Acta Scientific Pharmaceutical Sciences (ISSN: 2581-5423)

Volume 4 Issue 1 January 2020

Research Article

Simulation Comparison of Statistical Approaches and Procedures in Building SNP based Prediction Models for Drug Response

Wencan Zhang[1]*, Pingye Zhang[2], Feng Gao[3], Yonghong Zhu[4] and Ray Liu[5]

[1]*Takeda Develop Center, One Takeda PKWY, Deerfield, USA*

[2]*Merck, Lincoln Avenue, Rahway NJ, USA*

[3]*Biogen, Cambridge, USA*

[4]*Shanghai Henlius Biotech Inc, Shanghai, China*

[5]*Takeda, Cambridge, USA*

***Corresponding Author:** Wencan Zhang, Takeda Develop Center, One Takeda PKWY, Deerfield, USA.

DOI: 10.31080/ASPS.2020.04.0460

Received: December 13, 2019

Published: December 23, 2019

Abstract

Lack of replication on findings and missing heritability are two of the major challenges in Pharmacogenetics (PGx) studies related to developing predictive models for common disease prognosis and drug response. Recent innovations in statistical procedures and methodologies may help us understand and meet these challenges. We aimed using simulation based approaches with different prediction algorithms to compare their predictive accuracy. In our first simulation study, we compared four 1- step and one 2-step models built with five different approaches: Elastic Net (EN), Genome-wide Association Study (GWAS) + EN, Principal Component Regression (PCR), Random Forest (RF) and Support Vector Machine (SVM). The results showed that EN has the smallest test mean squared error (MSE), highest sensitivity and causal %. In the second simulation, we compared three 2-step approaches, GWAS+EN, GWAS+RF and GWAS+SVM. The GWAS+RF has the smallest test MSE and the best accuracy in picking up the seeded causal SNP variants. In the third simulation study, we compared two cross validation procedures: GWAS +EN vs. modified learn and confirm cross validation GWAS +EN (Modified CV GWAS+EN). The results showed that the latter approach has better prediction accuracy at the expense of a huge computational resource.

Keywords: Heritability; Sensitivity; SNPs

Introduction

Over the last decade, many new single nucleotide polymorphisms (SNPs) and SNP-harboring genomic regions have been identified with clinical importance by Genome-Wide Association Studies (GWAS). These variants may be used as biomarkers predictive of disease susceptibility or treatment response. Thus the predictive markers of genomics have been the integrated part of precision medicine. The concept of precision medicine has already been practiced in infectious diseases and oncology areas for years [1]. However, the identification of the core elements of precision medicine, pharmacogenomics signatures of disease/patient subset or response remains elusive in general. Lack of replication on findings and missing heritability [2-8] are two of the major challenges and implementation barriers to the application of pharmacogenomics findings in clinical practice such as building disease prognosis algorithm or response prediction models for common diseases, which are usually implicated by many genes with small effect at single gene level. Over the decade the exponential fall in the cost of genome-wide sequencing has led to the broad use of GWAS that can simultaneously examine genome-wide features instead of the traditional candidate gene approach, where the lead hits were found almost always false positive and only 2-6% of these leads can be replicated [9]. Thus one of the major goals in pharmacogenomics study is to detect a real correlation signal among SNPs that truly define drug response phenotype, ultimately leading the translation of this correlation information to benefit patients.

To build up a predictive model from genomic SNP data, it usually involves two steps/stages. First step is to scan and rank top SNPs to a manageable size through dimension reduction. Prior to a second step that involves predictive model building and subgroup identification [10], this step is usually done by a simple logistic regression or a trend test for a binary response, such as responder/non responder to a drug treatment or presence/absence of adverse event of special interest; or a generalized liner model for a continuous response, such as change from baseline for an efficacy measurement.. This first step is usually called feature selection. It may also be done in a modified tiered approach. This approach includes prioritized

feature selection on the 1st tier of pre-specified SNPs within genes of functional interest due to existing evidence and then followed by feature selection on the other 2nd tier SNPs. It is of more biological relevance for the SNPs identified from the targeted functional genes (tier 1) than the ones from the hypothesis-free GWAS (tier 2) to build the predictive model signatures.

There are different statistical and machine leaning methods for GWAS feature selection and predictive modeling. Cosgun., et al. (2011) applied three machine learning approaches: Random Forest Regression (RFR), Boosted Regression Tree (BRT) and Support Vector Regression (SVR) to the prediction of warfarin maintenance dose in a cohort of African Americans [11]. They showed that even though all three methods achieved better performance than the previously published reports, RFR had the best accuracy. Some PGx studies were retrospective and the collected and tested samples are often not from random sampling from the original clinical study. Therefore it is very difficult to use molecular biomarkers to interpret the clinical findings. The idea of cross validation and resampling strategy has been used in the PGx data analysis with different strategies. In addition, some statistical and machine learning methods have evolved for data analysis in PGx studies in GWAS and predictive model building [11-18].

With the development of innovative genomic test platforms, millions of SNP data across the genome are now readily available for association analysis. When the number of SNPs is very large, dimension reduction is inevitable before model building with manageable SNPs. A 2-stage approach is often the method of choice. Innovations in statistical procedures and methodologies can be very helpful to understand and meet those challenges in predictive model building. The objectives of the current study are to compare these new procedures and methods with three simulation themes used. In the first simulation theme, we compared five approaches: 1-step Elastic Net (EN), 2-step genome-wide association study (GWAS) + EN, 1-step Principal Component Regression (PCR), 1-step Random Forest (RF) and 1-step Support Vector Machine (SVM). For the second simulation theme, we compared three 2-step procedures, GWAS+EN, GWAS+RF and GWAS+SVM. In the third simulation theme, we compared two cross validation approaches: GWAS+EN and a modified learn and confirm cross validated GWAS+EN (i.e. Modified CV GWAS+EN).

Materials and Methods

Introduction to the statistical methods

Univariate association analysis

In all 2-step approaches in this study, the first step (stage) will use a univariate GLM association analysis to select for the top SNPs in Genome-wide association study (GWAS).

$y=\beta jxj+\varepsilon$ j=1, 2,…M

The null-hypnosis test for H0: βj=0. We order SNPs by their P-Values and only pick SNPs that pass the genome-wide significance level. We pick the top SNPs for the step (stage) two model building analysis. The GWAS is also could be used as a pre-screening step.

Elastic net (EN)

In the fitting of linear or logistic regression models, the elastic net is a regularized regression method that linearly combines the L1 and L2 penalties of the lasso and ridge methods [19]. The EN method has been used in both one and two step procedures. The decorrelation step leads to grouping effect and better prediction accuracy [19]. In addition, the decorrelation makes M>N possible.

The hyper-parameters associated with L1 and L2 penalties are trained using a Five-fold cross validation external to the predictive model building process to avoid potential bias in estimated test errors [20]. In the one step approach, we directly apply EN on all M SNPs and building predictive model.

EN has been used in both one and two step simulations.

Random forest (RF)

The Random Forest is a group of trees based on bootstrapped datasets [21]. The RF method has been used in both one and two step simulations. For B identically distributed variables (each with variance σ^2) with pair-wise correlation ρ. Variance of average = $\rho\sigma^2 + (1- \rho)\sigma^2/B$. Reduce ρ without increasing σ^2 too much. At each split, select a subset of features at random as candidates for splitting. Out of bag (OOB) error for each tree is computed based on samples not used in the bootstrapped dataset.

Generate variable importance list for all SNPs (generated only once). Iteratively fit RF, each time building a new forest after discarding lowest 30% of the SNPs used in the previous iteration, OOB error is computed for each iteration. Final prediction model is picked with the smallest number of SNPs whose Out-of-bag (OOB) error is within 1 standard error of the smallest OOB error of all forests.

RF has been used in both one and two step simulations.

Principal component regression (PCR)

In principal components regression (PCR), we use principal components analysis (PCA) to decompose the independent (x) variables into an orthogonal basis (the principal components), and select a subset of those components as the variables to predict y [22]. PCR method was only used in the one step analysis in the first simulation. Principal components (PC) are linear combinations of the SNPs. The 1st PC captures the most variance in X matrix. The top PCs would capture the majority of information in the data. Also, only top 3000 SNPs ranked by p-value are in the PCR method.

In the one step PCR approach, we apply PCA on all M SNPs and pick the top k PCs. We then use the k PCs for prediction. PCR is only evaluated in one step approach.

Support vector machine (SVM)

In machine learning, support vector machines (SVMs) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis [23]. SVM Ignores observations with residual errors less than a certain value ϵ.

The bigger the β is the more important this feature (SNP) is.

Compute β for each SNP, and order them by β (Order only once). Iteratively fit SVM each time building a new SVM after discarding lowest 10% of the SNPs used in the previous iteration. A 5-fold CV is used to get a CV error for each iteration. Final prediction model is picked with the smallest number of SNPs whose CV error is within 1 standard error of the smallest CV error of all SVM.

SVM has been used in both one and two step simulations.

Cross validation

A cross validation procedure was used in our simulation. In the model building process, we have one sample of individuals (training sample) to "learn" the prediction model. We can use another independent sample of individuals (testing sample) to evaluate how well the prediction power (test error) is for our prediction model. Cross validation (CV) can be used to estimate the test error using the training sample. It's just a technique to assess the prediction performance, we still use the entire training sample to train your prediction model and we do not waste any data.

A fivefold cross validation has been used in all simulations in this study. In our first and second simulations, we had following cross validation chart (Figure 1).

Figure 1: Standard cross validation flow chart.

Simulation one: one and two step comparisons

In our first simulation study, we compared five approaches: One step elastic net (EN), 2-step genome-wide association study (GWAS) + EN, one step principal component regression (PCR), one step random forest (RF) and one step support vector machine (SVM). In all one step approaches, the EN, PCR, RF and SVM are directly used in both feature selection and predictive model building on all SNP variants.

Settings for simulation one

In our first simulation, to take into account the original LD structure, we used a subset of the real whole genome data and only extracted the SNPs on Chr 1 (9,968 SNPs after QC from the Illumina Human Omni5Exome array (with ~ 4 million SNPs before QC) on 535 patients). We randomly select 5 SNPs as associated variants and use them to generate the phenotype:

$$y=\beta_1 x_1+ \beta_2 x_2+ \beta_3 x_3 + \beta_4 x_4+ \beta_5 x_5+e$$

$$e \sim N(0, \sigma^2)$$

The simulation has 300 for training and 235 for testing. Original LD structure is maintained. MAF for the causal variants: 5%,7%,8%,9% and16%. Base model: 5 associated variants together explain 20% of total variance. Top SNPs selected from GWAS (p)= 100. Top PCs selected (k) = 25 to account for more than 99% of total original variance. A 5 fold CV used to estimate the test error using training sample and 250 replicated data sets are generated.

Simulation Two: 2-step strategy comparisons

The second simulation compared three procedures, GWAS+EN, GWAS+RF and GWAS+SVM. The following settings are used.

Settings for simulation two and simulation three

Number of total genotyped SNPs (M) = 10,000. Number of causal variants (m) = 5, all with MAF=0.165. The coefficients for causal variants are 0.5,0.75,1,1.25,1.5 each and together explain 20% total variance. The remaining null markers are generated with MAF following a uniform distribution U (0.1, 0.4). Training sample size = Testing sample size = 300. Top SNPs selected from GWAS = 100. Select top 10 PCs for PCR. Iteratively fit Random Forest, each time building a new forest after discarding lowest 30%. Iteratively fit SVM, each time building a new SVM after discarding lowest 10%. There are 250 replicated data sets are generated and a 5 fold CV applied. These settings have also been used in simulation three.

Simulation three: additional cross validation considerations

A more sophisticated "learn and confirm" strategy was compared in simulation three. The purpose is finding a better way to conduct cross validation by having an extra validation (confirm) on the top SNPs already identified from the first step GWAS (learn), before a second step EN on model building. We considered two cross validation approaches. 1. GWAS+EN (as shown in the figure 1) and 2. A modified CV GWAS+EN. We used cross validation along with GWAS to stable the feature selection for the top variants (Figure 2).

Figure 2: Modified cross validation flow chart.

Results and Discussion

Results and discussion for simulation one

As shown in table 1, the comparison of the five approaches identified that the 1-step EN had the smallest test MSE (4.49) and the highest percent associated with the causal variants (0.14). However it also came up with a relatively higher training error (3.51). SVM had the highest sensitivity (0.74) and the smallest training error (0.04). Random forest approach had the second smallest test MSE (4.78). In addition, it had the second highest causal % (0.09). The 2-step approach (GWAS + CN) had the second highest sensitivity (0.64). The training error was biasedly down warded (underestimated) for SVM (0.04).

Approaches	Test MSE	Sensitivity	Causal %	Training error
EN	4.49	0.61	0.14	3.51
GWAS+EN	5.55	0.64	0.04	1.30
PCR	5.29	NA	NA	4.45
RF	4.78	0.51	0.09	0.73
SVM	5.41	0.74	0.01	0.04

Table 1: Results of simulation one.

Test MSE: Mean squared error on the testing sample.

Sensitivity: Number of causal SNPs in the final set /5.

Causal %: Number of associated SNPs in the final set/ Number of SNPs in the final set.

Training error: Prediction error on training sample.

In Simulation One, with manageable number of SNP variants (< 10,000) and sample size (535), 1-step methods, especially the EN and RF showed some advantages over the rest of the other methods.

Cosgun., *et al*. [11] applied three machine learning approaches: Random Forest Regression (RFR), Boosted Regression Tree (BRT) and Support Vector Regression (SVR) to the prediction of warfarin maintenance dose in a cohort of African Americans [11] and found R2 between the predicted and actual square root of warfarin dose in this model was on average 66.4% for RFR, 57.8% for SVR and 56.9% for BRT. Thus RFR had the best accuracy. Our results were consistent with Cosgun's study and had confirmed that RF is one of the better methods in prediction model building.

Summary for simulation one

When the phenotype-genotype model is generated by a linear model: EN and RF had better prediction accuracy than GWAS + EN. GWAS+EN approach may preferentially select in associated variants with the price of bringing more noise than EN. The training error was biasedly downwarded (underestimated) for SVM. The cross validation error was a good estimate of the true test error.

Results and discussion for simulation two

Table 2 showed the results from the second simulation. A new measurement, "no.diff " was introduced, where it means percent number of selected SNPs in final set- number of causal variants. A smaller number means a more accurate method. Our results suggested that GWAS+RF had the smallest test MSE (7.51). This results again confirmed findings from Cosgun., *et al*. [11] and even in the 2-step model building procedure, GWAS+RF had better accuracy than other methods.

Procedure	Test MSE	Sensitivity	Causal%	Training error
GWAS+EN	8.78	0.65	0.04	1.39
GWAS+RF	7.51	0.52	0.09	1.00
GWAS+SVM	10.02	0.48	0.05	2.00

Table 2: Results of simulation two on two stage approaches.

Test MSE**:** Mean squared error on testing sample.

Sensitivity**:** Number of causal SNPs in final set/number of causal SNPs.

Causal%**:** Number of causal SNPs in final set/number of selected SNPs in final set.

Training error: Prediction error on training sample.

Summary for simulation two

GWAS + Random Forest gives the best prediction accuracy among all 2-step strategies. GWAS + Random Forest tends to select in fewer number of SNPs with higher accuracy than the others.

Results and discussion on simulation three

The results of three procedure comparisons GWAS + EN and Modified CV GWAS + EN are shown in Table 3.

The results of the simulation in Table 3 showed that the Modified CV GWAS + EN have better prediction accuracy than GWAS + EN (MSE of 8.12 vs 8.79), and modest training error as well (3.1 vs 1.42)). Nevertheless, it came with huge computational cost. GWAS + EN had higher sensitivity than Modified CV GWAS + EN (0.66 vs 0.51).

The GWAS+EN procedure was a standard one (as shown in the Figure 1). The difference for the Modified CV GWAS+EN was that

Procedure	Test MSE	Sensitivity	Causal %	Training error
GWAS+EN	8.79	0.66	0.04	1.42
Modified CV GWAS+EN	8.12	0.51	0.04	3.10

Table 3: Results of simulation three on different cross validation considerations.

Test MSE: Mean squared error on testing sample.

Sensitivity: Number of causal SNPs in final set/number of causal SNPs.

Causal%: Number of causal SNPs in final set/number of selected SNPs in final set.

Training error**:** Prediction error on training sample.

we used a "learn and confirm" cross validation procedure along with GWAS to stable the feature selection for the top variants (Figure 2). The "learn and confirm" procedure was with an additional confirmation step on the selected top ranked SNPs in a different data set before building up the models. This strategy would be very similar to the model building procedure by Shigemizu., *et al.* (2014) with real type 2 diabetes data [12], in which an extra validation on the top identified SNPs was implemented before predictive model building. We recommend this procedure as it came with the best accuracy and gave additional stability and accuracy on the SNPs for the predictive model building.

Summary for the simulation three

Modified CV GWAS + EN had better prediction accuracy than GWAS + EN, but it came at huge computational cost. GWAS +EN had higher sensitivity than Modified CV GWAS + EN.

Conclusion

When the phenotype-genotype model is generated by a linear model:

1. One step EN has better prediction accuracy than GWAS + EN, with a manageable number of SNPs.
2. GWAS+EN is more likely to select in associated variants at the expense of selecting more noise than EN.
3. GWAS + Random Forest gives best prediction accuracy among all 2-step strategies.
4. Modified CV GWAS + EN (learn and confirm) has better prediction accuracy than GWAS + EN, with the burden of huge computational cost.

Acknowledgements

Useful discussions with Dr. Zheng Zha, Dr. Caiyan Li, Dr. Ling Wang and reviews by Dr. Yu-chen Su at Takeda Pharmaceutical Develop Center are highly appreciated.

Conflict of Interest

The project was carried out while Dr. Pingye Zhang was a summer intern at Takeda develop center at Deerfield, IL. USA. All other authors were Takeda employees at the time. The nature of the research is comparison of statistical methodologies and cross validation procedures, there is no conflict of interests.

Bibliography

1. Richard L Schilsky. "Personalized medicine in oncology: the future is now". *Nature Reviews Drug Discovery* 9 (2010): 363-366.
2. Schrodi SJ., *et al*. "Genetic-based prediction of disease traits: prediction is very difficult, especially about the future". *Frontiers in Genetics* 5 (2014).
3. Naomi R Wray., *et al*. "Pitfalls of predicting complex traits from SNPs". *Nature Reviews Genetics* 14.7 (2013): 507-515.
4. Sang Hong Lee., *et al*. "Estimating Missing Heritability for Disease from Genome-wide Association Studies". *The American Journal of Human Genetics* 88 (2011): 294-305.
5. Yang J., *et al*. "Common SNPs explain a large proportion of the heritability for human height". *Nature Genetics* 42 (2010): 565-569.
6. Visscher PM., *et al*. "A commentary on 'common SNPs explain a large proportion of the heritability for human height' by Yang et al (2010). *Twin Research and Human Genetics* 13 (2010): 517-524.
7. G SY Pang., *et al*. "Predicting potentially functional SNPs in drug-response genes". *Phamacogenomics* 10 (2009): 639-653.
8. YW Francis Lam. "Scientific Challenges and Implementation Barriers to Translation of Pharmacogenomics in Clinical Practice". *ISRN Pharmacology* (2013).
9. Lee SH., *et al*. "Estimating Missing Heritability for Disease from Genome-wide Association Studies". *American Journal of Human Genetics* 88 (2011): 294-305.
10. Thanh-Tung Nguyen., *et al*. "Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests". *BMC Genomics* 16 (2015): S5.
11. Erdal Cosgun., *et al*. "High-dimensional pharmacogenetic prediction of a continuous trait using machine learning techniques with application to warfarin dose prediction in African Americans". *Bioinformatics* 27 (2011): 1384-1389.
12. Daichi Shigemizu., *et al*. "The Construction of Risk Prediction Models Using GWAS Data and Its Application to a Type 2 Diabetes Prospective Cohort". *PLoS ONE* 9.3 (2014): e9254.
13. Charles Kooperberg., *et al*. "Risk Prediction using Genome-Wide Association Studies". *Genetics Epidemiology* 34.7 (2010): 643-652.

14. Zhi Wei., *et al*. "Large Sample Size, Wide Variant Advanced Machine-Learning Technique Boost Risk Prediction for Inflammatory Bowel Disease". *The American Journal of Human Genetics* 92 (2013): 1008-1012.

15. Xi Chen and Hemant Ishwaran. "Random forests for genomic data analysis". *Genomics* 99 (2012): 323-329.

16. Iris Schrijver., *et al*. "Opportunities and Challenges Associated with Clinical Diagnostic Genome Sequencing". *The Journal of Molecular Diagnostics* 14.6 (2012).

17. Rita M Cantor., *et al*. "Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application". *The American Journal of Human Genetics* 86 (2010): 6-22.

18. Li L., *et al*. "A multi-marker molecular signature approach for treatment-specific subgroup identification with survival outcomes". *The Pharmacogenomics Journal* 14 (2014): 439-45.

19. Zou H and Trevor T. "Regularization and Variable Selection via the Elastic Net". *Journal of the Royal Statistical Society Series B* 67.2 (2005): 301-320.

20. Christophe Ambroise and Geoffrey J McLachlan. "Selection bias in gene extraction on the basis of microarray gene-expression data". *PNAS* 99.10 (2002): 6562-6566.

21. Ho Tin Kam. "Random Decision Forests". Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14-16 August (1995): 278-282.

22. Jolliffe Ian T. "A note on the Use of Principal Components in Regression". *Journal of the Royal Statistical Society, Series C* 31 (1982): 300-303.

23. Cortes C and Vapnik V. "Support-vector networks". *Machine Learning* 20 (1995): 273-297.