**READ ME FIRST**

**CODE REPOSITORY for ‘Selecting Biomarkers for building optimal treatment selection rules using Kernel Machines’**

The repository folder contains three subfolders (along with the readme file).

The first subfolder (‘Functions’) contains all the functions required for running the analyses in R. This folder contains the following files:

1. exampleFun3 – contains function ‘kernlab.svm.new’ that runs the support vector machines procedure; ‘predict.svm.new’ that predicts class labels based on the fitted model; ‘caltheta.weights.l0’ and ‘caltheta.weights’ that computes theta (see main draft) with or without the biomarker cost.
2. predict\_fit – contains functions ‘predict.fit’ and ‘predict.fit.ind’ (the second one is applicable only for RFE while the first one is for the rest) that predicts performance of a given feature selection method.
3. Cv\_lambda – contains the function ‘cv.svm’ to run cross validation on the global SVM parameters.
4. global\_cv – contains the function ‘global.cv’ which is a generic CV method applicable to our feature selection to select the optimal tuning parameter for L0 or L1 cost in each relevant method (see draft).
5. decisionList – contains functions to run the decision list method of Zhang (2015).
6. scad – contains functions ‘scad.wsvm’ and ‘escad.wsvm’ to run SCAD and eSCAD on WSVM.
7. ssvc – contains function ‘zero.norm.svm’ to run the L0 SVM on WSVM (see draft).
8. rfe\_loss – contains function ‘rfe.loss’ to run the riskRFE procedure on WSVM (see draft).
9. l1svm – contains function ‘l1.svm’ to run the L1 SVM procedure on WSVM (see draft).
10. l0\_weston – contains function ‘l0.weston’ to run AROM on WSVM (see draft).
11. l0\_pen\_maldonado - contains functions ‘l0.pen.maldonado’ and ‘l1.pen.maldonado’ to run the L0 and L1 penalized nonlinear WSVM respectively (see draft).
12. l0\_fsv - contains function ‘l0.fsv’ to run the FSV procedure on WSVM (see draft).
13. knife - contains function ‘knife.hinge.matlab’ to run KNIFE on WSVM (see draft).
14. predict\_prev – contains function ‘predict.prev’ , used in real data analysis to predict cross-validated prevalence for disease for a given method.

The second subfolder (‘Simulations’) contains the following files:

1. simu\_setting\_linear1
2. simu\_setting\_linear2
3. simu\_setting\_nonlinear1
4. simu\_setting\_nonlinear2
5. simu\_setting\_nonlinear3
6. simu\_setting\_decisionlist
7. simu\_setting\_manyeffects
8. feature\_selection\_methods

The first 7 files invoke the seven simulation settings in the draft, while ‘feature\_selection\_methods’ is a generic file that contains code to run all the feature selection methods. Note that this file can be run only after data is generated using any of the 7 previous files.

The third subfolder (‘Data Analysis’) contains the following files:

1. SampleData – a .csv file containing sample data
2. data\_analysis – a code file to run our analyses on this data in R.

**REFERENCES**  
1. Zhang, Y., Laber, E. B., Tsiatis, A., and Davidian, M. (2015). Using Decision Lists to Construct Interpretable and Parsimonious Treatment Regimes. Biometrics, 71(4), 895-904.