ELSEVIER

Artificial Intelligence in Medicine 23 (2001) 277–293

Artificial Intelligence in Medicine

www.elsevier.com/locate/artmed

A sequential neural network model for diabetes prediction

Jin Park*, Dee W. Edington[1]

The University of Michigan, 1027 E. Huron, Ann Arbor, MI 48104-1688, USA

Received 10 April 2001; received in revised form 18 June 2001; accepted 25 June 2001

Abstract

This paper presents a neural network (NN) model to evaluate an existing Health Risk Appraisal (HRA)[2] for diabetes prediction over 3 years (1996–1998) based on a simulated learning algorithm on individual prognostic process, using the repeatedly measured HRAs of 6142 participants.

The approach uses a sequential multi-layered perceptron (SMLP) with backpropagation learning, and an explicit model of time-varying inputs along with the sequentially obtained prediction probability, which was obtained by embedding a multivariate logistic function for consecutive years.

The study captures the time-sensitive feature of associating risk factors as predictors to the occurrence of diabetes in the corresponding period. This approach outperforms the baseline classification and regression models in terms of gains (average profit: 0.18) and sensitivity (86.04%) for a test data.

The result enables a time-sensitive disease prevention and management program as a prospective effort. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Multi-layered perceptron; SMLP; Disease prediction; Backpropagation; HRA

1. Introduction

An artificial neural network (ANN or NN), is a modeling technique based on the observed behaviors of biological neurons, used to mimic the performance of the human system. In 1949, the psychologist Hebb [10] proposed his famous learning rule: the

* Corresponding author. Tel.: +1-734-647-7602; fax: +1-734-763-2206.

E-mail addresses: kddum@umich.edu (J. Park), dwe@umich.edu (D.W. Edington).

[1] Tel.: +1-734-647-7602; fax: +1-734-763-2206.

[2] Health Risk Appraisal: see Appendix for the definition and related methods for collection of data.

PII: S0933-3657(01)00086-0

strength of a *synapse*[3] between two neurons is increased by the repeated activation of one neuron by the other across these synapses.

The pioneering paper of McCulloch and Pitt [17] described the theory of neural networks. Then, in 1957, Rosenblatt [19] invented the *perceptron*, an artificial neuron, which *dendrities*[4] are replaced by inputs multiplied by weights. These weights can be associated with the synapses, where they determine the contribution of the individual inputs.

These weighted inputs are simply summed inside the neuron, which pass through a suitable threshold (activation). Similarly, the activated outputs from previous layers transfer to the output layer, passing through another activation, produces an output to simulate a desired output (target) at the end. By a learning algorithm, the neural net achieves learning by modifying weights proportional to the difference between the target and the gained output [3].

Various architectures of ANN have been used in different medical diagnoses [2,4,15,18,20,23], and the results were compared with physicians' diagnoses and existing classification methods. Many of them found that the ANN approach indeed, has more flexibility in modeling and reasonable accuracy in prediction. Maindonald [16] in his lecture notes, called attention to the new approaches of research training, such as ANN. Especially, across many different research areas, the insights of *evidence-based medicine* were emphasized. The result of a study by Shanker [20], that used neural networks to predict the onset of *diabetes mellitus*[5] in Pima Indian women, showed that neural network is a viable approach to classification.

Most of such papers lack survey in the selection of architecture. Discussion of its effectiveness and appropriateness to the objective and data were also insufficient to apply in specific diagnosis. Especially, the dichotomous classification of the onset of diabetes is not enough information for a disease management program since the progress of the disease would not be the same for all those who were classified as diabetics, and yet, the time of occurrence of the disease is also important to identify individual status and to implement an individualized intervention [1,18,20]. In addition, none of the current ANN approaches utilized Health Risk Appraisal (HRA) data in terms of worksite disease prediction [5,6,9,11,22].

We studied the result of choice of appropriate neural network architecture using SAS® Enterprise–Miner3.0 and user-defined implementations with Matlab5.3 for diabetes prediction and identification of its progression. HRA data were used to evaluate existing HRA variables and develop each person's index to predict his/her chance of having the disease over a relatively short time.

In the next section, we introduce the study population and their demographic information. In Section 3, for the comparison of models, a logistic regression model is used and a

[3] The point at which a nerve impulse passes from an axon of one neuron to the dendrite of another.

[4] A branched part of nerve cell that transmits impulses toward the cell body.

[5] Diabetes mellitus: a metabolic disease in which there is a deficiency or absence of insulin secretion by the pancreas. Diabetes mellitus occurs in two major forms: Type I, or insulin-dependent diabetes mellitus, and Type II, or non-insulin-dependent diabetes mellitus. In the study, due to the characteristics of the population (44–64 years old), Type II was mainly considered.

baseline multi-layered perceptron (MLP) without incorporating sequential prognostic information as classification model is implemented. Also, the sequential MLP model of interest is introduced. In Section 4, we illustrate the result using sequential multi-layered perceptron (SMLP) to compare with the other two models in terms of sensitivity and misclassification rate over a test set. We discuss the useful findings of learning with SMLP in Section 5 as regards predictors, and conclude with the significance of SMLP learning in Section 6. Suggestions for further research and the limitations for this investigation are in Section 7.

2. Methods

2.1. Study population

In this study, diabetes was examined especially since it has a relatively large incidence[6] rate in the US [8]. From a large manufacturing company in the US, we selected the case study population that met the following criteria:

1. chose a traditional indemnity/PPO insurance plan, 1993–1998;
2. continuously employed from 1993 to 1998;
3. age ranged between 45 and 64 as of 1998;
4. completed an HRA during the baseline year (1996);
5. had any diabetes case claim (ICD-9 = 250) in 1996, 1997, or 1998.

Persons with prior record of diabetes, were excluded from the study.

The total population that met the criteria was 3071 (Table 1). A matching number of non-diabetics (as of 1998) that met (1), (2), (3), and (4) was randomly selected from the same population with the same age and gender distribution as a control population.[7] Note that the population of self-reported diabetics without claims was excluded.

2.2. Variable selection and model

Input and target variables are shown in Table 2. More detailed modification and fitting methods are explained in respective sections.

Subjects' information in the HRA of the baseline year and 2 consecutive years' diabetes occurrence indicator were chosen for training. Forty HRA variables directly from the questionnaires and 11 calculated variables were available. For disease prediction of a given population, 40% random sample of the study population (2456 cases) was used for training, 30% for validation check, and the remaining 30% of the data (1843 cases) was used for testing for all cases [21].[8]

[6] Five percent of the participants according to the previous cumulative report.

[7] The total study population is 6142, with 6% difference more male in the study population.

[8] To reach a reliable prediction, data partition or bootstrap sampling was suggested [7]. Due to cross-missing inputs, slight different number of cases was used to maintain the 40:30:30% ratio.

Table 1
Classification of study sample

3 Years (1996–1998)	Male	Female	Total
Non-diabetics	1654 (26.93%)	1417 (23.07%)	3071
Diabetes	1663 (27.08%)	1408 (22.92%)	3071
Average age at 1996 (year)	49.6	49.7	6172
Diabetic case	1996	1997	1998
$N = 3071$	1147 (37.35%)	1087 (35.40%)	837 (27.25%)

3. Description of models and comparison

3.1. Overview of utilization of HRA variables and model selection process

Using HRA variables is a challenge since each variable attribute is subjective, compared to medical diagnostic data. Yet, such a full description of individual health status provides an overall idea of well-being, not necessarily for specific disease identification. In addition, HRA data are easy to obtain and are less costly than medical diagnostic data, and consist of biometric, behavioral, and attitudinal measures. While an HRA questionnaire seeks to identify an individual's status on selected precursors to provide the individual with an appraised personal wellness as an ongoing effort, it is reasonable to use evaluated HRA data for screening a certain disease in advance. Thus, individual effort on the modification of risk factors can follow [5,6,9].

HRA data have the unique features of self-reported variables, including fuzziness of membership and cross-missing information. In addition, since an HRA is a collection of health demographics, it has many variables (about 40–50) in different scales, including direct measures and evaluated measures. Some variables are distributed as highly skewed and time-sensitive. For a disease outcome prediction, a statistical parametric assumption often cannot be applied to an HRA data set. However, with non-linearity in ANN in a relaxed situation,[9] even higher dimensionality of parameters is allowed. When manifold input variables are to be studied for their associations, most existing classification procedures often fail to provide adequate results with significant reliability. In such cases, ANN is particularly useful to articulate the relationship in the data with an improvement of exactness, allowing non-linearity in associations within the data.

3.2. Multiple regression model

In multiple regressions, the choice of variables is important but with more than three independent variables, the result of regression is not an easy inference. The conventional way is to rely on individual and hierarchical significance of regression coefficients. When there is statistical significance, which is not in direct interpretable form, important

[9] Only assumption is independence in input variables and individual subjects, which is attained in large data sets.

Table 2
Variables[a]

Name	Definition	Type	Values
Age	Age at 1999	Numeric	45–64
Gender	Female/male	Categorical	0, 1
Alcohol consumption	Number of drinks per week	Numeric	0–99
Alcohol index	0–6/6–14/$\geq$14	Categorical	0–2
Achievable age	Calculated by modifying all factors that increase the overall risk	Numeric	37–72
Back pain	Existence of back pain	Categorical	0, 1
BMI, BMI_I	Weight (kg), height (cm)^2	Numeric	9–91
	BMI index over a range of BMI	Categorical	0–2
Described blood pressure	High/medium to low/not sure	Categorical	0–2
Preventive service	Weighted sum (any applicable preventive service), compared to national recommendation	Categorical	7–37
Bronchitis	Existence of bronchitis	Categorical	0, 1
Cancer	Existence of cancer	Categorical	0, 1
Cholesterol	High cholesterol > 239 mg/dl	Numeric	
Cholesterol range	High/low to medium/not sure	Categorical	0–3
Diet	Eat dietary fiber	Categorical	1–4
Frame	Small/medium/large	Categorical	1–3
Sleep	Hours of sleep per day: $\sim$6, 7, 8, $\geq$9	Categorical	1–4
Fatty food	Usual intake of fatty food	Categorical	1–4
Cigarette smoking	Current/past smoking status	Categorical	0–3
High BP (SBP, DBP, SBP–DBP)	SBP > 139, DBP > 89 (0–299/0–199/0–299)	Numeric	
First child birth	Time of first child birth	Categorical	0–2
Follow-up	Wish to be followed regarding the health status	Categorical	0–2
Planning to change	Wish to change of the health status	Categorical	0–2
Appraised age	Appraised health age based on the current risk compared to those of the same age/gender	Numeric	37–82
Ideal weight	National 1959 metropolitan standard	Numeric	82–248
Weight index	Index of % over from ideal weight (ratio)	Categorical	0.4–4.3
Absence days	absent days due to illness in the past year	Categorical	1–6
Instructed BP	Instructed to take medication	Categorical	1–4
Life satisfaction	Index of life-satisfaction	Categorical	0–3
Participation in HRA	Annual participation index (100–101–110–111)	Categorical	0–3
Personal loss	Index of personal-loss	Categorical	0–3
Exercise	Frequency of physical activity per week	Categorical	0–4
Perceived health	Perceived self health status	Categorical	1–4
Pregnant	Index of pregnancy	Categorical	1, 2
Number of total risk factors	Sum of measured high risks	Numeric	0–13
Social tie	Index of social-tie	Categorical	1–4
Stress score	Evaluated stress score	Numeric	9–35
Stroke	Existence of stroke	Categorical	0, 1
Seatbelt use	Use of seatbelt	Categorical	0–2
Wellness	Standardized score for being in good health	Numeric	50–100
Diabetes indicator	Any time occurrence	Categorical	0, 1
Year of diabetes	Year indicator of diabetes	Numeric	1996–1998

[a] Last two variables were used for target status of each observation in baseline/sequential models.

meaningful information is sacrificed in many cases [13]. Statistical significance often lacks in identifying meaningful covariates in the prediction of diabetes incidence using logistic regression analysis due to problem of multiple co-linearity and weak association. As in the neural net model, while diabetes indicator is used as a response variable, among many independent variables, and their interactions, selecting predictors to diabetes is a challenge. With an additional factor analysis, we have found the three groups of factors, the first group with age, appraised age and achievable age; second with weight and body mass index (BMI); and the last group with wellness and stress. Variations of the logistic regressions were applied with these selected factors.

3.3. Baseline neural network: multi-layered perceptron

For the baseline ANN architecture, a fully connected feed-forward NN with three hidden layers (three nodes per layer), one input layer, one output layer, and one target (diabetes incident indicator) were used. There was no direct connection to the target from input based on a priori information that diabetes onset is due to complex risk status, not dominated by one single factor (Fig. 1). The model allows most variables directly from the HRA as well as the calculated values (excluded appraised age, achievable age for reduction of co-linearity. For more detailed information on input variables, see Table 2).

Also, the model excludes variables with over 50% information missing. The choices of activation functions were hyperbolic tangent function at hidden layers and logistic function at the output layer. For the target, we chose a binary response indicating the disease occurrence anytime during the study period.

For the network training, a backpropagation algorithm was used based on the minimization of an error function, the average error between the actual output and the corresponding desired output (not depending on incremental time t_{i}). The risk probability at the output layer that corresponds to the posterior risk probability, is equivalent to a logistic function that represents the association of x_i through the network layers, and the association of x_i and desired output. On the other hand, the output from previous hidden layer (at t) is activated by hyperbolic tangent function, (a) $\tanh(y_t) = 1 - 2/(1 + \exp(2y_t))$, truncates the noise, ranges over $(-1, 1)$, where y_t has the form of weighted

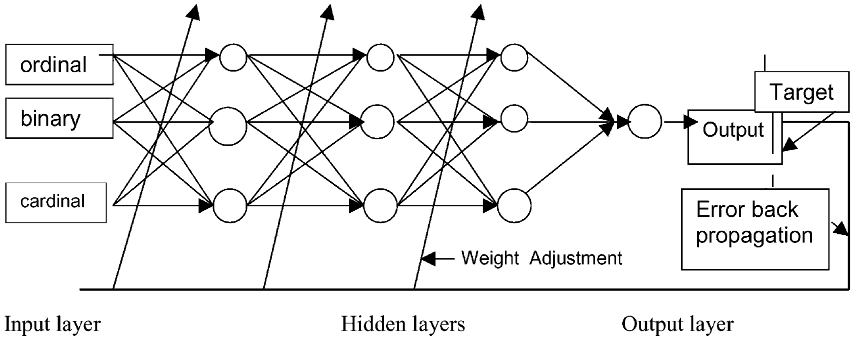

Fig. 1. Single prediction network with error-backpropagation.

sum of linear combination of inputs. Output $y_{m,t}$ (at mth neuron, at time t) at the output layer is: (b) $f(\alpha_m + \sum_l \omega_{lm} g_{3h}(\alpha_l + \sum_k \omega_{kl} g_{2h}(\alpha_k + \sum_j \omega_{jk} g_{1h}(\alpha_j + \sum_i \omega_{ij} x_i))))$, where $f(\cdot)$ is a logistic function to ensure its range (0, 1), and $g_{ih}(\cdot)$, $i = 1, 2, 3$, are the same hyperbolic tangent functions (in the realization of baseline MLP, $m = 1$).

For a single target value (y), the individual error-deviance function $L(y, \mu)$ becomes $-2\ln(1-\mu)$ if $y = 0$; $-2 \ln \mu$ if $y = 1$, where μ is the corresponding output value since the target value is distributed as binary. Error at each epoch, $E = L(y, \mu) = \sum_i f_i \sum_k L(y_{ik}, \mu_{ik})$. Thus, the global objective function to be minimized becomes $O = L(y, \mu)/\sum_i \sum_k f_i$, where $L(y_{ik}, \mu_{ik})$ is the individual error function, μ_{ik} the predicted output for the kth unit of the ith case, y_{ik} the output for the kth unit of the ith case and f_i is the frequency for the ith case.

While selection of a suitable architecture is the key for a robust NN model, an MLP architecture is good at detecting relevant inputs by reducing dimensionality while it still allows as many inputs as necessary. Since the hidden layer forms linear combinations of the inputs, the target will be approximated on the linear subspace spanned by the weight vectors. Even adding irrelevant inputs to the training data doesn't increase the number of hidden units required.

White [25,26] showed that most three hidden layered MLPs can generate the non-linearity of the objective function, which was shown in the model approach. Even with more nodes (4, 5, 6) at each hidden layer, the boundary of decision rule was not fully generated (average error rate =0.56) with fewer (2) hidden layers.

Through the very first few iterations, ANN learns reasonably fast and converges toward the learning of the system. Again, the iterative training is quite fast but training and validation data are not showing much difference in learning pattern. Also, after a few steps learning stops and stays almost at the same error bound over time at the presence of noise. Individual prognostic pattern was not detected since only the occurrence of the event in the proxy time was assessed, which may not be true especially with the binomial classes. In each class, the observation is clustered along a pattern, so, an over dispersion of the data is observed.

3.4. Sequential neural network with adaptive health status as a sequential input

In this section, we present the SMLP for timely classification of diabetes onset, which integrates the sequential information as a prognostic update via corrected probability.

Several modifications were made to the baseline ANN, including stratified randomized sampling with equal number of subjects in each group, random shuffling of inputs presented to the network[10] and, consideration of time-effect on each individual risk factor using sequential presentation of 3 years of HRAs and time dependent target (see Fig. 2). This was used as a base structure to a prognostic model.

In addition, an adaptive learning rate was applied to the sequential network, which is reasonable when the previously learned effect for a network should be reflected to the next learning. This also differentiates from the use of baseline sequentially (refer to Fig. 3b).

[10] The more randomization is complete, the better generalization of the learning is achieved [7].

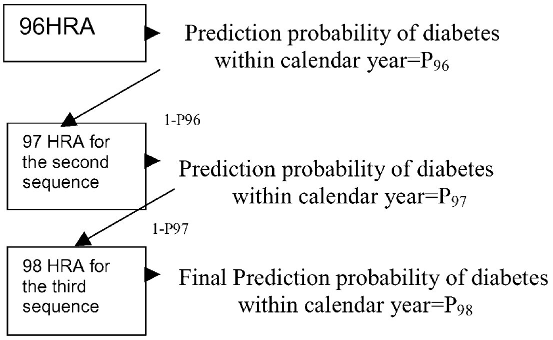
96HRA
Prediction probability of diabetes within calendar year=P96
1-P96
97 HRA for the second sequence
Prediction probability of diabetes within calendar year=P97
1-P97
98 HRA for the third sequence
Final Prediction probability of diabetes within calendar year=P98

Fig. 2. Sequential prediction of diabetes based on backpopagation learning. There is no feed-back from the forward prediction.

According to Ohno-Machado and Musen's sequential versus standard prognostic model for disease pattern recognition [18], standard prognostic approach usually oversimplifies the problem by choosing a single point in time to predict outcomes. With sequential modeling, we looked at the different stages of progress of people at risk according to HRA'96, HRA'97, and HRA'98, sequentially. Along with the HRAs, we sequentially

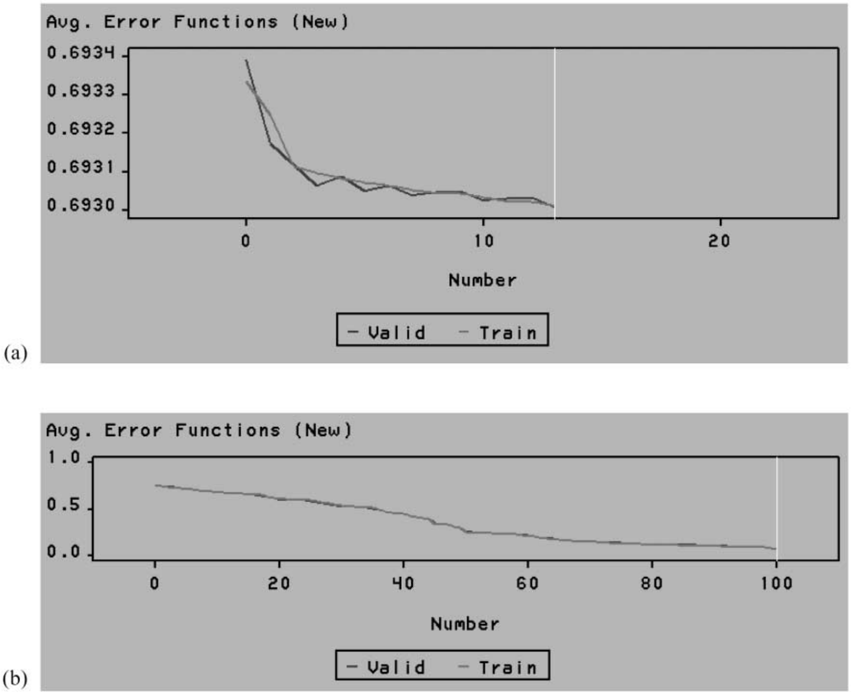

Fig. 3. (a) Average of error function over the iteration (one time): baseline. (b) Average of error function over the iteration (repeated years): baseline.

presented posterior probabilities predicted for diabetes. This utilizes the time-oriented database by considering biases of prediction due to diabetes occurrence in backward time intervals. Unlike recurrent network, the possibility that prediction in a future interval influences backward is excluded.

The time-sensitivity of covariates was considered as in the proportional hazard model but differentiates in the sense that authors considered them as any incremental changes in covariates in time interval. This way, we hope to recognize the individual pattern of prognosis of diabetes. In a long-term prediction, a slow progressing diabetes and fast progressing diabetes may have the same risk probability while the short-term sequential prediction differentiates these patterns, which provides more comprehensive information such as prediction with various prognosis patterns per year. Many previous researchers used aggregated inputs such as average risk factors over the time periods, which could be equivalent to standard proportional hazard model [2,4,9,20].

To achieve a robust solution, local optimization ran over time with random points each time. The architecture was built with the fully connected three hidden layers, one output layer and an input layer. The initial learning rate was randomly selected between 0.1 and 0.4. This learning rate is accelerated depending on the correct classification rate. The combination from standardized input to hidden layers is linear and activation functions were selected to be hyperbolic tangent and softmax,[11] at the hidden layer and the target layer, respectively.

With the sequential architecture, the initial net calculates diabetes occurrence probability based on HRA'96. Also, a time-variant variable was added, indicating the lapse from the first HRA till the first record of diabetes. The target is not just whether diabetes occurred within 3 years but an indicator of the person's diabetes claims case in a sequential manner from 1996. Since the time lapse from HRA collection in 1996 and claims record of 1996 was not always prospective, the 1996 case was considered as latent. An augmented data set with posterior probability of diabetes after 1996 (1-(probability of getting diabetes in 1996) and HRA in 1997) was sequentially used for the following year and so on.

An activation function for the hidden units was needed to introduce non-linearity[12] into the network as in the baseline MLP. A tanh(x) was used for the activation function for the hidden layer, which is suitable for the distribution of the target values. The softmax function, used in the output layer, is a function to make the sum of the outputs equal to one, so the interpretation of outputs as posterior probabilities is meaningful.

The decision regarding the potential diabetes advert is to estimate the posterior probability $P(\text{Class} = C_0|X_1 = x_1, X_2 = x_2, \dots, X_n = x_n, T)$, where T stands for the training data. For this reason, if the predicted probability is the outcome of incremental learning associated with each individual time fragment, the prediction can be applied to each potential case management.

Let the net input to each output unit be q_i, $i = 1, \dots, I$, where I is the number of units. Then, the softmax output P_i is

$$P_i = \frac{\mathrm{e}^{q_i}}{\sum_{i=1}^{I} \mathrm{e}^{q_i}}$$

[11] Multivariate logistic regression.

[12] Without non-linearity, the architecture would be the same as the perceptron with no hidden units, which may not solve the complicated problems as presented here [25].

The overall target function Y is multivariate class variable, according to the following rule: $Y_i(t, \delta) = 1$ if $\delta \in \Delta$ at time t, where $t = 1, 2, 3, 4$ and Δ is a classification class such that $\Delta = A$, when diabetes occurred within a year, B, when diabetes occurred after a year but within 2 years, C, when diabetes occurred after 2 years but within 3 years, D, when diabetes was not observed within 3 years.

However, note that the target for each single sequence is the development of diabetes within the year, and the probability of getting diabetes after the year is used along with the adaptive change of health status for the next sequence. Thus, the final output would be an estimated value of Y ((1, A), (2, B), (3, C), (4, D)) and each year's posterior prediction probability.

The structure of the proposed SMLP as described earlier, learns even from a small change in the input, and produces output according to the change. Furthermore, a sufficient representative training set can lead to a generalization of a given population.

4. Results

4.1. Regression and baseline MLP

After backward elimination steps, we have only main factors in the multiple logistic regression model, which is similar to the result of Sugimori et al. [24]. The result shows the statistical significance of the model with age, weight, BMI, cholesterol, and gender with or without log transformation of predictors (Table 3). Even though BMI and weight are in the same factor group, including both in the model explains the variation statistically (χ^2-test,$P \leq 0.001$). However, the prediction error (0.1381 and 0.2510, respectively) of the fitted model indicates that it loses some other information, which is not measured.

Table 4 is based on the 1996 HRA. The inference is the final prediction of diabetes onset within the 3 years of study period using baseline MLP with the combination of modal pattern and selected variables. As in the regression analysis, weight adjustments indicating internal association of factors, that were used for the final model, found that age, gender, weight, and BMI were good diabetes predictors. The most frequently associated factors to diabetes were high blood pressure, absence days, and indication of a past stroke. High

Table 3
Summary of logistic regression[a]

Summary of regression	Model 1	Model 2
Average square error	0.1371	0.2489
Miss classification rate	0.1889	0.4795
prediction error	0.1381	0.2510
SBC (=BIC)	2024	3390
Effect ($P < 0.05$) except age	0.15–0.8	0.18–0.234

[a] Model 1: logit Pred(DB) = a(Age) + b(Weight) + c(BMI) + d(Cholesterol) + e(Gender); Model 2: logit Pred(DB) = a(Age) + b(log(Weight)) + c(log(BMI)).

Table 4
Modal pattern (MLP)

Risk factors	Mode	Significant value	Predicted diabetes (%)
Absence days	1–3 Days	5–6 Days	62.60
Stroke	No	Yes	62.50
High blood pressure	Low/medium	High	62.23
Arthritis	No	Yes	57.30
Life satisfaction	Good	Not satisfied to fair	54.00
BMI_I	Medium	Obesity	52.94
Back pain	No	Yes	52.51
Diet	1–2 Servings per day	3–4 Servings per day	52.47
Cholesterol range	Low/medium	High	52.08
Planning to change	Yes	Yes	52.08
Weight index	Overweight	Overweight	51.31
Fatty food	3–4 Servings per week	3–4 Servings per week	51.27
Alcohol index	Medium	High	47.50

cholesterol, life dissatisfaction and obesity were also associated with diabetes prediction while medium level of alcohol consumption was associated with the non-diabetes prediction. However, the final misclassification rate was 0.4138, suggesting a search for an alternative model.

The average error over the training data did not converge fast enough, so validation data ended up at the average error level of 0.6 for the baseline MLP (Fig. 3a). In this case, the next steps to adjust the model are to find the co-linearity within the data (results in over-fitting), reduce data at a sufficient level, regularize the weight-decay to avoid over-fitting, and add more information associated with the target. Additional training data need to be added to obtain convergence below a reasonable error-bound.

As discussed in Section 3.3, the target in the baseline neural network does not reflect the time varying effect. This means that the time of occurrence of diabetes has little or no influence on the prediction. The result shows this as a classification model rather than prediction model since its correct prediction (classification) power decreases over the time period. In other words, it predicts very well the occurrence within a year, and its exactness diminishes gradually over time. However, risk of disease is as a result of progressive or accumulated events and it suggests a model, which accounts for the timely procedure of disease progress and its risk measured periodically (Table 5).

Table 5
Prediction error bound in baseline model and sequential model (confusion matrix with test data)

Confusion matrix		0 (%)	1 (%)	Final prediction error
Baseline MLP	From 0	99.6	0.4	0.4138
	From 1	89.7	10.3	
SMLP	From 0	99.7	0.3	0.0025
	From 1	13.96	86.04	

4.2. Sequential model

The values of the risk factors were used in the form as they were asked in the questionnaire. Thus, the suggestive cut-off values as in baseline MLP (high risk versus low risk) can be used as that for risk factors to diabetes. We found somewhat different cut-off values than the national standards, and they appeared to be population-sensitive. In the 3-year prediction (baseline NN), the linearity of the combination function makes the same result of using aggregated risk factors over the 3 years. However, as we discussed earlier, the outcome measures of the sequential ANN were intended to measure the temporal risk change contributing to the probability of each outcome over time, which is different from the previous model.

The risk factors were studied in a time-sensitive manner over 3 years. Diabetes incidences and the risk patterns over 3 years were tested with a SAS procedure (Proc. TRAJ over 3 years). The diabetes incidence indicator within a sequential time frame (annual) was set as the latent class parameter ($p > |T| = 0.022$). Fifteen percent of the repeated sample ($N = 1012$) demonstrated a decreasing Diabetes pattern. This shows the need for a sequential trace of risk change and its outcome as a prediction probability as in the SMLP.

The values in Table 6 show significant differences in prediction of diabetes over the 3-year-period and how the prediction achieves its robustness sequentially. As a result of the final adjustment of the time-sensitive risk factors, which have more impact on the prediction of diabetes in 1999, the prediction model achieved a quite stable prediction probability.

The sequential predictive NNs outperformed the existing linear/non-linear regression model and the baseline neural network in the sense of prediction power according to small changes in the time-varying inputs.

At a threshold of 20%, SMLP achieved 99.3% sensitivity for target 0 and 83.62% sensitivity for target 1, at a threshold of 50% and 86.04% sensitivity for target 1, and 99.7% for target 0 (test data). Overall misclassification rate for the test set ($N = 1939$) was 0.0025 ($N = 5$), and for validation data ($N = 1791$), misclassification rate was 0.0016 ($N = 3$). Sequentially, for the target in the first period, average error was 0.053, with the enhanced prognostic information, for the next period, it was improved as to 0.018, and for the last period, 0.009. Estimated gain over the baseline model, is illustrated in Fig. 4. Convergence rate (average error rate) and reliability (sensitivity/1-specificity) are depicted in Figs. 5 and 6, respectively.

In this case, there was much improvement of 1-specificity, which is a false positive. While the model maintains high sensitivity (close to 1), 1-specificity was also maintained

Table 6
Prediction probability over time

	Predicted DB = 0 (mean probability)	Predicted DB = 1 (mean probability)
Probability for DB (1996)	0.48	0.52
Probability for DB (1997)	0.139	0.86
Probability for DB (1998)	0.038	0.961

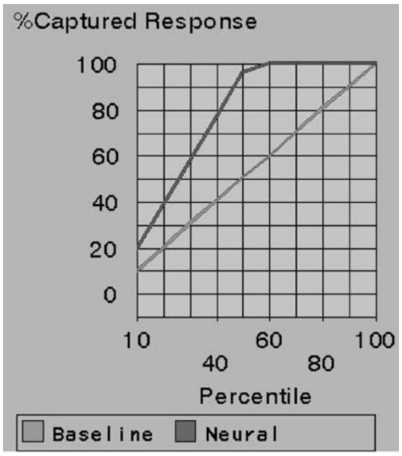

Fig. 4. Percentage of gains of SMLP over the baseline.

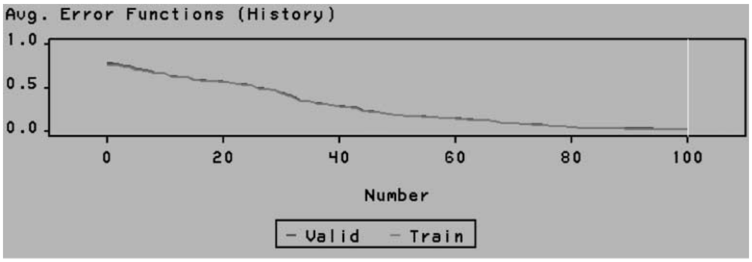

Fig. 5. Average error function (SMLP) for 3 years.

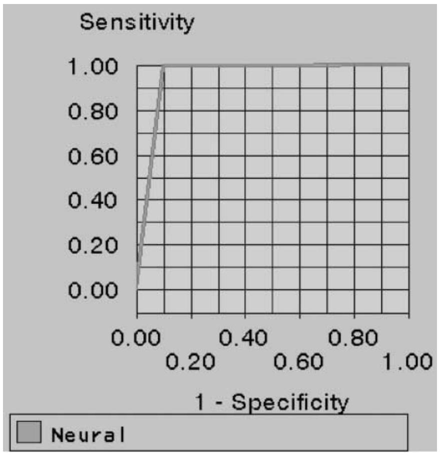

Fig. 6. Prediction power of SMLP.

lower than 0.4 positive S-growth curve. For the training set, the estimated error function decreased from 0.1 to 0.03 after only 933 presentations.

5. Discussion

The significance of the study is the time-sensitive feature of associating risk factors as predictors to the occurrence of diabetes. Time-sensitive HRA variables were sequentially used to predict diabetes within hierarchical time intervals. Thus, any significant changes of risk factors could be reflected to adjust the predictive model in a timely manner. Many of the existing studies in the literature, while attempting to discover the risk factors associated with the medical expenditure/health status, used aggregated values of risk factors rather than the time varying risks. In addition, by restricting domain hierarchically with assessed probability of getting diabetes, the approach successively distinguishes the difference of individual progresses (refer to Table 6).

The HRA variables were able to detect significant lifestyle and health status for the prediction of diabetes in the specified time period. Table 7 lists the significant risk factors to identify the disease-specified risks over a time period based on dual criteria: (1) from the

Table 7
Classification by factors and their contribution to the prediction

	Variable selection	Training class (%)	Test class (%)	Deduced predictability	*F*-ratio
Sex		–	–	–	–
Age		–	–	–	–
BMI	*D*[a]	–	–	0.0017	7.71
Frame	*D*	50.4	51.4	0.0066	3.56
Exercise	*D*	54	54.6	0.0017	0.44
Perceived health	*D*	52	52.7	0.0022	1.71
Smoking	*D*	54	57.3	0.0017	5.32
Smoking × alcohol	*D*	–	–	0.0047	10.10
Alcohol	*D*	–	–	0.0017	2.12
Diet	*D*	52	52	0.0005	0.37
Fatty food	*D*	55	52	0.0017	4.78
Sleep	*D*	50.5	55.9	0.0022	7.88
Absence days	*D*	59	53.9	0.0028	6.12
Preventive service	*D*	57.2	54.7	0.0028	3.47
Stress score	*D*	–	–	0.0012	1.23
Cholesterol range	*D*	54.0	55.3	0.0033	5.32
SBP–DBP	*D*	–	–	0.0039	9.05
Existing disease[b]	*D*	60.6	52.2	0.0017	0.89
High BP	*D*	60.6	60.7	0.0031	2.09
Planning to change	*D*	53.0	52.2	0.0028	1.02

[a] *D* stands for the factor dropped from the SMLP model at a time while maintaining the rest.

[b] Existing disease were appraised based on self-report of heart problem, stroke, cancer, or bronchitis (Table 2). The listed factors were chosen from the final prediction model after weight adjustments, not in order of significance.

model with minimum misclassification, while maintaining age and gender in the model, assess the wrong prediction rate by dropping one variable at a time and (2) using *F*-ratio, difference of sum of squared error (SSE) of the reduced model and the full model to the SSE of the full model, adjusted with appropriate model parameters. For example, existing disease (caner, bronchitis, stroke, heart problem) was associated with the prediction as much as 60.6% (training), and 52.9% for the testing data. By dropping it, the model predictability went down by 0.0017. Past stroke experience alone was associated with the prediction as high as 69.2% while it contributes to the predictability by 0.0031. (On the other hand, association of diabetes as predictor to stroke was found in Sugimori et al. [24].)

Smoking and alcohol at the same time appears more significant for the prediction than smoking or alcohol alone. BMI was assessed as more significant than its class variable (BMI_I, Table 2), sleeping hours, absence days due to illness per year were also significant. Of the variables related to high blood pressure, index variable (high if SBP $>$ 140 mmHg, DBP $>$ 89 mmHg, or taking any prescribed medication) was as much as 60.7% predicted as diabetics while the difference in SBP and DBP was significant contributor to the reduction of SSE. The listed factors were obtained from the final optimization of prediction error. The factors are not in order of significance due to dual criteria in addition to the percent with the predicted diabetes (age and gender were included as the control variable).

These factors were also assessed sequentially and the specified predictors consistently enhances predictability over time. The time-association of risk factors and their suggestive cut-off values are not presented since it is beyond the scope of the paper.

6. Conclusion

A new information system using NN improved the efficacy of early prognosis of future diabetes incidence. By the sequential process, prediction for a given period enhanced prediction accuracy in the following period, so this individually adjusted prediction increased the prediction power in the end. The suggested SMLP outperformed the baseline neural network and the regression model, in terms of final prediction error and sensitivity/1-specificity. In addition, the percentage gains over the baseline models tell us there is a good reason to use this alternative model in terms of early identification of individuals at risk for diabetes at an organization level.

The current work suggests that use of an ANN for other major diseases, such as cancer and heart disease, can enhance identification of individuals who are at risk for respective diseases in a time-sensitive manner. Studies have shown that the majority of health care costs were spent within the organization/intervention level for the top 20% risk people, and are highly associated with major risks as suggested above [5,8]. Thus, prediction of major risks may provide insight for healthcare cost distribution and utilization.

7. Shortcomings and direction for further research

The backpropagation algorithm is one of the efficient algorithms, which simulates error changes as weight changes. That is, we change each weight (including hidden units) by an amount proportionate to the rate at which the error changed with the weight. By using the backpropagation algorithm in SMLP, we hope to show how hidden units represent complex input patterns. Even though the algorithm (learning) may be similar to learning in the human brain, there is no guarantee of obtaining a global minimum point on the error surface since it is heavily affected by bad learning. Also, the algorithm has to back-propagate error (between the target and the output) through the same connection of weight in the reverse direction, which does not happen in the human cognition system. Moreover, there is no competition between good/bad learning to represent the existing pattern better, which can stimulate learning faster.

For prediction purposes, backpropagation works reasonably well, compared to logistics model. However, for an intervention, extracting features from the established ANN structure is quite difficult since it has non-linear association within the hidden units, and within the output units beyond three dimensions.

There were attempts to interpret the effects of individual input variables on the target and their association by means of weight decay, partial derivative (gradient of output with respect to inputs), and weight index. The interconnectivity can be explained with the simulated net weight but the given non-linearity interferes with its clear interpretation. However, there exist algorithms for extracting the knowledge from an NN such as Kohonen net into a set of rules. In a feed-forward NN, the rules can be built in a tree-like structure from each layer and we hope to build such a feature extraction method to parallel the suggested sequential prediction model. This paper intends to leave the significant sequential feature selection to the sequel that follows (manuscript under progress).

In addition, other major diseases such as heart problem or osteoporosis, are to be studied with similar architecture as the diabetes-prediction. To provide an appropriate time for an intervention before disease onset and how the early detection of risk factors contributes to the disease, medical expenditure (%), shall be addressed. In addition, since decision tree classification and self organizing map (SOM) were suggested as a more reliable solution for the prior probability [1,12,14], the reliability of the model will be assessed as well as using the ensemble methods such as boosting and bagging to integrate the results to achieve the robustness of the suggested prediction model.

Appendix. Health Risk Appraisal (HRA)

Health Hazard Appraisal was developed in the 1960s by Robbins and Hall and has given rise to a number of risk assessment devices that may be grouped as HRA methods. These aim to stimulate a person to modify his or her lifestyle by providing a quantitative estimate of the effects of a number of predictive factors (biological, lifestyle, family history) on his or her likelihood of dying prematurely. The university of Michigan's HRA asks 38 questions covering smoking, alcohol use, use of seat belt, dietary habit, hours of sleep and existing medical conditions. The HRA data used here is the outcome of the work site health promotion program to motivate more employees to be aware of their health status and to maintain a healthy status as well as to modify it by participating in available work-site intervention programs [Healthier People version 4.0: July, 1991].

References

[1] Andrews R, Geva S. Inserting and extracting knowledge from constrained error backpropagation networks. In: Proceedings of the 6th Australian Conference on Neural Networks. NSW, Sydney, 1995.

[2] Apolloni B, Avanzini G, Cesa-Bianchi N, Ronchini G. Diagnosis of epilepsy via backpropagation. In: Proceedings of the IJCNN, vol. 2. Washington, DC, 1990. p. 571–4.

[3] Bishop CM. Neural networks for pattern recognition. Oxford: Oxford University Press, 1995.

[4] Bounds DG, Lloyd PJ. MLP for low backpain. Proc IEEE: Int Conf NN 1988;2:481–9.

[5] Edington DW. Health behaviors and risk appraisal. Worksite Health Promot Econ 1995;97–115.

[6] Edington DW, Edington M, Yen L. The formula for proving your program's worth. Employee Service Manage 1989;31:12–7.

[7] Faraway J. Data splitting strategies for reducing the effect of model selection on inference. Comput Sci Stat 1998;30:332–41.

[8] Foxray N, Thamer M, Gardner E, Chan JK. Economic consequences of diabetes mellitus in the United States in 1997. Report from the American Diabetes Association, 1999.

[9] Gazmararian JA, Foxman B, Yen LT, Morgenstern H, Edington DW. Comparing the predictive accuracy of Health Risk Appraisal: The Centers for Disease Control versus Carter Center Program. Am J Public Health 1991;81:1296–301.

[10] Hebb DO. The organization of behaviors. New York: Wiley. 1949.

[11] Heyman B, Henriksen M, Maughan K. Probabilities and health risks: a qualitative approach. Social Sci Med 1998;47:1295–306.

[12] Ivanova I, Kubat M. Initialization of neural networks by means of decision trees. Knowledge-Based Syst 1995;8(6).

[13] Kleinbaum DG. Logistic regression: a self learning text. New York: Springer. 1998.

[14] Lawrence S, Burns IB, Tsoi AC, Giles CL. Neural network classification and prior class probabilities, lecture notes in computer science state-of-the-art surveys. New York: Springer. 1998.

[15] Lim CP, Harrison RF, Kennedy RL. Application of autonomous neural network systems to medical pattern classification tasks. Artif Intel Med 1997;11:215–39.

[16] Maindonald JH. New approaches to using scientific data statistics. Data Mining & Related Technologies in Research & Research Training. The Australian National University, 1998.

[17] McCulloch WS, Pitt WS. A logical calculation of the ideas immanent in nervous activity. Bull Math Biophys 1943;5:115–33.

[18] Ohno-Machado L, Musen MA. Sequential versus standard neural networks for pattern recognition: an example using the domain of coronary heart disease. Comput Biol Med 1997;27:267–81.

[19] Rosenblatt F. The perceptron: a perceiving and recognizing automation. Technical Report 85-46-1. Cornell Aeronautical Laboratory, 1957.

[20] Shanker MS. Using neural networks to predict the onset of diabetes mellitus. J Chem Inform Comput Sci 1996;36:35–41.

[21] Shao J. Linear model selection by cross-validation. J Am Stat Assoc 1993;88:486–94.

[22] Smith K, McKinlay S, Thorington B. The validity of Health Risk Appraisal instruments for assessing coronary heart disease risk. American Institute for Research, 1986.

[23] Stargren EI, Oberg BE. Predictive factors for one-year outcome of low-back pain and neck pain in patients treated in primary care: comparison between the treatment strategies chiropractic and physiotherapy. Pain 1998;77:201–7.

[24] Sugimori H, Miyakawa M, Yoshida K, Izuno T, Takahashi E, Tanaka C, et al. Health risk assessment for diabetes mellitus based on longitudinal analysis of MHTS database. J Med Syst 1998;22:27–32.

[25] White H. Connectionist nonparametric regression: multi-layer feed-forward networks can learn arbitrary mappings. Neural Networks 1990;3:535–50.

[26] White H. Nonparametric estimation of conditional quantiles using neural networks. In: Page C, LePage R, editors. Proceedings of the 23rd Symposium on the Interface: Computing Science and Statistics. ASA, Alexandria, 1992. p. 190–9.