Survival Prediction After Heart Failure

Li Shandross and Scott Hebert

2022-12-02

Contents

Abstract

[maximum 200 words]

Introduction

Cardiovascular diseases (CVDs) are a group of disorders of the heart and blood vessels which account for roughly 17 million deaths worldwide annually. CVDs are especially prevalent in industrialized countries, yet current evaluation of the disease progression in various CVDs, especially heart failure, remains lacking. Heart failure is one type of CVD that occurs when the heart fails to pump sufficient blood to the rest of the body. Prediction of heart failure outcome is of vital importance in clinical practice throughout the world but has not yielded promising results.

A study by Chicco & Jurman highlighted the potential of machine learning methods to provide physicians with better tools to predict heart failure patient outcomes. The authors analyzed a data set of 299 heart failure patient medical records originally released by Ahmad and colleagues. Chicco & Jurman investigated ten types of machine learning methods using all predictors in the dataset and performed feature selection to determine the most important predictors. Random forests yielded the best results of the ten techniques while feature selection showed several creatinine and ejection fraction to be the best predictors. A random forests model using only these top two predictors outperformed models with all available predictors (12 total), which also included age, anaemia, high blood pressure, blood creatinine phosphokinase, diabetes, blood platelets, sex, serum sodium, and smoking status.

This project utilizes the same dataset as Chicco & Jurman, but it utilizes a Bayesian logistic regression model in order to compare Bayesian methods to the machine learning methods of the reference paper.

Methods

Choice of models and model equations

We chose to compare four total models to better examine which aspects of the Bayesian models contribute to model accuracy.

First, a Bayesian logistic regression model with all predictors was formulated:

$y_i \sim Bern(\theta_i)$
$logit(\theta_i) = \beta_0 + \beta_1 a_i + \beta_2 m_i + \beta_3 h_i + \beta_4 k_i + \beta_5 d_i + \beta_6 e_i + \beta_7 p_i + \beta_8 x_i + \beta_9 c_i + \beta_{10} s_i + \beta_{11} g_i$

Then, a model with only the two predictors mentioned by the reference paper to be most important (known henceforth as the "reduced model") was created:

$y_i \sim Bern(\theta_i)$
$logit(\theta_i) = \beta_0 + \beta_1 e_i + \beta_2 c_i$

An intercept-only model was created for reference:

$y_i \sim Bern(\theta_i)$
$logit(\theta_i) = \beta_0$

Lastly, a model with horseshoe priors was formulated as a method of variable selection:

$y_i \sim Bern(\theta_i)$
$logit(\theta_i) = \beta_0 + \beta_1 a_i + \beta_2 m_i + \beta_3 h_i + \beta_4 k_i + \beta_5 d_i + \beta_6 e_i + \beta_7 p_i + \beta_8 x_i + \beta_9 c_i + \beta_{10} s_i + \beta_{11} g_i$
$\beta_0 \sim N(0, 1)$
$\beta_j | \lambda_j, \tau \sim N(0, \lambda_j \tau)$
$\lambda_j \sim C^+(0, 1),\ j = 1, \cdots, P$

$\tau \sim C^+(0, \tau_0)$ where $\tau_0 = \frac{p_0}{P - p_0} \frac{\sigma}{\sqrt{n}} = 0.025$
σ is approximated with pseudo variance $\tilde{\sigma}^2 = 1/\mu(1-\mu) = 2.142$ for a non-gaussian link

(Horseshoe priors are described further in the Horseshoe priors subsection.)

where:

ai refers to patient age in years,

mi refers to the presence of anaemia,

hi refers to the presence of high blood pressure,

ki refers to blood creatinine phosphokinase level in mcg/L,

di refers to the presence of diabetes,

ei refers to ejection fraction (percentage of blood leaving the heart upon each contraction),

pi refers to blood platelets in kiloplatelets/mL,

xi refers to sex (M/F),

ci refers to serum creatinine in mg/dL,

si refers to serum sodium in mEq/L,

gi refers to whether the patient smokes

Model implementation

Models are implemented using the packages `brms` and `rstanarm` with additional model checks performed using `arm`, `tidybayes`, and `bayesplot`. The horseshoe prior model is fit and checked using functions from `rstanarm` (unlike the other three models) because `brms` only supports the use of horseshoe priors for linear regression, not logistic regression.

Horseshoe priors

The horseshoe is a type of Bayesian prior (developed by Piironen and Vehtari) that serves as a shrinkage method to improve model fit. This prior is named for its U-shape that resembles a horseshoe and determines the constraints on coefficient estimates. Coefficients associated with predictors weakly supported by the data are shrunk very close to zero while coefficients more strongly supported by the data experience minimal shrinkage.

We chose to explore usage of the horseshoe prior for several reasons. First, results from the reference paper showed that the models with ejection fraction and serum creatinine as the only predictors outperformed models using all predictors. This suggests that constraining coefficient estimates of unimportant predictors may yield better prediction accuracy. Second, as the horseshoe prior only shrinks coefficients of unsupported variables towards zero, it provides an interesting compromise between the reduced model and the full model described above. Third, the horseshoe prior was out of the scope of the Applied Bayesian Modeling class, and this project serves as an opportunity to learn how to apply a new type of prior.

When specifying a horseshoe prior, it is necessary to make a prior guess at the number of relevant variables. Based on the work from Chicco & Jurman, we set this value equal to two. Other parameter values such as local and global degrees of freedom, global scale, etc. are chosen based on recommendations from "Sparsity information and regularization in the horseshoe and other shrinkage priors" by Piironen and Vehtari.

Initial checks and validation

Before running the models on the full dataset, we first performed an initial test of each model on a smaller sample of the data with few iterations for some preliminary validation. The models passed these initial checks, allowing us to proceed with the chosen four models.

We began fitting the models using 1000 iterations with 500 as warm up, spread across four chains. However, this created warnings of divergence and low bulk ESS for the horseshoe prior model, though other MCMC diagnostics like Rhat values of 1.0 and n_eff values greater than 225 indicated increasing warm-up and the number of iterations would be sufficient to fix the problem. We chose to manually increase certain defaults for all of the models for consistency; for example, we used 1,000 warmup iterations and set maximum tree depth to 20 (increased from the default value of 10).

In-sample checks

After the models were tuned and finalized, we performed in-sample checks. This included binned residual plots plotted against ejection fraction and serum creatinine (the two variables deemed most important by the reference paper the horseshoe prior model). A log transformation was completed on the serum creatinine variable in these plots for readibility.

Following our assessment of residuals, we performed posterior predictive checks using three summary statistics. These test quantities were created to evaluate any discrepancies between the model simulations and the true data in terms of predicted survival proportions under three scenarios: overall, among patients with a normal ejection fraction, and among patients with normal serum creatinine levels. A healthy range for ejection fraction is defined as greater than 40% in the paper by Chicco & Jurman while a healthy serum creatinine level is less than 1.2 mg/dL (Mayo Clinic). The test statistics are labeled as follows:

- T1: Proportion of survival
- T2: Proportion of survival among patients with an ejection fraction $> 40\%$
- T3: Proportion of survival among patients with serum creatinine < 1.2 mg/dL

Out-of-sample checks

We then perform leave-one-out (LOO) cross-validation on the four models to check for influential points, goodness of fit, and model comparison. We do not include a PSIS-LOO probability interval transform density because we are performing logistic regression, not linear regression, and we find issues with using this comparison.

Model comparison

To compare model prediction accuracy, we use the ELPD values from the LOO cross-validation, along with evaluation metrics suited for binary outcome data like Matthew's Correlation Coefficient (MCC), true positive rate, true-rate, accuracy, and ROC area under the curve. These additional metrics are taken from the reference paper and allow for comparison against Chicco & Jurman's machine learning models. The resulting values are calculated for models fit using all 299 observations, with 4000 draws from the posterior predictive distribution, metrics calculated for each draw and averaged. These are essentially training metrics since a model has seen all of the data, meaning accuracy may be somewhat inflated compared to the same metrics calculated on a test set.

Results

Fitted models

We obtain the following model fits for our four Bayesian logistic regression models.

Intercept-only model

$y_i \sim Bern(\theta_i)$
$logit(\theta_i) = -0.75$

Full model

$y_i \sim Bern(\theta_i)$
$logit(\theta_i) = 10.78 + 0.05a_i - 0.01m_i - 0.12h_i + 0.00k_i + 0.15d_i - 0.08e_i - 0.00p_i - 0.57x_i + 0.72c_i - 0.07s_i - 0.02g_i$

Reduced model

$y_i \sim Bern(\theta_i)$

Horseshoe model

$y_i \sim Bern(\theta_i)$
$logit(\theta_i) = 3.94 + 0.03a_i - 0.00m_i - 0.00h_i + 0.00k_i + 0.00d_i - 0.07e_i - 0.00p_i - 0.00x_i + 0.60c_i - 0.00s_i - 0.00g_i$

Horseshoe prior-induced shrinkage

As shown above by the fitted model, the horseshoe prior successfully shrunk most variables' coefficients to zero, except for the most relevant ones supported by the data. These important predictors included serum creatinine, ejection fraction, and age. While we provided the initial guess of two relevant predictors, the addition of age remains consistent with the reference paper results, which showed age as a potential third most important predictor. The shrinkage described is shown in the figure below, with the intercept term separate from the predictors given a difference in scale.

In-sample checks

The binned residuals plots generally looked okay, suggesting the current models were sufficient to proceed with, though the no predictor model may have an issue of over predicting survival at low ejection fraction levels (see Table 3 in the Appendix).

All four models showed very accurate predictions for survival proportion without restriction on type of patient (T1). However, predictions for survival among patients with a normal ejection fraction (T2) was overestimated by every model except for the intercept-only model, which under predicted survival for this group. However, the horseshoe prior model had the best posterior predictive p-value of 0.766 by about 0.07. For T3, the proportion of survival for patients with normal serum creatinine levels, all models performed more similarly by underestimating survival proportion with all predictive p-values less than 0.25.

Out-of-sample checks

Leave-One-Out (LOO) cross-validation was performed on all of the models with the following results. Aside from 3 out of the 299 examples in the full model and 1 example in the horseshoe model which were defined as "okay" (< 0.7), all of the validations ended up with every value being "good" (< 0.5). Additionally, the horseshoe prior model had the best ELPD (-122.1) of our four models, followed closely by the full model (-125.1). These best two models had ELPD values more than two standard deviations less than the reduced and intercept-only models.

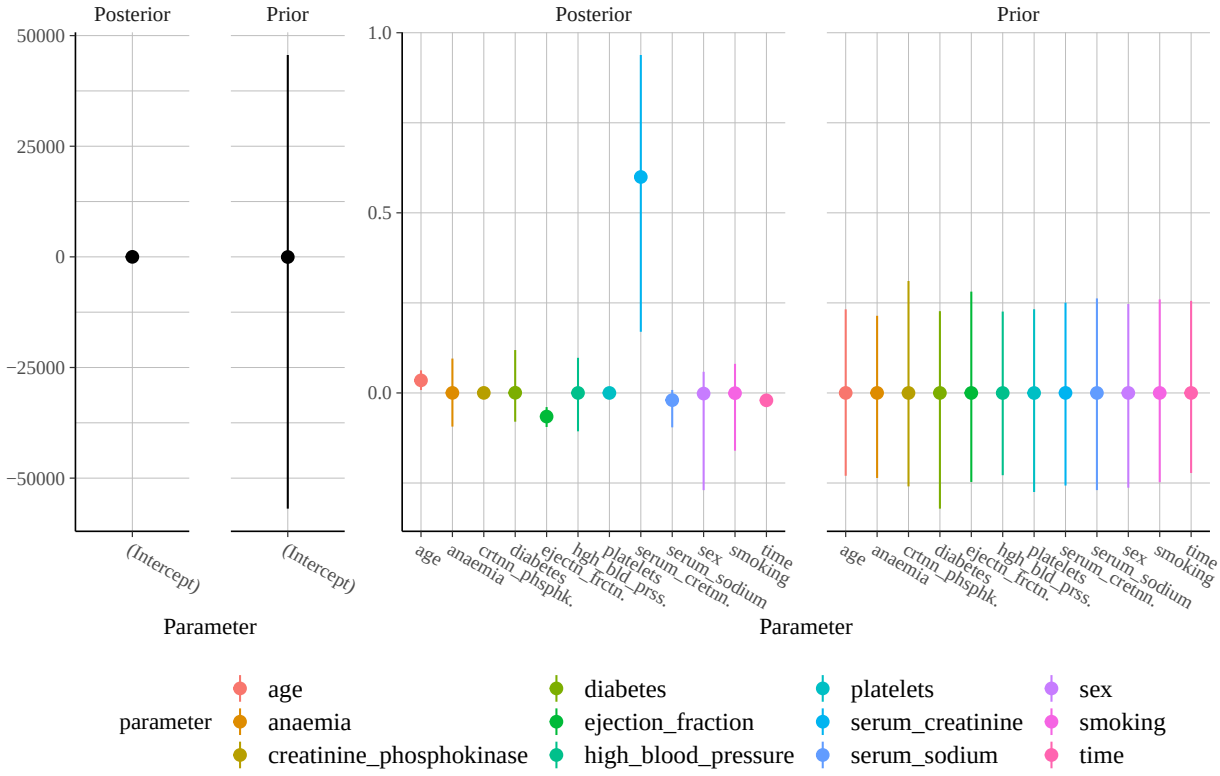

Figure 1: Posterior versus prior comparison for horseshoe prior model that shows parameter shrinkage.

Model comparison

Unlike with most of our previous in-sample and out-of-sample checks, the full model performs slightly better than the horseshoe prior model in terms of the prediction accuracy metrics shown in the table below. This is a little surprising, especially since it inconsistent with the findings from Chicco & Jurman. However, this may simply be a quirk of essentially only evaluating "training" set results, with test set results being consistent with the reference paper. Alternatively, Bayesian modeling may simply perform better with more predictors on this dataset compared to machine learning.

Comparisons against the top machine learning model's prediction accuracy show that both the horseshoe prior and full models perform competitively against this random forests model, beating it for all metrics shown except for true negative rate (and ELPD which are not given). As seen in Table 1 (and Table 2 in the Appendix), the gains in more often correctly identifying true positives seem to be the source of higher accuracy, AUC, and MCC values. Once again, though, our models' metric values are essentially training set results, which may be higher than those of test set results like that of the machine learning models.

Table 1: Survival prediction results of all models - mean of 4000 posterior samples

model	elpd	mean_mcc	mean_tpr	mean_tnr	mean_accuracy	auc
Horseshoe prior	-122.1	0.435	0.616	0.818	0.753	0.894
Full	-125.1	0.465	0.638	0.827	0.766	0.898
Reduced	-165.8	0.176	0.441	0.735	0.640	0.762
Intercept-only	-188.6	0.000	0.322	0.678	0.564	0.500

Conclusion and discussion

Appendix

Source code

Supplemental figures

Table 2: Survival prediction results of two-predictor reference paper models - mean of 100 iterations

	mcc	tpr	tnr	accuracy	auc
Random forests	0.418	0.541	0.855	0.585	0.698
Gradient boosting	0.414	0.550	0.845	0.585	0.792
SVM radial	0.348	0.519	0.816	0.543	0.667

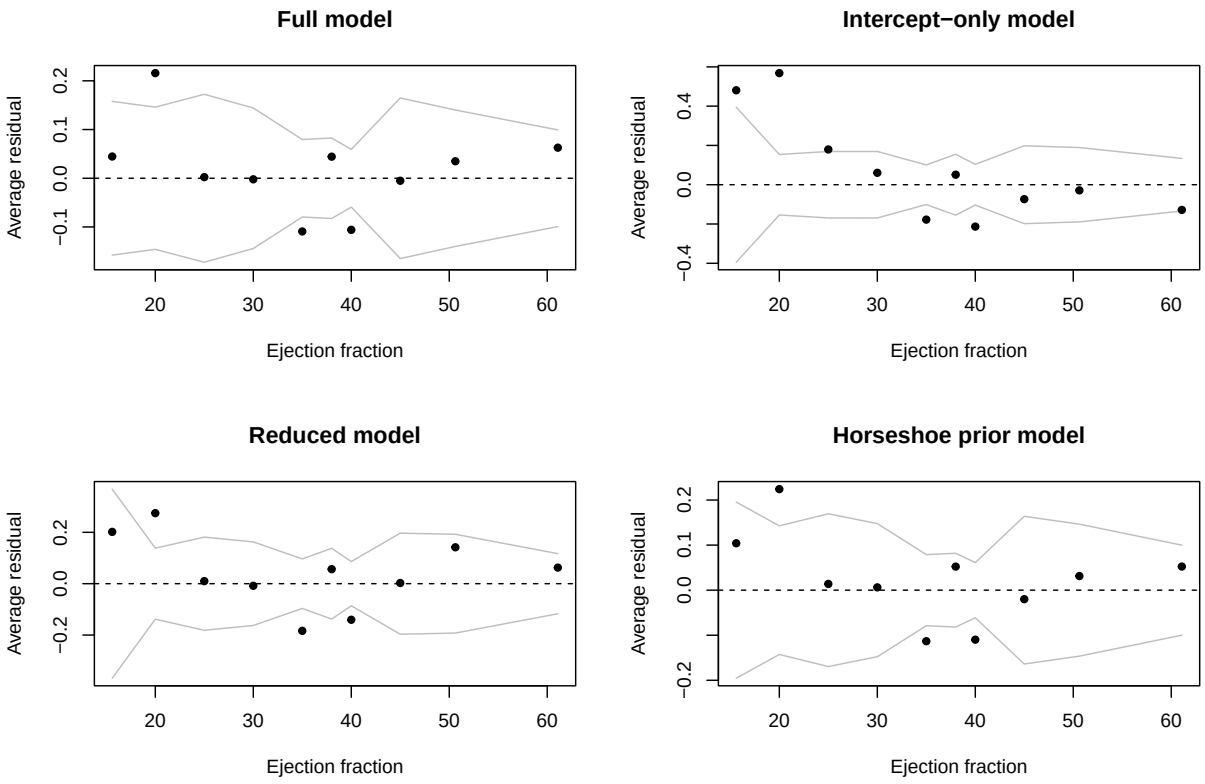

Figure 2: Binned model residuals plotted against ejection fraction

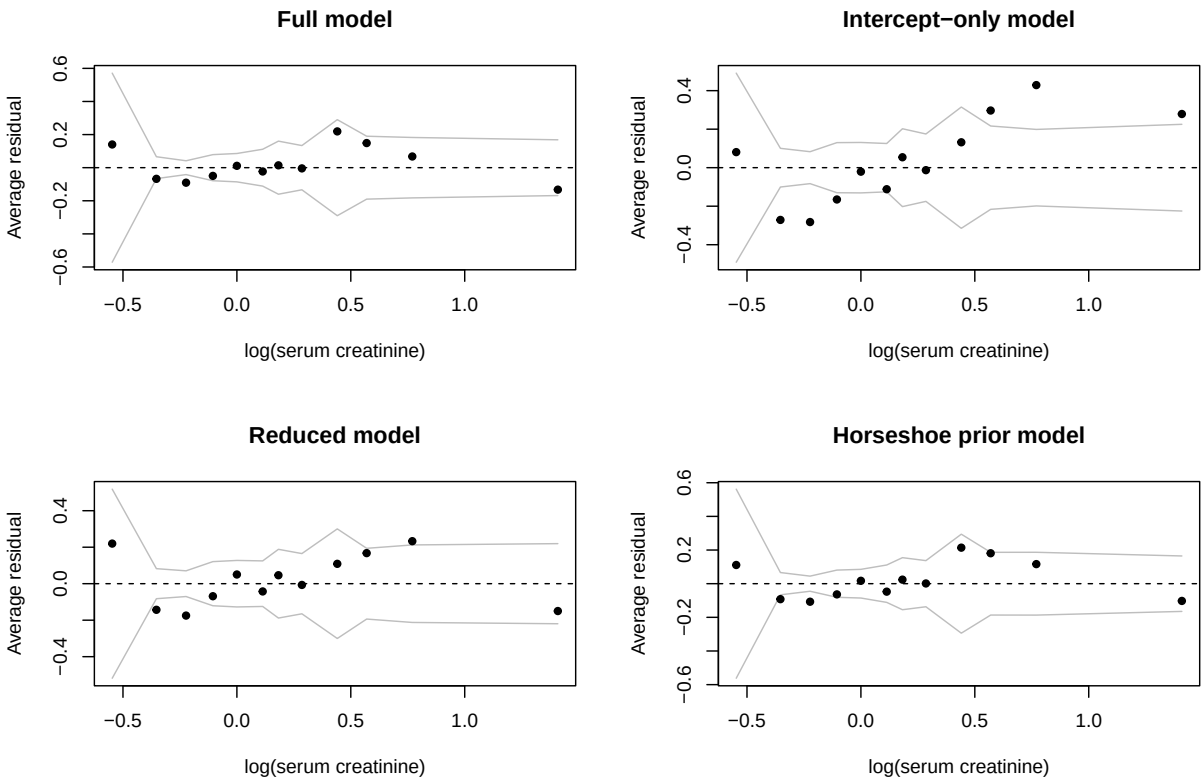

Figure 3: Binned model residuals plotted against log serum creatinine

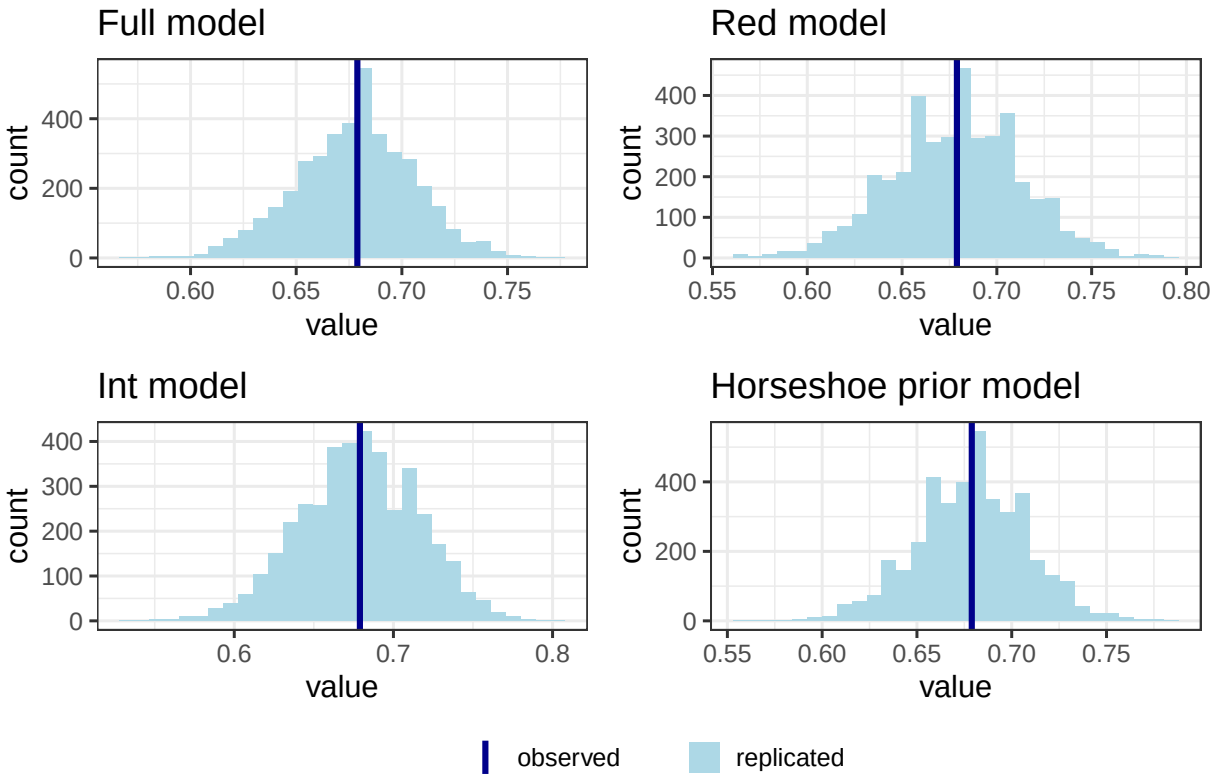

Figure 4: Survival proportion of all patience

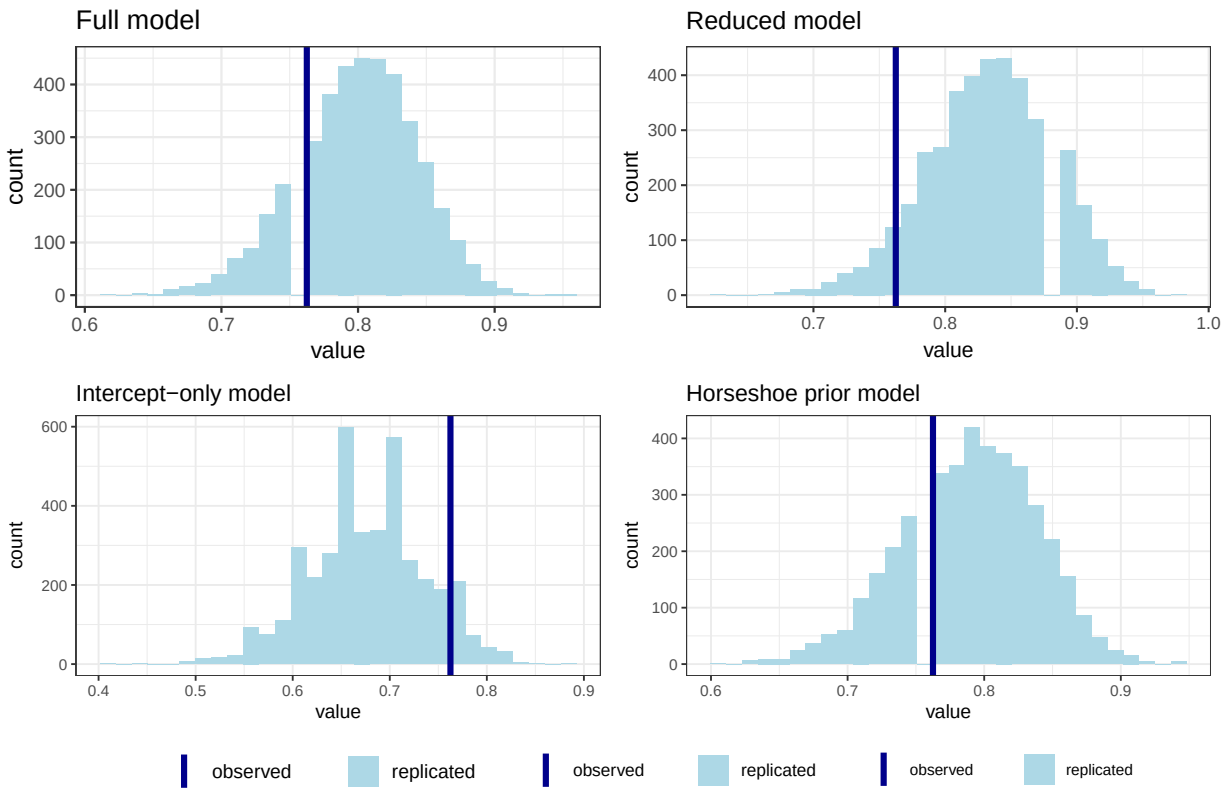

Figure 5: Survival proportion of patients with normal ejection fraction (over 40%)

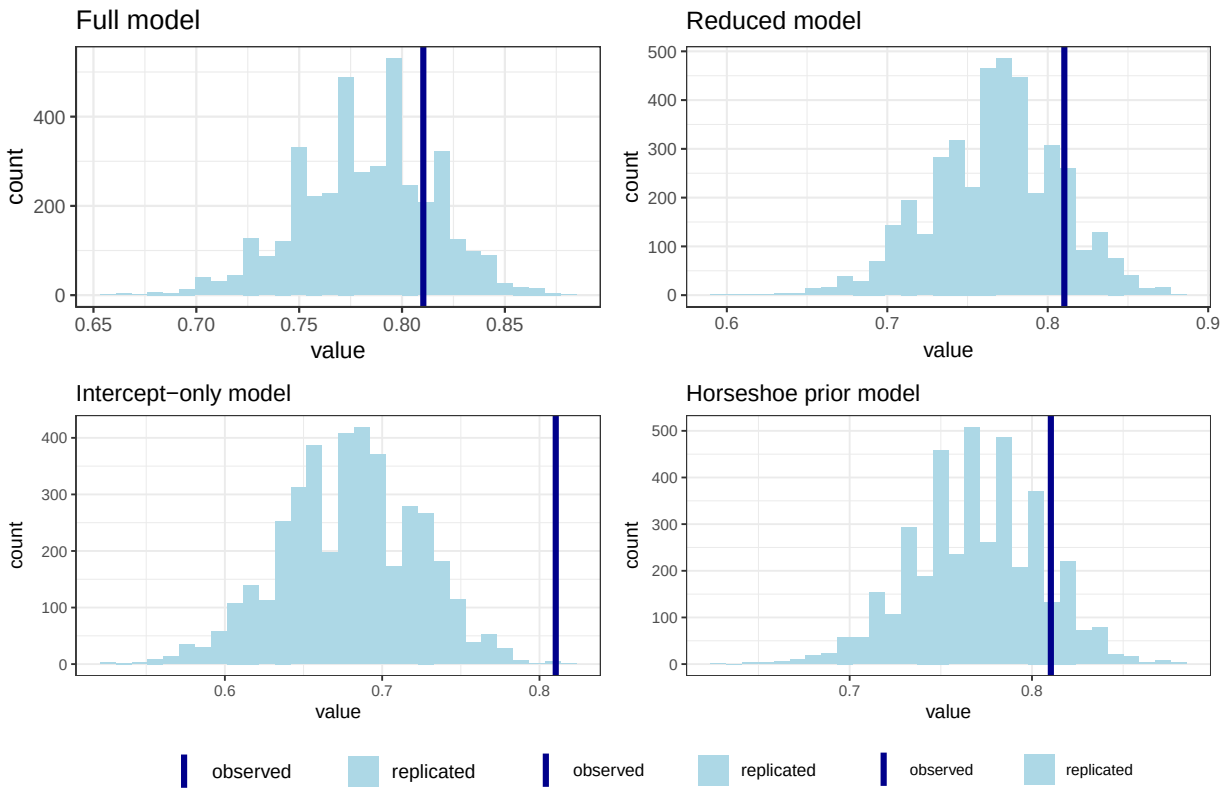

Figure 6: Survival proportion of patients with normal serum creatinine levels (under 1.2 mg/dL)

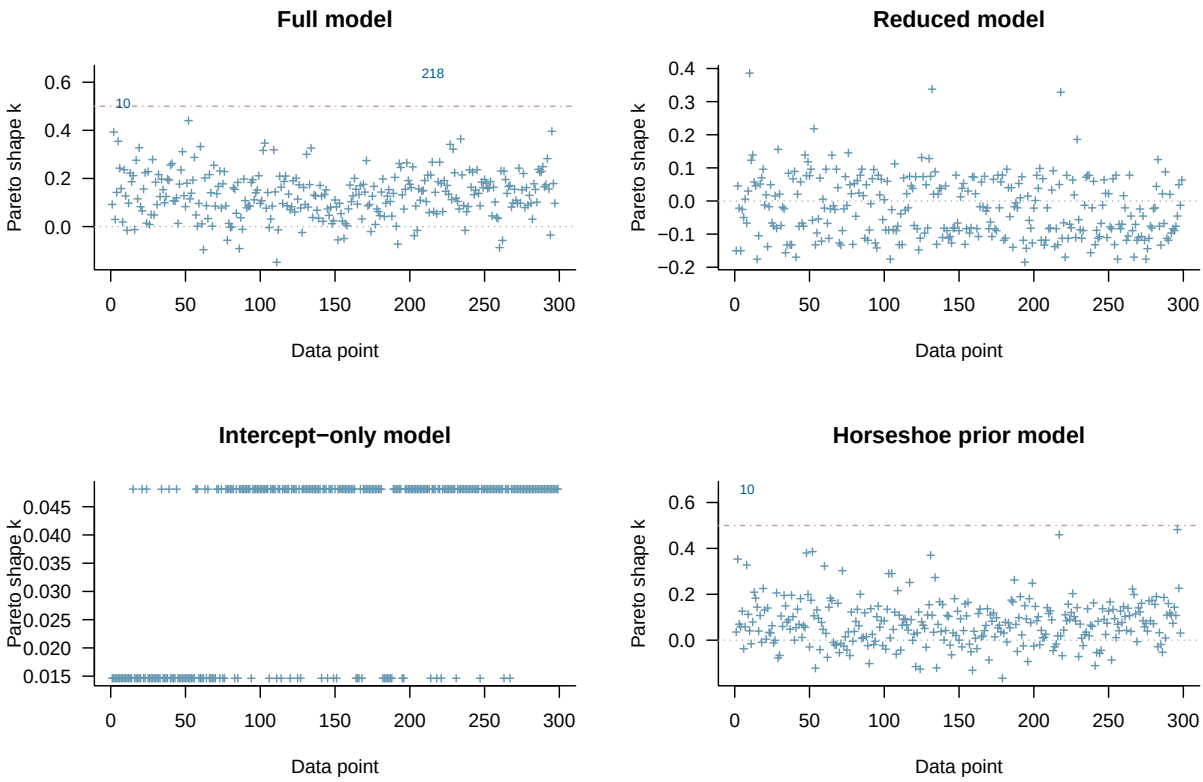

Figure 7: PSIS diagnostic plots for all models

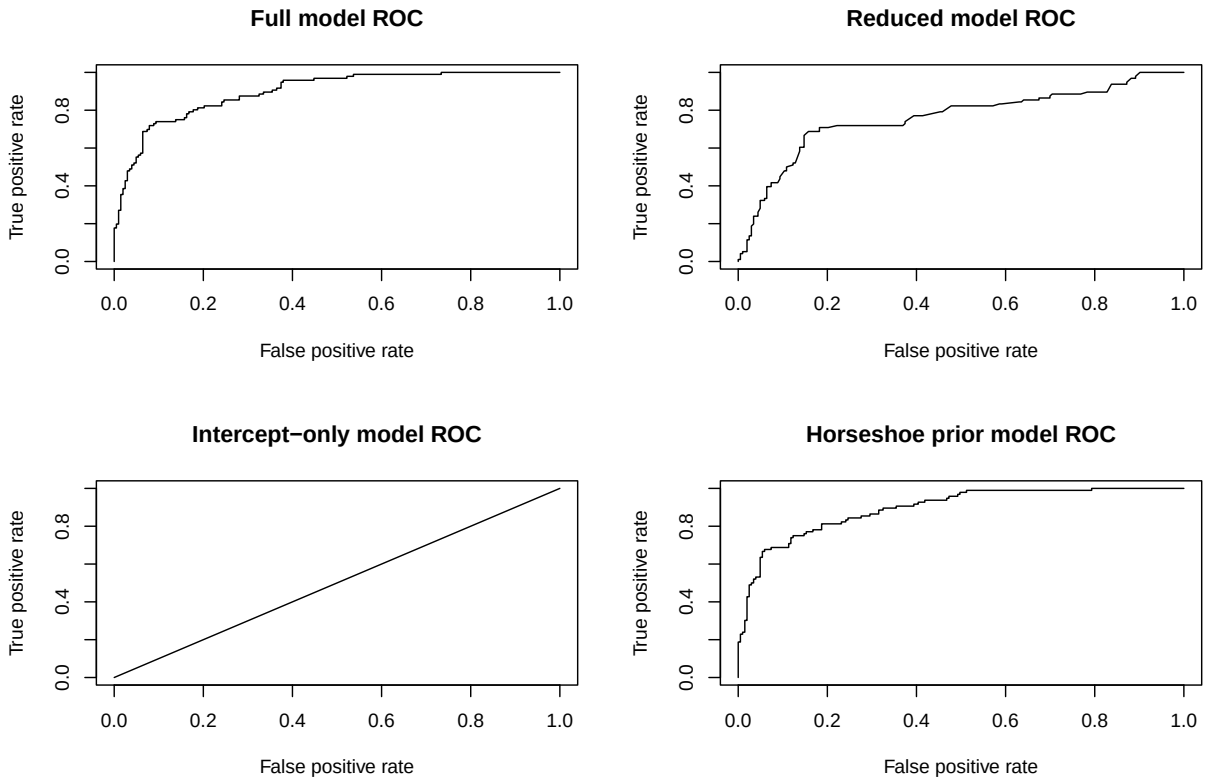

Figure 8: Receiver-Operating Characteristic (ROC) Curves for all models