BMJ open
accessible medical research

Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait—a cohort study

Bassam Farran,[1] Arshad Mohamed Channanath,[1] Kazem Behbehani,[2] Thangavel Alphonse Thanaraj[1]

To cite: Farran B, Channanath AM, Behbehani K, *et al*. Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait—a cohort study. *BMJ Open* 2013;**3**:e002457. doi:10.1136/bmjopen-2012-002457

▸ Prepublication history and additional material for this paper are available online. To view these files please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2012-002457).

Received 6 December 2012
Revised 31 March 2013
Accepted 12 April 2013

[1]Integrative Informatics, Dasman Diabetes Institute, Dasman, Kuwait
[2]Director-General, Dasman Diabetes Institute, Dasman, Kuwait

Correspondence to
Dr Thangavel Alphonse Thanaraj; Alphonse.Thangavel@dasmaninstitute.org

ABSTRACT

Objective: We build classification models and risk assessment tools for diabetes, hypertension and comorbidity using machine-learning algorithms on data from Kuwait. We model the increased proneness in diabetic patients to develop hypertension and vice versa. We ascertain the importance of ethnicity (and natives vs expatriate migrants) and of using regional data in risk assessment.

Design: Retrospective cohort study. Four machine-learning techniques were used: logistic regression, k-nearest neighbours (k-NN), multifactor dimensionality reduction and support vector machines. The study uses fivefold cross validation to obtain generalisation accuracies and errors.

Setting: Kuwait Health Network (KHN) that integrates data from primary health centres and hospitals in Kuwait.

Participants: 270 172 hospital visitors (of which, 89 858 are diabetic, 58 745 hypertensive and 30 522 comorbid) comprising Kuwaiti natives, Asian and Arab expatriates.

Outcome measures: Incident type 2 diabetes, hypertension and comorbidity.

Results: Classification accuracies of >85% (for diabetes) and >90% (for hypertension) are achieved using only simple non-laboratory-based parameters. Risk assessment tools based on k-NN classification models are able to assign 'high' risk to 75% of diabetic patients and to 94% of hypertensive patients. Only 5% of diabetic patients are seen assigned 'low' risk. Asian-specific models and assessments perform even better. Pathological conditions of diabetes in the general population or in hypertensive population and those of hypertension are modelled. Two-stage aggregate classification models and risk assessment tools, built combining both the component models on diabetes (or on hypertension), perform better than individual models.

Conclusions: Data on diabetes, hypertension and comorbidity from the cosmopolitan State of Kuwait are available for the first time. This enabled us to apply four different case–control models to assess risks. These tools aid in the preliminary non-intrusive assessment of the population. Ethnicity is seen significant to the predictive models. Risk assessments need to be developed using regional data as we demonstrate the applicability of the American Diabetes Association online calculator on data from Kuwait.

ARTICLE SUMMARY

Article focus

- To implement machine-learning-based classification models and risk assessment tools for diabetes, hypertension and comorbidity with data from Kuwait national health network.
- To assess the importance of ethnicity and of using regional data in risk assessment in a cosmopolitan state such as Kuwait.

Key messages

- Machine-learning-based classification models and risk assessment tools result in high accuracy and little uncertainty. Onsets of type 2 diabetes in general and in hypertensive population as well as of hypertension in general and in diabetic population are modelled.
- Two-stage aggregate calculators have dramatic increase in risk assessments.
- Ethnicity is very important to the predictive models; risk assessments developed using regional data outperform generalised global assessments.

Strengths and limitations of this study

- For the first time in the Middle East region (that has high incidence of diabetes), large-scale health data from Kuwait are available for research. Detailed classification models and risk assessment tools are made available.
- Integration of data from primary health centres and hospital records in the Kuwait Health Network is an ongoing task; as a result, data are not available on all items especially biochemical parameters.

INTRODUCTION

Incidence of diabetes, along with hypertension and other complications, is ever increasing worldwide. One in 10 adults suffers from diabetes, and 1 in 3 adults suffers from hypertension. A considerable portion of the world population suffers coexistent diabetes and hypertension. Diabetes leads to complications such as blindness, amputation and cardiovascular diseases.[1] Hypertension is directly responsible for 12.8% of all global death, and it causes around half of all deaths from stroke and heart diseases. With obesity levels increasing among young children and adolescents, type 2 diabetes and hypertension are starting to show in the young population—implying that such children will live with disorders that are usually associated with adults and the older population. The onset and prevalence of diabetes, hypertension and comorbidity are often seen in the prime working years of the affected population and these people live a lower quality of life during a significant portion of their productive years. This leads to decreasing productivity, increasing social costs and to placing a very high burden on the healthcare system.[2]

The global epidemic of diabetes has not spared the Arabian Gulf, particularly Kuwait that seems to have the highest prevalence in the peninsula.[3 4] Our recent report using nationwide data assesses the prevalence of type 2 diabetes at 33% (among Asian expatriates) and 25% (among natives), and of hypertension at 37% (among Asian expatriates) and 28% (among natives) in Kuwait.[5] In order to meet this challenge, efficient (preventive) strategies are needed to control risk factors like obesity, blood pressure, diet and inactivity. An effective way to address this issue is to have a non-intrusive preliminary screening tool that could identify the patients' risks for developing diabetes and/or hypertension. This can be used either on individual basis or on whole population level to identify groups of high-risk patients and subject them to preventive measures.

One-third of the population of Kuwait is composed of Kuwaiti natives and the remaining large proportion is composed of expatriates. This is very valuable, as it enables us to study the relationships between different ethnicities and their impact on risk factors and the development of diabetes.[4 6] Further, health informatics data are increasingly becoming huge as well as encompassing a large number of variants. Modelling intricate relationships across ethnicities and handling huge data require sophisticated techniques. A variety of computational and mathematical techniques has been deployed by researchers in the field to build not only predictive models but also physiological models for diabetes treatment. Techniques often used to build predictive models are logistic and Cox regression,[7] and those used to build physiological models include operations research methods to predict future glycaemia levels[8] in diabetic patients, compartmental modelling methods for blood glucose control[9] and computational simulations of blood glucose profiles.[10 11]

We implement in this study four machine-learning techniques to model diabetes and hypertension in Kuwaiti inhabitants. We further evaluate the performance of publicly available tools built with data from other ethnicities on data from Kuwait.

DATA, RESEARCH DESIGN AND METHODS

Data from Kuwait Health Network

Data for this study were taken from Kuwait Health Network, which is an initiative of Dasman Diabetes Institute in collaboration with the Ministry of Health and the Public Authority of Civil Information of the State of Kuwait. The network integrates health data from primary health centres with clinical data from different hospitals across Kuwait.

The data records are retrospective over the last 12 years. The ascertainment of diagnosis for diabetes and hypertension is through clinical diagnosis. The names and the civil identification numbers of the patients are anonymised before data are exported for use by researchers.

Data content

The current iteration of data contains 13 647 408 records associated with 300 489 hospital visitors labelled as diabetic/non-diabetic and hypertensive/non-hypertensive. Upon performing sanity checks, the final data set resulted in a total of 270 172 participants of which 74 134 are type 2 diabetic, 58 745 are hypertensive and 30 522 are comorbid. Ethnic distribution of the participants is Kuwaiti natives (55%), Asian expatriates (24%), Arab expatriates (16%) and expatriates from other countries (5%). The data include information on demography, anthropometry, vital signs, diagnosis and clinical laboratory measurements.

Caveats with data

The integration of data from primary health centres and hospital records in the Kuwait Health Network is an ongoing task; as a result, not all the data items are available for all the participants, thus limiting the sizes of the data sets, in certain instances, for using to model different disease states. The data on clinical measurements are partial at this stage, and this hinders the development of advanced models.

METHODS

Data mining and machine-learning calculations are performed using MATLAB (MATrix LABoratory). Four different techniques as described below are implemented.

Classification accuracy at best random classifier for a case–control data set

Classification Accuracy is defined as the proportion of correctly classified results in a population. The classification accuracy, A, of an algorithm c is given as

$$A_c = \frac{t}{N} 100\%$$

where t is the number of samples correctly classified and N is the total number of sample cases. We therefore

calculate the accuracy at best random classifier as the maximum of (d/N, nd/N), where d is the number of diabetics and nd the number of non-diabetics. This is the maximum achievable if a model is to predict all test points either as diabetic or non-diabetic.

Generalisation accuracy and cross validation

Since the data are not split into training or testing data, we resort to fivefold cross validation (CV), which is often used in the machine-learning community.[12 13] Fivefold CV is used to assess how well a classification model will generalise to an independent data set, and involves splitting the data set into five equal mutually exclusive subsets. Then, each of the subsets is used once for testing (with the other four being used for training). This process is repeated five times, with each of the five subsets being used exactly once for testing. The five results from the folds are then averaged to produce the generalisation accuracy.

Logistic regression

Logistic regression (LR) is a generalised linear model that estimates the probability of the occurrence of an event $\vec{x}$ by fitting data onto a logistic curve:

$$f(\vec{x}) = \frac{1}{1+e^{-\vec{a}\,\vec{x}}}$$

where $\vec{a}$ is the vector containing the regression coefficients. The number of regression coefficients is the same as the number of measurements we have for each of the hospital visitors—one coefficient for each independent variable. This statistical technique has excelled in the health domain[14] to capture relationships that exist among several independent variables and a binary output variable. We use fivefold CV to calculate the generalisation accuracy.

k-Nearest neighbours

This is perhaps the simplest classification algorithm, and involves, for each test point, finding the k-closest training points to it and labelling the test point by a majority vote.[15 16] For example, if a majority of the k-nearest training points to a new patient are diabetic (or hypertensive), then he/she will be classified as diabetic (hypertensive). To determine closeness, Euclidean distance is used in the case of continuous variables and Hamming distance for binary data and the former is defined as follows for vectors $\vec{p}$ and $\vec{q}$ of length N:

$$d_{\text{Euclidean}}(\vec{p},\vec{q}) = \sqrt{\sum_{i=1}^{N}(q_i - p_i)^2}$$

The Hamming distance for a binary string of length N is the number of positions for which the corresponding bits are different, that is, it is the population count (number of ones) in ($\vec{p}$ XOR $\vec{q}$). The best value for k (the nearest-neighbour count) is selected using fivefold CV as below: we take a set of possible values for k such as {4,5,6 and 7}, and, for each value in this set, we perform fivefold CV to obtain a generalisation accuracy. The value of k that yielded the highest accuracy is selected for use in our experiments.

Support vector machines

These are supervised learning algorithms that can be used for classification and regression. The standard formulation for support vector machine (SVM) learns from a set of input data (in our case, data associated with the hospital visitors that are diabetic or hypertensive, as the case may be) and predicts, for each new point, which of the two possible classes it belongs to. This is done by fitting a decision boundary between training points from the two different classes (a tutorial is available at http://www.cs.columbia.edu/~kathy/cs4701/documents/jason_svm_tutorial.pdf). SVMs' success lies in its ability to maximise the *margin*, which denotes the distance between an example and the decision boundary.[17] Then, since the unseen examples will be close to the training examples, the large margin ensures that the test cases are classified correctly. We use C-SVM, which is a formulation of the SVM that integrates a cost variable C. The cost variable controls the trade off between allowing training errors and forcing rigid margins. Default settings for the radial basis function kernel, σ=0.1, C=10 are used unless otherwise specified, where the Gaussian radial basis function used is defined as below:

$$K(\vec{p},\vec{q}) = \exp\left(\frac{\|\vec{p}-\vec{q}\|^2}{2\sigma^2}\right)$$

The variable σ is the width of the basis function, which determines the area of influence of the support vectors in the data space.

Multifactor dimensionality reduction

Multifactor Dimensionality Reduction (MDR) is a non-parametric and genetic model-free alternative to LR for detecting and characterising non-linear interactions among discrete genetic and environmental attributes. It is used to detect combinations of independent variables that interact to influence a dependent or class variable[18] (which assumes value of diabetic/non-diabetic or hypertensive/non-hypertensive in our case). The basis of the method is a constructive induction algorithm that converts two or more variables to a single attribute. Constructive induction is the process of transforming the original representation of hard concepts with complex interaction into a representation that highlights regularities. The ultimate goal of the algorithm is to create or discover a representation that aids the detection of non-linear interactions among the new attributes such that the overall prediction is better than that of the original representation. This technique has been successfully used in the medical field, with applications on

cancer research,[19] cardiovascular diseases[20] and diabetes.[21] We use the default configuration of the software (available on http://www.multifactordimensionality-reduction.org/), and only report the best performing model. The default settings are random seed=0, attribute count range=1–4, CV count=5, track top models=20, search type=exhaustive.

Risk assessment tools

Of the models mentioned above, k-NN is best suited for adaptation to output the result of classification in the form of 'low', 'borderline' and 'high' risk scores. By way of example of a 7-NN model, if, for a given test point, the number of diabetic patients within the k=7 closest neighbours is (0–1), the test patient is considered to be of 'low' risk; if (2–3), the test patient is considered to be of 'borderline' risk; and (4–7), the test patient is considered to be of 'high' risk. Various split schemas (as illustrated by an example presented in see online supplementary table S1) were tried and we chose the one that does not let high number of diabetic (or hypertensive) patients go undetected (ie, get assigned 'low' risk). This is because it is more dangerous to let a diabetic (or hypertensive) patient go unnoticed than to have a false alarm.

Different pathology conditions that are modelled

Classification models and risk assessment tools are developed for the following: (1) diabetes in general population; (2) diabetes in hypertensive patients; (3) hypertension in general population and (4) hypertension in diabetic patients. Further, a two-stage aggregate model for diabetes is built to take advantage of the models for diabetes in general population, and for diabetes in hypertensive population; a similar aggregate model is built in the case of hypertension also. These models and tools use only non-intrusive parameters such as height, weight, age, gender, ethnicity, hypertension and family history of hypertension and diabetes.

Choice of online risk assessment tools from other ethnicity for evaluating the applicability to Kuwaiti population

To evaluate the applicability of risk assessment tools developed with other data from other regions to data from Kuwait, we chose the diabetes risk test tool from the American Diabetes Association (http://www.diabetes.org/diabetes-basics/prevention/diabetes-risk-test/; last accessed 22 November 2012) that has been built using data available from within the USA.

RESULTS

Classification models for diabetes in general population

Classification models are built on a data set of 10 632 (2853 diabetic and 7779 non-diabetic) participants; these participants (chosen irrespective of their diagnosis for hypertension) have complete records of height, weight, age, gender, ethnicity, hypertension diagnosis and a family history of hypertension and diabetes. The best random classifier for the data set leads to an accuracy of 73.2%. Results below are obtained using fivefold CV, as are the results of the following subsections.

All of the four techniques perform almost equally well with a classification accuracy of up to 81.3% (table 1), which is significantly better than the best random classifier for the data set (at 73.2%). Classification accuracies obtained with individual models are 80.7% with LR, 81.3%±1.3% with SVM (RBF kernel, σ=0.1, C=10), 78.6%±0.85% with 9-NN and 78.30% with MDR.

Classification models for hypertension in general population

Classification models are built on a data set of 10 632 (6759 hypertensive and 3873 non-hypertensive) participants; these participants (chosen irrespective of their

Table 1 Performance of various classification models built for modelling diabetes and hypertension

Type of classification	N for case/control	Classification accuracy at the best random classifier (%)	Classification accuracy for the different models used (%)			
			LR	SVM	k-NN	MDR
(i) Diabetes in general population	2853/7779	73.2	80.7.	81.3±1.3	78.6±0.85	78.30
(ii) Diabetes in hypertensive population	1322/1382	51.1	70.9	87.4±1.1	75.6±2.7	72.1
(iii) Two-stage aggregate of (i) + (ii) – diabetes	2853/7779	73.2	N/A	84.9	88.2	N/A
(iv) Hypertension in general population	6759/3873	63.6	82.4	82.4±0.6	80.0±0.8	80.9
(v) Hypertension in diabetic population	2427/5994	71.2	80.1	80.8±1.3	76.0±1.4	67.3
(vi) Two-stage aggregate of (iv) + (v) – hypertension	1322/1382	51.1	N/A	95.3	90.3	N/A
Kuwait-specific data sets						
(i) Diabetes in general population	1334/4179	75.8	79.4	79.4	77.6	75.9
(ii) Hypertension in general population	3451/2062	62.6	80	79.9	76.8	77.9
Asian-specific data sets						
(i) Diabetes in general population	976/2061	67.9	84.3	84.3	81.4	83.6
(ii) Hypertension in general population	1933/1104	63.7	86.8	86.8	83.3	83.8

LR, logistic regression; SVM, support vector machine; k-NN, k-nearest neighbours; MDR, Multifactor Dimensionality Reduction.

diagnosis for diabetes) have complete records of height, weight, age, gender, ethnicity, diabetes diagnosis and a family history of hypertension and diabetes. Experiments are performed using the same setup as before, with the best random classifier achieving 63.6%. Fivefold CV in k-NN model gave an optimal k=7, yielding an 80±0.8% classification accuracy (see table 1), whereas SVM performed slightly better at 82.4±0.6% (RBF kernel, σ=0.01, C=100). All four techniques perform almost equally well with a classification accuracy of up to 82.4% much larger than the one obtained with the best random classifier for the data set (at 63.6%).

Classification models for diabetes in the hypertensive population and vice versa

Since hypertension and diabetes share many common predisposing factors, and that disposition to one increases the proneness to the other,[22 23] it is interesting to see how accurately the models can predict the onset of one disorder given the presence of the other.

Diabetes in the hypertensive population

Classification models are built on a data set of 2704 hypertensive participants, of which 1322 developed diabetes after the diagnosis for hypertension. The best random classifier for the data set achieved a classification accuracy of 51.1%. Fivefold CV results for k-NN (at k=6) and SVM (RBF kernel, σ=0.1, C=10) achieve accuracies of 75.6±2.7% and 87.4%±1.1%, respectively (table 1) both significantly higher than that achieved with the best random classifier (51.1%).

Hypertension in the diabetic population

Classification models are built on a data set of 8421 diabetic participants, of which 2427 developed hypertension after the diagnosis for diabetes. The best random classifier achieves a classification accuracy of 71.2% for the data set. Fivefold CV results for k-NN (k=10) and SVM (RBF kernel, σ=0.1, C=10) achieve accuracies of 76.0%±1.4% and 80.8%±1.3%, respectively, both higher than that achieved with the best random classifier.

The accuracies obtained with the best random classifiers for the above two data sets differ considerably at 71% for the hypertension in diabetic population and 51% for the diabetes in hypertensive population. This large difference is probably a reflection of differential intrinsic proneness for the two disorders—it is more often the case that hypertension develops after the onset of diabetes than vice versa.[23]

Two-stage aggregate models

In the previous sections, two types of models are demonstrated for each of diabetes and hypertension. Taking diabetes as an example, the two models are diabetes in general population, and diabetes in hypertensive population; a two-stage aggregate model can be built for diabetes by processing the data through these two component models (see figure 1A for the flow of data). Achieved classification accuracies from the aggregate model (for both diabetes and hypertension) built using the SVM and k-NN techniques ranged from 85% to 88% for diabetes and from 90% to 95% for hypertension (see table 1) are significantly higher than those obtained from the component models (at 76–79% for diabetes and 76–80% for hypertension).

Ethnicity in classification models

Kuwaiti natives and Asian expatriates have significant differences in prevalence and in trends associated with features (such as age at onset and body mass index) of diabetes and hypertension. In order to test the influence of ethnicity on the performance of the models, we performed the following two analyses:

1. Upon building separate classification models for Kuwaiti natives and Asian expatriates (table 1), we find that the classification algorithms are not performing equally well for the two ethnicities. The accuracy values obtained with the data set of Asian expatriates (eg, 84.3% for diabetes in general population and 86.8% for hypertension in general population using LR) are consistently higher than (a) those obtained with the data set of Kuwaiti natives by 5–8% (as LR obtained 79.4% and 80% for the diabetes and hypertension calculators respectively) and (b) those obtained with the overall data set (that includes participants from all ethnicities) by at least 3%. The Kuwaiti-specific data set does not show any improvement in accuracy over those obtained using data sets that include all ethnicities.
2. Upon building classification models for the overall set (that includes participants from all ethnicities) by excluding the ethnicity field, we find that the resultant classification accuracies are reduced by at least 6%. This indicates that the machine-learning techniques are capturing information from the ethnicity variable when included in the data set.

Parameters used by the classification models

With the outputs from the LR models, it is possible to examine the relative importance of parameters for prediction by looking at whether the associated coefficients are significantly different than 0. This is done by examining the p value associated with each coefficient, and if it less than 0.05, it can be concluded that the parameter is significant for classification. The variables that emerged from each of the modelled conditions are (1) Hypertension in diabetic population: body mass index (BMI), age and family history for diabetes; (2) Diabetes in hypertensive population: ethnicity and family history for hypertension; (3) diabetes in general population: BMI, age, gender, ethnicity, diagnosis for hypertension and family history for hypertension and (3) hypertension in general population: BMI, age, ethnicity and diagnosis for diabetes. A significant observation from the above results is that the data on hypertension are of significant predictive values for diabetes and vice

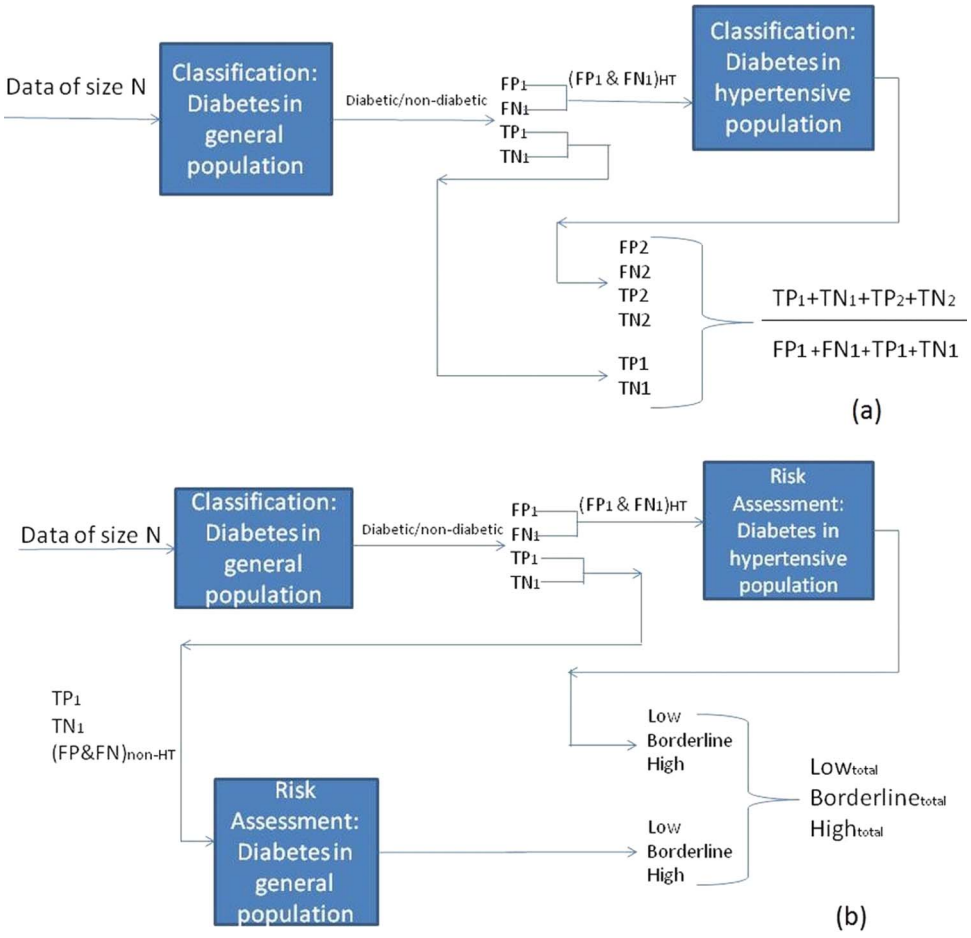

Figure 1 Illustration of the methodology and flow of data for two-stage aggregate classification model and the two-stage aggregate risk assessment tool for diabetes. (A) Illustration for the two-stage aggregate classification model for diabetes. A data set is passed through the classification model for diabetes in general population (ie, irrespective of the status on hypertension onset)—the output is classified as TP1, TN1, FP1 and FN1. Of the false-positives and false negatives, the ones that also have the affliction of hypertension are passed through the classification model for diabetes in hypertensive population—the output of the second model can be classified as TP2, TN2, FP2 and FN2. The combined classification accuracy of the aggregate model is then defined as (TP1+TP2+TN1+TN2)/(TP1+TN1+FP1+FP2). FP, false positives; TP, true positives; FN, false negatives; TN, true negatives; HT, hypertension. $(FP1 \text{ and } FN1)_{HT}$ indicates those patients who are tested false positives and false negatives and are hypertensive. (B) Illustration for the two-stage aggregate risk assessment tool for diabetes. A data set is passed through the classification model for diabetes in general population (ie, irrespective of the status on hypertension onset)—the output is classified as TP1, TN1, FP1 and FN1. Of the false positives and false negatives, the ones that also have the affliction of hypertension are passed through the risk assessment tool for diabetes in hypertensive population; of the false positives and false negatives, the non-hypertensive ones along with the true positives and true negatives are passed through the risk assessment tool for diabetes in general population. The combined risk assignment is the aggregate of risk assignments from the two component risk assessment tools. FP, false positives; TP, true positives; FN, false negatives; TN, true negatives; HT, hypertension. $(FP1 \text{ and } FN1)_{HT}$ indicates those patients who are tested false positives and false negatives and are hypertensive.

versa. This observation confirms that disposition to diabetes increases the proneness to develop hypertension and vice versa.

Risk Assessment Tools

Risk assessment tools are built for both diabetes (in the setting of diabetes in the general population) and hypertension (in the setting of hypertension in the general population). We develop separate risk assessment tools for the whole set (including all ethnicities) as well as for different ethnicities (Kuwaiti natives and Asian expatriates). The results are given in table 2, with the models developed in this study called IHBI. With the k-NN models, we see more diabetics classified higher up in risk level and more non-diabetics at the lower risk level. With the All Ethnicity assessment tool, 12.4% of the diabetics are assigned low risk as compared to 70.7% of the non-diabetics and 59.2% of the diabetics are assigned high risk as compared to 9.3% of the non-diabetics. Of the ethnic-specific assessment tools, the Asian ethnicity-specific tool is doing better than the overall tool: 9.6% of the diabetics are assigned low risk as compared to 73.6% of the non-diabetics and 75.5% of the diabetics are assigned high risk as compared to 9.8% of the non-

Table 2 Performance of the IHBI risk assessment tools (as built in this study) and ADA assessment tool for diabetes on Kuwaiti natives and Asian expatriates

Data set	Risk assignment by the ADA tool (%)		Risk assignment by the IHBI (k-NN) tool (%)		Risk assignment by the IHBI_Aggregate (k-NN) tool (%)	
	Diabetic patients	Non-diabetic patients	Diabetic patients	Non-diabetic patients	Diabetic patients	Non- diabetic patients (%)
All ethnicities (k=7,N=10632)						
'Low' risk	23.4	16.7	12.4	70.7	6.6	71.4
'Borderline' risk	32.7	32.2	28.4	20.0	18.9	23.7
'High' risk	43.9	51.1	59.2	9.3	74.5	4.9
Kuwaiti natives (k=8,N=5513)*						
'Low' risk	15.3	9.7	11.4	64.6	4.9	64.4
'Borderline' risk	38.1	44.2	31.6	24.4	30	25.5
'High' risk	46.6	46.1	57.0	10.9	65.2	10.2
Asians expatriates (k=7, N=3036) *						
'Low' risk	23.5	45.9	9.6	73.6	2.0	68.8
'Borderline' risk	16.8	11.3	14.9	16.6	9.6	19.1
'High' risk	59.6	42.8	75.5	9.8	88.4	12.2

*Split schema used is (0–1)—'low' risk; (2–)—'borderline' risk; (4–8)—'high' risk.
ADA, American Diabetes Association; k-NN, k-nearest neighbours.

diabetics. As a next step, we implemented the two-stage aggregate risk assessment tool. The flow of data and the methodology are as illustrated in figure 1B. The aggregate assessment tool gives even better performance (table 2): with the All Ethnicity risk assessment tool, up to 74.5% of diabetic patients are grouped into 'high' risk; as low as 4.9% of non-diabetics are grouped into 'high' risk; and with the Asian ethnicity-specific tool, it is even better with 88.4% of diabetic patients grouped as 'high' risk.

The performance of the risk assessment tools for hypertension is given in table 3. Both the types of risk assessment tools (the general one and the aggregate one) perform equally well in assigning 'high' risk to 92–94.8% of the hypertensive population (that includes all ethnicities); however, the assignment of non-hypertensive population to three classes of output is almost random (at around 30–37% each) with the exception of the Asian-specific tool that assigns 'low' risk to 49% of non-hypertensive population.

Table 3 Performance of the IHBI risk assessment tools for hypertension (as built in this study) on Kuwaiti natives and Asian expatriates

Data set	Risk assignment by the IHBI (k-NN) tool		Risk assignment by the IHBI_Aggregate (k-NN) tool	
	Hypertensive patients (%)	Non-hypertensive patients (%)	Hypertensive patients (%)	Non-hypertensive patients (%)
All Ethnicities (k=8,N=10632)				
'Low' risk	1.1	37.6	0.28	37.6
'Borderline' risk	6.5	31.8	4.9	31.9
'High' risk	92.4	30.6	94.8	30.5
Kuwaiti natives (k=8,N=5513)*				
'Low' risk	0.43	22.8	0.26	27
'Borderline' risk	4.9	34.7	5.6	38.1
'High' risk	94.6	42.5	94.2	34.9
Asian expatriates (k=8,N=3036)*				
'Low' risk	1.2	43.6	0.1	48.5
'Borderline' risk	4.1	27.4	2.6	26.1
'High' risk	94.7	29.1	97.3	25.5

*Split schema used is (0–1)—'low' risk; (2–3)—'borderline' risk; (4–8)—'high' risk.
k-NN, k-nearest neighbours.

Cross-applicability of risk assessment tools across different populations

We demonstrate that a risk assessment tool built with a specific regional data does not generalise and perform as well on other population groups, by evaluating the performance of the ADA online diabetes risk test tool (made available by American Diabetes Association), which is built using patients from the USA[24] (table 2). With the IHBI models, more diabetics are seen classified higher up in risk level (eg, 59.2% for the all-ethnicities calculator) and more non-diabetics at the lower risk level (70.7% for the all-ethnicities calculator), while with the ADA risk test tool, a random assignment is seen. Diabetic patients are not preferentially assigned 'high' risk nor are non-diabetic patients being preferentially assigned 'low' risk—44% of diabetics and 51% of non-diabetics are both assigned 'high' risk; and 23% of diabetics and 17% of nondiabetics are assigned 'low' risk. With Kuwaiti natives-specific data set, the ADA tool performs even more randomly with half of the diabetics as well as non-diabetics-assigned 'high' risk where as the IHBI models predicts 65% of the diabetics as 'high' risk and 64% of the non-diabetics as 'low' risk. Thus, the tools that are trained with data from elsewhere do not perform well on data from Kuwait.

DISCUSSION

The applicability of machine-learning techniques to differentiate type 2 diabetics from non-diabetic population and hypertensive patients from non-hypertensive ones is examined. The models are trained with data on non-intrusive basic parameters from the nationwide Kuwait Health Network on diabetes and hypertension. Classification accuracy, which measures the proportion of true results, is used as measure of the performance of each of the models. Accuracy values of >85% for correctly classifying diabetics from non-diabetics, and of >90% for correctly classifying hypertensive from non-hypertensive population are possible with the classification models built using the SVM and k-NN. The developed k-NN classification models are adapted to build risk assessment tools that output 'low' risk, 'borderline' risk and 'high' risk. Up to 75% of diabetics are being grouped into 'high' risk, and as few as 5% of non-diabetic patients are grouped into 'high' risk category. With the Asian ethnicity-specific tool, it is even better with 88.4% of the diabetic patients grouped as 'high' risk. Up to 94% of the hypertensive patients are grouped into 'high' risk by the ethnicity-independent tools; with the Asian ethnicity-specific tool, it is even better with 97% of hypertensive patients being grouped as 'high' risk.

Different pathology situations are modelled, namely diabetes in the general population (irrespective of the diagnosis for hypertension), diabetes in the hypertensive population, hypertension in the general population (irrespective of the diagnosis for diabetes) and hypertension in the diabetic population. Two-stage aggregate classification models, built combining both the models on diabetes or both the models on hypertension, perform far better than the individual models.

Ethnicity-specific models and risk assessment tools are built using either Kuwaiti natives or Asian expatriates; the models that are specific to Asian expatriates are doing better than those specific to Kuwaiti natives. An examination of the performance of the ADA online risk assessment tool on data from Kuwait (natives and Asian expatriates) indicates that the ADA tool performs almost in a random manner in distinguishing diabetics from non-diabetics in Kuwait. This implies that it is important to build 'local' or 'regional' assessment tools using local data.

LR models for diabetes identify hypertension diagnosis and family history of hypertension as significant predictors; in a similar fashion, the models for hypertension pick diabetes diagnosis and family history of diabetes as significant predictors. This is in agreement with the notion that disposition to diabetes increases the proneness to hypertension and *vice versa*.

Implications of using the developed prediction models in medical practice

In this paper, we show that predictive models built using basic non-intrusive data are able to identify patients at high risk for diabetes and hypertension. This becomes useful when applied in a public health setting. It would be advantageous to use the tool as a preliminary step to identify patients at high risk and to direct them for treatment (and research) purposes. These models can also be made available online, where concerned individuals can check their risk at home by answering simple questions such as their ethnicity, BMI and family history of diabetes. Those with higher risk can be advised to contact a medical professional, while lower risk patients can be advised of simple lifestyle changes. Up to 20–24% of Kuwaiti non-diabetic patients are identified as 'borderline' risk with our model. Without publicly available risk assessment tools, these patients would go unnoticed. In the future, should more robust biochemical data be available, more advanced models can be built as a second step in our study. Those identified as high risk from the basic models could be invited to enter biomarker values for a more detailed assessment.

Comparisons with other studies

Most of the available classification models and risk assessment tools for diabetes are based on LR.[7] The presented study reports on the applicability of machine-learning approaches. Models based on SVMs and k-NN give consistently high classification accuracies.

Prognostic measures (in terms of calibration and discrimination) help to evaluate validity of predictive models and to compare different published models. Discrimination describes the ability of the prediction model to distinguish patients at high risk of developing diabetes from those at low risk. We use the C-statistic to

measure discrimination, and since continuous outputs are required to plot the ROC, we show discrimination values for LR and SVM only. On the other hand, calibration measures the ability of the model to correctly estimate the absolute risks,[7] and we calculate it using the Hosmer-Lemeshow goodness of fit statistic[25] for the LR (since calibration calculations require the output to be a probability). The discrimination C-statistic for the LR and SVM models (that we developed for diabetes in general population) are seen as 0.820 and 0.831, respectively. These values are in good comparison with those reported for similar published models (using basic non-intrusive parameters similar to the ones used by models presented in this study) that range from 0.74 to 0.84.[7] The calibration p value for the presented LR model for diabetes in general population is evaluated as 0.135. A calibration p value of >0.05 means that the model is well calibrated, and a smaller value implies a poorly calibrated model.

Strengths and limitations of the study

The major strengths of this study are as follows: (1) for the first time in Kuwait, large amounts of health and medical data are available for research. Because of this, we have plenty of data to model the disorders of diabetes, hypertension and comorbidity. This translates into robust classification models and risk assessment tools that have little uncertainty. (2) Most of the classification models and risk assessment tools for diabetes are based on LR.[7] The presented study reports on the applicability of machine-learning approaches. Models based on SVMs and k-NN give consistently high classification accuracies.

The limitations of the study are as mentioned earlier under Data section. We further add that we considered only those patients with complete data for the predictors used in the models; it is possible that patients with missing data have different risk profiles as compared with patients included. However, the missing data are most often due to the reason that the integration of data by Kuwait Health Network is partial and ongoing.

CONCLUSIONS

Three main conclusions emerge from this study. First, using basic non-invasive parameters that are not laboratory-based, we are able to successfully predict, to a high degree of accuracy, the onset of diabetes and hypertension in patients in Kuwait, similar to what has been seen in other studies.[7] Second, we are able to model the increased proneness in diabetic patients to develop hypertension and vice versa. Aggregate models that combine individual ones on generalised population and on comorbid population enhance dramatically the predictive power. Third, in accordance with the literature, ethnicity plays a major role in determining diabetes and hypertension risk.[26 27 28] While developing classification models for patients in Kuwait, removing the ethnicity field from the data causes a drop of at least 6% in accuracy. This shows that the machine-learning techniques place a heavy weight on the ethnicity, as we would expect to see. Further supporting the claim on the need to train models with local data are results from evaluating the performance of the ADA's online diabetes risk test tool with data from Kuwait. Since the latter is built using patients in the USA, which naturally has a different ethnic demography to Kuwait, we see a large discrepancy in the results.

Acknowledgements The authors thank the International Scientific Advisory Board and the Ethics Committee at Dasman Diabetes Institute for approving the study and for discussions at the review meetings. The authors further thank Management of the institute for granting us access to the KHN data. The authors thank members of Kuwait-Scotland eHealth Innovation Network for useful discussions. Aridhia Informatics Ltd, Scotland is acknowledged for carving out research data export from their Informatics Layer to Kuwait Health Network for our use, and for many discussions on data quality, content and format. The IT department at the institute is acknowledged for its support to facilitate data sharing.

Contributors TAT undertook the study design, directed the reported work and directed the development of the manuscript. KB is responsible for setting up the Kuwait Health Network and access to the data; and is responsible for the research activities at the institute. BF performed the entire machine-learning algorithms and calculations as well as contributed to the manuscript. AMC handled data extraction, created the different data sets and performed the calculations for the data on the ADA online risk assessment tool. All authors have read and approved the final manuscript.

Funding This research received no specific grant from any funding agency in the public, commercial or non-profit sectors.

Competing interests None.

Ethics approval The study has been approved by the Ethics Committee at Dasman Diabetes Institute.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data are available.

REFERENCES

1. Morrish NJ, Wang SL, Stevens LK, *et al*. Mortality and causes of death in the WHO multinational study of vascular disease in diabetes. *Diabetolgia* 2001;44:s14–21.
2. Wild S, Roglic G, Green A, *et al*. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes Care* 2004;27:1047–53.
3. Badran M, Laher I. Type II diabetes mellitus in Arabic-speaking countries. *Int J Endocrinol* 2012;2012:e902873.
4. Alhyas L, McKay A, Majeed A. Prevalence of type 2 diabetes in the States of the co-operation council for the Arab States of the Gulf: a systematic review. *PLoS ONE* 2012;7:e40948.
5. Channanath AM, Farran B, Behbehani K, *et al*. State of diabetes mellitus, hypertension, and comorbidity in Kuwait—showcasing the trends as seen in native versus expatriate population. *Diabetes Care* 2013. In press.
6. Oldroyd J, Banerjee M, Heald A, *et al*. Diabetes and ethnic minorities. *Postgrad Med J* 2005;81:958:486–90.
7. Abbasi A, Peelen LM, Corpeleijn E, *et al*. Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study. *BMJ* 2012;345:e5900.
8. Grandinetti L, Pisacane O. Web based prediction for diabetes treatment. *Future Gen Comput Syst* 2011;27:139–47.
9. Parker RS, Doyle FJ III, Peppas NA. A model-based algorithm for blood glucose control in type I diabetic patients. *IEEE Trans Biomed Eng* 1999;46:148–57.
10. Lehmann ED, Deutsch T. A physiological model of glucose–insulin interaction in type 1 diabetes mellitus. *J Biomed Eng* 1992;14:235–42.
11. Nomura M, Shichiri M, Kawamori R, *et al*. A mathematical insulin-secretion model and its validation in isolated rat pancreatic islets perfusion. *Comput Biomed Res* 1984;17:570–9.
12. Geisser S. *Predictive inference*. New York, NY: Chapman and Hall, 1993.

13. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Mellish CS, ed. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 1995:1137–43.
14. Bhandari M, Joensson A. *Clinical research for surgeons*. Germany: Thieme Publishing Group, 2009.
15. Fix E, Hodges JL. Discriminatory analysis, nonparametric discrimination: Consistency properties. USAF School of Aviation Medicine, Randolph Field, Texas 1951, Report No: 4.
16. Cover TM, Hart PE. Nearest neighbor pattern classification. *IEEE Trans Inform Theory* 1967;13:21–7.
17. Cortes C, Vapnik VN. Support-vector networks. *Mach Learn* 1995;20:273–97.
18. Ritchie MD, Hahn LW, Roodi N, *et al*. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 2001;69:138–47.
19. Andrew AS, Karagas MR, Nelson HH, *et al*. DNA repair polymorphisms modify bladder cancer risk: a multi-factor analytic strategy. *Hum Hered* 2008;65:105–18.
20. Akagawa H, Narita A, Yamada H, *et al*. Systematic screening of lysyl oxidase-like (LOXL) family genes demonstrates that LOXL2 is a susceptibility gene to intracranial aneurysms. *Hum Genet* 2007;121:377–87.
21. Cho YM, Ritchie MD, Moore JH, *et al*. Multifactor-dimensionality reduction shows a two-locus interaction associated with type 2 diabetes mellitus. *Diabetologia* 2004;47:549–54.
22. Lago RM, Singh PP, Nesto RW. Diabetes and hypertension. *Nat Clin Pract Endocrinol Metabol* 2007;3:667.
23. Simonson DC. Etiology and prevalence of hypertension in diabetic patients. *Diabetes Care* 1988;11:821–7.
24. Bang H, Edwards AM, Bomback AS, *et al*. A self-assessment diabetes score: development, validation, and comparison with other diabetes risk-assessment scores. *Ann Intern Med* 2009; 151:775–83.
25. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15:361–87.
26. Abate N, Chandalia M. The impact of ethnicity on type 2 diabetes. *J Diabetes Complications* 2003;17:39–58.
27. Abate N, Chandalia M. Ethnicity and type 2 diabetes: focus on Asian Indians. *J Diabetes Complications* 2001;15:320–7.
28. Sukala WR, Page RA, Rowlands DS, *et al*. Exercise intervention in New Zealand Polynesian peoples with type 2 diabetes: Cultural considerations and clinical trial recommendations. *Australas Med J* 2012;5:429–35.

Supplemental Table S1: Illustration of adapting 7-NN classification model to perform risk assessment[@].

Risk assignment	Split (number of training points closest to the test point)	Count of diabetic patients assigned the risk specified in column 1	Count of nondiabetic participants assigned the risk specified in column 1
Split Schema 1			
"Low" risk	(0 to 1)	353	5501
"Borderline" risk	(2 to 3)	811	1558
"High" risk	(4 to 7)	1689	720
Total participants		2853	7779
Split Schema 2			
"Low" risk	(0 to 2)	718	6500
"Borderline" risk	(2 to 3)	1064	991
"High" risk	(4 to 7)	1071	288
Total participants		2853	7779

[@]**,**Two exemplary 'split' schemes are presented. It is expected that a high fraction of diabetic patients are assigned "high" risk and a low fraction are assigned "low" risk. Similarly, it is expected to see that a high fraction of nondiabetic participants are assigned "low" risk and that a low fraction are assigned "high" risk. It is clear that with split schema 2, more diabetics are assigned low and borderline classes, while more non-diabetics are being assigned to lower risk classes. Since it is not advisable that diabetic patients go undetected, split schema 1 is more (medically) acceptable, and it is the one that is used in the reported study. Similar experiments suggest using the same split for hypertension risk assessment as well.