.H., and Neoptolemos, J.P. (2016). Pancreatic cancer. Nat. Rev. Dis. Prim. *2*, 16022. https://doi.org/10.1038/nrdp.2016.22.
4. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA A Cancer J. Clin. *68*, 394–424. https://doi.org/10.3322/caac.21492.
5. Rawla, P., Sunkara, T., and Gaduputi, V. (2019). Epidemiology of Pancreatic Cancer: Global Trends, Etiology and Risk Factors. World J. Oncol. *10*, 10–27. https://doi.org/10.14740/wjon1166.
6. US Preventive Services Task Force; Owens, D.K., Davidson, K.W., Krist, A.H., Barry, M.J., Cabana, M., Caughey, A.B., Curry, S.J., Doubeni, C.A., Epling, J.W., Jr., et al. (2019). Screening for Pancreatic Cancer: US Preventive Services Task Force Reaffirmation Recommendation Statement. JAMA *322*, 438–444. https://doi.org/10.1001/jama.2019.10232.
7. Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. nature *486*, 207–214.
8. Huybrechts, I., Zouiouich, S., Loobuyck, A., Vandenbulcke, Z., Vogtmann, E., Pisanu, S., Iguacel, I., Scalbert, A., Indave, I., Smelov, V., et al. (2020). The Human Microbiome in Relation to Cancer Risk: A Systematic Review of Epidemiologic Studies. Cancer Epidemiol. Biomarkers Prev. *29*, 1856–1868. https://doi.org/10.1158/1055-9965.Epi-20-0288.
9. Bhatt, A.P., Redinbo, M.R., and Bultman, S.J. (2017). The role of the microbiome in cancer development and therapy. CA A Cancer J. Clin. *67*, 326–344. https://doi.org/10.3322/caac.21398.
10. Pushalkar, S., Hundeyin, M., Daley, D., Zambirinis, C.P., Kurz, E., Mishra, A., Mohan, N., Aykut, B., Usyk, M., Torres, L.E., et al. (2018). The pancreatic cancer microbiome promotes oncogenesis by induction of innate and adaptive immune suppression. Cancer Discov. *8*, 403–416.
11. Riquelme, E., Zhang, Y., Zhang, L., Montiel, M., Zoltan, M., Dong, W., Quesada, P., Sahin, I., Chandra, V., San Lucas, A., et al. (2019). Tumor microbiome diversity and composition influence pancreatic cancer outcomes. Cell *178*, 795–806.e12.
12. Barbour, S.E., Nakashima, K., Zhang, J.-B., Tangada, S., Hahn, C.-L., Schenkein, H.A., and Tew, J.G. (1997). Tobacco and smoking: environmental factors that modify the host response (immune system) and have an impact on periodontal health. Crit. Rev. Oral Biol. Med. *8*, 437–460.
13. Zhang, X., Liu, Q., Liao, Q., and Zhao, Y. (2020). Pancreatic Cancer, Gut Microbiota, and Therapeutic Efficacy. J. Cancer *11*, 2749–2758. https://doi.org/10.7150/jca.37445.
14. Risch, H.A. (2012). Pancreatic cancer: Helicobacter pylori colonization, N-Nitrosamine exposures, and ABO blood group. Mol. Carcinog. *51*, 109–118.
15. Wei, M.-Y., Shi, S., Liang, C., Meng, Q.-C., Hua, J., Zhang, Y.-Y., Liu, J., Zhang, B., Xu, J., and Yu, X.-J. (2019). The microbiota and microbiome in pancreatic cancer: more influential than expected. Mol. Cancer *18*. 97–15.
16. Brennan, C.A., and Garrett, W.S. (2016). Gut Microbiota, Inflammation, and Colorectal Cancer. Annu. Rev. Microbiol. *70*, 395–411. https://doi.org/10.1146/annurev-micro-102215-095513.
17. Kiss, B., Mikó, E., Sebő, É., Toth, J., Ujlaki, G., Szabó, J., Uray, K., Bai, P., and Árkosy, P. (2020). Oncobiosis and Microbial Metabolite Signaling in Pancreatic Adenocarcinoma. Cancers *12*, 1068. https://doi.org/10.3390/cancers12051068.

18. Yang, S., Li, X., Yang, F., Zhao, R., Pan, X., Liang, J., Tian, L., Li, X., Liu, L., Xing, Y., and Wu, M. (2019). Gut Microbiota-Dependent Marker TMAO in Promoting Cardiovascular Disease: Inflammation Mechanism, Clinical Prognostic, and Potential as a Therapeutic Target. Front. Pharmacol. *10*, 1360. https://doi.org/10.3389/fphar.2019.01360.

19. Huang, J.Y., Luu, H.N., Butler, L.M., Midttun, Ø., Ulvik, A., Wang, R., Jin, A., Gao, Y.-T., Tan, Y., Ueland, P.M., et al. (2020). A prospective evaluation of serum methionine-related metabolites in relation to pancreatic cancer risk in two prospective cohort studies. Int. J. Cancer *147*, 1917–1927. https://doi.org/10.1002/ijc.32994.

20. Mayers, J.R., Wu, C., Clish, C.B., Kraft, P., Torrence, M.E., Fiske, B.P., Yuan, C., Bao, Y., Townsend, M.K., Tworoger, S.S., et al. (2014). Elevation of circulating branched-chain amino acids is an early event in human pancreatic adenocarcinoma development. Nat. Med. *20*, 1193–1198. https://doi.org/10.1038/nm.3686.

21. Xie, G., Lu, L., Qiu, Y., Ni, Q., Zhang, W., Gao, Y.T., Risch, H.A., Yu, H., and Jia, W. (2015). Plasma metabolite biomarkers for the detection of pancreatic cancer. J. Proteome Res. *14*, 1195–1202. https://doi.org/10.1021/pr501135f.

22. Lee, J.-H., and Lee, J. (2010). Indole as an intercellular signal in microbial communities. FEMS Microbiol. Rev. *34*, 426–444.

23. Jaglin, M., Rhimi, M., Philippe, C., Pons, N., Bruneau, A., Goustard, B., Daugé, V., Maguin, E., Naudon, L., and Rabot, S. (2018). Indole, a signaling molecule produced by the gut microbiota, negatively impacts emotional behaviors in rats. Front. Neurosci. *12*, 216.

24. Ridlon, J.M., Kang, D.J., Hylemon, P.B., and Bajaj, J.S. (2014). Bile acids and the gut microbiome. Curr. Opin. Gastroenterol. *30*, 332–338.

25. Hylemon, P.B., Harris, S.C., and Ridlon, J.M. (2018). Metabolism of hydrogen gases and bile acids in the gut microbiome. FEBS Lett. *592*, 2070–2082.

26. Stoll, M.L., Kumar, R., Lefkowitz, E.J., Cron, R.Q., Morrow, C.D., and Barnes, S. (2016). Fecal metabolomics in pediatric spondyloarthritis implicate decreased metabolic diversity and altered tryptophan metabolism as pathogenic factors. Gene Immun. *17*, 400–405.

27. Pugin, B., Barcik, W., Westermann, P., Heider, A., Wawrzyniak, M., Hellings, P., Akdis, C.A., and O'Mahony, L. (2017). A wide diversity of bacteria from the human gut produces and degrades biogenic amines. Microb. Ecol. Health Dis. *28*, 1353881. https://doi.org/10.1080/16512235.2017.1353881.

28. Brunt, V.E., LaRocca, T.J., Bazzoni, A.E., Sapinsley, Z.J., Miyamoto-Ditmon, J., Gioscia-Ryan, R.A., Neilson, A.P., Link, C.D., and Seals, D.R. (2021). The gut microbiome–derived metabolite trimethylamine N-oxide modulates neuroinflammation and cognitive function with aging. GeroScience *43*, 377–394.

29. Xu, R., Wang, Q., and Li, L. (2015). A genome-wide systems analysis reveals strong link between colorectal cancer and trimethylamine N-oxide (TMAO), a gut microbial metabolite of dietary meat and fat. BMC Genom. *16*, S4–S9.

30. Han, S., Van Treuren, W., Fischer, C.R., Merrill, B.D., DeFelice, B.C., Sanchez, J.M., Higginbottom, S.K., Guthrie, L., Fall, L.A., Dodd, D., et al. (2021). A metabolomics pipeline for the mechanistic interrogation of the gut microbiome. Nature *595*, 415–420.

31. Zhou, W., Zhang, D., Li, Z., Jiang, H., Li, J., Ren, R., Gao, X., Li, J., Wang, X., Wang, W., and Yang, Y. (2021). The fecal microbiota of patients with pancreatic ductal adenocarcinoma and autoimmune pancreatitis characterized by metagenomic sequencing. J. Transl. Med. *19*, 215–312.

32. Matsukawa, H., Iida, N., Kitamura, K., Terashima, T., Seishima, J., Makino, I., Kannon, T., Hosomichi, K., Yamashita, T., Sakai, Y., et al. (2021). Dysbiotic gut microbiota in pancreatic cancer patients form correlation networks with the oral microbiota and prognostic factors. Am. J. Cancer Res. *11*, 3163–3175.

33. Half, E., Keren, N., Reshef, L., Dorfman, T., Lachter, I., Kluger, Y., Reshef, N., Knobler, H., Maor, Y., Stein, A., et al. (2019). Fecal microbiome signatures of pancreatic cancer patients. Sci. Rep. *9*, 16801–16812.

34. Kamiyama, S., Kuriyama, A., and Hashimoto, T. (2019). Edwardsiella tarda Bacteremia, Okayama, Japan, 2005–2016. Emerg. Infect. Dis. *25*, 1817–1823.

35. Yu, B., and Kumbier, K. (2020). Veridical data science. Proc. Natl. Acad. Sci. USA *117*, 3920–3929.

36. Sun, X., Jiao, X., Ma, Y., Liu, Y., Zhang, L., He, Y., and Chen, Y. (2016). Trimethylamine N-oxide induces inflammation and endothelial dysfunction in human umbilical vein endothelial cells via activating ROS-TXNIP-NLRP3 inflammasome. Biochem. Biophys. Res. Commun. *481*, 63–70. https://doi.org/10.1016/j.bbrc.2016.11.017.

37. Seldin, M.M., Meng, Y., Qi, H., Zhu, W., Wang, Z., Hazen, S.L., Lusis, A.J., and Shih, D.M. (2016). Trimethylamine N-Oxide Promotes Vascular Inflammation Through Signaling of Mitogen-Activated Protein Kinase and Nuclear Factor-κB. J. Am. Heart Assoc. *5*, e002767. https://doi.org/10.1161/jaha.115.002767.

38. Arlt, A., Schäfer, H., and Kalthoff, H. (2012). The 'N-factors' in pancreatic cancer: functional relevance of NF-κB, NFAT and Nrf2 in pancreatic cancer. Oncogenesis *1*, e35. https://doi.org/10.1038/oncsis.2012.35.

39. Missiroli, S., Perrone, M., Boncompagni, C., Borghi, C., Campagnaro, A., Marchetti, F., Anania, G., Greco, P., Fiorica, F., Pinton, P., and Giorgi, C. (2021). Targeting the NLRP3 Inflammasome as a New Therapeutic Option for Overcoming Cancer. Cancers *13*, 2297. https://doi.org/10.3390/cancers13102297.

40. Fahrmann, J.F., Schmidt, C.M., Mao, X., Irajizad, E., Loftus, M., Zhang, J., Patel, N., Vykoukal, J., Dennison, J.B., Long, J.P., et al. (2021). Lead-time trajectory of CA19-9 as an anchor marker for pancreatic cancer early detection. Gastroenterology *160*, 1373–1383.e6.

41. Prorok, P.C., Andriole, G.L., Bresalier, R.S., Buys, S.S., Chia, D., Crawford, E.D., Fogel, R., Gelmann, E.P., Gilbert, F., Hasson, M.A., et al. (2000). Design of the prostate, lung, colorectal and ovarian (PLCO) cancer screening trial. Contr. Clin. Trials *21*, 273S–309S.

42. Anderson, K.E., Mongin, S.J., Sinha, R., Stolzenberg-Solomon, R., Gross, M.D., Ziegler, R.G., Mabie, J.E., Risch, A., Kazin, S.S., and Church, T.R. (2012). Pancreatic cancer risk: Associations with meat-derived carcinogen intake in the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO) cohort. Mol. Carcinog. *51*, 128–137.

43. Wlodarska, M., Luo, C., Kolde, R., d'Hennezel, E., Annand, J.W., Heim, C.E., Krastel, P., Schmitt, E.K., Omar, A.S., Creasey, E.A., et al. (2017). Indoleacrylic acid produced by commensal peptostreptococcus species suppresses inflammation. Cell Host Microbe *22*, 25–37.e6.

44. Hubbard, T.D., Murray, I.A., Bisson, W.H., Lahoti, T.S., Gowda, K., Amin, S.G., Patterson, A.D., and Perdew, G.H. (2015). Adaptation of the human aryl hydrocarbon receptor to sense microbiota-derived indoles. Sci. Rep. *5*, 12689. https://doi.org/10.1038/srep12689.

45. Hezaveh, K., Shinde, R.S., Klötgen, A., Halaby, M.J., Lamorte, S., Ciudad, M.T., Quevedo, R., Neufeld, L., Liu, Z.Q., Jin, R., et al. (2022). Tryptophan-derived microbial metabolites activate the aryl hydrocarbon receptor in tumor-associated macrophages to suppress anti-tumor immunity. Immunity *55*, 324–340.e8. https://doi.org/10.1016/j.immuni.2022.01.006.

46. Roager, H.M., and Licht, T.R. (2018). Microbial tryptophan catabolites in health and disease. Nat. Commun. *9*, 3294. https://doi.org/10.1038/s41467-018-05470-4.

47. Huang, J.Y., Luu, H.N., Butler, L.M., Midttun, Ø., Ulvik, A., Wang, R., Jin, A., Gao, Y.T., Tan, Y., Ueland, P.M., et al. (2020). A prospective evaluation of serum methionine-related metabolites in relation to pancreatic cancer risk in two prospective cohort studies. Int. J. Cancer *147*, 1917–1927.

48. Jiao, L., Maity, S., Coarfa, C., Rajapakshe, K., Chen, L., Jin, F., Putluri, V., Tinker, L.F., Mo, Q., Chen, F., et al. (2019). A prospective targeted serum metabolomics study of pancreatic cancer in postmenopausal women. Cancer Prev. Res. *12*, 237–246.

49. Huang, J., Butler, L., Midttun, Ø., Wang, R., Jin, A., Gao, Y.-T., Ueland, P., Koh, W.-P., and Yuan, J.-M. (2017). Serum Choline, Methionine, Betaine, Dimethylglycine, and Trimethylamine-N-Oxide in Relation to Pancreatic Cancer Risk in Two Nested Case-Control Studies in Asian Populations (AACR).

50. Morgell, A., Reisz, J.A., Ateeb, Z., Davanian, H., Reinsbach, S.E., Halimi, A., Gaiser, R., Valente, R., Arnelo, U., Del Chiaro, M., et al. (2021). Metabolic Characterization of Plasma and Cyst Fluid from Cystic Precursors to Pancreatic Cancer Patients Reveal Metabolic Signatures of Bacterial Infection. J. Proteome Res. *20*, 2725–2738. https://doi.org/10.1021/acs.jproteome.1c00018.

51. Rogers, M.B., Aveson, V., Firek, B., Yeh, A., Brooks, B., Brower-Sinning, R., Steve, J., Banfield, J.F., Zureikat, A., Hogg, M., et al. (2017). Disturbances of the Perioperative Microbiome Across Multiple Body Sites in Patients Undergoing Pancreaticoduodenectomy. Pancreas *46*, 260–267. https://doi.org/10.1097/mpa.0000000000000726.

52. Whitcomb, D.C. (2004). Inflammation and Cancer V. Chronic pancreatitis and pancreatic cancer. Am. J. Physiol. Gastrointest. Liver Physiol. *287*, G315–G319.

53. Farrow, B., and Evers, B.M. (2002). Inflammation and the development of pancreatic cancer. Surg. Oncol. *10*, 153–169.

54. Dang, L., White, D.W., Gross, S., Bennett, B.D., Bittinger, M.A., Driggers, E.M., Fantin, V.R., Jang, H.G., Jin, S., Keenan, M.C., et al. (2009). Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. Nature *462*, 739–744. https://doi.org/10.1038/nature08617.

55. Gupta, V.K., Sharma, N.S., Durden, B., Garrido, V.T., Kesh, K., Edwards, D., Wang, D., Myer, C., Mateo-Victoriano, B., Kollala, S.S., et al. (2021). Hypoxia-Driven Oncometabolite L-2HG Maintains Stemness-Differentiation Balance and Facilitates Immune Evasion in Pancreatic Cancer. Cancer Res. *81*, 4001–4013. https://doi.org/10.1158/0008-5472.Can-20-2562.

56. Chen, K., Zhou, C., He, Y., Liu, J., and Yang, X. (2021). Metabolomics Profiling of EUS-FNA Sample Predicts Advanced Pancreatic Adenocarcinoma Prognosis. Research Square.

57. Sharma, A., Kandlakunta, H., Nagpal, S.J.S., Feng, Z., Hoos, W., Petersen, G.M., and Chari, S.T. (2018). Model to Determine Risk of Pancreatic Cancer in Patients With New-Onset Diabetes. Gastroenterology *155*, 730–739.e3. https://doi.org/10.1053/j.gastro.2018.05.023.

58. Pannala, R., Basu, A., Petersen, G.M., and Chari, S.T. (2009). New-onset diabetes: a potential clue to the early diagnosis of pancreatic cancer. Lancet Oncol. *10*, 88–95. https://doi.org/10.1016/s1470-2045(08)70337-1.

59. Pannala, R., Leirness, J.B., Bamlet, W.R., Basu, A., Petersen, G.M., and Chari, S.T. (2008). Prevalence and clinical profile of pancreatic cancer-associated diabetes mellitus. Gastroenterology *134*, 981–987. https://doi.org/10.1053/j.gastro.2008.01.039.

60. Theocharis, A.D., Tsara, M.E., Papageorgacopoulou, N., Karavias, D.D., and Theocharis, D.A. (2000). Pancreatic carcinoma is characterized by elevated content of hyaluronan and chondroitin sulfate with altered disaccharide composition. Biochim. Biophys. Acta *1502*, 201–206. https://doi.org/10.1016/S0925-4439(00)00051-X.

61. Katagiri, R., Goto, A., Nakagawa, T., Nishiumi, S., Kobayashi, T., Hidaka, A., Budhathoki, S., Yamaji, T., Sawada, N., Shimazu, T., et al. (2018). Increased Levels of Branched-Chain Amino Acid Associated With Increased Risk of Pancreatic Cancer in a Prospective Case-Control Study of a Large Cohort. Gastroenterology *155*, 1474–1482.e1. https://doi.org/10.1053/j.gastro.2018.07.033.

62. Yuan, C., Babic, A., Khalaf, N., Nowak, J.A., Brais, L.K., Rubinson, D.A., Ng, K., Aguirre, A.J., Pandharipande, P.V., Fuchs, C.S., et al. (2020). Diabetes, Weight Change, and Pancreatic Cancer Risk. JAMA Oncol. *6*, e202948. https://doi.org/10.1001/jamaoncol.2020.2948.

63. Canto, M.I., Almario, J.A., Schulick, R.D., Yeo, C.J., Klein, A., Blackford, A., Shin, E.J., Sanyal, A., Yenokyan, G., Lennon, A.M., et al. (2018). Risk of Neoplastic Progression in Individuals at High Risk for Pancreatic Cancer Undergoing Long-term Surveillance. Gastroenterology *155*, 740–751.e2. https://doi.org/10.1053/j.gastro.2018.05.035.

64. Petersen, G.M. (2016). Familial pancreatic cancer. Semin. Oncol. *43*, 548–553. https://doi.org/10.1053/j.seminoncol.2016.09.002.

65. Ohno, E., Hirooka, Y., Kawashima, H., Ishikawa, T., Kanamori, A., Ishikawa, H., Sasaki, Y., Nonogaki, K., Hara, K., Hashimoto, S., et al. (2018). Natural history of pancreatic cystic lesions: A multicenter prospective observational study for evaluating the risk of pancreatic cancer. J. Gastroenterol. Hepatol. *33*, 320–328. https://doi.org/10.1111/jgh.13967.

66. Prorok, P.C., Andriole, G.L., Bresalier, R.S., Buys, S.S., Chia, D., Crawford, E.D., Fogel, R., Gelmann, E.P., Gilbert, F., Hasson, M.A., et al. (2000). Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. Contr. Clin. Trials *21*, 273s–309s. https://doi.org/10.1016/s0197-2456(00)00098-2.

67. Fahrmann, J.F., Schmidt, C.M., Mao, X., Irajizad, E., Loftus, M., Zhang, J., Patel, N., Vykoukal, J., Dennison, J.B., et al. (2021). Lead-Time Trajectory of CA19-9 as an Anchor Marker for Pancreatic Cancer Early Detection. Gastroenterology *160*, 1373–1383.e6. https://doi.org/10.1053/j.gastro.2020.11.052.

68. Fahrmann, J.F., Bantis, L.E., Capello, M., Scelo, G., Dennison, J.B., Patel, N., Murage, E., Vykoukal, J., Kundnani, D.L., Foretova, L., et al. (2019). A Plasma-Derived Protein-Metabolite Multiplexed Panel for Early-Stage Pancreatic Cancer. J. Natl. Cancer Inst. *111*, 372–379. https://doi.org/10.1093/jnci/djy126.

69. Fahrmann, J.F., Irajizad, E., Kobayashi, M., Vykoukal, J., Dennison, J.B., Murage, E., Wu, R., Long, J.P., Do, K.A., Celestino, J., et al. (2021). A MYC-Driven Plasma Polyamine Signature for Early Detection of Ovarian Cancer. Cancers *13*, 913. https://doi.org/10.3390/cancers13040913.

70. Fahrmann, J.F., Vykoukal, J., Fleury, A., Tripathi, S., Dennison, J.B., Murage, E., Wang, P., Yu, C.Y., Capello, M., Creighton, C.J., et al. (2020). Association Between Plasma Diacetylspermine and Tumor Spermine Synthase With Outcome in Triple-Negative Breast Cancer. J. Natl. Cancer Inst. *112*, 607–616. https://doi.org/10.1093/jnci/djz182.

71. Johannes, F., Carla, P., and Amanda, W. (2021). A Blood-Based Polyamine Signature Associated with MEN1 Duodenopancreatic Neuroendocrine Tumor Progression 10.6084/m9.figshare.14639079.V1.

72. Vykoukal, J., Fahrmann, J.F., Gregg, J.R., Tang, Z., Basourakos, S., Irajizad, E., Park, S., Yang, G., Creighton, C.J., Fleury, A., et al. (2020). Caveolin-1-mediated sphingolipid oncometabolism underlies a metabolic vulnerability of prostate cancer. Nat. Commun. *11*, 4279. https://doi.org/10.1038/s41467-020-17645-z.

73. DeLong, E.R., DeLong, D.M., and Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics *44*, 837–845.

74. Behr, M., Kumbier, K., Cordova-Palomera, A., Aguirre, M., Ashley, E., Butte, A.J., Arnaout, R., Brown, B., Priest, J., and Yu, B. (2020). Learning epistatic polygenic phenotypes with Boolean interactions. Preprint at bioRxiv. https://doi.org/10.1101/2020.11.24.396846.

75. Dwivedi, R., Tan, Y.S., Park, B., Wei, M., Horgan, K., Madigan, D., and Yu, B. (2020). Stable discovery of interpretable subgroups via calibration in causal studies. Int. Stat. Rev. *88*, S135–S178.

76. Li, X., Tang, T.M., Wang, X., Kocher, J.-P.A., and Yu, B. (2020). A stability-driven protocol for drug response interpretable prediction (staDRIP). Preprint at arXiv. https://doi.org/10.48550/arXiv.2011.06593.

77. Candel, A., Parmar, V., LeDell, E., and Arora, A. (2016). Deep Learning with H2O. H2O (ai Inc), pp. 1–21.

78. Basu, S., Kumbier, K., Brown, J.B., and Yu, B. (2018). Iterative random forests to discover predictive and stable high-order interactions. Proc. Natl. Acad. Sci. USA *115*, 1943–1948.

79. Prentice, R.L., and Pyke, R. (1979). Logistic disease incidence models and case-control studies. Biometrika *66*, 403–411.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Pre-diagnostic sera from the Prostate Lung Colorectal and Ovarian (PLCO) Cancer Screening Cohort	PLCO Cohort	N/A
Plasmas from newly diagnosed resectable PDAC cases, healthy controls, and patients with chronic pancreatitis	Dana-Farber Cancer Institute/Brigham and Women's Hospital (DFCI/BWH), Beth Israel Deaconess Medical Center (BIDMC), and Columbia University Irving Medical Center (CUIMC).	N/A
Deposited data		
MetaboLights	This paper	https://www.ebi.ac.uk/metabolights/editor/MTBLS7260/descriptors
Other		
Acquity™ UPLC BEH amide, 100 Å, 1.7 μm 2.1× 100mm column	Waters Corporation, Milford, USA	catalog number: 176001908
Acquity™ UPLC HSS T3, 100 Å, 1.8 μm, 2.1 × 100mm column	Waters Corporation, Milford, USA	catalog number: 176001132
Ammonium formate (optima LCMS)	ThermoFisher, Waltham, MA, USA	catalog number: A11550
Formic Acid	Honeywell Fluka, Charlotte, NC, USA	catalog number: 60-006-17
LCMS Grade Acetonitrile	ThermoFisher, Waltham, MA, USA	catalog number: A955-4
LCMS Grade Methanol	ThermoFisher, Waltham, MA, USA	catalog number: A456-4
LCMS Grade Isopropanol	ThermoFisher, Waltham, MA, USA	catalog number: A461-4
Metabolomics Data Explorer database	https://sonnenburglab.github.io/Metabolomics_Data_Explorer/#/invivo	N/A

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be direct to and will be fulfilled by the lead contact, Johannes F. Fahrmann, Ph.D. (jffahrmann@mdanderson.org).

Materials availability

- This study did not generate new reagents.
- There are restrictions to the availability of human biospecimens due to existing MTA.

Data and code availability

- Relevant data supporting the findings of this study are available within the Article and Supplemental Materials.
- No new code was generated for this study.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

PLCO cohort

The PLCO Cancer Screening Trial is a randomized multicenter trial in the United States that aimed to evaluate the impact of early detection procedures for prostate, lung, colorectal and ovarian cancer on disease-specific mortality. All subjects involved in this study were enrolled with written consent as a criterion for eligibility to participate in the PLCO trial. Detailed information regarding the PLCO cohort is provided elsewhere.[66,67]

The study included 173 pancreatic cancer cases that were diagnosed within 5 years of blood draw and 863 matched non-cases from 10 participating PLCO study centers (Table 1). Pancreatic cancer cases were identified by self-report in annual mail-in surveys, state cancer registries, death certificates, physician referrals and reports from next of kin for deceased individuals. All medical and pathologic records related to pancreatic cancer diagnosis and supporting documentation were obtained and confirmed by PLCO staff. Pancreatic cancers were classified as localized, regional, distant, or unstaged using the National Cancer Institute Surveillance, Epidemiology, and End Results (SEER) historic staging system. Non-cases, alive at the time when the index case was diagnosed, were matched to cases at a ratio of 5:1 (non-case:case) based on the distribution of age, race, gender, and calendar date of blood draw in 2-month blocks within the case cohort.

Newly diagnosed pancreatic cancer cohort

An independent test set consisted of plasma samples from 99 patients with resected PDAC, 50 patients with chronic pancreatitis, and 100 healthy controls as previously described (Table S5).[67] Patients with pancreatic cancer provided informed written consent to blood collection pretreatment and to clinical data abstraction under DF/HCC (Dana-Farber/Harvard Cancer Center) protocol 12–013. Samples were collected under IRB approved local collection protocols at Dana-Farber Cancer Institute/Brigham and Women's Hospital (DFCI/BWH), Beth Israel Deaconess Medical Center (BIDMC), and Columbia University Irving Medical Center (CUIMC)). Healthy controls were recruited from DFCI/BWH and CUIMC and consisted of subjects undergoing screening colonoscopy or accompanying a non–blood-related patient to an appointment at a gastrointestinal cancer clinic. Healthy controls had no history of cancer in the 5 years before sample collection. Patients with pancreatic cancer and healthy controls were matched on gender and age at the time of blood collection. Patients with chronic pancreatitis (CP) were recruited from gastroenterology clinics at DFCI/BWH, BIDMC, and CUIMC. Patients were included if clinic notes from a gastroenterologist indicated a diagnosis of CP. Patients with pancreatic cancer or CP were not gender or age matched. Clinical data abstraction was performed identically across the sites with data uploaded to a password-protected REDCap database. All plasma samples were collected and processed according to a uniform, standardized protocol across the sites and patient groups.

METHOD DETAILS

Metabolomic analysis

Sample extraction

Serum and plasma metabolites were extracted from pre-aliquoted biospecimen (15 μL) with 45μL of LCMS grade methanol (ThermoFisher) in a 96-well microplate (Eppendorf). Plates were heat sealed, vortexed for 5 min at 750 rpm, and centrifuged at 2000 × g for 10 min at room temperature. The supernatant (30μL) was carefully transferred to a 96-well plate, leaving behind the precipitated protein. The supernatant was further diluted with 60μL of 100mM ammonium formate, pH3 (Fisher Scientific). For Hydrophilic Interaction Liquid Chromatography (HILIC) positive ion analysis, 15μL of the supernatant and ammonium formate mix were diluted with 195μL of 1:3:8:144 water (GenPure ultrapure water system, Thermofisher): LCMS grade methanol (ThermoFisher): 100mM ammonium formate, pH3 (Fisher Scientific): LCMS grade acetonitrile (ThermoFisher). For C18 analysis, 15μL of the supernatant and ammonium formate mix were diluted with 90μL water (GenPure ultrapure water system, Thermofisher) for positive ion mode. Each sample solution was transferred to 384-well microplate (Eppendorf) for LCMS analysis.

Untargeted metabolomic analyses

Untargeted metabolomics analysis was conducted on Waters Acquity UPLC system with 2D column regeneration configuration (I-class and H-class) coupled to a Xevo G2-XS quadrupole time-of-flight (qTOF) mass spectrometer as previously described.[68–71] Chromatographic separation was performed using HILIC (Acquity UPLC BEH amide, 100 Å, 1.7 μm 2.1 × 100mm, Waters Corporation, Milford, U.S.A) and C18 (Acquity UPLC HSS T3, 100 Å, 1.8 μm, 2.1 × 100mm, Water Corporation, Milford, U.S.A) columns at 45°C.

Quaternary solvent system mobile phases were (A) 0.1% formic acid in water, (B) 0.1% formic acid in acetonitrile and (D) 100mM ammonium formate, pH 3. Samples were separated on the HILIC using the following gradient profile at 0.4 mL/min flow rate: (95% B, 5% D) linear change to (70% A, 25% B and 5% D) over 5 min; 100% A for 1 min; and 100% A for 1 min. For C18 separation, the chromatography gradient was as follows at 0.4 mL/min flow rate: 100% A with a linear change to (5% A, 95% B) over 5 min; (95% B, 5% D) for 1 min; and 1 min at (95% B, 5% D).

A binary pump was used for column regeneration and equilibration. The solvent system mobile phases were (A1) 100mM ammonium formate, pH 3, (A2) 0.1% formic in 2-propanol and (B1) 0.1% formic acid in acetonitrile. The HILIC column was stripped using

90% A2 for 5 min at 0.25 mL/min flow rate, followed by a 2 min equilibration using 100% B1 at 0.3 mL/min flow rate. Reverse phase C18 column regeneration was performed using 95% A1, 5% B1 for 2 min followed by column equilibration using 5% A1, 95% B1 for 5 min at 0.4 mL/min flow rate.

Mass spectrometry data acquisition

Mass spectrometry data was acquired using 'sensitivity' mode in positive electrospray ionization mode within 50–800 Da range. For the electrospray acquisition, the capillary voltage was set at 1.5 kV (positive), sample cone voltage 30V, source temperature at 120°C, cone gas flow 50 L/h and desolvation gas flow rate of 800 L/h with scan time of 0.5 s in continuum mode. Leucine Enkephalin; 556.2771 Da (positive) was used for lockspray correction and scans were performed at 0.5s. The injection volume for each sample was 6 μL. The acquisition was carried out with instrument auto gain control to optimize instrument sensitivity over the samples acquisition time.

Data processing

LC-MS and LC-MSe data were processed using Progenesis QI (Nonlinear, Waters). Peak picking and retention time alignment of LC-MS and MSe data were performed using Progenesis QI software (Nonlinear, Waters). Data processing and peak annotations were performed using an in-house automated pipeline as previously described.[68–70,72] Annotations were determined by matching accurate mass and retention times using customized libraries created from authentic standards and by matching experimental tandem mass spectrometry data against the NIST MSMS, LipidBlast or HMDB v3 theoretical fragmentations. To correct for injection order drift, each feature was normalized using data from repeat injections of quality control samples collected every 10 injections throughout the run sequence. Measurement data were smoothed by Locally Weighted Scatterplot Smoothing (LOESS) signal correction (QC-RLSC) as previously described. Values are reported as ratios relative to the median of historical quality control reference samples run with every analytical batch for the given analyte.[68–70,72]

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analysis

Predictive performance estimates for individual microbial-related metabolites identified and quantified through metabolomic profiling of sera were assessed using receiver operating characteristic curve (ROC). Time-dependent ROC analyses were performed using pROC (version 1.15.3) in the R software environment (version 3.6.1, The R Foundation, https://www.r-project.org). The 95% confidence intervals (CI) for AUCs were estimated using the Delong method.[73] Corresponding 95% confidence intervals for odds ratios, adjusted odds ratios, specificity, sensitivity and the difference measurements were calculated using 1,000 bootstrap samples. Age, gender, BMI, and smoking status were included as covariates in the adjusted odds ratio.

Throughout the statistical analysis, we adhered to the PCS (Predictability, Computability and Stability) framework for veridical (trustworthy) data science,[35] which has proven valuable in many previous scientific discoveries including novel gene-gene interaction for the red-hair phenotype,[74] clinically-relevant subgroups in a randomized drug trial,[75] and interpretable drug response prediction.[76] For the modeling stage as in this paper, the PCS framework uses predictability as a reality check, and for reproducibility, it advocates for a stability analysis across different reasonable perturbations of the data and models that pass the prediction check. Under this framework, the entire PLCO specimen set was divided into (1) a Development Set that was used for training and tuning the models (Training Set) and model selection (Validation Set) and (2) a set-aside Test Set for obtaining an unbiased evaluation of the selected final model (Figure 2; Table S1). The Development Set consisted of case and non-case sera from seven of the ten PLCO study centers; the set-aside Test Set consisted of case and non-case sera from the remaining three PLCO study centers.

Seven different learning algorithms were evaluated including a deep learning model (fully-connected feedforward network), gradient boosting machine, auto-machine learning, iterative random forest, logistic regression with LASSO (L_1) regularization, logistic regression with ridge (L_2) regularization, and logistic regression models. Deep neural network, extreme gradient boosting, and auto machine learning algorithms were performed using the h2o package in R.[77] Iterative random forest was run using the iRF package in R.[78] To further evaluate model stability in accordance with PCS framework, data perturbations (e.g., via random selection and replacement) were introduced to the Development Set and the performance re-assessed. Based on AUC, a LASSO regression model with 3 selected microbial-associate metabolites (an indole-derivative, TMAO, and indoleacrylic acid) that showed the highest and most stable predictive performance was selected for subsequent testing in the set-aside Test Set as well as the independent newly diagnosed PDAC cohort.

To select the non-microbiome metabolites, the adjusted odds ratio and corresponding p value for each feature were calculated and corrected using Benjamini-Hochberg in the training set in which 12 metabolites showed an adjusted odds ratio greater than 1 with adjusted p value less than 0.05. Five out of 12 features yielded significant p values and adjusted odds ratio greater than 1 in the Validation Set. The prediction performance of the combined five non-microbiome features trained in the training set using logistic regression was evaluated against the microbiome metabolite panel and CA19-9 in the testing set.

For the combination of 3-marker microbial-related metabolite panel, non-microbiome metabolite panel and CA19-9, we fit a logistic regression with three separate predictors, one corresponding to each of the aforementioned features. This model was developed in the Development Set and validated in the set-aside Test Set.

Samples assayed via metabolomics herein reflect a nested case-control cohort that enriches for cases and, therefore, do not reflect the true risk of pancreatic cancer in the general population. In order to determine the 0.5%, 1%, 1.5% and 2% 5-year risk of pancreatic cancer, we thus adjust the estimates to reflect the entire PLCO study population using the approach of Prentice et al.[79] In this approach, a prospective logistic model is estimated from the case-control study that includes an offset term to the logistic model. The offset term is the logit of the prevalence in the population minus the logit of the prevalence in the analyzed dataset. Briefly, absolute risk values for each biomarker were estimated by calculating coefficients of a logistic regression in the training set and the intercept adjusted using the following equation:

$$Risk = \frac{\exp\left(\beta_0' + \beta_1 \times (model)\right)}{1 + \exp\left(\beta_0' + \beta_1 \times (model)\right)},$$

where

$$\beta_0' = \beta_0 - \log\left(\frac{P_{data}}{1 - P_{data}}\right) + \log\left(\frac{P_{Population}}{1 - P_{Population}}\right).$$

In this equation, β_0 is the intercept derived from logistic regression in the nested case-control within a cohort, P_{data} is the prevalence of the disease in our case-enriched dataset, $P_{Population}$ is the prevalence of the disease in the general population, *model* represents the predicted score derived from the selected model and β_1 is the corresponding coefficient for the model score.

ADDITIONAL RESOURCES

Microbial-associated metabolite database

To evaluate the association between the microbial-associated metabolites identified in the PLCO specimen sets with distinct microbial species, we used the Metabolomics Data Explorer database (https://sonnenburglab.github.io/Metabolomics_Data_Explorer/#/invivo) developed by Shuo Han and colleagues.[30] The database reports the metabolic profiles of 178 gut microorganism strains; microbiota-dependent metabolites were established in diverse biological fluids from gnotobiotic and conventionally colonized mice and traced back to the corresponding metabolomic profiles of cultured bacteria.[30]

Cell Reports Medicine, Volume *4*

Supplemental information

A blood-based metabolomic signature predictive of risk for pancreatic cancer

Ehsan Irajizad, Ana Kenney, Tiffany Tang, Jody Vykoukal, Ranran Wu, Eunice Murage, Jennifer B. Dennison, Marta Sans, James P. Long, Maureen Loftus, John A. Chabot, Michael D. Kluger, Fay Kastrinos, Lauren Brais, Ana Babic, Kunal Jajoo, Linda S. Lee, Thomas E. Clancy, Kimmie Ng, Andrea Bullock, Jeanine M. Genkinger, Anirban Maitra, Kim-Anh Do, Bin Yu, Brian M. Wolpin, Sam Hanash, and Johannes F. Fahrmann

Contents

Supplementary Table S1 (related to Table 3). Patient characteristics for the PLCO Development Set and the set-aside Test Set.

	Development Set					
	Training Set		Validation Set		Set-Aside Test Set	
	Non-cases	Cases	Non-cases	Cases	Non-cases	Cases
Total	494	102	142	33	225	37
Gender, N (%)						
Female	204 (41)	41 (40)	61 (43)	17 (52)	91 (40)	13 (35)
Male	290 (59)	61 (60)	81 (57)	16 (48)	134 (60)	24 (65)
Age At Randomization, N (%)						
<= 59	116 (23)	21 (21)	22 (15)	11 (33)	45 (20)	5 (14)
60-64	108 (22)	24 (24)	34 (24)	3 (9)	63 (28)	14 (38)
65-69	192 (39)	41 (40)	50 (35)	9 (27)	79 (35)	13 (35)
>= 70	78 (16)	16 (16)	36 (25)	10 (30)	38 (17)	5 (14)
Race, N (%)						
White	463 (94)	99 (97)	107 (75)	24 (73)	211 (94)	33 (89)
Black	22 (4)	3 (3)	2 (1)	1 (3)	6 (3)	2 (5)
Other	9 (2)	0 (0)	33 (23)	8 (24)	8 (4)	2 (5)

Supplementary Table S2 (related to Table 2). Selected microbial-associated metabolites and corresponding model coefficients in LASSO regression.

	Lasso selection	
Metabolite	**Selected in model**	**Coefficient**
AcetylCadaverine	-	-
5-hydroxy-L-tryptophan	-	-
5-methoxy-3-indoleacetic acid	-	-
Indole-3-lactic acid	-	-
Indoleacrylic acid	Yes	0.3653
Glycodeoxycholate	-	-
Indole-3-acetaldehyde	-	-
Indole-3-ethanol	-	-
Indole-derivative_2	Yes	0.5022
Indole-derivative_1	-	-
TMAO	Yes	0.2412
Deoxycholate	-	-
Indole-3-acetamide	-	-
Indole-3-acetate	-	-

Supplementary Table S3 (related to Table 2). Stability check of the LASSO regression using perturbed training data and evaluated on the Validation Set for the 3-marker microbial panel.

	Perturbations	AUC (95% CI)	Adj OR†
Lasso regression with 3 selected features	2 randomly selected centers	0.63 (0.44-0.82)	1.37 (0.89-2.09)
	2 randomly selected centers	0.73 (0.60-0.86)	2.33 (1.52-3.77)
	2 randomly selected centers	0.54 (0.41-0.68)	1.25 (0.90-1.73)
	2 randomly selected centers	0.55 (0.45-0.63)	1.27 (0.92-1.72)
	3 randomly selected centers	0.64 (0.54-0.73)	1.65 (1.23-2.24)
	300 random samples	0.60 (0.51-0.68)	1.40 (1.02-1.90)

† Age, gender, BMI, and smoking status were included as covariates in adjusted odds ratios (ORs)

Supplementary Table S4 (related to Table 3). Performance of the 3-marker microbial panel, the 5-marker non-microbial panel, and the combined (microbial+non-microbial) metabolite pannel amongst diabetic and non-diabetic individuals in the PLCO set-aside Test Set. † Age, gender, BMI, and smoking status were included as covariates in adjusted odds ratios (ORs); odds ratio per unit SD increase. N0: Number of non-cases, N1: Number of cases.

	3-marker microbial panel							
	Diabetics				**Non-Diabetic**			
	Sample Size	**AUC (95% CI)**	**Adj. OR (95% CI)†**	**P-value**	**Sample Size**	**AUC (95% CI)**	**Adj. OR (95% CI)†**	**P-value**
PLCO Testing Set	N0 = 14	0.62	0.8	0.77	N0 = 210	0.64	1.84	<0.001
	N1 = 4	(0.22-1.00)	(0.09-3.61)		N1 = 33	(0.53-0.77)	(1.32-2.61)	
All PLCO samples	N0 = 55	0.6	1.56	0.13	N0 = 805	0.62	1.5	<0.001
	N1 = 22	(0.46-0.74)	(0.88-2.95)		N1 = 150	(0.57-0.67)	(1.27-1.77)	
	5-marker non-microbial panel							
	Diabetics				**Non-Diabetic**			
	Sample Size	**AUC (95% CI)**	**Adj. OR (95% CI)†**	**P-value**	**Sample Size**	**AUC (95% CI)**	**Adj. OR (95% CI)†**	**P-value**
PLCO Testing Set	N0 = 14	0.65	1.93	0.43	N0 = 210	0.75	2.74	<0.001
	N1 = 4	(0.27-1.00)	(0.45-17.61)		N1 = 33	(0.65-0.84)	(1.83-4.32)	
All PLCO samples	N0 = 55	0.67	2.67	0.004	N0 = 805	0.74	2.95	<0.001
	N1 = 22	(0.52-0.82)	(1.44-5.72)		N1 = 150	(0.70-0.78)	(2.12-3.20)	
	Combined (microbial+non-microbial) Panel							
	Diabetics				**Non-Diabetic**			
	Sample Size	**AUC (95% CI)**	**Adj. OR (95% CI)†**	**P-value**	**Sample Size**	**AUC (95% CI)**	**Adj. OR (95% CI)†**	**P-value**
PLCO Testing Set	N0 = 14	0.65	1.7	0.52	N0 = 210	0.81	3.39	<0.001
	N1 = 4	(0.29-1.00)	(0.38-13.34)		N1 = 33	(0.72-0.89)	(2.19-5.61)	
All PLCO samples	N0 = 55	0.67	2.71	0.004	N0 = 805	0.76	2.79	<0.001
	N1 = 22	(0.53-0.81)	(1.44-5.84)		N1 = 150	(0.72-0.80)	(2.27-3.46)	

Supplementary Table S5 (related to Table 3). Patient and tumor characteristics for the newly-diagnosed PDAC cohort.

Variable	PDAC Case (N=99)		Chronic Pancreatitis (N=50)		Healthy Control (N=100)	
	No.	%	No.	%	No.	%
Institution						
DF/BWCC	69	70%	30	60%	94	94%
BIDMC	15	15%	15	30%	0	0%
CUMC	15	15%	5	10%	6	6%
Age (year), median (IQR)	69.8 (62.5-74.8)		65.4 (54.7-72.2)		63.7 (55.7-70.6)	
Gender						
Male	51	52%	33	66%	51	51%
Female	48	48%	17	34%	49	49%
Race						
White	94	95%	42	84%	84	86%
Black/African-American	0	0%	5	10%	5	5%
Asian	1	1%	0	0%	2	2%
Other	4	4%	3	6%	7	7%
Blood collection year						
2015-2016	19	19%	2	4%	0	0%
2017-2019	80	81%	48	96%	100	100%
Smoking Status						
Current Smoker	6	6%	11	22%	4	4%
Past smoker	50	51%	17	34%	42	42%
Never smoker	43	43%	22	44%	54	54%
BMI (kg/m^2), Meidan(IQR)	27.4 (24.0-30.0)		25.0 (22.8-27.6)		27.5 (24.3-32.0)	
Diabetes						
No	64	65%	23	46%	93	93%
Yes	35	35%	27	54%	7	7%
Etiology of chronic pancreatitis						

Alcohol	-	-	16	32%	-	-
Autoimmune	-	-	2	4%	-	-
Congenital anatomical variant	-	-	3	6%	-	-
Duct stricture or stones	-	-	7	14%	-	-
Idiopathic	-	-	21	42%	-	-
Other	-	-	1	2%	-	-
AJCC 8th edition staging pTNM[a]						
T0-2N0M0	15	24%	-	-	-	-
T3-4N0M0	2	3%	-	-	-	-
T1-4N1M0	28	45%	-	-	-	-
T1-4N2M0	17	28%	-	-	-	-
AJCC 8th edition staging ypTNM[b]						
T0-2N0M0	24	64%	-	-	-	-
T3-4N0M0	1	3%	-	-	-	-
T1-4N1M0	7	19%	-	-	-	-
T1-4N2M0	5	14%	-	-	-	-
PDAC recurrence						
No[c]	56	57%	-	-	-	-
Yes	43	43%	-	-	-	-

DF/BWCC: Dana-Farber/Brigham and Women's Cancer Center; BIDMC: Beth Israel Deaconess Medical Center; CUMC: Columbia University Medical Center
AJCC: American Joint Committee on Cancer, PDAC: Pancreatic ductal adenocarcinoma, BMI: Body mass index
[a]Patients who underwent up-front surgical resection
[b]Patients who received neoadjuvant treatment and then underwent surgical resection
[c]The median (IQR) follow-up time was 15.0 (7.2-23.2) months for patients without cancer recurrence

Supplementary Table S6 (related to Figure 2). Performance of all non-microbial metabolites in the PLCO Training and Validation Sets.

See excel file.

Supplementary Table S7 (related to Table 3). Selected non-microbial metabolites.

	Training - 5 centers		Validation- 2 centers	
Name	**Adj. Odds Ratio†**	***P-value (FDR) ‡***	**Adj. Odds Ratio†**	***P-value£***
Cholesterol glucuronide	1.735	<0.001	1.720	0.006
Galactosamine	1.749	<0.001	1.514	0.035
2-Hydroxyglutarate	1.857	<0.001	1.738	0.006
Erythritol	1.688	<0.001	1.532	0.030
Glucose	1.744	<0.001	1.662	0.018

† Age, gender, BMI, and smoking status were included as covariates in adjusted odds ratios (ORs); odds ratio per unit SD increase
‡ Benjamini and Hochberg adjusted p-values
£ Raw p-values

Supplementary Table S8 (related to Table 3). Performance of different learning models based on non-microbial metabolites and model stability check in the PLCO Validation Set. † Age, gender, BMI, and smoking status were included as covariates in adjusted odds ratios (ORs); odds ratio per unit SD increase.

Model	Hyperparameters	AUC (95% CI)	Adj OR†
Logistic regression	-	0.72 (0.63-0.81)	2.10 (1.04-2.90)
logistic regression with ridge (L2) regularization	Penalty weight = 0.18	0.69 (0.58-0.78)	1.74 (1.20-2.25)
logistic regression with LASSO (L1)	Penalty weight = 0.01, number of selected features = 4	0.71 (0.54-0.73)	2.08 (0.94-2.83)
Iterative Random Forest	Number of iterations = 3	0.60 (0.49-0.72)	1.44 (0.90-1.90)
Deep neural network model	Number of cross-validation folds = 6, hidden layers = 3 with 32 nodes in each layer	0.59 (0.48-0.68)	1.43 (0.95-2.10)
GBM	Number of trees = 42, max depth= 5	0.58 (0.46-0.67)	1.30 (0.93-1.87)
Auto ML	Selected model = randomized trees	0.66 (0.52-0.72)	1.85 (1.50-2.02)

	Perturbations	AUC (95% CI)	Adj OR†
Logistic regression with 5 selected features	2 randomly selected centers	0.71 (0.52-0.87)	2.10 (1.10-2.94)
	2 randomly selected centers	0.74 (0.61-0.91)	2.33 (1.42-4.10)
	2 randomly selected centers	0.69 (0.59-0.80)	2.11 (0.90-2.73)
	2 randomly selected centers	0.67 (0.45-0.85)	1.90 (1.12-2.72)
	3 randomly selected centers	0.60 (0.52-0.68)	1.65 (1.23-2.24)
	300 random samples	0.64 (0.55-0.71)	1.73 (1.52-2.20)

Supplementary Table S9 (related to Table 3). Performance of the 5-marker non-microbial panel in the PLCO set-aside Test Set and the entire specimen set.

Set-aside Test Set				
		5-marker non-microbial panel [a]		
Time to Dx	Sample Size	AUC. (95% CI)	Adj. OR † (95% CI)	*P-value*
[0-5)	N0 = 225 N1 = 37	0.74 (0.65 - 0.83)	2.72 (1.83 - 4.24)	<0.001
[0-2)	N0 = 225 N1 = 24	0.82 (0.72 - 0.92)	4.03 (2.41 - 7.32)	<0.001
[2-5)	N0 = 225 N1 = 13	0.59 (0.44 - 0.72)	1.32 (0.71 - 2.41)	0.36
Entire Set (Development + Set-aside Test Set)				
		5-marker non-microbial panel [a]		
Time to Dx	Sample Size	AUC. (95% CI)	Adj. OR † (95% CI)	*P-value*
[0-5)	N0 = 861 N1 = 172	0.74 (0.67 - 0.77)	2.59 (2.13 - 3.18)	<0.001
[0-2)	N0 = 861 N1 = 92	0.80 (0.75 - 0.85)	3.69 (2.83 - 4.91)	<0.001
[2-5)	N0 = 861 N1 = 80	0.65 (0.59 - 0.72)	1.74 (1.37 – 2.21)	<0.001

† Age, gender, BMI, and smoking status were included as covariates in adjusted odds ratios (ORs); odds ratio per unit SD increase
N0: number of non-cases
N1: number of cases
a: Non-microbial-related metabolite signature includes cholesterol glucuronide, hydroxyglutarate, galactosamine, glucose, and erythritol

Supplementary Table S10 (related to Figure 2 and Table 4). Performance of the combined metabolite panel plus CA19-9 stratified by diabetic status.

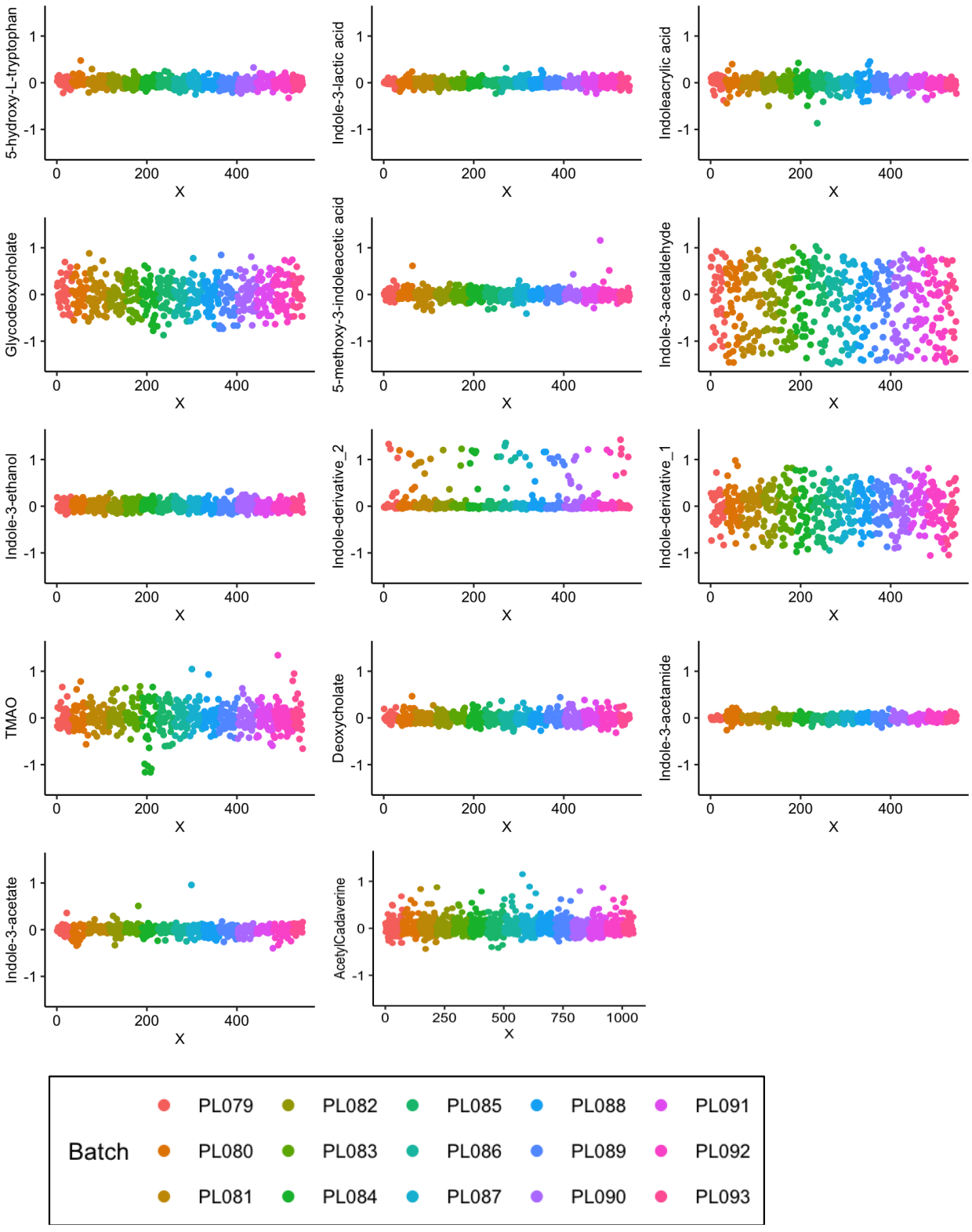
	Diabetics				Non-Diabetic			
5-marker non-microbial panel + 3-marker microbial panel + CA19.9	**Sample Size**	**AUC (95% CI)**	**Adj. OR (95% CI) †**	**P-value**	**Sample Size**	**AUC (95% CI)**	**Adj. OR (95% CI)†**	**P-value**
PLCO Testing Set	N0 = 14 N1 = 4	0.78 (0.50-1.00)	6.82 (1.14-210.61)	0.10	N0 = 210 N1 = 33	0.84 (0.76-0.92)	10.21 (4.55-26.61)	<0.001
All PLCO samples	N0 = 55 N1 = 22	0.71 (0.60-0.84)	3.75 (1.81-9.72)	0.001	N0 = 805 N1 = 150	0.80 (0.76-0.84)	9.54 (6.36-14.75)	<0.001

† Age, gender, BMI, and smoking status were included as covariates in adjusted odds ratios (ORs); odds ratio per unit SD increase

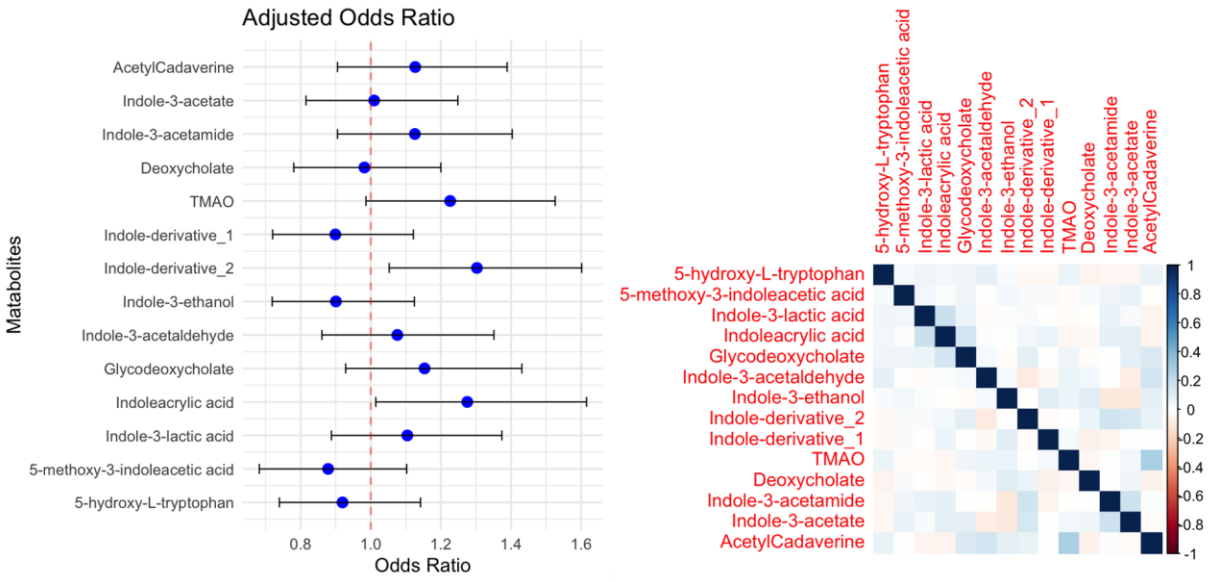
N0: number of non-cases

N1: number of cases

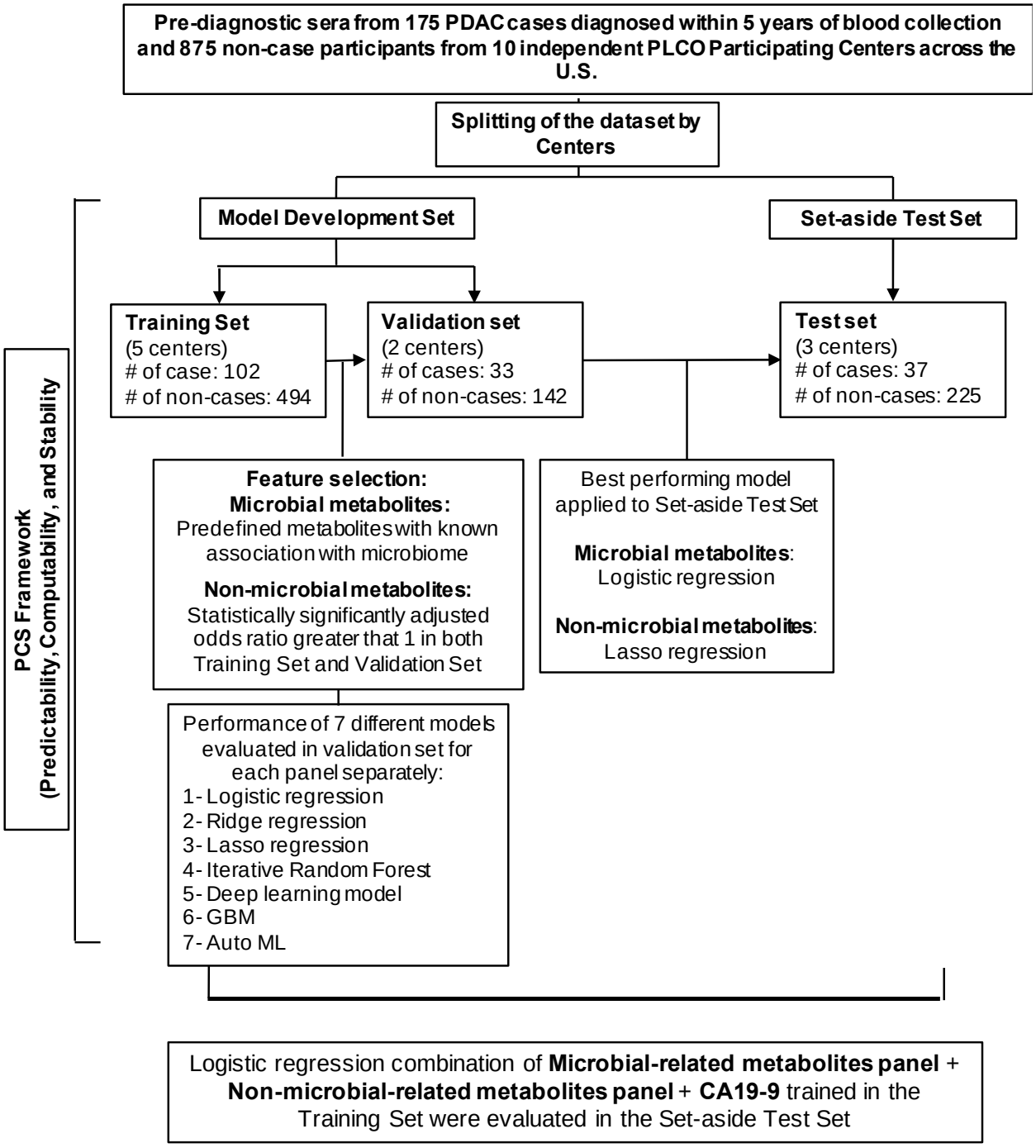
Supplementary Figure S1 (related to Figure 1 and Table 2). Distribution plots for detected microbial-related metabolites across analytical batches in the PLCO specimen set. X-axis represents individual specimens.



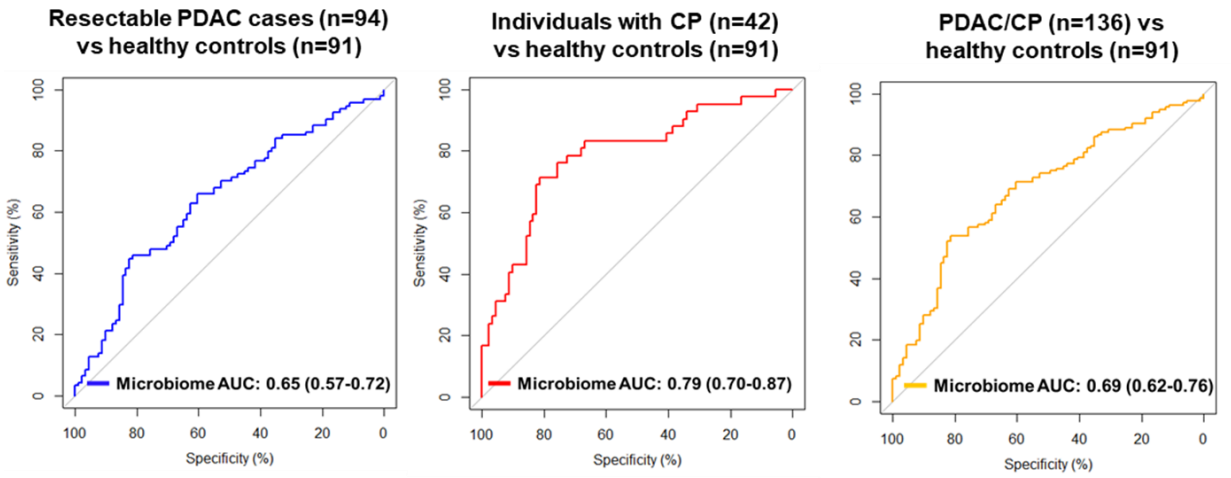
Supplementary Figure S2 (related to Figure 1 and Table 2). Odds ratios, adjusted odds ratios, and correlations for individual microbial-related metabolites for risk of pancreatic cancer in the Training Set. Gender, age, smoking status, and BMI were included as covariates in adjusted odds ratios.



Supplementary Figure S3 (related to Table 2). Workflow of analyses.



Supplementary Figure S4 (related to Table 3). Predictive performance of the 3-marker microbial panel in the independent newly-diagnosed PDAC cohort.
Abbreviation: CP- chronic pancreatitis. A subset samples were excluded due to insufficient sample volume or not having passed quality control criteria.



Odds Ratio (95% CI)		
Resectable PDAC cases vs healthy controls	Individuals with CP vs healthy controls	PDAC/CP vs healthy controls
1.55 (1.13-2.23)	2.83 (1.83-4.82)	2.07 (1.45-3.18)