Distribution Agreement

Signature:

__________________ __________________________
Feng Tian Date

Gene Profile Modeling and Integration for EWOC Phase I Clinical Trial Design While Fully Utilizing All Toxicity Information

By

Feng Tian
Master of Science in Public Health

Biostatistics and Bioinformatics

Zhengjia (Nelson) Chen, PhD
(Thesis Advisor)

Xiangqin Cui, PhD
(Reader)

Gene Profile Modeling and Integration for EWOC Phase I Clinical Trial Design While Fully Utilizing All Toxicity Information

By

Feng Tian

B.S.
Purdue University
2018

B.S.
China Agricultural University
2018

Thesis Committee Chair: Zhengjia (Nelson) Chen, PhD
Reader: Xiangqin Cui, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics and Bioinformatics
2020

Abstract

Gene Profile Modeling and Integration for EWOC Phase I Clinical Trial Design While Fully Utilizing All Toxicity Information

By Feng Tian

Background: Personalized medicine incorporating genomic profile has become the frontier in modern medicine. It is especially profound in optimizing healthcare for cancer patients because many gene mutations significantly affect cancer progression and efficacy of treatments. Personalized Maximum Tolerated Dose (pMTD) estimation in Phase I clinical trial is the initial key step to integrate genomic profile into personalized medicine.

Methods: Considering the limited sample size of phase I trails, selecting a small number of representative gene mutation profiles is required to keep the pMTD estimation valid. The main aim of this study is to achieve above goal by performing variable selections and comprehensive index construction using four common methods: model selection with logistic regression, regularization, principle components analysis, and random forest.

Results: The results of four methods are compared in the consistency of selected genes, the simplicity and the variety in dose estimation using EWOC-NETS (escalation with overdose control using normalized equivalent toxicity score) framework. We found that different methods are fairly consistent in selecting the important genes. The Elastic Net method is the optimal one to generate a model with simplicity and precise dose estimation in predicting tumor response.

Conclusion: For future pMTD estimation, it is a good idea to use Elastic Nets as a main reference and the common elements recommended by other mentioned methods as additional support to decide the required representative gene information to be incorporated into EWOC-NETS. The extracting and incorporation of summary genomic data will have great potential to improve treatment precision and trial efficacy.

Gene Profile Modeling and Integration for EWOC Phase I Clinical Trial Design While Fully Utilizing All Toxicity Information

By

Feng Tian

B.S.
Purdue University
2018

B.S.
China Agricultural University
2018

Thesis Committee Chair: Zhengjia (Nelson) Chen, PhD
Reader: Xiangqin Cui, PhD

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics and Bioinformatics
2020

ACKNOWLEDGEMENTS: Supported in part by NIH/NCI Grants No. 1 P50 CA217691-01A1(Z.C. and M.K.). Research reported in this publication was also supported in part by the Biostatistics Shared Resource Core of the University of Illinois Cancer Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Table of Contents

1. Introduction:

Many biomarkers, including proteins, lipid and genes are correlated with the effect of cancer treatments and also dose-toxicity relationships. Through integrating personal information into clinical trial modeling, precise medicine can be further developed[1]. Phase I cancer clinical trials are one initial important validation step for the application of drug. Considering that patients involved in cancer phase I clinical trials are mostly in an advanced cancer stage and are seeking for cure, the designs of the trials are supposed to minimize the number of individuals treated at low, non-effective dose and also with overdose[2]. To achieve that, Escalation with overdose control (EWOC) can be a good choice. EWOC is an adaptive Bayesian method to estimate the maximum tolerated dose during phase I clinical trials [4]. It shows the characteristic to minimize the predicted amount by which any given patient is underdosed. It also feature approaching the maximum tolerated dose (MTD)as rapidly as possible with the restriction that the posterior overdose probability is no more than a pre-specified feasible bound[3].

One drawback of EWOC model is that it uses a binary outcome variable to indicate the appearance of dose limited toxicity. As a result, much information is discarded in term of the multiple types of toxicities and their grades. Considering the nature of cancer phase1 clinical trial as a small sample of patients with limited information and therapeutic aim, it is important to completely and reasonably utilize the information obtained through trials to get the estimation of MTD quickly and precisely. In dealing with the loss of the toxicity information, one recommended way is to replace the traditional binary outcome in MTD estimation with the Normalized Equivalent Toxicity Score (NETS). NETS is a novel toxicity score that adjust the original toxicity score and weight all the toxicities for each individual. It is shown that this score can improve the accuracy of MTD estimation with the model called EWOC-NETS [4]. To further apply this model, additional covariates representing personal characteristic can be incorporated to help the MTD estimation for different subjects. Simulation results have shown that an additional binary categorical or standardized continuous variable in EWOC-NETS can improve the performance in estimating MTDs for different patients groups[1]. It is interesting to think about that what kinds of covariates should be involved in EWOC-NETS to further develop the personalized MTD estimation. One possible choice is the genetic information. Genetic information represents diversity and uniqueness. If genetic information can be integrated to EWOC-NETS, it might be possible to make a more precise and biologically plausible estimation for the MTD for each individual, which, is what desired for personalized medicine.

To integrate genetic information into the EWOC-NETS model, we can first select some dominant genes as predictors or combine the most valuable gene information as comprehensive index through modeling with the treatment response. Later, we can transform those potential candidates into covariates that provide helpful information in EWOC-NETS model. One challenge is that genetic markers are sometimes tightly correlated with each other, which can cause multicollinearity. Besides, the number of genes potentially involved can be relatively large compared to the sample size, which can cause overfitting. To overcome those problems, dimension reduction and the multicollinearity elimination are necessary to help minimize the loss of information. To achieve that, we applied four methods including model selections, regularization regression, principle components analysis(PCA) and random forest analysis. The performance of those methods in terms of the overlapping of recommended genes and the standard error of dose effect estimation in simulation will then be evaluated and compared to find the optimal method.

The manuscript is organized as follows. In section 2, we describe how this study is designed, what gene modeling methods are used and the concept of dose-toxicity model: EWOC-NETS with additional covariates. In section 3, we present the results of gene mutation profile modeling using different methods and make comparison with one simulation analysis under EWOC-NETS framework. The comparison results in terms of strength, weakness and similarity of those four methods are fully discussed in the section 3&4.

2.Method:

In this study, the real-world data were collected from 161 patients with rectum/colon cancer at Emory University. Investigators recorded their gender, smoking history, therapy, gene mutations and treatment response. This study mainly aims to model the gene mutations of cancer patients with treatment response to check consistency for different methods in selecting the candidate's genes and evaluate the performance of those methods in dose estimation in EWOC-NETS model.

The treatment response is a binary variable referred to the "Response Evaluation Criteria in Solid Tumors (RECIST)"[5]. Gene mutations are represented by binary variables: 1 means mutated and 0 means unmutated. The data were generated using the tumor sample from patients and the application of DNA sequencing techniques. 31 oncogenes or tumor suppressor gene were included as: KRAS, NRAS, BRAF, PIK3CA, P53, APC, ATM, SMAD2/4, BRCA 1/2, MUTYH, FBXW7, ASXl1, TET2, DNMT3A, RUNX1, FAM123B, arid1a_b, RB1, RNF43, PTEN, FLT3, ZNF_217, SOX_9, BCOR, CDK8, IRS2 Amp, SRC Amp, MYC Amp, CTNNB1 NOTCH1/3, BCL2 Amp. For each gene, the number of mutated ones range from 3 to 133. Considering that the treatment results reflect the toxicity and effect of the cancerous drug, the genes predictors that are most significant can be inspired as effective predictors in later EWOC-NETS model for better maximum tolerated dose estimation.

2.1 Methods for gene profile modeling

2.1.1 Logistic Model Selection

Firstly, descriptive statistics are generated for each gene and the response. The original response variable is categorized as 0,1,2,3 according to RECIST. It is then re-categorized to 0(unknown or stable/progressive disease) and 1(complete/ (very good) partial response). Then logistic regression is applied to model the association between this binary outcome on all the genes variables. Backward and stepwise selections are performed to decide the most significant genes to be retained in the final logistic model. Selections are first performed based on the significance test with the retaining criteria as $\alpha=0.1$. Then selections are performed using AIC, which denotes the goodness of fit for the model with penalization for the increasing number of predictors. After that, best subset selection is conducted with restrictions on the number of predictors for comparison. The scientific procedures can be referred from the SAS user guide[6].

For the significance test, it applies the partial F test to decide whether the additional covariates are significant with formula below to calculate the F statistics:

$$\mathrm{F} = \frac{\frac{SSE_{reduced} - SSE_{full}}{k}}{MSE_{full}} \sim X_K^2 \qquad (2.1)$$

k is the number of additional gene predictors in the full model.

For the AIC selection, the lowest AIC is preferred with definition as:

$$\mathrm{AIC} = 2\mathrm{k} - 2\ln(\tilde{L}) \qquad (2.2)$$

Where K is the number of parameters in the model and $\tilde{L}$ is the maximized likelihood function.

The best subset selection fits all the possible models and the best model with restricted number of predictors is selected according to certain criterion including: a Mallows' Cp close to the number of predictors, high R^2or high adjusted R^2 . Here we use the global score chi-square statistic as the comparison standard.

2.1.2 Regularization Methods

After the routine logistic regression selection, regression with regularizations are conducted. Lasso, Ridge and Elastic Nets regression are all performed with certain types of restrictions on the regression coefficients to help shrink the effect of unimportant genes.

For lasso regression, the coefficients estimations are restricted as below:

$$\min_{\beta_0, \beta} \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta_0 - \boldsymbol{X_i^T \beta})^2 \right\} \text{ subject to } \sum_{j=1}^{p} \left|\beta_j\right| \leq \mathrm{t} \qquad (2.3)$$

Above equation can be rewritten as the Lagrangian form as below:

$$\min_{\beta \in R^p} \left\{ \frac{1}{N} ||\boldsymbol{y} - \boldsymbol{X\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1 \right\} \qquad (2.4)$$

The regression coefficients are solved as $\tilde{\beta}^{lasso}$ with λ as the shrinkage parameter that control the amount of regularization and the size of the coefficients[7].

Similarly, for Ridge regression, below equation is solved to get $\tilde{\beta}^{ridge}$ [7]:

$$\min_{\beta \in R^p} \left\{ \frac{1}{N} \|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\} \tag{2.5}$$

For Elastic Net regression, the restriction for $\tilde{\beta}^{elastic\ net}$is shown as below[8]:

$$\min_{\beta \in R^p} \left\{ \frac{1}{N} \|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\beta\|_2^2 \right\} \tag{2.6}$$

In this study, the optimal parameter lambdas are selected with 10-fold cross-validation. Then genes with relatively large coefficients are retained in Lasso. Though the ridge regression cannot select predictors, it helps verify the results of Lasso through keeping the effective variables with relative large coefficients. Elastic Nets regression, which considered to be a combination of Lasso and Ridge, can further help compare the results. The retained genes in this part will be compared to the result of 2.1.1 to better decide the gene candidates for EWOC-NETS.

2.1.3 Principle Components Analysis

The third method is principle components analysis (PCA). PCA is a variable reduction procedure to deal with overfitting and correlated predictors. The predictors that contribute to the variety in the same direction will be incorporated into single artificial variables, which, are called principle components. Mathematically, a principle component is defined as the linear combination of optimally-weighted original variables[9].

Let components to be expressed as a l× 1 vector $t_i = (\ t_1, t_2, \ldots\ldots, t_l)_i$ and let the weight to calculate components scores to be defined as a l×p matrix of weight coefficient $w_k = (w_1, w_2, \ldots\ldots, w_p)_k$, where p is the number of parameter. X is defined as the data matrix. The weight for the first components can be calculated as:

$$w_1 = argmax_{\|w\|=1}(\|\boldsymbol{Xw}\|^2) \qquad (2.7)$$

Then the weight for kth principle components score can be calculated as:

$$w_k = argmax_{\|w\|=1}(\left\|\widehat{\boldsymbol{X_k}}\boldsymbol{w}\right\|^2) \qquad (2.8)$$

where $\widehat{X_k}$ is the matrix of X that subtract the first k-1 principle components;

Then, the complete principle components can be derived as:

$$\mathbf{T}=\mathbf{XW} \qquad (2.9)$$

After generating all the principle components, the next step is to decide how many components to be kept. The retained components are desired to adequately represent the majority of variance. The final components retained will be decided base on criterion including the eigenvalues (>1), the scree plot (first large break), the variance proportions for each component (5%-10%) and the interpretation convenience (simple structure).

2.1.4 Random Forest Analysis

The last method implemented in this study is the random forest tree analysis. It combines multiple decision trees as the building blocks, which does classification work using different features of the data. The key concept behind it is to let all the decisions trees to work as a committee and then get the final classification decision by checking the most frequently appeared results. In this way, random forest can perform an ensemble prediction. In this process, the importance of the features for accurate classification can be determined with simulations[10]. The importance of the predictors is decided through calculating the reduction in node impurity contributed by each variable and then weighted by the probability of achieving that node. The node probability can be computed by dividing the number of samples that reach the node with the total number of samples. The variable resulting in largest reduction of node impurity is the most important feature. Below are the corresponding equations:

$$\mathrm{R}Ffi_i(importance) = \frac{\sum_j normfi_{ij}}{\sum_{j\,\in all\,features, k \in all\,trees} normfi_{jk}} \qquad (2.10)$$

$$normfi_j = \frac{fi_j}{\sum_{j\,\in all\,features} normfi_j} \qquad (2.11)$$

, where $fi_{jk}\ is\ the\ frequency\ of\ label\ i\ in\ a\ node$

1000 simulations are conducted here with 70% training data and 30% validation data. Optimal splitting variable number is chosen first according to predictive accuracy for later simulation. The number of trees is selected to be 1000. The importance of genes in terms of the contribution to the accuracy is summarized to determine that which genes are more helpful to predict the true outcomes in validation set. The R package "randomForest" is applied to complete above procedures[11]. Results of this random forest analysis will be compared with the other three methods to check thatif there are any overlapped recommended genes, which have the potential to be effective in EWOC-NETS. The concepts of EWOC-NETS model with additional covariates can be viewed from below section 2.2.

2.2 EWOC-NETS model that utilize gene information

EWOC is an adaptive Bayesian model using a binary outcome to denote the appearance of dose limited toxicity (DLT). The detailed design can be viewed from [3]. The EWOC can be extended to EWOC-NETS by using the normalized equivalent score as the outcome variable. Basically, NETS is a novel way to represent the dose toxicity as a quasi-continuous variable. It utilizes all the toxicity information and help to get a more accurate estimated MTD compared to the traditional way, which, only use the worst toxicity for screening [4].

The first step to generate the score is to adjust the original grade of toxicity as below:

Original grade	Grade 0	Grade 1	Grade 2	Grade 3 Non-DLT	Grade 4 Non-DLT	Grade 3 DLT	Grade 4 DLT
Adjusted grade	0	1	2	3	4	5	6

Using the adjusted grade, NETS can be derived as:

$$S_i = \left[G_{i,max} - 1 + \frac{\exp\left\{\alpha + \beta\left(\sum_{j=1}^{J_i} \frac{w_{i,j} G_{i,j}}{G_{i,max}}\right)\right\}}{1 + \exp\left\{\alpha + \beta\left(\sum_{j=1}^{J_i} \frac{w_{i,j} G_{i,j}}{G_{i,max}}\right)\right\}} \right] \qquad (2.12)$$

Here, S_i is the NETS for the ith patient. $G_{i,max}$ is the worst adjusted toxicity grade for this individual. J is the total number of toxicities for each patient. $w_{i,j}$ is a specific weight of the jth toxicity for the ith individual. This weight, ranging from 0 to 1, is set to be 1 if it is uncorrelated with other toxicities and 0 if exactly a duplicate of other toxicities. α is the parameter to control the score interval. β is assumed to > 0 to make sure the non-decreasing of toxicity.

Considering the number of patients achieving DLT follow a quasi-distribution that is binomial distributed, which, belongs to the family of linear exponential distribution, the quasi maximum likelihood estimation for the MTD can be highly consistent[12, 13]. The detailed incorporation steps can be checked from [2].

Based on the framework of EWOC-NETS, additional covariates such as gender, race, or more microscopic, gene profile can be incorporated to estimate personalized MTDs. The first step to add new covariate is to add one categorical variable with values of 0 and 1. Follow the well-established EWOC-NETS setting, one additional categorical variable, with 0 as no gene mutation and 1 as gene mutation, will be incorporated as C. Then we can write the model as:

$$\mu_{S_i|X_i} = F(\beta_0 + \beta_1 X_i + \delta C_i) \qquad (2.13)$$

Here X_i is the assigned dose level. $\mu_{S_i|X_i}$ is called average normalized equivalent toxicity score. It is the expected value of NETS given dose level as X_i. C_i is the binary covariates providing gene mutation information and δ is the coefficient for C.

$\mu_{S_i|X_i}$ is also assumed to follow a logistic distribution, then reparameterization can be performed using following values: MTD: γ, the corresponding ANETS: θ, which show the worst acceptable toxicity level quantitatively with gene mutation. The start dose X_{min} in group 1 with no gene mutation and corresponding ANETS ρ_1. The start dose X_{min} in group 2 with gene mutation and corresponding ANETS ρ_2.Then e can write below equations:

$$\text{Logit}(\rho_1) = \beta_0 + \beta_1 X_{min} \qquad (2.14)$$

$$\text{Logit}(\rho_2) = \beta_0 + \beta_1 X_{min} + \delta \qquad (2.15)$$

$$\text{Logit}(\theta) = \beta_0 + \beta_1 \gamma + \delta \qquad (2.16)$$

In this way, all the coefficient can be rewritten as:

$$\beta_0 = logit(\rho_1) - \frac{X_{min}}{\gamma - X_{min}}[\gamma logit(\theta) - X_{min}\, logit(\rho_2)] \qquad (2.17)$$

$$\beta_1 = \frac{1}{\gamma - X_{min}}[logit(\theta) - logit(\rho_2)] \qquad (2.18)$$

$$\delta = \text{logit}(\rho_2) - logit(\rho_1) \qquad (2.19)$$

The prior distribution of (γ, ρ_1, ρ_2) can be assigned as non-informative uniform distribution in $[X_{min}, \ X_{max}] \times [0, \ \theta] \times [0, \theta]$. Then the quasi-Bernoulli likelihood is applied as mentioned in section 2.4.

$$\tilde{L}(\gamma, \rho_1, \rho_2 \ | D_k) = \prod_{i=1}^{k} F(\beta_0 + \beta_1 X_i + \delta C_i)^{S_i} * (1 - F(\beta_0 + \beta_1 X_i + \delta C_i))^{1-S_i} \qquad (2.20)$$

Now, reparametrize (23) with (20) – (22), and we can get the posterior distribution of (γ, ρ_1, ρ_2)as:

$$\pi_k(\gamma, \rho_1, \rho_2 \mid D_k) \propto \tilde{L}(\gamma, \rho_1, \rho_2 \mid D_k) * \mathrm{h}(\gamma, \rho_1, \rho_2,) \propto \tilde{L}(\gamma, \rho_1, \rho_2 \mid D_k) \qquad (2.21)$$

The corresponding posterior CDF of γ can thus be derived as $\pi_{k,C_{k+1}}(X_{k+1}|D_k)$, the details of this derivation can be checked from [1]. For the next patients, select the dose as $\pi^{-1}_{k,C_{k+1}}(\alpha)$ to control the probability of overdose. Upon the completion of the trial, the posterior median of dose will used as MTD estimate generated by MCMC sampler[14]. The derivation for additional continuous covariates and the simulation result can be viewed from [1]. This model can help estimate personalized MTDs for two groups with smaller standard error, bias and MSE compared to EWOC-NETS , which only estimates the marginal MTDs for all patients. Through combining the results from section 2.1 with selected candidates of gene predictors, simulation will be performed using this "EWOC-NETS with covariates" framework as follow: The logistic regression using tumor response as outcome and number of treatments as dose is fitted as an original model. Then the estimated probability will be regarded as the probability of DLT considering that DLT and complete response has a positive correlation. This probability of DLT can be further regarded as NETS, which is the normalized equivalent toxicity score using all the toxicity information ranging from 0 to 1. Finally, we fit logistic regression again using response as outcome and use number of treatments as dose, together with selected gene predictors as covariates and analyze that how the precision of dose estimation changes with the effect of gene information. The optimal method for gene profile integration can then be evaluated and decided based on the precision of dose estimation and the simplicity. The conclusion, thus, may applied later in practical, real-world EWOC-NETS trial analysis.

3. Results:

According to the supplementary table 1, among 161 patients, 59 individuals got positive treatment response and 102 individuals got negative response. For the gene mutations, most of them have a relatively small percentage of mutated expressions about 5% to 20%. There are also some unbalanced genes, which may not be good for prediction. With 31 genes that provide either similar or different information and a small sample of 161 patients as a phase I trial setting, integration and dimension reduction are required to be performed to get the most significant ones for later dose estimation. From the univariate association analysis shown in supplementary table 2, it can be seen that the mutation of BRAF, ATM, FLT3 and ZNF_217 are significantly associate with treatment response in a 0.05 significance level.

3.1 Factor Selection:

3.1.1 Logistic Model Selection

According to table 1, the models after backward selection based on significance test is:

Model 1:

Logit(p(Response=1)) = -0.7465+ 2.064*ATM – 1.924*BRAF + 1.152*NRAS + 1.059*FLT3

The model after stepwise selection based on significance test is:

Model 2:

Logit(p(Response=1)) = -0.9681+ 3.247*ZNF_217 + -2.4212*BRAF + 2.248*ATM + 1.2737*NRAS + 1.206*FLT3 + 1.195*FBXW7

Genes retained including ATM, BRAF, NRAS and FLT3 in model 1. Compared with the backward selection, two additional genes are kept in model 2, which are ZNF_217 and FBXW7. At the same time, the most significant gene change is from ATM to ZNF_217.

The model after backward/stepwise selection(table 3) base on AIC value is:

Model 3:

$$\mathrm{logit}(p(\mathrm{Response}=1)) = -0.543 + 2.12*\mathrm{ATM} -2.11*\mathrm{BRAF} + 1.186*\mathrm{FLT3}+ 1.063*\mathrm{FBXW7} + 1.021*\mathrm{NRAS}- 0.513*\mathrm{KRAS}$$

Gene predictors including ATM, BRAF, FLT3, FBXW7, NRAS and KRAS are retained.

Model 4:

$$\mathrm{Logit}(p(\mathrm{Response}=1)) = -0.662 + 3.29*\mathrm{ZNF\_217}- 2.75*\mathrm{BRAF}+ 2.36*\mathrm{ATM}+ 1.33*\mathrm{FLT3}+ 1.179*\mathrm{FBXW7}+ 1.147*\mathrm{NRAS} – 0.557*\mathrm{KRAS}$$

Gene predictors including ZNF_217, ATM, BRAF, FLT3, FBXW7, NRAS and KRAS are retained. The model of best subset with one to seven predictors are listed in table 2. It is highly consistent with the results of above backward and stepwise selection. These results suggest that when use AIC criteria for model comparison, more gene predictors are retained. Considering that AIC penalizes the number of predictors, a simpler model is preferred. This is contradictory to the results now, suggesting that AIC use different selection mechanism in comparing with significance test. Thus, further study is necessary for the characteristics of this type of binary gene mutation marker selection. Another noticeable point is that compared with backward selection, the most effective covariate changes are from ATM to ZNF_217 for both significance test method and AIC method. The corresponding confidence interval can be checked from table 1. Overall, backward selection shows a better performance to prevent the involvement of gene whose effect has large standard error here.

3.1.2 Regularization Methods:

According to table 3, for ridge regression, several genes with largest coefficients are similar to the previous results in 3.1.1 such as ZNF_217, BRAF, ATM and FBXW7. Some additional potential candidates that are different compared to 3.1.1are RUNX1 and

IRS2 AMP. For lasso regression, seven genes are kept using λ that minimizes the cross-validated sum of squared residuals. Those genes are NRAS, BRAF, ATM, FBXW7, FLT3, RUNX1 and ZNF_217. This result is also similar to the 3.1.1 and ridge regression. It is noticeable that RUNX1 is an additional recommended gene in LASSO and Ridge regression without appearing in the previous 3.1.1 results. There was only one gene, ZNF_217 kept in elastic net regression, which was also suggest to be the most effective gene predictor in stepwise selection using significance test, AIC, Ridge and Lasso. Above similarity in results indicates that the first two methods are fairly consistent although there are not exactly the same. These three regularization methods selected similar gene predictors, ZNF_217, BRAF, FBXW7, and ATM. As can be seen from table 3, some other genes including IRS2_Amp and RINX1 are also potential genes candidates that are retained in regularization with smaller coefficients, suggest less strength as the scale of each genes are the same as 0 and 1

3.1.3 Principle Components Analysis

After the first try of extraction, twelve components are retained. The foundamental retaining criteria is to make sure that the eigenvalue is larger than 1, which means that each components kept contribute larger variances compared to the original gene predictors. There is also no eigenvalue such as 0.99 to make this criteria ambiguous (Table 4). Another retaining criteria is to make sure that the variance proportion for retained components, over that total variance (31 here) are larger than 5% or totally larger 70%. This criteria is slightly violated since that started from the 7th components, the proportions are smaller than 5% and the total proportion is 63.9%, which smaller than 70%. (table 4). The scree plot shows the largest break appears after the third component.

At the same time, the first six components meet the criteria that “one component is supposed to have at least three loading variables.” Considering the overall above criterion, the first three components were selected for future EWOC-NETS simulation. After the additional PCA process with restriction of three components as shown in table 5, for the first components, BRAF and RNF43 take large loading proportion. For the second component, KRAS, ZNF_217 and SRC_Amp, BCOR, BCL2_amp all take large proportions. For the third components, FLT3 takes the largest proportion. The information above represent the most valuable variety of predictors, which is partly overlapped with previous result in 3.1.1 and 3.1.2 that the important effects are contributed by BRAF, ZNF_217, KRAS and FLT3. However, there are also some difference: some other genes, such as RNF43, BCOR and SRC_Amp, seem to represent major variance. In addition, the previous recommended gene ATM does not appear as the loading variable in those three retained components. Overall, PCA works somehow differently compared to previous methods. Considering that PCA value the variance more than the association without using outcome information, it might not be the first Choice so far for gene mutation data integration, though it still provides supportive information of effective genes.

3.1.4 Random Forest Analysis

The most important genes contributing to the accuracy for each simulation are recorded and their frequencies are also summarized after 1000 simulation as shown in figure.1. Four variables are decided as the optimal number to be sampled at each split in random forest analysis. In this simulation, ZNF_217 is the most important gene for over 500 times, followed by ATM, FLT3 and NRAS. This importance of genes is similar to previous model selection and regularization methods, supporting that ZNF_217, ATM, FLT3 can be commonly recommended for EWOC-NETS model candidates for potential accurate outcomes.

3.2 Incorporation of Gene Mutation Profile to EWOC-NETS:

The summarized simulation results can be seen from table 6. The intercept is the baseline log odds of get positive response or the log odds of DLT or the log of NETS/(1-NETS). The coefficient of dose denotes the log odds ratio when there are one unit increasing of dose.. It can be found that after incorporating genes variables into the model with dose as the main effect and the treatment response as the binary outcome, the standard error for the dose effect for all methods slightly increased in a small scale of (0.0037, 0.0146), which is fine with the aim to compare different methods and find the one with optimal precision for dose estimation. In this kind of situation, we can choose the model considering both precision and the simplicity of the model, which related to the overall efficiency of the trial. From table 6. it is found that when use Elastic Net or the PCA, the increasing of standard error is the smallest as 0.0037, suggesting that those two methods can be better choices that give relatively precise dose estimation. At the same time, the Elastic Net is more efficient for it results in a more simplified model compared with PCA while it estimates dose as precise as PCA. In a phase I trial, we must consider the cost of effort and time for patients so that they get the most appropriate dose in clinical trials. Thus, the Elastic Net regression seems to be the optimal method according to this study.

4. Conclusion and Discussion:

In conclusion, with good consistency for all methods, the elastic net method is the most beneficial one considering its simplified final model with the most précised estimation for dose effect. For the future generalized procedures of gene profile integration to the EWOC-NETS, the elastic nets can be used as a main tool to summary gene mutation profile. Other methods of model selection, regularization and random forest can then be used to further help provide supportive information with overlapped retained elements to verify the gene candidates that related to the treatment response. The PCA analysis with proper retained components can be suggestive reference but may not be very helpful considering that the outcome information is not used in this method With four methods applied in analysisthe highly overlapped genes can be used for further EWOC-NETS modeling process. In this specific study, ZNF_217, ATM, FLT3 are the recommended genes.

At the same time, it is surely necessary to repeat above analysis with more gene profile data to verify the conclusion. In addition, although using less gene information might cost less effort, we need to consider the interaction between genes. To further improve the procedure in this study, it might be a good idea to try to make the whole gene profile a comprehensive index, such as a quasi-continuous variable, and then include it in EWOC-NETS model. This can be a considerable way to estimated pMTDs using all gene information. At the same time, more information might be required to decide the correlation among genes as the weight for the single gene index . One previous idea suggests that to integrate gene profile data, genetic risk score can be used. It is similar to a propensity score with a previous univariate logistic regression analysis to calculated the odds ratio of gens as a weight[15]. Gene interactions involvement is surely worthy: previous research has shown that the interaction or the combination of cancer-related gene mutation can cause the cell to become easier to die, which inspired the clinical trial design to generate certain functionally gene interactions[16].

In this study, we also use tumor response instead of dose limited toxicity as outcome: this might be another supportive way to compared with the result of using dose limited toxicity as outcome to get a more considerable and comprehensive estimated personalized maximum tolerated dose. Another idea to help further develop the EWOC-NETS model with gene information is to use stratified randomized clinical trials to decide in what kinds of gene mutation groups, the patients are more sensitive to the drug[17]. In this way, we may find the groups that shows the most significant difference in dose effect, which help to verify the result of this study. From this study, we gain some inspiration of the most proper methods to select gene predictors for the dose estimation even when the gene mutation dataset is not large in a phase I clinical trial. For future plan, we plan to conduct a more comprehensive and intensive simulation with EWOC-NETS model. Considering that there are many gene expression pattern, setting the true maximum tolerated doses for each gene pattern can be time-consuming in simulation. However, it is valuable to conduct further investigation so that we can gain a more valid insight of the advantages and drawbacks for different gene integration methods in dealing with highly correlated gene mutation data.

5. References:

1. Chen, Z., et al., *Adaptive Estimation of Personalized Maximum Tolerated Dose in Cancer Phase I Clinical Trials Based on All Toxicities and Individual Genomic Profile*. PLoS One, 2017. **12**(1): p. e0170187.
2. Chen, Z., M. Tighiouart, and J. Kowalski, *Dose escalation with overdose control using a quasi-continuous toxicity score in cancer Phase I clinical trials.* Contemp Clin Trials, 2012. **33**(5): p. 949-58.
3. BABB, J., A. ROGATKO, and S. ZACKS, *Cancer phase I clinical trials: efficient dose escalation with overdose control.* Statistics in medicine, 1998. **17(10)**: p. p. 1103-1120.
4. Chen, Z., et al., *A novel toxicity scoring system treating toxicity response as a quasi-continuous variable in Phase I clinical trials.* Contemp Clin Trials, 2010. **31**(5): p. 473-82.
5. *NCI Dictionary of Cancer Terms: RECIST*. National Cancer Institute 2019 [cited 2019 08/01]; https://www.cancer.gov/publications/dictionaries/cancer-terms/def/recist].
6. Bruce Lund, M.A.S., Detroit MI, Wilmington DE, Charlotte NC, *Logistic Model Selection with SAS PROC's LOGISTIC, HPLOGISTIC.* 2017: p. 1-18.
7. Robert, T., *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society, 1996. **Series B.58**: p. 266-288.
8. Hui, Z.H., Trevor;, *Regularization and Variable Selection via the Elastic Net.* JRSSB, 2003. **67 (2)**: p. 310-320.
9. Jolliffe, I., *Principal Component Analysis* 2002: Springer Berlin Heidelberg.
10. Breiman, L., *Random Forests.* . 2001(October 2001): p. 5-32.
11. Liaw, A, *Documentation for R package randomForest*. 2012.
12. Gourieroux C, M.A., Trognon A. , *Pseudo maximum likelihood methods: theory.* . Econometrica 1984. **52**: p. 681–700.
13. McCullagh P, N.J., *Generalized Linear Models*. 1989: New York: Chapman and Hall; .
14. Tighiouart, M., A. Rogatko, and J.S. Babb, *Flexible Bayesian methods for cancer phase I clinical trials. Dose escalation with overdose control.* Stat Med, 2005. **24**(14): p. 2183-96.
15. Wang, C., et al., *Genetic risk score: Principle,methods and application.* Zhonghua liu xing bing xue za zhi = Zhonghua liuxingbingxue zazhi, 2015. **36**: p. 1062-1064.
16. Lord, C.J., A.N. Tutt, and A. Ashworth, *Synthetic lethality and cancer therapy: lessons learned from the development of PARP inhibitors.* Annu Rev Med, 2015. **66**: p. 455-70.
17. Freidlin, B., L.M. McShane, and E.L. Korn, *Randomized clinical trials with biomarkers: design issues.* J Natl Cancer Inst, 2010. **102**(3): p. 152-60.

6. Tables and figures

Table 1 Odds Ratio Estimates of four selection process

		Odds Ratio Estimates		
Method	**Effect**	**Estimate**	**95% Confidence limits**	**P-value**
Backward selection (Partial F test)	ATM	7.875	(1.320, 42.967)	0.0235
	BRAF	0.146	(0.016, 1.344)	0.0894
	FLT3	2.884	(0.946, 8.792)	0.0626
	NRAS	3.165	(0.844, 11.869)	0.0876
Stepwise selection (Partial F test)	ATM	9.471	(1.44, 62.293)	0.0193
	BRAF	0.089	(0.007, 1.137)	0.0627
	FBXW7	3.305	(0.871, 12.538)	0.0789
	FLT3	3.34	(1.076, 10.367)	0.0369
	NRAS	3.574	(0.934, 13.682)	0.0629
	ZNF_217	25.708	(1.917, 344.704)	0.0142
Backward selection (AIC)	ATM	8.289	(1.361, 50.480)	0.0218
	BRAF	0.1	(0.012, 1.175)	0.0687
	FBXW7	2.895	(0.767, 10.926)	0.1167
	FLT3	3.274	(1.042, 10.288)	0.0423
	KRAS	0.599	(0.293, 1.223)	0.1592
	NRAS	2.775	(0.716, 10.753)	0.1398
Stepwise selection (AIC)	ATM	10.605	(1.592, 70.648)	0.0147
	BRAF	0.064	(0.005, 0.873)	0.0393
	FBXW7	3.251	(0.850, 12.438)	0.0851
	FLT3	3.785	(1.188, 12.057)	0.0243
	KRAS	0.573	(0.273, 1.200)	0.1398
	NRAS	3.15	(0.805, 12.322)	0.0993
	ZNF_217	26.934	(2.001, 362.500)	0.013

Table 2 Model selected using best subset method

Regression Models Selected using best subset method		
Number of Variables	**Chi-Square Score**	**Variables Included in Model**
1	5.8512	ZNF_217
2	11.7115	ATM, ZNF_217
3	16.972	ATM, BRAF, ZNF_217
4	20.5411	ATM, BRAF, FLT3, ZNF_217
5	24.1398	ATM, BRAF, FBXW7, FLT3, ZNF_217
6	27.2078	ATM, BRAF, FBXW7, FLT3, NRAS, ZNF_217
7	28.9367	ATM, BRAF, FBXW7, FLT3, KRAS, NRAS, ZNF_217

*By default, the criterion used to determine the "best" subset in SAS is based on the global score chi-square statistic.

Table 3 Regularization methods with final models

		Final model	
Methods	**Lambda**	**Predictors retained**	**Equation**
RIDGR	Minimum	ALL	logit(p(Response=1)) = -0.630 + 0.154*KRAS – 0.175*BRAF + 0.247*ATM + 0.174*FBXW7 – 0.225*RUNX1 + 0.306*ZNF_217 + 0.150*IRS2_Amp
LASSO	Minimum	NRAS, BRAF, ATM, FBXW7, FLT3, RUNX1, ZNF_217	logit(p(Response=1)) = -0.684 + 0.297*NRAS – 0.529*BRAF + 0.816*ATM, + 0.383*FBXW7 + 0.406*FLT3 – 0.163*RUNX1 + 1.138*ZNF_217
Elastic Net	Minimum	ZNF_217	logit(p(Response=1)) = -0.548+ 0.0175*ZNF_217

* Number observations is 161
* Minimum is the lambda with smallest MSE through 10-fold cross validation;

Table 4 Eigenvalues for each component with their proportion to the total variance and loading information

Component	Eigenvalue	Proportion	Cumulative	Loading Biomarkers
1	2.588	0.084	0.083	BRAF, FLT3, APC, RNF43, CDK8, NOTCH_1/3
2	2.408	0.078	0.161	KRAS, ZNF_217, BCOR, SRC_Amp, BCL2_Amp
3	2.269	0.073	0.234	FLT3, IRS2_Amp, CDK8, MYC_Amp, RB1
4	1.926	0.062	0.297	RUNX1, RB1, PTEN, BCOR
5	1.768	0.057	0.354	ZNF_217, SRC_Amp, TET2, FAM123B
6	1.594	0.051	0.405	SOX9, SMAD2/4, PIK3CA
7	1.469	0.047	0.452	MUTYH, arid1a/b
8	1.303	0.042	0.494	ATM
9	1.222	0.039	0.534	NRAS
10	1.143	0.037	0.571	FBXW7
11	1.089	0.035	0.606	N/A
12	1.028	0.033	0.639	SMAD2/4, CTNNB1

* 12 components are retained according to "Eigenvalue>1 criteria"

*Total variance is 31 since there are 31 gene predictors

* The loading criteria is that the variance of a certain gene contribute to a component is over 40%

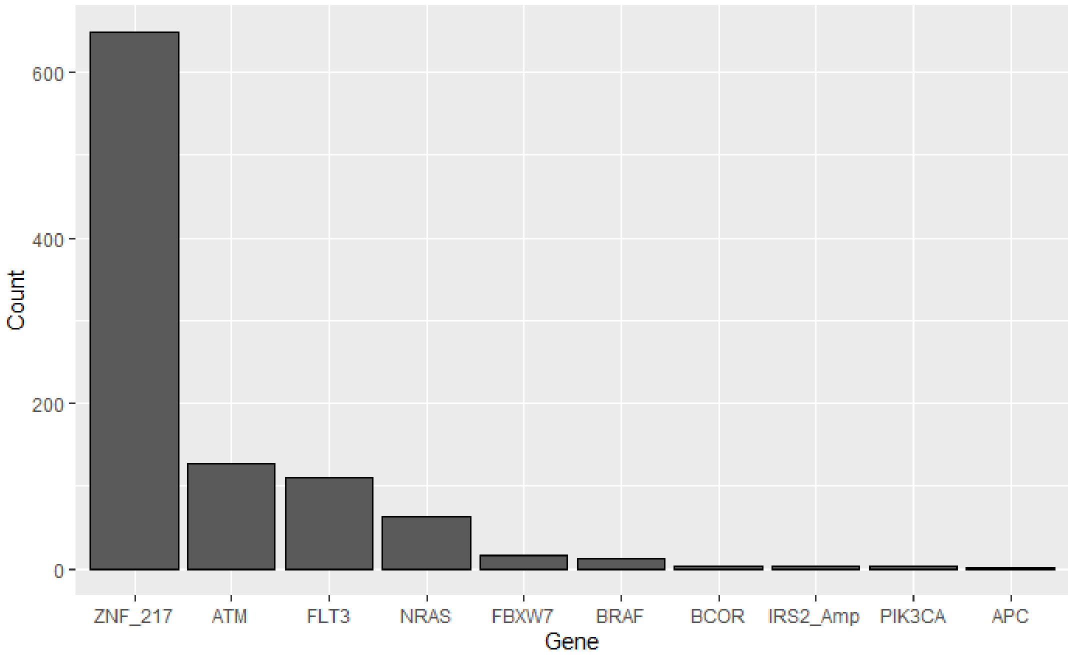
Table 5 Summary of PCA loading variables

		Final model
Methods	Component	Retained/important predictors
PCA	1	RNF43>BRAF>APC>NOTCH1/3>BRCA1/2>CTNNB1=MUTYH
	2	SRC_Amp>BCL2_Amp>ZNF_217=BCOR>ASXI1=KRAS>P53
	3	FLT3>CDK8>IRS2_Amp

Table 6 The result of simulation to compare the precision of dose estimation:

Method	Gene or Components	Intercept	SE	Dose	SE
Original fitting	Dose	-0.66	0.43	0.031	0.093
Backward (Partial F test)	Dose + ATM, BRAF FLT3, NRAS	-0.85	0.46	0.03	0.099
Stepwise (Partial F test)	Dose + ATM, BRAF, FBXW7, FLT3, NRAS, ZNF_217	-0.98	0.50	0.0090	0.11
Backward (AIC)	Dose + ATM, BRAF, FBXW7, FLT3, KRAS, NRAS	-0.73	0.49	0.049	0.010
Stepwise (AIC)	Dose + ATM, BRAF, FBXW7, FLT3, KRAS, NRAS, ZNF_217	-0.77	0.52	0.028	0.11
Ridge regression	Dose + ATM, BRAF, FBXW7, IRS2_Amp, KRAS, RUNX1, ZNF_217	-0.54	0.51	0.026	0.11
Lasso regression	Dose + ATM, BRAF, FBXW7, FLT3, NRAS, RUNX1, ZNF_217	-0.92	0.50	0.0066	0.11
Elastic Nets regression	Dose + ZNF_217	-0.65	0.44	0.012	0.097
PCA	Dose + first three components	-0.56	0.44	0.0055	0.097
Random forest	Dose + ATM, FLT3, NRAS, ZNF_217	-1.0	0.47	0.015	0.10

Figure 1 The importance of gene predictors in random forest tree analysis accuracy



*1000 simulations with two variables randomly sampled at each split and 1000 trees

7. Appendix:

Supplementary table 1 Descriptive statistics for all genes variables

Variable	Level	N (%) = 161
Response	0	102 (63.4)
	1	59 (36.6)
KRAS	0	73 (45.3)
	1	88 (54.7)
NRAS	0	151 (93.8)
	1	10 (6.2)
BRAF	0	150 (93.2)
	1	11 (6.8)
PIK3CA	0	133 (82.6)
	1	28 (17.4)
P53	0	33 (20.5)
	1	128 (79.5)
APC	0	28 (17.4)
	1	133 (82.6)
ATM	0	153 (95.0)
	1	8 (5.0)
SMAD2/4	0	142 (88.2)
	1	19 (11.8)
BRCA 1/2	0	153 (95.0)
	1	8 (5.0)

Variable	Level	N (%) = 161
MUTYH	0	156 (96.9)
	1	5 (3.1)
FBXW7	0	150 (93.2)
	1	11 (6.8)
ASXl1	0	157 (97.5)
	1	4 (2.5)
TET2	0	156 (96.9)
	1	5 (3.1)
DNMT3A	0	157 (97.5)
	1	4 (2.5)
RUNX1	0	157 (97.5)
	1	4 (2.5)
FAM123B	0	151 (93.8)
	1	10 (6.2)
arid1a/b	0	154 (95.7)
	1	7 (4.3)
RB1	0	158 (98.1)
	1	3 (1.9)
RNF43	0	157 (97.5)
	1	4 (2.5)

Variable	Level	N (%) = 161
PTEN	0	148 (91.9)
	1	13 (8.1)
FLT3	0	146 (90.7)
	1	15 (9.3)
ZNF_217	0	155 (96.3)
	1	6 (3.7)
SOX_9	0	148 (91.9)
	1	13 (8.1)
BCOR	0	152 (94.4)
	1	9 (5.6)
CDK8	0	148 (91.9)
	1	13 (8.1)
IRS2_Amp	0	153 (95.0)
	1	8 (5.0)
SRC_Amp	0	156 (96.9)
	1	5 (3.1)
MYC_amp	0	150 (93.2)
	1	11 (6.8)
CTNNB1	0	156 (96.9)
	1	5 (3.1)
NOTCH1/3	0	157 (97.5)
	1	4 (2.5)

Variable	Level	N (%) = 161
BCL2_amp	0	154 (95.7)
	1	7 (4.3)

Supplementary table 2 Univariate association for the treatment outcome with gene

			Response		
Covariate	**Statistics**	**Level**	**0 N=102**	**1 N=59**	**Parametric P-value***
KRAS	N (Col %)	0	44 (43.14)	29 (49.15)	0.460
	N (Col %)	1	58 (56.86)	30 (50.85)	
NRAS	N (Col %)	0	98 (96.08)	53 (89.83)	0.114
	N (Col %)	1	4 (3.92)	6 (10.17)	
BRAF	N (Col %)	0	92 (90.2)	58 (98.31)	**0.049**
	N (Col %)	1	10 (9.8)	1 (1.69)	
PIK3CA	N (Col %)	0	82 (80.39)	51 (86.44)	0.329
	N (Col %)	1	20 (19.61)	8 (13.56)	

			Response		
Covariate	**Statistics**	**Level**	**0 N=102**	**1 N=59**	**Parametric P-value***
P53	N (Col %)	0	22 (21.57)	11 (18.64)	0.658
	N (Col %)	1	80 (78.43)	48 (81.36)	
APC	N (Col %)	0	20 (19.61)	8 (13.56)	0.329
	N (Col %)	1	82 (80.39)	51 (86.44)	
ATM	N (Col %)	0	100 (98.04)	53 (89.83)	**0.021**
	N (Col %)	1	2 (1.96)	6 (10.17)	
SMAD2/4	N (Col %)	0	91 (89.22)	51 (86.44)	0.599
	N (Col %)	1	11 (10.78)	8 (13.56)	
BRCA 1/2	N (Col %)	0	96 (94.12)	57 (96.61)	0.483
	N (Col %)	1	6 (5.88)	2 (3.39)	
MUTYH	N (Col %)	0	99 (97.06)	57 (96.61)	0.874
	N (Col %)	1	3 (2.94)	2 (3.39)	

			Response		
Covariate	Statistics	Level	0 N=102	1 N=59	Parametric P-value*
FBXW7	N (Col %)	0	98 (96.08)	52 (88.14)	0.054
	N (Col %)	1	4 (3.92)	7 (11.86)	
ASXl1	N (Col %)	0	100 (98.04)	57 (96.61)	0.575
	N (Col %)	1	2 (1.96)	2 (3.39)	
TET2	N (Col %)	0	99 (97.06)	57 (96.61)	0.874
	N (Col %)	1	3 (2.94)	2 (3.39)	
DNMT3A	N (Col %)	0	100 (98.04)	57 (96.61)	0.575
	N (Col %)	1	2 (1.96)	2 (3.39)	
RUNX1	N (Col %)	0	98 (96.08)	59 (100)	0.123
	N (Col %)	1	4 (3.92)	0 (0)	
FAM123B	N (Col %)	0	97 (95.1)	54 (91.53)	0.365
	N (Col %)	1	5 (4.9)	5 (8.47)	

			Response		
Covariate	**Statistics**	**Level**	**0 N=102**	**1 N=59**	**Parametric P-value***
arid1a/b	N (Col %)	0	98 (96.08)	56 (94.92)	0.727
	N (Col %)	1	4 (3.92)	3 (5.08)	
RB1	N (Col %)	0	100 (98.04)	58 (98.31)	0.904
	N (Col %)	1	2 (1.96)	1 (1.69)	
RNF43	N (Col %)	0	99 (97.06)	58 (98.31)	0.624
	N (Col %)	1	3 (2.94)	1 (1.69)	
PTEN	N (Col %)	0	92 (90.2)	56 (94.92)	0.290
	N (Col %)	1	10 (9.8)	3 (5.08)	
FLT3	N (Col %)	0	96 (94.12)	50 (84.75)	**0.049**
	N (Col %)	1	6 (5.88)	9 (15.25)	
ZNF_217	N (Col %)	0	101 (99.02)	54 (91.53)	**0.016**
	N (Col %)	1	1 (0.98)	5 (8.47)	

Covariate	Statistics	Level	Response 0 N=102	 1 N=59	Parametric P-value*
SOX_9	N (Col %)	0	93 (91.18)	55 (93.22)	0.646
	N (Col %)	1	9 (8.82)	4 (6.78)	
BCOR	N (Col %)	0	95 (93.14)	57 (96.61)	0.355
	N (Col %)	1	7 (6.86)	2 (3.39)	
CDK8	N (Col %)	0	95 (93.14)	53 (89.83)	0.458
	N (Col %)	1	7 (6.86)	6 (10.17)	
IRS2_Amp	N (Col %)	0	99 (97.06)	54 (91.53)	0.120
	N (Col %)	1	3 (2.94)	5 (8.47)	
SRC_Amp	N (Col %)	0	99 (97.06)	57 (96.61)	0.874
	N (Col %)	1	3 (2.94)	2 (3.39)	
MYC_amp	N (Col %)	0	95 (93.14)	55 (93.22)	0.984
	N (Col %)	1	7 (6.86)	4 (6.78)	

			Response		
Covariate	**Statistics**	**Level**	**0 N=102**	**1 N=59**	**Parametric P-value***
CTNNB1	N (Col %)	0	98 (96.08)	58 (98.31)	0.433
	N (Col %)	1	4 (3.92)	1 (1.69)	
NOTCH1/3	N (Col %)	0	100 (98.04)	57 (96.61)	0.575
	N (Col %)	1	2 (1.96)	2 (3.39)	
BCL2_amp	N (Col %)	0	98 (96.08)	56 (94.92)	0.727
	N (Col %)	1	4 (3.92)	3 (5.08)	

* The parametric p-value is calculated by chi-square test.