International Journal of Social Science and Economic Research

ISSN: 2455-8834

Volume:08, Issue:09 "September 2023"

USING RISK FACTORS FOR DISEASE TO PREDICT PROBABILITY OF CONTRACTING A DISEASE IN A MACHINE LEARNING-BASED PRODUCT THAT RECOMMENDS INTERVENTIONS TO INCREASE HEALTH AND LONGEVITY

Jai Agarwal, Hrithik Jain, Taruna Agarwal, John Leddo, Divya Kulkarni, Sreeja Sambhuni, Ryan Park, Sanaa Karkera, Tasbih Bakshi, Varun Dasari, Navya Tuteja, Aadiv Aitharaju, Anish Reddy, Rohan Penmetsa, Aleena Ahmad, Michelle Kim, Arian Miran, Mason Elkas, Pavana Druthi Sudigali, Pavana Kruthi Sudigali, Rithvik Dandemraju, Vyasa Hari, Rohan Matta, Helena Gabrial, Varsha Mupparaju, Hardik Ramagiri, Aneesh Sreedhara, Pooja Somayajula, Amulya Gottipati, Eshwar Dokku, Aryan Dotiwalla, Shoumik Bisoi, Vihaan Cherukuri, Atharva Kulkarni, Arjun Erasani, Dheenav Rallabandi, Ranveer Kataria, Nithya Chilukuri, Prardhana Joy Thumma, Nikhit Rachapudi, Vivaan Sharma, David Sheng, Leon Fu, Kinnari Chaubal, Sid Vijay, Ashwin Tripathy, Jahnavi Guduru, Luke Richardson, Sudhit Sangela, Anaya Dandemraju, Aditya Nagar, Harini Puchakayala, Deepika Ravi, Celeste Cynkin, Sathvik Redrouthu, Smaran Pasupulati, Riya Pasupulati, Ritika Bishen, Sai Varun Konagalla, Vaishnavi Alapati, Vaishali Alapati, Krish Malik, Vikram Rudraraju, Saanvi Lamba, Ishwarya Ramineni, Aniketh Malipeddi, Yudhir Vasam, Aditya Devireddy, Sanah Shah, Vishal Renjith, Vaisnavi Malipeddi, Adan Eftekhari, Tanya Singhal, Kunal Singhal, Venkat Penmetsa, Talbot Biles, Ehaan Miran, Sushen Kilaru, Ayaan Sarkar, Sophia Nasibdar, Dashiell Michetti, Riya Srikumar, Medha Unnava, Collin O'Brien, Krisha Dotiwalla, Avilash Angirekula, Shaan Kosuru and Maruti Pariti

MyEdMaster, LLC

DOI: 10.46609/IJSSER.2023.v08i09.041 URL: https://doi.org/10.46609/IJSSER.2023.v08i09.041

Received: 10 September 2023 / Accepted: 23 September 2023 / Published: 4 October 2023

ABSTRACT

In previous papers, we have described a methodology of using meta-regression and Bayesian statistics to create a machine learning model that combines scientific research on wellness and longevity and responses from people's lifestyle questionnaires to make recommendations on what people can do to live longer and healthier. One of the goals of this software is to use data collected from logs of people's lifestyle choices to update the model and increasingly improve and personalize the recommendations made. One challenge faced here is that the elapsed time between lifestyle choices and the onset of disease may take years, making it difficult to make timely recommendations based on the effectiveness of what people are implementing in their daily lives. One way to address this challenge is to measure risk factors for disease rather than occurrence of disease itself as properly-selected risk factors may be more sensitive to changes in lifestyle choices while still being highly predictive of the risk of contracting major diseases. The present paper provides a methodology of using such risk factors in a machine learning model to recommend interventions for what people can do to live longer and healthier lives.

Introduction

Our project team has been working on a software program that takes the latest scientific research in wellness and longevity and people's responses to lifestyle surveys and uses machine learning to make recommendations to people on what they can do to live longer and be 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). As people use the software, they will enter their ongoing lifestyle habits and updated health outcomes, which the machine learning model will use to optimize its recommendations to the user based on an analysis of what works best for that person.

In a previous paper (Lu et al., 2023), we proposed a combination of meta-regression (van Houwelingen et al., 2002) and Bayesian statistics (Bishop & Tipping, 2003) to process the scientific research data and survey responses. In the first pass of this analysis, meta-regression was proposed as a way to combine the results of single interventions such as eating a particular type of food (e.g., vegetables) on a specific outcome (e.g., cardiovascular disease). The use of meta-regression was motivated by the fact that different research papers involve different sample sizes, project lengths, participant age ranges, ethnicities and genders, and report different outcomes (e.g., different degrees of risk reduction for the disease being investigated). Meta-regression was seen as a way to combine these into an overall predictive model that could estimate the reduction in risk of a particular disease that would result if a person incorporated a given level of the intervention into his or her lifestyle. Using meta-regression, a different predictive model would be created for each independent variable and each health outcome.

A major limitation of using meta-regression alone as the analytical method is that people are likely to combine different interventions in order to optimize lifespan and health outcomes. Someone who is committed to optimal health is likely to be interested in what foods to eat, what nutritional supplements to take, what types of exercises to perform, etc. This presents a challenge as most scientific research tends to focus on individual independent variables for the sake of experimental control. Researchers typically want to know what the effect of a specific independent variable is on a dependent variable and, therefore, introduce only that independent variable into the study and keep the other variables constant across conditions so as to increase the certainty that any change in the dependent variable can be attributed to the presence of the independent variable and not some confounding variable. For example, a scientist interested in studying the effects of a particular diet on weight loss is likely to hold exercise constant because if participants are given a program involving both diet and exercise, it becomes unclear whether the weight loss was due to diet or exercise or some combination/interaction of both.

People, on the other hand, who are seeking to increase their lifespans and improve their health are going to worry less about experimental control and more about increasing their outcomes. As a result, they are likely to want to try multiple interventions in order to reap the benefits of each. The challenge here is that there is less known about how to quantify the benefits of combinations of interventions than single interventions. For example, if one has a very good diet, how many supplements does the person really need to take?

In order to address this question, we conducted surveys of people, asking them about the different combinations of interventions they use and questions about their health (Lu et al., 2023). The goal was to collect data that would enable us to model how combinations of interventions impacted health. We proposed that Bayesian analysis could be used to integrate the questionnaire data with the meta-regression analysis of the research literature results. In this analysis, the meta-regression results would serve as the priors for the Bayesian models and the survey results would then be used to update these models.

Our ability to perform this integrated analysis is based on the premise that we have actual health outcome data. For example, if we want to predict the effects of exercise on the risk of heart disease, we need to have data on both a person's exercise program and whether they have or get heart disease. When conducting an analysis of research papers or questionnaire responses, this is not a problem as these papers or responses contain data on whether the people being reported on have heart disease.

However, the goal of our software is to continue to track people as they implement the recommendations the software provides and determine whether those recommendations are working by recording any changes in people's health outcomes. This 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, people 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 and 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)/

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 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, 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.

Table 1: Risk Factors Associated with Major Diseases

Heart disease	Stroke	Cancer	Diabetes	Depression	Arthritis
High blood pressure	High Blood Pressure	Tobacco Use	Family History	Genetic Predisposition	Age
High cholesterol	Smoking	Exposure to Ultraviolet (UV) Radiation	Excess Body Weight	Traumatic Life Events	Gender
Diabetes	Atrial Fibrillation	Unhealthy Diet	Unhealthy Diet	Brain Chemistry and Imbalances	Genetics
Smoking	Diabetes	Physical Inactivity and Obesity	Physical Inactivity	Chronic Illness or Health Conditions	Joint Injuries or Trauma
Unhealthy lifestyle	High Cholesterol	Exposure to Carcinogens	Age	Substance Abuse	Autoimmune Conditions

Inspection of the above table leads 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, "unhealthy diet" is a risk factor for both cancer and diabetes. However, "unhealthy diet" is hard to define in a way that can be readily measured. Moreover, diet itself can impact other risk factors such as cholesterol levels, which themselves are risk factors. In such cases, risk factors such as "unhealthy diet" need to be further defined. What may be a better approach in cases like these, where diet itself is a composite of all foods eaten, is to look at the component items that make up diet such as fruit and vegetable consumption and look at the relationship between those and diseases.

Another example is when a disease itself is a risk factor for another disease as is the case with diabetes, which is a risk factor for stroke and heart disease. In such cases, it may not be possible to measure the onset of diabetes in real time. Instead, one may have to quantify the risk of diabetes and once that risk is quantified, see how that risk impacts the risk of heat disease or stroke.

A third example is where a risk factor may be stated in overly broad terms, such as "unhealthy lifestyle," a risk factor for heart disease. In such cases, the term "unhealthy lifestyle" needs to be operationally defined so as to lend itself to be measurable and quantifiable when it comes to predicting diseases.

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 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 health outcomes can be recorded on a regular basis, thus providing dynamic predictions ofrisk of disease as a result of the daily 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.

Bishop,C. M. &Tipping,M.E.(2003).Bayesian regression and classification. *Nato Science Series sub Series III Computer And Systems Sciences*,190:267-288.

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.

van Houwelingen,H.C., Arends,L.R. &Stijnen,T.(2002). Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine*,21(4):589-624. doi:10.1002/sim.1040

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.

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.