Creation of a Machine Learning App to Facilitate Pancreatic Cancer Prediction

Reid Fleishman

All graphics, tables, charts, graphs, and images in this presentation were created by Reid Fleishman, unless otherwise stated under it

Prevalence of Pancreatic Cancer

- Deadly disease that is difficult to diagnose at an early stage (Siegel et al., 2020)

3%

5-yr Survival Rate for Late-Stage Diagnosis
(Siegel et al., 2020)

37%

5-yr Survival Rate for Early-Stage Diagnosis
(Siegel et al., 2020)

Early diagnosis is *key!*

- Effective screening for pancreatic cancer at an early stage is lacking (McGuigan et al., 2018; Poruk et al., 2013)

Need an effective way to aid in the diagnosis of pancreatic cancer at an early stage

Common Risk Factors of Pancreatic Cancer

Significant Risk Factors

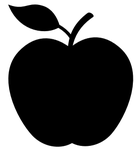
Age

Sex

Race

Smoking

BMI



Diabetes

Possible Risk Factors

Activity

Alcohol Consumption

Heart Conditions

Depression

Lowenfels & Maisonneuve, 2006; Arnold et al., 2009; Larsson et al., 2007; Gomez-Rubio et al., 2015; Muhammad et al., 2019; Everhart & Wright, 1995; Hassan et al., 2008; Ye et al., 2002; Lindgren et al., 2005; Zheng et al., 1993; Coughlin et al., 2000; Michaud et al., 2001; “Pancreatic Cancer Risk Factors,” n.d.)

Use of Machine Learning (ML) for Disease Prediction

Methods to Evaluate Models

$$\frac{TP}{TP + FN}$$

Sensitivity
(true positive rate)

$$\frac{TN}{TN + FP}$$

Specificity
(true negative rate)

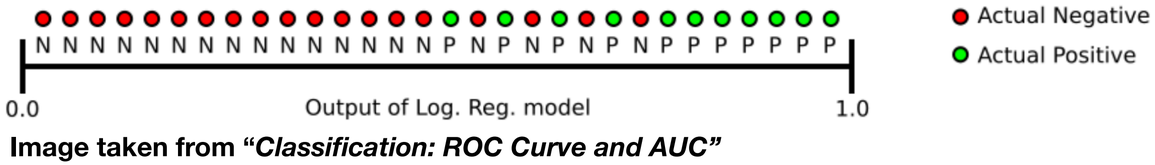


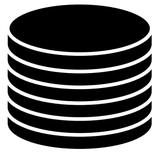
Image taken from "*Classification: ROC Curve and AUC*"

AUC
(area under the ROC curve)

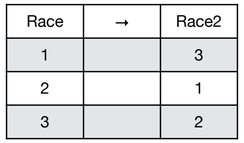
My Goals

1. Develop an effective machine learning model to predict one's risk of developing pancreatic cancer, improving upon previously-created models
    - Random Forest (RF), Decision Tree (DT), Boosted Trees (BT), Logistic Regression (LR), and Support Vector Machine (SVM)
    - Determine the most important variables for inclusion in a model
        - Specifically, determine the impact of including depression
2. Develop an iPhone app to facilitate data input and prediction output by doctors and/or patients

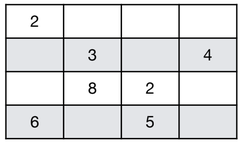
Dataset, Preprocessing, Training, and App



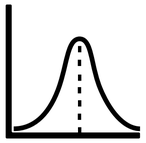
Data from IPUMS Health Surveys and PLCO Study (1993-2018)
752,527 participants (983 with PC)
17 variables



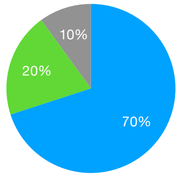
SPSS for re-coding certain variables



Missing value imputation
(Aljuaid & Sasi, 2016)



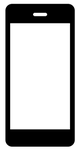
Normalize the data
(Muhammad et al., 2019; Roffman et al., 2018a, Roffman et al., 2018b)



Data randomly split into 70% training, 20% validation, 10% testing

Balance dataset using upsampling (0.13% PC Cases)
(Boyle, 2019)

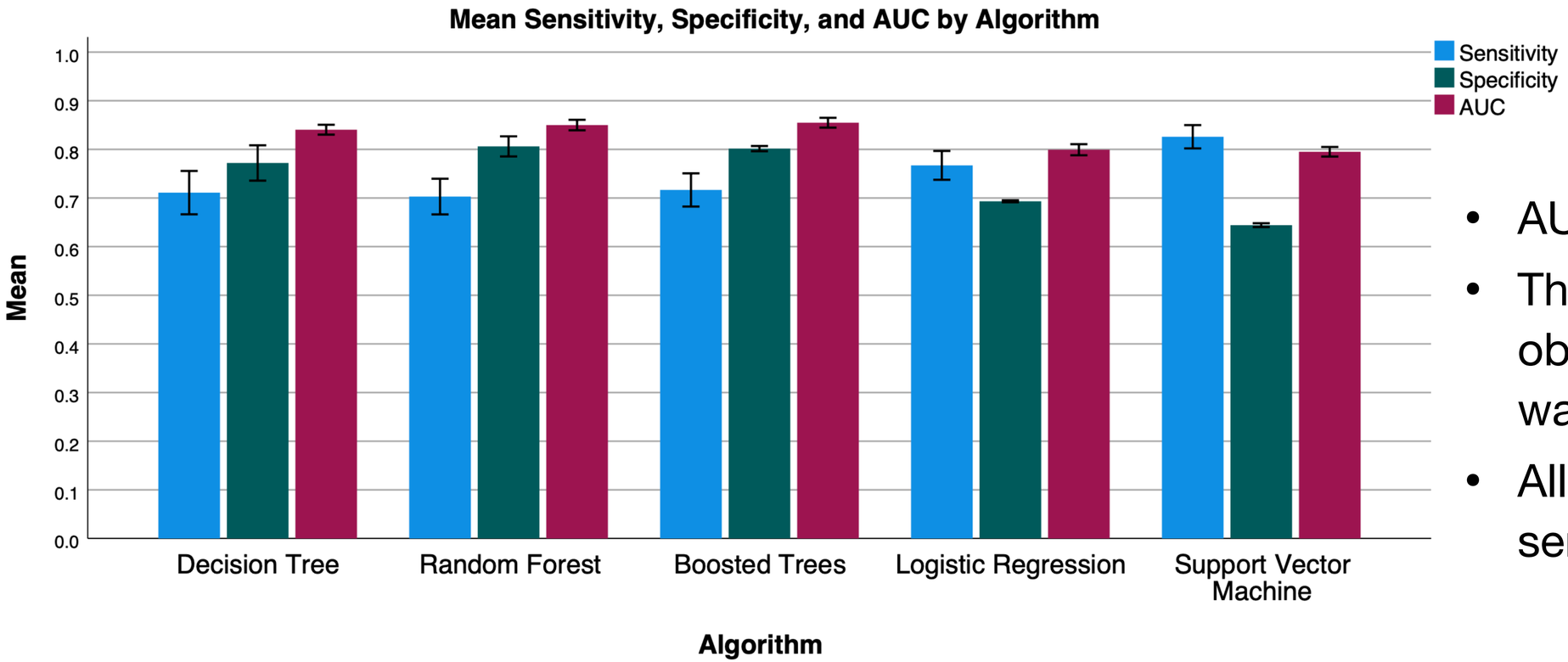
Trained using CreateML
(“Create ML,” n.d.)



App created using CoreML/SwiftUI
(“Core ML,” n.d.; “SwiftUI,” n.d.)

Results/Discussion

Model Performance

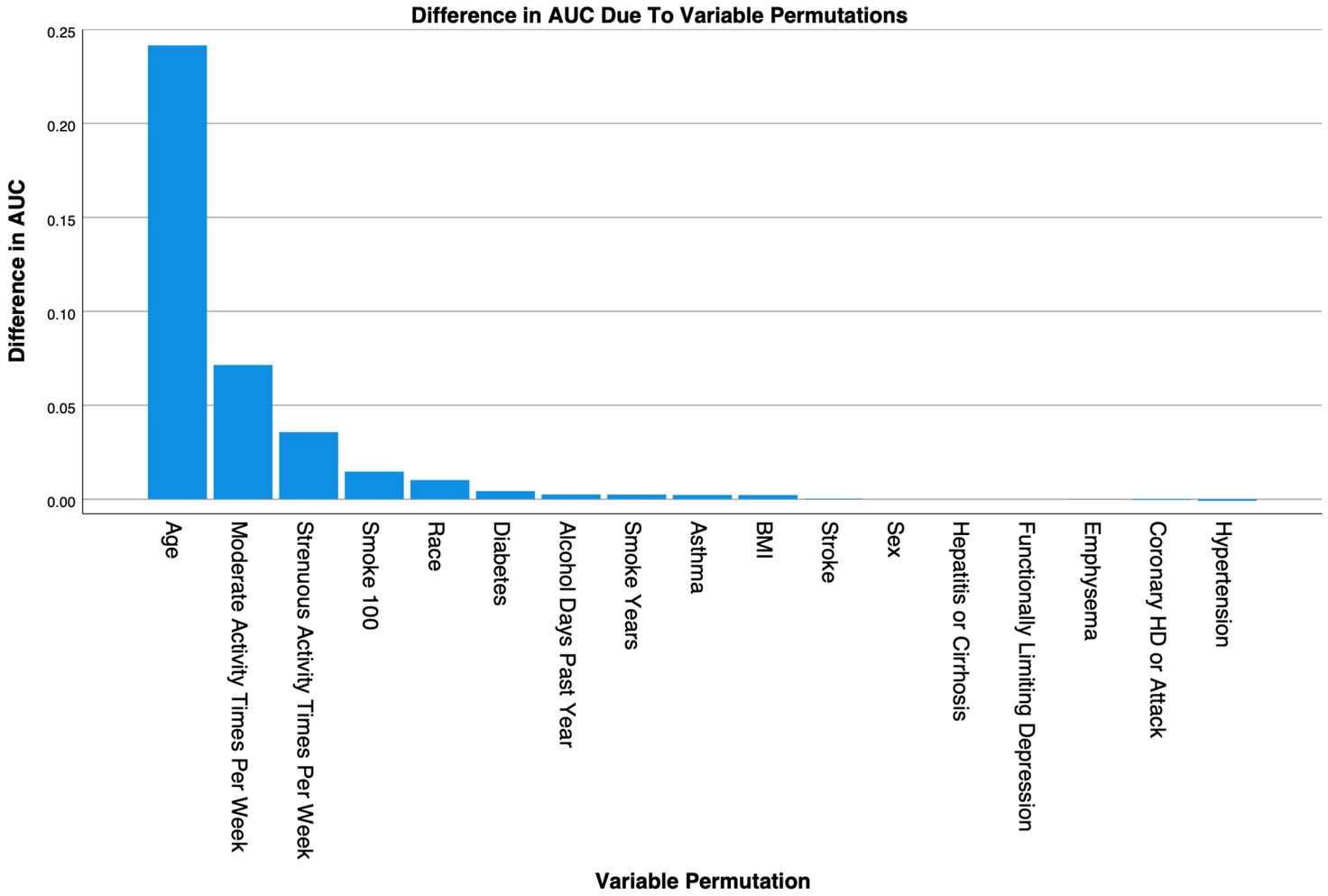


Data were obtained by evaluating the sensitivity, specificity, and AUC each model after training. Error bars represent the 95% confidence interval. n = 10 for all models.

- AUC of all models > ~0.8
- The highest-AUC model observed among the 10 runs was a BT, with AUC = 0.87
- All models varied in sensitivity and specificity

Results/Discussion

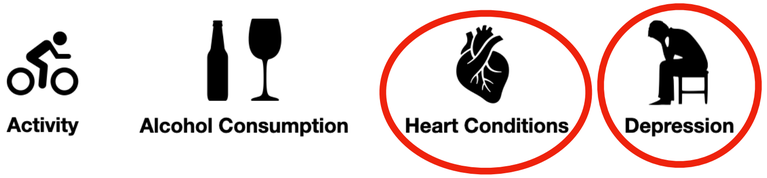
Variable Importance



- **Most Important:** Age, physical activity, smoking, race, and diabetes
- **Least Important:** Depression, emphysema, and heart conditions

Significant Risk Factors

Possible Risk Factors



Data were obtained by evaluating the difference in AUC between the final BT model and that from each variable permutation. The greater the difference in AUC, the greater importance that given variable has on predicting pancreatic cancer.

Comparison to Previous Pancreatic Cancer Models

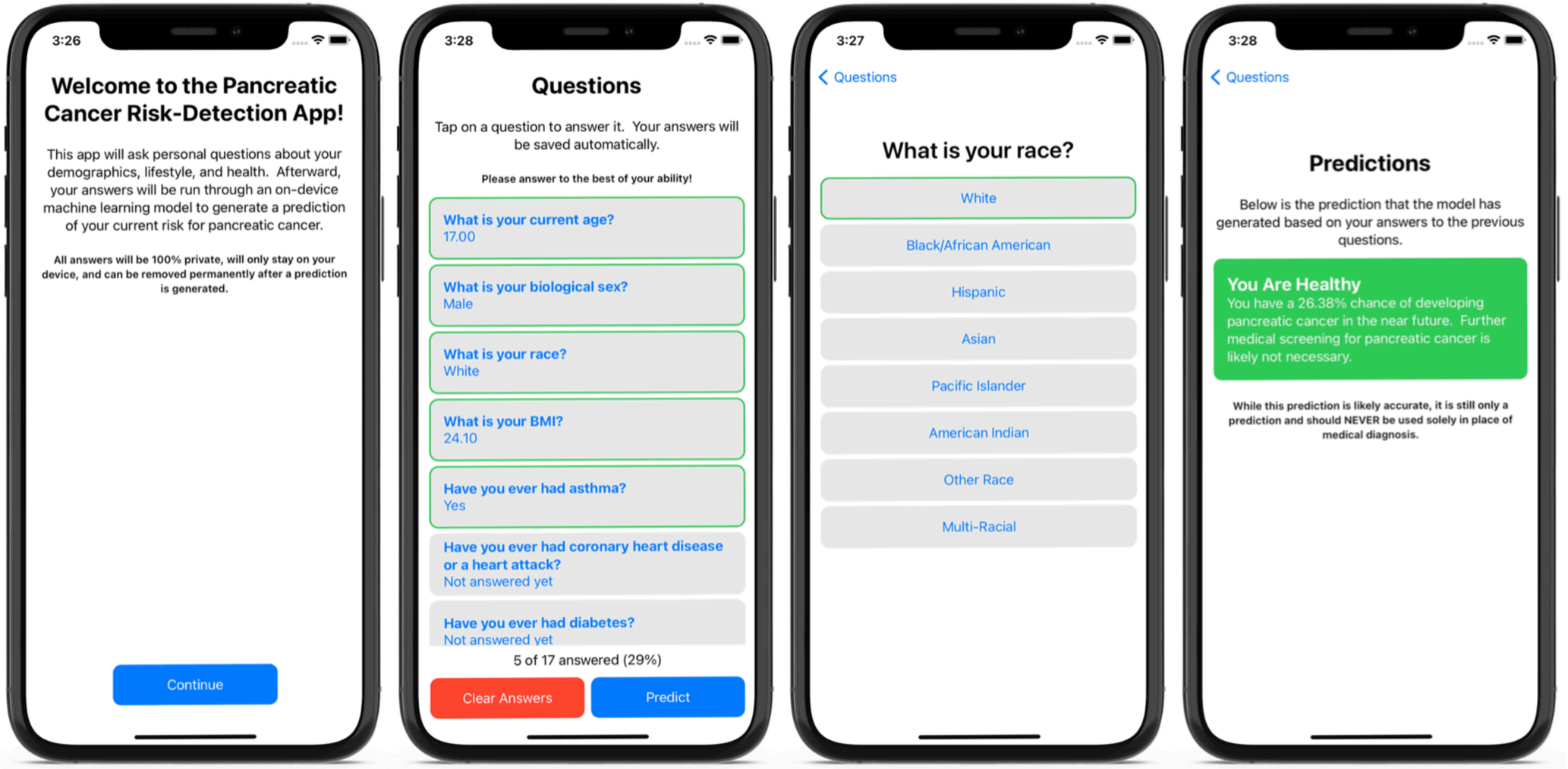
- The Boosted Trees model performed slightly worse than all previous pancreatic cancer models in terms of sensitivity and specificity, but better in terms of AUC

Comparison of the sensitivity, specificity, and AUC of the final BT model and that of previous pancreatic cancer models

Model	Sensitivity	Specificity	AUC
My BT	0.788	0.791	**0.870**
Muhammad et al., 2019 (ANN)	0.807	**0.807**	0.850
Hsieh et al., 2018 (ANN)	0.873	(not calculated)	0.642
Hsieh et al., 2018 (LR)	**0.998**	(not calculated)	0.707

Results/Discussion

iPhone App



Screenshots of the iPhone app created to facilitate user input of data and prediction output. The final BT model was embedded within the app. From left to right: welcome screen; question list; categorical answer view; final predictions.

Conclusions

- 5 algorithms are effective in using machine learning to discriminate between patients with and without pancreatic cancer. The tree-based models had the highest AUC.
- Age, physical activity, smoking, and race were the most influential risk factors
- Cannot evaluate the role of depression yet due to missing data and lack of information to determine a sudden onset
- An iOS application was created to facilitate data input and prediction output using the final BT model - potential use by doctors.

Future Research

- An Artificial Neural Network (ANN) should be developed as well
- Need more depression data and better criteria to distinguish between a sudden onset to further determine its importance
- Improvements to data upsampling (i.e., using SMOTE) (Brownlee, 2020)
- Improvements to the dataset - additional variables and data for more PC cases
- Once this is fully optimized, it could be an important tool assisting in the early diagnosis of pancreatic cancer by identifying appropriate candidates for further screening

Bibliography

- Aljuaid, T., & Sasi, S. (2016). Proper imputation techniques for missing values in data sets. *2016 International Conference on Data Science and Engineering (ICDSE)*. https://doi.org/10.1109/icdse.2016.7823957
- Arnold, L. D., Patel, A. V., Yan, Y., Jacobs, E. J., Thun, M. J., Calle, E. E., & Colditz, G. A. (2009). Are Racial Disparities in Pancreatic Cancer Explained by Smoking and Overweight/Obesity? *Cancer Epidemiology Biomarkers & Prevention, 18*(9), 2397–2405. https://doi.org/10.1158/1055-9965.epi-09-0080
- Lynn A. Blewett, Julia A. Rivera Drew, Miriam L. King and Kari C.W. Williams. *IPUMS Health Surveys: National Health Interview Survey, Version 6.4 [dataset]*. Minneapolis, MN: IPUMS, 2019. https://doi.org/10.18128/D070.V6.4
- Boyle, T. (2019, February 3). *Dealing with Imbalanced Data*. Medium. https://towardsdatascience.com/methods-for-dealing-with-imbalanced-data-5b761be45a18
- Brownlee, J. (2020, August 21). *SMOTE for Imbalanced Classification with Python*. Machine Learning Mastery. https://machinelearningmastery.com/smote-oversampling-for-imbalancedclassification/
- *Classification: ROC Curve and AUC. Machine Learning Crash Course*. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.
- *Core ML*. Apple Developer Documentation. https://developer.apple.com/documentation/coreml.
- Coughlin, S. S., Calle, E. E., Patel, A. V., & Thun, M. J. (2000). Predictors of pancreatic cancer mortality among a large cohort of United States adults. *Cancer Causes & Control, 11*, 915–923. https://doi.org/10.1023/a:1026580131793
- *Create ML*. Apple Developer Documentation. https://developer.apple.com/documentation/createml.
- Everhart, J., & Wright, D. (1995). Diabetes Mellitus as a Risk Factor for Pancreatic Cancer. *Jama, 273*(20), 1605–1609. https://doi.org/10.1001/jama.1995.03520440059037
- Gomez-Rubio, P., Zock, J.-P., Sharp, L., Hidalgo, M., Carrato, A., Ilzarbe, L., … Malats, N. (2015). Reduced risk of pancreatic cancer associated with asthma and nasal allergies. *Gut*. https://doi.org/10.1016/j.pan.2015.05.439
- Hassan, M. M., Li, D., El-Deeb, A. S., Wolff, R. A., Bondy, M. L., Davila, M., & Abbruzzese, J. L. (2008). Association Between Hepatitis B Virus and Pancreatic Cancer. *Journal of Clinical Oncology, 26*(28), 4557–4562. https://doi.org/10.1200/jco.2008.17.3526
- Hsieh, M. H., Sun, L.-M., Lin, C.-L., Hsieh, M.-J., Hsu, C.-Y., & Kao, C.-H. (2018). Development of a prediction model for pancreatic cancer in patients with type 2 diabetes using logistic regression and artificial neural network models. *Cancer Management and Research, 10*, 6317–6324. https://doi.org/10.2147/cmar.s180791
- Larsson, S. C., Orsini, N., & Wolk, A. (2007). Body mass index and pancreatic cancer risk: A metaanalysis of prospective studies. *International Journal of Cancer, 120*(9), 1993–1998. https://doi.org/10.1002/ijc.22535
- Lindgren, A. M., Nissinen, A. M., Tuomilehto, J. O., & Pukkala, E. (2005). Cancer pattern among hypertensive patients in North Karelia, Finland. *Journal of Human Hypertension, 19*(5), 373–379. https://doi.org/10.1038/sj.jhh.1001834
- Lowenfels, A. B., & Maisonneuve, P. (2006). Epidemiology and risk factors for pancreatic cancer. *Best Practice & Research Clinical Gastroenterology, 20*(2), 197–209. https://doi.org/10.1016/j.bpg.2005.10.001
- McGuigan, A., Kelly, P., Turkington, R. C., Jones, C., Coleman, H. G., & Mccain, R. S. (2018). Pancreatic cancer: A review of clinical diagnosis, epidemiology, treatment and outcomes. *World Journal of Gastroenterology, 24*(43), 4846–4861. https://doi.org/10.3748/wjg.v24.i43.4846
- Michaud, D. S., Giovannucci, E., Willett, W. C., Colditz, G. A., Stampfer, M. J., & Fuchs, C. S. (2001). Physical Activity, Obesity, Height, and the Risk of Pancreatic Cancer. *Jama, 286*(8), 921–929. https://doi.org/10.1001/jama.286.8.921
- Muhammad, W., Hart, G. R., Nartowt, B., Farrell, J. J., Johung, K., Liang, Y., & Deng, J. (2019). Pancreatic Cancer Prediction Through an Artificial Neural Network. *Frontiers in Artificial Intelligence, 2*. https://doi.org/10.3389/frai.2019.00002
- *Pancreatic Cancer Risk Factors*. American Cancer Society. https://www.cancer.org/cancer/pancreaticcancer/causes-risks-prevention/risk-factors.html.
- *PLCO*. The Cancer Data Access System. https://cdas.cancer.gov/plco/.
- Poruk, K. E., Firpo, M. A., Adler, D. G., & Mulvihill, S. J. (2013). Screening for Pancreatic Cancer. *Annals of Surgery, 257*(1), 17–26. https://doi.org/10.1097/sla.0b013e31825ffbfb
- Roffman, D. A., Hart, G. R., Leapman, M. S., Yu, J. B., Guo, F. L., Ali, I., & Deng, J. (2018). Development and Validation of a Multiparameterized Artificial Neural Network for Prostate 23 Cancer Risk Prediction and Stratification. *JCO Clinical Cancer Informatics, 2*, 1–10. https://doi.org/10.1200/cci.17.00119
- Roffman, D., Hart, G., Girardi, M., Ko, C. J., & Deng, J. (2018). Predicting non-melanoma skin cancer via a multi-parameterized artificial neural network. *Scientific Reports, 8*(1). https://doi.org/10.1038/s41598-018-19907-9
- Siegel, R. L., Miller, K. D., & Jemal, A. (2020). Cancer statistics, 2020. *CA: A Cancer Journal for Clinicians, 70*(1), 7–30. https://doi.org/10.3322/caac.21590
- *SwiftUI*. Apple Developer Documentation. https://developer.apple.com/documentation/swiftui/.
- Trevethan, R. (2017). Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. F*rontiers in Public Health, 5*. https://doi.org/10.3389/fpubh.2017.00307
- Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making, 19*. https://doi.org/10.1186/s12911-019-1004-8
- Ye, W., Lagergren, J., Weiderpass, E., Nyren, O., Adami, H.-O., & Ekbom, A. (2002). Alcohol abuse and the risk of pancreatic cancer. *Gut, 51*(2), 236–239. https://doi.org/10.1136/gut.51.2.236
- Zheng, W., Mclaughlin, J. K., Gridley, G., Bjelke, E., Schuman, L. M., Silverman, D. T., … Fraumeni, J. F. (1993). A cohort study of smoking, alcohol consumption, and dietary factors for pancreatic cancer (United States). *Cancer Causes & Control, 4*(5), 477–482. https://doi.org/10.1007/bf00050867