Comparisons of Statistical Approaches and Procedures in Building Predicting Models to Drug Response from SNPs through Simulation

Wencan Zhang[1], Pingye Zhang[2], Feng Gao[1] , Yonghong Zhu[1]

[1] Takeda Development Center, Deerfield, IL 60010
[2] University of Southern California

Abstract

Lack of replication on findings and missing heritability are two of the many challenges in Pharmacogenetics (PGx) studies in building predictive models for common disease prognosis and drug response. To build up a predictive model from genomic SNP (single nucleotide polymorphism) data usually needs a two-step process: first scan and rank top SNPs to a manageable size (feature selection) followed by a second step model building (two step/stage modeling). In our first simulation study, we compared one step and two step model building with five approaches: Elastic Net (EN), genome wide association study (GWAS) + EN, Principal Component Regression (PCR), Random Forest (RF) and Support Vector Machine (SVM). We used genotype data of 9,968 SNPs on Chromosome One by Illumina Infimium Omni5Exome array on 535 real samples. We randomly select 5 SNPs to generate a quantitative phenotype using a linear model and used a fivefold cross validation (CV) and 250 replications. The results have shown that EN has the smallest test MSE, highest sensitivity and causal %. In the second simulation, we compared three two-step approaches, GWAS+EN, GWAS+RF and GWAS+SVM. The GWAS+RF has the smallest test MSE (mean squared error) and best accuracy in picking up the seeded causal variants. In the third simulation study, we compared two cross validation procedures: GWAS +EN and modified CV GWAS +EN. The results show that the CV GWAS + EN has better prediction accuracy but at a huge computational cost.

Key Words: PGx, two step/stage predictive modeling, SNPs, cross validation, GWAS

INDRUDUCTION

Over the last decade, many new genetic associations of single nucleotide polymorphism (SNP) and SNP-located gene regions have been identified by Genome-Wide Association Studies (GWAS). The identified new variants could be used as biomarkers in prediction of the genomic contribution to the common complex disease. It is more interesting and important to find the biomarkers and genes that regulate the disposition and target pathways of drugs, to identify the sub-patient-population who is more likely to be a responder to a particular therapeutic treatment and demonstrate an improved efficacy profile. The concept of personalized medicine or more recently the precision medicine has already been used in oncology [1, 2]. Lack of replication on findings and missing heritability [2-8] are two of the many challenges and implementation barriers to translation of pharmacogenomics in clinical practice and building disease prognosis and drug response predictive models for common disease which are caused by many genes with small effects. The exponential fall in genome sequencing costs led to the use of GWAS studies which could simultaneously examine all candidate-genes in larger samples than the original finding, where the candidate-gene hits were found to almost always be false positives and only 2-6% replicate [23]. One of the major issues in pharmacogenomics is to detect a real correlation between the drug response phenotype with SNPs and utilize this correlation to benefit patients.

To build up a predictive model from genomic SNP data usually needs two steps / stages. First scan and rank top SNPs to a manageable size through dimension reduction, usually a simple logistic regression or a trend test for a binary response, such as responder/non responder to a drug treatment or adverse event, and a generalized liner model for a continuous response, such as change from baseline for a efficacy measure, before a second step predictive model building and subgroup identification [14]. The first step is usually called feature selection. This step could also be done in a tiered approach with tests on the SNPs within a specific list of targeted functional genes as the first tier and then GWAS on overall left over SNPs as the second tier. It is of more biological interests to the SNPs identified from the targeted functional genes (tier 1) than the ones from the hypothesis-free GWAS to build the predictive model signatures. Also, the identified rare variants may have very significant p-values but have limited prediction power in a pharmacogenomics setting due to smaller sample sizes from clinical trials.

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 [13] and RFR had the best accuracy, but all three techniques achieved better performance than the current published R(2).

Some pharmacogenomics studies are retrospective and using the samples collected from the clinical studies with different objectives and endpoints during the different phases of drug development. The idea of cross validation and re-sampling strategy have been used in the PGx data analysis with different strategies [9-14, 21]. To design a effective PGx statistical analysis plan, one needs to decide how many folds to divide the available data to training and validation data sets from the available subjects who have signed the PGx informed consent, also have blood samples actually collected and passed the QC for analysis. Many statistical and machine learning methods are available for data analysis in PGx studies in GWAS and predictive model building [11-15, 18]. Shigemizu et al (2014) had used a CV method with additional validation on the top SNPs identified at GWAS step [9].

Three simulation themes are used to reach our objectives of this study. In our first simulation study, we compared five approaches: One step Elastic Net (EN), two 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). The second simulation compared three procedures, GWAS+EN, GWAS+RF and GWAS+SVM. In our third simulation, we compared two cross validation approaches: GWAS+EN and a modified CV GWAS+EN.

2. MATERIALS AND METHODS

2.1 Introduction of the methodologies

2.1.1 Notations

N: Sample size

M: Total number of SNPs

m: number of associated variants

y_i: Continuous phenotype for i^{th} individual

x_{ij}: Genetic value for j^{th} SNP on i^{th} individual

X: $[x_1\ x_2\ \dots\ x_M]$ where $x_j=[x_{1j}\ x_{2j}\dots x_{Nj}]^T$

y: $[y_1\ y_2\ \dots\ y_N]^T$

β: $[\beta_1\ \beta_2 \ldots \beta_M]^T$

Suppose y is centered and x is standardized.

2.1.2 Univariate association analysis (step one)

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

$y=\beta_j x_j+\varepsilon \quad j=1,2,\ldots M$

The non-hypnosis test for H_0: $\beta_j=0$. 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.

2.1.3 Elastic Net (EN)

The EN method has been used in both one and two step procedures. The advantages of the Elastic net with d-correlation leads to grouping effect and better prediction accuracy. Also, the de-correlation make M>N possible.

Objective function

$$L(\lambda_1,\lambda_2,\beta)=|\mathbf{y}-\mathbf{X}\beta|^2+\lambda_2|\beta|^2+\lambda_1|\beta|_1$$

$$|\beta|^2=\sum_{j=1}^{p}\beta_j^2,$$

$$|\beta|_1=\sum_{j=1}^{p}|\beta_j|$$

Solution:

$$\hat{\beta}(OLS)=\arg\min_{\beta}\beta^T\left(X^TX\right)\beta-2y^TX\beta$$

$$\hat{\beta}(Ridge)=\arg\min_{\beta}\beta^T\left(\frac{X^TX+\lambda_2\mathbf{I}}{1+\lambda_2}\right)\beta-2y^TX\beta$$

$$\hat{\beta}(\text{lasso})=\arg\min_{\beta}\beta^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\beta-2\mathbf{y}^{\mathrm{T}}\mathbf{X}\beta+\lambda_1|\beta|_1.$$

$$\hat{\beta}=\arg\min_{\beta}\beta^{\mathrm{T}}\left(\frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda_2\mathbf{I}}{1+\lambda_2}\right)\beta-2\mathbf{y}^{\mathrm{T}}\mathbf{X}\beta+\lambda_1|\beta|_1.$$

De-correlation

$$\frac{X^TX+\lambda_2\mathbf{I}}{1+\lambda_2}=\begin{bmatrix}1 & \frac{\rho_{12}}{1+\lambda_2} & \cdots & \frac{\rho_{1M}}{1+\lambda_2}\\ & 1 & & \\ & & 1 & \frac{\rho_{M-1,M}}{1+\lambda_2}\\ & & & 1\end{bmatrix}$$

2.1.4 Random Forest (RF)

The RF method has been used in both one and two step simulations. The RF is a group of trees based on bootstrapped datasets. 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. A list of variable importance scores can be generated

2.1.5 Principal Component Regression (PCR)

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 Y. 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.

2.1.6 Support Vector Machine (SVM)

SVM has been used in both one and two step simulations. SVM Ignores observations with residual errors less than a certain value ϵ

$$H(\beta, \beta_0) = \sum_{i=1}^{N} V(y_i - f(x_i)) + \frac{\lambda}{2}\|\beta\|^2$$

Where

$$V_\epsilon(r) = \begin{cases} 0 & \text{if } |r| < \epsilon, \\ |r| - \epsilon, & \text{otherwise.} \end{cases}$$

$$f(x) = x^T\beta + \beta_0$$

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

2.2 Simulation one: one and two step comprisions

In our first simulation study, we compared five approaches: One step Elastic Net (EN), two 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. The two step approach , we first used the univeriate method as described in 2.1.2 and used EN in predictive model building. The details for the five approaches are as the following.

2.2.1 Five approaches

Approach number 1 (EN): Apply EN on all M SNPs and choose the selected SNPs for prediction.

Approach number 2 (GWAS+EN):

- Step-1: Apply GWAS on all M SNPs and pick top p SNPs
- Step-2: Apply EN on p top-picked SNPs.

Approach number 3 (PCR): Apply PCA on all M SNPs and pick the top k PCs, use the k PCs for prediction.

Approach number 4 (RF):

- 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.

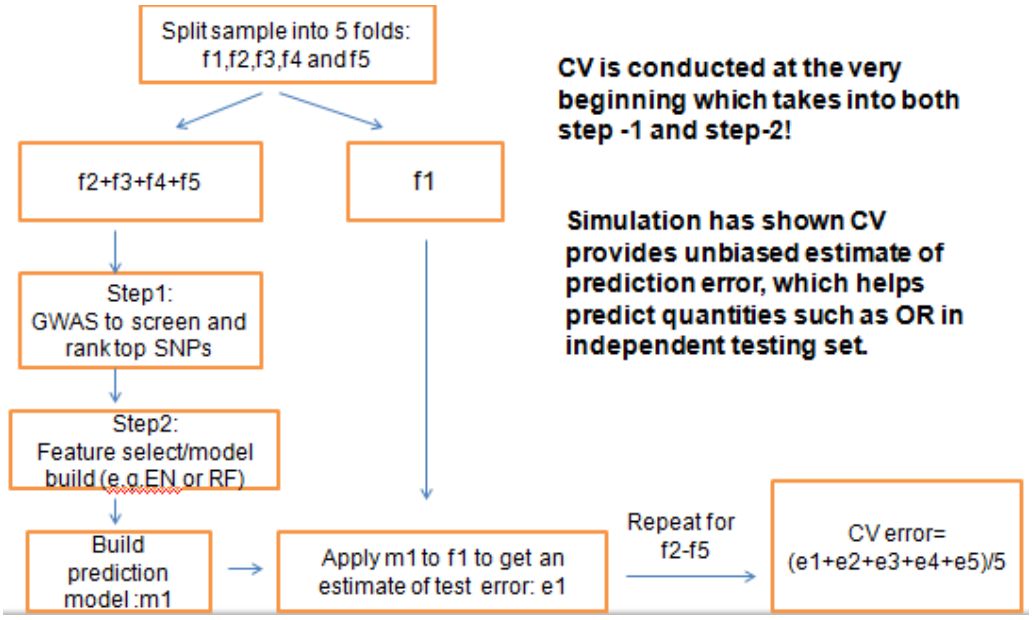
Approach number 5 (SVM):

- 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. 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.

2.2.2 Cross validation

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 (chart 1):

Chart 1. Standard cross validation.



- In the model building process, we have one sample of individuals (training sample) to "learn" the prediction model.
- If we have an independent sample of individuals (testing sample) we can use them to evaluate how well the prediction power (test error) is for your 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.

2.2.3 Settings for simulation one

In our first simulation, we used real genetic data to take into account the real LD structure extract SNPs on Chr 1 (9,968 SNPs after QC from the Illumina Infinium Human Omni5Exome

array on 535 patients). We randomly select 5 SNPs as associated variants and use them to generate the phenotype:

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

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

- The simulation has 300 for training and 235 for testing
- Original LD structure is maintained
- MAF for the causal variants: 5%,7%,8%,9%,16%
- Base model: 5 associated variants together explain **20%** of total variance
- Top SNPs selected from GWAS (p)= 100
- Top PCs selected (k) = 25
- 5 fold CV used to estimate the test error using training sample
- 250 replicated data sets are generated

2.3 Simulation Two: two-step strategy comparisons

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

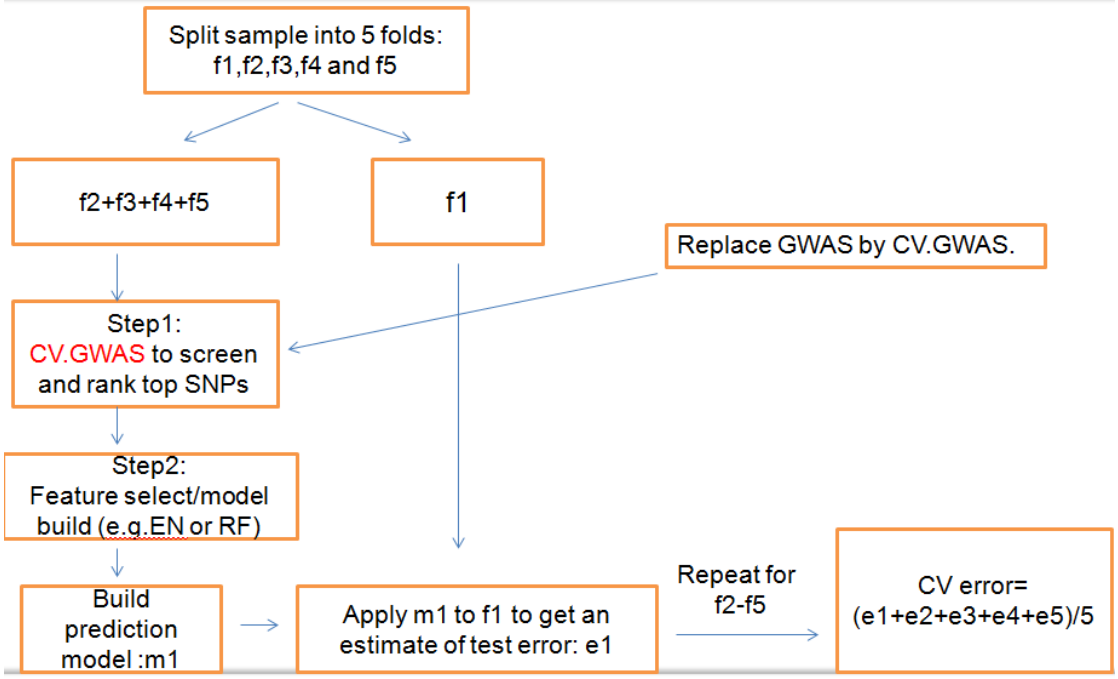
2.3.1 Settings for simulation two

- Number of total genotyped SNPs (M) = 10,000
- Number of causal variants (m) = 5, all with MAF=0.165 yielding %30 carriers. 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%
- 250 replicated data sets are generated
- 5 fold CV are applied

2.4 Simulation three: additional cross validation considerations

We also considered two cross validation approaches. 1. GWAS+EN (as shown in the Chart 1). And 2. A modified CV GWAS+EN. We used cross validation along with GWAS to stable the feature selection for the top variants (Chart 2).

Chart 2. A CV GWAS+EN procedure.



2.4.1 Settings for simulation three

- Number of total genotyped SNPs (M) = 10,000
- Number of causal variants (m) = 5, all with MAF=0.165 yielding %30 carriers. 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
- Top SNPs selected from CV.GWAS = 100
- Five fold CV are applied

3. RESULTS AND DISCUSSION

3.1.1 Results and discussion for simulation one

Table 1. The results of simulation one on comparison of 5 approaches

Approaches	Test MSE	Sensitivity	Causal %	CV error	Training error
EN	4.49	0.61	0.14	4.67	3.51
GWAS+EN	5.55	0.64	0.04	5.67	1.30
PCR	5.29	NA	NA	6.62	4.45
RF	4.78	0.51	0.09	4.82	0.73
SVM	5.41	0.74	0.01	7.23	0.04

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

CV error: correct cross validation error

Training error: prediction error on training sample.

As shown in table 1, the comparison of the five approaches identifies that one step EN with the smallest test MSE (4.49) and CV error (4.67), highest percent associated with the causal variants (0.14). But it comes up a relatively higher training error (3.51). SVM has the highest sensitivity (0.74) and the smallest training error (0.04). Random forest has the second smallest test MSE (4.78) and CV error (4.82). Also has the second highest Causal % (0.09). The two step approach has the second highest sensitivity (0.64). The training error is biased downward (underestimated) for SVM (0.04).

At least for this study with manageable number of SNP variants (< 10,000) and sample size (535), one step methods, especially the EN and RF showed some advantages over the rest methods.

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 [23] and found R(2) 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 had confirmed that RF is one of the better methods in prediction model building.

3.1.2 Summary for simulation one

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

1. EN and RF had a better prediction accuracy than GWAS + EN
2. GWAS+EN is more likely to select in associated variants with a price of selecting more noise than EN
3. The training error is biased downward (underestimated) for SVM
4. The cross validation error is a good estimate of the true test error

3.2 Results and discussion for simulation two

Table 2. Results of simulation two

Procedure	Test MSE	No. diff	Sensitivity	Causal%	Training error
GWAS+EN	8.78	87.44	0.65	0.04	1.39
GWAS+RF	7.51	24.53	0.52	0.09	1.00
GWAS+SVM	10.02	42.90	0.48	0.05	2.00

Test MSE: mean squared error on testing sample
No.diff: number of selected SNPs in final set- number of causal variants
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.

Table 2 shown the results from the second simulation. A new measurement, no.diff means percent number of selected SNPs in final set- number of causal variants. A smaller number means a more accurate method. It looks that GWAS+RF has the smallest test MSE (7.51) and a smaller difference (24.53). This results again confirmed findings from Cosgun et al (2011) [13] and even in a two step model building procedure that GWAS+RF has better accuracy than other methods.

3.2.1 Summary for simulation two

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

3.3 **Results and discussion on simulation three**

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

Table 3. Results of Simulation three on different cross validation considerations

Procedure	Test MSE	No. diff	Sensitivity	Causal%	Training error
GWAS+EN	8.79	87.15	0.66	0.04	1.42
Modified CV GWAS+EN	8.12	58.68	0.51	0.04	3.10

Test MSE: mean squared error on testing sample
No.diff: number of selected SNPs in final set-number of causal variants
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.

The purpose of simulation three is finding a better way to conduct cross validation by having an extra validate on the top SNPs already identified from the first step GWAS, before a second step EN on model building. The details of the three procedures are shown in the three charts (chart 1-chart 3). The results of the simulation are in table 3 showing CV GWAS + EN has better prediction accuracy than GWAS + EN (no.diff are 58.68 vs 87.15, table 3.), and modest training error (3.1). But it comes at huge computational cost. GWAS +EN has higher sensitivity than CV.GWAS + EN (0.66 vs 0.51).

The GWAS+EN procedure is a standard one (as shown in the Chart 1). The difference for the modified CV GWAS+EN is that we used cross validation along with GWAS to stable the feature selection for the top variants (Chart 2). This strategy would be very similar to the model building procedure by Shigemizu et al (2014) with real type 2 diabetes data [9] in which an extra validation on the top SNPs identified was implemented before predictive model building. We recommend this procedure because it comes with the best accuracy.

3.3.1 Summary for the simulation three

1. CV GWAS + EN has better prediction accuracy than GWAS + EN, but it comes at huge computational cost.
2. GWAS +EN has higher sensitivity than CV GWAS + EN.

4. CONCLUSIONS

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

1. EN has a better prediction accuracy than GWAS + EN
2. GWAS+EN is more likely to select in associated variants with a price of selecting more noise than EN
3. GWAS + Random Forest gives best prediction accuracy among all 2-step strategies
4. Modified CV GWAS + EN has better prediction accuracy than GWAS + EN, but it comes at huge computational cost.

Acknowledgements

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

References

[1] Richard L. Schilsky Personalized medicine in oncology: the future is now. NATURE REVIEWS. vOLUME 9 MAY 2010 363.

[2] . S. J. Schrodi, S. Mukerjee, Y. Shan et al. Genetic-based prediction of diseasetraits: prediction is very difficult, especially about the future. FrontiersinGenetics June (2014) Volume5 Article162 .2

[3] Naomi R. Wray, Jian Yang, Ben J. Hayes, Alkes L. Price, Mike E. Goddard, and Peter M. Visscher. Pitfalls of predicting complex traits from SNPs. Nat Rev Genet. 2013 July ; 14(7): 507–515. doi:10.1038/nrg3457.

[4] Sang Hong Lee, Naomi R. Wray, Michael E. Goddard, and Peter M. Visscher.

Estimating Missing Heritability for Disease from Genome-wide Association Studies. The American Journal of Human Genetics 88, 294–305, March 11, 2011.

[5] Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42, 565–569.

[6] Visscher, P.M., Yang, J., and Goddard, M.E. (2010). A commentary on 'common SNPs explain a large proportion of the heritability for human height' by Yang et al. (2010). Twin Res. Hum. Genet. 13, 517–524.

[7] G. SY Pang, Jingbo Wang, Zihua Wang and C. GL Lee. Predicting potentially functional SNPs in drug-response genes. Phamacogenomics (2009) 10(4), 639-653

[8] Y. W. Francis Lam. Scientific Challenges and Implementation Barriers to Translation of Pharmacogenomics in Clinical Practice. ISRN Pharmacology Volume 2013.

[9] Daichi Shigemizu, Testuo Abe, Takashi Morizono et al The Construction of Risk Prediction Models Using GWAS Data and Its Application to a Type 2 Diabetes Prospective Cohort. PLoS ONE March 2014 Volume 9 Issue 3 e9254.

[10] Charles Kooperberg, Michael LeBlanc, and Valerie Obenchain. Risk Prediction using Genome-Wide Association Studies. Genet Epidemiol. 2010 November ; 34(7): 643–652.

[11] Zhi Wei, Wei Wang, Jonathan Bradfield 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, 1008–1012, June 6, 2013.

[12] Xi Chen, Hemant Ishwaran. Random forests for genomic data analysis. Genomics 99 (2012) 323–329.

[13] Erdal Cosgun, Nita A. Limdi and Christine W. Duarte. High-dimensional pharmacogenetic prediction of a continuous trait using machine learning techniques with application to warfarin dose prediction in African Americans. Bioinformatics. Vol. 27 no. 10 (2011), pages 1384–1389.

[14] Thanh-Tung Nguyen, Joshua Zhexue Huang, Qingyao Wu, Thuy Thi Nguyen Mark Junjie Li. Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests. BMC Genomics 2015, 16(Suppl 2):S5

[15] Iris Schrijver, Nazneen Aziz, et al. Opportunities and Challenges Associated with Clinical Diagnostic Genome Sequencing. The Journal of Molecular Diagnostics, Vol. 14, No. 6, November 2012.

[16] Rita M. Cantor, Kenneth Lange, and Janet S. Sinsheimer. Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. The American Journal of Human Genetics 86, 6–22, January 8, 2010

[17] Li L, Guennel T, Marshall SL, Cheung LWK. A multi-marker molecular signature approach for treatment-specific subgroup identification with survival outcomes. *The Pharmacogenomics Journal,* 2014; 14(5): 439-45.

[18] Kochi, Y., Suzuki, A., Yamamoto, K. (2014). Genetic basis of rheumatoid arthritis: A current review. *Biochemical and Biophysical Research Communications*, **452**(2), 254-262.

[19] Qianchuan He and Dan-Yu Lin A variable selection method for genome-wide association studies. Bioinformatics. ol. 27 no. 1 2011, pages 1–8.

[20] Wei Z, Wang K, Qu H-Q, Zhang H, Bradfield J, et al. (2009) From Disease Association to Risk Assessment: An Optimistic View from Genome-Wide Association Studies on Type 1 Diabetes. PLoS Genet 5(10): e1000678. doi:10.1371/journal.pgen.1000678

[21] Xuexia Wang, Michael J Oldani,Xingwang Zhao,Xiaohui Huang, and Dajun Qian. A Review of Cancer Risk Prediction Models with Genetic Variants Cancer Inform. 2014; 13(Suppl 2): 19–28.

[22] Jia Kang, Judy Cho, and Hongyu Zhao. PRACTICAL ISSUES IN BUILDING RISK-PREDICTING MODELS FOR COMPLEX DISEASES. J Biopharm Stat. 2010 March ; 20(2): 415–440.

[23] Wikipedia, the free encyclopedia. Missing heritability problem. https://wikipedia.org/wiki/Missing_heritability_problem.