Distribution Agreement

Signature:

_______________________________ _______________

Xiaozhu Zhang Date

Predicting Combined Chemotherapeutic Agents' Efficacy Synergy via Multiple Regression Models and Cross Validation Technique, Upon Inter-Patient and Intra-Patient Levels

By

Xiaozhu Zhang
MSPH

Biostatistics and Bioinformatics Department

Zhaohui (Steve) Qin, PhD
Committee Chair

Xiangqin Cui, PhD
Committee Member

Predicting Combined Chemotherapeutic Agents' Efficacy Synergy via Multiple Regression Models and Cross Validation Technique, Upon Inter-Patient and Intra-Patient Levels

By

Xiaozhu Zhang

B.A., University of Iowa, 2017

Thesis Committee Chair: Zhaohui (Steve) Qin, 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 Department
2019

Abstract

Predicting Combined Chemotherapeutic Agents' Efficacy Synergy via Multiple Regression Models and Cross Validation Technique, Upon Inter-Patient and Intra-Patient Levels

By Xiaozhu Zhang

Background: Precise medicine is crucial to cancer treatment for minimizing potentially lethal side-effects and maximize drug efficacy, and accurately modeling the individual drug efficacy is the key step. However, previous studies mostly modeled single drug efficacy while combination chemotherapy is more frequently applied in practice. In this study, we compared and integrated several models and algorithms to predict individual multiple-drug-polymer response on both intra-patient and inter-patient levels. Eventually, we aim to push cancer treatment one step down to the road of precise medicine.

Methods: We are interested in three key variables: two drug dosages, gene expression, and gene mutation. By adding these variables one by one to the model and evaluating model performance, we can determine their relative importance in the prediction. Linear regression, ridge regression, lasso regression, elastic net regression and random forest algorithms are applied in model construction. The goodness of fit is evaluated through R-square value tested by 10-fold cross validation and leave one out cross validation. Model was built upon single cell line data as well as data composed of four cell lines' information to investigate models' ability of predicting synergy at inter-cell-line level and intra-cell-line level.

Results and Conclusion: Compared to baseline model in which dosage information are only explanatory variables, secondary model with added gene expression data generated significantly larger R-square. However, adding mutation data into final model did not improve model accuracy, and R-squares are nearly the same to secondary model. In addition, model built upon multiple cell lines were incompetent in predicting drug synergy. Among five regression methods, random forest algorithm consistently produces largest R-square in each model. 10-fold CV is proved to have better generality and LOOCV coupled with random forest algorithm built best model. In conclusion, this study proved feasibility of predicting multiple chemotherapeutic agents' efficacy synergy utilizing their dosage information and gene expression data with-in cell line. The efforts of adding mutation information returned result that lower than expectation. More information is needed to model the drug synergy among patients.

Keywords: Bioinformatics, Penalized regression, Machine learning, Cross Validation, Cancer, Two-drug efficacy synergy, Model comparison

Predicting Combined Chemotherapeutic Agents' Efficacy Synergy via Multiple Regression Models and Cross Validation Technique, Upon Inter-Patient and Intra-Patient Levels

By

Xiaozhu Zhang

B.A., University of Iowa, 2017

Thesis Committee Chair: Zhaohui (Steve) Qin, 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 Department
2019

TABLE OF CONTENTS

Introduction

With help of the whole genome sequencing, our understanding of various types of cancer improved greatly over past few decades [1]. However, despite the progress made in oncology, current biomarker technology is inadequate to serve the purpose of precision medicine to personalize cancer therapy and to extend the survival of patients [2]. Tailoring of medicine in oncology is of great importance, not only for its patient-oriented design to maximize the drug function, but also for minimizing the side effect of Chemotherapy ranging from slight hair loss to as severe as death [3]. For designing patient-oriented drugs, the ability to precisely predict chemotherapeutic response upon patients' individual clinical record is required. Multiple studies have built models with different algorithms trying to predict the drug response, and a few studies succeed to generate prediction with relatively high accuracy [4][5]. However, there're a few limitations to previous studies that impede their method to be applied in practical terms. First, many of previous studies utilized training and testing microarray data form different platforms, which makes the application of personalized medicine unrealistic [6]. Second and most importantly, current treatment for cancer frequently involves combination of several drugs, while most of previous studies focus on predicting efficacy of single drug. In human body, multiple chemotherapeutic agents interact with each and kill cancer cells. Main benefits of polymer-drug formation include lower chance of causing a drug-resistant tumor and allow to reduce dosage of single agent therefore lower the risk of generate side-effects, thus combination chemotherapy is considered cornerstone of contemporary therapeutic interventions. Nevertheless, the drug effect synergy of polymer could be synergic, addictive or antagonistic depending on the ratio of different drugs.

Wrong combination of drugs/drug ratios increase the likelihood of unwanted side effects. In other words, carefully exam the ratio of drugs and other factors that might affect the polymer synergy is key step of successful combination chemotherapy. Major obstacles for doing this including the following: complexity of chemotherapeutic drug response, as it would be affected by both congenital genetic traits as well as postnatal interaction with environment and personal lifestyle; scarce of clinical trial data with integrated clinical record leading to less power of the prediction; many patients receive more than one type of cancer treatment such as surgery and radiation therapy, which makes the prediction trickier and less accurate.

In this study, we compared and integrated several models and algorithms to predict individual multiple-drug-polymer response quantified by synergy index. Besides the ratio of dosage, we are interest in gene expression and mutation information as well, as these are commonly believed to have impact on drug effect. By including and excluding these key explanatory variables to the model, we can tell which variable weights than others in predicting the drug synergy. We built three models in total: baseline model which contains only dosage information as explanatory variable and synergy index as response factor; secondary model adds gene expression information as additional explanatory variable; final model further includes mutation information. For each model, we applied following algorithms: linear regression, ridge regressions, elastic net regression, lasso regression and random forest. Ridge, elastic net and lasso regression are implemented through R package "glmnet", and random forest algorithm is conducted via R package "caret". We trained our data upon data collect from single patient as well as from multiple patients combined, for testing the model performance on both within patient level and inter-patient level. We implemented 10-fold cross-validation (CV) and leave one out cross-validation (LOOCV) testing method on our models to test on our ability to capture the information of variability on drug response. This procedure is performed through R package “caret”.

To address the scarce of data availability issue, our gene expression information is collected via *in vitro* cell line drug response data employed from The Genomics of Drug Sensitivity in Cancer Project (GDSC) [7]. A cell line is a cell culture that proliferate permanently, and all the cells are originated from single individual, which assures the consistence of their gene expression and mutation level. GDSC project is an international cooperation project funded by Wellcome corporation, containing nearly 700 cancer cell lines and more than a hundred drug responses. Data published in GDSC website is rather comprehensive, including many aspects of cell line data. Except basic clinical information of patients who provided human cells in the GDSC project, other datasets we used in this study consist of following: expression data of RMA; Single-nucleotide polymorphism (SNP) that stands for genomic variants found in cell lines; Natural log half maximal inhibitory concentration (IC50), which counts for the single drug response. Another database we adopt into our study is COSMIC, which is short for the Catalogue of Somatic Mutations in Cancer [8]. It is the largest and the most comprehensive resource for exploring the impact of somatic mutations in human cancer in the world, and it outlines data in terms of structure, scope and content. COSMIC contains basically two kinds of data, one is high precision data, manually curated by experts, another one is genome-wide screen data. These two kinds of data together provide comprehensive coverage of cancer genomic landscape from somatic perspective. New and significant data are continually added in COSMIC and are available, people can search a gene, cancer type, and mutation through COSMIC. We recruited demographic information for cell lines, oncogene census and mutational information from COSMIC. Finally, synergy index and drug dosage ratio are obtained from Emory laboratory. In this dataset, combination of two drugs: Methotrexate and Vincristine are tested on thirteen cell lines and synergy of each combination of two drug dosage is recorded.

In conclusion, we evaluate different algorithms upon various explanatory variables, for prediction of drug effect synergy on both intra-patient and inter-patient levels. Eventually, we aim to push cancer treatment one step down to the road of precise medicine.

Methods

1. Data description.

In this study, we were interested in three key explanatory variables: Methotrexate and Vincristine, two cancer drugs' dosage ratios; cell line gene expression information; cell line gene mutation information. The dataset we acquired from Emory laboratory contain two drugs' dosage information in two formats: numerically in logarithm of mole that directly indicates dosage level of two drugs, and in a categorical indicator varies from 0 to 9 for Methotrexate Dose and 0 to 7 for Vincristine Dose, each number indicate a dosage level of a single drug (80 levels in total). The synergy index of two drugs in thirteen cell lines served as dependent factor in the model. Its value varied from -1 to 1, positive and negative value represent synergic and antagonistic correspondingly, and zero stood for addictive, meaning no interaction effect among two drugs. Expression data was collected from “RMA normalized expression data for Cell lines” dataset of GDC database. This dataset contained 17737 genes' expression level in 1019 cell lines. For the sake of computational efficiency without losing too much information, we picked a small group of oncogenes instead of using the whole somatic gene base. The selection of genes was based on the cancer gene census project of COSMIC. The cancer gene census listed genes that have mutations that believed to be related to cancer. Oncogenes enlisted in cancer gene census contain two tiers, standing for two level of relevance to cancer. Tier one gene showed strong, documented evidence that their activities were affected by cancer status, or mutations in these genes were related to cancer. For a gene to be listed in the tier two, it must show strong hint of being a role in cancer, but there was no extensive available evidence to support this relationship. There were in total 723 oncogenes selected through cancer gene census project, including 576 tier one oncogene and 147 tier two oncogenes. By cross-comparing two datasets, we chose 541 oncogenes to be studied in this project. Mutation information of selected genes were obtained from All Mutations in Census Genes dataset of COSMIC. We used a 0/1 dummy variable to indicate the mutation status among these oncogenes. Descriptive statistics showed that most of selected genes were mutated (98%).

2. Baseline model establishment:

$$\text{Synergy} \sim \text{Methotrexate dosage} + \text{Vincristine dosage}$$

First step, we built baseline model and evaluated the model performance. We had two purposes of doing this: First, since the baseline model was relatively small (only contained two explanatory variables), it was computational efficient for us to test and compare various algorithms and methods of model validation; Second, evaluation of model performance could be used as reference and to be compared to more complex models to evaluate the contribution of gene expression and mutation in predicting the two drug synergy. In this study, we used R-square as index for model accuracy. R-squared is a statistical measure of how close the data is to the fitted regression line, it's the percentage of the response variation that is explained by the fitted model. The higher R-squared is, the better the model fits the data. In addition to R-square, we also calculated Root-mean-square deviation, RMSE, and Mean absolute error. MAE and RMSE are the two most common methods to evaluate the accuracy of continuous variables and can express the average model prediction error. MAE is mean absolute error. It's the average over the sample of the absolute difference between the predicted value and the actual value, with every absolute difference the same weight. RMSE is the square root of the average of squared differences between the predicted value and the actual value.

We began with predicting the two-drug synergy with linear regression. The exact dosage (in moles) and categorical variable were both tested. We tested the model accuracy by conducting the leave one out cross validation (LOOCV) and 10-fold cross validation. Cross validation (CV) is used to estimate the performance of a machine learning model in general when used to predict on data which is not used for building the model. It results in less bias than other methods such as train/test split. K-fold CV means splitting data into k groups, using one of the k groups as the test group and the rest k-1 groups as training groups. LOOCV is a special case of K-fold CV where k equals the number of samples in the data set. Both procedures are performed using R package “caret”. The results generated from two formats of dosage are compared and the one with higher R-square is chosen for subsequent investigations.

To address the potential higher dimensional association between synergy and explanatory variables, we built models with penalized terms. In our case, the datasets had a large number of covariates relative to the number of samples. With such high number of features, it was possible that multicollinearities exist, and ordinary least square in linear regression cannot be defined because linear regression was usually for low dimension. Thus, it’s more accurate to add penalized terms. R package ‘glmnet’ can do penalization. There are basically three types of regressions with penalty: Ridge, Lasso and elastic net. Ridge regression has L2 penalty term, it has easier calculation, but it does not shrink parameters to zero, so we can’t use it to do variable selection. Ridge will retain all the features but will shrink the coefficients; therefore, the model will remain complex, which may lead to relatively poor model performance. Lasso regression has L1 penalty term, which is more difficult to calculate but it can shrink parameters to zero, which enables the extraction of important predictive variables and simplification of the model. Elastic net is a hybrid of Ridge regression and Lasso regression. We can alter a parameter to weight these regressions in a mixture. We employed “train” function of “caret” package to tune for best parameters (alpha, lambda) for elastic net regression. Similarly, a LOOCV and 10-fold CV were conducted for each of three regression models. Lastly, we implemented random forest algorithm through “caret” package to fit the model. Random forest model is an additive model that combines decisions from multiple base models to make a more accurate prediction. Base models are constructed by different subgroups of the data, and each base model is a simple decision tree, and is independent from others. Random forest randomly splits the data, and it finds the most important feature of each subset. Random forest regression is good for handling tabular data with numerical features, which fits our data sets in this study. In addition, random forest algorithm has following advantages in comparison with machine learning methods: fitting non-linear regression in each base model, finding non-linear interactions between the feature and the subject; reducing overfitting by paralleled base model construction. Furthermore, random forest runs efficiently on large data sets and provides a more accurate result. We applied LOOCV and 10-fold CV to test model performance. Model performances were compared based on R-square, RMSE and MAE.

3. Secondary model construction

$$\text{Synergy} \sim \text{Methotrexate dosage} + \text{Vincristine dosage} + \text{Gene1 expression} + \cdots + \text{Gene541 expression}$$

Oncogene expression information was added to explanatory variables in the second model by adding. For each of thirteen cell lines, there were 541 genes selected, therefore 541 expression values existed for each cell line. Due to the different structure between dosage and expression information, it was challenging to build the uniform matrix for the model establishment and prediction. We transformed the gene expression data from 541 rows into 541 columns with each representing expression information for a single gene. Next, we duplicated the value into 80 (8 times 10, corresponding to the level of synergy.) identical rows and then combined them with the dosage and synergy columns.

As described above, we built matrix for one of thirteen cell lines, MOLT-4, and conduct the synergy prediction. Then, we combined data from four cell lines: MOLT-4, CCRF-CEM, KONP-8 and NALM-6, into one large matrix and built the model using linear/ridge/lasso/elastic net regression and random forest algorithms again, and then predicted the drug synergy among different cell lines (patients). Once again, LOOCV and 10-fold CV were executed and R-square/RMSE/MAE values were recorded. By comparing the R-square as well as RMSE and MAE outcome from single cell line model to multiple cell line model, we can determine if the model was capable of train and predict synergy outcome on separate patients.

4. Final model and model comparison

$$\text{Synergy}\sim\text{Methotrexate dosage} + \text{Vincristine dosage} + \text{Gene1 expression} * \text{Gene1 mutation} + \cdots + \text{Gene541 expression} * \text{Gene541 mutation}$$

Lastly, we further refined our model by adding gene mutation information into our model. Mutation information was of great importance as it interact with gene expression and influence the drug efficacy among patients. Two challenges existed when adding mutation variable into our model, the binary nature of mutation for each gene and the same mutation across all cell lines. This made mutation information useless when predicting drug synergy among cell lines (patients). As result, we chose to limit application of mutation variable to single cell line level. In addition, most of genes we selected are also mutated, which is likely related to the overlapping definition of oncogene and mutated genes. In other words, genes in this study are selected through cancer gene census program, which use mutation status as one of criteria to determine oncogenes. We delivered two solutions to address this issue: First was to add mutation information to the single cell line matrix directly; second was to create an interaction term that calculated by mutation times expression, accounting for their linked function in testing drug efficacy. In other words, the second solution considered mutation as a criterion to further refine the gene we selected into our model, and to reduce noise.

Likewise, linear regression, ridge/lasso/elastic net regression and random forest algorithm were conducted in the new models separately, and LOOCV and 10-fold CV were operated to calculate R-square, RMSE, MAE. Once all models were evaluated, their performance were analyzed and compared to each other. We chose models with higher R-squares over those with lower r-squares. Among models that generates similar R-squares, we chose the one with lower RMSE and MAE levels.

Results

Table 1 – Evaluation of Numerical and Categorical Format of Dosage Variable

Format	Validation	R-squared	RMSE	MAE
Numerical	10-fold	0.385371	0.01830717	0.01024578
	LOOCV	0.07421187	0.01965982	0.01035824
Categorical	10-fold	0.4127761	0.01909684	0.01081816
	LOOCV	0.009370495	0.02012892	0.01015256

We first built linear regression on baseline model, where synergy index is explained by dosage information of two drugs. Numerical (log mole) and categorical format of dosage are tested and compared for choosing the better one to be used in later steps. (Table 1.) R-square, RMSE and MAE calculated by both 10-fold CV and LOOCV for above models is summarized in table. Although both numerical and categorical variable provides similarly good RMSE and MAE (less than 0.1), categorical dosage generates better R-Square and thus provide more accurate synergy prediction. We believe the reason for this is that log mole dosage for both drugs cluster around zero, which introduce multicollinearity and negatively influence the accuracy of our model. To best avoid multicollinearity, we only employ dosage in its categorical form in later steps. Noticeably, R-square calculated through 10-fold cv are significantly larger than it from LOOCV.

Table 2. – Baseline Model – Performance Evaluation

$$\text{Synergy} \sim \text{Methotrexate dosage} + \text{Vincristine dosage}$$

Regression	Validation	R-squared	RMSE	MAE
Linear	10-fold	0.4127761	0.01909684	0.01081816
	LOOCV	0.009370495	0.02012892	0.01015256
Ridge	10-fold	0.19862460	0.01945838	0.01336863
	LOOCV	NA	0.01931	0.01094
Lasso	10-fold	0.3701960	0.01837342	0.01186131
	LOOCV	NA	0.01114	0.01114
Elastic Net	10-fold	0.3687883	0.01817660	0.01167680
	LOOCV	NA	0.01975	0.01115
Random Forest	10-fold	0.4229945	0.01918930	0.01082013
	LOOCV	0.014526237	0.01960475	0.01051307

(Table 2.) Performance evaluation for linear regression, ridge regression, lasso regression and random forest upon baseline model are summarized in table 2. First thing to notice is that LOOCV validation method generated NAs for Ridge, Lasso and Elastic net regression. In the penalized regressions, the formula of calculating R-squared is $R^2 = 1 - \frac{E(\hat{y}-y)}{V(y)}$, but for LOOCV, there is only one subject in each fold. In this case, the variance for each fold is zero. Thus, R-squared reaches infinity and have no meaning with variance equal to zero. Even for linear regression and random forest, where LOOCV did generate valid result, R-square calculated are smaller compared to 10-fold cross validation. [insert] RMSE and MAE for all algorithms are relatively small (less than 0.03), indicating small variance among predicted and actually synergy. However, this is partially due to the small absolute values of our synergy indexes, which produced small difference in prediction. For the baseline model, best R-square is 0. 4229945, calculated through random forest algorithm. Linear regression also carried out good result as baseline model is relatively simple in formation, plus we addressed linear dependence issue by adopting categorical dosage variables. Penalized regressions are designed to overcome potential multicollinearity problem or more explanatory variables than its sample size, thus failed to show advantage in baseline model. Although the highest R-square is out-standing, dose information along is inadequate for predicting the inter-drugs synergy.

Table 3. – Secondary Model – Performance Evaluation

$$\text{Synergy} \sim \text{Methotrexate dosage} + \text{Vincristine dosage} + \text{Gene1 expression} + \cdots + \text{Gene541 expression}$$

Regression	Validation	R-squared	RMSE	MAE
Linear	10-fold	0.5109134	0.03107704	0.01950833
	LOOCV	0.3433941	0.03530442	0.01964376
Ridge	10-fold	0.50000159	0.03183830	0.02056731
	LOOCV	NA	0.02059	0.02059
Lasso	10-fold	0.6779864	0.02864726	0.01753968

	LOOCV	NA	0.01447	0.01447
Elastic Net	10-fold	0.6774070	0.02852505	0.01735685
	LOOCV	NA	0.01453	0.01453
Random Forest	10-fold	0.6038140	0.02755692	0.01339132
	LOOCV	0.8125041	0.04291264	0.02236781

(Table. 3) Table 3 summarizes the performance evaluation for second model, where we added expression data as additional explanatory data. Like table.2, LOOCV generated NA for penalized regressions. In the secondary model, RMSE and MAE calculated from each algorithm are larger than those in Table 2, yet still less than 0.05. Interestingly, both the largest and smallest error of secondary model come from random forest algorithm, by means of LOOCV and 10-fold CV correspondingly. After including additional information, R-square of linear regression decreased while that of other four regressions increased. This is likely related to the introduction of multicollinearity. Compared to penalized regressions and random forest algorithm, linear regression does not address linear dependence among variables and thus its result is negatively influenced. On the other contrary, penalized regressions' performance improved greatly. Penalized regression is best for model with large number of variables and it adjusts for multicollinearity by adding penalty terms. Among ridge, lasso and elastic net regression, lasso regression is good at picking up small signal from high noise, thus it gives largest R-square and smallest error. As for elastic net regression, we tuned level of alpha with a sequence from 0 to 1 by 0.1. Largest R-square is obtained when alpha is equal to 1, which again proved that lasso regression has better goodness of fit, and this also explains R-square and errors are roughly the same for lasso and elastic net regression. Maximum R-square reaches 0.8125041 at random forest algorithm, tested by LOOCV. Unlike it in baseline model, LOOCV carries out test result with better accuracy in more complicated model. Another thing to be noticed here is that in random forest algorithm, RMSE/MAE disagree with R-square, meaning that LOOCV generates better R-square but higher error rates, which indicates the trade-off between goodness-of-fit and prediction accuracy among LOOCV and 10-fold CV. In general, we conclude that gene expression data improved model performance in terms of prediction accuracy.

Table 4 – Secondary model –Multiple Cell Lines

Regression	Validation	R-squared	RMSE	MAE
Linear	10-fold	0.08942684	0.05052688	0.0264831
	LOOCV	0.01737584	0.05440216	0.02534709
Ridge	10-fold	0.06536312	0.04857385	0.02116718
	LOOCV	NA	0.02113	0.02113
Lasso	10-fold	0.10299321	0.04840805	0.02049762
	LOOCV	NA	0.02042	0.02042
Elastic Net	10-fold	0.11129450	0.04859496	0.02078302
	LOOCV	NA	0.02039	0.02039
Random Forest	10-fold	0.13825679	0.05340509	0.02204032
	LOOCV	0.049178822	0.05254014	0.02245384

So far, models were built upon data collect from single cell line. Results are promising, which proves the feasibility of intra-cell-line prediction of drug synergy. To investigate models’ ability of utilizing local data record to predict new patients expected two-drug synergy, we built secondary model using data collected from four cell lines and evaluated the prediction summarized on table (Table.4). The strength of inter-cell-line synergy prediction is assessed by comparing result from table 3 and 4. According to table 4, result is apparently not as good as it in table 3. Largest R-square is 0.13825679, implying that only 14% variance of response variable (synergy) is explained by the model. RMSE and MAE are small in general, but larger than it in table 3. When model composed of data form four cell lines, LOOCV carries out smaller R-square than 10-fold CV since data set is larger, which introduces more variances. In conclusion, inter-cell-line synergy prediction is incompetent at this stage.

Table 5 – Final Model – Performance Evaluation

$$\begin{aligned}\text{Synergy} &\sim \text{Methotrexate dosage} + \text{Vincristine dosage} + \text{Gene1 expression} \\ &* \text{Gene1 mutation} + \cdots + \text{Gene541 expression} * \text{Gene541 mutation}\end{aligned}$$

Regression	Validation	R-squared	RMSE	MAE
Linear	10-fold	0.5109134	0.03107704	0.01950833
	LOOCV	0.3433941	0.03530442	0.01964376
Ridge	10-fold	0.50000159	0.03183830	0.02056731
	LOOCV	NA	0.02059	0.02059
Lasso	10-fold	0.6779864	0.02864726	0.01753968
	LOOCV	NA	0.01447	0.01447
Elastic Net	10-fold	0.6774070	0.02852505	0.01735685

	LOOCV	NA	0.01453	0.01453
Random Forest	10-fold	0.6029322	0.02752697	0.01337908
	LOOCV	0.8110620	0.04292179	0.02237318

To add gene mutation data into model, we offered two solutions: first is to directly combine binary indicators for mutation status to secondary model, and resulting model would be as below:

$$\begin{aligned} \text{Synergy} \sim & \text{Methotrexate dosage} + \text{Vincristine dosage} + \text{Gene1 expression} + \cdots \\ & + \text{Gene541 expression} + \text{Gene1 mutation} + \cdots + \text{Gene541 mutation} \end{aligned}$$

However, this model is incapable of addressing the inner link between genes' expression and mutation status. In other words, individual gene's expression and mutation are considered two separate explanatory variables in this model. Therefore, we preformed second solution, which is to create interaction terms for gene expression and mutation. By doing this, we further refined the genes we selected by times zero to expressions of genes that are not mutated. As result, our final model is illustrated below:

$$\begin{aligned} \text{Synergy} \sim & \text{Methotrexate dosage} + \text{Vincristine dosage} + \text{Gene1 expression} \\ & * \text{Gene1 mutation} + \cdots + \text{Gene541 expression} * \text{Gene541 mutation} \end{aligned}$$

And its result is summarized in (Table. 5). Prediction performance for above model is identical to the secondary model (Table.3) expect for random forest algorithm. Percentage of mutation status in the genes we selected through cancer gene census program was unproportionally distributed (534 mutated and 7 not mutated), which partially explains the resemblance of table 3 and 5. More importantly, the 7 genes we eliminated are originally linearly dependent upon other variables and are ignored by penalized regressions and cross validation featured by "caret" package. Maximum R-square is 0.8110620, which is calculated through random forest algorithm tested by LOOCV. Again, this value makes little difference to table 3. Due to the limited number of unmutated genes, mutational information did not provide much improvement to the model.

Discussion

In this study, we employed two-drug dosage information, gene expression and mutation information to construct models via linear regression, ridge, lasso and elastic net regression and random forest algorithm for predicting two-drug efficacy synergy. Multiple dimension of comparison and evaluation were conveyed on explanatory variables, algorithms and means of cross validations. Our results are primary evaluated by R-squares, as it measures goodness of fit of models. Error terms such as RMSE, MAE were calculated as well and served as references.

By comparing the result from baseline model, secondary model and final model, we conclude that gene expression data improves model performance greatly, indicating strong relationship between gene expression and two-drug synergy. However, adding mutation information barely changed model performance. The possible reason for this includes uneven percentage of mutated and unmutated genes. More specifically, genes we included in this study are selected by crossing selection between GDSC whole gene expression dataset and cancer gene census. According to the cancer gene census program, most if not all somatic cells involve cancer via mutation. This is to say that interaction between mutation and gene expression already exist in our model, which introduces potential selection bias. Another limitation of mutation information applied in this study is that the binary indicator does not provide information about the nature of specific mutation. For example, larger scale mutation such as frameshift mutation tremendously change gene function and expression level, while other mutation like single nucleotide polymorphism (SNP) may cause little to none impact. In addition, changing binary indicator to counts of mutation in each gene is an alternative solution. Mutation counts quantifies scale of mutation, and intuitively influence gene expression and drug efficacy. Due to limited time and resources, we did not have access to such data, but future studies should consider above characteristic of mutation to gain more precise understanding of relationship between mutation and drug efficacy to improve model accuracy.

Within each model, we evaluate performance of regressions and machine learning algorithm. As illustrated in table 2, 3 and 5, random forest algorithm consistently produced largest R-squares. As we discussed in method part, random forest algorithm is good at finding non-linear relationships, and is specifically designed for large dataset with large number of numerical variables. Among penalized regressions, lasso regression utilized its L1 penalty terms to extract important variables out of huge noise, thus provided better result than ridge and elastic net regression. Linear regression provided the second largest R-square in baseline model, as we manually eliminated multicollinearity by choosing categorical form of dosage variables. All result was carried out by 10-fold cross validation and LOOCV. 10-fold CV calculated larger R-square in linear regression, as LOOCV does not function properly in linear regression. LOOCV is excelling in secondary and final model, especially for random forest algorithm. However, this is a trade-off between accuracy and precision (R-square and RMSE/MAE).

Lastly, we compared model performance on models composed of single cell line data versus models built upon four cell lines (Table.3 and Table.4). Largest R-square of single cell line model is 0.8125041, provided by random forest algorithm through LOOCV, and through 10-fold CV tested on random forest, largest R-square of four-cell-line model is 0.13825679, which is significantly smaller. R-square calculated through other regressions provide similar results. We also observed that error terms (RMSE/MAE) for single cell line is relatively small compare to four cell line models. In conclusion, models we built is competent in predicting two-drug synergy within individual cell line, but our intension to predict inter-cell-line polymer drug efficacy turned out below expectation. Noticeably, we did not include mutation information in multiple cell line model because the binary indicator for mutation status is uniform among all cell lines. With access to cell line individualized mutational information, we expect improved results. Besides, although we addressed multicollinearity by performing penalized regression and non-linear regression such as random forest, conducting variables selection is another possibility to improve model performance. For the sake of computational efficiency, step-wise variable selection is not recommended. Instead, Principal components analysis (PCA) should be considered in future studies as it could reduce number of variables and possibly produce more accurate models.

There are a few more limitations to this study, such as the overlapping criteria of mutational gene and oncogene introduces selection bias. If had enough computational power, adding whole genome expression and mutation ought to further improve our model performance. Another than three key variables we interested in this study, more elements such as admission time, medical record, personal lifestyle and socioeconomical status could all be influential to the effect of drugs and influence model performance specifically for predicting inter-patient drug efficacy.

Acknowledgment

The author thanks to The Laboratory for Molecular Immunoengineering for providing methrotrexate and vincristine synergy data.

Bibliography

[1] M. J. et al. (2012). Cancer of the ampulla of Vater: analysis of the whole genome sequence exposes a potential therapeutic vulnerability. *Genome Med. 2012; 4(7): 56.*

[2] Mishra A, Verma M (2010). Cancer biomarkers: are we ready for the prime time?. *Cancers (Basel). 2010 Mar 22;2(1):190-208. doi: 10.3390/cancers2010190.*

[3] Jiang Y, Wang M (2010). Personalized medicine in oncology: tailoring the right drug to the right patient. *Biomark Med. 2010 Aug;4(4):523-33. doi: 10.2217/bmm.10.66.*

[4] Geeleher P, Cox NJ, Huang RS (2014). Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biol. 2014 Mar 3;15(3):R47. doi: 10.1186/gb-2014-15-3-r47.*

[5] H.M. Bøvelstad et al. (2007). Predicting survival from microarray data—a comparative study. *Bioinformatics, Volume 23, Issue 16, 15 August 2007, Pages 2080–2087, https://doi.org/10.1093/bioinformatics/btm305.*

[6] Shi L et al. (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol. 2006 Sep;24(9):1151-61.*

[7] Yang W et al. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res. 2013 Jan;41(Database issue):D955-61. doi: 10.1093/nar/gks1111. Epub 2012 Nov 23.*

[8] Catalogue Of Somatic Mutations In Cancer. *https://cancer.sanger.ac.uk/cosmic*.