Scientific Research Publishing

Applied Mathematics, 2018, 9, 954-974
http://www.scirp.org/journal/am
ISSN Online: 2152-7393
ISSN Print: 2152-7385

Methodology for Constructing a Short-Term Event Risk Score in Heart Failure Patients

Kévin Duarte[1,2*], Jean-Marie Monnez[1,2,3], Eliane Albuisson[4,5,6]

[1]CNRS, INRIA, Institut Elie Cartan de Lorraine, Université de Lorraine, Nancy, France
[2]CHRU Nancy, INSERM, Université de Lorraine, CIC, Plurithématique, Nancy, France
[3]IUT Nancy-Charlemagne, Université de Lorraine, Nancy, France
[4]Institut Elie Cartan de Lorraine, Université de Lorraine, CNRS, Nancy, France
[5]CHRU Nancy, BIOBASE, Pôle S2R, Université de Lorraine, Nancy, France
[6]Faculté de Médecine, InSciDenS, Université de Lorraine, Nancy, France
Email: *k.duarte@chru-nancy.fr

How to cite this paper: Duarte, K., Monnez, J.-M. and Albuisson, E. (2018) Methodology for Constructing a Short-Term Event Risk Score in Heart Failure Patients. *Applied Mathematics*, **9**, 954-974.
https://doi.org/10.4236/am.2018.98065

Received: May 2, 2018
Accepted: August 26, 2018
Published: August 29, 2018

Abstract

We present a methodology for constructing a short-term event risk score in heart failure patients from an ensemble predictor, using bootstrap samples, two different classification rules, logistic regression and linear discriminant analysis for mixed data, continuous or categorical, and random selection of explanatory variables to build individual predictors. We define a measure of the importance of each variable in the score and an event risk measure by an odds-ratio. Moreover, we establish a property of linear discriminant analysis for mixed data. This methodology is applied to EPHESUS trial patients on whom biological, clinical and medical history variables were measured.

Keywords

Ensemble Predictor, Linear Discriminant Analysis, Logistic Regression, Mixed Data, Scoring, Supervised Classification

1. Introduction

In this study, we focus on the problem of constructing a short-term event risk score in heart failure patients based on observations of biological, clinical and medical history variables.

Numerous event risk scores in heart failure patients have been proposed in recent years, but one aspect is particularly important to consider in the construction of a score and in the relevance of the results obtained. This concerns the choice of classification models whose conditions of use may be restrictive. The most currently used classification models in these studies are logistic regres-

sion and Cox proportional hazard model. Quoting for example the Seattle Heart Failure Model (SHFM) risk score [1] and the Seattle Post Myocardial Infarction Model (SPIM) risk score [2] which allow respectively predicting survival in chronic and post-infarction heart failure patients:

- SHFM risk score was derived in a cohort of 1153 patients with ejection fraction < 30% and New York Heart Association (NYHA) class III to IV and validated in 5 other cohorts of patients with similar characteristics. Area under ROC curve (AUC) at 1 year was 0.725 in resubstitution and ranged from 0.679 to 0.810 in the 5 validation cohorts.
- SPIM risk score was derived in a cohort of 6632 patients from the Eplerenone Post-Acute Myocardial Infarction Heart Failure Efficacy and Survival Study (EPHESUS) trial [3] and validated on a cohort of 5477 patients. AUC at 1 year was 0.742 in derivation and 0.774 in validation.

These two risk scores were developed using Cox proportional hazard model and characteristics available at baseline as explanatory variables. Overall, there are several limitations to using these risk scores. They were constructed using only data available at baseline. However, as many studies include inclusion criteria based on clinical or biological parameters measured at baseline, it is possible that some variables are not present in the score due to these inclusion criteria. For example, patients were included in the EPHESUS trial only if their potassium level at baseline was less than 5 mmol/L. This is a reason why potassium is not present in the SPIM score although this is an important parameter which moreover may evolve considerably over time. Concerning the model, the Cox proportional risk model assumes the proportionality of risks, an important condition not always obtained and verified.

In this study, we used a new approach:

- we develop a methodology for constructing a short-term event (death or hospitalization) risk score, taking into account the most recent values of the parameters and therefore the closest values of an event, in order to generate alerts and eventually immediately modify drug prescription; using EPHESUS trial data, we could only construct a score at 1 month in order not to have too few patients with event in the learning sample; but with the same methodology, a score could be constructed at a closer time;
- we use an ensemble predictor, that is more stable than a predictor built on a single learning sample, using bootstrap samples; this allows an internal validation of the score using AUC out-of-bag (OOB); moreover, we use two classification methods, logistic regression and linear discrimination analysis, and, in order to avoid overlearning, for each predictor we use a random selection of explanatory variables, after testing other methods of selection that did not give better results, the number of drawn variables being optimized after testing all possible choices;
- furthermore, our method of construction can be adapted to data streams: when patient data arrives continuously, the coefficients of variables in the score function can be updated online.

In the next section, we present how we defined the learning sample using the available data from EPHESUS trial and the list of explanatory variables used. In the third section, we state a property of linear discriminant analysis (LDA) for mixed data, continuous or categorical. In the fourth section, after presenting the methodology used to build a risk score and to reduce its variation scale from 0 to 100, we define a measure of the importance of variables or groups of correlated variables in the score and a measure of the event risk by an odds-ratio. In the fifth section, we describe the results obtained by applying our methodology to our data. The paper ends with a conclusion.

2. Data

The database at our disposal was EPHESUS, a clinical trial that included 6632 patients with heart failure (HF) after acute myocardial infarction (MI) complicated by left ventricular systolic dysfunction (left ventricular ejection fraction < 40%) [3]. All patients were randomly assigned to treatment with eplerenone 25 mg/day or placebo.

In this trial, each patient was regularly monitored, with visits at the inclusion in the study (baseline), 1 month after inclusion, 3 months later, then every 3 months until the end of follow-up. At each visit, biological, clinical parameters or medical history were observed. In addition, all adverse events (deaths, hospitalizations, diseases) that occurred during follow-up were collected.

To define the learning sample used to construct the short-term event risk score, we made the following working hypothesis: based on biological, clinical measurements or medical history on a patient at a fixed time, we sought to assess the risk that this patient has a short-term HF event. The individuals considered are couples (patient-month) without taking into account the link between several couples (patient-month) concerning the same patient. Therefore, it was assumed that the short-term future of a patient depends only on his current measures.

Firstly, we did a full review of the database in order to:

- identify the biological and clinical variables that were regularly measured at each visit,
- determine the medical history data that we could update from information collected during the follow-up.

We were thus able to define a set of 27 explanatory variables whose list is presented in **Figure 1**. Estimated plasma volume derived from Strauss formula (ePVS) was defined in [4]. Estimated glomerular filtration rate (eGFR) was assessed using three formulas [5] [6] [7]. The different types of hospitalization were defined in supplementary material of [3].

Then, we defined the response variable as the occurrence of a composite short-term HF event (death or hospitalization for progression of HF). In order to have enough events, we defined the short term as being equal to 30 days. Patient-months with a follow-up of less than 30 days and no short-term HF event during this incomplete follow-up period, were not taken into account.

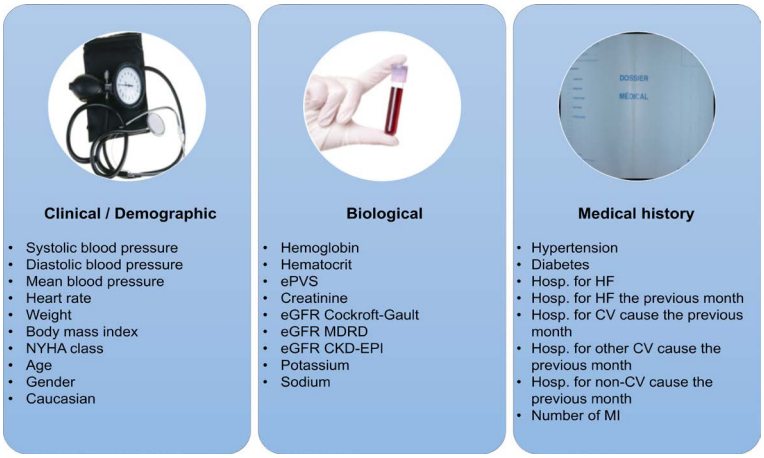

Figure 1. List of variables.

There were finally 21,382 patient-months from 5937 different patients whose 317 with short-term HF event and 21,065 with no short-term event.

3. Property of Linear Discriminant Analysis of Mixed Data

Denote A' the transposed of a matrix A.

In case of mixed data, categorical and continuous, a classical method to perform a discriminant analysis is:

1) perform a preliminary factorial analysis according to the nature of the data, such as multiple correspondence factorial analysis (MCFA) [8] for categorical data, multiple factorial analysis (MFA) [9] for groups of variables, mixed data factorial analysis (MDFA) [10], ... ;

2) after defining a convenient distance, perform a discriminant analysis from the set of values of principal components, or factors.

See for example the DISQUAL (DIScrimination on QUALitative variables) method of Saporta [11], which performs MCFA, then LDA or quadratic discriminant analysis (QDA).

Denote as usual T the total inertia matrix of a dataset partitioned in classes, W and B respectively its intraclass and interclass inertia matrix.

We show hereafter that when performing LDA with metrics T^{-1} or W^{-1}, it is not necessary to perform a preliminary factorial analysis and LDA can be directly performed from the raw mixed data.

Metrics W^{-1} will be used in the following but can be replaced by T^{-1}.

Let $I=\{1,2,\cdots,n\}$ a set of n individuals, partitioned in q disjoint classes $I_1,\cdots,I_q$. Denote $n_k=card(I_k)$, p_{ki} the weight of i^{th} individual of class I_k $(i=1,\cdots,n_k;k=1,\cdots,q)$ and $P_k=\sum_{i=1}^{n_k}p_{ki}$ the weight of I_k, with $\sum_{k=1}^{q}P_k=1$. p quantitative variables or indicators of modalities of categorical variables, denoted $x^1,\cdots,x^p$, are observed on these individuals. Suppose that there exists no affine relation between these variables, especially for each categorical variable an indicator is removed.

For $j=1,\cdots,p$, denote x_{ki}^j the value of x^j for i^{th} individual of class I_k.

Denote x_{ki} the vector $\left(x_{ki}^1 \cdots x_{ki}^p\right)'$ and g_k the barycenter of the elements x_{ki} for $i \in I_k$:

$$g_k = \frac{1}{P_k}\sum_{i\in I_k} p_{ki}x_{ki}. \tag{1}$$

Intraclass inertia (p,p) matrix W is supposed invertible:

$$W = \sum_{k=1}^{q}\sum_{i=1}^{n_k} p_{ki}\left(x_{ki}-g_k\right)\left(x_{ki}-g_k\right)'. \tag{2}$$

A currently used distance in LDA $d_{W^{-1}}(a,b)$ between two points *a* and *b* in $\mathbb{R}^p$ is such that:

$$d_{W^{-1}}^2(a,b) = (a-b)'W^{-1}(a-b). \tag{3}$$

Suppose we want to classify an individual knowing the vector *a* of values of $x^1,\cdots,x^p$. Principle of LDA is to classify it in I_k such that $d_{W^{-1}}^2(a,g_k)$ is minimal.

Consider now new variables $y^1,\cdots,y^m$ affine combinations of $x^1,\cdots,x^p$, with $m \geq p$, such that:

$$y_{ki} = Ax_{ki} + \beta, \tag{4}$$

with $y_{ki} = \left(y_{ki}^1 \cdots y_{ki}^m\right)'$, A a (m,p) matrix of rank p and β a vector in $\mathbb{R}^m$.

Denote h_k the barycenter of vectors y_{ki} in $\mathbb{R}^m$ for $i \in I_k$:

$$h_k = \frac{1}{P_k}\sum_{i\in I_k} p_{ki}y_{ki} = \frac{1}{P_k}\sum_{i\in I_k} p_{ki}\left(Ax_{ki}+\beta\right) = Ag_k + \beta, \tag{5}$$

$$y_{ki} - h_k = A\left(x_{ki}-g_k\right). \tag{6}$$

Let Z the intraclass inertia (m,m) matrix of $\{y_{ki}, i=1,\cdots,n_k; k=1,\cdots,q\}$:

$$Z = \sum_{k=1}^{q}\sum_{i\in I_k} p_{ki}\left(y_{ki}-h_k\right)\left(y_{ki}-h_k\right)' = AWA'. \tag{7}$$

The rank of Z is equal to the rank of A, $p \leq m$. For $m > p$, the (m,m) matrix Z is not invertible. Then use in this case the pseudoinverse (or Moore-Penrose inverse) of Z, denoted Z^+, which is equal to the inverse of Z when $m = p$, to define the pseudodistance denoted d_{Z^+} in $\mathbb{R}^m$. The denomination pseudodistance is used because Z^+ is not positive definite. Remind the definition of a pseudoinverse and two theorems [12].

Definition Let A a (k,l) matrix of rank *r*. The pseudo-inverse of A is the unique (l,k) matrix A^+ such that:

1) $AA^+A = A$,

2) $A^+AA^+ = A^+$,

3) $\left(AA^+\right)' = AA^+$,

4) $\left(A^+A\right)' = A^+A$.

Theorem 1 Maximal rank decomposition

Let A a (k,l) *matrix of rank r. Then there exist two full-rank* (r) *matrices, F of dimension* (k,r) *and G of dimension* (r,l) $(rg(F) = rg(G) = r)$ *such that*

$A = FG$.

Theorem 2 Expression of A^+

Let $A = FG$ *a full-rank decomposition of A. Then* $A^+ = G'\left(F'AG'\right)^{-1}F'$.

Prove now:

Proposition 1 $d^2_{Z+}\left(Aa+\beta, Ab+\beta\right) = d^2_{W^{-1}}\left(a,b\right)$.

Proof. $Z = \left(AW\right)A'$. AW and A' are of full-rank *p*. Applying theorem 2 yields:

$$Z^+ = A\left(\left(AW\right)' AWA'A\right)^{-1}\left(AW\right)' \tag{8}$$

$$= A\left(A'A\right)^{-1}\left(WA'AW\right)^{-1}\left(AW\right)' \tag{9}$$

$$= A\left(A'A\right)^{-1}W^{-1}\left(A'A\right)^{-1}A'. \tag{10}$$

$$A'Z^+A = W^{-1}. \tag{11}$$

Note that, when $m = p$, A is invertible and $Z^+ = \left(AWA'\right)^{-1} = Z^{-1}$.

$$d^2_{Z^+}\left(Aa+\beta, Ab+\beta\right) = \left(A\left(a-b\right)\right)' Z^+\left(A\left(a-b\right)\right) = \left(a-b\right)' W^{-1}\left(a-b\right). \quad \square$$

Thus:

Proposition 2 *Let A a* $\left(m,p\right)$ *matrix,* $m > p$ *, of rank p and for* $k = 1,\cdots,q$ *,* $i = 1,\cdots,n_k$ *,* $y_{ki} = Ax_{ki} + \beta$ *. The results of LDA of the dataset* $\left\{x_{ki}, k = 1,\cdots,q, i = 1,\cdots,n_k\right\}$ *with the metrics* W^{-1} *on* $\mathbb{R}^p$ *are the same as those of LDA of the dataset* $\left\{y_{ki}, k = 1,\cdots,q, i = 1,\cdots,n_k\right\}$ *with the pseudometrics* $Z^+ = \left(AWA'\right)^+$.

Applications

Denote x_i^j the value of the variable x^j for individual *i* belonging to *I*, $i = 1,\cdots,n$, $j = 1,\cdots,p$ and $x_i = \left(x_i^1 \cdots x_i^p\right)'$ the vector of values of $\left(x^1,\cdots,x^p\right)$ for individual *i*. Denote p_i the weight of individual *i*, such that $\sum_{i=1}^{n} p_i = 1$. To perform a factorial analysis of the dataset $\left\{x_i, i = 1,\cdots,n\right\}$, the difference between two individuals *i* and *i'* is measured by a distance $d\left(i,i'\right)$ defined on $\mathbb{R}^p$ associated to a metrics *M*, such that

$$d^2\left(i,i'\right) = \left(x_i - x_{i'}\right)' M\left(x_i - x_{i'}\right). \tag{12}$$

Denote *X* the $\left(n,p\right)$ matrix whose element $\left(i,j\right)$ is x_i^j . Denote *D* the diagonal $\left(n,n\right)$ matrix whose element $\left(i,i\right)$ is p_i .

Perform a factorial analysis of $\left(X,M,D\right)$, for instance principal component analysis (PCA) for continuous variables or MCFA for categorical variables or MDFA for mixed data. Suppose *X* of rank *p*. Denote $u_j = \left(u_j^1 \cdots u_j^p\right)'$ a unit vector of the j^{th} principal axis. Denote $c^j = XMu_j = \left(c_1^j \cdots c_n^j\right)'$ the j^{th} principal component. Denote *U* the $\left(p,p\right)$ matrix $\left(u_1 \cdots u_p\right)$ and *C* the $\left(n,p\right)$ matrix $\left(c^1 \cdots c^p\right) = XMU$; as $u_1,\cdots,u_p$ are *M*-orthonormal, $U'MU = I$ and

$$C = XMU \Leftrightarrow X = CU' \Leftrightarrow \text{for } i = 1,\cdots,n, x_i = Uc_i \tag{13}$$

$$\Leftrightarrow \text{for } i = 1,\cdots,n, c_i = U'Mx_i. \tag{14}$$

Using the metrics of intraclass inertia matrix inverse, LDA from *C* is equivalent to LDA from *X*.

Suppose now that the variable $x^{p+1} = 1 - x^p$ is introduced; when x^p is the

indicator of a modality of a binary variable, x^{p+1} is the indicator of the other modality. Then:

$$\begin{pmatrix} x_i^1 \\ \vdots \\ x_i^p \\ x_i^{p+1} \end{pmatrix} = \begin{pmatrix} u_1^1 & \cdots & u_p^1 \\ \vdots & \ddots & \vdots \\ u_1^p & \cdots & u_p^p \\ -u_1^p & \cdots & -u_p^p \end{pmatrix} \begin{pmatrix} c_i^1 \\ \vdots \\ c_i^p \end{pmatrix} + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \tag{15}$$

Denote X_1 the $(n, p+1)$ matrix whose element (i, j) is x_i^j. LDA from C with the metrics of intraclass inertia matrix inverse is equivalent to LDA from X_1 with the metrics of intraclass inertia matrix pseudoinverse.

For instance:

1) If $x^1, \cdots, x^p$ are continuous variables, LDA from X is equivalent to LDA from C obtained by PCA, such as normed PCA, or generalized canonical correlation analysis (gCCA) [13] and MFA which can be interpreted as PCA with specific metrics.

2) If $x^1, \cdots, x^p$ are indicators of modalities of categorical variables, and if MCFA is performed to obtain C, LDA from C with the metrics of intraclass inertia matrix inverse is equivalent to LDA from X with the metrics of intraclass inertia matrix pseudoinverse.

3) Likewise, if $x^1, \cdots, x^p$ are continuous variables or indicators of modalities of categorical variables, and if MDFA [10] is performed to obtain C, LDA from C with the metrics of intraclass inertia matrix inverse is equivalent to LDA from X with the metrics of intraclass inertia matrix pseudoinverse. In this case, other metrics can also be used, such as that of Friedman [14] or that of Gower [15].

4. Methodology for Constructing a Score

4.1. Ensemble Methods

Consider the problem of predicting an outcome variable y, continuous (in the case of regression) or categorical (in the case of classification) from observable explanatory variables $x^1, \cdots, x^p$, continuous or categorical.

The principle of an ensemble method [16] [17] is to build a collection of N predictors and then aggregate the N predictions obtained using:

- in regression: the average of predictions $\widehat{y_i}$;
- in classification: the rule of the majority vote or the average of the estimations of a posteriori class probabilities.

The ensemble predictor is expected to be better than each of the individual predictors. For this purpose [16]:

- each single predictor must be relatively good,
- single predictors must be sufficiently different from each other.

To build a set of predictors, we can:

- use different classifiers,
- and/or use different samples (e.g. by bootstrapping, boosting, randomizing outputs) [17] [18] [19],
- and/or use different methods of variables selection (e.g. ascending, stepwise,

shrinkage, random) [20] [21] [22] [23],

- and/or in general, introduce randomness into the construction of predictors (e.g. in random forests [24], randomly select a fixed number of variables at each node of a classification or regression tree).

In Random Generalized Linear Model (RGLM) [25], at each iteration,

- a bootstrap sample is drawn,
- a fixed number of variables are randomly selected,
- the selected variables are rank-ordered according to their individual association with the outcome variable y and only the top ranking variables are retained,
- an ascending selection of variables is made using Akaike information criterion (AIC) [26] or Bayesian information criterion (BIC) [27].

Tufféry [28] wrote that logistic models built from bootstrap samples are too similar for their aggregation to really differ from the base model built on the entire sample. This is in agreement with an assertion by Genuer and Poggi [16]. However, Tufféry suggests the use of a method called "random forest of logistic models" introducing an additional randomness: at each iteration,

- a bootstrap sample is drawn,
- variables are randomly selected,
- an ascending variables selection is performed using AIC [26] or BIC [27] criteria.

Note that this method is in fact a particular case of RGLM method.

Present now the method used in this study to check the stability of the predictor obtained on the entire learning sample.

4.2. Method of Construction of an Ensemble Predictor

The steps of the method for constructing an ensemble predictor are presented in the form of a tree (**Figure 2**).

At first step, n_1 classifiers are chosen.

At second step, n_2 bootstrap samples are drawn and are the same for each classifier.

At third step, for each classifier and each bootstrap sample, n_3 modalities of random selection of variables are chosen, a modality being defined either by a number of randomly drawn variables or by a number of predefined groups of correlated variables, which are randomly drawn, inside each of which a variable is randomly drawn.

At fourth step, for each classifier, each bootstrap sample and each modality of random selection of variables, one method of selection of variables is chosen, a stepwise or a shrinkage (LASSO, ridge or elastic net) method.

This yields a set of $n_1 \times n_2 \times n_3$ predictors, which are aggregated to obtain an ensemble predictor.

4.3. Choices Made

To assess accuracy of the ensemble predictor, the percentage of well-classified is currently used. But this criteria is not always convenient, especially in the

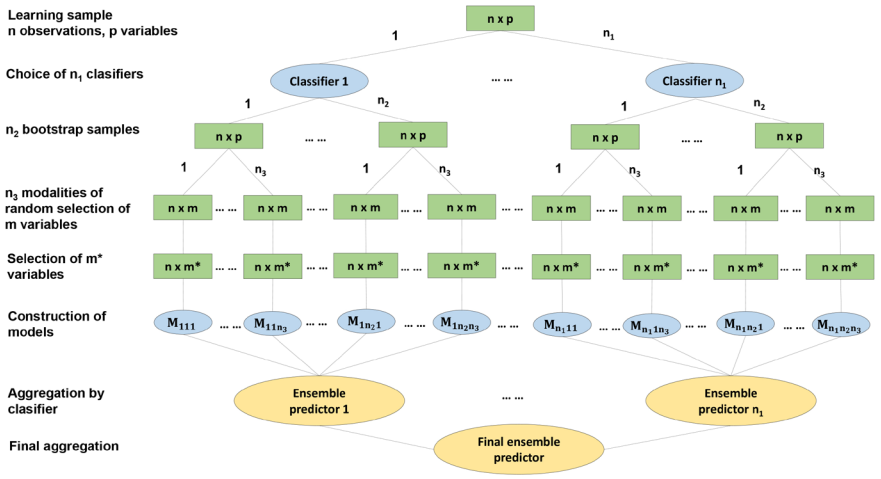

Figure 2. General methodology for the construction of a score.

present case of unbalanced classes. We decided to use AUC. AUC in resubstitution being usually too optimistic, we used AUC OOB [29]: for each patient, consider the set of predictors built on the bootstrap samples that do not contain this patient, *i.e.* for which this patient is "out-of bag", then aggregate the corresponding predictions to obtain an OOB prediction.

Two classifiers were used: logistic regression and LDA with metrics W^{-1}. Other classifiers were tested but not retained because of their less good results, such as random forest-random input (RF-RI) [24] or QDA. The k-nearest neighbors method (k-NN) was not tested, because it was not adapted to this study due to the presence of very unbalanced classes with a too small class size.

1000 bootstrap samples were randomly drawn.

Three modalities of random selection were retained, firstly a random draw of a fixed number of variables, secondly and thirdly a random draw of a fixed number of predefined groups of correlated variables followed by a random draw of one variable inside each drawn group. The number of variables or of groups drawn was determined by optimization of AUC OOB.

Fourth step did not improve prediction accuracy and was not retained.

4.4. Construction of an Ensemble Score

Denote n the total number of patient-months and p the number of variables. Denote x_i^j the value of variable x^j for patient-month i, $i=1,\cdots,n$, $j=1,\cdots,p$. Each patient-month i is represented by a vector $x_i=\left(x_i^1\cdots x_i^p\right)'$ in $\mathbb{R}^p$.

4.4.1. Aggregation of Predictors

In the case of two classes Ω_1 and Ω_0, whose barycenters are respectively denoted g_1 and g_0, Fisher linear discriminant function

$$S_1(x)=\left(x-\frac{g_1+g_0}{2}\right)'W^{-1}(g_1-g_0)=\alpha_1'x+\beta_1 \tag{16}$$

can be used as score function. For logistic regression, the following score function can be used:

$$S_2(x) = \ln \frac{P(\Omega_1 \mid X = x)}{P(\Omega_0 \mid X = x)} = \alpha_2' x + \beta_2. \tag{17}$$

Remind that, in the case of a multinormal model with homoscedasticity (covariance matrices within classes are equal), when $P(\Omega_1) = P(\Omega_0)$, logistic model is equivalent to LDA [17]; indeed:

$$S_2(x) = \ln \frac{P(\Omega_1 \mid X = x)}{P(\Omega_0 \mid X = x)} = \ln \frac{P(\Omega_1)}{P(\Omega_0)} + S_1(x) = S_1(x). \tag{18}$$

So we used the following method to aggregate the obtained predictors:

1) the score functions obtained by LDA are aggregated by averaging; denote now S_1 the averaged score;

2) likewise the score functions obtained by logistic regression are aggregated by averaging; denote S_2 the averaged score;

3) a combination of the two scores, $\lambda S_1 + (1-\lambda) S_2$ is defined, $0 \le \lambda \le 1$; a value of λ that maximizes AUC OOB is retained; denote S_0 the optimal score obtained by this method.

If s is an optimal cut-off, the ensemble classifier is defined by:

$$\text{If } S_0(x) > s, \ x \text{ is classified in } \Omega_1; \tag{19}$$

$$\text{if not, } x \text{ is classified in } \Omega_0. \tag{20}$$

4.4.2. Definition of a Score from 0 to 100

The variation scale of the score function $S_0(x)$ was reduced from 0 to 100 using the following method. Denote:

$$S_0(x) = \alpha_0' x + \beta_0 = \sum_{j=1}^{p} \alpha_0^j x^j + \beta_0. \tag{21}$$

Denote for $j = 1, \cdots, p$:

$$P_j = \left|\alpha_0^j\right| \left(\max_{1 \le i \le n} x_i^j - \min_{1 \le i \le n} x_i^j \right) \tag{22}$$

and

$$P = \sum_{j=1}^{p} P_j = \sum_{j=1}^{p} \left|\alpha_0^j\right| \left(\max_{1 \le i \le n} x_i^j - \min_{1 \le i \le n} x_i^j \right). \tag{23}$$

Let m^j the minimal value of the variable x^j if $\alpha_0^j > 0$, or its maximal value if $\alpha_0^j < 0$.

Denote $S(x)$ the "normalized" score function, with values from 0 to 100, defined by:

$$S(x) = \frac{100}{P} \sum_{j=1}^{p} \alpha_0^j \left(x^j - m^j \right) \tag{24}$$

$$= 100 \sum_{j=1}^{p} \frac{\alpha_0^j \left(x^j - m^j \right)}{\sum_{k=1}^{p} \left|\alpha_0^k\right| \left(\max_{1 \le i \le n} x_i^k - \min_{1 \le i \le n} x_i^k \right)} \tag{25}$$

$$= \alpha' x + \beta, \text{ with } \begin{pmatrix} \beta \\ \alpha_1 \\ \vdots \\ \alpha_p \end{pmatrix} = \begin{pmatrix} -\frac{100}{P}\sum_{j=1}^{p} \alpha_0^j m^j \\ 100\frac{\alpha_0^1}{P} \\ \vdots \\ 100\frac{\alpha_0^p}{P} \end{pmatrix}. \quad (26)$$

4.4.3. Measure of Variables Importance

Explanatory variables are not expressed in the same unit. To assess their importance in the score, we used "standardized" coefficients, multiplying the coefficient of each variable in the score by its standard deviation. These coefficients are those associated with standardized variables and are directly comparable. For all variables, the absolute values of their standardized coefficient, from the greatest to the lowest, were plotted on a graph. The same type of plot was used for groups of correlated variables, whose importance is assessed by the sum of absolute values of their standardized coefficients.

4.4.4. Risk Measure by an Odds-Ratio

Define a risk measure associated to a score s by an odds-ratio $OR_1(s)$:

$$OR_1(s) = \frac{P(Y=1 \mid S>s)}{P(Y=0 \mid S>s)} \frac{P(Y=0)}{P(Y=1)} = \frac{P(S>s \mid Y=1)}{P(S>s \mid Y=0)} = \frac{Se(s)}{1-Sp(s)}. \quad (27)$$

An estimation of $OR_1(s)$, also denoted $OR_1(s)$, is $\frac{n_1}{n_0} \times \frac{N_0}{N_1}$ with $n_k = \#\{S>s\} \cap \{Y=k\}$ and $N_k = \#\{Y=k\}$, $k=0,1$.

Note that:

- $OR_1(s)$ decreases when $Se(s)$ decreases and $Sp(s)$ is constant. In practice, the decrease will be much smaller when there are many observations;
- $OR_1(s)$ is not defined when $Sp(s)$ is equal to 1.

For these reasons, the following definition can also be used:

$$OR_2(s) = \max_{t \le s:\, OR_1(t) < \infty} OR_1(t). \quad (28)$$

Note that OR_1 is the slope y/x of the line joining the origin to the point (x,y) of the ROC curve. In the case of an "ideal" ROC curve, supposed continuous above the diagonal line, assuming that there is no vertical segment in the curve, this slope increases from point $(1,1)$, corresponding to the minimal value of score, to point $(0,0)$, corresponding to its maximal value; the case of a vertical segment (*Se* decreases, *Sp* is constant), occurring when the score of a patient with event is between those of two patients without event, is particularly visible in the case of a small number of patients and also justifies the definition of OR_2, whose curve fits that of OR_1.

For very high score values, when n_0 or n_1 are too small, the estimation of OR_1 is no longer reliable. A reliability interval of the score could be defined, depending on the values of n_0 and n_1.

5. Results

5.1. Pre-Processing of Variables

5.1.1. Winsorization

To avoid problems related to the presence of outliers or extreme data, all continuous variables were winsorized using the 1st percentile and the 99th percentile of each variable as limit values [30]. We chose this solution because of the large imbalance of the classes (317 patients with event against 21,065 with no event, so there is a ratio of about 1 to 66). The elimination of extreme data would have led to decrease the number of patients with event.

5.1.2. Transformation of Variables

Among qualitative variables, two are ordinal: the NYHA class with 4 modalities and the number of myocardial infarction (no. MI) with 5 modalities. In order to preserve the ordinal nature of these variables, we chose to use an ordinal encoding. For NYHA, we therefore associated 3 binary variables: NYHA ≥ 2, NYHA ≥ 3 and NYHA ≥ 4. In the same way, for the no. MI, we considered 4 binary variables: no. MI ≥ 2, no. MI ≥ 3, no. MI ≥ 4 and no. MI ≥ 5.

On the other hand, continuous variables were transformed in the context of logistic regression. For each continuous variable, a linearity test was performed using the method of restricted cubic splines with 3 knots [31]. A cubic spline restricted with 3 knots is composed of a linear component and a cubic component. Linearity testing is to test, under the univariable logistic model, the nullity of the coefficient associated with the cubic component. To do this, we used the likelihood ratio test. The results of linearity tests are given in **Table 1** (p-value 1).

Table 1. Linearity tests and transformation of continuous variables.

Variable	p-value 1	Transformation function $f(x)$	p-value 2
Hemoglobin	0.090		
Hematocrit	0.007	x^{-2}	1.00
ePVS	0.69		
Creatinine	0.21		
eGFR Cockroft-Gault	<0.0001	$\ln(x)$	0.40
eGFR MDRD	<0.0001	$x^{-0.5}$	0.79
eGFR CKD-EPI	0.005	$\ln(x)$	0.90
Sodium	0.056		
Potassium	<0.0001	$(x-4.6)^2$	0.47
Heart rate	<0.0001	$(x-60)^2$	0.91
Systolic BP	<0.0001	$(x-140)^2$	0.34
Diastolic BP	<0.0001	$(x-84)^2$	0.49
Mean BP	<0.0001	$(x-102)^2$	0.66
Weight	0.090		
BMI	0.060		
Age	0.64		

At 5% level, linearity was rejected for 9 of 16 continuous variables. For each of these 9 variables, we represented graphically the relationship between the logit (natural logarithm of the ratio probability of event/probability of non-event) and the variable. An example of graphical representation is given for potassium: we observe a quadratic relationship between the logit and the potassium (**Figure 3**). In agreement with the relationship observed, we applied a simple, monotonous or quadratic transformation function to each of the 9 variables. The transformation function applied to each variable is given in **Table 1**.

For hematocrit and the three variables of eGFR, the relationship is clearly monotonous. So we considered some simple monotonic transformation functions as $f(x)=x^a$ with $a\in\{-2,-1,-0.5,0.5,1,2\}$ or $f(x)=\ln(x)$, then we retained for each variable the transformation for which the likelihood under univariable logistic model was maximal (minimal p-value).

For other variables not checking linearity, namely potassium, the three blood pressure measures (systolic, diastolic and mean), and heart rate, the relationship between the logit and the variable was rather quadratic. We therefore applied a quadratic transformation function $(X-k^*)^2$ with k^* an optimal value determined by maximizing likelihood under univariable logistic model. To compare, we also used the criterion of maximal AUC to determine an optimal value. These results are presented in **Table 2**. Notice that the optimal values determined by the two methods are the same for systolic BP, diastolic BP and heart rate and are very close for potassium and mean BP.

Also note that the transformation applied to potassium allows to take into account both hypokalemia and hyperkalemia, two different clinical situations pooled here that may increase the risk of death and/or hospitalization measured by the score.

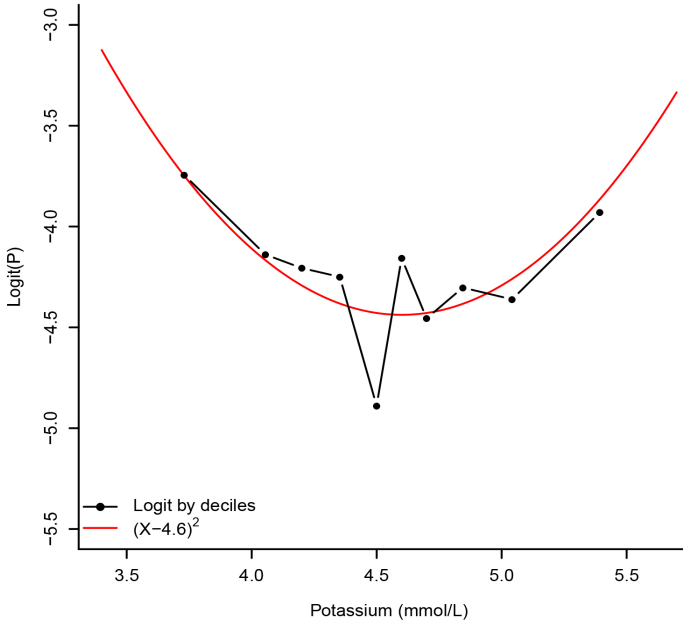

Figure 3. Relationship between potassium and logit of probability of event.

Table 2. Quadratic transformations.

Variable	"Raw" variable X	Criterion 1 Maximizing likelihood for $(X-k^*)^2$		Criterion 2 Maximal AUC for $(X-k^*)^2$	
	AUC	k^*	AUC	k^*	AUC
Systolic BP	0.5818	140	0.5995	140	0.5995
Diastolic BP	0.5834	84	0.5970	84	0.5970
Mean BP	0.5915	102	0.6091	101	0.6094
Potassium	0.5312	4.6	0.5665	4.7	0.5676
Heart rate	0.6473	60	0.6521	60	0.6521

To verify that the transformation of the variables was good, a linearity test for each transformed variable was performed according to the previously detailed principle. All tests are not significant at the 5% level (see Table 1, p-value 2).

5.2. Ensemble Score

5.2.1. Ensemble Score by Logistic Regression

As a first step, we applied our methodology with the following parameters:

- use of a single classification rule, logistic regression ($n_1 = 1$),
- draw of 1000 bootstrap samples ($n_2 = 1000$),
- random selection of variables according to a single modality ($n_3 = 1$).

Three modalities for the random selection of variables were defined:

- 1st modality: random draw of m variables among 32,
- 2nd modality: random draw of m groups among 18, then one variable from each drawn group,
- 3rd modality: random draw of m groups among 24, then one variable from each drawn group.

The groups of variables considered for each modality are presented in Table 3. For modalities 2 and 3, we formed groups of variables based on correlations between variables. For the second modality, we gathered for example in the same group hemoglobin, hematocrit and ePVS because of their high correlations. For the third modality, the same groups were used, except for the two variables linked to hospitalization for HF, the four variables linked to the no. MI and the three variables related to the NYHA class, for which each binary variable was considered as a single group.

For each modality, an ensemble score was built for all possible values of m and the one that gave maximal AUC OOB was selected. In Table 4 are reported the results obtained for each modality with the optimal m. The best result was obtained for the third modality, with AUC OOB equal to 0.8634.

The ensemble score by logistic regression, denoted $S_2(x)$, obtained by averaging the three ensemble scores that we constructed, gave slightly better results, with AUC OOB of 0.8649.

5.2.2. Ensemble Score by LDA for Mixed Data

The same methodology was used by simply replacing the classification rule (logistic regression) by LDA for mixed data and keeping the same other settings.

Table 3. Composition of groups of variables.

Variables	Modality 1	Modality 2	Modality 3
Systolic BP	-		
Diastolic BP	-	Blood pressure	Blood pressure
Mean BP	-		
Heart rate	-	-	-
Weight	-	Obesity	Obesity
BMI	-		
NYHA ≥ 2	-		-
NYHA ≥ 3	-	NYHA	-
NYHA ≥ 4	-		-
Age	-	-	-
Gender	-	-	-
Caucasian	-	-	-
Hemoglobin	-		
Hematocrit	-	Hematology	Hematology
ePVS	-		
Creatinine	-		
eGFR Cockroft-Gault	-	Renal function	Renal function
eGFR MDRD	-		
eGFR CKD-EPI	-		
Potassium	-	-	-
Sodium	-	-	-
Hypertension	-	-	-
Diabetes	-	-	-
Hosp. for HF	-	Previous hosp. for HF	-
Hosp. for HF the previous month	-		-
Hosp. for CV cause the previous month	-	-	-
Hosp. for other CV cause the previous month	-	-	-
Hosp. for non CV cause the previous month	-	-	-
No. MI ≥ 2	-		-
No. MI ≥ 3	-	No. MI	-
No. MI ≥ 4	-		-
No. MI ≥ 5	-		-

Table 4. Results obtained by logistic regression.

Parameters	AUC in resubstitution	AUC OOB
Modality 1 $m = 19$	0.8716	0.8616
Modality 2 $m = 14$	0.8688	0.8611
Modality 3 $m = 8$	0.8691	0.8634
Ensemble score	0.8728	0.8649

Again, for each modality, we searched the optimal m parameter. The obtained results are presented in Table 5.

As for logistic regression, the best results were obtained for the third modality, with AUC OOB equal to 0.8638.

Table 5. Results obtained by LDA for mixed data.

Parameters	AUC in resubstitution	AUC OOB
Modality 1 $m = 12$	0.8679	0.8614
Modality 2 $m = 5$	0.8673	0.8631
Modality 3 $m = 7$	0.8690	0.8638
Ensemble score	0.8707	0.8654

The ensemble score by LDA, denoted $S_1(x)$, yielded better results with AUC OOB equal to 0.8654.

5.2.3. Ensemble Score Obtained by Synthesis of Logistic Regression and LDA

The final ensemble score denoted $S_0(x)$, obtained by synthesis of the two ensemble scores $S_1(x)$ and $S_2(x)$ presented previously, provided the best results with AUC equal to 0.8733 in resubstitution and 0.8667 in OOB.

This ensemble score corresponds to the one obtained by applying our methodology with the following parameters:

- two classification rules are used, logistic regression and LDA for mixed data ($n_1 = 2$),
- 1000 bootstrap samples are drawn ($n_2 = 1000$),
- m variables are randomly selected according to three modalities ($n_3 = 3$).

The scale of variation of the score function $S_0(x)$ was reduced from 0 to 100 according to the procedure described previously. We denote this "normalized" score $S(x)$.

In **Table 6**, we present the "raw" and "standardized" coefficients associated with each of the variables in the score function $S_0(x)$ and the "normalized" score function $S(x)$.

5.2.4. Importance of Variables in the Score

To have a global view of the importance of the variables in the "normalized" score, we represented on a graph the absolute value of standardized coefficient associated with each variable, from the largest value to the smallest (see **Figure 4**). Note that the most important variables are heart rate, NYHA class ≥ 3 and history of hospitalization for HF in the previous month. On the other hand, variables such as weight, no. MI ≥ 5 or BMI do not play a large part in the presence of others.

The same type of graph was made to represent the importance of the groups of variables in configuration 2 defined by the sum of the absolute values of the "standardized" coefficients associated with the variables of the group, from the largest sum to the smallest (see **Figure 4**). Note that the two most influential groups are "NYHA" (NYHA ≥ 2, NYHA ≥ 3 and NYHA ≥ 4) and "History of hospitalization for HF" (hospitalization for HF in the previous month and hospitalization for HF during life). Three important groups follow: "Hematology" (ePVS, hemoglobin, hematocrit), "Heart rate" and "Renal function" (creatinine and three formulas of eGFR). The least important groups of variables are "Obesity" (weight, BMI) and "Gender".

Table 6. Ensemble score.

Variables	Ensemble score $S_0(x)$		Ensemble score "normalized" $S(x)$	
	coefficient	Standardized coefficient	coefficient	Standardized coefficient
Constant	−0.210	−0.210	44.60	44.60
Hemoglobin	−0.0580	−0.0871	−0.478	−0.717
Hematocrit^{-2}	314.00	0.0442	2590.00	0.364
ePVS	0.131	0.107	1.07	0.877
Creatinine	0.00349	0.0964	0.0287	0.794
Ln (eGFR Cockroft-Gault)	−0.0940	−0.0396	−0.774	−0.326
eGFR MDRD$^{-0.5}$	−0.892	−0.0183	−7.34	−0.151
Ln(eGFR CKD-EPI)	−0.175	−0.0590	−1.44	−0.486
Sodium	−0.0232	−0.0861	−0.191	−0.709
(Potassium-4.6)^2	0.301	0.0889	2.48	0.732
(Heart rate-60)^2	0.000696	0.221	0.00572	1.82
(Systolic BP-140)^2	0.000125	0.0729	0.00103	0.600
(Diastolic BP-84)^2	0.0000985	0.0220	0.000810	0.181
(Mean BP-102)^2	0.000201	0.0545	0.00165	0.448
Weight	0.0000258	0.000374	0.000212	0.00308
BMI	0.00196	0.00844	0.0161	0.0695
Age	0.00449	0.0506	0.0370	0.416
Caucasian	−0.162	−0.0455	−1.33	−0.374
Male	0.0434	0.0195	0.357	0.161
Hypertension	0.136	0.0665	1.12	0.547
Diabetes	0.0904	0.0422	0.744	0.347
Hosp. for HF	0.549	0.175	4.52	1.44
Hosp. for HF the previous month	1.53	0.185	12.60	1.52
Hosp. for CV cause the previous month	0.403	0.168	3.31	1.38
Hosp. for non-CV cause the previous month	0.361	0.0486	2.97	0.400
Hosp. for other CV cause the previous month	0.104	0.0205	0.852	0.169
No. MI ≥ 2	0.0840	0.0377	0.692	0.310
No. MI ≥ 3	0.118	0.0323	0.973	0.266
No. MI ≥ 4	0.242	0.0342	1.99	0.281
No. MI ≥ 5	0.0443	0.00370	0.365	0.0304
NYHA ≥ 2	0.309	0.150	2.54	1.23
NYHA ≥ 3	0.612	0.194	5.04	1.60
NYHA ≥ 4	1.65	0.142	13.60	1.16

5.2.5. Risk Measure by an Odds-Ratio

We represented the variation of n_0, n_1, $Se(s)$, $1-Sp(s)$, $OR_1(s)$ and $OR_2(s)$ according to the score s (Table 7). For score values $s > 49.1933$, n_1 is less than or equal to 30. Thus, beyond this threshold value 49.1933, OR_1 is no longer very reliable. We therefore defined as reliability interval of the OR_1 and OR_2 functions $[0;49.1933]$.

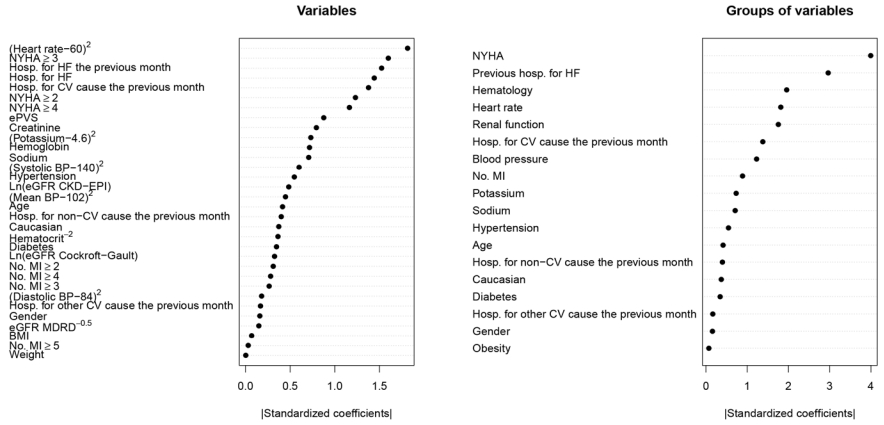

Figure 4. Importance of variables and groups of variables.

Table 7. Variation of n_0, n_1, $Se(s)$, $1-Sp(s)$, $OR_1(s)$ and $OR_2(s)$ according to the values of score s.

s	n_0	n_1	$Se(s)$	$1-Sp(s)$	$OR_1(s)$	$OR_2(s)$
s^* = 23.7094	4527	250	0.7918	0.2149	3.6844	3.6844
11.8489	19683	317	1.0000	0.9344	1.0702	1.0702
13.7320	17684	316	0.9968	0.8395	1.1874	1.1874
15.1105	15684	316	0.9968	0.7446	1.3388	1.3388
16.3630	13686	314	0.9905	0.6498	1.5245	1.5245
17.6044	11689	311	0.9811	0.5549	1.7679	1.7679
18.9050	9697	303	0.9558	0.4604	2.0762	2.0762
20.4525	7709	291	0.9180	0.3660	2.5081	2.5081
22.3007	5729	271	0.8549	0.2720	3.1428	3.1428
24.7670	3766	234	0.7382	0.1788	4.1278	4.1278
28.8573	1822	178	0.5615	0.0865	6.4884	6.4884
33.2656	872	128	0.4038	0.0414	9.7431	9.8363
38.2403	414	86	0.2713	0.0197	13.7706	13.7706
49.1933	70	30	0.0978	0.0033	29.4283	31.5217
55.1424	28	22	0.0694	0.0014	50.4112	50.4112
58.0352	14	16	0.0505	0.0007	70.8812	74.7575

We represented the variation of odds-ratio OR_1 and OR_2 in this reliability interval (**Figure 5**). By reading the graph, for a patient with a score of 40 for example, $\frac{P(Y=1|S>40)}{P(Y=0|S>40)}$ is about 15 times higher than $\frac{P(Y=1)}{P(Y=0)}$.

6. Conclusions and Perspectives

In this article, we presented a new methodology for constructing a short-term event risk score in heart failure patients, based on an ensemble predictor built using two classification rules (logistic regression and LDA for mixed data), 1000 bootstrap samples and three modalities of random selection of variables. This score was normalized on a scale from 0 to 100. AUC OOB is equal to 0.8667. Note

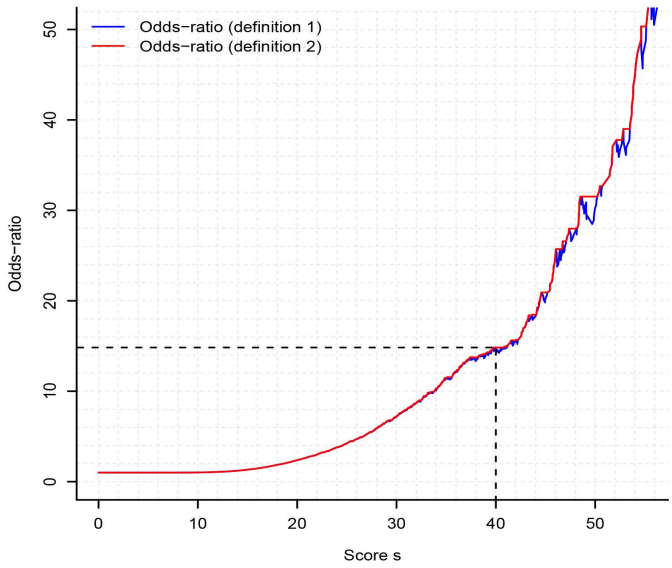

Figure 5. Risk measure by an odds-ratio.

that an important variable such as potassium that does not appear in other scores (as SPIM risk score) is taken into account in this score.

Moreover, we defined a measure of the importance of each variable and each group of variables in the score and defined an event risk measure by an odds-ratio.

Due to the nature of the data available (data obtained from the EPHESUS study), we had to define the short term to 30 days in order to have enough patients with HF event. It would be better to have data of patients with shorter intervals, in order to have data the closest possible of an event and eventually improve the quality of the score. When such data will be available, it will be interesting to apply the same methodology to construct a new score.

Furthermore, we proved a property of linear discriminant analysis for mixed data.

Finally, this methodology can be adapted to the case of a data stream. Suppose that new data for heart failure patients arrives continuously. Data can be allocated to bootstrap samples using Poisson bootstrap [32]. The coefficients of each variable in each predictor based on logistic regression or binary linear discriminant analysis can be updated online using a stochastic gradient algorithm. Such algorithms are presented in [33] for binary LDA and [34] for logistic regression; they use online standardized data in order to avoid a numerical explosion in the presence of extreme values. Thus the ensemble score obtained by averaging can be updated online. To the best of our knowledge, it is the first time that this problematics is studied in this context.

Acknowledgements

Results incorporated in this article received funding from the Investments for the Future program under grant agreement No ANR-15-RHU-0004.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Levy, W.C., Mozaffarian, D., Linker, D.T., *et al.* (2006) The Seattle Heart Failure Model: Prediction of Survival in Heart Failure. *Circulation*, **113**, 1424-1433. https://doi.org/10.1161/CIRCULATIONAHA.105.584102

[2] Ketchum, E.S., Dickstein, K., Kjekshus, J., *et al.* (2014) The Seattle Post Myocardial Infarction Model (SPIM): Prediction of Mortality after Acute Myocardial Infarction with Left Ventricular Dysfunction. European Heart Journal: *Acute Cardiovascular Care*, **3**, 46-55. https://doi.org/10.1177/2048872613502283

[3] Pitt, B., Remme, W., Zannad, F., *et al.* (2003) Eplerenone, a Selective Aldosterone Blocker, in Patients with Left Ventricular Dysfunction after Myocardial Infarction. *New England Journal of Medicine*, **348**, 1309-1321. https://doi.org/10.1056/NEJMoa030207

[4] Duarte, K., Monnez, J.M., Albuisson, E., Pitt, B., Zannad, F. and Rossignol, P. (2015) Prognostic Value of Estimated Plasma Volume in Heart Failure. *JACC*: *Heart Failure*, **3**, 886-893. https://doi.org/10.1016/j.jchf.2015.06.014

[5] Cockcroft, D.W. and Gault, H. (1976) Prediction of Creatinine Clearance from Serum Creatinine. *Nephron*, **16**, 31-41. https://doi.org/10.1159/000180580

[6] Levey, A.S., Coresh, J., Balk, E., *et al.* (2003) National Kidney Foundation Practice Guidelines for Chronic Kidney Disease: Evaluation, Classification, and Stratification. *Annals of Internal Medicine*, **139**, 137-147. https://doi.org/10.7326/0003-4819-139-2-200307150-00013

[7] Levey, A.S., Stevens, L.A., Schmid, C.H., *et al.* (2009) A New Equation to Estimate Glomerular Filtration Rate. *Annals of Internal Medicine*, **150**, 604-612. https://doi.org/10.7326/0003-4819-150-9-200905050-00006

[8] Lebart, L., Morineau, A. and Warwick, K. (1984) Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices. Wiley, New York.

[9] Escofier, B. and Pagès, J. (1990) Multiple Factor Analysis. *Computational Statistics and Data Analysis*, **18**, 121-140. https://doi.org/10.1016/0167-9473(94)90135-X

[10] Pagès, J. (2004) Analyse Factorielle de Données Mixtes. *Revue de Statistique Appliquée*, **52**, 93-111.

[11] Saporta, G. (1977) Une Méthode et un Programme d'Analyse Discriminante sur Variables Qualitatives. Analyse des Données et Informatique, Inria, 201-210.

[12] Rotella, F. and Borne, P. (1995) Théorie et Pratique du Calcul Matriciel. Editions Technip.

[13] Carroll, J.D. (1968) A Generalization of Canonical Correlation Analysis to Three or More Sets of Variables. *Proceedings of the* 76*th Annual Convention of the American Psychological Association*, Washington DC, 227-228.

[14] Friedman, J.H. and Meulman, J.J. (2004) Clustering Objects on Subsets of Attributes (with Discussion). *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), **66**, 815-849. https://doi.org/10.1111/j.1467-9868.2004.02059.x

[15] Gower, J.C. (1971) A General Coefficient of Similarity and Some of its Properties. *Biometrics*, **27**, 857-871. https://doi.org/10.2307/2528823

[16] Genuer, R. and Poggi, J.M. (2017) Arbres CART et Forêts Aléatoires, Importance et

Sélection de Variables. https://arxiv.org/pdf/1610.08203v2.pdf

[17] Hastie, T., Tibshirani, R. and Friedman, J. (2009) The Elements of Statistical Learning. Springer, New York. https://doi.org/10.1007/978-0-387-84858-7

[18] Efron, B. and Tibshirani, R.J. (1994) An Introduction to the Bootstrap. CRC Press, Boca Raton.

[19] Breiman, L. (1996) Bagging Predictors. *Machine Learning*, **24**, 123-140. https://doi.org/10.1007/BF00058655

[20] In Lee, K. and Koval, J.J. (1997) Determination of the Best Significance Level in Forward Stepwise Logistic Regression. *Communications in Statistics-Simulation and Computation*, **26**, 559-575. https://doi.org/10.1080/03610919708813397

[21] Wang, Q., Koval, J.J., Mills, C.A. and Lee, K.I.D. (2007) Determination of the Selection Statistics and Best Significance Level in Backward Stepwise Logistic Regression. *Communications in Statistics-Simulation and Computation*, **37**, 62-72. https://doi.org/10.1080/03610910701723625

[22] Bendel, R.B. and Afifi, A.A. (1977) Comparison of Stopping Rules in Forward "Stepwise" Regression. *Journal of the American Statistical Association*, **72**, 46-53.

[23] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B* (*Methodological*), **58**, 267-288. http://www.jstor.org/stable/2346178

[24] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-35. https://doi.org/10.1023/A:1010933404324

[25] Song, L., Langfelder, P. and Horvath, S. (2013) Random Generalized Linear Model: A Highly Accurate and Interpretable Ensemble Predictor. *BMC Bioinformatics*, **14**, 5. https://doi.org/10.1186/1471-2105-14-5

[26] Akaike, H. (1998) Information Theory and an Extension of the Maximum Likelihood Principle. In: Parzen, E., Tanabe, K. and Kitagawa, G., Eds., *Selected Papers of Hirotugu Akaike*, Springer Series in Statistics (Perspectives in Statistics), Springer, New York, 199-213.

[27] Schwarz, G. (1978) Estimating the Dimension of a Model. *The Annals of Statistics*, **6**, 461-464. https://doi.org/10.1214/aos/1176344136

[28] Tufféry, S. (2015) Modélisation Prédictive et Apprentissage Statistique avec R. Editions Technip.

[29] Breiman, L. (1996) Out-of-Bag Estimation. https://www.stat.berkeley.edu/~breiman/OOBestimation.pdf

[30] Dixon, W.J. (1960) Simplified Estimation from Censored Normal Samples. *The Annals of Mathematical Statistics*, **31**, 385-391. https://doi.org/10.1214/aoms/1177705900

[31] Royston, P. and Sauerbrei, W. (2007) Multivariable Modeling with Cubic Regression Splines: A Principled Approach. *Stata Journal*, **7**, 45-70.

[32] Oza, N.C. and Russell, S. (2001) Online Bagging and Boosting. *Proceedings of Eighth International Workshop on Artificial Intelligence and Statistics*, Key West, 4-7 January 2001, 105-112.

[33] Duarte, K., Monnez, J.M. and Albuisson, E. (2018) Sequential Linear Regression with Online Standardized Data. *PLoS ONE*, **13**, e0191186. https://doi.org/10.1371/journal.pone.0191186

[34] Monnez, J.M. (2018) Online Logistic Regression Process with Online Standardized Data.