Preventive Medicine Reports 32 (2023) 102129

Contents lists available at ScienceDirect

Preventive Medicine Reports

journal homepage: www.elsevier.com/locate/pmedr

Risk prediction of chronic diseases with a two-stage semi-supervised clustering method

Zaixing Mao [a,*], Yasufumi Fukuma [a], Hisashi Tsukada [a], Satoshi Wada [b]

[a] SAI, KK. Wako, Saitama, Japan
[b] RIKEN Center for Advanced Photonics, RIKEN, Wako, Saitama, Japan

A B S T R A C T

Early detection of chronic diseases such as cardiovascular disease (CVD) and diabetes can make the difference between life and death. Previous studies have demonstrated the feasibility of disease diagnosis and prediction using machine learning and disease-indicating biomarkers. The aim of this study is to develop a method to detect the risk of future disease even when disease-indicating biomarker readings are in the normal range. Data from the US Centers for Disease Control and Prevention (CDC) National Health and Nutrition Examination Surveys (NHANES) are used for this study. A two-stage semi-supervised K-Means (SSK-Means) clustering approach was developed to identify the underlying risk of each individual and categorize them into high or low-risk groups for CVD and diabetes. Our developed method of classification can identify groups as high risk or low risk, even if they would have been considered normal using traditional biomarker threshold criteria. For CVD, the SSK-Means clustering results showed that individuals over 30 years of age in the high-risk group were almost twice as likely to develop CVD as individuals in the low-risk group. For diabetes, the SSK-Means clustering results showed that individuals over 50 years in the high-risk group have at least two times the risk of developing diabetes compared with individuals in the low-risk group.

1. Introduction

Cardiovascular disease (CVD) was the leading cause of death in the United States in 2019, accounting for over 650,000 deaths (Kochanek et al., 2020). In 2010, the American Heart Association (AHA) proposed a definition of ideal cardiovascular health behaviors and health factors in attempts to reduce CVD mortality and improve cardiovascular health (Lloyd-Jones et al., 2010). Diabetes can lead to diabetic retinopathy, which is the most common cause of new cases of blindness in adults aged 20–74 years (Fong, 2004). Early detection of these diseases can be life-saving.

2. Related works

2.1. Lsimitations of disease-indicating biomarkers

Several previous studies have demonstrated the feasibility of disease diagnosis and prediction using machine learning techniques and disease-indicating biomarkers (Pasha et al., 2020; Zriqat et al., 2017; Ashiquzzaman et al., 2017; Soltani and Jafarian, 2016). Disease-indicating biomarkers, such as cholesterol level for CVD (Pekkanen et al., 1990) and glycohemoglobin for diabetes (Krishnamurti and Steffes, 2001), are powerful in diagnosing the current condition of the human body. However, a normal value in a disease-indicating biomarker does not necessarily mean that the risk of future disease development is low, as disease development may still be at an early stage. Relying solely on these biomarkers could lead to false-negative detections.

2.2. Lsimitations of current disease risk factor analysis.

Analysis of disease risk factors is essential for disease prevention and control. It is widely accepted that lifestyle habits (McBride, 1992) and obesity (Krauss et al., 1998) are risk factors for cardiovascular disease, but their association with the disease are often assumed independently and may not consider correlations with other parameters. Machine learning methods (Park et al., 2019) can analyze multiple risk factors, but they typically require labeled data from low-risk and diseased individuals. However, defining low-risk individuals can be ambiguous, so “not diseased” is often used as a proxy. This approach may lead to contaminated training data and lower performance in disease risk assessment because currently healthy individuals may be at high risk of developing disease in the future.

* Corresponding author.
E-mail address: zmao@sensor-ai.com (Z. Mao).

https://doi.org/10.1016/j.pmedr.2023.102129
Received 12 June 2022; Received in revised form 17 January 2023; Accepted 3 February 2023
Available online 6 February 2023

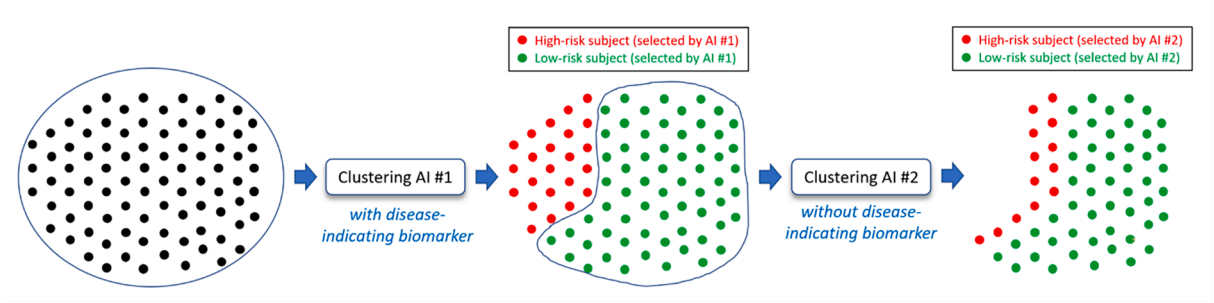

Fig. 1. Overview workflow.

3. Overview of approach

The aim of this study is to improve conventional models of disease risk factors, even for individuals whose disease-indicating biomarker levels are within the normal range. As shown in Fig. 1, our approach consists of two steps: First, an initial clustering AI with disease-indicating biomarkers is used to divide subjects into high- and low-risk groups. This ensures that subjects whose disease-indicating biomarker levels are in the abnormal range are identified as high-risk subjects. Next, for the subjects in the low-risk group selected by the first clustering AI, a second clustering AI without disease-indicating biomarkers is used to further identify the high- and low-risk subjects. By excluding disease-indicating biomarkers from the analysis, this step can capture subjects whose disease-indicating biomarker values are within the normal range but who have other measurements similar to those of a diseased patient.

Semi-supervised K-Means (SSK-Means) clustering is a multivariate analysis method that has the advantage of not requiring labeled data for each class type. In the context of risk factor analysis, only "diseased" labeled data is needed and there is no need to use "not diseased" as a proxy for "low risk". Because of these advantages, SSK-Means is chosen as the machine learning model for this analysis.

4. Data

Data used in this study are from the Centers for Disease Control and Prevention (CDC) National Health and Nutrition Examination Surveys (NHANES). Results from the 2013–2014 NHANES (Centers for Disease Control and Prevention, 2013) were used for training and development of the risk assessment algorithm. Results from a separate 2011–2012 NHANES survey (Centers for Disease Control and Prevention, 2012) are used for validation. This study was based on a publicly available anonymized databases, and thus exempt from ethical compliance.

4.1. Preprocessing

Each NHANES study contains over 100 files of different measurements. However, many of these measurements have very few participants or change from survey to survey. To ensure sufficient study statistics for both the training and validation data sets, the following 19 data files will be selected for study: demographic variables and sample weights, blood pressure, body measurements, urinary albumin and creatinine levels, complete blood count with 5-part differential, HDL cholesterol, total cholesterol, folate, glycohemoglobin, hepatitis A, hepatitis B: Core antibody, s-surface antigen and hepatitis D antibody, vitamin D, vitamin B12, diabetes disease, kidney disease, food security, medical conditions, smoking habits, and sleep disorders. Each data file contains a different number of examinations, and the total number of individual measurements from the 20 data files is 48.

Table 1
Measurements removed by the first filter.

Target Disease Type	Filtered Measurements	
	Ground Truth Related	Disease Indicating Biomarkers
CVD	Presence of cardio-vasculature disease is true if any of the following inspection items is true: Ever told had congestive heart failure? Ever told you had coronary heart disease? Ever told you had angina/angina pectoris? Ever told you had heart attack?	• High-density lipoprotein cholesterol. Total cholesterol.
Diabetes	Presence of diabetes is true if the following inspection items is true: Doctor told you have diabetes.	• Glycohemoglobin.

Table 2
Measurements removed by the second filter.

Target Disease Type	Filtered Measurements		
	highly correlated with other measurements	Irrelevant	By design
CVD	• *epi*-25-hydroxyvitamin D3 • BMI • Food didn't last?	• Segmented neutrophils percent	• Age • Gender
Diabetes	• BMI	• Red blood cell folate • Couldn't afford balanced meals? • Worried run out of food?	• Age • Gender

4.2. Measurements filtering

Measurements are filtered through two stages to remove disease-indicating biomarkers (for clustering AI #2 only) and improve risk separation performance.

In the first filtering stage, for each target disease type, the corresponding medical conditions (which serve as ground truth) and disease-indicating biomarkers (for clustering AI #2 only) are removed. In this work, two disease types are studied: CVD and diabetes. Measurements that are filtered out for each target disease type are summarized in Table 1.

The second stage filter is used to remove measurements that are highly correlated with other measurements, irrelevant to the target disease or that, such as age, may lead to undesirable results. Age is a highly correlated factor to many diseases. As age increases, the likelihood of developing a disease naturally increases. Thus, if age is used as one of the input parameters, a risk assessment algorithm will simply

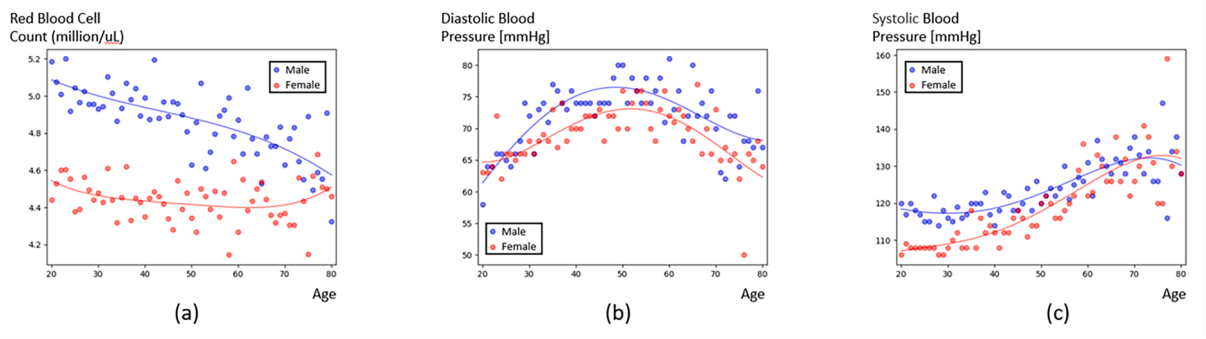

Fig. 2. Polynomial fitting for age-dependent parameters. (a): the distribution of red blood cell count. (b): the distribution of diastolic blood pressure. (c) the distribution of systolic blood pressure. The median value for each age point is represented as dots. Red for female and blue for male. The fitted quartic polynomial are shown as red and blue lines. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

classify young people as low risk. This risk assessment algorithm can determine the emerging risks, but it cannot predict the future risks. While age and gender are not included as clustering terms, they are utilized in the subsequent risk prediction.

Measurements filtered out by the second filter for each target disease type are summarized in Table 2.

After the second filter, a subject is removed from the study if there is missing data in any of the remaining measurements. This process reduced the dataset size from 10,907 to 8238 for the CVD analysis and 8264 for the diabetes analysis. During data collection, some measurements may have been erroneously recorded. To exclude such data points, outlier data points are removed if the value of an measurement (excluding questionnaires) is greater than the 99th percentile. This process reduced the dataset size to 7508 for the CVD analysis and 5389 for the diabetes analysis. The number of measurements after all filtering steps are: 36 for the CVD analysis and 40 for the diabetes analysis.

4.3. Input parameter gender-age-dependency removal

In cross-sectional studies, the basic assumption for predicting future disease risk is that the individual's measures will not change significantly if they maintain their current lifestyle. However, some biomarkers are inherently gender and age-dependent and change with age.

To remove the age dependence of the different parameters, we first estimate the normative age dependence of each parameter, using only healthy subjects. As shown in Fig. 2, the median value for each age point is represented as dots. Red for female and blue for male. For each gender-age-dependent parameter, the median value is fitted against the age distribution using a quartic polynomial and shown as the red and blue lines.

Next, gender-age-adjusted parameters are calculated for each individual by subtracting the nominal value at their age calculated with the corresponding fitted polynomial from their unadjusted parameter values.

4.4. Standardization

By its very nature, the measurement range of each measurement can vary widely. These differences in range will cause issues for the clustering algorithm which is based on distance computation. The final step of preprocessing is a standardization process in which each continuous input is adjusted for mean and scaled to unit variance. The standardized z based on the input x is defined as:

$z_i = \frac{x_i - \overline{x}}{\sigma}$, where x_i is input x of subject i, $\overline{x} = \frac{1}{N}\sum_{i=0}^{N} x_i$ is the average value of x of all subjects, $\sigma^2 = \frac{1}{N}\sum_{i=0}^{N}(x_i - \overline{x})^2$ is the variance of x of all subjects.

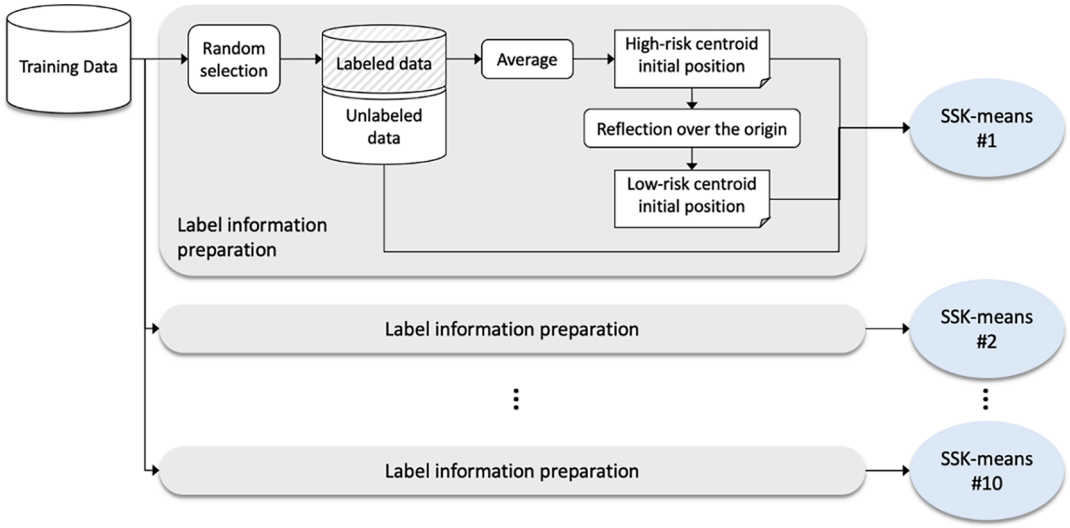

Fig. 3. Training procedure.

Table 3
Data split of training and validation data.

Target Disease Type	Training Data	Validation Data
CVD	3864	3639
Diabetes	2765	2624

5. Method

5.1. Semi-supervised K-Means clustering

K-Means cluster (Lloyd, 1982 Mar) analysis is a method used to group data into subgroups based on similarity. K, the desired number of clusters, is set to 2 in our study, which refers to the high and low-risk groups. The procedure of K-Means clustering can be summarized as follows:

In the hyperdimensional space of input variables:

1) Initialize the position of 2 centroids.
2) Assign each data point to its nearest centroid.
3) For each centroid, calculate the location of the center of the data points assigned to it. Then move the centroid to the central position.
4) Repeat steps 2) and 3) until termination conditions are met, e.g. after a fixed number of iterations or when the centroids stop moving.

K-Means clustering can divide people into two groups, but it has no inherent meaning for centroids and cannot control how the data is divided. To improve K-Means, we can use semi-supervised K-Means (SSK-Means) (Arthur and Vassilvitskii, 2006), which initializes centroids with labeled data and allows reassignment of labeled data points to different centroids.

In SSK-Means, centroids are initialized with labeled data and the average position of data points belonging to the same label is used to define the initial centroids. This allows the clustering algorithm to focus on relevant features and gives inherent meaning to the centroids. For example, if a centroid location is initialized by diabetics, that centroid is automatically given the meaning "high risk of diabetes".

During training, it is possible to reassign labeled data points to different centroids. However, in this study, we do not allow reassignment of labeled data because we assume there is not a significant amount of mislabeled data.

The goal of this study is to divide people into high and low- risk groups. While it is clear to assign people who are already ill to the high-risk group, it is more difficult to determine who should be assigned to the low-risk group. Healthy people are not necessarily at low risk of getting sick.

An advantage of SSK-Means is that it uses mainly labeled data in the initialization phase, using the average value of the labeled data to determine the initial positions of the centroids. Since the input variables have been standardized to an average value of 0, the low-risk centroid is set as the reflection of the high-risk centroid around the origin.

Table 4
Number of subjects in high and low-risk groups after each clustering step in CVD risk assessment.

	After Clustering AI #1	After Clustering AI #2
Number of subjects in high-risk group	1415	1430
Number of subjects in low-risk group	2224	2209

Table 5
The mean values and standard deviations of key biomarkers in different risk-groups in CVD risk assessment.

Group names	Group description	HDL cholesterol [mg/dL]	Red blood cell folate [ng/mL]	Urine albumin-creatinine ratio [mg/g]
High-risk group I	High-risk subjects selected by clustering AI #1	48.6 ± 13.8	490 ± 241	17.1 ± 21.7
High-risk group II	Additional high-risk subjects selected by clustering AI #2	65.2 ± 11.0	514 ± 189	13.4 ± 13.0
Low-risk group	Low-risk subjects	54.9 ± 14.5	472 ± 187	7.9 ± 6.9

The training procedure is summarized in Fig. 3. First, a small fraction of the diseased subjects is randomly selected as labeled data and used to set the initial positions of the high-risk and low-risk centroids. The labeled and unlabeled data, along with the centroid initial positions, are then sent to the SSK-Means system for training.

The training procedure is repeated 10 times to produce an ensemble of SSK-Means algorithms. The classification results of these algorithms are combined and averaged to make the final prediction. The effect of the amount of labeled data on the separation power is investigated by varying the percentage of labeled data from 0% to 30%. Detailed results can be found in the Appendix.

For each disease, the disease label is determined by the ground truth information listed in Table 1. The detailed training and validation data split is summarized in Table 3.

After training, parameters in the 10 SSK-Means algorithms are frozen for application. The validation procedure is summarized in Fig. 4.

6. Results

6.1. Cardiovascular risk assessment

The results for the SKK-Means clustering are shown in Table 4. After clustering AI #2, an additional 15 individuals were selected as high-risk individuals without using cholesterol information for the assessment.

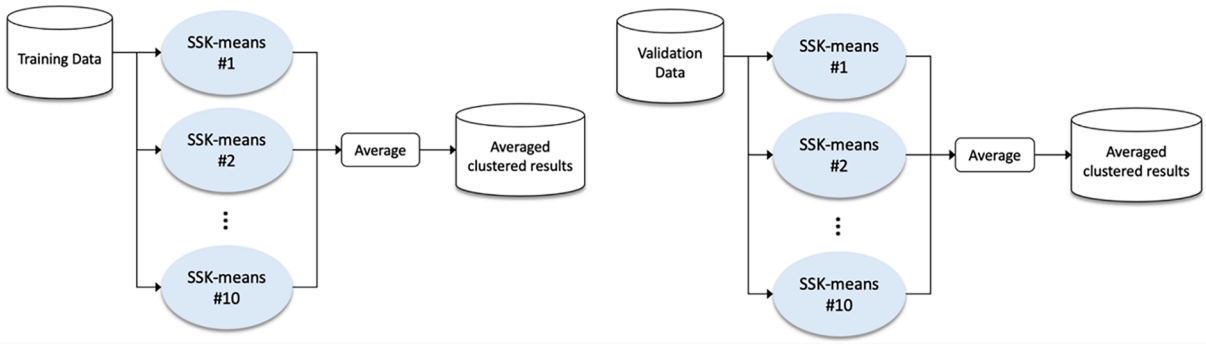

Fig. 4. Validation procedure. Left: performance evaluation using training data. Right: performance evaluation using validation data.

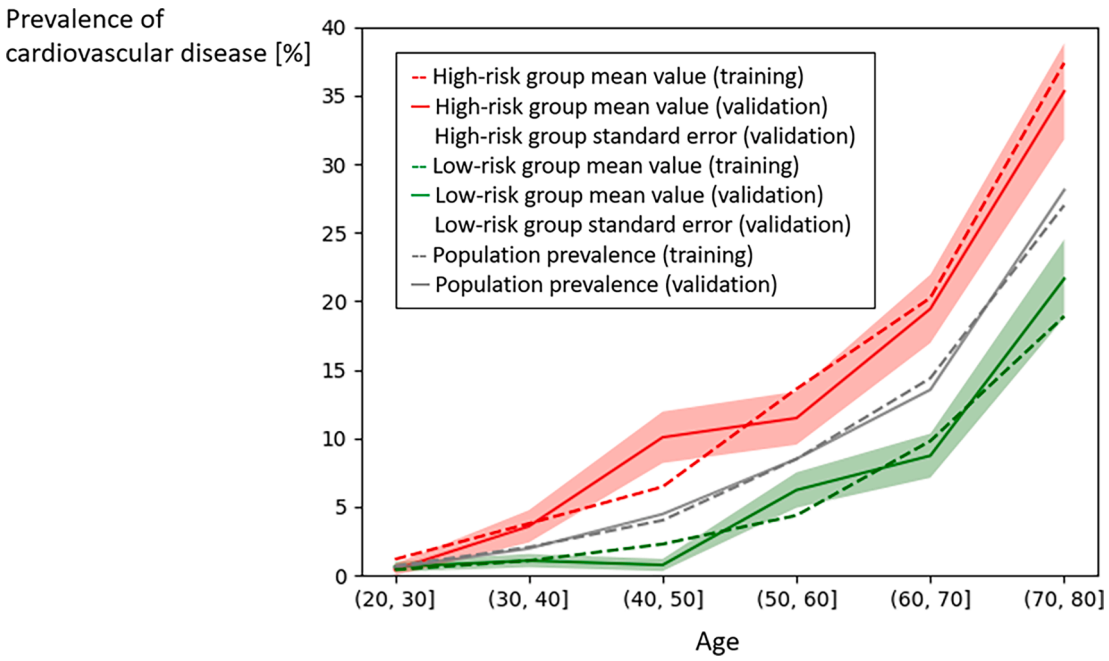

Fig. 5. Prevalence of CVD vs age. Red dashed line, mean value of high-risk group in the training data. Green line, mean value of low-risk group in the training data. Red line, mean value of high-risk group in the validation data. Red band, standard error of high-risk group in the validation data. Green line, mean value of low-risk group in the validation data. Green band, standard error of low-risk group in the validation data. 10% of the labeled data is used as ground truth for SKK-Means clustering. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 6
Number of subjects in high and low-risk groups after each clustering step in diabetes risk assessment.

	After Clustering AI #1	After Clustering AI #2
Number of subjects in high-risk group	628	744
Number of subjects in low-risk group	1996	1880

Table 7
The mean values and standard deviations of key biomarkers in different risk-groups of the diabetes analysis.

Group names	Group description	Glycohemoglobin [%]	Vitamin B12 [pmol/L]	BMI [kg/m^2]
High-risk group I	High-risk subjects selected by clustering AI #1	6.4 ± 1.6	414 ± 197	35.5 ± 6.7
High-risk group II	Additional high-risk subjects selected by clustering AI #2	5.4 ± 0.4	409 ± 199	31.3 ± 3.9
Low-risk group	Low-risk subjects	5.5 ± 0.6	432 ± 206	26.0 ± 4.0

The mean values and standard deviations of key biomarkers in the different risk groups are shown in Table 5. As expected, lower HDL cholesterol levels are generally associated with a higher risk of heart disease, and subjects in high-risk group I have lower HDL cholesterol levels on average. However, the most interesting group of subjects is the high-risk group II. Their high HDL cholesterol levels seem to indicate that they are at low risk for CVD. However, a closer look at the other measurements reveals a significant difference between them and the low-risk group. For example, they appear to have the highest folate concentration in red blood cells within the three groups. Recent studies have shown that high folate concentrations in red blood cells are significantly associated with an increased risk of coronary heart disease (Peng and Wang, 2017). Also, the urine albumin-creatinine ratio, which is known to correlate positively with CVD risk (Gerstein et al., 2001), is higher in the high-risk group II than in the low-risk group.

Although not statically significant, the difference in means seems to indicate that clustering AI #2 may have selected individuals at higher risk for CVD despite high HDL cholesterol levels. In the following text, the combination of high-risk group I and high-risk group II is referred to as the high-risk group.

The results of SKK-Means clustering for CVD are shown in Fig. 5. Clustering separates data into high and low-risk groups and predicts future risk based on age and disease prevalence. Cardiovascular prevalence increases with age, but there is a significant separation between high and low-risk groups for those over 30 years old. This indicates that people in the high-risk group in their 20 s may have more than twice the risk of developing heart-related problems when they are in their 40 s. Numerical values of Fig. 5 can be found in the Appendix.

6.2. *Diabetes risk assessment*

The results for the SKK-Means clustering are shown in Table 6. After clustering AI #2, an additional 116 individuals were selected as high-risk individuals without using Glycohemoglobin information for the assessment.

Table 7 presents the means and standard deviations of key biomarkers in the different risk groups. As expected, subjects in the high-risk group I have, on average, higher glycohemoglobin levels than low-risk subjects. However, subjects in high-risk group II appear to have lower glycohemoglobin levels, indicating that they may be at low risk for diabetes. Despite this, other measurements show a significant difference between high-risk group II and the low-risk group. For instance, they have the lowest vitamin B12 levels among the three groups, and recent studies have linked vitamin B12 deficiency with diabetes mellitus (Kibirige and Mwebaze, 2013). Additionally, the high-risk group II has a higher BMI level compared to the low-risk group, which is known to be a strong risk factor for diabetes (Narayan et al., 2007).

The difference in means, while not statistically significant, appears to

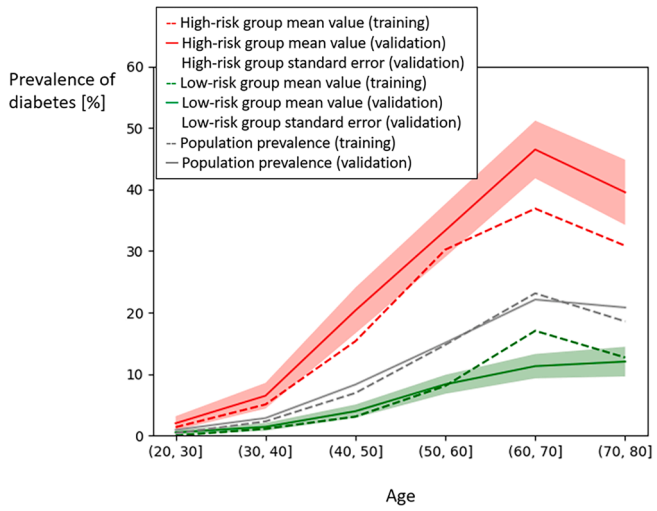

Fig. 6. Diabetes prevalence vs age. Same coloring scheme as Fig. 5.

indicate that clustering AI #2 may have identified individuals with a higher risk of diabetes despite having low glycohemoglobin levels.

The results of prevalence versus age for diabetes are summarized in Fig. 6. Similar to CVD, diabetes prevalence also shows an upward trend with age. As shown, individuals over 30 years in the high-risk group have at least two times the risk of developing diabetes compared with individuals in the low-risk group. Numerical values of Fig. 6 can be found in the Appendix.

7. Discussion

In this study, we demonstrated the feasibility of a two-stage, semi-supervised clustering method for detecting future risk of cardiovascular disease and diabetes. The first stage was developed to use disease-indicating biomarkers to maximize detection sensitivity when subjects' disease-indicating biomarkers already have abnormal levels. The second stage was designed to further reduce false-negative results by intentionally ignoring the effects of disease-indicating biomarkers and focusing on the other biomarkers. Disease risk detection performance is evaluated and validated using publicly available NHANES datasets.

To the best of our knowledge, this is the first study to incorporate such a multi-stage semi-supervised approach for future risk assessment.

In this study, it was assumed that the parameters for classifying risk groups would not change significantly if the individual maintained his or her lifestyle. To ensure this, the parameters were age-gender-adjusted to remove any age dependence. Based on this assumption, our study demonstrated the feasibility of predicting future disease risk using cross-sectional data.

In the future, we plan to conduct a longitudinal study to evaluate the effectiveness of our classification method and determine whether individuals can reduce their risk of disease by improving relevant parameters. If successful, this could benefit the health management of all humanity.

8. Limitations

This analysis is based on cross-sectional data collected by CDC. The limitation of cross-sectional studies is that although we can determine the correlation between our multivariate risk factor and the disease, the causality between the disease and the multivariate risk factor is unknown. Because the data are based on self-report, subjects may misreport or be unaware of their disease state. To address these limitations, we plan to use longitudinal data labeled by medical experts in the future.

Funding statement

The funder, Topcon Corporation, provided support in the form of salaries for authors ZM and HT, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors SW participated in study design, decision to publish, and preparation of the manuscript.

The funder, SAI KK, provided support in the form of salaries for authors YF, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.pmedr.2023.102129.

References

Arthur, David, and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Stanford, 2006.

Ashiquzzaman, A., Tushar, A.K., Islam, M.R., Shon, D., Im, K., Park, J.H., et al., 2018. Reduction of overfitting in diabetes prediction using deep learning neural network. In: InIT Convergence And Security 2017. Springer, Singapore, pp. 35–43.

Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, [2011-2012] [https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2011].

Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, [2013-2014] [https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2013].

Fong, D.S., et al., 2004. Retinopathy in diabetes. Diabetes Care 27 (suppl_1), s84–s87.

Gerstein, H.C., Mann, J.F., Yi, Q., Zinman, B., Dinneen, S.F., Hoogwerf, B., Hallé, J.P., Young, J., Rashkow, A., Joyce, C., Nawaz, S., 2001. Albuminuria and risk of cardiovascular events, death, and heart failure in diabetic and nondiabetic individuals. JAMA 286 (4), 421–426.

Kibirige, D., Mwebaze, R., 2013. Vitamin B12 deficiency among patients with diabetes mellitus: is routine screening and supplementation justified? J. Diab. Metab. Disord. 12 (1), 1–6.

Kochanek, K.D., Xu, J.Q., Arias, E., 2019. Mortality in the United States 2019. NCHS Data Brief, no 395. National Center for Health Statistics, Hyattsville, MD, p. 2020.

Krauss, R.M., Winston, M., Fletcher, B.J., Grundy, S.M., 1998. Obesity: impact on cardiovascular disease. Circulation 98 (14), 1472–1476.

Krishnamurti, U., Steffes, M.W., 2001. Glycohemoglobin: a primary predictor of the development or reversal of complications of diabetes mellitus. Clin. Chem. 47 (7), 1157–1165.

Lloyd, S., 1982. Least squares quantization in PCM. IEEE Trans. Inform. Theory. 28 (2), 129–137.

Lloyd-Jones, D.M., Hong, Y., Labarthe, D., Mozaffarian, D., Appel, L.J., Van Horn, L., et al., 2010. Defining and setting national goals for cardiovascular health promotion and disease reduction: the American Heart Association's strategic Impact Goal through 2020 and beyond. Circulation 121 (4), 586–613.

McBride, P.E., 1992. The health consequences of smoking: cardiovascular diseases. Med. Clin. North Am. 76 (2), 333–353.

Narayan, K.V., Boyle, J.P., Thompson, T.J., Gregg, E.W., Williamson, D.F., 2007. Effect of BMI on lifetime risk for diabetes in the US. Diab. Care 30 (6), 1562–1566.

Park, H.C., Lee, Y.K., Cho, A., Han, C.H., Noh, J.W., Shin, Y.J., et al., 2019. Diabetic retinopathy is a prognostic factor for progression of chronic kidney disease in the patients with type 2 diabetes mellitus. PloS One 14 (7), e0220506.

Pasha SN, Ramesh D, Mohmmad S, Harshavardhan A. Cardiovascular disease prediction using deep learning techniques. InIOP Conference Series: Materials Science and Engineering 2020 Dec 1 (Vol. 981, No. 2, p. 022006). IOP Publishing.

Pekkanen, J., Linn, S., Heiss, G., Suchindran, C.M., Leon, A., Rifkind, B.M., et al., 1990 Jun 14. Ten-year mortality from cardiovascular disease in relation to cholesterol level among men with and without preexisting cardiovascular disease. New Engl. J. Med. 322 (24), 1700–1707.

Peng, Y., Wang, Z., 2017. Red blood cell folate concentrations and coronary heart disease prevalence: A cross-sectional study based on 1999–2012 National Health and Nutrition Examination Survey. Nutrit. Metab. Cardiovasc. Dis. 27 (11), 1015–1020.

Soltani, Z., Jafarian, A., 2016. A new artificial neural networks approach for diagnosing diabetes disease type II. Int. J. Adv. Comput. Sci. Appl. 1 (7), 89–94.

Zriqat, I. A., Altamimi, A. M., & Azzeh, M. (2017). A comparative study for predicting heart diseases using data mining classification methods. arXiv preprint arXiv: 1704.02799.