Identifying Clinical Detection in Disease Context by Computational Mining of PubMed

Xiaoxue Fu*, Yi Zhou
Shanghai University of Medicine & Health Sciences, Shanghai

*Corresponding author: xiaoxuefu@sumhs.edu.cn

Abstract. To propose a novel informatics-based strategy for identifying candidate clinical tests for pancreatic cancer. We implement experiments on 6483, 27372, 28449 and 13017 publications respectively associated with risk, diagnosis, treatment and prognosis of this disease searched from PubMed and two query sets composed of 3880 clinical detection terms and 109 genes. By integrating gene-disease database, clinical information of tests, we determine the most likely terms related to the risk, diagnosis, treatment and prognosis of the disease and rank them according to our statistical scheme. We found 21, 42, 38 and 28 detection terms respectively related to risk, diagnosis, treatment and prognosis of pancreatic cancer. For risk assessment, “Adiponectin”, “Cholecystokinin (CCK)” and “Interleukin 2” with high New Term Frequency Inverse Document Frequency (NTFIDF) imply they take on greater importance to the documents where they occur in the corpus. “Carotene, Beta”, “Albumin, Serum” and “Prothrombin Fragment 1+2” with high Inverse Document Frequency (IDF) weightings indicate they provide more information. Furthermore, by incorporating 109 genes into query sets, we discovered “IGF Binding Protein-1 (IGFBP-1)” with the highest NTFIDF (0.088407) whereas “IGF Binding Protein-1 (IGFBP-1)”, “Leptin”, “Interleukin 2”, “5-Methyltetrahydrofolate” and “Gram” possessed high IDF (2.72). Similarly, for diagnosis, treatment and prognosis, “Cholecystokinin (CCK)”, “Gram”, “Neurotensin” gain the highest IDF while “Gastrin, Serum”, “Neurotensin”, “Gastrin, Serum” have the highest NTFIDF correspondingly. The findings suggest a unique opportunity for laboratory and clinical research and should be further validated in prospective research.

Keywords: clinical; genes; cancer; diagnosis; risk; detection terms.

1. Introduction
Pancreatic cancer is a highly malignant tumor with a poor prognosis [1]. Most patients (80%) present with inoperable advanced pancreatic cancer at the time of initial diagnosis [2]. Although considerable progress has been made in improving cancer survival rates over the past few decades, the 5-year survival rate for pancreatic cancer has remained mostly unchanged, rising only from 3.0% in 1975 to 5.4% in 2005 (according to data from surveillance, epidemiology and end results (SEER))[3]. Therefore, death rates from pancreatic cancer might be improved if this disease could be detected and diagnosed at an earlier stage [4]. At the beginning of the 21st century, the estimated number of pancreatic cancers
throughout the world was 110,000, with an estimated global mortality rate of 98% [5]. It is likely that some apparent long-term survivors were, in fact, misdiagnosed [6].

Currently, prevention or early diagnosis at a curable stage is exceedingly difficult [7]; patients rarely exhibit symptoms and tumors do not display sensitive and specific markers to aid detection [8-12]. Pancreatic cancers also have few prevalent genetic mutations [13-15]; the most commonly mutated genes are KRAS, CDKN2A (encoding p16), TP53 and SMAD4 [16-20] — none of which are currently druggable. Indeed, therapeutic options are limited and progress in drug development is impeded because most pancreatic cancers are complex at the genomic [21-23], epigenetic and metabolic levels, with multiple activated pathways and crosstalk evident [24-30]. Furthermore, the multilayered interplay between neoplastic and stromal cells in the tumor microenvironment challenges medical treatment [31-33]. Fewer than 20% of patients have surgically resectable disease [34]; However, neoadjuvant therapies might shift tumors towards resectability [35-37]. Although newer drug combinations and multimodal regimens in this setting, as well as the adjuvant setting, appreciably extend survival, ~80% of patients will relapse after surgery and ultimately die of their disease [38-39]. Therefore, considering life quality and entire survival is crucial.

As current therapeutic results are so dismal, it is essential to develop new methods for mining clinical tests related to this disease in order to implement preventive and screening strategies that can reduce the burden of this lethal cancer [40]. The goal of this paper is to provide support for risk, diagnosis, treatment and prognosis of pancreatic cancer, especially in personalized medicine. Furthermore, this method can also be used for other cancers.

2. Methods

2.1. Indexing-Retrieval model for publications of PubMed

We used literature from PubMed from the year 2002 to 2017 as the data source. By typing “pancreatic cancer risk”, “pancreatic cancer diagnosis”, “pancreatic cancer treatment” and “pancreatic cancer prognosis” into PubMed, we received 6483, 27372, 28449 and 13017 publications respectively. After the XML files were downloaded from PubMed directly or generated from using the tool of EFETCH [41], we imported them into a relational database in PostgreSQL. In this study, we downloaded only the PubMed-ID lists of the documents and used EFETCH to produce XML files, which is faster than using the browser to download a single XML file and has the advantage of using multiprocessing in the following step to insert all files of type XML into database. Some PubMed-IDs may change over time hence we will generate a completely new database and a new full text index once a year.

We obtained the information of 3880 clinical tests from Mayo Medical Laboratories’ Web site, which was considered as the first query set. In addition, 109 genes concerned with pancreatic cancer were acquired from the gene-disease database Orphanet [42].

The Indexing-Retrieval modeling indexed all titles, abstracts, keywords, MeSH terms and substances from the year 2002 to 2017, downloaded as XML files from PubMed. After completing the step of generating the full text index, the programme retrieved it with the query terms. The overview of the scheme conducted with literature in PubMed is shown in Fig.1. By typing “pancreatic cancer risk”, “pancreatic cancer diagnosis”, “pancreatic cancer treatment” and “pancreatic cancer prognosis” into PubMed, we obtain 6483, 27372, 28449 and 13017 publications respectively. After the XML files are downloaded from PubMed directly or generated by using the tool of EFETCH, we import them into a relational database in PostgreSQL. The results manifest how many terms were discovered and which terms co-occur in the same abstract or title sorted by PubMed-IDs.
2.2. Statistical analysis
In information retrieval, TF-IDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus [43]. The TF-IDF value increases proportionally to the number of times a word appears in the document and is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general. TF-IDF is one of the most popular term-weighting schemes today; 83% of text-based recommender systems in digital libraries use TF-IDF [44].

Variations of the TF-IDF weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query. One of the simplest ranking functions is computed by summing the TF-IDF for each query term; many more sophisticated ranking functions are variants of this simple model. The weight of a term that occurs in a document is simply proportional to the term frequency [45]. The specificity of a term can be quantified as an inverse function of the number of documents in which it occurs [46]. The TF-IDF is the product of two statistics, term frequency and inverse document frequency. In the case of the term frequency $t_f(t,d)$, the simplest choice is to use the raw count of a term $t$ in a document $d$, i.e. the number of times that term $t$ occurs in document $d$. If we denote the raw count by $f(t,d)$, then the simplest $t_f$ scheme is $t_f(t,d) = f(t,d)$. Other possibilities are included in the research [47]. Here, according to our data, term frequency calculation is adjusted for the following equation:

$$t_f(t,D) = \text{ave} \left( \frac{f(t,d_i)}{|d_i|} \right) \quad t \in d_i, d_i \in D \quad i = 1,2,3,\ldots$$  

Fig 1. Flow Diagram Outlining the Methodology of Indexing-Retrieval Model.
Where $|d_i|$ equals the number of words in $d_i$, $f(t,d_i)$ indicates the raw count of a term $t$ in a document $d_i$. The term frequency of term $t$ in corpus $D$ is the mean of all $f(t,d_i)$.

The inverse document frequency (idf) is a measure of how much information the word provides, that is, whether the term is common or rare across all documents. It is the logarithmically scaled inverse fraction of the documents that contain the word, obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient.

$$idf(t, D) = \log \frac{N}{|\{d \in D, t \in d\}|} \quad (2)$$

In above formula, $N$ equals total number of documents in the corpus $D$, that is, $N=|D|$; $|\{d \in D, t \in d\}|$ is number of documents where the term $t$ appears ($tf(t,d_i) \neq 0$); This is the traditional model of implementing IDF. We will discuss extensions of this algorithm along with an analysis of TF-IDF according to our own results. We will introduce the improved model next. We applied it for each term according to our estimate of producing the query according to the model. We would like to estimate TF-IDF of each term as follows:

$$tf - idf(t, D) = \begin{cases} 
\log(1 + \frac{N}{|\{d \in D, t \in d\}|}) & \text{weighting for how much information term } t \text{ provides} \\
(tf(t, D) \times \log \frac{N}{|\{d \in D, t \in d\}|}) & \text{weighting for how important a term is to the corpus} 
\end{cases} \quad (3)$$

In this formula, the first term represents the discrimination power of term $t$ for the collection and the second term expresses the popularity of term $t$ among the users.

3. RESULTS

3.1. Relevant clinical detection terms to pancreatic cancer

By implementing the Indexing-Retrieval model on the collections and query set of clinical detection terms, the outputs show 6483, 27372, 28449 and 13017 documents are respectively indexed and 3880 clinical-test terms are searched. And we discover 21, 42, 38 and 28 terms correspondingly.

3.2. Ranking the terms based on new weighting model

Let us now think about the results from running the improved weighting model on our data. In this case, terms with high IDF value show that they provide more knowledge. Nevertheless, terms having high new TF-IDF prove they could be more important to the collection. We tested our new TF-IDF scheme on four corpus for risk, diagnosis, treatment and prognosis of pancreatic cancer with query set of 3880 clinical testing terms (Fig. 2.). By implementing the new improved TF-IDF model approach on the candidate terms, we rank them according to two weights of TF-IDF and idf separately. TF-IDF indicates the importance of each term to the collection and idf specifies the amount of information each term provides. Fig. 2. is described for risk, diagnosis, treatment and prognosis correspondingly.
Fig 2. Candidate Detection Terms Ranked Based on an Improved Weighting Model.
3.3. Gene-related testing terms generated and ranked by incorporating gene-disease database and clinical information of tests

We now attempt to improve our probability estimates, which should yield better retrieval performance. To achieve this goal, we integrate disease-gene database and clinical information of clinical testing into the novel weighting model to improve the performance.

We also use the above four collections as the empirical corpus. A gene-list document including 109 genes is extracted from aforesaid database Orphanet to be applied as an additional query set. After retrieving the documents with the two query sets, including clinical tests and genes, we receive 7, 11, 8, 6 gene-related detection terms to risk, diagnosis, treatment and prognosis of pancreatic cancer. The results are summarized in Fig. 3. By implementing our informatics-based approach on two query sets and weighting model, we rank candidate detection terms respectively. TF-IDF indicates the importance of each term to the collection and IDF manifests the amount of information each term provides. Fig. 3. is shown for risk, diagnosis, treatment and prognosis accordingly.
Fig 3. Gene-Related Clinical Detection Terms Found for Pancreatic Cancer by Computational Mining of PubMed.

4. DISCUSSION
We identify a relationship between clinical detection and diseases by computational mining of biomedical literature. Literature of PubMed associated with clinical tests and pancreatic cancer and rank them accordingly. We support this association incorporating gene-disease database and clinical information of clinical tests into the new improved statistical model.

The overall 5-year survival for pancreatic cancer has changed little over the past few decades, and pancreatic cancer is predicted to be the second leading cause of cancer-related mortality in the next decade in Western countries. The use of bio-marker to help define treatment and the potential of neoadjuvant therapies also offer opportunities to improve outcomes [48]. Given the dismal prognosis for patients with pancreatic cancer, it is important that additional rigorous investigation will be necessary to support and extend these interesting findings. Methods for extracting biomedical facts from the scientific literature have improved considerably, and the associated tools will probably soon be used in
many laboratories to automatically annotate and analyze the growing number of system-wide experimental data sets. Owing to the increasing body of text and the open-access policies of many journals, literature mining is also becoming useful for both hypothesis generation and biological discovery [49].

This study has limitations that warrant consideration. We index and retrieve literature to determine related clinical tests to the disease of pancreatic cancer, which may introduce bias into our analysis. We therefore incorporate genes involved in this disease and clinical information of clinical tests to optimize the consequences. More important, in this study, we were not sufficiently powered to undertake case-ascertainment and recruitment for this study and future studies should examine the effect of clinical detection on the disease. A further limitation is that the data resources we used. Although we incorporated gene data from the disease-gene database, more data should be included in the future to obtain more useful results. This way, we will improve the accuracy of our results and reduce data limitations. The other non-cancer genes may also have potential in cancer therapy.50 Finally, our analysis is retrospective and does not establish a causal association between the clinical tests and the disease.

Scientific literature is a powerful resource in cancer research. In particular, through the informatics approaches to analyze publications of PubMed, we demonstrate the data aggregation of clinical tests and gene-disease database for pancreatic cancer, quantified the risk, diagnosis, treatment and prognosis of pancreatic cancer, as well as other cancers, identified clinical detection for pancreatic cancer and began the first screening trials for pancreatic cancer. The lists of clinical tests and gene-disease database will expand in the future. In consequence, personalized treatments that target these testing defects and improved screening of high-risk individuals provide an opportunity to reduce the mortality from this cancer.

Acknowledgments
I would like to express my gratitude to corresponding author Yi Zhou who helped me during the writing of this thesis. I also owe a special debt of gratitude to all the professors in Foreign Languages Institute, from whose devoted teaching and enlightening lectures I have benefited a lot and academically prepared for the thesis.

References
[1] Hee Seung Lee, Chan Young Jang, Sun A Kim, et al. Combined use of CEMIP and CA 19-9 enhances diagnostic accuracy for pancreatic cancer. Scientific Reports 2018; volume 8;Article number: 3383.
[2] Ryan, D. P., Hong, T. S. & Bardeesy, N. Pancreatic adenocarcinoma. N Engl J Med 2014; 371, 2140–2141.
[3] Howlader N, et al. SEER Cancer Statistics Review 2010; 1975–2005.
[4] Cleary SP, et al. Prognostic factors in resected pancreatic adenocarcinoma: analysis of actual 5-year survivors. J. Am. Coll. Surg. 2004; 198: 722–731.
[5] Parkin, D. M., Bray, F., Ferlay, J. & Pisani, P. Estimating the world cancer burden: Globocan 2000. Int. J. Cancer 2001; 94, 153–156.
[6] Carpelan-Holmstrom, M. et al. Does anyone survive pancreatic ductal adenocarcinoma? A nationwide study re-evaluating the data of the Finnish Cancer Registry. Gut 2005; 54, 385–387.
[7] Jorg Kleeff, Murray Korc, Minoti Apte, Carlo La Vecchia, Colin D. Johnson, Andrew V. Biankin, Rachel E. Neale, et al. Pancreatic cancer. Nature Reviews Disease Primers 2016; volume 2; Article number: 16022.
[8] Ferlay, J. et al. GLOBOCAN 2012: cancer incidence and mortality worldwide: IARC CancerBase No. 11. International Agency for Research on Cancer [online] http://globocan.iarc.fr, 2013.
[9] Rahib, L. et al. Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. Cancer Res. 2014; 74, 2913–2921.
[10] Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2015. CA Cancer J. Clin. 2015;65, 5–29.
[11] He, J. et al. 2564 resected periampullary adenocarcinomas at a single institution: trends over three decades. HPB (Oxford) 2014; 16, 83–90.
[12] Rosato, V. et al. Population attributable risk for pancreatic cancer in Northern Italy. Pancreas 2015; 44, 216–220.
[13] Boffetta, P., Boccia, S. & La Vecchia, C. (eds) A Quick Guide to Cancer Epidemiology. Springer 2014.
[14] International Agency for Research on Cancer (IARC). Cancer incidence in five continents. Vol. X. IARC [online] http://ci5.iarc.fr, 2013.
[15] Malvezzi, M. et al. European cancer mortality predictions for the year 2015: does lung cancer have the highest death rate in EU women? Ann. Oncol. 2015; 26, 779–786. This work demonstrates that pancreatic cancer is a major neoplasm without favourable incidence and mortality trends over the past decades, and its prognosis remains dismal.
[16] Cancer Research UK. Pancreatic cancer statistics. Cancer Research UK 2015 [online] www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/pancreatic-cancer.
[17] Anderson, K., Mack, T. & Silverman, D. in Cancer of the Pancreas (eds Schottenfeld, D. & Fraumeni, J.) Oxford Univ. Press 2006; 721–762.
[18] Bosetti, C. et al. Cigarette smoking and pancreatic cancer: an analysis from the International Pancreatic Cancer Case–Control Consortium (Panc4). Ann. Oncol. 2012; 23, 1880–1888.
[19] Parkin, D. M. 2. Tobacco-attributable cancer burden in the UK in 2010. Br. J. Cancer 2011; 105, S6–S13.
[20] Whiteman, D. C. et al. Cancers in Australia in 2010 attributable to modifiable factors: introduction and overview. Aust. N. Z. J. Public Health 2015; 39, 403–407.
[21] Boffetta, P., Hecht, S., Gray, N., Gupta, P. & Straif, K. Smokeless tobacco and cancer. Lancet Oncol. 2008; 9, 667–675.
[22] Behrens, G. et al. Physical activity and risk of pancreatic cancer: a systematic review and meta-analysis. Eur. J. Epidemiol. 2015; 30, 279–298.
[23] Genkinger, J. M. et al. Central adiposity, obesity during early adulthood, and pancreatic cancer mortality in a pooled analysis of cohort studies. Ann. Oncol. 2015; 26, 2257–2266.
[24] Larsson, S. C. & Wolk, A. Red and processed meat consumption and risk of pancreatic cancer: meta-analysis of prospective studies. Br. J. Cancer 2012; 106, 603–607.
[25] Rohrmann, S. et al. Meat and fish consumption and risk of pancreatic cancer: results from the European Prospective Investigation into Cancer and Nutrition. Int. J. Cancer 2013; 132, 617–624.
[26] Bosetti, C. et al. Nutrient-based dietary patterns and pancreatic cancer risk. Ann. Epidemiol. 2013; 23, 124–128.
[27] MacMahon, B., Yen, S., Trichopoulous, D., Warren, K. & Nardi, G. Coffee and cancer of the pancreas. N. Engl. J. Med. 1981; 304, 630–633.
[28] Tramacere, I. et al. Alcohol drinking and pancreatic cancer risk: a meta-analysis of the dose–risk relation. Int. J. Cancer 2010; 126, 1474–1486.
[29] Lucenteforte, E. et al. Alcohol consumption and pancreatic cancer: a pooled analysis in the International Pancreatic Cancer Case–Control Consortium (PanC4). Ann. Oncol. 2012; 23, 374–382.
[30] Bosetti, C. et al. Diabetes, antidiabetic medications, and pancreatic cancer risk: an analysis from the International Pancreatic Cancer Case–Control Consortium. Ann. Oncol. 2014; 25, 2065–2072.
[31] Chari, S. T. et al. Pancreatic cancer-associated diabetes mellitus: prevalence and temporal association with diagnosis of cancer. Gastroenterology 2008; 134, 95–101.
[32] Bosetti, C. et al. Ulcer, gastric surgery and pancreatic cancer risk: an analysis from the
International Pancreatic Cancer Case–Control Consortium (PanC4). Ann. Oncol. 2013; 24, 2903–2910.

[33] Turati, F. et al. Family history of cancer and the risk of cancer: a network of case–control studies. Ann. Oncol. 2013; 24, 2651–2656.

[34] Rizzato, C. et al. ABO blood groups and pancreatic cancer risk and survival: results from the PANcreatic Disease ReseArch (PANDoRA) consortium. Oncol. Rep. 2013; 29, 1637–1644.

[35] Hezel, A. F., Kimmelman, A. C., Stanger, B. Z., Bardeesy, N. & Depinho, R. A. Genetics and biology of pancreatic ductal adenocarcinoma. Genes Dev. 2006; 20, 1218–1249.

[36] Tanaka, M. Thirty years of experience with intraductal papillary mucinous neoplasm of the pancreas: from discovery to international consensus. Digestion 2014; 90, 265–272.

[37] Preis, M. & Korc, M. Signaling pathways in pancreatic cancer. Crit. Rev. Eukaryot. Gene Expr. 2011; 21, 115–129.

[38] Pasca di Magliano, M. et al. Common activation of canonical Wnt signaling in pancreatic adenocarcinoma. PLoS ONE 2007; 2, e1155.

[39] Tanizaki, J., Okamoto, I., Sakai, K. & Nakagawa, K. Differential roles of trans-phosphorylated EGFR, HER2, HER3, and RET as heterodimerisation partners of MET in lung cancer with MET amplification. Br. J. Cancer 2011; 105, 807–813.

[40] Gore, A. J., Deitz, S. L., Palam, L. R., Craven, K. E. & Korc, M. Pancreatic cancer-associated retinoblastoma 1 dysfunction enables TGF-β to promote proliferation. J. Clin. Invest. 2014; 124, 338–352.

[41] Döring K, Grüning BA, Telukunta KK, Thomas P, Günther S. PubMedPortable: A Framework for Supporting the Development of Text Mining Applications. PLoS ONE 2016; 11(10): e0163794.

[42] Hui Yang, Peter N Robinson & Kai Wang. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nature Methods 2015; volume12, pages841–843.

[43] Rajaraman, A.; Ullman, J.D. (2011). "Mining of Massive Datasets (PDF). Data Mining 2011;1–17. doi:10.1017/CBO9781139058452.002. ISBN 978-1-139-05845-2.

[44] Breitinger, Corinna; Gipp, Bela; Langer, Stefan. "Research-paper recommender systems: a literature survey". International Journal on Digital Libraries 2005; 17 (4): 305–338. doi:10.1007/s00799-015-0156-0. ISSN 1432-5012.

[45] Luhn, Hans Peter. "A Statistical Approach to Mechanized Encoding and Searching of Literary Information" (PDF). IBM Journal of research and development 2015; 1 (4): 315.

[46] Spärck Jones, K. "A Statistical Interpretation of Term Specificity and Its Application in Retrieval". Journal of Documentation 1972; 28: 11–21.

[47] Manning, C.D.; Raghavan, P.; Schutze, H. "Scoring, term weighting, and the vector space model". Introduction to Information Retrieval 2005; 100.

[48] John P. Neoptolemos, Jörg Kleeff, Patrick Michl, et al. Therapeutic developments in pancreatic cancer: current and future perspectives. Nature Reviews Gastroenterology & Hepatology. 2018; volume 15, pages333–348.

[49] Lars Juhl Jensen, Jasmin Saric & Peer Bork. Literature mining for the biologist: from information retrieval to biological discovery. Nature Reviews Genetics. 2006; volume 7, pages 119–129.

[50] Bollacker, Kurt D.; Lawrence, Steve; Giles, C. Lee. "CiteSeer: An Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications". Proceedings of the Second International Conference on Autonomous Agents 1998; 116–123