Journal of the American Medical Informatics Association, 24(1), 2017, 145–152
doi: 10.1093/jamia/ocw069
Advance Access Publication Date: 21 June 2016
Research and Applications

Research and Applications

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

David L Masica,[*,1,8] **Marco Dal Molin,**[*,2,8] **Christopher L Wolfgang,**[3,5,8] **Tyler Tomita,**[1] **Mohammad R Ostovaneh,**[4] **Amanda Blackford,**[7] **Robert A Moran,**[4] **Joanna K Law,**[4] **Thomas Barkley,**[2] **Michael Goggins,**[4,5,8] **Marcia Irene Canto,**[4] **Meredith Pittman,**[2,8] **James R Eshleman,**[8,9] **Syed Z Ali,**[2] **Elliot K Fishman,**[6] **Ihab R Kamel,**[6] **Siva P Raman,**[6] **Atif Zaheer,**[6] **Nita Ahuja,**[3] **Martin A Makary,**[3] **Matthew J Weiss,**[3] **Kenzo Hirose,**[3] **John L Cameron,**[3] **Neda Rezaee,**[3] **Jin He,**[3] **Young Joon Ahn,**[3] **Wenchuan Wu,**[3] **Yuxuan Wang,**[8,9] **Simeon Springer,**[8,9] **Luis L Diaz Jr,**[3,8,9] **Nickolas Papadopoulos,**[8,9] **Ralph H Hruban,**[2,5,8,9] **Kenneth W Kinzler,**[8,9] **Bert Vogelstein,**[8,9] **Rachel Karchin,**[†,1,5,8] **and Anne Marie Lennon**[†,3,4,8]

[1]Department of Biomedical Engineering and the Institute for Computational Medicine, The Johns Hopkins University, Baltimore, Maryland, [2]Departments of Pathology, [3]Departments of Surgery, [4]Departments of Medicine, [5]Departments of Oncology, [6]Departments of Radiology, [7]Departments of Biostatistics and Bioinformatics, [8]Departments of the Sol Goldman Pancreatic Cancer Research Center, [9]Departments of the Ludwig Center and Howard Hughes Medical Institute at the Sidney Kimmel Cancer Center, The Johns Hopkins Medical Institutions, Baltimore, Maryland. *Drs Masica and Dal Molin contributed equally as first authors, and †Drs Lennon and Karchin contributed equally as senior authors

Correspondence to Anne Marie Lennon, MD, PhD: The Ludwig Center and Howard Hughes Medical Institute at the Sidney Kimmel Cancer Center, The Johns Hopkins Medical Institutions,1800 Orleans St, Baltimore, Maryland 21287, USA; amlennon@jhmi.edu; fax: (410) 614-8337; Rachel Karchin, PhD, The Department of Biomedical Engineering and the Institute for Computational Medicine, The Johns Hopkins University, 3400 N. Charles St, Baltimore, Maryland 21218, USA; karchin@jhu.edu; fax: (410) 516-5294.

Received 7 December 2015; Revised 2 March 2016; Accepted 7 April 2016

ABSTRACT

Objective: Our objective was to develop an approach for selecting combinatorial markers of pathology from diverse clinical data types. We demonstrate this approach on the problem of pancreatic cyst classification.

Materials and Methods: We analyzed 1026 patients with surgically resected pancreatic cysts, comprising 584 intraductal papillary mucinous neoplasms, 332 serous cystadenomas, 78 mucinous cystic neoplasms, and 42 solid-pseudopapillary neoplasms. To derive optimal markers for cyst classification from the preoperative clinical and radiological data, we developed a statistical approach for combining any number of categorical, dichotomous, or continuous-valued clinical parameters into individual predictors of pathology. The approach is unbiased and statistically rigorous. Millions of feature combinations were tested using 10-fold cross-validation, and the most informative features were validated in an independent cohort of 130 patients with surgically resected pancreatic cysts.

Results: We identified combinatorial clinical markers that classified serous cystadenomas with 95% sensitivity and 83% specificity; solid-pseudopapillary neoplasms with 89% sensitivity and 86% specificity; mucinous cystic neoplasms with 91% sensitivity and 83% specificity; and intraductal papillary mucinous neoplasms with 94%

sensitivity and 90% specificity. No individual features were as accurate as the combination markers. We further validated these combinatorial markers on an independent cohort of 130 pancreatic cysts, and achieved high and well-balanced accuracies. Overall sensitivity and specificity for identifying patients requiring surgical resection was 84% and 81%, respectively.

Conclusions: Our approach identified combinatorial markers for pancreatic cyst classification that had improved performance relative to the individual features they comprise. In principle, this approach can be applied to any clinical dataset comprising dichotomous, categorical, and continuous-valued parameters.

Key words: pancreatic cyst, mucinous cyst, IPMN, MOCA, combination marker, clinical model, composite marker

BACKGROUND AND SIGNIFICANCE

Single-parameter markers of pathology are typically unreliable or have low accuracy. Composite markers, defined as comprising multiple diagnostic features or parameters (synonymous with "combination" or "multiparameter" marker), and methods for their discovery have received recent and increased attention.[1–5] When designed properly, composite markers increase statistical power relative to the individual features they comprise and improve important diagnostic parameters such as sensitivity and specificity.

The potential to increase statistical power is appealing, but the derivation of optimal multiparameter markers can be hampered by large combinatorics (possible number of ways to combine individual parameters). For instance, 20 individual parameters can be combined to generate $>10^6$ unique combinations. Another important consideration is a subtle one: the addition of features that indicate a disease is present cannot increase the *specificity* of the resulting composite marker. Combining features that indicate a disease is present can increase the sensitivity of the composite marker, but increased specificity can only result from features that indicate the disease is absent. Therefore, a simultaneous increase of both sensitivity and specificity requires the composite marker to include 2 types of features, a subset of which indicates that the disease is present and others that indicate the disease is absent; this requirement exponentially increases the combinatorics problem. Because the probability of falsely inferring association increases with every feature tested, searches for composite markers require strict control of the false discovery rate. These searches greatly benefit from large cohorts amenable to cross-validation so that markers can be selected and validated in numerous independent groupings of the data.

The challenge of evaluating patients with pancreatic cysts is one area where combinatorial analysis could be helpful. Advances in cross-sectional imaging have resulted in the frequent detection of pancreatic cysts, but these advances have outpaced our ability to reliably distinguish benign cysts from those with significant malignant potential, presenting a dilemma for clinical management.[6] On the one hand, invasive pancreatic ductal adenocarcinomas have the lowest 5-year survival rate among all cancers,[7] indicating a clear opportunity to save lives by resecting pancreatic cysts with even a modest risk of progressing to invasive cancer. On the other hand, many cysts are entirely benign or have a minimal risk of progression, making conservative management (ie, surveillance) the preferred option.[7] Thus, there is a need for sensitive and specific ways to distinguish high-risk from low-risk cysts prior to surgery.

The 4 most common pancreatic cyst types highlight the abovementioned challenge. Serous cystadenomas (SCAs) are nonmucinous lesions that can develop anywhere in the pancreas.[8] Virtually all SCAs are benign, and surveillance is recommended for asymptomatic patients.[9] Mucinous cystic neoplasms (MCNs) are solitary, well-defined cysts typically located in the body or tail of the pancreas, and are more frequently present in middle-aged women.[6] Because MCNs can progress to invasive carcinoma, resection rather than lifelong surveillance is recommended for this type of disease.[10] Solid-pseudopapillary neoplasms (SPNs) are rare indolent neoplasms that are seen almost exclusively in young women.[6,11] Because of their locally invasive and metastatic potential, SPNs should undergo surgical resection. Intraductal papillary mucinous neoplasms (IPMNs) are the most commonly observed type of neoplastic pancreatic cyst, and have the potential to progress from low- to intermediate- to high-grade dysplasia and ultimately to invasive cancer.[12] However, this progression occurs in only a small number of patients with IPMNs.[13] Ideally, IPMNs with high-grade dysplasia or an associated invasive adenocarcinoma should be considered for resection, while asymptomatic IPMNs with low- or intermediate-grade dysplasia may undergo surveillance.[10] Thus, markers that can reliably distinguish cyst type and IPMN cyst grade are of critical importance for directing treatment in patients with pancreatic cysts.

Three types of preoperative features are currently used to predict the type of pancreatic cyst: clinical characteristics, imaging characteristics, and cyst fluid analysis.[6] Single features have been employed to identify cyst type. An example of this is cyst fluid carcinoembryonic antigen (CEA), which is often used to differentiate IPMNs and MCNs from other cyst types. However, recent studies have shown that cyst fluid CEA alone has an accuracy of just over 60% for distinguishing IPMNs and MCNs from other types of cysts.[14] Because of the low sensitivities and specificities of individual features, a combination of features is often used to analyze pancreatic cysts. In one study, combining cyst fluid CEA, cytology, and molecular markers increased the accuracy from 62.5% for CEA alone, and to 72.9% for the combined markers.[14]

The Multivariate Organization of Combinatorial Alterations (MOCA) algorithm is a bioinformatics approach for identifying individual and composite genetic markers of phenotype.[15–17] The algorithm can rapidly test millions of combinations of markers while conservatively controlling the false discovery rate and selecting the most predictive and recurrent feature combinations from cross-validation testing. MOCA has previously been used to identify genes important in glioblastoma progression[15] and to predict response to anticancer therapeutics.[16] Here, we further develop MOCA to consider diverse types of clinical data, and select composite markers of cyst type and grade from a large cohort of patients with surgically resected pancreatic cysts.

MATERIALS AND METHODS

This study was approved by the Johns Hopkins School of Medicine Institutional Review Board.

Patients

Consecutive patients with SCAs, SPNs, MCNs, or IPMNs who had undergone surgical resection at a single institution between January 1995 and January 2014 were identified from a prospective surgical database. Medical records were reviewed and general demographics and symptoms documented.

Imaging and pathology

Preoperative imaging was identified and relevant features documented. When patients had more than one imaging modality performed, the largest size documented on computed tomography (CT) or magnetic resonance imaging (MRI) was used for the cyst size and main pancreatic duct diameter. Dilation of the main pancreatic duct was defined as a diameter $\geq$5 mm. Communication between the main pancreatic duct and the cyst was preferentially assessed with MRI, or, if this was not available, with CT or endoscopic ultrasound (EUS). A solid component included the presence of a mural nodule, enhancing solid component, or thickened wall seen on any imaging modality. Multiple cysts were defined as the presence of more than one cyst within the pancreas, which were anatomically separate from each other. Surgically resected cysts were classified pathologically based on the 2010 World Health Organization classification.[18]

Selection of clinical features and composite markers

We used our previously published MOCA algorithm[16] to select composite clinical markers of pancreatic cyst type and grade. For this work, a required improvement to the algorithm was the ability to accept and combine dichotomous, categorical, and continuous-valued parameters into individual predictors. The addition of categorical parameter handling is requisite for considering clinical features that have 3 or more possible classes, which is common.

We restricted individual feature selection to those that were recorded in at least 50% of patients (see Supplemental Table S1). Restricting selection to features recorded in $\geq$50% of the cases was enforced: first, to prevent any potential bias toward features that were recorded in only a small subset of the total cohort, thereby having the potential to result in inflated performance statistics. Second, it is difficult to compare the performance of features that are assessed over very different sample sizes. Third, features recorded in a small subset of the cohort could reduce the power of the 10-fold cross-validation.

All composite markers were selected using 10-fold cross-validation, and false discoveries were controlled using the Benjamini and Hochberg method. MOCA has been described previously.[15–17] A detailed description of the MOCA implementation used in the study is provided in *Supporting Information*.

Independent validation of composite markers

To validate performance independently from that obtained during the 10-fold cross-validation selection process, we tested top-performing composite markers in an independent and non-overlapping cohort of 130 patients with surgically resected pancreatic cysts (patient demographics and cyst characteristics are presented in Supplemental Table S1).

RESULTS AND DISCUSSION

General demographics: marker-selection dataset

A total of 1036 patients with surgically resected pancreatic cysts were identified, of whom 10 with an IPMN were excluded because the cyst was incidentally identified in the surgical specimen after resection of a different pancreatic neoplasm. The final marker-selection cohort consisted of 1026 patients, which included 584 IPMNs, 322 SCAs, 78 MCNs, and 42 SPNs. The general demographics, symptoms, clinical features, and pathological diagnoses are detailed in Supplemental Table S1. Complete data were not available in all patients. To account for this deficit, the percentages presented in this table reflected the number of patients who were positive for the feature divided by the total number of patients for which the data of that type were available. This 1026-patient cohort is referred to as the *marker-selection cohort* throughout.

General demographics: validation dataset

Clinical data from an independent and non-overlapping cohort of 130 patients[19] were used as a validation dataset. That cohort included 96 patients with IPMNs, 12 with SCAs, 12 with MCNs, and 10 with SPNs. Patient demographics and cyst characteristics for this cohort are presented in Supplemental Table S2. This 130-patient cohort is referred to as the *validation cohort* throughout.

Individual clinical features that distinguish cyst types

We used MOCA to select individual clinical features for discriminating the 4 cyst types from the 1026-patient marker-selection cohort (Table 1). In general, these individual clinical features had low balanced accuracy (defined as the arithmetic mean of sensitivity and specificity), providing high specificity but low sensitivity, or vice versa. The features in Table 1 also highlight the fact that highly significant *Q*-values (*P*-values corrected for multiple testing) do not necessarily translate to good clinical predictors. Many of the features have *Q*-values between 10^{-10} and 10^{-5}, but have sensitivities or specificities that are too low for clinical utility.

No SCA had main pancreatic duct dilation, nor was there any observed main pancreatic duct communication, giving each of these features 100% sensitivity (Table 1); however, specificity was low in both cases because these features were uncommon in SPNs and MCNs. The age feature had the highest balanced accuracy of any of the SPN features (Table 1). SCAs and SPNs, and in particular MCNs, are more common in women, giving the gender feature moderate to high sensitivity for these 3 cyst types. However, the high prevalence of these 3 cyst types in women necessarily means that the associated specificities will be low for any 1 of the cyst types. The majority of patients over 65 years of age had IPMNs (Table 1). The highest balanced accuracy of any feature in Table 1 resulted from IPMN classification with the main pancreatic duct communication feature.

Composite markers that distinguish cyst types

Next, we used MOCA to select the composite markers that best distinguished cyst type in the marker-selection cohort. Each composite marker contained up to 6 features indicative of a particular cyst type and up to 6 features that excluded that cyst type. We required that indicative markers be “present” and that exclusionary markers be “absent” for the composite marker to be scored as positive for a particular cyst type (Table 2). The most informative composite marker could in principle be composed of any mixture of present and absent features.

The composite markers listed here had the highest balanced accuracy for identifying each cyst type, selected using 10-fold cross-validation from the *marker selection* cohort. Further validation in the independent *validation* cohort is also shown. A composite

Table 1. Selected individual features for discriminating pancreatic cysts by type

Cyst type	Clinical Feature	Sensitivity (%)	Specificity (%)	*Q*-value
SCA	No communication with MPD	100	60	3.6×10^{-9}
	No MPD dilation	100	34	3.6×10^{-9}
	Single cyst	91	42	4.0×10^{-9}
	Sex (female)	71	44	6.1×10^{-5}
SPN	Age <55 years	90	77	2.2×10^{-10}
	Single cyst	100	35	2.1×10^{-6}
	Sex (female)	83	40	5.5×10^{-3}
	Location, body/tail	65	63	5.3×10^{-3}
MCN	Sex (female)	94	42	8.9×10^{-10}
	Location, body/tail	80	67	8.9×10^{-10}
	Single cyst	100	37	8.9×10^{-10}
	No communication with MPD	100	38	8.9×10^{-10}
IPMN	Age ≥65 years	67	66	6.9×10^{-10}
	Multiple cysts	39	95	6.9×10^{-10}
	Location, head/ uncinate/neck	61	68	6.9×10^{-10}
	Communication with MPD	78	100	6.9×10^{-10}

MPD = main pancreatic duct. Sensitivity and specificity for identifying each of 4 cyst types using the corresponding features are listed. *Q*-values are Fisher's exact 2-tailed *P*-values corrected using the Benjamini and Hochberg false discovery rate.

Table 2. Top composite markers for discriminating the four pancreatic cyst types.

Cyst type	Any of these *Present*	All of these *absent*	Marker selection (*n* = 1026) (sensitivity, specificity)	Validation (*n* = 130) (sensitivity, specificity)
SCA	Age ≥25	Abdominal pain Communication with MPD MPD dilation	95%, 83%	100%, 71%
SPN	Age <55	Jaundice Multifocal Weight loss	89%, 86%	89%, 85%
MCN	Age <75	Male Communication with MPD Multifocal	91%, 83%	90%, 78%
IPMN	Age ≥85 Communication with MPD MPD dilation Abdominal pain	None	94%, 90%	75%, 88%

marker was scored positive for the corresponding cyst type if any 1 of the clinical features in the second column is present and all features from the third column are absent. MPD = main pancreatic duct.

The SCA composite marker described in Table 2 accurately identified 95% of SCA cases in our cohort while accurately excluding 83% of all non-SCAs. The high specificity of this composite marker resulted from the fact that it accurately excluded IPMNs, which classically communicate with the main pancreatic duct and can cause main pancreatic duct dilation or pancreatitis. Similarly, the composite marker excluded SPNs by implementing a low age threshold.

The MOCA algorithm revealed that pancreatic cysts identified in patients under 55 years of age were likely to be SPNs provided the patient did not have multiple cysts, jaundice, or weight loss. A composite marker based on these features had a sensitivity of 89% and a specificity of 86% for SPNs (Table 2). MOCA also showed that patients were most likely to have MCNs if their age was <75 years, provided they were not male, there was no communication between their cysts and the main pancreatic duct, and the patient did not have multiple cysts. The composite MCN marker based on these features had a sensitivity and specificity of 91% and 83%, respectively (Table 2). Its high specificity was achieved by accurately rejecting IPMNs, which communicate with the main pancreatic duct, often have multiple cysts, and are usually found in older individuals. Finally, MOCA classified cysts as IPMNs if the patient's age was ≥85, if there was communication with the main pancreatic duct, if the main pancreatic duct was dilated, or if the patient had abdominal pain. The composite IPMN marker based on these features had a sensitivity of 94% and specificity of 90%.

The markers in Table 2 were selected using 10-fold cross-validation with strict correction for multiple testing, and the requirement that any composite marker eligible for further analysis had to be selected independently in all 10 data splits. To further validate our model, we applied the markers from Table 2 to the independent validation cohort of 130 patients.[19] The results from this validation are shown in the last column in Table 2. The SCA marker validated with a 3.5% decrease in balanced accuracy, maintaining high sensitivity and reasonable specificity. Both the SPN and MCN markers validated with nearly identical sensitivities and specificities. The IPMN marker achieved a balanced accuracy of 82%, which represents a 10% decrease relative to that achieved during cross-validation selection.

During composite marker derivation, the MOCA algorithm tests millions of composite markers, of which thousands can be statistically significant (see Supplementary Materials and Methods). Statistically significant composite markers of a single cyst type can share many similar individual features. For example, 2 composite markers can differ only by the presence or absence of a single feature or by a small difference in continuous-valued features such as age. Therefore, it was instructive to decompose the top-performing composite markers into their individual features to determine the relative contribution of each feature to the marker's performance. Figure 1A shows the occurrence of the most frequently observed individual clinical features in the 100 composite markers with the highest balanced accuracy for identifying IPMNs.

Among these 100 IPMN markers, communication with the main pancreatic duct was the most frequently identified feature, contributing to each of the 100 composite markers (Figure 1A); this result is not surprising considering the high balanced accuracy of the communication feature shown in Table 1. A dilated main pancreatic duct was the second most informative feature, selected in 70% of the top composite IPMNs. Sixty-three percent of the top composite IPMN markers included age, where the minimum age selected was 80 years, just slightly younger than the age cutoff of 85 selected by the top composite marker for IPMNs (Table 2). The multiple cysts feature was selected in 38% of the top composite markers, but this feature was not in the top-performing marker. Abdominal pain and

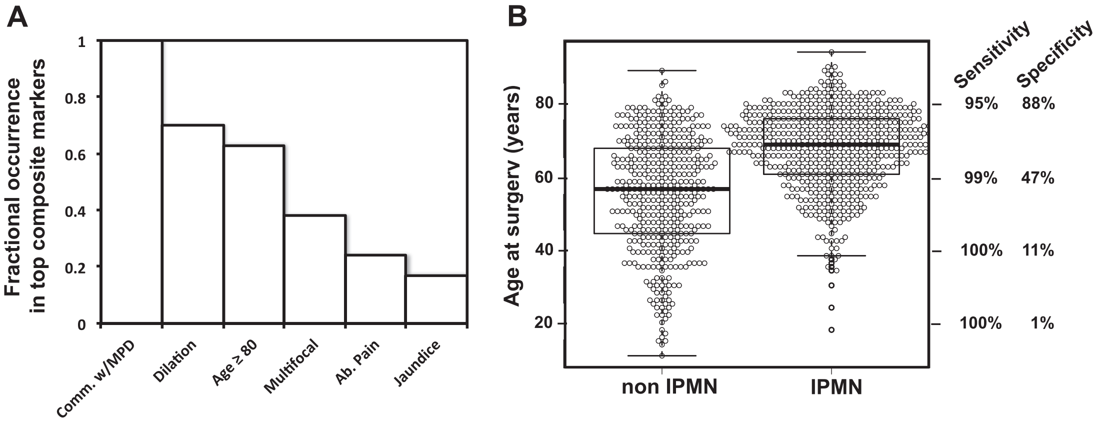

Figure 1. Contribution of individual clinical features to the identification of IPMNs. (**A**) The top 100 composite markers for identifying IPMNs decomposed into their individual clinical features. (**B**) Distribution of age and corresponding variable sensitivity and specificity (right *x*-axis) for varying age in the context of the composite IPMN marker from Table 2. MPD = main pancreatic duct. Comm. w/MPD = communication with the main pancreatic duct. Ab. Pain = abdominal pain. The sensitivities and specificities were calculated in the context of the corresponding composite marker (Table 2).

Table 3. Selected individual clinical features for identifying high-risk IPMNs among all IPMNs

Clinical Feature	Sensitivity (%)	Specificity (%)	*Q*-value
Jaundice	24	95	8.6×10^{-9}
Weight loss	35	85	4.0×10^{-7}
MPD dilation	58	75	8.6×10^{-9}
Diabetes	25	85	1.6×10^{-2}
Mural nodule	37	83	1.9×10^{-4}

Q-values represent Fisher's exact 2-tailed *P*-values corrected using the Benjamini and Hochberg false discovery rate. MPD = main pancreatic duct.

jaundice were also informative symptoms for IPMNs when found to be present.

Continuous-valued features such as age and cyst size can be useful because they can provide more information than the mere presence or absence of a feature (as with binary-valued features). As an example, the box and whisker plot in Figure 1B shows a distinct distribution of ages separating IPMNs from non-IPMNs. Figure 1B also shows how the performance of the top-performing composite IPMN marker (Table 2) varied with age. Between ages 80 and 60 years, the sensitivity of the composite marker increased by 4%, while the specificity dropped by 41%; this would be acceptable only if a highly sensitive test was desired. From 60 to 40 years, the imbalance is high, with a sensitivity increase of only 1% at the expense of a 36% decrease in specificity.

Composite markers can help identify cysts that should be surgically excised

Management of IPMNs depends on whether clinical and radiologic features suggest the presence of high-risk IPMNs (defined as those with high-grade dysplasia or associate invasive adenocarcinoma). MOCA selected composite clinical markers for this high-risk group in the subset of patients who were pathologically proven to have an IPMN. Of the 584 IPMNs, 310 (58%) were high-risk.

Significant individual features for identifying high-risk IPMNs generally had high specificity but low sensitivity (Table 3). In particular, jaundice, weight loss, diabetes mellitus, and presence of a solid component had specificities of 95%, 85%, 85%, and 83%, respectively. This result indicates that, in this cohort, these features are almost exclusive to patients with high-risk IPMNs; however, the low sensitivities (24 – 58%) indicate that none of these individual features were present in the majority of patients with high-risk IPMNs. The main pancreatic duct dilation feature was slightly more balanced, present in the majority of high-risk IPMNs and absent from the majority of low- and intermediate-grade cases.

Table 4. Top composite marker for identifying high-risk IPMNs among all IPMNs

Classification	Any of these present	All of these absent	Marker selection (*N* = 584) (sensitivity, specificity)	Validation (*N* = 96) (sensitivity, specificity)
High-grade/ invasive IPMN	Jaundice MPD dilation Cyst diameter ≥4 cm	None	81%, 61%	70%, 69%

The top composite marker was defined as the one with the highest balanced accuracy for identifying high risk IPMNs, selected using 10-fold cross validation from the *marker selection* cohort. Further validation in the independent *validation* cohort is also shown. The composite marker was scored as positive if any 1 of the 4 clinical features was present. MPD = main pancreatic duct.

The composite marker with the highest balanced accuracy selected for identifying high-risk IPMNs combined cyst size, jaundice, and main pancreatic duct dilation, and achieved a sensitivity of 81% and specificity of 61% (Table 4). Thus, in the marker-selection cohort, 81% of all high-risk IPMNs occurred in patients who had jaundice, main pancreatic duct dilation, or a cyst larger ≥4 cm; 39% of low- or intermediate-grade IPMNs had at least 1 of these features. When applied to the independent validation cohort, the balanced accuracy decreased by 1.5%, with decreased sensitivity and increased specificity (Table 4).

Figure 2A shows the ensemble distribution of individual features contributing to the 100 top-performing composite markers for identifying high-risk IPMN. Cyst size, main pancreatic duct dilation, and jaundice were informative features, in agreement with the composition of the top composite marker. Weight loss and presence of a

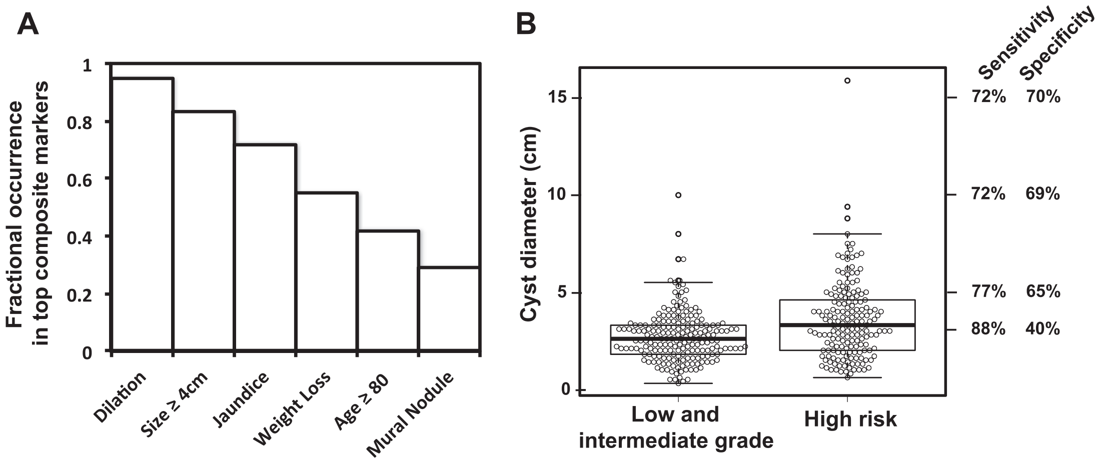

Figure 2. Distribution of individual clinical features among top-performing composite markers. (**A**) The top 100 composite markers for identifying high-risk IPMNs were decomposed into their individual clinical features. (**B**) Distributions of cyst diameter among IPMNs and the corresponding sensitivity and specificity for either of these 2 features in the context of the top-performing composite marker listed. The sensitivities and specificities were calculated in the context of the corresponding composite marker (Table 4).

mural nodule were informative features during composite marker selection and also highly specific individual features (Table 3); however, these features were not part of the top-performing composite marker. Age was a moderately informative feature during selection; however, age conferred no predictive benefit to the top-performing composite marker for high-risk IPMNs.

Figure 2B shows the distribution of the continuous-valued cyst-size feature from the top-performing composite marker for identifying high-risk IPMNs, and how the performance of the composite marker changes as this feature is varied. As predicted by the composite markers, patients with high-risk IPMNs tended to have larger cysts than those with low- or intermediate-grade IPMNs (Figure 2B). In this cohort, low- and intermediate-grade IPMNs had a median size of 2.6 cm, with a reasonably narrow IQR. High-risk IPMNs had a median cyst size of 3.3 cm, with a distribution skewed toward larger diameters (Figure 2B).

One of the most important characteristics of a marker for pancreatic cysts is the identification of those cysts that require resection. We used the markers from Tables 2 and 4 to discriminate cysts that would generally be considered not to require resection (SCA, low- and intermediate-grade IPMNs) from those that require resection (SPN, MCNs, high-risk IPMNs). The composite marker had a sensitivity of 84% and a specificity of 81% for identifying the cysts that required surgery in our 1026-patient marker-selection cohort. In the independent validation cohort, these markers discriminated cysts requiring resection from those not requiring resection with 77% sensitivity and 75% specificity.

CONCLUSIONS

For this study we developed a computational approach for deriving composite markers from diverse clinical data types, and selected markers of pancreatic cyst type (Table 2) and IPMN grade from the largest single series of surgically resected pancreatic cysts published to date. The composite markers presented here each comprised well-known individual features for their respective cyst types and grades. Indeed, the association between high-risk IPMNs and "worrisome" clinical symptoms, the distinctly young age associated with SPN diagnosis, the relative absence of MCNs in males, and the prevalence of main pancreatic duct dilation and communication in patients with IPMNs have all been previously described.[6,10,11,20] One of the major novel aspects of our study involved forging these well-known individual observations into composite markers whose performance was superior to the features in isolation.

Another novel aspect is the use of advanced bioinformatics methods to quantitatively assess the importance of each feature evaluated and the relative performance of millions of different composite markers. This was achieved through a nonbiased computational approach that utilized cross-validation and strict control of false-positive discoveries. Notably, other statistically rigorous multivariate feature-selection methods exist, and they produce a type of classifier distinct from the composite markers derived by MOCA. Some commonly used approaches include Adaptive Boosting,[21] Elastic Net Regression,[22] and Random Forests.[23] Although these methods can achieve good classification performance, they produce classifiers that can be difficult to interpret. The "black box" nature of these machine-learning approaches is fundamentally limiting to the pursuit of biological insight and hypothesis generation. Rather than producing complex algorithmic classifiers, MOCA derives simple human-interpretable markers that are amenable to comparison with clinical literature and hypothesis generation.

The key composite markers discriminating pancreatic cysts by type have reasonably balanced accuracies (Table 2). It is important to consider the relative prevalence of each cyst type when interpreting the performance of these composite markers. As an example, a marker for identifying SCAs will have a minimum specificity of 84% if it can at least exclude all IPMNs; this is because 84% of all non-SCAs in this cohort are IPMNs. Indeed, the ability to exclude diagnosis, and in turn maintain high specificity, is integral to the design of composite markers such as those presented in this work. However, it should be recognized that marker performance would vary as a function of the distribution of cyst types and grades for a given cohort. This problem exists for any imperfect marker, and is important for considering translational potential.

This study has limitations with respect to pancreatic cyst classification. It is retrospective, and complete data were not present on all patients. Molecular markers were not included in this study, but have shown promise in other studies, and their inclusion could further enhance the accuracy of the markers presented here.[18,24,25] Although we used 10-fold cross-validation with subsequent validation in an independent cohort, all possible data bias could not be corrected. For example, patients from both cohorts all underwent surgical resection, but standard guidelines dictate surveillance, rather than resection, for most SCAs and non-high-risk IPMNs. This suggests that SCAs and low- and intermediate-grade IPMNs in this study might have features that are not perfectly typical of their population averages. Given that surgical pathology is the most reliable method for verifying cyst type and grade, this is not an easy bias to overcome. Longitudinal studies where benign cysts are followed for progression over many years might help address this limitation.

It is also informative to consider limitations in the computational methodology and implementation. While the results are encouraging and selected markers achieve relatively high diagnostic performance, this single test case is not sufficient to guide expectations regarding performance in other disease classification problems. MOCA is inherently a data analysis method, and performance will be subject to the common limitations of data quality, quantity, and efficiency of algorithmic search strategies (ie, searching the possible combinations of markers). Furthermore, the size imbalance between our cross-validation (marker-selection) cohort and the holdout validation cohort leaves uncertainty with respect to the translational potential of the presented composite markers. These markers need further testing in large cohorts before their real clinical utility can be considered.

The method presented here contributes a statistically rigorous and unbiased approach to deriving composite clinical markers of pathology. Of great practical importance, the method combines individual clinical parameters that indicate the presence of a disease with those that indicate the absence of the disease, which is required to optimize sensitivity and specificity simultaneously. In principle, MOCA can be applied to any dataset comprising binary, categorical, and continuous-valued clinical parameters, and could complement traditional and *ad hoc* marker-selection approaches in many contexts and clinical settings.

FUNDING

This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research, the Sol Goldman Center for Pancreatic Cancer Research, the Virginia and D.K. Ludwig Fund for Cancer Research, Susan Wojcicki and Dennis Troper, the Michael Rolfe Foundation, and National Institutes of Health grant P50 CA62924.

COMPETING INTERESTS

Drs Diaz, Papadopoulos, Kinzler, and Vogelstein B. are founders of Personal Genome Diagnostics, Inc., and PapGene, Inc. Drs Vogelstein B. and Kinzler are also members of the Scientific Advisory Board of Syxmex-Inostics. These companies and others have licensed technologies from Johns Hopkins, of which Drs Diaz, Hruban, Klein, Goggins, Papadopoulos, Kinzler, and Vogelstein are inventors and receive royalties from these licenses. The terms of these arrangements are being managed by the university in accordance with its conflict of interest policies.

Drs Masica, Dal Molin, Blackford, Wolfgang, Tomita, Ostovaneh, Moran, Law, Barkley, Canto, Pittman, Eshleman, Ali, Fishman, Kamel, Raman, Zaheer, Ahuja, Makary, Weiss, Hirose, Cameron, Rezaee, He, Ahn, Wu, Wang, Springer, Karchin, and Lennon have no disclosures.

CONTRIBUTORS

Drs Masica, Dal Molin, Blackford, Tomita, Ostovaneh, Pittman, Raman, Moran, Law, Barkley, Rezaee, He, Ahn, Wu, Karchin, Lennon, Hruban, Kinzler, and Vogelstein were involved in the study conception, performed data acquisition, and were involved in editing and final approval of the manuscript.

Drs Wolfgang, Canto, Eshleman, Ali, Fishman, Kamel, Zaheer, Ahuja, Makary, Weiss, Hirose, Cameron, Wang, Springer, Goggins, Diaz, and Papadopoulos were involved in the conception and design of the study, data interpretation, editing, and final approval of the manuscript.

REFERENCES

1. Man M, Close SL, Shaw AD, *et al*. Beyond single-marker analyses: mining whole genome scans for insights into treatment responses in severe sepsis. *Pharmacogenomics J*. 2013;13(3):218–226.
2. McMahon CG, Lamont JV, Curtin E, *et al*. Diagnostic accuracy of heart-type fatty acid-binding protein for the early diagnosis of acute myocardial infarction. *Am J Emerg Med*. 2012;30(2):267–274.
3. Valva P, Casciato P, Diaz Carrasco JM, *et al*. The role of serum biomarkers in predicting fibrosis progression in pediatric and adult hepatitis C virus chronic infection. *PLoS One*. 2011;6(8):e23218.
4. Su Kim D, Choi YD, Moon M, *et al*. Composite three-marker assay for early detection of kidney cancer. *Cancer Epidemiol Biomarkers Prev.* 2013;22(3): 390–398.
5. French B, Saha-Chaudhuri P, Ky B, Cappola TP, Heagerty PJ. Development and evaluation of multi-marker risk scores for clinical prognosis. *Stat Methods Med Res*. 2012;1–17.
6. Lennon AM, Wolfgang C. Cystic neoplasms of the pancreas. *J Gastrointest Surg*. 2013;17(4):645–653.
7. Howlander N, Noone A, Krapcho M, *et al*. SEER Cancer Statistics Review, 1975-2011. *Secondary SEER Cancer Stats Rev. 1975-2011*, 2014.
8. Galanis C, Zamani A, Cameron JL, *et al*. Resected serous cystic neoplasms of the pancreas: a review of 158 patients with recommendations for treatment. *J Gastrointest Surg*. 2007;11(7):820–826.
9. Jaïs B, Rebours V, Malleo G, *et al*. Serous cystic neoplasm of the pancreas: a multinational study of 2622 patients under the auspices of the International Association of Pancreatology and European Pancreatic Club (European Study Group on Cystic Tumors of the Pancreas). *Gut*. 2015;1–7.
10. Tanaka M, Fernandez-Del Castillo C, Adsay V, *et al*. International consensus guidelines 2012 for the management of IPMN and MCN of the pancreas. *Pancreatology.* 2012;12(3):183–197.
11. Law JK, Ahmed A, Singh VK, *et al*. A systematic review of solid-pseudopapillary neoplasms: are these rare lesions? *Pancreas*. 2014;43(3):331–337.
12. Lennon AM, Manos LL, Hruban RH, *et al*. Role of a multidisciplinary clinic in the management of patients with pancreatic cysts: a single-center cohort study. *Ann Surg Oncol*. 2014;21(11):3668–3674.
13. Maguchi H, Tanno S, Mizuno N, *et al*. Natural history of branch duct intraductal papillary mucinous neoplasms of the pancreas: a multicenter study in Japan. *Pancreas*. 2011;40(3):364–370.
14. Al-Haddad M, Dewitt J, Sherman S, *et al*. Performance characteristics of molecular (DNA) analysis for the diagnosis of mucinous pancreatic cysts. *Gastrointest Endosc*. 2014;79(1):79–87.
15. Masica DL, Karchin R. Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. *Cancer Res*. 2011;71(13):4550–4561.
16. Masica DL, Karchin R. Collections of simultaneously altered genes as biomarkers of cancer cell drug response. *Cancer Res*. 2013;73(6):1699–1708.
17. Li L, Masica D, Ishida M, *et al*. Human bile contains MicroRNA-laden extracellular vesicles that can be used for cholangiocarcinoma diagnosis. *Hepatology*. 2014.
18. Bosman FT, World Health Organization, International Agency for Research on Cancer. *WHO Classification of Tumours of the Digestive*

System. 4th ed. Lyon: International Agency for Research on Cancer; 2010.

19. Springer S, Wang Y, Molin MD, Masica DL, Jiao Y, Kinde I, Blackford A, *et al*. A combination of molecular markers and clinical features improve the classification of pancreatic cysts. *Gastroenterology*. 2015;149(6):1501–1510.
20. Suzuki Y, Atomi Y, Sugiyama M, *et al*. Cystic neoplasm of the pancreas: a Japanese multi-institutional study of intraductal papillary mucinous tumor and mucinous cystic tumor. *Pancreas*. 2004;28(3):241–246.
21. Rätsch G, Onoda T, Müller K-R. Soft Margins for AdaBoost. *Machine Learning*. 2001;42(3):287–320.
22. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J Royal Stat Soc*. 2005;67(2):301–320.
23. Breiman L. Random forests. *Machine Learning*. 2001;45(1):5–32.
24. Wu J, Matthaei H, Maitra A, *et al*. Recurrent GNAS mutations define an unexpected pathway for pancreatic cyst development. *Sci Transl Med*. 2011;3(92):92ra66.
25. Wu J, Jiao Y, Dal Molin M, *et al*. Whole-exome sequencing of neoplastic cysts of the pancreas reveals recurrent mutations in components of ubiquitin-dependent pathways. *Proc Natl Acad Sci USA*. 2011;108(52): 21188–21193.