Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220404

Characterization of Type 2 Diabetes Using Counterfactuals and Explainable AI

Marta LENATTI [a,1], Alberto CARLEVARO[a,b], Karim KESHAVJEE[c], Aziz GUERGACHI[d,e], Alessia PAGLIALONGA[a] and Maurizio MONGELLI[a]
[a] *National Research Council of Italy (CNR), Institute of Electronics, Information Engineering and Telecommunications (IEIIT), Italy;*
[b] *University of Genoa, Department of Electrical, Electronics and Telecommunications Engineering and Naval Architecture (DITEN), Italy;*
[c] *University of Toronto, Institute of Health Policy, Management and Evaluation, Dalla Lana School of Public Health, Canada*
[d] *Ryerson University, Ted Rogers School of Management, Toronto, Canada*
[e] *York University, Department of Mathematics and Statistics, Toronto, Canada*

Abstract. Type 2 diabetes mellitus is a metabolic disorder of glucose management, whose prevalence is increasing inexorably worldwide. Adherence to therapies, along with a healthy lifestyle can help prevent the onset of disease. This preliminary study proposes the use of explainable artificial intelligence techniques with the aim of (i) characterizing diabetic patients through a set of easily interpretable rules and (ii) providing individualized recommendations for the prevention of the onset of the disease through the generation of counterfactual explanations, based on minimal variations of biomarkers routinely collected in primary care. The results of this preliminary study parallel findings from the literature as differences in biomarkers between patients with and without diabetes are observed for fasting blood sugar, body mass index, and high-density lipoprotein levels.

Keywords. Diabetes, Counterfactual Explanations, eXplainable AI

1. Introduction

An increasing number of adults are at risk of developing Type 2 Diabetes (T2D) worldwide, correlating with trends in population aging and sedentary lifestyles. In 2021, an estimated 61.4 million (9.2%) of adults in Europe were living with diabetes, 90% of whom had T2D [1]. Interventions aimed at reducing the risk of T2D and delay disease progression can help reduce its massive burden (European diabetes-related expenditure: US$ 189 billion). To support T2D prediction and prevention, several classification methods have been proposed (e.g., [2]). In the context of explainable artificial intelligence, which is gaining increasing attention in the medical field [3], counterfactual explanations (CEs) [4] are a local technique that searches for minimal distance variations in the input features space (e.g., biomarkers), that would lead to a change in the output class (e.g., diabetes diagnosis). Recently, different approaches for CEs generation have been proposed (e.g., deep learning based [5] or graph based [6]). For example,

[1] Corresponding Author: Marta Lenatti, CNR-IEIIT, c/o Politecnico di Milano, Piazza Leonardo da Vinci, 32, I-20133 Milan, Italy; Phone: +02 23999648; E-mail:marta.lenatti@ieiit.cnr.i

Dhurandhar et al [7] analyzed local explanations applied to brain images to understand why certain subjects were classified as autistic and others were not. To our knowledge, none of the previously proposed methodologies have assessed CE using a rule-based description of the output classes to provide personalized treatments.

The aim of this study was to introduce and characterize CEs as a method to develop individualized diabetes prevention recommendations. We investigate biomarkers, and change in biomarkers, that can help reduce the individual risk of T2D onset and progression.

2. Methods

The dataset used in this study includes 1857 records extracted from primary care electronic medical records (EMRs) from a multi-disease national database, the Canadian Primary Care Sentinel Surveillance Network (CPCSSN, http://cpcssn.ca/). Specifically, the dataset includes 428 subjects with T2D (class "T2D"; age: mean= 56, range=18-90; gender: 227 F, 201 M) and 1429 subjects without T2D (class "No T2D"; age: mean= 63, range=18-83; gender: 916 F, 513 M). The prevalence of T2D in our dataset is 23% and it reflects the prevalence of diabetes in older adults (around 20% in adults > 60 years) [1]. For each patient, seven biomarkers were considered: systolic blood pressure (sBP), body mass index (BMI), low-density lipoprotein (LDL), high-density lipoprotein (HDL), triglycerides (TG), fasting blood sugar (FBS), and age. A Two Class Support Vector Data Descriptor (TC-SVDD) [8] was trained on 80% of the dataset by applying stratification and a natively explainable method, the Logic Learning Machine (LLM) [9], was used to extract a rule-based description of the two classification regions, following the approach presented in [10]. Counterfactual explanations were generated from a set of 120 randomly sampled points within the SVDD region describing T2D subjects. The sample size was defined following preliminary analysis to ensure adequate coverage of the whole age range, including the tails of the distribution (i.e., <50 years and >75 years). The rationale in CE generation was to find the minimum perturbations that should be applied to each biomarker to allow a subject initially classified as diabetic by the TC-SVDD, to be classified as non-diabetic. Given a binary classification problem characterized by two classes S_1 and S_2 and an observation $\boldsymbol{x} \in S_1$, our goal is to determine the minimum variation $\Delta\boldsymbol{x}^*$, so that the point $\boldsymbol{x}^* = \boldsymbol{x} + \Delta\boldsymbol{x}^*$ belongs to the opposite class, by solving the following minimization problem:

$$\min_{\Delta\boldsymbol{x} \in \mathbb{R}^n} d\big(\boldsymbol{x}, (\boldsymbol{x} + \Delta\boldsymbol{x})\big) \; subject\ to \; \|(\boldsymbol{x} + \Delta\boldsymbol{x}) - \boldsymbol{a}_2\|^2 \leq R_2^2 \wedge \|(\boldsymbol{x} + \Delta\boldsymbol{x}) - \boldsymbol{a}_1\|^2 \geq R_1^2$$

where R_1^2, R_2^2, $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ are the radii and the centers of the spheres for class S_1 and S_2, respectively, as defined by the TC-SVDD. The algorithms were implemented in MATLAB (version 2021a).

3. Results

The TC-SVDD achieved the following classification performance on the test set: accuracy = 0.77, sensitivity = 0.65, and specificity = 0.83. Table 1 shows the rules extracted by the LLM from the two regions identified by the TC-SVDD, and the rule

covering (i.e., the percentage of records in the output class that satisfy the rule). Consistent with the diagnostic criteria for T2D [1], the two rules that best characterize the two classes (i.e., the rules with higher covering), are based on the FBS value alone. In particular, high FBS values (i.e., higher than 6.89 mmol/L) are associated with class "T2D" (rule #1) whereas low FBS values (i.e., lower than 5.1 mmol/L) are associated with class "No T2D" (rule #5). The remaining records are covered by more complex rules, i.e., rules based on more than one condition.

Table 1. Rules characterizing the two SVDD regions and rule covering (Cov).

#	Class	Rule	Cov%
1	T2D	FBS>6.89	52.70
2	T2D	FBS>4.91 ∧ 48< age ≤80 ∧ 98< sBP ≤163 ∧ 21< BMI ≤48 ∧ LDL≤1.88	19.90
3	T2D	6.33 < FBS ≤ 6.91 ∧ age > 37 ∧ 100 < sBP ≤ 145	16.55
4	T2D	6.04 < FBS ≤ 6.31 ∧ age ≤ 76 ∧ sBP ≤ 152 ∧ 22 < BMI ≤ 48 ∧ TG > 0.17	8.16
5	NoT2D	FBS ≤ 5.1	47.46
6	NoT2D	FBS ≤ 7.23 ∧ sBP ≤ 111	18.22
7	NoT2D	FBS ≤ 6.68 ∧ age > 50 ∧ BMI > 24 ∧ LDL > 3.61	15.57
8	NoT2D	FBS ≤ 7.87 ∧ 147 < sBP ≤ 182 ∧ LDL > 1.09	15.51

An example of counterfactual (class "No T2D") generated from a record of a T2D patient under the constraint of minimal distance is shown below:

- *T2D patient:* Age = 73 years; sBP = 113 mmHg; BMI = 29.4 kg/m2; LDL = 2.3 mmol/L; HDL = 1.82 mmol/L; TG = 0.83 mmol/L; FBS = 6.9 mmol/L
- *Counterfactual*: Age = 73 years; sBP = 113 mmHg; BMI = 27.9 kg/m2; LDL = 2.38 mmol/L; HDL = 3 mmol/L; TG = 1.07 mmol/L; FBS = 4.89 mmol/L

Small, yet meaningful changes can be found between the observation as the diabetic patient is associated with higher FBS and BMI, stable sBP and lower HDL, LDL, and TG, with respect to the counterfactual, allowing for the two observations to be associated with different classes, hence to be described by different rules, as shown in Table 1.

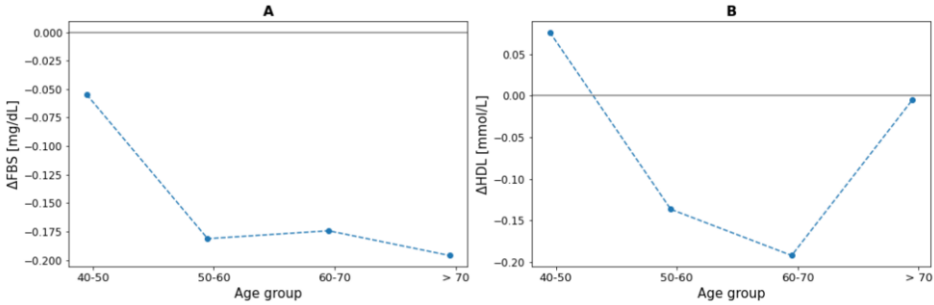

Figure 1. Average change in FBS (panel A) and HDL (panel B) associated with a variation in the output class from class "T2D" to class "No T2D" as a function of the age group.

The average change observed for each biomarker, computed as the difference between the value in the counterfactual (i.e., class "No T2D") and the value in the actual T2D patient is shown in Figure 1 as a function of age (i.e., 40-50 years, 50-60 years, 60-70 years, and 70 or more years). Figure 1(A) shows that the average change in FBS is negative, indicating an improvement in FBS from a diabetic subject to a non-diabetic one. The change in FBS tends to increase with age, ranging, on average, from -0.2 to -1.2 mmol/L. Vice versa, Figure 1(B) shows that the average variation for HDL is positive,

as non-diabetic subjects are characterized by higher values (i.e., better values) with respect to diabetic ones. Regarding HDL, the variation is quite stable for different age groups, around 0.20-0.30 mg/dL.
The average variation observed across the 120 generated counterfactuals is: ΔAge = 0.01 years; ΔsBP = −0.17 mmHg; ΔBMI = −0.14 kg/m2; ΔLDL = 0.53 mmol/L; ΔHDL = 0.26 mmol/L; ΔTG = 0.12 mmol/L; ΔFBS = −0.88 mmol/L.

4. Discussion

The extraction of rules from the two classification regions identified by the TC-SVDD allows us to obtain a comprehensive, readable view of the observed pathology, that can be easily interpreted and validated by clinicians, even if they have no prior knowledge in the field of artificial intelligence. Counterfactual explanations allow us to move from a global to a local approach, tailored on an individual basis.

Although the counterfactual generation process searches, by its definition, for minimal variations that determine a change in output class, the results obtained allow us to appreciate differences in biomarkers between the two classes, in line with the available knowledge. Indeed, high FBS is a peculiar characteristic of diabetic subjects, as T2D is a metabolic disorder of glucose regulation resulting in hyperglycemia. A trend towards higher BMI in diabetic patients is consistent with the known link between obesity and T2D. For this reason, maintaining a healthy lifestyle characterized by moderate exercise can help in the prevention of diabetes. Moreover, low HDL concentrations have been associated with higher risk of developing T2D [11], possibly representing one of the risk factors for cardiovascular disease. In contrast, a counterintuitive result is related to the presence of slightly higher (i.e., worse) LDL and TG values in nondiabetic patients, compared with diabetic ones. This result can be explained by the fact that diabetic and pre-diabetic subjects are very often prescribed LDL-lowering drugs like statins [12] to control hypercholesterolemia. Further research on larger data should characterize different stratifications, such as by gender, as the variations obtained in biomarkers might be different for male or female subjects. In addition, causal relationships among biomarkers should also be assessed, for example using causal inference approaches.

The proposed method is characterized by a degree of uncertainty in the classification, because counterfactuals are sought in the proximity of the decision boundary, defined by the SVDD, thus only decreasing the risk of developing diabetes by a small amount (the minimal amount needed for changing the output class). A safer definition of subjects with low probabilities of developing diabetes, would be the search of counterfactuals from “reliable SVDD regions”, i.e., by defining regions with a reduced false negative rate, as in [13]. The reliable approach is based on the relaxation of the definition of counterfactuals generated at minimum distance, as they will be searched within spherical regions of smaller size, enabling a more precise identification of the opposite class. Specifically, the TC-SVDD will define a region surrounding non-diabetic subjects, minimizing the presence of diabetic subjects within that region, according to the imposed false negative rate. In this way, personalized prevention strategies could be devised through more effective changes in biomarkers, so that the risk of developing the disease can be substantially reduced. Also, it should be noted that characterization of diabetic patients is based on a combination of controllable characteristics (biomarkers that can be manipulated through therapies or lifestyle changes) and non-controllable characteristics (e.g. features that are not manipulable by definition such as age, genetic factors, and

family history). Therefore, the search for realistic counterfactuals should be performed by perturbing only controllable variables and keeping non-controllable variables fixed.

5. Conclusions

This preliminary study aims to demonstrate the efficacy of a new method for CE generation based on SVDD on biomarker variation and diabetes diagnosis. Specifically, the search for minimal variations in biomarkers, able to change the classification of the subject from diabetic to non-diabetic allows us to highlight a trend (decrease or increase) in biomarkers, consistent with current knowledge. Further research is needed to validate the proposed method in different datasets (e.g., with varying size and number of output classes) and for different chronic diseases (e.g., COPD). This approach can be extended in the future to a broader set of biomarkers including lifestyle indicators (e.g., diet, alcohol consumption, smoke), comorbidities and medications and to a broader definition of counterfactual obtained perturbing controllable features only, at progressively increasing distance (i.e., progressively decreasing risk of developing diabetes), with the ultimate goal of personalizing prevention strategies on an individual basis.

Acknowledgments: The work was supported by Fondazione Compagnia di San Paolo, scientific research call 2019 (Bando 2019–2020 per progetti di ricerca scientifica presentati da enti genovesi): project "Advances in pneumology via ICT and data analytics" (PNEULYTICS)

References

[1] International Diabetes Federation (2021). IDF Diabetes Atlas 10th edition.
[2] Silva KD, Lee WK, Forbes A, Demmer RT, Barton C, Enticott J. Use and performance of machine learning models for type 2 diabetes prediction in community settings: A systematic review and meta-analysis. International Journal of Medical Informatics. Elsevier BV; 2020. p. 104268.
[3] Hakkoum H, Abnane I, Idri A. Interpretability in the medical field: A systematic mapping and review study. Applied Soft Computing; 2022. 117. p. 108391
[4] Wachter S, Mittelstadt BDM, and Russell C. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. Harvard Journal of Law and Technology. 2018;31(2):841–87.
[5] Nemirovsky D, Thiebaut N, Xu Y, Gupta A, "Providing actionable feedback in hiring marketplaces using generative adversarial networks," WSDM '21. Association for Computing Machinery, 2021.
[6] Poyiadzi R, Sokol K, Santos-Rodríguez R, Bie TD, Flach PA, "Face: Feasible and actionable counterfactual explanations," AAAI/ACM Conference on AI, Ethics, and Society, 2020.
[7] Dhurandhar A, Chen PY, Luss R, Tu CC, Ting P, Shanmugam K, Das P. Explanations based on the missing: towards contrastive explanations with pertinent negatives. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18), 2018.
[8] Huang G, Chen H, Zhou Z, Yin F, Guo K. Two-class support vector data description. Pattern Recognition. Elsevier BV; 2011. p. 320–329.
[9] Muselli M, Ferrari E. Coupling Logical Analysis of Data and Shadow Clustering for Partially Defined Positive Boolean Function Reconstruction. IEEE Trans. Knowl. Data Eng. IEEE; 2011. p. 37–50.
[10] Carlevaro A, Mongelli M. Reliable AI Through SVDD and Rule Extraction. Lecture Notes in Computer Science. Springer International Publishing; 2021. p. 153–171.
[11] Wilson PWF. Prediction of Incident Diabetes Mellitus in Middle-aged Adults. Arch Intern Med. American Medical Association (AMA); 2007. p. 1068.
[12] Feingold KR. Cholesterol Lowering Drugs. 2021 Mar 30. In: Feingold KR, Anawalt B, Boyce A,Chrousos G, de Herder WW, Dhatariya K, et al, editors. Endotext.

[13] Carlevaro A, Mongelli M. A New SVDD Approach to Reliable and eXplainable AI. IEEE Intell. Syst.IEEE; 2021. p. 1–1.