

# Logistic regression and classification and regression trees (CART) in acute myocardial infarction data modeling



Václav Faltus<sup>1</sup>, Zdeněk Monhart<sup>2</sup>

<sup>1</sup>Centre of Biomedical Informatics, ICS AS CR, v.v.i., Prague 8, Czech Republic

<sup>2</sup>Municipal Hospital Znojmo, Department of Internal Medicine, Czech Republic

faltus@euromise.cz



## Introduction

Cardiovascular risk factors and their increasing number are commonly integrated into an estimation of outcomes in patients with acute coronary syndrome. In this work we compare classification and regression trees (CART) and logistic regression in modeling the in-hospital mortality. The considered predictor variables are the five traditional risk factors (RF) (diabetes mellitus, hypertension, hyperlipidaemia, smoking and previous IM status) and six drug groups (heparin, aspirin, betablocker, statin, ACEI/ARB and thienopyridin). We also compare the predictive accuracy of logistic regression with that of regression trees.

## Methods

Our data are available on a sample of patients with acute myocardial infarction consecutively admitted to six municipal hospitals in the Czech Republic during the years 2003–2006. Our study sample is obtained by yearly retrospective chart reviews. The registry hospitals are: Čáslav, Kutná Hora and Znojmo in years 2003–2006, Jindřichův Hradec and Písek in 2004, Chrudim in years 2005–2006. All of them are non-PCI hospitals from geographically different rural regions of the Czech Republic and collaborate with different PCI centers.

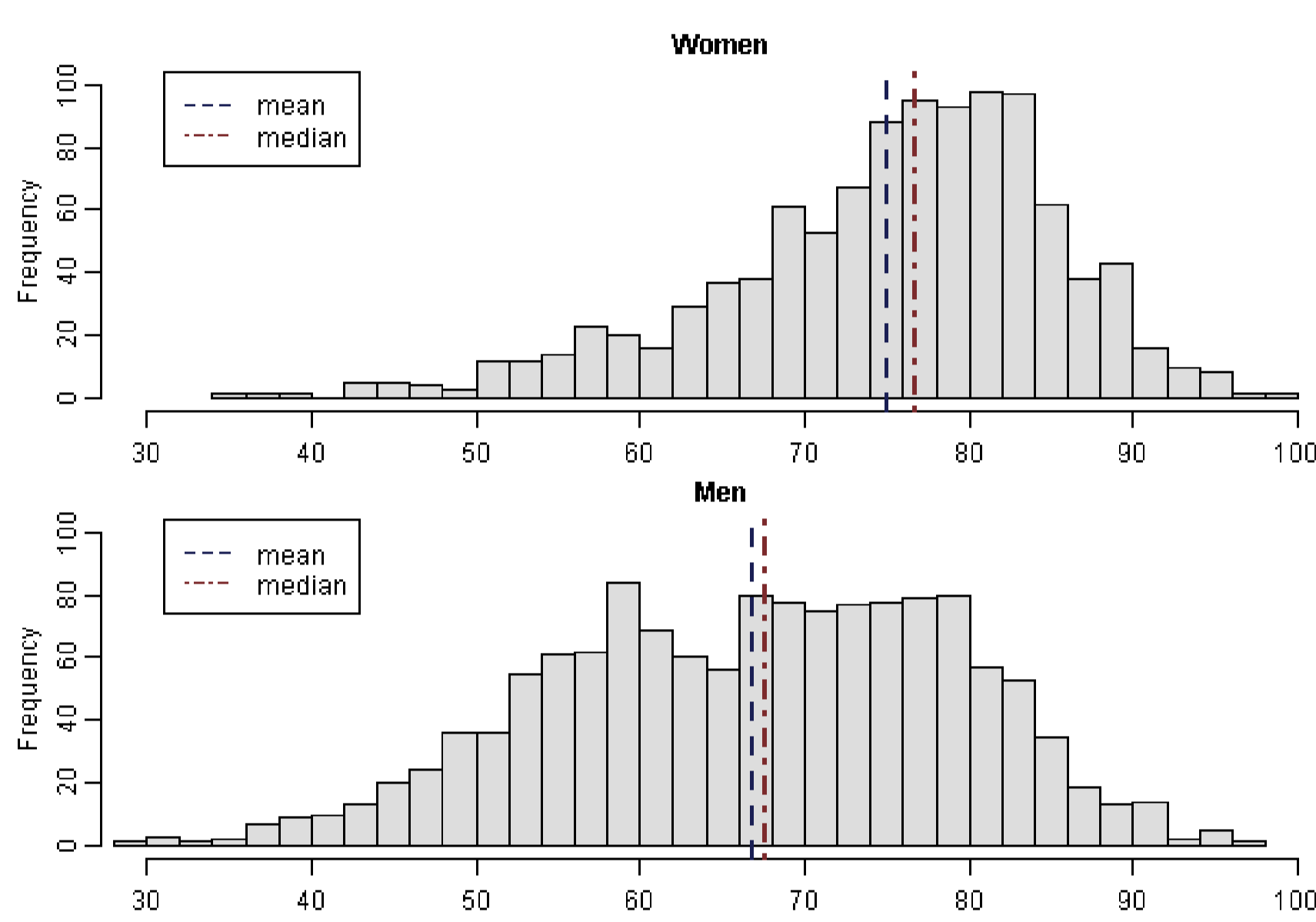


Figure 1: Histograms of age according to gender.

In total there is 2415 (244 omitted) patients with aMI in our sample. Women (1057, 43.77%), are in average older (Figure 1) than men (1358, 56.23%) and they are less frequently smokers (Figure 3) than men. As the smoking and gender variables are so highly correlated to the age variable and both often in-significant in logistic regression in-hospital mortality modeling, we do not consider gender as predictor variable and use the smoking risk factor variable only beyond the number of present risk factors. We define *rf4* the number of present risk factors (RF) as indicated in Figure 3. Because of rather large amount of missing data in *thienopyridin* variable, we do not use it when considering separate predictors. Next we define *am5* and *am6* the number of administered medications (AM) as indicated in Figure 2.

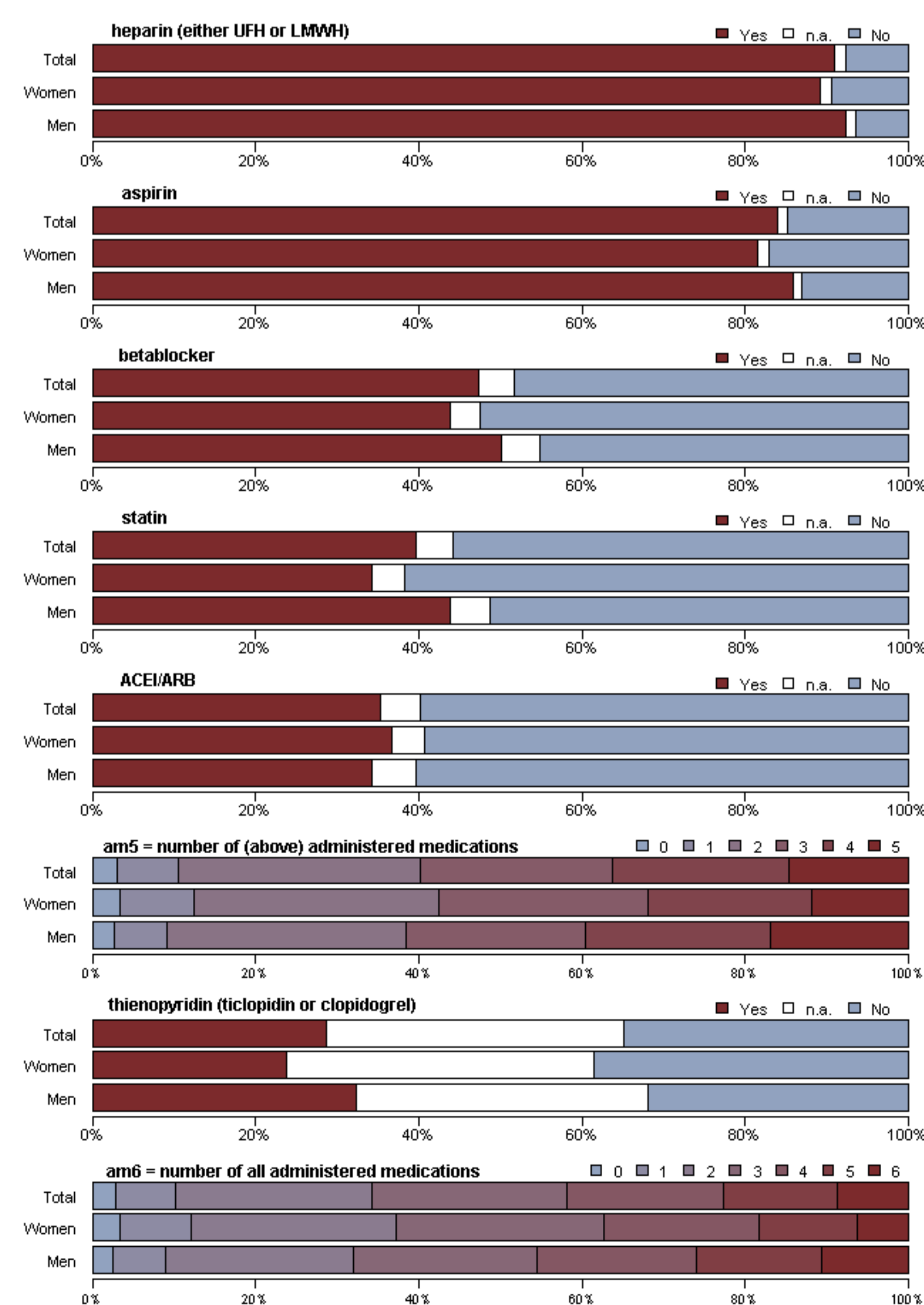


Figure 2: Acute pharmacotherapy (administered within 24 hours after admission). Plots nr. 6 and 8 represent the created predictors *am5* and *am6*, respectively.

## Comparison of predictive models

The four considered models are *CART(counts)* :  $exitus \sim age + rf4 + am6 + smoking$ , *CART(separate)* :  $exitus \sim age + all\ RF\ (except\ smoking) + all\ AM\ (except\ thienopyridin)$ , *Logistic reg.(counts)* :  $exitus \sim age + rf4 + am6 + smoking$ , and *Logistic reg.(separate)* :  $exitus \sim age + all\ RF\ (except\ smoking) + all\ AM\ (except\ thienopyridin)$ . We use repeated split-sample validation to compare the predictive accuracy of the CART and the logistic regression. The data are randomly divided into derivation and validation components. The derivation and validation samples consist of 70 % and 30 % of the data. Each model is then fit on the derivation sample and the predictions are obtained for each subject in the validation sample using the model derived on the derivation sample. The predictive accuracy of each model is summarized by the area under the ROC curve.

The model area under the ROC curve is obtained for both the derivation and validation samples. Some other characteristics of the predictive ability of the models are given too. We use the generalized  $R_N^2$  index of Nagelkerke and the Brier's score. The area under the ROC curve, the generalized  $R_N^2$  index and the Brier's score were computed using the *val.prob* function from the *Design* [1] package for *R* [2]. The regression tree models are fit the *tree* function from the *tree* [3] package.

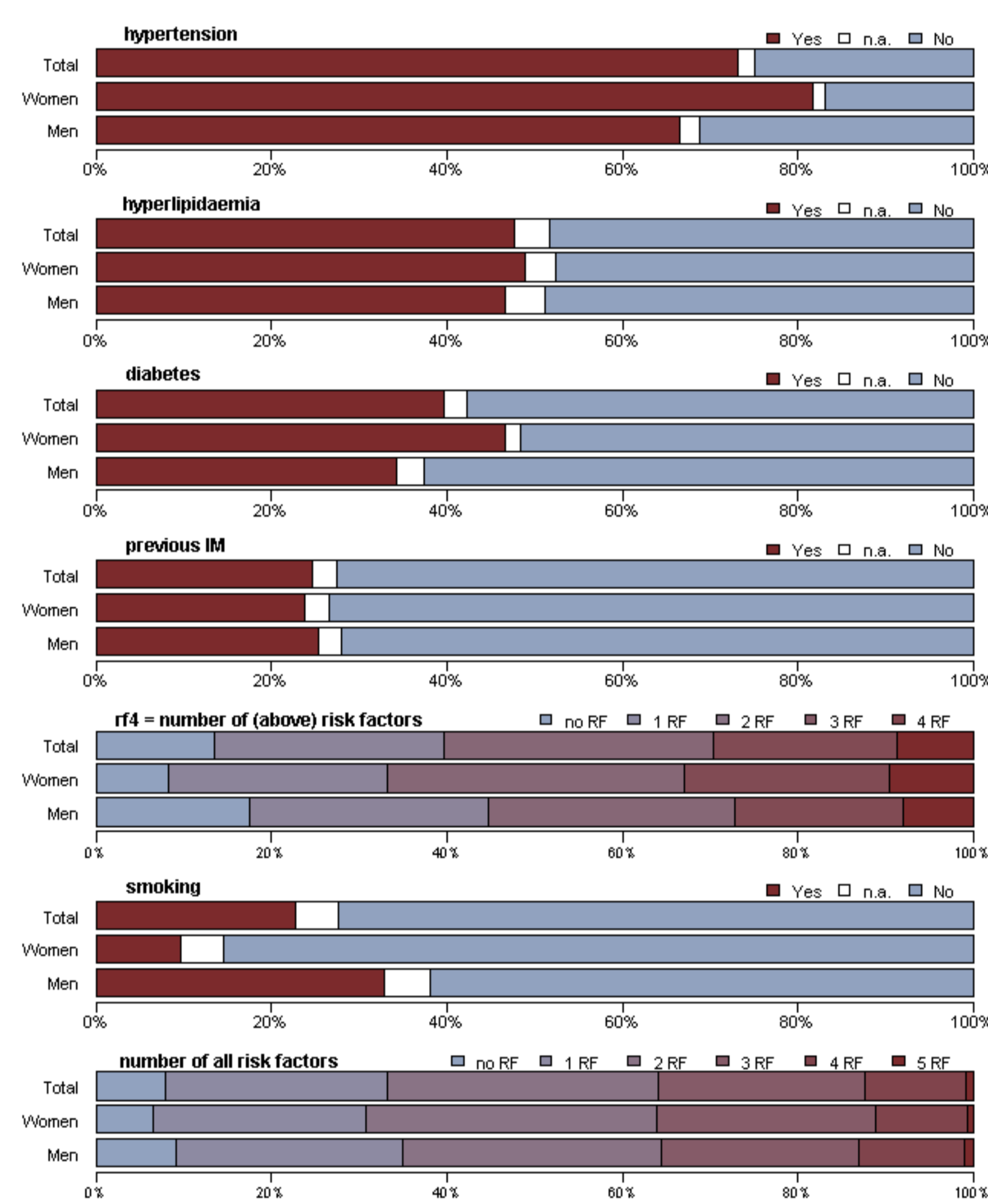


Figure 3: Traditional cardiac risk factors. Plot nr. 5 represents the created predictor variable *rf4*.

## Results

The means of area under ROC curve,  $R_N^2$  index and Brier's score, computed for the 1000 validation samples, are reported in Table 1. The mean ROC for the regression tree model using counts of administered medication and present risk factors is 0.698, while the mean ROC of the regression tree model using considered predictors separately is 0.687. The mean ROC for the logistic regression using counts is 0.782, while the mean ROC for the logistic regression using predictors separately is 0.795. Both logistic regression models clearly surpass the regression tree models in terms of predictive accuracy. The logistic regression model using predictors separately has slightly higher predictive accuracy than logistic regression model using counts. When using regression trees, the model based on counts shows higher predictive accuracy than the model based on the predictors separately.

|                           | ROC: derivation sample | ROC: validation sample | $R_N^2$ : validation sample | Brier's score: validation sample |
|---------------------------|------------------------|------------------------|-----------------------------|----------------------------------|
| CART: (counts)            | 0.716                  | 0.698                  | 0.116                       | 0.090                            |
| CART: (separate)          | 0.705                  | 0.687                  | 0.103                       | 0.090                            |
| Logistic reg.: (counts)   | 0.782                  | 0.761                  | 0.159                       | 0.079                            |
| Logistic reg.: (separate) | 0.795                  | 0.777                  | 0.179                       | 0.084                            |

Table 1: Average values of area under ROC curve,  $R_N^2$  index and Brier's score for each modeling approach.

The density estimates of the area under ROC curve,  $R_N^2$  index and Brier's score in the 1000 validation datasets for each modeling approach are given in Figure 4. Evidently, the distributions of ROC curve areas for the regression tree models

is shifted to smaller values of the ROC area, showing poorer performance compared to that of logistic regression models. Similar result was reported in [4] but in our situation the overlap of the CART and Logistic regression estimated density is larger, indicating more similar predictive performance of both approaches.

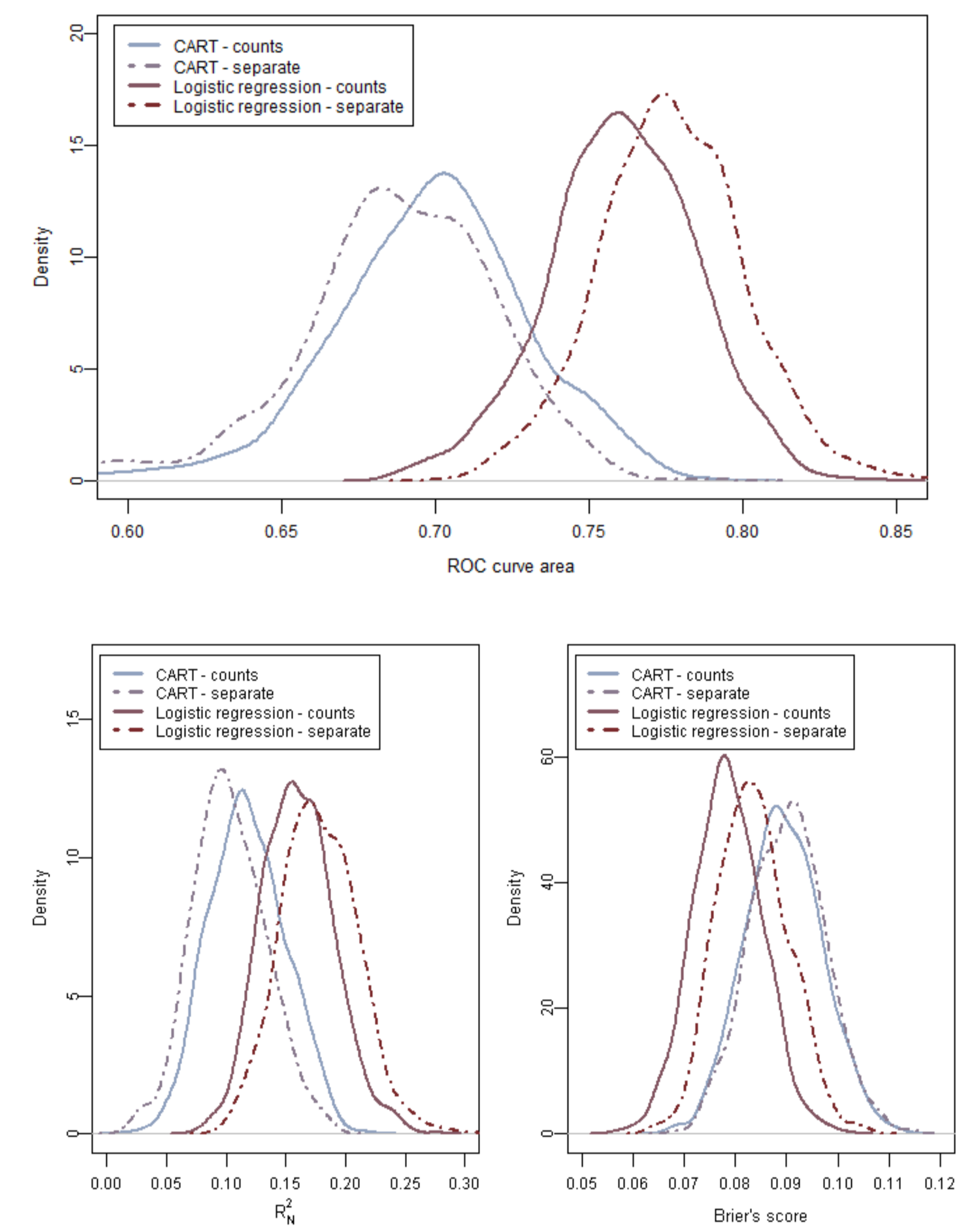


Figure 4: Density estimates of ROC area,  $R_N^2$  index and Brier's score. The displayed plots were truncated to intervals with non-zero density estimates.

## Discussion

We have demonstrated that regression tree method did not predict in-hospital mortality as accurately as did the logistic regression. The predictive performance of logistic regression was higher than that of regression tree method and this result did not depend on taking the predictor variables either as counts or separately.

Relatively better performance of logistic regression suggests that there is a linear relationship between the log-odds of in-hospital mortality and considered predictor variables. Former and current analyses of our data sample showed that there are important interactions but their inclusion did not improve the logistic regression model, thus we did not consider them. Also, the regression tree model based on counts of predictor variables showed slightly better performance than that based on predictor variables separately. It is known that regression trees have problems with capturing the additive relationships. Because of the better performance of the regression tree model based on counts of predictor variables, with respect to the model based on predictors separately, we believe that this fact also contributed to the poorer performance of regression tree models in our sample.

## References

- [1] Frank E. Harrell Jr. *Design: Design Package*. <http://biostat.mc.vanderbilt.edu/s/Design>, 2007.
- [2] R Development Core Team. *R: A language and environment for statistical computing*. *R Foundation for Statistical Computing*, 2008.
- [3] Brian Ripley. *tree: Classification and regression trees*. 2007.
- [4] Peter C. Austin. A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. *Statist. Med.*, 26: 2937–2957, 2007.
- [5] Nicola J. Crichton, John P. Hinde, and Jonathan Marchini. Models for Diagnosing Chest Pain: Is CART Helpful? *Statist. Med.*, 16: 717–727, 1997.
- [6] Leo Breiman, Jerome Friedman, Charles J. Stone, and R. A. Olshen. *Classification and Regression Trees*. *Chapman & Hall/CRC*, 1998.
- [7] Alan Agresti. *Categorical Data Analysis*, 2nd Edition. *John Wiley & Sons, Ltd.*, 2002.

## Acknowledgement

The work was supported by the grant 1M06014 of the Ministry of Education, Youth and Sports of the Czech Republic.