ELSEVIER

Journal of Clinical Epidemiology 66 (2013) 398−407

Journal of Clinical Epidemiology

ORIGINAL ARTICLES

Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes

Peter C. Austin[a,b,c,*], Jack V. Tu[a,b,d], Jennifer E. Ho[e,f,g], Daniel Levy[e,f,h], Douglas S. Lee[a,b,i]

[a]*Institute for Clinical Evaluative Sciences, G105, 2075 Bayview Ave, Toronto, Ontario, Canada*
[b]*Institute of Health Management, Policy and Evaluation, University of Toronto, Suite 425, 155 College St., Toronto, Ontario, Canada*
[c]*Dalla Lana School of Public Health, University of Toronto, 6th floor, 155 College St., Toronto, Ontario, Canada*
[d]*Division of Cardiology, Sunnybrook Schulich Heart Centre and Faculty of Medicine, University of Toronto, 2075 Bayview Ave., Toronto, Ontario, Canada*
[e]*National Heart, Lung, and Blood Institute's Framingham Heart Study, 73 Mt. Wayte Ave., Framingham, MA, USA*
[f]*Center for Population Studies, National Heart, Lung, and Blood Institute, 31 Center Dr., Bethesda, MD 20892, USA*
[g]*Department of Medicine, Section of Cardiovascular Medicine, Boston University, 88 East Newton St., C-818, Boston, MA 02118, USA*
[h]*Department of Medicine, School of Medicine, Boston University, 72 E. Concord St., Boston, MA 02118, USA*
[i]*Department of Medicine, University Health Network and Faculty of Medicine, University of Toronto, Room 4NU-482, 200 Elizabeth St., Toronto, Ontario, Canada*

Accepted 25 November 2012; Published online 4 February 2013

Abstract

Objective: Physicians classify patients into those with or without a specific disease. Furthermore, there is often interest in classifying patients according to disease etiology or subtype. Classification trees are frequently used to classify patients according to the presence or absence of a disease. However, classification trees can suffer from limited accuracy. In the data-mining and machine-learning literature, alternate classification schemes have been developed. These include bootstrap aggregation (bagging), boosting, random forests, and support vector machines.

Study Design and Setting: We compared the performance of these classification methods with that of conventional classification trees to classify patients with heart failure (HF) according to the following subtypes: HF with preserved ejection fraction (HFPEF) and HF with reduced ejection fraction. We also compared the ability of these methods to predict the probability of the presence of HFPEF with that of conventional logistic regression.

Results: We found that modern, flexible tree-based methods from the data-mining literature offer substantial improvement in prediction and classification of HF subtype compared with conventional classification and regression trees. However, conventional logistic regression had superior performance for predicting the probability of the presence of HFPEF compared with the methods proposed in the data-mining literature.

Conclusion: The use of tree-based methods offers superior performance over conventional classification and regression trees for predicting and classifying HF subtypes in a population-based sample of patients from Ontario, Canada. However, these methods do not offer substantial improvements over logistic regression for predicting the presence of HFPEF.

Keywords: Boosting; Classification trees; Bagging; Random forests; Classification; Regression trees; Support vector machines; Regression methods; Prediction; Heart failure

Conflict of interest statement: The authors declare that there is no conflict of interest.

Funding: This study was supported by the Institute for Clinical Evaluative Sciences (ICES), which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). The opinions, results, and conclusions reported in this article are those of the authors and are independent from the funding sources. No endorsement by ICES or the Ontario MOHLTC is intended or should be inferred. This research was supported by an operating grant from the Canadian Institutes of Health Research (CIHR) (MOP 86508). Dr Austin is supported in part by a Career Investigator award from the Heart and Stroke Foundation. Dr Tu is supported by a Canada Research Chair in Health Services Research and a Career Investigator Award from the Heart and Stroke Foundation. Dr Lee is a clinician−scientist of the CIHR. The data used in this study were obtained from the Enhanced Feedback for Effective Cardiac Treatment (EFFECT) study. The EFFECT study was funded by a CIHR Team Grant in Cardiovascular Outcomes Research.

* Corresponding author. Institute for Clinical Evaluative Sciences, G1 06, 2075 Bayview Avenue, Toronto, Ontario M4N 3M5, Canada. Tel.: +1 416 480 6131; fax: +1 416 480 6048.

E-mail address: peter.austin@ices.on.ca (P.C. Austin).

0895-4356
http://dx.doi.org/10.1016/j.jclinepi.2012.11.008

What is new?

Key findings

- Modern data-mining and machine-learning methods offer advantages for predicting and classifying heart failure (HF) patients according to disease subtype: HF with preserved ejection fraction (HFPEF) and HF with reduced ejection fraction compared with conventional regression and classification trees.
- Conventional logistic regression performed at least as well as modern methods from the data-mining and machine-learning literature for predicting the probability of the presence of HFPEF in patients with HF.

What this adds to what was known?

- Boosted trees, bagged trees, and random forests do not offer an advantage over conventional logistic regression for predicting the probability of disease subtype in patients with HF.

What is the implication and what should change now?

- Conventional logistic regression should remain a standard tool in the analyst's toolbox when predicting disease subtype in patients with HF.
- Analysts interested in classifying HF patients according to disease subtype should use ensemble-based methods rather than conventional classification trees.

1. Introduction

There is an increasing interest in using classification methods in clinical research. Classification methods allow one to assign subjects to one of a mutually exclusive set of states. Accurate classification of disease states (disease present/absent) or of disease etiology or subtype allows subsequent investigations, treatments, and interventions to be delivered in an efficient and targeted manner. Similarly, accurate classification of disease states permits more accurate assessment of patient prognosis.

Classification trees use binary recursive partitioning methods to partition the sample into distinct subsets [1–4]. Although their use is popular in clinical research, concerns have been raised about the accuracy of tree-based methods of classification and regression [2,4]. In the data-mining and machine-learning literature, alternatives to and extensions of classical classification trees have been developed in recent years. Many of these methods involve aggregating classifications over an ensemble of classification trees. For this reason, many of these methods are referred to as ensemble methods. Ensemble-based methods include bagged classification trees, random forests, and boosted trees. Alternate classification methods include support vector machines (SVMs).

In patients with acute heart failure (HF), there are two distinct subtypes: HF with preserved ejection fraction (HFPEF) and HF with reduced ejection fraction (HFREF). The distinction between HFPEF and HFREF is particularly relevant in the clinical setting. Although the treatment of HFREF is based on a multitude of large randomized clinical trials, the evidence base for the treatment of HFPEF is much smaller and more focused on related comorbid conditions [5]. Although the overall prognosis appears to be similar within the two subtypes of HF, there are important differences in cause-specific mortality, which would be relevant in risk stratification and disease management [6]. The diagnosis of HFREF versus HFPEF is ideally made using results from echocardiography. Although echocardiography should ideally be done in all HF patients at some point in their clinical care, this test is not always performed even in high-resource regions, and treatment decisions may need to be made before echocardiographic data are available. In one US Medicare cohort, more than one-third of HF patients did not undergo echocardiography in hospital [7].

The present study had two objectives. First, to compare the accuracy of different methods for classifying HF patients according to two disease subtypes, HFPEF vs. HFREF, and for predicting the probability of patients having HFPEF in a population-based sample of HF patients in Ontario, Canada. Second, to compare the accuracy of the prediction of the presence of HFPEF using methods from the data-mining literature with that of conventional logistic regression.

2. Methods for classification and prediction

In this section, we describe the different methods that will be used for classification and prediction. For classification, we restrict our attention to binary classification in which subjects are classified as belonging to one of two possible categories. Our case study will consist of patients with acute HF that is further classified as HF with preserved ejection fraction (HFPEF) and HF with reduced ejection fraction (HFREF). By prediction, we mean prediction of the probability of an event or of being in a particular state. In our case study, this will be the predicted probability of having HFPEF. We consider the following classification methods: classification trees, bagged classification trees, random forests, boosted classification trees, and SVMs. For prediction, we consider the following methods: logistic regression, regression trees, bagged regression trees, random forests, and boosted regression trees.

2.1. Classification and regression trees

Classification and regression trees use binary recursive partitioning methods to partition the sample into distinct subsets [3]. At the first step, all possible dichotomizations of all continuous variables (above vs. below a given threshold) and of all categorical variables are considered. Using each possible dichotomization, all the possible ways of partitioning the sample into two distinct subsets are considered. The binary partition that results in the greatest reduction in impurity is selected. This process is then repeated iteratively until a predefined stopping rule is satisfied. For classification, a subject's class can be determined using the status that was observed for the majority of subjects within that subset to which the given subject belongs (i.e., classification by majority vote). For prediction, the predicted probability of the event for a given subject can be estimated using the proportion of subjects who have the condition of interest among all the subjects in the subset to which the given subject belongs.

Advocates for classification and regression trees have suggested that these methods allow for the construction of easily interpretable decision rules that can be easily applied in clinical practice. Furthermore, it has been suggested that classification and regression tree methods are adept at identifying important interactions in the data [8–10] and in identifying clinical subgroups of subjects at very high or very low risk of adverse outcomes [11]. Advantages of tree-based methods are that they do not require the specification of the parametric nature of the relationship between the predictor variables and outcome. Additionally, assumptions of linearity that are frequently made in conventional regression models are not required for tree-based methods.

We grew classification and regression trees using the tree function from the *tree* package for the R statistical programming language [12,13]. In our study, we used the default criteria in the *tree* package for growing regression trees: at a given node, the partition was chosen that maximized the reduction in deviance; the smallest permitted node size was 10; and a node was not subsequently partitioned if the within-node deviance was less than 0.01 of that of the root node. Once the initial regression tree had been grown, the tree was pruned. The optimal number of leaves was determined by identifying the tree size that minimized the tree deviance when 10-fold cross-validation was used in the derivation sample.

2.2. Bagging classification or regression trees

Bootstrap aggregation or bagging is a generic approach that can be used with different classification and prediction methods [4]. Our focus is on bagging classification or regression trees. Repeated bootstrap samples are drawn from the study sample. A classification or regression tree is grown in each of these bootstrap samples. Using each of the grown regression trees, classifications or predictions are obtained for each study subject. Finally, for each study subject, a prediction is obtained by averaging the predictions obtained from the regression trees grown over the different bootstrap samples. For each study subject, a final classification is obtained by a majority vote across the classification trees grown in the different bootstrap samples. We used the bagging function from the *ipred* package for the R statistical programming language to fit bagged regression trees [14]. All parameter values were set to the default values in the bagging function. In our application of bagging, we used 100 bootstrap samples.

2.3. Random forests

The Random Forests approach was developed by Brieman [15]. The Random Forests approach is similar to bagging classification or regression trees with one important modification. When one is growing a classification or regression tree in a particular bootstrap sample, at a given node, rather than considering all possible binary splits on all candidate variables, one only considers binary splits on a random sample of the candidate predictor variables. The size of the set of randomly selected predictor variables is defined before the process. When fitting random forests of regression trees, we let the size of the set of randomly selected predictor variables be $\lfloor p/3 \rfloor$, where p denotes the total number of predictor variables and $\lfloor\ \rfloor$ denotes the floor function. When fitting random forests of classification trees, we let the size of the set of randomly selected predictor variables be $\sqrt{p}$ (these are the defaults in the R implementation of random forests). We grew random forests consisting of 500 regression or classification trees. Predictions or classifications are obtained by averaging predictions across the regression trees or by majority vote across the classification trees, respectively. We used the randomForest function from the *RandomForest* package for R to estimate random forests [16]. All parameter values were set to their defaults.

2.4. Boosting

One of the most promising extensions of classical classification methods is boosting. Boosting is a method for combining "the outputs from several 'weak' classifiers to produce a powerful 'committee'" [4]. A weak classifier has been described as one whose error rate is only slightly better than random guessing [4]. Breiman has suggested that boosting applied with classification trees as the weak classifiers is the "best off-the-shelf" classifier in the world [4].

When focusing on classification, we used the *AdaBoost.M1* algorithm proposed by Freund and Schapire [17]. Boosting sequentially applies a weak classifier to a sequence of reweighted versions of the data, thereby producing a sequence of weak classifiers. At each step of the sequence, subjects who were incorrectly classified by the previous classifier are weighted more heavily than those who were correctly classified. The classifications from this sequence of weak classifiers are then combined through

a weighted majority vote to produce the final prediction. The reader is referred elsewhere for a more detailed discussion of the theoretical foundation of boosting and its relationship with established methods in statistics [18,19]. The AdaBoost.M1 algorithm for boosting can be applied with any classifier. However, it is most frequently used with classification trees as the base classifier [4]. Even using a ''stump'' (a stump is a classification tree with exactly one binary split and two terminal nodes or leaves) as the weak classifier has been shown to produce substantial improvement in prediction error compared with a large classification tree [4]. Given the lack of consensus on optimal tree depth, we considered four versions of boosted classification trees: classification trees of depths 1, 2, 3, and 4 as the base classifiers. For each method, we used sequences of 100 classification trees. We used the ada function from the *ada* package for R for boosting classification trees, which implements the AdaBoost.M1 algorithm [20].

Generalized boosting methods adapt the above algorithm for use with regression rather than with classification [4,21]. We considered four different base regression models: regression trees of depths 1, 2, 3, and 4. These have also been referred to as regression trees with interaction depths 1 through 4. For each method, we considered sequences of 10,000 regression trees. We used the gbm function from the *gbm* package for boosting regression trees [22].

2.5. Support vector machines

SVMs are based on the fact that with an appropriate function to a sufficiently high dimension, data from two categories can always be separated by a hyperplane [23]. An SVM is the separating hyperplane that maximizes the distance from the nearest subjects with and without the outcome. Subjects are then classified according to which side of the hyperplane they lie on. Readers are referred elsewhere for a more extensive treatment of SVMs [4,23]. We used the svm function from the *e1071* package for R [24].

2.6. Logistic regression

Finally, conventional logistic regression can be used to obtain predicted probabilities of being in a particular state or of the occurrence of a specific outcome. Unlike the methods described previously, logistic regression results in only a predicted probability of an event and not a binary classification. We used the lrm function from the *Design* package for the R statistical programming language to estimate the logistic regression models [25].

3. Methods

3.1. Data sources

The Enhanced Feedback for Effective Cardiac Treatment (EFFECT) Study was an initiative to improve the quality of care for patients with cardiovascular disease in Ontario, Canada [26,27]. The EFFECT study consisted of two phases. During the first phase, detailed clinical data on patients hospitalized with HF between April 1, 1999, and March 31, 2001, at 103 acute care hospitals in Ontario, Canada, were obtained by a retrospective chart review. During the second phase, data were abstracted on patients hospitalized with HF between April 1, 2004, and March 31, 2005, at 96 Ontario hospitals. Data on patient demographics, vital signs, and physical examination at presentation, medical history, and results of laboratory tests were collected for this sample.

In the EFFECT study, detailed clinical data were available on 9,943 and 8,339 patients hospitalized with a diagnosis of HF during the first and second phases of the study, respectively. After excluding subjects with missing data on key variables and for whom ejection fraction could not be determined, 3,697 and 4,515 subjects were available from the first and second phases, respectively, for inclusion in the present study. The first and second phases of the EFFECT study will be referred to as the EFFECT-1 and EFFECT-2 samples, respectively (these were referred to as the EFFECT Baseline sample and EFFECT Follow-up sample, respectively, in the original EFFECT publication [27]).

For the purposes of our analyses, only participants with available left ventricular ejection fraction (LVEF) assessment by cardiac imaging were included. Participants were classified as having HFPEF (LVEF, $>45\%$) or HFREF (LVEF, $\leq 45\%$). This distinction is clinically relevant as the treatment for HFPEF and HFREF is distinct: whereas the treatment of HFREF with beta-blockers, angiotensin-converting enzyme inhibitors, and aldosterone blockers is well-substantiated, the treatment of HFPEF is much less defined and focuses more on underlying comorbid conditions [5].

As candidate variables for classifying HF subtype or predicting the presence of HFPEF, we considered 34 variables denoting demographic characteristics, vital signs, presenting signs and symptoms, results of laboratory investigations, and previous medical history. These variables are listed in Table 1.

In each of the two samples, the Kruskal–Wallis and chi-square tests were used to compare continuous and categorical baseline characteristics, respectively, between patients with HFPEF and those with HFREF. Furthermore, characteristics were compared between patients in the EFFECT-1 and those in the EFFECT-2 samples.

3.2. Comparison of predictive ability of different regression methods

We examined the predictive accuracy of each method using the EFFECT-1 sample as the model derivation sample and the EFFECT-2 sample as the model validation sample. Using each prediction method, a model was developed for

Table 1. Comparison of patients with HFPEF and HFREF in the EFFECT-1 and EFFECT-2 samples

	EFFECT-1 sample			EFFECT-2 sample		
Variables	**HFREF (N = 2,529)**	**HFPEF (N = 1,168)**	***P*-value**	**HFREF (N = 2,776)**	**HFPEF (N = 1,739)**	***P*-value**
Age (yr)	75.0 (66.0–81.0)	77.0 (70.0–83.0)	<0.001	76.0 (67.0–82.0)	79.0 (71.0–85.0)	<0.001
Male (%)	1,547 (61.2)	423 (36.2)	<0.001	1,686 (60.7)	643 (37.0)	<0.001
Heart rate on admission (bpm)	96.0 (78.0–113.0)	90.0 (74.0–110.0)	<0.001	94.0 (76.0–112.0)	86.0 (70.0–105.0)	<0.001
Systolic blood pressure on admission (mm Hg)	142.0 (122.0–164.0)	156.0 (134.0–180.0)	<0.001	140.0 (120.0–161.0)	150.0 (131.0–173.0)	<0.001
Respiratory rate on admission	24.0 (20.0–30.0)	24.0 (20.0–28.0)	0.853	24.0 (20.0–28.0)	22.0 (20.0–28.0)	0.156
History of hypertension (%)	1,245 (49.2)	675 (57.8)	<0.001	1,784 (64.3)	1,264 (72.7)	<0.001
Diabetes mellitus (%)	919 (36.3)	374 (32.0)	0.01	1,055 (38.0)	661 (38.0)	0.997
Current smoker (%)	415 (16.4)	130 (11.1)	<0.001	382 (13.8)	157 (9.0)	<0.001
History of coronary artery disease (%)	1,339 (52.9)	345 (29.5)	<0.001	1,519 (54.7)	579 (33.3)	<0.001
Atrial fibrillation (%)	594 (23.5)	396 (33.9)	<0.001	752 (27.1)	619 (35.6)	<0.001
Left bundle branch block (%)	530 (21.0)	55 (4.7)	<0.001	540 (19.5)	109 (6.3)	<0.001
Any ST elevation (%)	371 (14.7)	70 (6.0)	<0.001	169 (6.1)	35 (2.0)	<0.001
Any T-wave inversion (%)	905 (35.8)	319 (27.3)	<0.001	864 (31.1)	414 (23.8)	<0.001
Neck vein distension (%)	1,575 (62.3)	671 (57.4)	0.005	1,826 (65.8)	1,115 (64.1)	0.254
s3 (%)	351 (13.9)	92 (7.9)	<0.001	245 (8.8)	72 (4.1)	<0.001
s4 (%)	122 (4.8)	44 (3.8)	0.149	97 (3.5)	39 (2.2)	0.017
Rales >50% of lung field (%)	299 (11.8)	101 (8.6)	0.004	361 (13.0)	206 (11.8)	0.253
Pulmonary edema (%)	1,298 (51.3)	588 (50.3)	0.579	1,750 (63.0)	1,042 (59.9)	0.036
Cardiomegaly (%)	1,012 (40.0)	393 (33.6)	<0.001	1,377 (49.6)	691 (39.7)	<0.001
Cerobrovascular disease/ transient ischemic attack (%)	379 (15.0)	189 (16.2)	0.349	432 (15.6)	326 (18.7)	0.005
Previous AMI (%)	1,145 (45.3)	239 (20.5)	<0.001	1,281 (46.1)	401 (23.1)	<0.001
Peripheral arterial disease (%)	376 (14.9)	132 (11.3)	0.003	404 (14.6)	206 (11.8)	0.01
Chronic obstructive pulmonary disease (%)	362 (14.3)	206 (17.6)	0.009	577 (20.8)	378 (21.7)	0.446
Dementia (%)	124 (4.9)	68 (5.8)	0.242	168 (6.1)	133 (7.6)	0.036
Cirrhosis (%)	23 (0.9)	12 (1.0)	0.731	19 (0.7)	13 (0.7)	0.806
Cancer (%)	282 (11.2)	132 (11.3)	0.893	292 (10.5)	186 (10.7)	0.851
Hemoglobin	12.7 (11.3–14.1)	12.3 (10.7–13.5)	<0.001	12.7 (11.2–14.0)	12.1 (10.7–13.4)	<0.001
White blood cell count	8.9 (7.1–11.4)	9.1 (7.1–11.5)	0.527	8.7 (7.0–11.4)	8.8 (7.0–11.4)	0.901
Sodium	139.0 (136.0–141.0)	139.0 (136.0–141.0)	0.574	139.0 (136.0–141.0)	139.0 (136.0–142.0)	0.681
Glucose	7.7 (6.1–11.2)	7.2 (6.0–10.1)	<0.001	7.4 (6.0–10.5)	7.1 (5.9–9.6)	0.004
Urea	8.3 (6.1–12.2)	7.6 (5.6–11.4)	<0.001	8.3 (6.2–12.1)	8.1 (5.9–11.5)	0.006
Creatinine	108.0 (86.0–142.0)	98.0 (76.0–136.0)	<0.001	110.0 (87.0–146.0)	100.0 (79.0–134.0)	<0.001
eGFR (mL/min/1.73 m^2)	55.1 (39.3–71.9)	57.3 (39.2–75.8)	0.06	54.9 (38.9–70.9)	54.6 (39.7–74.2)	0.301
Potassium	4.2 (3.9–4.6)	4.2 (3.8–4.6)	0.008	4.2 (3.9–4.6)	4.2 (3.8–4.6)	<0.001

Abbreviations: AMI, acute myocardial infarction; HFPEF, heart failure with preserved ejection fraction; HFREF, heart failure with reduced ejection fraction; EFFECT, Enhanced Feedback for Effective Cardiac Treatment.

Dichotomous variables are reported as N (%), whereas continuous variables are reported as median (25th percentile–75th percentile). The Kruskal–Wallis and chi-square tests were used to compare continuous and categorical baseline characteristics, respectively, between patients with HFPEF and those with HFREF.

predicting the probability of HFPEF using the subjects in the EFFECT-1 sample. We then applied the developed model to each subject in the EFFECT-2 sample to estimate that subject's predicted probability of having HFPEF. Note that the derivation and validation samples consist of patients from the same jurisdiction (Ontario). Furthermore, most acute hospitals that cared for HF patients were included in both these two data sets. However, the derivation and validation samples are separated temporally (1999/2000 and 2000/2001 vs. 2004/2005). The study design ensured that there was very little overlap in patients between the two study periods.

The tree-based methods considered all 34 variables described in Table 1 as candidate predictor variables. Two separate logistic regression models were fit to predict the probability of the presence of HFPEF. First, we fit a logistic regression model that contained all 34 variables as main effects. No variable reduction was performed. Restricted cubic

splines (cubic splines that are linear in the tails) with four knots were used to model the relationship between each continuous covariate and the log-odds of having HFPEF [28]. Second, we fit a logistic regression model that contained 14 predictor variables that previously had been identified as important predictors of HF disease subtypes using data from the Framingham Heart Study (age, sex, heart rate, systolic blood pressure, history of coronary heart disease, history of hypertension, diabetes mellitus, current smoker, hemoglobin, eGFR, atrial fibrillation, left bundle branch block, any ST elevation, and any T-wave inversion) [29].

Predictive accuracy was assessed using two different metrics. First, we calculated the area under the receiver operating characteristic curve (AUC), which is equivalent to the *c*-statistic [28,30]. Second, we calculated the Brier Score [28] (mean squared prediction error), which is defined as $\frac{1}{N}\sum_{i=1}^{N}(\widehat{P}_i - Y_i)^2$, where N denotes the sample size, $\widehat{P}_i$ is the predicted probability of the outcome, and Y_i is the observed outcome (1/0). We used the val.prob function from the *Design* package to estimate these two measures of predictive accuracy.

We also examined the calibration of the predictions obtained using each method. For each method, subjects in the validation sample were divided into 10 groups defined by the deciles of the predicted probability of the presence of HFPEF. Within each of the 10 groups, the mean predicted probability of HFPEF was compared with the observed probability of having HFPEF.

3.3. Comparison of accuracy of classification for different classification methods

Classification models were developed that considered all 34 variables described in Table 1 as potential predictor variables. As described previously, accuracy of classification was assessed using the EFFECT-1 sample as the model derivation sample and the EFFECT-2 sample as the model validation sample. For each subject in the validation sample, a true HF subtype was observed (HFPEF vs. HFREF), and a classification was obtained (HFPEF vs. HFREF) for each classification method developed in the EFFECT-1 sample. The accuracy of classification was assessed using sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) [31].

4. Results

4.1. Description of study sample

Comparisons of baseline characteristics between patients with and without preserved ejection fraction in the EFFECT-1 and EFFECT-2 samples are reported in Table 1. In each of the EFFECT-1 and EFFECT-2 samples, there were statistically significant differences in 24 of the 34 baseline covariates between subjects with HFPEF and subjects with HFREF. Comparisons of baseline characteristics of patients in the EFFECT-1 sample with those of patients in the EFFECT-2 sample are reported in Table 2. There were significant differences in 20 of the 34 baseline covariates between the two samples. Importantly, the proportion of patients with HFPEF was modestly higher in the EFFECT-2 sample than it was in the EFFECT-1 sample (31.6% vs. 38.5%). The higher proportion of patients with HFPEF in the EFFECT-2 sample could reflect a higher average age and greater prevalence of risk factors such as hypertension and atrial fibrillation over time, which are more commonly associated with HFPEF.

4.2. Comparison of predictive ability of different regression methods

The predictive accuracy of the different methods for predicting the probability of the presence of HFPEF is reported in Table 3. The AUC in the EFFECT-2 sample of the different models developed in the EFFECT-1 sample ranged from a low of 0.683 for the regression tree to a high of 0.780 for the nonparsimonious logistic regression model. Boosted regression trees of depths 3 and 4 had AUCs that were very similar to those of the nonparsimonious logistic regression model (0.772 and 0.774, respectively). The Brier Score in the EFFECT-2 sample of the different models developed in the EFFECT-1 sample ranged from a high of 0.2152 for the regression tree to a low of 0.1861 for the nonparsimonious logistic regression model.

For both measures of predictive accuracy, the use of conventional regression trees resulted in predicted probabilities of the presence of HFPEF with the lowest accuracy. A nonparsimonious logistic regression resulted in the greatest out-of-sample predictive accuracy when using the EFFECT-2 sample as the validation sample. Boosted regression trees of depths 3 and 4 had predictive accuracy that approached that of the nonparsimonious logistic regression model.

The calibration of each of the prediction methods is described graphically in the Figure. Although all methods tended to underestimate the probability of the presence of HFPEF, the two logistic regression models and the random forests resulted in estimates that displayed the best calibration. The underestimation of the predicted probability of HFPEF or miscalibration was most likely because of the differences in the prevalence of HFPEF between the two samples. As noted previously, the prevalence of HFPEF was modestly higher in the EFFECT-2 sample than it was in the EFFECT-1 sample.

4.3. Comparison of accuracy of classification for different classification methods

The sensitivity, specificity, PPV, and NPV of the different classification methods are reported in Table 4. The sensitivity in the EFFECT-2 sample of the different models

Table 2. Comparison of EFFECT-1 and EFFECT-2 samples

Variables	EFFECT-1 sample ($N = 3,697$)	EFFECT-2 sample ($N = 4,515$)	*P*-value
HFPEF (%)	1,168 (31.6)	1,739 (38.5)	<0.001
Age (yr)	75.0 (68.0–82.0)	77.0 (68.0–83.0)	<0.001
Male (%)	1,970 (53.3)	2,329 (51.6)	0.124
Heart rate on admission (bpm)	94.0 (77.0–112.0)	91.0 (74.0–110.0)	<0.001
Systolic blood pressure on admission (mm Hg)	147.0 (126.0–170.0)	144.0 (124.0–166.0)	<0.001
Respiratory rate on admission	24.0 (20.0–30.0)	23.0 (20.0–28.0)	<0.001
History of hypertension (%)	1,920 (51.9)	3,048 (67.5)	<0.001
Diabetes mellitus (%)	1,293 (35.0)	1,716 (38.0)	0.005
Current smoker (%)	545 (14.7)	539 (11.9)	<0.001
History of coronary artery disease (%)	1,684 (45.6)	2,098 (46.5)	0.407
Atrial fibrillation (%)	990 (26.8)	1,371 (30.4)	<0.001
Left bundle branch block (%)	585 (15.8)	649 (14.4)	0.067
Any ST elevation (%)	441 (11.9)	204 (4.5)	<0.001
Any T-wave inversion (%)	1,224 (33.1)	1,278 (28.3)	<0.001
Neck vein distension (%)	2,246 (60.8)	2,941 (65.1)	<0.001
s3 (%)	443 (12.0)	317 (7.0)	<0.001
s4 (%)	166 (4.5)	136 (3.0)	<0.001
Rales >50% of lung field (%)	400 (10.8)	567 (12.6)	0.015
Pulmonary edema (%)	1,886 (51.0)	2,792 (61.8)	<0.001
Cardiomegaly (%)	1,405 (38.0)	2,068 (45.8)	<0.001
Cerobrovascular disease/transient ischemic attack (%)	568 (15.4)	758 (16.8)	0.081
Previous AMI (%)	1,384 (37.4)	1,682 (37.3)	0.865
Peripheral arterial disease (%)	508 (13.7)	610 (13.5)	0.762
Chronic obstructive pulmonary disease (%)	568 (15.4)	955 (21.2)	<0.001
Dementia (%)	192 (5.2)	301 (6.7)	0.005
Cirrhos (%)	35 (0.9)	32 (0.7)	0.233
Cancer (%)	414 (11.2)	478 (10.6)	0.376
Hemoglobin	12.5 (11.2–13.9)	12.4 (11.0–13.8)	0.035
White blood cell count	9.0 (7.1–11.4)	8.8 (7.0–11.4)	0.105
Sodium	139.0 (136.0–141.0)	139.0 (136.0–141.0)	0.399
Glucose	7.5 (6.0–10.9)	7.3 (6.0–10.1)	<0.001
Urea	8.1 (5.9–12.0)	8.2 (6.1–11.9)	0.371
Creatinine	105.0 (84.0–140.0)	106.0 (84.0–142.0)	0.679
eGFR (mL/min/1.73 m^2)	56.0 (39.3–73.2)	54.8 (39.3–72.0)	0.219
Potassium	4.2 (3.9–4.6)	4.2 (3.9–4.6)	0.236

Abbreviations: AMI, acute myocardial infarction; HFPEF, HF with preserved ejection fraction; EFFECT, Enhanced Feedback for Effective Cardiac Treatment.

Dichotomous variables are reported as *N* (%), whereas continuous variables are reported as median (25th percentile–75th percentile). The Kruskal–Wallis and chi-square tests were used to compare continuous and categorical baseline characteristics, respectively, between patients in the two phases of the EFFECT sample.

developed in the EFFECT-1 sample ranged from a low of 0.378 for the random forest to a high of 0.500 for the boosted classification tree of depth 4. Specificity ranged from a low of 0.820 for the conventional classification tree and the boosted classification trees of depth 4 to a high of 0.897 for the random forest. PPV ranged from a low of 0.616 for the classification tree to a high of 0.696 for the random forest. The NPV ranged from a low of 0.697 for the random forest to a high of 0.726 for the boosted classification tree of depth 2.

5. Discussion

Classification plays an important role in modern clinical research. The objective of binary classification schemes or algorithms is to classify subjects into one of two mutually exclusive categories based on their observed characteristics. In clinical research, a common binary classification is diseased/nondiseased, different disease subtypes, or disease etiology. Classification trees are a commonly used binary classification method. In the data-mining and machine-learning fields, improvements to classical classification trees have been developed. Many of these methods involve aggregating classifications across a set of classification trees. There is limited research comparing the performance of different classification/prediction methods for predicting the presence of disease, disease etiology, or disease subtype.

We compared the performance of modern classification and regression methods with classification and regression trees to classify patients with HF into one of two mutually exclusive categories (HFPEF vs. HFREF) or to predict the probability of the presence of HFPEF. We found that modern classification methods offered improved performance over conventional classification trees for classifying HF patients according to disease subtype. Several observations warrant comment. First, when focusing on predicting the probability of the presence of HFPEF, conventional

Table 3. Accuracy of prediction in EFFECT-2 sample

Prediction method	AUC or *c*-statistic	Brier score
Regression tree	0.683	0.2152
Bagged regression tree	0.733	0.2079
Random forest	0.751	0.1959
Boosted regression tree (depth 1)	0.752	0.2049
Boosted regression tree (depth 2)	0.768	0.1962
Boosted regression tree (depth 3)	0.772	0.1933
Boosted regression tree (depth 4)	0.774	0.1918
Logistic regression (full model)	0.780	0.1861
Logistic regression (simple model)	0.766	0.1914

Abbreviations: EFFECT, Enhanced Feedback for Effective Cardiac Treatment; AUC, area under the receiver operating characteristic curve.

regression trees had lower predictive accuracy compared with all other methods that we examined. Second, logistic regression had the best predictive accuracy for predicting the presence of HFPEF. Third, when focusing on classification, boosted classification trees of depth 4 had the highest sensitivity. Fourth, random forests had the highest specificity for classifying patients according to disease subtype.

The present study had a very limited focus: comparing the ability of different methods to predict or classify disease subtype in patients hospitalized with HF in Ontario, Canada. Our conclusions about the relative performance of different classification and prediction methods should be restricted to this patient population and this specific classification scheme (i.e., HFPEF vs. HFREF). Readers should not conclude that logistic regression will have superior predictive ability compared with ensemble-based methods in all settings and for all conditions or outcomes. However, recent studies in patients with cardiovascular disease merit discussion. A recent study that examined the predictive ability of ensemble-based methods for predicting the probability of short-term mortality in patients hospitalized with either acute myocardial infarction or HF found that ensemble methods resulted in improved predictive accuracy compared with conventional regression trees [32]. However,

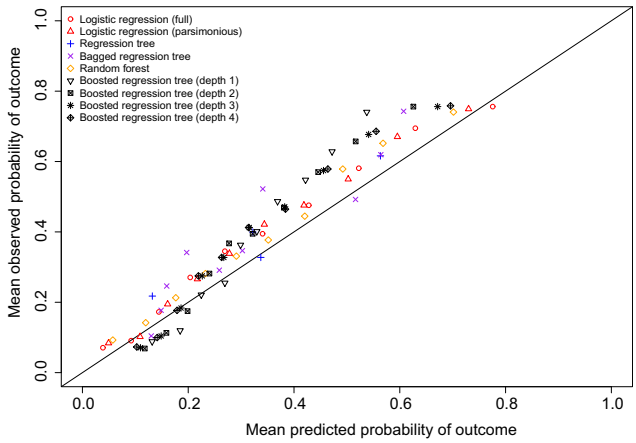

Figure. Calibration of prediction methods in the EFFECT Follow-up sample. EFFECT, Enhanced Feedback for Effective Cardiac Treatment.

ensemble-based methods did not result in improved predictive performance compared with conventional logistic regression. In a different study focusing on classifying patients with HF according to mortality outcomes, boosted classification trees were found to result in minor to modest improvement in accuracy of classification compared with conventional classification trees [33].

Comparisons similar to the above have been conducted by other authors. Wu et al. [34], comparing the performance of logistic regression, boosting, and SVMs to predict the subsequent development of HF, found that the former two methods had comparable performance, whereas the latter method had the poorest performance. Maroco et al. [35] compared 10 different classifiers for predicting the evolution of mild cognitive impairment with dementia. They concluded that random forests and linear discriminant analysis had the best performance for predicting progression to dementia. Although these two studies focused on disease incidence, a third study compared three methods for predicting survival in patients with breast cancer [36]. They found that a decision tree resulted in the greatest accuracy, followed by artificial neural networks, with logistic regression resulting in the lowest accuracy. In an extensive set of analyses, Caruana and Niculescu-Mizil [37] compared the performance of 10 prediction/classification algorithms on 11 binary classification problems using eight performance metrics. They found that bagged trees, random forests, and neural networks resulted in the best average performance across the different metrics and data sets. In general, they found that conventional classification trees and logistic regression had inferior performance to the best-performing methods. In a related study, Caruana et al. [38] examined the effect of dimensionality (i.e., the number of available predictor variables) on the relative performance of different classification algorithms. They found that as dimensionality increases, the relative performance of the different algorithms changes. They also observed that random forests tended to perform well across all dimensions.

There are several limitations to the present study. First, as noted previously, our conclusions are limited to the relative performance of methods for classification/prediction of disease subtype in patients hospitalized with HF. Our conclusions are not intended to be generalized to other patient populations or other outcomes and conditions. Second, for some of the prediction and classification methods, we used the default settings in the given statistical software package for estimating the given model (e.g., regression trees). Similarly, for random forests, we used the default setting for the size of the random sample of predictor variables that was considered at each split of a given tree. However, for other methods, no such default specification existed. In particular, for boosted classification and regression trees, there is limited research on the optimal depth of the fitted trees. Furthermore, it is possible that the optimal tree depth may vary across settings and outcomes. Because of limited research on optimal tree depth, we grew four

Table 4. Sensitivity and specificity of classification in EFFECT-2 sample

Classification method	Sensitivity	Specificity	Positive predictive value	Negative predictive value
Classification tree	0.462	0.820	0.616	0.709
Bagged classification tree	0.451	0.849	0.653	0.712
Random forest	0.378	0.897	0.696	0.697
Boosted classification tree (depth 1)	0.453	0.876	0.695	0.719
Boosted classification tree (depth 2)	0.491	0.847	0.667	0.726
Boosted classification tree (depth 3)	0.492	0.828	0.642	0.722
Boosted classification tree (depth 4)	0.500	0.820	0.635	0.724
Support vector machines	0.401	0.887	0.690	0.703

Abbreviation: EFFECT, Enhanced Feedback for Effective Cardiac Treatment.

different sets of boosted trees, with tree depths of 1, 2, 3, and 4. For this reason, concluding that boosted trees had the best performance among the different modern prediction methods risks resulting in an incorrect conclusion because comparable tuning parameters were not varied for the other prediction methods. Our finding that boosted trees of depth 4 tended to have superior performance compared with those of other depths merits replication in other data sets, in other settings, and for other outcomes. However, it should be noted that boosted trees of this depth performed well for predicting cardiovascular mortality in an earlier study [32].

In summary, we found that modern, flexible tree-based methods from the data-mining and machine-learning literature offer substantial improvement in prediction and classification of HF subtype compared with conventional classification and regression trees. However, conventional logistic regression was able to more accurately predict the probability of the presence of HFPEF among patients with HF compared with the methods proposed in the data-mining and machine-learning literature.

References

[1] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Boca Raton: Chapman & Hall/CRC; 1998.

[2] Austin PC. A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. Stat Med 2007;26:2937−57.

[3] Clark LA, Pregibon D. Tree-based methods. In: Chambers JM, Hastie TJ (eds) *Statistical models in S*. Chapman & Hall: New York, NY, 1993:377−419.

[4] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Data mining, inference, and prediction. New York, NY: Springer-Verlag; 2001.

[5] Hunt SA, Abraham WT, Chin MH, Feldman AM, Francis GS, Ganiats TG, et al. 2009 focused update incorporated into the ACC/AHA 2005 Guidelines for the Diagnosis and Management of Heart Failure in Adults: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines: developed in collaboration with the International Society for Heart and Lung Transplantation. Circulation 2009;14:e391−479.

[6] Lee DS, Gona P, Vasan RS, Larson MG, Benjamin EJ, Wang TJ, et al. Relation of disease pathogenesis and risk factors to heart failure with preserved or reduced ejection fraction: insights from the framingham heart study of the national heart, lung, and blood institute. Circulation 2009;24:3070−7.

[7] Masoudi FA, Havranek EP, Smith G, Fish RH, Steiner JF, Ordin DL, et al. Gender, age, and heart failure with preserved left ventricular systolic function. J Am Coll Cardiol 2003;42:217−23.

[8] Sauerbrei W, Madjar H, Prompeler HJ. Differentiation of benign and malignant breast tumors by logistic regression and a classification tree using Doppler flow signals. Methods Inf Med 1998;37:226−34.

[9] Gansky SA. Dental data mining: potential pitfalls and practical issues. Adv Dental Res 2003;17:109−14.

[10] Nelson LM, Bloch DA, Longstreth WT Jr, Shi H. Recursive partitioning for the identification of disease risk subgroups: a case-control study of subarachnoid hemorrhage. J Clin Epidemiol 1998;51:199−209.

[11] Lemon SC, Roy J, Clark MA, Friedmann PD, Rakowski W. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. Ann Behav Med 2003;26:172−81.

[12] R Core Development Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2005.

[13] Ripley B. Tree: classification and regression trees. [1.0−28]. 2010.

[14] Peters A, Hothorn T. ipred: improved predictors. [0.8−8]. 2009.

[15] Breiman L. Random forests. Machine Learn 2001;45:5−32.

[16] Liaw A, Wiener M. Classification and regression by randomForest. R News 2002;2:18−22.

[17] Freund Y, Schapire R. Experiments with a new boosting algorithm. In: Machine learning: Proceedings of the Thirteenth International Conference. San Francisco, California: Morgan Kauffman; 1996: 148−56.

[18] Buhlmann P, Hathorn T. Boosting algorithms: regularization, prediction and model fitting. Stat Sci 2007;22:477−505.

[19] Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (with discussion). Ann Stat 2000;28:337−407.

[20] Culp M, Johnson K, Michailidis G. ada: ada: an R package for stochastic boosting. [2.0−2]. 2010.

[21] McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychol Methods 2004;9:403−25.

[22] Ridgeway, G. gbm: generalized boosted regression models. [1.6−3.1]. 2010.

[23] Duda RO, Hart PE, Stork DG. Pattern classification. New York: Wiley-Interscience; 2001.

[24] Dimitriadou E, Hornik K, Leisch F, Meyer D, Weingessel A. e1071: misc functions of the Department of Statistics (e1071), TU Wien. [1.5−24]. 2010.

[25] Harrell FE. Design: design package. [2.3−0]. 2009.

[26] Tu J, Donovan LR, Lee DS, Austin PC, Ko DT, Wang JT, et al. Quality of cardiac care in Ontario—phase 1. 1. Toronto, ON: Institute for Clinical Evaluative Sciences; 2004.

[27] Tu JV, Donovan LR, Lee DS, Wang JT, Austin PC, Alter DA, et al. Effectiveness of public report cards for improving the quality of cardiac care: the EFFECT study: a randomized trial. J Am Med Assoc 2009;302:2330−7.

[28] Harrell FE Jr. Regression modeling strategies. New York, NY: Springer-Verlag; 2001.

[29] Ho JE, Gona P, Pencina MJ, Tu JV, Austin PC, Vasan RS, et al. Discriminating clinical features of heart failure with preserved vs. reduced ejection fraction in the community. Eur Heart J 2012;33:1734−41.
[30] Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York, NY: Springer; 2009.
[31] Zhou X, Obuchowski N, McClish D. Statistical methods in diagnostic medicine. New York: Wiley-Interscience; 2002.
[32] Austin PC, Lee DS, Steyerberg EW, Tu JV. Regression trees for predicting mortality in patients with cardiovascular disease: what improvement is achieved by using ensemble-based methods? Biom J 2012;54:657−73. http://dx.doi.org/10.1002/bimj.201100251.
[33] Austin PC, Lee DS. Boosted classification trees result in minor to modest improvement in the accuracy in classifying cardiovascular outcomes compared to conventional classification trees. Am J Cardiovasc Dis 2011;1:1−15.
[34] Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med Care 2010;48(6, Suppl 1):S106−13.
[35] Maroco J, Silva D, Rodrigues A, Guerreiro M, Santana I, de Mendonca A. Data mining methods in the prediction of dementia: a real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BioMed Cent Res Notes 2011;4:1−14. http://dx.doi.org/10.1186/1756-0500-4-299.
[36] Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med 2004;34:113−27. http://dx.doi.org/10.1016/j.artmed.2004.07.002.
[37] Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. International Conference on Machine Learning. ICML '06 Proceedings of the 23rd international conference on Machine learning, 161−168. New York, NY: ACM; 2006.
[38] Caruana R, Karampatziakis N, Yessenalina A. An empirical evaluation of supervised learning in high dimensions. International Conference on Machine Learning. ICML '08 Proceedings of the 25th international conference on Machine learning, 96−103. New York, NY: ACM; 2008.