Personalized Medicine: Application To A Breast Cancer Study

by

Chenxin Yang

BS in Pharmacy, The Second Military Medical University, China, 2013

Submitted to the Graduate Faculty of

the Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Chenxin Yang

It was defended on

April 15, 2019

and approved by

Chaeryon Kang, PhD, Assistant Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Douglas Landsittel, PhD, Professor, Department of Biomedical Informatics
School of Medicine, University of Pittsburgh

Thesis Advisor: Gong Tang, PhD, Associate Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Personalized Medicine: Application To A Breast Cancer Study

Chenxin Yang, MS

University of Pittsburgh, 2019

Abstract

In randomized clinical trials, investigators compare the clinical outcomes among treatment arms and make claims on the effectiveness of experimental treatments versus the standard ones. Recent developments in biotechnology and associated biomarkers have led to advances in evaluating heterogeneous patient response and the relationship between treatment responses and certain biomarkers. Precision medicine, therefore, is becoming very popular in the healthcare industry. It is of great public health significance that proper implementation of precision medicine leads to informed and efficient decision making and patient management in clinical practice. Traditionally discovery of a predictive marker of treatment benefit is performed via a test of the interaction term between treatment and the marker of interest in a regression model that predicts the clinical outcome of interest. Recently a new paradigm has been proposed by redefining the search for predictive markers, as the search for an optimal individualized treatment rule (ITR) on treatment selection. Here we describe this new approach and apply those methods to a breast cancer study to identify clinical and genomic markers that are predictive of treatment benefit. The R package "*personalized*" was used in the implementation. Application of some of these methods does identify optimal ITRs that lead to improved outcomes based on the empirical estimates. However, validation via random splitting of training and testing datasets suggested that the findings may be resulted from over-fitting. These ITR-based methods provide a powerful tool for us to identify predictive markers for treatment response, but caution should be taken especially with high-dimensional marker data.

Table of Contents

List of Tables

List of Figures

Preface

I am heartily thankful to my advisor, Dr. Gong Tang, whose encouragement, guidance and support from the initial to the final level enabled me to develop an understanding of this topic. I also owe a deep sense of gratitude to my committee members, Dr. Douglas Landsittel and Dr. Chaeryon Kang, for their keen interest and generous help. This thesis would not be possible without their kind and selfless support.

I would also like to express my deep sense of thanks to the Department of Biostatistics of Pitt Public Health, where I spend very meaningful and fulfilled two years. I am genuinely thankful to every faculty and staff for their heart-warming help and co-operation throughout my study period.

1.0 Introduction

1.1 Individualized Treatment Rules

In randomized clinical trials, study participants are randomly assigned to a control arm and one or several experimental arms. The control arm represents standard regimen or practice and the experimental arms represent the new or alternative regimens that are suspected to be superior to the control arm in a certain outcome. When the superiority of the experimental regimen is demonstrated at the conclusion of a study, the experimental regimen will replace the control regimen and become the new standard practice. More studies will be developed to either confirm the finding or test whether this newly developed standard can be further improved. This ever-lasting procedure has been the blue print for drug development and policy improvement in past decades. In recent years it has been recognized that patients may respond to the same treatment differently, which is caused by the heterogeneity among different individuals. It has becoming an emerging issue especially with the availability of large amount of biomarkers and genomic markers from recent vast development in biotechnology. For example, breast cancer patients with hormone receptor-positive tumors benefit from hormonal treatments such as tamoxifen and anastrozole but those with hormone receptor-negative tumors do not.[1,2] Herceptin reduces the risk of recurrence by 50% in breast cancer patients with Her2-positive tumors but its benefit in patients with Her2-negative tumors is minimal if any.[3] Accounting for this heterogeneity represents a significant challenge which has motivated the trend toward personalized medicine over recent years.[4, 5] Models that use individual characteristics to predict the optimal treatment that achieves the best

response are critical for patients and treating physicians to make informed decisions on their treatment selection.

Assuming larger outcomes are preferable, optimal individualized treatment rules (ITRs) are the treatment assignment rules that assign the treatment which maximizes the outcome of the overall population based on their characteristics. The term ITRs is derived initially from dynamic treatment regimes, a concept proposed by Murphy (2003)[6] that focuses more on the change of treatment according to participants characteristics along time. A number of studies have been conducted on this topic. Qian and Murphy (2011)[7] used l_1-penalized least squares to estimate the optimal ITRs. Zhao et al. (2012) estimated optimal ITR using an outcome-weighted method with the hinge loss used. Chen et al. (2017)[8] provided a general framework for subgroup identification under different scenarios. Other studies focused on the methodology behind subgroup identification. For example, Zhou et al.[9] proposed a different method to address variable selection in finding optimal ITRs using residual weighted learning (RWL).

An R package, "*personalized*", has been recently developed to implement some of the aforementioned methods for estimating the optimal ITRs under various choices of loss functions and providing subgroup identification.[10]

1.2 A Sub-study of the NSABP Protocol B-41 Trial

The data considered in this thesis is a subset from the National Surgical Adjuvant Breast and Bowel Project (NSABP) B-41 clinical trial, a three-armed, randomized, phase III study.[12] 529 women participants with early stage operable HER2-positive breast cancer were enrolled from July 16, 2007 to June 30, 2011. All participants were over 18 years old and had ECOG performance

status of 0 or 1 at study entry. Participants were randomly assigned to receive one of three treatment regimens with 1:1:1 ratio. Each participant would at first receive four cycles of standard doxorubicin 60 mg/m² and cyclophosphamide 600 mg/m² intravenously on day 1 every 3 weeks followed by four cycles of weekly paclitaxel (80 mg/m²) intravenously on days 1, 8, and 15, every 4 weeks. Concurrently with weekly paclitaxel, participants would receive either trastuzumab weekly, lapatinib daily, or weekly trastuzumab plus lapatinib daily until surgery. All participants would receive trastuzumab after surgery until the completion of 52 weeks of HER2-targeted therapy. At the end of neoadjuvant treatments of 6 to 7 months, breast surgery (lumpectomy or mastectomy) was performed and participants' pathological response status was ascertained. Pathological complete response (pCR) is defined as the absence of any invasive component in the resected breast specimen and absence of cancer on H&E evaluation of all resected lymph nodes following completion of neoadjuvant therapy (ypT0/Tis ypN0) among participants with HER2-positive tumors and under neoadjuvant HER2-targeting regimens. The aim of NSABP B-41 trial is to compare the two lapatinib-containing treatment arms to the trastuzumab-only arm with pCR as the primary endpoint.

Trastuzumab is the first human epidermal growth factor receptor 2 (HER2) targeting monoclonal antibody that was approved by the United States Food and Drug Administration (FDA) as a first-line treatment, combined with paclitaxel, for HER2-positive breast cancer. for the treatment of HER2-overexpressing metastatic breast cancer (MBC).[13,14] Lapatinib is a small molecule inhibitor of epidermal growth factor receptor (EGFR) and HER2. It has been shown that trastuzumab can block the signal transaction and thus suppress the overexpressing of HER2.[15] A study has shown that lapatinib has synergy effects with trastuzumab.[16] Therefore, it is expected that trastuzumab in combination with lapatinib would have more promising outcomes.

Among 529 women enrolled in the trial, 519 participants had their pathological response determined. Breast pathological complete response was noted in 93 (52·5%, 95% CI 44·9–59·5) of 177 participants in the trastuzumab group, 91 (53·2%, 45·4–60·3) of 171 participants in the lapatinib group (p=0·9852); and 106 (62·0%, 54·3–68·8) of 171 participants in the combination group (p=0·095). Based on these results, the study concluded that substitution of lapatinib for trastuzumab did not improve pCR. Although combined HER2-targeted therapy led to higher pCR than the regimen with trastuzumab-alone in addition to chemotherapy, the difference was not statistically significant. Given that participants achieving pCR have much better prognosis in terms of long-term outcomes, discovery of useful prognostic and predictive clinical and genomic markers for pCR is imperative for developing rules that optimize treatment benefit for individual participants.

In a correlative study on B-41, the Nanostring PAM50 assay was performed on core biopsy samples from 271 study participants prior to neoadjuvant treatments: 94 of them received trastuzumab, 95 of them received lapatinib, and the remaining 82 subjects received a combination of trastuzumab and lapatinib. The PAM50 assay (using the PAM50-RUO CodeSet) simultaneously measures the expression levels of 72 target sequences, including eight endogenous invariant mRNA targets, six positive quality control targets, and eight negative quality control targets consisting of probes with no sequence homology to human RNA. At the end, expression levels from 50 cancer-related genes were recorded and used to determine the intrinsic subtype of the tumors among four categories: Luminal A, Luminal B, Her2-enriched and Basal.[17] To better understand whether any of the clinical or genomic markers may predict treatment benefit in pCR, we considered two sets of analyses: (1) trastuzumab-containing regimens (176 participants in total) versus lapatinib only (95 participants), and (2) trastuzumab combined with lapatinib (82

participants) versus trastuzumab alone (94 participants). For both comparisons, the recommended treatment is compared with the treatment received to determine the optimal treatment, using treatment effects. In the following context, we will denote one treatment arm as the "control arm" and the other treatment arm as the "treatment arm" for convenience.

In this thesis, we applied the R package "*personalized*" to identify an optimal individualized treatment rule and determine participant subgroups who would have achieved optimal pCR had they followed the estimated optimal treatment rule using data from these 271 NSABP B-41 trial participants. For each of the two comparisons, we define one treatment regimen as the treatment and another as the control for brevity. In the comparison between trastuzumab combined with lapatinib and trastuzumab, the former is the treatment; in the comparison between trastuzumab containing regimens and lapatinib, trastuzumab containing regimens are the treatment. The rule is defined based on a benefit score that is an estimated function of the markers: if the benefit score of a study participant is higher than the cut-point, she will be recommended the treatment; otherwise, the control will be recommended. Subgroup identification is accomplished using three different sets of potential markers: (1) clinical characteristics, including age, race, lymph node status, estrogen-receptor status, human epidermal growth factor receptor 2 (HER2) status and tumor size, (2) genomic markers, including breast cancer subtype and various genes, and (3) both clinical characteristics and genomic markers.

This thesis is organized as follows. In Section 2, we review and compare different methodologies regarding value function approximation, especially the selection of loss functions. In Section 3, we apply the methodology discussed in Section 2 on the NSABP B-41 trial data using the "*personalized*" package in R. In Section 4, problems and future studies are discussed. Related R code is given in the Appendix A.

2.0 Optimal Individualized Treatment Rules

Randomized clinical trials are designed to compare the efficacy of new treatment regimens to that of a standard regimen in terms of average effects across some participant population. When there is strong evidence to support the superiority of one of more regimens to the standard one, investigators would claim a positive trial and recommend treating similar participants with the newly found superior regimen(s) in future practice. This has been the model for drug development in many fields of medicine, especially cancer, in past decades. Investigators understand that participant response to treatments may not be homogeneous, that is, certain participants would benefit from replacing the standard treatment by the new treatment but other participants may not. Especially in the treatment of cancer participants, newer treatments such as new antibodies, second or third generations of chemotherapies and immunotherapies often lead to unexpected complications or deadly adverse events. It becomes imperative to identify markers that can be used to determine which participants would benefit from the new treatment and which participants would not. The determination of a predictive marker, a patient attribute that can be used to predict differential benefit from one treatment over another, is traditionally done via testing the interaction between treatment and the marker or participant characteristic under consideration.[18] In the past, study design for screening predictive markers of treatment benefit were hampered by two issues: (1) a marker study usually requires a quadrupled sample size comparing to a similar superiority clinical trial on efficacy; (2) there are very few useful clinical markers. Recent developments in microarray and RNA/DNA sequencing technologies provide an unique opportunity for investigators to explore predictive utility of the vast amount of genomic signatures. Demands for genomic predictive markers of treatment benefit has since stimulated statistical methodologies for

providing valid and effective discoveries other than the traditional approach by testing interactions in past few years. Recently Qian and Murphy (2011)[7] and Zhao et al. (2012)[19] provided a framework for defining ITR and proposed statistical methods to estimate optimal ITRs via optimizing a loss function. The methodology of R package "*personalized*" is mainly based on the methods built in this framework by further developments proposed by Xu et al. (2015)[11] and Chen et al. (2017)[8]. In the following context, we will use their notations and introduce the framework, recently proposed methods and application under the setting of two treatment arms and a binary outcome.

2.1 Individualized Treatment Rule (ITR)

Assuming that in a randomized clinical trial with two treatments, T, coded as -1 and 1, data on the binary outcome $Y \in \{0,1\}$ and potential markers $Z = \{Z_1, Z_2, \dots, Z_d\}$ are collected from n subjects. Let $\Lambda = \{-1,1\}$, an ITR is a map D from the space of Z to Λ: $T = D(Z) \in \Lambda$. Let P^D represent the distribution of (T, Y, Z) under a given ITR D, $E^D(Y)$ be the expected value of the outcome Y under P^D, a larger value of $E^D(Y)$ then indicates better outcomes, assuming larger outcomes are more desirable. $E^D(Y)$ can also be called the value function and be written as $v(D)$. An optimal ITR is a treatment assignment rule D that maximizes a value function should D be implemented. With observed data $\{Z, T, Y\}$, let $\pi = Pr(T = 1|Z)$, a popular value function is:

$$v(D) = E^D(Y) = \int Y \, dP^D = \int Y \frac{dP^D}{dP} dP = \int Y \frac{I(T=D(Z))}{pr[T|Z]} dP = E\left[\frac{I(T=D(Z))}{T\pi+\frac{1-T}{2}} Y\right], \quad (1)$$

where, $I()$ is an indicator function. In randomized clinical trials, π is a known constant. In observational studies, $\pi(Z)$ is usually estimated from a regression model of T on Z.

Subsequently, an optimal ITR, D^*, is a rule that maximizes this value function:

$$E^{D^*}(Y) = \max\{E^D(Y), all\ D\}$$

From

$$E(Y|T=1,\mathrm{D}) + E(Y|T=-1,\mathrm{D}) - v(D) = E\left[\frac{I\left(T \neq D(Z)\right)}{T\pi + \frac{1-T}{2}} Y\right], (2)$$

When Y is nonnegative, the optimal ITR D^* should minimize the right hand side of the above equation, which can be assessed as a weighted classification error for classifying T using Z. For a given data, the optimal ITR can be estimated by minimizing the empirical value:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{Y_i}{T_i\pi + \frac{1-T_i}{2}} I\left(T_i \neq D(Z_i)\right), (3)$$

with $D(\mathrm{Z})$ regarded as a classifier on the space of Z. This can be further written as

$$\frac{1}{n}\sum_{i=1}^{n} \frac{Y_i}{T_i\pi + \frac{1-T_i}{2}} I\left(T_i \neq sign\left(f(Z_i)\right)\right), (4)$$

where $f()$ is a function from the space of Z to the set of real numbers, R; $sign(t)$=1 if $t > 0$ and 0 if $t < 0$, and $D(\mathrm{Z})$ is defined as $sign\left(f(\mathrm{Z})\right)$. Zhao et al. (2012)[19] linked maximizing the value function to minimizing a weighted misclassification error with classifying each subject into -1 or 1 according to the sign of $f(Z)$. If an observation i is associated with a large weight at $\frac{Y_i}{T_i\pi + \frac{1-T_i}{2}}$, the optimal classifier $sign\left(f(\mathrm{Z})\right)$ will tend to assign this subject to the observed T_i. Observations associated with small weights will be assigned to the group that is opposite to T_i. Minimizing (4) is computationally challenging since it is a weighted sum of 0-1 loss functions which are neither continuous nor convex.

2.2 Selection of The Loss Function

A surrogate hinge loss function with a penalty term was adopted by Zhao et al. (2012)[19]:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{T_i\pi+\frac{1-T_i}{2}}\left(1-T_i f(Z_i)\right)^{+}+\lambda_n\|f\|^2, \quad (5)$$

where $\mathrm{t}^{+}=\max(t,0)$ and $\|f\|$ is some norm for function f. The penalty term $\lambda_n\|f\|^2$ was added for model regularization. One could use a linear combination of Z to represent f: $f(Z)=\beta_0+\beta_1 Z_1+\cdots+\beta_p Z_p$ or $f(Z)=\sum_{i=1,2,\ldots,n}\beta_i K(Z,Z_i)+\beta_0$, where $K(.,.)$ is a pre-determined kernel function. The estimated ITR will be $sign(f(Z;\hat{\beta}))$ with $\hat{\beta}$ obtained from minimizing (5). Zhao et al. (2012)[19] named this method as outcome weighted learning (OWL) and showed that the empirical risk function (4) under the estimated optimal ITR would converge to the risk function (2) under the optimal ITR under some regularity conditions.

Chen et al. (2017)[8] provided a general framework for subgroup identification via ITR by extending to other types of loss function. Assume that the counterfactual outcomes from a subject are $Y^{(1)}$ and $Y^{(-1)}$, had the subject taken treatment $T=1$ or $T=-1$, respectively. In practice, only one of the two counterfactual outcomes can be observed from a participant, $I(T=1)Y^{(1)}$ and $I(T=-1)Y^{(-1)}$ can be used to denote the participant's outcome result, where $I(\cdot)$ is the indicator function for each value of Y. The propensity score $\Pr(T=1|Z)=\pi(Z)$ is always known in randomized trials, and can be estimated in observational studies. Assuming $f(Z)$ is a function of baseline covariates that can be used to predict treatment assignment via $D(Z)=sign(f(\mathrm{Z}))$, then for a given loss function $M(y,v)$ that satisfy the following two assumptions: (1) $M_v(y,v)=$

$\frac{\partial M(y,v)}{\partial v}$ is increasing in v for any given y, (2) $U(y) \equiv M_v(y,0)$ is monotone in y, Chen et al. (2017)[8] considered minimizing the risk function $l_W(f) = E\big(l_W(f,Z)\big)$ where:

$$l_W(f,z) = E\left[\frac{M\{Y,Tf(Z)\}}{T\pi(Z)+(1-\mathrm{T})/2}|Z=z\right]$$

$$= E[M\{Y,f(Z)\}|T=1,Z=z] + E[M\{Y,-f(Z)\}|T=-1,Z=z].$$

Chen et al. (2017)[8] proposed to define the optimal ITR as $\{signf_{W0}(Z)\}$ that minimizes the above risk function, where $f_{W0}(Z) = argmin_f l_W(f)$.

For the OWL method by Zhao et al. (2012)[19], the loss function M(y,v)=y max{1-v,0}. A couple of other loss functions were suggested by Chen et al. (2017)[8], for example, M(y,v)=(y-v)^2 for continuous outcome y and M(y,v)=-[yv-log{1+exp(-v)}] for binary outcome y.

Using an approximation similar to the OWL method with $f_W = \sum_{k=1}^{K}\beta_k B_k(Z)$, where $B_k(Z), k = 1,\dots,K$ are some basis functions, one can estimate β by minimizing the empirical version of the loss function. Then use $\widehat{f_W} = \sum_{k=1}^{K}\widehat{\beta_k}B_k(Z)$ as the benefit score to define the optimal ITR as sign($\widehat{f_W}$). In practice when the dimension of Z is large, Chen et al. (2017)[8] recommended applying regularization on the variable selection via adding a penalty term.

In randomized clinical trials, where the treatment assignment is independent from Z, one could readily estimate the improvement conditional on treatment by following quantities:

$$E(Y|D^*(Z)=1,\mathrm{T}=1) - E(Y|D^*(Z)=1,\mathrm{T}=-1),$$

and

$$E(Y|D^*(Z)=-1,\mathrm{T}=-1) - E(Y|D^*(Z)=-1,\mathrm{T}=1),$$

where $D^*()$ is the estimated optimal ITR.

2.3 Validation

In practice, investigators will try various basis functions and assess the performance of the estimated optimal ITR by comparing the empirical risk function under various settings. As with any complex modeling of high dimensional data, over-fitting of the data may lead to overly enthusiastic results. As one approach to reduce the potential for overfitting, Huling and Yu (2018)[10] recommended an internal validation procedure by randomly splitting the observed data into a training set and a testing set. In each split, the optimal ITR developed from the training set is applied to the testing test and assess the improved performance in outcome had the estimated optimal ITR been followed. For example, for a randomized clinical trial, one may identify subgroups in the testing data $E_{j,k} = \{i: T_i = j, D(z_i) = k\}, j, k = -1,1$. Then assess the stability of results by comparing the empirical means of the response variable between $E_{-1,-1}$ and $E_{1,-1}$, and between $E_{-1,1}$ and $E_{1,1}$. If the markers under consideration are informative in predicting treatment benefit, one would observe consistent improvement in average response between patients who were assigned to the estimated optimal ITR and those who were not across these numerous random splittings. Failure in producing a consistent pattern of improvement may indicate that there is no strong evidence from the data to support the predictive utility of those markers. However, the lack of evidence may due to two possibilities: (1) the markers are not informative and the proposed algorithm for obtaining an optimal ITR is over-fitting the data, (2) the decrease in sample size in the random splitting practice leads to reduced power for detecting an optimal ITR based on those markers.

2.4 The "personalized" Package in R

The "*personalized*" package is built by Huling and Yu (2018) under the framework proposed by Chen et. Al (2017).[8] This package provides a quick and convenient way for subgroup identification. For a given data with covariate Z, outcome of interest Y and treatment $T = 1$ or $T = -1$, users are able to use the covariates Z to predict the optimal treatment between $T = 1$ and $T = -1$ that can maximize the expected outcome of interest under the given treatment assignment rule. A propensity score function, which is a function that uses subject covariates to predict the probability of treatment T = 1, needs to be specified before subgroup identification. Then, the main build-in function of this package, *fit.subgroup()*, can be used to identify treatment subgroups. Within this function, loss function for benefit score calculation and the cut-off point for treatment assignment can be selected. Many loss functions that satisfies the assumptions in Section 2.2 are available: square loss $M(y, v) = (y = v)^2$, the logistic loss $M(y, v) = y * log(1 + exp\{-v\})$, the hinge loss $M(y, v) = y * \max(0, 1 - v)$ etc. Loss function can be specified under the option *loss* within *fit.subgroup()* function. Under each loss function, the option *cutpoint* allows users to define the cut-off value for treatment recommendation. The cutpoint can be a constant or a quantile of benefit scores. The average outcomes within each subgroup by treatment they received are then reported in the output, together with the improvement of outcome following the subgroup separation and the range of benefit scores.

These results can be biased estimates because the implemented ITR is estimated from the same dataset with high-dimensional marker data, and the imbedded variable selection and model regularization would lead to overfitting. Validation is therefore necessary. Using the build-in function *validate.subgroup()*, unbiased results can be obtained by bootstrap bias correction or repeated training/testing splitting by choosing between *method = "boot"* or *method =*

"training_test_replication". For boostrap bias correction, a statistic is first estimated by the training data, then the bias with regard to that statistic is estimated by bootstrap samples extracted with replacement using whole data. Then a bias-corrected statistic can be obtained by these two values. For repeated training/testing splitting method, data is randomly partitioned into training and testing sets at a ratio that can be defined arbitrarily by users, the average outcome values for the training set, thus the predicted average outcomes given the same covariates under different treatment, are then estimated by the empirical average of outcomes in the testing set. This method allows us to obtain unbiased assessment on the implementation of the estimated optimal ITR in the study population. Many replications are performed for both methods.

Plots of average outcomes within each subgroup by treatment status are available after *fit.subgroup()* or *validate.subgroup()*. Available options are boxplot, density plot, conditional plot and interaction plot specified by "*type* = " option. Boxplots reflect the range and quantiles of average outcomes; density plots describe the distribution of the average outcomes; conditional plots show the relationship between benefit scores and the smoothed mean outcomes conditional on treatment received; interaction plots display the interaction of average outcomes for different subgroups and treatment status. Figure 1 to Figure 4 show an example of the four kind of graphs from a simulation study with a binary outcome.

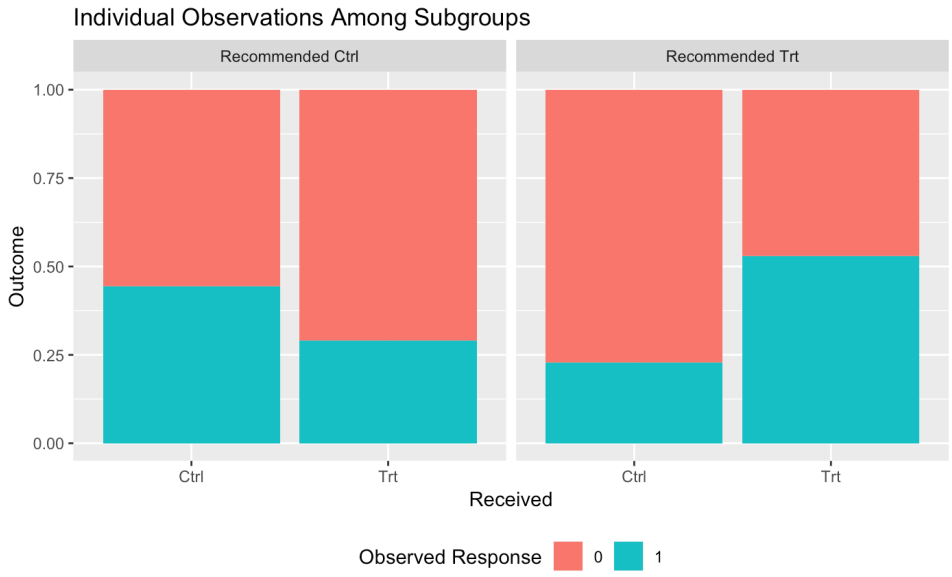

Figure 1 Example Boxplot

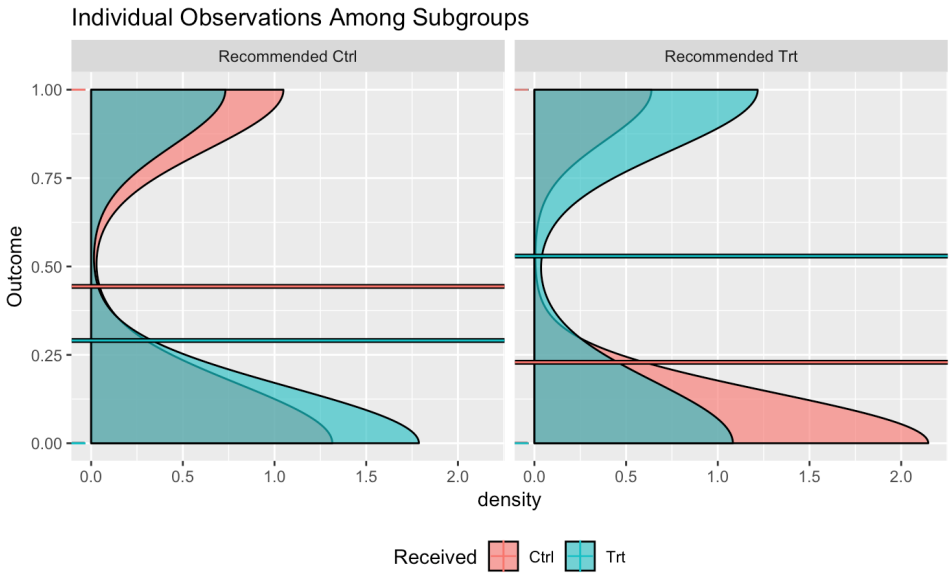

Figure 2 Example Density Plot

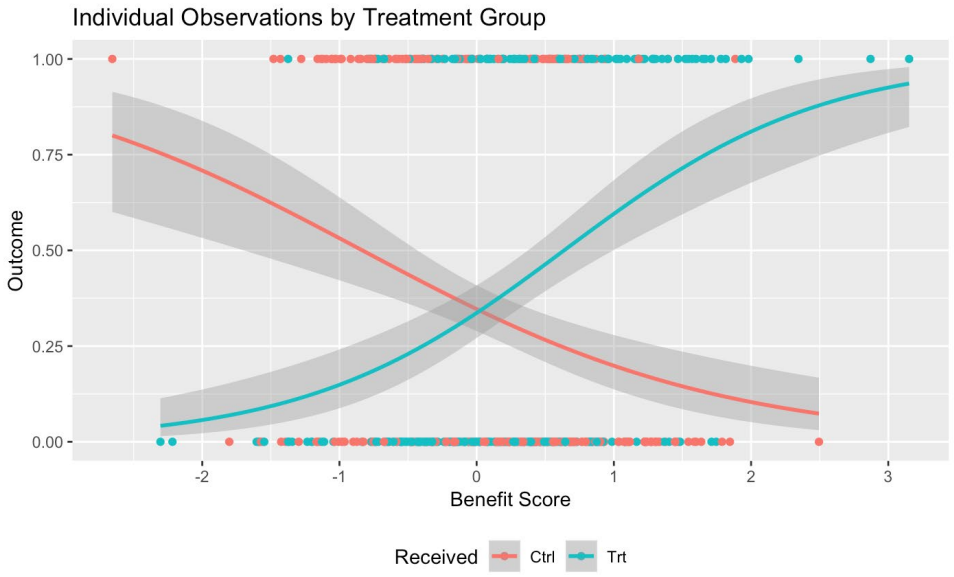

Figure 3 Example Conditional Plot

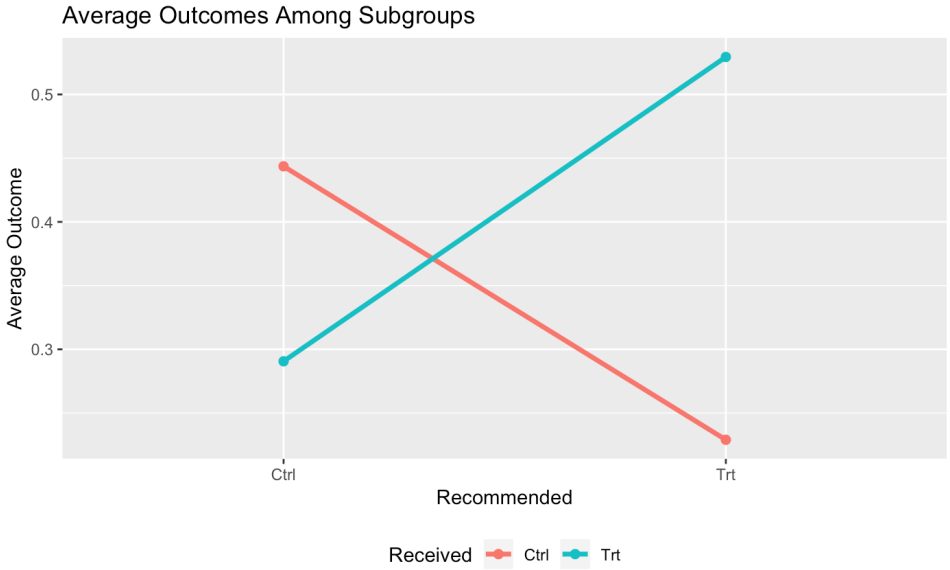

Figure 4 Example Interaction Plot

One convenient way to check the improvement from following an ITR would be look at the conditional plot and interaction plot. As Figure 3 shows, outcome for the control group (red line) decreases as benefit score increases; outcome for the treatment group (blue line) increases as benefit score increases. A non-parallel pattern of conditional plot like this indicates that following

the ITR improves the overall outcome. In Figure 4, a crossed graph means that patients have higher outcome, hence better response, when they received recommended treatment, suggesting benefit from ITRs. For the sake of brevity, this thesis will mainly discuss interaction plots as an indicator for improvement from ITRs.

3.0 Application to The NSABP B-41 Study

3.1 Analytical Methods

The NSABP B-41 trial data was collected from 271 adult participants with operable HER2-positive breast cancer. During their neoadjuvant chemotherapy, 82 of these participants received trastuzumab, 95 participants received a combination of trastuzumab and lapatinib, and the rest 94 participants received lapatinib only. Mastectomy or lumpectomy surgery was then performed after the treatment. Collected baseline covariates includes clinical markers such as age, race, and genomic markers, such as breast cancer subtype and cancer related gene expression. The outcome of interest, also the primary endpoint of NSABP B-41 trial, is pCR that coded as 1 or 0, with 1 represents complete response, 0 otherwise. Participants with a benefit score higher than the cutpoint would be assigned to the treatment, otherwise, the control would be recommended. Under each subgroup of treatment recommendation, the average outcome of participants who received the recommendation and the average outcome of participants who did not received the recommendation are compared, then treatment effects conditional on subgroups, $E[Y|T = Ctrl, T = Recom] - E[Y|T \neq Ctrl, T = Recom]$ for the control, and $E[Y|T = Trt, T = Recom] - E[Y|T \neq Trt, T = Recom]$ for the treatment, are computed.

The purpose of this thesis is to examine if there is the application of ITRs application on the NSABP B-41 trial data would improve the outcome. As previous studies have proven, trastuzumab has significantly better efficacy than lapatinib on treating HER2 positive breast cancer.[12,20] Our focus is thus the comparison of trastuzumab plus lapatinib (T + L) versus trastuzumab. We also compared treatment effects between trastuzumab containing treatment, thus

trastuzumab alone or trastuzumab plus lapatinib (trastuzumab-containing regimens), and lapatinib in order to detect participants with gene types that are more sensitive to lapatinib.

Three types of models are constructed using the variables in the data, clinical marker models, genomic marker models and overall models. The clinical marker models include age, race, lymph node status, HER2 gene status, estrogen-receptor status and tumor size. Due to the fact that *fit.subgroup* cannot handle missing data, subjects with missing values have to be removed from the analysis (29 out of 176, 16.5%, subjects removed for T + L VS Trastuzumab and 53 out of 271, 19.6%, subjects for trastuzumab containing regimens VS lapatinib). To avoid overfitting in later validation process, we choose to remove variables with more than 30 missing values. Here, tumor grade is removed from clinical marker models because it has too many missing values. The gene models are constructed by breast cancer subtype and gene expression of 58 genes. The overall models use all covariates in clinical markers and genomic markers. Generalized additive model (GAM) only applies to continuous variables, but categorical variables are included in each of these model. Therefore, we did not use GAM as the loss function. The hinge loss is used instead. For each of the three kinds, four scenarios are applied: (1) subgroup identification using the logistic loss; the median value of benefit scores is set as the cutoff point for treatment recommendation. (2) subgroup identification using the logistic loss; 0 is set as the cutoff point for treatment recommendation. (3) subgroup identification using the hinge loss and Gaussian, thus Radial Basis Function (RBF) kernel; the median value of benefit scores is set as the cutoff point for treatment recommendation. (4) subgroup identification using the hinge loss and Gaussian (RBF) kernel; 0 is set as the cutoff point for treatment recommendation. The models under the logistic loss uses the loss function $M(y, v) = y * log(1 + exp\{-v\})$, which is specified by setting the option *loss* to

"*logistic_loss_lasso*", while the hinge loss uses the loss function $M(y, v) = y * \max(0, 1 - v)$ specified by "*owl_hinge_loss*".

Treatment effects of the two recommendation groups might balance out and cause a seemingly overall beneficial treatment effect value when treatment effects of the two subgroups have similar absolute values but of different signs. To better examine the gain from individualized treatment assignment, treatment effects conditional on subgroups are reported instead. Positive treatment effects for both groups means the following the recommendation would improve the outcome.

To evaluate overoptimism of the estimated treatment rule and to obtain unbiased results, we perform validation by splitting the data into the training sets and the testing sets repeatedly. Here, we use 25% of the data as the training sets, and the rest 75% as the testing sets. The data is first randomly partitioned into these two sets, then subgroup treatment effects for the training sets are estimated by empirical average values of treatment effects in the testing sets. As the number of replications increases, the average results for all replications would approach the real values, and we can obtain unbiased estimates of the average outcomes for the training set. The number of replications is set to 100 in this study.

3.2 Results

3.2.1 T + L Versus Trastuzumab

In this section, we compare treatment effects between T + L and trastuzumab. Participants in the treatment group are those who received trastuzumab plus lapatinib during neoadjuvant

therapy (82 subjects), and participants in the control group received only lapatinib during neoadjuvant therapy (94 subjects). After deleting subjects with missing values, there are 72 participants left in T + L group, and 75 participants in Trastuzumab group.

We first use clinical markers to find the optimal individualized treatment. For the sake of simplicity, we use the numbers of scenarios stated in Section 3.1 to describe each setting. We first assess the results of optimal ITRs using the data itself without validation. Under each scenario, participants who received the recommended treatment have larger outcomes, and positive treatment effects for each subgroup are obtain, meaning the individualized treatment assignment improve the outcomes. For clinical marker models, the results of scenario (1) and (2) are shown in Table 1. Cut-points are specified in the parentheses after row names. The number in each cell is the average outcome of participants under that situation, and n marks the number of participants in each category. When we set the median value of benefit scores as the cutpoint, 78 participants are assigned to the control and 69 participants are assigned to the treatment. When 0 is used as the cut-point, 70 participants are assigned to the control and 77 participants are assigned to the treatment In both subgroups, treatment effects are positive. Less participants are assigned to the treatment in scenario (1). Higher average outcomes and treatment effect for each treatment group are achieved in scenario (1). Figure 5 to 8 show boxplot, density plot, interaction plot and conditional plot for scenario (1), respectively. It is clearer in the graphs that participants always have higher outcomes when they received the recommendation than when they received the other treatment. From Figure 7, the interaction plot, we can measure the gain of ITRs by the difference of average outcomes for each treatment. Figure 9 shows the comparison between scenario (1) and scenario (2), with the graph to the left representing scenario (1) and the graph to the right representing scenario (2). It can be observed that the blue line representing received treatment to

the right has a smaller slope, meaning that the improve of outcomes is smaller in scenario (2) when participants follow the treatment recommendation. This corresponds to what we observes in Table 1. In following analysis, we will only report the comparison of interaction plots of different scenarios since they are more straightforward.

Table 1 T + L (Treatment) VS. Trastuzumab (Control), Clinical Markers, the Logistic Loss

Cut-off		Recommended Control	Recommended Treatment
Median	Received Control	0.65 (n = 40)	0.4 (n = 35)
	Received Treatment	0.53 (n = 38)	0.62 (n = 34)
	Improvement	0.12 (n = 78)	0.22 (n = 69)
0	Received Control	0.69 (n = 36)	0.38 (n = 39)
	Received Treatment	0.56 (n = 34)	0.58 (n = 38)
	Improvement	0.14 (n = 70)	0.19 (n = 77)

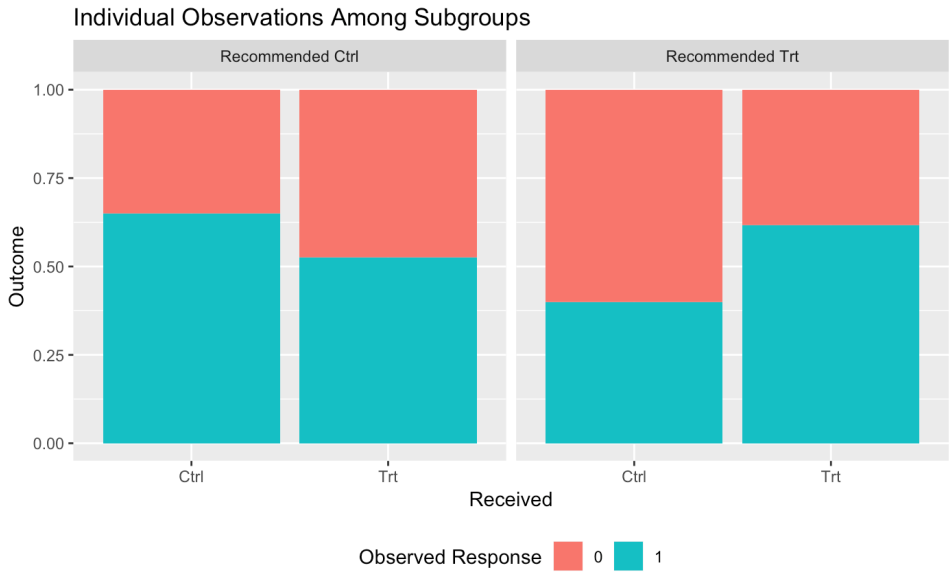

Figure 5 T + L (Treatment) VS. Trastuzumab (Control), Clinical Markers, Scenario(1), Boxplot

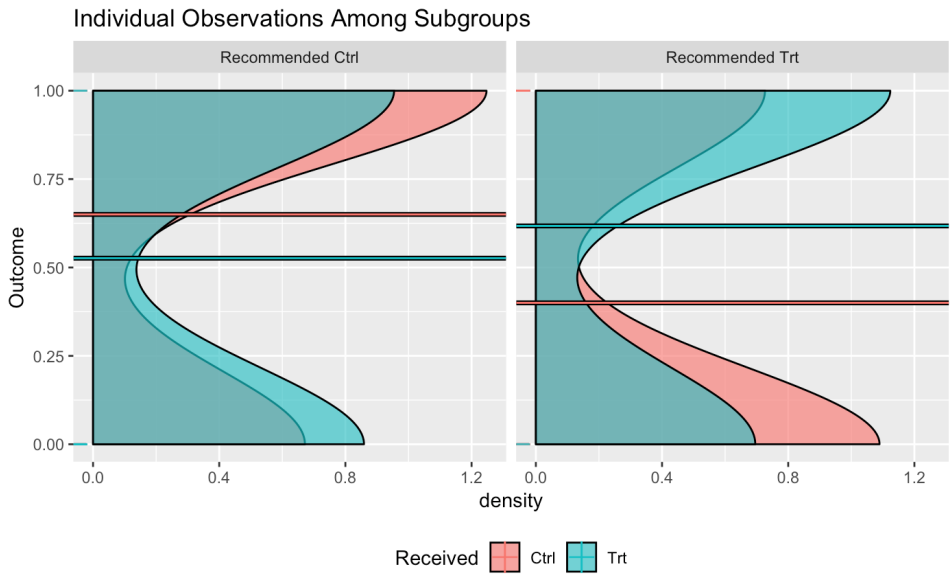

Figure 6 T + L (Treatment) VS. Trastuzumab (Control), Clinical Markers, Scenario(1), Desnsity Plot

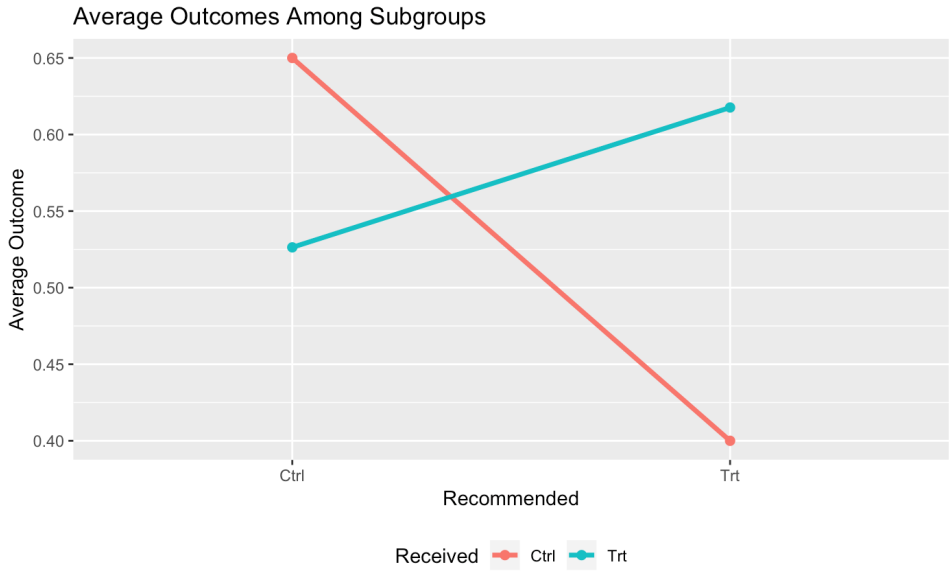

Figure 7 T + L (Treatment) VS. Trastuzumab (Control), Clinical Markers, Scenario(1), Interaction Plot

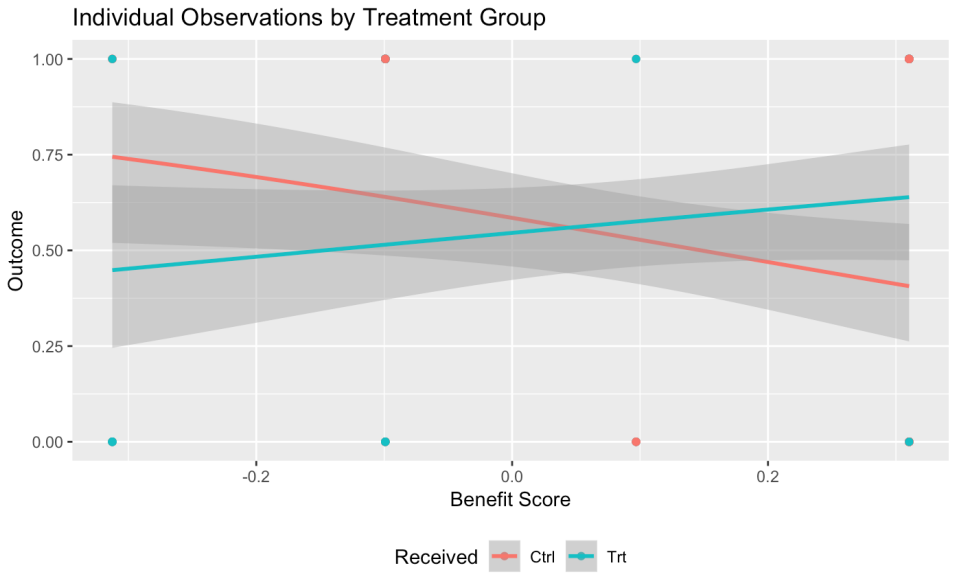

Figure 8 T + L (Treatment) VS. Trastuzumab (Control), Clinical Markers, Scenario(1), Conditional Plot

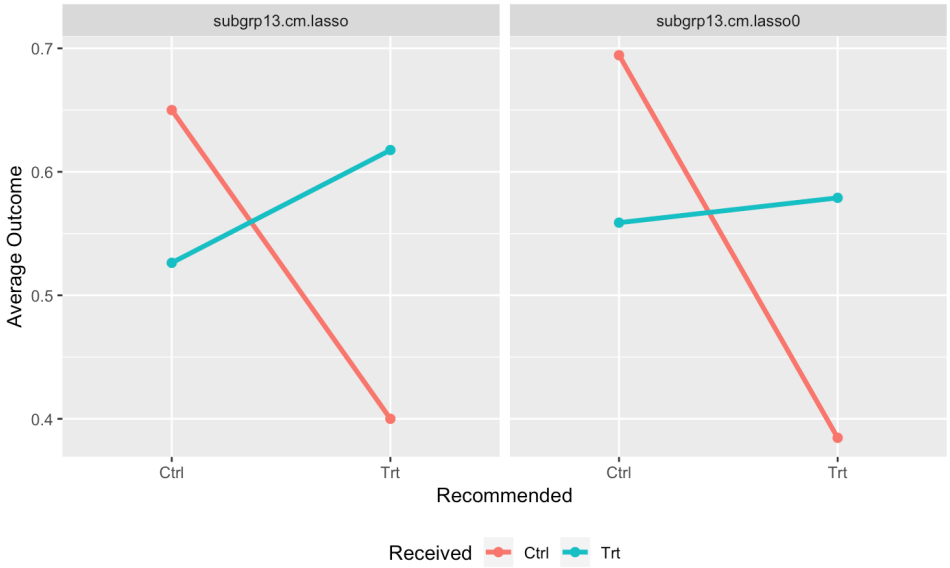

Figure 9 T + L (Treatment) VS. Trastuzumab (Control), Clinical Markers, Scenario (1) &(2)

Next, we examine the results using the hinge loss and Gaussian kernel. Table 2 shows the results of scenario (3) and scenario (4). The change of cutpoint does not significantly influence resulted treatment effects. More participants are assigned to the treatment group when 0 is set as the cut-point. The comparison plot of different cutpoints (Figure 10) shows that using 0 as the cut-

point (the graph to the right) leads to better improvement of the outcome. Figure 11 and Figure 12 show the comparisons between different loss functions when the cutpoint remains the same. The graphs in Figure 11 is resulted from the clinical marker models using the logistic loss (left) and the hinge loss (right). The median value of benefit scores is set as the cutpoint for both models. The graphs in Figure 12 shows the comparison between the logistic loss (left) and the hinge loss (right) for clinical marker models using 0 as the cutpoint. The hinge loss separate the participants in a more strict way, since the average outcomes of participants who received the same treatment in both subgroups differs more distinctively in scenario (3) and (4).

Table 2 T + L (Treatment) VS. Trastuzumab (Control), Clinical Markers, the Hinge Loss

Cut-off		Recommended Control	Recommended Treatment
Median	Received Control	0.67 (n = 49)	0.27 (n = 26)
	Received Treatment	0.32 (n = 25)	0.70 (n = 47)
	Improvement	0.35 (n = 74)	0.43 (n = 73)
0	Received Control	0.71 (n = 49)	0.19 (n = 26)
	Received Treatment	0.16 (n = 19)	0.72 (n = 53)
	Improvement	0.56 (n = 68)	0.52 (n = 79)

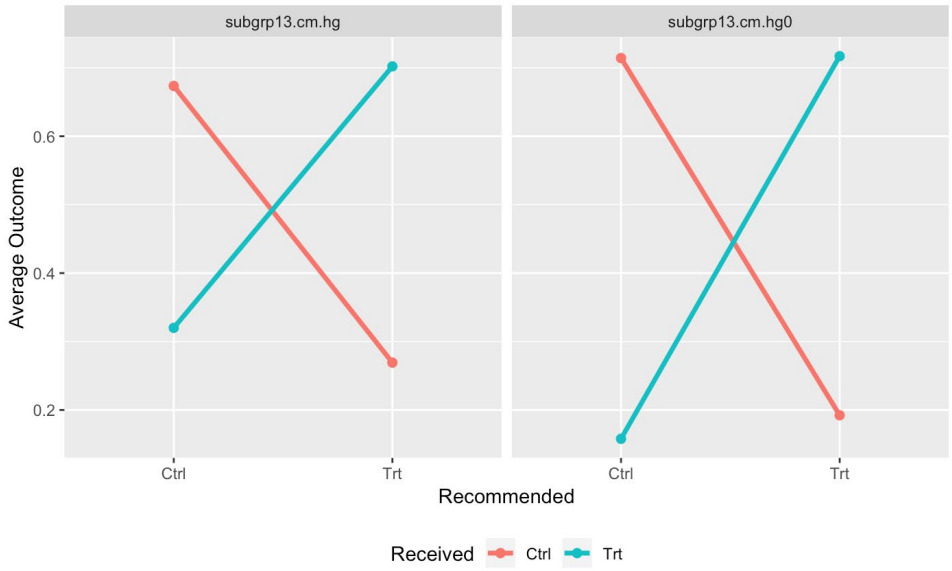

Figure 10 T + L (Treatment) VS. Trastuzumab (Control), Clinical Markers, Scenario (3) &(4)

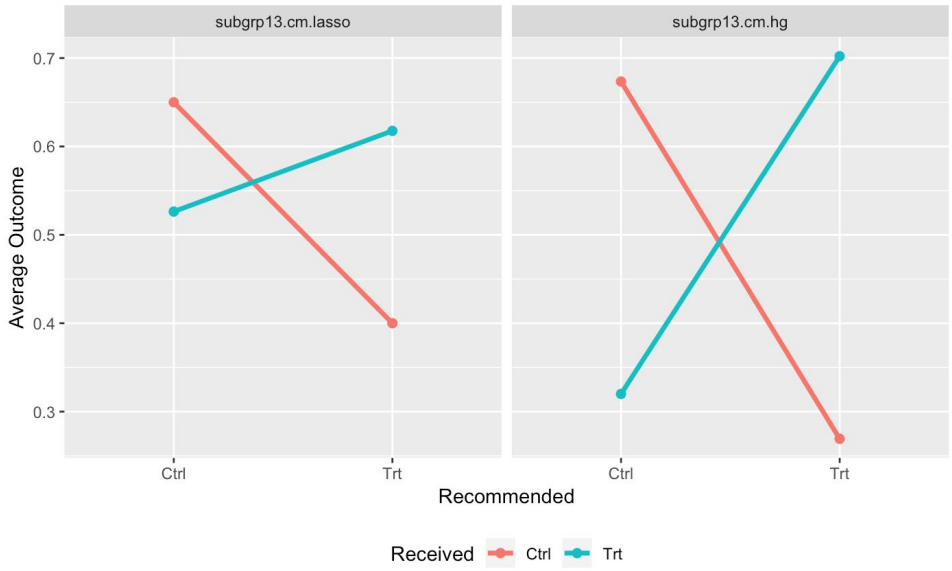

Figure 11 T + L (Treatment) VS. Trastuzumab (Control), Clinical Markers, Scenario (1) &(3)

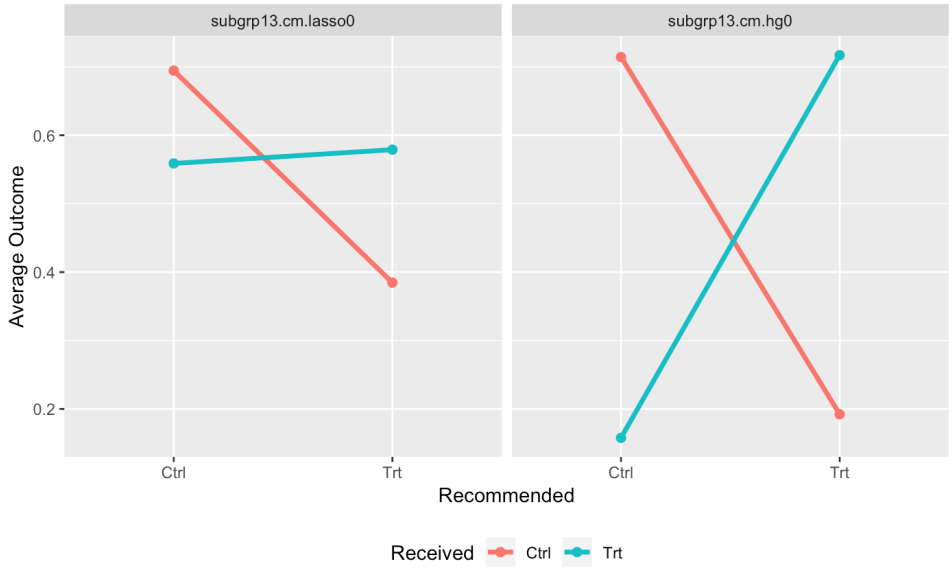

Figure 12 T + L (Treatment) VS. Trastuzumab (Control), Clinical Markers, Scenario (2) &(4)

The individualized treatment rule is then evaluated using genomic markers. The estimated treatment effects are shown in Table 3. Although both subgroups have positive treatment effects, the values under scenario (1) are smaller than values under scenario (2). Besides, the number of participants in the control recommended group varies greatly. Most participants (137/176) are assigned to the treatment under scenario (2), and the distribution of participants is less balanced within each subgroup. As shown in Figure 13, the estimated average outcomes of both subgroups varies more dramatically for participants who received the control, but they remain relatively stable for participants who received the treatment.

Table 3 T + L (Treatment) VS. Trastuzumab (Control), Genomic Markers, the Logistic Loss

Cut-off		Recommended Control	Recommended Treatment
Median	Received Control	0.60 (n = 47)	0.47 (n = 47)
	Received Treatment	0.56 (n = 41)	0.56 (n = 41)

Table 3 Continued

	Improvement	0.04 (n = 88)	0.09 (n = 88)
0	Received Control	0.84 (n = 25)	0.42 (n = 69)
	Received Treatment	0.57 (n = 14)	0.56 (n = 68)
	Improvement	0.27 (n = 39)	0.14 (n = 137)

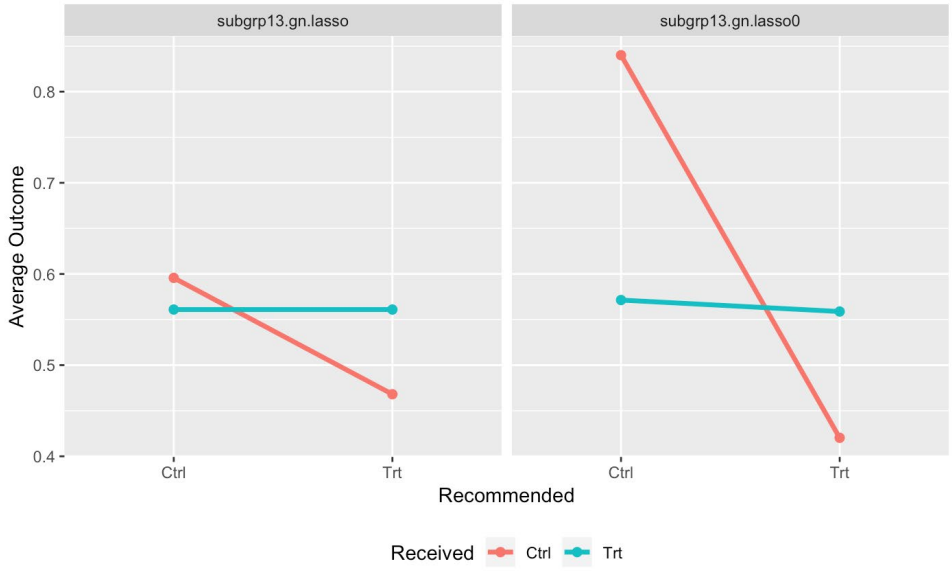

Figure 13 T + L (Treatment) VS. Trastuzumab (Control), Genomic Markers, Scenario (1) & (2)

We then change the loss function to the hinge loss, using Gaussian kernel. Similarly, participants are evenly assigned to both treatments under scenario (3) when the median benefit score is chosen as the cutpoint, but more participants are recommended the treatment when 0 is the cutpoint. The comparison plot of interactions (Figure 14) shows that under scenario (4), the average outcome of participants who received the treatment but assigned to the control and that of participants who received the treatment and assigned to the treatment are closer to each other (right), compared with scenario (3) (left). For the control recommended subgroup, this difference is smaller. When the four scenarios are compared together, the hinge loss seems to separate the

data better, especially for participants who received the treatment. As shown in the comparison plot between scenario (1) to the left and scenario (3) to the right (Figure 15), and the comparison plot between scenario (2) to the left and scenario (4) to the right (Figure 16), the blue line that representing participants who received T + L have steeper slopes in graphs to the right. This means that the average outcome of these participants in each subgroups differs more significantly than that of participants who received trastuzumab, the control.

Table 4 T + L (Treatment) VS. Trastuzumab (Control), Genomic Markers, the Hinge Loss

Cut-off		Recommended Control	Recommended Treatment
Median	Received Control	0.68 (n = 65)	0.21 (n = 29)
	Received Treatment	0.26 (n = 23)	0.68 (n = 59)
	Improvement	0.42 (n = 88)	0.47 (n = 88)
0	Received Control	0.75 (n = 51)	0.28 (n = 43)
	Received Treatment	0.41 (n = 22)	0.62 (n = 60)
	Improvement	0.34 (n = 73)	0.34 (n = 103)

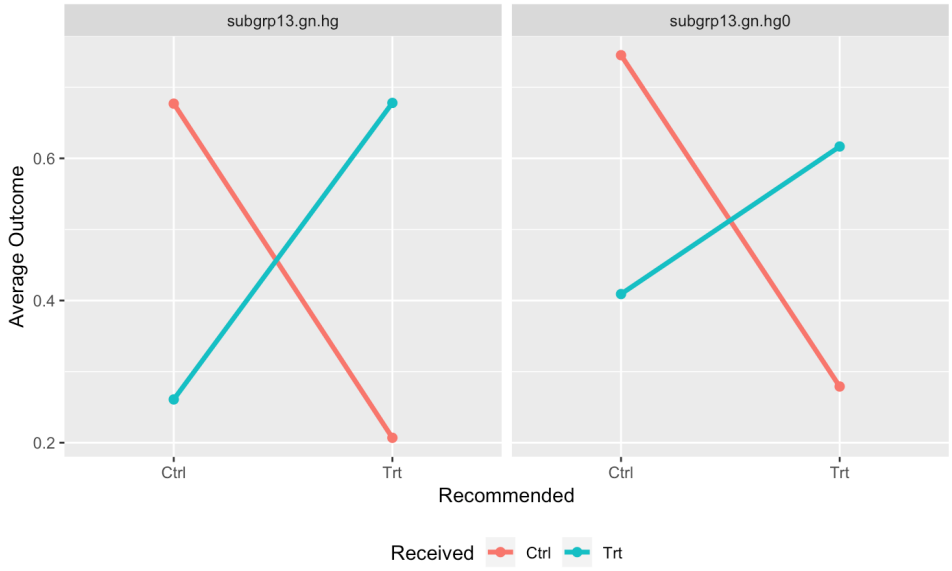

Figure 14 T + L (Treatment) VS. Trastuzumab (Control), Genomic Markers, Scenario (3) & (4)

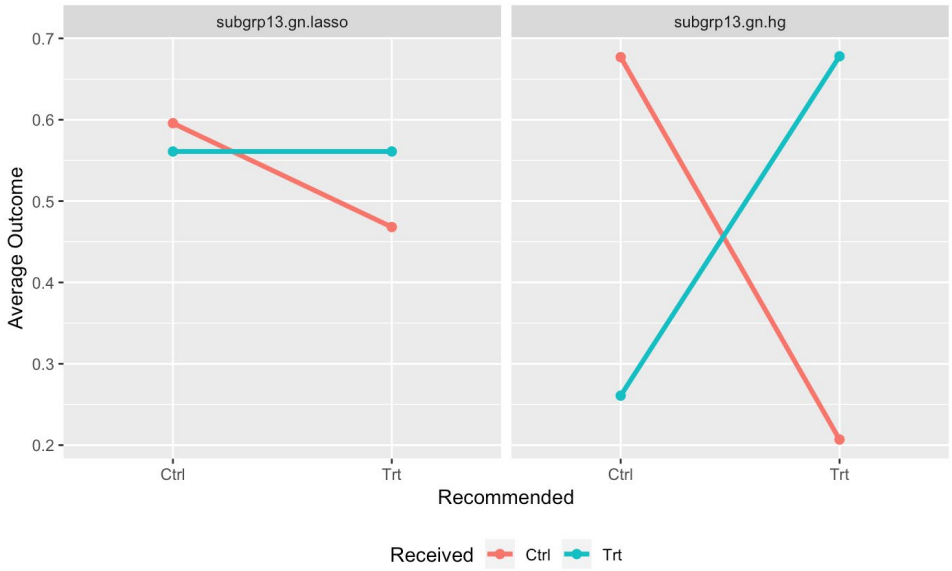

Figure 15 T + L (Treatment) VS. Trastuzumab (Control), Genomic Markers, Scenario (1) & (3)

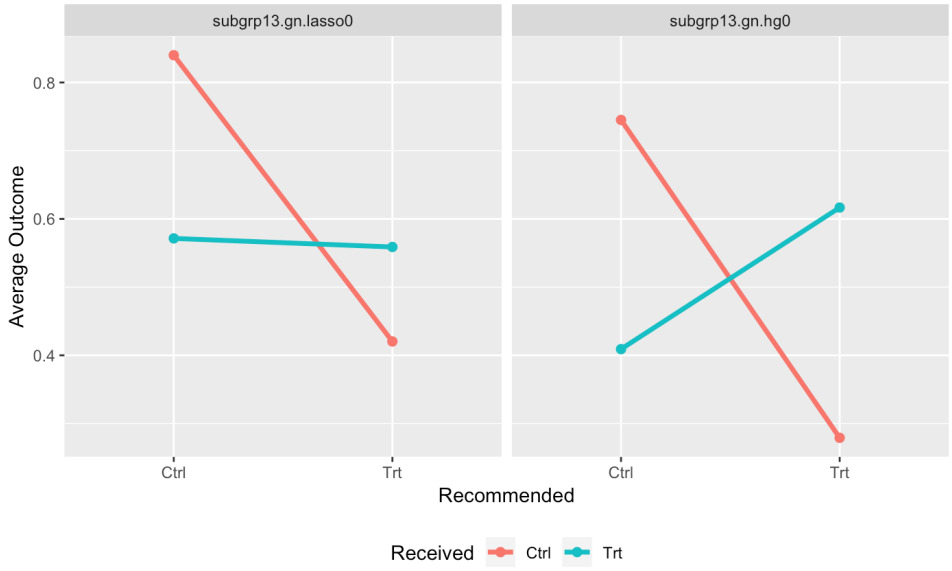

Figure 16 T + L (Treatment) VS. Trastuzumab (Control), Genomic Markers, Scenario (2) & (4)

All covariates, both clinical markers and genomic markers, are then combined together to evaluate the gain from ITR under the same scenarios. For scenario (1) and (2), the choice of cutpoint yields great influence to treatment effects. 74 participants are assigned to the control under scenario (1), but only 5 participants are in the same subgroup under scenario (2). Almost all participants are assigned to the treatment under scenario (2) (See Table 5). Participants are unevenly distributed for the control recommended group under scenario (2), which leads to a high treatment effect 0.75 since the only participant in the category "received the treatment but recommended the control" has an outcome of 0. In Figure 17, the graph to the right represents scenario (2). The steep slope of the blue line representing participants who received T + L is resulted from the above mentioned uneven distribution. The red lines in both graphs represent participants who received the control, and they are approximately parallel.

Table 5 T+L (Treatment) VS Trastuzumab (Control), All Covariates, the Logistic Loss

Cut-off		Recommended Control	Recommended Treatment
Median	Received Control	0.64 (n = 42)	0.39 (n = 33)
	Received Treatment	0.56 (n = 32)	0.58 (n = 40)
	Improvement	0.08 (n = 74)	0.18 (n = 73)
0	Received Control	0.75 (n = 4)	0.52 (n = 71)
	Received Treatment	0 (n = 1)	0.58 (n = 71)
	Improvement	0.75 (n = 5)	0.06 (n = 142)

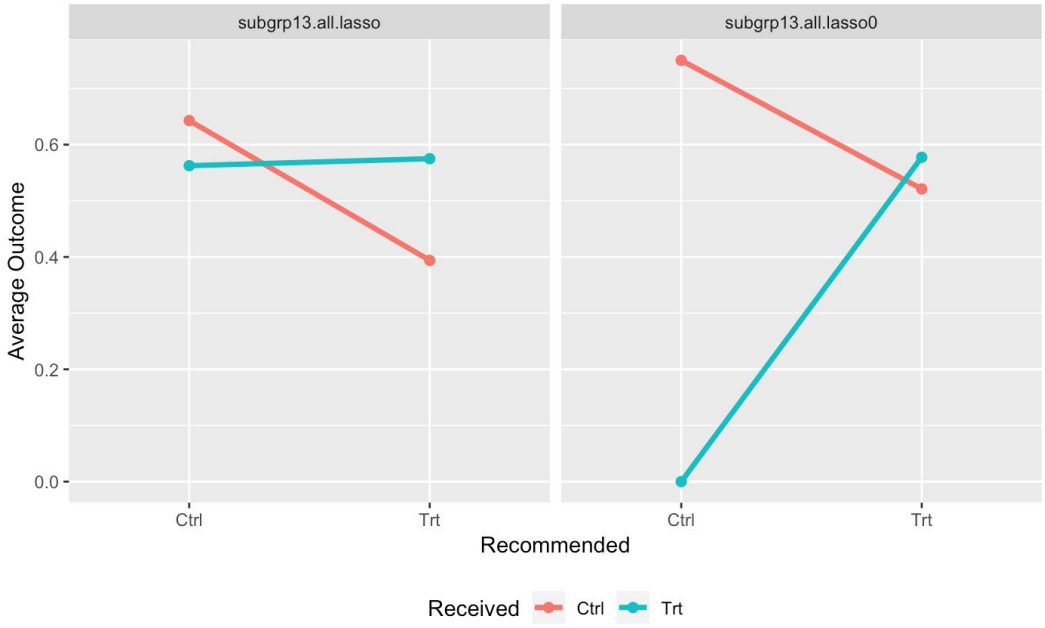

Figure 17 T+L (Treatment) VS Trastuzumab (Control), All Covariates, Scenario (1) & (2)

We then move on to scenario (3) and (4), where the loss function is the hinge loss using Gaussian kernel. Overall, no significant change is observed as the cutpoint varies. The median benefit score is more restricted than 0 as the cutpoint, since more people under scenario (4) are

assigned to the treatment. In Figure 18, the graph to the left shows the interaction of average outcomes under scenario (3), and the graph to the right shows the interaction of average outcomes under scenario (4). The average outcome of participants who received the control in both subgroups remains relatively stable when the cutpoint varies, but that of participants who received the treatment is a little higher under scenario (3), meaning a slightly better separation of T + L sensitive participants. Similarly, Figure 19 and Figure 20 show the comparison of treatment effects of different loss function using same cutpoints. Compared to scenario (1) in Figure 19 (left), scenario (3) to the right of Figure 19 better improves the outcome, since the average outcomes of two subgroups differ greatly in the plot to the right. When 0 is set as the cutpoint, scenario (2) achieves better improvement for the treatment (Figure 20to the left), but scenario (3) achieves better improvement for the control (Figure 20 to the right).

Table 6 T + L (Treatment) VS. Trastuzumab (Control), All Covariates, the Hinge Loss

Cut-off		Recommended Control	Recommended Treatment
Median	Received Control	0.70 (n = 54)	0.095(n = 21)
	Received Treatment	0.25 (n = 20)	0.69 (n = 52)
	Improvement	0.45 (n = 74)	0.60 (n = 73)
0	Received Control	0.80 (n = 41)	0.21 (n = 34)
	Received Treatment	0.25 (n = 12)	0.63 (n = 60)
	Improvement	0.55 (n = 53)	0.43 (n = 94)

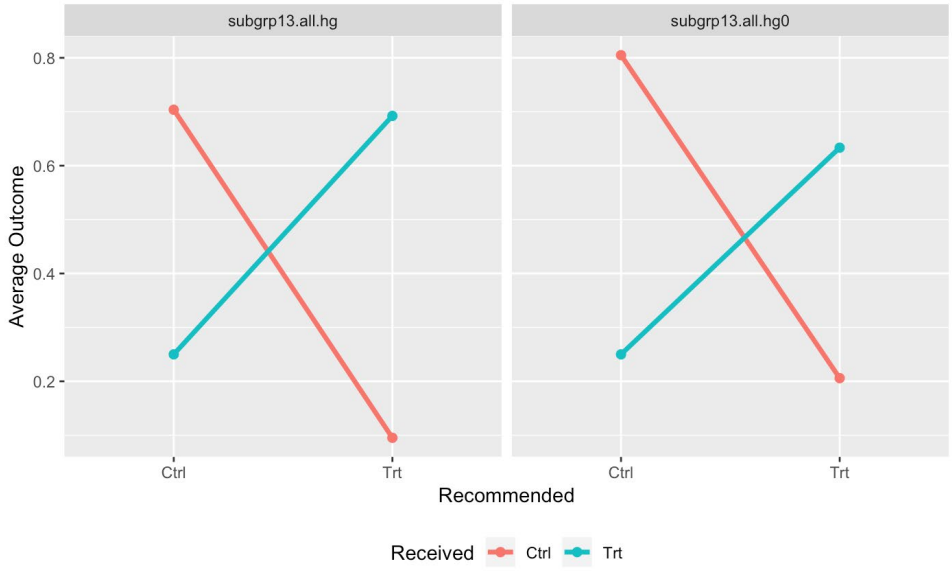

Figure 18 T + L (Treatment) VS. Trastuzumab (Control), All Covariates, Scenario (3) & (4)

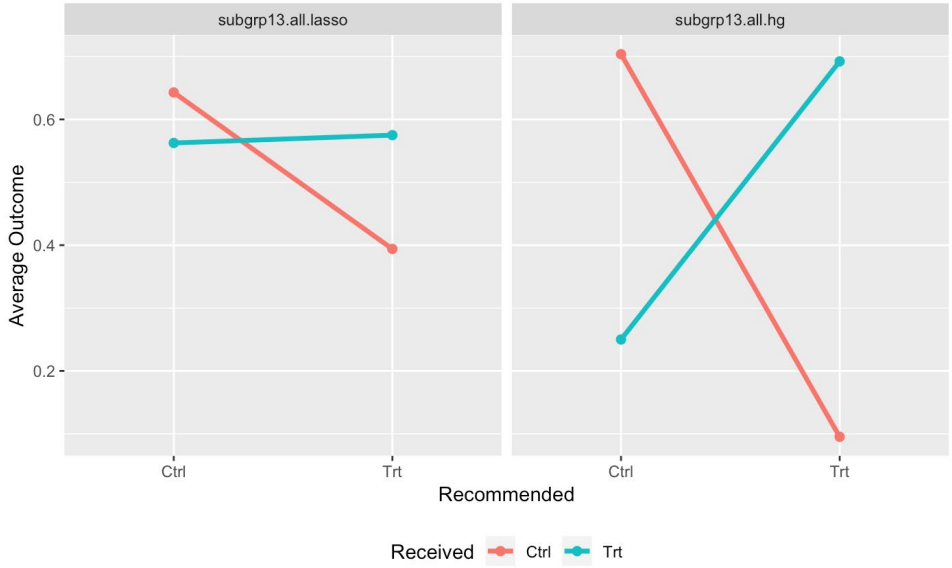

Figure 19 T + L (Treatment) VS. Trastuzumab (Control), All Covariates, Scenario (1) & (3)

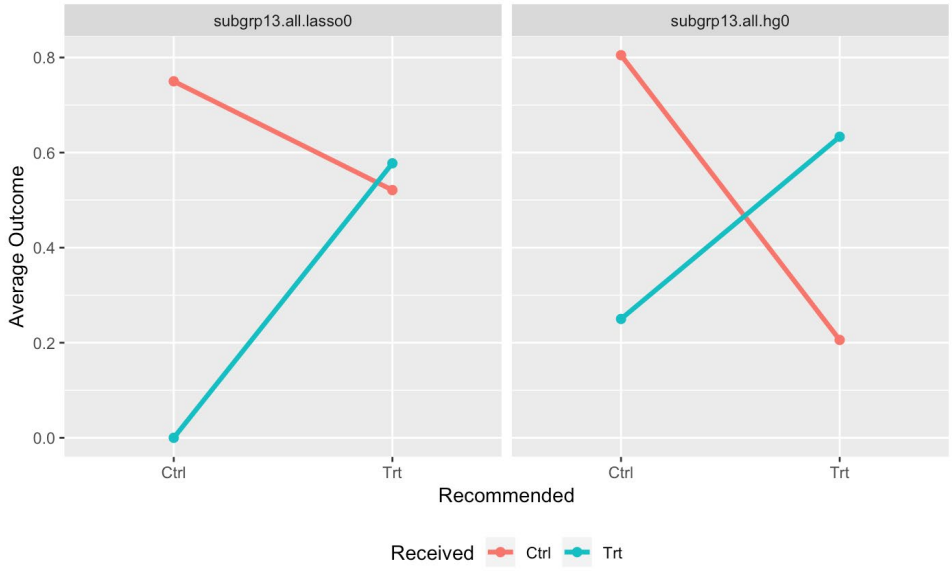

Figure 20 T + L (Treatment) VS. Trastuzumab (Control), All Covariates, Scenario (2) & (4)

Overall, the subgroup identification results under all scenarios look promising that the application of ITRs to the NSABP B-41 trail data can improve the outcome both for those who received trastuzumab combined with lapatinib and for those who received trastuzumab alone. However, as previous stated, above results can be biased estimates of treatment effects, since the comparison of treatment effects of same participants under different treatments is not conducted. In order to obtain unbiased results, we then conduct validations. The results displayed below are average values of the 100 replications, with SE indicating the standard error of that value among all replications. This also leads to non-integer values of sample sizes (n) in each category.

For clinical marker models using the logistic loss, the results of subgroup separation change significantly after validation. Under both scenarios, treatment effects of control recommendation groups become negative, meaning participants have better outcomes when not following the recommendation. Although treatment effects remain the same for treatment recommendation groups, their values are small (Table 7). Besides, in each subgroup, participants who received the treatment always have higher outcomes. This can also be reflected in Figure 21, with the plot to

the left representing scenario (1) and the plot to the right representing scenario (2). For all participants, those who are recommended the control have higher outcomes on average under both scenarios. This seems contrary to what we observed that patients always have higher outcome on average when receiving the treatment. However, if we further examine the scale of the interaction plot, the difference of average outcomes between two subgroups are pretty small, less or around 0.1. This difference is likely to be caused by some noise and therefore can be ignored.

Table 7 T + L (Treatment) VS. Trastuzumab (Control), Validated Clinical Markers, the Logistic Loss

Cut-off		Recommended Control	Recommended Treatment
Median	Received Control	0.57 (SE = 0.1395, n = 15.07)	0.47 (SE = 0.1814, n = 3.74)
	Received Treatment	0.59 (SE = 0.1039, n = 14.44)	0.54 (SE = 0.1509, n = 3.75)
	Improvement	-0.02 (SE = 0.1353, n = 29.51)	0.07 (SE = 0.2889, n = 7.49)
0	Received Control	0.56 (SE = 0.1542, n = 6.66)	0.44 (SE = 0.1311, n = 12.56)
	Received Treatment	0.58 (SE = 0.2023, n = 6.76)	0.56 (SE = 0.1562, n = 11.02)
	Improvement	-0.02 (SE = 0.1908, n = 13.42)	0.12 (SE = 0.1712, n = 23.58)

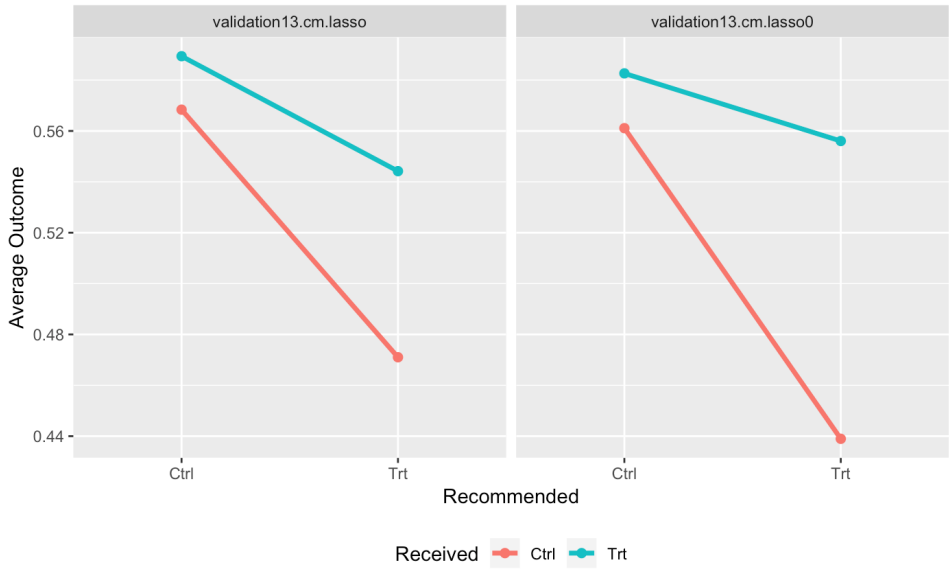

Figure 21 T + L (Treatment) VS. Trastuzumab (Control), Validated Clinical Markers, Scenario (1) & (2)

We then examine the validated results for clinical marker models using the hinge loss. Under both scenario (3) and (4), we correctly assign participants into two subgroups in a way that participants in both subgroups have positive treatment effects, thus higher outcome values when follow the recommendations (Table 8). Appear in the interaction plots of average outcomes (Figure 22), two lines representing different treatment status crossed in both plot, meaning that for some patients, receiving the control would lead to better response than the treatment. However, this improvement is very small. Under both cutpoints, the improvement is less than 0.05, which is almost negligible.

Table 8 T + L (Treatment) VS. Trastuzumab (Control), Validated Clinical Markers, the Hinge Loss

Cut-off		Recommended Control	Recommended Treatment
Median	Received Control	0.60 (SE = 0.1592, n = 9.42)	0.45 (SE = 0.1288, n = 9.51)
	Received Treatment	0.56 (SE = 0.1389, n = 9.58)	0.65 (SE = 0.1073, n = 8.49)
	Improvement	0.04 (SE = 0.2078, n = 19)	0.20 (SE = 0.1682, n = 18)
0	Received Control	0.59 (SE = 0.1659, n = 5.83)	0.47 (SE = 0.1271, n = 13.15)
	Received Treatment	0.56 (SE = 0.1441, n = 6.76)	0.69 (SE = 0.1307, n = 11.26)
	Improvement	0.02 (SE = 0.2419, n = 12.59)	0.22 (SE = 0.2015, n = 24.41)

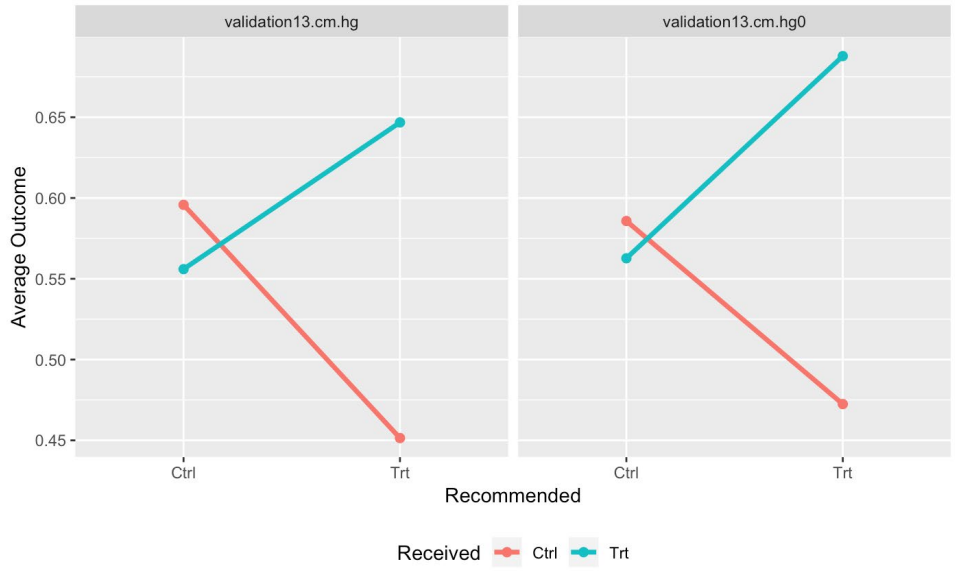

Figure 22 T + L (Treatment) VS. Trastuzumab (Control), Validated Clinical Markers, Scenario (3) & (4)

Genomic marker models using the logistic loss do not achieve meaningful separation after validation. Under both scenario (1) and (2), participants always have better outcomes when they receive the treatment, trastuzumab plus lapatinib, regardless of the recommended treatment (Table 9). As the cutpoint varies, the gain of ITR for does not vary much for the control recommendation. Under scenario (2), the absolute value of treatment effect decreases for the treatment recommendation, but it increases for the control group. Also, the number of participants that is assigned to the treatment increases. For all subgroups under both scenarios, participants who are recommended the control have higher outcomes on average. This is especially apparent under scenario (1), within participants who received the control, trastuzumab only (Figure 23).

Table 9 T + L VS Trastuzumab, Validated Gene Models, the Logistic Loss

Cut-off		Recommended Control	Recommended Treatment
Median	Received Control	0.58 (SE = 0.1217, n = 14.36)	0.43 (SE = 0.1592, n = 8.84)
	Received Treatment	0.60 (SE = 0.164, n = 11.6)	0.56 (SE = 0.1099, n = 9.2)
	Improvement	-0.02 (SE = 0.1702, n = 25.96)	0.13 (SE = 0.1602, n = 18.04)
0	Received Control	0.57 (SE = 0.2208, n = 7.81)	0.49 (SE = 0.1404, n = 16.03)
	Received Treatment	0.60 (SE = 0.206, n = 6.19)	0.55 (SE = 0.1063, n = 13.97)
	Improvement	-0.04 (SE = 0.2155, n = 14)	0.05 (SE = 0.1611, n = 30)

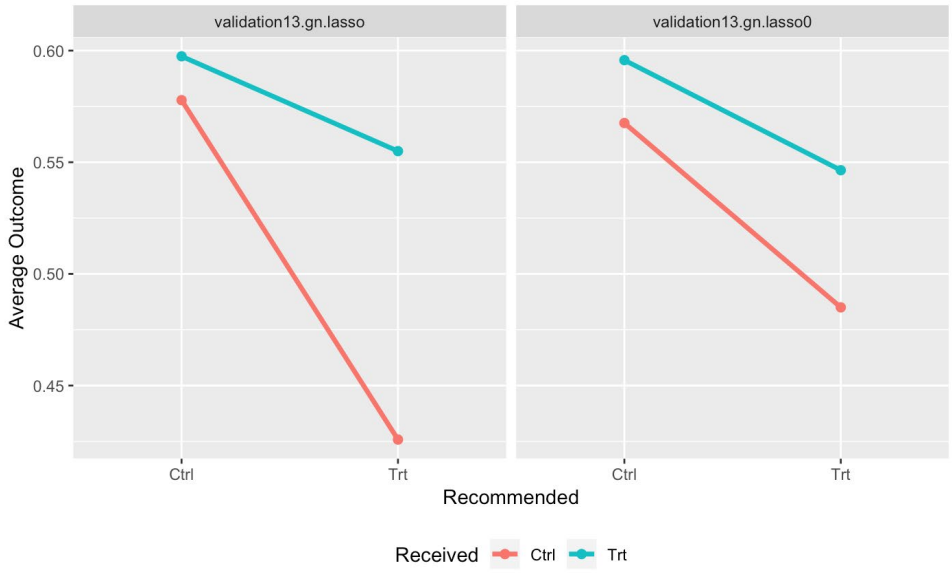

Figure 23 T + L (Treatment) VS. Trastuzumab (Control), Validated Genomic Markers, Scenario (1) & (2)

The hinge loss cannot differentiate subgroups based on genomic markers of this data, either. Participants who received the treatment always obtain higher better outcomes on average. Thus, treatment effects for the control recommendation is always negative, but it is always positive for the treatment recommendation (Table 10). Interpretation to the interaction plot is similar:

patients always benefit more from the treatment. Participant who are assigned the control have slightly higher average outcomes than the treatment group under the hinge loss, which is likely to be cause by noise, too (Figure 24).

Table 10 T + L (Treatment) VS. Trastuzumab (Control), Validated Gene Models, the Hinge Loss

Cut-off		Recommended Control	Recommended Treatment
Median	Received Control	0.58 (SE = 0.1299, n = 11.48)	0.49 (SE = 0.1485, n = 11.86)
	Received Treatment	0.64 (SE = 0.1224, n = 10.52)	0.57 (SE = 0.1242, n = 10.14)
	Improvement	-0.07 (SE = 0.1754, n = 22)	0.08 (SE = 0.161, n = 22)
0	Received Control	0.54 (SE = 0.1444, n = 9.57)	0.47 (SE = 0.143, n = 14.1)
	Received Treatment	0.58 (SE = 0.104, n = 9.52)	0.56 (SE = 0.1442, n = 10.81)
	Improvement	-0.04 (SE = 0.176, n = 19.09)	0.08 (SE = 0.1787, n = 24.91)

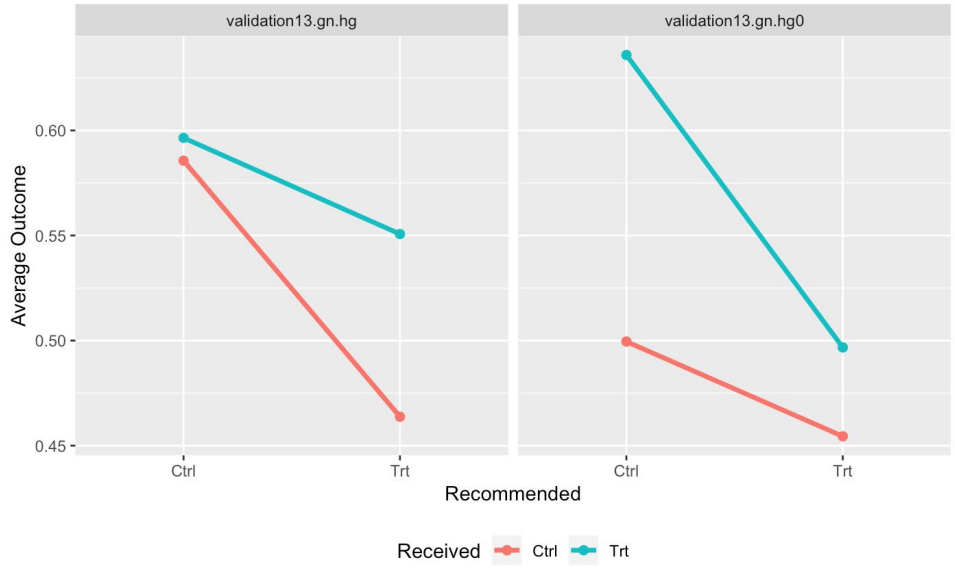

Figure 24 T + L (Treatment) VS. Trastuzumab (Control), Validated Genomic Markers, Scenario (3) & (4)

Next, we assess the gain from ITR using all covariates. Under both scenarios, receiving the treatment always leads to higher average outcomes. In terms of the treatment recommendation, 0

is more lenient than the median value of benefit scores as the cutpoint, since around 28 participants are assigned to the treatment under scenario (2), while only around 11 participants are assigned to the treatment under scenario (1). The interaction between the recommended treatment and average outcomes can be clearly observed in Figure 25. Although differences of average outcomes exist between subgroups, the lines representing the treatment is above the control line, indicating no improvement from following ITRs.

Table 11 T + L (Treatment) VS. Trastuzumab (Control), Validated Overall Models, the Logistic Loss

Cut-off		Recommended Control	Recommended Treatment
Median	Received Control	0.59 (SE = 0.1294, n = 13.72)	0.51 (SE = 0.1884, n = 5.74)
	Received Treatment	0.67 (SE = 0.1006, n = 11.79)	0.56 (SE = 0.1086, n = 5.75)
	Improvement	-0.09 (SE = 0.195, n = 25.51)	0.05 (SE = 0.1784, n = 11.49)
0	Received Control	0.46 (SE = 0.2364, n = 5)	0.56 (SE = 0.1582, n = 14)
	Received Treatment	0.63 (SE = 0.231, n = 4.37)	0.59 (SE = 0.1358, n = 13.63)
	Improvement	-0.20 (SE = 0.2266, n = 9.37)	0.03 (SE = 0.2061, n = 27.63)

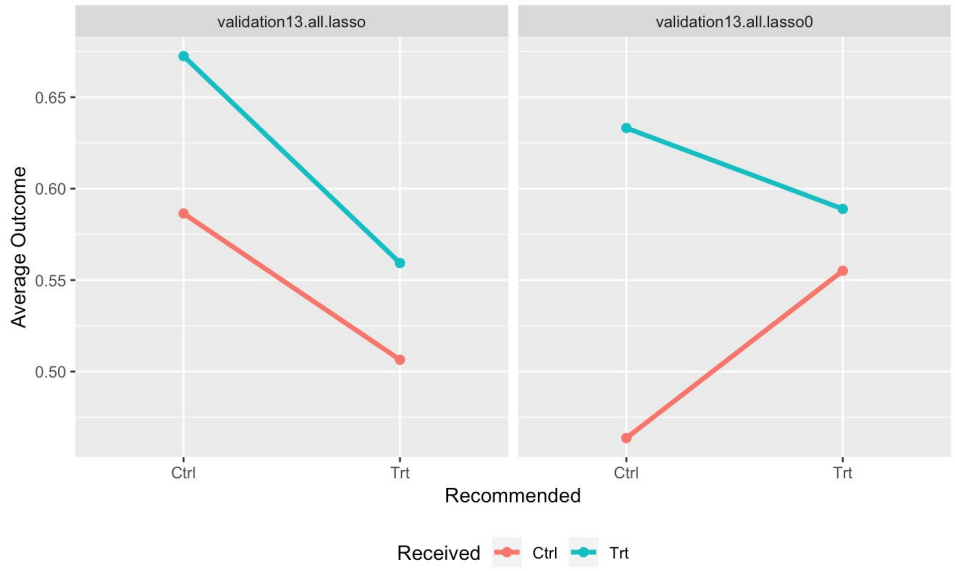

Figure 25 T + L (Treatment) VS. Trastuzumab (Control), Validated Overall Models, Scenario (1) & (2)

The overall model was then fitted under the hinge loss. No significant gain is observed either. Interpretation of the resulted table is similar to previous results. One thing worth noticing is that under scenario (4), both subgroup achieve positive improvement. Nonetheless, for the control subgroup, this improvement is only 0.001, which might be caused by noise (Table 12). Appearing on the interaction plot, two lines for scenario (4) are overlapped on control subgroup (Figure 26) and no meaningful improvement can be observed.

Table 12 T + L (Treatment) VS. Trastuzumab (Control), Validated Overall Models, the Hinge Loss

Cut-off		Recommended Control	Recommended Treatment
Median	Received Control	0.60 (SE = 0.1534, n = 9.93)	0.50 (SE = 0.149, n = 9.28)
	Received Treatment	0.65 (SE = 0.1621, n = 9.07)	0.54 (SE = 0.1578, n = 8.72)
	Improvement	-0.05 (SE = 0.2225, n = 19)	0.04 (SE = 0.2335, n = 18)
0	Received Control	0.64 (SE = 0.1808, n = 4.98)	0.48 (SE = 0.1763, n = 14.1)
	Received Treatment	0.64 (SE = 0.1671, n = 5.48)	0.53 (SE = 0.1312, n = 12.44)
	Improvement	0.001 (SE = 0.2282, n = 10.46)	0.06 (SE = 0.2051, n = 26.54)

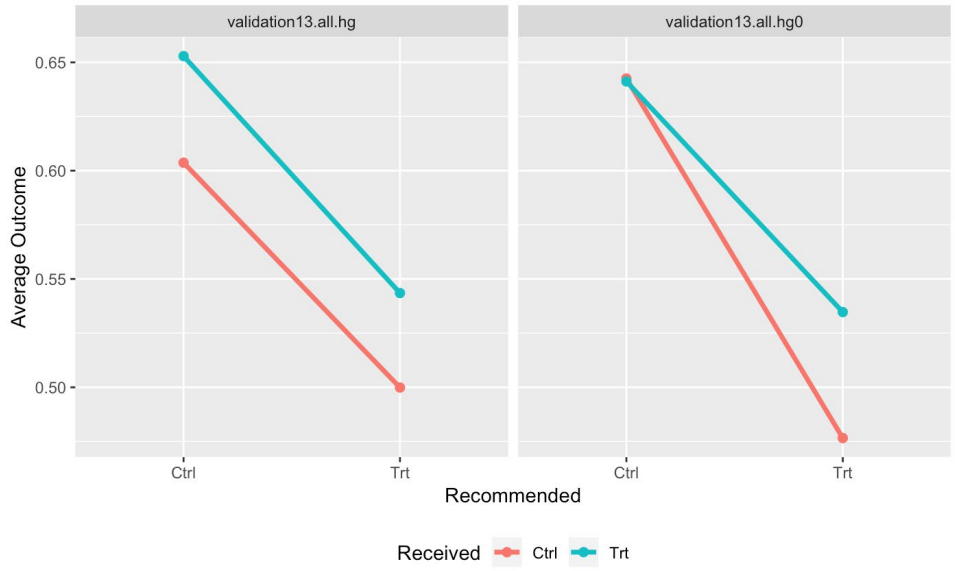

Figure 26 T + L (Treatment) VS. Trastuzumab (Control), Validated Overall Models, Scenario (3) & (4)

3.2.2 Trastuzumab-containing Regimens Versus Lapatinib

It has been proven that trastuzumab has better efficacy than lapatinib.[12,20] The comparison between trastuzumab-containing regimens treatment and lapatinib is thus not very clinically meaningful. This analysis is done to assess if there is any participants have particular characteristics that are more sensitive to lapatinib. No significant improvement by following ITR is observed in any of the models we proposed under the four scenarios. In all models, patients in treatment subgroup always obtain higher outcomes on average than patients in control subgroup. For the sake of brevity, here we only report the validated subgroup identification results when all covariates are included.

Overall, individualized subgroup identification does not improve participants' outcome when comparing treatment effects of trastuzumab-containing regimens (the treatment) and lapatinib (the control). Table 13 shows that for both cutpoints under scenario (1) and (2), participants always have higher outcomes on average when receiving the treatment, irrespective

of the treatment recommendation. More directly, in the treatment and outcome interaction plot (Figure 27), the blue line is above the red line in both plots, meaning that the average outcome for participants who received the treatment are higher than that of participants who received the control in both treatment recommendation subgroups.

Table 13 Trastuzumab-containing Regimens (Treatment)VS Lapatinib (Control), Validated Overall Models, the Logistic Loss

Cut-off		Recommended Control	Recommended Treatment
Median	Received Control	0.45 (SE = 0.1778, n = 10.38)	0.42 (SE = 0.157, n = 7.9)
	Received Treatment	0.49 (SE = 0.1467, n = 20.05)	0.61 (SE = 0.1131, n = 16.67)
	Improvement	-0.04 (SE = 0.1721, n = 30.43)	0.19 (SE = 0.1585, n = 24.57)
0	Received Control	0.25 (SE = 0.3048, n = 1.31)	0.43 (SE = 0.1258, n = 16.54)
	Received Treatment	0.31 (SE = 0.2841, n = 3.64)	0.57 (SE = 0.0707, n = 33.51)
	Improvement	-0.19 (SE = 0.3003, n = 4.95)	0.14 (SE = 0.1516, n = 50.05)

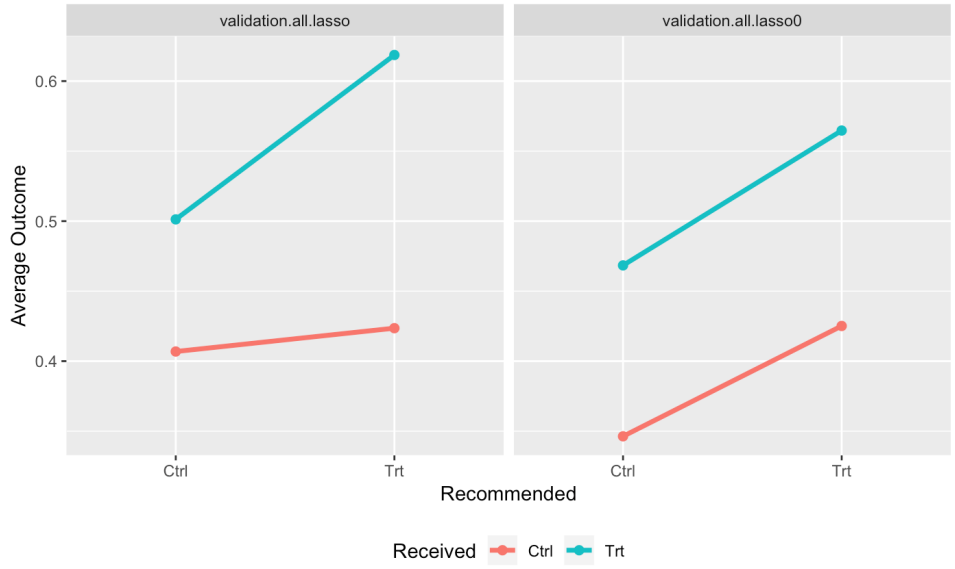

Figure 27 Trastuzumab-containing Regimens (Treatment)VS Lapatinib (Control), Validated Overall Models, Scenario (1) & (2)

Results for the hinge loss are similar to what we obtain under the logistic loss. Participants who received the treatment have higher average outcomes in both scenarios. Hence, treatment effects are always positive for the treatment recommendation, but always negative for the control recommendation. In the interaction plots for the two scenarios (Figure 28), the blue lines that representing treatment received group are above the red lines that representing control received group in both plots, indicating no improvement from ITRs.

Table 14 Trastuzumab-containing Regimens (Treatment)VS Lapatinib (Control), Validated Overall Models, the Hinge Loss

Cut-off		Recommended Control	Recommended Treatment
Median	Received Control	0.42 (SE = 0.1555, n = 8.47)	0.45 (SE = 0.1596, n = 9.16)
	Received Treatment	0.56 (SE = 0.1108, n = 19.53)	0.58 (SE = 0.1158, n = 17.84)
	Improvement	-0.14 (SE = 0.2134, n = 28)	0.13 (SE = 0.2174, n = 27)
0	Received Control	0.06 (SE = 0.0962, n = 0.36)	0.39 (SE = 0.1071, n = 17.64)
	Received Treatment	0.60 (SE = 0.1624, n = 1.15)	0.54 (SE = 0.0766, n = 35.85)
	Improvement	-0.55 (SE = 0.0906, n = 1.51)	0.15 (SE = 0.1365, n = 53.49)

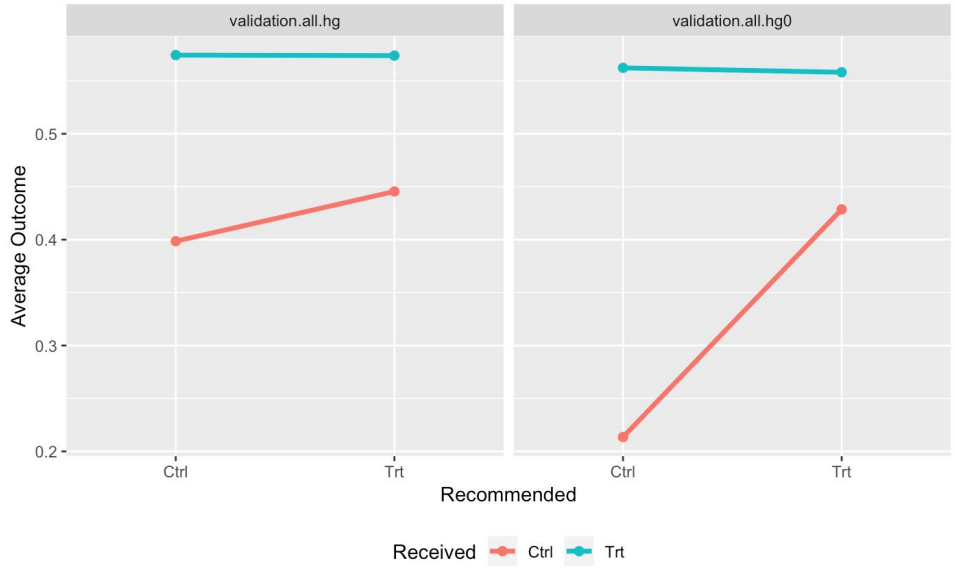

Figure 28 Trastuzumab-containing Regimens (Treatment)VS Lapatinib (Control), Validated Overall Models, Scenario (3) & (4)

4.0 Discussion

With recent advance in biotechnology, large amount of biomarkers from different platforms become available in the past decade. Precision medicine emerges as an important issue that these markers could help to identify patient heterogeneity in treatment response and lead to better patient management and more effective treatment regimens. Individualized treatment rule provides a useful perspective about precision medicine and obtaining an optimal ITR from existing data supplies a natural application in precision medicine. Following several methods that attempt to define and estimate an optimal ITR, Chen et al. (2017)[8] proposed a general framework for the optimality of an ITR under various choices of loss functions or risk functions, and subsequent estimation and inference procedure. Huiling and Yu (2018)[10] incorporated this framework into an R package, "*personalized*". We applied this package to analyze a sub-study of the NSABP B-41 study. [Robidoux et al., 2013] It was demonstrated that both clinical markers and genomic markers from the PAM50 panel could lead to much improved and promising patient management scheme had the estimated optimal ITR been applied to the same group of patients with the amount of improvement in the pCR varying from 0.43 to 0.56 in various subgroups, as shown in Tables 2, 4 and 6 where the hinge loss function was applied. However, internal validation via repeatedly random splitting into training data sets and testing data sets did not produce consistent improvement in the pCR had the estimated optimal ITRs been applied to the testing datasets. The ITR-based methods provide a powerful tool to identify predictive treatment markers and optimal scheme for marker-directed treatment assignment, and lead to precision medicine in practice. However, the performance of the developed ITRs need to be validated via external validation before their application.

Appendix A Example R Code

Below is an example of the workflow of subgroup identification using the "*personalized*" package in R. The process shown here is using the clinical markers model to compare treatment efficacy between T + L and Trastuzumab under the logistic loss.

```
library(personalized)
mydata_full = read.csv("filepath", header = T, stringsAsFactors = T)
sapply(mydata_full, class)
factors = c(4,8,9,10,11,12,106,108:110,112,114:130,132,133:135,195:198)
mydata_full[,factors] = lapply(mydata_full[,factors], factor)
groups = c(1,3)
mydata_13 = mydata_full[mydata_full$TRT %in% groups,]  #select out group 1 and group 3

levels(mydata_13$TRT)
levels(mydata_13$TRT) = c(0,NA,1)  # 0 for Trastuzumab, 1 for Trastuzumab+Lapatinib
levels(mydata_13$RACE)
levels(mydata_13$RACE) = c(1,0,0,0,NA)   # 1 for White, 0 for non-white
levels(mydata_13$ER)
levels(mydata_13$ER) = c(1,0)  # 1 for positive, 0 for negative
levels(mydata_13$LymphNodeInv)
levels(mydata_13$LymphNodeInv) = c(1,0)  # 1 for positive, 0 for negative
levels(mydata_13$HER2IHC)
levels(mydata_13$HER2IHC) = c(0,0,0,1,NA)   # 1 for strong, 0 for weak
levels(mydata_13$Subtype_BX)
levels(mydata_13$Subtype_BX) = c(0,1,0,0)
mydata_13$TRT = as.numeric(as.character((mydata_13$TRT)))
mydata_13$PCRBRNode = as.numeric(as.character((mydata_13$PCRBRNode)))
mydata_13$RACE = as.numeric(as.character((mydata_13$RACE)))
```

```
mydata_13$LymphNodeInv = as.numeric(as.character((mydata_13$LymphNodeInv)))
mydata_13$ER = as.numeric(as.character((mydata_13$ER)))
mydata_13$HER2IHC = as.numeric(as.character((mydata_13$HER2IHC)))
mydata_13$Subtype_BX = as.numeric(as.character(mydata_13$Subtype_BX))

mydata_cm131 =
mydata_13[,c("TRT","PCRBRNode","AGE","RACE","LymphNodeInv","ER","HER2IHC","MCSIZ",
                            "pseudo_ID")]
summary(mydata_cm131$TRT==1)
mydata_cm13 = na.omit(mydata_cm131)
summary(mydata_cm13$TRT==1)
x_cm131 = data.matrix(mydata_cm13[,c(3:9)],rownames.force = NA)
x_cm13 = x_cm131[,1:6]
trt_cm13 = as.factor(mydata_cm13[,c(1)])
levels(trt_cm13) = c("Ctrl", "Trt")   # Trt: Trastuzumab+Lapatinib group; Ctrl:
Lapatinib group
trt_cm13 = as.character(trt_cm13)
y_cm13 = as.numeric(mydata_cm13[,c(2)])

gene_type13 = mydata_13[c(133,137:194,200)]
x_gn131 = data.matrix(gene_type13, rownames.force = NA)
x_gn13 = x_gn131[,1:59]
trt_gn13 = as.factor(mydata_13[,c(4)])
levels(trt_gn13) = c("Ctrl", "Trt")
trt_gn13 = as.character(trt_gn13)
y_gn13 = as.numeric(mydata_13$PCRBRNode)

x_all131 = merge(x_cm131,x_gn131,by = "pseudo_ID")
x_all13 = data.matrix(x_all131[,2:66])

propensity.func.13 <- function(x, trt) 82/(82+94)
```

```
# Find patients subgroups using clinical markers with the lasso

# Since propensity score is constant, augmentation function is not necessary
# fit.subgroup function does not work when there's missing data
set.seed(123)
subgrp13.cm.lasso <- fit.subgroup(x = x_cm13, y = y_cm13, trt = trt_cm13,
                                  propensity.func = propensity.func.13,
                                  loss = "logistic_loss_lasso", method = "weighting",
                                  cutpoint = "median", larger.outcome.better = TRUE,
                                  retcall = TRUE,nfolds = 10)
summary(subgrp13.cm.lasso)

set.seed(123)
subgrp13.cm.lasso0 <- fit.subgroup(x = x_cm13, y = y_cm13, trt = trt_cm13,
                                   propensity.func = propensity.func.13,
                                   loss = "logistic_loss_lasso", method = "weighting",
                                   cutpoint = 0, larger.outcome.better = TRUE,
                                   retcall = TRUE,nfolds = 10)
summary(subgrp13.cm.lasso0)

# Compare received and recommended treatment
received13.trt.lasso.cm = data.frame(subgrp13.cm.lasso$trt.received)
recommended13.trts.lasso.cm = data.frame(subgrp13.cm.lasso$recommended.trts)
compare13.trt.lasso.cm = cbind(received13.trt.lasso.cm, recommended13.trts.lasso.cm)
compare13.trt.lasso.cm

received13.trt.lasso0.cm = data.frame(subgrp13.cm.lasso0$trt.received)
recommended13.trts.lasso0.cm = data.frame(subgrp13.cm.lasso0$recommended.trts)
compare13.trt.lasso0.cm = cbind(received13.trt.lasso0.cm,
recommended13.trts.lasso0.cm)
compare13.trt.lasso0.cm
```

```
# Summarize significant covariates
print(summarize.subgroups(subgrp13.cm.lasso), p.value = 0.05)
print(summarize.subgroups(subgrp13.cm.lasso0), p.value = 0.05)

# Cross validation for clinical markers model using multiple replications method
class(subgrp13.cm.lasso)
validation13.cm.lasso <- validate.subgroup(subgrp13.cm.lasso, B = 100,
                         method = "training_test_replication",train.fraction = 0.75)
                         validation13.cm.lasso

validation13.cm.lasso0 <- validate.subgroup(subgrp13.cm.lasso0, B = 100,
                          method = "training_test_replication",train.fraction = 0.75)
                          validation13.cm.lasso0

# Plots of patient outcomes conditional on treatment
plotCompare(subgrp13.cm.lasso,subgrp13.cm.lasso0,type = "boxplot")
plotCompare(subgrp13.cm.lasso,subgrp13.cm.lasso0,type = "density")
plotCompare(subgrp13.cm.lasso,subgrp13.cm.lasso0,type = "interaction")
plotCompare(subgrp13.cm.lasso,subgrp13.cm.lasso0,type = "conditional")

plotCompare(validation13.cm.lasso,validation13.cm.lasso0,type = "boxplot")
plotCompare(validation13.cm.lasso,validation13.cm.lasso0,type = "density")
plotCompare(validation13.cm.lasso,validation13.cm.lasso0,type = "interaction")
plotCompare(validation13.cm.lasso,validation13.cm.lasso0,type = "conditional")
```

Bibliography

[1] Fisher, B., Fisher, E. R., Costantino, J., Redmond, C., Poisson, R., Bowman, D., . . . Others. (1989). A randomized clinical trial evaluating tamoxifen in the treatment of patients with node-negative breast cancer who have estrogen-Receptor–Positive tumors. The New England Journal of Medicine, 320(8), 479-484. doi:10.1056/NEJM198902233200802

[2] Harris, L., Fritsche, H., Mennel, R., Norton, L., Ravdin, P., Taube, S., . . . Bast, R. C. (2007). American society of clinical oncology 2007 update of recommendations for the use of tumor markers in breast cancer. Journal of Oncology Practice, 3(6), 336-339. doi:10.1200/JOP.07685043

[3] Romond, E. H., Perez, E. A., Bryant, J., Suman, V. J., Geyer, C. E., Davidson, N. E., . . . Wolmark, N. (2005). Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. The New England Journal of Medicine, 353(16), 1673-1684. doi:10.1056/NEJMoa05212

[4] Lesko, L. J. (2007). Personalized medicine: Elusive dream or imminent reality? Clinical Pharmacology & Therapeutics, 81(6), 807-816. doi:10.1038/sj.clpt.6100204

[5] Piquette-Miller, M., & Grant, D. M. (2007). The art and science of personalized medicine. Clinical Pharmacology & Therapeutics, 81(3), 311-315. doi:10.1038/sj.clpt.6100130

[6] Murphy, S. A. (2003). Optimal dynamic treatment regimes. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 65(2), 331-366. doi:10.1111/1467-9868.00389

[7] Qian, M., & Murphy, S. A. (2011). performance guarantees for individualized treatment rules. The Annals of Statistics, 39(2), 1180-1210. doi:10.1214/10-AOS864

[8] Chen, S., Tian, L., Cai, T., & Yu, M. (2017). A general statistical framework for subgroup identification and comparative treatment scoring: A general framework for subgroup identification and scoring. Biometrics, 73(4), 1199-1209. doi:10.1111/biom.12676

[9] Zhou, X., Mayer-Hamblett, N., Khan, U., & Kosorok, M. R. (2017). Residual weighted learning for estimating individualized treatment rules. Journal of the American Statistical Association, 112(517), 169-187. doi:10.1080/01621459.2015.1093947

[10] Huling, J. D., & Yu, M. (2018). Subgroup identification using the personalized package.

[11] Xu, Y., Yu, M., Zhao, Y., Li, Q., Wang, S., & Shao, J. (2015). Regularized outcome weighted subgroup identification for differential treatment effects. Biometrics, 71(3), 645-653. doi:10.1111/biom.12322

[12] Robidoux, A., Prof, Tang, G., PhD, Rastogi, P., MD, Geyer, C. E., Prof, Azar, C. A., MD, Atkins, J. N., MD, . . . Wolmark, N., Prof. (2013). Lapatinib as a component of neoadjuvant therapy for HER2-positive operable breast cancer (NSABP protocol B-41): An open-label, randomised phase 3 trial. Lancet Oncology, the, 14(12), 1183-1192. doi:10.1016/S1470-2045(13)70411-X

[13] Romond, E. H., Perez, E. A., Bryant, J., Suman, V. J., Geyer, C. E., Davidson, N. E., . . . Wolmark, N. (2005). Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. The New England Journal of Medicine, 353(16), 1673-1684. doi:10.1056/NEJMoa052122

[14] Nahta, R., & Esteva, F. J. (2006). HER2 therapy: Molecular mechanisms of trastuzumab resistance. Breast Cancer Research : BCR, 8(6), 215-215. doi:10.1186/bcr1612

[15] Valabrega, G., Montemurro, F., & Aglietta, M. (2007). Trastuzumab: Mechanism of action, resistance and future perspectives in HER2-overexpressing breast cancer. Annals of Oncology, 18(6), 977-984. doi:10.1093/annonc/mdl475

[16] Xia, W., Gerard, C. M., Liu, L., Baudson, N. M., Ory, T. L., & Spector, N. L. (2005). Combining lapatinib (GW572016), a small molecule inhibitor of ErbB1 and ErbB2 tyrosine kinases, with therapeutic anti-ErbB2 antibodies enhances apoptosis of ErbB2-overexpressing breast cancer cells. Oncogene, 24(41), 6213-6221. doi:10.1038/sj.onc.1208774

[17] Wallden, B., Storhoff, J., Nielsen, T., Dowidar, N., Schaper, C., Ferree, S., . . . Parker, J. S. (2015). Development and verification of the PAM50-based prosigna breast cancer gene signature assay. BMC Medical Genomics, 8(1), 54. doi:10.1186/s12920-015-0129-6

[18] Paik, S., Tang, G., Shak, S., Kim, C., Baker, J., Kim, W., . . . Wolmark, N. (2006). Gene expression and benefit of chemotherapy in women with node-negative, estrogen Receptor–Positive breast cancer. Journal of Clinical Oncology, 24(23), 3726-3734. doi:10.1200/JCO.2005.04.7985

[19] Zhao, Y., Zeng, D., Rush, A. J., & Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. Journal of the American Statistical Association, 107(499), 1106-1118. doi:10.1080/01621459.2012.695674

[20] Brower, V. (2014). Adjuvant trastuzumab improves survival in early breast cancer. Lancet Oncology, the, 15(13), e589-e589. doi:10.1016/S1470-2045(14)71110-6

[21] Keerthi, S. S., & Lin, C. (2003). Asymptotic behaviors of support vector machines with gaussian kernel. Neural Computation, 15(7), 1667-1689. doi:10.1162/089976603321891855