10-year CVD Risk Prediction and Minimization via Inverse Classification

Chen Yang*
Health Informatics Program
The University of Iowa
Iowa City,IA 52242
chen-yang@uiowa.edu

W. Nick Street†
Management Sciences Department
The University of Iowa
Iowa City,IA 52242
nick-street@uiowa.edu

Jennifer G. Robinson
Departments of Epidemiology and Medicine
The University of Iowa
Iowa City,IA 52242
jennifer-g-robinson@uiowa.edu

ABSTRACT

Cardiovascular diseases (CVD) remain the leading cause of death around the world. In past decades, many preventive strategies have been recommended to reduce the risk of CVD. However, current CVD risk prediction schemes are not targeted to personalized and optimized recommendations. The goal of this study was to better identify individuals at high risk of a CVD event, and recommend an optimal set of risk factor changes that could reduce the risk of long-term CVD events. We identified 100 demographic, lab, lifestyle, and medication variables for 12907 individuals who participated to the ARIC study and had no CVD events at baseline. We examined the prognostic performance of these features in isolation and ranked them based on mutual information. Then we combined those features to build predictive models using k-nearest neighbor prediction to estimate the 10-year CVD risk for each individual. Our feature-ranking method agreed with traditional risk factors identified by a domain expert. Our approach was successful in identifying cases with high risk and performed as well as traditional methods. Then we applied inverse classification to find the personalized optimal changes to reduce 10-year CVD risk. We also created a personalized package of five optimal changes for each individual to reduce their 10-year CVD risk. This approach can be applied to other chronic disease risk prediction and personalized recommendations, and may be useful to both health care providers and patients in making personalized health care recommendations and decisions.

*This author thanks the University of Iowa Graduate College fellowship for the support.

†The corresponding author.

IHI'12, January 28–30, 2012, Miami, Florida, USA.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and Medical Sciences; I.2.1 [**Computing Methodologies**]: Artificial Intelligence—*Applications and Expert Systems, Medicine and science*

General Terms

Algorithms, Experimentation

Keywords

CVD, Data Mining, Inverse Classification, Risk Prediction

1. INTRODUCTION

Cardiovascular diseases (CVD) include coronary heart disease (CHD), stoke, and peripheral artery disease (PAD). CVDs are the leading cause of death and health care spending both in developed and developing countries across the world [9, 12, 2]. Lifestyles are related with CVD risks. Many research studies have provided evidence that adopting a healthy lifestyle can reduce CVD risk [7, 13]. King *et al.* [7] reported that 1344 (8.5%) of participants in the ARIC study had four healthy lifestyles at visit one, and 970 (8.4%) of the remainder had newly adopted a healthy lifestyle six years later. The total mortality and CVD events were lower for new adopters than for the others.

Preventive approaches in the literature typically focus on those with the highest identifiable risk. Remington *et al.* [11] discussed an approach that tries to identify high-risk individuals through screening and to provide intensive interventions that reduce the risk of disease – for instance, scanning for hypertension and using medications to reduce blood pressure. We wish to extend this idea by identifying the best way to reduce risk for *any* individual.

The main idea of our approach is to build a classifier – a predictive model for estimating the probability of CVD onset – and to work backward through this model, identifying the factor(s) that most affect the risk. As such, we frame the lifestyle recommendation task as an *inverse classification* problem [1]. Inverse classification focuses on the "action oriented" decision features for a test point to perform optimization or decision support. The goal is to determine values for these features that lead the targeted record into the desired predicted class. Our application is to find optimal changeable variable values that reduce an individual's predicted CVD risk to as low a value as possible.

This paper is organized as follows. In Section 2, we introduce related work. Section 3 discusses the dataset and our preprocessing steps. Methods including inverse classification, risk prediction and optimization, and validation are addressed in Section 4. Section 5 presents experimental results. The last sections provides discussion and conclusions.

2. RELATED WORK

Traditional lifestyle recommendations emphasize the population level. The 2006 AHA Scientific Statement [8] provided the following recommendations: "to balance caloric intake and physical activity to achieve and maintain a healthy body weight; consume a diet rich in vegetables and fruits; choose whole-grain, high-fiber foods; consume fish, especially oily fish, at least twice a week; limit intake of saturated fat to <7% of energy, trans fat to <1% of energy, and cholesterol to <300 mg/day by choosing lean meats and vegetable alternatives, fat-free (skim) or low-fat (1% fat) dairy products..." However, it does not take into consideraton the individual's health conditions and personalized choices. Individuals are usually eager to know the single best lifestyle choice they can make in order to reduce their disease risks. Of course, they want to have other options, and to know how much they can benefit from changing their lifestyle. The Framingham score scheme has been widely used by physicians to estimate their patients' 10-year CVD risks by looking up their risk score in a table that maps traditional CVD risk factors like HDL-C, SBP, LDL, age, gender, and others into a real number score and risk probabilities [16]. Chambless *et al.* [3] estimated risk prediction functions for coronary heart disease based on the ARIC study. With additional nontraditional risk factors and markers of subclinical disease, these factors substantially improved prediction of CHD for men, although less so for women.

However, individuals cannot always change these traditional risk factors directly. It may be difficult for them to adopt the recommendations and take actions in their lifestyle. Thus, it's hard for them to achieve their health goals. What's more, they still could not answer the "greedy questions" addressed previously like what is the best thing they can change? How much they can benefit from this change regarding risk reduction?

Classification is one of the most well-studied algorithms in data mining. A classifier is built and trained with instances that have known outcomes. When a new instance is provided, the classifier could assign it to the best-fitting outcome class. A typical example of classification in the medical domain is breast cancer diagnosis, classifying a breast mass as either benign or malignant [10]. In this work we use k-nearest neighbor (k-NN) classification [5]. The k-NN algorithm assigns the query point label by using the weighted votes of its k nearest neighbors. Chi *et al.* [4] used this approach as follows. They developed the Prediction and Optimization-Based Decision Support System (PODSS) algorithm to build a decision support system by using k-NN. The lifestyle variables were changed in a personalized fashion to reduce CVD risk. A novel validation approach was also proposed.

In this study we expand on the inverse classification approach taken in [4] to identify individuals with high 10-year CVD risk, and provide them a package of optimized personalized recommendations to minimize their 10-year CVD risks.

3. DATASET AND PREPROCESSING

We used the Atherosclerosis Risk in Communities (ARIC) study datasets [14]. The Cohort Component and the Community Surveillance Component consist of four communities. The Cohort Component began in 1987 and subjects were examined every three years. ARIC recruited around 4,000 individuals aged between 45 and 64 for each of the four communities.

The total sample size is 15,792. The first screen(baseline) period is 1987-89, and the follow-up periods are 1990-92, 1993-95, and 1996-98. The Community Surveillance Component is the investigation of the community-wide occurrence of hospitalized myocardial infraction and coronary heart disease deaths in men and women aged 35-84 years. Only patients without a CVD event before baseline (1989-89) were included. The 10-year CVD outcomes, including CHD and stroke, were defined based on the Community Surveillance Component.

We selected 100 demographic,laboratory tests, lifestyle, and medication variables by combining feature selection and domain knowledge. We discretized all continuous variables into five equal size groups. Table 1 summarizes the variables. Most of the variables in the table were chosen directly from ARIC dataset except the total of sport hours variable. The ARIC survey asked its participants for their four most common activities and hours per week. The total of sport hours were the sum of all listed activity times. For the measure of physical activity, the ARIC dataset contains information about popular sports and leisure-time activities for individuals. The outcome label of dataset is binary, 1 if a patient has any CVD event in ten years, 0 if not. CHD is defined as the presence of any of the following diagnoses: probable MI, definite MI, suspect MI, missing pain (ECG and/or enzyme diagnosis), definite fatal CHD, definite MI, and possible fatal CHD. Stroke is defined as definite TIB (definite brain infarction, Thrombotic), probable TIB, possible stroke of undetermined type, undocumented fatal cases with stroke codes, and out-of-hospital deaths with stroke codes [4].

The ARIC dietary questionnaire contained variables of the frequency of consumption of various foods over the previous year. From these questions we computed the daily consumption of fruits, vegetable, white carbs, and cereals.

There are missing data in almost all the features in our dataset. k-NN imputation is a robust and sensitive methods to handle missing data [15]. Thus, we applied k-NN imputation to deal with our missing data. The k-NN imputation method searches for an instance's k nearest neighbors based on similarity measurement (we used Euclidean distance) and fills the missing value with the distance-weighted average of its corresponding k nearest neighbors' values. Additionally, all the features were normalized into the domain between 0 and 1, avoiding Euclidean distance domination of large value attributes.

We chose 12907 individuals without CVD at baseline, and the corresponding characteristics are provided in Table 1. In this table, the term cardinality means the average number of values per attribute. We discretized the continuous variables in order to reduce the computational complexity of the inverse classification problem. The University of Iowa IRB has approved this research study.

4. METHODS

4.1 Risk Prediction Algorithm

We now introduce some definitions. Suppose that D^{train} denotes the training dataset, and D^{test} denotes the testing dataset. The training set contains N individuals denoted by $X_1, ..., X_N$. Every individual contains a set V of variables ($|V|$ means the size of set V) and one binary class variable C with values C_1 and C_0. Then we divide the V variables set into three disjoint sets. Let L denote lifestyle variables which can be changeable. M denotes medications, and U denotes unmodifiable variables, and $V = L \cup M \cup U$. Each variable V_j has k possible values, denoted as $V_{j1}, ..., V_{jk}$. For variable in L and M, we apply an exhaustive approach to search all their possible values, and save the optimal changes. The CVD probability for each individual was calculated by using

$$p(C_i|Q) = \sum_{q \in Neighbors(Q)} \frac{I(q_C, C_i) * W(Dist(q,Q))}{\sum_{q \in Neighbors(Q)} W(Dist(q,Q))}. \quad (1)$$

Here q_C is the class membership of query instance Q. Define *Neighbors(Q)* as Q's k nearest neighbors in D^{train}. $W()$ is the distance-weighted function: $\frac{1}{Dist(q,Q)}$. The indicator function I is defined as

$$I(C_i, C_j) = \begin{cases} 1 & \text{iff } C_i = C_j \\ 0 & \text{iff } C_i \neq C_j. \end{cases} \quad (2)$$

We apply Euclidean distance in our similarity measurement. The distance between two individuals is defined as

$$Dist(Q_i, Q_j) = \sqrt{\sum_{d \in V} w(d) * \delta(Q_{id}, Q_{jd})^2}. \quad (3)$$

The function δ is defined as

$$\delta(Q_{id}, Q_{jd}) = \begin{cases} |Q_{id} - Q_{jd}| & \text{if } d \text{ is numeric} \\ 0 & \text{iff } Q_{id} = Q_{jd} \ \& \ d \text{ is nominal} \\ 1 & \text{iff } Q_{id} \neq Q_{jd} \ \& \ d \text{ is nominal.} \end{cases} \quad (4)$$

The weight function for each feature related to the outcome label is defined by mutual information

$$w(d) = \sum_{d \in V} \sum_{c \in C} p(c,d) \log \left(\frac{p(c,d)}{p(c)\, p(d)} \right). \quad (5)$$

4.2 Inverse Classification and Optimization

The inverse classification algorithm searches for the minimum required changes for one instance to be reclassified into a desired class or at least to move as close as possible to that class. In our application, we used the inverse classification algorithm to answer the following questions: What is the single best change to reduce the individual's 10-year CVD risk? What is the optimal set of five changes to reduce the risk? We split our dataset into two equal-sized groups. Our algorithm was then evaluated using a validation method described in Section 4.3. Finally we transformed the data to create the inverse classification problem of identifying cases that would advance to CVD with higher risk, and solved it using k-NN. For our tests we chose the K-nearest neighbors parameter $k = 90$ for risk predication, and $k = 5$ for missing value imputation. We apply inverse classification to provide answers for the previously addressed problems.

Our second goal was to combine the features to achieve the best identification of higher 10-year CVD risk cases. Then, we will find the best action for each individual to minimize the 10-year CVD risk. We formulate this as the following optimization process: For each individual, we searched through one's lifestyle L and medication M variable spaces, tried all feasible discrete values of the variables, saved the optimal single best change and the optimal set of five changes, targeting minimizing 10-year CVD risk defined by equation 1. Recommendation packages of size greater than one were found with a procedure similar to stepwise feature selection, i.e., the best second choice is found in the context of the first best choice, and so on. Algorithm 1 shows the details of our procedure.

Algorithm 1 The 10-year CVD risk prediction and optimization algorithm.

```
Algorithm. CVDRiskPredictionandOptimization(
    Query Q, Training set, Test set)
//parameters
//k1=90 # of neighbors for prediction
//k2=5 # of neighbors for imputation
//S= package size
Begin
For i = 1 to N //number of individuals
  Q' = Q
  For j = 1 to S
    For e in changeable variables
      For p in feasible values(e)
        set value of e to p
        impute imputable variables
        compute probability & save if best so far
      End p
    End e
    update Q' with best variable change
    reset best
  End j
  Test Q'
End i
```

4.3 Validation Method

We tested the performance of the algorithm in a fashion similar to cross-validation, that is, by dividing the dataset into two groups, one for optimization and one for validation, and performing a statistical comparison on their predicted outcomes. We employed a leave-one-out test methodology as depicted in Figure 1. One case was left aside for testing, and a model was built using the remaining cases. The model was then applied to the test case, and the result was recorded. The process was repeated, using each case as a test point. Note that we are turning back the clock, and individuals do not have the chance to actually follow the recommendations provided by our algorithm. Ideally we would have actual clinical trials to validate the outcomes. However, we searched each individual Q's possible changes, and saved the best version of Q'. We then apply Q' to the test set to validate its risk reductions. We implemented this approach through splitting the dataset into two parts as training and testing as described previously. By obtaining both Q and Q's probabilities, we estimated the 10-year CVD risk reductions if they would have followed the recommendations.

5. RESULTS

Table 1: Selected top risk factors of 10-year CVD from ARIC dataset via mutual information

Variable	Variable_names	weight_of_variable_by_MI	Modifiable
DIABTS02	Diabetes	0.01126282	No
MSRA08F	Blood Sugar REG MED past 2 weeks	0.011209538	Medication
HYPERT05	Hypertension	0.009123591	No
GENDER	Sex	0.009044474	No
MSRA08A	High BP MED in past 2 weeks	0.00892728	Medication
WSTHPR01	Waist-to-hip ratio	0.008699026	No
HDL01	Re-calibrated HDL CHOL. in mg/dl	0.007586154	Imputable
HEMA09	Fibrinogen value	0.007219591	No
TRGSIU01	Total triglycerides in mmol/L	0.006746994	Imputable
LVHSCR01	LVH numeric left ventricular hypertrophy	0.006096085	No
CIGTYR01	Cigarette years of smoking	0.005659599	No
CHMA15	Uric acid (MG-DL)	0.005179445	No
SBPA21	2nd AND 3rd systolic BP average	0.005070901	Imputable
LIPA07	Apolipoprotein B (MG-DL)	0.005006796	No
ANTZ04	Weight to the nearest LB	0.004832679	Imputable
CHMA09	Creatinine (MG-DL)	0.004794068	No
INTPLQ01	Plaque in either internal carotid	0.004688219	No
INTPS01	Plaque/shadowing in either internal	0.004667392	No
CHMA16	Insulin (UU-ML)	0.004443084	No
V1AGEZ1	Age at Visit 1	0.004064866	No
LIPA06	Apolipoprotein AI (MG-DL)	0.003948518	No
ANTZ07A	Waist girth to nearest CM	0.003943412	No
HMTA03	White blood count Q3	0.003902625	No
HMTA01	Hematocrit Q1	0.003827918	No
CHOL	Dietary cholesterol (mg)	0.003744852	Lifestyle
LDL02	Re-calibrated LDL CHOL. in mg/dl	0.003354329	Imputable
FEV1FVC1	FEV(1)/FVC predicted (%)	0.003338196	No
EVRSMK01	Ever smoked cigarettes	0.003320216	No
HMTA02	Hemoglobin	0.003248267	No
ANTZ01	Standing height to nearest CM	0.0028815	No
BMI01	Body mass index in KG/(M*M)	0.002858455	No
RACEGRP1	Race	0.002851487	No
P_MFAT	Monounsaturated fatty acid (%kcal)	0.002720198	Lifestyle
SBPA22	2nd and 3rd diastolic BP average	0.002668821	Imputable
ANTA05B	Triceps measure 2 to nearest MM	0.002583063	No
HEMA17	VWF value	0.002556059	No
P_TFAT	Total fat (%kcal)	0.00251638	Lifestyle
MENOPS01	Menopausal status	0.002422043	No
TFAT	Total fat (g)	0.002391889	Lifestyle
AFAT	Animal fat (g)	0.002373213	Lifestyle
HOM62	Combined family income group	0.002368665	No
HEMA07	VIII: C Value	0.001963566	No
SFAT	Saturated fatty acid (g)	0.001862782	Lifestyle
P_CARB	Carbohydrate(%kcal)	0.001840513	Lifestyle
TCHSIU01	Total cholesterol in mmol/L	0.001759625	Imputable
HOM32	Number of cigarettes per day	0.001693043	Lifestyle
SODI	Sodium	0.001610455	Lifestyle
fruits	Fruits	0.001489372	Lifestyle
MSRA08E	Blood thinning MED in past 2 weeks	0.001439642	No
PAD01	Peripheral Artery Disease (PAD)	0.00142109	No
P_AFAT	Animal fat (%kcal)	0.001418874	Lifestyle
CHMA11	Magnesium (MEQ-L)	0.001374897	No
LIPA08	Lipoprotein Lp (a) Data (μg/mL)	0.001215193	No
P_SFAT	Saturated fatty acid (%kcal)	0.001192481	Lifestyle
HMTA06	Neutrophil bands	0.001018264	No
PAD02	Pad02(Peripheral Artery Disease, V1-Def 2., Abi<.9 For Both Genders)	0.00093554	No
ELEVEL01	Education level	0.000920468	No
DTIA66	How often eat liver	0.00091408	No
PFAT	Polyunsaturated fatty acid (g)	0.000894985	Lifestyle
ANTA06B	Subscaps measures 2 to nearest MM Q6B	0.000789857	No
CHMA08	Urea nitrogen (MG-DL)	0.000785857	No
CHMA06	Potassium (MMOL-L)	0.00077702	No
MSRA08C	Heart rhythm control medication past 2 weeks	0.000604278	No
VFAT	Vegetable fat (g)	0.000587033	Lifestyle
HMTA04	Platelet count	0.000480867	No
MOMHXDIA	Maternal history of diabetes	0.000462751	No
CHOLMD01	Cholesterol lowering medication use	0.000425656	Lifestyle
DFIB	Dietary fiber (g)	0.000391932	Lifestyle
PFTA26	FEV(1) (liters)	0.000389209	No
HMTA07	Lymphocytes	0.000335151	No
DTIA51	Consumed dark or grain breads	0.000304759	Lifestyle
CHMA13	Albumin (GM-DL)	0.000199372	No
MSRA10	Take pain MEDS in past 2 weeks	0.000183089	No
OMEGA	Omega fatty acid w20:5 and w22:6 (g)	0.000157228	Lifestyle
CHMA05	Sodium(Na) (MMOL-L)	0.000156103	No
DTIA55	Consumed nuts	0.00013099	Lifestyle
ABI04	Ankle-brachial index ABI	0.000113244	No
ALCO	Alcohol intake (g) per day	8.79401E-05	Lifestyle
FAMHXCHD	Family history of CHD	8.19693E-05	No
DADHXCHD	Paternal history of CHD	8.19693E-05	No
CHMA10	Calcium	7.40834E-05	Lifestyle
P_PFAT	Polyunsaturated fatty acid (%kcal)	5.88378E-05	Lifestyle
ECGMA31	Heart rate	5.86203E-05	No
cereals	Cereals	5.72606E-05	Lifestyle
MOMHXSTR	Maternal history of stroke	5.47992E-05	No
MOMHXCHD	Maternal history of CHD	5.19767E-05	No
P_VFAT	vegetable fat (%kcal)	4.20168E-05	Lifestyle
PFTA24	FVC (liters)	3.547E-05	No
fish	Fish	3.21844E-05	Lifestyle
P_PROT	Protein (%kcal)	2.76504E-05	Lifestyle
MSRA09	Aspirin in past 2 weeks	2.47184E-05	Medication
SPORT_HRS	Sport hours	2.35069E-05	Lifestyle
FAMHXSTR	Family history of stroke	1.06002E-05	No
DADHXSTR	Paternal history of stroke	1.06002E-05	No
white_carbs	White_carbs	6.47352E-06	Lifestyle
CARB	Carbohydrate (g)	2.66155E-06	Lifestyle
FAMHXDIA	Family history of diabetes	1.72746E-06	No
DADHXDIA	Paternal history of diabetes	1.72746E-06	No
vegetables	vegetables	6.54653E-08	Lifestyle

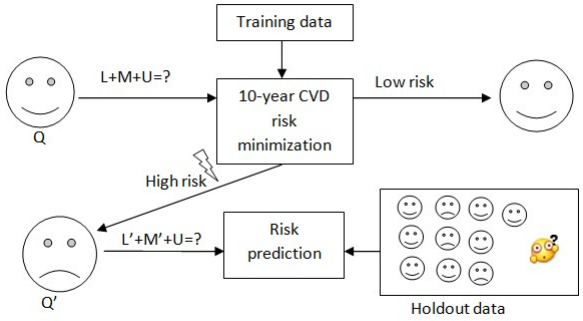

Figure 1: The validation method

Table 1 shows the attribute importance for 10-year CVD risk prediction. These variables were selected from the ARIC study via mutual information feature selection method. To evaluate the performance of our classifier, Figure 2 shows ROC curve obtained from our mutual information based K-NN classifier. The AUC is 0.7456, and k=90 for risk prediction.

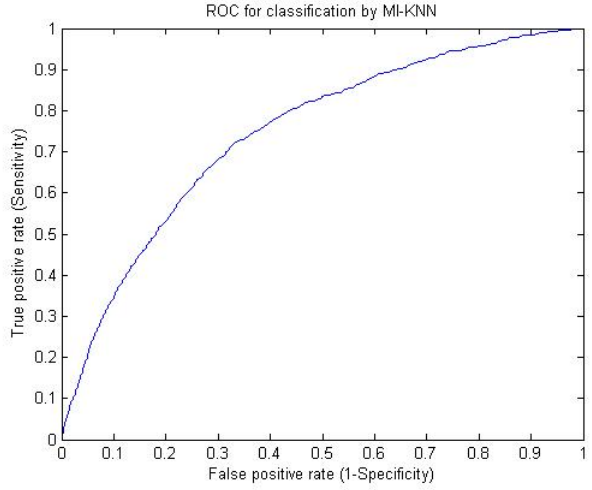

Figure 2: ROC curve obtained from the mutual information based K-NN classifier

We can provide a single best personalized recommendation list for each individual. For instance, the following individual's best single change is to reduce animal fat(g). This person's original value of animal fat(g) was 86.19, discretized to a value in the highest level 1. The five levels of the variable are 0, 0.25, 0.50, 0.75, and 1. As shown in Table 2, if the individual changed this variable from level 1 to level 0, the maximum absolute 10-year CVD risk would reduce 7.96%, for a relative risk reduction of 44.28%. Note the original 10-year CVD risk of the individual is 17.98%. 10.02% was the optimized 10-year CVD risk after single variable change. If the individual changed this variable from level 1 to level 0.50, the absolute 10-year CVD risk would reduce 5.87%, for a relative risk reduction of 32.66%.

We report the results of packages of five variable changes to optimally lower the 10-year CVD risk. We excluded cases for which our method could not help lower their risks. For example, those individuals with CVD probabilities of zero, and those with very small risk probabilities were excluded. We also did not include those recommendations that reduce the individuals' risk to zero, considering these extreme cases to be impractical in real life. Feasible changes were defined by our clinical expert as follows. If someone is on a medication, they need to stay on it. Individuals will not be recommended to increase their traditional unhealthy lifestyle variable values, and individuals will not be recommended to decrease their traditional healthy variable values. For instance, if individuals are not current smokers, they will not be recommended to smoke; if individuals are current smokers, they will not be recommended to increase the number of cigarettes.Table 3 shows the popular recommended changes.

Figure 3 shows individuals' 10-year CVD absolute risk probabilities of original vs. minimized risks via a package of five personalized recommendations. For the original risk, the statistics are as follows: min: 1.03%; mean: 9.92%; median: 8.61%; max: 38.64%. For the minimized risk, the statistics are as follows: min: 1.00%; mean: 4.84%; median: 3.42%; max: 27.33%. For the absolute reductions, the statistics are as follows: min: 0%; mean: 5.08%; median: 4.29%; max: 31.77%. For the relative risk reductions, the statistics are as follows: min: 0%; mean: 50.62%; median: 51.55%; max: 96.23%.

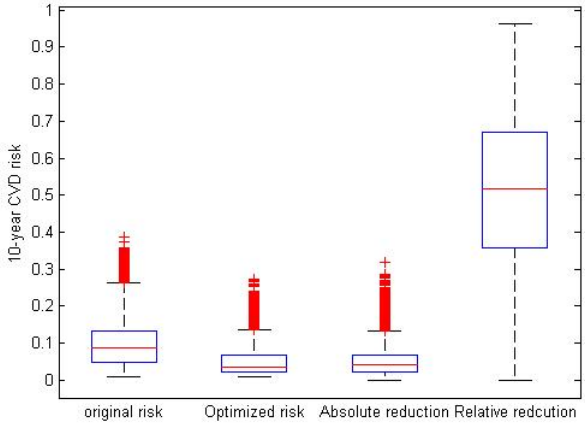

Figure 3: 10-year CVD absolute risk optimization via a package of five recommendations

An example is shown in Table 4, and the max reductions are cumulative. If this individual changes their total fat(g) level from level 0.5 to level 0.25, then the 10-year CVD absolute risks will reduce 8.66% (33.89% relative). Based on this change, if they also reduce their animal fat intake from level 1 to level 0, their absolute risk will reduce 8.77% (34.32%). If they follow all the changes, they will reduce their absolute risk from original the 25.55% to 11.10%. The relative risk reduction is 56.56%.

6. DISCUSSION AND CONCLUSION

Table 2: A list of all feasible changes and their absolute 10-year risks of a single variable for an individual

User_ID	BestVariable	OriginalValue	FeasibleChange	RiskReduction(Relative)	OriginalRisk	OptimizedRisk
1	Animal fat(g)	1	0	**7.96(44.28)%(max)**	17.98%	10.02%
			0.25	6.91(38.45)%		11.07%
			0.50	5.87(32.66)%		12.11%
			0.75	5.86(32.59)%		12.12%

Table 3: The most popular recommended changes ordered by frequencies

Variables	Variable names	Frequency counts	Percentage	Change directions
CHOL	Dietary cholesterol (mg)	3076	12.51%	Decrease
TFAT	Total fat (g)	2727	11.09%	Decrease
P_MFAT	Monounsaturated fatty acid (%kcal)	2436	9.91%	Decrease
P_TFAT	Total fat (%kcal)	1897	7.72%	Decrease
Fruits	Fruits	1678	6.83%	Increase
SODI	Sodium (mg)	1591	6.47%	Decrease
AFAT	Animal fat (g)	1561	6.35%	Decrease
MSRA08F	Blood sugar REG MED past 2 weeks?	1329	5.41%	Take
HOM32	Number of cigarettes per day	1301	5.29%	Decrease
P_AFAT	Animal fat (%kcal)	1261	5.13%	Decrease
P_SFAT	Saturated fatty acid (%kcal)	1149	4.67%	Decrease
MSRA08A	High BP MED in past 2 weeks?	815	3.32%	Take
VFAT	Vegetable fat (g)	739	3.01%	Increase
PFAT	Polyunsaturated fatty acid (g)	671	2.73%	Increase
DFIB	Dietary fiber (g)	640	2.60%	Increase
CHOLMD01	Cholesterol lowering medication use	558	2.27%	Take
OMEGA	Omega fatty acid w20:5 and w22:6 (g)	336	1.37%	Increase
DTIA51	Consumed dark or grain breads	281	1.14%	Increase
ALCO	Alcohol intake (g) per day	175	0.71%	Decrease
Cereals	Cereals	115	0.47%	Increase
P_PFAT	Polyunsaturated fatty acid (%kcal)	80	0.33%	Increase
P_PROT	Protein (%kcal)	38	0.15%	Increase
P_VFAT	Vegetable fat (%kcal)	34	0.14%	Increase
MSRA09	Asprin in past 2 weeks	29	0.12%	Take
Fish	Fish	23	0.09%	Increase
SPORT_HOURS	Sport hours	15	0.06%	Increase
DTIA55	Consumed nuts	13	0.05%	Increase
P_CARB	Carbohydrate(%kcal)	8	0.03%	Decrease
White_Carbs	White carbs	5	0.02%	Decrease
CARB	Carbohydrate (g)	1	0.00%	Decrease

In this study, we obtained dataset from ARIC Study, and proposed personalized lifestyle and medication recommendations via inverse classification. The accuracy of prediction by our classifier was measured as AUC=0.75, consisted with widely accepted Framingham Heart Study with AUC=0.78, showing better AUC=0.72 than the prediction in ARIC study presented by Grundy et.al[6]. For each individual, our algorithm provides the one best change and the package of five adjustments to minimize their 10-year CVD risks. King et al [7] found that 1344 (8.5%) individuals in ARIC study visit 1 had the following 4 lifestyle habits: five or more fruits and vegetables daily, regular exercise, BMI 18.5-29.9 kg/m2, and no current smoking. Six years later, there were 970 (8.4%) individuals had newly adopted a healthy lifestyle. They also concluded that by adopting one healthy lifestyle in middle age can reduce CVD risks by an estimated 35%. Our approach is different from other research studies that were performed by traditional Cox regression survival analysis approach. Yang et al. [17] proposed a data mining approach that improved on the traditional Cox regression model in their MPGN type II patients survival analysis. This data mining approach would likely improve the accuracy and significance of the results. It can guide health care providers to develop more individualized treatment actions based on identifying of individuals who are at increased risk of CVD.

The findings from this study could be used to identify high-risk individuals of CVD prevention. This is a very important issue in public health area to prevent CVD. Since current medical prognostic prediction for the CVD is largely based on the practitioner's personal experience, this study provides a step toward data-driven clinical decision support. The model also adds information to aid decision making. The patient can decide which risk factor changes they may wish to undertake first. This approach could also provide personalized packages of healthy changes for high-risk individuals to minimize their CVD risks.

Our results show that changing the lifestyle and medication variables in the reverse classification approach can determine which patients are more likely to progress to CVD in

Table 4: Optimal package of five recommendations for individual 5 to minimize their 10-year CVD risks

Best_Variable	original_value	best_value	max_reduction	original_prob	after_opt	relative_reduction
Total fat (g)	0.5	0.25	0.0866	0.2555	0.1689	0.3389
Animal fat (g)	1	0	0.0877	0.2555	0.1678	0.3432
Monounsaturated fatty acid (%kcal)	0.75	0	0.1006	0.2555	0.1549	0.3939
Total fat (%kcal)	0.75	0	0.1335	0.2555	0.1220	0.5227
Saturated fatty acid (%kcal)	0.75	0	0.1445	0.2555	0.1110	0.5656

10 years. To make the process more realistic, we could limit the search to all the feasible changes for this individual. Future work would include modeling the risk prediction and optimization over time, and deal with confounding factors that mislead the algorithm to produce improper recommendation like increase smoking. The medication problem is more complicated, taking blood-pressure-lowering medication as an example, those individuals who had high blood pressure would be prescribed to take medications to reduce their blood pressure, and this group of individuals had higher risk than the blood pressure normal group without medication. Additionally, there were few untreated high blood pressure individuals in our dataset. Even more, the effect of lowering blood pressure through medication is complicated itself. Thus, these make the recommendations for medication become tricky like not taking blood-pressure-lowering medication. CVD is a set of very common diseases with many risk factors and long processes. The mutual information based k-NN classifier can be used to predict CVD risks and can be applied to identify high-risk individuals. We used a reverse classification approach to provide optimized recommendations that could be useful to develop clinical interventions and specific individual care plans. Our inverse classification approach is novel to provide CVD risk predictions and optimizations compared to the conventional hazard ratio risk prediction scores for personalized and optimized recommendations in such applications. The results are also generalizable to other diseases.

7. ACKNOWLEDGMENTS

The Atherosclerosis Risk in Communities (ARIC) study is conducted and supported by the NHLBI in collaboration with the ARIC study investigators. This research was performed by using a limited access dataset obtained by the NHLBI and does not necessarily reflect the opinions or views of the ARIC Study or the NHLBI.

8. REFERENCES

[1] C. Aggarwal, C. Chen, and J. Han. The inverse classification problem. *Journal of Computer Science and Technology*, 25(3):458–468, 2010.

[2] D. L. Bhatt, P. G. Steg, E. M. Ohman, A. T. Hirsch, Y. Ikeda, J.-L. Mas, S. Goto, C.-S. Liau, A. J. Richard, J. Röther, P. W. F. Wilson, and for the REACH Registry Investigators. International prevalence, recognition, and treatment of cardiovascular risk factors in outpatients with atherothrombosis. *JAMA: The Journal of the American Medical Association*, 295(2):180–189, 2006.

[3] L. Chambless, A. Folsom, A. Sharrett, P. Sorlie, D. Couper, M. Szklo, and F. Nieto. Coronary heart disease risk prediction in the atherosclerosis risk in communities (ARIC) study. *Journal of Clinical Epidemiology*, 56(9):880–890, 2003.

[4] C.-L. Chi, W. N. Street, and J. G. Robinson. Comparing predictive effects to create an individualized lifestyle recommendation. *Journal of Biomedical Informatics*, under review.

[5] T. Cover and P. Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27, 1967.

[6] S. M. Grundy, R. B. D'gostino, Sr, L. Mosca, G. L. Burke, P. W. Wilson, D. J. Rader, J. I. Cleeman, E. J. Roccella, J. A. Cutler, and L. M. Friedman. Cardiovascular risk assessment based on us cohort studies. *Circulation*, 104(4):491–496, 2001.

[7] D. King, A. Mainous III, and M. Geesey. Turning back the clock: Adopting a healthy lifestyle in middle age. *The American Journal of Medicine*, 120(7):598–603, 2007.

[8] A. Lichtenstein, L. Appel, M. Brands, M. Carnethon, S. Daniels, H. Franch, B. Franklin, P. Kris-Etherton, W. Harris, B. Howard, et al. Diet and lifestyle recommendations revision 2006: A scientific statement from the American Heart Association Nutrition Committee. *Circulation*, 114(1):82, 2006.

[9] D. Lloyd-Jones, Y. Hong, D. Labarthe, D. Mozaffarian, L. Appel, L. Van Horn, K. Greenlund, S. Daniels, G. Nichol, G. Tomaselli, et al. 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, 2010.

[10] O. Mangasarian, W. Street, and W. Wolberg. Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, 43(4):570–577, 1995.

[11] P. Remington, R. Brownson, and M. Wegner. *Chronic Disease Epidemiology and Control*. Number Ed. 3. American Public Health Association, 2010.

[12] V. Roger, A. Go, D. Lloyd-Jones, R. Adams, J. Berry, T. Brown, M. Carnethon, S. Dai, G. de Simone, E. Ford, et al. Heart disease and stroke statistics–2011 update: A report from the American Heart Association. *Circulation*, 123(4):e18, 2011.

[13] M. Stampfer, F. Hu, J. Manson, E. Rimm, and W. Willett. Primary prevention of coronary heart disease in women through diet and lifestyle. *New England Journal of Medicine*, 343(1):16–22, 2000.

[14] The ARIC Investigators. The atherosclerosis risk in communities (ARIC) study: design and objectives. *American Journal of Epidemiology*, 129(4):687–702, 1989.

[15] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520, 2001.

[16] P. Wilson, R. D'Agostino, D. Levy, A. Belanger, H. Silbershatz, and W. Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837, 1998.

[17] C. Yang, N. W. Street, D.-F. Lu, and L. Lanning. A data mining approach to mpgn type II renal survival analysis. In *Proceedings of the 1st ACM International Health Informatics Symposium*, IHI '10, pages 454–458, New York, NY, USA, 2010. ACM.