

Detection of Urinary Biomarkers for Early Diagnosis of Pancreatic Cancer by Data Analysis

Chi Le^{1,†}, Yucheng Liu^{2,†}, Fangyi Tian^{3,†} and Yang Xu^{4,†}

¹ZJU-UoE Institute, Zhejiang University, Hangzhou, Zhejiang, China

²College of Animal Sciences & Technology, Huazhong Agricultural University, Wuhan, Hubei, China

³Basic Medical School, Capital Medical University, Beijing, China

⁴College of Life Sciences, Nanjing Agricultural University, Taizhou, Jiangsu, China

†These authors contributed equally


Keywords: Pancreatic Cancer, Diagnosis, Urinary Biomarkers.


Abstract: In this sample-structured document, neither the cross-linking of float elements and bibliography nor metadata/copyright information is available. The sample document is provided in “Draft” mode and to view it in the final layout format, applying the required template is essential with some standard steps.


According to data released by the American Cancer Society in 2019, the mortality rate caused by pancreatic cancer ranks fourth among malignant tumors. By 2030, the incidence of Pancreatic ductal adenocarcinoma (PDAC) will continue to increase and may become the second leading cause of death among all tumor diseases. If the tumor could be detected and resected at an early stage, the survival rate of PDAC patients will be greatly improved. However, symptoms rarely show until the cancer reaches its advanced stage and most of the available treatments are palliative. Therefore, most patients have reached the advanced stage of cancer when they are diagnosed and thus having poor prognoses. Therefore, we are interested in the early detection, prediction and diagnosis of pancreatic cancer, and we will discuss which factors are related to pancreatic cancer in the following parts.


We collected a total of 590 samples in which 7 attributes, age, CA 19–9 (Carbohydrate antigen199), creatinine, LYVE1 (Lymphatic Vessel Endothelial Hyaluronic Acid Receptor 1), REG1B (regenerating islet-derived 1 beta), TFF1 (Recombinant Trefoil Factor 1) and REG1A (Recombinant Human Regenerating Islet-Derived Protein 1-alpha) were selected as our independent variables. The dependent variable Y is diagnosis which indicates whether a participant has pancreatic cancer. Logistic regression and lasso regression were used to construct a model for the prediction of pancreatic cancer. All analyses above were performed using R software, version 4.1.1.

We finally found that the distributions of Blood plasma levels of CA 19–9 monoclonal antibody, creatine, LYVE1, REG1B, TFF1 and REG1A are all positive skewed and asymmetrical. In addition, people's illness is significantly related to age, creatine, LYVE1, REG1B, TFF1 and REG1A. However, the level of CA 19-9 monoclonal antibody in the human body is not so significantly correlated with the corresponding human disease. After selecting appropriate methods and analyzing a large amount of data, according to the regression results, etc., we can conclude that the incidence of PDAC disease is significantly related to age and gender. Based on this, in the follow-up research, it has provided the possibility for early prediction and disease prevention and control of PDAC based on age and gender, and also provided new ideas for the pharmaceutical, treatment and daily care of the disease.

^a <https://orcid.org/0000-0003-4414-3073>

^b <https://orcid.org/0000-0002-3095-9904>

^c <https://orcid.org/0000-0003-0536-6127>

^d <https://orcid.org/0000-0001-9482-5887>

1 INTRODUCTION

The pancreas is an organ located in the abdomen and it plays an important role in converting the food we eat into energy for body's activities. Pancreatic cancer is one of the most common digestive tract malignancies. It begins in the tissue of pancreas and is an extremely deadly type of cancer, which is ranked as fourth leading cause of cancer-related mortality in western countries (Zeng *et al.*, 2019). 90% of pancreatic malignancies are pancreatic ductal adenocarcinoma (PDAC) (He *et al.*, 2014). Symptoms rarely shows until the cancer reaches its advanced stage and most of the available treatments are palliative (Adamska, Domenichini and Falasca, 2017). With over 80% of cases diagnosed at advanced stages, PDAC patients have a median survival of 5-6 months, and the overall 5-year survival rate is less than 10% because patients are diagnosed too late (Arnold *et al.*, 2019). However, if we can detect and resect the tumor at an early stage of PDAC, the survival rate of cancer patients can be greatly improved.

Therefore, we are interested in the early detection, prediction, and diagnosis of PDAC, and trying to find out what factors are related to pancreatic cancer. Also, the development of non-invasive diagnosis to detect early PDAC becomes an urgent need. Non-invasive diagnostic technology can avoid or reduce pancreatic biopsy to identify PDAC fibrosis early, and can perform dynamic monitoring, which has important clinical application value. Nevertheless, there is no reliable, non-invasive screening test to detect PDAC accurately and those methods are expensive (Brezgyte *et al.*, 2021). By contrast, detecting biomarkers for the diagnosis of PDAC is minimally invasive and relatively cheap.

Serum CA19-9, the only biomarker in clinical practice currently, is less specific and sensitive for screening purposes and is mainly used to monitoring treatment response (Ballehaninna and Chamberlain, 2012). Previous studies have found that urine, an alternative biological fluid has many advantages, such as accumulation of biomarkers at higher concentrations so that the biomarkers are easy to be detected. Certain urinary metabolites can indicate malignancy of various organs, possibly reflecting the metabolic effects of cancer (Dinges, *et al.*, 2019). Creatinine is a product of muscle metabolism and is primarily cleared by the kidneys (Delanaye, Cavalier and Pottel, 2017). Therefore, biostatistical methods such as model analysis can be used to identify more effective and stable biomarkers and provide a reference basis for early-stage PDAC detection and diagnosis, clinical practice, related treatment and so on.

In the existing literature, a regression model PancRISK have been developed using three protein biomarkers to detect pancreatic cancer and classify PDAC patients. Researchers of this experiment creatively replaced REG1A with REG1B, showing the ability of our urinary panel to distinguish control individuals and patients with benign hepatobiliary diseases from early stage PDAC patients with specificity and sensitivity >85%. On the basis of these studies, our team explored the previous research to make some bold assumptions and continued to use PancRISK with other methods to analyze data basing on the original database to evaluate whether PDAC is related to age, gender and other factors.

2 METHODS

2.1 Data Source

The data was selected from the Kaggle platform. Kaggle is an open online platform, mainly for developers, data scientists and anyone in need to provide a platform for holding machine learning competitions, hosting databases, and writing and sharing code. The data comes from the data set in a paper published by Silvana Debernardi and colleagues in the journal PLOS Medicine on December 10, 2020. The paper and the complete data set are open access.

590 clinical specimens were obtained from different centers: Barts Pancreas Tissue Bank, University College London, University of Liverpool, Spanish National Cancer Research Center, Cambridge University Hospital, and University of Belgrade.

590 urine specimens were assayed and there were three groups: 183 individuals who had no pancreatic diseases in control group (group 1), 208 patients who had benign hepatobiliary diseases in benign group (group 2) and 199 PDAC patients before treatments (group 3). 50.7% specimens were obtained from female individuals.

Among these three types of samples, benign samples included 119 CP cases, 54 gallbladder diseases, 20 cystic lesions of the pancreas, and 15 cases with abdominal pain and gastrointestinal symptoms suggestive of pancreatic origin.

In group 3, PDAC patients had 6 stages: 102 I-II (IA, IB, IIA) and 97 III-IV (IIIB, III, IV)

Besides, there were 67 patients in urine specimens with common urological tract malignancies: 18 patients with prostate cancers (PC) (median age 69 years, range 52–83), 29 patients with renal cell carcinoma (RCC) (median age 67 years, range 20–

85), and 20 patients with bladder transitional cell cancer (TCC) (median age 65 years, range 44–81). The restriction is that the number of I–IIA PDAC samples is low ($n=27$). In addition, the study used samples collected from control individuals as replacements for the lack of specimens from individuals with hereditary predisposition to PDAC.

350 matched plasma specimens for samples (92 control, 108 benign, and 150 PDAC).

Our dataset Urinary biomarkers for pancreatic cancer was downloaded from Kaggle, which was uploaded in 2020.

2.2 Variable Measuring

The required urine and plasma samples were collected from multiple centers after the respective institutional review board approvals, and the potential impact of bacterial growth on urine biomarkers was tested with 20 mg/ml boric acid. Finally, the samples were maintained at a low temperature of -80°C . Commercially sourced ELISA kits were used for assaying the biomarkers: lymphatic vessel endothelial hyaluronan receptor 1 (LYVE1), trefoil factor 1 (TFF1), regenerating family member 1 beta (REG1B) and plasma CA19-9. Each measurement was run in duplicate, and further repeats were performed when there was a discrepancy. The FLUOstar Omega Microplate Reader was used to determine optical density. The Roche platform (Cobas 601E [ECLIA] technology) at The Doctors Laboratory in London was used to measure plasma CA19-9. Urine creatinine was determined at the Clinical Biochemistry Laboratory of the University of Westminster using an ILab Aries analyser from Instrumentation Laboratory.

All the research staff who performed the experiments did not know about the sample diagnosis.

All protein concentration data were natural-log-transformed and mean-centred.

2.3 Data Analysis

The values obtained from open access were analyzed by exploratory data analysis firstly. At the beginning, there were 14 original attributes and we finally selected 7 attributes age, CA 19–9, creatinine, LYVE1, REG1B, TFF1 and REG1A as our independent variables. The other 7 attributes are confounders. The dependent variable Y is diagnosis which indicates whether a participant has pancreatic cancer. To meet the requirements of logistic regression, all the category variables were coded by 0 and 1. After that, we chose to use random forest

algorithm to interpolate all the not available data in the dataset. To explore the relationship between independent variable and response, boxplots were chosen to demonstrate the independent variables visually (Fig 1.). It should be noted that all the independent variables except the creatinine are greater in the pancreatic cancer group significantly. The χ^2 test is used to test the correlation between sex and diagnosis. The t-test is employed for the purpose of testing the correlation between all independent variables and dependent variables. To further explore influence of interactions between the independent variables, P-value was calculated.

In logistic regression, the logit function of p is used to modeling the log odds of response variable as a suitable transformation. In our study, we found that some variables were highly correlated with others, in which case their interactions are possible to differently effect the response compared to single independent variable. According to the result of correlation analysis, we kept all the significantly-correlated interactions of two variables. Lasso analysis were performed to further select useful variables and interactions. Next, we performed best subset selection to identify the best model that contains a given number of independent variables. We finally selected the model based on best subset selection and the Akaike Information Criterion (AIC) value, and then evaluated the model with confusion matrix.

3 RESULTS

During the process of exploratory data analysis (EDA), we first chose age, plasma CA 19-9, creatinine, LYVE1, REG1A, REG1B and TFF1 as independent variables and diagnosis as our dependent variables. After viewing characteristics of all variables (Table 1.), we found that the distributions of most of them were skewed, which means that they were asymmetrical. Take plasma CA 19_9 as an example, the maximum of its blood plasma level is 31000, but more than 88% participants are smaller than 1,000. The boxplots also verified this point (Fig 1.). In the comparison of plasma CA 19_9, there are much more outliers in PDAC group than the PDAC-free group, which indicates it might be a significant measurement for PDAC. Also, in variables age, LYVE1, REG1A, REG1B and TFF1, the minimums, maximums, medians, first quartile, and third quartile of PDAC patients group seem to be greater.

Table 1: Characteristics of Variables (N=590).

Characteristics	N	Percent(%)	
Sex			
Male	301	51.0	
Female	289	49.0	
Diagnosis			
Pancreatic-cancer-free	391	66.3	
Pancreatic cancer	199	33.7	
Continuous variable	N	Percent(%)	Range
Age			
>=50	446	75.6	26-89
<50	144	24.4	
Plasma CA19_9			
>=1000	69	11.7	0-31000.0
<1000	521	88.3	
Creatine			
>=2	39	0.1	0-4.1
<2	551	99.9	
Lymphatic vessel endothelial hyaluronan receptor 1 (LYVE1)			
>=10	32	0.1	0-23.9
<10	558	99.9	
Regenerating family member 1 beta (REG1B)			
>=700	18	0.1	0-1403.9
<700	572	99.9	
Trefoil factor 1 (TFF1)			

>=1000	95	16.1	0-13344.3
<1000	495	83.9	
Regenerating family member 1 alpha (REG1A)			
>=1000	97	16.4	0-13200.0
<1000	493	83.6	

What's more, both male and female participants are nearly 50%. More than three quarters of the participants are older than 50 years old. For the diagnosis of pancreatic cancer, about one-third of the participants are patients.

In the regression part, we decided to select the model with 5 predictors: $\text{Logit}(\text{diagnosis}) = 0.65 + 0.0077\text{Age} + 0.057\text{LYVE1} + 0.00033\text{REG1B} + 0.00011\text{REG1A} - 0.000074(\text{creatinine} : \text{REG1A})$. The colon means the interactions between two variables. The confusion matrix showed the accuracy of prediction made by the model (Table 2). According to the confusion matrix, the accuracy is 0.85 which represents the proportion of correctly identified samples. The sensitivity is 0.73 which represents the proportion of actual positive samples identified correctly. The precision is 0.82, which represents the proportion of predict positive samples identified correctly. The specificity is 0.91, which represents the proportion of actual negative samples identified correctly. These four factors indicate that our model fits well with the practical condition.

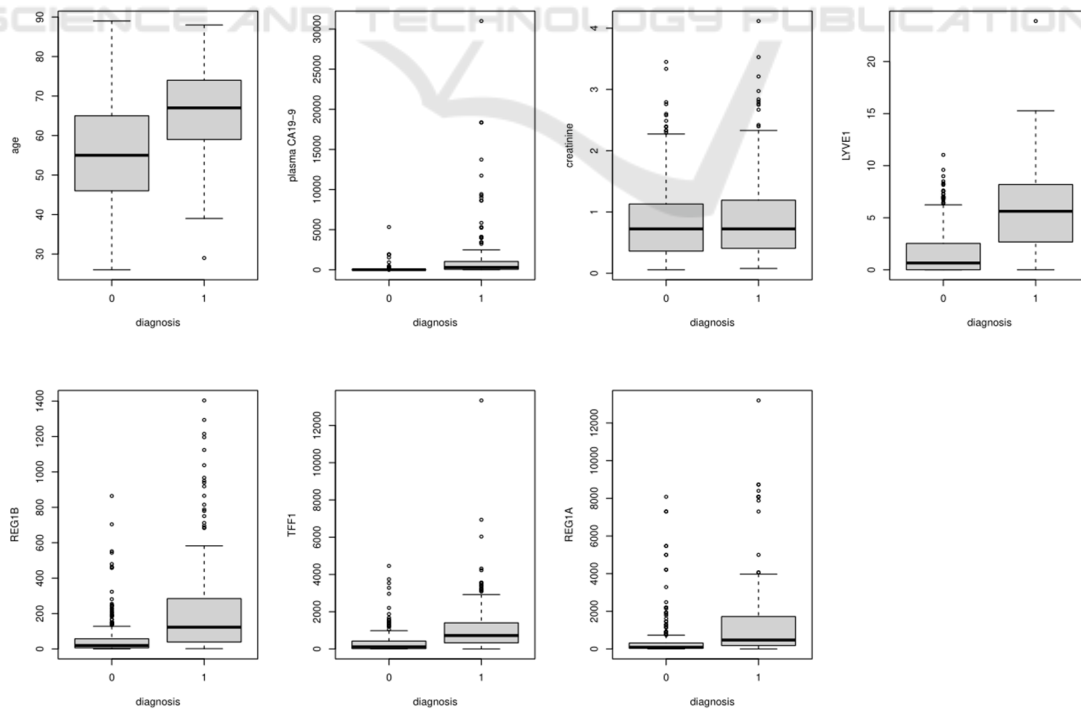


Figure 1: The Boxplots of Associations Between Independent Variables X and Dependent Variable Y Diagnosis.

The boxplots were used to demonstrate all the independent variables X visually and verify the positive skewed trend.

Table 2: Confusion matrix of the model.

	Reference	
predict	0	1
0	105	17
1	10	45

After fitting the best lambda, we create a confusion matrix to evaluate the accuracy of our modeling. Our data are divided into two parts in which the training part contains 70 percent of the data and the test part contains 30 percent. The reference means the true value and the prediction represents the value that the model predicted.

4 DISCUSSION

Pancreatic cancer is a highly malignant tumor of the digestive system, and the molecular mechanism of its occurrence and progression is still uncertain. In this article, we are interested in the early detection, prediction and diagnosis of pancreatic cancer. We have analyzed and discussed again based on the data of previous researchers, trying to explore which factors are related to pancreatic cancer, but it still has certain limitation.

We detect five urinary biomarkers in this study. Lymphatic vessel endothelial hyaluronan receptor 1 (LYVE1) is a receptor that binds to both soluble and immobilised hyaluronan. LYVE1 plays an important role in lymphatic hyaluronan transport and tumor metastasis. Regenerating family member 1 beta (REG1B) belongs to a family of glycoproteins and may promote regeneration of pancreatic islets. Regenerating family member 1 alpha (REG1A) is a protein which is highly similar to REG1B (Frappart and Hofmann, 2020). Trefoil factor 1 (TFF1) is a 6.5 kDa secreted protein that belongs to a family of gastrointestinal secretory peptides. It is expressed predominantly in normal gastric mucosa and involved in the regeneration and repair of urinary tract. TFF1 plays an important role in the development of cancer. Creatinine is a protein which is a product of muscle metabolism and is primarily cleared by the kidneys.

There are still many factors that are not included in the database that can still affect the incidence and prediction of PDAC to a large extent. Firstly, HER2 may play an important role in the occurrence and development of pancreatic ductal adenocarcinoma in elderly patients. The overexpression rate of HER2 may be related to gender, but its mechanism needs

further study (Ballehaninna and Chamberlain, 2012). Secondly, we still have a lot to learn from in research methods. In known studies, including drawing survival curves based on the Kaplan-Meier method, comparing survival time differences using Log-rank test, multivariate Cox regression analysis to assess the risk factors affecting patient survival, etc., can be used to obtain better results. good result. In future research, we will continue to work hard to bring better research and results.

5 CONCLUSION

In our work, it can be concluded that age, LYVE1, REG1A, REG1B, and the interaction between creatinine and REG1A are the key predictors for the diagnosis of pancreatic cancer. Their performances are successfully validated by confusion matrix. Furthermore, we plan to search for more clinical datasets to verify our model and apply our logistic regression approach to more available datasets of cardiovascular diseases and other types of cancer.

REFERENCES

- Adamska, A., Domenichini, A., & Falasca, M. (2017). Pancreatic Ductal Adenocarcinoma: Current and Evolving Therapies. *International journal of molecular sciences*, 18(7), 1338. <https://doi.org/10.3390/ijms18071338>
- Arnold, M., Rutherford, M. J., Bardot, A., Ferlay, J., Andersson, T. M., Myklebust, T. Å., Tervonen, H., Thursfield, V., Ransom, D., Shack, L., Woods, R. R., Turner, D., Leonfellner, S., Ryan, S., Saint-Jacques, N., De, P., McClure, C., Ramanakumar, A. V., Stuart-Panko, H., Engholm, G., ... Bray, F. (2019). Progress in cancer survival, mortality, and incidence in seven high-income countries 1995-2014 (ICBP SURVMARK-2): a population-based study. *The Lancet. Oncology*, 20(11), 1493–1505. [https://doi.org/10.1016/S1470-2045\(19\)30456-5](https://doi.org/10.1016/S1470-2045(19)30456-5)
- Ballehaninna, U. K., & Chamberlain, R. S. (2012). The clinical utility of serum CA 19-9 in the diagnosis, prognosis and management of pancreatic adenocarcinoma: An evidence based appraisal. *Journal of gastrointestinal oncology*, 3(2), 105–119. <https://doi.org/10.3978/j.issn.2078-6891.2011.021>
- Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6), 394–424. <https://doi.org/10.3322/caac.21492>

- Brezgyte, G., Shah, V., Jach, D., & Crnogorac-Jurcevic, T. (2021). Non-Invasive Biomarkers for Earlier Detection of Pancreatic Cancer-A Comprehensive Review. *Cancers*, *13*(11), 2722. <https://doi.org/10.3390/cancers13112722>
- Delanaye, P., Cavalier, E., & Pottel, H. (2017). Serum Creatinine: Not So Simple!. *Nephron*, *136*(4), 302–308. <https://doi.org/10.1159/000469669>
- Dinges, S.S., Hohm, A., Vandergrift, L.A. et al. Cancer metabolomic markers in urine: evidence, techniques and recommendations. *Nat Rev Urol* *16*, 339–362 (2019). <https://doi.org/10.1038/s41585-019-0185-3>
- Frappart, P. O., & Hofmann, T. G. (2020). Pancreatic Ductal Adenocarcinoma (PDAC) Organoids: The Shining Light at the End of the Tunnel for Drug Response Prediction and Personalized Medicine. *Cancers*, *12*(10), 2750. <https://doi.org/10.3390/cancers12102750>
- He, X. Y., & Yuan, Y. Z. (2014). Advances in pancreatic cancer research: moving towards early detection. *World journal of gastroenterology*, *20*(32), 11241–11248. <https://doi.org/10.3748/wjg.v20.i32.11241>
- Makawita, S., Dimitromanolakis, A., Soosaipillai, A., Soleas, I., Chan, A., Gallinger, S., Haun, R. S., Blasutig, I. M., & Diamandis, E. P. (2013). Validation of four candidate pancreatic cancer serological biomarkers that improve the performance of CA19.9. *BMC cancer*, *13*, 404. <https://doi.org/10.1186/1471-2407-13-404>
- Newton, J. L., Allen, A., Westley, B. R., & May, F. E. (2000). The human trefoil peptide, TFF1, is present in different molecular forms that are intimately associated with mucus in normal stomach. *Gut*, *46*(3), 312–320. <https://doi.org/10.1136/gut.46.3.312>
- Siegel, R. L., Miller, K. D., & Jemal, A. (2018). Cancer statistics, 2018. *CA: a cancer journal for clinicians*, *68*(1), 7–30. <https://doi.org/10.3322/caac.21442>
- Zeng, S., Pöttler, M., Lan, B., Grützmann, R., Pilarsky, C., & Yang, H. (2019). Chemoresistance in Pancreatic Cancer. *International journal of molecular sciences*, *20*(18), 4504. <https://doi.org/10.3390/ijms20184504>