5-2014

Patient rule induction method for subgroup identification given censored data.

Patrick James Trainor
University of Louisville

PATIENT RULE INDUCTION METHOD FOR SUBGROUP IDENTIFICATION GIVEN CENSORED DATA

By

Patrick James Trainor
B.S., Seattle University, 2012

A Thesis
Submitted to the Faculty of the
School of Public Health and Information Sciences of the University of Louisville
in Partial Fulfilment of the Requirements
for the Degree of

Master of Science

Department of Biostatistics and Bioinformatics
University of Louisville
Louisville, Kentucky

May 2014

PATIENT RULE INDUCTION METHOD FOR SUBGROUP IDENTIFICATION GIVEN CENSORED DATA

By

Patrick James Trainor
B.S., Seattle University, 2012

A Thesis Approved on

April 10, 2014

by the following Thesis Committee:

Shesh Rai

Kiseop Lee

Maiying Kong

ABSTRACT

PATIENT RULE INDUCTION METHOD FOR SUBGROUP IDENTIFICATION GIVEN CENSORED DATA

Patrick J. Trainor

April 10, 2014

The identification of subgroups in clinical studies is an important aspect of personalized medicine. In order to develop tailored therapeutics, the factors that characterize subgroups with differential prognosis, response to treatment, and incidence of adverse events or toxicities must be elucidated. We present a generalization of a statistical learning algorithm, Patient Rule Induction Method (PRIM), that is well suited for this task given a right-censored time-to-event outcome measure. This algorithm works to recursively partition a covariate space into mutually exclusive boxes that can be utilized to define subgroups. Conceptually the algorithm is similar to classification and regression trees but rather than satisfying the goal of minimizing overall prediction error, PRIM works to find the extrema of the response surface. The algorithm's performance in prognostic subgroup identification is demonstrated with simulation studies and a case study using data from the Framingham Heart Study. We find that the algorithm has much utility as it provides a set of easy to interpret rules that define subgroups with maximal (minimal) survival or differential response to an intervention as measured by a survival outcome.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

The identification of subgroups in clinical studies is a task of critical importance to field of personalized medicine. Tezak et al. (2010) define personalized medicine as "providing the right intervention or therapy, at the right dose, for the right person, at the right time, by understanding the individual's biology." In order to develop tailored therapeutics, an understanding of factors influencing an individual's disease course and response to a treatment must be developed. We define predictive factors as factors that can be used to predict whether an individual will respond to treatment and prognostic factors as those that can be used to estimate an individual's chance of disease recovery or recurrence (NCI, 2014). The study of prognostic factors is a common goal in epidemiology, while the study of predictive factors is mostly conducted in a pharmacological setting. Epidemiological studies, such as the Framingham Heart Study discussed in this paper, furnish many examples in which prognostic factors have been used to identify subgroups with better or worse prognosis than the overall population. However, while it has been well established that there are many factors that cause significant variability in the response to clinical therapeutics and interventions, clinical studies in pharmacology and in a pharmaceutical setting are rarely designed to elucidate such predictive factors. Clinical efficacy and safety studies are typically designed to demonstrate an overall response to a therapeutic or intervention and not to identify subgroups of responders or non-responders. Such variability in responsiveness may lead to sub-par clinical and dosing guidelines, or may lead to the "failure" of a clinical study to demonstrate an overall treatment effect, when subpopulations of responders do exist. Cardiovascular medicine furnishes many

such examples. Activation of the drug clopidogrel has been shown to be dependent on the presence of certain genetic polymorphisms which results in differential responsiveness as measured by the inhibition of platelet reactivity (Luchessi et al., 2013; Frelinger et al., 2013). This phenomenon of differential responsiveness and variability in rates of adverse events has been documented with other cardiovascular drugs such as warfarin, amiodarone, β-blockers, and statins (Johnson, 2013; Janse et al.,1998). The concept of personalized medicine hopes to characterize such differences as a function of genetics, patient factors, family history, concomitant medications, and concomitant diseases.

The prevailing methodology for identifying subgroups using predictive or prognostic factors in clinical studies has been a regression approach. In order to identify factors that influence prognosis, researchers typically build multivariable regression models and then use a significance based approach to identify the covariates that influence a patient's prognosis. Likewise, to identify predictive factors, researchers typically build multivariable models that incorporate treatment by covariate interactions. If a treatment by covariate interaction term is deemed significant, then it is concluded that the covariate is a predictive factor. There are significant flaws with using a regression based approach such as this for subgroup identification. The most fundamental is that the goal of regression modeling is to develop models that fit the data well globally. Yet often a subgroup with better prognosis or enhanced response to treatment is a "bump" in the response surface which may be smoothed out by regression models. Consequently, we desire a methodology for finding bumps in a response surface for prognostic or predictive subgroup identification. In this paper we present a generalization of a statistical learning algorithm, Patient Rule Induction Method (Friedman and Fisher, 1999), that is well suited for such a task. Our modified algorithm has been tailored for finding subgroups given a right censored time-to-event response variable, since this is a common outcome measure in clinical studies.

The layout of this paper is as follows: In section 2 we discuss the original formulation of the PRIM algorithm and why it is ideal for subgroup identification. We then introduce our generalization of the algorithm for prognostic subgroup identification given a right censored time-to-event response variable in section 3. We then present a further generalization so that the algorithm can be used for predictive subgroup identification. Section 4 details a simulation study that demonstrates the suitability of the algorithm for prognostic subgroup identification. In section 5 we demonstrate prognostic subgroup identification using the Framingham Heart Study data. Finally we discuss our results and the algorithms suitability for subgroup identification in section 6.

2 PATIENT RULE INDUCTION METHOD

While prognostic and predictive subgroup identification serve different purposes, the goals of both analyses are equivalent. In prognostic subgroup identification, we want to find the region of the covariate space (the set of all possible covariate values) in which a response variable takes its maximum or minimum values. For example, if the response variable is a time-to-event outcome measure, then prognostic subgroup identification means finding the subgroup which has greatest (or worst) survival. With predictive subgroup identification, we want to find the region of the covariate space in which the difference in response is maximal between study arms. In the context of a time-to-event outcome measure, we are trying to identify the subgroup in which the difference in survival between intervention and control is greatest. In either case, the object of subgroup identification is the identification of the region of the covariate space in which the maxima of a response variable is found. The PRIM algorithm was developed for such a task. The algorithm is similar to regression trees (Segal, 1988) in that the covariate space is recursively partitioned using a binary split criteria. However, while regression trees split nodes using the minimization of the sum of squares between fitted and observed values of the response variable, the PRIM algorithm uses a spit criteria designed to maximize the response variable in one of the child nodes. The final solution given by the PRIM algorithm is a sequential partitioning of the covariate space into disjoint regions with the goal that the response variable takes its maximum in the first partition and then decreases monotonically thereafter. The rules that define the individual partitions then can be used to define subgroups.

In this section, Patient Rule Identification Method for a numerical response variable is presented as was originally detailed by Friedman and Fisher (1999). While they developed PRIM methodology for categorical and continuous response variables, a generalization to right-censored response variables is needed and is presented in this paper. The covariate space can be written as an external direct product of input variables: $S = S_1 \times S_2 \times \ldots \times S_p$. With such a formulation, $x_{i,j} \in S_j$ is the realization of the jth covariate for the ith study subject. Thus, for a clinical study, data will be of the form $\{\mathbf{x}_i, y_i\}$ where $\mathbf{x}_i$ will be a length p vector of covariates for the ith subject and y_i will be a clinical response variable. In general the PRIM algorithm proceeds by two subroutines (top-down peeling and bottom-up pasting). A single iteration of both routines generates a partition of the covariate space with the object of the partitioning being maximization of the response variable.

2.1 TOP-DOWN PEELING

The top down peeling process begins with a p-dimensional hypercube $B_1 = S$ that covers all of the data. The top-down peeling algorithm then proceeds as follows:

> **Generate** sub-boxes $b_{j-} = \{\mathbf{x} : x_j < x_{j\theta}\}$ and $b_{j+} = \{\mathbf{x} : x_j > x_{j(1-\theta)}\}$ for each continuous covariate $j \in \{1, 2, \ldots, p\}$.
>
> **Generate** sub-boxes $b_{j,m} = \{\mathbf{x} : x_j \neq s_m\}$ for each m level of categorical covariate $j \in \{1, 2, \ldots, p\}$.
>
> **Generate** set of sub-boxes $\mathcal{C}_1(b) = \{b_{j-}, b_{j+}, b_{j,m} : j \in \{1, 2, \ldots, p\}\}$.
>
> **Define** $\mathcal{C}_2(b) = \{b : b \in \mathcal{C}_1(b) \text{ and } \omega(B_1 - b) > \omega_0\}$ where $\omega(b) = \frac{1}{n}\sum_{i=1}^{n} 1(\mathbf{x}_i \in b)$.
>
> **If** $\mathcal{C}_2(b) \neq \emptyset$:
>
> > **Choose** $b^* = \arg\max_{b \in \mathcal{C}_2(b)} \text{ave}\{y_i : \mathbf{x}_i \in B_1 \setminus b\}$.

Update $B_1 \leftarrow B_1 \setminus b^*$.

Else Stop.

In Figure 1 we illustrate how candidate sub-boxes are generated given two continuous covariates and a binary response variable. For covariates X_1 and X_2 we generate two sub-boxes, one with the upper θ quantile of the data removed (b_{j+}) and one with the lower θ quantile of the data removed (b_j-). Since the response variable is binary in this case, we select b^* to be removed from B_1 according to the criteria that we want to maximize the proportion of responders in $B_1 \setminus b^*$.

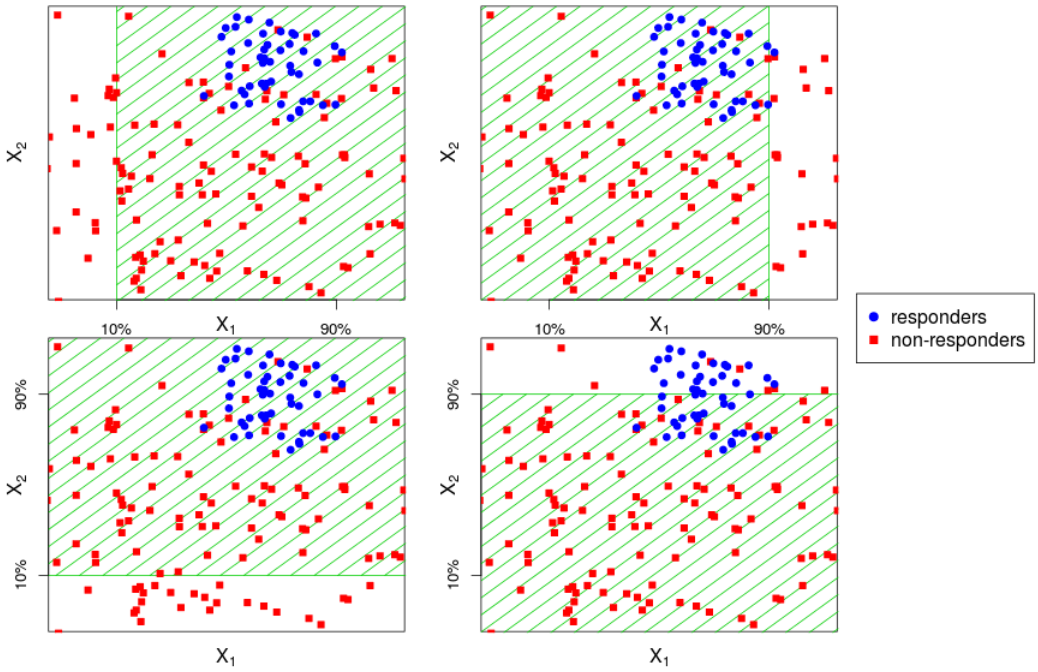

Figure 1: Example candidate sub-boxes for classification problem with two covariates given $\theta = .1$.

After selecting b^*, B_1 is updated, $B_1 \leftarrow B_1 \setminus b^*$, and then the peeling routine begins a new iteration using the updated hypercube. The peeling routine continues until no subsequent peeling steps can be taken without the empirical support of B_1 falling below a pre-specified level ω_0. This iterative procedure is illustrated in Figure 2.

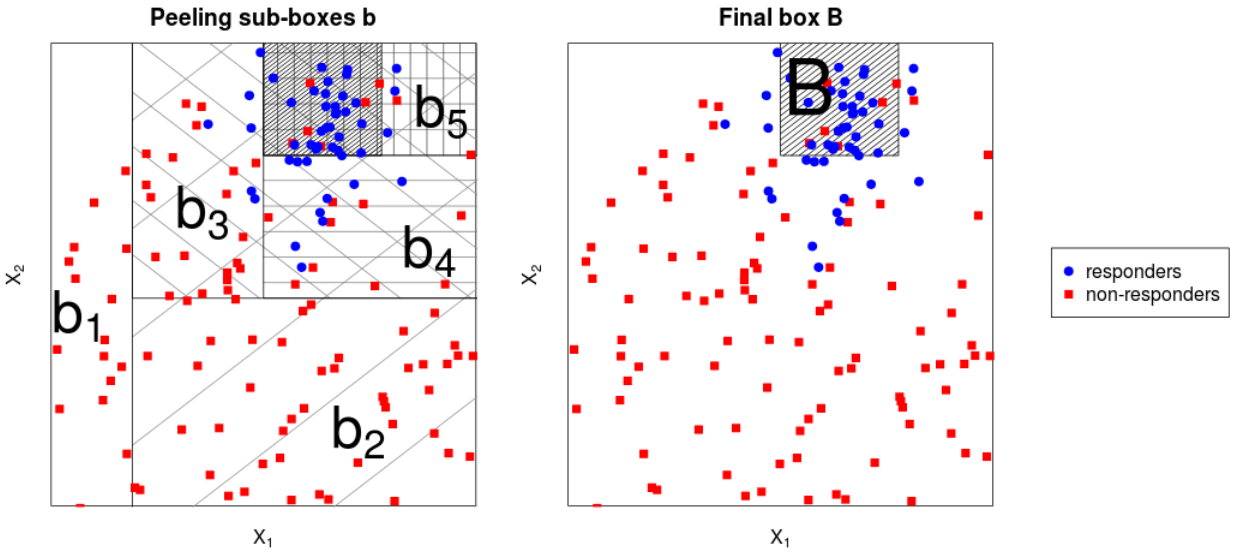

Figure 2: Example peeling for classification problem with two covariates given $\theta = .1$.

2.2 BOTTOM-UP PASTING

In order to improve upon the performance of the top-down peeling algorithm Friedman and Fisher (1999) propose a second algorithm known as bottom-up pasting. This algorithm proceeds as follows:

> **Generate** sub-boxes b_{j-} from the left boundary of B_1 for variable j, to the ϕ quantile to the left, and b_{j+} from the right boundary of B_1 for variable j, to the ϕ quantile to the right for continuous variables $j \in \{1, 2, \ldots, p\}$.
>
> **Generate** sub-boxes $b_{j,m} = \{\mathbf{x} : x_j = s_m\}$ for each previously deleted m level of categorical covariate $j \in \{1, 2, \ldots, p\}$.
>
> **Generate** set of sub-boxes $\mathcal{C}_1(b) = \{b_{j-}, b_{j+}, b_{j,m} : j \in \{1, 2, \ldots, p\}\}$.
>
> **Choose** $b^* = \arg\max_{b \in \mathcal{C}_2(b)} \text{ave}\{y_i : \mathbf{x}_i \in B_1 \cup b\}$.
>
> **If** $\text{ave}\{y_i : \mathbf{x}_i \in B_1 \cup b^*\} >> \text{ave}\{y_i : \mathbf{x}_i \in B_1\}$:

Update $B_1 \leftarrow B_1 \cup b^*$

Else Stop.

An example pasting routine is depicted in Figure 3.

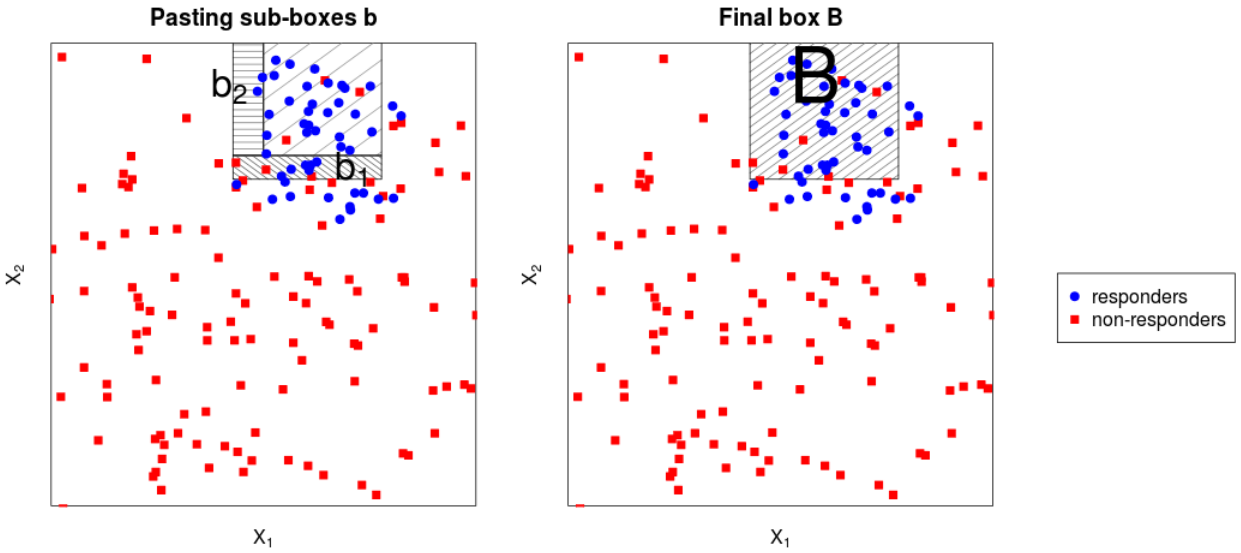

Figure 3: Example pasting for classification problem with two covariates.

2.3 COVERING

After the first iteration of the peeling and pasting procedures, the result is a hypercube $B \subsetneq S$ that was constructed with the goal of optimizing the mean of the response variable y. The peeling and pasting routine is then repeated on the remaining data belonging to $S \backslash B_1$, yielding a new hypercube B_2. In this way a sequence of hypercubes $\{B_1, B_2, \ldots, B_K\}$ is generated that partitions the covariate space and ideally has the property that $\text{ave}\{y_i | \mathbf{x}_i \in B_1\} > \text{ave}\{y_i | \mathbf{x}_i \in B_2\} > \ldots > \text{ave}\{y_i | \mathbf{x}_i \in B_K\}$.

2.4 USING PRIM FOR SUBGROUP IDENTIFICATION

Provided that the sequence of hypercubes has the property that the mean of the response variable is decreasing, then B_1 is an estimate of the region of the covariate

space where the response variable takes its maximum values. As a concrete example, if the response variable was level of low-density lipoprotein (LDL) then we can use the rules that define the hypercube B_1 in order to estimate the subgroup with the highest levels of LDL. Since the rules were generated by an algorithm that iteratively shaves off and possibly adjoins quantiles for the covariates, the rules that define B_1 are cleanly specified. Using the LDL example, the rules defining membership in B_1 might be: $\{i : \text{sex}_i = \text{male}, 55 < \text{age}_i < 65, \text{BMI}_i > 27.8, \text{SBP}_i > 150\}$.

3 GENERALIZATION OF PRIM TO RIGHT-CENSORED DATA

The original formulation of the PRIM algorithm is well suited for either numerical or categorical data. Given numerical data, the binary spit criteria utilized is the maximization of the response variable estimated by calculating means in the child nodes. For categorical data the binary split criteria is the maximization of the probability of class membership estimated by calculating proportions in the child nodes. Given instead a right-censored time-to-event response variable the study data for the ith subject consists of $\{\boldsymbol{X}_i, (t_i, \delta_i)\}$ where δ_i is a censoring indicator. To generalize the PRIM algorithm to this type of data a suitable split criteria must be selected. We propose using the Tarone-Ware class of test statistics (Tarone and Ware, 1977; Kalbfleisch and Prentice, 2002) as such a split criteria.

3.1 TARONE-WARE STATISTICS

In this section we present a brief discussion of the Tarone-Ware class of statistics. To test for the equality of r failure distributions, $F_1(t), F_2(t), \ldots, F_r(t)$, the failure times for the sample pooled over the r strata are ordered: $t_1 < t_2 < \ldots < t_k$. At time t_j, d_j denotes the number of failures for the pooled sample and n_j denotes the total number of subjects from the pooled sample at risk at immediately prior to t_j. Further, $d_{i,j}$ and $n_{i,j}$ denote the number of failures and number at risk priort to t_j in the ith strata where $i \in \{1, 2, \ldots, r\}$. Then for each failure time in the pooled sample the data can be arranged into $2 \times r$ contingency tables:

	strata 1	strata 2	...	strata r	
failures	$d_{1,j}$	$d_{2,j}$	...	$d_{r,j}$	d_j
survivors	$n_{1,j} - d_{1,j}$	$n_{2,j} - d_{2,j}$	...	$n_{r,j} - d_{r,j}$	$n_j - d_j$
at risk	$n_{1,j}$	$n_{2,j}$	...	$n_{r,j}$	n_j

With this formulation, the rank statistics form a r length vector, $\mathbf{v} = (v_1, v_2, \ldots v_r)^T$, with entries given by:

$$v_i = \sum_{j=1}^{k} W(t_j) \left(d_{i,j} - \frac{n_{i,j} d_j}{n_j} \right)$$

where $W(t_i)$ is a non-negative weight function described in the preceding paragraph. The estimated covariance matrix of the rank statistics has the following entries:

$$V_{i,i} = \sum_{j=1}^{k} \left\{ W^2(t_j) \frac{n_{i,j}(n_j - n_{i,j}) d_j (n_j - d_j)}{n_j^2 (n_j - 1)} \right\}$$
$$V_{i,l} = \sum_{j=1}^{k} \left\{ W^2(t_j) \frac{-n_{i,j} n_{l,j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)} \right\}.$$

To use a statistic from the Tarone-Ware class of statistics test to compare the survival of the jth group with the lth group, $j \neq l$, the following test statistic can be constructed (Klein and Moeschberger, 1997):

$$Z_{j,l} = \frac{(v_j - v_l)}{\sqrt{V_{jj} + V_{ll} - 2V_{jl}}}.$$

This test statistic follows a $N(0, 1)$ distribution.

3.2 TARONE-WARE STATISTICS AS PRIM BINARY SPLIT CRITERIA

In our simulation studies we used the weight function $W(t_i) = 1$, which reduces the test statistic to the well known log-rank test statistic. To use PRIM with such a split

criteria, we propose the following top-down peeling algorithm:

Generate sub-boxes $b_{j-} = \{\mathbf{x} : x_j < x_{j\theta}\}$ and $b_{j+} = \{\mathbf{x} : x_j > x_{j(1-\theta)}\}$ for each continuous covariate $j \in \{1, 2, \ldots, p\}$.

Generate sub-boxes $b_{j,m} = \{\mathbf{x} : x_j \neq s_m\}$ for each m level of categorical covariate $j \in \{1, 2, \ldots, p\}$.

Generate set of sub-boxes $\mathcal{C}_1(b) = \{b_{j-}, b_{j+}, b_{j,m} : j \in \{1, 2, \ldots, p\}\}$.

Define $\mathcal{C}_2(b) = \{b : b \in \mathcal{C}_1(b) \text{ and } \omega(B_1 - b) > \omega_0\}$ where $\omega(b) = \frac{1}{n}\sum_{i=1}^{n} 1(\mathbf{x}_i \in b)$.

If $\mathcal{C}_2(b) \neq \emptyset$:

Compute $Z_{B_1 \setminus b, B_1}$ for all sub-boxes b.

Choose $b^* = \arg\max_{b \in \mathcal{C}_2(b)} Z_{B_1 \setminus b, B_1}$.

Update $B_1 \leftarrow B_1 \setminus b^*$.

Else Stop.

For the bottom-up pasting routine, we proceed similarly–the distinction between a continuous response variable and a right-censored time-to-event variable being computing $Z_{B_1 \cup b, B_1}$ for all sub-boxes b and then pasting on the sub-box b^* to B_1 if it results in greater survival.

3.3 PROGNOSTIC VS. PREDICTIVE SUBGROUP IDENTIFICATION

As detailed thus far, using a Tarone-Ware statistic for a split criteria adapts PRIM for finding subgroups with maximal survival in a clinical study. This is the goal of prognostic subgroup identification. However, to adapt PRIM for predictive subgroup

identification with right-censored time-to-event variables, we need to modify the split criteria even further. In this case we are searching for the region of the covariate space in which the difference in survival between a treatment group and a control group is maximal.

Generate sub-boxes $b_{j-} = \{\mathbf{x} : x_j < x_{j\theta}\}$ and $b_{j+} = \{\mathbf{x} : x_j > x_{j(1-\theta)}\}$ for each continuous covariate $j \in \{1, 2, \ldots, p\}$.

Generate sub-boxes $b_{j,m} = \{\mathbf{x} : x_j \neq s_m\}$ for each m level of categorical covariate $j \in \{1, 2, \ldots, p\}$.

Generate set of sub-boxes $\mathcal{C}_1(b) = \{b_{j-}, b_{j+}, b_{j,m} : j \in \{1, 2, \ldots, p\}\}$.

Define $\mathcal{C}_2(b) = \{b : b \in \mathcal{C}_1(b) \text{ and } \omega(B_1 - b) > \omega_0\}$ where $\omega(b) = \frac{1}{n}\sum_{i=1}^{n} 1(\mathbf{x}_i \in b)$.

If $\mathcal{C}_2(b) \neq \emptyset$:

Compute $Z_{treatment,control}$ for all sub-boxes b.

Choose $b^* = \arg\max_{b \in \mathcal{C}_2(b)} Z_{treatment,control}$.

Update $B_1 \leftarrow B_1 \setminus b^*$.

Else Stop.

4 SIMULATION STUDIES

In order to evaluate the PRIM algorithm's ability to find subgroups with best prognosis in a clinical study, we conducted simulation studies. For each of the simulated study subjects we assumed both failure time and censoring time followed an exponential distribution, i.e. $T \sim \exp(\lambda)$ and $C \sim \exp(\xi)$. Then $\log(T) \equiv -\log(\lambda) + W$, where W has an extreme value distribution with pdf $f(w) = \exp(w - e^w)$. So we assume that each subject's failure time has the same error distribution W. To create proportional hazards we define $\lambda(\mathbf{X}) = \lambda_0 \exp(\mathbf{X}\boldsymbol{\beta})$, where $\mathbf{X}_i = (X_{1,i}, X_{2,i}, \ldots, X_{i,p})$, a vector of covariate realizations for the ith subject and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ denotes pre-specified effects. We set out to evaluate weather the rules defining the hypercube after one iteration of the PRIM algorithm behave in a predictable fashion. We hypothesized that algorithm's sensitivity and specificity in identifying subgroups with best prognosis would increase with sample size, and decrease with increasing scale of the error distribution. For $n = 50, 100, 200$ and 1000 and $\sigma = 0.1, 0.5, 1.0$ and 2.0 we generated study data using the following procedure:

1. **Generate** random w_i such that W follows an extreme value distribution with scale parameter σ.

2. **Generate** random $X_{1,i}$ and $X_{2,i}$ so that each X follows a standard normal distribution.

3. **Specify** effects $\boldsymbol{\beta} = (1, 1)$ and baseline hazard rate $\lambda_0 = 1/10$.

4. **Compute** failure time t_i for ith subject using $\log(t_i) = -\log\left[\lambda_0 \exp(\mathbf{X}_i\boldsymbol{\beta})\right] + w_i$.

5. **Generate** random censoring time c_i for the ith subject subject such that $C \sim \exp(1/20)$.

We then have data of the form $\{\boldsymbol{X}_i, (t_i, \delta_i)\}$. The peeling subroutine of the PRIM algorithm was then conducted using algorithm parameters $\theta = 0.15$ and $\omega_0 = 0.1$. At the conclusion of the peeling subroutine we then should have a 2-dimensional box B_1 that contains about $1/10$ of the original data. Since the hazard for the ith subject is specified by $\lambda_0 \exp(\mathbf{X}_i\boldsymbol{\beta})$, we expect that since $\mathbf{X}\beta \sim N(0, 2)$, if a subject has $X_1 + X_2$ less than the 10% percentile of $N(0, 2)$ then that subject should belong to B_1.

4.1 SIMULATION STUDY RESULTS

Results of the prognostic identification simulation study are shown in Table 1. As expected, the algorithm's sensitivity and specificity are inversely related to the scale of the extreme value error distribution. For error distribution scale parameters $\sigma = 2, 1, 0.5$, it is confirmed that the algorithm's sensitivity does increase with increasing sample size. However, we do not observe this effect as the effect of the error distribution on failure time becomes negligible ($\sigma = 0.1$).

Table 1: Prognostic identification simulation study

		Sensitivity		Specificity	
Error dist. σ	N	mean	sd	mean	sd
2	50	0.339	0.287	0.900	0.032
	100	0.439	0.258	0.915	0.029
	200	0.522	0.212	0.930	0.024
	1000	0.633	0.084	0.958	0.009
1	50	0.492	0.298	0.907	0.037
	100	0.572	0.218	0.928	0.024
	200	0.633	0.150	0.942	0.017
	1000	0.664	0.049	0.962	0.005
0.5	50	0.636	0.284	0.903	0.046
	100	0.664	0.187	0.932	0.024
	200	0.682	0.118	0.947	0.014
	1000	0.674	0.042	0.963	0.005
0.1	50	0.766	0.216	0.889	0.062
	100	0.755	0.151	0.925	0.034
	200	0.727	0.100	0.945	0.018
	1000	0.680	0.040	0.963	0.004

One possible explanation warranting exploration for this effect is that it is due to the algorithm approximating level sets of a response variable with hyper-cubes instead of geometric shapes that conform to the response variable surface. Figures 4 and 5 illustrate this fact. For both figures 1000 observations were randomly generated such that two covariates X and Y follow a standard normal distribution. The observations are plotted with their sum on a third axis. A blue plane is added at the 85% of the sum $X+Y$. A theoretical PRIM peel solution is shown as a salmon plane. Figure 5 is

a rotation of Figure 4 illustrating the projection of the data above the 85% percentile onto the theoretical PRIM solution. As can be seen, a portion of the data above the 85th quantile fails to be captured in the PRIM solution, leading to diminished sensitivity.

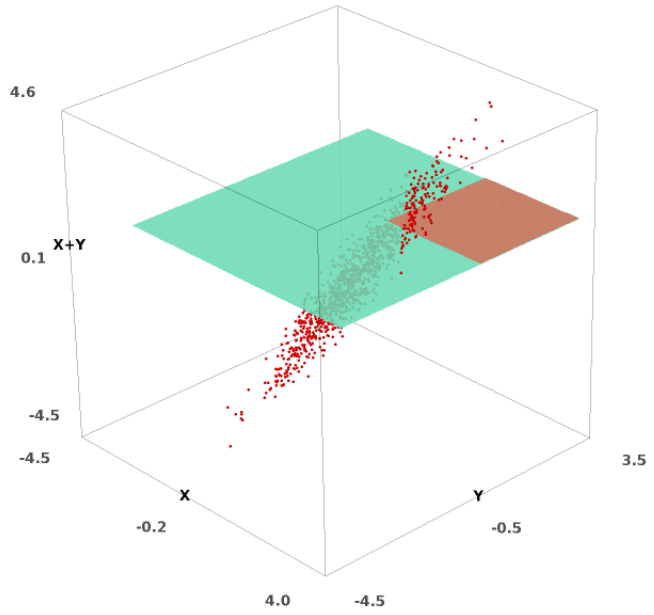

Figure 4: Theoretical PRIM solution to 85th quantile of $X + Y$ approximation.

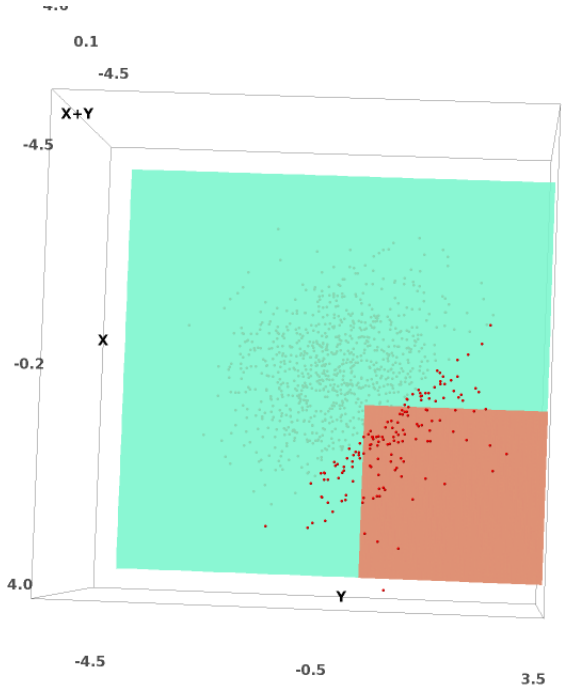

Figure 5: View of $X + Y$ projected onto the theoretical PRIM solution.

5 FRAMINGHAM CASE STUDY

We utilized the PRIM algorithm for prognostic subgroup identification using a subset of the Framingham Heart Study provided as a public use teaching dataset by the National Heart, Lung, and Blood Institute. This renowned study has been carried on for over 65 years and has contributed greatly to the understanding of cardiovascular diseases and their risk factors (Mahmood et al., 2013). The data utilized consisted of potential cardiovascular risk factors, disease markers, and time-to-event outcomes for 4,434 study subjects. We applied the PRIM algorithm peeling subroutine to the data to induce 5 boxes consisting of roughly half of the data. The response variable was the right censored time from first study examination to myocardial infarction. The covariates we considered which all represent measurements taken at the first study examination were:

- Sex (Male/Female)
- Total serum cholesterol (mg/dL)
- Age
- Systolic blood pressure (mmHg)
- Diastolic blood pressure (mmHg)
- Current smoker (Yes/No)
- Body mass index
- Currently prescribed blood pressure medications (Yes/No)

- Heart Rate
- Serum glucose (mg/dL)
- History of coronary heart disease (Yes/No)

We applied the algorithm with parameters $\theta = 0.15$ and $\omega_0 = \frac{1}{11-k}$, where k indexes the sequence of boxes. In this way, each of the boxes should have roughly 1/10 of the total data. We then can use the rules defining each of the boxes to define subgroups, for which we hypothesize that measures of survival will decrease monotonically with the index of the boxes or equivalently the subgroups defined by the boxes.

5.1 RESULTS OF FRAMINGHAM CASE STUDY

The results of applying the PRIM algorithm's peeling subroutine to the Framingham data are shown in Table 2 through Table 8. In Table 2, the covariates used in the analysis and their contributions to the first three boxes induced by the algorithm are presented. In the rightmost column the covariate is listed, followed by the covariate type, (Factor=F, Numeric=Num), at which iteration during the peeling routine the covariate was used as a split criteria, and the factor level or quantile that was deleted. For example, from the table we see that in Box 1 "Previous CHD" is the first split criteria utilized and that the factor level deleted is "Yes". The next split criteria used in the construction of Box 1 is the covariate "Sex" for which the level deleted is "Male". The first numeric covariate used as a split criteria (used as the fourth split) is "Age" for which the upper quantile was removed. This covariate is again used as a split criteria for the 8th split–again removing the upper quantile. Dashes are used to denote that a covariate was not used as a split criteria in the construction of the box.

Table 2: Framingham box induction steps

Variable	**Type**	**Box 1**	**Deleted**	**Box 2**	**Deleted**	**Box 3**	**Deleted**
Sex	F	2	Male	2	Male	2	Male
Total Cholesterol	Num	6, 7	Upper	6	Upper	4	Upper
Age	Num	4, 8	Upper	4	Upper	-	-
Systolic BP	Num	5	Upper	5, 7, 8, 10, 11	Upper	5	Upper
Diastolic BP	Num	-	-	-	-	-	-
Current Smoker	F	9	Yes	-	-	7	Yes
BMI	Num	-	-	-	-	-	-
BP Meds	F	3	Yes	3	Yes	3	Yes
Heart Rate	Num	-	-	-	-	-	-
Serum Glucose	Num	-	-	9	Upper	6	Upper
Previous CHD	F	1	Yes	1	Yes	1	Yes

Table 3 then illustrates the rules defining the first box induced by the algorithm. In this table, "Old Range" corresponds to the acceptable range of covariate values before this box is induced. "New Range" corresponds to the new range of values after the box has been induced. Covariates never used as split criteria have their ranges omitted for interpretablity. A study subject has membership in this box if the following requirements are met: Female, non-smoker, no history of coronary heart disease, not taking blood pressure medication, total serum cholesterol between 135 and 253 mg/dL, age between 32 and 52 years old at enrolment, and systolic blood pressure between 83.5 and 149.0 mmHg. Using this information we can define a subgroup that we hypothesize will have the "best" prognosis in terms of survival.

Table 3: Framingham box 1

Variable	**Type**	**Old Range**	**New Range**
Sex	F	M, F	F
Total Cholesterol	Num	113-696	135-253
Age	Num	32-70	32-52
Systolic BP	Num	83.5-295.0	83.5-149.0
Diastolic BP	Num	-	-
Current Smoker	F	Y, N	N
BMI	Num	-	-
BP Meds	F	Y, N	N
Heart Rate	Num	-	-
Serum Glucose	Num	-	-
Previous CHD	F	Y, N	N

Tables 4-5 provide rules defining the next two subgroups. We hypothesize that survival will decrease monotonically with each following box induced.

Table 4: Framingham box 2

Number	Variable	Type	Old Range	New Range
1	Sex	F	M, F	F
2	Total Cholesterol	Num	113-696	135-287
3	Age	Num	32-70	32-61
4	Systolic BP	Num	83.5-295.0	83.5-124.0
5	Diastolic BP	Num	-	-
6	Current Smoker	F	-	-
7	BMI	Num	-	-
8	BP Meds	F	Y, N	N
9	Heart Rate	Num	-	-
10	Serum Glucose	Num	40-394	47-89
11	Previous CHD	F	Y, N	N

Table 5: Framingham box 3

Number	Variable	Type	Old Range	New Range
1	Sex	F	M, F	F
2	Total Cholesterol	Num	113-696	143-302
3	Age	Num	-	-
4	Systolic BP	Num	83.5-295.0	85.5-164.0
5	Diastolic BP	Num	-	-
6	Current Smoker	F	Y, N	N
7	BMI	Num	-	-
8	BP Meds	F	Y, N	N
9	Heart Rate	Num	-	-
10	Serum Glucose	Num	40-394	45-97
11	Previous CHD	F	Y, N	N

In Table 6 the first five boxes induced by the algorithm (covering roughly half of the data) are compared using a log-rank test. As hypothesized the observed number of events is monotonically increasing. A three way comparison of the first three boxes and a two way comparison of just the first two boxes is shown in Tables 7 and 8, respectively.

Table 6: Comparison of boxes 1-5

Box	**N**	**Observed**	**Expected**	**p-value**
1	409	4	30.6	$2.24 * 10^{-17}$
2	412	8	29.8	
3	409	27	27	
4	406	44	26	
5	421	59	28.5	

Table 7: Comparison of boxes 1-3

Box	**N**	**Observed**	**Expected**	**p-value**
1	409	4	13.7	$8.69 * 10^{-7}$
2	412	8	13.3	
3	409	27	12	

Table 8: Comparison of boxes 1 and 2

Box	**N**	**Observed**	**Expected**	**p-value**
1	409	4	6.08	0.23
2	412	8	5.92	

6 DISCUSSION

In this paper we present how the PRIM algorithm can be used to identify subgroups in clinical studies given a right-censored time-to-event response variable. We have demonstrated how to apply this algorithm for prognostic subgroup identification, and have explained how the algorithm could be further extended for use in predictive subgroup identification.

Some of the advantages of using the PRIM algorithm for subgroup identification are readily apparent from the simulation studies and Framingham case study. First, the algorithm is well suited for the task of subgroup identification because it is designed to find regions of the covariate space in which a response variable has its extrema. This is an important concept as this is usually the goal of subgroup identification–to find the subpopulation with the best (worst) prognosis, best (worse) response to treatment, highest (lowest) rates of adverse events or toxicities. A regression approach has the flaw that regression models may fit the data well globally, while simultaneously "smoothing" out the bumps in the response surface that PRIM finds. The Framingham case study illustrates that the rules defining a box induced by the algorithm readily translate into clinically meaningful subgroup definitions.

The PRIM algorithm is not free of disadvantages, however. The greatest disadvantage of the algorithm is that it is a greedy algorithm. Greedy algorithms make the "best" local choice at each iteration, which may yield a final solution that is sub-optimal–for a good discussion of this problem see Gutin et al. (2002). One promising solution to this problem would be to build an ensemble learner using an aggregation of bootstrapped PRIM models. However, such an ensemble model would

obscure the interpretability of the rules defining boxes, making defining subgroups more challenging. Additionally, the PRIM algorithm suffers from nominal covariates dominating splits if large proportions of the data share the same factor level. For example, in all three of the first induced Framingham boxes, "Sex" is used as the second split criteria. In this instance, the candidate sub-boxes for removal for the numerical covariates such as "Age" are based on removing 15% of the data (since $\theta = 0.15$ in our analysis), while the candidate sub-boxes for removal using the "Sex" covariate contain roughly 50% of the data. Thus, the test statistics for comparing the candidate sub-boxes based on factor levels of "Sex" are likely to be greater than for those for "Age" by virtue of the choice of the algorithm parameter θ. One possible way solution to this problem would be to build an ensemble learner that restricts the number of covariates considered at each split. This approach would be similar to the construction of random forests (Hastie et al., 2008).

REFERENCES

Abu-Hanna, A., Nannings, B., Dongelmans, D., and Hasman, A. (2010). Prim versus cart in subgroup discovery: When patience is harmful. *Journal of Biomedical Informatics*, 43:701–708.

Dyson, G., Frikke-Schmidt, R., Nordestgaard, B., Tybjaerg-Hansen, A., and Sing, C. (2007). An application of the patient rule-induction method for evaluating the contribution of apolipoprotein e and lipoprotein lipase genes to predicting ischemic heart disease. *Genetic Epidemiology*, 31:515–527.

Frelinger, A., Bhatt, D., Lee, R., Mulford, D., Wu, J., Nudurupati, S., Nigam, S., Lampa, M., Brooks, J., Barnard, M., and Michelson, A. (2013). Clopidogrel pharmacokinetics and pharmacodynamics vary widely despite exclusion or control of polymorphisms, noncompliance, diet, smoking, co-medications, and pre-existent variability in platelet function. *Journal of the American College of Cardiology*, 61:872–879.

Friedman, J. and Fisher, N. (1999). Bump hunting in high-dimensional data. *Statistics and Computing*, 9:123–143.

Gutin, G., Yeo, A., and Zverovich, A. (2002). Traveling salesman should not be greedy: domination analysis of greedy-type heuristics for the tsp. *Discrete Applied Mathematics*, 117:81–86.

Harrington, D. and Fleming, T. (1982). A class of rank test procedures for censored survival data. *Biometrika*, 69:553–566.

Hastie, T., Tibshirani, R., and Friedman, J. (2008). *The Elements of Statistical Learning*. Springer.

Janse, M., Malik, M., and Camm, A. (1998). Identification of post acute myocardial infarction patients with potential benefit from prophylactic treatment with amiodarone. *European Heart Journal*, 19:85–95.

Julian, D., Camm, A., Frangin, G., Janse, M., Munoz, A., Schwartz, P., and Simon, P. (1997). Randomised trial of effect of amiodarone on mortality in patients with left-ventricular dysfunction after recent myocardial infarction: Emiat. *The Lancet*, 349:667–674.

Kalbfleisch, J. and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data*. John Wiley and Sons.

Kehl, V. and Ulm, K. (2004). Responder identification in clinical trials with censored data. *Journal of Biomedical Informatics*, 50:1338–1355.

Klein, J. and Moeschberger, M. (1997). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer.

Lipkovich, I., Dmitrienko, A., Denne, J., and Enas, G. (2011). Subgroup identification based on differential effect search: A recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*, 30:2601–2621.

Luchessi, A., Silbiger, V., Hirata, R., Lima-Neto, L., Cavichioli, D., Iniguez, A., Bravo, M., Bastos, G., Sousa, A., Brion, M., Carracedo, A., and Hirata, M. (2013). Pharmacogenomics of anti-platelet therapy focused on peripheral blood cells of coronary arterial disease patients. *Clinica Chimica Acta*, 425:9–17.

Mahmood, S., Levy, D., Vasan, R., and Wang, T. (2013). The framingham heart study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet*, pages 999–1008.

Nannings, B., Abu-Hanna, A., and Jonge, E. (2008). Applying prim (patient rule induction method) and logistic regression for selecting high-risk subgroups in very-elderly icu patients. *International Journal of Medical Informatics*, 77:272–279.

National Cancer Institute (2014). *NCI Dictionary of Cancer Terms*.

Peto, R. and Peto, J. (2010). Asymptotically efficient rank-invariant test procedures. *Journal of the Royal Statistical Society*, 135:185–198.

Polonik, W. and Wang, Z. (2010). Prim analysis. *Journal of Multivariate Analysis*, 101:525–540.

Segal, M. (1988). Regression trees for censored data. *Biometrics*, 44:35–47.

Tarone, R. and Ware, J. (1977). On distribution-free tests for equality of survival distributions. *Biometrika*, 64:156–160.

Tezak, Z., Kondratovich, M., and E., M. (2010). US FDA and personalized medicine: In vitro diagnositc regulatory perspective. *Personalized Medicine*, 7:517–530.

CURRICULUM VITA

Patrick Trainor
936 S. 1st St.
Louisville, Kentucky 40203
(775) 830-7511

OBJECTIVE

My objectives are to earn a PhD in biostatistics with a focus in bioinformatics and clinical informatics, contribute to interdisciplinary applied statistics projects, and to conduct original research in the fields of statistical learning and clinical trials methodology.

EDUCATION

BS Mathematics
Seattle University, Seattle, WA, 2012

MS Biostatistics
University of Louisville, Louisville, KY expected May 2014

MA Mathematics
University of Louisville, Louisville, KY expected August 2014

RESEARCH EXPERIENCE

Graduate Assistant (Volunteer) 2012-Current
Biostatistics Shared Facility, J.G. Brown Cancer Center, Louisville, KY
Projects I have collaborated on:

- Development of predictive models for adverse surgical outcomes in total laparopscopic hysterectomies.
- Current project #1: PCR-based detection of HPV types in benign oral mucosal lesions.
- Current project #2: A comparison of Saphenous Vein Graft Intervention with and without the use of Distal Embolic Protection Device.

Consultant & Intern 2012-2013
Theravance Inc.

- Design and simulation of dedicated cardiovascular safety studies with time to event endpoints.
- Developed national incidence estimates of paralytic ileus and conducted an analysis of comorbid diagnoses and procedures.
- Other applied statistical projects as needed.

Undergraduate Research Assistant 2010-2012
Seattle University

- Conducted a retrospective spatial-statistical analysis of a Jeffrey Pine beetle epidemic that took place in the Lake Tahoe basin region from 1991-1996.
- Developed generalized linear models for predicting the probability of mortality in Jeffrey Pine trees during a pine beetle epidemic.

TEACHING EXPERIENCE

Graduate Teaching Assistant 2012-Current
University of Louisville

Math Laboratory Assistant 2009-2011
Seattle University

OTHER PROFESSIONAL EXPERIENCE

Financial Analyst 2012
Centennial Mortgage, Inc., Seattle, WA
- Developed relational databases for financial data.
- Developed software for preliminary underwriting analysis for multifamily and healthcare mortgage refinances.

Squad Leader 2007-2013
4th Anti-Terrorism Battalion
United States Marine Corps (Reserve)

COMPUTING EXPERIENCE

Programming Languages: Python, Ruby, PHP, R, MATLAB/Octave, SQL, Visual Basic
Statistical Software: SAS, R, SPSS
Bioinformatics Software: Bioconductor
Mathematical Software: SAGE, MATLAB/Octave, Mathematica
Document Markeup: LaTeX
Database: PostgreSQL, MySQL, Microsoft Access
Operating Systems: UNIX, various flavors of Linux, and Windows

AWARDS

- 1st Place Graduate Research Competition in Mathematics, 2013 Kentucky Academy of Science Meeting
- Janet E. Mills Award for Undergraduate Research in Mathematics, Seattle University

TALKS GIVEN

Patient Rule Induction Method for Identifying Subgroups in Clinical Studies. Kentucky Academy of Science Annual Meeting, Morehead, Kentucky, 2013.

Statistical Considerations for Dedicated Cardiovascular Safety Studies. Theravance, Inc. seminar, San Francisco, California, 2013.

Regression Models for Mortality in Environmental Science. Northwest Undergraduate Mathematics Symposium, Portland, Oregon, 2012.

RELEVANT GRADUATE LEVEL COURSES

- Advanced Probability Theory
- Advanced Mathematical Statistics
- Biostatistical Methods I/II
- Categorical Data Analysis
- Survival Analysis
- Statistical Programming
- Introduction to Clinical Trials
- Clinical Trials Statistics Laboratory
- Independent Study in Bioinformatics
- Independent Study in Statistical Learning
- Differential Equations and Dynamical Systems
- Abstract Algebra
- Real Analysis