Learning from Data: Classification of Adverse Event by Biomarkers

Bin Zhuo

February 03, 2018

1 Introduction

Two hundred (200) subjects were enrolled in a clinical trial and took a pill daily for a total of eight weeks. Twenty biomarker were measured for each participant, both at baseline (Week 0) and on a weekly basis postdose. Some subjects experienced adverse event (AE) during the course of the study, and dosing was immediately terminated for those who had AE although collection of biomarker data continued. Data collected up to the time point of AE or till the end of study are used to the questions of interest as follows:

(1) Find biomarkers that are changing as a result of treatment

(2) Predict AE based on historical biomarker data

(3) Identify biomarkers that are related to AE

2 Method

I applied non parametric tests to the cleaned data (see Section 2.1) to address (1), and machine learning tools to the dimensionally reduced data (see Section 2.3) to explore (2) and (3).

The following packages were used for this project. Some of the code is suppressed in this report, and available in the source file.

```
## Load the packages
library(dplyr)   # package for data manipulation
library(tidyr)   # package for data manipulation
library(ggplot2) # package for visualization
library(caret )  # package for classification
library(xtable)  # pakcage for tabulation
```

2.1 Data Cleaning and Manipulation

Throughout this report, `AE` is coded as 1 if a subject experienced AE during the study, and 0 otherwise. Then change from baseline (`CHG`), fold change (`FC`) as well as log FC (`logFC`) are calculated for each biomarker at each time point, using the following transformation

$$CHG_i = M_i - M_0, \quad FC_i = \frac{M_i + 0.1}{M_0 + 0.1}, \quad logFC_i = log(FC_i)$$

where M_i is the value of marker M at time i, M_0 the corresponding baseline value at Week 0.

This is the first few lines of the cleaned data.

```
##   SUBJECT EVENT.TIME MARKER.NAME BL WEEK MARKER.VALUE       FC CHG
## 1       1          4         M01  6    1            8 1.327869   2
## 2       1          4         M02  1    1            2 1.909091   1
## 3       1          4         M03 24    1           30 1.248963   6
## 4       1          4         M04  9    1            9 1.000000   0
## 5       1          4         M05  8    1            9 1.123457   1
## 6       1          4         M06  9    1           11 1.219780   2
##       logFC AE
## 1 0.2835753  1
## 2 0.6466272  1
## 3 0.2223133  1
## 4 0.0000000  1
## 5 0.1164104  1
## 6 0.1986707  1
```

2.2 Explanatory Analysis

The first step of explanatory analysis is to visualize the predictors by outcome (whether subject experienced AE). As an attempt, I tried box plot (for individual observations) and line plot (for averaged biomarker by time and outcome), shown in Fig 1 and Fig 2. Fig 1 visually presents FC of all the biomarker data, with a few potential outliers for most of the biomarkers, suggesting a log transformation may be beneficial. Since some of the outliers were much larger in magnitude, log-transformation does not appear to alleviate the concern. In the appendix, I created figures similar to Fig 1 and Fig 2, but using log transformed FC (see Fig X1 and Fig X2). Therefore, the outliers were included in the analysis and no diagnosis was done to them.

Fig 2 gives a more clear picture of how each biomarker is changing over time, and how they differ between the two outcomes (i.e., 1 for AE and 0 otherwise). I have the following observations:

- There is strong positive correlation between time and average FC of biomarkers 'M1-M3'.
- Biomarkers 'M3', 'M7', 'M8', 'M13', 'M16', 'M17' and 'M20' appear to be well separated by the dichotomized outcomes, suggestive of their relation to AE.

Fig 3 shows a heatmap of the pairwise correlation between the biomarkers, suggesting no apparent correlation exist among the fold change of most biomarkers. The maximum correlation is between `M1` and `M2`, with value 0.2788.

Fig 1: Biomarkers over time by outcome

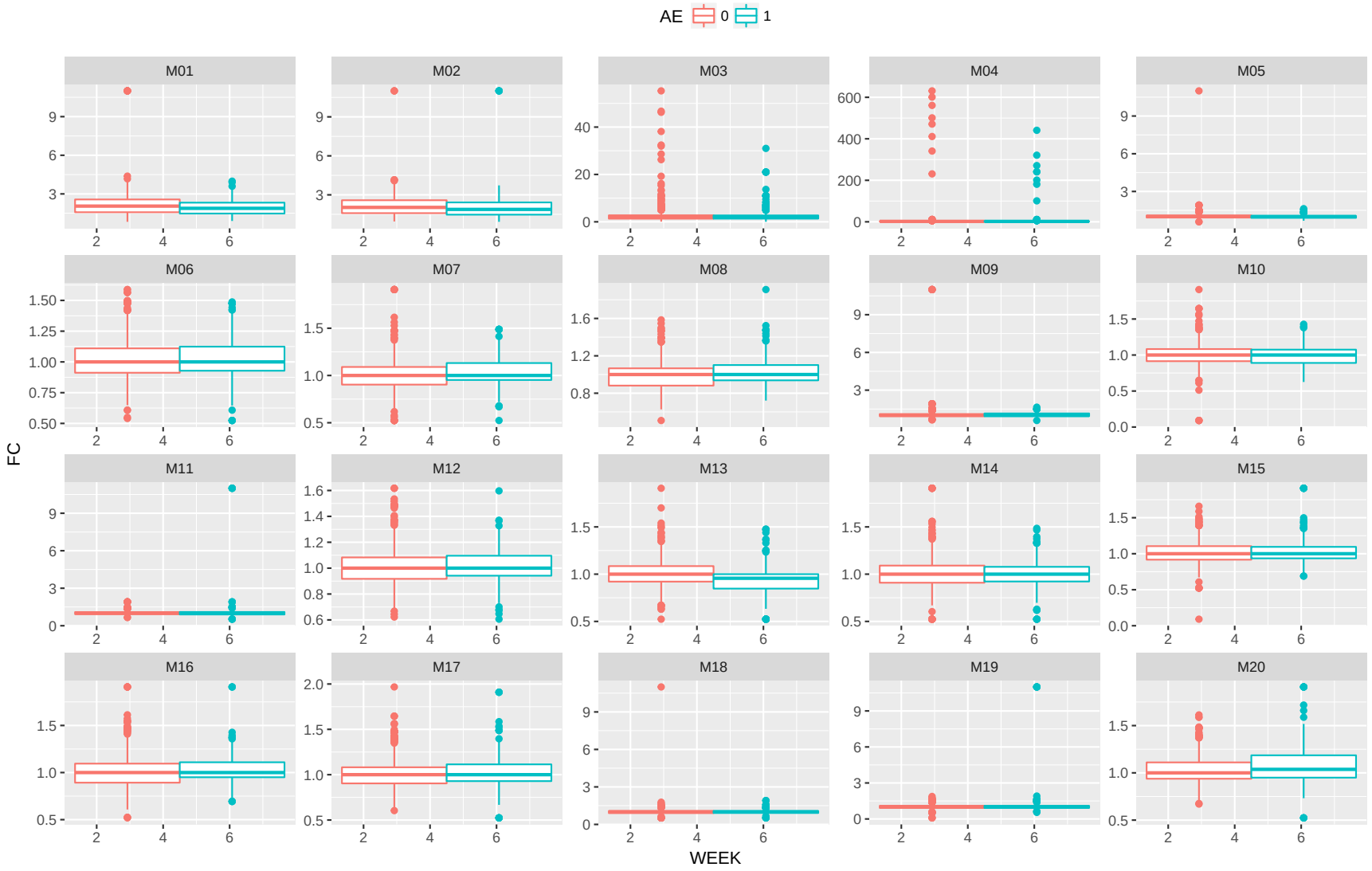


Fig 2: Mean of Biomarker over time by outcome

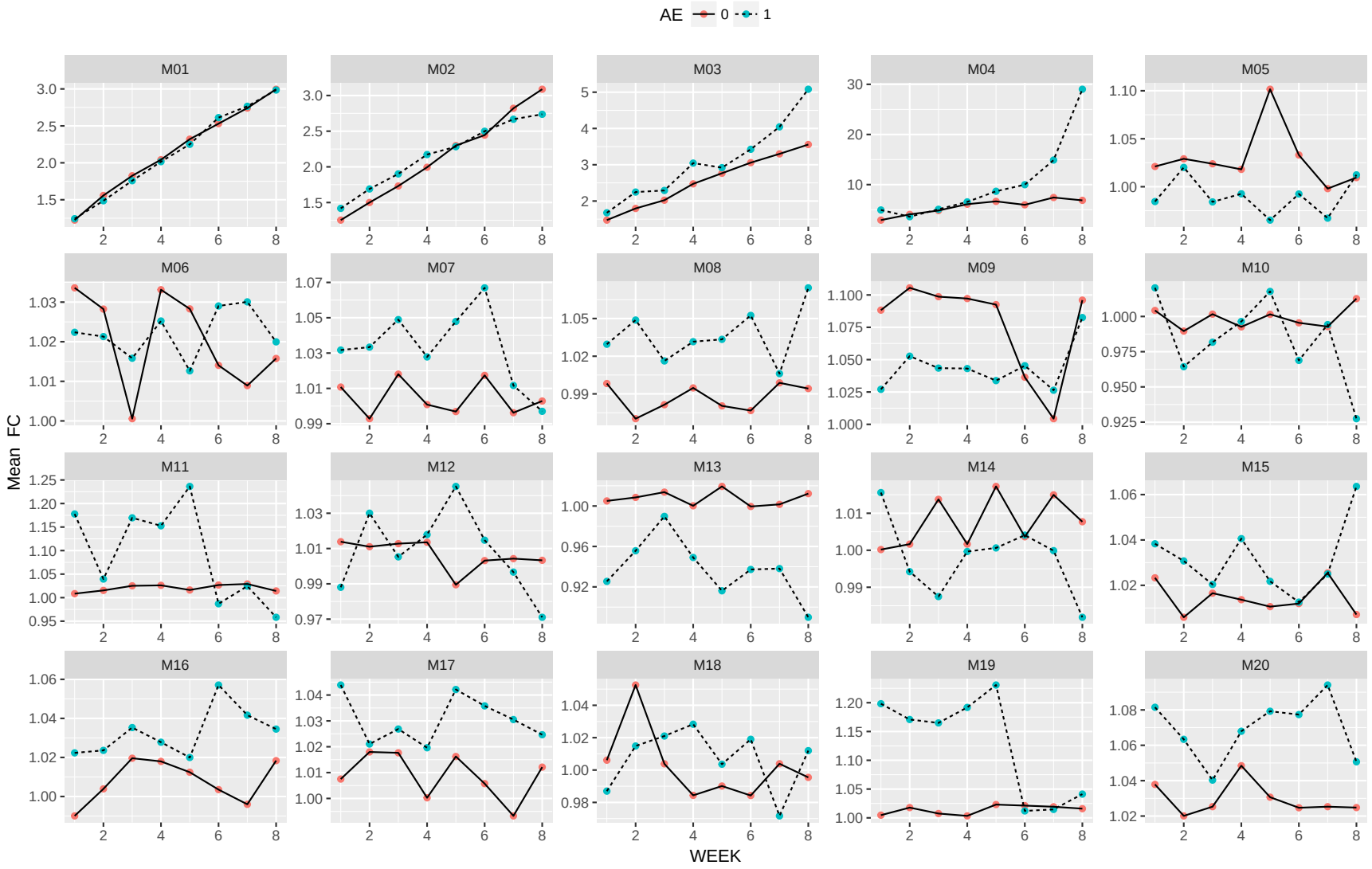
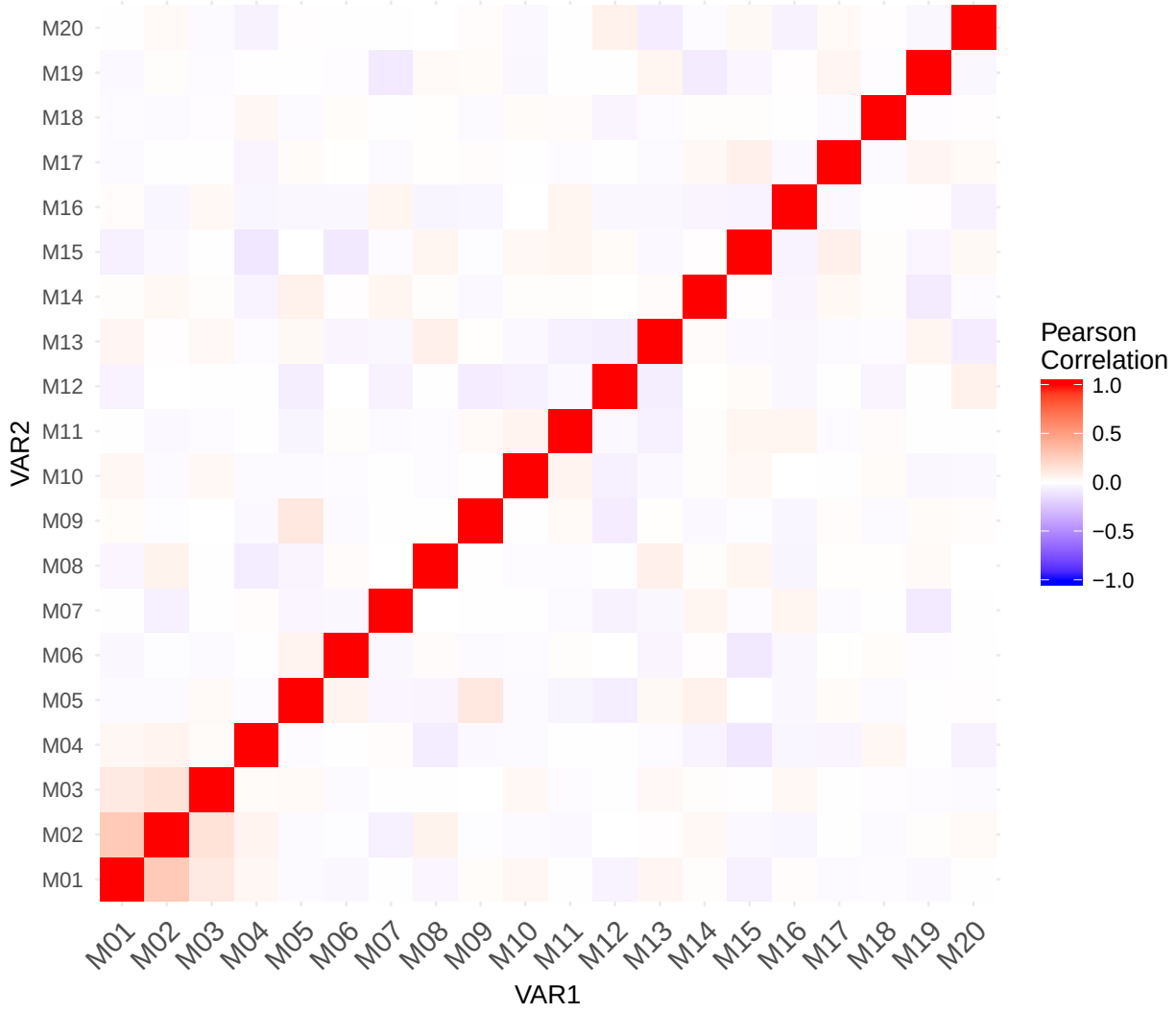


Fig 3: Pairwise Correlation of Biomarkers (FC)



2.3 Dimension Reduction

This data has feature that the biomarkers were collected at multiple time points, but the response variable was obtained only at the end of the study or when the subject was dropped from dosing. However, many of the machine learning methods are developed for cross-sectional rather than longitudinal data analysis. Therefore I was motivated to use dimension reduction techniques to map the longitudinal data into one dimensional space. Specifically, I applied principle component analysis (PCA) to each of the biomarker, and used the score of the first component as the new predictor for that biomarker.

Let X_{ij} be a measurement taken at time j for a given biomarker of subject i, and $\mathbb{X}_{\mathbb{N}\times\mathbb{T}} = (X_{ij})$. Therefore, by first component of PCA, I'm trying to find a vector α of length T, such that

$$\alpha = \arg\max \frac{(\alpha\mathbb{X})'(\alpha\mathbb{X})}{\alpha'\alpha}$$

Essentially, PCA is to find a linear combination of the longitudinal measurements that retains max variation in the data. Since some subjects have censored data when they were dropped before the study ended, the PCA procedure is done by grouping subjects based on the number of measurement they had. The following table gives a summary of subjects falling into each category.

NO AE	WEEK 4	WEEK 5	WEEK 6	WEEK 7	WEEK 8
140	15	10	8	10	17

Table 1: Number of subjects experienced AE by time

I pulled subjects who experienced AE at Week 8 and those who did not have AE together for PCA analysis, as they have the same number of measurements for each biomarker. As a result, the data cleaned at the first step is collapsed into a data of 200×21, where each row represents a subject, and column 1 is the outcome, and each of the rest columns the score of PCA. This data is the input for the machine learning models, to be described next.

2.4 Classification

After the dimensionally reduced data is obtained, various machine learning tools are readily applicable. The data was split by 3:1 into `training` and `testing` set, where the former was used to train the best classifier and the latter to test the accuracy of the classifier trained. Ten-fold cross-validation was used in the training process.

```
# split the data into training and testing set
inTraining <- createDataPartition(final$AE, p = .75, list = FALSE)
training <- final[ inTraining,]
testing  <- final[-inTraining,]


fitControl <- trainControl(# 10-fold CV
                 method = "repeatedcv",
                 number = 10,
                 # repeated ten times
                 repeats = 10)
```

2.4.1 Candidate Classifiers

In this report, I chose the following methods to train the model:

- Boosted Logistic Regression (BLR)
- Random Forest (RF)
- Support Vector Machine with Linear Kernel (SVM-LK)
- Stochastic Gradient Boosting (GBM)

2.4.2 Parameter Tuning

The following table gives a description of how each classifier was tuned.

Table 2: Parameter Configuration and Tuning

Model	Tuning Parameter	Search Grid
BLR	`nIter`: # Boosting Iterations	`10 to 300 by 10`
RF	`mtry`: # Randomly Selected Predictors	`1 to 15 by 1`
SVM-LK	`cost`: Cost	`0.01, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10`
GBM	`interaction.depth`: Max Tree Depth	`1 to 5 by 1`
	`n.trees`: # Boosting Iterations	`50 to 500 by 50`
	`shrinkage`: Shrinkage	`0.001, 0.01, 0.1`
	`n.minobsinnode`: Min. Terminal Node Size	`10 to 15 by 1`

2.4.3 Model Performance Evaluation

The classifiers are evaluated by accuracy and Cohen's Kappa score (Cohen, 1968). Kappa can take value between -1 and 1, with 0 indicating no agreement between the rates, and larger positive values better agreement. Landis and Koch (Landis and Koch, 1977) discussed in detail about the usage of kappa. Kappa can usually improve the quality of the final model in problems where there are a low percentage of samples in one class. In this report, a kappa value of ≥ 0.4 (moderate to almost perfect agreement) is desired. Specificity and sensitivity were not used to evaluate the model performance during the training step, but were presented in the `testing` data.

2.5 Biomarker Identification

All of the four classifiers produce variable importance metric, which will be used to identify biomarkers of interest. The biomarkers that have higher importance metric will be considered as related to AE.

3 Results

In this part, I responded to each of the questions listed in Section 1, and summarized the main findings from the data.

3.1 Identifying Biomarkers Responding to Treatment

To find biomarkers that are changing as a result of treatment, I first defined what is "change". Because each subject was dosed daily for 8 weeks, "change" was interpreted in the following two ways:

C1: Absolute change: whether 'CHG' or 'logFC' is different from 0, or 'FC' is different from 1, since both are relative to baseline

C2: Change of trend: whether a given biomarker is changing over time because of repeated dose.

As has shown in Fig 1, there are quite a few outliers for each of the biomarkers. To avoid the consequence of outliers, I chose non-parametric approaches which does not depend on the absolute value of the biomarkers. Throughout this report, I chose `FC` as the proper transformation, but the result should not change in the context of non-parametric test. I created an binary variable `IND` with two possible values: `IND = 1` if `FC > 1` and `IND = -1` if `FC <= 1`. For C1, I used *Wilcoxon Signed Rank test* to evaluate whether `FC` is different from 1; For C2, I used *Chi-square test* to examine independence, that is, whether `FC` is changing over time. Table 3 shows the p-values of each biomarker for C1 and C2.

Not surprisingly, biomarkers `M1-M4` are significantly related to time, and all biomarkers have significant absolute change (which may not be of interest). Of course, Chi square test only concludes that these 4

biomarkers are not independent of time. Although we could further test linearity between `M1-M4` and time using linear models, it was not of primary interest and thus not done here.

MARKER.NAME	p.Chisq.test	p.Wilcox.test
M01	0.0000	0.0000
M02	0.0000	0.0000
M03	0.0000	0.0000
M04	0.0000	0.0000
M05	0.8663	0.0000
M06	0.3182	0.0000
M07	0.3895	0.0000
M08	0.9661	0.0000
M09	0.8601	0.0000
M10	0.8762	0.0000
M11	0.7623	0.0000
M12	0.6228	0.0000
M13	0.4823	0.0000
M14	0.9764	0.0000
M15	0.8838	0.0000
M16	0.2538	0.0000
M17	0.9699	0.0000
M18	0.6501	0.0000
M19	0.8869	0.0000
M20	0.6074	0.0000

Table 3: Non-parametric test of C1 and C2

3.2 Predicting the Occurrence of AE with Biomarker data

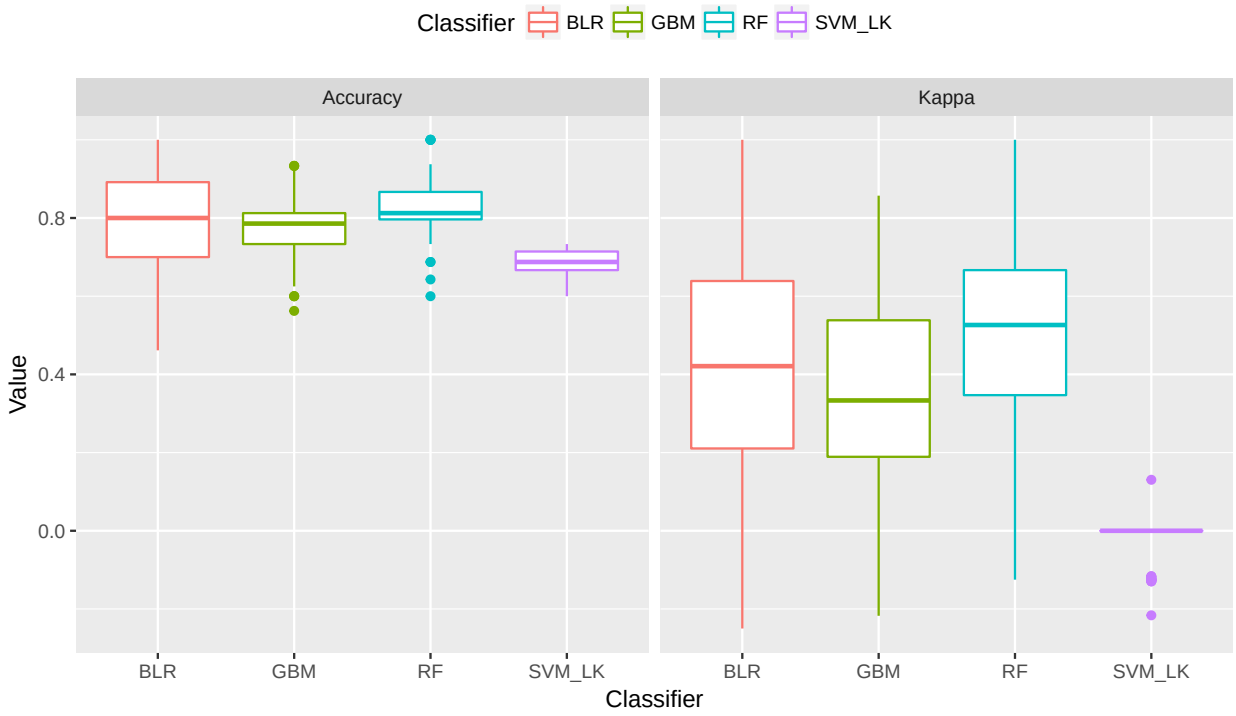
The best tuning parameters were identified from the best models for each classifier, with results listed in Table 4 below.

Table 4: Parameter Configuration and Tuning

Model	Tuning Parameter	Optimal Value
BLR	`nIter`	16
RF	`mtry`	7
SVM-LK	`cost`	0.01
GBM	`interaction.depth`	1
	`n.trees`	50
	`shrinkage`	0.1
	`n.minobsinnode`	10

Next, I examined the performance of all the classifiers with accuracy and kappa. The following boxplot (Fig 4) displays the descriptive statistics of accuracy and kappa. In general, the performance of BLR, GBM and RF are comparable, but SVM-LK is significantly lower. For those three methods, the mean accuracy achieved in the `training` set ranges from 0.7788 to 0.8284, and the mean kappa from 0.3681 to 0.5343. The best classifier in terms of both accuracy and kappa is RF.

Fig 4: Boxplot of accuracy and Kappa



How accurate can these classifiers be in the `testing` set? Prediction outcome for each classifier is obtained by using the corresponding model with optimal tuning parameter. **Table 5** shows that RF, GBM achieves the highest performance in the `testing` set, with an accuracy of 0.9.

Classifier	Accuracy	AccuracyNull	AccuracyPValue	Kappa
BLR	0.88	0.71	0.01	0.72
RF	0.90	0.70	0.00	0.76
SVM_LK	0.70	0.70	0.57	0.00
GBM	0.90	0.70	0.00	0.76

Table 5: Prediction Accuracy of Classifiers

Fig 5 shows the predicted versus actual outcome for each of the trained classifier. Note that for BLR, the total number of outcome could be less than the size of training set (not necessarily shown in this figure). Tracking down to the training stage, it may be because of the cross-validation procedure. Since only 30% subjects had AE, it was possible that at one re-sampled data, the outcomes of `AE` were all 0, which trigged the warning "`There were missing values in resampled performance measures`". At this time, I wasn't able to fix this problem in the code.

3.3 Identifying Biomarkers related to AE

The nice feature of these models implemented in R package `caret` is that they can return a metric of variable importance. **Table 6** below is the ranking of biomarkers by their relative importance in the corresponding model. Biomarkers "M03, M08, M09, M13" have higher average ranking in general, and thus are worth more investigation in practice. While other biomarkers are highly ranked in one method, they do not appear to be important in other methods. Ultimately, choosing the best biomarker needs biological experts' input.

Biomarker	BLR_rank	RF_rank	SVM_LK_rank	GBM_rank
M01	7	5	7	6
M02	11	7	11	10
M03	2	3	2	4
M04	19	1	19	1
M05	13	20	13	11
M06	12	18	12	12
M07	15	19	15	13
M08	3	10	3	5
M09	8	2	8	2
M10	4	11	4	14
M11	17	12	17	15
M12	10	13	10	16
M13	1	4	1	3
M14	20	17	20	17
M15	18	14	18	18
M16	5	8	5	19
M17	6	15	6	8
M18	14	16	14	7
M19	9	6	9	9
M20	16	9	16	20

Table 6: Rank of Variable Importance by Different Classifiers

4 Conclusion and Discussion

In this project, I applied dimension reduction and machine learning tools to study 20 biomarkers and their relation to AE, using data collected from a clinical study. I used nonparametric test to find biomarkers that are changing as a result of treatment. I found that all biomarkers have significant change from Week 0, and additionally, `M1-M4` show increasing trend with time. I leveraged PCA to reduce the longitudinal data into one dimension, after which the data is suitable as input for classification. I implemented four machine learning techniques (i.e., RF, BLR, SVM-LK, and GBM) to train the data, and achieved an accuracy of 0.8284 in the `training` set by RF, and that of 0.9 by RF, GBM in the `testing` set. I used the variable importance ranking generated from each classifier to identify the biomarkers that are potentially related to AE. I suggested biomarkers M03, M08, M09, M13 be investigated for further evidence of relation to AE.

While these classifiers gives somewhat satisfactory performance, there still may be room to improve. For example, SVM with other kernels (e.g., polynomial kernel, exponential string kernel) may render improved performance over SVM-LK. On the other hand, many other tools for classification are available and worth trying, for example, SIDES (Lipkovich, Dmitrienko et al., 2017).

Apart from dimension reduction technique used here, other attempts have been made to address classification problems in the context of longitudinal data. For example, one may opt to use data from a single time point, or to average the longitudinal data over time, but either way would suffer from substantial information loss. While keeping only the first component of PCA in this study also causes loss of information, the loss is minimized in the sense that first component retains the maximum variation in the data.

For transformation of data, I tried all of the three forms `FC`, `logFC`, `CHG`. As a matter of fact, `FC` gives the highest accuracy in both training and testing set. While `logFC` is slightly better than `CHG`, their accuracy drops about 10% as compared to `FC`. This evaluation can be easily done by switching `valuetype` between the three forms in the `pca_all()` function before PCA was done.

Interestingly, Chen and Bowman (Chen and DuBois Bowman, 2011) developed a support vector machine classifier for this unique type of problem. The basic idea is that the separating function (or hyperplane) of SVM is defined as

$$h(\mathbf{x}) = \mathbf{w} * (\mathbf{x}\beta^T) + b$$

instead of $h(\mathbf{x}) = \mathbf{w} * \mathbf{x}^T + b$ that is traditionally used in SVM. An iterative procedure is used to estimate the parameter vector β and α by quadratic programming. Due to time constrain, this method was not considered in this report, but is definitely worth trying in the future.

4.1 Appendix

Fig X1: Biomarkers over time by outcome

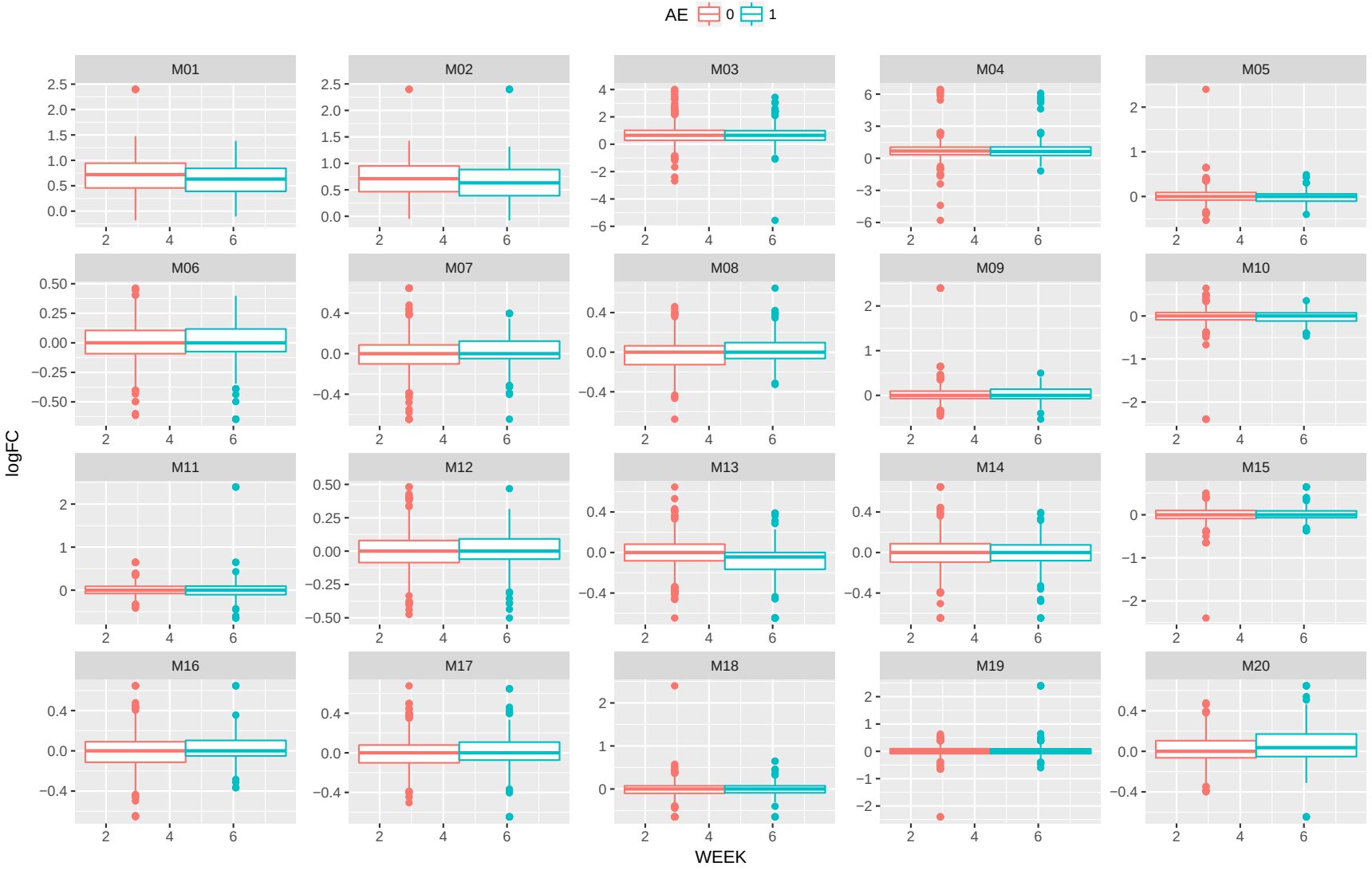
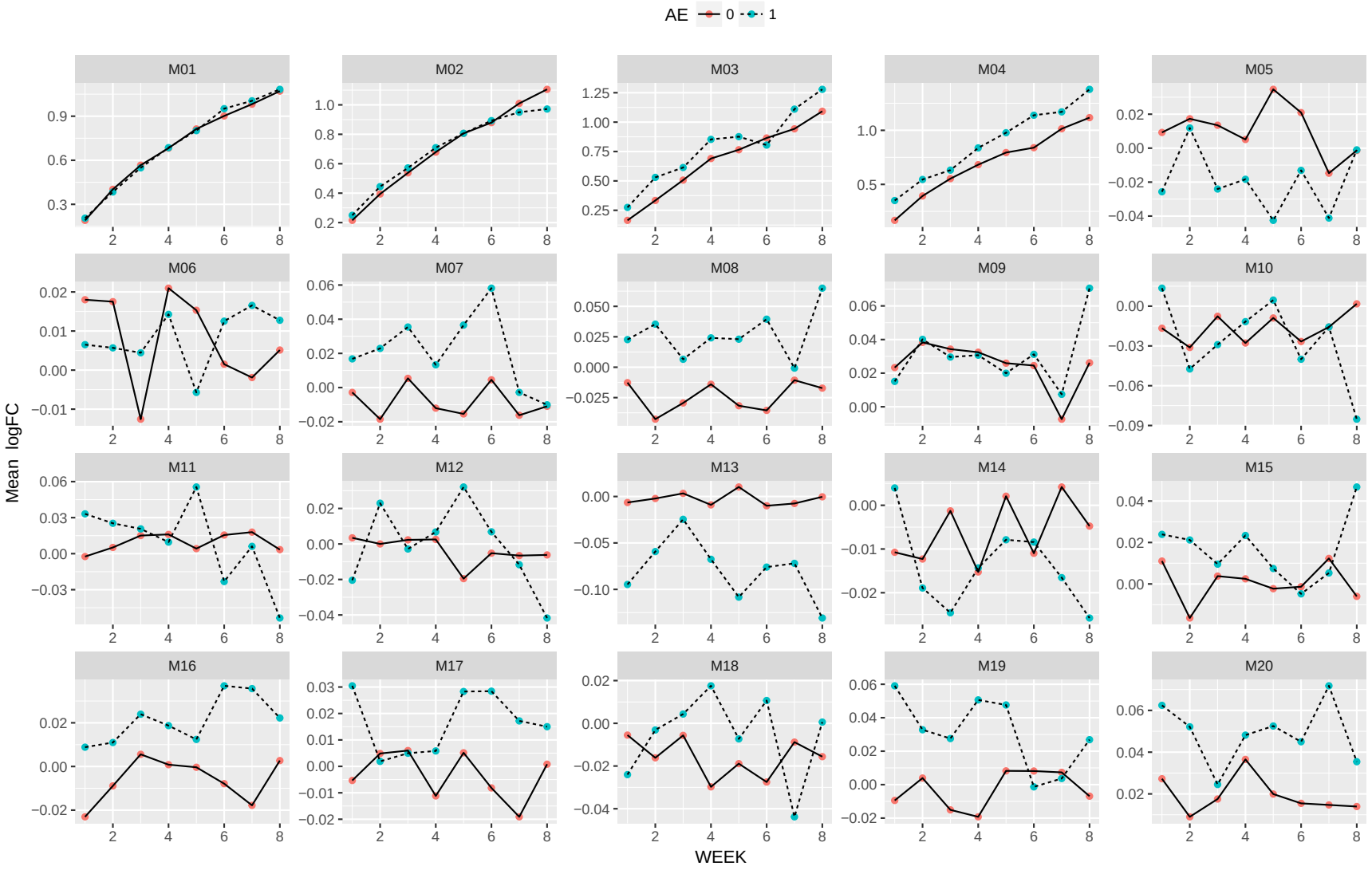


Fig X2: Mean of Biomarker over time by outcome



References

Chen, Shuo and F DuBois Bowman. 2011. "A novel support vector classifier for longitudinal high-dimensional data and its application to neuroimaging data." *Statistical Analysis and Data Mining: The ASA Data Science Journal* 4(6):604–611.

Cohen, Jacob. 1968. "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit." *Psychological bulletin* 70(4):213.

Landis, J Richard and Gary G Koch. 1977. "The measurement of observer agreement for categorical data." *biometrics* pp. 159–174.

Lipkovich, Ilya, Alex Dmitrienko et al. 2017. "Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials." *Statistics in Medicine* 36(1):136–196.