Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Array Platform Modeling and Analysis (B)

Detection of Pancreatic Cancer Biomarkers Using Mass Spectrometry

Kiyoun Kim[1,*], Soohyun Ahn[1,*], Johan Lim[1], Byong Chul Yoo[2], Jin-Hyeok Hwang[3] and Woncheol Jang[1]

[1]Department of Statistics, Seoul National University, Seoul, Republic of Korea. [2]Research Institute and Hospital, National Cancer Center, Goyang, Republic of Korea. [3]Department of Internal Medicine, Seoul National University Bundang Hospital, Seongman, Republic of Korea. *These authors contributed equally as first authors of this work.

ABSTRACT

BACKGROUND: Pancreatic cancer is the fourth leading cause of cancer-related deaths. Therefore, in order to improve survival rates, the development of biomarkers for early diagnosis is crucial. Recently, diabetes has been associated with an increased risk of pancreatic cancer. The aims of this study were to search for novel serum biomarkers that could be used for early diagnosis of pancreatic cancer and to identify whether diabetes was a risk factor for this disease.

METHODS: Blood samples were collected from 25 patients with diabetes (control) and 93 patients with pancreatic cancer (including 53 patients with diabetes), and analyzed using matrix-assisted laser desorption/ionization-time of flight mass spectrometry (MALDI-TOF/MS). We performed preprocessing, and various classification methods with imputation were used to replace the missing values. To validate the selection of biomarkers identified in pancreatic cancer patients, we measured biomarker intensity in pancreatic cancer patients with diabetes following surgical resection and compared our results with those from control (diabetes-only) patients.

RESULTS: By using various classification methods, we identified the commonly splitting protein peaks as m/z 1,465, 1,206, and 1,020. In the follow-up study, in which we assessed biomarkers in pancreatic cancer patients with diabetes after surgical resection, we found that the intensities of m/z at 1,465, 1,206, and 1,020 became comparable with those of diabetes-only patients.

KEYWORDS: biomarker, classification, mass spectrometry, pancreatic cancer

SUPPLEMENT: Array Platform Modeling and Analysis (B)

CITATION: Kim et al. Detection of Pancreatic Cancer Biomarkers Using Mass Spectrometry. *Cancer Informatics* 2014:13(S7) 45–53 doi: 10.4137/CIN.S16341.

RECEIVED: September 9, 2014. **RESUBMITTED:** November 11, 2014. **ACCEPTED FOR PUBLICATION:** November 16, 2014.

ACADEMIC EDITOR: JT Efird, Editor in Chief

TYPE: Original Research

FUNDING: WJ's work was supported by grant no. 12-2013-010 from SNUBH Research fund and the Basic Science Research program through the National Research Foundation (NRF) of Korea grant by the Ministry of Education (No. 2013R1A1A2010065. JHH's work was supported by grant no. 06-2010-067, 12-2013-010 and 02-2014-034 from SNUBH Research Fund. JL's work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (The Ministry of Science, ICT and Future Planning, MSIP) (No. 2011-0030810). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: jhhwang@snubh.org, wcjang@snu.ac.kr

Paper subject to independent expert blind peer review by minimum of two reviewers. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

Introduction

Pancreatic cancer is a lethal malignancy representing the fourth leading cause of cancer-related deaths.[1] Despite many therapeutic and diagnostic trials, the five-year pancreatic cancer survival rate remains at around 5%, largely because of late diagnosis, resistance to chemotherapy or radiation therapy, and a high recurrence rate even after surgery.[1] Therefore, to improve the clinical outcome of this disease, it will be necessary to develop early detection methods, make precise selection of surgical candidates, and use optimal treatment strategies.

Using biomarkers is a potentially effective method for diagnosing pancreatic cancer early and thereby improving prognosis. However, biomarker screening for early detection is currently unavailable. Carbohydrate antigen 19-9 (CA 19-9) is the only pancreatic cancer biomarker and has been widely used for diagnosis, monitoring of therapeutic response, and prognosis; however, it has several limitations when used to detect small pancreatic cancer.[2] In previous studies, blood metabolites were detected as low-mass ions (LMIs) that reflect the pathological changes in a cancer; therefore, LMI

data obtained by mass spectrometry (MS) has potential as a screening tool for early detection of cancer.[3–5] Over the last decade, researchers have also attempted to use MS data to find biomarkers for pancreatic cancer.[6–9]

In addition to using novel biomarkers, efforts to improve clinical outcomes have involved screening for pancreatic cancer in high-risk populations. Accordingly, an epidemiological study identified patients with new-onset diabetes as having a higher risk of pancreatic cancer,[10] and this finding was supported by a meta-analysis that associated diabetes with an increased risk of the disease.[11] Although questions remain about whether diabetes is a risk factor or the result of pancreatic cancer, it is evident that diabetes, especially new-onset diabetes, is strongly associated with this cancer. In this study, we used matrix-assisted laser desorption/ionization-time of flight mass spectrometry (MALDI-TOF/MS) with the aim of finding new metabolites representing novel serum biomarkers that could differentiate between diabetes associated with pancreatic cancer and diabetes alone.

Recent developments in proteomics have provided novel methods for detecting biomarkers.[12,13] A comprehensive review of MS data analysis is provided by Hilario et al.[14] Briefly, a typical cancer-related experiment involving MS requires the completion of several preprocesses that are crucial to successful analysis of MS data. The final and the most important step is the classification of cancer status based on the preprocessed data. Ge and Wong[6] applied ensemble methods to proteomics data in an investigation of pancreatic cancer, and they recommended the use of ensemble methods over a single decision tree.

For technical and biological reasons, missing values are commonly observed in MS data. The simplest solution to this problem is the removal of all variables (m/z values in MS data) with missing values. This strategy may not be ideal when the number of missing values is high and a significant proportion of variables are affected. Very few previous papers address the issues of missing values in MS data[15]; however, a comprehensive overview of imputation methods for missing values can be found in the work of Karpievitch et al.[16] and its associated references. These authors state that missing values can be categorized into two groups: missing completely at random (MCAR) and abundance-dependent missing values. The latter are related to censoring because of the detection limit of the instrument. However, in practice, it is not easy to identify the types of missing values, and both types often exist in a data set. For MCAR values, a common method is to generate imputations from the normal distribution with a sample mean and variance for each row, but this method often suffers from variance underestimation because the sample variance calculation does not include missing values. Missing values because of censoring cause more complicated issues and require advanced imputation techniques. Karpievitch et al.[16] proposed a two-stage method in which they first identified the type of missing value and, depending on the type, used different imputation techniques.

Considering the approach of Karpievitch et al, which is heavily dependent on the classification of missing types, we addressed the question of imputing missing values without knowing their type. In MS, the majority of missing values are usually due to censoring, and so we used a lower value (in our case zero) as the mean of the normal distribution generating imputations. Furthermore, it is more realistic to assume unequal variance at m/z values. Therefore, to estimate the variance as a function of an m/z value, we fitted the LOcal regrESSion (LOESS) model (a nonparametric regression model) to the sample variances. With a relatively low mean value and precise variance estimates, our imputation was at least comparable to the imputations generated from the normal distribution with sample mean and variance. In addition to using our imputation technique, we handled missing values with dichotomization. We also proposed a clinical validation method for the biomarkers we found.

We organize the remainder of this article as follows. In Section 2, we describe the data set and statistical models, focusing on the imputation of missing values. In Section 3, we present the analysis of our results and address the validation of our findings. Finally, in Section 4, we discuss our conclusions.

Material and Methods

Patients and samples. Blood samples were collected from 53 patients with pancreatic cancer and diabetes, and 40 patients with pancreatic cancer without diabetes between September 2009 and December 2011 at Seoul National University Bundang Hospital. For resectable cases, blood samples were collected before surgery, 14 days after surgery, and then during follow-up meetings with intervals of 3, 6, and 12 months. Resected pancreatic cancer patients received gemcitabine- or 5-fluorouracil-based chemotherapy as an adjuvant treatment during follow-up. Unresectable pancreatic cancer patients took medication to relieve their symptoms. Blood samples from 25 diabetic patients without pancreatic cancer were also collected between December 2009 and July 2010 at the same hospital. Diabetic patients took various diabetic medications, including insulin. The 25 diabetic patients were used as the control group, while the 53 patients with both pancreatic cancer and diabetes were the case group. We used six replicates for each blood serum sample in these two groups, and so the total number of spectra used was 468. The patients with pancreatic cancer only were used for clinical validation by comparing them with the control patients after surgical resection. There was no gender difference between the control and case groups. However, patients with pancreatic cancer were significantly older than those without pancreatic cancer (70 vs. 57 years). Categorizing by the stage of disease, there were 34%, 34%, and 32% of patients in stages II, III, and IV, respectively. In patients with pancreatic cancer, preoperative jaundice was noted in 8 patients and 23 patients received curative intent surgery. This study was approved by the human

subjects committee of the Seoul National University Bundang Hospital, and patients gave their written, informed consent to participate in the research. The study followed the ethical guidelines of the 1975 Declaration of Helsinki.

Data preprocessing. All the serum samples were processed using an identical procedure as follows. Approximately 5–10 mL of blood was obtained from each patient. Blood was drawn into serum-separating tubes and centrifuged at 2,000 × *g* for 10 minutes at room temperature. The serum was removed and transferred to a capped polypropylene tube in aliquots, and the samples were stored at –70 °C. Sera (25 μL) were vortexed with 100 μL methanol/chloroform (2:1, v/v), and then incubated for 10 minutes at the room temperature. Subsequently, this mixture was centrifuged at 6,000 × *g* for 10 minutes at 4 °C. The supernatant was completely dried in a concentrator for 1 hour, and then resuspended in 30 μL of 50% acetonitrile/0.1% trifluoroacetic acid (TFA) and vortexed for 30 minutes. The methanol/chloroform extract was mixed (1:12, v/v) with an α-cyano-4-hydroxycinnamic acid solution in 50% acetonitrile/0.1% TFA, and 1 μL of the mixture was spotted on the MALDI target for analysis. Our mass spectra data represent the average of 20 accumulated spectra. To minimize the experimental error, variable factors, including focus mass, laser intensity, target plate, and data acquisition time, were tested. Ideal focus mass and laser intensity were fixed at 500 *m*/*z* and 5,000 *m*/*z*, respectively, and each sample was analyzed at least five times using these settings and different extractions and data acquisition times.

The raw MS spectra were measured between 9 and 2,500 Da, and had different lengths, with a mean average length of 165,977. Preprocessing plays a crucial role in analyzing MS data. Typical data preprocessing steps are as follows:

1. Transformation: Often the increments in *m*/*z* values are not constant but approximately proportional to the *x*-axis values. A logarithm transformation can be used to calibrate the scale of the *m*/*z* values.
2. Smoothing and baseline subtraction: Electronic or chemical noise produces background fluctuations. To remove these background noises, a nonparametric regression method can be used to estimate the background intensity values, which are then subtracted from the transformed data. After background correction, a smoothing procedure is required to smooth the effect of the isotopic envelope in the data.
3. Normalization: Systematic differences exist in the total amount of desorbed and ionized proteins; a normalization procedure is used to appropriately adjust these differences.
4. Peak detection: In each spectrum, the peaks that correspond to proteins must be identified.
5. Peak alignment: Owing to chemical and electronic noise, the same biological peak (representing a specific molecule) may not match in all spectra. Therefore, by aligning the peaks of the different spectra, it is possible to appropriately match peaks.

There are several tools available for data preprocessing, but we opted to use the R package MALDIquant.[17] Figure 1 shows the results of the preprocessing steps with a raw spectrum.

Classifications and missing values. After preprocessing, we applied four difference classification methods: Classification And Regression Tree (CART), bagging, random forest, and lasso.[18–21]

Using CART, a decision tree method, we first made successive splits on the *m*/*z* values to divide the feature space into separate regions. With each region, we used the most commonly occurring class as the prediction. While CART is easily understood, it can be unstable and produce an accuracy that is not comparable to other classification methods.

Bagging and random forest are ensemble methods that combine several algorithms to produce better prediction results than can be obtained by using each individual algorithm. Bagging draws samples repeatedly from the original training data with replacement and following a uniform probability distribution. Therefore, in these bootstrapped samples, it is possible to observe the same data more than once. We obtained predictions by using a decision tree on each bootstrapped sample, and then we aggregated these predictions. By using multiple samples, bagging can reduce the variance of predictions. Random forests also use bootstrapped sampled data sets, but only use a limited random selection of *m* predictors (*m*/*z* values in our study) to build decision trees. By using this method, it is possible to reduce the correlation between decision trees. In random forests, often $m = \sqrt{p}$, where p is the number of predictors.

In our study, another possible approach for making predictions was using logistic regression with *m*/*z* values as input and disease status as output. However, we could not use traditional logistic regression because the number of predictors was greater than the sample size. As an alternative, the lasso is a variable selection method of (logistic) regression. The lasso shrinks the estimated coefficient toward zero using the log-likelihood with a penalty function, the sum of the absolute values of the estimated coefficients. By using this penalty term, the lasso can produce a sparse model with selected nonzero predictors.

To evaluate each classifier, we randomly split the data into a training set (37 cases and 18 controls) and a test set (16 cases and 7 controls). We built each classifier based on the training set and estimated the prediction error with the test set. Among the training data, we found 6,860 missing values of intensity, ie, 21.7% of data were missing. Because the amount of missing was not negligible, we considered imputation for replacing missing values. A simple method of imputation would be to replace the missing values with a small predefined value, as used in GeneSpring MS software. However, using the same value for every imputation would

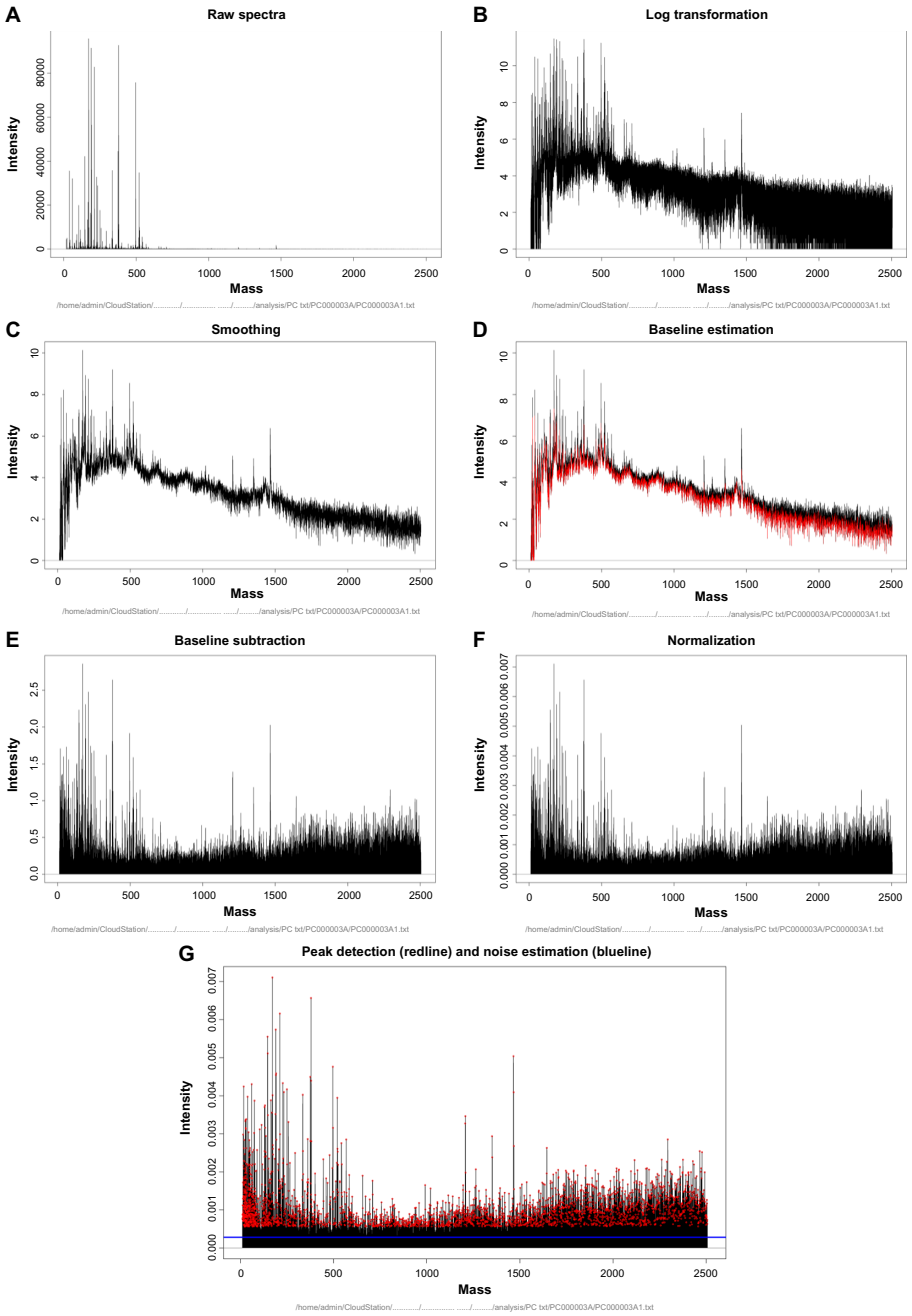

Figure 1. Preprocessing.

result in a reduction in variance for the peaks with missing values. Considering that one of the main goals of preprocessing is to stabilize variance across peaks (because stabilization of variance may affect classification if the classifier is sensitive to the scale of variances), using the same value for imputation may not be the optimal choice.

Therefore, we considered and applied the following methods for handling missing values. First, we converted the data into a binary structure using PPC (peak probability contrast).[13] The PPC is a procedure that constructs a vector of binary features for common peaks in preprocessing. Each feature is set to 1 if a peak has a height greater than given a threshold. The missing values are to set to zero. By using a binary structure, the issue of variance stabilization is eliminated.

We also applied imputation by adding random noise. We generated random noise from the normal distribution with a mean of zero and variance σ^2, and imputed the absolute values of this noise. Here σ^2 is the variance of intensity, and to estimate σ^2, we proposed two nonparametric methods, (i) LOESS and (ii) MAD (median absolute deviation) estimator, in wavelet regression. The LOESS is a nonparametric regression method that fits a simple linear model to localized subsets of the data in order to construct a nonlinear model. To estimate σ^2 with the LOESS, we first calculated sample variances for each peak and fitted the LOESS to these

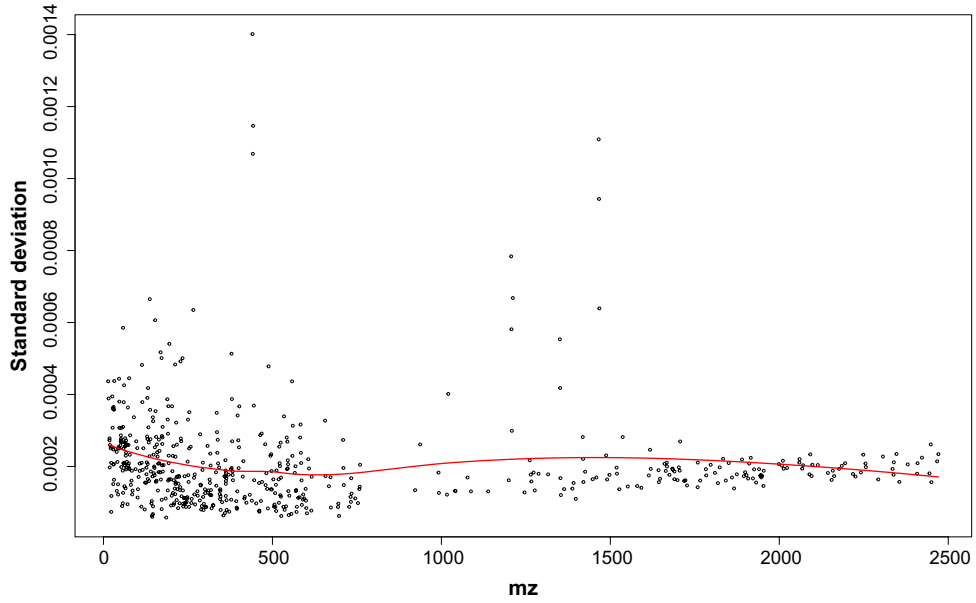

Figure 2. σ estimation using LOESS.

variances. Figure 2 gives the variance estimates obtained by using LOESS. For MAD, we fitted a wavelet regression model to the spectra and computed a robust standard deviation estimator.[22]

Clinical validation. To validate the selection of biomarkers, we applied two different approaches. First, we tested whether the intensity of the selected metabolites in pancreatic cancer patients with diabetes was reduced after surgical resection to the intensity level observed in patients with diabetes only. If the metabolites that we found were true biomarkers, we would expect that the intensity level between two groups would be similar after a certain period. Second, we investigated whether specific metabolites were related to diabetes by comparing post-surgery results obtained from patients with pancreatic cancer only with results obtained from patients with diabetes only. For both tests, we used two-sample *t*-tests and computed the distances between the comparable groups over time to identify the elapsed time, from surgery to recovery, at which the selected metabolites were statistically significant.

Results

Preprocessing. Having randomly split the 78 samples (53 cases and 25 controls) into training (55 samples: 37 cases and 18 controls) and test sets (23 samples: 16 cases and 7 controls) as described above, we used MALDIquant to preprocess each sample (six replicates per sample = 468 raw spectra). We log-transformed the intensities and conducted smoothing and baseline subtraction using SNIP (statistics-sensitive non-linear iterative peak-clipping).[23] For normalization, we used TIC (total ion current).[24] To identify peaks, we found the sites (m/z values) where intensity was more than twice the size of the noise estimate. Once peaks were identified, they were aligned by using the self-calibrated warping (SCW) algorithm.[25] To merge the spectra produced by the six replicates of each sample, we used a minimum intrasample percentage of presence of 50%. The number of peaks ranged from 1,894 to 2,184, while the mean average length of raw spectra was 165,977. To extract peaks from each group (case and control), we used a minimum intersample percentage of presence of 50%. As a result, we produced an intensity matrix with 574 peaks.

Classification. After preprocessing, we found that more than 20% of intensity values were missing. Thus, we used the three methods discussed above to handle missing values, and then we applied four classification methods. We repeated all procedures 100 times, beginning with the random splitting of test and training data sets, and computed prediction error rates each time. Figure 2 shows the LOESS fit of σ. The results are summarized in Table 1.

Bagging and random forest (ensemble methods) outperformed the lasso, and these methods were more successful when PPC was used to handle missing values. This is likely due to the binary structure of PPC as both bagging and random forest are based on binary decision trees. We found that using PPC also reduced standard errors. Comparing the two imputation methods, the LOESS method was slightly better than wavelet regression in terms of the prediction error and standard error for each classifier. The improved performance of LOESS could be due to the fact that it estimated σ^2 directly and provided a slightly different σ^2 for each peak. Boxplots of prediction errors for each classifier are given in Figure 3. As expected, CART was unstable and had the highest prediction error. By using median comparisons, we determined that the other three classification methods performed at similar levels.

In order to summarize the accuracy of each method, we present receiver operating characteristic (ROC) curves for each imputation method and for each of the four classifiers

Table 1. Prediction error for each method. Standard errors are given in parenthesis.

METHOD	PPC	LOESS	WAVELET
CART	0.1339 (0.0828)	0.1365 (0.0757)	0.1326 (0.0743)
Bagging	0.0735 (0.0520)	0.0839 (0.0593)	0.0939 (0.0673)
Random Forest	0.0757 (0.0537)	0.0839 (0.0616)	0.0917 (0.0702)
Lasso	0.1004 (0.0592)	0.0826 (0.0667)	0.0865 (0.0638)

(Fig. 4). These ROC curves are based on 100 randomly split test and train sets.

In addition to the ROC curves, we use the AUC (area under the curve) to investigate the performance of each method. Figure 5 shows boxplots of AUC obtained for each classifier with the different imputation methods. The results are also based on 100 randomly split test and train data sets.

Because bagging and random forest with PPC outperformed other methods when handling missing values (Table 1), we primarily focused on the classification results obtained using these methods. Accordingly, we found that the following were common candidates for biomarkers: *m/z* 175.0852, 287.0996, 402.1181, 1,020.5152, 1,206.5410, 1,207.5488, 1,208.5621, 1,465.6118, 1,466.6155, 1,467.5927, and 1,468.6177. To validate the significance of these candidates, we plotted the raw spectra for the given *m/z* values for pancreatic cancer patients with diabetes and patients with diabetes only. Differences in intensities were observed at *m/z* 1,465 1,468, 1,206 1,208, and 1,020. Figure 6 shows the difference of intensities at *m/z* 1,465 and 1,206.

To select biomarkers, we computed variable importance measures by calculating the mean decrease in prediction accuracy in bagging and random forest, and then selected the top 10 as candidates for biomarkers. For the lasso, the selected variables in the final model were considered as candidates. Based on the common candidates from each of the classifiers, we identified five *m/z* sites as potential candidates for biomarkers. Because some of these sites were too close together to be considered as separate sites and a protein can be distributed in the range of several *m/z* values, we appropriately merged these five sites into three *m/z* sites (namely 1,465, 1,206, and 1,020).

Among these three mass ions, 1,465 *m/z* was identified as a fibrinogen alpha chain by using MALDI-MS/MS analysis and a Swiss-Prot database search. However, we failed to identify the other two mass ions (1,206 and 1,020 *m/z*).

Clinical validation. Among patients with pancreatic cancer, 33 opted to have curative surgery. If the biomarkers we found are not solely related to diabetes, but are instead true biomarkers for pancreatic cancer, the prognosis after the surgery should be similar to that of the control group with diabetes only. Additionally, we compared the control group with patients only suffering from pancreatic cancer (no diabetes) after they had undergone surgery in order to identify metabolites that are solely related to pancreatic cancer. Thus, we collected 35 samples from 14 pancreatic cancer patients with diabetes and 30 samples from 19 pancreatic cancer-only patients for analysis. We compared the pancreatic cancer with diabetes and diabetes-only groups using a two-sample *t*-test. At each time *t*, we used the blood samples obtained after time *t* to compute test statistics. Results are presented in Table 2, where *n* represents the sample size of the pancreatic cancer patients with diabetes after time *t*.

At week 1, there were significant differences between the two groups. However, as time progressed, these differences

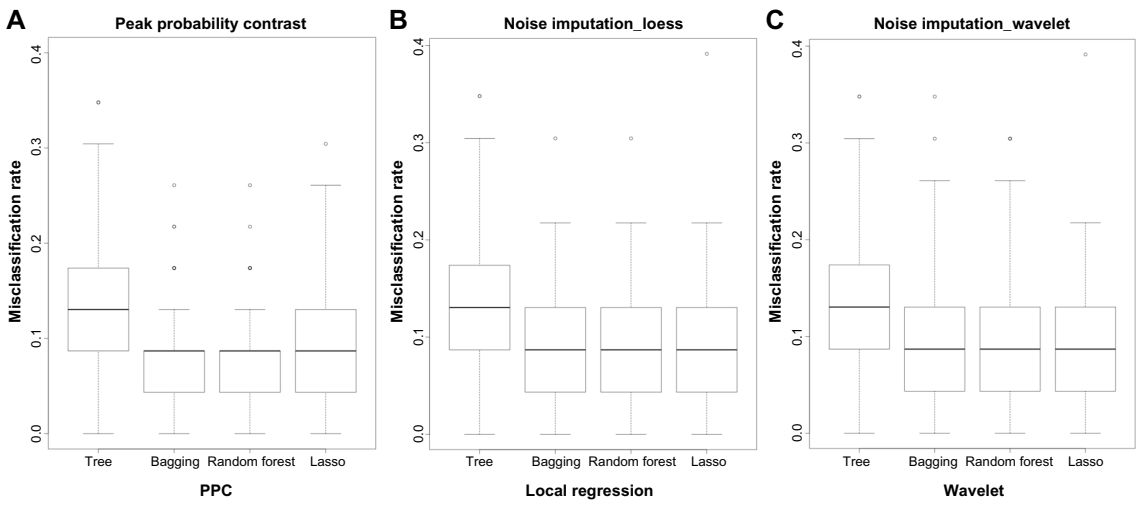

Figure 3. Boxplots of misclassification rates for each imputation methods with four classifiers.

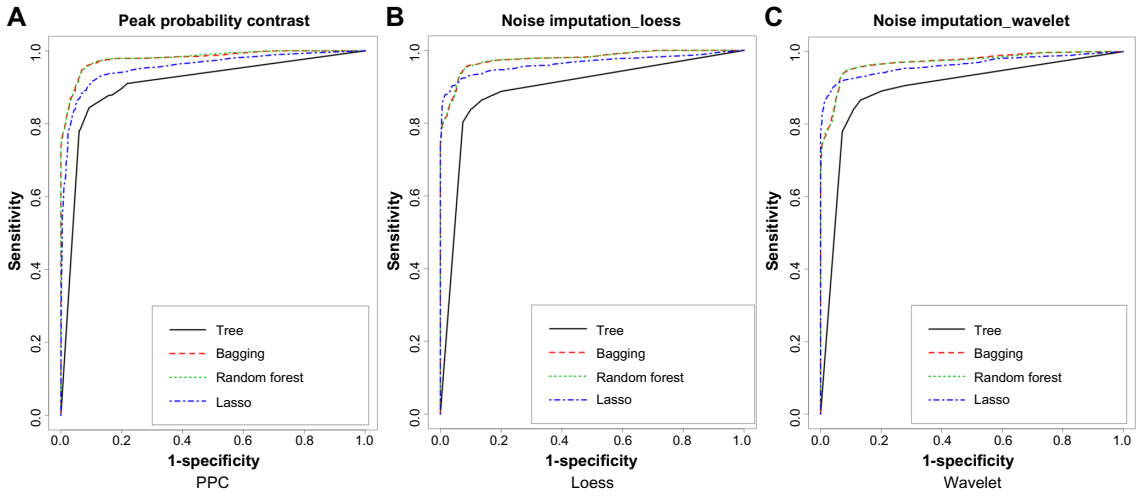

Figure 4. ROC curves for each imputation methods with four classifiers.

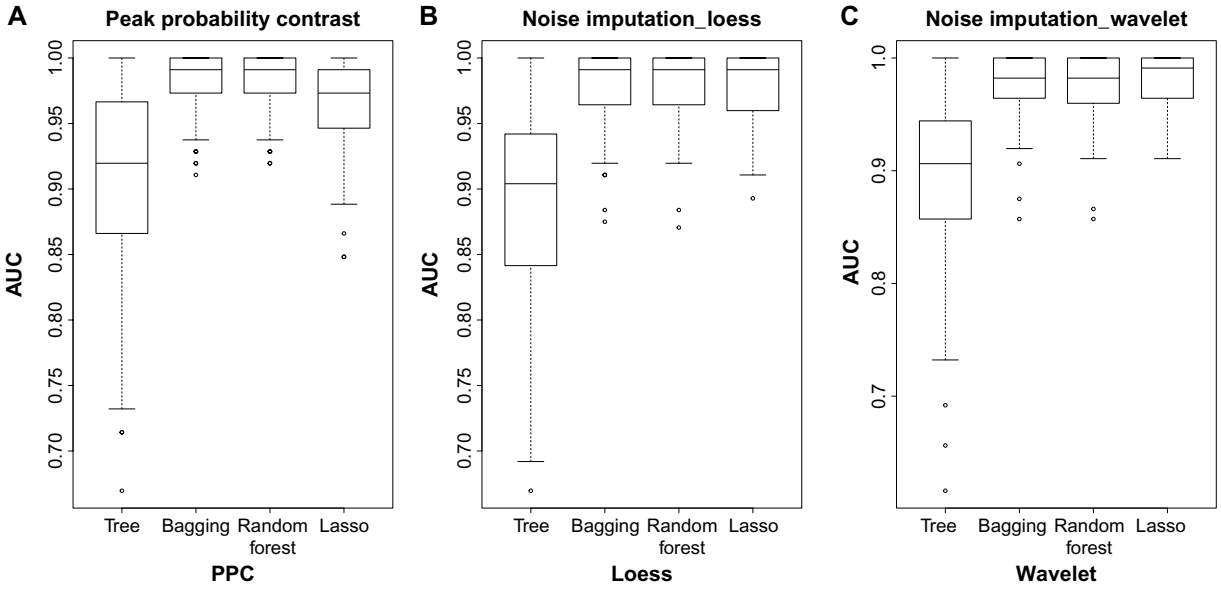

Figure 5. Boxplots for AUC.

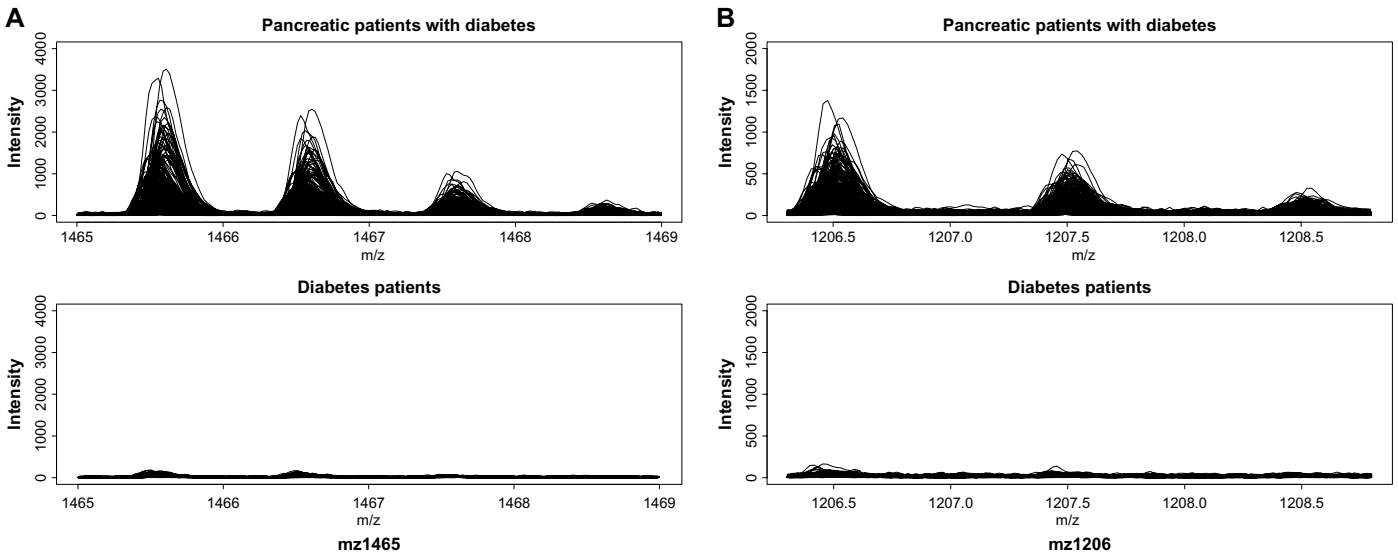

Figure 6. Intensities of biomarkers between control and case.

Table 2. Differences in the potential biomarkers in patients with pancreatic cancer and diabetes vs. those with diabetes only: *P*-value and sample size over time (week).

WEEK	m/z1465*	m/z1206*	m/z1020*	*n*
1	0.000	0.000	0.000	14
2	0.000	0.000	0.000	13
3	0.001	0.000	0.093	9
4–13	0.001	0.000	0.086	9
14–6	0.002	0.000	0.123	9
17	0.003	0.000	0.142	8
18	0.013	0.002	0.261	7
19–20	0.046	0.007	0.253	6
21	0.054	0.013	0.268	6
22–30	0.071	0.023	0.218	6
31	0.096	0.073	0.271	6
32 +	0.114	0.089	0.274	6

Table 3. Differences in the potential biomarkers in patients with pancreatic cancer without diabetes vs. those with diabetes only: *P*-value and sample size over time (week).

TIME	m/z1465*	m/z1206*	m/z1020*	*n*
1	0.000	0.000	0.005	18
2	0.000	0.000	0.010	16
3	0.000	0.000	0.466	10
4–5	0.000	0.000	0.831	10
6–10	0.000	0.001	0.762	9
11–17	0.000	0.002	0.576	8
18	0.001	0.003	0.537	8
19–30	0.002	0.010	0.469	7
31	0.008	0.030	0.413	6
32 +	0.017	0.043	0.723	5

apparently diminished. At *m/z* 1,465, the difference was no longer statistically significant after 21 weeks, while it took 31 and 3 weeks for the loss of statistical significance in *m/z* 1,206 and 1,020, respectively.

We also compared the pancreatic cancer-only patients with the diabetes-only controls. Table 3 summarizes the result.

Similar to results shown in Table 2, the difference between the two groups became statistically nonsignificant after three weeks at *m/z* 1,020. However, in contrast, differences at *m/z* 1,465 and 1,206 remained statistically significant throughout the examination period up to 32 weeks. This could be explained if the metabolites at *m/z* 1,465 and 1,206 were related to diabetes as well as pancreatic cancer.

We considered the dissimilarity between the two groups for each hypothesis given above, with the aim of validating our results. We used distances between the two groups (clusters) to assess their dissimilarity, and these were computed using three different linkage methods (single, complete, and average) in hierarchical cluster analyses. Figure 7 shows how the patterns of distances changed over time. For all three linkage methods, as time progressed, the distance between diabetes only and pancreatic cancer with diabetes was apparently smaller than the distance between diabetes only and pancreatic cancer without diabetes.

Discussion

In this paper, we analyzed MS data to find biomarkers for pancreatic cancer. From a methodological standpoint, we proposed three methods for handling missing values in MS data, and we compared their performance when they were combined with popular classification methods. We found that PPC worked best when we used tree-based classification methods. We also suggested a clinical validation method for biomarker identification. We identified three metabolites as possible biomarkers. Of these, *m/z* 1,020 apparently discriminated patients with pancreatic cancer and diabetes from those with diabetes only.

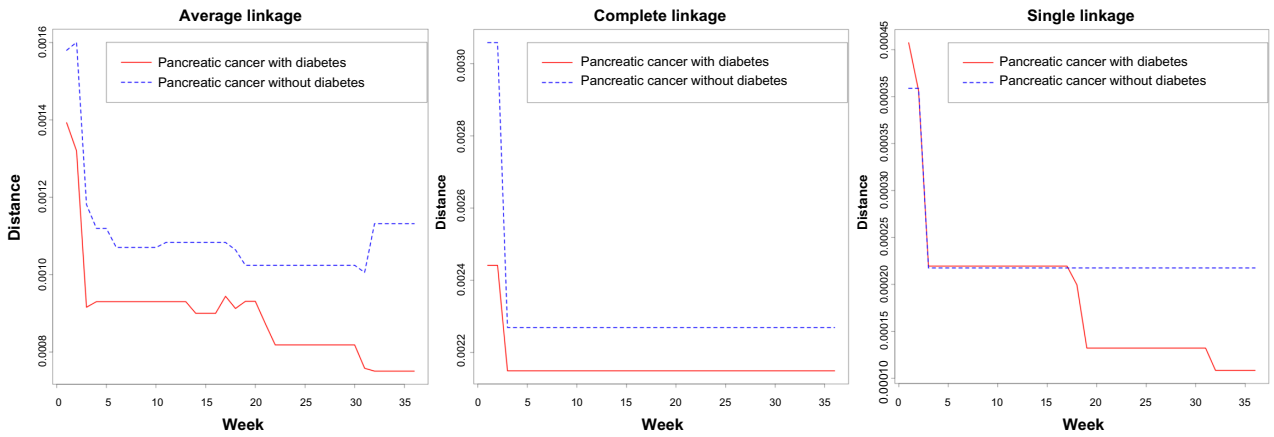

Figure 7. Distances between pancreatic cancer with diabetes and without diabetes groups with 3 different linkage methods.

Author Contributions

Conceived and designed the experiments: BCY, JHH. Analyzed the data: KK, SA, JL, WJ. Wrote the first draft of the manuscript: KK, SA, JHH, WJ. Contributed to the writing of the manuscript: KK, SA, JL, BCY, JHH, WJ. Jointly developed the structure and arguments for the paper: JL, JHH, WJ. Made critical revisions and approved final version: JL, JHH, WJ. All authors reviewed and approved the final manuscript.

REFERENCES

1. Siegel R, Ma J, Zou Z, Jemal A. Cancer statistics, 2014. *CA Cancer J Clin*. 2014;64:9–29.
2. Ruckert F, Pilarsky C, Grutzmann R. Serum tumor markers in pancreatic cancer–recent discoveries. *Cancers (Basel)*. 2010;2:1107–24.
3. Chen T, Xie G, Wang X, et al. Serum and urine metabolite profiling reveals potential biomarkers of human hepatocellular carcinoma. *Mol Cell Proteomics*. 2011;10:M110.004945.
4. Asiago VM, Alvarado LZ, Shanaiah N, et al. Early detection of recurrent breast cancer using metabolite profiling. *Cancer Res*. 2010;70:8309–18.
5. Yoo BC, Kong SY, Jang SG, et al. Identification of hypoxanthine as a urine marker for non-Hodgkin lymphoma by low-mass-ion profiling. *BMC Cancer*. 2010;10:55.
6. Ge G, Wong GW. Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles. *BMC Bioinformatics*. 2008;9:275.
7. Sugimoto M, Wong DT, Hirayama A, Soga T, Tomita M. Capillary electrophoresis mass spectrometry-based saliva metabolomics identified oral, breast and pancreatic cancer-specific profiles. *Metabolomics*. 2010;6:78–95.
8. Bathe OF, Shaykhutdinov R, Kopciuk K, et al. Feasibility of identifying pancreatic cancer based on serum metabolomics. *Cancer Epidemiol Biomarkers Prev*. 2011;20:140–7.
9. He S, Cooper HJ, Ward DG, Yao X, Heath JK. Analysis of premalignant pancreatic cancer mass spectrometry data for biomarker selection using a group research optimizer. *T I Meas Control*. 2012;34(6):674–88.
10. Chari ST, Leibson CL, Rabe KG, Ransom J, de Andrade M, Petersen GM. Probability of pancreatic cancer following diabetes: a population-based study. *Gastroenterology*. 2005;129:504–11.
11. Ben Q, Xu M, Ning X, et al. Diabetes mellitus and risk of pancreatic cancer: a meta-analysis of cohort studies. *Eur J Cancer*. 2011;47:1928–37.
12. Wu B, Abbott T, Fishman D, et al. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*. 2003;19(13):1636–43.
13. Tibshirani R, Hastie T, Narasimhan B, et al. Sample classification from protein mass spectrometry by peak probability contrasts. *Bioinformatics*. 2004;20(17):3034–44.
14. Hilario M, Kalousis A, Pellegrini C, Müller M. Processing and classification of protein mass spectra. *Mass Spectrom Rev*. 2006;25:409–49.
15. Hrydziuszko O, Viant MR. Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline. *Metabolomics*. 2012;8:S161–74.
16. Karpievitch YV, Dabney AR, Smith RD. Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics*. 2012;13(suppl 16):S5.
17. Gibb S, Strimmer K. MALDIquant: a versatile R package for the analysis of mass spectrometry data. *Bioinformatics*. 2012;28(17):2270–1.
18. Breiman L, Friedman JH, Olshen RA, et al. *Classification and Regression Trees*. UK: Chapman and Hall; 1983.
19. Breiman L. Bagging predictors. *Mach Learn*. 1996;24(2):123–40.
20. Breiman L. Random forest. *Mach Learn*. 2001;45(1):5–32.
21. Tibshirani R. Regression shrinkage and selection via the lasso. *J Royal Stat Soc B*. 1996;58(1):267–88.
22. Donoho DL, Johnstone IM. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*. 1994;81:425–55.
23. Ryan CG, Clayton E, Griffin WL, Sie SH, Cousens DR. SNIP, a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications. *Nucl Instrum Methods Phys Res B*. 1988;34(3):396–402.
24. Sauve Anne C, Speed Terence P. Normalization, Baseline Correction and Alignment of High-Throughput Mass Spectrometry Data. *Procedings Gensips*. 2004.
25. He QP, Wang J, Mobley JA, Richman J, Grizzle WE. Self-calibrated warping for mass spectra alignment. *Cancer Inform*. 2011;10:65–82.