Journal of Biomedical Informatics xxx (2012) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

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

Individualized patient-centered lifestyle recommendations: An expert system for communicating patient specific cardiovascular risk information and prioritizing lifestyle options

Chih-Lin Chi [a,*], W. Nick Street [b], Jennifer G. Robinson [c], Matthew A. Crawford [a]

[a] Center for Biomedical Informatics, Harvard Medical School, 10 Shattuck Street, Boston, MA 02115, USA
[b] Management Sciences Department and Interdisciplinary Graduate Program in Informatics, S210 Pappajohn Business Building, The University of Iowa, Iowa City, IA 52242, USA
[c] Department of Epidemiology, University of Iowa, Lipid Research Clinic, University of Iowa, USA

ARTICLE INFO

Article history:
Received 17 February 2011
Accepted 27 July 2012
Available online xxxx

Keywords:
Individualized lifestyle recommendation
Decision support systems
Patient centered medicine
Machine learning
k-nearest neighbor
Optimization

ABSTRACT

We propose a proof-of-concept machine-learning expert system that learned knowledge of lifestyle and the associated 10-year cardiovascular disease (CVD) risks from individual-level data (i.e., Atherosclerosis Risk in Communities Study, ARIC). The expert system prioritizes lifestyle options and identifies the one that maximally reduce an individual's 10-year CVD risk by (1) using the knowledge learned from the ARIC data and (2) communicating for patient-specific cardiovascular risk information and personal limitations and preferences (as defined by variables used in this study). As a result, the optimal lifestyle is not only prioritized based on an individual's characteristics but is also relevant to personal circumstances.

We also explored probable uses and tested the system in several examples using real-world scenarios and patient preferences. For example, the system identifies the most effective lifestyle activities as the starting point for an individual's behavior change, shows different levels of BMI changes and the associated CVD risk reductions to encourage weight loss, identifies whether weight loss or smoking cessation is the most urgent change for a diabetes patient, etc. Answers to the questions noted above vary based on an individual's characteristics. Our validation results from clinical trial simulations, which compared original with the optimal lifestyle using an independent dataset, show that the optimal individualized patient-centered lifestyle consistently reduced 10-year CVD risks.

1. Background

Diet patterns and healthy behaviors are important for disease prevention, especially for cardiovascular health. What lifestyle choices are healthy? The AHA Scientific Statement [19] recommends aiming for a healthy body weight, desired lipid profile, normal blood pressure, physical activity, and not smoking, etc. It also suggests specific thresholds for diet and lifestyle such as the limitation of saturated fat intake to <7% of energy, cholesterol intake <300 mg, and total fat intake between 25% and 35% of energy.

These are generalized recommendations optimized for the whole population without considering individuals' particular circumstances and characteristics. For behavior interventions, such as lifestyle changes, an individual's commitment and willingness to change play an important role in the success of such interventions. Tailoring individualized strategies in lifestyle changes has been broadly discussed as a way to promote health because they are more personally relevant. The personal relevance leads to the individual taking a more active role in their health, reading the literature more thoroughly, and discussing with others more often. (e.g., [14,17,8,25,21]). Such a tailoring approach has been shown to be an effective method for most studies of nutrition interventions and partial studies of physical activity to promote health [14].

The basic idea of tailored interventions is to use communication, medication, or other types of treatments that are specific for an individual or a group to improve health or change behaviors [1]. The proposed decision support system provides four key functions for tailored interventions (individualized lifestyles): (1) The system predicts 10-year CVD risks based on an individual's characteristics. We consulted a preventive clinic doctor to identify a set of variables that can be easily obtained and are often considered in clinics (Table 1) for lifestyle recommendations. (2) The system prioritizes lifestyle options and actively identifies the lifestyle that would most reduce CVD risks, based on the individual's characteristics. (3) Through communication to collect personally relevant limitations (such as personal food and time limitations) and preferences, the system actively identifies the most effective lifestyle based on these parameters to produce effective and easily-complied individualized lifestyle recommendations. (4) The system

* Corresponding author. Fax: +1 617 432 0693.
E-mail address: Chih-Lin_Chi@hms.harvard.edu (C.-L. Chi).

1532-0464/$ - see front matter © 2012 Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.jbi.2012.07.011

Table 1
Variables from ARIC data used for lifestyle recommendation.

#	Types	Variables [values]
Lifestyle		
1	NUM	Body mass index [23.35, 25.75, 28.06,31.38]
2	NUM	Alcohol intake (g) per day [0, 9.43]
3	Nominal	Smoking status [0, 1]
4	NUM	Total activity hours per week [3, 5, 7, 10]
5	NUM	Carbohydrate (g) [128.79, 166.19, 203.57, 258.52]
6	NUM	Dietary cholesterol (mg) [147.5, 200.23, 256.84, 337.56]
7	NUM	Dietary fiber (g) [10.52, 14.12, 17.82, 22.93]
8	NUM	Protein (% kcal) [14.52, 16.68, 18.67, 21.08]
9	NUM	Saturated fatty acid (% kcal) [9.55, 11.25, 12.68, 14.37]
10	NUM	Total fat (% kcal) [27.32, 31.41, 34.71, 38.42]
Characteristics		
11	NUM	Cigarette years of smoking [0, 280, 660]
12	Nominal	Cholesterol lowering medication use [0, 1]
13	Nominal	Diabetes [0, 1]
14	Nominal	Education level [(1) grade school or 0 years education, (2) high school, but no degree, (3) high school graduate (4), vocational school, (5) college (6) graduate school or professional school]
15	Nominal	sex [0, 1]
16	NUM	HDL cholesterol in mg/dl [37.56, 45, 52.97, 64.52]
17	Nominal	Hypertension [0, 1]
18	NUM	LDL cholesterol in mg/dl [105, 126, 145, 168]
19	Nominal	Menopausal status [(1) primary amenorrhea, (2) premenopause, (3) perimenopause, (4) post, natural, (5) post, surgical, (6) unknown ovarian status]
20	Nominal	Race [B: black, N: non-black]
21	NUM	Total cholesterol in mmol/L [4.65, 5.22, 5.74, 6.39]
22	NUM	Total triglycerides in mmol/L [0.82, 1.08, 1.41, 1.94]
23	NUM	age [48, 52, 56, 60]
24	Nominal	High blood pressure medication in past 2 weeks [yes, no, unknown]
25	NUM	2nd and 3rd Systolic blood pressure average [106, 115, 123, 135]
26	NUM	2nd and 3rd Diastolic blood pressure blood pressure average [65, 70, 76, 82]
27	NUM	Standing height to nearest CM [160, 165, 171, 177]
28	NUM	Waist girth to nearest CM [85, 93, 99, 107]
29	NUM	Hip girth to nearest CM [97, 101, 105, 111]
30	NUM	Heart rate [58, 63, 68, 75]
31	NUM	White blood count [4.6, 5.4, 6.3, 7.4]
32	NUM	Apolipoprotein AI (MG-DL) [107, 122, 137, 157]
33	NUM	Apolipoprotein B (MG-DL) [69, 83, 97, 116]
34	NUM	APOLP (A) DATA (UG-ML) [19, 43, 86, 175]
35	NUM	Creatinine (MG-DL) [0.9, 1, 1.1, 1.2]

allows the individual to choose or prioritize what they want to do first based on how much the change would benefit them as an individual, rather than a one size fits all approach of general recommendation.

Typical expert systems contain a knowledge source and a mechanism for problem solving that returns a response based on a query. The knowledge source of most expert systems, such as the famous example of MYCIN [24], consists of rules derived from direct input from domain experts and evidence from the literature. However, acquiring and maintaining knowledge in this form is time- and labor-intensive [29].

Other systems avoid the knowledge acquisition problem using machine-learning methods for inference, and are thus based exclusively on data. Supervised machine learning involves constructing a mapping from independent variables, or features, to known outcomes. The resulting decision function transfers values of independent variables into a predicted value of the dependent variable.

Decision functions exist in several forms such as linear or nonlinear functions, tree, rules, and data. In this project, we used a lazy learning approach *k*-Nearest Neighbor (*k*-NN) [13] as the supervised classifier. Lazy learning – also called instance-based, case-based, or memory-based learning – builds a prediction model specifically for the query case. For example, a *k*-NN classifier finds the *k* closest training cases to decide the label for a query case. Lazy learning algorithms show three types of properties [4]. First, the classifiers defer processing of the output until a query case appears. Second, their responses combine the training with the query information. Third, they discard the constructed answer and any intermediate results.

Eager learning algorithms, such as support vector machines (SVMs), artificial neural networks (ANNs), and decision trees, compile data in advance and use it to construct a predictive model. They use this global model to give responses to all queries. Thus, the compilation process is called training, which is unnecessary in lazy learning methods. Instead of training, lazy learning algorithms need to store all the training cases, and find the closest training cases to the query case for prediction. Compared to black box approaches such as SVMs and ANNs, lazy learning is more interpretable. However, lazy learning is sensitive to irrelevant attributes [3]. Feature weighting methods such as mutual information [31], conceptually similar to feature selection methods, can solve the problem.

We modified the prediction and optimization-based decision support system (PODSS) algorithm [10] to build this expert system. The PODSS algorithm allows the choice of any classification method that is deemed appropriate for a particular problem. We choose *k*-NN as our classifier for both conceptual and clinical reasons. Conceptually, CVD risks are computed from people with similar characteristics. This transparency allows us to easily visualize and understand how CVD risks are computed, and why the system concluded that a particular change would result in a risk reduction. Clinically, due to the nature of data storage for prediction, a patient can visualize people's (de-identified) lifestyle and outcomes. We expect such a visualization mechanism will enable observational learning and further help healthy lifestyle adoption. In fact, in social cognitive theory [6,12], observing other people's behavior is one important element in self-efficacy that improves an individual's confidence in successfully carrying out a behavior.

Predictive performance is evaluated by comparing true with predictive labels. This is a process called validation. In order to implement this idea, one needs to separate a dataset into two subsets, one for training and the other one for testing. Training set is used to generate a predictive model, and the predictive performance is evaluated on the testing set. Leave-one-out [27] is an extreme example with low variance and bias but high computation requirement. In leave-one-out, a dataset, which consists of n data points, is separated into 1 testing data point and $n-1$ training data points. The same process is repeated for n times, each of which has a different testing data point.

A variety of machine learning methods have been used to construct the knowledge base for expert systems. As mentioned, in nearest-neighbor or case-based prediction [28], the knowledge consists solely of previous cases, including the problem, the solution, and the outcome, stored in a central location, called the case library. To obtain the solution for a new case, one simply identifies the stored case(s) most similar to the problem, and the proposed solution can be adapted from the retrieved case. More generally, machine learning (ML)-based expert systems are able to give recommendations that are generated by non-linear forms of knowledge, and are easily updated by simply adding new cases. However, the use of ML in expert systems has been limited (see, e.g., [5,30] typically involving rule induction). Systems built on artificial neural networks include Liau et al. [18] for oil distillation and Song and Kusiak [26] for boiler control. Chi et al. [10] built a hospital-selection expert system that combined support vector machines with optimization.

Generally, the types of solutions that can be structured by expert systems can be divided into selection and construction [11]. For selection, several action sets have been pre-determined, and the solution is the most promising action set. On the other hand,

Please cite this article in press as: Chi C-L et al. Individualized patient-centered lifestyle recommendations: An expert system for communicating patient specific cardiovascular risk information and prioritizing lifestyle options. J Biomed Inform (2012), http://dx.doi.org/10.1016/j.jbi.2012.07.011

construction needs to construct a set of actions from scratch. In order to avoid infeasible solutions, constraints can regulate the solution construction. In the hospital-selection problem [10], each hospital has a unique set of characteristics (e.g. teaching status, bed size, volume for surgeries, etc.) that make up a pre-determined action set and one has to select the most promising action set that represents a real hospital; we cannot construct an ideal hospital using characteristics of different ones. Therefore, it is a selection problem. On the other hand, the lifestyle problem in this project was construction. There were no pre-determined lifestyle options, so we needed to construct a set of lifestyle values (e.g., BMI $\leqslant$ 23.35, quit smoking, and physical activity $\geqslant$ 10 h/week) Then we determined the combination that maximally reduced an individual's cardiovascular disease risk. We note that both hospital-selection and lifestyle expert systems were built by the same principle, prediction and optimization. Due to the requirement of different solution types, we needed to choose a different optimization tool to better suit the problem nature.

2. Method

We used the revised PODSS algorithm [10] to construct this expert system. The system learned the knowledge of lifestyle with the associated CVD risks from the ARIC data (Section 2.1) using *k*-NN prediction models (described in Section 2.2). Missing data was imputed by the expected distance between a query and a training case (Section 2.3). The individualized patient-centered lifestyle was identified by optimization approaches (Section 2.4). When complying with a recommended lifestyle, the lifestyle coupled with biological changes interactively influenced CVD risks. Section 2.5 discusses this issue and uses the combined changes to predict an individual's modified CVD risk. Finally, we discuss a validation approach using clinical trial simulations (Section 2.6).

2.1. Data preparation

Knowledge for the system was extracted from the data of the Atherosclerosis Risk in Communities (ARIC) study [15]. This study contains a Cohort Component and a Community Surveillance Component for four communities. The Cohort Component began in 1987 and subjects were examined every 3 years. ARIC recruited around 4000 individuals aged 45–64 from each of the four communities. As a result, the total sample size is 15,792.

The 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 infarction and coronary heart disease (CHD) deaths in men and women aged 35–84 years. Patients with any cardiovascular disease (CVD) event before the baseline (1987–89) are excluded. The sample size in this study is 13,006. 10-year CVD outcomes (including both CHD and stroke) are defined using the Community Surveillance Component.

We asked a preventive clinic doctor (one of the co-authors) to decide variables that can be obtained from usual care. The data was obtained from patient self-report on questionnaires or examination by study protocol. We discretized all variables based on an equal population (i.e., similar population size in each discrete value) in order to simplify the problem. Variables are classified into patient characteristics and lifestyle. Patient characteristics describe a patient and are fixed. On the other hand, lifestyle variables are changeable and can be changed to improve health. Table 1 summarizes both types of variables and their cutpoints for discretization. For example, there are five discrete values for body mass index, less than 23.35, between 23.35 and 25.75, between 25.75 and 28.06, between 28.06 and 31.38, more than 31.38 based on quantiles. The ARIC survey asked participants the four most common activities and hours per week they perform, and total activity hours is the sum of all activity time. The outcome is binary, whether a patient has any CVD event (CHD event and stroke) over the 10 years of follow-up. CHD is defined as any of the following diagnoses: probable myocardial infarction (MI), definite MI, suspect MI, definite fatal CHD, definite MI, and possible fatal CHD. Stroke is defined as definite Thrombotic (TIB) (brain infarction, Thrombotic), probable TIB, possible stroke of undetermined type, undocumented fatal cases with stroke codes, and out-of-hospital deaths with stroke codes.

2.2. k-Nearest neighbor

k-Nearest Neighbor (*k*-NN) does not compile a universal predictive model in advance. It postpones induction until classification. In other words, it stores all the training data and predicts by utilizing the distance-weighted true classes of the query's *k* nearest neighbors [31]. The choice of the *k* influences the system's performance on CVD risk prediction. A small *k* causes overfitting and, more importantly, in many cases, the expert system cannot find lifestyle recommendation because of the difficulty to find similar persons with both good and bad outcomes. On the other hand, a large *k* may underfit, but the probability transition and CVD risk estimation are much smoother. Since we are focused on producing high-quality recommendations, we use data size as the *k* for this pilot project.

The model is described as

$$p(c_j|q) = \frac{\sum_{x\in K_q} 1(x_c = c_j)\cdot K(d(x,q))}{\sum_{x\in K_q} K(d(x,q))}, \tag{1}$$

where $c_j \in 1,\ldots,J$ is one of the J possible classes and x_c is the class membership of query q. 1() is 1 iff the argument is true, and $1(x_c = c_j)$ defines the specific class to which a query q belongs. K is the distance weight function, and K_q is the set of q's k nearest neighbors among the training data. The distance function between q and x is defined as

$$d(x,q) = \left(\sum_{f\in F} w(f)\cdot\delta(x_f,q_f)^r\right)^{\frac{1}{r}}, \tag{2}$$

where F is the feature set. In this project, we define $r = 2$ (i.e., Euclidean distance). $\delta()$ is defined as follows.

$$\delta(x_f,q_f) = \begin{cases} |x_f - q_f|, & f \text{ is numeric} \\ 0, & f \text{ is categorical and } x_f = q_f \\ 1, & f \text{ is categorical and } x_f \neq q_f \end{cases} \tag{3}$$

$w(f)$ is the feature weighting function which is defined as Eq. (4). We use mutual information between the feature and the class variable as the weight of feature f. v is a value of a feature and V_f is the value set of f. The purpose of providing such weights to features is similar to feature selection, in which the aim is to identify a set of features that contribute most information to the class prediction. *k*-NN is particularly vulnerable to useless or misleading features, so some form of feature selection is necessary. In our case, we apply higher weights ($w(f)$) to features that give more information about the class, instead of simply selecting them or not.

$$w(f) = \sum_{v\in V_f}\sum_{c_j\in J} p(c_j, x_f = v)\cdot\log\frac{p(c_j, x_f = v)}{p(c_j)\cdot p(x_f = v)} \tag{4}$$

2.3. Handling missing values

Missing values are very common in medical data. They may result from unwillingness to answer questions, the non-inclusion of

Please cite this article in press as: Chi C-L et al. Individualized patient-centered lifestyle recommendations: An expert system for communicating patient specific cardiovascular risk information and prioritizing lifestyle options. J Biomed Inform (2012), http://dx.doi.org/10.1016/j.jbi.2012.07.011

tests, or other reasons. Missing value imputation (e.g., compute the mean or mode) is a common approach. In this project, we impute distance measures instead of missing values. We use expected distance to impute the distance between a query and a training case.

There are two possible scenarios: either a query or a training case has a value missing, or both are missing. For the first scenario, the value of either a query or a training case is known, and we compute the expected distance measure of matching this known value. For example, a feature has three categorical values [*r*, *g*, *b*] whose probabilities are [0.4, 0.25, 0.35], respectively. When the known value of either a query or a training case is b, and the other is missing, the expected distance is $0.4 \times (1) + 0.25 \times (1) + 0.35 \times (0) = 0.6$. On the other hand, if the three values are numeric, [−1, 0, 1], and the known value is 1, the expected distance is $0.4 \times (|1 - (-1)|) + 0.25 \times (|1 - 0|) + 0.35 \times (|1 - 1|) = 1.05$.

For the second scenario, both values are missing, The probability for values *r* to *r*, *r* to *g*, *r* to *b*, *g* to *r*, *g* to *g*, *g* to *b*, *b* to *r*, *b* to *g*, and *b* to *b* are 0.16, 0.1, 0.14, 0.1, 0.0625, 0.0875, 0.14, 0.0875, and 0.1225, respectively. The expected distance for a categorical variable is $0.16 \times (0) + 0.1 \times (1) + 0.14 \times (1) + 0.1 \times (1) + 0.0625 \times (0) + 0.0875 \times (1) + 0.14 \times (1) + 0.0875 \times (1) + 0.1225 \times (0) = 0.655$. The expected distance for a numeric variable is $0.16 \times (|-1-(-1)|) + 0.1 \times (|-1-0|) + 0.14 \times (|-1-1|) + 0.1 \times (|0-(-1)|) + 0.0625 \times (|0-0|) + 0.0875 \times (|0-1|) + 0.14 \times (|1-(-1)|) + 0.0875 \times (|1-0|) + 0.1225 \times (|1-1|) = 0.935$.

2.4. Optimization: the healthiest plan

The key idea of PODSS is generate a recommendation with predicted result close to the desired result. In other words, we want to optimize the confidence of a good outcome. Fig. 1 shows a stylized classification problem with predictions of a patient with three different lifestyle choices. "+" stands for free from CVD event and "−" stands for CVD event. The lifestyle with the prediction "A−" is not a good choice. Although lifestyle with prediction "A+" is better, the predicted score (confidence) is low. "A*" is the best choice because we predict no CVD with high confidence. A high-confidence lifestyle recommendation (with the lowest expected CVD probability) is generated by comparing various patient-lifestyle pairs (Fig. 2).

We note that the number of combinations of all lifestyle values is a huge number (e.g., 5 BMI values × 3 alcohol values × 5 cholesterol levels × ...). For practical reasons, the system should return a fast lifestyle recommendation after receiving a query patient's information. Thus, a heuristic optimization method is chosen for this project. There are many discrete optimization techniques, such as genetic algorithms, tabu search, and simulated annealing. [9] Among these, hill climbing can quickly find an good answer (local optimum), but usually not the best (global optimum).

There are two types of optimization in this project. The first one finds the best value for a single lifestyle component (e.g., each lifestyle variable in Table 1) that can minimize one's CVD risk. The formulation is described as

$$\begin{aligned} &\text{minimize} \quad p(x_1 \cup x_{2ij}) \\ &\text{subject to} \quad i = 1, \ldots, |x_2| \\ &\qquad\qquad j = 1, \ldots, |S_i| \end{aligned} \tag{5}$$

where x_{2ij} represents one lifestyle component *i* with the value *j*. *S* represents the set of possible values for *i*. The objective is to find the best value *j* of the single lifestyle choice *i* for a patient with characteristic vector x_1. *p* is the decision function as described in (1). We can simply use exhaustive search to try all possible values of each lifestyle variable since all variables have been discretized.

The second scenario finds the combination of several lifestyle components (e.g., quit smoking, 3 h physical activity/week, etc.) that minimize one's CVD risk. The formulation is described as follows:

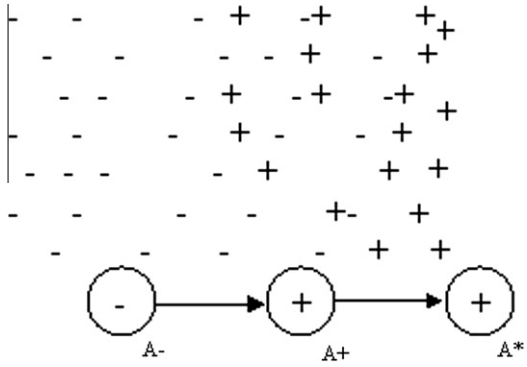

Fig. 1. Optimizing confidence of prediction.

$$\begin{aligned} &\text{minimize} \quad p(x_1 \cup x_2) \\ &\text{subject to} \quad x_{2i} \in S, \quad i = 1, \ldots, |x_2| \end{aligned} \tag{6}$$

where x_2 represents one's lifestyle vector, and the returned x_2 is the best lifestyle vector for an individual. In order to return x_2 immediately, we use forward selection [16] to solve the problem. We start with an empty lifestyle vector, and then include a lifestyle component in each iteration given x_1 and the previously included lifestyle components. Finally, we can construct the entire x_2 vector.

Patients may not want to comply with lifestyle recommendations or may ignore certain recommendations for many personal reasons. One possible way to improve adherence is to involve a patient's participation. As a result, a lifestyle recommendation can satisfy one's preferences and real-world limitations. For example, a patient is too busy to do physical activities more than 5 h/week. In another example, a patient wants to compare CVD reductions for losing 2 and 5 body mass index (BMI) points and decide on the best choice (based on effects and efforts). Optimization based on preference can generate healthy lifestyle recommendations subject to an individual's personal tendencies. In Eq. (6), we use constraints or change certain x_2 to x_1 (e.g., exercise no more than 3 h/week) to incorporate patient preferences into the optimization process. One can also provide several sets of personal preferences, compare the effects and convenience of each resulting lifestyle recommendation, and then decide the best one (see Fig. 3).

2.5. Updating patient characteristics

Variables are interrelated. Certain patient characteristics should change with new lifestyles, e.g., cholesterol level changes with the change of saturated fat and dietary cholesterol intake. If these changes are not made, the change of CVD probability resulting from a lifestyle change is unrealistically small. To solve these problems, we revise a patient's characteristics by replacing them with predicted values. In other words, we predict a patient's biological changes due to the new lifestyle.

We asked a preventive clinic medical doctor (one of our authors) to identify characteristics that would change with lifestyle. High density lipoprotein (HDL), hypertension, low density lipoprotein (LDL), total cholesterol, total triglycerides, 2nd and 3rd systolic blood pressure, 2nd and 3rd diastolic blood pressure, waist girth, hip girth, apolipoprotein AI, and apolipoprotein B were identified.

The implementation of this idea is very intuitive. As described previously, a set of new lifestyle values is created based on all patient characteristics. Then, we use distance-weighted *k*-NN with mutual information to predict the above nine variables given the

Please cite this article in press as: Chi C-L et al. Individualized patient-centered lifestyle recommendations: An expert system for communicating patient specific cardiovascular risk information and prioritizing lifestyle options. J Biomed Inform (2012), http://dx.doi.org/10.1016/j.jbi.2012.07.011

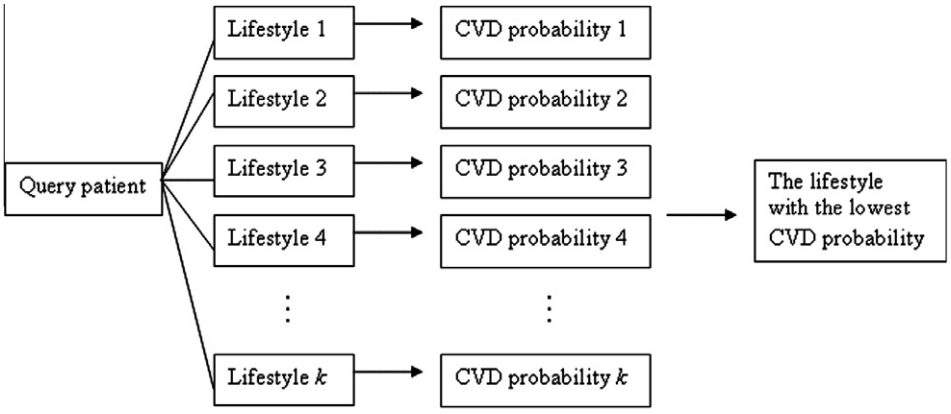

Fig. 2. Optimization process: Identifying the best lifestyle.

new lifestyle values and the patient characteristics that cannot change with the new lifestyle (e.g., education level). We use the notation $\hat{x}_1$ to represent predicted values of these variables.

2.6. Validation method

Validation is difficult for the proposed problem. In our dataset, no real patients ever used the expert system, so there are no comparison targets. We designed a clinical trial simulation algorithm to validate the system using the concept of comparison against other models [22] coupled with experimental designs[2,23]. The aim is to compare CVD risks, which are estimated using holdout data (data independent of the data used for training the system), between the original lifestyle and the recommended lifestyle.

Fig. 4 illustrates this clinical trial simulation method. We stratify and randomly assign 50% of the cases to train the expert system, and the remaining 50% as holdout data to validate recommendations. A query patient obtains a lifestyle recommendation from the expert system. CVD risks ($Risk_0$ and $Risk_1$) of the patient (x_1) with the original lifestyle (x_2) and the recommended lifestyle (x_2^*) are estimated using k-NN on the holdout data. The patient with two different lifestyles and CVD risks will then be assigned to Q and Q'.

We repeat the same process for every query patient. As a result, Q includes a set of CVD risks of the original lifestyle, and Q' includes a set of CVD risks of the recommended lifestyle. Finally, we compare CVD risks between Q and Q' groups and examine whether the risk reduction (from the original to recommended lifestyle) is significantly greater than 0.

We note that in this clinical trial simulation setting, we are comparing two sets of CVD risks for the same set of patients. Except for different lifestyle, all conditions and parameters of patients are identical between Q and Q' groups. Clinical trial designs aim to minimize differences between groups and seek for unbiased comparison. Comparison between inequivalent groups poses difficulty to understand true influence of the target of interest such as a treatment approach. For example, in a clinical trial, the treatment group is 10-year younger than the placebo group. Although outcome of the treatment group is significantly better than the placebo group, we are not sure whether the 'improvement' is the result of treatment or age. Thus, the clinical trial simulation mechanism proposed above naturally rules out biases and allows understand the CVD risk reduction due to lifestyle change.

2.7. The PODSS algorithm

We modified the PODSS algorithm from Chi et al. [10] for the lifestyle problem as summarized in Fig. 5. This algorithm recommends customized lifestyle changes $\left(x_{2\tilde{j}}^*\right)$ to a query patient $x_1 \cup \hat{x}_1 \cup x_{2\tilde{j}}$, and then estimates the changes of the patient characteristics $\left(x_1 \cup \hat{x}_1^*\right)$. Finally, the algorithm predicts and compares CVD probabilities before and after lifestyle changes.

Each iteration of leave-one-out defines a query patient $x_1 \cup \hat{x}_1 \cup x_{2\tilde{j}}$ and two subsets (training and validation) (Steps 1 and 2). A k-NN classifier (Step 4) with mutual information (obtained by Eq. (4), Step 3) identifies the best lifestyle by Eq. (5) or (6). After identification of the best lifestyle, a set of k-NN classifiers (Step 5) using the features $x_1 \cup x_{2\tilde{j}}$ predict the new patient's characteristics $\hat{x_1}^*$ based on the query patient's characteristics set that cannot change with lifestyle and the patient's new lifestyle $x_1 \cup x_{2\tilde{j}}^*$. Step 6 uses a k-NN classifier from the validation set and predicts CVD probabilities for the original query patient

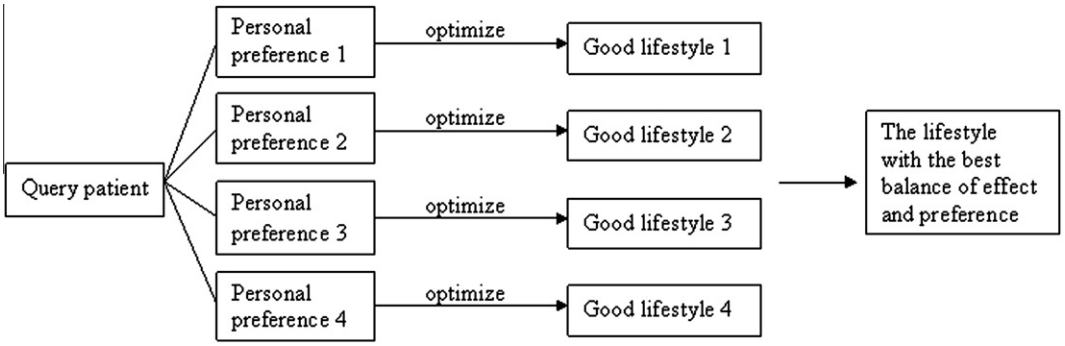

Fig. 3. Optimization based on preference. Optimal lifestyle recommendation is chosen from considering both the preference and effect of each lifestyle.

Please cite this article in press as: Chi C-L et al. Individualized patient-centered lifestyle recommendations: An expert system for communicating patient specific cardiovascular risk information and prioritizing lifestyle options. J Biomed Inform (2012), http://dx.doi.org/10.1016/j.jbi.2012.07.011

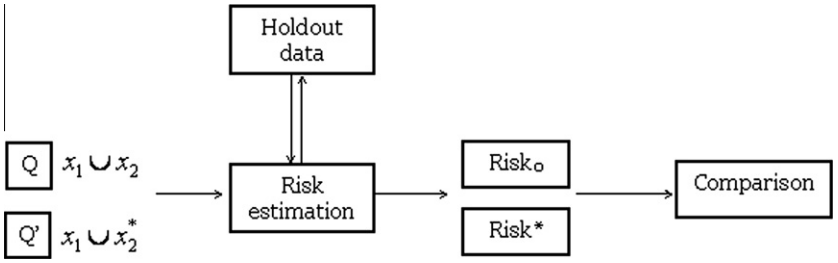

Fig. 4. The concept of using clinical trial simulations as the validation approach.

$\left(x_1 \cup \hat{x}_1 \cup x_{2\bar{j}}\right)$ and the new query patient $\left(x_1 \cup \hat{x}_1^* \cup x_{2\bar{j}}^*\right)$. Finally, the system returns the best lifestyle and new characteristics of the query patient (Step 7), along with the optimized CVD probability.

3. Results

Table 2 compares the original and recommended lifestyles. The "Original" column shows the average CVD risk of patients with their original lifestyles, and "Recommended" shows the average CVD risk of patients with recommended lifestyle changes. The number in each cell is estimated by the hold-out validation set. Each subject receives two type of recommendations, the best single lifestyle changes ("Predict-Single" column) and multiple lifestyle changes ("Predict-Multiple" column). In the first scenario, the system recommends the best single lifestyle component change (the change with the maximum CVD-risk reduction) to each user. In the second scenario, each user receives a recommendation of multiple lifestyle component changes (more than one components in most cases). It does not make sense to incorporate a lifestyle with little reduction on CVD risk. Thus, we set a very small threshold (CVD probability $\leqslant 0.0000000001\%$) to decide whether to include a further lifestyle component. The number of lifestyle components

Input

Training data D, $D_i = x_1 \cup \hat{x}_1 \cup x_{2\bar{j}}$, where x_1 is the patient characteristics that do not change with lifestyle, $\hat{x}_1$ is the patient characteristics that change with lifestyle, and $i \in$ total data points. $x_{2\bar{j}}$ is the lifestyle variable set chosen by the patient, $j \in$ all possible lifestyle choices (or all possible combinations of lifestyle components)

Outputs

Query patient with new lifestyle and new patient characteristics $x_1 \cup \hat{x}_1^* \cup x_{2\bar{j}}^*$, and original and optimized CVD probabilities

Recommending steps

1 Split the larger portion into training and validation sets

2 Define a query patient $x_1 \cup \hat{x_1} \cup x_{2\bar{j}}$ from leave-one-out

3 Obtain mutual information from the training set

4 k-NN classifier with the training set and identify the best lifestyle $x_{2\bar{j}}^*$ by Equation 5 or 6

5 k-NN classifiers and predict $\hat{x_1}^*$ based on the query patient's characteristics set that cannot change with lifestyle and the patient's new lifestyle, $x_1 \cup x_{2\bar{j}}^*$

6 Estimate CVD probabilities for $x_1 \cup \hat{x}_1 \cup x_{2\bar{j}}$ and $x_1 \cup \hat{x}_1^* \cup x_{2\bar{j}}^*$ by validation set

7 Return the new query patient $x_1 \cup \hat{x}_1^* \cup x_{2\bar{j}}^*$ and CVD probabilities

Fig. 5. The modified PODSS algorithm.

Please cite this article in press as: Chi C-L et al. Individualized patient-centered lifestyle recommendations: An expert system for communicating patient specific cardiovascular risk information and prioritizing lifestyle options. J Biomed Inform (2012), http://dx.doi.org/10.1016/j.jbi.2012.07.011

varies based on patients. In general, query patients with many bad behaviors (e.g., smoking, too much total fat) need more changes than the ones with fewer bad behaviors. The row "Avg relative CVD reduction" show the average CVD probability reduction $\frac{P(x_1 \cup \hat{x}_1 \cup x_{2j}) - P(x_1 \cup \hat{x}_1^* \cup x_{2j}^*)}{P(x_1 \cup \hat{x}_1 \cup x_{2j})}$ of all query patients. The number indicates relative risk reduction of the new lifestyle, and the number in parenthesis shows standard deviation. The row "*P*-value improvement" examines whether each recommendation is better than "Predicted-Ori". *P*-values in both cases are very small. There are two reasons for the small *p*-value. First, the data size is big (13,006 data points). Second, CVD risks of individuals are consistently reduced.

Fig. 6 summarizes the single best lifestyle changes for all subjects. In these figures, we allow each subject to receive only one best lifestyle change in order to examine the system. The *x*-axis represents interval values and the *y*-axis represents number of times the single lifestyle change was recommended. The top three recommended lifestyles are reduce cholesterol intake, quit smoking, and lose weight. We note that we did not set constraints to force recommendations (e.g. we did not tell the system that a smoker has to quit smoking); instead, the automatically-captured knowledge tells the system what to recommend. All recommendations come from the system's predictions based directly on the data.

There is no single best lifestyle change for carbohydrates, and very few recommendations involving protein. The effect of activity hours is under-estimated because an activity level of 0 h is not recorded, although there are lots of blank values. Thus, we are unable to distinguish 0 h of activity from a missing value in the dataset, and we treat all of them as missing values, which comprises 1/3 of the dataset. Recorded values start from 1 h of activity per week. In other words, subjects with values did at least 1 h activity per week, and the algorithm cannot recognize the difference of CVD risks between doing activity and without activity. Thus, the negative effect of less activity was underestimated.

We use patient P48 as a case study to show patient-centered lifestyle recommendations in Tables 3 and 4. In Table 3, the "Original" column shows P48's original lifestyle. P48's CVD probability is 9.55%, which is higher than the average of 8% (Table 2). There are two recommendations, "Whole plan" and "Modified whole plan". In both columns, cells with a number indicate the recommendation of a change. An empty cell indicates no recommendation of change, e.g., carbohydrate.

The first recommendation, "Whole plan", is identified by the system based on the lowest CVD probability. A small threshold (1e–12) was set to determine whether a lifestyle component should be included. The system does not include carbohydrate in this plan because of very small CVD-risk reduction.

Relative CVD probability reduction (19.9%) of the "Whole plan" is high, but it is hard to follow. One may want to modify a lifestyle plan based on personal preferences and real-world constraints. The "Modified whole plan" shows such an example. Losing 8 BMI points (from >31.38 to 23.35) is very difficult for P48, as is fiber consumption. In addition, P48 does not have much time for activity, but is willing to make some lifestyle changes for good health. A new plan ("Modified whole plan") is generated based on P48's preferences and constraints. The relative CVD probability reduction of the new plan is still high (18.53% relative CVD reduction), and the new plan is easier for the patient to follow. We note that healthcare providers do not recommend increasing alcohol intake due to the lack of randomized trials and also potential for addiction. However, from observational data, moderate drinkers are healthier than non-drinkers or those who drink >2 drinks/day. Therefore, AHA Scientific Statement [19] suggests if alcoholic beverages are consumed, they should be limited to no more than two drinks for men and one drink for women per day. Since the purpose of this manuscript is proof of concept, one important goal is, without providing any knowledge, the machine should be able to learn the healthy lifestyle knowledge by itself. However, when using this system in a clinical setting, we may need to integrate additional clinical concerns with the learned knowledge. In both whole-plan scenarios, the influences of protein and alcohol are extremely small. In order to use a consistent threshold as described in Section 3, we still keep these recommendations in this table.

Table 2
Comparison between true and predicted outcomes.

	Original	Recommended	
		Predict-single	Predict-multiple
Avg probability	8%	7.27%	7.17%
Avg relative CVD reduction	–	8.65% (5.49%)	9.9% (5.45%)
P-value improvement	–	$\ll 0.0005$	$\ll 0.0005$

In another scenario (Table 4), P48 may only want simple and effective lifestyle changes instead of the whole plan. The system generates "package of three" lifestyle changes for P48 including weight loss, total fat reduction, and smoking cessation. P48's CVD probability will be 7.66% (i.e., 99.47% of total possible reduction of the whole plan). P48 wants to find out if CVD risks differ by BMI level and then decide a reasonable one to follow. Columns "lower BMI 1–3" shows risks in different BMI levels.

Losing weight is difficult for P48, so the patient wants to see the compensatory plan for not losing enough BMI (in the range of 28.06–31.38). Unfortunately, many lifestyle changes cannot compete with losing more weight ("BMI 3 compensation"). BMI loss, smoking cessation, and total fat reduction are the most important changes for P48, and losing weight is more effective than other lifestyle changes. For other individuals, the story may be different.

Table 5 shows the proportions of six poor lifestyle habits, smoking, obesity, over-intake of cholesterol, over-intake of saturated fat, over-intake of total fat, and insufficient activity, for hypertension patients, diabetes patients, and smokers. The proportions of the three type of subjects in the entire sample are 31.9%, 8.2%, and 43.6%, respectively. Obesity is determined by whether a subject's BMI is greater than 30 [7]. Too much cholesterol, saturated fat, and total fat are determined by cutpoints 300 mg, 7% energy, and 35% energy as suggested by current dietary recommendations [19]. There is no suggested cutpoint to decide insufficient activity. In this project, the cutpoint is 5 h, which is the median activity value.

Table 6 summarizes three most frequent single lifestyle recommendations, quit smoking, cholesterol control, and weight control, for smokers. The proportions of the three groups of smokers are 54.4%, 37.2%, and 4.8%, respectively. For this examination, all smokers received only one lifestyle change recommendation. In other words, this lifestyle change can reduce the most CVD risk for an individual. Although the single lifestyle recommendation is not practical, we use it to examine how the system reasons. Each column represents a recommendation. The majority (54.4%) of smokers are recommended to quit smoking, 37.2% should control their cholesterol, and 4.8% should control their weight (BMI). The result is very surprising because 100% smoking cessation would be expected. For the group of cholesterol control, the possible explanation is smoking exacerbates the effect of total cholesterol and high-density lipoprotein cholesterol [20]. For the group that receives the recommendation of control weight prior to smoking cessation, 89% of this group has BMI >31.38 (the top BMI value), and the rest have BMI ranged from 28.06 to 31.38. Obesity has worse health consequences than simply being overweight, and weight loss is very important in this group. This might be a surprising finding that bears future investigation – it implies that obesity is worse than smoking for cardiovascular health in middle-aged people.

Please cite this article in press as: Chi C-L et al. Individualized patient-centered lifestyle recommendations: An expert system for communicating patient specific cardiovascular risk information and prioritizing lifestyle options. J Biomed Inform (2012), http://dx.doi.org/10.1016/j.jbi.2012.07.011

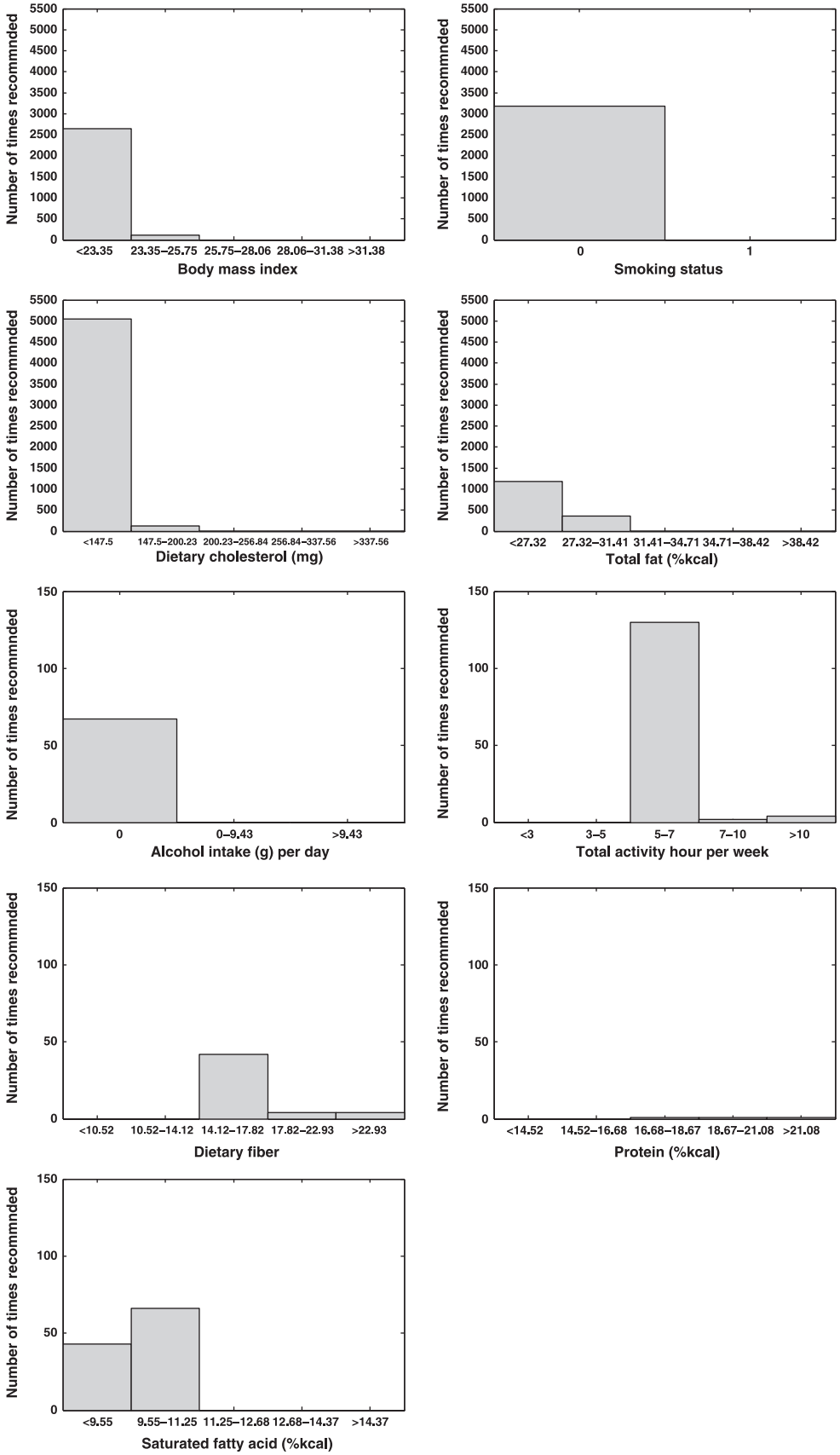

Fig. 6. The number of times a single best lifestyle change was recommended.

Please cite this article in press as: Chi C-L et al. Individualized patient-centered lifestyle recommendations: An expert system for communicating patient specific cardiovascular risk information and prioritizing lifestyle options. J Biomed Inform (2012), http://dx.doi.org/10.1016/j.jbi.2012.07.011

Table 3
Customized lifestyle recommendation for case P48.

Lifestyle components	Original	Whole plan	Modified whole plan
BMI	>31.38	<23.35	28.06–31.38
Alcohol (g/day)	0	>9.43	>9.43
Smoking	Yes	No	No
Sport (h/week)	<3	>10	3–5
Carbohydrate (g)	<128.79		
Cholesterol (mg)	147.5–200.23	<147.5	<147.5
Fiber (g)	<10.52	>22.93	10.52–14.12
Protein (% kcal)	>21.08	<14.52	<14.52
Saturated fat (% kcal)	>14.37	<9.55	<9.55
Total fat (% kcal)	>38.42	<27.32	<27.32
CVD risk	9.55%	7.65% (19.9% relative CVD risk reduction)	7.78% (18.53% relative CVD risk reduction)

Table 4
Package of three lifestyles with different levels of BMI and the compensatory plan for not losing enough BMI.

Lifestyle components	A package of three	Lower BMI 1	Lower BMI 2	Lower BMI 3	BMI 3 compensation
BMI	<23.35	23.35–25.75	25.75–28.06	28.06–31.38	28.06–31.38
Alcohol (g/day)					>9.4
Smoking	No	No	No	No	No
Sport (h/week)					> 10
Carbohydrate (g)					
Cholesterol (mg)					< 147.5
Fiber (g)					> 22.93
Protein (% kcal)					< 14.52
Saturated fat (% kcal)					< 9.55
Total fat (% kcal)	< 27.32	< 27.32	< 27.32	< 27.32	< 27.32
CVD risk	7.66% (99.47% of total possible reduction)	7.71% (96.84% of total possible reduction)	7.75% (94.74% of total possible reduction)	7.8% (92.1% of total possible reduction)	7.78% (93.16% of total possible reduction)

Table 5
Distribution of smoking, cholesterol intake, obesity, saturated fat, total fat, and insufficient activity in those with hypertension, diabetes, or smokers.

Poor lifestyle habits of the three types of subjects	Hypertension (%)	Diabetes (%)	Smokers (%)
Smoking	22.6	22.4	100
Obese	39.6	51.9	19.4
Over cholesterol	28.5	33.3	32.2
Over saturated fat	94.8	95	95.8
Over total fat	36.3	43.5	41.5
Insufficient activity	28.1	28.2	25.2

Table 6
Poor lifestyle habits of smokers for the three most frequently recommended single lifestyle changes (quit smoking, cholesterol control, and weight control).

Poor lifestyle habits of the three groups of smokers	Quit smoking (54.4% smokers) (%)	Cholesterol control (37.2% smokers) (%)	Weight control (4.8% smokers) (%)
Smoking	100	100	100
Obese	11.8	20.8	94.2
Over cholesterol	5.5	77.2	8.4
Over saturated fat	93.4	98.9	94.8
Over total fat	28.1	56.3	35.7
Insufficient activity	25.9	23.7	24

In order to understand how the system reasons, we analyze characteristics of subjects in each different group in Table 6. Compared with "Smokers" in Table 5, we can see the first group of subjects are less obese (11.8%), and consume less cholesterol (5.5%), saturated fat (93.4%), and total fat (28.1%), so quit smoking is certainly the first recommendation. The second group of subjects are slightly more obese (20.8%), consumes more cholesterol (77.2%), saturated fat (98.9%), and total fat (56.3%), especially many subjects have a cholesterol intake problem, and they receive recommendations to control cholesterol. The third group of subjects are very obese (94.2%), but they have less cholesterol (8.4%), saturated fat (94.8%), and total fat (35.7%) intake. Thus, they are recommended to control their weight.

Clinically, weight control is usually recommended for diabetes and hypertension patients, and weight control is indeed the first recommendation for diabetes patients (Table 7) and and the second recommendation for hypertension patients (Table 8). In addition, the system shows that lowering cholesterol is among the top two recommendations for both diseases.

Because the recommendation is individualized, the results can also be viewed from the opposite direction. The individualized property identifies which subjects benefit the most from certain lifestyle changes. Table 9 shows observed CVD probabilities for patients who receive a single recommendation to reduce BMI. Each patient's BMI loss recommendation varies. Some receive the recommendation to lose 1 BMI interval, and others are recommended to lose more. (There are five intervals for BMI (see Table 1), and

Table 7
Poor lifestyle habits of individuals with diabetes for the three most frequently recommended single lifestyle changes (weight control, cholesterol control, and total fat control).

Poor lifestyle habits of the three groups of diabetes	Weight control (42.8% diabetes) (%)	Cholesterol control (39.7% diabetes) (%)	Total fat control (9.8% diabetes) (%)
Smoking	29.4	42.8	34
Obese	78.7	40.5	11.5
Over cholesterol	10.5	70.9	4.8
Over saturated fat	91.4	99.1	100
Over total fat	32.1	50	86.5
Insufficient activity	32.3	24.9	26.9

Please cite this article in press as: Chi C-L et al. Individualized patient-centered lifestyle recommendations: An expert system for communicating patient specific cardiovascular risk information and prioritizing lifestyle options. J Biomed Inform (2012), http://dx.doi.org/10.1016/j.jbi.2012.07.011

Table 8
Poor lifestyle habits of individuals with hypertension for the three most frequently recommended single lifestyle changes (cholesterol control, weight control, and quit smoking).

Poor lifestyle habits of the three groups of hypertension	Cholesterol control (44.1% hypertension) (%)	Weight control (33.2% hypertension) (%)	Quit smoking (13.4% hypertension) (%)
Smoking	38.9	18.8	100
Obese	30.6	71.6	8
Over cholesterol	58.7	6.6	2.5
Over saturated fat	98.4	92.1	86.6
Over total fat	43.4	24.1	15.6
Insufficient activity	26.3	30.5	24.8

Table 9
CVD probabilities.

Recommendation	Frequency	Baseline CVD probability (%)
Lose 1 BMI interval	66	4.55
Lose 2 BMI intervals	359	8.36
Lose 3 BMI intervals	838	6.32
Lose 4 BMI intervals	1479	10.62

lose 1 BMI interval means reduce BMI to the next interval.) On average, patients who receive the recommendation to lose more BMI intervals have higher BMI values. We expect that subjects with smaller BMI value have smaller CVD risk, and, again, we use this knowledge to examine the system.

Table 9 summarizes the fraction of CVD events for subjects in the four BMI interval recommendations. As expected, patients who are recommended to lose more BMI points tend to have higher CVD risk except for patients who are recommended to lose two BMI levels. LDL for this group is very high although they do not consume much cholesterol and fat.

4. Conclusion

This paper applies a revised PODSS algorithm to learn patterns from patient-level data (ARIC) and generate customized lifestyle recommendations that lead to the lowest predicted CVD risk based on one's preferences. The central idea of the PODSS aims to identify the best match between an intervention and a patient. To implement this idea, we used a predictive model to capture the relationship between interventions, patients, and CVD outcomes, and then used optimization techniques to identify the intervention-patient pair that results in the lowest CVD probability under the constraint of patient preference. To build a realistic expert system, we had extensive discussions with a prevention clinical doctor and used real-world scenarios to test the system (e.g., Table 4 shows the relationship between BMI and CVD to promote weight loss and identify the "package of three" lifestyle changes to help efficiently reduce CVD risk). However, the most relevant parameters for a given individual were, in all cases, learned by the system, based on the outcomes of similar individuals in the training set.

To be more realistic, we may need to include variables such as socio-economic factors: a glut of fast food options, a dearth of grocery stores with fresh produce, restricted access to recreational facilities, etc. In this proof-of-concept project, we determined a set of variables that are easily obtained and often considered for lifestyle recommendation in clinics. The ARIC study is a reasonable choice of dataset since it has a fairly extensive set of measured variables from which to choose. In general, if we had more information available, we could make use of it easily without causing overfitting (see discussion of feature weighting in Section 2.2). We also note that features which are not directly measured, such as the socio-economic factors may in fact be implicitly included in our analysis. Measurement of features such as diet and exercise. means that one's "neighbors" in our feature space might in fact also be neighbors in the broader geographic sense, i.e., people from similar backgrounds living in similar conditions. Hence we are at least giving our system the opportunity to learn what is important to risk prediction, even in some cases where it is not directly measured.

CVD events from ARIC data are coded as dichotomous outcomes (0, 1), and we use a function to learn and smooth out these events based on the event distribution. An example is the prediction of a logistic regression model, which learns from dichotomous outcomes and produces a probability estimate of a CVD event. A similar principle can be used to our prediction model, *k*-NN, which provides a probability estimate based on similar individuals. As a simplified example, an individual that has 2 CVD cases among their 100 nearest neighbors would have a risk prediction of 2%. The computation in our model is more complicated, applying weights for both the nearness of the neighbors and the importance of the various features in order to obtain more accurate risk estimation.

Because of the different philosophy in computational approaches, lifestyle studies using statistical methods such as ARIC or Framingham Risk models have several properties significantly different from the proposed machine learning approach. The former focus on variable interpretation and causal relationships; therefore, identifying confounding variables for unbiased interpretation is crucial. Significant risk factors identified in the ARIC and Framingham Risk models are variables that explain the most CVD risk for the population. On the other hand, machine learning focuses on prediction, and our machine learning model uses this focus to guide the selection of the lifestyle that minimizes an individual's CVD risk. To improve predictive performance, including more variables that contribute information for prediction of the class is a common strategy, and therefore, our model incorporates more variables (including those considered confounders in statistical models) than ARIC and Framingham Risk models. Because of the difference between interpretation and prediction, including confounding variables is less important in a machine learning model. Since the *k*-NN model is highly nonlinear and can learn interactions among variables, including a feature whose effect is dependent on another, or one that has an effect only on a small number of patients, can still improve the overall predictive performance.

In addition, the difference in computational approaches also results in very different applications. The variables that the ARIC and Framingham Risk models identified are the lifestyle variables optimized for most people, which may be dramatically different from particular individuals or subgroups. For example, lifestyle recommendations for a 200-lb athlete and a 200-lb couch potato should differ. Instead of identifying the lifestyle optimized for most people, our approach optimized the lifestyle for each individual based on CVD risk prediction that is estimated by similar individuals' CVD outcomes.

We anticipate the proposed system will serve as an interactive system to collect an individual's characteristics, limitations, and preferences. The system uses these data to predict CVD risks of various lifestyle scenarios to help physicians and patients to decide

Please cite this article in press as: Chi C-L et al. Individualized patient-centered lifestyle recommendations: An expert system for communicating patient specific cardiovascular risk information and prioritizing lifestyle options. J Biomed Inform (2012), http://dx.doi.org/10.1016/j.jbi.2012.07.011

the most effective and easily-complied lifestyle choice. We anticipate such an interactive decision process (among patients, physicians, and the system) will increase patients' involvement and participation, and subsequently may improve a person's commitment to healthy lifestyle and influence them to actually change behaviors.

PODSS uses *k*-NN classifiers to capture nonlinear knowledge directly from a dataset and then automatically applies the knowledge to the decision support system. Complex interactions among variables can be recorded in a non-linear function, and then the function is used to generate recommendations based on one's characteristics. In other words, the recommended lifestyle is the best output (lowest CVD) of the function that records complex interactions among variables. Such a lazy learning approach stores all data and predicts a query case by using the outcomes of similar cases. For the problem with a large dataset, e.g., GWAS study (thousands of individuals and millions of SNPs), we need to pay special attention to large data size handling and computational speed. Possible solutions include reducing the number of features, sophisticated data structures for faster query processing, or a change in classifiers. The data size in this study is much smaller, so we did not experience such computational issues. In our experience, a lifestyle recommendation is generated in a few seconds.

Performing optimization requires comparing the effects of multiple lifestyle combinations for each individual. In order to provide quick recommendation, we choose hill climbing as the optimization approach. Although the recommended lifestyle is a local optimum (a good lifestyle but not necessarily the global best among all combinations) for a patient, for practical reasons, this optimization approach is the best choice.

Our proof-of-concept approach is a potential machine learning method for use in the domains of behavior changes, patient-centered medicine, personalized medicine, and comparative effectiveness research for individuals or patient subgroups. In future work, we will apply this approach to patient-centered comparative effectiveness research studies. Genetic variation data can be incorporated into PODSS as a part of the patient characteristics. Identifying the best treatment option by comparing effects and harms for individuals is an interesting direction. We also plan to identify individuals and subgroups that most benefit from a certain treatment option. This approach can be applied to identify the best lifestyle and drug combinations to reduce risk factor levels and CVD risks for individuals. We will further extend our approach to identify the best strategy to prevent multiple diseases (e.g., CVD, hypertension, and diabetes).

References

[1] NCI definition of tailored intervention; 2010. <http://www.cancer.gov/dictionary?cdrid=561723> (accessed 04.12).
[2] Adelman L. Experiments, quasi-experiments and case studies: a review of empirical methods for evaluating decision support systems. IEEE Trans Syst Man Cybern 1991;21(2):293–301.
[3] Aha DW. Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. Int J Man–Mach Stud 1992;36(2):267–87.
[4] Aha DW. Lazy learning. Kluwer Academic Publishers; 1997.
[5] Bali RK, Feng DD, Burstein F, Dwivedi AN. Introduction to the special issue on advances in clinical and health-care knowledge management. IEEE Trans Inform Biomed 2005;9(2):157–61.
[6] Bandura A. Social foundations of thought and action: a social cognitive theory. Englewood Cliffs, NJ: Prentice Hall; 1986.
[7] Body Mass Index (BMI). The description of BMI; 2009. <http://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html> (accessed 04.09).
[8] Brug J, Oenema A, Campbell M. Past, present, and future of computer-tailored nutrition education. The Am J Clin Nutr 2003;77(Suppl 4):1028S–34S.
[9] Burke EK, Kendall G, editors. Search methodologies. Springer; 2006.
[10] Chi C-L, Street WN, Ward MM. Building a hospital referral expert system with a prediction and optimization-based decision support system algorithm. J Biomed Inform 2008;41(2):371–86.
[11] Clancey WJ. Heuristic classification. Artif Intell 1985;27(3):289–350.
[12] Colleen A, Joseph S, Susan R, Wayne F. Health behavior models. The Int Electron J Health Educat 2000;3:180–93 (Special Issue).
[13] Duda RO, Hart PE, Stork DG. Pattern classification. second ed. NY: John Wiley and Sons, Inc.; 2001.
[14] Enwald HP, Huotari ML. Preventing the obesity epidemic by second generation tailored health communication: an interdisciplinary review. J Med Internet Res 2010;12(2):e24.
[15] Atherosclerosis Risk in Communities Study (ARIC). The project description and data; 2008. <http://www.cscc.unc.edu/aric/> (accessed 04.08).
[16] John GH, Kohavi R, Pfleger K. Irrelevant features and the subset selection problem. In: Proceedings of the 11th international conference on machine learning. San Francisco, CA. Association for Computing Machinery; 1994. p. 121–9.
[17] Kreuter MW, Wray RJ. Tailored and targeted health communication: strategies for enhancing information relevance. Am J Health Behav 2003;27(Suppl 3):S227–32.
[18] Liau LC-K, Yang TC-K, Tsai M-T. Expert system of a crude oil distillation unit for process optimization using neural networks. Expert Syst Appl 2004;26(2):247–55.
[19] Lichtenstein AH, Appel LJ, Brands M, Carnethon M, Daniels S, et al. Diet and lifestyle recommendations revision 2006: a scientific statement from the American Heart Association Nutrition Committee. Circulation 2006;114(1):82–96.
[20] Nakamura K, Barzi F, Huxley R, Lam T-H, Suh I, Woo J, Kim HC, Feigin VL, Gu D, Woodward M. Does cigarette smoking exacerbate the effect of total cholesterol and high-density lipoprotein cholesterol on the risk of cardiovascular diseases? Heart 2009;95:909–16.
[21] Oenema A, Tan F, Brug J. Short-term efficacy of a web-based computer-tailored nutrition intervention: Main effects and mediators. Ann Behav Med 2005;29(1):54–63.
[22] O'Keefe RM. Expert system verification and validation: a survey and tutorial. Artif Intell Rev 1993;7:3–42.
[23] Shadish WR, Cook TD, Campbell DT. Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin company; 2001.
[24] Shortliffe EH, Davis R, Axline SG, Buchanan BG, Green CC, Cohen SN. Computer-based consultations in clinical therapeutics: explanation and rule acquition capabilities of the MYCIN system. Comput Biomed Res 1975;8(4):303–20.
[25] Skinner CS, Campbell MK, Rimer BK, Curry S, Prochaska JO. How effective is tailored print communication? Ann Behav Med 1999;21(4):290–8.
[26] Song Z, Kusiak A. Optimization of temporal processes: a model predictive control approach. IEEE Trans Evol Comput 2009;13(1):169–79.
[27] Tan P-N, Steinbach M, Kumar V. Introduction to data mining. Addison Wesley; 2005.
[28] Watson I, Marir F. Case-based reasoning: a review. The Knowl Eng Rev 1994;9(4):355–81.
[29] Watson ID, Basden A, Brandon PS. The client centered approach: expert system maintenance. Expert Syst 1992;9(4):189–96.
[30] Weiss SM, Buckley SJ, Kapoor S, Damgaard S. Knowledge-based data mining. In: Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining. Washington, DC, 2003. Association for Computing Machinery; 2003. p. 456–61.
[31] Wettschereck D, Aha DW, Mohri T. A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. Artif Intell Rev 1997;11:273–314.

Please cite this article in press as: Chi C-L et al. Individualized patient-centered lifestyle recommendations: An expert system for communicating patient specific cardiovascular risk information and prioritizing lifestyle options. J Biomed Inform (2012), http://dx.doi.org/10.1016/j.jbi.2012.07.011