Scientific Research Publishing

Applied Mathematics, 2021, 12, 627-653
https://www.scirp.org/journal/am
ISSN Online: 2152-7393
ISSN Print: 2152-7385

Construction of Parsimonious Event Risk Scores by an Ensemble Method. An Illustration for Short-Term Predictions in Chronic Heart Failure Patients from the GISSI-HF Trial

Benoît Lalloué[1,2*], Jean-Marie Monnez[1,2], Donata Lucci[3], Eliane Albuisson[4,5,6]

[1]Université de Lorraine, CNRS, Inria (Project-Team BIGS), IECL (Institut Elie Cartan de Lorraine), Nancy, France
[2]Inserm U1116, Centre d'Investigation Clinique Plurithématique 1433, Université de Lorraine, Nancy, France
[3]ANMCO Research Center, Florence, Italy
[4]Université de Lorraine, CNRS, IECL (Institut Elie Cartan de Lorraine), Nancy, France
[5]DRCI, CHRU de Nancy, Vandœuvre-lès-Nancy, France
[6]Faculté de Médecine, Département Grand Est de Recherche en Soins Primaires, Vandœuvre-lès-Nancy, France
Email: *b.lalloue@gmail.com, jean-marie.monnez@univ-lorraine.fr, donata.lucci@anmco.it, eliane.albuisson@univ-lorraine.fr

How to cite this paper: Lalloué, B., Monnez, J.-M., Lucci, D. and Albuisson, E. (2021) Construction of Parsimonious Event Risk Scores by an Ensemble Method. An Illustration for Short-Term Predictions in Chronic Heart Failure Patients from the GISSI-HF Trial. *Applied Mathematics*, **12**, 627-653.
https://doi.org/10.4236/am.2021.127045

Received: April 24, 2021
Accepted: July 18, 2021
Published: July 21, 2021

Abstract

Selecting which explanatory variables to include in a given score is a common difficulty, as a balance must be found between statistical fit and practical application. This article presents a methodology for constructing parsimonious event risk scores combining a stepwise selection of variables with ensemble scores obtained by aggregation of several scores, using several classifiers, bootstrap samples and various modalities of random selection of variables. Selection methods based on a probabilistic model can be used to achieve a stepwise selection for a given classifier such as logistic regression, but not directly for an ensemble classifier constructed by aggregation of several classifiers. Three selection methods are proposed in this framework, two involving a backward selection of the variables based on their coefficients in an ensemble score and the third involving a forward selection of the variables maximizing the AUC. The stepwise selection allows constructing a succession of scores, with the practitioner able to choose which score best fits his needs. These three methods are compared in an application to construct parsimonious short-term event risk scores in chronic HF patients, using as event the composite endpoint of death or hospitalization for worsening HF within 180 days of a visit. Focusing on the fastest method, four scores are constructed, yielding out-of-bag AUCs ranging from 0.81 (26 variables) to 0.76 (2 variables).

Keywords

Ensemble Score, Ensemble Methods, Scoring, Variable Selection, Heart Failure

1. Introduction

1.1. Ensemble Scores

In [1], we proposed a method for constructing an event risk-score based on the use of an ensemble method and applied the latter to construct a short-term event risk score in heart failure (HF) patients. The principle of an ensemble method is to build a collection of predictors and thereafter aggregate the predictions [2], a well-known example being the random forests method [3]. An ensemble predictor is expected to be better than each of the individual predictors, provided that 1) each single predictor is relatively good and 2) single predictors are sufficiently different from each other [4]. Our goal is now to define a methodology for constructing a parsimonious event risk score using an already-defined ensemble method. This methodology and the ensemble method defined in [1] are then used to define a short-term event risk score in chronic HF patients in order to provide an example of concrete application (Section 1.3).

1.2. Selection of Variables

The more the variables contained in a model, the more complicated its use in particular in clinical practice. Therefore, a balance must be found between increasing the number of variables to allow for a better statistical fit and keeping this number sufficiently small to facilitate practical application. With the increased number of potential predictors in the medical field (through the use of "big data" from both electronic medical records and the increasing number of available biomarkers), the need for the statistical selection of variables also increases, particularly if the goal is to continue building parsimonious and effective models. For HF, variables can be selected using a literature review in order to assess which variables are the most clinically relevant [5] [6]. This often constitutes a preliminary step before using various methods of statistical selection. Among the statistical methods, a simple method is to retain only the significant variables derived from univariate analyses [7] [8] [9] or from a full multivariate model [10] [11]. Slightly more elaborate methods such as stepwise selection can also be used [12] [13]. Finally, certain studies select variables with more complex methods, using bootstrapping [14], random forests and decision trees [15] [16] or other selection methods [17].

Since the primary goal in the present study is to construct a score using an already-defined ensemble method, some of the above selection methods are not applicable in this setting. For example, the likelihood ratio test based on a probabilistic model can be used to achieve a stepwise selection for a given classifier such as logistic regression, but not directly for a classifier constructed by aggre-

gation of several classifiers. Other selection criteria must therefore be defined in this framework. Given this context, this article presents in Section 2 a methodology for constructing parsimonious event scores combining a stepwise selection of variables and the use of ensemble scores. In particular, we define herein three methods, two of which involve a backward selection based on the variables' coefficients in an ensemble score, and the third involving the combination of a forward selection using the area under the ROC curve (AUC) as criterion and an ensemble score. Due to the stepwise selection, a succession of scores is constructed which allows the user to choose which of the latter yields the best balance between performance and the number of variables.

1.3. Application to Chronic Heart Failure Patients Scoring

As a concrete illustration, these three methods of construction of parsimonious scores are compared according to AUC and processing time in an application aimed at constructing short-term event risk scores in chronic heart failure (CHF) patients. Heart failure is a global and major cause of mortality and morbidity [18] [19]. The association between HF outcomes (death, hospitalization, device implantation, transplantation, etc.) and a large number of variables (whether demographic, clinical, biochemical, biomarkers, etc.) has been widely highlighted in the medical literature. A common approach to usefully synthesize the information provided by this large number of predictor variables is to create a risk score aimed at predicting the probability of adverse events. Many predicting scores and models have already been published: in a recent literature review, Di Tanna *et al.* [20] identified 58 risk-prediction models for HF in 40 articles published between 2013 and 2018. Among these articles, 11 studies used logistic regressions (mostly binary and multivariate) and 22 Cox regressions (mostly multivariate and stepwise). A much larger number of these models have furthermore been published over the last three decades [21] [22] [23] [24]. Scores using other methods, such as machine learning methods, are rarer although increasingly proposed nowadays [16] [25] [26]. Some studies aiming to predict HF events have used various forms of ensemble methods without designating the latter as such, for example by constructing multiple imputed datasets, drawing bootstrap samples on each of these datasets, and subsequently building models on each sample prior to their aggregation [13] [14] [27].

In [1], Duarte *et al.* used their proposed methodology to construct a short-term event risk score in HF patients, using an ensemble method involving two classification rules (logistic regression and linear discriminant analysis), bootstrap samples as well as introducing random selections of variables in the construction of predictors. We used herein this methodology and constructed parsimonious scores using the methods defined in Section 2. The application for short-term predictions in CHF patients is presented in Section 3, and a discussion in Section 4.

2. Methodology

In this section, a methodology for constructing parsimonious event risk scores

combining a stepwise selection of variables with ensemble scores is presented. Each method consists of two phases, first a preselection of variables per classifier, second a stepwise construction of ensemble scores.

2.1. Preliminary Exclusion of Variables

Univariate tests (Wilcoxon test for continuous variables and Fisher's exact test for categorical variables) are first used to test the association between the response variable and each explanatory variable. Variables with a p-value greater than 0.2 are excluded.

2.2. Construction of an Ensemble Score for a Binary Outcome

The methodology detailed in Duarte *et al.* [1] is adapted to construct the scores. Basically, an ensemble method is used, where several models are built using various classification methods, different samples and different variable selections, and are subsequently aggregated in a unique score by weighted averaging. This method can be described in seven phases, as follows:

1) n_1 classifiers are chosen.

2) n_2 bootstrap samples are drawn from the working sample. Each bootstrap sample is used n_1 times (each sample is used by each classifier).

3) n_3 modalities of random selection of variables are chosen, "modality" representing a means to select the variables.

4) $n_1 n_2 n_3$ models are built, each using a different combination of classifiers, bootstrap samples and modalities of selection of variables.

5) A first aggregation by classifiers is performed. The coefficients of the models are averaged to yield n_1 intermediate scores.

6) The coefficients of the intermediate scores are normalized such that the scores themselves are between 0 and 100, using the same method as in Duarte *et al.* ([1], Subsection 4.4.2).

7) The final score is constructed by taking a convex combination of the intermediate scores maximizing the AUC OOB (AUC on out-of-bag samples).

The AUC OOB (AUC on out-of-bag samples) is computed as follows: for a given statistical unit, the scores obtained from bootstrap samples that do not include this statistical unit are aggregated to obtain an OOB prediction. By applying this method for all statistical units, the OOB predictions for the entire sample are used to compute the AUC OOB.

The search of an optimal set of coefficients of the convex combination of the intermediate scores may be achieved in a discrete subset of the set $A = \left\{ \left(\alpha_1, ..., \alpha_{n_1} \right) : \alpha_1 + ... + \alpha_{n_1} = 1 \right\}$. We used this method in the application. This search may take too much time due to the number of elements of A. Otherwise, the simplest way is to use $A_1 = \left\{ (1,0,...,0), ..., (0,0,...,1) \right\}$, thus to choose the classifier among the n_1 classifiers which maximizes the AUC OOB. Note that in Super Learner ([28] among others), the coefficients α_i are determined using cross validation and a least square regression technique.

Compared to the methodology presented in Duarte *et al.* [1] the normalization of the coefficients is carried out before rather than after the final aggregation. The latter change is made to balance the intermediate scores in the event that their raw coefficients would have different orders of magnitude.

2.3. Preselection of Variables and Construction of Parsimonious Scores

As the number p of explanatory variables after the first exclusion of variables still remains too large to create a parsimonious score, a second phase is added in order to preselect a fewer number of variables. Three different methods with an additional preselection are proposed and their results compared. In Method 1, any adapted preselection of variables can be performed for each of the n_1 classifiers and the sets of preselected variables are united in one set; then, a backward construction of scores is performed. In Method 2, a backward construction of scores is performed with a random selection of variables at each step. In Method 3, a forward preselection of variables for one of the classifiers or for each of the classifiers using the AUC in resubstitution as criterion is performed followed by a forward construction of scores using the AUC OOB as criterion.

2.3.1. Method 1

Preselection of variables: For each of the n_1 classifiers, any adapted preselection of variables can be performed. Thus, n_1 sets of preselected variables are created. The union of these n_1 sets is used as initial preselection. Let s be the number of preselected variables.

Backward construction of scores: For $i = 1, 2, \ldots, s$, at step i: an ensemble score is constructed from $j = s - i + 1$ variables (*i.e.*, for $i = 1$, $j = s$; for $i = s$, $j = 1$), using the method described in 2.2 with n_1 classifiers, n_2 bootstrap samples, n_3 modalities of random selection of variables. The variable with the lowest normalized and standardized coefficient in absolute value in this score is excluded for the step $i + 1$ (backward selection).

This allowed determining the evolution of the AUC OOB according to the number of selected variables, as well as the order of removal of the variables. Parsimonious scores with few variables can be chosen among this sequence of s scores.

2.3.2. Method 2

Preselection of variables: No initial preselection of variables is performed; all of the p explanatory variables are included.

Backward construction of scores: For $i = 1, 2, \ldots, p$, at step i: an ensemble score is constructed from $j = p - i + 1$ variables (*i.e.*, for $i = 1$, $j = p$; for $i = p$, $j = 1$), using the method described in 2.2 with n_1 classifiers, n_2 bootstrap samples, n_3 modalities of random selection of variables. The variable with the lowest normalized and standardized coefficient in absolute value in this score is excluded for the step $i + 1$.

Again, this process allows determining the evolution of the AUC OOB according to the number of selected variables, as well as the order of removal of the variables, and parsimonious scores with few variables can be chosen among this sequence of p scores.

2.3.3. Method 3

Forward preselection of variables: A forward preselection using AUC as criterion is performed for one of the classifiers or each of the classifiers. For a given classifier, let t denote a stopping time; for $i = 1, 2, \ldots, t$, at step i: $i - 1$ variables denoted $V_1, \ldots, V_{i-1}$ are available from step $i - 1$, for every set of variables $V_1, \ldots, V_{i-1}, V_j$ with $j \neq 1, \ldots, i-1$, a classification is performed on the entire sample without bootstrapping; the variable, denoted V_i, yielding the maximal AUC in resubstitution is included, provided that the AUC significantly increases using DeLong's test; otherwise, the inclusion of variables is stopped.

Note that the AUC can be computed as long as there is a prediction for each statistical unit, without assumption on the manner with which this prediction was obtained.

Forward construction of scores: For each classifier, for $i = 1, 2, \ldots, t$, at step i: an intermediate score using the i preselected variables for this classifier, is constructed, using n_2 bootstrap samples (the same for all of the classifiers) and n_3 modalities of random selection of variables. The n_1 intermediate scores using the same number of preselected variables are aggregated in a final score by combining their predictions for each statistical unit as described in 2.2.

2.4. Comparison Criteria between the Methods

The area under the ROC curve (AUC) for the out-of-bag (OOB) estimations is used as internal validation and as the main criterion to compare the different scores. Several AUC OOB are studied: the AUC OOB for the intermediate scores and, mainly, the AUC OOB for the global score. Sensitivity (Se) and specificity (Sp) corresponding to the highest Youden index (Se + Sp − 1), as well as the number of selected variables and processing time, are also taken into account.

3. Application for Short-Term Predictions in CHF Patients

3.1. Choices Made for the Construction of the Ensemble Scores

Herein, two classifiers (n_1 = 2), linear discriminant analysis (LDA), which is equivalent to linear regression on binary outcomes, and logistic regression (LR) were chosen to construct the ensemble scores. The number of bootstrap samples was, n_2 = 1000. Two modalities of random selection of variables were chosen, n_3 = 2: namely, one modality consisted in randomly drawing a defined number of variables; the other in randomly drawing a defined number of groups of related variables (correlated or linked by construction) and, for each selected group, randomly draw one variable. The groups of related variables used in the application are shown in the Supplementary Material A.1.

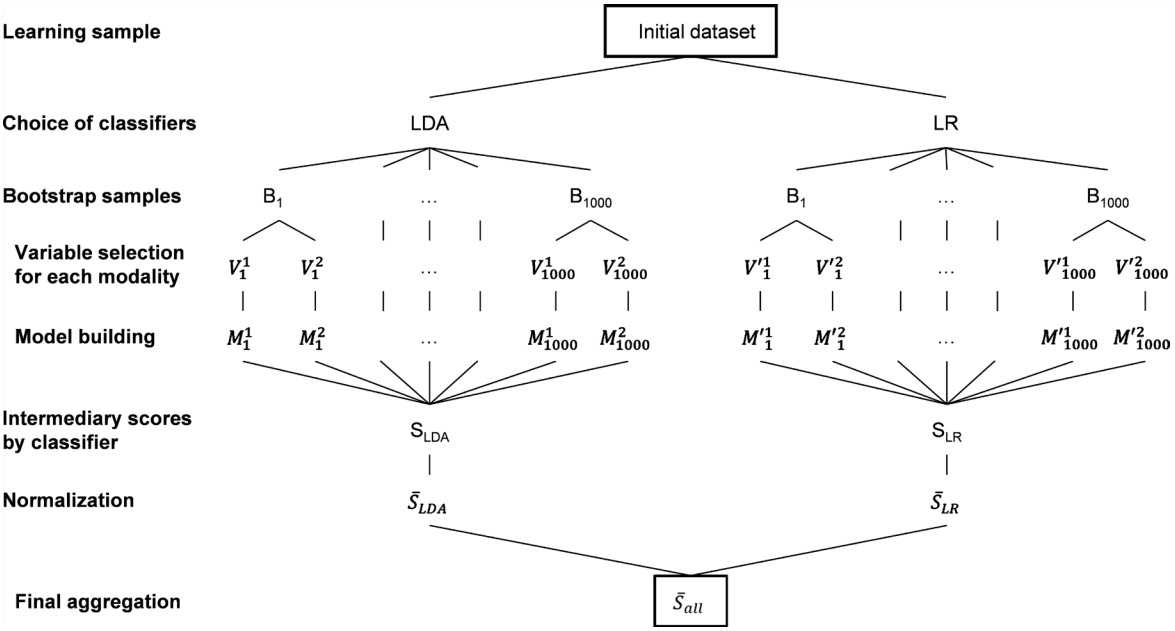

Figure 1. Methodology of construction of the ensemble score.

The score constructed for linear discriminant analysis is denoted S_{LDA} and the one for logistic regression S_{LDA}. The two normalized scores are denoted $\overline{S}_{LDA}$ and $\overline{S}_{LR}$, and the final score $\overline{S} = \lambda \overline{S}_{LDA} + (1-\lambda)\overline{S}_{LR}$ $(0 \le \lambda \le 1)$ (see Figure 1). The optimal value of λ was determined by testing values from 0 to 1 using incremental 0.01 steps and selecting the value maximizing the AUC OOB.

3.2. Choices Made for the Preselection of Variables

For Method 1, a stepwise preselection using the Akaike Information Criterion (AIC) was performed on the working sample, without bootstrapping, both for LDA and for LR. Note that herein, the AIC can be used as criterion since both LDA and LR are probabilistic models.

For Method 3, the results presented used a forward preselection with LR (Method 3a). Results obtained using a forward preselection using both LR and LDA (Method 3b) or using only LDA (Method 3c) are available as Supplementary Material (Part B).

3.3. Description of the Data

3.3.1. Description of the Original Data

The data used in this study are derived from the GISSI-HF trial: a multicenter, randomized, double-blind, placebo-controlled trial designed to assess the effect of n − 3 polyunsaturated fatty acids in patients with CHF. The detailed protocol and main results of this trial have already been described elsewhere [29] [30].

Eligible patients were adult men and women with clinical evidence of HF of any cause, with a New York Heart Association (NYHA) class II-IV, and having had a left ventricular ejection fraction (LVEF) measured within 3 months prior to enrolment. Patients with a LVEF greater than 40% had to have been admitted

at least once to hospital for HF in the preceding year to meet the inclusion criteria. In addition to contraindications linked to the studied treatment, exclusion criteria included acute coronary syndrome or revascularization procedure within the preceding 1 month; and planned cardiac surgery expected to be performed within 3 months after randomization.

After randomization and the baseline visit, patients underwent scheduled visits at 1, 3, 6, 12 months and every 6 months thereafter until the end of the trial. Data collected at baseline included patient description, medical history, etiology of HF, LVEF measurements, electrocardiogram data, clinical and cardiovascular examination, blood chemistry tests, pharmacological treatments and dietary habits. During the follow-up visits, collected data consisted of patient description, clinical and cardiovascular examination, LVEF measurement, electrocardiogram data, blood chemistry tests (only at 1, 3, 6, 12, 24, 36 and 48 months), pharmacological treatment (including the study treatment) and dietary habits. Events of interest were also recorded. The entire GISSI-HF trial included 7046 eligible and randomized patients, with the final sample analyzed in [30] and comprised of 6975 patients.

The present study used a subsample of the GISSI-HF data containing 1231 patients with N-terminal prohormone brain natriuretic peptide (NT-proBNP) measurements. The dataset included baseline and follow-up visits for these patients, as well as their associated health events.

(Patient, visit) couples were used herein as statistical units, *i.e.* each observation was associated to a patient for a given visit. We assumed that the short-term future of a patient was only dependent on the most recent measurements. Thus, the links between several couples pertaining to the same patient were not taken into account, as in [1] [31]. This yields an initial sample of 12,882 (patient, visit) couples.

3.3.2. Variables Pre-Processing

Several variables were derived from the available data, either for the follow-up visits (when values were available at baseline but not for the follow-up) or for all visits: mean blood pressure (BP) (1/3 * systolic BP + 2/3 * diastolic BP); estimated plasma volume (ePVS) ((100-hematocrit)/hemoglobin as defined in [32]); estimated glomerular filtration rate (eGFR) (using the MDRD formula [33]); age and body mass index (BMI). Binary variables for the therapeutic classes of drugs were also derived from detailed information pertaining to pharmacological treatments in order to indicate the consumption of ACE-inhibitors, beta-blockers, calcium antagonists or diuretics.

Categorical variables were recoded as binary dummy variables. In particular, in the case of ordinal variables (*i.e.* NYHA class and peripheral edema), an ordinal encoding was used, namely constructing the binary variables NYHA ≥ II, NYHA ≥ III and NYHA ≥ IV and, similarly, peripheral edema ≥ ankles, peripheral edema ≥ knee, peripheral edema ≥ above.

Since some variables were only available at baseline but were unlikely to

change over time (e.g. sex), their values were copied for follow-up visits. Similarly, certain medical history variables available at baseline (such as previous acute myocardial infarction (AMI), previous stroke, angina pectoris, coronary artery bypass graft (CABG), previous hospitalization for worsening HF) were copied for follow-up visits and, when possible, updated using the information from the events.

NT-proBNP values were only measured at baseline and at the 3-months follow-up. Due to the importance of this variable in the literature [17] [34] [35] [36], it was decided to retain and interpolate its value for the other visits as follows: the value for the 1-month follow-up visit was computed as 2/3 * (baseline value) + 1/3 * (3-months value). Value of the 3-months visit was copied for the subsequent visits.

Lastly, the response variable was defined as the occurrence of a composite event (death for worsening HF or hospitalization for worsening HF) within 180 days of a visit.

3.3.3. Exclusion of Variables and Observations

Since the laboratory tests for measuring blood parameters were performed only at baseline, 1, 3, 6, 12, 24, 36 and 48 months, only the observations corresponding to these visits were retained. Incomplete observations (with missing values) were also excluded.

Several variables not relevant to this study were excluded (e.g. "technical variables", such as identification numbers or dates, or "intermediary variables" used to build other variables, such as the cause of death or drug doses), as well as variables with more than 1000 missing values. The remaining variables and the groups of related variables are shown in the Supplementary Material A.1.

Six binary variables with univariate p-value greater than 0.2 (Fisher's exact test) were excluded: gender being "female", main cause of HF being "hypertension" or "other", history of coronary angioplasty, left ventricular hypertrophy, pathological Q waves.

3.4. Winsorization and Transformation of the Variables

In order to eliminate outliers without excluding the associated observations, all continuous variables were winsorized: all values lower than the 1st percentile (respectively greater than the 99th percentile) were set to the value of the 1st percentile (resp. the 99th percentile). This method was used to avoid excluding more observations, since the number of cases was already small compared to the controls and to avoid reducing the number of patients with event.

Continuous variables were then transformed to satisfy the linearity assumption of logistic regression. For each continuous variable, a similar method to that described in Duarte *et al.* [1] was used. First, the restricted cubic splines method with 3 knots was used to test the linearity assumption for each variable under the univariate logistic model: using a likelihood ratio test, the nullity of the coefficient associated with the cubic component of the spline was tested [14] [17].

Then, for each variable with a significantly non-null coefficient with a 5% threshold, a graphical representation of the links between the variable and the logit was performed. If the relationship was monotonous, simple monotonic transformations of the form $f(x)=x^a$ with $a \in \{-2,-1,-1/2,1/2,1,2\}$ or $f(x)=\ln(x)$ were tested. If the relationship was not monotonous, quadratic transformations of the form $f(x)=(x-k)^2$ were tested, with k situated between the minimum and the maximum of the variable by incremental 0.1 steps. To determine the values of a or k, all possible values were tested and the transformation which yielded a non-significant p-value for the linearity test and a minimal p-value for the test of nullity of the coefficient in univariate logistic regression was retained. Eleven of 23 continuous variables had a significantly non-null coefficient associated with the cubic component of the restricted cubic spline when tested. Among these eleven variables, six (mean blood pressure, eGFR, triglycerides, cholesterol HDL, total cholesterol and NT-proBNP) had a monotonic relationship with the logit. All except eGFR and NT-proBNP had x^{-2} for optimal transformation, while the optimal transformation for eGFR and NT-proBNP was $1/x$ and $\ln(x)$ respectively. The remaining five variables (BMI, systolic blood pressure, hematocrit, uricemia and LVEF) had a quadratic relationship with the logit and were transformed accordingly. After the transformation, the coefficient associated with the cubic component of the spline was non-significantly different from 0 for each of the transformed variables.

This transformed dataset was used for Methods 1, 2 and the LR intermediate score of Method 3a. A similar technique was used on a duplicate dataset for the LDA intermediate score of Method 3a, but with transformation of the variables in order to satisfy the linearity assumption for linear regression; fifteen variables were transformed: ten were transformed using a quadratic $(x-k)^2$ transformation (BMI, systolic blood pressure, diastolic blood pressure, mean blood pressure, hematocrit, hemoglobin, ePVS, serum sodium, uricemia, total cholesterol and LVEF); three using an inverse square x^{-2} transformation (eGFR, triglycerides, cholesterol HDL); one using a square transformation (serum creatinine); and one using a square root transformation (NT-proBNP).

The p-values of the tests, before and after transformation, as well as the transformation functions applied to the variables both for the LR and for the LDA are available as Supplementary Material (Part C).

3.5. Working Sample

Given the large imbalance between cases and controls, the sample was balanced by duplicating each case 15 times. This is equivalent to giving each case fifteen times more weight than a control. Preliminary analyses (not shown) showed that using a sample that was rebalanced in this manner resulted in better performance compared to using the unbalanced sample.

After the exclusions, the working sample consisted in 11,411 observations of 62 explanatory variables, with 5595 (duplicated) events and 5816 non-events.

Summary statistics of the sample prior to data management (winsorization, transformation of the variables and sample balancing) are available in Supplementary Material A.2. Summary statistics of the sample after winsorization and sample balancing, but before the transformation of the variables, are provided in Supplementary Material A.1.

3.6. Results for the Preselections of Variables by the Three Methods

The detailed preselections with their corresponding AUC are given in Table 1. The numbers of variables needed to obtain a given AUC OOB for each of the three methods are provided in Table 2.

For Method 1, 50 variables were preselected during the stepwise selection phase, after which the maximum AUC OOB was obtained for the score using 49 variables. The total runtime for the first method was approximately 1h30 (5 min for the two stepwise preselections and 1h25 for the backward selection using scores).

Comparatively, for Method 2, the maximum AUC OOB corresponded to the score using 58 variables. The total runtime of the second method was approximately 1h35 (exclusively for the backward selection using scores).

For Method 3a, the logistic forward preselection yielded 26 variables, mostly clinical or biological, after which the AUC no longer increased significantly. The total runtime of the third method was approximately 1h05 minutes if all the scores were constructed (less than 5 min for the preselection and 30 min for each of the successions of scores). However, unlike the other two methods, it is not mandatory to construct all of the scores with Method 3a and one could construct only one score after the preselection of variables. In this case, the total runtime would be reduced to less than 10 min (less than 5 min for the preselection and 2 - 5 min to construct one score).

Preselected variables were extremely similar between all 3 methods. For Methods 1 and 2, three variables were needed to obtain an AUC OOB greater than 0.75 (for Method 3a, only two were needed). Among these variables, two were common to all methods: NT-proBNP and NYHA ≥ III. In order to obtain an AUC OOB above 0.78, all methods necessitated eight variables, seven of which were common to the three methods: NT-proBNP, NYHA ≥ III, Glycemia, systolic blood pressure, beta-blockers, peripheral edema ≥ "above" and NYHA ≥ II. Lastly, for an AUC OOB threshold of 0.80, Methods 1 and 2 necessitated 17 variables, while Method 3a necessitated 15. In this case, 13 variables were common to the three methods: added to the six aforementioned variables were cholesterol HDL, heart rate, uricemia, third heart sound, bilirubin and paroxystic atrial fibrillation. Globally, the three selections were very similar.

For a fixed number of variables, the three methods yielded extremely similar AUC OOB, even when the selections of variables themselves were different. Since Method 3a generally yielded the best AUC OOB for a given number of selected variables and with a faster runtime, only the results for parsimonious

Table 1. Preselections of variables obtained with the three methods and corresponding AUC OOB of the associated scores.

	Method 1		Method 2		Method 3a			
	Variables	AUC OOB*	Variables	AUC OOB*	Variables	AUC OOB** (LR part)	AUC OOB** (LDA part)	AUC OOB*** (all)
1	NT-proBNP	0.7246	NT-proBNP	0.7246	NT-proBNP	0.7246	0.7246	0.7246
2	NYHA ≥ III	0.7482	NYHA ≥ III	0.7482	NYHA ≥ III	0.7482	0.7523	0.7523
3	Periph. edema ≥ "above"	0.7547	Heart rate	0.7529	NYHA ≥ II	0.7550	0.7579	0.7579
4	Glycemia	0.7620	Systolic BP	0.7591	Glycemia	0.7621	0.7642	0.7642
5	Systolic BP	0.7671	NYHA ≥ II	0.7647	Periph. edema ≥ "above"	0.7687	0.7688	0.7694
6	Beta-blockers	0.7730	Beta-blockers	0.7696	Beta-blockers	0.7731	0.7714	0.7736
7	NYHA ≥ II	0.7787	Glycemia	0.7764	Systolic BP	0.7791	0.7761	0.7792
8	Cholesterol HDL	0.7829	Periph. edema ≥ "above"	0.7810	Cholesterol HDL	0.7835	0.7796	0.7835
9	Mean BP	0.7827	Cholesterol HDL	0.7852	Paroxystic AF	0.7864	0.7831	0.7867
10	Diastolic BP	0.7840	Uricemia	0.7885	Uricemia	0.7902	0.7866	0.7904
11	Heart rate	0.7861	Bilirubin	0.7912	Bilirubin	0.7925	0.7876	0.7926
12	Uricemia	0.7897	Diuretics	0.7913	Implantable defibrillator	0.7948	0.7908	0.7950
13	Third heart sound	0.7922	Previous AMI	0.7932	Neoplasia	0.7966	0.7924	0.7968
14	Bilirubin	0.7950	Paroxystic AF	0.7953	Third heart sound	0.7984	0.7947	0.7985
15	Previous AMI	0.7967	Third heart sound	0.7982	Heart rate	0.8001	0.7963	0.8002
16	Paroxystic AF	0.7988	LVEF	0.7990	Previous AMI	0.8020	0.7977	0.8020
17	Implantable defibrillator	0.8010	Triglycerides	0.8006	Triglycerides	0.8038	0.7993	0.8038
18	Neoplasia	0.8027	Neoplasia	0.8028	LVEF	0.8052	0.8010	0.8052
19	LVEF	0.8045	Ascitis	0.8038	Hypertension	0.8067	0.8021	0.8067
20	Triglycerides	0.8064	Implantable defibrillator	0.8060	Mitral insufficiency	0.8080	0.8040	0.8080
21	Diuretics	0.8070	Hemoglobin	0.8058	Smoker or ex-smoker	0.8091	0.8053	0.8091
22	Ascitis	0.8085	ePVS	0.8061	Ascitis	0.8104	0.8060	0.8104
23	Mid-apical pulmonary rales	0.8091	Hematocrit	0.8070	Periph. edema ≥ "ankles"	0.8116	0.8069	0.8116
24	Smoker or ex-smoker	0.8099	Smoker or ex-smoker	0.8080	NYHA ≥ IV	0.8119	0.8071	0.8119
25	Mitral insufficiency	0.8108	Mitral insufficiency	0.8086	BMI	0.8130	0.8084	0.8130
26	Hypertension	0.8121	BMI	0.8103	Mid-apical pulmonary rales	0.8137	0.8084	0.8137
27	BMI	0.8131	Hypertension	0.8119				
28	Periph. edema ≥ "ankles"	0.8144	Previous hosp. for worsening HF	0.8119				
29	Periph. edema ≥ "knee"	0.8151	Mid-apical pulmonary rales	0.8127				
30	CABG	0.8157	Diabetes	0.8127				
31	Calcium antagonists	0.8161	CABG	0.8133				
32	Previous hosp. for worsening HF	0.8166	Periph. edema ≥ "knee"	0.8136				

Continued

33	Bundle branch block	0.8170	Diastolic BP	0.8137
34	NYHA ≥ IV	0.8176	NYHA ≥ IV	0.8145
35	Serum sodium	0.8178	Bundle branch block	0.8147
36	Diabetes	0.8180	Calcium antagonists	0.8153
37	COPD	0.8181	Total cholesterol	0.8149
38	Previous stroke	0.8184	Mean BP	0.8151
39	Years of school education	0.8187	COPD	0.8150
40	Age	0.8186	Periph. edema ≥ "ankles"	0.8163
41	Weight	0.8186	Atrial fibrillation	0.8166
42	Serum creatinine	0.8185	Cause of HF = "not known"	0.8165
43	eGFR	0.8186	Previous stroke	0.8167
44	Total cholesterol	0.8184	Aortic stenosis	0.8167
45	Aortic stenosis	0.8185	Age	0.8164
46	Cause of HF = "not known"	0.8186	Angina pectoris	0.8164
47	Atrial fibrillation	0.8187	Years of school education	0.8166
48	Pulmonary rales	0.8186	Waiting for cardiac transplantation	0.8168
49	Basal pulmonary rales	0.8188	Serum sodium	0.8170
50	Transient ischemic attack	0.8187	Definitive pace maker	0.8171
51			Basal pulmonary rales	0.8168
52			Weight	0.8169
53			eGFR	0.8169
54			Transient ischemic attack	0.8170
55			Hepatomegaly	0.8165
56			Pulmonary rales	0.8168
57			ECG evaluation	0.8167
58			ACE-inhibitors	0.8172
59			Serum creatinine	0.8168
60			Serum potassium	0.8170
61			CVP > 6 cm H20	0.8170
62			Cause of HF = "cardiomyopathy"	0.8167

*AUC OOB obtained for the score including the variable in the row as well as all previous variables. **The AUC OOB of these columns were obtained by building an intermediate score using only LDA (respectively LR) for the linear part (resp. logistic part) from the selected variables. ***The AUC OOB of this column was obtained by constructing a full ensemble score with the same number of variables for both LDA and LR, using the optimal λ for each score. ACE: angiotensin-converting enzyme; AF: atrial fibrillation; AMI: acute myocardial infarction; AUC OOB: area under the ROC curve out-of-bag; BMI: body mass index; BP: blood pressure CABG: coronary artery bypass graft; COPD: chronic obstructive pulmonary disease; CVP: central venous pressure; eGFR: estimated glomerular filtration rate; ePVS: estimated plasma volume; HDL: high-density lipoprotein; HF: heart failure; LVEF: left ventricular ejection fraction; NT-proBNP: N-terminal prohormone brain natriuretic peptide; NYHA: New York Heart Association.

scores constructed by this method are given at the end of this section.

3.7. Results for Parsimonious Scores Constructed by Method 3a

Four scores constructed by Method 3a were particularly studied: the score including all variables selected by the forward preselection, denoted S3.26 (the number of the method and the number of variables used), and three "parsimonious" scores, denoted S3.15, S3.8 and S3.2, which yielded an AUC OOB above certain thresholds (0.80, 0.78 and 0.75). To attain these thresholds, 15, 8 and 2 variables were respectively needed. The AUC OOB with $\lambda = 0.5$ and the optimal λ, as well as the optimal sensitivity and specificity according to the maximum Youden index of these four scores are given in Table 3.

Score S3.2 had an AUC OOB of 0.7523 with an optimal $\lambda = 1$ (*i.e.* only LDA

Table 2. Number of variables needed to obtain an AUC above given thresholds.

AUC OOB	Method 1	Method 2	Method 3a	Number of variables common to all methods
≥0.750	3	3	2	2
≥0.760	4	5	4	2
≥0.770	6	7	6	4
≥0.780	8	8	8	7
≥0.790	13	11	10	9
≥0.800	17	17	15	13
≥0.810	25	26	22	21

Note: even if the methods necessitated the same number of variables to obtain a given AUC, the variables themselves may not be the same.

Table 3. Summary of the characteristics of the parsimonious scores constructed using Method 3a.

Score designation	S3.26		S3.15		S3.8		S3.2	
Data	Working sample defined in Section 3.3. Variables transformed differently for the linear intermediate score and the logistic intermediate score.							
Number of bootstrap samples	1000							
Number of variables used	26		15		8		2	
Number of modalities	2							
λ value	$\lambda = 0.5$	$\lambda = 0$ (optimal)	$\lambda = 0.5$	$\lambda = 0.09$ (optimal)	$\lambda = 0.5$	$\lambda = 0.06$ (optimal)	$\lambda = 0.5$	$\lambda = 1$ (optimal)
AUC OOB of the LDA	0.8084		0.7963		0.7796		0.7523	
AUC OOB of the LR	0.8137		0.8001		0.7835		0.7482	
AUC OOB of the final score	0.8121	0.8137	0.7996	0.8002	0.7830	0.7835	0.7502	0.7523
Sensitivity*	0.861	0.823	0.759	0.724	0.713	0.748	0.810	0.826
Specificity*	0.611	0.651	0.689	0.719	0.707	0.675	0.551	0.547
Maximum Youden index	0.472	0.474	0.448	0.443	0.420	0.423	0.361	0.373

*Sensitivity and specificity associated with the maximum value of the Youden index.

was used). Score S3.8 had an AUC OOB of 0.7835 with an optimal $\lambda = 0.06$. Score S3.15 had an AUC OOB of 0.8001 with an optimal $\lambda = 0.09$. Finally, the full score including all preselected variables had an AUC OOB of 0.8137 with an optimal $\lambda = 0$ (*i.e.* only LR was used). It is interesting to note that for score S3.2, only LDA was used while for score S3.26 only LR was used. Thus, both classifiers are useful.

4. Discussion

4.1. Methodological Discussion

In this article, we presented and compared different methods of construction of parsimonious ensemble scores, with the construction of short-term event scores for CHF as a concrete illustration. Parsimonious scores were obtained by combining stepwise selections of variables and the use of an ensemble score. Since classic criteria of stepwise selection based on probabilistic models cannot be used in the case of an ensemble score, we proposed using a criterion based on the absolute values of the coefficients of variables in an ensemble score and a second criterion based on the AUC.

An advantage of a stepwise selection of predictors is that it allows automatically building a succession of scores and therefore choosing which of the latter has the best balance between performance and the number of variables, according to the desired quality objectives. Once this choice is made, the selected score can be used as a "classic" score. The use of an ensemble method to construct this score also provides confidence in the stability and performance of the results. Indeed, ensemble methods generally yield better results than a single predictor, provided that the predictors constituting the ensemble perform sufficiently well individually and are sufficiently different from each other [4]. The downside is that since the method relies on estimating a large number of models before their aggregation, this approach takes longer than estimating a single model. However, in the present context, it is only necessary to perform this procedure once to obtain the selection of variables and their associated coefficients, after which a simple linear combination is sufficient to obtain the score for any new observation.

Other selection methods could have been tested, for example by building all possible ensemble scores at each step with one more variable than in the previous step, keeping only the variable yielding the largest increase in AUC OOB. However, this would have entailed a lengthy processing time due to the large number of ensemble scores to construct and preliminary results (not shown) conclude that they would not have yielded a better performance than the presented methods. In the application, variants of Method 3 could also be used, e.g. preselecting variables using LDA as opposed to LR. Summarized results for these alternative methods are presented in the Supplementary Material.

4.2. Application Discussion

Regarding the variables used, when applying our method to the construction of a

short-term score in patients with CHF, the most predictive variable was systematically NT-proBNP, which is a well-known predictor of HF [10] [17] [20] [34] [35] [36]. Other explanatory variables, such as NYHA class, systolic blood pressure, LVEF, BMI, beta-blocker medication, uricemia, atrial fibrillation, heart rate or smoking status, have also often been selected in other studies [11] [13] [17] [20] [22] [23] [36]. Note that in a previous study on the same 1231 patients from the GISSI-HF trial with NT-proBNP, Barlera *et al.* [10] constructed a mortality predictive score using a Cox model and 14 variables: NT-proBNP, hs-cTnT, NYHA class, age, COPD, systolic blood pressure, diabetes, eGFR, sex, uricemia, LVEF, hemoglobin, BMI and aortic stenosis. In the present study, certain variables used in a number of scores were included in the original set of variables but were not selected in the final scores, such as age, gender, diabetes, serum creatinine, eGFR, hemoglobin or serum sodium. Sex was not significant in univariate analysis. The remainder of these variables were not retained during the forward AUC preselection phase in Method 3a. However, it should be noted that these variables were selected in Methods 1 and 2, generally in the second half of the selection. Finally, the preselection of Method 3a also included less common variables such as glycemia, peripheral edema, cholesterol HDL, bilirubin, implantable defibrillator, neoplasia, triglycerides, mitral insufficiency, as well as history of AMI, hypertension or ascites.

All variables included in the parsimonious scores S3.15, S3.8 and S3.2 are easily available from either the patient's medical history (paroxystic atrial fibrillation, previous AMI, implantable defibrillator, neoplasia), the patient's drug consumption (beta-blockers), a clinical examination (NYHA class, peripheral edema, heart rate, blood pressure, third heart sound), or laboratory blood tests (NT-proBNP, glycemia, cholesterol HDL, bilirubin, uricemia, triglycerides).

To our knowledge, no study has presented a score for short-term (180 days) events in CHF. Therefore, comparing the performance of our scores with others in the literature is difficult. Recent existing scores were generally constructed to predict long-term events for CHF patients, often at 1 or 2 years [10] [12] [17] [36] [37] and sometimes longer [9] [13], or to predict either short- or long-term events for acute HF patient [1] [35]. For instance, regarding CHF:

- In Voors *et al.* [14], several models were compared to predict different outcomes in CHF patients. Their models using 15 or 9 variables (including NT-proBNP) to predict a composite endpoint of all-cause mortality or HF hospitalization yielded an AUC of 0.71 or 0.69 in derivation, respectively. Herein, scores S3.15 and S3.8 obtained 0.80 and 0.78 AUC OOB values, respectively.
- The AUC of score S3.8 is similar to that of the score proposed by Spinar *et al.* [36] to assess the 2-year prognosis of CHF (all-cause mortality, heart transplantation, device implantation), which yielded an AUC of 0.79 without cross-validation nor external validation for a model using 7 variables.
- The MAGGIC risk score [13], which has been shown to feature one of the best accuracies to predict 1-year mortality in CHF patients in Canepa *et al.*

[37] using 13 variables and subsequently studied on many validation cohorts, had an AUC between 0.64 and 0.74 in the studies without NT-proBNP [31] [35] [37] [38], and of 0.74 with NT-proBNP [35]. Note that the AUC for the composite endpoint of death and hospitalization, as used in the current study, is generally lower than the AUC for all-cause death only. Score S3.8 achieved an AUC OOB of 0.78 using 8 variables.

The main limitation of our application study is that only one dataset was used in our tests. However, the present work is mostly a "proof of concept" of the usefulness of the presented methods of construction of parsimonious ensemble scores.

5. Conclusion

Variables selection methods based on a probabilistic model can be used to achieve a stepwise selection for a given classifier such as logistic regression, but not directly for a classifier constructed by aggregation of several classifiers. In this article, we have proposed to construct parsimonious ensemble scores using sample balancing, several classifiers, bootstrap samples and stepwise variable selection methods in this setting. As a concrete application, we constructed a short-term event (death or hospitalization for HF at 180 days) score for CHF patients, yielding satisfactory AUC values with respect to other scores in other HF patients' populations. The methods proposed and tested in this article can be reproduced on any delay, any set of variables and any other settings (other types of HF or other diseases) as long as there is a sufficient number of cases, *i.e.* a sufficiently large training dataset. Applications on other datasets and comparisons with other methods should be conducted in order to confirm the interest of the proposed methods.

Acknowledgements

The authors thank Mr. Pierre Pothier for editing this manuscript. Results incorporated in this article received funding from the investments for the Future program, France 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] 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

[2] Hastie, T., Tibshirani, R. and Friedman, J. (2009) The Elements of Statistical Learning. 2nd Edition, Springer, Berlin. https://doi.org/10.1007/978-0-387-84858-7

[3] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/A:1010933404324

[4] Genuer, R. and Poggi, J.-M. (2017) Arbres CART et Forêts aléatoires, Importance et sélection de variables. https://hal.archives-ouvertes.fr/hal-01387654

[5] Upshaw, J.N., Konstam, M.A., van Klaveren, D., Noubary, F., Huggins, G.S. and Kent, D.M. (2016) Multistate Model to Predict Heart Failure Hospitalizations and All-Cause Mortality in Outpatients with Heart Failure with Reduced Ejection Fraction: Model Derivation and External Validation. *Circulation*: *Heart Failure*, **9**, Article ID: e003146. https://doi.org/10.1161/CIRCHEARTFAILURE.116.003146

[6] Senni, M., Parrella, P., De Maria, R., Cottini, C., Böhm, M., Ponikowski, P., *et al.* (2013) Predicting Heart Failure Outcome from Cardiac and Comorbid Conditions: The 3C-HF Score. *International Journal of Cardiology*, **163**, 206-211. https://doi.org/10.1016/j.ijcard.2011.10.071

[7] Bhandari, S.S., Narayan, H., Jones, D.J.L., Suzuki, T., Struck, J., Bergmann, A., *et al.* (2016) Plasma Growth Hormone Is a Strong Predictor of Risk at 1 Year in Acute Heart Failure. *European Journal of Heart Failure*, **18**, 281-289. https://doi.org/10.1002/ejhf.459

[8] Ramírez, J., Orini, M., Mincholé, A., Monasterio, V., Cygankiewicz, I., de Luna, A.B., *et al.* (2017) Sudden Cardiac Death and Pump Failure Death Prediction in Chronic Heart Failure by Combining ECG and Clinical Markers in an Integrated Risk Model. *PLoS ONE*, **12**, e0186152. https://doi.org/10.1371/journal.pone.0186152

[9] Xu, X.-R., Meng, X.-C., Wang, X., Hou, D.-Y., Liang, Y.-H., Zhang, Z.-Y., *et al.* (2018) A Severity Index Study of Long-Term Prognosis in Patients with Chronic Heart Failure. *Life Sciences*, **210**, 158-165. https://doi.org/10.1016/j.lfs.2018.09.005

[10] Barlera, S., Tavazzi, L., Franzosi, M.G., Marchioli, R., Raimondi, E., Masson, S., *et al.* (2013) Predictors of Mortality in 6975 Patients with Chronic Heart Failure in the Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico-Heart Failure Trial: Proposal for a Nomogram. *Circulation*: *Heart Failure*, **6**, 31-39. https://doi.org/10.1161/CIRCHEARTFAILURE.112.967828

[11] Lupón, J., de Antonio, M., Vila, J., Peñafiel, J., Galán, A., Zamora, E., *et al.* (2014) Development of a Novel Heart Failure Risk Tool: The Barcelona Bio-Heart Failure Risk Calculator (BCN Bio-HF Calculator). *PloS ONE*, **9**, e85466. https://doi.org/10.1371/journal.pone.0085466

[12] Levy, W.C., Mozaffarian, D., Linker, D.T., Sutradhar, S.C., Anker, S.D., Cropp, A.B., *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

[13] Pocock, S.J., Ariti, C.A., McMurray, J.J.V., Maggioni, A., Køber, L., Squire, I.B., *et al.* (2013) Predicting Survival in Heart Failure: A Risk Score Based on 39 372 Patients from 30 Studies. *European Heart Journal*, **34**, 1404-1413. https://doi.org/10.1093/eurheartj/ehs337

[14] Voors, A.A., Ouwerkerk, W., Zannad, F., van Veldhuisen, D.J., Samani, N.J., Ponikowski, P., *et al.* (2017) Development and Validation of Multivariable Models to Predict Mortality and Hospitalization in Patients with Heart Failure. *European Journal of Heart Failure*, **19**, 627-634. https://doi.org/10.1002/ejhf.785

[15] Krumholz, H.M., Chaudhry, S.I., Spertus, J.A., Mattera, J.A., Hodshon, B. and Herrin, J. (2016) Do Non-Clinical Factors Improve Prediction of Readmission Risk?: Results From the Tele-HF Study. *JACC*: *Heart Failure*, **4**, 12-20. https://doi.org/10.1016/j.jchf.2015.07.017

[16] Desai, R.J., Wang, S.V., Vaduganathan, M., Evers, T. and Schneeweiss, S. (2020)

Comparison of Machine Learning Methods with Traditional Models for Use of Administrative Claims with Electronic Medical Records to Predict Heart Failure Outcomes. *JAMA Network Open*, **3**, Article ID: e1918962. https://doi.org/10.1001/jamanetworkopen.2019.18962

[17] Simpson, J., Jhund, P.S., Lund, L.H., Padmanabhan, S., Claggett, B.L., Shen, L., *et al.* (2020) Prognostic Models Derived in PARADIGM-HF and Validated in ATMOSPHERE and the Swedish Heart Failure Registry to Predict Mortality and Morbidity in Chronic Heart Failure. *JAMA Cardiology*, **5**, 432-441. https://doi.org/10.1001/jamacardio.2019.5850

[18] Metra, M. and Teerlink, J.R. (2017) Heart Failure. *The Lancet Journal*, **390**, 1981-1995. https://doi.org/10.1016/S0140-6736(17)31071-1

[19] Orso, F., Fabbri, G. and Maggioni, A.P. (2017) Epidemiology of Heart Failure. In: Bauersachs, J., Butler, J. and Sandner, P., Eds., *Heart Failure*, Springer International Publishing, Cham, 15-33. https://doi.org/10.1007/164_2016_74

[20] Di Tanna, G.L., Wirtz, H., Burrows, K.L. and Globe, G. (2020) Evaluating Risk Prediction Models for Adults with Heart Failure: A Systematic Literature Review. *PloS ONE*, **15**, e0224135. https://doi.org/10.1371/journal.pone.0224135

[21] Alba, A.C., Agoritsas, T., Jankowski, M., Courvoisier, D., Walter, S.D., Guyatt, G.H., *et al.* (2013) Risk Prediction Models for Mortality in Ambulatory Patients with Heart Failure: A Systematic Review. *Circulation: Heart Failure*, **6**, 881-889. https://doi.org/10.1161/CIRCHEARTFAILURE.112.000043

[22] Ouwerkerk, W., Voors, A.A. and Zwinderman, A.H. (2014) Factors Influencing the Predictive Power of Models for Predicting Mortality and/or Heart Failure Hospitalization in Patients With Heart Failure. *JACC: Heart Failure*, **2**, 429-436. https://doi.org/10.1016/j.jchf.2014.04.006

[23] Rahimi, K., Bennett, D., Conrad, N., Williams, T.M., Basu, J., Dwight, J., *et al.* (2014) Risk Prediction in Patients with Heart Failure: A Systematic Review and Analysis. *JACC: Heart Failure*, **2**, 440-446. https://doi.org/10.1016/j.jchf.2014.04.008

[24] Ferrero, P., Iacovoni, A., D'Elia, E., Vaduganathan, M., Gavazzi, A. and Senni, M. (2015) Prognostic Scores in Heart Failure—Critical Appraisal and Practical Use. *International Journal of Cardiology*, **188**, 1-9. https://doi.org/10.1016/j.ijcard.2015.03.154

[25] Chicco, D. and Jurman, G. (2020) Machine Learning Can Predict Survival of Patients with Heart Failure from Serum Creatinine and Ejection Fraction Alone. *BMC Medical Informatics and Decision Making*, **20**, Article No. 16. https://doi.org/10.1186/s12911-020-1023-5

[26] Bazoukis, G., Stavrakis, S., Zhou, J., Bollepalli, S.C., Tse, G., Zhang, Q., *et al.* (2020) Machine Learning versus Conventional Clinical Methods in Guiding Management of Heart Failure Patients—A Systematic Review. *Heart Failure Reviews*, **26**, 23-24. https://doi.org/10.1007/s10741-020-10007-3

[27] O'Connor, C.M., Whellan, D.J., Wojdyla, D., Leifer, E., Clare, R.M., Ellis, S.J., *et al.* (2012) Factors Related to Morbidity and Mortality in Patients with Chronic Heart Failure with Systolic Dysfunction: The HF-ACTION Predictive Risk Score Model. *Circulation: Heart Failure*, **5**, 63-71. https://doi.org/10.1161/CIRCHEARTFAILURE.111.963462

[28] Polley, E. and van der Laan, M. (2010) Super Learner in Prediction. U.C. Berkeley Division of Biostatistics Working Paper Series. https://biostats.bepress.com/ucbbiostat/paper266

[29] Tavazzi, L., Tognoni, G., Franzosi, M.G., Latini, R., Maggioni, A.P., Marchioli, R., *et*

al. (2004) Rationale and Design of the GISSI Heart Failure Trial: A Large Trial to Assess the Effects of N-3 Polyunsaturated Fatty Acids and Rosuvastatin in Symptomatic Congestive Heart failure. *European Journal of Heart Failure*, **6**, 635-641. https://doi.org/10.1016/j.ejheart.2004.03.001

[30] GISSI-HF Investigators (2008) Effect of N-3 Polyunsaturated Fatty Acids in Patients with Chronic Heart Failure (the GISSI-HF Trial): A Randomised, Double-Blind, Placebo-Controlled Trial. *The Lancet Journal*, **372**, 1223-1230. https://doi.org/10.1016/S0140-6736(08)61239-8

[31] Kwon, J.-M., Kim, K.-H., Jeon, K.-H., Lee, S.E., Lee, H.-Y., Cho, H.-J., *et al*. (2019) Artificial Intelligence Algorithm for Predicting Mortality of Patients with Acute Heart Failure. *PloS ONE*, **14**, e0219302. https://doi.org/10.1371/journal.pone.0219302

[32] 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

[33] Levey, A., Bosch, J., Lewis, J., Greene, T., Rogers, N. and Roth, D. (1999) A More Accurate Method to Estimate Glomerular Filtration Rate from Serum Creatinine: A New Prediction Equation. *Annals of Internal Medicine*, **130**, 461-470. https://doi.org/10.7326/0003-4819-130-6-199903160-00002

[34] Braunwald, E. (2008) Biomarkers in Heart Failure. *The New England Journal of Medicine*, **358**, 2148-2159. https://doi.org/10.1056/NEJMra0800239

[35] Khanam, S.S., Choi, E., Son, J.-W., Lee, J.-W., Youn, Y.J., Yoon, J., *et al*. (2018) Validation of the MAGGIC (Meta-Analysis Global Group in Chronic Heart Failure) Heart Failure Risk Score and the Effect of Adding Natriuretic Peptide for Predicting Mortality after Discharge in Hospitalized Patients with Heart Failure. *PLoS ONE*, **13**, e0206380. https://doi.org/10.1371/journal.pone.0206380

[36] Spinar, J., Spinarova, L., Malek, F., Ludka, O., Krejci, J., Ostadal, P., *et al*. (2019) Prognostic Value of NT-proBNP Added to Clinical Parameters to Predict Two-Year Prognosis of Chronic Heart Failure Patients with Mid-Range and Reduced Ejection Fraction—A Report from FAR NHL Prospective Registry. *PloS ONE*, **14**, e0214363. https://doi.org/10.1371/journal.pone.0214363

[37] Canepa, M., Fonseca, C., Chioncel, O., Laroche, C., Crespo-Leiro, M.G., Coats, A.J.S., *et al*. (2018) Performance of Prognostic Risk Scores in Chronic Heart Failure Patients Enrolled in the European Society of Cardiology Heart Failure Long-Term Registry. *JACC*: *Heart Failure*, **6**, 452-462. https://doi.org/10.1016/j.jchf.2018.02.001

[38] Rich, J.D., Burns, J., Freed, B.H., Maurer, M.S., Burkhoff, D. and Shah, S.J. (2018) Meta-Analysis Global Group in Chronic (MAGGIC) Heart Failure Risk Score: Validation of a Simple Tool for the Prediction of Morbidity and Mortality in Heart Failure with Preserved Ejection Fraction. *Journal of the American Heart Association*, **7**, Article ID: e009594. https://doi.org/10.1161/JAHA.118.009594

Supplementary Material for "Construction of Parsimonious Event Risk Scores by an Ensemble Method. An Illustration for Short-Term Predictions in Chronic Heart Failure Patients from the GISSI-HF Trial"

A1. Summary Statistics of the Sample before Data Management

Table A1. Descriptive statistics of the explanatory variables after winsorization and sample balancing performed before transformation of the variables.

Variables		Groups of related variables	Mean (SD) or N (%)
Female[b,d]		-	2227 (19.5%)
Age[a,g]		-	68.10 (10.20)
Years of school education[d,g]		-	6.92 (3.65)
Weight[g]		*Obesity*	75.87 (14.33)
BMI[a,g]			26.96 (4.48)
Smoker or ex-smoker[b,d]		-	6645 (58.2%)
Heart Rate[g]		-	72.49 (13.38)
Diastolic blood pressure[g]		*Blood pressure*	76.28 (10.17)
Systolic blood pressure[g]			125.21 (19.41)
Mean blood pressure[a,g]			92.58 (12.17)
NYHA class[c] (ref: "NYHA I")	≥*II*	*NYHA*	10837 (95.0%)
	≥*III*		3061 (26.8%)
	≥*IV*		242 (2.1%)
Peripheral edema[c,d] (ref: "No")	≥*Ankles*	*Peripheral edema*	1768 (15.5%)
	≥*Knee*		316 (2.8%)
	≥*Above*		159 (1.4%)
Main cause of HF [b] (ref: "Ischemic")	*Cardiomyopathy*	-	3126 (27.4%)
	Hypertension	-	1726 (15.3%)
	Other	-	346 (3.0%)
	Not known	-	175 (1.5%)
Ascites[b,d]		-	147 (1.3%)
Hepatomegaly[b,d]		-	2188 (19.2%)
Mitral insufficiency[b,d]		-	5461 (47.9%)
CVP > 6 cm H20[b,d]		-	1139 (10.0%)
Basal pulmonary rales[b,d]		-	1732 (15.2%)
Mid-apical pulmonary rales[b,d]		-	79 (0.7%)
Pulmonary rales[b,d]		-	599 (5.2%)
Aortic stenosis[b,d]		-	315 (2.8%)
Third heart sound (S_3)[b,d]		-	2177 (19.1%)
Hematocrit[g]		*Hematology*	40.16 (4.53)
Hemoglobin[g]			13.40 (1.60)
ePVS[a,g]			4.57 (0.92)

Continued

Serum creatinine[g]	*Renal function*	1.27 (0.44)
eGFR[a,g,h]		64.08 (22.63)
Serum potassium[g]	-	4.48 (0.50)
Serum sodium[g]	-	139.49 (3.33)
Uricemia[g]	-	6.43 (1.94)
Triglycerides[g]	-	137.92 (84.01)
Cholesterol HDL[g]	*Cholesterol*	47.58 (13.19)
Total Cholesterol[g]		175.10 (44.48)
Bilirubin[g]	-	0.84 (0.42)
Glycemia[g]	-	122.98 (46.60)
NT-proBNP[f,g]	-	1856.60 (2194.91)
Diabetes mellitus[b,d]	-	3481 (30.5%)
Hypertension[b,d]	-	6470 (56.7%)
Previous AMI[b,e]	-	5421 (47.5%)
Previous stroke[b,e]	-	643 (5.6%)
Previous hosp. for worsening HF[b,e]	-	6526 (57.2%)
Angina pectoris[b,e]	-	2060 (18.1%)
Coronary angioplasty[b,d]	-	1478 (13.0%)
Transient ischemic attack (TIA)[b,d]	-	1228 (10.8%)
COPD[b,d]	-	2348 (20.6%)
CABG[b,e]	-	2847 (24.9%)
Implantable defibrillator[b,d]	-	1020 (8.9%)
Paroxystic AF[b,d]	-	2756 (24.2%)
Neoplasia[b,d]	-	592 (5.2%)
Definitive pace maker[b,d]	-	1944 (17.0%)
Waiting for cardiac transplantation[b,d]	-	122 (1.1%)
LVEF[d,g]	-	32.58 (10.05)
Bundle branch block[b]	-	3883 (34.0%)
Atrial fibrillation[b]	-	2087 (18.3%)
Left ventricular hypertrophy[b]	-	1885 (16.5%)
Pathological Q waves[b]	-	2236 (19.6%)
Normal ECG evaluation[b]	-	415 (3.6%)
ACE-inhibitors[a,b]	-	8782 (77.0%)
Beta-blockers[a,b]	-	7430 (65.1%)
Calcium antagonists[a,b]	-	803 (7.0%)
Diuretics[a,b]	-	10813 (94.8%)

[a]derived variable; [b]binary variable encoding; [c]ordinal encoding; [d]baseline value copied to follow-up visits; [e]baseline value copied to follow-up visits and updated when possible; [f]interpolated values; [g]winsorized variable. SD: standard deviation; BMI: body mass index; NYHA: New York Heart Association; HF: heart failure; CVP: central venous pressure; ePVS: estimated plasma volume; eGFR: estimated glomerular filtration rate; HDL, high-density lipoprotein; AMI: acute myocardial infarction; COPD: chronic obstructive pulmonary disease; CABG: coronary artery bypass graft; AF: atrial fibrillation; LVEF: left ventricular ejection fraction; ACE: angiotensin-converting enzyme.

Table A2. Descriptive statistics of the explanatory variables prior to the data management phase.

Variables		Groups of related variables	Mean (SD) or N (%)
Female[b,d]		-	1219 (19.7%)
Age[a]		-	66.94 (10.76)
Years of school education[d]		-	7.00 (3.77)
Weight		*Obesity*	76.25 (14.76)
BMI[a]			26.96 (4.46)
Smoker or ex-smoker[b,d]		-	3425 (55.3%)
Heart rate		-	70.22 (13.25)
Diastolic blood pressure		*Blood pressure*	77.37 (10.29)
Systolic blood pressure			127.16 (18.75)
Mean blood pressure[a]			93.96 (12.07)
NYHA[c] (ref: "NYHA I")	*≥II*	*NYHA*	5671 (91.6%)
	≥III		1017 (16.4%)
	≥IV		46 (0.74%)
Peripheral edema[c,d] (ref: "No")	*≥Ankles*	*Peripheral edema*	732 (11.8%)
	≥Knee		106 (1.7%)
	≥Above		33 (0.5%)
Main cause of HF[b] (ref: "Ischemic")	*Cardiomyopathy*	-	1866 (30.2%)
	Hypertension	-	956 (15.4%)
	Other	-	178 (2.9%)
	Not known	-	133 (2.1%)
Ascites[b,d]		-	21 (0.3%)
Hepatomegaly[b,d]		-	914 (14.8%)
Mitral insufficiency[b,d]		-	2647 (42.8%)
CVP > 6 cm H20[b,d]		-	467 (7.5%)
Basal pulmonary rales[b,d]		-	738 (11.9%)
Mid-apical pulmonary rales[b,d]		-	51 (0.8%)
Pulmonary rales[b,d]		-	263 (4.3%)
Aortic stenosis[b,d]		-	105 (1.7%)
Third heart sound (S_3)[b,d]		-	945 (15.3%)
Hematocrit		*Hematology*	40.58 (4.35)
Hemoglobin			13.60 (1.56)
ePVS[a]			4.47 (0.89)
Serum creatinine		*Renal function*	1.21 (0.42)
eGFR[a]			67.59 (22.79)
Serum potassium		-	4.47 (0.50)

Continued

Serum sodium	-	139.65 (3.45)
Uricemia	-	6.39 (1.84)
Triglycerides	-	147.31 (110.90)
Cholesterol HDL	*Cholesterol*	49.15 (13.62)
Total cholesterol		180.67 (44.55)
Bilirubin	-	0.81 (0.56)
Glycemia	-	119.61 (46.53)
NT-proBNP[f]	-	1312.57 (1978.60)
Diabetes mellitus[b,d]	-	1535 (24.8%)
Hypertension[b,d]	-	3390 (54.8%)
Previous AMI[b,e]	-	2649 (42.8%)
Previous stroke[b,e]	-	293 (4.7%)
Previous hosp. for worsening HF[b,e]	-	3068 (49.6%)
Angina pectoris[b,e]	-	968 (15.6%)
Coronary angioplasty[b,d]	-	792 (12.8%)
Transient ischemic attack (TIA)[b,d]	-	500 (8.1%)
COPD[b,d]	-	1074 (17.4%)
CABG[b,e]	-	1321 (21.3%)
Implantable defibrillator[b,d]	-	488 (7.9%)
Paroxystic AF[b,d]	-	1174 (19.0%)
Neoplasia[b,d]	-	242 (3.9%)
Definitive pace maker[b,d]	-	824 (13.3%)
Waiting for cardiac transplantation[b,d]	-	38 (0.6%)
LVEF[d]	-	33.56 (9.74)
Bundle branch block[b]	-	2007 (32.4%)
Atrial fibrillation[b]	-	911 (14.7%)
Left ventricular hypertrophy[b]	-	1031 (16.7%)
Pathological Q waves[b]	-	1200 (19.4%)
Normal ECG evaluation[b]	-	289 (4.7%)
ACE-inhibitors[a,b]	-	4848 (78.3%)
Beta-blockers[a,b]	-	4434 (71.6%)
Calcium antagonists[a,b]	-	509 (8.2%)
Diuretics[a,b]	-	5689 (91.9%)

[a]derived variable; [b]binary variable encoding; [c]ordinal encoding; [d]baseline value copied to follow-up visits; [e]baseline value copied to follow-up visits and updated when possible; [f]interpolated values. SD: standard deviation; BMI: body mass index; NYHA: New York Heart Association; HF: heart failure; CVP: central venous pressure; ePVS: estimated plasma volume; eGFR: estimated glomerular filtration rate; HDL, high-density lipoprotein; AMI: acute myocardial infarction; COPD: chronic obstructive pulmonary disease; CABG: coronary artery bypass graft; AF; atrial fibrillation; LVEF: left ventricular ejection fraction; ACE: angiotensin-converting enzyme.

A2. Alternative Methods: Description and Results

A2.1. Method 3b

Preselection of variables: Two forward preselections using AUC as criterion were performed, one for logistic regression (LR) and the other for linear discriminant analysis (LDA). Let t denote a stopping time. For $i=1,2,...,t$: at step i: $i-1$ variables denoted $V_1,...,V_{i-1}$ were available from step $i-1$. For every set of variables $V_1,...,V_{i-1},V_j$ with $j\neq 1,...,i-1$, a logistic regression (respectively a linear regression) was performed on the entire sample without bootstrap. The variable, denoted V_i, yielding the maximal AUC in resubstitution was included for the step $i+1$, provided that the AUC significantly increased using DeLong's test; otherwise, the inclusion of variables was stopped ($t=i$). The preselection using logistic regression (respectively LDA) was used to build an intermediate LR score (respectively an intermediate LDA score). Note that the number of preselected variables and the preselected variables themselves may differ between the two preselections.

Note that, contrary to the preselection phase of Method 1 with AIC, there is no need in this instance for a probabilistic model. Indeed, the AUC can be computed as long as there is a prediction, without assumption on the manner with which this prediction was obtained.

Construction of intermediate scores: For each classifier, intermediate scores using only the associated selected variables were constructed, using 1000 bootstrap samples (the same for both classifiers) and two modalities of selection of variables (all variables or all groups of related variables). Since the preselection was performed separately for both classifiers, intermediate scores may not use the same variables.

Construction of final scores: The two intermediate scores were aggregated in a final score by averaging their prediction for each statistical unit. Since the intermediate scores in this method were constructed independently from each other on two different sets of variables, there were multiple ways to combine the latter. In this instance, intermediate scores using the same number of preselected variables by classifier were aggregated in a final score.

A2.2. Method 3c

Preselection of variables: A forward preselection using AUC as criterion was performed using LDA. Let t denote a stopping time. For $i=1,2,...,t$: at step i: $i-1$ variables denoted $V_1,...,V_{i-1}$ were available from step $i-1$. For every set of variables $V_1,...,V_{i-1},V_j$ with $j\neq 1,...,i-1$, a linear regression was performed on the entire sample without bootstrap. The variable, denoted V_i, yielding the maximal AUC in resubstitution was included for the step $i+1$, provided that the AUC significantly increased using DeLong's test; otherwise, the inclusion of variables was stopped ($t=i$).

Note that, contrary to the preselection phase of Method 1 with AIC, there is no need in this instance for a probabilistic model. Indeed, the AUC can be computed as long as there is a prediction for each statistical unit, without assumption on the manner with which this prediction was obtained.

Construction of intermediate scores: For each classifier, intermediate scores using only the preselected variables, with the transformations corresponding to the classifier (see Subsection 3.2.4), were built, using 1000 bootstrap samples (the same for both classifiers) and two modalities of selection of variables (all variables or all groups of related variables).

Construction of final scores: The two intermediate scores using the same number of preselected variables were aggregated in a final score by averaging their prediction for each statistical unit.

Table A3. Results.

Method 3b					Method 3c			
Variables (LR part)	AUC OOB** (LR part)	Variables (LDA part)	AUC OOB** (LDA part)	AUC OOB*** (all)	Variables	AUC OOB** (LR part)	AUC OOB** (LDA part)	AUC OOB*** (all)
NT-proBNP	0.7246	NT-proBNP	0.7246	0.7246	NT-proBNP	0.7246	0.7246	0.7246
NYHA ≥ III	0.7482	NYHA ≥ III	0.7523	0.7523	NYHA ≥ III	0.7482	0.7523	0.7523
NYHA ≥ II	0.7550	Glycemia	0.7591	0.7625	Glycemia	0.7559	0.7591	0.7592
Glycemia	0.7621	NYHA ≥ II	0.7642	0.7642	NYHA ≥ II	0.7624	0.7642	0.7642
Periph. edema ≥ "above"	0.7687	Paroxystic AF	0.7683	0.7721	Paroxystic AF	0.7669	0.7683	0.7685
Beta-blockers	0.7731	Systolic BP	0.7725	0.7801	Systolic BP	0.7721	0.7725	0.7733
Systolic BP	0.7791	Beta-blockers	0.7770	0.7818	Beta-blockers	0.7781	0.7770	0.7789
Cholesterol HDL	0.7835	Cholesterol HDL	0.7807	0.7856	Cholesterol HDL	0.7823	0.7807	0.7829
Paroxystic AF	0.7864	Uricemia	0.7839	0.7886	Uricemia	0.7860	0.7839	0.7865
Uricemia	0.7902	Third heart sound	0.7866	0.7916	Third heart sound	0.7883	0.7866	0.7888
Bilirubin	0.7925	Periph. edema ≥ "above"	0.7891	0.7935	Periph. edema ≥ "above"	0.7923	0.7891	0.7926
Implantable defibrillator	0.7948	Implantable defibrillator	0.7912	0.7956	Implantable defibrillator	0.7941	0.7912	0.7945
Neoplasia	0.7966	Neoplasia	0.7930	0.7975	Neoplasia	0.7958	0.7930	0.7961
Third heart sound	0.7984	Triglycerides	0.7949	0.7990	Triglycerides	0.7979	0.7949	0.7981
Heart rate	0.8001	Heart rate	0.7965	0.8009	Heart rate	0.7998	0.7965	0.7999
Previous AMI	0.8020	Bilirubin	0.7980	0.8026	Bilirubin	0.8019	0.7980	0.8019
Triglycerides	0.8038	Previous AMI	0.7995	0.8038	Previous AMI	0.8036	0.7995	0.8036
LVEF (baseline)	0.8052	LVEF	0.8011	0.8052	LVEF	0.8052	0.8011	0.8052
Hypertension	0.8067	Mitral insufficiency	0.8028	0.8069	Mitral insufficiency	0.8064	0.8028	0.8064
Mitral insufficiency	0.8080	Diuretics	0.8039	0.8082	Diuretics	0.8070	0.8039	0.8070
Smoker or ex-smoker	0.8091	Hypertension	0.8051	0.8094	Hypertension	0.8084	0.8051	0.8084
Ascitis	0.8104	Smoker or ex-smoker	0.8060	0.8105	Smoker or ex-smoker	0.8093	0.8060	0.8093
Periph. edema ≥ "ankles"	0.8116	Periph. edema ≥ "ankles"	0.8069	0.8117	Periph. edema ≥ "ankles"	0.8101	0.8069	0.8101
NYHA ≥ IV	0.8119	Ascitis	0.8081	0.8121	Ascitis	0.8119	0.8081	0.8119
BMI	0.8130	BMI	0.8093	0.8131	BMI	0.8131	0.8093	0.8131
Mid-apical pulm. Rales	0.8137							

*AUC OOB obtained for the score including the variable in the row as well as all previous variables. **The AUC OOB of these columns were obtained by building an intermediate score using only LDA (respectively LR) for the linear part (resp. logistic part) from the selected variables. ***The AUC OOB of these columns was obtained by building a full ensemble score with the same number of variables for both LDA and LR, using the optimal λ for each score. BMI: body mass index; NYHA: New York Heart Association; LVEF: left ventricular ejection fraction.

A3. Transformation of the Variables for the Logistic Regression and the Linear Discriminant Analysis

Table A4. P-values of the linearity tests before and after transformation for LR and LDA.

Variable	*For logistic regression*			*For linear discriminant analysis*		
	p-value before	*Transformation*	*p-value after*	*p-value before*	*Transformation*	*p-value after*
Age	0.364			0.853		
Years of school education	0.449			0.462		
Weight	0.280			0.267		
BMI	0.006	$(x-27.8)^2$	0.051	0.004	$(x-28.0)^2$	0.059
Heart rate	0.149			0.806		
Diastolic blood pressure	0.704			0.291		
Systolic blood pressure	<0.001	$(x-142.0)^2$	0.756	<0.001	$(x-142.0)^2$	0.133
Mean blood pressure	0.028	x^{-2}	0.516	0.003	$(x-108.7)^2$	0.707
Hematocrit	0.001	$(x-43.4)^2$	0.051	<0.001	$(x-43.4)^2$	0.220
Hemoglobin	0.068			0.005	$(x-15.3)^2$	0.222
ePVS	0.242			0.034	$(x-3.3)^2$	0.300
Serum creatinine	0.376			0.007	x^2	0.352
eGFR	0.004	x^{-1}	0.648	<0.001	x^{-2}	0.487
Serum potassium	0.067			0.056		
Serum sodium	0.055			0.031	$(x-142.3)^2$	0.455
Uricemia	<0.001	$(x-6.7)^2$	0.383	<0.001	$(x-6.7)^2$	0.975
Triglycerides	0.023	x^{-2}	0.672	0.009	x^{-2}	0.220
Cholesterol HDL	0.009	x^{-2}	0.819	0.001	x^{-2}	0.404
Total cholesterol	0.011	x^{-2}	0.230	0.001	$(x-192.5)^2$	0.051
Bilirubin	0.210			0.800		
Glycemia	0.924			0.609		
NT-proBNP	<0.001	$\mathrm{Ln}(x)$	0.407	<0.001	$x^{0.5}$	
LVEF	<0.001	$(x-42.8)^2$	0.228	<0.001	$(x-42.7)^2$	0.959

See next tables for abbreviations.