International Journal of Social Science and Economic Research

ISSN: 2455-8834

Volume:09, Issue:06 "June 2024"

Using Risk Factors for Disease to Predict Probability of Contracting a Disease in Young Adults: Application to a Machine Learning-based Product That Recommends Lifestyle Changes to Increase Health and Longevity

Taruna Agarwal, Jai Agarwal, Hrithik Jain, John Leddo, Dev Agarwal, Arjun Erasani, Yousef Malik, Jackson Dunnington, Adarsh Iruvanti, Shaurya Jain, Avyan Illapakurthy, Arnav Illapakurthy, Gabriela Duffy, Shirali Vasireddy, Rishabh Jha, Deniz Gurbuz, Avishi Anurag, Vishal Renjith, Harman Sabharwal, Aryan Sajjad, Arsal Sajjad, Dharana Cherukuwada, Abhinav Tewtia, Dhruv Shrivastava, Harini Natarajan, Ishwarya Ramineni, Sambhav Jain, Krish Tripathy, Mitra Manikandan,Veda Rangareddyvari, Saanvi Lamba, Prardhana Joy Thumma, Arian Miran, Collin O'Brien, Rithvik Dandemraju, Harrison Biear, Siddharth Mannepalli, Ryan Park, Raiyan Amiruddin, Jahnavi Sajja, Julie Yu, Kareem Abdelwahab, Ashna Melanathuru, Gabriel Chavarria, Neil Saboo, Ehaan Miran, Ananth Namilae and Sophia Nasibdar

METY Technology, Inc.

DOI: 10.46609/IJSSER.2024.v09i06.026 URL: https://doi.org/10.46609/IJSSER.2024.v09i06.026

Received: 27 June 2024 / Accepted: 8 July 2024 / Published: 15 July 2024

ABSTRACT

In previous papers, we have described My Youthspan, a software program that takes the latest scientific research in wellness and longevity and uses data science and machine learning to make personalized recommendations for people to live longer and be healthier. My Youthspan is targeted towards adults 40 and older, many of whom have legitimate concerns about the risk of contracting serious age-related diseases. Currently, we are developing a similar product for young adults, aged 18-40. While most young adults do not face serious diseases such as cancer or cardiovascular disease, the United States Centers for Disease Control reports that 54% of young adults have one or more chronic health conditions. While these chronic health conditions, such as obesity or high blood pressure, generally are not immediately life-threatening, they tend to be risk factors for more serious conditions later on such as stroke or heart disease. The goal of the present paper is to describe a methodology for quantifying these risk factors and then showing how they can be used in evaluating how lifestyle changes may reduce the risk in young adults of later contracting major age-related disease.

ISSN: 2455-8834

Volume:09, Issue:06 "June 2024"

Introduction

In previous papers, we have discussed our ongoing research and efforts to create a software product, called My Youthspan, that takes the latest research on wellness and longevity and uses data science and machine learning to make personalized recommendations on what people can do to live longer and healthier (Jain et al., 2023; Lu et al., 2023). The idea behind the machine learning component is that an initial model will be built based on a review of hundreds of scientific results, which will be combined and be able to make personalized recommendations based on a person's characteristics such as age, gender, individual and family medical history, etc. and lifestyle habits (e.g., what they eat, whether and how much they exercise, how much sleep they get, whether and how much they smoke or drink alcohol). One feature of that software is a daily log that users fill out to record the health interventions they do in terms of diet, exercise nutritional supplements, stress management, and sleep and updates to their health. The software uses this information to make ongoing predictions of user life expectancy and risk of contracting major diseases such as cancer or cardiovascular disease. The initial My Youthspan product is targeted to people 40 and over, since these tend to be the ones most susceptible to age-related diseases and other consequences of aging.

While My Youthspan is targeted to older adults, the United States' Centers for Disease Control (CDC) reports that 54% of young adults, aged 18-35, have one or more chronic health conditions themselves. The types of chronic health conditions that afflict young adults, e.g., obesity, high blood pressure, high cholesterol, also afflict older adults. Often, the reverse is not true. Young adults generally do not suffer from diseases like cancer, heart disease, stroke that older adults experience and that may be life-threatening. The good news with many of the chronic conditions that young people experience is that these tend not to be immediately life threatening and may not even substantially impact a young adult's quality of life. Indeed, the chronic health condition may not even be psychologically meaningful. If a young person learns that his or her total cholesterol is 220 instead of below 200, what does that mean to him/her? Most likely, the young person feels healthy and vital and has the energy to do the necessary things in life without having to make accommodations for this health condition. Therefore, it may be hard to convince the young adult to do something about this high cholesterol, either taking a drug (which may have side effects and cost money) or changing his or her lifestyle, such as giving up foods s/he likes.

The bad news with many of the most common chronic conditions that young people experience is that they are often risk factors for the more serious, life-threatening conditions that people develop as they get older. A person with high cholesterol as a young adult is more likely to get a heart attack later in life if the high cholesterol is left untreated. A person who has high blood pressure as a young adult is more likely to suffer a stroke.

ISSN: 2455-8834

Volume:09, Issue:06 "June 2024"

Currently, we are creating a version of our My Youthspan software that is targeted to young adults, aged 18-40. As with our regular My Youthspan product, we want to help young adults live longer and healthier lives, which includes avoiding major age-related diseases like cancer or heart disease. As with our regular My Youthspan product, the product for young adults takes the latest scientific research on wellness and longevity and uses data science and machine learning to make personalized recommendations to people on what lifestyle changes they can make in areas such as diet, exercise, taking nutritional supplements, managing stress and healthy sleep so that they can live longer and healthier. The software also provides a daily log, so that users can track their lifestyle activities and health outcomes. These activity and health outcome data are used to refine the machine learning models that make the recommendations.

However, this tracking approach potentially presents a problem. Developing heart disease or cancer is typically not an instantaneous outcome but one that evolves over a period of years and often unbeknownst to the person who is developing the condition. Therefore, when tracking health outcomes, it is expected that on a week-to-week or month-to-month basis, young adults would be listing "none" when asked about whether they have heart disease or cancer, regardless of what lifestyle choices they are making.

This can potentially distort a machine learning model seeking to learn the relationship between lifestyle choices and health outcomes. If the input is always "none" with regard to a chronic or serious health outcome, the model then learns that lifestyle choices do not impact risk of disease. This is not only misleading but potentially dangerous as people may conclude that their lifestyle choices are healthy when they are not.

One way to address this problem is to measure predictors of disease that are both measurable and are responsive, within reasonable timeframes, to lifestyle choices. If a quantitative relationship can be established between the risk factor and the disease, then the software can provide updated estimates of risk of disease that are based on changes in the observed risk factors.

Using disease indicators offers some potential advantages:

1. Enhanced predictive accuracy. Disease indicators, such as biomarkers and physiological measurements, can offer a more granular view of an individual's health status than just measuring whether or not a person has a disease. This finer level of detail can lead to improved predictive accuracy (Bhadra et al., 2019).

2. Early disease detection. Using disease indicators can allow models to detect individuals at risk before actual symptoms begin to show. This can enable earlier interventions and preventative measures (Chen et al., 2020).

ISSN: 2455-8834

Volume:09, Issue:06 "June 2024"

3. Personalized risk assessment. Disease indicators enable the customization of risk assessments, allowing for more personalized healthcare recommendations and interventions (Hosny et al., 2018).

In deciding to use indicators as a way to predict diseases, there are some challenges that need to be taken into consideration. These include:

1. Data availability and quality. The availability and quality of disease indicator data can pose significant challenges. Ensuring access to comprehensive and reliable datasets is crucial (Drew et al., 2019).

2. Interpretability. Models incorporating disease indicators may be less interpretable, making it challenging for healthcare professionals to understand the rationale behind risk predictions (Bahrampour et al., 2019).

3. Privacy Concerns. The use of sensitive health data raises privacy and ethical concerns. Appropriate data anonymization and protection measures must be in place (Chen et al., 2020).

4. Bias and Generalizability. Models trained on indicator data may be prone to bias if the data is not representative. Ensuring model generalizability across diverse populations is essential (Rajkomar et al., 2019).

Using disease indicators has shown promise in the field of medical machine learning. For example, Hosny et al. (2018) developed a machine learning model that utilized disease indicators, including genetic markers and early symptoms, to predict the risk of cardiovascular disease. Their model achieved superior performance in early detection compared to traditional risk assessment methods. Additionally, Drew et al. (2019) explored the integration of electronic health records and imaging data as disease indicators to predict the risk of specific cancers. They demonstrated that combining these diverse data sources resulted in more accurate risk assessments. Chen et al. (2020) addressed the privacy concerns associated with disease indicator data by proposing federated learning techniques that allow models to be trained across multiple healthcare institutions without sharing patient-specific data. The remainder of the present paper is devoted to showing how disease indicators will be used in the young adult longevity and wellness software being developed by METY Technology.

Using Disease Indicators to Predict Risk of Contracting Diseases

There are two major steps for incorporating risk factors into a predictive model for estimating the risk of contracting diseases. The first is to identify the relevant predictors of diseases. Generally,

ISSN: 2455-8834

Volume:09, Issue:06 "June 2024"

this can be done from a review of the literature, and, fortunately, there tends to be a consensus on what these risk factors are. The table below shows several major diseases and five known major risk factors for each.

All cause mortality

Table 1: Risk Factors Associated with All Cause Mortality

Risk Factor	Relative Risk Ratio
Chronic Diseases	1.75
Smoking, < 10 cigarettes/day	1.3
Smoking, 10-19 cigarettes/day	1.8
Smoking 20-30 cigarettes/day	2.09
Smoking >= 40 cigarettes/day	2.78
BMI	1.29 for each 5 kg/m^2 increase above 25

Cancer (all forms)

Table 2: Risk Factors Associated with Cancer

Risk Factor	Relative Risk Ratio
Smoking, < 10 cigarettes/day	1.2
Smoking, 1-20 cigarettes/day	1.47
Smoking, 21-30 cigarettes/day	6.88
Smoking, >30 cigarettes/day	7.52
Alcohol, males	1.1 if consume 2 drinks/day
Alcohol, females	1.1 if consume 1 drink/day
Infectious diseases	1.13
BMI	1.07 for each 4.78 kg/m^2 increase above 25
Physical inactivity	ovarian cancer, 1.29
	colon cancer, 1.25
	endometrial cancer, 1.29
	breast cancer, 1.08
	prostate cancer, 1.08
	rectal cancer, 1.07

ISSN: 2455-8834

Volume:09, Issue:06 "June 2024"

Depression

Table 3: Risk Factors Associated with Depression

Risk Factor	Relative Risk Ratio
Substance/drug abuse	1.5
Smoking	1.02 for each cigarette smoked per day
Alcohol, moderate	1.16
Alcohol, heavy	1.22

Diabetes

Table 4: Risk Factors Associated with Diabetes

Risk Factor	Relative Risk Ratio
BMI	1.015 for every 1 kg/m^2 above 25
Physical inactvity	1.52 for people with BMI 18.5-24.9
	1.65 for people with BMI >= 25
Smoking, <20 cigarettes/day	1.25
Smoking, 20-25 cigarettes/day	1.61
Smoking > 25 cigarettes/day	1.94
Alcohol, heavy drinker	2.18

Stroke

Table 5: Risk Factors Associated with Stroke

Risk Factor	Relative Risk Ratio
High blood pressure, systolic	2 for every 20 mmHg above 115
High blood pressure, diastolic	2 for every 10 mmHg above 75
Smoking	1.12 for every 5 cigarettes/day
Cholesterol, 5-5.9 mmol/L	1.05
Cholesterol, 6-6.9 mmol/L	1.16
Cholesterol, >=7 mmol/L	1.22

ISSN: 2455-8834

Volume:09, Issue:06 "June 2024"

Cardiovascular Disease

Table 6: Risk Factors Associated with Cardiovascular Disease

Risk Factor	Relative Risk Ratio
High blood pressure, systolic	1.05 for each 10 mmHg above 120
High blood pressure, diastolic	1.04 for each 5 mmHg above 80
Cholesterol, total, male, below 120 mg/dL	0.9
Cholesterol, total, male, 120-159 mg/dL	1.17
Cholesterol, total, male, 160-199 mg/dL	1
Cholesterol, total, male, 200-239 mg/dL	0.833
Cholesterol, total, male, 240-279 mg/dL	1.2
Cholesterol, total, male, >=280 mg/dL	1.5
Cholesterol, total, female, below 120 mg/dL	0.9
Cholesterol, total, female, 120-159 mg/dL	1
Cholesterol, total, female, 160-199 mg/dL	1.25
Cholesterol, total, female, 200-239 mg/dL	1.25
Cholesterol, total, female, 240-279 mg/dL	1
Cholesterol, total, female, >=280 mg/dL	1.25
Smoking, 1-3 cigarettes/day	1.04
Smoking, 4-6 cigarettes/day	0.96
Smoking, 7-9 cigarettes/day	1.35
Smoking, 10-14 cigarettes/day	1.42
Smoking, 15-24 cigarettes/day	1.7
Smoking, >=25 cigarettes/day	2.12

Inspection of the above tables lead to some relevant observations when it comes to incorporating risk factors into a predictive model of diseases. First, some risk factors, such as high blood pressure or high cholesterol, overlap multiple diseases. This makes these important ones to collect data on, since these data can be used across different predictive models. Second, some screening is necessary as some risk factors are hard to quantify or operationally define. For example, “chronic diseases” is a risk factor for all-cause mortality and “infectious diseases” is a risk factor for cancer. However, “chronic diseases” and “infectious diseases” are hard to define in a way that can be readily measured. What may be a better approach in cases like these is to look at individual chronic or infectious diseases and their relationships to all-cause mortality and cancer.

Once the risk factors have been selected and appropriately defined so that they can be measured, the next step is to quantify the relationship between the risk factor and disease. In some cases, there may be a straightforward formula that can accept the quantified risk factor and output the

ISSN: 2455-8834

Volume:09, Issue:06 "June 2024"

risk of the disease. One example of this is the Framingham Risk Score that predicts the 10-year risk of coronary heart disease from a person's systolic blood pressure. Here, the formula is risk = $1 - (0.987^{(SBP-120)})$, where SBP stands for systolic blood pressure (Whelton et al., 2018) .

In other cases, different levels of the risk factor get lumped into categories and then the relationship between risk factor categories and disease is quantified. An example of this is shown in Wilson et al. (1998). Wilson et al. also seek to quantify the risk of coronary heart disease using blood pressure as a risk factor. Here, Wilson et al. developed four categories of systolic blood pressure: normal, high normal, hypertension stage 1, and hypertension stage 2-4. People were categorized based on their systolic blood pressures, with <130 considered normal, 130-139 considered high normal, 140-159 considered hypertension 1 and above 160 considered hypertension 2-4. Instead of calculating absolute risk of coronary heart disease, Wilson et al. calculate relative risk ratios of getting coronary heart disease with those in the normal category considered the reference point. Accordingly, the relative risk ratios for different levels of blood pressure are 1 for normal, 1.32 for high normal, 1.73 for hypertension 1 and 1.92 for hypertension 2-4.

Conclusion

The process of using risk factors to build a machine learning model to estimate the likelihood of contracting a disease involves several steps. First, the literature is reviewed to establish what risk factors are most important to/diagnostic of the diseases of interest. Next, the risk factors need to be evaluated to ensure that they are sufficiently operationalized so that quantifiable data regarding them can be measured. This step may involve some operationalization on the part of the researchers.

Once the risk factors have been selected and quantified, there needs to be a method for relating the data measured from the risk factor to a prediction of the likelihood of disease. In an idealized case, a formula can be constructed, similar to the Framingham Risk Score, that uses the actual measurement to quantify the risk. In cases where this is not possible, categorical data can be used instead, whereby the raw data is assigned to a category and then the category is associated with a prediction of likelihood of occurrence of the disease.

Once this model is developed, the machine learning model is trained on data that pairs interventions (such as diet, exercise, etc.) and the associated changes in risk factor levels (e.g., blood pressure) and these changes in risk factor levels are used to stand in for likelihood of contracting a particular disease. In doing so, changes in health outcomes can be recorded on a regular basis, thus providing dynamic predictions of risk of disease as a result of the daily

ISSN: 2455-8834

Volume:09, Issue:06 "June 2024"

interventions a person is using. This enables people to fine tune their lifestyle choices, and, in doing so, optimize their health and longevity.

References

Bahrampour, S., Ramirez, L., Azimi, J., & Davidson, N. (2019). Interpretability in Machine Learning: An Overview of Transparency and Explainability in AI. arXiv preprint arXiv:1903.03894.

Bhadra, A., Saha, S., & Singh, D. (2019). Prediction of Type 2 Diabetes using Machine Learning Algorithms. Journal of Health and Medical Informatics, 10(1), 1-9.

Chen, M., Zhou, X., He, T., & Huang, Z. (2020). Federated Learning in Healthcare: A Review and Case Studies. arXiv preprint arXiv:2007.07835.

Drew, B. J., & Reid, C. L. (2019). Early Detection of Cancer: Evaluation of a Machine Learning Model using Clinical Notes in Electronic Health Records. Journal of Oncology Practice, 15(6), e531-e538.

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. W. L. (2018). Artificial Intelligence in Radiology. Nature Reviews Cancer, 18(8), 500-510.

Jain, H., Redrouthu, S., Agarwal, J., Agarwal, T., Leddo, J. et al. (2023). A Machine Learning-based Lifespan Calculator. International Journal of Social Science and Economic Research, 8(7), 2102-2108.

Lu, T., Yuan, Y., Agarwal, J., Agarwal, T., Jain, H., Leddo, J. et al. (2023). A Meta-regression and Bayesian Regression Framework for Combining Results of Scientific Research and Surveys of People's Lifestyles to Make Recommendations on What Interventions Will Help Them Live Longer and Healthier. International Journal of Social Science and Economic Research, 8(3), 524-531.

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2019). Ensuring Fairness in Machine Learning to Advance Health Equity. Annals of Internal Medicine, 170(10), 681-682.

Whelton, P. K., Carey, R. M., Aronow, W. S., et al. (2018). 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Hypertension, 71(6), e13-e115.

ISSN: 2455-8834

Volume:09, Issue:06 "June 2024"

Wilson, P.W., D'Agostino, R.B., Levy, D., Belanger, A.M., Silbershatz, H. & Kannel, W.B. (1998). Prediction of coronary heart disease using risk factor categories. Circulation. 97(18):1837-1847.

