Research Article

Statistics
in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Web-based Supporting Material for "A comparative study of variable selection methods in the context of developing psychiatric screening instruments" by F. Lu and E. Petkova

Feihan Lu[a], Eva Petkova[bc]

1. Table of Contents

2. eResults

For the development of psychiatric screening instruments based on items in existing questionnaires, various statistical methods can be applied, including *t*-test, CART, Random Forest, LASSO, Elastic Net. In addition, we propose a new method Imputed-LASSO, which combines Random Forest imputation and LASSO, when there are missing data. We

[a] *Department of Statistics, Columbia University, New York, NY 10027.*
[b] *Department of Child and Adolescent Psychiatry, New York University, New York, NY 10016.*
[c] *Nathan S. Kline Institute for Psychiatric Research, Orangeburg, NY 10962.*
[*] *Correspondence to: Room 1005 SSW, MC 4690, 1255 Amsterdam Avenue, New York, NY 10027, USA. E-mail: fl2238@columbia.edu*

Contract/grant sponsor: NIMH grants R01 MH089390-01 and RC MH089721-01

performed a comprehensive simulation study in order to investigate the performance of these six variable/item selection methods with respect to variable selection and future data prediction. Our simulations centerer around three key aspects: (1) the characteristics of predictors (i.e., correlations and interactions among predictors and unobservability of certain true variables); (2) pattern of missing feature data (i.e., complete data, equal proportion of missingness across predictors (**M1**), and unequal proportion of missingness (**M2**) and (3) prevalence of cases in the training data set (i.e., 0.5 and 0.3). In this web-based supplementary material, we present the plots extracted from the original paper and discuss the corresponding results.

2.1. Prediction Performance for M2 with Equal Prevalence

Figure 1 shows the performance with respect to prediction of the methods, in the case when data with equal prevalence of cases and non-cases and under unequal proportion of missing feature case **M2**. The results shares similarity with that under equal prevalence of missingness across variables **M1**. Best prediction was obtained with two-sample *t*-test under scenarios C3, C4 and C5, and with Imputed-LASSO under other scenarios. This shows the advantage of two-sample *t*-test and Imputed-LASSO in future data prediction under this condition. Note that Random Forest uses all predictors in the data set to do prediction, and thus its prediction performance is only a benchmark and not to be compared with other methods.

2.2. Selection Performance for Complete data and M1 with Unequal Prevalence

Under the condition of unequal prevalence of cases and non-cases, linear models (GLM, LASSO and Elastic Net), CART and Random Forest selected true variables with less probability than equal prevalence situations, and the reduction was especially significant for CART. Two-sample *t*-test and Imputed-LASSO, in the contrary, were not affected severely by missingness or unbalanced prevalence and they still selected variables well for both complete data and **M1**. Figure 2 and 3 show the selection results in this situation.

2.3. Prediction Performance for Complete data and M1 with Unequal Prevalence

For both complete data and **M1** data, the prediction performance of the methods were strongly affected by unbalanced prevalence in that further reduction in error rate was limited. For complete data, best prediction was achieved most of the time with Elastic Net; For **M1**, Imputed-LASSO outperformed other methods in most scenarios, suggesting the merit of Imputed-LASSO not only in producing good variable selection, but also in providing good future data prediction under this case. Figure 4 and 5 show the boxplots of predictive errors and AUC's for complete data and **M1** with unequal prevalence of cases.

2.4. Selection performance for Imputed-GLM, Imputed-LASSO and Imputed-Elastic Net under with **M2** *with equal prevalence*

For incomplete data (**M1** and **M2**), after Random Forest imputation, GLM or Elastic Net could also be applied instead of LASSO. To comparing these three approaches, a small simulation study in the case of equal case prevalence and **M2** missingness is carried out. The imputed versions of these regularized regerssion methods behaved similarly to their performance on complete data. Figure 6 and 7 show their performance with respect to variable selection and future data prediction.

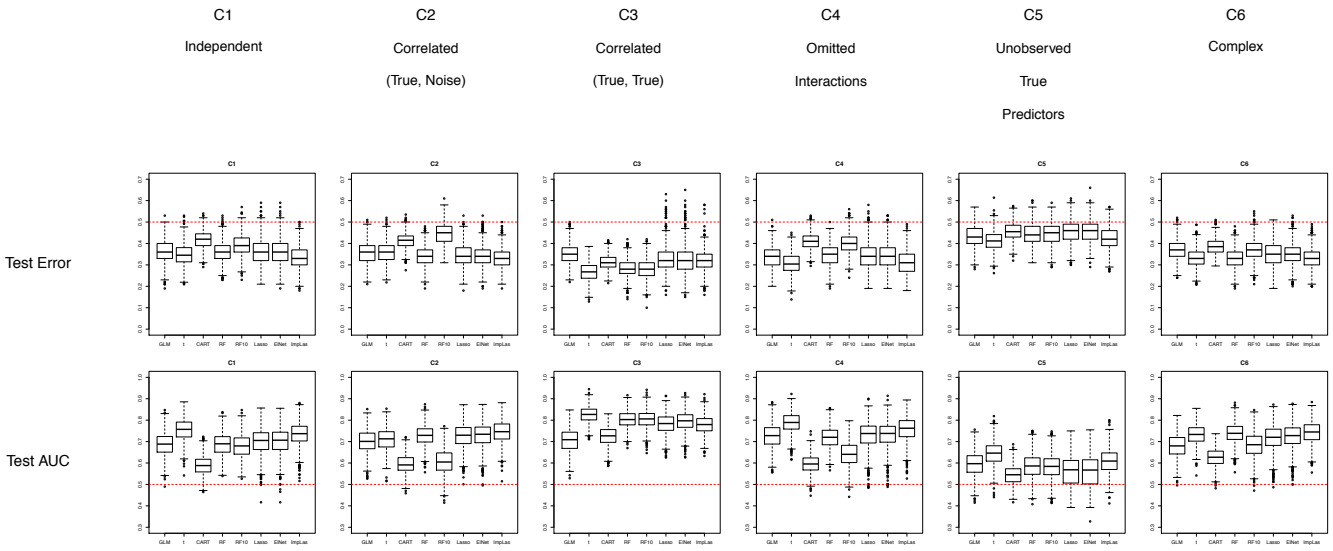

Figure 1. Unequal proportion of missing data in predictors **M2** and equal prevalence of cases and non-cases in the training data set: performance of the methods with respect to misclassification error and AUC in the test data set under the six scenarios C1 to C6. *RF* refers to Random Forest; *RF10* refers to Random Forest with 10 variables selected; *ElNet* refers to Elastic Net; *ImpLas* refers to Imputed-LASSO. Note: Random Forest should not be compared with other methods.

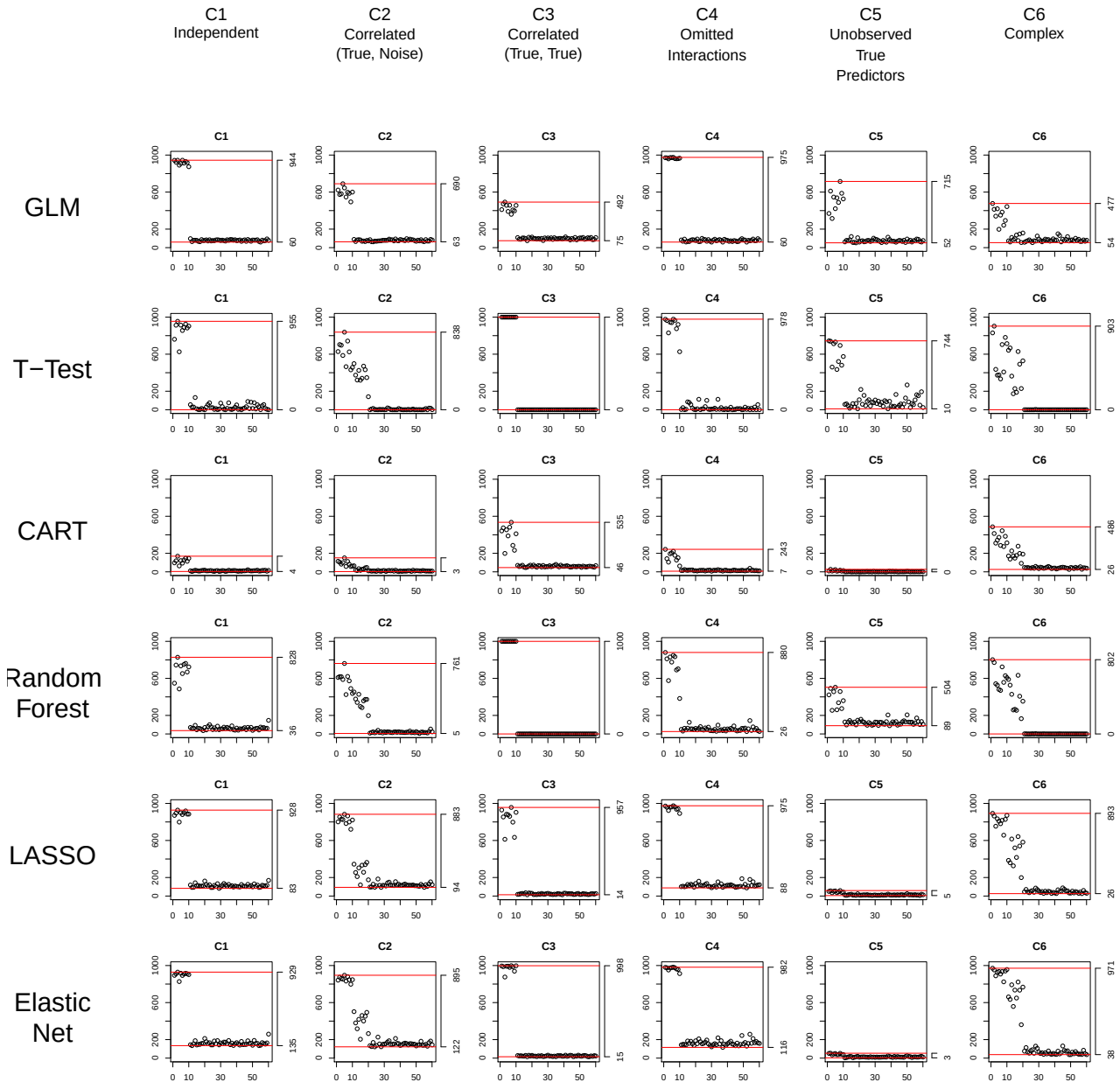

Figure 2. Complete data and unequal prevalence of cases and non-cases in the training data set: performance of the methods with respect to selecting correct variables under the six scenarios C1 to C6.

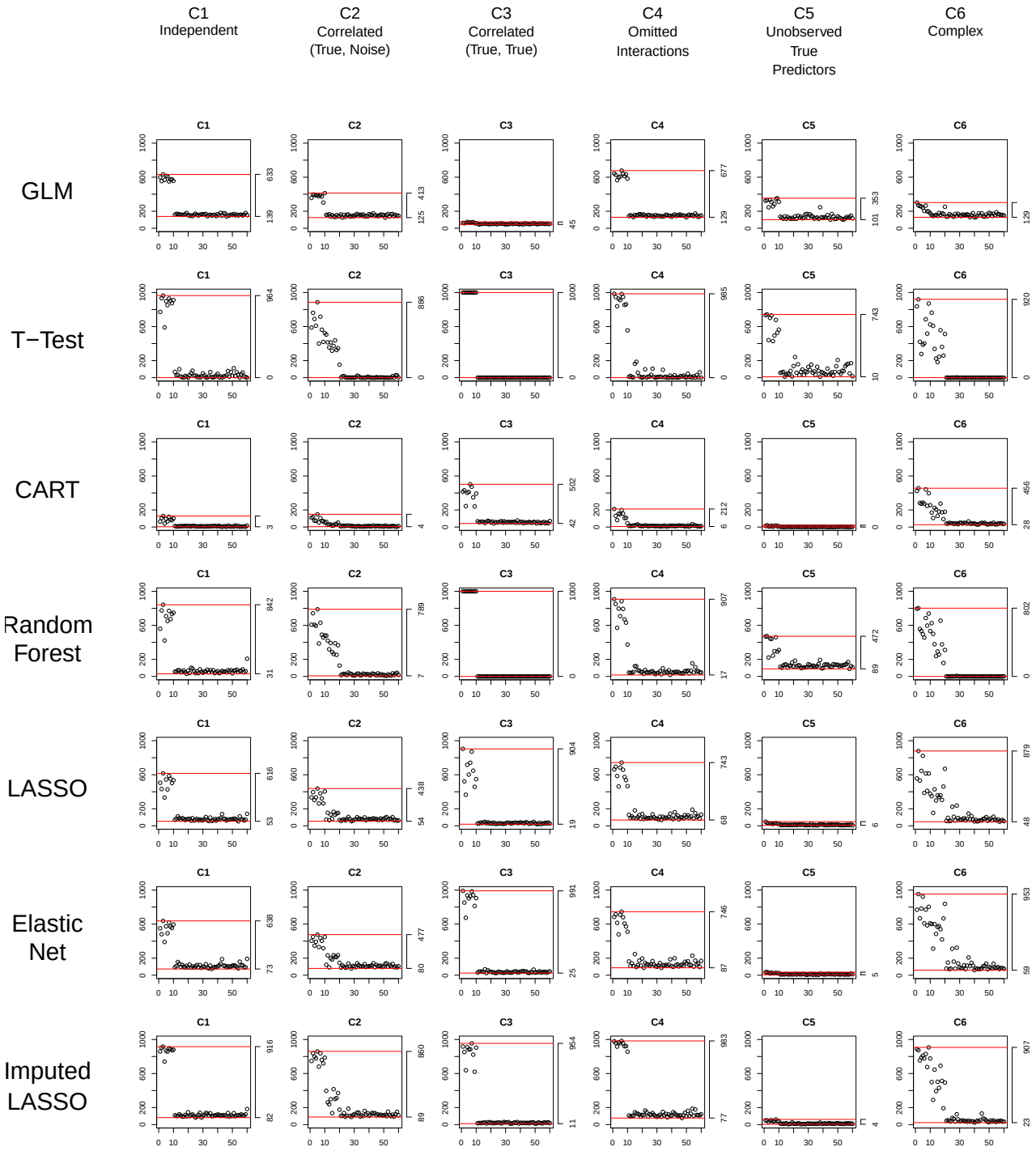

Figure 3. Equal proportion of missing data in predictors **M1** and unequal prevalence of cases and non-cases in the training data set: performance of the methods with respect to selecting correct variables under the six scenarios C1 to C6.

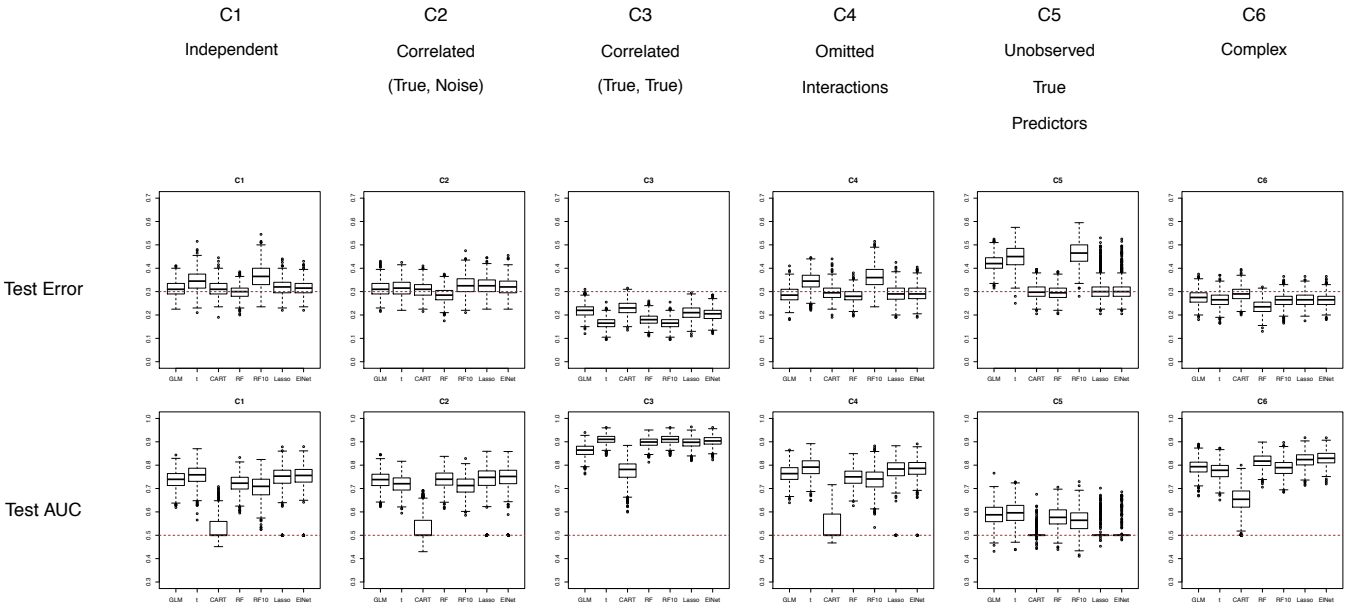

Figure 4. Complete data and unequal prevalence of cases and non-cases in the training data set: performance of the methods with respect to misclassification error and AUC in the test data set under the six scenarios C1 to C6. *RF* refers to Random Forest; *RF10* refers to Random Forest with 10 variables selected; *ElNet* refers to Elastic Net. Note: Random Forest should not be compared with other methods.

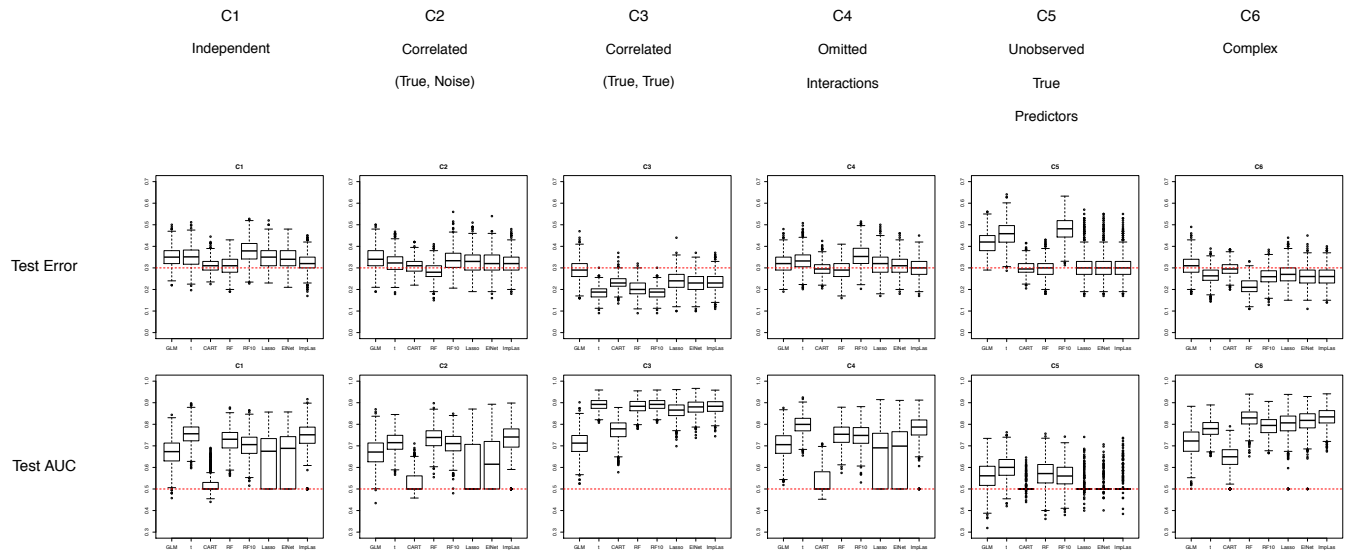

Figure 5. Equal proportion of missing data in predictors **M1** and unequal prevalence of cases and non-cases in the training data set: performance of the methods with respect to misclassification error and AUC in the test data set under the six scenarios C1 to C6. *RF* refers to Random Forest; *RF10* refers to Random Forest with 10 variables selected; *ElNet* refers to Elastic Net,*ImpLas* refers to Imputed-LASSO. Note: Random Forest should not be compared with other methods.

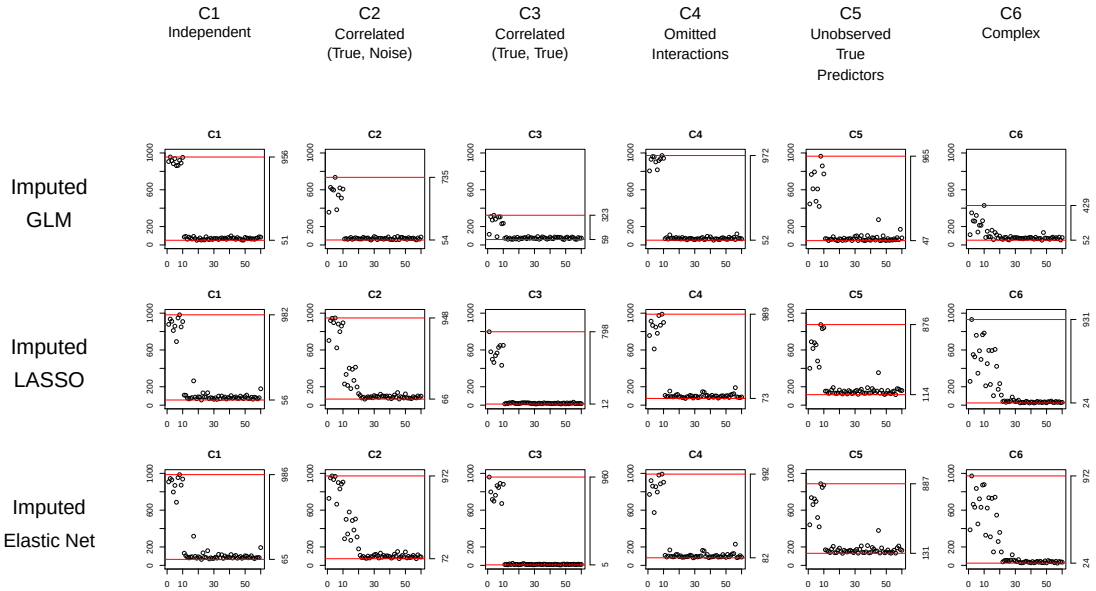

Figure 6. Selection performance for Imputed-GLM, Imputed-LASSO and Imputed-Elastic Net under equal proportion of missing data in predictors **M2** and equal prevalence of cases and non-cases in the training data set.

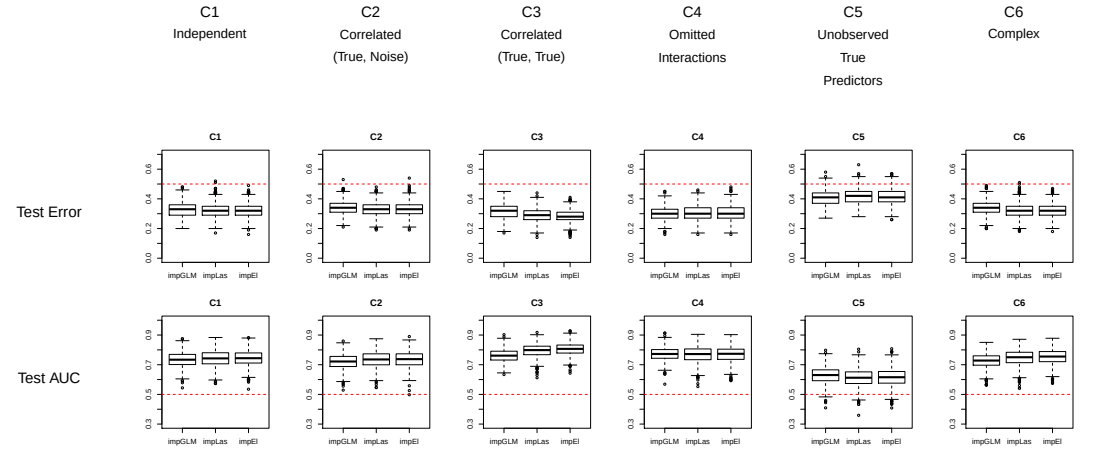

Figure 7. Prediction performance for Imputed-GLM, Imputed-LASSO and Imputed-Elastic Net under equal proportion of missing data in predictors **M2** and equal prevalence of cases and non-cases in the training data set.