# Using Probe Electrospray Ionization Mass Spectrometry and Machine Learning for Detecting Pancreatic Cancer with High Performance

Wen Y Chung[1†], Elon Correa[2†], Kentaro Yoshimura[3], Ming-Chu Chang[4], Ashley Dennison[1], Sen Takeda[3#], Yu-Ting Chang[4#*]

[1]Department of Hepatobiliary and Pancreatic Surgery, Leicester General Hospital, UK.

[2]School of Computing, Science and Engineering University of Salford, UK.

[3]Department of Anatomy and Cell Biology, Faculty of Medicine, University of Yamanashi, Japan.

[4]Division of Gastroenterology and Hepatology, Department of Internal Medicine, National Taiwan University Hospital, College of Medicine, National Taiwan University, Taipei, Taiwan.

[†] Authors share co-first authorship

[#]Authors share co-senior authorship

*Corresponding author:

Yu-Ting Chang, M.D., M.S., Ph.D.

Department of Internal Medicine, National Taiwan University Hospital, College of Medicine, National Taiwan University;

No. 7 Chung Shan South Road, Taipei 100, Taiwan

Tel.: 886-2-23123456 ext. 63561

E-mail: yutingchang@ntu.edu.tw

**One Sentence Summary:** Advances in pancreatic cancer diagnosis can be facilitated by the parallel development of mass spectrometry technique and machine learning which when conflated produce a significant improvement in diagnostic accuracy up to 92%.

## ABSTRACT

A rapid blood-based diagnostic modality to detect pancreatic cancer with high accuracy is an unmet medical need. The study aimed to validate a unique diagnosis system using Probe Electrospray Ionization Mass Spectrometry (PESI-MS) and Machine Learning to the diagnosis of pancreatic cancer. For a case-control study, peripheral blood samples were collected from a total of 322 consecutive patients with pancreatic ductal adenocarcinoma (PDAC) and 265 controls with a family history of PDAC. All the control subjects were followed up for at least 2 years and confirmed free of pancreatic malignancy. Five µl of serum samples from controls and PDAC patients were analyzed using PESI-MS system. The mass spectra from each specimen were then fed into machine learning algorithms to discriminate between control and cancer cases. A total of 587 serum samples were analyzed. The sensitivity of the machine learning algorithm using PESI-MS profiles only to identify PDAC cases is 90.8 % with specificity of 91.7% (95% CI 83.9%-97.4% and 82.8%-97.7% respectively). Combined PESI-MS profiles with age and CA19-9 as predictors, the accuracy for earlier stage of PDAC (stage 1 or 2) is 92.9% and for advanced stage (stage 3 or 4) is 93% (95% CI 86.3-98.2; 87.9-97.4 respectively). PESI-MS profiles combined with machine learning is an approach that achieved very high accuracy in screening human serum samples for PDAC. The accuracy, low-cost, and simplicity of the technique provides an opportunity to detect PDAC at an early stage and must be applicable to the examination of at-risk populations.

## INTRODUCTION

The outlook for pancreatic ductal adenocarcinoma (PDAC) remains dismal due to lack of diagnostics effective for early detection, which leads to the diagnosis at advanced stage. As the majority of patients presented with advanced disease the 5-year survival rates remain below 10%(*1*). To improve the prognosis of PDAC will require earlier detection, ideally utilising diagnostic technique that could be employed in at risk populations such as those with diabetes (*2*). Currently serum carbohydrate antigen19-9 (CA19-9) is the most widely used diagnostic test but, due to false-positivity in patients with cirrhosis, chronic pancreatitis, cholangitis, other malignancies and false-negativity in Lewis blood-type negative patients, cannot be recommended for general screening(*3*). CA19-9 is reported to discriminate between PDAC patients and healthy controls or benign pancreatic disease with a sensitivity of 78.2-80.3% and a specificity of 80.2-82.8%(*4, 5*). To improve survival and reduce healthcare expenditure, it is estimated that a new diagnostic method will require a minimum sensitivity of 88% at a specificity of 85%(*6*).

Advances in cancer diagnosis can be facilitated by the parallel development of different techniques which when conflated produce a significant improvement in diagnostic accuracy. Both mass spectrometry and artificial intelligence are now established techniques in the medical armamentarium and machine learning, an application of artificial intelligence (AI) that endows systems with the ability to automatically learn and improve from experience, has overcome some of the prejudices inherent within medicine and is proving capable of solving some of the difficulties

associated with large and complex datasets. Mass spectrometry is a powerful technique for the rapid molecular diagnosis of cancer and probe electrospray ionization (PESI) analysis uses a very fine low invasive needle to achieve direct mass spectrometry. The probe needle directly collects a small amount of tissue without pretreatment(*7*) and detects sets of low-molecular-weight metabolites and lipids in specimens which provide important information for disease detection. For biological samples it is one of the commonly used and successful ambient ionization techniques(*8, 9*). PESI-MS has been shown to discriminate renal cell carcinoma, hepatocellular carcinoma and head and neck squamous cell carcinoma from surrounding normal tissue(*10*). Those results support the premise that PESI-MS is a versatile and promising technique for the identification of a number of different malignancies and when combined with AI as a potential screening approach in high risk populations. Recent studies from Germany, the USA and France demonstrated that AI algorithms perform better than dermatologists at detecting skin cancer(*11*). We recently demonstrated the potential of PESI-MS combined with partial least squares-logistic regression as a diagnostic system in animal studies and squamous cell carcinoma(*10*). The major attraction of PESI-MS is that it is a simple, rapid and inexpensive technique which can be easily automated and requires minimal sample preparation(*12*). The present study aims to demonstrate that PESI-MS combined with AI can diagnose PDCA using human serum samples, facilitating early diagnosis in the clinical setting, potentially improving patient outcomes and reducing healthcare costs.

## RESULTS

### HIGH-PERFORMANCE METHOD FOR DIAGNOSING PDAC

The data analysis based on this cohort shows that PESI-MS coupled with SVM classifier identified over 90% of PDAC cases as well as over 90% of the non-cancer cases with a 91.7% of sensitivity (95% CI 91.6-91.8) and a 90.4% of specificity (95% CI 90.2-90.6) respectively based only on the mass spectra from PESI-MS (Table 1 "spectra only as predictors", negative ion mode). The performance of this diagnostic approach exceeds the expectation required for novel methods for PDAC screening and represents a valuable tool for the screening of PDAC.

In order to improve the accuracy, we have added several parameters as predictors for diagnosis (Table 1). In this mode, we considered the PESI-MS spectra and the factors "age" and "CA19-9" level as predictors. Inclusion of these two parameters drastically improved the sensitivity to 95.1% (95% C.I: 95.0-95.2) at the cost of a decreased specificity to 89% (95% C.I: 88.8-89.2) (Table 1).

### DISCRIMINATING BETWEEN THE HEALTHY CONTROLS AND DIFFERENT STAGES OF PDAC

All the results presented in this section are based only on PESI-MS peaks as predictors (age and CA19-9 excluded). We tested if our system is also able to discern the changes in metabolites according to the cancer stage. A partial least squares-discriminant analysis (PLS-DA) algorithm (Figure 1A) was employed here to discriminate between (a) "Control", (b) "Earlier PDAC stages" (stages 1 + 2) and (c)

“Advanced PDAC stages” (stages 3 + 4) based on PES-MS spectra only as predictors. The dashed lines drawn around the cluster groups represent the 95% confidence interval for the respective cluster. PLS-DA was used because it also computes scores vectors that can be used to display a scatter plot of the general clustering of samples. The model was built using the combination of both negative and positive ion modes to compensate each other for achieving more accurate discrimination. Based on the biochemical information captured by PESI-MS, the subjects spontaneously grouped into control, early and advanced stage PDAC. At a glance, the control (black triangles) and cancer samples (blue and red circles) seem to form two well defined and clearly distinct clusters. The results suggest that PESI-MS is able to detect subtle alterations in spectra due to progression of PDAC. Comparing the separation of earlier (blue circles) with advanced (red circles) cancer stages, the scatter plot also shows two distinct clusters but with a slightly higher degree of overlap. This implies that different levels of metabolic changes between earlier and advanced cancer stages have also been detected by PESI-MS.

SVM was also applied to discriminate between healthy control patients and individual stages of PDAC (Fig. 1B). The results show that, on average, 91.4% of the PDAC stage 1 patients were identified and that the higher the PDAC stage the higher the sensitivity of the model; over 92% to 97.6% for stages 2, 3 and 4. Table S2 in supplementary information shows results of SVM models discriminating between control and PDAC stages.

## COMBINING INFORMATION FROM BOTH MASS SPECTROMETRY ION MODES

As MS ion mode is likely to capture slightly different molecular signatures from samples and the data analysis of each dataset presented slightly different classification results, a combination of both datasets should be beneficial for an automated cancer diagnostic process. Therefore, PESI-MS data from both ion modes were combined and a new specialized ML algorithm, based on SVM, was individually trained and tested to diagnose PDAC. Table 2A presents the average classification results obtained for the automated cancer diagnosis over 1,000 independent bootstrap cross-validations. The results shown in Table 2A suggest that the combination of both ion mode datasets is beneficial and the information gained improves the performance of the models. On this combined dataset, PESI-MS coupled with SVM detected with cancer cases 90.8% of sensitivity and 91.7% of specificity (95% C.I.83.9-97.4 and 82.8-97.7) respectively.

## HEALTHY CONTROLS VS CANCER STAGES ON THE COMBINED MASS SPECTROMETRY ION MODES

Surgical resection is the only potentially curative treatment for PDAC(*13*). Based on the TNM cancer staging system developed by the American Joint Committee on Cancer, pancreatic cancers detected in stages 1 and sometimes 2 are potentially operable. The results in Table 2B show that PESI-MS coupled with machine learning distinguished between healthy controls and subjects with earlier stage of PDAC

(stages 1 or 2) with sensitivity of 81.2% and specificity of 96.8% (95% CI:57.6-95.4 and 92.5-100) respectively. Although our method does not formally stage the disease, it is clearly able to detect differences between early and advanced tumors. The results for the discrimination between healthy controls and subjects with advanced stage PDAC (stages 3 or 4) are also reported in Table 3B.

## DISCUSSION

In this study we have successfully developed a novel screening system for pancreatic cancer by simultaneously attaining high sensitivity and specificity that exceeds the criteria required for PDAC diagnosis and screening. This was realized by combination of unique ionization method for mass spectrometry (PESI-MS) and machine learning that utilizes all the spectral data obtained by PESI-MS. Since our system does not extract the conspicuous spectral peaks from the datasets, judgment of cancer is based on the broad collections of spectra even if there are dozens of peaks relatively higher contributing to the diagnosis. Taking the diagnostic accuracy up to 91.2%, this method must open a new avenue in the diagnosis of PDAC ever attained by other methods. Another important point that has to be paraphrased here is the inexplicability of exact pathophysiological mechanism for drawing a diagnosis. Since this method relies on the collections of spectral peaks telling us the diagnosis possibly due to changes in metabolism, we cannot definitely narrow down the molecules responsible for explaining the molecular mechanisms underlying the cancer. While another important direction is to identify the molecules for diagnosis the PDAC, the focus of this study is validating the performance of new diagnostic technique. As machine learning does not necessarily require us to annotate the molecules underlying biologic pathways or modifiable risk factors that are associated with PDAC(*14*). We can employ simple and cost-effective mass spectrometer whose resolution is not high enough for identifying the molecules.

Data-driven machine learning produces consistent results if significant “learning”

of biochemical changes is achieved by a large cohort of PDAC and healthy controls (N=587 in this study). PDAC has been shown to take up to a decade following the initial mutation that gives opportunity for earlier diagnosis(*15*). Furthermore, 2-3 year window is open to detect PDAC's at stage 0 and 1. Although we do not register the patients at stage 0 or carcinoma in situ in our cohort, this technique should be applicable at these stages, considering this method can achieve an accuracy of 92.2% for the detection of early stages of PDAC including stage 1. Therefore, our system is promising in PDAC diagnosis that will revolutionize the routine of pancreatic diagnosis.

Metabolomics, including lipidomics, is a feasible way to identify metabolites responsible for PDAC detection(*16*). Mayers and co-workers had shown that a metabolic biomarker signature with 9 plasma metabolites plus CA19-9 differentiated PDAC from chronic pancreatitis with diagnostic accuracy of 90.6% (95% CI 84.9% to 94.6%)(*17*). In this study although PESI-MS principally measures metabolites and lipids we also identified the top 30 discriminating factors, in negative and positive modes in the supplementary appendix (figures S2 and S3). Those peaks include lipid profiles, phospholipids, sphingolipids and cholesterol sulfates that have also been identified in other studies(*18*). Our system does not deviate from molecular prediction method while it focuses on the fingerprint of responsible molecules.

The serum samples analyzed by PESI-MS do not require any pretreatments such as desalting, fractionation or enrichment. In addition, PESI-MS does not require large amount of sample for diagnosis, even sub-picoliters of samples are sufficient for

analysis. In addition to these superior attributes, the advantages of our system lie in low invasiveness, robustness and rapidity of measurements, and comprehensive resolution of substances in the samples.

There are limitations associated with the study mainly related to its retrospective design. It must be acknowledged that given a current accuracy of 92.9% there is still potential for further improvement of machine learning in PDAC diagnosis. Because our study population was a unique cohort from Taiwan, in other populations specific optimization will be required. The new diagnostic technique needs validation in several independent prospective world-wide cohorts.

Improving the outcome of PDAC will require a method which enables early and accurate diagnosis. Although the incidence of pancreatic cancer is too low to justify whole population screening, the accuracy of 92.9% for earlier PDAC stage combined PESI-MS and machine learning potentially justifies the evaluation of high-risk groups. Inherited pancreatic cancer syndromes and familial pancreatic cancer are the logical targets but other asymptomatic groups could be considered(*19, 20*). The most obvious of these is new-onset diabetes and a study from the Mayo Clinic demonstrated that diabetes has a 40% prevalence in PDAC and is frequently new-onset(*21*).

In conclusion, AI and machine learning have begun to enter the field of cancer diagnostics. Our study clearly demonstrates the feasibility of developing a diagnostic test with a comprehensive metabolite profiling MS platform plus machine learning that can detect PDAC with greater accuracy than has previously been achieved with either conventional tumor markers or a metabolic signature. There is a need for prospective,

real-world clinical evaluation of the diagnostic approach rather than only retrospective assessment of performance. The next stage is a large-scale diagnostic accuracy study among the at-risk populations where the test is intended to be employed.

## MATERIAS AND METHODS

### STUDY POPULATION

Between January 2005 and December 2017 after obtaining written informed consent and before treatment at the National Taiwan University Hospital (NTUH), peripheral blood samples were collected from a total of 322 patients (age 63.6 ± 13.0, female = 45.3%, male = 54.7%) with cytological and/or pathological confirmation of PDAC. All the patients' demographic data, including age, gender, serological studies, image studies, survival data, and clinical presentation were collected. Peripheral blood was collected from 265 high risk individuals (HRI) age 46.8 ± 14.8 (female 59.6%, male 40.4%) with a family history of PDAC, participating in a pancreatic cancer screening program at the NTUH between January 2005 and December 2015(*22*). Data from control subjects included a detailed family history, a full physical examination, blood sampling and magnetic resonance imaging (MRI). All of the control subjects were followed up for at least 2 years and confirmed free of pancreatic malignancy. Serum collected from PDAC patients and controls was stored at -80˚C until PESI-MS analysis. The study was reviewed and approved by the Institutional Review Board of NTUH (No. 201301048RIND) and University of Yamanashi (No. 645). Table 3 shows demographic data of the study subjects.

### PESI-MS ANALYSIS

A PESI-MS system (installation of PESI on a single quadrupole mass spectrometer compartment of LCMS-2020; Shimadzu, Kyoto, Japan) was used in the

study. Five µl of serums were added to 95 µl of 50% ethanol in a 1.5-ml tube and vortexed for 2 min. Samples were centrifuged at 150,000 × g for 1 minute, and the resultant supernatant collected. Sample analysis by PESI-MS was performed as described previously(*23*). All analyses were performed in both positive and negative ion modes for each specimen. The mass spectra from each specimen was generated using LabSolutions software (Ver. 5.82 SP1; Shimadzu). PESI is a discontinuous ionization method and differs from other ESI methods. The measurement was performed for 2 minutes with a time window of 10 sec that demonstrates a stable continuous ionization at maximal intensity (ca.50, 000). This window was spread out into 10 consecutive sets of spectral data. The *m/z* of our analysis ranges from 10 to 2,000, which is divided into 1990 bins, each of which corresponds to a unitary mass by taking account of the mass spectral accuracy of Shimadzu PESI-MS 2020. Each bin contains information on m/z as well as the peak intensity and they are used to construct the database.

**DATA PROCESSING AND MACHINE LEARNING**

To demonstrate the suitability of PESI-MS to detect significant biochemical differences between PDAC and control subjects, a specialized machine learning (ML) algorithm was trained and tested on each PESI-MS ion mode dataset. The positive ion mode dataset contained 583 subjects (318 PDAC and 265 controls) and the negative ion mode 587 subjects (322 PDAC and 265 controls). The first objective of the algorithm was to accurately distinguish between cancer and control samples. The candidate models were trained and tested using 1,000 independent repetitions of a

bootstrap cross-validation process (the average over the 1,000 independent models is reported). Bootstrap is a re-sampling technique that can be applied as cross-validation to estimate the performance of a model. The method randomly splits the data into training and test sets (see appendix for more information). An independent and automated ML classification model is then trained on the training partition of the data and tested on the test partition. This whole process including the random splitting of the data plus model training and validation, is independently repeated 1,000 times. The performance of each model is recorded and the average classification results over the 1,000 repetitions is calculated and reported in the results tables.

The ML algorithm used to build the classification model was a support vector machines (SVM)(*24*) and all algorithms have been implemented in R Statistical Package(*25*). During the model selection phase of the data analysis, other ML algorithms, such as random forest(*26*), partial least squares-discriminant analysis (PLS-DA)(*27*) and convolutional neural networks (CNN)(*28*) have been tested but SVM produced the best overall performance on the datasets available. In particular, the CNN models seemed to be overkill for binary classification (see results in supplementary appendix). SVM are supervised learning algorithms used for regression analysis and classification. By using different kernels, SVM can easily perform non-linear classification. SVM results reported here are all based on a radial basis kernel as it produced the best performance results on the present dataset. As the number of subject samples is large enough and the data do not present too many missing values, the different partitions of the data used for analysis have always been

selected in such a way that there are no subjects with missing values (listwise deletion)(*29*).

For each ion mode data (positive and negative) two different types of models were extensively trained and tested. The first model type used only the PESI-MS peaks as predictors (independent variables) to discriminate cancer and control cases. The objective is to show that the information from PESI-MS analysis of peripheral blood samples alone can discriminate between healthy controls and all 4 stages of PDAC at least as well as the tumor marker CA19-9. The second used the PESI-MS peaks (as model type 1) plus age and CA19-9 values as predictors. The sensitivity, specificity, and accuracy of each model to discriminate PDAC from controls were calculated. Figure 1 depicts the workflow of sample collection, sample analysis and data analysis processes used in this study.

**SAMPLE EXCLUSION CRITERIA**

Serum samples from all 587 subjects were analyzed via PESI-MS in both negative (n = 587 samples) and positive (n = 583 samples) ion modes – 4 samples (spectra) from the positive ion mode were considered inconsistent and excluded from the analysis. When adding CA19-9 as predictors for diagnosis, samples with missing CA19-9 values were excluded. The numbers then become 515 samples in negative ion mode (i.e., 587 excluding 20 PDAC & 52 control samples) and 512 samples in positive ion mode (i.e., 583 excluding 19 PDAC & 52 control samples). To form the combined positive and negative ion mode data, we only considered samples with CA19-9 values

recorded in both ion modes (n = 512).

**REFERENCES**


1. *Cancer survival in England: adult, stage at diagnosis and childhood – patients followed up to 2016* (2017 https://www.ons.gov.uk/releases/cancersurvivalinengland).
2. M. Ilic, I. Ilic, Epidemiology of pancreatic cancer. *World J Gastroenterol* **22**, 9694-9705 (2016).
3. M. J. Duffy, C. Sturgeon, R. Lamerz, C. Haglund, V. L. Holubec, R. Klapdor, A. Nicolini, O. Topolcan, V. Heinemann, Tumor markers in pancreatic cancer: a European Group on Tumor Markers (EGTM) status report. *Annals of oncology : official journal of the European Society for Medical Oncology* **21**, 441-447 (2010).
4. K. E. Poruk, D. Z. Gay, K. Brown, J. D. Mulvihill, K. M. Boucher, C. L. Scaife, M. A. Firpo, S. J. Mulvihill, The clinical utility of CA 19-9 in pancreatic adenocarcinoma: diagnostic and prognostic updates. *Current molecular medicine* **13**, 340-351 (2013).
5. J. C. Gui, W. L. Yan, X. D. Liu, CA19-9 and CA242 as tumor markers for the diagnosis of pancreatic cancer: a meta-analysis. *Clinical and experimental medicine* **14**, 225-233 (2014).
6. O. Ghatnekar, R. Andersson, M. Svensson, U. Persson, U. Ringdahl, P. Zeilon, C. A. Borrebaeck, Modelling the benefits of early diagnosis of pancreatic cancer using a biomarker signature. *International journal of cancer* **133**, 2392-2397 (2013).
7. K. Yoshimura, L. C. Chen, D. Asakawa, K. Hiraoka, S. Takeda, Physical properties of the probe electrospray ionization (PESI) needle applied to the biological samples. *J Mass Spectrom* **44**, 978-985 (2009).
8. K. Yoshimura, M. K. Mandal, M. Hara, H. Fujii, L. C. Chen, K. Tanabe, K. Hiraoka, S. Takeda, Real-time diagnosis of chemically induced hepatocellular carcinoma using a novel mass spectrometry-based technique. *Anal Biochem* **441**, 32-37 (2013).
9. L. S. Eberlin, R. J. Tibshirani, J. Zhang, T. A. Longacre, G. J. Berry, D. B. Bingham, J. A. Norton, R. N. Zare, G. A. Poultsides, Molecular assessment of surgical-resection margins of gastric cancer by mass-spectrometric imaging.

*Proc Natl Acad Sci U S A* **111**, 2436-2441 (2014).

10. K. Ashizawa, K. Yoshimura, H. Johno, T. Inoue, R. Katoh, S. Funayama, K. Sakamoto, S. Takeda, K. Masuyama, T. Matsuoka, H. Ishii, Construction of mass spectra database and diagnosis algorithm for head and neck squamous cell carcinoma. *Oral Oncol* **75**, 111-119 (2017).
11. H. HA, F. C, S. R, T. F, B. T, B. A, K. A, H. AB, T. L, E. A, U. L, Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology* **29**, 1836-1842 (2018).
12. C. Ho, C. Lam, M. Chan, R. Cheung, L. Law, L. Lit, K. Ng, M. Suen, H. Tai, Electrospray Ionisation Mass Spectrometry: Principles and Clinical Applications. *Clinical Biochemist Reviews* **24**, 3-12 (2003).
13. D. P. Ryan, T. S. Hong, N. Bardeesy, Pancreatic Adenocarcinoma. *New England Journal of Medicine* **371**, 1039-1049 (2014).
14. A. Rajkomar, J. Dean, I. Kohane, Machine Learning in Medicine. *N Engl J Med* **380**, 1347-1358 (2019).
15. S. Yachida, S. Jones, I. Bozic, T. Antal, R. Leary, B. Fu, M. Kamiyama, R. H. Hruban, J. R. Eshleman, M. A. Nowak, V. E. Velculescu, K. W. Kinzler, B. Vogelstein, C. A. Iacobuzio-Donahue, Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467**, 1114-1117 (2010).
16. T. Kobayashi, S. Nishiumi, A. Ikeda, T. Yoshie, A. Sakai, A. Matsubara, Y. Izumi, H. Tsumura, M. Tsuda, H. Nishisaki, N. Hayashi, S. Kawano, Y. Fujiwara, H. Minami, T. Takenawa, T. Azuma, M. Yoshida, A novel serum metabolomics-based diagnostic approach to pancreatic cancer. *Cancer Epidemiol Biomarkers Prev* **22**, 571-579 (2013).
17. J. Mayerle, H. Kalthoff, R. Reszka, B. Kamlage, E. Peter, B. Schniewind, S. Gonzalez Maldonado, C. Pilarsky, C. D. Heidecke, P. Schatz, M. Distler, J. A. Scheiber, U. M. Mahajan, F. U. Weiss, R. Grutzmann, M. M. Lerch, Metabolic biomarker signature to differentiate pancreatic ductal adenocarcinoma from chronic pancreatitis. *Gut* **67**, 128-137 (2018).
18. S. Riedl, B. Rinner, M. Asslaber, H. Schaider, S. Walzer, A. Novak, K. Lohner, D. Zweytick, In search of a novel target - phosphatidylserine exposed by non-apoptotic tumor cells and metastases of malignancies with poor

treatment efficacy. *Biochimica et biophysica acta* **1808**, 2638-2645 (2011).

19. H. T. Lynch, J. F. Lynch, S. J. Lanspa, Familial pancreatic cancer. *Cancers* **2**, 1861-1883 (2010).
20. A. Stoita, I. D. Penman, D. B. Williams, Review of screening for pancreatic cancer in high risk individuals. *World J Gastroenterol* **17**, 2365-2371 (2011).
21. S. T. Chari, C. L. Leibson, K. G. Rabe, L. J. Timmons, J. Ransom, M. de Andrade, G. M. Petersen, Pancreatic cancer-associated diabetes mellitus: prevalence and temporal association with diagnosis of cancer. *Gastroenterology* **134**, 95-101 (2008).
22. M. C. Chang, C. H. Wu, S. H. Yang, P. C. Liang, B. B. Chen, I. S. Jan, Y. T. Chang, Y. M. Jeng, Pancreatic cancer screening in different risk individuals with family history of pancreatic cancer-a prospective cohort study in Taiwan. *Am J Cancer Res* **7**, 357-369 (2017).
23. K. Yoshimura, L. C. Chen, H. Johno, M. Nakajima, K. Hiraoka, S. Takeda, Development of Non-proximate Probe Electrospray Ionization for Real-Time Analysis of Living Animal. *Mass spectrometry* **3**, S0048 (2014).
24. C. Cortes, V. Vapnik, Support-vector networks. *Machine Learning* **20**, 273-297 (1995).
25. R. D. C. Team. (R Foundation for Statistical Computing, Vienna, Austria).
26. L. Breiman, in *Machine Learning*. (Springer, 2001), vol. 45, pp. 5-32.
27. M. Barker, W. Rayens, Partial least squares for discrimination. *Journal of Chemometrics* **17**, 166-173 (2003).
28. H. H. Aghdam, E. J. Heravi, *Guide to Convolutional Neural Networks*. (Springer International Publishing, 2017), pp. 282.
29. P. D. Allison, *Missing Data*. Quantitative Applications in the Social Sciences (Sage Publications, 2001), pp. 104.
30. B. Efron, Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika* **68**, 589-599 (1981).

## Acknowledgments

We would like to thank Ayumi Iizuka, Setsuko Fukuta and Neil Loftus for the help of PESI-MS analysis and Taiwan Pancreas Foundation's support for the international collaborative project.

## Contributors

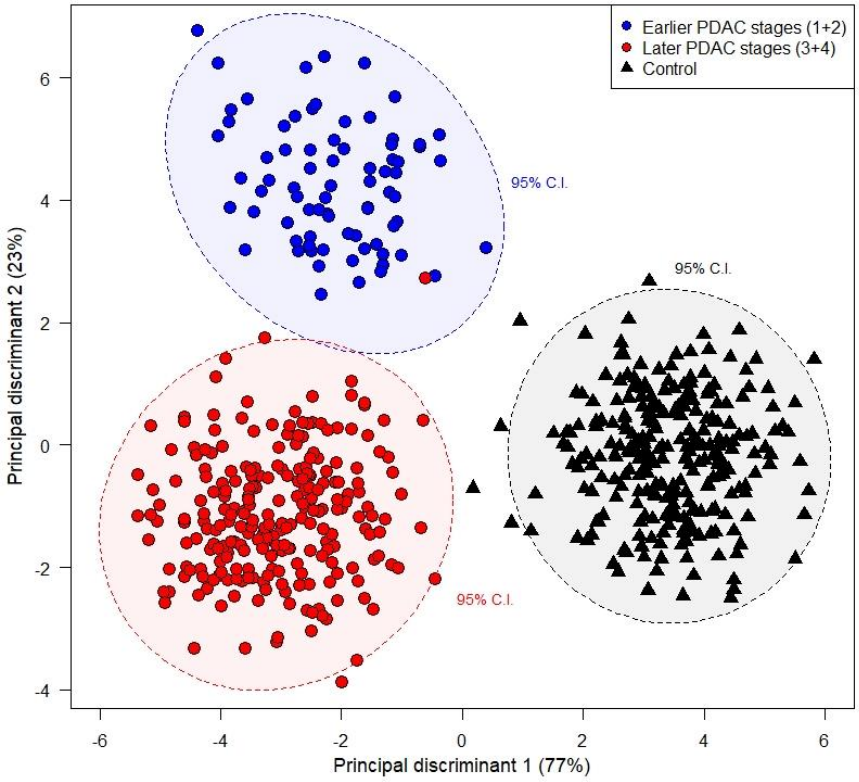
WYC contributed to the overall design, execution of the experiment and data analyses. EC contributed to data analyses and interpretation, particularly the design and analysis of the statistical analyses, and drafting and writing of the manuscript. KY contributed to the execution of the experiment and data analyses. MCC contributed to data collection and data analysis. AD contributed to the overall design, interpretation of results, and drafting and writing of the manuscript. ST contributed to the overall design, supervised the execution of the experiment, provided advice on aspects of the statistical analyses, and commented on manuscript drafts. YTC contributed to the overall design, data collection, execution of the experiment, interpretation of results, drafting and writing of the manuscript and confirms that all authors have seen and approved the final text. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted

## Declaration of interests

All authors declare no competing interests.

## Figures

A)




B)

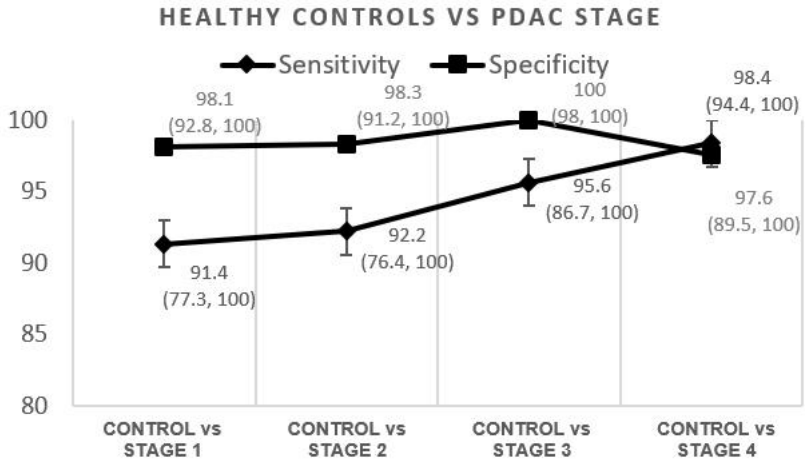



Figure 1: (1A) PLS-DA scores plot for the combined data (- and + ion modes). Separation between samples from healthy controls and stage I+II and stage II+IV of PDAC. The model was based on the PESI-MS spectra only as predictors. The numbers between parentheses represent % of variance from the data explained by each principal discriminant. The dashed lines drawn around the clusters represent

95% CI for the respective cluster. (1B) SVM discrimination results (%) for healthy controls vs individual stages of PDAC with 95% CI between parentheses.

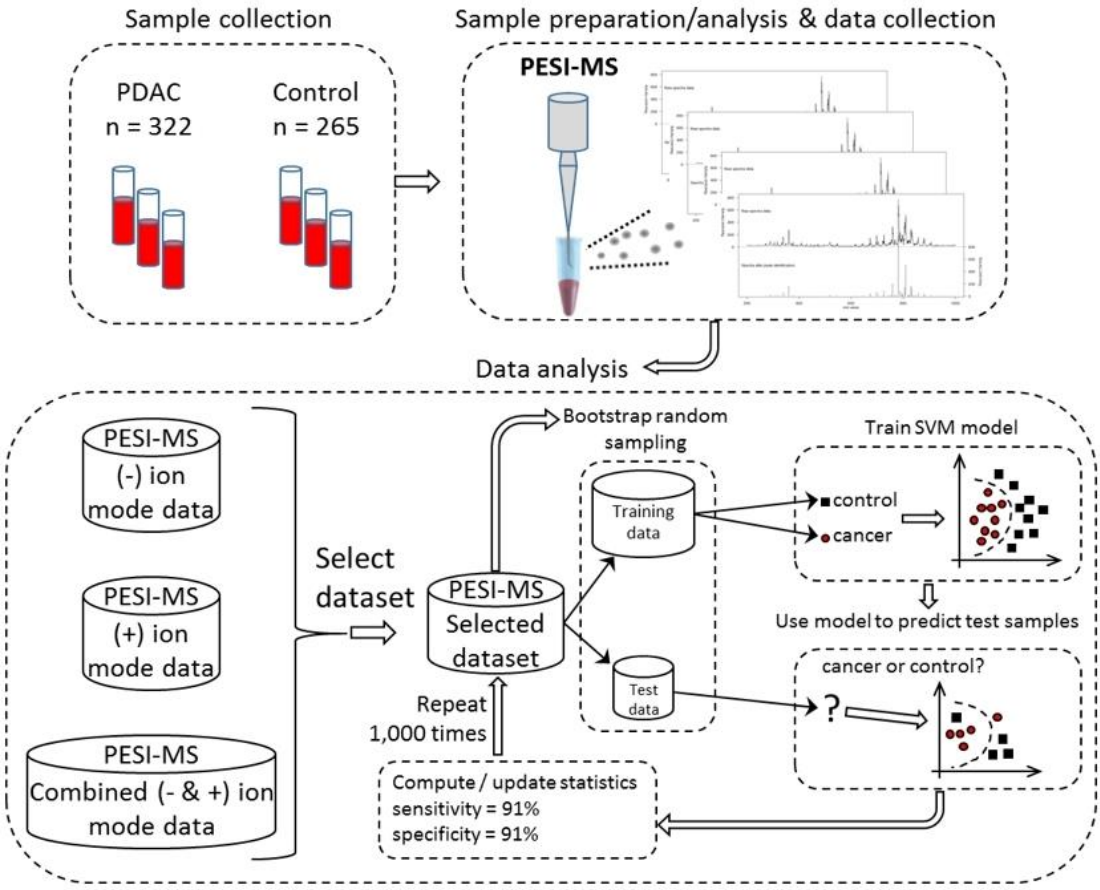



Figure 2: Workflow of the data collection, sample analysis and data analysis processes.

## Tables

**Table 1.** Control vs Cancer: average automated cancer diagnosis classification results obtained over 1,000 independent models built using bootstrap cross-validation.

| | | SPECTRA ONLY AS PREDICTORS | | SPECTRA + AGE + CA19-9 AS PREDICTORS | |
|---|---|---|---|---|---|
| | **Ion mode** | **Negative** | **Positive** | **Negative** | **Positive** |
| | Control | 265 | 265 | 213 | 213 |
| | Cancer | 322 | 318 | 302 | 299 |
| | Total | 587 | 583 | 515 | 512 |
| CLASSIFICATION | **Accuracy %** | 91.1 | 89.1 | 92.8 | 92.7 |
| | **(95% C.I.)** | (91, 91.2) | (89, 89.2) | (92.7, 92.9) | (92.6, 92.8) |
| | **Sensitivity %** | 91.7 | 91.2 | 95.1 | 95.9 |
| | **(95% C.I.)** | (91.6, 91.8) | (91, 91.4) | (95, 95.2) | (95.8, 96) |
| | **Specificity %** | 90.4 | 86.7 | 89 | 87.3 |
| | **(95% C.I.)** | (90.2, 90.6) | (86.4, 87) | (88.8, 89.2) | (87, 87.6) |

**Table 2. (2A)** Prediction Combining Decisions from Negative and Positive Ion Modes: average automated classification results obtained over 1,000 independent bootstrap cross-validations. **(2B)** Cancer Stage Prediction Combining Decisions from Negative and Positive Ion Modes: average automated classification results obtained over 1,000 independent bootstrap cross-validations.

| **(2A)** | | COMBINED SPECTRA (Negative + Positive ion modes) ONLY AS PREDICTORS |
|---|---|---|
| | Control | 265 |
| | Cancer | 318 |
| | Total | 583 |
| CLASSIFICATION | **Accuracy %** | 91.2 |
| | **(95% C.I.)** | (86.7, 95.8) |
| | **Sensitivity %** | 90.8 |
| | **(95% C.I.)** | (83.9, 97.4) |
| | **Specificity %** | 91.7 |
| | **(95% C.I.)** | (82.8, 97.7) |

| **(2B)** | | COMBINED SPECTRA (Negative + Positive ion modes) + AGE + CA19-9 AS PREDICTORS | |
|---|---|---|---|
| | | * Control vs **Earlier** PDAC stages | Control vs **Advanced** PDAC stages |
| | Control | 213 | 213 |
| | Cancer | 68 | 231 |
| | Total | 281 | 444 |
| CLASSIFICATION | **Accuracy %** | 92.9 | 93 |
| | **(95% C.I.)** | (86.3, 98.2) | (87.9, 97.4) |
| | **Sensitivity %** | 81.2 | 92.7 |
| | **(95% C.I.)** | (57.6, 95.4) | (84.2, 100) |
| | **Specificity %** | 96.8 | 93.4 |
| | **(95% C.I.)** | (92.5, 100) | (86.2, 100) |

* Cohen's kappa coefficient for the “Control vs Earlier PDAC stages” model is $k = 0.8$.

**Table 3**. The demographic data of 322 patiensts with pancreatic ductal adenocarcinoma (PDAC) and 265 high risk controls (HRCs)

| | PDAC (n=322) | HRC (n=265) |
|---|---|---|
| Age (years) | 63.675 (24.4-91.7) | 46.847 (17.7-88.2) |
| Gender | | |
| Female | 146 (45.3%) | 158 (59.6%) |
| Male | 176 (54.7%) | 107 (40.4%) |
| Body-mass index* | 22.04 (13.3-42.7) | 23.24 (15.6-37.7) |
| Pancreatic cancer location | | |
| Head (head/neck/uncinate)** | 150 (46.6%) | |
| Body | 111 (34.5%) | |
| Tail | 61 (18.9%) | |
| Stage | | |
| I | 11 (3.4%) | |
| II | 66 (20.5%) | |
| III | 60 (18.6%) | |
| IV | 185 (57.5%) | |
| CA-19-9*** | 4558.46 (0.5-66526.69) | 15.21 (0-269.4) |
| CEA**** | 63.309 (0.1-5783.59) | 1.734 (0-9.37) |
| OS (month) | 11.35 (0.07-94.22) | |

**Supplementary Materials**

**Bootstrap resampling**

Bootstrap is a re-sampling technique that can be applied as cross-validation to estimate the performance of a model. The method randomly splits the data into training and test sets. Bootstrap does this by randomly selecting, with replacement, *N* samples from a set containing exactly *N* samples. All selected samples, including the repetitions, are then used as training set and the non-selected samples (never seen by the model) are used as test set(*30*) effectively having all samples analyzed (*N* = total number of subjects “or PESI-MS spectra” available) in a bag. A single sample is then taken out of the bag randomly and its number noted. This sample now forms part of the training data, and the sample is placed back into the bag. This random sample picking process is repeated until *N* samples are in the training set. Some samples will be used multiple times, and on average, for each bootstrap partition 63.2% of all of the samples will be selected for training with the remaining 36.8% used as the test set.

**Pancreatic cancer data – convolutional neural network model (deep learning)**

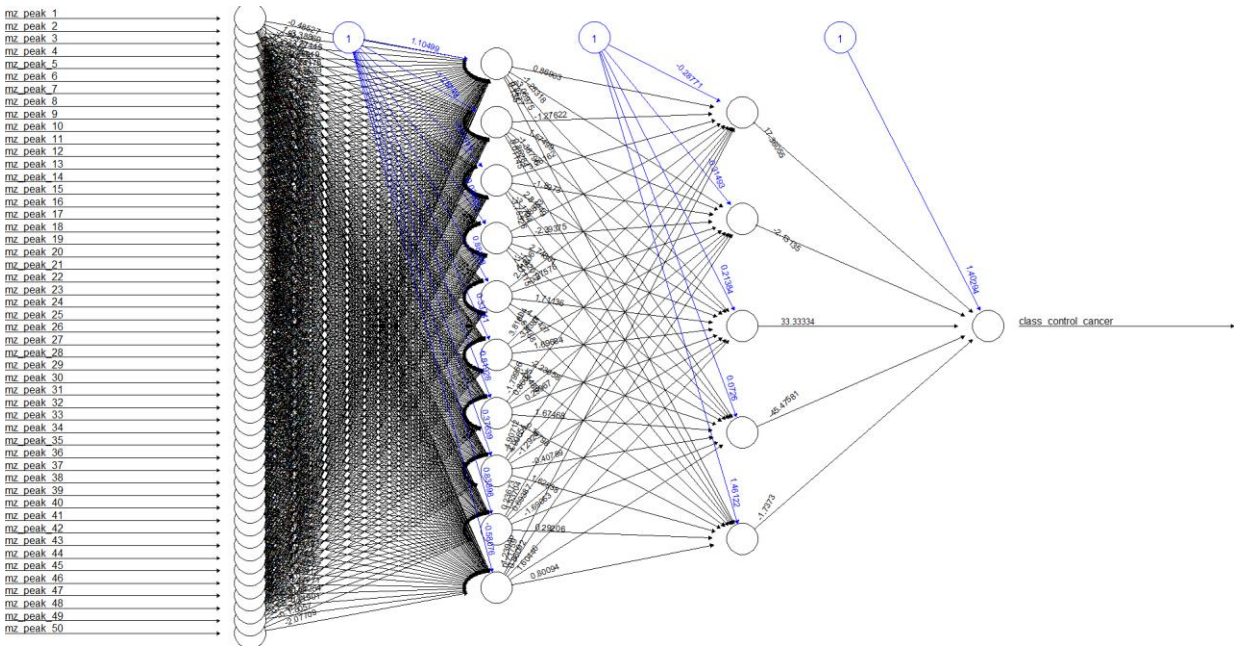



**Figure S1:** Partial representation of a sub convolutional neural network (the full one is too large to plot) with the best classification results obtained using CNN shown in Table A1.

**Table S1:** Results averaged over 1,000 independent modles from classification using convolutional neural network.

| control vs cancer (combined dataset positive + negative data) | |
|---|---|
| Accuracy | 81.6% |
| Sensitivity | 86.1% |
| Specificity | 73.3% |

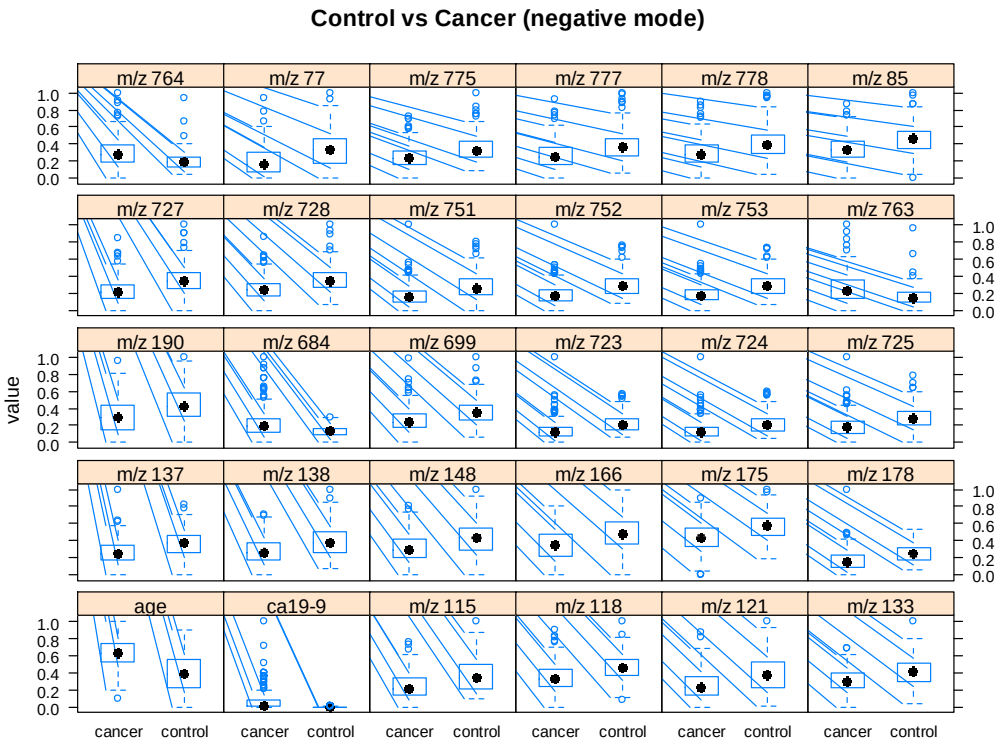



Figure S2: Top 30 discriminating factors (PESI-MS peaks plus age and CA19-9) for healthy controls vs PDAC using the negative ion mode data.

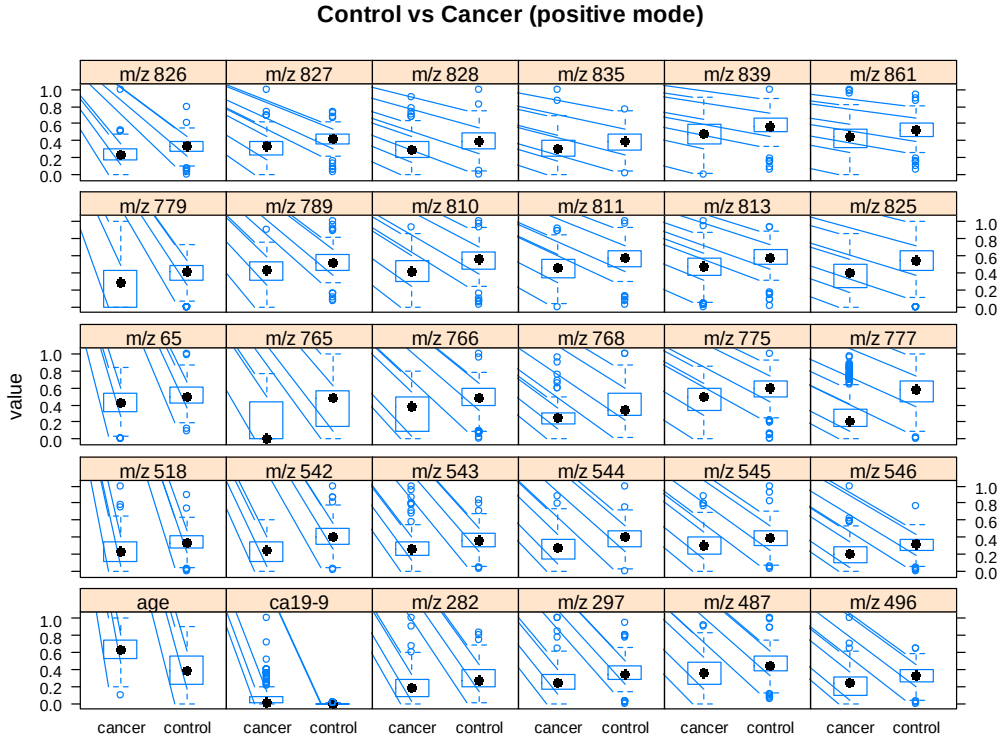



Figure S3: Top 30 discriminating factors (PESI-MS peaks plus age and CA19-9) for healthy controls vs PDAC using the positive ion mode data.

**Table S2:** Control vs Cancer Stages: average automated classification results obtained over 1,000 independent models built using bootstrap cross-validations.

| | | **SPECTRA ONLY AS PREDICTORS** | | | |
|---|---|---|---|---|---|
| | | Control vs **Earlier** PDAC stages | | Control vs **Advanced** PDAC stages | |
| | **Ion mode** | **Negative** | **Positive** | **Negative** | **Positive** |
| | Control | 265 | 265 | 265 | 265 |
| | Cancer | 77 | 76 | 245 | 242 |
| | Total | 342 | 341 | 510 | 507 |
| CLASSIFICATION | **Accuracy %** | 87.1 | 88.1 | 91.3 | 89.4 |
| | **(95% C.I.)** | (86.9, 87.3) | (87.9, 88.3) | (91.2, 91.4) | (89.3, 89.5) |
| | **Sensitivity %** | 85.9 | 90.2 | 89.1 | 86.2 |
| | **(95% C.I.)** | (85.5, 86.3) | (89.8, 90.6) | (88.9, 89.3) | (85.9, 86.5) |
| | **Specificity %** | 88.5 | 86.2 | 93.4 | 92.5 |
| | **(95% C.I.)** | (88.1, 88.9) | (85.8, 86.6) | (93.2, 93.6) | (92.2, 92.8) |
| | | **SPECTRA + AGE + CA19-9 AS PREDICTORS** | | | |
| | | Control vs **Earlier** PDAC stages | | Control vs **Advanced** PDAC stages | |
| | **Ion mode** | **Negative** | **Positive** | **Negative** | **Positive** |
| | Control | 213 | 213 | 213 | 213 |
| | Cancer | 68 | 68 | 234 | 231 |
| | Total | 281 | 281 | 447 | 444 |
| CLASSIFICATION | **Accuracy %** | 90.9 | 91.6 | 91.2 | 91.9 |
| | **(95% C.I.)** | (90.7, 91.1) | (91.4, 91.8) | (91.1, 91.3) | (91.8, 92.0) |
| | **Sensitivity %** | 88 | 89.3 | 90.4 | 93.6 |
| | **(95% C.I.)** | (87.6, 88.4) | (89.0, 89.6) | (90.2, 90.6) | (93.4, 93.8) |
| | **Specificity %** | 94 | 94.1 | 92.1 | 90.3 |
| | **(95% C.I.)** | (93.7, 94.3) | (93.8, 94.4) | (91.9, 92.3) | (90.1, 90.5) |