Miran *et al. BMC Res Notes (2021) 14:184*
https://doi.org/10.1186/s13104-021-05596-7

BMC Research Notes

RESEARCH NOTE

Open Access

A model-agnostic approach for understanding heart failure risk factors

Seyed M. Miran*, Stuart J. Nelson and Qing Zeng-Treitler

Abstract

Objective: Understanding the risk factors for developing heart failure among patients with type 2 diabetes can contribute to preventing deterioration of quality of life for those persons. Electronic health records (EHR) provide an opportunity to use sophisticated machine learning models to understand and compare the effect of different risk factors for developing HF. As the complexity of the model increases, however, the transparency of the model often decreases. To interpret the results, we aimed to develop a model-agnostic approach to shed light on complex models and interpret the effect of features on developing heart failure. Using the HealthFacts EHR database of the Cerner EHR, we extracted the records of 723 patients with at least 6 yeas of follow up of type 2 diabetes, of whom 134 developed heart failure. Using age and comorbidities as features and heart failure as the outcome, we trained logistic regression, random forest, XGBoost, neural network, and then applied our proposed approach to rank the effect of each factor on developing heart failure.

Results: Compared to the "importance score" built-in function of XGBoost, our proposed approach was more accurate in ranking the effect of the different risk factors on developing heart failure.

Keywords: Explainable AI, Model-agnostic approach, Heart failure

Introduction

Heart Failure (HF) is a serious problem for public health and the economy in the United States[1–3]. Coronary artery disease and hypertension significantly increase the risk of developing HF [4, 5]. Results from the Framingham Heart Study showed that the risk of developing HF among patients with hypertension is up to three times more than that of normotensive people [6]. atrial fibrillation [7], chronic obstructive pulmonary disease (COPD) [8, 9], chronic kidney disease (CKD) [10], anemia [12], asthma [11, 13], arthritis [14], depression [15], and cancer [16] have also been identified to correlate with HF.

Many research studies have compared the effect of different comorbidities on developing HF [17]. The results of these studies, however, are not always consistent. Coronary heart disease, diabetes, and hypertension have been identified as the most important risk factors [18]. Levy and his colleagues [6] monitored 5,143 people for 20.1 years and reported that hypertension was the biggest risk factor for developing HF. In a separate study, the authors found that coronary heart disease had the biggest effect on developing HF [19]. Diabetes and HF, which are often called "twin epidemics", are highly correlated [20]. Comparing the effect of the above-mentioned comorbidities on the risk of developing HF among patients with type 2 diabetes is thus of particular interest for prevention and treatment purposes.

Electronic Health Records (EHR) are increasingly employed by clinical researchers. Using EHR data, machine learning algorithms can help to understand the effect of different comorbidities on the clinical outcome [21, 22]. In the current study, we used HealthFacts, an EHR-based database developed and maintained by Cerner Corporation, to understand and

*Correspondence: miran@gwu.edu
Biomedical Informatics Center, School of Medicine and Health Sciences, George Washington University, Washington, DC, USA

interpret the effect of age and different well-known comorbidities on developing HF among patients with type 2 diabetes.

Different modeling methods, such as logistic regression, or machine learning methods, such as random forest, XGBoost, and neural networks, can be used for our binary classification task of developing HF. Given the increasing amount of EHR data, more sophisticated models may improve the assessment of the risk of a clinical outcome. However, the most powerful predictive models, such as the ensemble model and deep neural network, are not interpretable [23]. Identifying the most significant factors for the risk of HF among patients with diabetes requires improvement of the global interpretability of machine learning algorithms.

Different researchers from different fields of study have sought to understand the effect of individual features on an outcome when the underlying relationship is nonlinear and nonmonotonic [24, 25]. One approach is to perturb the values of a given feature with random numbers while keeping the other features fixed in order to investigate the relationship between the change of the response and to calculate the "situational importance" of each feature [26, 27]. The main difficulty with this approach are that many randomly generated values do not happen in the real world, and that the approach does not consider the underlying dependent relationship between variables.

The Shapley value, based on a study by Shapley decades ago [28], has been used to investigate the global effect of each individual feature on the response. For a feature, its Shapley value is the average marginal contribution of that feature across all possible combinations of all features. Although the Shapley value can help rank the importance of the features, it is computationally very expensive especially when the number of features is large. Like most of the permutation-based algorithms, its other drawback is that it may include unrealistic instances in the computation in cases where the features are correlated.

Another model-agnostic approach to compare and understand the global effect of features on the response is through the use of a "global surrogate." The idea behind this is to train an interpretable model (such as a linear model) on the features and their predictions from the black-box model to approximate the underlying model. Local interpretable model-agnostic explanations (LIME) is a method that is capable of providing both local and global interpretability using this approach [29].

Although all the methods mentioned above can shed light on a black-box model, all of them have some degree of randomness and do not necessarily lead to the same result when repeated for the same underlying model. The LIME and Shapley value methods are very computationally expensive. Developing an approach that is deterministic in results with less computational burden can make a significant contribution to the field.

Main text

Methods

Data procurement and preparation

In this research study, we used the Health Facts database, which has more than 400 million encounters from 689 different hospitals. We first identified patients with type 2 diabetes using ICD-9 and ICD-10 codes, while requiring at least one encounter without the diabetes ICD codes before the first diabetes visit. The date of the first encounter with the ICD diabetes codes was considered as the index date. We excluded patients who had developed HF prior to the index date. We required that each patient have at least six years of follow up with at least one encounters each year. Most patients in the database do not have long term follow-up, thus we could identify only 723 patients. The average age for patients who developed HF was 68 years old (min = 40, max = 90) and the average age for patients who did not develop HF was 64.37 years old (min = 40, max = 90). The cohort characteristics at the end of the follow-up time can be seen in Additional file 1: Table S1.

Model building

Since the data size was relatively small, we bootstrapped it with replacement and increased the size to 10,723. We randomly selected 80% of the encounters for training and 20% for testing purpose. Using the Sklearn Python package, we fitted logistic regression, random forest, XGBoost models, and used the Keras python package to fit a deep neural network model to the data. We tuned hyperparameters through five-fold cross-validation. For the neural nets, the best model by the ROC-AUC measurement, was a fully connected network with four hidden layers and different nodes, ranging from 5 to 22 nodes, implementing sigmoid and ReLu activation functions. To prevent overfitting, dropout at a rate of 0.25 was applied to each hidden layer. Finally, using the test data, we found that ROC-AUC for XGBoost was greater than the other three methods. The only transparent method, logistic regression, was outperformed by the black-box methods. The ROC-AUC of the different models are as follows: Logistic regression: 0.62; Neural network: 0.68; Random forest: 0.74; XGBoost: 0.91. In order to understand the effect of different features on the risk of HF, we need to interpret the black-box models.

Effect score

We developed a model-agnostic metric called "effect score" to compare and interpret the effect of each feature on the risk of the clinical outcome. This method calculates how the logit of the output changes if the current value of feature $\boldsymbol{i}$ $(\boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{c}})$ changes from a chosen reference value $\left(\boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{r}}\right)$ for that feature, e.g., a normal value for a lab test. In order to lower the chance of incorporating unrealistic possibilities by using random values, the value of feature $\boldsymbol{i}$ is replaced only with other observed values for the same feature. Given that there are $\boldsymbol{m}$ features and $\boldsymbol{n}$ observations in the system, the algorithm for computing the effect score for feature $\boldsymbol{x_i}$ can be seen in Table 1.

Results and discussion

Since the XGBoost model resulted in the highest area under the ROC curve, we used that model for understanding HF risk factors. As there was only one continuous variable, "age", we only used the "effect score" algorithm for comparing the effects of categorical variables on the risk of HF. The result of the algorithm can be seen in Table 2.

As it can be seen in Table 2, the highest score = 1.75 is achieved with ischemic heart disease, thus it is considered the most influential individual risk factor for developing HF. This score can be interpreted as meaning that if all other features for *patienti* and *patientj* are the same and *patienti* has ischemic heart disease but *patientj* has not, the logit of developing HF for *patienti* is on average 1.75 more than that of *patientj*. The other scores can be interpreted in a similar fashion.

Table 2 Effect scores for categorical features

Feature	Effect score
Ischemic heart disease (IHD)	1.75
Hypertension (HTN)	1.68
Atrial fibrillation (AF)	1.52
Chronic obstructive pulmonary disease (COPD)	1.52
Cancer	1.48
Chronic kidney disease (CKD)	1.40
Anemia	1.36
Asthma	1.34
Arthritis	1.02
Depression	1.00

One might argue that the "effect score" is more helpful if the underlying model is a neural network model. Ensemble methods, including XGBoost, can provide an "importance score" that ranks the degree of influence of different features on the output. Difficulty with the "importance score" occurs when the predictors are correlated, like those in this case; a result that leads noninfluential predictors to be preferred to significant ones [30]. When we applied the importance metric through Sklearn library in Python, this issue was observed in our use case as well. The ranking of the features from that method is as follows: CKD, COPD, anemia, IHD, Depression, AF, HTN, cancer, and Arthritis. Using this metric, hypertension, long known as a very important factor in the development of HF [6, 18, 19] has less importance than depression. It is highly doubtful that depression is clinically a more significant risk factor than hypertension. This ranking is not consistent with the current clinical.

knowledge. Additionally, it should be noted that the Sklearn approach provides no avenue for interpretation of the meaning of the scores.

Another use of our proposed approach is to understand the nonlinear relationship between a continous feature and the output. In Fig. 1, it can be seen that, as it was expected [31], the risk of HF increases as the age increases. Each point denotes the difference between the logit of risk of HF for *patienti* with that of *patientj*, who has the same comorbidities but is 40 years old. The LOWESS (locally weighted scatterplot smoothing) technique was used to show the trend between the age and risk of HF on the population level. The figure suggests the relationship between the age and the risk of HF is linear. The diversion of observations at the same age indicates that there is a correlation between the age and different comorbid conditions.

Table 1 Different steps to compute the effect score

Step number	Steps
1	fit a machine learning model (for a neural network model, the activation function of the output layer needs to be a sigmoid function)
2	determine a reference value x_i^r
3	$es_{i,j} = logit\left(f\left(x_1^j, \ldots, x_i^j, \ldots, x_n^j\right)\right) - logit(f(x_1^j, \ldots, x_i^r, \ldots, x_n^j)\Big)$ - where $f(.)$ is the prediction of the probability of the positive class by the model - if $x_{i,j} = x_{i,k}$, then consider the average of them
4	$ES_i = \sum_{j=1}^{n} \lvert es_{i,j} \rvert$
5	For a continuous feature, plot $es_{i,j}$ against the value of $x_{i,j\epsilon n\{obs\}}$ to depict the effect of x_i at different values on the output with respect to the reference
6	Rank ES_i of categorial and continuous features to compare strength of different features

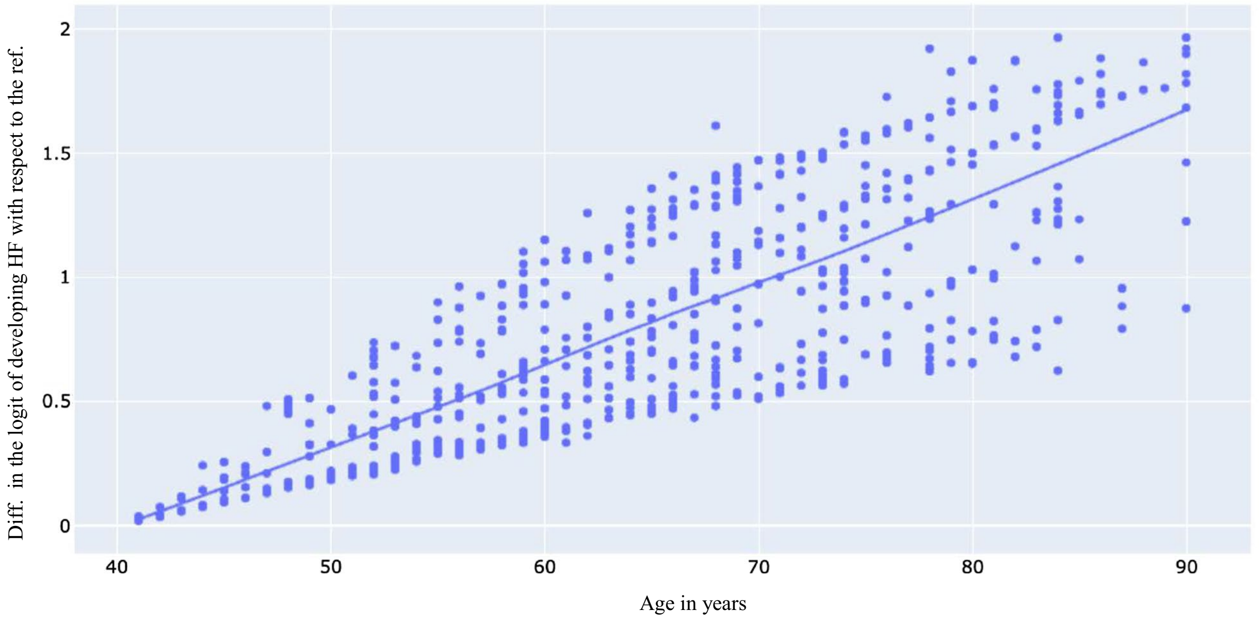

Fig. 1 Visualization of relationship between age and heart failure

Summary and conclusion

Enhancing interpretation of machine learning models can contribute to a better understanding of clinical events. In this paper, we propose a model-agnostic approach to explore how a complex machine learning model can be used to investigate risk factors associated with developing HF among patients with type 2 diabetes. Our approach enables a researcher to interpret the global effect of each individual feature on the outcome; compare the significance of different individual categorical variables; appreciate a nonlinear nonmonotonic relationship between a continuous feature and outcome; visualize the effect of each individual observation on the outcome; and fit a trendline through locally weighted scatterplot smoothing to understand the global effect of that feature.

Limitations

- In future work, the proposed algorithm should be implemented using a higher quality EHR database with larger patient samples.
- This algorithm should not be used to compare the effect of a continuous variable on the outcome with that of a categorical variable.
- When the number of levels of two categorical variables are far different, the results from this algorithm may be biased.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13104-021-05596-7.

Additional file1: **Table S1** Patient characteristics.

Acknowledgements
Not applicable

Authors' contributions

Study design and implementation: SMM, SJN, QZ. Output evaluation: SMM, SJN, QZ. Manuscript composition: SMM, SJN, QZ. All authors read and approved the final manuscript.

Funding
The authors are partially funded by the following grants: NIH 1UL1TR001876; VA 1I01 HX002422; VA 1I01HX002679; NIH R21LM012929.

Availability of data and materials

The datasets used as input are the property of Cerner corporation, from whom they are available.

Declarations

Ethics approval and consent to participate

No patients were contacted, only de-identified data from the EHR database were used.

Consent for publication

Not applicable.

Competing interests

Authors do not have any competing interests.

Received: 3 January 2021 Accepted: 5 May 2021
Published online: 17 May 2021

References

1. Benjamin EJ, et al. Heart disease and stroke Statistics-2019 update a report from the American Heart Association. Circulation. 2019. https://doi.org/10.1161/HHF.0b013e318291329a.
2. Heidenreich PA, et al. Forecasting the impact of in the United States: a policy statement from the American Heart Association. Circ Heart Fail. 2013;6(3):606–19.
3. Adams KF Jr, et al. Characteristics and outcomes of patients hospitalized for heart failure in the United States: rationale, design, and preliminary observations from the first 100,000 cases in the Acute Decompensated Heart Failure National Registry (ADHERE). Am Heart J. 2005;149(2):209–16.
4. Fox K, et al. Coronary artery disease as the cause of incident heart failure in the population. Eur Heart J. 2001;22(3):228–36.
5. Gheorghiade M, Bonow RO. Chronic heart failure in the United States: a manifestation of coronary artery disease. Circulation. 1998;97(3):282–9.
6. Levy D, Larson MG, Vasan RS, Kannel WB, Ho KK. The progression from hypertension to congestive heart failure. JAMA. 1996;275(20):1557–62.
7. Mackenzie J. Diseases of the Heart, ed 3. London: H. Frowde; 1913.
8. Le Jemtel TH, Padeletti M, Jelic S. Diagnostic and therapeutic challenges in patients with coexistent chronic obstructive pulmonary disease and chronic heart failure. J Am Coll Cardiol. 2007;49(2):171–80.
9. Ni H, Nauman DJ, Hershberger RE. Managed care and outcomes of hospitalization among elderly patients with congestive heart failure. Arch Intern Med. 1998;158(11):1231–6.
10. Heywood JT, et al. High prevalence of renal dysfunction and its impact on outcome in 118,465 patients hospitalized with acute decompensated heart failure: a report from the ADHERE database. J Cardiac Fail. 2007;13(6):422–30.
11. Spence RK. The economic burden of anemia in heart failure. Heart Fail Clin. 2010;6(3):373–83.
12. Groenveld HF, et al. Anemia and mortality in heart failure patients: a systematic review and meta-analysis. J Am Coll Cardiol. 2008;52(10):818–27.
13. Sun D, et al. A history of asthma from childhood and left ventricular mass in asymptomatic young adults: the Bogalusa Heart Study. JACC Heart Fail. 2017;5(7):497–504.
14. Khalid U, et al. Incident heart failure in patients with rheumatoid arthritis: a nationwide cohort study. J Am Heart Assoc. 2018;7(2):e007227.
15. Williams SA, Kasl SV, Heiat A, Abramson JL, Krumholz HM, Vaccarino V. Depression and risk of heart failure among the elderly: a prospective community-based study. Psychosom Med. 2002;64(1):6–12.
16. Bowles EJA, et al. Risk of heart failure in breast cancer patients after anthracycline and trastuzumab treatment: a retrospective cohort study. J Natl Cancer Inst. 2012;104(17):1293–305.
17. Bozkurt B, et al. Contributory risk and management of comorbidities of hypertension, obesity, diabetes mellitus, hyperlipidemia, and metabolic syndrome in chronic heart failure: a scientific statement from the American Heart Association. Circulation. 2016;134(23):e535–78.
18. Dunlay SM, Weston SA, Jacobsen SJ, Roger VL. Risk factors for heart failure: a population-based case-control study. Am J Med. 2009;122(11):1023–8.
19. He J, Ogden LG, Bazzano LA, Vupputuri S, Loria C, Whelton PK. Risk factors for congestive heart failure in US men and women: NHANES I epidemiologic follow-up study. Arch Intern Med. 2001;161(7):996–1002.
20. Dhingra R, Vasan RS. Diabetes and the risk of heart failure. Heart Fail Clin. 2012;8(1):125–33.
21. Kohane IS. Using electronic health records to drive discovery in disease genomics. Nat Rev Genet. 2011;12(6):417.
22. Blouin R, Hall M, Saydah S. Using administrative databases to identify cases of chronic kidney disease: a systematic review. Emerg Health Threats J. 2011;4(1):7175. https://doi.org/10.3402/ehtj.v4i0.7175.
23. Aggarwal CC. Neural networks and deep learning. Springer. 2018;10:978–83.
24. Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. Knowl Inf Syst. 2014;41(3):647–65.
25. Shao Y, Cheng Y, Shah RU, Weir CR, Bray BE, Zeng-Treitler Q. Shedding light on the black box: explaining deep neural network prediction of clinical outcomes. J Med Syst. 2021;45(1):1–9.
26. Achen CH. Interpreting and using regression. California: Sage; 1982.
27. Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. J Comput Graph Stat. 2015;24(1):44–65.
28. Shapley LS. A value for n-person games. In: Kuhn HW, Tucker AW, editors. Contributions to the theory of game. Princeton: Princeton University Press; 1953. p. 307–17.
29. Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016. pp. 1135–44.
30. Toloşi L, Lengauer T. Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics. 2011;27(14):1986–94. https://doi.org/10.1093/bioinformatics/btr300.
31. Ho K, Pinsky JL, Kannel WB, Levy D. The epidemiology of heart failure: the Framingham Study. J Am Coll Cardiol. 1993;22(4):A6-13.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.