Supporting information for:

"A novel approach for selecting combinatorial clinical markers of pathology applied to a large retrospective cohort of surgically resected pancreatic cysts"

Algorithmic overview

MOCA begins by converting user-provided data to a binary (dichotomous) format (see Figure S1, below). To estimate the significance and performance of individual parameters for predicting phenotype, MOCA populates a two-by-two contingency table with the phenotype and each clinical parameter individually, and calculates the Fisher's exact two-tailed P-value, sensitivity, and specificity (see example calculation in Figure S2, below); in this study, the phenotype is any one of the pancreatic cyst types or grades. All P-values are corrected using the Benjamini and Hochberg false discovery rate (FDR).

A

	$Patient_1$	$Patient_2$	$Patient_3$	$Patient_4$	$Patient_5$	$Patient_6$	$Patient_7$	$Patient_8$	$Patient_9$
IPMN	1	0	1	1	1	0	1	1	0

B

	$Patient_1$	$Patient_2$	$Patient_3$	$Patient_4$	$Patient_5$	$Patient_6$	$Patient_7$	$Patient_8$	$Patient_9$
Pancreatitis	0	0	1	0	0	0	1	1	0
Smoking history	Never	Used to	Never	Never	Never	Current	Current	Used to	Never
Age	77	51	81	72	79	41	72	69	59

C

	$Patient_1$	$Patient_2$	$Patient_3$	$Patient_4$	$Patient_5$	$Patient_6$	$Patient_7$	$Patient_8$	$Patient_9$
IPMN	1	0	1	1	1	0	1	1	0
Pancreatitis	0	0	1	0	0	0	1	1	0
Never smoked	1	0	1	1	1	0	0	0	1
Used to smoke	0	1	0	0	0	0	0	1	0
Current smoker	0	0	0	0	0	1	1	0	0
Age ≥ 75	1	0	1	0	1	0	0	0	0

Figure S1: Toy example illustrating the conversion of user-provided data types during MOCA analysis. Figure S1A is an example of a phenotype feature, showing nine hypothetical patients whose phenotype is either IPMN positive ("1") or IPMN negative ("0"). Figure S1B illustrates the three possible types of data variables, which are binary, categorical, and continuous. In this toy example, *Pancreatitis* is a binary feature, because a patient is diagnosed as having acute pancreatitis ("1") or not ("0"). *Smoking history* is an example of categorical variable, because in this toy example it can be one of three values (*Never*, *Used to*, or *Current*). The *Age* feature is an example of a continuous variable. Figure S1C illustrates how MOCA would convert and merge the example data from S1A and S1B. Because MOCA analysis ultimately requires binary variables, the phenotype (IPMN vs. non-IPMN) and pancreatitis features are unchanged (first two rows in S1C). The three-category smoking feature is converted to three binary features, each representing one of the three possible smoking statuses. For continuous-valued data such as age, MOCA applies many thresholds to binarize (i.e., dichotomize) the variable. Here, for simplicity, we show only the threshold defining patients ≥ 75 years of age as positive for the age feature, whereas patients < 75 years of age are defined as negative for the age feature (final row, S1C).

MOCA composite marker derivation

Composite markers are those that combine multiple individual clinical parameters. MOCA derives composite markers by randomly selecting sets of individual parameters, and testing every possible combination of the selected parameters using Boolean logic operations (see Figures S2 and S3, below). Next, each of the resulting composite markers is compared to the phenotype of interest, and the corresponding P-value, sensitivity, and specificity are recorded (Figures S2C and S2D). This process, of randomly selecting parameters and comparing every parameter combination with the phenotype of interest, is repeated 10,000 times; every 1,000 iterations, composite markers with the top 1% balanced accuracy (arithmetic mean of sensitivity and specificity) are decomposed and appended back to the initial parameter pool. Therefore, as the calculation progresses, the probability of selecting the most informative parameters for composite marker derivation increases. This optimization process rapidly converges on composite markers with optimal performance for predicting the phenotype of interest. For every tested marker, the Fisher's exact two-tailed P-value is corrected using the Benjamini and Hochberg FDR; only markers with an FDR-corrected P-value < 0.05 were considered for further analysis.

For this study, each random sampling (see previous paragraph) included six individual parameters. Therefore, for a single cycle of composite marker selection, MOCA tested 6.3 X 10^5 composite markers (see equation S1).

$$\textit{Equation S1}: \left(\sum_{r=1}^{n} \frac{n!}{(n-r)!r!}\right) \cdot 10{,}000 \ (where\ n = 6)$$

For each phenotype tested, MOCA selected composite markers using two cycles. First, composite markers were derived using the Boolean logic union operation. Second, the composite marker with the highest balanced accuracy from that selection process was used to initiate a cycle to form combinations using the Boolean difference operation. Toy examples of composite markers resulting from the union operation are shown in Figures S3A and S3B. An example of a final marker resulting from the inclusion of the difference operation is shown in Figure S3C. For each phenotype, the total tested combinations were 1.26 X 10^6 (Equation S1, for two cycles).

A

	$Patient_1$	$Patient_2$	$Patient_3$	$Patient_4$	$Patient_5$	$Patient_6$	$Patient_7$	$Patient_8$	$Patient_9$
Pancreatitis	0	0	1	0	0	0	1	1	0
Age ≥ 75	1	0	1	0	1	0	0	0	0
Pancreatitis + Age ≥ 75	1	0	1	0	1	0	1	1	0

B

	$Patient_1$	$Patient_2$	$Patient_3$	$Patient_4$	$Patient_5$	$Patient_6$	$Patient_7$	$Patient_8$	$Patient_9$
IPMN	1	0	1	1	1	0	1	1	0
Pancreatitis + Age ≥ 75	1	0	1	0	1	0	1	1	0

C

	IPMN	non-IPMN
Pancreatitis + Age ≥ 75	5	0
$\overline{\text{Pancreatitis + Age ≥ 75}}$	1	3

D

Fisher's two-tailed P-value	Sensitivity	Specificity
0.048	83.3%	100%

Figure S2: Example of MOCA-derived composite feature and corresponding statistical significance and performance for predicting phenotype. The "+" symbol denotes the Boolean logic union (OR) operation. Figure S2A shows two individual clinical parameters from Figure S1C combined using the Boolean logic union operation to form a composite parameter. The union operation (denoted by "+" symbol) defines a composite marker as positive if any of the constituent markers are positive. For instance, $Patient_5$ is positive for the *Pancreatitis + Age ≥ 75* composite marker because $Patient_5$ is positive for at least one of the constituent markers (in this case, *Age ≥ 75*). Conversely, $Patient_9$ is negative for the composite marker, because $Patient_9$ does not have pancreatitis and is not ≥ 75 years (denoted as $\overline{\textit{Pancreatitis + Age ≥ 75}}$). Figure S2B shows the IPMN phenotype parameter and the composite clinical parameter, and S2C is a two-by-two contingency table populated by each of these features. This table indicates that: five patients were both IPMN positive and positive for the composite marker (i.e., true positives); zero IPMN-negative patients were positive for the composite marker (false positives); one IPMN-positive patient is negative for the composite marker (false negative); and three patients were both IPMN negative and negative for the composite marker (true negatives). Figure S2D is the Fisher's two-tailed P-value that results from the contingency table in S2C, and the corresponding statistical sensitivity and specificity.

For each phenotype, markers were selected using a 10-fold cross-validation strategy. For instance, when selecting composite makers for identifying SCA-positive patients, 921 patients were used to select composite markers (training set) and the remaining 103 used to assess the predictive performance of the selected markers (testing set). This process was then repeated for each of 10 data splits, assuring that each sample was present in only one of the testing sets, and that approximately the same number of

samples for a given phenotype were included in each of 10 testing data partitions (e.g., 33 of the 322 SCAs in each of the first nine test sets, and 25 SCAs in the final test set). For each phenotype, any composite marker considered for further analysis was required to be selected in each of the ten cross-validation calculations. This requirement assures that only the most predictive markers from 10 separate calculations are returned for further analysis. Because 1.26 X 10^6 composite markers were tested for each of ten 10-fold cross-validation calculations, 1.3 X 10^7 composite makers were tested for each of five phenotypes considered in this study (5.15 X 10^7 total composite markers tested). In total, calculations required approximately one-hour compute time on a 256-core cluster running 2.60-2.66 GHz AMD Opteron and Intel Xeon processors.

MOCA is freely available for nonprofit use and can be downloaded at: http://karchinlab.org/apps/appMoca.html.

A

	$Patient_1$	$Patient_2$	$Patient_3$	$Patient_4$	$Patient_5$	$Patient_6$	$Patient_7$	$Patient_8$	$Patient_9$
P_1	0	0	1	0	0	0	1	1	0
P_2	1	0	1	0	1	0	0	0	0
$P_1 + P_2$	1	0	1	0	1	0	1	1	0

B

	$Patient_1$	$Patient_2$	$Patient_3$	$Patient_4$	$Patient_5$	$Patient_6$	$Patient_7$	$Patient_8$	$Patient_9$
P_1	0	0	1	0	0	0	1	1	0
P_2	1	0	1	0	1	0	0	0	0
$\overline{P_1 + P_2}$	0	1	0	1	0	1	0	0	1

C

	$Patient_1$	$Patient_2$	$Patient_3$	$Patient_4$	$Patient_5$	$Patient_6$	$Patient_7$	$Patient_8$	$Patient_9$
P_1	0	0	1	0	0	0	1	1	0
P_2	1	0	1	0	1	0	0	0	0
P_3	1	1	1	0	1	0	1	0	0
$(P_1 + P_2) - P_3$	0	0	0	0	0	0	0	1	0

Figure S3: Interpreting the MOCA Boolean logic operations utilized in this study. Figure S3A illustrates the union of hypothetical parameters P_1 and P_2; the composite marker $P_1 + P_2$ is positive if either P_1 or P_2 is positive (the OR operation; similar to Figure S2A). The Boolean logic NOR operation is the opposite of the OR operation, where $\overline{P_1 + P_2}$ is negative if either P_1 or P_2 is positive (Figure S3B). The three-parameter marker in Figure S3C is positive if either P_1 or P_2 is positive, provided P_3 is not positive; therefore, only $Patient_8$ is positive for the $(P_1 + P_2) - P_3$ composite marker.

	All patients (N=1026)	N* (%)	IPMN (N=584)	SCA (N=322)	MCN (N=78)	SPN (N=42)
Age at surgery (years), median (IQR)	66 (54-74)	1026 (100)	69 (61-76)	62 (50-71)	51 (40-58)	36 (25-43)
Female sex, N (%)	623 (60.72)	1026 (100)	287 (49.14)	228 (70.81)	73 (93.59)	35 (83.33)
Race, N (%)						
White	844 (82.34)		519 (88.87)	241 (74.84)	57 (74.03)	27 (64.29)
African American	96 (9.37)	1025 (99.9)	34 (5.82)	42 (13.04)	13 (16.88)	7 (16.67)
Other	85 (8.29)		31 (5.31)	39 (12.11)	7 (9.09)	8 (19.05)
Symptoms, N (%)						
Abdominal pain	296 (29.11)	1024 (99.8)	262 (44.86)	14 (4.44)	17 (22.37)	3 (7.14)
Weight loss	172 (16.93)	1017 (99.1)	150 (25.73)	20 (6.35)	2 (2.63)	0
Acute pancreatitis	28 (2.75)	1024 (99.8)	15 (2.57)	2 (0.63)	9 (11.84)	2 (4.76)
Jaundice	102 (10.03)	1024 (99.8)	89 (15.24)	10 (3.17)	3 (3.95)	0
Diabetes	163 (16.38)	1013 (98.7)	117 (20.45)	37 (11.75)	7 (9.09)	2 (6.45)
Cyst size (cm), median (IQR)	3.15 (2.1-4.6)	658 (64.1)	2.9 (2-3.9)	3.6 (2.3-5.4)	4 (3-6.7)	4 (3-6)
Cyst location, N (%)^						
Head/uncinate	378 (52.65)		280 (66.19)	79 (43.17)	9 (11.69)	10 (28.57)
Neck	41 (5.71)	707 (68.9)	29 (6.86)	10 (5.46)	1 (1.3)	1 (2.86)
Body/tail	338 (47.08)		144 (34.04)	105 (57.38)	66 (85.71)	23 (65.71)
Multiple cysts, N (%)	188 (25.44)	739 (72)	172 (38.74)	16 (8.65)	0	0
MPD communication, N (%)	256 (34.32)	746 (72.7)	255 (78.22)	0	0	1 (3.13)
MPD dilation > 5mm, N (%)	174 (20.74)	839 (81.8)	169 (40.24)	0	2 (3.08)	3 (9.38)
Mural nodule, N (%)	190 (28.4)	669 (65.2)	106 (26.63)	57 (31.84)	14 (20)	13 (59.09)
CEA>192 ng/ml, N (%)	25 (41.67)	60 (5.8)	16 (59.26)	0	9 (64.29)	0
Grade of dysplasia/invasive cancer, N (%)						
Low	138 (13.5)		81 (13.97)	Not applicable	57 (73.08)	Not applicable
Intermediate	198 (19.37)	1026 (100)	189 (32.59)	Not applicable	9 (11.54)	Not applicable
High	137 (13.41)		129 (22.24)	Not applicable	8 (10.26)	Not applicable
Invasive cancer	185 (18.1)		181 (31.21)	Not applicable	4 (5.13)	Not applicable

Table S1: Cyst and patient characteristics for the 1026-patient marker-selection cohort. *Data recorded in N (%) number of patients; IPMN: Intraductal papillary mucinous neoplasm; MCN: Mucinous cystic neoplasm; SCA: Serous cystadenoma; SPN: Solid-pseudopapillary neoplasm; IQR: Interquartile range; MPD: main pancreatic duct; CEA: Carcinoembryonic antigen.

	All Samples N = 130	IPMN* N = 96	MCN N =12	SCA N =12	SPN N =10
Sex (n=130^)					
Female - no. (%)	83 (64)	52 (54)	12 (100)	9 (75)	10 (100)
Race (n=130^)					
African American - no. (%)	7 (5)	4 (4)	1 (8)	1 (8)	1 (10)
White - no. (%)	114 (88)	87 (91)	10 (83)	8 (67)	9 (90)
Other - no. (%)	9 (7)	5 (5)	1 (8)	3 (25)	0 (0)
Age at surgery (n=130^)					
Years - mean (SD)	62.3 (17)	69.2 (10)	45.5 (14)	54.9 (15)	24.1 (4)
Symptoms (n=130^)					
Abdominal pain - no. (%)	23 (18)	21 (22)	2 (17)	0 (0)	0 (0)
Pancreatitis - no. (%)	16 (12)	14 (15)	2 (17)	0 (0)	0 (0)
Jaundice - no. (%)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Weight loss - no. (%)	6 (5)	5 (5)	1 (8)	0 (0)	0 (0)
Diabetes - no. (%)	25 (19)	21 (22)	1 (8)	2 (17)	1 (10)
Cyst size (n=130^)					
cm - median (IQR)	3.4 (2)	3 (2)	5.2 (4)	4 (2)	4 (2)
Cyst location (n=130^)+					
Head or Uncinate – no. (%)	56 (45)	51 (55)	0 (0)	1 (8)	4 (40)
Neck – no. (%)	16 (13)	13 (14)	0 (0)	2 (17)	1 (10)
Body or Tail – no. (%)	65 (52)	42 (45)	12 (100)	9 (75)	5 (50)
Multiple cysts (n=124^)					
Yes - no. (%)	31 (25)	29 (31)	0 (0)	1 (8)	1 (11)
Communication with MPD (n=107^)					
Yes - no. (%)	39 (36)	38 (46)	1 (14)	0 (0)	0 (0)
Mural Nodule (n=123^)					
Yes - no. (%)	32 (26)	19 (20)	3 (30)	2 (17)	8 (100)
CEA > 192 ng/mL (n=51^)					
Yes - no. (%)	30 (59)	24 (60)	6 (86)	0 (0)	0 (0)
IPMN Histotype (n=95^)					
Gastric - no. (%)	-	64 (67)	-	-	-
Intestinal - no. (%)	-	11 (11)	-	-	-
Pancreatobiliary - no. (%)	-	9 (9)	-	-	-
Oncocytic - no. (%)	-	3 (3)	-	-	-
Mixed - no. (%)°	-	8 (8)	-	-	-
Grade of Dysplasia/Invasive Cancer in IPMNs, and MCNs (n=108^)					
Low - no. (%)	25 (23)	15 (16)	10 (83)	-	-
Intermediate - no. (%)	49 (45.5)	47 (49)	2 (17)	-	-
High - no. (%)	22 (20.5)	22 (23)	0 (0)	-	-
Invasive Cancer - no. (%)	12 (11)	12 (12)	0 (0)	-	-

Table S2: Cyst and patient characteristics for the 130-patient validation cohort. *Includes one ITPN. ^Indicates the number of patients in whom data on this variable was available. +All sites where a pancreatic cyst was located were documented. Since some of the cysts extended to more than one site, this resulted in a number of locations that was greater than the number of cysts. °More than one histologic subtype in the same lesion. SD = standard deviation. IQR = Interquartile range.