

Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes – ELSA-Brasil: accuracy study

Comparação de algoritmos de aprendizagem de máquina para construir um modelo preditivo para detecção de diabetes não diagnosticada – ELSA-Brasil: estudo de acurácia

André Rodrigues Olivera^I, Valter Roesler^{II}, Cirano Iochpe^{III}, Maria Inês Schmidt^{III}, Álvaro Vigo^{IV}, Sandhi Maria Barreto^V, Bruce Bartholow Duncan^{III}

Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre (RS), Brazil

^IMSc. IT Analyst, Postgraduate Computing Program, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre (RS), Brazil.

^{II}PhD. Professor, Postgraduate Computing Program, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre (RS), Brazil.

^{III}PhD. Professor, Postgraduate Epidemiology Program and Hospital de Clínicas, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre (RS), Brazil.

^{IV}PhD. Professor, Postgraduate Epidemiology Program, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre (RS), Brazil.

^VPhD. Professor, Department of Social and Preventive Medicine & Postgraduate Program in Public Health, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte (MG), Brazil.

KEY WORDS:

Supervised machine learning.
Decision support techniques.
Data mining.
Models, statistical.
Diabetes mellitus, type 2.

PALAVRAS-CHAVE:

Aprendizado de máquina supervisionado.
Técnicas de apoio para a decisão.
Mineração de dados.
Modelos estatísticos.
Diabetes mellitus tipo 2.

ABSTRACT

CONTEXT AND OBJECTIVE: Type 2 diabetes is a chronic disease associated with a wide range of serious health complications that have a major impact on overall health. The aims here were to develop and validate predictive models for detecting undiagnosed diabetes using data from the Longitudinal Study of Adult Health (ELSA-Brasil) and to compare the performance of different machine-learning algorithms in this task.

DESIGN AND SETTING: Comparison of machine-learning algorithms to develop predictive models using data from ELSA-Brasil.

METHODS: After selecting a subset of 27 candidate variables from the literature, models were built and validated in four sequential steps: (i) parameter tuning with tenfold cross-validation, repeated three times; (ii) automatic variable selection using forward selection, a wrapper strategy with four different machine-learning algorithms and tenfold cross-validation (repeated three times), to evaluate each subset of variables; (iii) error estimation of model parameters with tenfold cross-validation, repeated ten times; and (iv) generalization testing on an independent dataset. The models were created with the following machine-learning algorithms: logistic regression, artificial neural network, naïve Bayes, K-nearest neighbor and random forest.

RESULTS: The best models were created using artificial neural networks and logistic regression. These achieved mean areas under the curve of, respectively, 75.24% and 74.98% in the error estimation step and 74.17% and 74.41% in the generalization testing step.

CONCLUSION: Most of the predictive models produced similar results, and demonstrated the feasibility of identifying individuals with highest probability of having undiagnosed diabetes, through easily-obtained clinical data.

RESUMO

CONTEXTO E OBJETIVO: Diabetes tipo 2 é uma doença crônica associada a graves complicações de saúde, causando grande impacto na saúde global. O objetivo foi desenvolver e validar modelos preditivos para detectar diabetes não diagnosticada utilizando dados do Estudo Longitudinal de Saúde do Adulto (ELSA-Brasil) e comparar o desempenho de diferentes algoritmos de aprendizagem de máquina.

TIPO DE ESTUDO E LOCAL: Comparação de algoritmos de aprendizagem de máquina para o desenvolvimento de modelos preditivos utilizando dados do ELSA-Brasil.

MÉTODOS: Após selecionar 27 variáveis candidatas a partir da literatura, modelos foram construídos e validados em 4 etapas sequenciais: (i) afinação de parâmetros com validação cruzada (*10-fold cross-validation*); (ii) seleção automática de variáveis utilizando seleção progressiva, estratégia “*wrapper*” com quatro algoritmos de aprendizagem de máquina distintos e validação cruzada para avaliar cada subconjunto de variáveis; (iii) estimação de erros dos parâmetros dos modelos com validação cruzada; e (iv) teste de generalização em um conjunto de dados independente. Os modelos foram criados com os seguintes algoritmos de aprendizagem de máquina: regressão logística, redes neurais artificiais, *naïve* Bayes, K vizinhos mais próximos e floresta aleatória.

RESULTADOS: Os melhores modelos foram criados utilizando redes neurais artificiais e regressão logística alcançando, respectivamente, 75,24% e 74,98% de média de área sob a curva na etapa de estimação de erros e 74,17% e 74,41% na etapa de teste de generalização.

CONCLUSÃO: A maioria dos modelos preditivos produziu resultados semelhantes e demonstrou a viabilidade de identificar aqueles com maior probabilidade de ter diabetes não diagnosticada com dados clínicos facilmente obtidos.

INTRODUCTION

Type 2 diabetes is a chronic disease characterized by the body's inability to efficiently metabolize glucose, which increases blood glucose levels and leads to hyperglycemia. This condition is associated with a wide range of serious health complications affecting the renal, neurological, cardiac and vascular systems, and it has a major impact on overall health and healthcare costs.¹

Recent studies have estimated that around 415 million people have diabetes, and that the number of cases may increase to 642 million by 2040. In addition, approximately half of these individuals are not aware of their condition, which may further intensify the negative consequences of the disease. Diabetes was the main cause of death of nearly five million people in 2015, and it has been estimated that by 2030 it will become the seventh largest cause of death worldwide.²⁻⁴

It is believed that diabetes, like other noncommunicable chronic diseases, is mainly caused by behavioral factors such as poor diet and physical inactivity. Early interventions aimed towards creating lifestyle changes, with or without associated pharmacological therapies, have been proven effective in delaying or preventing type 2 diabetes and its complications. This has led many countries to invest in national programs to prevent this disease. To reduce costs and amplify the results, population-level interventions need to be combined with interventions that are directed towards individuals who are at high risk of developing or already having diabetes,⁵ so as to focus interventions, at the individual patient level, on those for whom they are most appropriate.

To this end, over recent years, a series of clinical prediction rules have been developed to identify individuals with unknown diabetes or those at high risk of developing diabetes.⁵⁻⁹ However, few of these rules have been drawn up using the most recently developed machine-learning techniques, which potentially have the ability to produce algorithms of greater predictive ability than those developed through the technique most commonly used to date, i.e. multiple logistic regression.

OBJECTIVE

This paper presents the development and comparison of predictive models created from different machine-learning techniques with the aim of detecting undiagnosed type 2 diabetes, using baseline data from the Longitudinal Study of Adult Health (ELSA-Brasil).

METHODS

These analyses were performed on data from the baseline survey (2008-2010) of ELSA-Brasil, a multicenter cohort study that had the main aim of investigating multiple factors relating to adult health conditions, including diabetes and cardiovascular diseases. The study enrolled 15,105 public servants aged between

35 and 74, at six public higher-education and research institutions in different regions of Brazil between 2008 and 2010, as has been previously reported in greater detail.^{10,11} The institutional review boards of the six institutions at which the study was conducted gave their approval, and written informed consent was obtained from all participants.

All analyses were performed using R version 3.2.3. The source codes used in the analysis are freely available.

Dataset and preliminary variable selection

Data from the ELSA study baseline were used to create the predictive models. At this baseline, the 15,105 participants were evaluated through interviews, clinical examinations and laboratory tests. The interviews addressed educational achievement; characteristics and composition of home and family; dietary habits; alcohol drinking habits; smoking habits; presence of dyslipidemia or hypertension; physical activity at leisure; sleep quality; medical history; and medication use, among other topics. The examinations involved anthropometric measurements and blood and urine tests, among others. The study generated more than 1500 variables for each participant at baseline, as described previously.^{10,11}

A total of 1,473 participants were excluded from the present analyses because they had self-reported diabetes. Another three participants were excluded because some information required for characterizing undiagnosed diabetes was missing. An additional 1,182 participants (8.7%) were excluded from the analyses because data relating to other variables were missing. Among the remaining 12,447 participants, 1,359 (11.0%) had undiagnosed diabetes.

Undiagnosed diabetes was considered present when, in the absence of a self-report of diabetes or use of anti-diabetic medication, participants had fasting glucose levels ≥ 126 mg/dl, glucose levels ≥ 200 mg/dl two hours after a standard 75 g glucose load or had glycated hemoglobin (HbA1c) $\geq 6.5\%$.

Through procuring variables in the ELSA dataset that were similar to those investigated in previously published predictive models for detecting diabetes or in situations of high risk of developing diabetes, we selected 27 diabetes risk factors for analysis. Any variables that implied additional costs beyond those of filling out a questionnaire and performing basic anthropometry, such as clinical or laboratory tests, were excluded so that the model obtained could be applied using a straightforward survey and simple anthropometric measurements. The final variable subset was validated by experts, and this resulted in the subset of 27 candidate variables described in **Table 1** and **Table 2**. **Table 2** also presents the analysis target variable of prevalent diabetes "a_dm".

The original dataset was randomly divided into two parts in the ratio of 70:30. The first part (training/validation dataset) was used for parameter and cutoff tuning, automatic variable selection and error estimation using cross-validation; the second part (test

Table 1. Numerical variables

Variable identity	Minimum	Median	Mean	Maximum	SD	Variable description
a_cons_est_nacl	0.14	9.67	11.65	3533.00	42.57	Estimated daily salt consumption in grams
a_rcq	0.40	0.89	0.89	1.27	0.09	Waist-hip ratio
a_rendapercapita	27.63	1410.90	1756.34	7884.50	1436.67	Per capita family net income in R\$
afia7	0.00	0.00	0.12	7.00	0.68	Bicycle use for transport (days/week)

SD = standard deviation.

Table 2. Categorical variables, including the target variable “a_dm”

Variable identity	Number of levels	Frequency in each level	Description	Possible values
a_ativfisica	3	1: 9523; 2: 1735; 3: 1189	Physical activity during leisure time	1 = weak; 2 = moderate; 3 = strong
a_binge	2	0: 10764; 1: 1683	Sporadic excessive alcohol drinker	0 = no; 1 = yes
a_consdiafrutas	2	0: 5434; 1: 7013	Daily consumption of fruits	0 = no; 1 = yes
a_consdiaverduras	2	0: 6003; 1: 6444	Daily consumption of vegetables	0 = no; 1 = yes
a_dm	2	1: 11088; 0: 1359	Diabetes mellitus	0 = yes; 1 = no
a_escolar	4	1: 619; 2: 773; 3: 4219; 4: 6836	Education	1 = middle school not completed or less; 2 = middle school completed; 3 = high school completed; 4 = university undergraduate course completed
a_fumante	3	0: 7212; 1: 3619; 2: 1616	Smoker	0 = never smoked; 1 = former smoker; 2 = smoker
a_gidade	4	1: 2899; 2: 5077; 3: 3320; 4: 1151	Age group	1 = 35 to 44 years; 2 = 45 to 54 years; 3 = 55 to 64 years; 4 = 65 to 74 years
a_imc2	4	1: 122; 2: 4705; 3: 5011; 4: 2609	Four-level body mass index	1 = underweight; 2 = eutrophic; 3 = overweight; 4 = obese
a_medanthipert	2	0: 9232; 1: 3215	Use of antihypertensive drugs	0 = no; 1 = yes
a_medoutahip	2	0: 12367; 1: 80	Use of other antihypertensive drugs	0 = no; 1 = yes
a_medredlip	4	0: 11122; 1: 1117; 2: 97; 3: 111	Use of lipid-lowering drugs	0 = no use; 1 = use of statins; 2 = use of others; 3 = use of more than one class
a_sfhfprem	2	0: 12371; 1: 76	Self-reported heart failure (< 50 years of age)	0 = no; 1 = yes
a_sfmiprem	2	0: 12386; 1: 61	Self-reported myocardial infarction (< 50 years of age)	0 = no; 1 = yes
a_sfrevprem	2	0: 12402; 1: 45	Self-reported revascularization (< 50 years of age)	0 = no; 1 = yes
a_sfstkprem	2	0: 12373; 1: 74	Self-reported stroke (< 50 years of age)	0 = no; 1 = yes
a_sintsono	2	0: 8321; 1: 4126	Sleep quality	0 = no; 1 = yes
a_sitconj	5	1: 8248; 2: 2028; 3: 1283; 4: 474; 5: 414	Marital status	1 = married; 2 = divorced; 3 = single; 4 = widowed; 5 = other
claa2	2	0: 8397; 1: 4050	Pain/discomfort in the legs while walking (Q2)	0 = no; 1 = yes
diea133	3	0: 1038; 1: 11136; 2: 273	Coffee consumption (Q133)	0 = no; 1 = yes, with caffeine; 2 = yes, decaffeinated
hfda07	2	0: 3271; 1: 9176	Hypertension, family history (Q7)	0 = no; 1 = yes
hfda11	2	0: 7879; 1: 4568	Diabetes, family history (Q11)	0 = no; 1 = yes
hmpa08	2	0: 8205; 1: 4242	High cholesterol (Q8)	0 = no; 1 = yes
rcta8	2	1: 5566; 2: 6881	Sex	1 = male; 2 = female

dataset) was used for generalization tests. The models created were evaluated in terms of area under the receiver operating characteristic curve (AUC), sensitivity, specificity and balanced accuracy (arithmetic mean of sensitivity and specificity). The machine-learning algorithm families of logistic regression, artificial neural network (multilayer perceptron/backpropagation), Bayesian network (naïve Bayes classifier), instance-based learning (*K*-nearest neighbor) and ensemble (random forest) were used.

Machine-learning algorithms

The machine-learning algorithms are briefly described below:

*Logistic regression*¹² is a well-established classification technique that is widely used in epidemiological studies. It is generally used as a reference, in comparison with other techniques for analyzing medical data.

*Multilayer perceptron/backpropagation*¹³ is the principal artificial neural network algorithm. When there is no hidden layer on the network, this algorithm is equivalent to logistic regression, but it can solve more difficult problems with more complex network architectures. The price of using complex architectures is that it produces models that are more difficult to interpret. Additionally, it can be computationally expensive.

*Naïve Bayes classifier*¹⁴ is a type of Bayesian network that, despite enormous simplicity, is able to create models with high predictive power. The algorithm works well with heterogeneous data types and also with missing values, because of the independent treatment of each predictor variable for model construction.

*K-nearest neighbor (instance-based learning)*¹⁵ is a classical instance-based learning algorithm in which a new case is classified based on the known class of the nearest neighbor, by means of a majority vote. This type of algorithm is also called lazy learning because there is no model building step and the entire computing procedure (i.e. the search for the nearest neighbor) is performed directly during the prediction. All the cases (training/validation dataset) need to be available during the prediction.

*Random forest*¹⁶ is a machine-learning algorithm from the “ensemble” family of algorithms,¹⁷ which creates multiple models (called weak learners) and combines them to make a decision, in order to increase the prediction accuracy. The main idea of this technique is to build a “forest” of random decision “trees” and use them to classify a new case. Each tree is generated using a random variable subset from the candidate’s predictor variables and a random subset of data, generated by means of bootstrap. This algorithm also can be used to estimate variable relevance.

Data preparation

Standardization of numerical variables

Transformation between different data types (categorical or numerical) was performed by means of binarization or discretization,

when necessary. In binarization, a categorical variable with n levels is transformed into $n - 1$ dummy variables that have values equal to “1” when the case belongs to the level represented by the dummy variable or “0” otherwise.

In discretization, a numerical variable is transformed into a categorical variable by defining a set of cutoff points for that variable, such that the ranges of values between the cutoff points correspond to the levels of the categorical variable. The Ameva algorithm¹⁸ was used to find the best cutoff points for each numerical variable.

General process of model construction and evaluation

The models were built, evaluated and compared using four sequential steps:

1. Parameter tuning;
2. Automatic variable selection;
3. Error estimation; and
4. Generalization testing in an independent dataset.

The complete process is depicted in **Figure 1**. First, manual variable preselection was performed as described above (“Manual Variable Selection” box in the **Figure**). After that, 30% of the dataset (“Test” dataset in the **Figure**), containing 3,709 complete cases, was separated for generalization testing, while the other part (“Training/Validation” dataset in the **Figure**), containing 8,738 complete cases, was used as the dataset for the first three steps of the process.

The first step in model building (“Parameter Tuning” box in **Figure 1**) evaluated each machine-learning algorithm with different sets of configurable parameters of the algorithm by means of tenfold cross-validation, repeated three times. In tenfold cross-validation, the dataset (training/validation) is divided into ten parts, of which nine are used for training (selecting) a model and the tenth for validation of this model. This process is repeated to calculate the validation measurements, such as AUC, while varying the part of the dataset used for validation each time. Finally, the mean of the validation measurements across repeats is calculated. The results from this step (“Best Parameters” item in the **Figure**), containing the best parameters and cutoffs for classification for each algorithm, were used in the next steps.

The second step (“Automatic Variable Selection” box in **Figure 1**) generated four different variable subsets using different algorithms and cross-validation (using only the best settings found in the preceding step), with the wrapper strategy and a forward selection search for automatic variable selection. The best variable subsets found in this step (“Best Variable Subsets” item in **Figure 1**) were used in the next steps.

The third step (“Error Estimation” box in **Figure 1**) used cross-validation to obtain more reliable estimates of the performance

of different learning schemes, using the best settings and subsets obtained in the preceding steps.

Finally, the last step (“Generalization Testing” box in Figure 1) evaluated models using only the learning scheme that obtained the best performance for each algorithm in the test dataset that had not previously been used.

The following sections describe each step in more details.

Parameter tuning

This first step in model building evaluated each algorithm with different parameter configurations to find out which parameter configuration produced the best results for each algorithm and data type conversion used. The parameters tested for each algorithm are listed in Table 3.

Because of the wide range of possible values for the parameters, a search strategy was adopted. At first, a limited set of values for

each parameter was selected, and each combination of parameters was evaluated by means of tenfold cross-validation, repeated three times, thus generating 30 models. Each instance of machine learning was tested with and without data discretization. The results from

Table 3. Parameters analyzed in parameter tuning

Algorithm	Parameter	Description
Artificial neural network	Size	Number of neurons on hidden layer
	Decay	Weight decay
	Skip	Direct link between input and output neurons
Logistic regression	Epsilon	Convergence tolerance value
Naïve Bayes	Laplace	Real number to control Laplace smoothing
K-nearest neighbor	Minvotes	Minimum votes to define a decision
	k	Number of neighbors considered
Random forest	Ntree	Number of random trees generated

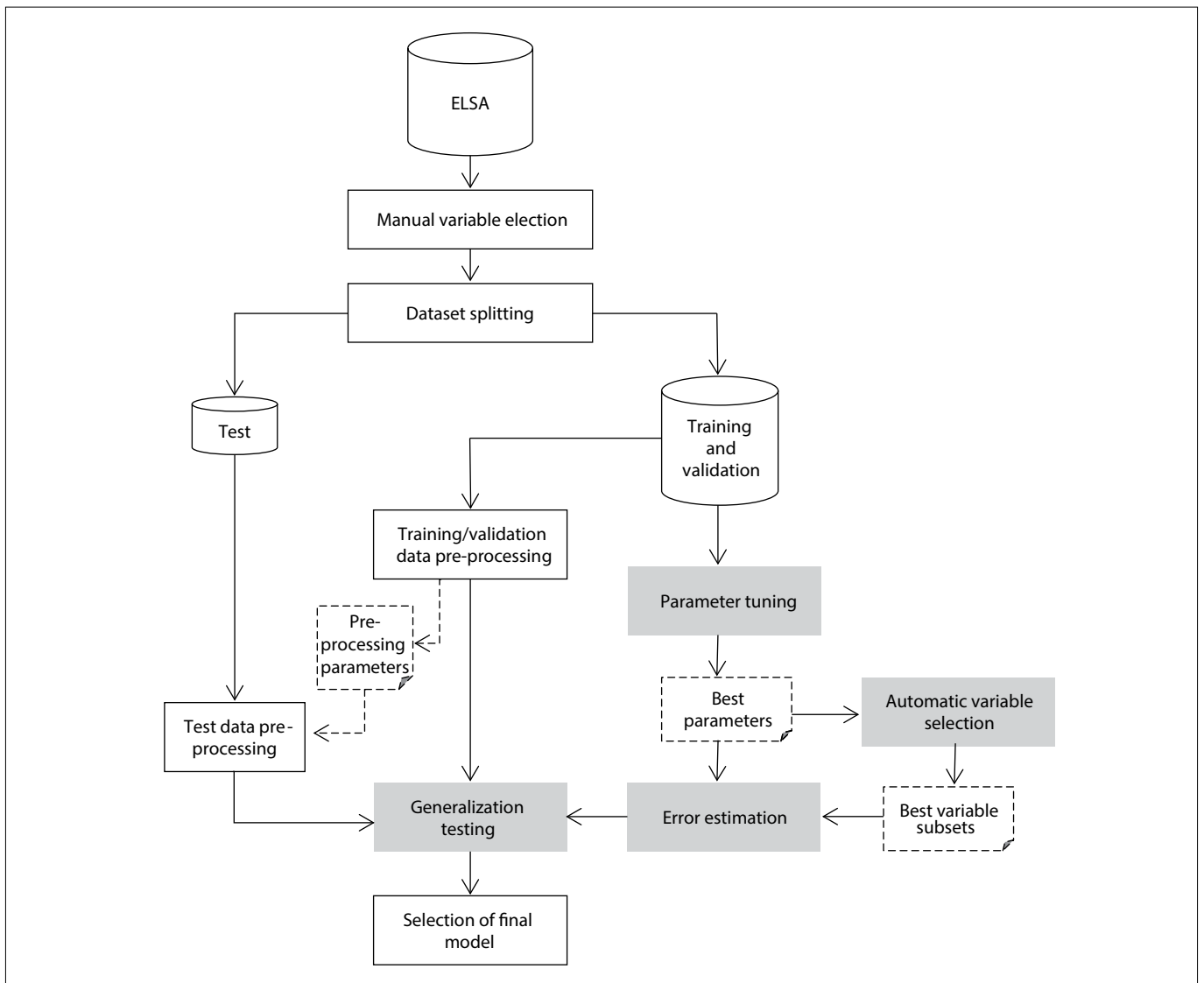


Figure 1. General process of model construction and evaluation.

the 30 models generated in each test were averaged. The parameter configuration that produced the best mean AUC was chosen. Moreover, a set of different cutoffs (predefined by the analyst) to generate the classification was evaluated to find out which produced the best classification on average between the 30 models in terms of balanced accuracy.

After that, the results were analyzed and, when necessary, new parameter values and/or cutoff points were selected for new tests. In this case, the new values were selected around the values from which the best results had been obtained up to that moment. The idea was to start testing a sparse range of values and then decrease the granularity of the values in order to avoid trying values that were very likely to produce poor results. This search stopped when there was no increase in the predictive power of the models that had been created using the specific machine-learning algorithm and data type conversion evaluated.

Automatic variable selection

The automatic variable selection step had the aim of finding subsets from the 27 candidate variables that could increase the performance of the predictive models, compared with other models created using different sets of candidate variables.

These subsets of variables were generated using the wrapper strategy.¹⁹ In this strategy, models are created and evaluated by means of a machine-learning algorithm and a validation method, such as cross-validation, using different subsets of variables. The subset from which the best performance is achieved, in terms of a criterion such as AUC, is chosen as the best subset. Because of the large number of possible subsets, a heuristic search was used to generate the variable subset candidates that were more likely to create better models, thereby optimizing the process. The main advantage of this method compared with other strategies is that it evaluates multiple variables together. The drawback is that, because it depends on a machine-learning algorithm to create/evaluate models, it is possible that the subset of variables that produces the best results using one algorithm can produce bad results when using another algorithm or another parameter setting for the same algorithm.

Four machine-learning algorithms were used: logistic regression, artificial neural network, K-nearest neighbor and naïve Bayes classifier. The random forest algorithm was not included because it already performs an embedded variable selection. The forward selection search strategy was used in modeling because it tends to generate smaller subsets. The same validation technique (tenfold cross-validation repeated three times), decision criterion (mean AUC) and dataset partition that had been used in the parameter tuning step were used again in this step. The best parameter settings obtained in the parameter tuning step were used to configure the parameters of the

algorithms for this step. Each machine-learning technique generated a distinct subset of variables. The subsets thus generated were used in the next step.

Error estimation

The error estimation step evaluated each machine-learning algorithm using the parameters obtained in the first step and the subsets generated in the second step, in addition to the original variable subset containing all the candidate variables. This step also served to evaluate the use of discretization. The evaluation was done through tenfold cross-validation, which was repeated ten times to get more reliable prediction performance estimates.

Generalization testing

Finally, one model was generated from the training/validation dataset for each algorithm, using the best results from the preceding step. These best models were then evaluated (hold-out evaluation) in the test set, since this generalization testing has the aim of evaluating model behavior when faced with data that was not used in its creation. The results from this evaluation serve as a quality measurement for these models.

Development of an equation for application of the results

The model with best results from generalization testing was used to create a web tool to apply the questionnaire in practice. The prediction from the logistic regression model for any given individual is calculated by multiplying that individual's value for each variable in the model by the coefficient derived from the model for that variable, and then summing the results and transforming this sum into a probability of undiagnosed diabetes using the logistic function. If this probability is above the predetermined cutoff (here, 11%), the individual is classified as positive (at high risk of undiagnosed diabetes); or otherwise, as negative.

RESULTS

Study sample

Among the 12,447 ELSA participants included in this study, 5,566 (44.67%) were men. The participants were between 35 and 74 years old; the largest proportion (5,077) was in the group between 45 and 54 years old; 6,836 (54.92%) had a complete university education or more; 5,011 (40.26%) were overweight and 2,609 (20.96%) were obese. Using the World Health Organization definition (fasting glucose \geq 110 mg/dl and/or 2 hour post-load glucose \geq 140 mg/dl), 5,539 (44.5%) presented intermediate hyperglycemia. Other details about the study sample can be found in Table 1 and Table 2.

Parameter tuning

The best parameter configuration for each data type conversion of each algorithm is depicted in **Table 4**.

The first and second columns of **Table 4** present the name of the algorithm and whether discretization was used, respectively. The third column shows the values of the parameter configuration that provided the best result for the machine-learning algorithm and data type conversion of each row. The next four columns present basic statistics (mean, standard deviation, first and third quartiles and cutoff points, respectively) of the AUC obtained in the cross-validation. The eighth column shows the cutoff that provided the mean best balanced accuracy (BA) and the last two columns shows the mean balanced accuracy and its standard deviation.

Table 4 shows each machine-learning algorithm with its different data type conversions, sorted in descending order in terms of AUC and balanced accuracy for each algorithm and data type conversion.

Although defining which algorithms produce better results was not the objective of this step, it was possible to gain an initial insight into their predictive powers. In this regard, the best results were produced by artificial neural networks and logistic regression with mean AUC of 75.24% (row 1) and 74.98% (row 3), respectively.

Table 4 also shows the impact in terms of performance, when discretization was used in each machine-learning algorithm. For example, performance decreased (around 1% overall and almost 3% in the case of random forest) when the data were discretized in the models generated by all the algorithms except naïve Bayes. In general, the performance behavior of the machine-learning algorithms and conversion remained similar for the next steps.

Another result that can be seen in most cases is the impact on the choice of the parameter settings, caused by the conversion used. For example, the best performance of the artificial neural network algorithm was achieved without data conversion and with size = 175 (i.e. 175 neurons in the hidden layer). However,

when discretization was used, the best parameter setting changed to size = 100.

The best parameter setting achieved was used to configure the five algorithms used for the automatic variable selection step, as well as in further steps.

Results from automatic variable selection

The automatic variable selection step generated four distinct subsets of variables as shown in **Table 5** (rows 1 to 4): *lr-fs*, created with logistic regression (*fs* in the name stands for “forward selection”); *ann-fs*, created with an artificial neural network; *knn-fs*, created with K-nearest neighbor; and *nb-fs*, created with a naïve Bayes algorithm.

Table 5. Variable subsets generated in automatic variable selection

Subset	Best mean AUC	Number of variables	Variable names
ann-fs	75.48%	14	a_ativfisica, a_binge, a_escolar, a_gidade, a_imc2, a_medanthipert, a_medredlip, a_rcq, a_rendapercapita, a_sfhfprem, diea133, hfda07, hfda11, rcta8.
lr-fs	75.44%	11	a_ativfisica, a_binge, a_escolar, a_gidade, a_imc2, a_medanthipert, a_rcq, diea133, hfda07, hfda11, rcta8.
knn-fs	74.94%	12	a_binge, a_escolar, a_gidade, a_imc2, a_medanthipert, a_medoutahip, a_rcq, a_sfmiprem, a_sfstkpem, hfda07, hfda11, rcta8.
nb-fs	74.47%	10	a_ativfisica, a_binge, a_escolar, a_gidade, a_imc2, a_medanthipert, a_rcq, afia7, diea133, hfda11.

Table 4. Results from parameter tuning

Algorithm	Data conversion	Parameters	AUC (mean)	AUC (SD)	AUC (1q)	AUC (3q)	Cutoff	BA (mean)	BA (SD)
Artificial neural network	–	Size = 175; decay = 2; skip = false	75.24%	1.87%	73.91%	76.77%	0.12	69.04%	2.36%
Artificial neural network	Discretization	Size = 100; decay = 3; skip = true	74.16%	1.94%	72.71%	74.87%	0.11	67.95%	1.71%
Logistic regression	–	Epsilon = 0.01	74.98%	1.81%	73.83%	76.27%	0.11	68.46%	2.37%
Logistic regression	Discretization	Epsilon = 0.01	74.01%	1.98%	72.61%	74.98%	0.11	67.74%	1.98%
K-nearest neighbor	–	Neighbor = 475	74.45%	2.05%	72.96%	75.56%	0.1	68.59%	1.99%
K-nearest neighbor	Discretization	Neighbor = 275	73.60%	2.11%	72.31%	74.87%	0.09	67.55%	2.30%
Naïve Bayes	Discretization	Laplace = 0.001	73.67%	2.26%	72.21%	75.04%	0.09	68.09%	2.11%
Naïve Bayes	–	Laplace = 1	73.23%	2.58%	71.85%	74.52%	0.31	67.74%	2.57%
Random forest	–	Ntree = 7,000	72.90%	1.94%	71.77%	74.26%	0.13	67.24%	2.06%
Random forest	Discretization	Ntree = 4,300	70.85%	2.12%	69.41%	72.56%	0.12	65.55%	2.11%

AUC = area under the ROC curve; SD = standard deviation; 1q/3q = first and third quartiles; BA = balanced accuracy.

The first column of **Table 5** shows the identifier name of the subset, the second column presents the AUC achieved by the variable subset that was chosen for each algorithm, the third shows the number of variables of each subset and the fourth presents these variable names.

The dataset partitions used for this step were the same as used in the parameter tuning step. Thus, it is possible to gain an insight into the performance improvement in terms of AUC when using a variable subset instead of using all the variables from the original dataset. Furthermore, merely the fact that a smaller subset was used to create the models is already an advantage because this makes the model and its application much simpler.

Because of the nature of the wrapper strategy, it can be expected that each machine-learning algorithm will present better results when using the variable subset created by the algorithm itself. However, in the next step all the subsets were tested with all the algorithms.

Results from error estimation

The aim of this step was to obtain more reliable error estimates regarding algorithm performance. For this reason, 10 repetitions were used instead of 3, for the repeated tenfold cross-validation, thus generating 100 models instead of 30 for each test.

The machine-learning algorithms were tested using the best parameters found in the first step (depicted in **Table 4**), with the variable subsets generated in the second step (described in **Table 5**), as well as with the original set of variables. Performance was tested with and without discretization.

Table 6 describes the best results obtained for each machine-learning algorithm, variable subset and data conversion used. Respectively, the columns represent the name of machine-learning algorithm used; data type conversion; variable subset; AUC mean, standard deviation (SD) and first and third quartiles achieved in cross-validation; and mean and standard deviation of the balanced accuracy (BA).

Like in the results from the parameter tuning step, the artificial neural network algorithm and logistic regression achieved the best results. Without data conversion, these algorithms produced similar models, with AUC of 75.45% (row 1) and 75.44% (row 4), respectively, each using the variable subset generated with its own algorithm, as expected. K-nearest neighbor and naïve Bayes also reached good results, with AUC of close to 75%. The best results with the naïve Bayes classifier were obtained using a subset of variables other than *nb-fs*. This was possible because the variable subset search with this algorithm used discretized data following the best results from parameter tuning, while the best result in the current phase was without variable transformation.

Finally, as in the parameter tuning step, random forest produced the worst results. Independent of the subset of variables, this algorithm showed a worse yield in terms of mean AUC.

Table 6 also shows the impact of using a specific variable subset, compared with the best results obtained from the models generated using the original variable set. This difference is very small: around 0.25% better using the variable subset instead of all the original variables for the artificial neural network models. The results obtained with a subset of variables were slightly better (around 0.5%) than the original with logistic regression and K-nearest neighbor models. The best naïve Bayes classifier model result from using a variable subset was more than 1% better than the best result from using all the variables. Finally, random forest models produced the best results using all of the available variables.

Results from generalization testing

In generalization testing, the best learning scheme (which includes data type conversion used, parameter setting, classification cutoff and variable subset) found for each algorithm in the preceding step was evaluated in the test dataset, which had been separated at the beginning of the process and had not been used until this step.

Table 7 shows the best results obtained in the error estimation phase together with the results obtained in generalization testing.

All the algorithms maintained good performance in generalization testing. The biggest loss of performance in relation to the error estimate step, as assessed from changes in the AUC, was 1.64% for the K-nearest neighbor algorithm. The artificial neural network, logistic regression and naïve Bayes had performance losses of 1.30%, 1.03% and 0.80%, respectively. The least loss in generalization testing was 0.458%, achieved by the random forest algorithm, which produced the worst performance in terms of AUC of all the algorithms. Nevertheless, the worst result was an AUC of 72.35%.

Since the best result from this step in terms of AUC (74.41%) was obtained using logistic regression, and given the easy interpretation and applicability of this model, logistic regression was chosen to be used to create the diabetes risk assessment tool.

Web tool proposed for detecting undiagnosed diabetes

Finally, the model generated using the logistic regression algorithm in the generalization test was selected to build a web tool for detecting undiagnosed diabetes. This model produced sensitivity of 68% and specificity of 67.2%. The prototype interface of the tool is shown in **Figure 2**. Since the model was constructed and probably would be used in Brazil, the tool was created in Portuguese.

The final coefficients of the equation generated are described in **Table 8**.

New cases can be classified using this model, as follows:

1. Standardize the value of the only numerical variable (*a_rcq*) by subtracting the training mean (0.8889311) from the value

- and dividing the result by the training standard deviation (0.08615528).
2. Binarize the categorical variables;
 3. Calculate the sum of the variables created in the preceding steps using the coefficients from **Table 8**;
 4. Add to this sum the value of the intercept term, described in the first row of **Table 8**;
 5. Calculate the probability of undiagnosed diabetes for a given individual = $1/(1+e^{-x})$, where x equals the sum resulting from the preceding steps.
- If the probability is greater than 0.11, then classify the individual as presenting high risk of having undiagnosed diabetes; otherwise, classify the individual as presenting low risk.

Table 6. Error estimation results

Algorithm	Transformation	Variable subset	AUC (mean)	AUC (SD)	AUC (1q)	AUC (3q)	BA (mean)	BA (SD)
Artificial neural network	-	ann-fs	75.45%	1.96%	74.18%	76.96%	69.36%	2.17%
Artificial neural network	-	lr-fs	75.42%	1.99%	74.07%	77.04%	69.47%	2.14%
Artificial neural network	-	knn-fs	75.35%	1.98%	74.06%	76.85%	68.90%	2.09%
Artificial neural network	-	nb-fs	75.33%	2.05%	74.01%	76.95%	69.23%	2.30%
Artificial neural network	-	original	75.20%	1.96%	73.93%	76.79%	69.00%	2.20%
Logistic regression	-	lr-fs	75.44%	1.98%	74.00%	77.04%	69.30%	2.12%
Logistic regression	-	nb-fs	75.35%	2.02%	73.97%	76.93%	68.93%	2.07%
Logistic regression	-	ann-fs	75.35%	1.96%	74.09%	76.96%	68.91%	2.07%
Logistic regression	-	knn-fs	75.32%	1.95%	74.02%	77.00%	68.76%	2.10%
Logistic regression	-	original	74.94%	1.97%	73.58%	76.53%	68.41%	2.26%
K-nearest neighbor	-	knn-fs	74.98%	2.13%	73.54%	76.83%	68.52%	2.14%
K-nearest neighbor	-	ann-fs	74.80%	2.23%	73.51%	76.59%	68.74%	2.04%
K-nearest neighbor	-	lr-fs	74.77%	2.20%	73.22%	76.69%	68.63%	2.36%
K-nearest neighbor	-	nb-fs	74.68%	2.17%	73.15%	76.43%	68.64%	2.07%
K-nearest neighbor	-	original	74.44%	2.32%	72.99%	76.34%	68.52%	2.14%
Naïve Bayes	-	lr-fs	74.85%	2.20%	73.30%	76.56%	68.95%	2.17%
Naïve Bayes	-	ann-fs	74.71%	2.23%	73.23%	76.43%	68.79%	2.21%
Naïve Bayes	-	knn-fs	74.66%	2.19%	73.20%	76.39%	68.58%	2.14%
Naïve Bayes	Discretization	nb-fs	74.49%	2.12%	72.97%	76.11%	68.15%	2.06%
Naïve Bayes	Discretization	original	73.75%	2.35%	72.16%	75.53%	68.14%	2.15%
Random forest	-	original	72.81%	2.32%	71.61%	74.35%	67.06%	2.34%
Random forest	-	ann-fs	72.10%	2.24%	70.63%	73.79%	64.59%	2.33%
Random forest	-	knn-fs	71.75%	2.40%	70.05%	73.50%	59.72%	2.43%
Random forest	-	lr-fs	70.62%	2.53%	68.92%	72.33%	61.85%	2.56%
Random forest	-	nb-fs	70.42%	2.47%	68.69%	72.24%	61.19%	2.26%

AUC = area under the ROC curve; SD = standard deviation; 1q/3q = first and third quartiles; BA = balanced accuracy.

Table 7. Generalization testing results compared with those of the error estimation step

Algorithm	Error estimation		Generalization			
	AUC	BA	AUC	BA	Sensitivity	Specificity
Logistic regression	75.44%	69.30%	74.41%	67.62%	67.99%	67.24%
Artificial neural network	75.45%	69.36%	74.17%	67.78%	66.25%	69.30%
Naïve Bayes	74.85%	68.95%	74.06%	68.52%	74.94%	62.1%
K-nearest neighbor	74.98%	68.52%	73.34%	67.76%	70.97%	64.55%
Random forest	72.81%	67.06%	72.35%	67.50%	67.74%	67.24%

AUC = area under the ROC curve; BA = balanced accuracy.

1 Medidas antropométricas

IDADE

 anos

ALTURA

 m

PESO

 kg

SEXO

 Masculino Feminino

CIRCUNFERÊNCIA DO QUADRIL

 cm

CIRCUNFERÊNCIA DA CINTURA

 cm

2 Escolaridade

QUAL SEU NÍVEL DE ESCOLARIDADE?

 Fundamental incompleto
 Fundamental completo
 Superior incompleto
 Superior completo

3 Hábitos Comportamentais

JÁ CONSUMIU BEBIDAS ALCOÓLICAS?

 Sim
 Não

NOS ÚLTIMOS 12 MESES, COM QUE FREQUÊNCIA VOCÊ CONSUMIU 5 OU MAIS DOSES DE QUALQUER TIPO DE BEBIDA ALCOÓLICA EM UM PERÍODO DE 2 HORAS?

 Duas vezes por dia ou mais
 Praticamente todos os dias
 Uma a duas vezes por semana
 Duas ou três vezes por mês
 Somente em ocasiões especiais
 Nunca

ATUALMENTE CONSUME BEBIDAS ALCOÓLICAS?

 Sim
 Não

QUANTOS DIAS POR SEMANA VOCÊ FAZ CAMINHADAS NO SEU TEMPO LIVRE?

 dias

NOS DIAS QUE VOCÊ FAZ ESSAS CAMINHADAS, QUANTO TEMPO NO TOTAL ELAS DURAM POR DIA?

 min

QUANTOS DIAS POR SEMANA VOCÊ FAZ ATIVIDADES FÍSICAS FORTES, FORA AS CAMINHADAS, NO SEU TEMPO LIVRE?
Exemplo: correr, fazer ginástica de academia, pedalar em ritmo rápido, praticar esportes competitivos, etc.

 dias

NOS DIAS QUE VOCÊ FAZ ESSAS ATIVIDADES FÍSICAS FORTES, QUANTO TEMPO NO TOTAL ELAS DURAM POR DIA?

 min

QUANTAS VEZES POR SEMANA VOCÊ PRÁTICA ATIVIDADES FÍSICAS MÉDIAS, FORA AS CAMINHADAS, NO SEU TEMPO LIVRE?
Exemplo: nadar, ou pedalar em ritmo médio, praticar esportes por diversão, etc.

 dias

NOS DIAS QUE VOCÊ FAZ ESSAS ATIVIDADES FÍSICAS MÉDIAS, QUANTO TEMPO NO TOTAL ELAS DURAM POR DIA?

 min

4 Histórico Médico

CONSUME CAFÉ?

 Sim, com cafeína
 Sim, descafeinado
 Não

UTILIZA MEDICAMENTOS HIPOLIPEMIANTES (MEDICAMENTOS PARA CONTROLE DE COLESTEROL)?

 Não / Não lembro
 Uso de estatinas
 Uso de outros
 Mais de um tipo

UTILIZA DE MEDICAMENTO ANTI-HIPERTENSIVO?

 Sim
 Não / Não lembro


SUA MÃE, SEU PAI OU ALGUM DE SEUS IRMÃOS OU IRMÃS TEVE OU TEM DIABETES (AÇÚCAR ALTO NO SANGUE E/OU PRESENTE NA URINA)?

 Sim
 Não / Não lembro

SUA MÃE, SEU PAI OU ALGUM DE SEUS IRMÃOS OU IRMÃS TEVE OU TEM HIPERTENSÃO (PRESSÃO ALTA)?

 Sim
 Não / Não lembro

Resultado:



O algoritmo utilizado para chegar a esse resultado foi baseado na dissertação de mestrado de André Rodrigues Oliveira intitulada "Comparação de algoritmos de aprendizagem de máquina para construção de modelos preditivos de diabetes não diagnosticado", sob orientação de Václav Rošter e Cirano Iochpe. Essa dissertação está disponível em: <https://www.lfugs.br/bitstream/handle/70183/40847/00099126.pdf?sequence=1>

Este trabalho foi efetuado com apoio do CNPq (Conselho Nacional de Pesquisa e Desenvolvimento), no âmbito do Edital Universal 14/2013.

Se quiser fazer algum comentário, pode enviar mail para roesler@inf.lfugs.br.

Figure 2. Prototype for a web interface for the risk equation.

Table 8. Coefficients from logistic regression model

Binarized variable	Coefficient
(Intercept)	-1.6929
rcta82	0.1826
a_gidade2	0.6458
a_gidade3	0.9566
a_gidade4	1.0548
a_escolar2	-0.2023
a_escolar3	-0.3556
a_escolar4	-0.6952
diea1331	-0.2811
diea1332	-0.1339
a_binge1	0.2614
a_ativfisica2	-0.1071
a_ativfisica3	-0.3266
a_imc22	-1.0311
a_imc23	-0.8642
a_imc24	-0.3796
a_rcq	0.5417
a_medanthipert1	0.4137
hfda071	-0.1386
hfda111	0.3666

DISCUSSION

We created predictive models for detecting undiagnosed diabetes using data from the ELSA study with different machine-learning algorithms. The best results were achieved through both an artificial neural network and logistic regression, with no relevant difference between them.

Generally, most of the algorithms used achieved mean AUCs greater than 70%. The best algorithm (logistic regression) produced an AUC of 74.4%. Since these test dataset values are superior to the AUCs of several other scores that were previously validated in other populations,²⁰ this score shows potential for use in practice.

The generalization testing showed the results from asking a population similar to that of ELSA some simple questions. Out of 403 individuals from the testing dataset who had diabetes and did not know about their condition, 274 were identified as positive cases (68.0% sensitivity) using the model generated through the logistic regression algorithm. The web tool prototype for detecting undiagnosed diabetes could be refined for use in Brazil.

The methods and concepts for building predictive models for use in healthcare, as well as the challenges and difficulties faced when analyzing healthcare data, have been well described.¹⁷⁻²³ Many groups have published predictive models for detecting undiagnosed diabetes. Although several groups have reported AUCs above 0.80, these values generally reduce to < 0.70 when tested on independent samples.²⁰ Differences in predictive power across studies can be ascribed to different characteristics relating to the different datasets, and to different techniques and methods for building and evaluating the models. The characteristics that may vary across

studies include the definition of the target variable, model objectives and candidate variables, among others. These models are generally constructed using conventional statistical techniques such as logistic regression and Cox regression. Systematic reviews^{5,16,24-26} present several such studies: some, like ours, have focused on predicting undiagnosed diabetes; while others have focused on individuals at high risk of developing incident diabetes.

Use of machine-learning techniques is still new in this field.²⁷⁻²⁹ The main studies have compared the results obtained through using a specific technique with the results obtained through logistic regression. One report³⁰ described creation of pre-diabetes risk models using an artificial neural network and support-vector machines that were applied to data from 4,685 participants in the Korean National Health and Nutrition Examination Survey (KNHANES), collected between 2010 and 2011. In comparison with results³¹ from logistic regression on the same dataset, the models created using support-vector machines and an artificial neural network produced slightly better results.

Two other reports^{32,33} also compared artificial neural networks with logistic regression for creating predictive diabetes models. In the first, models created using artificial neural networks on data from 8,640 rural Chinese adults (760 of them with diabetes) produced better results (AUC = 89.1% ± 1.5%) than models created using logistic regression (AUC = 74.4% ± 2.1%). In the second, a radial basis function artificial neural network that was applied to data from 200 people (100 cases with diabetes and 100 with pre-diabetes) at 17 rural healthcare centers in the municipality of Kermanshah, Iran, showed better results than logistic regression and discriminant analysis, for identifying those with diabetes. Another study³⁴ comparing diabetes models created using data from 2,955 women and 2,915 men in the Korean Health and Genome Epidemiology Study (KHGES) showed similar results from logistic regression and naïve Bayes, although naïve Bayes showed better results with unbalanced datasets. Finally, another study³⁵ used data from 6,647 participants (with 729 positive cases) in the Tehran Lipid and Glucose Study (TLGS) and created models with decision trees reaching 31.1% sensitivity and 97.9% specificity (balanced accuracy was around 64.5%),³⁶ for detecting increased blood glucose levels.

In summary, use of machine-learning techniques may prove to be a viable alternative for building predictive diabetes models, often with good results, but rarely with notably superior results, compared with the conventional statistical technique of logistic regression.

CONCLUSION

Comparison between different techniques showed that all of them produced quite similar results from the same dataset used, thus demonstrating the feasibility of detecting undiagnosed diabetes through easily-obtained clinical data. The predictive algorithm for identifying individuals at high risk of having

undiagnosed diabetes — based only on self-reported information from participants in ELSA-Brasil, from which the highest AUC (0.74) was obtained when tested on a part of the sample that had not been used for its derivation — was a logistic regression equation. However, the machine-learning techniques of artificial neural network, naïve Bayes, k-nearest neighbor and random forest all produced AUCs that were similar or slightly smaller.

REFERENCES

- Glauber H, Karnieli E. Preventing type 2 diabetes mellitus: a call for personalized intervention. *Perm J*. 2013;17(3):74-9.
- Beagley J, Guariguata L, Weil C, Motala AA. Global estimates of undiagnosed diabetes in adults. *Diabetes Res Clin Pract*. 2014;103(2):150-60.
- Guariguata L, Whiting DR, Hambleton I, et al. Global estimates of diabetes prevalence for 2013 and projections for 2035. *Diabetes Res Clin Pract*. 2014;103(2):137-49.
- International Diabetes Federation. *IDF Diabetes Atlas*. 7th ed. Brussels: International Diabetes Federation; 2015. Available from: <http://www.diabetesatlas.org>. Accessed in 2017 (Feb 20).
- Buijsse B, Simmons RK, Griffin SJ, Schulze MB. Risk assessment tools for identifying individuals at risk of developing type 2 diabetes. *Epidemiol Rev*. 2011;33:46-62.
- Thooputra T, Newby D, Schneider J, Li SC. Survey of diabetes risk assessment tools: concepts, structure and performance. *Diabetes Metab Res Rev*. 2012;28(6):485-98.
- 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.
- Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med*. 2011;9(1):103.
- Noble D, Mathur R, Dent T, Meads C, Greenhalgh T. Risk models and scores for type 2 diabetes: systematic review. *BMJ*. 2011;343:d7163.
- Schmidt MI, Duncan BB, Mill JG, et al. Cohort Profile: Longitudinal Study of Adult Health (ELSA-Brasil). *Int J Epidemiol*. 2015;44(1):68-75.
- Aquino EM, Barreto SM, Bensenor IM, et al. Brazilian Longitudinal Study of Adult Health (ELSA-Brasil): objectives and design. *Am J Epidemiol*. 2012;175(4):315-24.
- Hosmer DW, Lemeshow S. *Applied logistic regression*. 2nd ed. Hoboken: Wiley; 2005.
- Haykin SO. *Neural networks and learning machines*. 3rd ed. Upper Saddle River: Prentice Hall; 2008.
- Friedman N, Geiger D, Goldszmidt M. Bayesian Network Classifiers. *Machine Learning*. 1997;29(2-3):131-63. Available from: http://www.cs.technion.ac.il/~dang/journal_papers/friedman1997Bayesian.pdf. Accessed in 2017 (Feb 20).
- Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*. 1967;13(1):21-7. Available from: <http://ieeexplore.ieee.org/document/1053964/>. Accessed in 2017 (Feb 20).
- Breiman L. Random forests. *Machine Learning*. 2001;45(1):5-32. Available from: http://download.springer.com/static/pdf/639/art%253A10.1023%252FA%253A1010933404324.pdf?originUrl=http%3A%2F%2Flink.springer.com%2Farticle%2F10.1023%2FA%3A1010933404324&token2=exp=1487599835~acl=%2Fstatic%2Fpdf%2F639%2Fart%25253A10.1023%25252FA%25253A1010933404324.pdf%3ForiginUrl%3Dhttp%253A%252F%252Flink.springer.com%252Farticle%252F10.1023%252FA%253A1010933404324*~hmac=ba7626571c8b7a2e4710c893c3bc243eb963021f7bbf0e70ef0fe0a27344e28d. Accessed in 2017 (Feb 20).
- Kotsiantis SB, Zaharakis ID, Pintelas PE. Machine learning: a review of classification and combining techniques. *Artif Intell Rev*. 2006;26(3):159-90. Available from: http://www.cs.bham.ac.uk/~pxt/IDA/class_rev.pdf. Accessed in 2017 (Feb 20).
- Gonzalez-Abril L, Cuberos FJ, Velasco F, Ortega JA. Ameva: An autonomous discretization algorithm. *Expert Systems with Applications*. 2009;36(3):5327-32. Available from: <http://sci2s.ugr.es/keel/pdf/algorithm/articulo/2009-Gonzalez-Abril-ESWA.pdf>. Accessed in 2017 (Feb 20).
- Guyon I, Elisseeff A. An introduction to variable and feature selection. *Journal of Machine Learning Research*. 2003;3:1157-82. Available from: <http://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf>. Accessed in 2017 (Feb 20).
- Brown N, Critchley J, Bogowicz P, Mayige M, Unwin N. Risk scores based on self-reported or available clinical data to detect undiagnosed type 2 diabetes: a systematic review. *Diabetes Res Clin Pract*. 2012;98(3):369-85.
- Bellazi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. *Int J Med Inform*. 2008;77(2):81-97.
- Brown DE. Introduction to data mining for medical informatics. *Clin Lab Med*. 2008;28(1):9-35, v.
- Harrison JH Jr. Introduction to the mining of clinical data. *Clin Lab Med*. 2008;28(1):1-7, v.
- Koh HC, Tan G. Data mining applications in healthcare. *J Healthc Inf Manag*. 2005;19(2):64-72.
- Lavrac N. Selected techniques for data mining in medicine. *Artif Intell Med*. 1999;16(1):3-23.
- Obenshain MK. Application of data mining techniques to healthcare data. *Infect Control Hosp Epidemiol*. 2004;25(8):690-5.
- Yoo I, Alafairet P, Marinov M, et al. Data mining in healthcare and biomedicine: a survey of the literature. *J Med Syst*. 2012;36(4):2431-48.
- Barber SR, Davies MJ, Khunti K, Gray LJ. Risk assessment tools for detecting those with pre-diabetes: a systematic review. *Diabetes Res Clin Pract*. 2014;105(1):1-13.
- Shankaracharya, Odedra D, Samanta S, Vidyarthi AS. Computational intelligence in early diabetes diagnosis: a review. *Rev Diabet Stud*. 2010;7(4):252-62.
- Choi SB, Kim WJ, Yoo TK, et al. Screening for prediabetes using machine learning models. *Comput Math Methods Med*. 2014;2014:618976.
- Lee YH, Bang H, Kim HC, et al. A simple screening score for diabetes for the Korean population: development, validation, and comparison with other scores. *Diabetes Care*. 2012;35(8):1723-30.

32. Wang C, Li L, Wang L, et al. Evaluating the risk of type 2 diabetes mellitus using artificial neural network: an effective classification approach. *Diabetes Res Clin Pract.* 2013;100(1):111-8.
33. Mansour R, Eghbal Z, Amirhossein H. Comparison of artificial neural network, logistic regression and discriminant analysis efficiency in determining risk factors of type 2 diabetes. *World Applied Sciences Journal.* 2013;23(11):1522-9. Available from: [https://www.idosi.org/wasj/wasj23\(11\)13/14.pdf](https://www.idosi.org/wasj/wasj23(11)13/14.pdf). Accessed in 2017 (Feb 20).
34. Lee BJ, Ku B, Nam J, Pham DD, Kim JY. Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes. *IEEE J Biomed Heal Inform.* 2014;18(2):555-61.
35. Ramezankhani A, Pournik O, Shahrabi J, et al. Applying decision tree for identification of a low risk population for type 2 diabetes. *Tehran Lipid and Glucose Study. Diabetes Res Clin Pract.* 2014;105(3):391-8.
36. Golino HF, Amaral LS, Duarte SF, et al. Predicting increased blood pressure using machine learning. *J Obes.* 2014;2014:637635.

Acknowledgements: We thank the ELSA-Brasil participants who agreed to collaborate in this study

Sources of funding: ELSA-Brasil was supported by the Brazilian Ministry of Health (Science and Technology Department) and the Brazilian Ministry of Science and Technology (Study and Project Financing Sector and CNPq National Research Council), with the grants 01 06 0010.00 RS, 01 06 0212.00 BA, 01 06 0300.00 ES, 01 06 0278.00 MG, 01 06 0115.00 SP, 01 06 0071.00 RJ and 478518_2013-7; and by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), AUXPE PROEX 2587/2012

Conflict of interest: None

Date of first submission: November 22, 2016

Last received: January 19, 2017

Accepted: February 1, 2017

Address for correspondence:

Bruce Bartholow Duncan
Programa de Pós-Graduação em Epidemiologia e Hospital de Clínicas,
Universidade Federal do Rio Grande do Sul (UFRGS)
Rua Ramiro Barcelos, 2.600/414
Porto Alegre (RS) — Brasil
CEP 90035-003
E-mail: bbduncan@ufrgs.br