Development of a Predictive Dynamic Model to Assess Pancreatic Cancer Risk

Nirosha Rathnayake[1], MS, Evi Farazi[2], PhD, Danstan Bagenda[3], PhD , Xiaoyue Zoe Cheng[4], PhD, Quan Ly[5], MD,
[1]Department of Biostatistics, UNMC, [2]Department of Epidemiology, UNMC,[3]Department of Epidemiology, UNMC, [4]Department of Mathematics, UNO, [5]Department of Division of Surgical Oncology, Department of Surgery, UNMC

Abstract

Background: Pancreatic cancer poses a significant burden in the United States and worldwide due to its late diagnosis and poor survival. It shows disease heterogeneity, which most likely reflects its complex etiology, including both genetic and environmental factors. The latter affect the genome by promoting the generation of mutations and the epigenome through epigenetic modifications. Thus, a patient's genomic profile reflects a multitude of environmental exposures, which may explain tumor heterogeneity. Many studies have focused on identifying environmental risk factors for pancreatic cancer and have pointed to tobacco smoking, alcohol consumption, occupational exposures, obesity, diet, past medical history, drugs, and diabetes. Most of these factors are studied in isolation even though patients are exposed to many of these factors concurrently.

Objective: The proposed work was designed to utilize Bayesian network analysis to identify novel risk factors and interactions of these in pancreatic cancer as well as develop a predictive model for pancreatic cancer risk based on environmental exposures, lifestyle and medical conditions information.

Methods: The data set included 1990 pancreatic cancer cases, 346 controls, and 2022 high risk individuals collected from the Pancreatic Cancer Registry (PCCR) and Great Plains Health Informatics Database (GPHID), respectively, from UNMC, Pittsburgh, NorthShore, Alabama, and Michigan. The clinical and environmental data was computationally analyzed using Bayesian networks (BN) through algorithms from information theory such as minimum description length (MDL) and mutual information (MI) using BayesiaLab software[1]. Missing values were assessed using Structural Expectation Maximization (Structural EM) algorithm. We obtained two different models; 1. for cases and controls, 2. for controls and high risk individuals and examined the prediction accuracy of both models.

Results: The current results were obtained using unsupervised learning technique based on the total minimum description length (MDL) score, which has to be the lowest for the better model. BN contained many of the previously reported factors for the pancreatic cancer such as tobacco smoking, alcohol consumption, occupation exposures, other medical conditions, BMI. While Most of the supplement types have a protective effect on pancreatic cancer.

While the case-control BN model correctly predicts 381 cases out of 416 with 91.6% true positive rate (TPR), and 34 controls out of 40 with 85% TNR, control-high risk BN model performs better in predicting the correct assignment of the group having prediction accuracy of 97.8% for controls and 94.4% for high risk individuals. The predictive importance of the risk factors is shown in the plot 1.

Conclusions: The results showed the promise of these machine learning methods in generating a prediction model for pancreatic cancer risk and also both models perform better with the predictin.

Introduction

The American Cancer Society estimates that 48,960 new cases of pancreatic cancer and 40,560 deaths occurred in 2015.[3] The deadly nature of this cancer is attributed to late diagnosis; 53% patients have metastatic disease at diagnosis with a 2.4% 5-year survival rate.[2] Pancreatic cancer shows disease heterogeneity,[4,5] which likely reflects the complex disease etiology, involving both genetic and environmental factors. These factors affect the genome through the generation of mutations and epigenetic changes. A patient's exposure to a multitude of environmental factors may explain tumor heterogeneity.[6,7]

Many studies have shown influence of tobacco smoking, alcohol consumption, occupational exposures, obesity, diet, past medical history, drugs, and diabetes in pancreatic cancer development.[8-22] Most of these factors are studied in isolation even though individuals are exposed to many of them concurrently. Cancer is a complex disease not caused by a single factor but rather a multitude of factors that interact in different ways. Nutrition is becoming increasingly accepted as a factor playing a role in carcinogenesis. However, it would be overly simplistic to state that bad nutrition habits alone would result in the development of cancer. Not all individuals with bad nutrition habits or who smoke go on to develop cancer, pointing to the complex gene-environment interactions that differ in each individual and determine the course of a disease like cancer. For example, diet has been shown to influence the immune system and the immune system influences cancer development.[23] Therefore, it is possible that an individual with a compromised immune system due to other medical conditions may be more susceptible to the effects of diet on cancer development compared to an individual with an intact immune system. This is an example of an interaction between two risk factors and how it can affect cancer development. In reality, interactions between multiple risk factors are at play in the cancer process. Some interactions are known but others are not and need to be investigated in order to have a clearer understanding of the complex nature of cancer development that appears to be unique in each individual.

BNs are non-parametric probabilistic models and have performed well in predictive studies. The BN probability table depends on two main parameters, the number of parents and the number of states of the parent and child nodes. The number of states can be determined using equal distance, normalized equal distance, equal frequency, k-means or tree based method, however it can also be our choice based on our expert knowledge.

Specific Aims

- Develop a predictive model for pancreatic cancer risk based on environmental exposures, lifestyle, and medical history using Bayesian Network Analysis.
- Identify novel risk factors and risk factor interactions in pancreatic cancer development using Bayesian Network Analysis thus it would permit the identification of individuals at risk for pancreatic cancer and lead to earlier diagnosis and improved survival.

Methods

The data set included 1990 pancreatic cancer cases, 346 controls, and 2022 high risk individuals collected from the Pancreatic Cancer Registry (PCCR) and Great Plains Health Informatics Database (GPHID), respectively, from the five centers listed above. The clinical and environmental data were analyzed using Bayesian networks (BN) through algorithms from information theory such as MDL and MI using BayesiaLab software[1]. BayesiaLab requires to discretize all the continuous variables and if necessary aggregate discrete variables before perform any learning techniques. We discretized continuous variables and aggregated some of the discrete variables such as country of birth, alcohol, diet, supplement consumption, smoking, and education level. The BN was obtained using unsupervised learning estimated the links between the variables based on minimum description length (MDL), which is a score-based algorithm.

$$MDL(B,D) = \alpha DL(B) + DL(D|B)$$

The BN was obtained by MWST algorithm based on the MDL. The MDL is a two-feature score that describes the number of bits required for representing a "model" and "data given this model". In machine learning, these two terms will be the Bayesian network and log-likelihood of the data given the model respectively. [1] The minimum value of MDL gives the best solution to the data set. First, the data set split into two sets, 20% of the data was randomly selected for the test set and the remainder 80% of the data was considered to train the model by taking 70% to fit the model and 30% to evaluate the model. The model was refined by performing 10-folds cross validation on the trained model to alleviate any sampling bias that may occur from randomly selected validation set (Table 1.2 & 1.4). In addition to the graphical diagrams generated by BNs (Fig 1), which could provide information regarding risk factors and interactions among them, the degree of dependency among these variables could be expressed with probabilistic terms (parameters) quantitatively by obtaining the conditional probabilities. The importance of predictive variables when making predictions for cases was estimated based on mutual information (MI), which was estimated by marginal and conditional entropy (Fig. 1.2). MI is similar to correlation and covariance using in traditional statistical analysis to determine the relative importance between variables.

Results

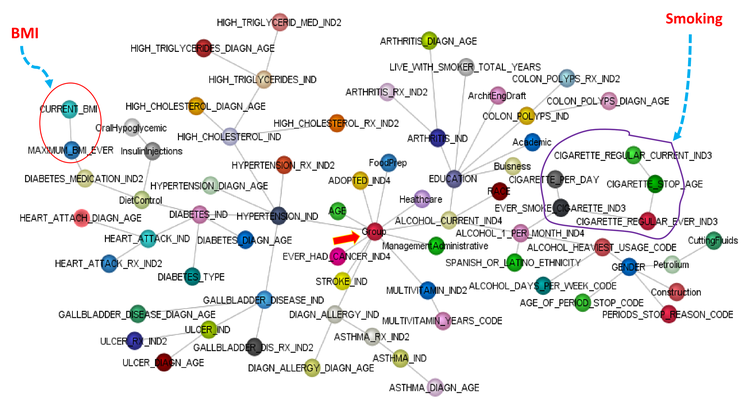

Fig. 1. Bayesian network using unsupervised machine learning

- The associated variables are connected in the network (Fig. 1), while the most correlated variables to the group has a direct connection to group variable, that has two categories (cases & controls).
- The table 1.1 shows that the model has correctly classified 1099 cases out of 1120, 113 controls out of 188 with precision of 98.1% and 60.1%, respectively.
- The model performs well on the test data set giving 91.6% precision for the cases and 85.0% for the controls.

Case-Control Model Evaluation using 10-fold Cross Validation			
$RMSE = 0.254, R^2 = 0.482$		True	
		Cancer (1120)	Control (188)
Predicted	Cancer (1174)	1099 (98.1%)	75 (39.9%)
	Controls (134)	21 (1.9%)	113 (60.1%)
The overall precision = 92.7%. Of the 1120 cancer cases of the test data set, 21 were incorrectly classified (FPR=1.9%) & of the 188 controls, 75 were incorrectly misclassified (FNR=39.9%).			

Table 1.1: confusion matrix for case-control BN model

Prediction Accuracy of Case-Control Model			
		True	
		Cancer (416)	Control (40)
Predicted	Cancer (387)	381 (91.6%)	6 (15.0%)
	Controls (69)	35 (8.4%)	34 (85.0%)
On new data set, the prediction accuracy of the cancer group is 91.6% and in the control group 85.0%			

Table 1.2: Testing case-control BN model

- Some of the health related variables that increases the risk of PC will be hypertension, high cholesterol, diabetes, gall bladder disease, and heart attack etc, with the highest increase in probability of 10.8% from hypertension and 3.33% from high cholesterol.
- The model also picks that it is more likely to have the high increase in probability in older age group (> 59 years) compared to other young age groups.
- Also, multivitamin supplement is more likely to be protective and there is no risk from allergies.

Results Ctd.

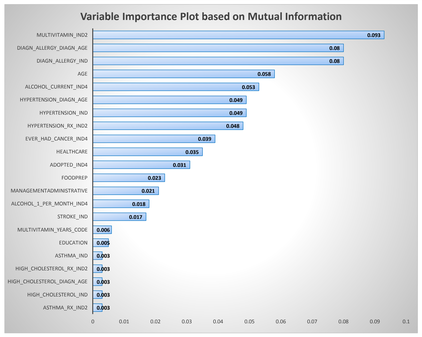

Fig. 1.2 variable importance plot

- Figure 1.2 shows the variable importance of making predictions for the cases using case-control BN model. While some of the important variables reduce the risk, the other will increase the risk of PC development.
- The both model predicts that the increase in likelihood in cases, as well as high risk group is higher in smoking consumption than alcohol consumption.
- The second BN model is based on control and high risk individuals data. The model performs well with an overall precision of 93.5%, and RMSE of 0.237 (Fig. 1.3).

Control-High Risk Model Evaluation using 10-fold Cross Validation			
$RMSE = 0.237, R^2 = 0.531$		True	
		Controls (184)	High Risk(1142)
Predicted	Controls (120)	109 (59.2%)	11 (1.0%)
	High Risk (1206)	75 (40.8%)	1131 (99.0%)
The overall precision = 93.5%. Of the 184 controls of the test data set, 75 were incorrectly classified (FPR = 40.8%) & of the 1142 high risk individuals, 11 were incorrectly misclassified (FNR = 1.0%).			

Table 1.4: Confusion matrix for control-high risk BN model

- The trained model was tested using the test data set which correctly predicted with a prediction accuracy of 97.8% for controls and 94.4% for high risk (Fig. 1.4).

Prediction Accuracy of Control-High Risk Model			
		True	
		Controls (416)	High Risk (40)
Predicted	Controls (387)	45 (97.8%)	23 (5.6%)
	High Risk (69)	1 (2.2%)	387 (94.4%)
On new data set, the prediction accuracy of the cancer group is 97.8% and in the control group 94.4%			

Fig. 1.4 Testing control-high risk BN model

- Based on the test data set the misclassification rate is 5.3% and the precision is 94.7%.
- Furthermore, the BN model can be used to obtain the conditional probability which gives the probability increase in PC. It also shows the interaction between the risk factors.
- The likelihood increase in cases given an individual with hypertension and age >59 years old is higher (12.18%) than an individual with hypertension (10.75%).

Conclusions

The misclassification rate is low in both models, representing that accurate predictions on unseen data sets. Identifying individuals at higher risk for pancreatic cancer development would enable clinicians to more effectively screen this small subpopulation. This will result in early diagnosis and improved disease outcomes. Furthermore, the model will provide new information on modifiable risk factors of pancreatic cancer, which would contribute to disease prevention. Having examine to relationship between high risk individuals and controls is very useful for clinician to direct the individuals for PC screening, hence it will help to detect PC early.

References

1. Stefan Conrady, Lionel Jouffe, Bayesian Networks & BayesiaLab, A practical introduction for researchers, 2015.
2. Cotterchio, M., Lowcock, E., Hudson, T. J., Greenwood, C., & Gallinger, S. (2014). Association between allergies and risk of pancreatic cancer. *Cancer Epidemiology and Prevention Biomarkers*.
3. American Cancer Society. Cancer Facts and Figures 2015. Atlanta: American Cancer Society; 2015.
4. Collisson EA, Sadanandam A, Olson P, Gibb W, Truitt M, Gu S, Cooc J, Weinkle J, Kim G, et al. Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. Nature Medicine. 2011. 17: 500 – 503.
5. Bailey P, Chang D, et al. Genomic analyses identify molecular subtypes of pancreatic cancer. Nature. 2016. 531: 47 – 52.
6. Fenga. Occupational exposure and risk of breast cancer. Biomed. Rep. 2016. 4(3):282-292.
7. Krakowsky RH and Tollefsbol TO. Impact of Nutrition on Non-Coding RNA Epigenetics in Breast and Gynecological Cancer. Front Nutr. 2015. 2:16.
8. Maisonneuve P and Lowenfels A. Risk factors for pancreatic cancer: a summary review of meta-analytical studies.
9. Iodice S, Gandini S, Maisonneuve P et al. Tobacco and the risk of pancreatic cancer: a review and meta-analysis. Langenbecks Arch Surg. 2008. 393: 535 – 545.
10. Lynch S, Vrieling A, Lubin J et al. Cigarette smokinh and pancreatic cancer: a pooled analysis from the pancreatic cancer cohort consortium. Am J Epidemiol. 2009. 170: 403 – 413.
11. Genkinger J, Spiegelman D, Anderson K et al. Alcohol intake and pancreatic cancer risk: a pooled analysis of fourteen cohort studies. Cancer Epidemiol Biomarkers Prev. 2009. 18: 765 – 776.
12. Dong J, Zou J, Yu X. Coffee drinking and pancreatic cancer risk: a meta-analysis of cohort studies. World J Gastroenterol. 2001. 17: 1204 – 1210.
13. Genkinger J, Li R, Spiegelman D et al. Coffee, tea, and sugar-sweetened carbonated soft drink intake and pancreatic cancer risk: a pooled analysis of 14 cohort studies. Cancer Epidemiol Biomarkers Prev. 2012. 21: 305 – 318.
14. Collins J, Esmen N, Hall T. A review and meta-analysis of formaldehyde exposure and pancreatic cancer. Am J Ind Med. 2001. 39: 336 – 345.
15. Ojajarci A, Partanen T, Ahlbom A et al. Risk of pancreatic cancer in workers exposed to chlorinated hydrocarbon solvents and related compounds: a meta-analysis. Am J Epidemiol. 2001. 153: 841 – 850.
16. Stolzenberg-Solomon R, Jacobs E, Arslan A et al. Circulating 25-hydroxyvitamin D and risk of pancreatic cancer: cohort consortium vitamin D pooling project of rarer cancers. Am J Epidemiol. 2010. 172: 81 – 93.
17. Berrington de Gonzalez A, Sweetland S, Spencer E. A meta-analysis of obesity and the risk of pancreatic cancer. Br J Cancer. 2003. 89: 519 – 523.
18. Bae J, Lee E, Guyatt G. Citrus fruit intake and pancreatic cancer risk. The quantitative analysis of case-control and cohort studies. Cancer Epidemiol. 2012. 36: 60 – 67.
19. Larsson S, Wold A. Red and processed meat consumption and risk of pancreatic cancer: meta-analysis of prospective studies. Br J Cancer. 2012. 106: 603 – 607.
20. Huxley R, Ansary-Moghaddam A, Berrington de Gonzalez A et al. Type II diabetes and pancreatic cancer: a meta-analysis of 36 studies. Br J Cancer. 2005. 92: 2076 – 2083.
21. Gandini S, Lowenfels A, J'Affee E. Allergies and the risk of pancreatic cancer: a meta-analysis with review of epidemiology and biological mechanisms. Cancer Epidemiol Biomarkers Prev. 2005. 14: 1908 – 1916.
22. Larsson S, Giovanucci E, Bergvist L et al. Aspirin and non-steriodal anti-inflammatory drug use and risk of pancreatic cancer: a meta-analysis. Cancer Epidemiol Biomarkers Prev. 2006. 15: 2561 – 2564.
23. Veldhoen M, Ferreira C. Influence of nutrient-derived metabolites on lymphocyte immunity. Nature Medicine. 2015. 21 (7): 709-718. doi: 10.1038/nm.3894.
24. Li, D. (2012). Diabetes and pancreatic cancer. *Molecular carcinogenesis*, *51*(1), 64-74.
25. Rahman, F., Cotterchio, M., Cleary, S. P., & Gallinger, S. (2015). Association between alcohol consumption and pancreatic cancer risk: A case-control study. *PloS one*, *10*(4), e0124489.

Acknowledgements

- This work was supported by a pilot grant from the Cancer Prevention and Control Program at the Fred and Pamela Buffet Cancer Center.
- Dr. Evi Farazi (UNMC, Epidemiology)
- Dr. Danstan Bagenda (UNMC; Anesthesiology)
- Dr. Xiaoyue Zoe Cheng (UNO; Mathematics)
- Dr. Jane Meza (UNMC; Biostatistics)
- Dr. Quan Ly (UNMC – PI of PCCR)
- Dr. Oleg Shats (UNMC – Assistant Director for Cancer Informatics at FPBCC)