COMBINATORIAL PATTERN-BASED SURVIVAL ANALYSIS WITH APPLICATIONS IN BIOLOGY AND MEDICINE

by

ANUPAMA RAJASEKHARA REDDY

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Operations Research

Written under the direction of

Peter L. Hammer and Endre Boros

And approved by

_____

_____

_____

_____

_____

New Brunswick, New Jersey

[October, 2009]

ABSTRACT OF THE DISSERTATION

COMBINATORIAL PATTERN-BASED SURVIVAL ANALYSIS WITH APPLICATIONS IN BIOLOGY AND MEDICINE

By Anupama Rajasekhara Reddy

Dissertation Directors:
Peter L. Hammer and Endre Boros

In the current era of targeted therapies and personalized medicine, survival analysis (predicting survival time of patients) is a very important problem. Survival analysis is similar to regression except for the presence of censored observations (observations with incomplete survival time information). We propose to use a combinatorial pattern-based methodology, Logical Analysis of Data (LAD), for survival analysis. LAD is a two-class classification method. In this thesis we extend LAD for survival analysis in various ways. Our first approach is to define high- and low-risk patients, and reduce the problem to two-class classification. This approach is particularly useful for datasets with a large number of samples, and small number of features. In datasets where the feature space is high-dimensional (for example, gene expression data), we first used an unsupervised clustering approach to identify robust clusters in the data, the hypothesis being that the different clusters are associated with different survival profiles. We present a linear programming model to predict survival. Finally, we develop a new method, Logical Analysis of Survival Data (LASD), and validate it on a kidney cancer dataset. Ensemble methods are presented to improve the robustness of LASD.

Acknowledgement

I am very grateful to Dr. Peter Hammer for introducing me to the exciting field of data mining, and in particular to Logical Analysis of Data. He has had a very strong influence on me, and I feel honored to be one of his students. I thank Dr. Endre Boros for being my advisor and for supporting, and guiding me all throughout my Ph.D. I am grateful to Dr. Gabriela Alexe for immediately accepting to be my mentor after Dr. Hammer's tragic death. Thanks Gabriela for launching me into the interesting field of Bioinformatics, and for being an encyclopedia of knowledge! I am very thankful to Dr. Gyan Bhanot for truly *educating* me. Thank you for mentoring me in various very exciting projects, and also for your encouragement and friendship.

I am indebted to Dan Stratila and M. K. Jeong for agreeing to be on my thesis committee. I also thank Stanley Hazen, Marie-Luise Brennan, Kimryn Rathmell, Rose Brannon, Shridar Ganesan, Honghui Wang, Tiberius Bonates, Chris Huang, Huiqing Liu, Sandor Szalma, Joseph Irgon, Michael Seiler, Erhan Bilal, and Louis-Philippe Kronek, for collaborations on the various projects described in this thesis. I especially thank Tiberius, Marie, Rose and Louis-Philippe for their friendship.

I would also like to thank my friends and colleagues: Malvika Surendra, Vimla Gulabani, Shilpa Shanbhag, Bijita Majumdar, Noam and Michal Goldberg, Katie D'Agosta, Clare Smietana, and Terry Hart.

I am very lucky to have such a wonderful family. My parents, Meera and Rajasekhara and brother Kartik have supported and encouraged me throughout my graduate life, and most of all given me a lot of love and affection. I thank my in-laws Anikó, József, Melinda and Gergely Papp for their unconditional love and support.

Lastly, but mostly I am grateful to my husband Dávid for his endless love, patience and motivation. Thank you for listening to my research ideas, for giving me useful feedback, and also for proof reading my thesis. I do not have words to thank you enough.

Table of contents

List of tables

List of illustrations

Chapter 1

Introduction

In this thesis we study survival analysis in medicine and biology using a pattern-based methodology called Logical Analysis of Data (LAD). Survival analysis involves predicting survival time for patients based on a set of variables. This is a very important problem for the biomedical community, especially in the current era of targeted therapies and personalized medicine. Logical Analysis of Data is a rule-based or pattern-based classification methodology, which uses Boolean logic, optimization, and combinatorics to identify signatures separating two labeled classes. It has been successfully applied to a wide array of problems. In this thesis we propose and analyze several approaches to build meaningful prognostic models for survival prediction using LAD, and develop a novel methodology for survival analysis. All these approaches are motivated by real world biomedical problems proposed by medical experts.

Survival analysis is an important branch of statistics and data-mining which involves predicting survival or failure time. Survival analysis differs from regression (prediction of continuous outcome) mainly because of the presence of censored samples. A sample is considered censored if it has incomplete survival time information. For example, in the medical context, observations are patients, who are usually enrolled in the study when they come to the clinic or start treatment and are observed over time for events such as death, heart attack, recurrence of cancer, etc. Patients are censored if either

they discontinued coming to the hospital (moved out of town, stopped answering the phone, etc.), or the study ended without their experiencing the event. In most experiments, censored patients represent a large proportion of the samples, so that simply eliminating them and reducing the problem to a regression analysis is not practical, as it would have a significant impact on the statistical power of the experiment.

In Chapter 2, we describe a study involving cardiovascular patients where survival information was collected for a period of three years [1-3]. The goal of this study was to develop an accurate risk score to predict one-year mortality or myocardial infarction (heart attack) based on measurements in the blood, in order to provide a score which would help clinicians to determine appropriate treatment. Because this was a study with a relatively short duration (three years), and it was known that blood is a dynamic medium with predictive capability only for short periods of time, we selected a one-year cut-off for defining high- and low-risk patients. This reduced the problem to a two-class classification problem, for which it was straightforward to develop LAD patterns to distinguish the classes. In this analysis, we developed a series of models to predict risk for patients, which could be used in an efficient and inexpensive manner in the clinic for which we present complete “in-silico to in-vitro” analysis in the chapter.

Chapter 3 is a study to predict long term mortality based on clinical measurements for patients undergoing coronary artery bypass surgery (CABG) [4, 5]. The main challenges in this dataset were that only 25% of the patients had an event during the study, the data was noisy, and the variables had low predictive power. As a result, we were forced to focus on patients at extreme risk, for which there was sufficient signal in the data collected. Based on experimental validation analysis within the dataset, we

defined a high-risk class as those patients who died within three years and a low-risk class as those who survived beyond nine years. We also defined a score to identify *'confusing patients'* in the data. These are the high-risk (low-risk) patients who seem to have very similar measurements as patients of the opposite class but for whom the variables in the dataset do not have sufficient predictive power. We define a statistical score to identify these *'confusing patients',* and build patterns using LAD for the non-confusing patients, which are able to classify them into the extreme high and low-risk subsets. In addition, we identified the most important variables and patterns predictive of patients at extreme risk.

Chapter 4 is a study identifying subtypes or clusters in kidney cancer [6]. A cluster is defined as a subset of samples which are similar to each other compared to other samples. Usually cancers are classified based on tissue of origin, morphology, histological, radiological and pathological analysis. More recently, as a result of the observation that cancers with similar presentations with respect to these measures seem to have divergent outcomes under identical treatment regimens, there is a sense in the community that one might learn more by looking within the tumors and analyzing their molecular footprint, which is defined by the genes/proteins/microRNA they express and the specific genetic mutations they have accumulated to grow and metastatize. Since microarray technology is the most developed of all the ways to determine a molecular profile for tumors, it is therefore becoming common practice to identify subtypes of cancers based on gene-expression profiling. In such an analysis, each subtype is defined as the set of patients with a similar pattern of gene expression compared to normal samples, under the presumption that this altered profile of expression is responsible for

the disease and drives disease progression. Since we do not have an understanding of the disease process the classification into clusters is an unsupervised problem. The hypothesis to be tested is that the subtypes which are identified by such an unsupervised analysis would also have differences in clinical outcome (such as disease free survival). The dataset we analyzed consisted of gene-expression data for the most common morphological subtype of kidney cancer, called clear cell Renal Cell Carcinoma or ccRCC. We applied a robust algorithm called consensus ensemble clustering to identify subtypes from gene expression data, and then used LAD patterns to distinguish between the identified subtypes. The subtypes and LAD signatures distinguishing them were validated on external datasets. This analysis validated the hypothesis that subtypes based on gene-expression patterns not only exist, but are also useful to classify patients into distinct survival classes, with significant potential clinical impact.

In Chapter 5 we discuss new supervised methods for predicting continuous outcome and also for handling censored patients [7-9]. First we develop a simple linear programming model for survival regression. We then present a new method, called Logical Analysis of Survival Data (LASD), which uses the principles of LAD to create a continuous risk score. The method involves building patterns at every time-point t in the data when an event occurs. These patterns distinguish samples which experienced the event before time t (high-risk for time t) from those samples which had the event after t (low-risk for time t). A pattern-specific score is defined and computed as the area under the Kaplan-Meier (survival) curve for the samples satisfied by the pattern. Finally, we compute a patient-specific score as the average of the pattern scores for patterns covering the patient which we present as an estimator of risk or survival time. We also present

ensemble methods to improve the performance and robustness of LASD and illustrate the performance of these methods on a kidney cancer dataset (the same dataset discussed in Chapter 4).

In Chapter 6 we summarize the fundamentals of classification problems, LAD, classification performance metrics, survival analysis, a summary of our conclusions and main contributions, some possible directions for future research and concluding remarks.

1.1 Logical Analysis of Data

Logical Analysis of Data (LAD) is a combinatorics, optimization and Boolean logic-based data-mining algorithm to solve two-class classification problems. The input dataset, $S \in \mathbb{R}^m$, consists of *samples* or *observations* in two disjoint classes, $S = S^+ \cup S^-$, where S^+ is the set of *positive samples*, and S^- is the set of *negative samples*, and $S^+ \cap S^- = \emptyset$. In two-class classification problems the main task is to distinguish between the positive and negative samples based on variables measured (*i.e.,* coordinates). The key ingredient of the LAD algorithm is the identification of patterns or rules in the data, which distinguish positive samples from negative samples. These patterns are then used to define a function which allows the classification of new or unseen samples. The main concepts of LAD were introduced in the late 1980s [10, 11] and subsequently, it was applied successfully to a wide array of problems in the fields of medicine, finance, social sciences, etc. [12-16]. The main elements of the LAD algorithm are discussed below:

1.1.1 Discretization and support set selection

The LAD algorithm works on binary data. A standard step in analyzing numerical variables with LAD is discretization (*i.e.*, transformation of a numerical variable to discrete levels without losing predictive power). This step consists of finding cut-points

$(c_1, c_2, ...)$ for each numerical variable v. These are simply interpretated as a sequence of putative "high" and "low" threshold values which can be collectively used to build a global classification model over all variables.

The problem of discretization is well studied, and there exist many powerful methods discussed in the survey papers [17, 18]. Discretization techniques are divided into two main categories, based on whether they use class or outcome information: supervised (chi-merge, khiops, information gain, etc.), and unsupervised (equal width, equal frequency) methods. Another possibility for smaller datasets is simply to use all possible cut-points to discretize each numerical variable. As we shall see later, cut-points are the basis for the synthesis of general rules that can be used for classification and prediction purposes.

To each variable vv and cut-point c we associate an *indicator variable I(vv, c)* defined by:

$$I(v, c) = \begin{cases} 1 & \text{for } v > c \\ 0 & \text{for } v \leq c \end{cases} \qquad (1.1)$$

Transforming the data from discrete levels to indicator variables, results in a binary dataset. For each variable, virtually any numerical value can be considered a cut-point. However, we focus on identifying cut-points with a high distinguishing power.

A *support set* is defined as a smallest subset of binary variables which can distinguish every pair of positive and negative samples in the data. Support sets can be identified by solving a minimum set covering problem. Given a binary dataset with m variables, we denote by $y_i,\ i = 1, \ldots, m$, selector variables (binary) to indicate whether the corresponding variables in the data are retained in the support set. We denote by p_i

and n_i the i^{th} coordinate of $p \in S^+$, and $n \in S^-$ respectively. A support set can be obtained by solving the following optimization problem:

$$\begin{aligned} &\textbf{Minimize } \sum_{i=1}^{m} y_i \\ &\textbf{subject to } \sum_{i=1}^{m} (p_i \oplus n_i)\, y_i \geq 1 \qquad \forall\, p \in S^+, n \in S^- \\ &\qquad y_i \in \{0,1\},\ i = 1, \ldots, m \end{aligned} \tag{1.2}$$

where $(p_i \oplus n_i) = 1$ when $p_i \neq n_i$, and is 0 otherwise. The right hand side of the constraint is changed to integers larger than 1 to increase the robustness of the support set selected. In some datasets, indicator variables alone can separate positive observations from the negative ones. For large datasets, solving the set covering model to optimality is computationally very expensive, and instead we use a greedy heuristic by a process of forward selection of binary indicator variables.

1.1.2 Combinatorial patterns

In this section we show how one or more indicator variables can be used in combination to produce rules that can define homogenous subgroups of interest within the data. While an indicator variable can partially predict the outcome by relating the high or low values of a variable with a specific outcome, the simultaneous use of more than one indicator variable allows for the definition of more complex rules that can be used for the precise classification of an observation. Such rules are called *combinatorial patterns* (or simply patterns), and can be regarded to be indicative of a specific class. A pattern *P* is a subcube of $\{0,1\}^m$, where *m* is the number of variables; they can also be described as conjunctions of indicator variables. Patterns define homogeneous subgroups of observations with distinctive characteristics. These subgroups have a distribution of positive and negative samples which is significantly different compared to the original

population. If an observation satisfies the conditions imposed by the definition of a pattern, we shall say that the observation is *covered by* that pattern.

A *positive (negative) pattern* is defined as a combination of indicator variables which covers a large proportion of positive (negative) observations and a minority of the negative (positive) ones. A *pure pattern* is one which covers only samples of one class and none of the other. Important characteristics of a pattern are defined below:

- *Degree*, the number of variables or conditions involved in the definition of the pattern;
- *Positive prevalence*, the proportion of positive observations covered by the pattern;
- *Negative prevalence*, the proportion of negative observations covered by the pattern;
- *Positive homogeneity*, the proportion of positive observations among all those observations covered by the pattern;
- *Negative homogeneity*, the proportion of negative observations among all those observations covered the pattern.

LAD patterns are generated by exhaustive search by enumerating all possible combinations of a given degree, prevalence and homogeneity. The above parameters (degree, prevalence and homogeneity) are to be calibrated in the model building stage using cross-validation experiments.

1.1.3 Classification model

An LAD model is simply a collection of positive and negative patterns of given characteristics (degree, homogeneity, and prevalence), with the property that every observation in the dataset is covered by at least k of the patterns, where k is a parameter to be tuned for maximizing the accuracy (see the next section for the definition of

accuracy and related performance measures for classification models). Ideally, the positive patterns of a model would cover exclusively positive observations, while the negative patterns would cover only negative observations. We use a set-covering model similar to the one described above (1.2), to identify a minimum subset of the patterns in the classification model. In cases when the dataset is large, and there are a large number of patterns, we use a greedy heuristic to select patterns into an LAD model.

1.1.4 Discriminant score and prediction

A *discriminant scoring function, d(s),* for a sample *s* is defined to be the difference between the proportion of positive patterns and negative patterns covering *s.* Let us denote by π and ν the number of positive and negative patterns covering *s*. Let *P* (*N*) be the number of positive (negative) patterns in the LAD model.

$$d(s) = \frac{\pi}{P} - \frac{\nu}{N} \qquad (1.3)$$

This score goes between -1 and +1. Samples are classified to be positive (negative) if their discriminant score is greater than zero (is less than zero). There are two situations in which a sample has score zero: (i) none of the patterns cover the sample, (ii) equal proportions of positive and negative patterns cover the sample. This is usually a rare occurrence, and such samples are left unclassified.

1.1.5 Software Implementations

There exist several implementations of the Logical Analysis of Data algorithm: Datascope [19], Ladoscope [20], Cap-LAD [21], etc. In this thesis, we use mainly Datascope and Ladoscope.

1.2 Performance measures for classification

There are several parameters which can be used to measure the quality of a classification model. There are two main categories of measures: (i) when the prediction of a classification model is the class, (ii) when the prediction is a probability or continuous score. In the former case when the model predicts class, a simple measure of accuracy is the proportion of correctly classified samples. This measure is biased when the sizes of the classes ($|S^+|, |S^-|$) are unbalanced. Let us denote by S_C^+ (S_C^-), S_U^+ (S_U^-) the set of correctly classified positive (negative) samples, and unclassified positive (negative) samples (discriminant score = 0), respectively. Below we define performance measures that are used in this thesis, and whose behavior is not affected by unbalanced class distribution.

- The *sensitivity* of a classification model (also called positive predictive value, PPV) is defined as the proportion of correctly classified positive cases

$$Sensitivity = \frac{|S_C^+|}{|S^+|} \qquad (1.4)$$

- S*pecificity* (negative predictive value, NPV) is defined as the proportion of correctly classified negative samples.

$$Specificity = \frac{|S_C^-|}{|S^-|} \qquad (1.5)$$

- *Accuracy* of a model is defined as the average of sensitivity and specificity and takes into account unclassified samples.

$$Accuracy = \frac{Sensitivity + 0.5\frac{S_U^+}{S^+} + Specificity + 0.5\frac{S_U^-}{S^-}}{2} \qquad (1.6)$$

Another measure that is used in the thesis is the *Receiver Operating Characteristics* (ROC) curve [22]. This is used when the prediction is a continuous score,

or probability instead of a predicted class. This is a plot of 1-Specificity versus Sensitivity for all possible predicted scores. *Area under the ROC curve* (AUC) is defined as the integral of the ROC curve (ranges from 0 to 1). The diagonal indicates no predictive value, with an AUC of 0.5. An AUC of 1 is considered as a perfect measure. AUCs below 0.5 indicate poor performance (inverting the predicted outcome gives a better performance).

The performance of a classification model is evaluated in two possible ways: (i) bootstrapping: model is built on a random subset of the data, and is tested on the remaining data, and this procedure is repeated many times. (ii) *k*-fold cross-validation: data is divided randomly into *k* parts, one part is left out as test set, and the model is built on the remaining *k*-1 parts. This is repeated by selecting each of the *k* parts as the test set once. The mean and standard deviation of the accuracy on the test sets is computed.

1.3 Other classification methods

1.3.1 Random forests

Random forests is an ensemble classification method, introduced by Breiman [23]. The main feature in this method is to build "many" decision tree models on bootstrapped data selected from the data. Consider a dataset of *N* samples and *M* variables. *N* samples are selected randomly with replacement to form a *bag* set. The remaining samples form the *out-of-bag* set. Decision tree models are built on the bag set on *m* ($m<N$) variables selected randomly from the *M* variables. The model is tested on the out-of-bag samples. This procedure is repeated many times. Random forests is a very powerful method, and is known to outperform several other classification methods.

1.3.2 Support vector machines

Support vector machines was introduced by Vapnik [24]. It is an optimization based classification method which classifies samples by constructing hyperplanes to separate the two classes with the objective of maximizing the margin between the *support vectors*. SVM uses a *kernel* function to first project the samples to a higher-dimensional space where the classes can be separated by a hyperplane. Soft margin SVM models were introduced to avoid overfitting (permits models to make errors).

1.4 Survival Analysis

Survival analysis is an important branch of statistics and data mining which involves predicting survival or failure time. Survival analysis is also referred to as life data analysis, failure analysis, deterioration modeling or reliability theory depending on the field of its application. In the medical context, observations are patients, they are enrolled in the study when they come for a treatment and are observed for events. Some examples of events in this context are death, heart attack, recurrence of cancer. The variables or attributes in the study are collected at the time when they enroll for the study. *Survival/failure time*, also known as *time to event*, is the time from the patients' enrollment in the study until the occurrence of the event being studied. In this paper, we focus on developing prognostic methods based on previously recorded observations.

Survival analysis differs from regression (prediction of continuous outcome) mainly because of the presence of censored data. A sample is considered to be censored if it has incomplete survival time information. There are different types of censoring: right-censoring, left-censoring, right-truncation and left-truncation. We will concentrate on right-censored survival analysis problems since they constitute majority of the situations. Right-censoring occurs when the patient did not have an event until the end of the study.

These studies are usually conducted for a fixed period of time (e.g. 10 or 15 years), during which patients are observed for the event of interest. Some patients do not experience the event during this period. Such patients are said to be right-censored and their censoring time serves as a lower bound for the time to event. If the study was conducted for long enough periods, then the event would be observed in all patients. One approach for handling censored data is to disregard them from the study; this would result in a classical regression problem. However, this is a naïve approach, because usually a large proportion of the patients are censored. Using information about censored patients for predicting the survival time is the main motivation behind solving survival problems.

More recent techniques include using classification and regression trees [25] for estimating survival functions such as relative risk tree [26], neural networks [27], naïve Bayes classifiers [28], splines [29], etc.

Meta-classifiers, such as bagging [30], with survival decision trees as base classifiers is presented in several papers [31-35]. A general flexible framework for survival ensemble techniques is described by Hothorn *et al.* [36].

Another commonly used method is that of transforming a survival analysis problem into a classification problem involving the prediction of patients at high/low risk of having the event, or good/bad clinical prognosis. This has been shown in a number of publications [12, 37-39].

1.4.1 Notations and Problem description

The random variables $\mathcal{X}$ and $\mathcal{C}$ represent the time to event and censoring time respectively. Observed variables are represented by a triplet: (T, Δ, Z) where $T = \min(\mathcal{X}, \mathcal{C})$, the *censoring status* Δ is equal to 1 if the event occurred and is 0 otherwise

and Z is a vector of attributes. T is the survival time which can be the actual time to event or a lower bound of it, depending on the censoring status.

The dataset consists of a sample of n observations from the above observed variables denoted by $(t_i, \delta_i, z_i)_{i=1\ldots n}$. Note that z_i is a vector of attributes. A basic quantity to describe the survival problem is the survival function $S(t) = P\ (T > t)$ i.e. the probability to survive until time t. Let us denote by $\hat{S}(t)$ an estimated survival function.

The main problem in survival analysis is the estimation of the survival function, or the expected value of the survival function, based on information from attributes. In this paper, we develop models to derive an estimator $\hat{S}(t|z_i)$ of the survival function for each observation.

There are several existing statistical and data-mining techniques to predict the survival time in the presence of censored data. Below we briefly discuss some of these methods.

1.4.2 Kaplan-Meier or Product-limit survival function

Product-limit estimates of the survival function (survival probability as a function of time) were developed by Kaplan and Meier in 1958 [40], also known as Kaplan-Meier (KM) estimates. They are univariate estimates of the survival function, based on the entire range of the data, which do not take into account effects of covariates. Let d_t be the number of observations who experience an event at time *t*, and let Y_t be the number of observations who are at risk at time *t* (who have event at time *t* or later); then the formula for the Kaplan-Meier (KM) estimator is given as:

$$\hat{S}(t) = \prod_{\{i|t_i \le t\}} (1 - \frac{d_{t_i}}{Y_{t_i}}), \qquad (1.7)$$

where t_i's are the times when an observation experiences an event. The main assumptions of KM or product-limit estimators are (i) censoring does not depend on prognosis, and (ii) survival does not depend on the time when the patient joined the study. These are standard assumptions, which are meaningful and reasonable, and are required for all existing methods.

1.4.3 Cox proportional hazards regression

Cox proportional hazards regression proposed by Cox in 1972 [40] is a standard statistical technique for modeling the effect of attributes on the survival time. The Cox model predicts hazard rate, i.e., event rate at time *t* conditional on survival until *t*. The main assumptions of the Cox model are that the hazard rate has a log-linear relationship with the attributes, and that the ratio of the hazard rates of any two observations in the dataset depends only on the attributes, but is independent of their survival times (proportional hazards assumption). However, this model does not make any assumptions about the shape or distribution of the hazard function. Thus, it is considered to be a semi-parametric method.

1.4.4 Random survival forests

Random survival forests (RSF) is an extension of random forests (RF) for survival analysis. The main differences in RSF and RF are in (i) the rules for splitting nodes, and (ii) computation of survival probability for the leaves of the decision trees. Logrank test (described below), logrank score, approximate logrank test and conservation of event rules are available as options for node splitting. For each of the terminal or leaf nodes, cumulative hazard estimate is computed as a function of time. For a new sample, the final prediction is the average of the cumulative hazard estimates predicted in each of the bootstrapped trees.

1.4.5 Logrank test

Logrank test is used to analyze the differences in survival distributions for groups of samples. The null hypothesis for the logrank test is that there is no difference in survival probabilities for two or more groups at any of the time points in the study. At every time point when there was an event, the observed and the expected event probability is computed for each of the groups. The χ^2 statistic is used for the logrank test, the degrees of freedom being computed as number of groups minus one. The logrank test assumes that the KM survival curves for the different groups do not cross each other. This can be checked by simply plotting the KM curves for the different groups.

1.4.6 Performance measures for survival analysis

Standard performance measures used for regression models (e.g. squared error, correlation coefficient, etc.) cannot be used for survival problems due to the presence of censored samples. In the literature, many publications consider different measures to evaluate goodness of fit of the estimated survival function. The most classical one is concordance accuracy or concordance index or c-index [41] The c-index is a rank statistic which is equivalent to the area under the ROC curve (AUC) when censored samples are excluded. More recent proposals for performance measures include the Brier score [42], Sep and D measures [43], and quadratic loss functions [36].

We will use c-index as a performance measure mainly because we are interested in the problem of ranking observations according to their event-risk.

1.4.6.1 Concordance index

The c-index is a rank statistic which measures the ability of a model to rank the observations in order of the risk. C-index evaluates the proportion of correctly ordered pairs of observations. A pair of observations is said to be correctly ordered if the

observation with the shorter survival time is predicted to be at higher risk when compared to the one with longer time to event. Note that not all pairs are considered for this computation. For example, a pair of censored observations is not relevant because we cannot say which one has higher event-risk.

We denote by $\hat{\rho}_i$ a risk estimation for observation i. A pair of observations (i, j) is *concordant* (*semi-concordant*) if:

- $\delta_i = 1$ (i experiences an event),
- $t_i < t_j$
- $\hat{\rho}_i > \hat{\rho}_j$ ($\hat{\rho}_i = \hat{\rho}_j$).

We denote by N_{rp} the total number of relevant pairs, N_{cp} the number of concordant pairs, and by N_{sp} the number of semi-concordant pairs. The formula for c-index is:

$$c_{index} = \frac{N_{cp} + \frac{1}{2}N_{sp}}{N_{rp}} \tag{1.8}$$

The c-index ranges between 0 and 1, where 1 means perfect ranking and 0.5 is equivalent to random ranking.

The main drawback of c-index is that it may overestimate models which produce false negative errors [4]. A model which assigns time monotonic risk values for the uncensored samples in the training data and which assigns the *lowest* risk values for the censored samples in the training data would have a perfect c-index in any cross-validation experiment on the training data. This model would be considered "c-index perfect." However this model may assign *low* risk value to unseen patients which resemble the censored samples in the training data, even if these patients could be in fact of higher risk.

Chapter 2

Prediction of One-year Myocardial Infarction and Death using blood based parameters: In-silico to In-vitro[1]

2.1 Introduction

Cardiovascular or heart disease is the number one cause of death and disability worldwide, including the United States. Each year, there are more deaths in the United States due to cardiovascular disease than cancer. Myocardial infarction (MI) or heart attack is a pathologic process in which heart muscle dies due to lack of oxygen supply. This happens when the blood vessels that supply oxygen to the heart become blocked. In general, cardiovascular diseases are very complex in their mechanism. There are many known risk factors for heart disease: clinical parameters (age, gender, hypertension, family history etc.), life-style related parameters (stress, diet, exercise, obesity, etc.), and genetic factors (polymorphisms in genes such as APOE, LDLR, APOA1, MPO, etc.). There have been several recent advances in understanding the mechanisms of cardiovascular disease, but accurate prediction of risk for cardiovascular patients still remains a challenge. It is of great importance to identify patients at high-risk in order to provide them the option of more aggressive forms of treatment and anticipate possible surgical interventions [44, 45]. Similarly, it is also important to identify patients at low-

[1] Based on collaborations with Stanley Hazen, Marie-Luise Brennan (Cleveland Clinic). This chapter is part of one submitted manuscript [1], and two working papers [2, 3].

risk so as to make the best use of health care resources and not to overburden the patient with unnecessary treatment. Current clinical risk assessment tools include the use of algorithms developed from epidemiology-based studies of untreated primary prevention populations. They are, however, limited in their application to a higher risk and medicated cardiology outpatient setting which constitutes an increasing percentage of patients seen in the healthcare environment [46]. An area of active investigation is the incorporation of combinations of biological markers, genetic polymorphisms, and noninvasive imaging approaches for additive prognostic value [47-50]. Despite considerable interest, efforts to incorporate more holistic array-based phenotyping technologies (e.g. genomic, proteomic, metabolomic, expression array) for improved cardiac risk stratification remain in their infancy and have yet to be translated into efficient and robust platforms amenable to the high throughput demands of clinical practice.

Blood is a complex but integrated sensor of physiologic homeostasis (equilibrium). Perturbations in blood composition and blood cell function are seen in both acute and chronic inflammatory conditions. Elevated white blood cell count (both neutrophils and monocytes) has long been associated with cardiovascular mortality [51, 52]. Myeloperoxidase, an abundant white blood cell granule protein [53], has been mechanistically linked with multiple stages of cardiovascular disease [54], including modification of lipoproteins [55-57], creation of lipid mediators [58], regulation of protease cascades [59, 60], and modulation of nitric oxide bioavailability and vascular tone [61, 62]. Systemic myeloperoxidase levels are increased in patients presenting with chest pain [60] and suspected acute coronary syndromes [63].It has been noticed that

such patients often experience short-term adverse cardiovascular events after initial diagnosis of their condition. Similarly, numerous mechanistic and epidemiological ties exist between various components and activities of circulating red blood cells and platelets with processes critical to both vascular homeostasis and cardiovascular disease [64-67]. We hypothesized that a peroxidase-based hematology analyzer (that rapidly generates routine blood tests called complete blood cell count (CBC) and differential analysis) would concomitantly provide a broad spectrum of novel data relevant to the evaluation of cardiovascular risk in subjects.

Risk stratification for cardiovascular outcomes has traditionally relied upon the use of individual risk factors either alone or in additive combinations. Given the multi-factorial nature of cardiovascular disease, detection of combinatorial patterns indicative of both high and low risk holds promise for improving accuracy of risk stratification, treatment decisions and outcomes. The application of most existing array-based platforms for the analysis of blood components remains in the research domain. Herein we report the development of the PEROX risk score for the accurate prediction of one-year non-fatal MI and death risk using clinical and laboratory data available routinely in an outpatient cardiac setting, combined with white blood cell (WBC), peroxidase-, red blood cell (RBC)- and platelet-related parameters obtained during high throughput performance of a complete blood count with differential analysis.

For this study, blood was collected from patients who were stable, without chest pain, and were undergoing elective diagnostic cardiac catheterization at the Cleveland Clinic. Hematology analyses were performed on the blood to obtain white blood cell, peroxidase, red blood cell and platelet-related parameters. The patients were followed up

after the procedure and information about any death or myocardial infarction event was recorded. This presented a survival analysis problem which involved prediction of risk of death or myocardial infarction. The main challenges in this dataset were: (i) that only 6% of the dataset had an event (myocardial infarction or death) and (ii) that this was a relatively short study (maximum follow up time was 3 years). The approach that we chose to follow was to build a classification model to distinguish between patients who had an event within 1 year (high-risk patients) and those patients who did not have an event within 1 year (low-risk patients). The reason for choosing the 1-year cut-off are: a) the effects in blood are dynamic which means that an increase in predictive power requires that we chose lower cut-offs; b) events that occur at later times can be attributed to other factors, such as changes in diet, exercise, etc. c) There was a very small proportion of events at time points prior to 1 year. Hence it is necessary to chose a 1-year time point to be able to properly define a 2-class classification problem. The Logical Analysis of Data (LAD) methodology was used to identify patterns to distinguish high-risk from low-risk patients.

In this chapter, results are presented in two sections: In Section 2.2 (PEROX Risk Score) we present the results on a risk score developed using clinical, laboratory and peroxidase-based hematology analyzer parameters. This section develops a holistic approach which provides accurate prediction of high and low-risk patients. The main problem here is that this model cannot be easily translated into effective clinical use. In Section 2.3 (CHRP(PEROX) & CHRP Risk Scores) we present two models built on the same protocol as the PEROX model. In the first model, the variables used were only from the peroxidase-based hematology analyzer (CHRP(PEROX) risk score); in the second

model we used variables from any generalized hematology analyzer (CHRP risk score). Both of these models can be adapted easily and effectively for use in the clinic. Of the two, the CHRP risk score is more easily applicable in a clinical setting because it has variables built on a generalized hematology analyzer which is easily available to the clinician. In addition, even though the CHRP(PEROX) and CHRP risk scores have lower accuracies compared to the PEROX risk score, they have superior prognostic power when compared to existing cardiovascular risk scores. We compare risk scores built on these models with traditional cardiovascular risk factors used in assessing the risk of a patient, and show that the CHRP(PEROX) and CHRP based risk scores have high independent prognostic values when compared with traditional risk factors.

2.2 PEROX Risk Score

2.2.1 Analysis

Subjects missing any hematology analyzer variable were excluded from the study. Hematology analyses from 7,369 subjects (out of an initial cohort of 7,466 subjects) were available for analysis. Imputation based on median value within deciles of age and per gender was performed if any clinical or laboratory variable was missing. The initial dataset was stratified based on whether a patient experienced an event (non-fatal MI or death) within one-year following enrollment. Randomization using a uniform distribution method was performed to randomly select 80% of patients (Derivation Cohort) for model building. The remaining 20% (Validation Cohort) was set aside for model testing and validation. To assess trends in the data, traditional statistical analyses were performed. Mean and median differences were assessed with Student's t-test and Mann-Whitney test, respectively. A p-value of <0.05 was considered significant. Association between

variables was assessed by Spearman's correlation. Hazard ratios were generated for variables or logarithmically transformed variables (if not normally distributed). Cluster analysis of data was performed to assess whether there were identifiable clusters within the data for death or MI outcomes.

Logical Analysis of Data (LAD) was used to identify high and low-risk patterns, and to build a model predictive of risk for death or MI at the one-year time point. Variables to be included in the model were selected based on clinical significance, and reproducibility (for hematology parameters) as monitored in inter-day and intra-day replicates. We tried several different methods for discretizing the numerical variables. We had to make a tradeoff between using a large number of cut-points (to increase the predictive power), and fewer cut-points (to take into account the noise inherent in the measurements made by the cytometer, and to reduce over-fitting). Finally we selected three equal frequency cut-points, because they optimize the cross-validation accuracy and also make sense biologically. Feature selection was based on using set-covering model (1.2) in each of the groups (WBCs, RBCs and Platelets) separately to identify support-sets. This feature selection step allows us to identify a set of variables which together are highly predictive, while reducing the use of redundant variables. Variables included in the PEROX risk score model are listed in Tables 1 and 2. We built separate models for Death and MI high and low-risk data. Risk scores for both models were computed, and finally averaged to get the combined risk score for death/MI in 1 year. The reason for building separate models is that the biology and mechanisms for the two events are different, and we get better results by optimizing the individual models and then combining the risk scores. The constraint that we used for building patterns was degree 2,

and 10% minimum prevalence (coverage of samples of the same class). Patterns were generated and extensively tuned for both homogeneity and prevalence to obtain best accuracy on cross-validation experiments. The PEROX risk score was calculated as the scaled LAD discriminant score (range: 0-100).

2.2.2 Model Validation and Risk Score Comparisons

Clinical utility of the PEROX risk score for stratification of patients into high-, medium- and low-risk categories was based on tertiles (3 equal frequency cut-points) of the one-year PEROX risk score in the Derivation Cohort. Figure 2.1 shows a plot of the Kaplan-Meier curves for the predicted risk groups in the Validation cohort. Comparison of the survival distributions in these three groups was made using the log-rank test. We also present the risk plot with mean risk score in the three groups on the x-axis and the % of events on the y-axis. Cubic splines (with 95% confidence intervals) were used to draw smooth curves through these points to examine the relationship between the mean PEROX risk score in the three risk groups and one-year event rates.

Receiver operating characteristic (ROC) curves were plotted for one-year death, MI, and combined death or MI events for the validation cohort using risk scores assigned by the PEROX risk score model. ROC curves were also plotted for the Adult Treatment Panel III (ATP III), Reynolds Risk Score, and Duke angiographic scoring systems [68-70].

2.2.3 Results

The population examined had a mean age of 64 ± 11 years, 68% were male, 90% were Caucasian, and 69% had history of cardiovascular disease at the time of enrollment. Clinical and laboratory parameters used in the development of the PEROX risk score are shown in Table 2.1A. Traditional cardiac risk factors and laboratory measurements were

essentially similar in the derivation and validation cohorts. One-year event rates for incident non-fatal MI or death, individually, and as a composite, did not significantly differ between the derivation and validation groups (p=0.37 for MI; p=0.50 for death; p=1.00 for MI or death).

Traditional risk factors and comprehensive hematology variables are associated with cardiovascular risk. The hazard ratios of traditional cardiac risk factors, laboratory measurements, and clinical characteristics for predicting incident one-year risk for non-fatal MI and death are shown in Table 2.1A. Significant hazard ratios associated with death included age, hypertension, history of smoking, diabetes, fasting blood glucose, HDL cholesterol, creatinine, and C-reactive protein level. Surprisingly, elevations in total cholesterol, LDL cholesterol, and triglycerides, and reduced diastolic blood pressure and body mass index were associated with decrease in risk. These associations likely reflect confounding by indication bias whereby patients with a higher prevalence of co-morbidities are more likely to be taking medication or undergoing aggressive interventions. As expected, increased risk for incident one-year non-fatal MI was associated with diabetes, hypertension, elevated systolic blood pressure, fasting blood glucose, C-reactive protein, and creatinine concentration, while decreased risk was associated with higher HDL cholesterol levels (Table 2.1A).

The hazard ratios of hematology measurements used in the PEROX risk score for predicting incident one-year risk for non-fatal MI or death are shown in Table 2.1B. Multiple significant hazard ratios were observed between various leukocyte, erythrocyte, and platelet parameters and incident one-year risks for non-fatal MI and death, consistent

with multiple prior individual reported associations with various hematological parameters [64-67].

Table 2.1A. Clinical and Laboratory Parameters in the PEROX Model

Abbreviations: MI, myocardial infarction; HR, hazard ratio; CI, confidence interval.
Data are shown as mean ± standard deviation for normally distributed variables, median (interquartile range) for non-normally distributed variables, or number in category (percent of total in category). Hazard ratios were calculated per standard deviation (for normally distributed variables). For variables with non-normal distribution (creatinine, potassium, c-reactive protein), values were log transformed and hazard ratios calculated per log of standard deviation. *p <0.05

	Derivation Cohort (N = 5,895)	**Validation Cohort (N = 1,474)**	**Death 1 year HR (95% CI)**	**MI 1 year HR (95% CI)**
Traditional Risk Factors				
Age (years)	64.1 ± 11.3	64.1 ± 10.9	1.88 (1.65-2.14)*	1.14 (0.99-1.32)
Male – n (%)	4,021 (68)	1,024 (69)	0.93 (0.73-1.18)	1.21 (0.88-1.66)
Hypertension – n (%)	4,335 (74)	1,075 (73)	1.67 (1.24-2.25)*	1.53 (1.07-2.19)*
Current smoking – n (%)	770 (13)	162 (11)*	0.90 (0.63-1.29)	1.28 (0.87-1.89)
History of smoking – n (%)	3,869 (66)	995 (68)	1.35 (1.04-1.74)*	0.90 (0.67-1.20)
Diabetes mellitus – n (%)	2,054 (35)	544 (37)	2.09 (1.66-2.62)*	1.55 (1.17-2.06)*
Laboratory Measurements				
Fasting blood glucose (mg/dl)	111 ± 47	112 ± 43	1.23 (1.13-1.33)*	1.27 (1.16-1.39)*
Creatinine (mg/dl)	1.1 (0.8-1.1)	1.1 (0.8-1.1)	1.57 (1.48-1.67)*	1.22 (1.09-1.37)*
Potassium (mmol/l)	4.2 (4.0-4.5)	4.2 (4.0-4.5)	1.10 (1.04-1.17)*	0.97 (0.84-1.12)
C-reactive protein (mg/dl)	3.0 (1.7-5.9)	3.0 (1.6-5.5)	1.92 (1.71-2.16)*	1.21 (1.05-1.40)*
Total cholesterol (mg/dl)	176 ± 43	178 ± 43	0.71 (0.62-0.81)*	0.93 (0.80-1.07)
LDL cholesterol (mg/dl)	100 ± 36	101 ± 36	0.78 (0.69-0.89)*	0.97 (0.84-1.13)
HDL cholesterol (mg/dl)	46 ± 14	46 ± 14	0.84 (0.74-0.95)*	0.71 (0.60-0.84)*
Triglycerides (mg/dl)	160 ± 119	163 ± 120	0.82 (0.71-0.96)*	1.07 (0.96-1.19)
Clinical Characteristics				
Systolic blood pressure (mm Hg)	135 ± 21	136 ± 22*	0.96 (0.85-1.07)	1.17 (1.02-1.34)*
Diastolic blood pressure (mm Hg)	75 ± 12	75 ± 13	0.81 (0.73-0.90)*	0.97 (0.85-1.12)
Body mass index (kg/m^2)	30 ± 6	30 ± 6	0.78 (0.68-0.89)*	0.90 (0.78-1.05)
Aspirin use – n (%)	4,270 (72)	1,087 (73)	0.64 (0.51-0.81)*	0.93 (0.68-1.27)
Statin use – n (%)	3,450 (59)	869 (59)	0.82 (0.65-1.03)	0.70 (0.53-0.92)*

Table 2.1B. Peroxidase-based Hematology Parameters in PEROX Model

Abbreviations: MI, myocardial infarction; HR, hazard ratio; CI, confidence interval; RBC, red blood cell; Hgb, hemoglobin. Data are shown as mean ± standard deviation for normally distributed variables, or median (interquartile range) for non-normally distributed variables. Some variables have no unit of measure associated with them. Median for peroxidase X sigma was zero, therefore, mean is shown. Hazard ratios were calculated per standard deviation (for normally distributed variables). For variables with non-normal distribution, values were log transformed and hazard ratios calculated per log of standard deviation.

	Derivation Cohort	Validation Cohort	Death 1 Year HR (95% CI)	MI I Year HR (95% CI)
White Blood Cell Related				
White blood cell count ($x10^3/\mu l$)	6.50 ± 2.19	6.51 ± 2.22	1.31 (1.21-1.42)*	1.04 (0.91-1.20)
Neutrophil count ($x10^3/\mu l$)	4.39 ± 1.97	4.42 ± 1.94	1.37 (1.26-1.48)*	1.01 (0.88-1.16)
Lymphocyte count ($x10^3/\mu l$)	1.54 ± 0.76	1.52 ± 0.86	0.73 (0.62-0.86)*	1.02 (0.89-1.16)
Monocyte count ($x10^3/\mu l$)	0.35 ± 0.18	0.35 ± 0.17	1.13 (1.09-1.16)*	1.06 (0.96-1.16)
Eosinophil count ($x10^3/\mu l$)	0.21 ± 0.15	0.21± 0.18	1.11 (1.03-1.19)*	1.05 (0.93-1.18)
Basophil count ($x10^3/\mu l$)	0.05 ± 0.03	0.05 ± 0.03	1.09 (0.98-1.21)	1.07 (0.94-1.22)
Number of peroxidase saturated cells ($x10^3/\mu l$)	0.82 (0.30-1.53)	0.80 (0.30-1.50)	1.00 (0.89-1.12)	1.06 (0.91-1.23)
Neutrophil cluster mean x	61.7 ± 6.0	61.7 ± 6.3	0.96 (0.86-1.06)	0.97 (0.85-1.11)
Neutrophil cluster mean y	70.0 ± 6.0	70.0 ± 6.4	1.01 (0.90-1.14)	0.95 (0.84-1.07)
Ky	97.36 ± 2.38	97.25 ± 2.41	0.97 (0.86-1.09)*	0.90 (0.78-1.04)
Peroxidase x sigma	0.01 ± 0.12	0.01 ± 0.12	1.10 (1.03-1.18)*	1.06 (0.96-1.18)
Peroxidase y mean	18.1 ± 0.7	18.1 ± 0.7	1.61 (1.46-1.77)*	1.10 (0.96-1.27)
Peroxidase y sigma	8.11 ± 1.07	8.12 ± 1.05	1.79 (1.61-1.99)*	1.16 (1.01-1.33)*
Lobularity index	1.9 (1.0-2.1)	1.9 (1.0-2.1)	0.92 (0.83-1.01)	1.03 (0.89-1.20)
Lymphocyte/large unstained cell threshold	45.0 ± 1.6	45.1 ± 1.6	1.16 (1.08-1.24)*	1.07 (1.00-1.17)
Perox d/D	0.9 (0.9-1.0)	0.9 (0.9-1.0)	0.91 (0.85-0.97)*	1.16 (0.85-1.56)
Blasts (%)	0.77 ± 0.49	0.77 ± 0.49	1.34 (1.22-1.47)*	1.07 (0.93-1.23)
Polymorphonuclear ratio (%)	1.0 (0.99-1.0)	1.0 (0.99-1.0)	0.77 (0.65-0.90)*	0.99 (0.84-1.15)
Neutrophil x channel mode	27.5 ± 3.6	27.4 ± 3.7	0.91 (0.82-1.02)	1.08 (0.93-1.25)
Mononuclear central x channel	14.1 (13.0-15.0)	14.1 (13.0-15.0)	0.80 (0.74-0.88)*	1.12 (0.95-1.32)
Mononuclear central y channel	14.5 ± 1.1	14.5 ± 1.1	0.79 (0.73-0.87)*	1.04 (0.89-1.20)
Mononuclear polymorphonuclear valley	18.0 (18.0-20.0)	18.0 (18.0-20.0)	0.69 (0.61-0.77)*	1.06 (0.94-1.21)
Red Blood Cell Related				
RBC count ($x10^6/\mu l$)	4.30 ± 0.52	4.33 ± 0.52	0.59 (0.53-0.66)*	0.93 (0.81-1.08)
Hematocrit (%)	40.9 ± 6.2	41.0 ± 4.2	0.51 (0.45-0.59)*	0.78 (0.65-0.93)*
Mean corpuscular hgb (MCH; pg)	30.4 ± 2.1	30.3 ± 2.0	0.83 (0.75-0.92)*	1.03 (0.89-1.19)
Mean corpuscular hgb conc. (MCHC; g/dl)	33.4 ± 5.7	33.4 ± 5.7	0.86 (0.80-0.92)*	0.91 (0.82-1.01)
RBC hgb concentration mean (CHCM; g/dl)	35.1 ± 1.3	35.2 ± 1.3	0.53 (0.49-0.59)*	0.90 (0.78-1.04)
RBC distribution width (RDW; %)	13.4 ± 1.2	13.4 ± 1.2	1.48 (1.42-1.55)*	1.26 (1.14-1.40)*
Hgb distribution width (HDW; g/dl)	2.7 ± 0.3	2.7 ± 0.3	1.52 (1.39-1.66)*	1.26 (1.12-1.43)*
Hgb content distribution width (CHDW; pg)	3.8 ± 0.4	3.8 ± 0.4	1.44 (1.37-1.51)*	1.19 (1.07-1.33)*
Normochromic/Normocytic RBC count ($x10^6/\mu l$)	3.65 ± 0.39	3.66 ± 0.39	0.64 (0.60-0.68)*	0.89 (0.78-1.01)
Macrocytic RBC count ($x10^6/\mu l$)	0.01 (.01-.03)	0.01 (.01-.03)	1.76 (1.55-2.00)*	1.03 (0.89-1.20)
Hypochromic RBC count ($x10^6/\mu l$)	0.006 (0.001-0.002)	0.005 (0.001-0.002)	1.12 (0.99-1.27)	1.18 (1.00-1.38)
Platelet Related				
Plateletcrit (PCT; %)	0.18 ± 0.05	0.18 ± 0.06	1.15 (1.04-1.27)*	0.99 (0.85-1.14)
Mean platelet concentration (MPC; g/dl)	27.1 ± 1.7	27.0 ± 1.7	0.75 (0.68-0.83)*	0.97 (0.84-1.12)
Platelet conc. distribution width(PCDW; g/dl)	5.6 ± 0.4	5.7 ± 0.4	0.95 (0.84-1.06)	0.95 (0.83-1.01)
Large platelets ($x10^3/\mu l$)	4 (3-6)	4 (3-6)	1.10 (0.94-1.28)	1.10 (0.91-1.34)
Platelet clumps ($x10^3/\mu l$)	41.5 ± 37.1	42.4 ± 36.1	1.00 (1.00-1.00)	1.00 (1.00-1.00)

Patterns identify patient risk for myocardial infarction (MI) or death. High-risk patterns (Table 2.2A) were satisfied by patients that were more likely to experience death (>3.6-fold risk) or myocardial infarction (>1.4-fold risk) over the ensuing year, and low-risk patterns (Table 2.2B) were observed in patients less likely to experience death (<0.34-fold risk) or myocardial infarction (<0.57-fold risk). Remarkably, in general, patterns that were predictive of high- or low-risk for death demonstrate different general composition of variables compared to variables that are included in high- and low-risk patterns for MI.

Unique discriminating patterns in those who died included variables derived from multiple RBC- and WBC (peroxidase)-related parameters, as well as the level of C-reactive protein. High-risk patterns for MI included a wider variety of variables, including multiple RBC, WBC (peroxidase) and platelet parameters, traditional risk factors, and blood chemistries (Table 2.2A). Variables common to both high-risk death and MI patterns included age, hypertension, mean RBC hemoglobin concentration, hemoglobin concentration distribution width, hypochromic erythrocyte cell count, and perox Y sigma (a peroxidase-based measure of neutrophil size distribution). Variables that were shared between low-risk patterns for both death and MI risk included C-reactive protein levels, absolute neutrophil count, mean platelet concentration (a flow cytometry determined index of platelet granule content) and monocyte/polymorphonuclear valley (a measure of separation among clusters of peroxidase-containing cell populations). In general, the low-risk patterns for incident one-year death and MI risk are dominated by multiple diverse hematology analyzer variables of all three blood cell types and age.

Table 2.2A. High-risk Patterns in PEROX Model for One-year Death or Myocardial Infarction.

For each pattern, the number of patients satisfied by it (N), the event rate, hazard ratio (HR) and 95% confidence interval (CI) are shown for the Derivation cohort.

Death High Risk	Pattern	N	Death Rate	HR (95% CI)
1	Hemoglobin content distribution width > 3.93, & Cell hgb concentration mean < 35.07	815	13%	4.94 (3.88-6.30)
2	Hypochromic RBC count > 189, & Hemoglobin content distribution width > 3.93	658	13%	4.47 (3.48-5.73)
3	Mean corpuscular hgb concentration < 34.38, & Perox d/D < 0.89	466	14%	4.46 (3.42-5.81)
4	Hypochromic RBC count > 189, & Macrocytic RBC count > 192	588	13%	4.37 (3.39-5.64)
5	Mean corpuscular hgb concentration < 33.00, & Monocyte cluster X center < 14.38	422	14%	4.37 (3.33-5.74)
6	Age > 67, & Hematocrit < 36.45	515	13%	4.08 (3.13-5.32)
7	Monocyte/polymorphonuclear valley < 18.50, Perox cluster Y axis sigma > 9.48	474	13%	3.85 (2.93-5.07)
8	Monocyte cluster X center < 14.38, & Perox cluster Y axis mean > 19.02	494	12%	3.68 (2.80-4.85)
9	C-reactive protein > 13.75, & History of hypertension	531	12%	3.63 (2.77-4.76)
MI High Risk	**Pattern**	**N**	**MI Rate**	**HR (95% CI)**
1	Mean platelet component concentration > 27.89, & Potassium < 3.85	332	5%	2.17 (1.33-3.56)
2	Triglycerides < 130, & Age > 76	464	5%	1.94 (1.23-3.04)
3	RBC distribution width > 13.83, & Lymphocyte count > 1.75	371	5%	1.93 (1.18-3.17)
4	Hypochromic RBC count > 56, & Diabetes	1,212	4%	1.91 (1.37-2.68)
5	Body mass index < 24.7, & Neutrophil count < 3.58	446	4%	1.91 (1.20-3.03)
6	Systolic blood pressure > 150, & Hypertension	1,163	4%	1.89 (1.35-2.66)
7	Polymorphonuclear cluster x axis mode > 29.87, & RBC distribution width > 13.22	729	4%	1.80 (1.22-2.67)
8	Hgb concentration distribution width > 2.69, & Perox cluster y axis sigma > 8.59	842	4%	1.79 (1.23-2.61)
9	Platelet concentration component distribution width < 5.39, & Mean RBC hgb concentration < 34.69	870	4%	1.79 (1.23-2.60)
10	Mean RBC hemoglobin > 32.60, & Male	500	4%	1.78 (1.13-2.81)
11	Lymphocyte count < 0.96, & Potassium > 4.4	387	4%	1.73 (1.04-2.87)
12	Platelet concentration distribution width > 6.04, & Monocyte count > 0.46	119	4%	1.7 (0.71-4.06)
13	Neutrophil y cluster mean < 71.19, & Current smoker	447	4%	1.69 (1.04-2.74)
14	Mean platelet concentration > 23.19, & Basophil count > 0.12	178	3%	1.36 (0.61-3.03)

Table 2.2B. Low-risk Patterns in PEROX Model for One-year Death or Myocardial Infarction.

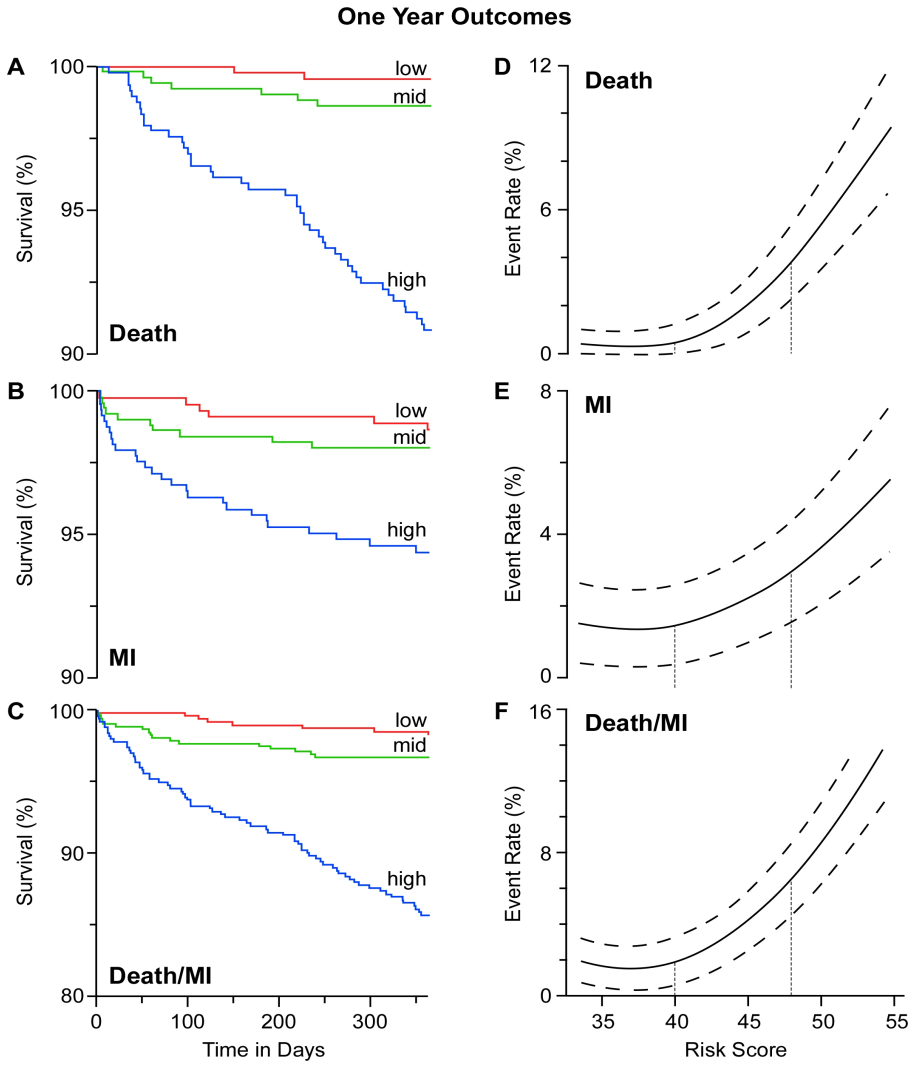
For each pattern, the number of patients satisfied by it (N), the event rate, hazard ratio (HR) and 95% confidence interval (CI) are shown for the Derivation cohort.

Death Low Risk	Pattern	N	Death Rate	HR (95% CI)
1	Cell hgb concentration mean > 35.07, & Hematocrit > 42.25	1,443	1%	0.18 (0.10-0.31)
2	Number of macrocytic cells < 192, & Age < 67	2,283	1%	0.22 (0.15-0.32)
3	Cell hgb concentration mean > 35.07, & RBC count > 4.42	1,494	1%	0.24 (0.15-0.38)
4	Mean platelet component concentration > 27.52, & Age < 67	1,651	1%	0.24 (0.16-0.38)
5	Perox cluster Y axis sigma < 8.10, & Age < 67	1,982	1%	0.26 (0.17-0.38)
6	C-reactive protein < 4.0, & Hematocrit > 42.25	1,688	1%	0.26 (0.17-0.40)
7	Hematocrit > 42.25, & Perox d/D > 0.89	1,972	1%	0.27 (0.18-0.40)
8	Monocyte/polymorphonuclear valley > 18.50, & Age < 67	1,750	1%	0.27 (0.18-0.41)
9	Cell hgb concentration mean > 35.07, & White blood cell count < 5.86	1,436	1%	0.30 (0.19-0.46)
10	Neutrophil count < 3.96, & Age < 67	1,697	2%	0.34 (0.23-0.49)
MI Low Risk	**Pattern**	**N**	**MI Rate**	**HR (95% CI)**
1	No history of cardiovascular disease, & RBC distribution width < 13.22	919	1%	0.31 (0.15-0.63)
2	Lymphocyte/Large unstained cell threshold < 44.50, & Percent blasts < 0.51	946	1%	0.34 (0.17-0.66)
3	Systolic blood pressure < 134, & Basophil count < 0.03	743	1%	0.34 (0.16-0.73)
4	Number of platelet clumps > 41, & Glucose < 92.5	782	1%	0.37 (0.18-0.76)
5	Hemoglobin concentration distribution width < 2.69, & Number of hypochromic cells < 14	891	1%	0.41 (0.22-0.77)
6	Number of hypochromic cells < 14, & Neutrophil count < 5.83	1,159	1%	0.43 (0.25-0.74)
7	Monocyte cluster x center < 12.70, & Neutrophil y cluster mean > 69.30	841	1%	0.44 (0.23-0.82)
8	Monocyte/polymorphonuclear valley > 14.50, & Creatinine < 0.75	910	1%	0.44 (0.24-0.81)
9	No history of cardiovascular disease, & Systolic blood pressure < 134	756	1%	0.44 (0.23-0.86)
10	Number of peroxidase saturated cells < 0.01, & Neutrophil count < 4.69	781	1%	0.47 (0.25-0.90)
11	High density lipoprotein > 59, & Mean platelet concentration < 28.56	830	1%	0.49 (0.27-0.90)
12	Monocyte cluster x center < 12.70, & C-reactive protein < 5.31	896	1%	0.49 (0.27-0.88)
13	Monocyte cluster x center < 12.70, & Basophil count < 0.07	961	1%	0.54 (0.31-0.93)
14	No history of cardiovascular disease, & Neutrophil x cluster mean < 66.07	1,261	2%	0.57 (0.36-0.92)

The PEROX risk score predicts incident one year risks for non-fatal MI and death. Within the Derivation Cohort, ROC curve for the PEROX risk score for the one-year death, MI and the composite of death/MI demonstrated an area under the curve of 80%, 66% and 75%, respectively. For the composite endpoint, the cut-point which maximizes the accuracy was identified, and this was virtually identical to the top tertile within the Derivation Cohort. Within the Validation Cohort, the PEROX risk score demonstrated comparable results to that observed in the Derivation Cohort for the prediction of the one-year endpoints of death, non-fatal MI, or the composite death/MI were 83%, 66%, and 76%, respectively.

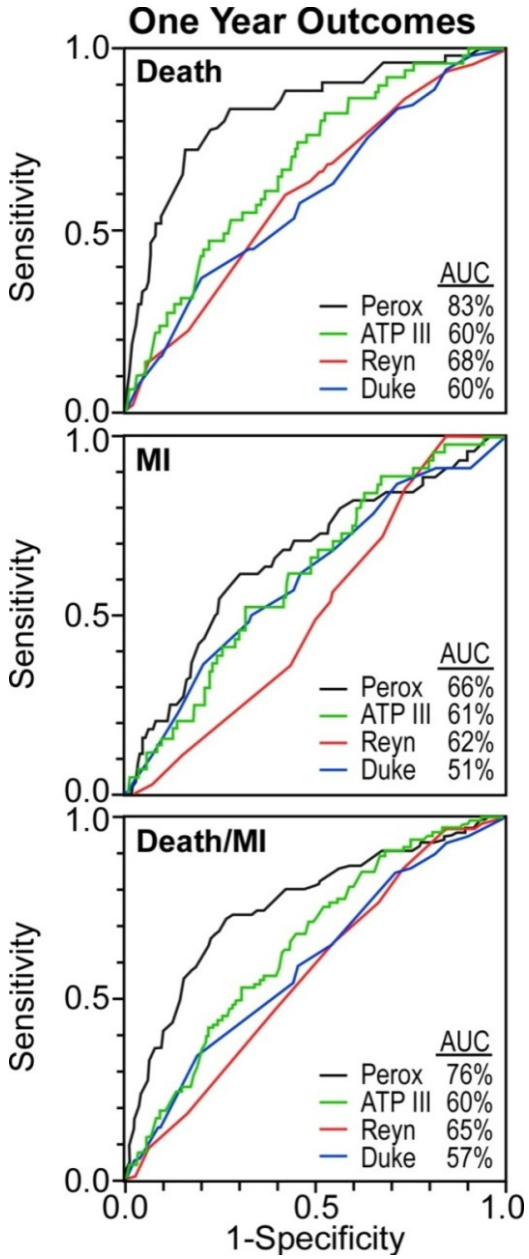
KM curves for the high, medium and low risk groups in the validation cohort are presented in Figures 2.1A-C. The log-rank test p-value < 0.001 for each individual outcome was used and shows that the survival distributions of these groups are significantly different. Figure 2.1D-F demonstrates the relationship between PEROX risk score vs. one-year event rates within the Validation Cohort. Strong tight positive associations were noted between increasing risk score and risk for experiencing non-fatal MI, death or the composite adverse outcome.

Figure 2.1. Kaplan-Meier Curves and Composite Risk for One-year Outcomes Based on Tertiles of PEROX Score in Validation Cohort.
Kaplan-Meier curves for cumulative probability of death (A), myocardial infarction (B), or either event (C) according to low, medium, and high tertiles of PEROX risk score. Spline curves (solid line) with 95% confidence intervals (dashed line) showing association between cumulative event (Y axis) for death (D), myocardial infarction (E), and death or myocardial infarction (F), for PEROX risk score (X axis) are shown. Vertical dotted lines indicate the tertile cut-points.



Relative performance of the PEROX risk score for accurate risk assessment and classification of patients. Next, the PEROX risk score performance within the Validation Cohort was compared with alternative clinically validated risk algorithms including ATP III, Reynolds, and Duke angiographic scoring systems using ROC curve analyses (Figure 2.2). The PEROX risk score demonstrated superior prognostic accuracy for one-year death (AUC=83%), MI (AUC=66%) or the composite of death or MI (AUC=76%) compared with each of the traditional scoring systems examined (Figure 2.2). The utility of the PEROX risk score in predicting outcomes within the Validation Cohort was also examined within primary prevention (patients with a history of coronary artery disease) and secondary prevention (patients with no prior history of coronary artery disease) subgroups, as well as patients stratified based upon presence versus absence of diabetes. Similar prognostic accuracy for the PEROX risk score was observed within these subgroups within the Validation Cohort as indicated by the area under the curve for the primary and secondary prevention subpopulations (77% and 74%, respectively), and those with and without diabetes (74% and 75%, respectively). Separate analyses of the traditional risk scores within the primary and secondary prevention subpopulations demonstrated similar results to that observed within the entire Validation Cohort (i.e. markedly reduced AUC compared to the PEROX risk score).

Figure 2.2. Comparison of Classification of One-year Death (A), Myocardial Infarction (B) and Death or Myocardial Infarction (C) According to PEROX risk score, and Validated Clinical Using Risk Scores in Validation Cohort.
Receiver operator characteristics curves plotting sensitivity (X axis) and 1-specificity (Y axis) are shown for PEROX (N=1,474 patients; black line), ATP III (N=1,474 patients; green line), Reynolds Risk (N=1,403 patients; red line), and Duke Angiographic Risk (n=1,129 patients; blue line) scores. For each death, myocardial infarction and either outcome (Death/MI), inset within the figure is the area under the curve (AUC) for each risk score.



The potential clinical utility of the PEROX risk score was next compared to traditional risk algorithms in stratifying patients into risk groups. The overall population distribution between low-, intermediate- and high-risk categories for both ATP III and the PEROX risk score for the composite outcome of non-fatal MI or death are illustrated in the top panel in Table 2.3. Of note, PEROX risk score categories were defined by tertiles and thus approximately equal proportions of subjects within the entire cohort are stratified into each risk bin. In contrast, 45% of subjects were categorized as having low (<10% 10 year) ATP III risk, and only 19% of the cohort were stratified to high (>20% 10 year) risk. The one-year event rate for non-fatal MI or death was 11% versus 5% among subjects stratified within high versus low ATP III risk categories, a risk gradient of over 2-fold. By comparison, the one-year event rate for non-fatal MI or death among subjects stratified within high versus low PEROX risk categories was 14% versus 2%, a risk gradient of over 7-fold.

Of those who experienced death or MI within one-year, only 31% were identified as high risk by ATP III with 69% of the population misclassified (Table 2.3). In contrast, 70% of subjects experiencing death or MI within one-year were identified within the top high-risk tertile of the PEROX risk score. Thus, within a non-symptomatic cardiology patient population, the PEROX risk score more accurately classified subjects at high risk than the traditional ATP III global risk score. Of those subjects within the cohort classified as low ATP III risk, 36% were misclassified and experienced an event, yet this number drops to 8% of subjects in the low-risk PEROX tertile. The reasons for the high event rate within the "low-risk" ATP III group likely reflect the inability to use this score in higher risk cardiology patients, which includes secondary prevention subjects and

those whose lipid and blood pressure levels are normalized by aggressive use of medication. In separate analyses, we performed similar comparisons within subpopulations of the Validation Cohort stratified based upon primary versus secondary prevention status at the time of enrollment. Again, PEROX risk tertiles demonstrated improved accuracy in the correct classification of risk at both high and low ends of the spectrum for both primary prevention and secondary prevention subjects (Table 2.3). Finally, in additional analyses we explored the impact of diabetes on comparisons between the PEROX risk score and ATP III since the latter treats diabetes as a risk equivalent (automatic 20% 10 year cardiovascular risk) but does not assign points for this in the initial risk score calculation. Performance analyses of subjects with and without diabetes for both PEROX and ATP III risk scores showed that the PEROX risk score consistently demonstrated improved risk stratification at both high and low ends of the risk spectrum in both non-diabetics and diabetics alike (Table 2.3).

Table 2.3. Risk Stratification Using ATP III and PEROX risk score.
The population was stratified by PEROX score (based on tertile) and by ATP III (based on low (<10%), medium (10 to <20%) and high (≥20%) 10 year cardiovascular event rate). The % of events indicates total percent of major adverse cardiac events in the strata.

		ATP III	PEROX
% Population	Low Risk	45%	31%
	Medium Risk	36%	35%
	High Risk	19%	34%
Primary and Secondary Prevention			
High Risk	% Events in High-risk Strata	31%	70%
	Event Rate in High-risk Strata	11%	14%
Low Risk	% Events in Low-risk Strata	36%	8%
	Event Rate in Low-risk Strata	5%	2%
Primary Prevention			
High Risk	% Events in High-risk Strata	19%	52%
	Event Rate in High-risk Strata	8%	13%
Low Risk	% Events in Low-risk Strata	48%	18%
	Event Rate in Low-risk Strata	3%	2%
Secondary Prevention			
High Risk	% Events in High-risk Strata	34%	75%
	Event Rate in High-risk Strata	11%	14%
Low Risk	% Events in Low-risk Strata	32%	6%
	Event Rate in Low-risk Strata	6%	2%
Diabetic			
High Risk	% Events in High-risk Strata	29%	78%
	Event Rate in High-risk Strata	13%	14%
Low Risk	% Events in Low-risk Strata	36%	5%
	Event Rate in Low-risk Strata	8%	3%
Non-Diabetic			
High Risk	% Events in High-risk Strata	33%	63%
	Event Rate in High-risk Strata	10%	13%
Low Risk	% Events in Low-risk Strata	35%	12%
	Event Rate in Low-risk Strata	4%	2%

2.3 CHRP(PEROX) and CHRP Risk Scores

We now test the hypothesis that using only information generated from the analysis of whole blood with a hematology analyzer during the performance of a traditional CBC with differential including peroxidase-based measurements, high- and low-risk patterns may be identified allowing for the development of a Peroxidase-based Comprehensive Hematology Risk Profile (CHRP(PEROX)), a single laboratory value that accurately predicts incident risks for non-fatal MI and death in subjects. This model is very similar to the PEROX model, with the exception of the inclusion of clinical and laboratory measures. We also built another model which uses only parameters from any generalized hematology analyzer which does not use any of the parameters derived from peroxidase. The hypothesis here was that we can accurately predict risk of one year myocardial infarction or death based solely on the components of blood. The goal is to test whether CHRP and CHRP(PEROX) risk scores can be used by clinicians along with other cardiovascular risk factors to assess the treatment for a patient.

2.3.1 Analysis

We use the same dataset as discussed above in the PEROX section with a Derivation Cohort with N=5,895 subjects and a Validation Cohort with N=1,473 subjects. However, in this model, we use only a subset of the variables used for building the PEROX model. The methodology used in this case is very similar to the PEROX model, with the exception that instead of using the entire Derivation cohort for building the models, we use as high risk samples all the samples with events and as low-risk samples those with maximum stenosis < 50% (controls). For the case of building LAD model for Death we had 242 cases, 1678 controls, and for MI 148 cases patients, and 1694 controls. We can

build more robust patterns if we restrict the dataset to the samples at extreme risk. The model is tested on the entire Validation cohort.

2.3.2 Results

The next stage in the analysis was to take these findings and attempt to simplify and make them clinically implementable by using only parameters routinely available from i) whole blood analysis on a i) peroxidase-based or ii) non peroxidase-based hematology analyzer. Using whole blood analysis on a peroxidse based hematology analyzer, 25 high-risk and 34 low-risk binary patterns were identified using the Derivation Cohort. These patterns were distilled down, the CHRP (PEROX) risk score which manifested highly accurate prognostic value. The input variable list used in CHRP (PEROX) was then simplified to include only variables that are generated using a general (non-peroxidase-based) hematology analyzer, we developed the CHRP risk score, which consisted of 19 high-risk and 24 low-risk patterns.

Independent prospective testing of the CHRP(PEROX), CHRP risk scores within the Validation Cohort revealed superior predictive power (72%, 71%, respectively) for prediction of one-year risk of death or MI compared with traditional cardiovascular risk factors, laboratory tests, as well as clinically established risk scores including Adult Treatment Panel III (60%), Reynolds (64%), and Duke angiographic (63%) scoring systems. Even though, these risk scores have lower predictive value than the PEROX risk score (76%), they can be translated to clinical use much more effectively because they only require variables from the hematology analyzer and do not require input of data such as demographics, lipid panel etc. Superior prognostic accuracy for prediction of 1 year incident MI and death was also observed with CHRP in both primary and secondary prevention subgroups, diabetics and non-diabetics alike, and even amongst those with no

evidence of significant coronary atherosclerotic burden (< 50% stenosis in all major coronary vessels) at time of recent cardiac catheterization. Table 2.4 presents details of the ROC accuracy for CHRP(PEROX) and CHRP risk score for all patients, and also in the primary and secondary cohorts for the validation set We also present the ROC accuracy for several other traditional cardiovascular risk measures in the same table for comparison. Table 2.5 and 2.6 presents the unadjusted and adjusted hazard ratios for CHRP(PEROX) and CHRP risk scores respectively, compared to the important and known cardiovascular risk measures. This table shows both the importance of these risk scores individually, and also that they provide independent prognostic value when computed with other risk factors in a multivariate logistic regression model. The list of high and low-risk patterns, ROC curves, KM and risk profile plots for both the risk scores are given in the Appendix.

Table 2.4. Area under the ROC curve (%) for CHRP (PEROX) and traditional cardiovascular risk parameters

	Dth/MI-1	Dth-1	MI-1
CHRP(PEROX)	72.3	77.3	65.2
CHRP(PEROX) – primary prevention	76.0	78.5	70.1
CHRP(PEROX) – secondary prevention	70.5	62.3	76.6
CHRP	70.5	77.5	60.9
CHRP – primary prevention	82.6	80.8	84.8
CHRP – secondary prevention	68.2	76.4	57.5
Age	62.7	68.2	54.7
Male	49.6	47.6	51.7
Diabetis mellitus	57.0	57.8	55.6
Hypertension	57.2	55.4	59.3
Current smoking	50.8	50.1	52.5
Past smoking	51.2	54.4	46.8
Total cholesterol	48.5	47.8	50.1
Low density lipoprotein	48.3	47.4	50.3
High density lipoprotein	45.2	49.2	39.6
Triglycerides	52.1	47.2	58.9
Glucose	55.9	52.8	58.6
Creatinine	64.5	67.9	57.9
HemoglobinA1C	50.5	47.5	54.4
History of cardiovascular disease	59.2	58.9	59.1
History of myocardial infarction	58.5	57.9	59.2
History of revascularisation	58.0	57.6	58.0
History of stroke	54.1	56.6	51.6
Maximum stenosis $\geq$ 50	59.6	59.5	59.3

Table 2.5. Hazard ratio of CHRP (PEROX) and traditional cardiovascular risk measures for tertiles

	1st tertile	2nd tertile	3rd tertile
CHRP (PEROX)	≤37.94	38.23-49.09	>49.17
Unadjusted	1	1.95 (1.43-2.68)	6.34 (4.79-8.40)
Adjusted†	1	1.71 (1.24-2.36)	4.98 (3.71-6.69)
Age	≤59.34	>59.34, ≤ 70	>70
Unadjusted	1	1.53 (1.18-1.98)	2.59 (2.04-3.28)
Adjusted†	1	1.36 (1.04-1.78)	1.88 (1.45-2.43)
LDL	≤82	>82, ≤ 110.8	>110.8
Unadjusted	1	0.67 (0.54-0.84)	0.75 (0.61-0.93)
Adjusted†	1	0.81 (0.65-1.02)	1.06 (0.85-1.33)
HDL	≤39	>39, ≤ 49	>49
Unadjusted	1	0.84 (0.68-1.04)	0.72 (0.58-0.91)
Adjusted†	1	0.91 (0.73-1.13)	0.80 (0.64-1.01)
Gender	Female	Male	
Unadjusted	1	1.05 (0.87-1.28)	
Adjusted†	1	0.94 (0.77-1.16)	
Hypertension	No	Yes	
Unadjusted	1	1.60 (1.27-2.02)	
Adjusted†	1	1.17 (0.93-1.48)	
Current Smoking	No	Yes	
Unadjusted	1	1.03 (0.79-1.35)	
Adjusted†	1	1.25 (0.93-1.68)	
Past Smoking	No	Yes	
Unadjusted	1	1.13 (0.93-1.37)	
Adjusted†	1	0.95 (0.77-1.17)	
Diabetes	No	Yes	
Unadjusted	1	1.79 (1.50-2.14)	
Adjusted†	1	1.40 (1.16-1.68)	

†Adjusted models contain CHRP(PEROX), age, LDL, HDL, gender, hypertension, current smoking, past smoking, and diabetes.

Table 2.6. One-Year Hazard Ratios of CHRP and traditional cardiovascular risk measures for entire cohort.

	1st tertile	2nd tertile	3rd tertile
CHRP	≤34.43	>34.98, ≤ 48.58	≥49.12
Unadjusted	1	1.80 (1.32-2.47)	5.22 (3.95-6.88)
Adjusted†	1	1.61 (1.17-2.22)	4.07 (3.04-5.46)
Age	≤59.34	>59.34, ≤ 70	>70
Unadjusted	1	1.53 (1.18-1.98)	2.59 (2.04-3.28)
Adjusted†	1	1.39 (1.06-1.82)	1.94 (1.50-2.51)
LDL	≤82	>82, ≤ 110.8	>110.8
Unadjusted	1	0.67 (0.54-0.84)	0.75 (0.61-0.93)
Adjusted†	1	0.79 (0.63-0.99)	1.06 (0.84-1.32)
HDL	≤39	>39, ≤ 49	>49
Unadjusted	1	0.84 (0.68-1.04)	0.72 (0.58-0.91)
Adjusted†	1	0.87 (0.70-1.08)	0.75 (0.59-0.94)
Gender	Female	Male	
Unadjusted	1	1.05 (0.87-1.28)	
Adjusted†	1	1.12 (0.91-1.38)	
Hypertension	No	Yes	
Unadjusted	1	1.60 (1.27-2.02)	
Adjusted†	1	1.13 (0.89-1.43)	
Current Smoking	No	Yes	
Unadjusted	1	1.03 (0.79-1.35)	
Adjusted†	1	1.21 (0.90-1.63)	
Past Smoking	No	Yes	
Unadjusted	1	1.13 (0.93-1.37)	
Adjusted†	1	0.99 (0.81-1.22)	
Diabetes	No	Yes	
Unadjusted	1	1.79 (1.50-2.14)	
Adjusted†	1	1.44 (1.19-1.73)	

†Adjusted models contain CHRP, age, LDL, HDL, gender, hypertension, current smoking, past smoking, and diabetes.

2.4 Discussion

Studies by our group [71] and Buffon, *et al.* [72] previously implicated intracellular peroxidase content of leukocytes in cardiovascular risk stratification of patients. Based upon these preceding observations and the numerous mechanistic links between myeloperoxidase [54, 73], monocytes [74, 75] and neutrophils [76] for atherosclerosis and acute coronary syndromes, we hypothesized that data derived from a peroxidase-based hematology analyzer would harbor clinically useful information related to cardiovascular disease prognosis. As the analyses unfolded, it became clear that additional clinical information from alternative hematology measures in addition to peroxidase/leukocyte related parameters provided significant additional prognostic value.

Review of the components contributing to the high- and low-risk patterns that contribute to the PEROX risk score reveals that a striking number of RBC-related phenotypes, and a small number of platelet-related parameters as well, provide prognostic value in identifying individuals at both increased and decreased risk for near term adverse cardiac events. The present studies thus reveal that alterations in multiple subtle phenotypes within WBC, RBC and platelet lineages in blood reflect processes linked to vascular health and cardiovascular risk. Moreover, each of these hematopoietic lineages shows numerous mechanistic links to cardiovascular disease pathogenesis and involvement in acute complications.

The present results also indicate that addition of a common peroxidase-based hematology analyzer that rapidly generates a complete blood cell count (CBC) and differential to the clinical assessment of a stable cardiology patient can be used to dramatically improve the accuracy with which subjects can be risk classified at both the high- and low-risk ends of the spectrum. The holistic hematology analyzer data collected

provides a broad spectrum of novel data from which can be recognized patterns, like fingerprints, providing clinically relevant information to the evaluation of cardiovascular risk in subjects.

The performance of the PEROX risk score in patients was surprisingly accurate given the population examined was comprised of stable subjects and the relatively short endpoint of one-year outcomes used. This contrasts with ATP III and Reynolds risk scores, both of which were developed using a long-term (10 year) outcome. In fairness to both ATP III and Reynolds, it is important to also note that neither of these traditional risk scores was developed in high-risk and heavily medicated populations, but rather, in predominantly untreated healthy populations for the purpose of community risk screening. However, the very nature of how these clinical tools were developed limits their clinical utility in treated populations in which co-morbidities are intervened upon. Another surprising finding in the present studies is how the PEROX score, which does not include angiographic data, outperforms the Duke score, which includes angiographic measures of cardiovascular disease. This observation strongly underscores the growing appreciation that atherosclerosis is more than a plumbing problem - it is a systemic disease - with parameters in the blood combined with biochemical profiles of systemic inflammation being strongly linked to disease pathogenesis. While some of the patterns identified as low-and high-risk traits within subjects are of unclear biological meaning, the majority are comprised of elements with recognizable mechanistic connections to disease pathogenesis. Moreover, as a group, all patterns reported appear to be robust, reproducible and present in multiple independent samplings of the cohort. The identification of reproducible high- and low-risk patterns amongst the clinical, laboratory

and hematological parameters monitored further indicates the presence of underlying complex relationships between multiple hematology parameters, clinical and metabolic parameters, and cardiovascular disease pathogenesis.

Much interest focuses on the idea that array-based phenotyping will play an ever increasing role in the future of preventive medicine, serving as a powerful method to improving risk classification of subjects and, ultimately, individualized tailoring of therapies. Rather than utilize research-based arrays (genomic, proteomic, metabolomic, expression array) that are no doubt powerful and extremely useful, we decided instead to utilize a robust, clinically validated high-throughput workhorse of clinical laboratory medicine, a hematology analyzer. The hematology analyzer selected had the added advantage of being a flow cytometer that uses *in situ* peroxidase cytochemical staining for identifying and quantifying leukocytes, and was therefore based on a phenotypic screen relevant to disease pathogenesis.

While the PEROX risk score developed here should only be considered proof of concept, the holistic approach taken illustrates a powerful message - that in the outpatient cardiology clinic setting using only clinical information routinely available plus a drop of blood, utilization of a broad phenotypic array based approach can permit rapid development of a precise risk score that provides markedly improved prognostic value of near-term relevance. Several alternative populations will be particularly interesting to examine using the PEROX risk score. For example, it will be of interest to explore whether risk stratification of a healthy community-based population is accurately predicted by the PEROX risk score, or whether addition of traditional risk factors that failed to remain significant in the present cohort (e.g. LDL cholesterol) would provide

added risk assessment in a predominantly untreated population. Similarly, examination of patients presenting with chest pain and suspected acute coronary syndromes represents a particularly attractive cohort to monitor, given the high throughput nature of the hematology analyzer (seconds per sample). One might hypothesize that additional platelet parameters, for example, might add to rapid risk screening in such a cohort. The results from the present studies suggest that expanded use of more comprehensive hematology analyzer profiling of blood holds promise for improved risk assessments and monitoring of therapeutic responses in the future.

We have also addressed the issue of developing a risk score that could be translated for effective clinical use. Using only data from a peroxidase-based hematology analyzer for whole blood analysis generates a spectrum of data from which high and low risk patterns can be identified for predicting a subject's risk for experiencing major adverse cardiac events. A composite single value was built based upon these patterns, the peroxidase based-Comprehensive Hematology Risk Profile (CHRP(PEROX)), which accurately predicts incident risks for non-fatal MI and death in subjects, and accurately classifies patients for both high and low near-term (one year) cardiovascular risks. Multivariate logistic regression analysis shows that the CHRP(PEROX) is a strong predictor of risk independent of traditional cardiac risk factors and laboratory markers in subjects. Moreover, CHRP(PEROX) provides strong prognostic value even within subjects who show no significant angiographic evidence of atherosclerosis on recent cardiac catheterization.

Further we built another risk score: CHRP, which uses only parameters from a generalized hematology analyzer, which has comparable accuracy to the CHRP(PEROX)

risk score. Additionally, the CHRP also shows additional prognostic value when modeled with traditional risk factors using multivariate logistic regression. The most advantageous feature is that generation of the CHRP risk score does not require any special instruments or tests.

CHRP(PEROX) and CHRP risk scores can be very effectively translated to clinical use by using an additional software patch in existing hematology analyzer. This can be a very cheap and effective method for clinicians to assess one-year cardiovascular risk for patients. Moreover, existing risk scores, like the ATP III, Reynolds, and Duke scores can be used in a complementary fashion with the CHRP(PEROX) and CHRP risk scores to identify patients at high and low risk with higher prognostic value.

Chapter 3

Identification of extremal risk groups in patients undergoing coronary artery bypass graft (CABG) surgery[2]

3.1 Introduction

Blockage in the coronary arteries is the most common cause for myocardial infarction (MI). Coronary artery bypass grafting (CABG) is an invasive heart surgery which involves rerouting or bypassing blocked coronary arteries (arteries which supply oxygen to the muscles of the heart) with arteries grafted from other parts of the body. Several studies like the Framingham study [77] use traditional clinical cardiovascular risk factors for building long-term (10-year) risk profiles. In another study, exercise stress testing variables were used to predict cardiovascular risk using Logical Analysis of Data (LAD) patterns [78]. In this study, we want to predict the mortality risk after CABG surgery based on clinical measurements. The hypothesis here is that we can identify long term mortality risk for patients based on clinical variables collected at the time of the CABG. The goal would be to identify the most predictive variables to gain a better understanding of the disease.

The dataset consists of 15,586 patients who underwent coronary artery bypass grafting. All of these patients had isolated CABG, meaning that that this was the only major cardiac surgical procedure done. These patients were also "primary", meaning that this was their first bypass surgery. These surgeries were elective, as opposed to

[2] Based on collaborations with Gabriela Alexe (Broad Institute of MIT & Harvard), Endre Boros (RUTCOR, Rutgers University), Michael Lauer, and Eugene Blackstone (Cleveland Clinic). This chapter is part of three working manuscripts [4], [5], [79].

emergencies, patients were clinically stable and were referred for surgery for treatment of angina pectoris (chest pain due to the heart getting inadequate blood supply during times of stress) and documented severe coronary artery disease. The clinical measurements were collected at the time of surgery. This study was conducted from 1990 until 2003. The patients were followed up after surgery to collect mortality information. For each patient several clinical measurements were recorded along with time to death or censoring. Table 3.1 presents a description of the clinical covariates used for this study. Patients were marked as censored if they did not have an event until the end of the study or they relocated and did not come to the hospital, etc. The time to censoring for patients is the last time when the hospital could contact them. Of the 15,586 patients there were 3,854 deaths in this data.

Table 3.1. Clinical measurements collected for 15,586 patients who underwent coronary artery bypass surgery (1990-2003).

	Description
Dead	Death occurred during follow-up; 0: no, 1: yes
iv_dead	Interval of follow-up for death or censoring in years.
Age	Age at the time of surgery
Male	Gender, 1: male, 0: female
iv_opyrs	Interval between 1990 and the surgery in years
Crcl	Calculated creatinine clearance
hx_cva	History of stroke; 0: no, 1: yes
Black	Ethnicity is black; 0: no, 1: yes
Asian	Ethnicity is asian; 0: no, 1: yes
Hispanic	Ethnicity is hispanic; 0: no, 1: yes
Bmi	Body Mass Index
Ita	Number of internal thoraxic artery grafts placed

iptca_l6	Percutaneous procedure less than 6 hours prior to surgery; 0: no, 1: yes
ithrl_l6	thrombolytic therapy less than 6 hours prior to surgery; 0: no, 1: yes
Recentmi	Recent myocardial infarction. 0: none; 1: >21 days ago; 2: 8 - 21 days ago; 3: 1 - 7 days ago; 4: 6 - 24 hours ago; 5: < 6 hours ago.
Unstbang	Unstable angina prior to surgery. 0: no, 1: yes.
Stabling	Stable angina prior to surgery. 0: no, 1: yes.
Ccs	Canadian Cardiac Society Class of angina; ranges from 0 to 4, where 0 is no angina and 4 is very severe angina.
Nyha	New York Heart Association Class, a measure of dyspnea; ranges from 1 to 4, where 1 implies no real dyspnea, whereas 4 is severe dyspnea.
Nyhagccs	Variable indicating that NYHA class is worse than CCS class; 0: no, 1: yes.
Iabphemo	An intra-aortic balloon pump was placed prior to surgery for hemodynamic instability; 0: no, 1: yes.
Iabpptca	Angio-aortic balloon pump was put in as an elective for support of a patient undergoing a percutaneous procedure; 0: no, 1: yes.
iabp_ang	Angio-aortic balloon pump was placed because of unstable angina; 0: no, 1: yes.
Iabp_oth	An intra-aortic balloon pump was inserted prior to surgery for an indication other than the three just listed(iabphemo,iabpptca,iabp_ang); 0: no, 1: yes.
Insulin	Insulin treated diabetes; 0: no, 1: yes; missing values (3%) are set to 0.
Niddm	Non insulin treated diabetes. 0: no, 1: yes; missing values (only about 3%) are set to 0.
hx_htn	History of hypertension; 0: no, 1: yes.
hx_pvd	Presence of peripheral vascular disease; 0: no, 1: yes.
Smoking	Smoking history. 0: never; 1: past only; 2: current.
hx_ptca	Prior history of a percutaneous revascularization; 0: no, 1: yes.
Lmt	Maximum percent stenosis of the left main coronary artery
Lad	Maximum percent stenosis of the left anterior descending coronary artery
Rca	Maximum percent stenosis of the right coronary artery
Lcx	Maximum percent stenosis of the left circumflex coronary artery
lvf_cath	This is a measure of severity of left ventricular dysfunction. 1:normal left ventricular function, whereas 4:severe left ventricular dysfunction, 2.5:missing.
miss_lvf	Indicates that data on LV function are missing; 0:no, 1:yes.

The main difficulties with this dataset were the lack of informative variables and the fact that there are very few events (25%). For analysis, the dataset was first split into training and test sets based on random stratified sampling. The entire dataset was split

into 240 groups based on matching sets of clinical variables, and then $1/10^{th}$ of the data was chosen randomly from each of the groups to form the test set The reason for using random stratified sampling, is to ensure that the test set is representative of the training set. The aim of this study is to build prognostic models for predicting a patient's mortality risk, identifying important individual and combinatorial features (patterns) for distinguishing patients at high risk from those at low risk.

The data was extremely noisy, i.e. the clinical measurements for the high and low risk patients were very similar. In order to extract the main signal from the data, we designed a score to identify "confusing patients". These are high (low)-risk patients who have very similar measurements to a large proportion of patients in the opposite class, i.e. the low (high)-risk group. Once we had eliminated the confusing patients, we applied the Logical Analysis of Data (LAD) methodology to identify combinatorial patterns of high degree to distinguish between extremal risk groups (those patients who died very early on in the study, compared to those who lived until very long). A high-risk patient is defined as one who died before 3 years after CABG, while a low-risk patient is one who survived for at least 9 years after CABG. The reason for choosing these thresholds is because most patients with events >3 and <9 years had very similar profiles. Note that those patients who were censored before 3 years are not part of the data.

3.2 Methods and Results

We designed the following experiment: Identify confusing samples, i.e. samples belonging to one class, but which look very similar to a large proportion of samples in the other class. These samples tend to confuse the classification model and make the results

less robust. We eliminate these confusing patients and build prognostic models to predict mortality-risk.

Our first approach was to build an ensemble of two-class classification models for different pairs of high and low-risk groups. We defined a high-risk(h) class as those observations who had an event before h years. While, a low-risk(l) class as those patients who survived at least l years. We built Logical Analysis of Data (LAD) and Support Vector Machines (SVM) models for all pairs of high-risk(h) vs. low-risk(l) classes, where h=1,…,7 and l=7,…14. Thus for each of the LAD and SVM classifiers we built 56 models. From each model, we recorded a risk score for each patient (ranges from 0 to 1), the risk level increasing as the score increases. Finally, we integrate these scores by applying separately Cox regression on scores generated by LAD and SVM. The concordance accuracy (c-index) of the predicted survival on the test set was 71%, 73% for LAD, and SVM respectively. When we apply only Cox regression on the data, we obtain 75% c-index.

We noticed here that the performance for the ensemble models proposed by us does not work as well as Cox regression using c-index as the performance measure. The reason for this could be (i) c-index is a biased measure, and gives a perfect score to those models which give the highest risk for short time censored samples, which actually may have died sooner; (ii) variables have low predictive power.

Thus, we decided to focus on building a model to distinguish patients at extreme high-risk from extreme low-risk. We used accuracy as the performance measure in this case. Based on the ensemble of models described above, we selected high-risk(3) vs. low-risk(9) for classification. We eliminated confusing patients, and then built a classifier

using LAD to distinguish between the remaining high-risk(3) and low-risk(9) patients. Finally, we identified the top individual and combinatorial features based on prognostic value. In the following, we examine each of these step in more detail.

3.2.1 Identification of confusing samples

In order to identify the confusing observations in the data, we define the following score function, which is higher for samples which are more confusing. We want the score function for sample *i* to have the following properties.

- score is directly proportional to the number of samples of the opposite class surrounding it (within a circle of radius *T* around it);
- it is inversely proportional to the distance between *i* and samples of the opposite class which are close to *i*;
- it is inversely proportional to the time at which the sample *i* experienced an event (death). Samples with very early events should be penalized more if they are confusing.

Let *i* be a given high-risk (low-risk) patient in the dataset *S*. Let *D(i,j)* be the Euclidean distance of *i* from another patient *j*. Let us define the neighborhood $N_S(i)$ of patient *i* as the group of low-risk (high-risk) patients $j \in S$ whose distance from *i* is less than some fixed threshold *T*: $N_S(i) \coloneqq \{ j : D(i,j) < T \}$.

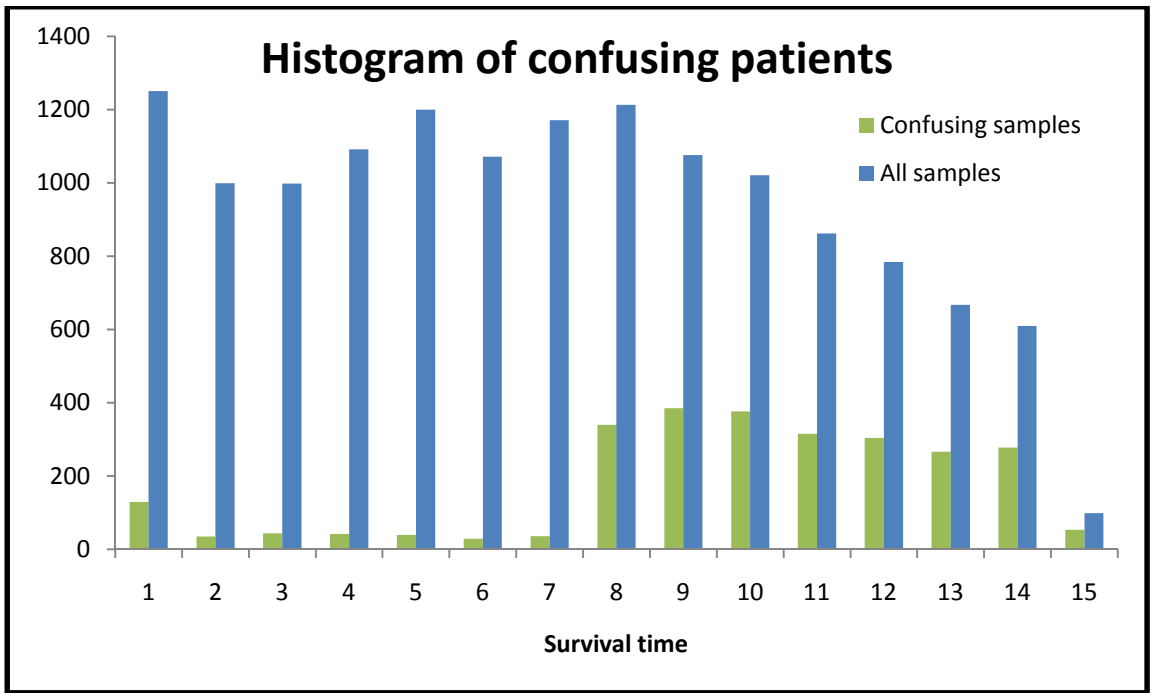
Let *M(h,l)* be a subset of the data consisting of patients who died before time *h* (high-risk patients), and patients who survived beyond time *l* (low-risk patients). We define the following score function based on the above properties.

$$Score(i) = \sum_{h=1}^{7} \sum_{l=7}^{14} \frac{|N_S(i)|}{\max_{k \in M(h,l)} |N_S(k)|} \cdot \left(1 - \text{mean}(\{\text{D(i, j)}|\text{j} \in \text{N}_\text{S}\text{(i)}\})\right) \cdot \frac{1}{\text{h}}$$

The data is normalized by dividing each variable by its maximum value before computing the distances. We use the threshold T = 0.5 for identifying neighbors. All high-risk observations with a score > 0.05, low-risk observations with score > 0 are considered to be confusing. We have chosen these particular parameters to show proof of concept that removing confusing patients enables us to build robust models. Ideally these parameters would have to be validated by running cross-validation experiments.

Using the above score function, we have identified 2,671 confusing patients in the training data. These confusing observations correspond to 208 high-risk (time to event≤3 years and event=1) and 1,592 low-risk (time to event>9 years) patients. The high-risk confusing patients had on average 25.22 low-risk samples in their neighborhoods, while the low-risk confusing samples had 4.31 high-risk samples in their neighborhoods. Figure 3.1 is a plot of the histogram of confusing patients (green) compared to all samples (blue) to show the relative distribution of confusing patients.

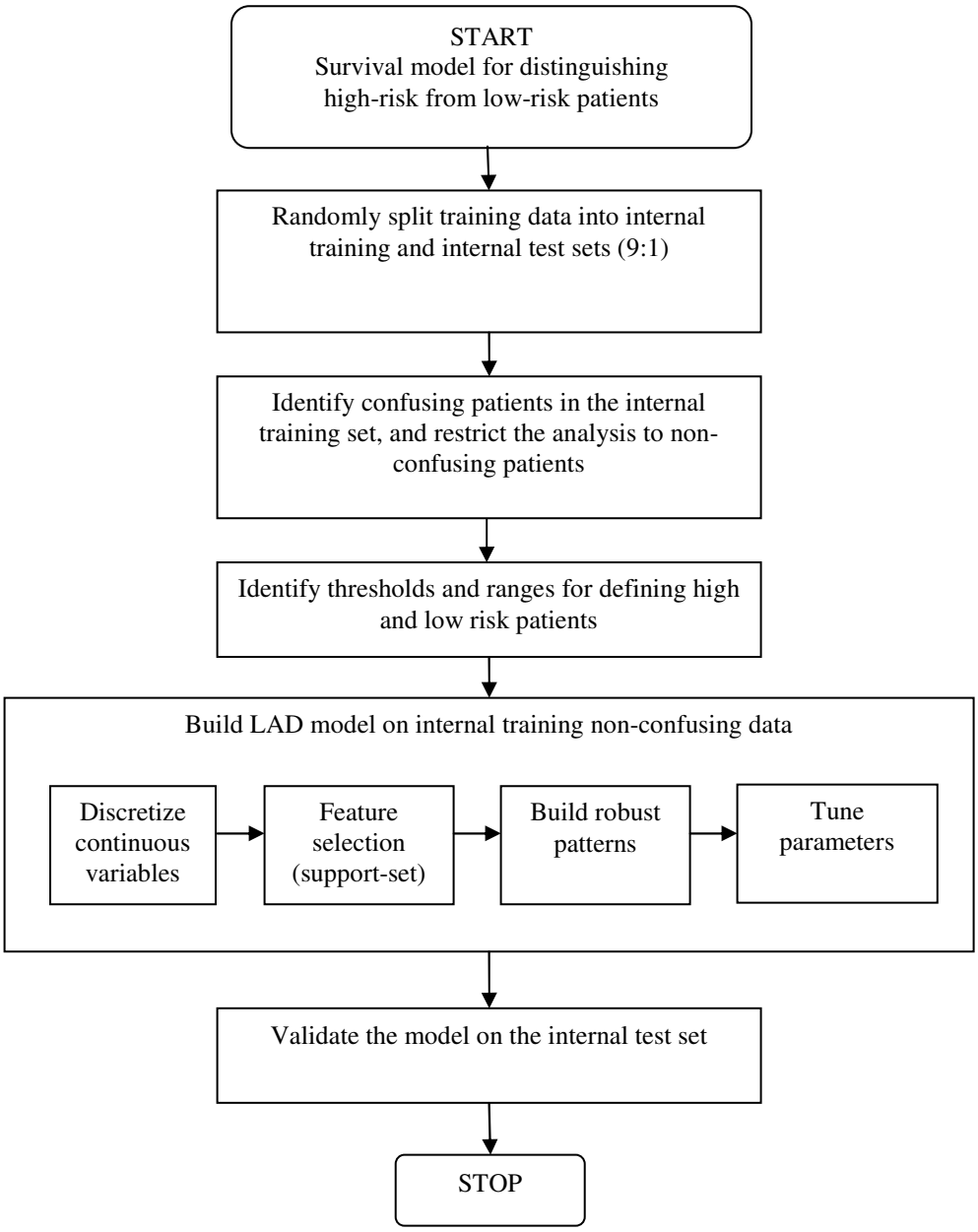
Figure 3.1. Histogram of survival time among the group of confusing patients.



3.2.2 Classification of high-risk versus low-risk groups

The training set is randomly split into an internal training set and an internal test set. The proportion of the split for internal training to internal test set is the same as the split between the original training and test sets (9:1). The reason for stratifying the training data into internal training and test sets is because we wanted to maintain the original test set for external validation, and use it only once we finalized the models. The confusing patients identified on the internal training set are removed. The analysis below is restricted to the non-confusing patients. We focused on building LAD classification models to distinguish between patients who died before 3 years (high-risk(3)) vs. patients who survived beyond 9 years (low-risk(9)). The parameters of the LAD model are tuned by running 3-fold cross-validation experiments. The flow chart in Figure 3.2 summarizes the steps used for predicting patients at extreme risk for (high-risk at 3 years vs. low risk at 9 years).

Figure 3.2. Flow chart for the procedure of building LAD survival model



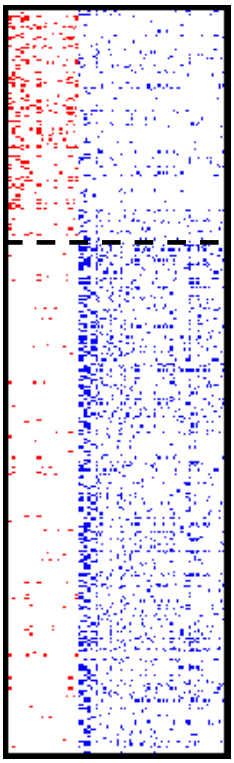
3.2.3 Performance of high-risk(3) vs. low-risk(9) LAD model

The LAD classification model built to classify high-risk(3) vs. low-risk(9) training data consists of 34 high-risk patterns and 70 low-risk patterns. Each of the high-risk patterns covers on an average 14% of the high-risk(3) patients, and 1.9% of the low-risk(9)

patients, while a low-risk pattern covers on an average 11% of the low-risk(9) patients, and 2.7% of the high-risk(3) patients.

Figure 3.3 is a heat map of the LAD patterns. The rows correspond to patients and the columns correspond to patterns, and a cell (i,j) is colored red if a high-risk pattern j covers patient i, and blue if a low-risk pattern covers this patient. The dotted line separates the high- and low-risk patients. From this visualization of the model on the test set, we can see that the model can accurately distinguish the high-risk patients from the low-risk ones. This shows that we have a very strong classification model.

Figure 3.3. Heat map of LAD patterns on high-risk(3) vs low-risk(9) test data.
The rows correspond to patients and the columns to patterns, and a cell (*i*,*j*) is colored red if a high-risk pattern *j* covers patient *i*, and blue if a low risk pattern covers the patient. The dotted line separates the high and low-risk patients.



The classification accuracy, sensitivity and specificity of the high-risk(3) versus low-risk(9) LAD model on internal training set, internal test set, on 10-fold cross-validation experiments, external test set and also on the full training set (including confusing patients) are presented in Table 3.2. We also built classification models using Support Vector Machines (SVM) and Random Forests (RF) for the same high-risk(3) vs. low-risk(9) data and recorded the classification accuracy, specificity and sensitivity (shown in Table 3.3). The parameters for SVM and RF were tuned extensively using cross-validation experiments. Note that while the accuracies of LAD, RF and SVM models on cross-validation and the internal test set are not significantly different, the LAD model has significantly higher sensitivity values when compared with the SVM, and RF models. The LAD model has more balanced positive and negative predictive value compared to SVM and RF models. This is an important point to be noted, that in case of unbalanced datasets (datasets with significantly different class sizes) most classification methods focus on maximizing the number of correctly predicted observations, while LAD gives equal importance to classifying samples correctly in both classes.

Table 3.2. Classification accuracy for high-risk(3) vs. low-risk(9) data for LAD

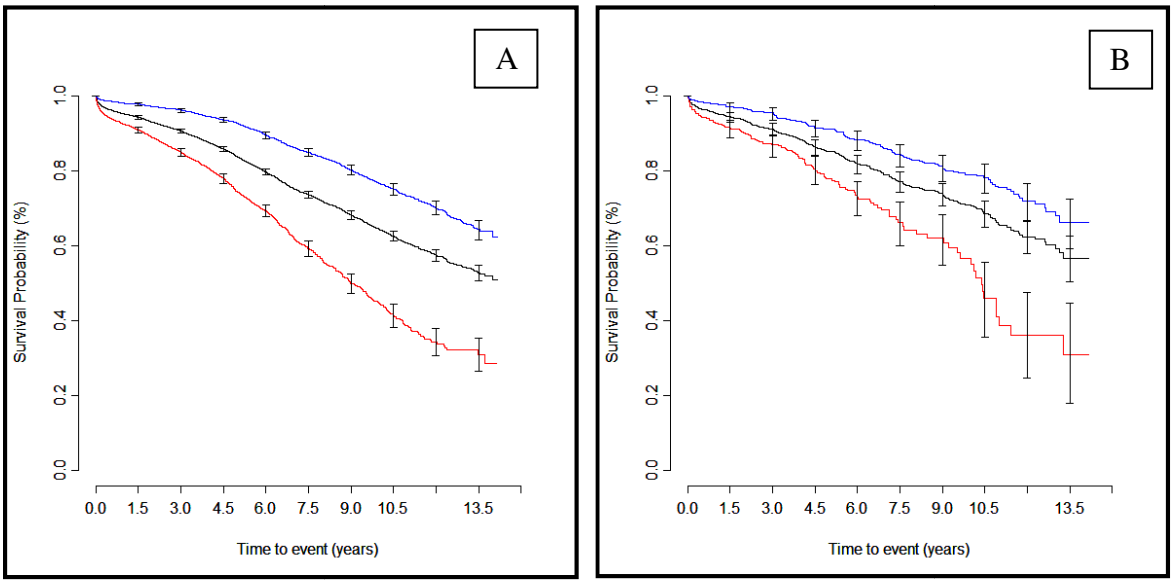
	Logical Analysis of Data		
	Accuracy	Specificity	Sensitivity
Internal Training	81%	82%	76%
Cross-validation	80%	84%	73%
Internal Test set	79%	82%	73%
Test set	76%	80%	70%
All Training data (including confusing)	79%	80%	75%

Table 3.3. Classification accuracies for high-risk(3) vs low-risk(9) for SVM and RF

	Support Vector Machines			Random Forests		
	Accuracy	Specificity	Sensitivity	Accuracy	Specificity	Sensitivity
Internal Training	83%	97%	70%	79%	97%	62%
Cross-validation	80%	95%	64%	80%	97%	63%
Internal Test set	78%	96%	59%	77%	98%	58%
Test set	75%	98%	52%	75%	97%	52%
All Training data (including confusing)	77%	98%	56%	76%	97%	56%

Kaplan-Meier survival curves of the groups predicted by the LAD model to be in predicted high, medium and low-risk groups are shown in Figures 3.4A and 3.4B for the training and test sets respectively. In Figure 3.4, the red (blue) survival plot corresponds to the predicted high-risk (low-risk) group, while the black plot corresponds to the survival curve for the entire population (baseline). We can observe that the high risk survival curves are below the baseline curve, while the low-risk curves are above that of the baseline. The high confidence bands for the test set could be due to the smaller size of the test set

Figure 3.4. Kaplan-Meier curves for the predicted high and low risk groups in (A) training data, and (B) test data



3.2.4 Identification of important combinatorial features

The LAD classification model was built by using patterns of degree 3, i.e.: they involve at most 3 variables in their defining conditions. Strong high and low-risk patterns built to distinguish between high-risk(3) vs. low-risk(9) data are shown in Table 3.4 and Table 3.5 respectively. We selected those patterns which had the highest homogeneity (coverage of patients of the same class when compared to the coverage of the other class) and prevalence (proportion of coverage of patients of the same class).

Table 3.4. Important high-risk(3) patterns

	Pattern description	Homogeneity	Prevalence
H1	ita $\leq$ 1 & miss_lvf = 1 & hx_htn = 1	0.70	0.27
H2	ithrl_l6 = 0 & age > 75 & hx_htn = 1	0.72	0.24
H3	Iabphemo = 0 & age > 75 & hx_htn = 1	0.72	0.23
H4	stablang = 1 & unstbang = 0 & iptca_l6 = 0	1.00	0.19
H5	Unstbang = 1 & iptca_l6 = 0	1.00	0.22
H6	Miss_lvf = 1 & (crcl < 55 or crcl >= 60) & hx_pvd = 1	0.86	0.11

Table 3.5. Important low-risk(9) patterns

	Pattern description	Homogeneity	Prevalence
L1	Stabling = 0 & lvf_cath = 1 & hx_pvd = 0	0.90	0.40
L2	Unstbang = 0 & ita > 0 & lvf_cath = 1	0.91	0.39
L3	Stabling = 0 & unstbang = 0 & age $\leq$ 55	0.93	0.27
L4	Age $\leq$ 70 & lvf_cath = 1	0.90	0.36
L5	lvf_cath = 1 & hx_htn = 0	0.90	0.19
L6	iptca_l6 = 0 & ita = 2 & hx_cva = 0	0.94	0.21

3.2.5 Identification of important individual features

We set up several different criteria for evaluating the importance of individual features. The variables were evaluated for significance based on the following tests:

1. Principal component analysis (PCA): We selected the top 25% of up and down-regulated features in principal components (eigenvectors) which explain at least 85% of the variation in the data. This test selects sets of variables with the highest variation in the data, while reducing the redundancy. Note that PCA is unsupervised (does not take class information into account). PCA is discussed in more detail in Chapter 5.
2. LAD patterns of degree 1: A variable was considered important if it was involved in strong degree 1 LAD patterns for distinguishing high-risk(3) from low-risk(9).samples. This is a test of the predictive power of a variable considered alone.
3. LAD patterns of degree 2: A variable was considered important if it was involved in strong degree 2 LAD patterns for distinguishing high-risk(3) from low-risk(9). This is a test of the predictive power of a variable considered in combination with another.

4. Cox regression: We selected the top 25% up-regulated, and down-regulated variables based on the hazard ratio predicted using Cox regression.
5. Linear regression: We used the log(time to event) as the independent variable and ran linear regression only for patients who had an event (death), and again selected the top 25% up- and down-regulated variables.

Each of the above tests measures complementary sets of information. Combining all the tests gives us a very robust meta-test for selecting important features. A feature is considered to be important if its role is significant in at least 3 of the above 5 tests. Given the large size, noisy nature of the data, we performed 3-fold permutation tests for the above procedure, and selected only those variables which are consistently important in all the permutation tests. Table 3.6 lists the important variables along with their associated cut-points and an indication of whether they are up/down regulated.

Table 3.6. Important features along with associated cut-points and indication of up/down regulation

Feature	Cut-point	Up/down regulated
hx_cva	1	Up
hx_htn	1	Up
hx_pvd	1	Up
hx_ptca	1	
lvf_cath	1	Down
lvf_cath	2	Down
Age	< 55	
Age	[55,60)	Down
Age	[60,65)	Down
Age	[65,70)	Down
Age	[70,75)	Down
Age	≥ 75	Up
Male	1	
Bmi	[20-23)	Down
Ita	0	Up
Ita	1	Down

Crcl	<40	Up
Unstbang	1	Up
Stabling	1	Up
Insulin	1	Up
Niddm	1	Down
miss_lvf	1	Up
Ccs	2	Up
Nyha	3	
Lmt	[0-20)	Down
Rca	[0-10)	Down

3.2.6 Discovery of new classes in the high-risk(3) & low-risk(9) dataset

We have identified two distinct groups of patients among the data of high-risk(3) and low-risk(9) patients. These groups were first identified when we projected the patients onto their first two principle components using PCA. We further analyzed the clusters by applying consensus clustering [79] techniques (discussed in detail in Chapter 5). Figure 3.5 shows a 2-D plot of the clusters projected onto their first 2 principal components. The 2 new classes correspond to patients in the green (cluster 1) and orange (cluster 2) colors. The high and low risk patients in the two classes are marked by "+" and "circle" symbol respectively. To get a better look at the classes, we have made a 3-D plot of the first 3 PCs (Figure 3.6).

Figure 3.5. Plot of the 2 new classes discovered in hr(3) & lr(9) data

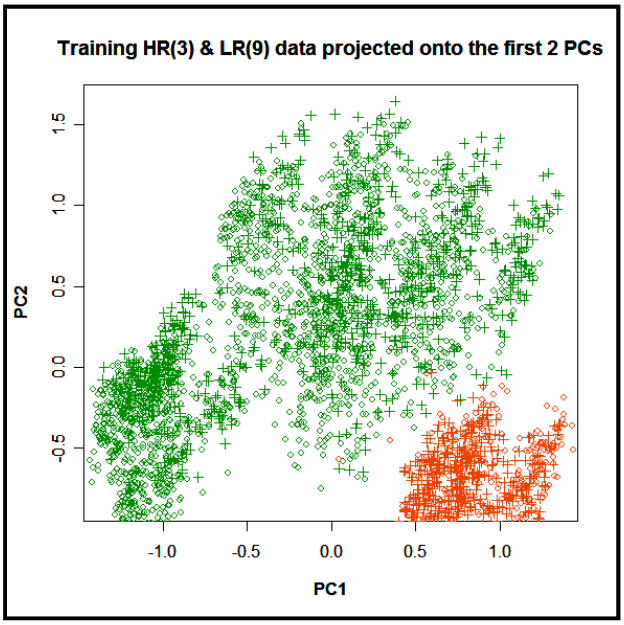
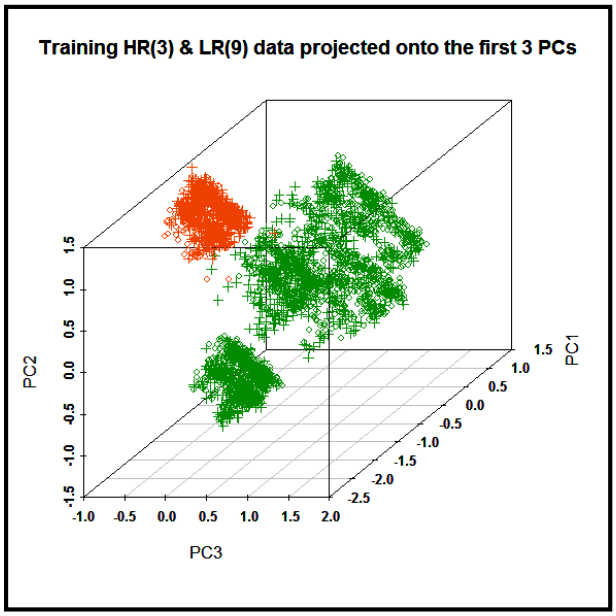


Figure 3.6. 3D plot of the 2 new classes discovered in hr(3) & lr(9) data



Cluster 1 consists of 792 high-risk patients and 1,848 low-risk patients, while Cluster 2 consists of 205 high-risk, 603 low-risk patients. The next step is to identify markers to distinguish between the two clusters. Variable NYHA (New York Heart

Association class, measure of dyspnea) distinguishes between the two clusters. All patients in Cluster 1 are characterized by NYHA=1, while patients in Cluster 2 by NYHA>1. Some important high and low risk patterns associated within Cluster 1 and 2 are listed in Tables 3.7, and 3.8. Stablang (stable angina prior to surgery), unstablang (unstable angina prior to surgery), ita (number of internal thoraxic arteries) and crcl (clearance creatinine) play important roles in distinguishing between the two clusters.

Table 3.7. Important high and low-risk patterns identified in Cluster 1

	Pattern description	Homogeneity	Prevalence
H1	stablang = 0 & unstablang = 1	1.00	0.29
H2	Nyha > 1 & unstablang = 1	1.00	0.29
H3	stablang = 1 & ithrl_l6 = 0	1.00	0.22
H4	ccs < 3 & age > 75 & lvf_cath > 1	0.76	0.20
L1	stablang = 0 & unstablang = 0 & lvf_cath = 1	0.90	0.47
L2	unstablang = 0 & ita > 0 & hx_htn = 0	0.91	0.35
L3	ita > 0 & age $\leq$ 75 & hx_htn = 0	0.90	0.33

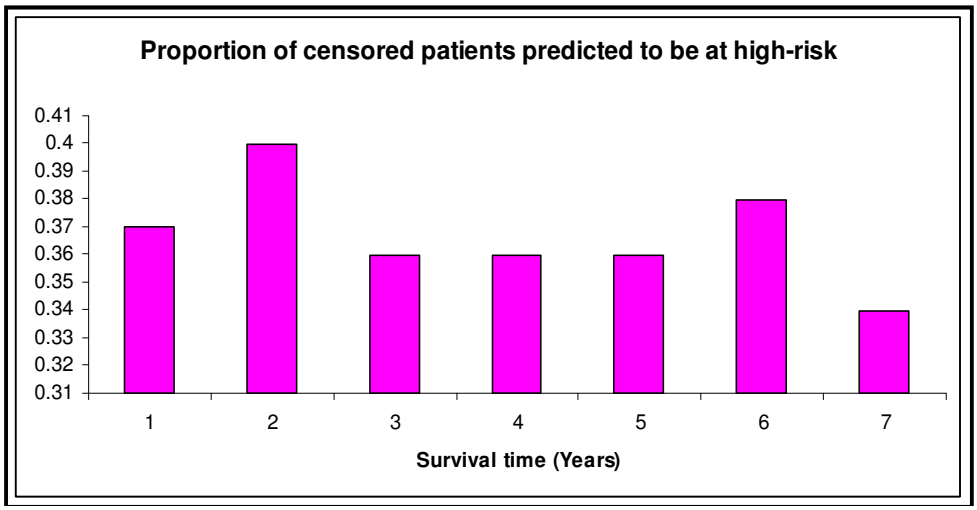
Table 3.8. Important high and low-risk patterns identified in Cluster 2

	Pattern description	Homogeneity	Prevalence
H1	crcl $\leq$ 40	0.65	0.27
H2	lvf_cath > 2 & ita = 0 & lmt > 20	0.69	0.21
H3	stablang = 1	1.00	0.12
H4	Miss_lvf = 1 & lmt $\geq$ 20 & lcx > 60 & lcx $\leq$ 80	0.67	0.13
L1	ita > 0 & stablang = 0 & miss_lvf =0	0.90	0.73
L2	hx_pvd = 0 & ita > 0 & stablang = 0	0.91	0.69
L3	hx_htn = 0 & hx_pvd = 0 & lmt $\leq$ 20	0.91	0.27
L4	hx_pvd = 0 & age $\leq$ 55	0.91	0.27
L5	hx_pvd = 0 & ita > 0 & crcl > 40	0.90	0.68

3.2.7 Prediction of risk for short-time censored patients

We use the LAD model built on the high-risk(3) vs. low-risk(9) data to predict the mortality risk for the short-time censored patients. We define short-time censored patients as those whose survival time is below 7 years, but who didn't have an event. Figure 3.7 shows a plot of the distribution of the proportion of those censored patients who are predicted to be high risk patients.

Figure 3.7. Plot of the distribution of censored patients predicted to be at high risk by the LAD model



3.3 Conclusions

One of the main challenges with analyzing this data was that the high-risk patients cannot be distinguished accurately from the low-risk ones based only on the clinical variables measured. The reason could be attributed to the fact that heart disease is very complex in its mechanism and it is known that there are many risks for heart disease: mainly genetic polymorphisms and many other measures (for example: the hematology measurements in blood, from Chapter 2, diet and exercise) which are not considered in this study. Using only the clinical measures we are unable to predict some of the patients (whom we called as confusing). In this paper, we present a simple score function to identify these confusing patients. Once we remove these confusing patients, we can distinguish more robustly the remaining high and low-risk patients.

Given the small number of events in the dataset, we chose to focus on the extreme risk patients. A threshold of 3 years for defining the high-risk group and 9 years for the low-risk group was selected based on optimizing for concordance accuracy. We build

LAD patterns of degree 3 to distinguish between the high and low-risk patients. The heat map of the patterns in the model covering the patients in the test set indicates that the LAD model is of high quality. The LAD model has an accuracy of 80% on cross-validation and 79% on the test set. The LAD model also has balanced sensitivity and specificity performance when compared to the RF and SVM models. Kaplan-Meier survival curves show that the predicted high risk patients have survival curves which are below the baseline curve, while that for low risk is above the baseline survival curve, as expected.

Most classification algorithms focus on minimizing the total error (proportion of incorrectly classified samples). In case of unbalanced and skewed datasets (significantly different sizes of the classes) such algorithms usually fail. In the case of our data, there are only 25% deaths in the entire dataset But these are very important events, we do not want to predict incorrectly samples with high-risk for mortality. One of the advantages of LAD is that it minimizes the error separately in the two classes.

We have identified some very strong high and low risk patterns involving the combinations of at most 3 variables which have high prevalence and high homogeneity. These patterns can be considered as medical hypotheses and should be further investigated. Further, in order to identify important variables individually, we have used a very robust scheme based on performing robust permutation analysis on five complementary tests. The important promoters of high-risk are: history of cardiovascular disease, history of hypertension, age > 75, no internal thoraxic arteries, clearance creatinine < 40, unstable angina prior to surgery, stable angina prior to surgery, insulin treated diabetes, indicator for missing data on LVF (this could be a secondary measure

for the severity of the disease), Canadian cardiac society score for heart disease. The important down-regulated variables are: age between 55-75, body mass index between 20-23, 1 internal thoraxic artery, non-insulin treated diabetes, maximum stenosis on left main coronary artery $\leq$ 20%, maximum stenosis on right coronary artery $\leq$ 10%

In the high-risk(3) group we have discovered two new subclasses which can be distinguished by the variable NYHA (New York heart association score). Important risk markers are stable angina prior to surgery, unstable angina prior to surgery, number of internal thoraxic arteries and clearance creatinine.

For short-time censored patients, we have used the high-risk(3) vs. low-risk(9) LAD model to predict whether a short-time censored patient is at high risk or not. It would be very interesting to check the current status of these patients in the database. Most of these patients were censored because the study ended. If these patients are followed up in the future, this data would be a true “external” validation for the LAD prognostic model.

Chapter 4

Molecular Stratification of Clear Cell Renal Cell Carcinoma Reveals Distinct Subtypes and Survival Patterns[3]

4.1 Introduction

Renal-cell carcinoma (RCC) is the most common cancer in the adult kidney, making up 3% of all malignancies in the United States. Each year, more than 50,000 men and women are diagnosed with RCC and about 12,000 of them die from this disease. Histopathologically, there are four major recognized subtypes of RCC: 75% are clear-cell, 15-20% papillary, 5% chromophobe, and ~1% unclassified. Each variant of RCC presents different genetic alterations, clinical behavior, and response to therapy.

Clear cell RCC (ccRCC) tumors have a distinctive histology, with an abundant cytoplasm rich in lipids and glycogen. The role of inactivation of the VHL gene, which occurs in upwards of 90% of ccRCC tumors [80], and subsequent effects on VEGF (vascular endothelial growth factor) signaling pathway in the pathogenesis of ccRCC is well established and has led to the development of new treatments targeting the VEGF-mediated signaling pathway [81-83]. Despite the prevalence of VHL mutation, ccRCC

[3] Based on collaborations with Gyan Bhanot and Michael Seiler (BioMaPs, Rutgers University), Kimryn Rathmell and Rose Brannon (University of North Carolina). This chapter is part of a submitted manuscript [6].

tumors have a wide range of natural histories and responses to VEGF-targeted therapy [84, 85].

Using the Fuhrman classification system, which primarily has prognostic significance in early stage tumors, ccRCC is classified as low, intermediate, or high grade by the analysis of tumor cell morphology. Early stage, low grade (Fuhrman grade 1) tumors tend to have less cytological atypia and better disease-free survival after resection when compared to high grade ccRCC [86]. High grade (Fuhrman grade 4) tumors are similarly prognostic, although they have an extremely high risk ratio of RCC-related death [86]. In contrast, intermediate (Fuhrman grade 2 or 3) grade tumors, which comprise the bulk of renal tumors greater than 2.5 cm (included in clinical stage T1 up to 7cm), have a fairly unpredictable risk for development of distant disease. As no adjuvant treatment has yet been demonstrated to have efficacy in this malignancy, an intensive screening program for recurrent disease is applied to virtually every patient in this group. This group of intermediate grade large masses is among the most difficult to manage, as up to a third of these patients will go on to develop metastatic disease and little in the tumor histology or staging information is helpful in predicting this devastating outcome.

Even in the untreated metastatic setting, clinical tumor behavior ranges from indolent, with barely perceptible disease progression over months or years, to highly aggressive and rapidly lethal. The molecular basis of this diversity in histologic grade, clinical behavior, and response to VEGF-targeted and other biologic therapy is unclear, and makes ccRCC a ripe target for studies investigating the molecular nature of these heterogeneities.

Several studies have used gene expression arrays to characterize ccRCC [87, 88]. An early study looked at 29 ccRCC and showed that unsupervised clustering separates them into two classes, which roughly correlate with long term outcome [89]. Another study identified a potential gene expression signature for aggressive clinical behavior in ccRCC by analysis of gene expression profiles of a set of 10 non-aggressive (low grade), 9 aggressive (mostly high grade), 9 metastatic, and 12 normal kidney samples [90]. The authors found that unsupervised clustering showed clear separation of normal kidney from all cancer specimens, and within the cancer specimens there were two clusters containing mostly non-aggressive and aggressive/metastatic specimens respectively. This data suggests that low grade and high grade ccRCC have distinct underlying biology. Most recently, a study on 27 ccRCC tumors, including 16 metastatic tumors, identified a three-gene prognostic signature based on unsupervised clustering and cross-validation with previous studies. The largest ccRCC study, performed on 177 samples [91], found gene expression signatures for five subclasses of ccRCC which were correlated with survival. However, none of these studies focused on the indeterminate intermediate tumors or were able to identify any clear distinguishing biological features separating the classes they identified.

Gene expression analyses have provided insight into the clinical heterogeneity of other solid tumors. In particular, for breast cancer, unsupervised clustering of gene expression data [92, 93] has identified breast cancer subtypes with distinct gene expression profiles correlated with recurrence risk and survival. Supervised learning methods applied to gene expression data have resulted in FDA approved gene panels predictive of risk for breast cancer recurrence [94, 95]. We have recently developed

methods based on unsupervised consensus ensemble clustering [79, 96-98] which successfully identified distinct genetic subtypes of ER+, HER2+ and Basal-like breast cancer that correlated with significantly differential risks of long term recurrence.

To understand ccRCC, we applied these unsupervised consensus clustering methods to mRNA expression data from a set of 52 ccRCC samples. Our methods identified two robust and distinct subtypes of ccRCC with distinct gene expression signatures and dysregulated pathways. These subtypes, which we call clear cell A (ccA) and clear cell B (ccB), can be determined from the expression of a highly specific gene set, making application to other tumors or data sets practical. Additionally, we used Logical Analysis of Data (LAD) to find signatures for groups of genes altered in each subtype. We applied our methods of identifying ccA and ccB subtypes to the published datasets of (i) 177 clear cell tumors [91], and (ii) 21 ccRCC [99] for validating the clusters. We found that the two subtypes predict for very different survival profiles, with ccA having a significantly better outcome than ccB. Additionally, we resolved the biological pathways differentiating the two classes of ccRCC: the better prognosis ccA cluster overexpresses angiogenesis and classic RCC genes, while the poor prognosis ccB cluster overexpesses genes in the Wnt signaling pathway. We also show that our results correlate well with stage and grade. This method can be especially useful for predicting the class for intermediate grade tumors, since they are the most unpredictable, and knowing the prognosis and underlying biological cause will provide great assistance for clinicians and drug discovery alike.

4.2 Materials and methods

4.2.1 Samples

Tumors specimens from 52 clear cell RCC patients and normal tissue from 18 samples were collected by the UNC tissue procurement core facility from consenting patients undergoing radical or partial nephrectomy (surgical removal of kidney) for RCC during the period of 1994 – 2008. Specimens were analyzed by a pathologist for clinical information (Table 4.1) and quality assurance, flash frozen in liquid nitrogen, and accessed with appropriate institutional IRB approvals. The validation sample sets consists of (i) 177 samples from Zhao *et al.* [91], (ii) 21 samples from Gordon *et al.*[99]. For Zhao *et al.* dataset of 177 samples we obtained updated survival information.

Table 4.1. Demographics and clinical characteristics for the UNC cRCC tumors

Characteristic	Subgroup	n/N (%)
Arrays	Clear Cell tumors	49/70 (70)
	Independent replicates	3/70 (4)
	Normals	18/52 (26)
Tumor size (cm)	Median (min-max)	4.5 (1.8-17)
Tumor stage	T1a (0-4 cm)	19/49 (39)
	T1b (4-7 cm)	17/49 (35)
	T2 (>7 cm)	8/49 (16)
	T3a (invasion of adrenal gland or perinephretic fat)	2/49 (4)
	T3b (invasion of renal vein)	3/49 (6)
Nuclear grade	Grade 1	2/49 (4)
	Grade 2	34/49 (69)
	Grade 3	13/49 (27)

4.2.2 Gene Expression Analysis

Gene expression microarray data was obtained using Agilent Whole Human Genome (4x44k) Oligo Microarrays (Santa Clara, CA) and processed in the UNC Genomics Core.

We retained only those variables which have < 30% missing values. Remaining missing data was imputed *k*-nearest neighbors method (k = 10) using Significance Analysis of Microarrays (SAM) [100].

The samples originated from 3 groups, so there could be a potential bias in gene expression measurements. We use "distance weighted discriminant" method (DWD) [101] to remove systemic biases in the data. Each sample was then standard normalized (subtract the mean of the array and divide by the standard deviation). For the Zhao *et al.* validation set, gene expression data from 177 ccRCC specimens was collected as published [91]. This data was tabulated and imputed using the methods described above. These samples originated from 10 different runs, and thus we merged the different groups using DWD, and normalized as above. Gene expression data from Gordon *et al.*, was collected as published [99]. Raw data from unflagged variables was retrieved using GeneSpring software and normalized using the methods described above.

4.2.3 Pathway Analysis

A genetic pathway is a set of interactions occurring between groups of genes which are functionally related. They work together as a network to achieve some aggregate function for the cell. A genetic mutation disrupting the function of one gene in a pathway breaks the connection between genes acting before and after the mutant gene, and may lead to a dysregulated pathway.

For pathway analysis, two-class unpaired SAM was performed with 100 permutations on the full variable set Heat maps were created using Cluster 3.0 [102] and Java Treeview [103]. Differentially regulated genes were analyzed in DAVID Bioinformatics Database [104] for functional annotation. The gene ontologies and pathways selected had p-value and False Discovery Rate (FDR) <0.05.

4.2.4 Principal Component Analysis (PCA)

PCA is a feature reduction technique that is commonly used to reduce large feature sets to those which are most informative [105, 106]. From the eigenvectors we identify the features whose coefficients are in the top 25% by absolute value in these vectors. These features are retained for further analysis because they represent most of the variation in the data. Using this method, in the UNC dataset, we identified 20 eigenvectors and 281 features which were retained for further analysis.

4.2.5 Unsupervised Consensus Ensemble Clustering

This is a technique [97] that identifies robust clusters in the data. This method was applied to normalized expression data projected on to the features identified by PCA. The samples were divided into k=2, 3, and 4 clusters using bootstrap averaging over both features and samples. We also averaged over two clustering techniques, k-means [107] and Self-Organizing Map (SOM) [108]. This makes the results insensitive to data and clustering method used. *k*-means is a centroid-based clustering method which starts by randomly assigning k centroid vectors in feature space and then iteratively proceeds over the following two steps until convergence: a) Use the Euclidean distance measure between the sample vector and the centroid vectors to assign each sample to the nearest centroid vector and, b) Move each centroid location to the center of the samples assigned to it. SOM is a type of neural network which works by training a rectangular grid of nodes represented by vectors in feature space. As each input vector is considered, the node which best matches the input vector and each node in its neighborhood is adjusted towards the input vector.

4.2.6 Logical Analysis of Data (LAD)

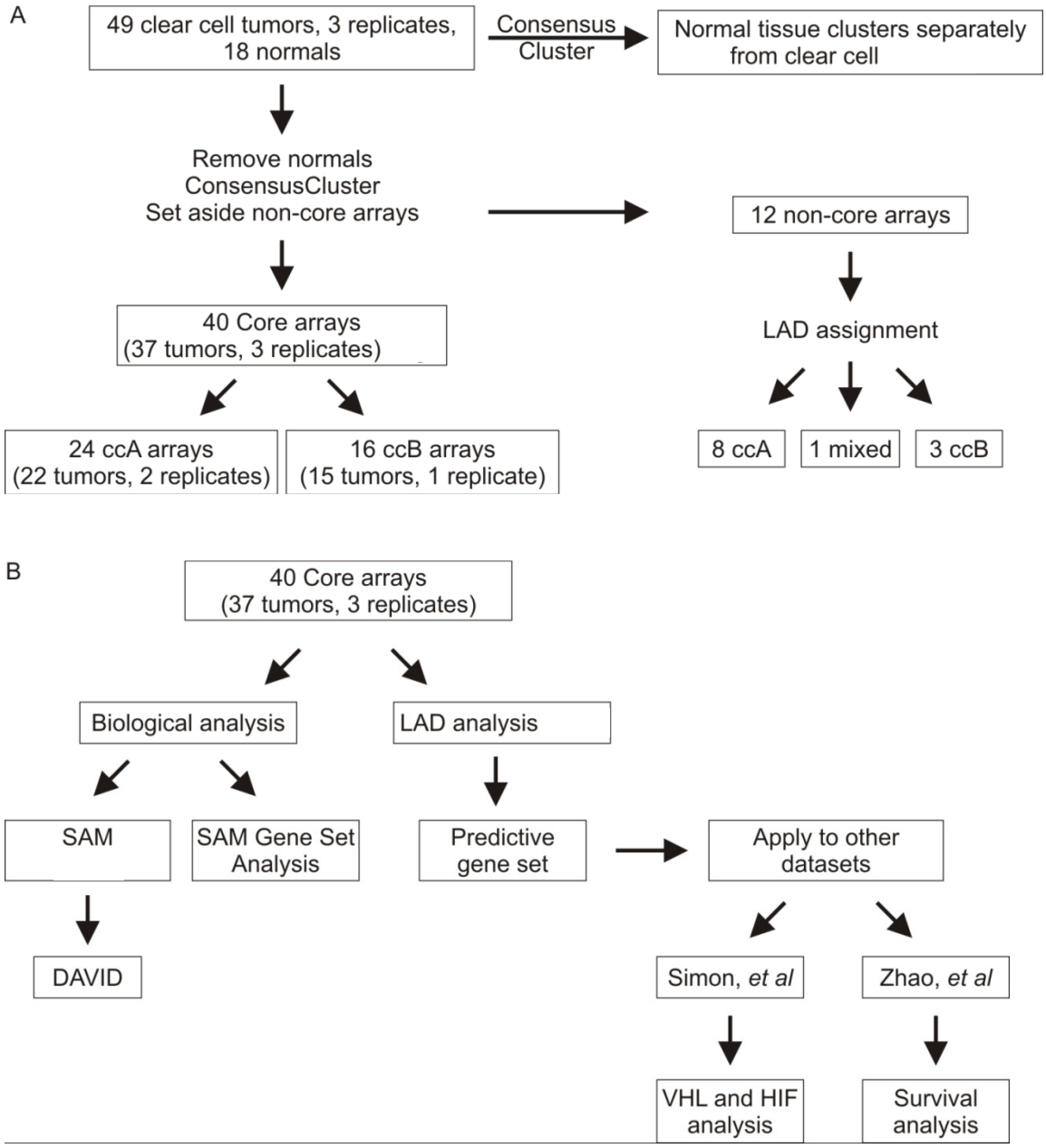
Once we identify robust clusters in the data, the next step is to identify patterns to distinguish them. We use LAD patterns to distinguish ccA from ccB. Using LAD, we identified all patterns that: a) classified the samples into high/low expression threshold using a single cut-point at median expression value, and b) built LAD models with equal proportion of positive and negative patterns.

There were two main goals for building LAD patterns: (i) identify those genes (or combinations of genes) which separate the clusters identified using clustering, (ii) robustly predict cluster membership for external datasets (including those generated in other labs, and from different chips for microarray analysis). The latter is not an easy problem, since the variable gene expression levels in external datasets maybe very different and incomparable. To resolve this issue, we used normalization based on samples. The key assumption here is that most of the genes are housekeeping (do not vary in tumors and normals), and using this method of normalization, genes which are affected by the cancer will have large deviations from the mean. To make the predictions robust, we computed confidence levels by running 100 bootstraps (consisting of 80% of the patterns from the entire set), and the LAD discriminant score was computed for each bootstrapped sample. The final LAD score was computed as the average of 100 runs, and the confidence level was computed as the % of times the sample was predicted to be in ccA or ccB. For the final prediction: samples with confidence levels < 0.75 are left as unclassified. This approach was used for predicting the cluster membership for non-core samples as well as for the validation datasets.

4.3 Results

Figure 4.1 is a flowchart of the analyses performed in this study. We used the following steps: (1) Check if normal samples separate out from the tumors by clustering, (2) Run consensus clustering on ccRCC and identify stable "core" clusters, (3) Use LAD to distinguish between the core clusters and identify predictive patterns, (4) Identify pathways enriched by the clusters (5) Validate the LAD patterns on external datasets, (6) Compare differences in survival and clinical characteristics in the predicted clusters. Details of results in each step are explained below.

Figure 4.1. Flow chart diagram depicts the order of analyses.
(A) Delineation of steps taken to identify ccRCC subtypes. (B) Diagram of analyses to characterize and validate identified subtypes.

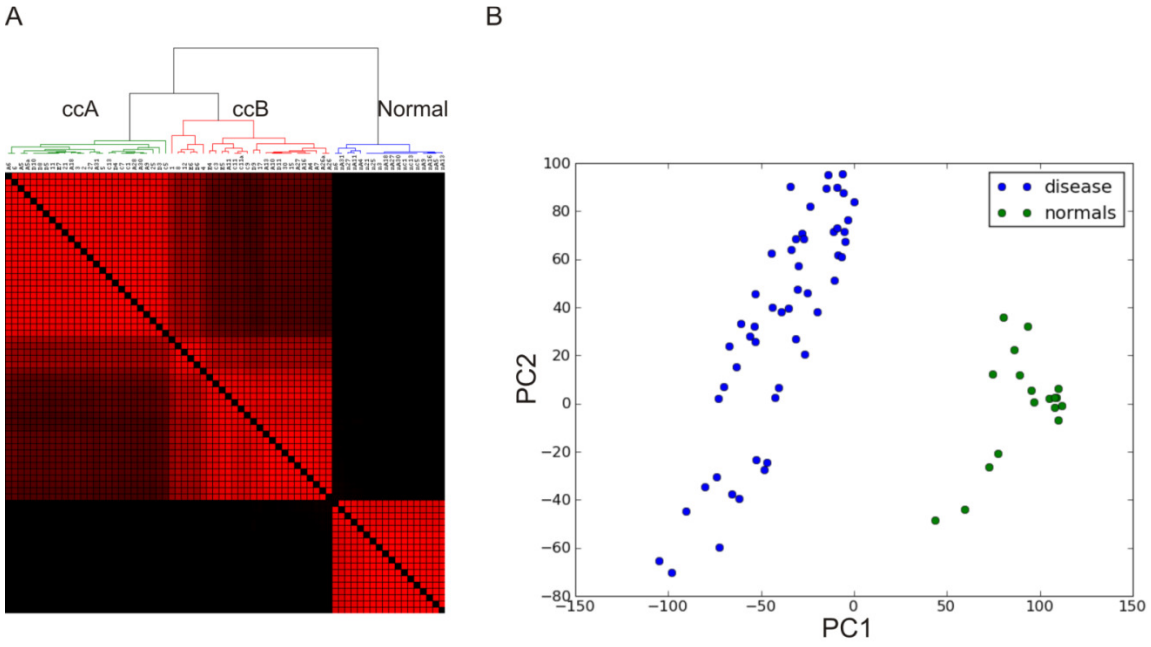


4.3.1 Identification of subtypes of ccRCC

High quality gene expression data was obtained for 52 samples of ccRCC (with three samples processed twice for internal quality control) and 18 normal samples. To determine the optimal number of clusters present in the data, unsupervised consensus ensemble clustering was performed. ConsensusCluster software was developed and used to separately run k-Means and Self-Organizing Maps 300 times using random sub-samplings of the arrays and gene variables. The results for each k (k = number of clusters) were combined into a consensus matrix, which visualizes the fraction of times two samples are clustered together. Bootstrapping along the sample and feature space rendered these results robust (insensitive) to data perturbation.

Consensus clustering was perfomed on all the samples (including normals). The data split very clearly into 3 clusters. Normals separated out from the tumors, and the tumors split into two clusters (Figure 4.2A). Figure 4.2B shows a PCA plot of data projected onto the first two principal components.

Figure 4.2. The two ccRCC subtypes are distinct from normal kidney tissue.
(A) Both consensus matrix and (B) PCA plot (scatter plot of the top 2 eigenvectors – PC1, PC2) show the complete delineation between the clear cell tumors and corresponding normal kidney tissue removed from ccRCC patients. These results verify that the subtypes do not arise from errors in the expression levels due to contamination from normal tissue.

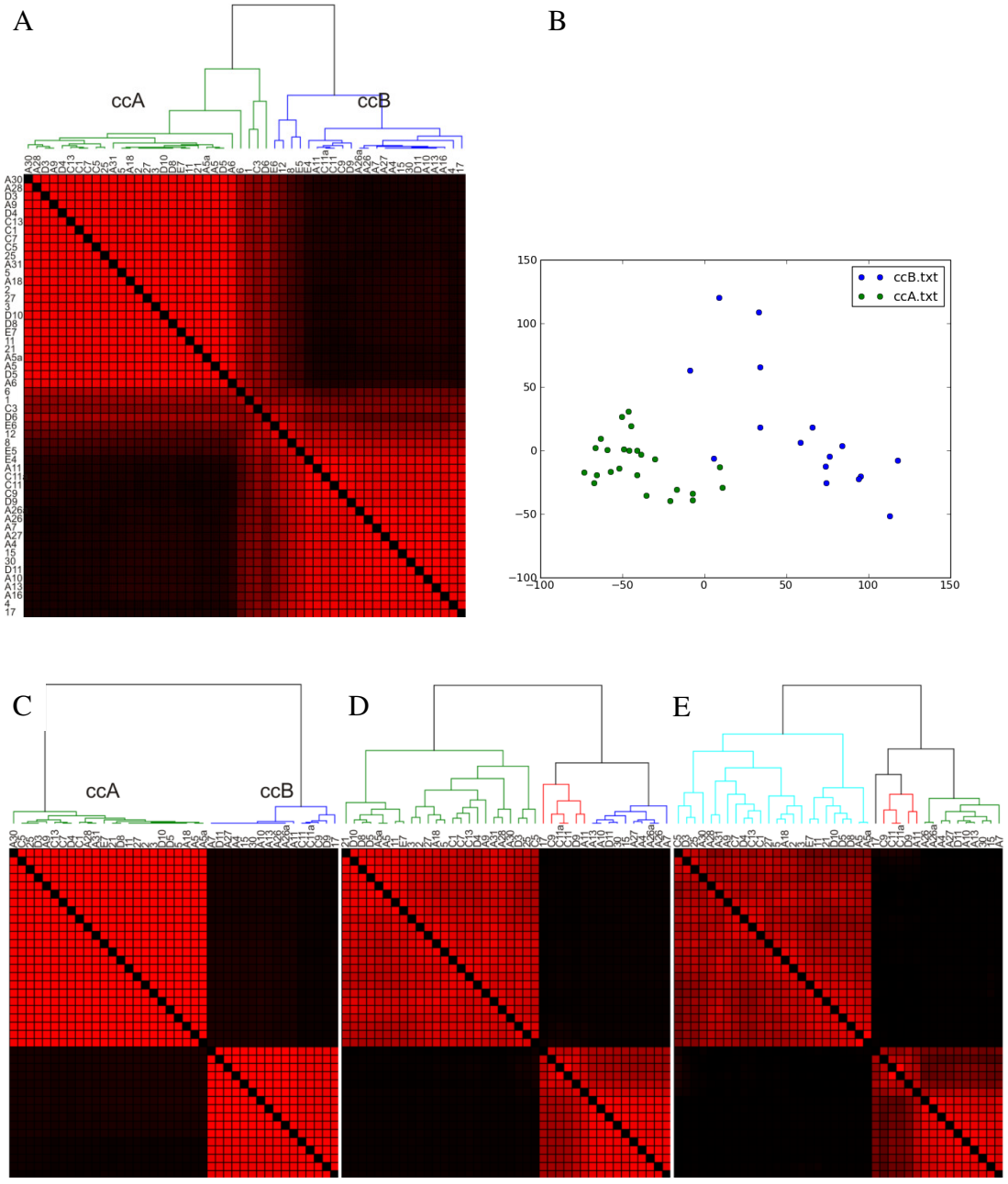


To identify clean subsets of samples within each cluster, groups of samples that either remained together in the same cluster or split cleanly into two clusters across bootstrapped trees were assigned to "core clusters". Those arrays whose membership shifted under bootstrap analysis or when k was increased to k+1 were set aside for later classification. The final core clusters in our dataset included 40 samples of the original 52. This membership permits the tumors with best patterned features to define the cluster boundaries. Figure 4.3 shows a color coded map of the consensus matrices for these core cluster samples for k = 2, 3, 4. The red areas represent regions where the tumor samples on the rows and columns belong to the same cluster. As demonstrated in Figure 4.3A, the 52 samples split into two robust and stable subtypes of ccRCC. The core clusters for k=2

show remarkable stability even when the number of permitted clusters is increased to k=3 and k=4 (Figure 4.3C-E respectively), suggesting that the optimal number of robust clusters in this dataset is two.

Neither cluster is caused by inclusion of normal tissue in the RNA extraction (Figure 4.2). Normal kidney clearly assorts independently of either cluster, and maintains independent features regardless of the k assignment. These analyses demonstrate that ccRCC can be optimally clustered into two distinct subtypes (ccA and ccB), which are defined based on purely molecular characteristics of the tumors and are statistically robust and reproducible.

Figure 4.3. Consensus matrices demonstrate the presence of only two core clusters of intermediate grade ccRCC.
Consensus matrix heat maps demonstrate the presence of two clusters within all clear cell tumors (A) and invariance of the two ccRCC core clusters using (B) PCA plot shows the data for the two clusters projected onto top two PCs (C) k=2, (D) k=3, and (E) k=4 cluster assignments for each cluster method. Red areas identify samples clustered together across the bootstrap analysis. The two subtypes are stable (retain their sample membership) even when we instruct the algorithm to identify more than 2 clusters.

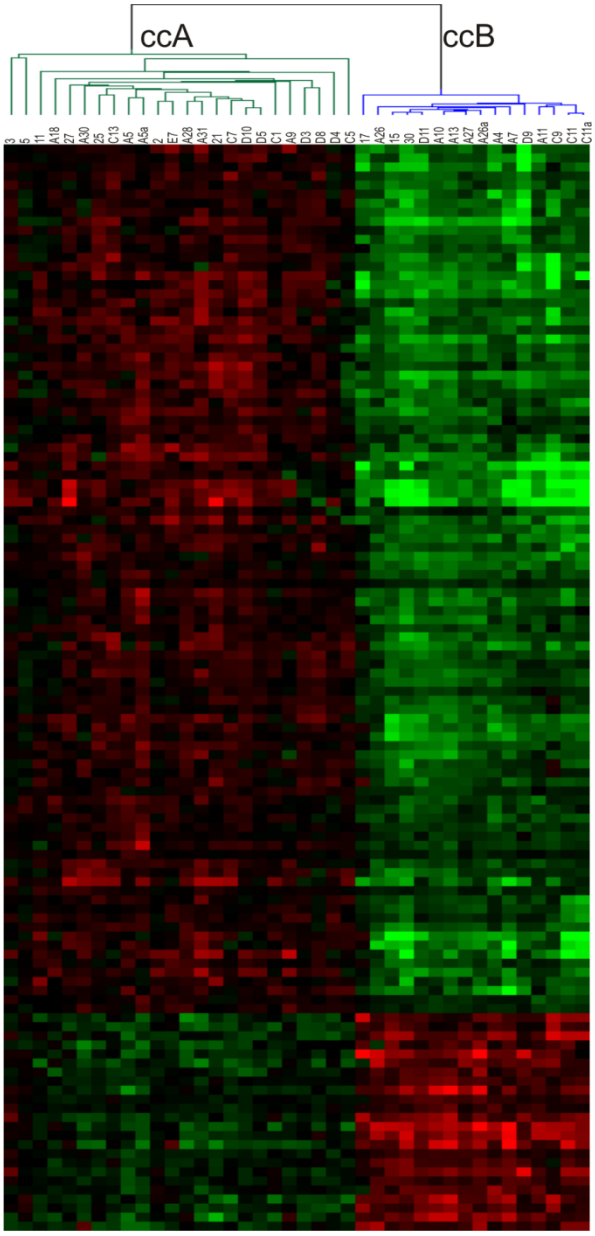


4.3.2 Use of LAD to delineate gene set to stratify ccRCC into ccA and ccB

Given the large size of the data, we reduce the entire set of variables using t-test (p-value $< 10^{-5}$). This reduced the data to 1,075 variables. We used only one cut-point for each gene -- median value. We identified a robust support-set by solving set-covering problems (1.2) in leave-one-out experiments (experiments by leaving each sample out once and rebuilding the patterns). The union of the support-sets in the leave-one-out experiments resulted in 120 features (Table 4.S1). A heat map of these variables on the core ccA and ccB clusters is plotted in Figure 4.4. Strong LAD patterns of degree 1 and 2 were identified, and equal subsets of positive and negative patterns were chosen in the model. Degree 1 model has 160 positive and negative patterns (80 each), and degree 2 model has 236 positive and negative patterns (118 each). The LAD patterns were validated using leave-one-out experiments, which were then used to predict the subtype of the original samples (accuracy was 100%, p-value = 0.0).

These patterns were used to predict the 11 non-core samples excluded from the ConsensusCluster, patterns predicted cluster membership for 10 of the 11 samples, 8 ccA and 3 ccB.

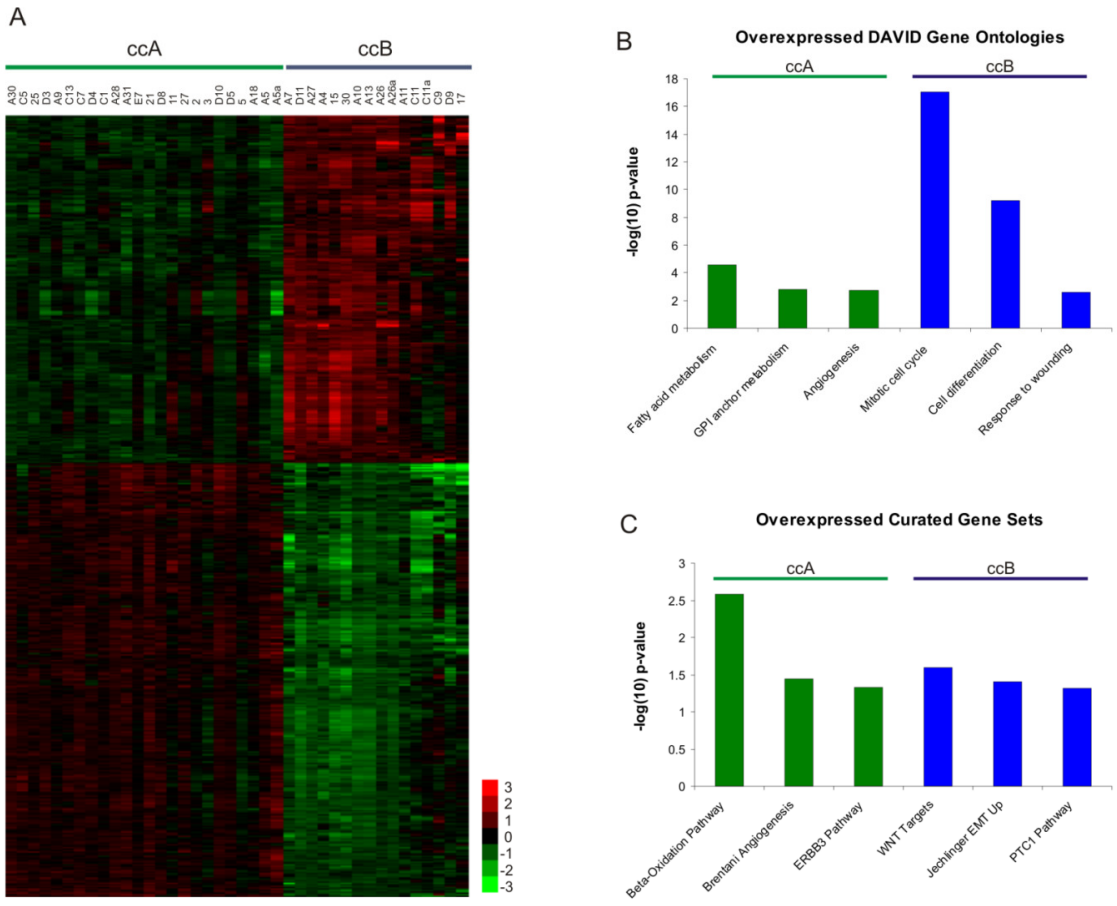
Figure 4.4. Heat maps show the clustering of ccA and ccB core by LAD variables. Gene expression data for core arrays and 120 variables selected using Logical Analysis of Data (LAD).



4.3.3 Analyzing pathway differences between two core clusters

An advantage of the consensus cluster is the opportunity to identify biological differences between the genetically defined ccRCC subtypes ccA and ccB. SAM was applied to the full variable set at a false discovery rate of zero (FDR=0); 6361 variables differentiated ccA from ccB. DAVID was used to analyze the variable sets for over-expressed categories with a p-value and FDR < 0.05. The most important Gene Ontologies associated with cluster ccA were fatty acid metabolism, GPI anchor metabolism, angiogenesis, beta oxidative, and ERBB3 pathways. Interestingly, pathways normally highlighted as typical of ccRCC (angiogenesis) were more highly expressed in the ccA subclass. Further substantiating this notion, a number of genes such as EPAS1, PDGFD, EGLN3, HIG2, and CA9 that are tightly correlated with certain aspects of VHL inactivation and HIF (hypoxia inducible factor) signaling are overexpressed in ccA relative to ccB. In contrast, core cluster ccB overexpressed genes associated with activation of the mitotic cell cycle, cell differentiation, response to wounding, and the Wnt receptor signaling pathway. These data suggest that ccA and ccB portray different dominant biologic pathways, resulting in distinct patterns of gene expression, and potentially distinct modes of clinical behavior.

Figure 4.5. Pathway analysis of subtypes shows that ccA and ccB are highly dissimilar.
(A) Heat map of the 6361 variables differentially expressed between ccA and ccB as determined by SAM analysis (FDR<0.004). (B) Select gene ontologies from the further analysis of these variables elucidate the differences between the clear cell subtypes. (C) SAM Gene Set Analysis with MSigDB curated gene sets confirm DAVID results and provide further evidence that ccB is more aggressive.



4.3.4 Validation of ccRCC subtypes and variables in the Zhao *et al.* dataset

To validate the cluster patterns in independent tumor data sets and identify clinically meaningful differences between the clusters, we analyzed first the previously described Zhao *et al.* dataset of 177 ccRCC microarrays [91]. Table 4.2 lists the clinical characteristics for the 177 tumors. For prediction, we combined the two patterns sets (degree 1 & 2) and removed equivalent and duplicated patterns. Since these two datasets

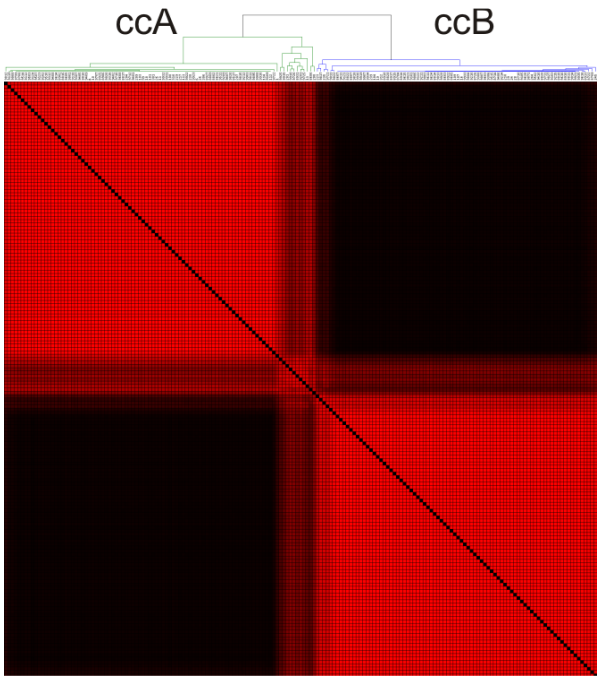
are generated using different chips, they don't have the same set of variables. In the validation dataset we identified matching variables (111 concordant variables, representing 67 genes), and reduced the pattern sets to include only the matching variables (33 ccA and 30 ccB patterns). The prediction score, cluster assignment and confidence levels for the individual samples are in Table 4.S3. Out of the 177 ccRCC tumors: 83 tumors were predicted to be ccA, 60 in ccB and 34 were left unclassified.

Table 4.2. Demographics and clinical characteristics for the 177 ccRCC tumors in Zhao *et al.* validation set

Clinical variables	
Age (mean $\pm$ std. dev.)	65.2 $\pm$ 10.9 years
Gender (%)	Male (58%)
Grade (%)	1 (5%), 2(19%), 3(53%), 4(23%)
Stage (%)	1 (28%), 2(16%), 3(23%), 4(33%)
Survival (median)	44 months

Another independent analysis for validation was running consensus clustering on the matching 111 variables for the 177 ccRCC in the Zhao *et al.* validation set This showed that there were clearly 2 clusters in the data (Figure 4.6). When increased degrees of freedom were permitted (k = 3, 4, 5), two dominant clusters remained. The cores were labeled as ccA or ccB by comparing their gene expression levels to those in the corresponding clusters in the training data set

Figure 4.6. Validation of LAD variables in Zhao *et al* data[91] show the existence of two ccRCC clusters.
Consensus matrix of 177 ccRCC tumors determined by 111 variables corresponding to the 120 LAD variables. Two distinct clusters are visible, validating the ability of the LAD variable set to classify ccRCC tumors into ccA or ccB subtypes.



When compared with the cluster assignment predicted by LAD patterns, we found that there was a large concordance (86% of ccA, 90% of ccB). These analyses validate both our assertion that there are two genetic subtypes of ccRCC and that the genes and patterns we have identified can robustly distinguish the subtypes across datasets and gene expression platforms.

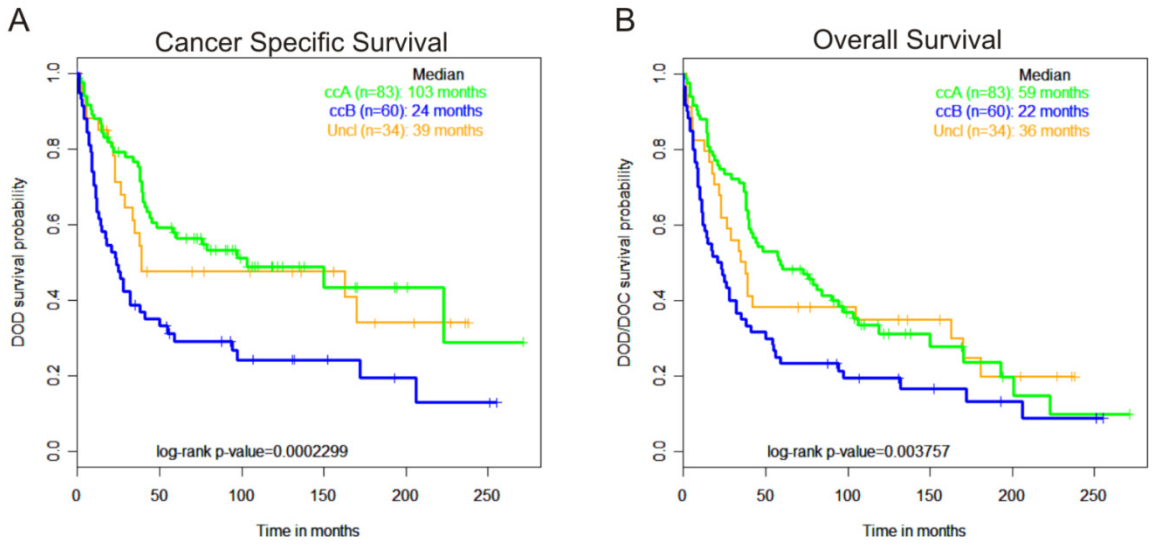
4.3.5 Cores ccA and ccB have different survival outcomes

The previous data confirmed the presence of two genetically defined subtypes of ccRCC; therefore, we turned our attention to identifying differences in clinical behavior associated with these genetic signatures. Analyzing the core samples from the original set of tumors, there was no significant correlation of the two subtypes with tumor size, renal

vein invasion, extracapsular extension, or Fuhrman grade 2 vs. 3 (p-value > 0.27). In the larger validation dataset of intermediate grade tumors, no significant correlation was observed with age, gender, tumor stage, or performance status (p-value > 0.32).

Cancer specific survival, DOD (death due to disease) and overall survival, DOD + DOC (death due to disease + death due to other causes) for the LAD predicted ccA and ccB classes are plotted using Kaplan-Meier curves (Figure 4.7). For cancer specific survival, the ccA subtype was found to have a significant survival advantage over ccB patients (median survival of 8.58 years vs. 2 years, log-rank p-value=0.0002). At the five year time point, 56% of ccA patients and only 29% of ccB patients had survived (Figure 4.7A). This difference in survival outcome is equally pronounced when analyzing overall survival. Figure 4.7B shows a significantly greater survival for ccA patients over ccB patients (median survival of 4.91 years vs. 1.83 years, log-rank p-value=0.004). At 5 years, survival for ccA patients is 48%, while it is only 23% for ccB patients (Figure 4.7B).

Figure 4.7. Classification of tumors from Zhao *et al.* data [91] using LAD.
177 ccRCC tumors were individually assigned to ccA, ccB or unclassified by LAD prediction analysis, and cancer specific (A) or overall survival (B) were calculated via Kaplan-Meier curves. The ccB subtype had a significantly decreased survival outcome compared to ccA, while unclassified tumors had an intermediate survival time (log rank p-value<0.004).



4.3.6 ccA/ccB subtype contributes to patient risk analysis

When all 177 tumors from the larger data set were assigned to subtypes, 71% of low grade (grade 1) tumors clustered as ccA and 68% of high grade tumors (grade 4) clustered as ccB tumors from the tumors which were assigned to classes. As low grade ccRCC tumors tend toward good prognosis, and high grade tumors toward poor prognosis [86], this result was expected and logical.

To determine whether the LAD score (discriminant score) added prognostic value to current clinical measures, a Cox multivariate analysis was performed (Table 4.3). The hazard ratio (HR) for LAD score, stage, grade, and performance are presented in Table 4.3 along with the 95% confidence interval. To see the added prognostic value, we present the results as unadjusted (univariate) and adjusted (multivariate Cox regression) for comparison. Since stage is an important predictor of survival, we also repeat the

adjusted and unadjusted results after stratifying for stage. We see that the HR for the LAD score (4.87) is very significant and is higher than the HR for other variables. In the adjusted model, the HR for LAD score is comparable to stage and is higher than grade and performance. When we stratify by stage, the LAD score still has very high prognostic power in both the unadjusted and adjusted form. The hazard ratio for the combined model (2.72) is highly significant.

Table 4.3. Hazard Ratio (HR) along with the 95% confidence interval (CI) for the predicted LAD score, Stage, Grade and Performance status.
"Unadjusted" refers to the HR for each of the variables individually, as opposed to "adjusted", which refers to HRs computed using a multivariate Cox regression model for all the variables together.

	Unadjusted HR (95% CI)	Adjusted HR (95% CI)	Unadjusted HR (95% CI) Stratified by Stage	Adjusted HR (95% CI) Stratified by Stage
LAD score	4.87 (2.13 – 11.1)	2.67 (1.14 – 6.25)	3.64 (1.54 – 8.62)	2.75 (1.17 – 6.47)
Grade	2.20 (1.60 – 3.02)	1.61 (1.11 – 2.13)	1.65 (1.17 – 2.34)	1.61 (1.12 – 2.31)
Performance	1.82 (1.43 – 2.31)	1.45 (1.13 – 1.87)	1.46 (1.14 – 1.86)	1.44 (1.12 – 1.85)
Stage	3.32 (2.54 – 4.34)	2.93 (3.85 – 2.93)		

4.3.7 ccRCC subtypes and HIF mutation status

Using the LAD patterns, we also validated our patterns on dataset published by Gordan *et al.* [99]. Out of the 21 samples, we predicted 10 to be ccA, 6 to be ccB and left 5 samples unclassified. This dataset has information for the status of HIF gene for each patient (whether it was homozygous wild type, heterozygous or mutant form). We checked the status of HIF mutations in the predicted subtypes, however we didn't find significant differences between the clusters (Table 5.S4).

4.4 Discussion

This analysis demonstrates that unsupervised consensus clustering algorithms can identify distinct classifications of tumors based purely on genetic pattern finding algorithms. This unique analysis provides a powerful biostatistical method to discriminate genetically distinct types of tumors that may be informative of tumor biology and/or influence tumor behaviour. In the specific model of clear cell renal cell carcinoma, a tumor type with one known defining genetic lesion present in virtually every tumor, two distinct subclasses of ccRCC (ccA and ccB) were identified and characterized by divergent patterns of gene expression and highly significant differences in long term survival.

The clinical heterogeneity and previous analyses of gene expression data [90, 91, 99, 109, 110] have suggested that there are at least two genetic subtypes of ccRCC. The analysis presented in this study has demonstrated that there are likely *only* two primary subtypes of ccRCC by showing that more subtypes are not stable under bootstrap analysis (Figure 4.3**).** Although the 177 tumor array analysis was initially described containing five discrete classifications, using our discriminating variable set the consensus clustering analysis failed to yield this number of clusters. It is possible that the 52 initial samples may have been insufficient to support the classification of other minor subtypes which may exist within ccA and ccB or that the restricted number of overlapping genes between the two platforms limited the flexibility of analysis. Our belief, however, is that this apparent discrepancy may reflect the high statistical discriminating power of the classification variable set to identify ccA and ccB.

A fundamental problem in the genetic analysis of human tumors is that the measurement of genetic noise in making pairwise comparisons across thousands of

independent and dependent variables. Our novel use of Principal Component Analysis permits a meaningful reduction of the number of variables and, when used cooperatively with consensus clustering and LAD analysis, is able to identify stable clusters and patterns of gene expression. This analysis of ccRCC demonstrates that this method is a highly robust and functional system for classification of biological samples into categories with meaningful clinical features.

In ConsensusCluster analysis, a "Core Class" or "Core Cluster" is defined by finding a non-overlapping set of samples that are distinguishable from each other with high accuracy, independent of sample or feature perturbation. This accuracy can be user defined (for example, a p-value for leave-one-out validation), and excludes some samples to reduce biological noise and permit highly congruent pattern analysis. Once these classes are defined, the left out samples can be predicted using LAD patterns (defined as simple cut-point rules on collections of gene expression levels). LAD analysis assigns these patterns to the classes. A robust approach was introduced to predict cluster membership using confidence levels generated by bootstrapping.

This method of tumor analysis permits a refined analysis of samples into genetically defined classes, the end result being predictive gene signatures useful for classification of samples outside of the primary analysis. We have demonstrated this procedure by applying the results derived on our initial ccRCC set of samples to extend our subtype assignments to independent datasets generated on dissimilar platforms.

Consensus ensemble clustering makes our results robust to both sample and feature bootstrap analysis. This rigorous technique addresses a common criticism of microarray analysis that it is often impossible to reproduce and generalize results found

on one dataset to other datasets. Indeed, the robustness of our predictions allowed us to identify a stable set of features with predictive clinical value over and above conventional clinical features.

Previously defined sets of predictive variables developed based on training set modules for ccRCC share some features with this predictive variable set, although most are non-overlapping. In particular, platelet derived growth factor (PDGF) isoforms have been repeatedly identified in both our set and that of Takahashi, *et al* [111]. This observation is notable given the activity and widespread use of drugs with PDGF-receptor inhibitory activity in ccRCC. Additionally, the diversity of variable sets potentially provides a unique opportunity for multivariate analysis which might provide the most clinically valuable prognostic tools, and additionally inform us of new features of ccRCC biology. The specific features which are derived as variables with predictive value in this analysis, themselves may provide clues as to the underlying biology of the clear cell clusters. One notable candidate identified in this analysis is NPR3, a gene whose expression in tumors may be informative of the kidney cell of type of origin of subtypes of ccRCC or may represent a component of HIF signaling [112].

The subtypes ccA and ccB showed a significant difference in survival outcome of patients after nephrectomy, with ccA patients having a better prognosis. Analysis using SAM and DAVID revealed that the better prognosis ccA group relatively overexpressed genes associated with hypoxia, angiogenesis, fatty acid metabolism, and the ERBB3 pathways, whereas ccB tumors overexpressed a potentially more aggressive panel of genes that regulate EMT, the cell cycle, and wound healing pathways. Interestingly, the ccA tumors overexpress genes “classically” associated with ccRCC. These gene sets

associated with some components of the hypoxia response pathway and angiogenic processes are well known to be broadly dysregulated in clear cell RCC; therefore, it is intriguing that this subset is relatively more highly expressed in a single subtype of ccRCC, ccA. However, *VHL* mutation and subsequent activation of the hypoxia response pathway is so highly correlated with ccRCC that many of these pathways are expected to be upregulated in ccRCC across the board. We have identified *VHL* mutations in both clusters, as expected. This conundrum presents numerous possibilities: ccB subtypes may have acquired additional genetic events in addition to *VHL* pathway events, which contribute to a more immature and aggressive phenotype. Alternatively, understanding the types of VHL mutations present in ccA vs ccB may be instructive, and this and other biological queries are rich areas for future investigation.

Finally, our robust panel of 120 variables, representing 110 genes, whose expression levels can classify tumor samples based on an LAD score into ccA and ccB subtypes with high accuracy, may provide a valuable resource for clinical decision making regarding frequency of surveillance or choices for adjuvant therapy in the future. This panel provides the basis for the development of an RT-PCR (reverse transcription polymerase chain reaction; method used to accurately measure RNA levels) based assay which could be applied to formalin fixed tissues to assign subtypes of ccRCC to individual tumor specimens, which would require validation by a prospective clinical trial.

Chapter 5

Logical Analysis of Survival Data[4]

5.1 Introduction

In the current era of evolving targeted therapies for cancer, and their increased use in the adjuvant settings, it becomes more important than ever to be able to precisely assign prognostic risk for death, metastasis or recurrence of cancer. Cancer is the cause of one in eight deaths worldwide, it causes more deaths than AIDS, tuberculosis, and malaria combined. Cancer is the second leading cause of death in economically developed countries (following heart diseases) and the third leading cause of death in developing countries (following heart diseases and diarrhoeal diseases) [49]. The known causes for cancer are either environmental factors (tobacco, radiation from toxins, etc.), or genetic factors (inherited mutations, immune system, etc.). One of the main goals of cancer research is to identify mutated genes (dysregulated genes) that are causally implicated in carcinogenesis [113]. The hope is that once we identify the genetic cause for cancer, we can design drugs to target them. Research efforts on cancer have increased exponentially in the recent years. Most studies involve collecting tumor samples from patients at the time of surgery or biopsy and then generate and analyze gene-expression data, single nucleotide polymorphisms, copy number variation, etc.

[4] Based on collaborations with Endre Boros (RUTCOR, Rutgers University), Gyan Bhanot (BioMaPs, Rutgers University) and Louis-Philippe Kronek (GSCOP, Grenoble). This chapter is part of two published paper [7, 9], and a submitted manuscript [8].

Survival analysis involves predicting survival time (or time to event) for samples based on recorded variables. The outcome in this case is continuous, thus it looks very similar to regression analysis but with the main difference that some of the samples are censored. A sample is considered to be censored if it has incomplete time to event information. There are different types of censoring, however we will concentrate on right-censored survival analysis problems since they constitute majority of the situations. Right-censoring occurs when the patient did not have an event until the end of the study. These studies are usually conducted for a fixed period of time (e.g. 10 or 15 years), during which patients are observed for the event of interest. Some patients do not experience the event during this period. Such patients are said to be right-censored and their censoring time serves as a lower bound for their survival time. If the study was conducted for long enough periods then the event would be observed in all patients. A naïve approach for handling censored data is to disregard them from the study (reducing the problem to classical regression analysis). In most studies, a very large proportion of samples are censored, so we cannot disregard them. We would like to use their information until the time when it was known that they survived.

In the previous chapters we saw a few approaches to build models for survival analysis. In Chapters 2 and 3 we selected thresholds for defining high- and low-risk patients, and then the problem reduced to building a classification model to distinguish between the high and low-risk patients. In Chapter 4 we identified subtypes by robust clustering techniques, and showed that the subtypes have very different survival profiles. In this chapter we develop a new supervised prognostic method, Logical Analysis of Survival Data (LASD), based on the principles of LAD, to predict survival time and also

to handle censored patients. This algorithm is based on defining high- and low-risk classes for every time-point t when an event occurs in the dataset, and then building patterns to distinguish them. Each pattern is associated with a risk score, computed as the area under the survival curve for patients covered by that pattern. Finally a patient-specific score is computed as the average of the pattern scores for patterns covering that observation.

We discuss first a simple linear programming model to predict survival. Then we discuss the algorithm for LASD in detail. Finally, we discuss ensemble methods to improve the performance and robustness of LASD. We illustrate our proposed approaches on the clear cell renal cell carcinoma dataset [91] discussed in Chapter 4. This dataset has microarray gene expression measurements for 177 tumor samples. The outcome that we are interested in predicting is death due to the cancer. We used the subtypes predicted in the previous chapter, and built prognostic models on the subtypes separately. Cancers are classified based on their anatomical location, instead of causality and function. Stratifying the samples into subclasses based on similar gene-expression levels enables us to make more accurate predictions, by optimizing the models separately in each of the subtypes. While predicting survival for new or unseen samples, we use a two-step approach: (i) predict the subtype, (ii) predict survival time for the patient in the corresponding subtype.

We present three main approaches for predicting survival time for samples: (1) Linear programming survival (LPS) model (2) Logical Analysis of Survival Data (LASD), an extension of LAD to predict survival risk for samples, (3) Bagging LASD models. We present below details on these algorithms.

5.2 Linear Programming Survival (LPS) model

The main problem in survival analysis is that of predicting time to event based on a set of relevant covariates or attributes. Linear programming approaches have been presented by Mangasarian [114] in the context of diagnosis and prognosis of breast cancer. This model does not take into account log transforming the survival time, which is the usual modeling assumption for survival regression. We incorporate this assumption into our model, and also propose a different objective function, based on concordance index. Another motivation for our model is that Cox proportional hazards regression has the proportional hazards assumption (ratio of the hazard rates for any two samples is independent of time) which is often violated in practice. The LPS model we propose is not based on this assumption, so it has more general applicability.

Let t_i be the time to event (survival or censoring time) for patient i, and δ_i be the censoring status ($\delta_i = 1$ corresponds to the event and $\delta_i = 0$ implies censoring). Let us denote by $z_{i,j}$ the value for covariate z_j for patient *i.* We assume a linear relationship between the logarithm of survival time (time to event) and the variables. This is a standard assumption in survival analysis.

$$\log(t) = \beta_0 + \sum_{j=1}^{m} z_j \beta_j + \varepsilon, \tag{5.1}$$

where, β_0 is the constant term, β_j's are coefficients for the features, and ε is the "error" or "residual" term due to the effect of other factors on survival time. We want to estimate β_0 and β_j's based on the features. This is a standard linear regression problem which can be solved by minimizing the sum of squared errors. In the case of survival regression, the complication arises due to the presence of censored observations. For the censored observations, we know that their time to event is a lower bound on their survival

time, thus we can replace the equality constraint in (1) by an inequality ($\leq$). We use the following constraints to characterize the survival regression problem:

$$\log(t_i) = \beta_0 + \sum_{j=1}^{m} z_{i,j}\beta_j + \varepsilon_i\,, \quad \forall\,\{i\,|\delta_i = 1\} \tag{5.2}$$

$$\log(t_i) \leq \beta_0 + \sum_{j=1}^{m} z_{i,j}\beta_j + \varepsilon_i, \quad \forall\,\{i \mid \delta_i = 0\} \tag{5.3}$$

Apart from minimizing the squared error, we would also like to maximize the c-index measure. For the c-index measure, over-estimation and under-estimation errors are not equivalent. The relative importance of each of these errors is a function of the time to event t_i. For each observation i, let us introduce the following notations:

- y_i: Predicted log-survival time, with estimates $\widehat{\beta_0}$ and $\widehat{\beta_J}$

$$y_i = \widehat{\beta_0} + \sum_{j=1}^{m} z_{i,j}\widehat{\beta_J} \tag{5.4}$$

- D_i: number of events before t_i

$$D_i = |\{\, j \mid \delta_j = 1, t_j < t_i\}| \tag{5.5}$$

- R_i: number of observations with time to event > t_i

$$\mathrm{R_i} = |\{\, k \mid\ t_k > t_i\}| \tag{5.6}$$

- ε_i^-: underestimation error

$$\varepsilon_i^- = \mathrm{Max}\,(\log(t_i) - \mathrm{y_i}, 0) \tag{5.7}$$

- ε_i^+: overestimation error

$$\varepsilon_i^+ = \mathrm{Max}\,(\mathrm{y_i} - \log(t_i)\,, 0) \tag{5.8}$$

Recall that in the computation of the c-index (from Chapter 1), we count the proportion of correctly ordered pairs of observations. For a given observation i, assuming that all other observations are well ordered, if we under-estimate the prediction y_i, then D_i represents the maximum number of non-concordant pairs in the c-index computation. Similarly, if we over-estimate y_i, then R_i is the maximum number of non-concordant

pairs. This is the motivation for using the weighted sum of the under and over-estimation errors with weights D_i and R_i respectively, as the objective function for the LPS model. Note that overestimation errors are not penalized for censored observation ($\delta_i = 0 => \varepsilon_i^+ = 0$). The formulation for the LPS model is shown below:

$$\begin{aligned} &\textbf{Minimize } \textstyle\sum_i D_i\varepsilon_i^- + R_i\varepsilon_i^+ \\ &\textbf{Subject to}: \\ &\beta_0 + \textstyle\sum_{j=1}^m z_{i,j}\beta_j + \varepsilon_i^- \geq \log(t_i) \qquad \forall\, i \\ &\beta_0 + \textstyle\sum_{j=1}^m z_{i,j}\beta_j - \varepsilon_i^+ \leq \log(t_i) \qquad \forall\, \{i \mid \delta_i = 1\} \\ &\varepsilon_i^+ \geq 0, \varepsilon_i^- \geq 0\,, \beta_0 \in \mathbb{R}, \beta_j \in \mathbb{R}, j = 1, \dots, m \end{aligned} \tag{5.9}$$

These models were solved using the linear programming solver MINOS [115] on NEOS servers [116]. We use the simplex algorithm for solving LPS, because it returns a basic feasible solution with many variables at 0, in contrast to interior point methods.

5.3 Logical Analysis of Survival Data

In this section we present a new methodology, Logical Analysis of Survival Data (LASD), which is a rule-based method, for estimating the survival distribution and event-risk for observations. This method is an extension of Logical Analysis of Data (LAD) for survival analysis problems. LASD is based on Boolean logic and is designed to handle binary attributes. The problem of transforming numerical attributes to binary ones, *i.e.*, discretization, is well studied, and there exist many powerful methods discussed in the survey papers [17, 18]. For gene-expression datasets, a simple discretization into high and low levels based on the median cut-point for each variable works very well in practice. Patterns built on such datasets are much more robust and can be reproduced and validated more effectively on external datasets (generated on different chips and in laboratories with different protocols).

In the following, we introduce the concept of logical survival patterns and describe a procedure to generate them. Then we present an estimation of survival function for observations based on pattern coverage. Finally we develop a method to build a survival model from the set of survival patterns.

5.3.1 Logical survival patterns

5.3.1.1 Definition

Let *n* be the number of samples and *m* the number of variables. We denote by $z_{i,k}$ the value of the k^{th} component of feature vector, $z_i \in \{0,1\}^n$ for sample i. A *logical survival pattern* is defined as a sub-cube of $\{0,1\}^n$ containing observations with homogenous survival properties. In other words, it is a set of conditions of the form $z_{i,k} = 0$ or $z_{i,k} = 1$, on some features which characterize observations with similar time to event. We define the degree, of a pattern P, $\deg(P)$, as the number of conditions necessary to characterize P (*n* minus the dimension of the sub-cube). We denote by $c_1, c_2, \dots, c_{\deg(P)}$ the set of conditions defining P. An observation i is said to be covered by pattern P if $z_{i,j}, (j = 1, \dots, \deg(P))$, satisfies the following Boolean function:

$$c_1\ AND\ c_2\ AND\ \dots AND\ c_{\deg(P)} \tag{5.10}$$

We denote by $Cov(P)$ the set of observations covered by P and by $P \backslash c$ the pattern based on P without considering the condition c.

Now we introduce two specific kinds of survival patterns. We define the *high-risk pattern at time t (HRP_t)* as a pattern covering only those observations which experience an event before t, *i.e.*, $t_i \leq t$ and $\delta_i = 1$. In a similar way, the *low-risk pattern at time t (LRP_t)* is defined as a pattern covering only those observations which experience an

event after t, *i.e.* $t_i > t$. Note that, HRP_t covers only observations that experience an event, while LRP_t covers both events and censored observations.

In most practical problems, especially in biology and medicine, this definition of survival patterns is strict and unrealistic, in practice we allow a small proportion of errors. Later, we relax the definition to: HRP_t is a pattern covering a major proportion of observations that experience an event before t, and a very small proportion of errors (observations that experience an event after t). A similar relaxation applies for LRP_t.

5.3.1.2 Pattern generation

The main idea of pattern generation is to build both HRP_t and LRP_t for the entire range of event times. We use the idea of maximum pattern generation [117] which is defined for a binary classification problem. This algorithm considers a reference observation and builds a pure pattern (covering only observations of the same class) that has the largest coverage containing it. Generally, the constraint of pure pattern is relaxed to allow a small proportion of observations of the opposite class. We apply this idea in the survival context for a reference observation by looking for maximum HRP_t and LRP_t which covers it and with the largest coverage.

Our pattern generation consists of generating these maximum survival patterns for each observation with an event in the dataset. Such a procedure is computationally very expensive. Our algorithm is based on an efficient heuristic, which is also described for classification problems [117]. We will describe briefly the procedure to build one pattern. Then, we will introduce some useful notations to fully describe the procedure.

We consider a reference observation i that experienced an event ($\delta_i = 1$) at time t_i. We illustrate the procedure for generating a HRP_{t_i} pattern for this observation. A pattern is built by backward selection. We start with a pattern that consists of conditions

on all attributes which satisfy the reference observation. With a greedy procedure we remove conditions from the pattern description based on maximizing the *separability power* of the resulting pattern (for a precise definition, see below). The separability power ensures that the pattern is getting closer to observations we want to cover and farther from the others. At the end of each iteration, the number of observations covered increases. We stop removing conditions from the pattern description when the number of errors (observations of the other class) exceeds a user defined parameter that we call *fuzziness*. In the case of an HRP_{t_i}, the incorrectly covered observations are observations with a time to event greater than t_i. For the case of generating an LRP_{t_i} then the incorrectly covered observations are observations which experienced an event ($\delta_i = 1$) before time t_i.

We introduce the following notations: Observations which had an event before time t will be denoted by $O^{\leq t} = \{ i \mid t_i \leq t \; and \; \delta_i = 1\}$ and $O^{<t} = \{ i \mid t_i < t \; and \; \delta_i = 1\}$. In a symmetric way, we represent observations which survived beyond time t by $O^{\geq t} = \{i | t_i \geq t\}$ and $O^{>t} = \{i | t_i > t\}$. So an ideal HRP_t covers only observations from the set $O^{\leq t}$ and not from $O^{>t}$. One can note that observations with $t_i \leq t \; and \; \delta_i = 0$ are not in either set For a HRP_t, we call the set $O^{\leq t}$ as positive observations and $O^{>t}$ as negative observations. The reverse remark holds for LRP_t.

Disagreement between observation O_i and a pattern P, denoted by $D_{i,P}$, is defined as the number of conditions in pattern P not satisfied by observation i.

Disagreement between set of observations S and pattern P, denoted by $D_{S,P}$, is defined as the weighted sum of $D_{O_i,P}$ where $i \in S$. The weights are the absolute differences between t_i and t.

Separability power of pattern P_t, $Sep(P_t)$ is defined as the ratio of the disagreement between pattern P_t and the negative observation set and disagreement of P_t with the positive observation set

$$Sep(P_t) = \frac{D_{O^{>t},P_t}}{D_{O^{\leq t},P_t}} \ when\ P_t\ is\ \ HRP_t \tag{5.11}$$

$$Sep(P_t) = \frac{D_{O^{<t},P_t}}{D_{O^{\geq t},P_t}} \ when\ P_t\ is\ LRP_t \tag{5.12}$$

$Sep(P_t)$ is the ability of the pattern to distinguish between high- and low-risk observations at time *t*. The higher is the value of $Sep(P_t)$, the better is the separability power of pattern P_t, resulting in more homogeneous survival characteristics of the set of observations covered by P_t. For instance, a high separability power for a HRP_t means that $D_{O^{>t},P}$ is high and/or $D_{O^{\leq t},P}$ is small.

Figure 5.1 describes the full heuristic to generate the high-risk survival pattern ($HRP_{t_{ref}}$) for a given dataset, reference observation (with time to event t_{ref}) and fuzziness. The algorithm for generating a low-risk survival pattern ($LRP_{t_{ref}}$) is symmetric to the one described above. In the next section we will use these survival patterns to estimate the survival distribution of observations.

Figure 5.1. Heuristic algorithm for generating survival patterns

Generate maximum pattern: $\boldsymbol{HRP_{t_{ref}}}$
Input: i_{ref}, fuzziness
Step1. $HRP_{t_{ref}} \coloneqq (z_1 = z_{ref,1})\ AND\ (z_2 = z_{ref,2})\ AND$ ... AND ... $(z_i = z_{ref,i})\ AND$ ... $(z_n = z_{ref,n})$
Step2. Find a condition c_k from $HRP_{t_{ref}}$ such that $HRP_{t_{ref}} \backslash c_k$ is still a $HRP_{t_{ref}}$ pattern (*i.e.,* number of low-risk observations at time *t* covered by $HRP_{t_{ref}}$ is ≤ fuzziness) and $Sep(HRP_{t_{ref}} \backslash c_k)$ is maximum.
Step3. If such a c_k exists then $HRP_{t_{ref}} \coloneqq HRP_{t_{ref}} \backslash c_k$ and go to step 2 else return $HRP_{t_{ref}}$.

5.3.2 Survival function estimator

In this section, we estimate the survival function of an observation based on the set of generated survival patterns. It is important to note here that an observation can be covered by an arbitrary number of patterns. We introduce the following concepts:

Estimated baseline survival function, $\hat{S}_B(t)$, is estimated as the Kaplan-Meier (KM) estimate of all the observations in training set

Estimated pattern survival function, $\hat{S}_P(t)$, is computed as the KM estimate of $Cov(P)$, set of observations covered by pattern *P*.

Estimated survival function of an observation, $\widehat{S}(t|z_i)$, covered by patterns $P_1, \dots, P_K$ is estimated as:

$$\widehat{S}(t|z_i) = \frac{\sum_{j=1,\dots,K} \hat{S}_{P_j}(t) + \hat{S}_B(t)}{K+1} \tag{5.13}$$

For this estimation, $\hat{S}_B(t)$ is used as baseline. Then this baseline is corrected for observations based on pattern coverage. For an observation we predict the survival function to be the mean of the survival estimates of all patterns covering it for each time

point including the baseline survival. The baseline estimate is the default estimate for an observation not covered by any of the patterns. The event-risk of an observation is evaluated as the area under the predicted survival curve.

5.3.3 Survival model

With the pattern generation method that we discussed, we generate a large set of patterns (one high and one low-risk patterns for each observation with an event). In order to reduce redundant information, we want to identify from this large set a small subset of patterns that has a high performance.

We define a *survival model* as a minimal subset of patterns which covers all observations with an event. The motivation here is to reduce redundant information and to find a simple model to allow practical interpretation by experts in the field of the application.

We use the c-index as the objective function for finding such a subset A set covering model with the c-index as an objective function is a computationally expensive procedure. Instead of solving it to optimality, we use a greedy heuristic to find such a model. This heuristic is based on forward selection. We start with an empty set of patterns, and add patterns one by one to maximize the c-index. This procedure terminates when we have covered all observations with an event in the dataset at least a fixed number of times (≥ 1), or the amount of increase in the objective function, when we select another pattern to be added to the model, is smaller than some user defined stopping parameter. The smaller the increase in the objective function, the lower is the new pattern's contribution to the model. This also addresses the problem of over-fitting, which is frequently encountered when analyzing biomedical datasets.

5.4 Bagging LASD models

In most cases, medical datasets are "noisy" due to heterogeneity intrinsic in the nature of events, thus models built on these datasets tend to have lower accuracies, and may not be robust or applicable to new or unseen data. In order to overcome these problems, we apply the concepts of Bagging for LASD. Bagging, i.e. bootstrap aggregating, is a meta-algorithm or ensemble method [23], which improves the stability and robustness of classification models. Bagging improves results mainly when the data is noisy, and the perturbed models have uncorrelated error distributions.

Bagging involves randomly partitioning the training dataset into a "bag" set and an "out-of-bag" set. The regression or classification method is applied on the bag set and predictions are made on the out-of-bag set This procedure is repeated several times, each time sampling with a uniform probability distribution with replacement. The predictions on out-of-bag set are then aggregated by weighted voting, averaging, etc. to get the final prediction.

For bagging LASD models, we aggregate the results by taking a weighted average of the predicted risk scores, where the weights are the bag accuracies [118]. The output of bagging LASD is an ensemble of LASD models. To predict the risk score for a new observation, we aggregate the results of predictions of all models in the ensemble.

5.5 Results

We illustrate the results of the proposed algorithms on the kidney cancer dataset [91] which was presented in Chapter 4. This dataset consists of 177 samples with microarray gene-expression measurements. Missing entries were imputed using k-nearest neighborhood method (k=10). Distance weighted discriminant (DWD) [101] was used to

remove systematic bias between groups of samples which were analyzed in different batches, and then the data was standard normalized by the array. In Chapter 4 we clustered this data into two subtypes ccA and ccB based on gene-expression patterns. These subtypes were analyzed separately.

Each variable was converted into a binary variable by using the median as a cut-point. The reason for selecting only one cut-point per variable is because gene-expression data is very noisy, and has low reproducibility (patterns identified in one dataset do not work well on another dataset generated using a different microarray chip, lab, etc.). Log-rank tests were used for feature selection. To increase the robustness of the selected variables, these logrank tests were run in 1000 bootstrapped experiments, each time randomly selecting 75% of the data. Finally, variables were selected if the p-value < 0.05 for ccA, (<0.025 for ccB) in at least 75% of the tests. These thresholds were selected to ensure that we obtain less than 100 important variables for building survival patterns. There were 41 and 79 binary variables selected in ccA and ccB respectively.

After feature selection, we ran our proposed methods and also other algorithms (Cox regression [119] and Random survival forests [120]) on the selected variables. LPS and Cox regression were performed on the selected variables and validated by bootstrapping 25 times. The parameters in the Cox model were tuned and optimized for c-index in training data. LASD patterns were built for each of the subtypes, pattern coverage was analyzed, and patient-specific scores were computed. We analyzed the accuracy of (i) LASD with model selection, denoted by LASD (model) and (ii) LASD without model selection, denoted by LASD (all patterns). These results were validated based on five five-folding experiments. Bagging was applied to the LASD patterns

(without model selection) in 100 bootstrapped experiments. In Bagging, no additional validation is required since Bagging already involves bootstrapping. The cross-validation estimate is the out-of-bag accuracy. To compare the performance of Bagging LASD, we ran Random survival forests (RSF) also for 100 bootstrapped trees. All the statistical analyses were run using R.2.4.1 [121].

The concordance index (c-index) for LPS, Cox regression, LASD (all patterns), LASD (model), Bagging LASD and RSF is presented in Table 5.1 for ccA and ccB subtypes. The main results of this chapter are discussed below.

Prediction results are more accurate after stratifying data into subtypes. To prove this point empirically, we also built LASD patterns on the entire dataset of 177 samples (without identifying subtypes) and the results were cross-validated by running five five-folding experiments. Using LASD (all patterns), concordance accuracy was 0.659, while with LASD (model) it was 0.677. When we used Bagging, the concordance accuracy increased to 0.695. This is much lower than the accuracies of the models built separately on ccA and ccB (Table 5.1). In general we expect more accurate results when we build models on robust clusters identified in the data. For validating results on an external dataset, we first predict the subtype that the samples belong to, and then apply the LASD patterns corresponding to the respective subtype to predict survival.

Table 5.1. Cross-validation results (concordance index and 95% confidence interval) of the proposed methods: Linear Programming Survival (LPS) model, Logical Analysis of Survival Data (LASD) for all patterns and model selection, and Bagging LASD.
Results from Cox proportional hazards regression and Random survival forests are also presented for comparison. The analyses were run separately on the two subtypes, ccA and ccB. Concordance accuracy is computed as mean accuracy of the cross-validation experiments.

	ccA	**ccB**
Cox	0.658 $\pm$ 0.033	0.516 $\pm$ 0.033
LPS	0.645 $\pm$ 0.036	0.550 $\pm$ 0.035
LASD (all patterns)	0.758 $\pm$ 0.036	0.721 $\pm$ 0.023
LASD (model)	0.732 $\pm$ 0.037	0.731 $\pm$ 0.025
Bagging LASD		
Out-of-Bag accuracy	0.759 $\pm$ 0.014	0.740 $\pm$ 0.013
Final prediction	0.749	0.776
Random survival forests		
Out-of-Bag accuracy	0.757 + 0.006	0.742 $\pm$ 0.004
Final prediction	0.74	0.761

LPS model has similar predictive power as Cox regression. They are both modeled as regression problems where the independent variables are functions of the survival time, and are linearly related to the covariates. The c-index for the LPS and Cox regression on the bootstrapping experiments are similar. The LPS model does not have the proportional hazard assumption, unlike the Cox model, and thus can be used for a wider range of applications.

LASD performs significantly better than LPS and Cox regression. Using LASD and building high degree patterns to characterize high and low-risk patients proves to be more accurate than Cox regression. This shows the high degree of complexity involved in progression and metastasis of tumors. Using model selection for LASD results in using

fewer patterns for computing survival risk, and eliminating patterns that contribute marginally to the results. Moreover, the results of LASD (model) are comparable to LASD (all patterns). LASD (model) has fewer patterns and provides higher accuracies when compared with Cox regression, which make it very advantageous for use in predicting survival.

LASD patterns have distinct survival profiles. The model for ccA consists of 9 high-risk and 8 low-risk patterns, while for ccB 16 high-risk and 6 low-risk patterns (Table 5.2A and 5.2B). These patterns cover patients which have survival distributions very different compared to the baseline, as indicated by the significant log-rank p-values. Figure 5.2 shows the KM plots for LASD patterns in Table 5.2. High risk patterns are colored red, and low risk pattern are green. It is clear from the plot that high-risk patterns cover mostly patients with early events, while low-risk patterns cover mostly patients who had late events. Figure 5.3 consists of plots of heat map of the patterns in Table 5.2. The patterns correspond to the rows and samples to the columns. The samples are ordered by their survival time. The horizontal color bar indicates the censoring status for patient (blue indicates event, and grey indicates censoring). The vertical color bar indicates the type of pattern (red indicates high-risk, and blue indicates low-risk). From the coverage heat maps and the KM curves it is evident that the patterns have distinct survival profiles.

Table 5.2A. Description of survival patterns in LASD model for ccA samples.
High-risk patterns (HR1-HR9) are those which characterize patients at risk for an event at time *t*, while low-risk patterns (LR1-LR8) characterize patients which survived beyond time *t*. "↑" represents up-regulation (≥ median) and "↓" down-regulation (< median) based on whether the gene is above or below the median. Note that all the patterns have a very significant log-rank p-value ($p<0.001$). Time represents the reference time *t* for which the pattern was built, and the score represents the area under the KM curve for the patients covered by the pattern.

Pattern ID	Time	Score	Logrank p-value	Description
HR1	4	2.5	3.11E-14	LOC286052 ↑ & ATPAF1 ↑ & UCP3 ↑ & N4BP3 ↓
HR2	10	4.96	0.00E+00	Hs.100912 ↓ & BPHL ↓ & MR1 ↑ & KCNJ8 ↑ & Hs.102471 ↑ & Hs.102471 ↑
HR3	15	5.5	0.00E+00	ATPAF1 ↑ & CEP57 ↑ & BPHL ↓ & MR1 ↑ & Hs.102471 ↑ & Hs.102471 ↑
HR4	14	6.8	7.77E-16	ATPAF1 ↑ & ATPAF1 ↓ & Hs.100912 ↓ & BPHL ↓ & LOX ↑
HR5	23	7.33	1.11E-16	ASNSD1 ↑ & MR1 ↑ & Hs.102471 ↑ & Hs.102471 ↑ & Hs.102471 ↓
HR6	19	7.6	3.45E-14	ATPAF1 ↑ & Hs.100912 ↓ & CEP192 ↓ & MR1 ↑ & LOX ↓ & KCNJ8 ↑
HR7	25	9.5	9.47E-14	ATPAF1 ↓ & MAPT ↑ & ASNSD1 ↑ & LOX ↓ & Hs.102471 ↓
HR8	38	15.5	4.67E-12	ATPAF1 ↑ & CEP192 ↓ & BPHL ↓ & KCNJ8 ↑
HR9	48	20.2	4.73E-14	CEP192 ↓ & MAPT ↑ & LOX ↓
LR1	4	186.55	1.87E-03	LOC286052 ↓
LR2	15	186.94	2.98E-03	Hs.100912 ↑ & Hs.102471 ↓
LR3	44	187.37	9.51E-03	ATPAF1 ↓ & LOX ↓
LR4	10	193.95	1.71E-04	Hs.102471 ↑
LR5	38	198.84	9.07E-04	BPHL ↑ & ASNSD1 ↓
LR6	6	232.85	3.62E-04	DCUN1D3 ↓ & KBTBD3 ↓
LR7	58	239.8	3.75E-04	CEP57 ↓ & BPHL ↓
LR8	150	249.02	3.16E-04	ATPAF1 ↓ & LOX ↑ & Hs.102276 ↓ & Hs.102471 ↓

Table 5.2B. Description of survival patterns in LASD model for ccB samples.
High-risk patterns (HR1-HR16) are those which characterize patients at risk for an event at time t, while low-risk patterns (LR1-LR6) characterize patients which survived beyond time t. "↑" represents up-regulation and "↓" down-regulation based on whether the gene is above or below the median. Note that all the patterns have a very significant log-rank p-value (p<1e-5). Time represents the reference time t for which the pattern was built, and the score represents the area under the KM curve for the patients covered by the pattern.

Pattern ID	Time	Score	Log-rank p-value	Description
HR1	2	1.25	0.00E+00	Hs.102471 ↑ & Hs.102572::Hs.602127 ↓ & Hs.103183::Hs.596971 ↓ & ZNF384 ↓ & Hs.103426 ↓ & C1orf174 ↑
HR2	2	1.4	0.00E+00	MR1 ↑ & MAN1A1 ↑ & PPARA ↓ & Hs.103183::Hs.596971 ↓ & FKBP9 ↑
HR3	4	2	0.00E+00	Hs.102471 ↑ & MAN1A1 ↑ & PPARA ↓ & COPE ↑ & C1orf174 ↑
HR4	6	2.5	0.00E+00	LIG3 ↓ & C1orf166 ↓ & Hs.102471 ↑ & MAN1A1 ↑ & PPARA ↓ & PPARA ↓
HR5	6	2.71	0.00E+00	LIG3 ↓ & C1orf166 ↓ & BPHL ↓ & MR1 ↑ & LOX ↑ & Hs.102471 ↑
HR6	7	2.75	0.00E+00	C1orf166 ↓ & BPHL ↓ & MR1 ↑ & PPARA ↓ & ZNF384 ↓
HR7	8	3.45	0.00E+00	ASTE1 ↓ & C1orf166 ↓ & BPHL ↓ & PPARA ↓ & ZNF384 ↓
HR8	9	4	5.29E-08	MR1 ↓ & LOX ↑ & PPARA ↓ & FAM104A ↑ & C1orf174 ↑
HR9	10	4.63	2.05E-10	MR1 ↓ & Hs.102471 ↑ & Hs.102572 ↓ & MAN1A1 ↑
HR10	11	5.08	1.11E-16	C1orf166 ↓ & MAN1A1 ↑ & C1orf174 ↑
HR11	17	6.64	3.93E-12	C1orf166 ↓ & BPHL ↓ & ZNF384 ↓ & FAM104A ↑
HR12	15	7.87	3.49E-09	LIG3 ↓ & Hs.102607 ↓ & PSMA1 ↑
HR13	24	8.14	7.77E-16	LIG3 ↓ & ASTE1 ↓ & ARL6IP4 ↓
HR14	29	9.86	1.45E-05	BPHL ↓ & Hs.102471 ↑ & FAM104A ↑
HR15	34	12.10	3.07E-07	MCTS1 ↑ & Hs.103334::Hs.202872 ↓ & FAM104A ↑
HR16	172	36.04	4.60E-05	PSMA1 ↑
LR1	13	150.60	2.00E-05	C1orf166 ↑ & Hs.103334::Hs.202872 ↑
LR2	14	185.63	3.78E-07	KBTBD3 ↑ & Hs.102735 ↓ & PPARA ↑
LR3	34	195.03	5.69E-08	KBTBD3 ↑ & TTC5 ↑ & Hs.102735 ↓
LR4	50	204.54	8.74E-08	Hs.100912 ↓ & Hs.102471 ↓ & TTC5 ↑
LR5	172	230.96	2.92E-06	Hs.102735 ↓ & CUL4B ↓ & Hs.103822 ↑ & FKBP9 ↓
LR6	206	242.75	6.75E-06	CUL4B ↓ & Hs.103183::Hs.596971 ↑ & Hs.103822 ↑ & FKBP9 ↓

Figure 5.2. Plot of Kaplan-Meier survival curve for patterns for (A) ccA samples, (B) ccB samples.
Red curves represent the high-risk patterns, and green represents low-risk patterns. Log-rank test for each of the patterns is highly significant (p-value < 0.001 for ccA and p-value < 0.00001 for ccB).

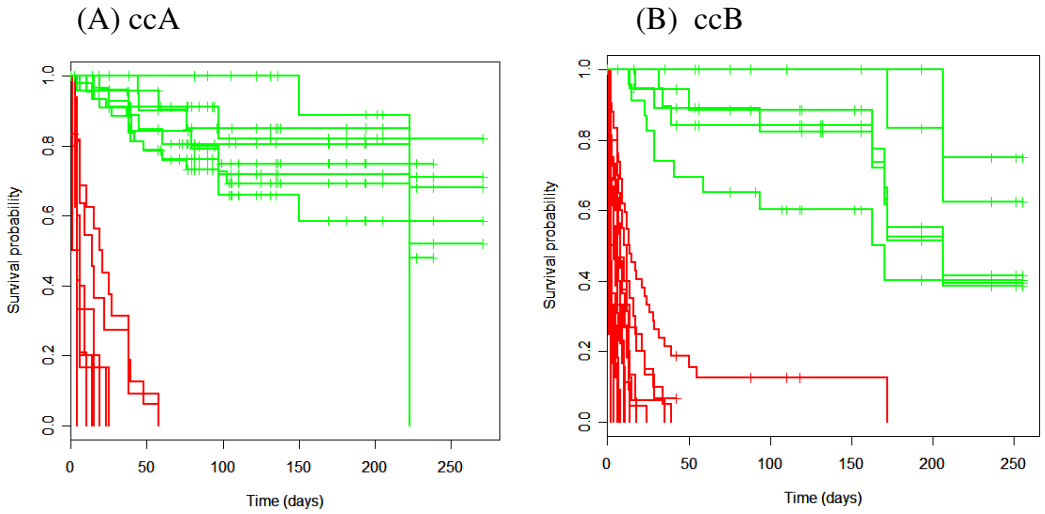
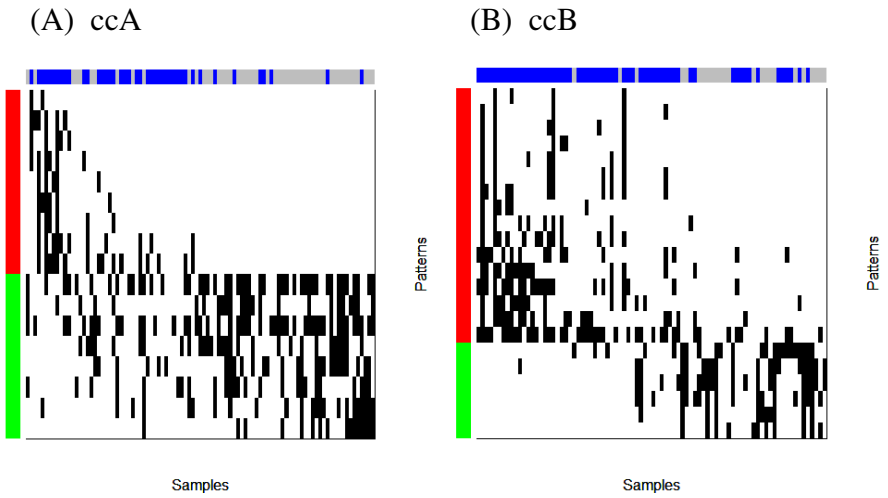


Figure 5.3. Heat map for the patterns in Table 2A for ccA samples (A), and Table 2B for ccB samples (B).
The patterns (P_i) are along the rows, and samples (S_j) are along the columns (ordered by their survival time). A cell (i,j) is colored black if pattern P_i covers sample S_j, and is colored white otherwise. The horizontal color bar represents the censoring status for the patients (grey = censored, and blue = event), the vertical color bar is red for high-risk patterns and green for low-risk patterns. The samples and patterns are sorted by increasing order of their survival times, and survival scores respectively.



Bagging improves robustness of LASD substantially. Accuracy of Bagging LASD models is slightly higher than that of LASD, but mainly the confidence intervals have reduced, improving the robustness of the predictions. Bagging LASD provides comparable results to Random survival forests, one of the most powerful ensemble methods as shown in a recent paper [120]. Note that the out-of-bag accuracies for Bagging LASD and RSF cannot be directly compared, since RSF provides out-of-bag accuracy of the k^{th} bootstrap as aggregation of results from the first k trees, while Bagging LASD, provides accuracy of the out-of-bag samples for the k^{th} tree. This is why the 95% confidence interval for RSF is much lower than that of Bagging LASD. We provide the OOB accuracy for comparison with the cross-validation accuracy with LASD and Cox regression. The main results here is the final prediction accuracy (aggregate of the out-of-bag predictions for all the bootstrapped trees).

Bagging is a technique known to perform well in practice and improve the predictions of a base learner. Bagging results in building hundreds of models, and thus it is considered to be a black box. In order to identify important variables in the bootstrapped models, we compute the importance scores for variables as their frequency of occurrence in the patterns (Table 5.3).

Table 5.3A. Top 20 variables based on importance score for ccA subtype.
This is computed as the mean frequency of occurrence of the variable in high, low-risk patterns in the Bagging LASD model for ccA samples.

Unigene cluster	Gene name	Importance score
Hs.714295		0.214
Hs.648565	ATF1	0.202
Hs.89497	LMNB1	0.188
Hs.531081	LGALS3	0.176
Hs.591957	DKFZp761E198	0.176
Hs.705395::Hs.703245		0.168

Hs.133892::Hs.713685		0.164
Hs.483564	PFDN1	0.164
Hs.44235	C13orf1	0.134
Hs.658510		0.130
Hs.194698	CCNB2	0.123
Hs.422662	VRK1	0.112
Hs.557550	NPM1	0.100
Hs.108106	UHRF1	0.090
Hs.657339	LOC440295	0.072
Hs.648565	ATF1	0.071
Hs.90756	KLB	0.070
Hs.540469		0.068
Hs.654389	CUX1	0.066
Hs.124696	BDH2	0.064

Table 5.3B. Top 20 variables based on importance score for ccB subtype.
This is computed as the average of frequency of occurrence of the variable in high, low-risk patterns in the Bagging LASD model for ccB samples.

Unigene cluster	Gene name	Importance score
Hs.22047	LOC388588	0.218
Hs.126137::Hs.705753		0.209
Hs.654668	ARHGAP26	0.157
Hs.81907	C5orf33	0.139
Hs.518475	RFC4	0.132
Hs.298023	AQP5	0.129
Hs.584801	SFRS2	0.122
Hs.664750		0.103
Hs.709753		0.103
Hs.605712		0.101
Hs.12967	SYNE1	0.097
Hs.591852	ADAM9	0.097
Hs.371823	PRDM2	0.095
Hs.74052		0.094
Hs.80305	ARHGAP19	0.094
Hs.7099	PIGG	0.093
Hs.568613	SLC25A33	0.088
Hs.662923		0.087
Hs.181173	GLB1L	0.085
Hs.372082	TNRC6B	0.080

Risk scores can be used to stratify the patients into risk groups. Figure 5.4 is a plot of predicted survival vs. actual survival time. The censored samples are marked as "+". This shows some trend, but the results are not very accurate. When we stratify the patients into 2 risk groups based on the median survival score, the survival distributions of the two risk groups are highly significant (p-value = 4E-10 for ccA, p-value = 9E-8 for ccB; Figure 5.5).

Figure 5.4. Plot of LASD survival score vs. actual survival time (in log scales). Censored samples are marked with a "+". There is a clear trend in the survival scores, but the individual scores are not very accurate.

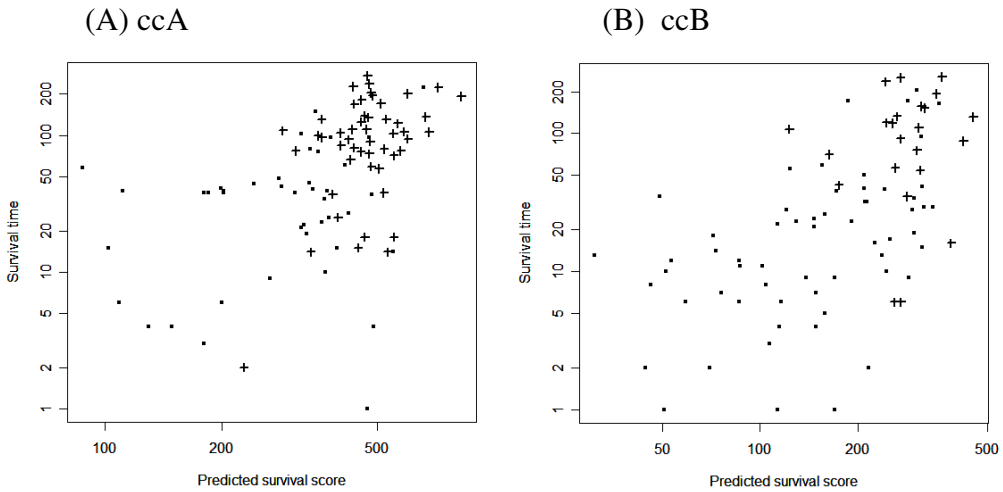
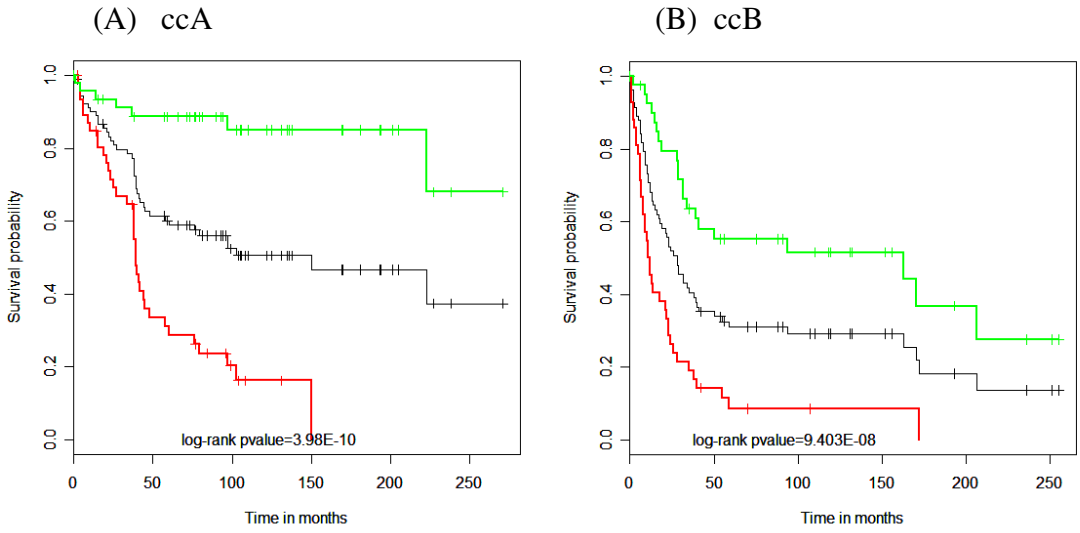


Figure 5.5. Risk stratification of patients into two groups based on the median score. The two groups have very different survival profiles (p-value = 4E-10, 9E-08 for ccA, ccB respectively).



LASD survival score is highly predictive when compared with clinical parameters (stage, grade and performance). Stage, grade and performance are clinical parameters measured for tumors, and are used in the clinic by oncologists to assess survival risk for

patients. We compute hazard ratios for the LASD score, stage, grade and performance individually (unadjusted), and also together in a multivariate Cox model (adjusted). The reason for making the unadjusted and adjusted measurements of hazard ratio is to show that the LASD score is not only accurate (individually), but also provides additional prognostic value for assessing risk for patients. Table 5.4 provides presents the results for the adjusted and unadjusted setting (hazard ratios and associated p-values).

Table 5.4. Hazard ratio (p-value) is computed for LASD prediction, and clinical parameters (stage, grade and performance) individually (unadjusted), and in a multivariate Cox regression model (adjusted).
LASD has a very significant hazard ratio not only individually, but also in the adjusted model, proving that it provides additional prognostic value for risk assessment of ccRCC tumors in ccA and ccB subtypes.

(3A) ccA	Unadjusted HR (p-value)	Adjusted HR (p-value)
LASD prediction	10.5 (2.3E-07)	10.29 (6E-07)
Stage	0.913 (0.47)	1.00 (0.98)
Grade	1.58 (0.033)	1.84 (0.016)
Performance	1.83 (0.00032)	1.77 (0.012)

(3B) ccB	Unadjusted HR (p-value)	Adjusted HR (p-value)
LASD prediction	4.1 (6E-07)	3.362 (4E-05)
Stage	0.795 (0.044)	0.938 (0.058)
Grade	1.86 (0.0018)	1.425 (0.093)
Performance	1.5 (0.00066)	1.252 (0.074)

LASD score have high predictive value for intermediate grade and stage tumors. Concordance accuracy for tumors only for intermediate grade = 0.78, 0.73, and for intermediate stage = 0.61, 0.8 for ccA and ccB subtypes, showing that the LASD scores can be used to predict survival for intermediate grade and stage (which have high error rates with current staging and grading methods).

5.6 Conclusion and discussion

We have described novel techniques to predict survival risk of tumor samples using clear cell renal cell carcinoma as a model tumor system. This cancer type is particularly well suited to this analysis, as a large percentage of patients display intermediate features of

disease aggressiveness, for which risk prediction in the clinic fails using the traditional clinical parameters (stage, grade and performance). Additionally, renal cell carcinoma (RCC) has undergone resurgence in interest owing to the rapid influx of effective therapies which can slow the growth of the disease, and may prove useful in preventing the recurrence of disease when used in the adjuvant setting.

The linear programming survival (LPS) model provides similar performance when compared to Cox regression, with the added advantage that it doesn't have the proportional hazards assumption, which is one of the main drawbacks of Cox regression. In fact this assumption is violated in many real life datasets. The other advantages are that it can reduce the dimensions of the dataset, and provide a support-set which can be used by LASD to predict event-risk with high accuracy. This model provides a basic framework for solving survival regression problems. It is very flexible, and it is possible to solve survival problems with other kinds of censoring by modifying the constraints. We could also incorporate medical knowledge, and special features of the dataset into the model. For example: we might want to give a high cost function for errors on early events, or for intermediate grade tumors, etc.

LASD is an accurate prognostic tool for the estimation of survival functions and event-risk for patients. The main advantage of LASD is that compared to classic statistical tools it can detect interactions between variables, i.e. patterns, without any prior hypotheses. Survival patterns are meaningful characterizations of groups of observations which are homogenous in terms of survival. Survival patterns HRP_t characterize a high-risk population at time *t*, while LRP_t characterize a low risk population at *t*.

Survival patterns can be represented by simple rules. They are transparent objects, and can be easily understood by medical experts and biologists. They are very useful biological research hypotheses and can be interpreted and further investigated by the experts. For instance, in the case of gene-expression profiles, patterns can detect novel interactions of genes (gene-expression signatures), which are linked with survival.

A survival pattern explains only partial information. In order to explain the entire dataset, we need a survival model, i.e. a group of survival patterns. Patterns can be combined in an infinite number of ways to build a model. In this paper, we present a general framework where the emphasis is on building concise models containing both high and low-risk patterns. For the estimation of the survival distribution, each pattern in the model has the same weight. This general framework can be modified based on the particular application, in terms of the performance and simplicity in expression. LASD algorithm is very easy to tune. It has only one parameter (fuzziness) for generating survival patterns, and an additional parameter (stopping criterion) for building the model.

In order to obtain accurate and stable predictors, we can use the concepts of bagging on LASD models. In general, ensemble methods provide much better performance when compared with the base classifiers. Based on the ensemble LASD models, we analyze the importance of attributes. Since an ensemble is equivalent to a black-box (lacks the transparency of an LASD model), these importance scores are particularly important in understanding the role of attributes in the prediction.

The Fuhrman grading system provides valuable histologic insight into patient prognosis for patients with either low or high grade tumors. However, intermediate Fuhrman grade ccRCC tumors (grades 2 and 3) are difficult for pathologists and

clinicians to classify into risk categories. Clinical stage also fails to accurately provide information regarding the risk of death from cancer for patients with completely resected tumors which displayed criteria of stage 2 or 3. Even with stage 4 metastatic disease, the highly variable natural history of this disease makes it difficult to determine how quickly to initiate a long term course of therapy with the intent to stabilize the disease growth. Thus, patients and their physicians have the potential to benefit tremendously from accurate strategies to assign risk of cancer death.

In particular, in this era of evolving targeted therapies, and their increased use in the adjuvant settings, it becomes more important than ever to be able to precisely assign prognostic risk for death from cancer. Our results suggest that the method we describe may be very valuable to urologists and oncologists in assigning a more accurate risk score to intermediate grade patients and this may in turn help clinicians determine the most appropriate therapy for an individual patient.

Chapter 6

Conclusions

Logical Analysis of Data (LAD) is a two-class classification method. In this thesis we have extended the use of LAD for survival analysis. One of the main features of LAD that makes it an attractive data mining algorithm is its ability to identify high degree patterns in the data characteristic to samples in a particular class. Predicting survival time using genetic and molecular information is receiving a lot of attention given the large amounts of data being collected for building associations for time to events like death, metastasis or recurrence of cancer. The analyses discussed in this thesis have been motivated by real medical problems: cardiovascular disease and cancer, which are the top two causes for deaths in the United States. In this thesis we have investigated survival analysis for these events. We have shown a variety of different approaches based on using LAD:

- For the case when there are few events in the data, and the study was conducted for a short period, we used a cut-point in time to define high- and low-risk samples (two classes), and then built patterns using LAD to separate the two classes. The resulting LAD discriminant score correlated well with event risk, and was used to stratify patients into risk groups.
- For a long term study, when we had few variables with low predictive power and few events occurred in the data, we focused on separating the extreme risk

groups. We also developed a novel method to identify samples which contribute most to noise in the data, and remove them from the data, in order to build robust prognostic models.

- For the case when we have microarray gene-expression data and no outcome information, we first identified robust clusters in the data and then built LAD patterns to distinguish between the clusters. Clusters or subtypes of the data in this case have a biological meaning, i.e., they have similar gene-expression signatures. We validate our subtypes on an external dataset where survival information is available, by applying the LAD patterns developed on the original data. The main result here was that the subtypes not only have different genetic signatures but also very different survival profiles.
- For survival analysis problems, which have many informative features and long term survival information, we developed extensions of LAD to build high degree survival patterns and developed a survival function for each sample based on the patterns covering the sample. We also presented a simple linear programming approach to compute event risk.

6.1 Contributions

In Chapter 2, we have provided complete analysis “in-silico to in-vitro”. We start from the PEROX risk score to show proof of concept, and then developed the CHRP(PEROX) risk score which can be translated for use in the clinic using a specialized peroxidase-based hematology analyzer, and finally CHRP risk score, which can be generated using any hematology analyzer. The CHRP(PEROX) and CHRP risk scores can be very effectively translated to clinical use by using an additional software patch in hematology

analyzers. This can be a very cheap and effective method for clinicians to assess one-year cardiovascular risk for patients. Moreover, existing risk scores, like the ATP III, Reynolds, and Duke scores can be used in a complementary fashion with the CHRP(PEROX) and CHRP risk scores to identify patients at high and low risk with higher accuracy.

We have defined a simple function to identify confusing samples in the data, specifically for the case when the variables have low predictive capability. Confusing samples are those high (low)-risk observations which have very similar measurements to observations of the opposite class. The models built on the non-confusing data are robust, and have higher accuracies. Confusing samples are a consequence of low predictive power of attributes in the data. Another way to approach this problem is to use more informative attributes. In the future we could combine the use of clinical variables along with hematology variables, or genetic attributes.

Gene expression datasets are known to be noisy, and usually data produced in different laboratories, using different chips and protocols cannot be compared. In this thesis (Chapter 4), we have presented some protocols to compare different gene-expression datasets. It is the usual practice to normalize variables or genes (variables) across observations. This method needs two main assumptions: (i) large number of observations in the data, and (ii) similar class distribution for the different datasets that we are comparing. Both of the above assumptions are often violated in practice. Usually gene-expression studies consist of a few hundred samples, and ~25,000-50,000 variables. Also, different datasets usually come from different hospitals, where it is very likely that the proportions of early and late events are significantly different. The latter assumption

of similar class distribution is not valid for the simple case when the test set consists of a single patient coming to the hospital. Instead we proposed to standard normalize the samples. This approach is simple and does not require the assumption of large sample set or class distribution. The assumption here is that most of the genes are house-keeping (they do not vary between the tumors and normals), the genes which are affected by the cancer have a large deviation from the expression of the house-keeping genes, i.e. they are either turned on or off compared to the others. Second, we proposed using a single cut-point (median) for each variable for binarization. This ensures that we do not overfit the data, and that the results are reproducible in other datasets.

We have described novel techniques to predict survival risk of tumor samples using clear cell renal cell carcinoma as a model tumor system. We propose a simple linear programming survival (LPS) model for predicting event-risk for patients, using a model where the independent variable is event time, and log(time) is linearly related to the variables. The main advantages of this model over Cox regression, which is the most popular approach for survival analysis, is that unlike the Cox model, it doesn't have the restrictive proportional hazards assumption. This model provides a basic framework for solving survival regression problems. Future work includes modeling different kinds of censoring, and cost functions into this model. We also want to explore using this approach for feature selection before applying LASD.

We have proposed a new method, LASD, for the estimation of survival functions and event-risk for patients. The main advantage of LASD contrary to classic statistical tools it can detect interactions between variables, i.e. patterns, without any prior hypotheses. This method is based on building patterns at every time-point *t* in the data

when an event occurs to distinguish samples which experienced an event before time *t* (high-risk) versus those sample which had an event after *t*. A pattern-specific score is computed as the area under the Kaplan-Meier curve for the samples satisfied by the pattern. Finally, we compute a patient-specific score as the average of the pattern scores for patterns covering the patient. We also present ensemble methods to improve the performance and robustness of LASD. The current implementation of this algorithm is much slower than that of decision trees. Future work involves building efficient data structures and algorithms to accelerate pattern generation.

In the LASD models, a large proportion of the patterns selected are usually associated with low survival times, because we use concordance accuracy as a measure for selecting patterns into the model. This measure computes the proportion of pairs of samples which are ranked in the correct order by the predicted risk score. Samples that have a very early event contribute to a large proportion of such pairs (since they are compared with all the samples with larger survival). Therefore, our procedure is strongly biased towards accurately identifying samples with potentially early events. While this bias towards early events is clearly important for determining the initial course of therapy, it may be less appropriate for decisions regarding full course of treatment or for the health insurance industry. In these cases, determining risk for later time periods or computing an average risk across all time periods may be more applicable. Indeed, if the patient survives to time *t*, the clinician would want to estimate the risk for potential events for all time periods after *t*. Such methods to assess a dynamic risk as a function of *t* for clinical use are easily devised, using straightforward extensions of the method we illustrate here. For example, one could simply choose patterns more biased towards times

greater than *t*. Similarly, to define an average risk score of interest to the health insurers, one might select patterns uniformly spaced in time. Alternately, it is also possible to use performance measures other than the concordance accuracy within the overall framework we describe here. Overall, our basic analytical method is highly malleable to a variety of different questions, needs, and end users.

6.2 Future work

Pharmaceutical companies are very interested in many aspects of drug discovery, especially finding new targets for their existing drugs. For future projects, LASD could be used for drug discovery and predicting sensitivity of known drugs in different diseases. We are currently working with a major pharmaceutical company to identify potential targets for drug "X" which is known to work on inhibiting and reducing the growth of a particular cancer. We are trying to build LASD patterns which correlate with drug sensitivity as measured by IC50 levels (drug concentration required to reduce the cell growth by 50%). This is a continuous score. We are using gene-expression microarray data measured on different cancer cell lines. The outcomes (IC50 levels) are also measured on these cell lines. The hypothesis here is that the gene-expression patterns we build across cell lines for IC50 sensitivity will correspond to signatures of sensitivity to drug X. We can use these patterns to predict potential targets among other cancers. These patterns can also be used to understand and study the biological mechanisms of drug X.

Currently, genetic studies are focused on genome wide association studies (GWAS) and analyzing high-throughput genetic data. These studies consist of measuring single nucleotide polymorphisms (SNPs) at uniformly spaced positions all over the

genome, usually ~500,000-1,000,000 SNPs. Public databases are available for GWAS for several cancers, like TCGA [122] for ovarian cancer, glioblastoma and lung cancer, CGEMs [123] for prostate and breast cancer samples, and GEO [124] for breast cancer data. These studies have case-control data or even survival outcomes. We are currently working on using patterns to discover interactions in the SNPs which correlate with survival. We are also developing novel pattern-based feature selection methods to identify a support set from the large set of variables.

There are also several public sources of data for complementary sets of molecular and genetic information (SNPs, mRNA, microRNA, methylation patterns, copy number variations, etc.) for the same samples for better understanding of the mechanism of tumors. Analyzing individual data sources is inefficient and does not take into account the interactions between the different sources of data. Also, adding all the sources together and building one model is not very effective, as it will lead to results which are not robust. Network-based integrated approaches are the preferred methodologies. The goal here is to develop a novel, mechanistic-based dynamic network model for a single disease with data of various sources. Identifying important nodes (genes/proteins) and edges (interactions) of this network will provide better targets for the treatment of the disease. On the other hand, interrogating (destabilizing) known targets in a network will result in many changes that can be potentially used as pharmacodynamic markers. We will apply standard tests to identify the topological properties of the network (scale free or random, degree distribution, connectedness, clustering coefficient, stability to perturbation, etc). By comparing the network in the disease state to that in the normal state, we will identify not only the obvious changes (loss of regulation at genes which are

shut down or amplified in disease), but also understand how these changes alter the dynamic properties of the network.

Bibliography

[1] M. L. Brennan, A. Reddy, D. M. Brennan, A. Hsu, S. A. Mann, P. L. Hammer, and S. L. Hazen, "Comprehensive peroxidase-based hematological profiling for the prediction of one-year myocardial infarction and death," *Manuscript submitted for publication*, 2008.

[2] A. Reddy, M.L. Brennan, D.M. Brennan, A. Hsu, S.A. Mann, P.L. Hammer, and S. L. Hazen, "Comprehensive Peroxidase-based Hematology Risk Profile (CHRP(PEROX)): risk predictor for one year myocardial infarction and death based on peroxidase-based hematology parameters from data available during a routine cardiac outpatient visit," *Manuscript under preparation,* 2009.

[3] A. Reddy, M.L. Brennan, D.M. Brennan, A. Hsu, S.A. Mann, P.L. Hammer, and S. L. Hazen, "Comprehensive Hematology Risk Profile (CHRP): risk predictor for one year myocardial infarction and death based on general hematology parameters from data available during a routine cardiac outpatient visit," *Manuscript under preparation,* 2009.

[4] G. Alexe and A. Reddy, "Survival ensemble models using Logical Analysis of Data for predicting long term mortality for patients undergoing coronary artery bypass surgery," *Manuscript under preparation*, 2007.

[5] G. Alexe, A. Reddy, and E. Boros, "Prediction of extremal risk patients undergoing coronary artery bypass surgery," *Manuscript under preparation*, 2007.

[6] A. R. Brannon, A. Reddy, M. Seiler, R. Pruthi, E. Wallen, B. Ljungberg, H. Zhao, J. D. Brooks, S. Ganesan, G. Bhanot, and W. K. Rathmell, "Molecular stratification of clear cell renal cell carcinoma using consensus clustering reveals distinct subtypes and survival patterns," *Manuscript submitted for publication*, 2009.

[7] A. R. Kronek L.P., "Logical Analysis of Survival Data: prognostic survival models by detecting high degree interactions in right-censored data," *Bioinformatics,* vol. 24, pp. i248-i253, 2008.

[8] A. Reddy and L.-P. Kronek, "Optimization tools for Operations Research," 2008.

[9] A. Reddy, A.R. Brannon, M. Seiler, J. Irgon, B. Ljungberg, H. Zhao, J.D. Brooks, W.K. Rathmell, S. Ganesan, and G. Bhanot, "A predictor for survival in intermediate grade clear cell renal cell carcinoma," *The 2009 International Conference on Bioinformatics & Computational Biology*, Las Vegas, 2009.

[10] Y. Crama, Hammer PL, Ibaraki T. , "Cause-effect relationships and partially defined Boolean functions.," *Ann. of Op. Res.,* vol. 16, pp. 299-326, 1988.

[11] E. Boros, P. L. Hammer, T. Ibaraki, A. Kogan, E. Mayoraz, and I. Muchnik, "An Implementation of Logical Analysis of Data," *IEEE Transaction on Knowledge and Data Engineering,* 2000.

[12] S. Alexe, E. Blackstone, P. L. Hammer, H. Ishwaran, M. S. Lauer, and C. E. Pothier Snader, "Coronary Risk Prediction by Logical Analysis of Data," in *Annals of Operations Research*, 2002.

[13] G. Alexe, S. Alexe, L. A. Liotta, M. Reiss, and P. L. Hammer, "Ovarian cancer detection by logical analysis of proteomic data," *Proteomics,* pp. 766-83, 2004.

[14] A. B. Hammer, P. L. Hammer, and I. Muchnik, "Logical analysis of Chinese labor productivity patterns," *Annals of Operations Research,* pp. 165-177, 1999.

[15] P. L. Hammer and T. O. Bonates, "Logical Analysis of data -- An overview: From combinatorial optimization to medical applications," *Annals of Operations Research,* pp. 203-225, 2006.

[16] A. Reddy, H. Wang, H. Yu, T. O. Bonates, V. Gulabani, J. Azok, G. Hoehn, P. L. Hammer, A. E. Baird, and K. C. Li, "Logical Analysis of Data (LAD) model for the early diagnosis of acute ischemic stroke," *BMC Med Inform Decis Mak,* vol. 8, p. 30, 2008.

[17] H. Liu, F. Hussain, C. L. Tan, and M. Dash, "Discretization: An enabling technique," *Data Mining and Knowledge Discovery,* pp. 393-423, 2004.

[18] S. Kotsiantis and D. Kanellopoulus, "Discretization Techniques: A recent survey," *GESTS International Transactions on Computer Science and Engineering,* vol. 32, pp. 47-58, 2006.

[19] S. Alexe, "Datascope," in *Implementation of Logical Analysis of Data methodology*.

[20] P. Lemaire, "Ladoscope," in *Implementation of Logical Analysis of Data method*.

[21] T. O. Bonates, P.L. Hammer, and A. Kogan, "Maximum patterns in datasets," *Discrete Appl. Math.,* vol. 156, 2008.

[22] D. M. Green and S. J.M, *Signal detection theory and psychophysics*. New York: John Wiley and Sons Inc., 1966.

[23] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, 2004.

[24] V. N. Vapnik, *The Nature of Statistical Learning Theory (Information Science and Statistics)*: Springer, 1999.

[25] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Belmont: Wadswoth International Group, 1984.

[26] M. LeBlanc and J. Crowley, "Relative Risk Trees for Censored Survival Data," *Biometrics,* vol. 48, pp. 411-425, June 1992.

[27] B. D. Ripley and R. M. Ripley, "Neural networks as statistical methods in survival analysis," *Artificial Neural Networks: Prospects for Medicine,* 1998.

[28] B. Zupan, J. Demsar, M.W. Kattan, J.R. Beck, and I. Bratko, "Machine learning for survival analysis: a case study on recurrence of prostate cancer," *Artificial Intelligence in Medicine,* vol. 20, pp. 59-75, 2000.

[29] O. Intrator and C. Kooperberg, "Trees and splines in survival analysis," *Statistical methods in medical research,* pp. 237-261, 1995.

[30] L. Brieman, "Bagging predictors," *Machine learning,* vol. 24(2), pp. 123-140, 1996a.

[31] T. Hothorn, B. Lausen, A. Benner, and M. Radespiel-Troeger, "Bagging survival trees," *Statistics in Medicine,* vol. 23, pp. 77-91, 2004.

[32] L. Breiman and A. Cutler, 2002.

[33] I. Ruczinski, C. Kooperberg, and M. LeBlanc, "Logic Regression," *Journal of Computational and Graphical Statistics,* pp. 475-511, 2003.

[34] H. Ishwaran, E. Blackstone, C. Pothier, and M. Lauer, "Relative risk forests for exercise heart rate recovery as a predictor of mortality," *Journal of American Statistical Association,* 2004.

[35] H. Ishwaran, U.B. Kogalur, E.H. Blackstone, M.S. Lauer, "Random survival forests," *Annals of Applied Statistics,* vol. 2, 2008.

[36] T. Hothorn, P. Buhlmann, S. Dudoit, A. Molinaro, M. van Der Laan, "Survival ensembles," *Biostatistics,* pp. 355-373, 2006.

[37] G. Alexe, S. Alexe, D. E. Axelrod, T. O. Bonates, I. L. Lozina, M. Reiss, and P. L. Hammer, "Breast cancer prognosis by combinatorial analysis of gene expression data," *Breast Cancer Research,* 2006.

[38] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, and et.al., "Diffuse large Bcell lymphoma outcome prediction by gene-expression profiling and supervized machine learning," *Nature Medicine,* pp. 68-74, 2002.

[39] M. J. van de Vijer, Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, and e. al., "A gene expression signature as a predictor of survival in breast cancer," *New England Journal of Medicine,* pp. 1999-2009, 2002.

[40] J. P. Klein and M. L. Moeschberger, *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer, 2003.

[41] F. E. Harrell, K. L. Lee, R. M. Califf, D. B. Pryor, and R. A. Rosati, "Regression modelling strategies for improved prognostics," *Statistics in Medicine,* vol. 3, pp. 143-152, 1983.

[42] E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher, "Assessment and comparison of prognostic classification schemes for survival data," *Statistics in Medicine,* pp. 2529-2545, 1999.

[43] P. Royston and W. Sauerbrei, "A new measure of prognostic separation in survival data," *Statistics in Medicine,* pp. 723-748, 2004.

[44] M. Naghavi, E. Falk, H. S. Hecht, M. J. Jamieson, S. Kaul, D. Berman, Z. Fayad, M. J. Budoff, J. Rumberger, T. Z. Naqvi, L. J. Shaw, O. Faergeman, J. Cohn, R. Bahr, W. Koenig, J. Demirovic, D. Arking, V. L. Herrera, J. Badimon, J. A. Goldstein, Y. Rudy, J. Airaksinen, R. S. Schwartz, W. A. Riley, R. A. Mendes, P. Douglas, and P. K. Shah, "From vulnerable plaque to vulnerable patient--Part III:

Executive summary of the Screening for Heart Attack Prevention and Education (SHAPE) Task Force report," *Am J Cardiol,* vol. 98, pp. 2H-15H, Jul 17 2006.

[45] A. S. Maisel, V. Bhalla, and E. Braunwald, "Cardiac biomarkers: a contemporary status report," *Nat Clin Pract Cardiovasc Med,* vol. 3, pp. 24-34, Jan 2006.

[46] R. See, J. B. Lindsey, M. J. Patel, C. R. Ayers, A. Khera, D. K. McGuire, S. M. Grundy, and J. A. de Lemos, "Application of the screening for Heart Attack Prevention and Education Task Force recommendations to an urban population: observations from the Dallas Heart Study," *Arch Intern Med,* vol. 168, pp. 1055-62, May 26 2008.

[47] T. J. Wang, P. Gona, M. G. Larson, G. H. Tofler, D. Levy, C. Newton-Cheh, P. F. Jacques, N. Rifai, J. Selhub, S. J. Robins, E. J. Benjamin, R. B. D'Agostino, and R. S. Vasan, "Multiple biomarkers for the prediction of first major cardiovascular events and death," *N Engl J Med,* vol. 355, pp. 2631-9, Dec 21 2006.

[48] S. Kathiresan, O. Melander, D. Anevski, C. Guiducci, N. P. Burtt, C. Roos, J. N. Hirschhorn, G. Berglund, B. Hedblad, L. Groop, D. M. Altshuler, C. Newton-Cheh, and M. Orho-Melander, "Polymorphisms associated with cholesterol and risk of cardiovascular events," *N Engl J Med,* vol. 358, pp. 1240-9, Mar 20 2008.

[49] R. Detrano, A. D. Guerci, J. J. Carr, D. E. Bild, G. Burke, A. R. Folsom, K. Liu, S. Shea, M. Szklo, D. A. Bluemke, D. H. O'Leary, R. Tracy, K. Watson, N. D. Wong, and R. A. Kronmal, "Coronary calcium as a predictor of coronary events in four racial or ethnic groups," *N Engl J Med,* vol. 358, pp. 1336-45, Mar 27 2008.

[50] T. A. Gaziano, C. R. Young, G. Fitzmaurice, S. Atwood, and J. M. Gaziano, "Laboratory-based versus non-laboratory-based method for assessment of cardiovascular disease risk: the NHANES I Follow-up Study cohort," *Lancet,* vol. 371, pp. 923-31, Mar 15 2008.

[51] J. Danesh, R. Collins, P. Appleby, and R. Peto, "Association of fibrinogen, C-reactive protein, albumin, or leukocyte count with coronary heart disease: meta-analyses of prospective studies," *Jama,* vol. 279, pp. 1477-82, May 13 1998.

[52] J. S. Rana, S. M. Boekholdt, P. M. Ridker, J. W. Jukema, R. Luben, S. A. Bingham, N. E. Day, N. J. Wareham, J. J. Kastelein, and K. T. Khaw, "Differential leucocyte count and the risk of future coronary artery disease in healthy men and women: the EPIC-Norfolk Prospective Population Study," *J Intern Med,* vol. 262, pp. 678-89, Dec 2007.

[53] S. Sugiyama, Y. Okada, G. K. Sukhova, R. Virmani, J. W. Heinecke, and P. Libby, "Macrophage myeloperoxidase regulation by granulocyte macrophage colony-stimulating factor in human atherosclerosis and implications in acute coronary syndromes," *Am J Pathol,* vol. 158, pp. 879-91, Mar 2001.

[54] S. J. Nicholls and S. L. Hazen, "Myeloperoxidase and cardiovascular disease," *Arterioscler Thromb Vasc Biol,* vol. 25, pp. 1102-11, Jun 2005.

[55] E. A. Podrez, D. Schmitt, H. F. Hoff, and S. L. Hazen, "Myeloperoxidase-generated reactive nitrogen species convert LDL into an atherogenic form in vitro," *J Clin Invest,* vol. 103, pp. 1547-60, Jun 1999.

[56] R. Zhang, M. L. Brennan, Z. Shen, J. C. MacPherson, D. Schmitt, C. E. Molenda, and S. L. Hazen, "Myeloperoxidase functions as a major enzymatic catalyst for initiation of lipid peroxidation at sites of inflammation," *J Biol Chem,* vol. 277, pp. 46116-22, Nov 29 2002.

[57] L. Zheng, M. Settle, G. Brubaker, D. Schmitt, S. L. Hazen, J. D. Smith, and M. Kinter, "Localization of nitration and chlorination sites on apolipoprotein A-I catalyzed by myeloperoxidase in human atheroma and associated oxidative impairment in ABCA1-dependent cholesterol efflux from macrophages," *J Biol Chem,* vol. 280, pp. 38-47, Jan 7 2005.

[58] A. K. Thukkani, J. McHowat, F. F. Hsu, M. L. Brennan, S. L. Hazen, and D. A. Ford, "Identification of alpha-chloro fatty aldehydes and unsaturated lysophosphatidylcholine molecular species in human atherosclerotic lesions," *Circulation,* vol. 108, pp. 3128-33, Dec 23 2003.

[59] S. J. Weiss, G. Peppin, X. Ortiz, C. Ragsdale, and S. T. Test, "Oxidative autoactivation of latent collagenase by human neutrophils," *Science,* vol. 227, pp. 747-9, Feb 15 1985.

[60] A. T. Askari, M. L. Brennan, X. Zhou, J. Drinko, A. Morehead, J. D. Thomas, E. J. Topol, S. L. Hazen, and M. S. Penn, "Myeloperoxidase and plasminogen activator inhibitor 1 play a central role in ventricular remodeling after myocardial infarction," *J Exp Med,* vol. 197, pp. 615-24, Mar 3 2003.

[61] J. P. Eiserich, S. Baldus, M. L. Brennan, W. Ma, C. Zhang, A. Tousson, L. Castro, A. J. Lusis, W. M. Nauseef, C. R. White, and B. A. Freeman, "Myeloperoxidase, a leukocyte-derived vascular NO oxidase," *Science,* vol. 296, pp. 2391-4, Jun 28 2002.

[62] J. A. Vita, M. L. Brennan, N. Gokce, S. A. Mann, M. Goormastic, M. H. Shishehbor, M. S. Penn, J. F. Keaney, Jr., and S. L. Hazen, "Serum myeloperoxidase levels independently predict endothelial dysfunction in humans," *Circulation,* vol. 110, pp. 1134-9, Aug 31 2004.

[63] S. Baldus, C. Heeschen, T. Meinertz, A. M. Zeiher, J. P. Eiserich, T. Munzel, M. L. Simoons, and C. W. Hamm, "Myeloperoxidase serum levels predict risk in patients with acute coronary syndromes," *Circulation,* vol. 108, pp. 1440-5, Sep 23 2003.

[64] M. Tonelli, F. Sacks, M. Arnold, L. Moye, B. Davis, and M. Pfeffer, "Relation Between Red Blood Cell Distribution Width and Cardiovascular Event Rate in People With Coronary Disease," *Circulation,* vol. 117, pp. 163-168, Jan 15 2008.

[65] S. G. Thompson, J. Kienast, S. D. Pyke, F. Haverkate, and J. C. van de Loo, "Hemostatic factors and the risk of myocardial infarction or sudden death in patients with angina pectoris. European Concerted Action on Thrombosis and

Disabilities Angina Pectoris Study Group," *N Engl J Med,* vol. 332, pp. 635-41, Mar 9 1995.

[66] P. E. Morange, C. Bickel, V. Nicaud, R. Schnabel, H. J. Rupprecht, D. Peetz, K. J. Lackner, F. Cambien, S. Blankenberg, and L. Tiret, "Haemostatic factors and the risk of cardiovascular death in patients with coronary artery disease: the AtheroGene study," *Arterioscler Thromb Vasc Biol,* vol. 26, pp. 2793-9, Dec 2006.

[67] J. Danesh, R. Collins, R. Peto, and G. D. Lowe, "Haematocrit, viscosity, erythrocyte sedimentation rate: meta-analyses of prospective studies of coronary heart disease," *Eur Heart J,* vol. 21, pp. 515-20, Apr 2000.

[68] "Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, And Treatment of High Blood Cholesterol In Adults (Adult Treatment Panel III)," *Jama,* vol. 285, pp. 2486-97, May 16 2001.

[69] P. M. Ridker, J. E. Buring, N. Rifai, and N. R. Cook, "Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score," *Jama,* vol. 297, pp. 611-9, Feb 14 2007.

[70] L. R. Smith, F. E. Harrell, Jr., J. S. Rankin, R. M. Califf, D. B. Pryor, L. H. Muhlbaier, K. L. Lee, D. B. Mark, R. H. Jones, H. N. Oldham, and *et al.*, "Determinants of early versus late cardiac death in patients undergoing coronary artery bypass graft surgery," *Circulation,* vol. 84, pp. III245-53, Nov 1991.

[71] R. Zhang, M. L. Brennan, X. Fu, R. J. Aviles, G. L. Pearce, M. S. Penn, E. J. Topol, D. L. Sprecher, and S. L. Hazen, "Association between myeloperoxidase levels and risk of coronary artery disease," *Jama,* vol. 286, pp. 2136-42, Nov 7 2001.

[72] A. Buffon, L. M. Biasucci, G. Liuzzo, G. D'Onofrio, F. Crea, and A. Maseri, "Widespread coronary inflammation in unstable angina," *N Engl J Med,* vol. 347, pp. 5-12, Jul 4 2002.

[73] D. A. Morrow, M. S. Sabatine, M. L. Brennan, J. A. de Lemos, S. A. Murphy, C. T. Ruff, N. Rifai, C. P. Cannon, and S. L. Hazen, "Concurrent evaluation of novel cardiac biomarkers in acute coronary syndrome: myeloperoxidase and soluble CD40 ligand and the risk of recurrent ischaemic events in TACTICS-TIMI 18," *Eur Heart J,* vol. 29, pp. 1096-102, May 2008.

[74] J. Loscalzo, "The macrophage and fibrinolysis," *Semin Thromb Hemost,* vol. 22, pp. 503-6, 1996.

[75] M. Navab, G. M. Ananthramaiah, S. T. Reddy, B. J. Van Lenten, B. J. Ansell, G. C. Fonarow, K. Vahabzadeh, S. Hama, G. Hough, N. Kamranpour, J. A. Berliner, A. J. Lusis, and A. M. Fogelman, "The oxidation hypothesis of atherogenesis: the role of oxidized phospholipids and HDL," *J Lipid Res,* vol. 45, pp. 993-1007, Jun 2004.

[76] T. Naruko, M. Ueda, K. Haze, A. C. van der Wal, C. M. van der Loos, A. Itoh, R. Komatsu, Y. Ikura, M. Ogami, Y. Shimada, S. Ehara, M. Yoshiyama, K.

Takeuchi, J. Yoshikawa, and A. E. Becker, "Neutrophil infiltration of culprit lesions in acute coronary syndromes," *Circulation,* vol. 106, pp. 2894-900, Dec 3 2002.

[77] P. W. Wilson, R. J. Garrison, R. D. Abbott, and W. P. Castelli, "Factors associated with lipoprotein cholesterol levels. The Framingham study," *Arterioscler Thromb Vasc Biol,* vol. 3, pp. 273-281, May 1, 1983 1983.

[78] M. S. Lauer, S. Alexe, C. E. Pothier Snader, E. H. Blackstone, H. Ishwaran, and P. L. Hammer, "Use of the logical analysis of data method for assessing long-term mortality risk after exercise electrocardiography," *Circulation,* vol. 106, pp. 685-90, Aug 6 2002.

[79] G. Alexe, G. S. Dalgin, R. Ramaswamy, C. DeLisi, and G. Bhanot, "Data Perturbation Independent Diagnosis and Validation of Breast Cancer Subtypes Using Clustering and Patterns," *Cancer Informatics,* vol. 2, pp. 243-274, 2006.

[80] M. L. Nickerson, E. Jaeger, Y. Shi, J. A. Durocher, S. Mahurkar, D. Zaridze, V. Matveev, V. Janout, H. Kollarova, V. Bencko, M. Navratilova, N. Szeszenia-Dabrowska, D. Mates, A. Mukeria, I. Holcatova, L. S. Schmidt, J. R. Toro, S. Karami, R. Hung, G. F. Gerard, W. M. Linehan, M. Merino, B. Zbar, P. Boffetta, P. Brennan, N. Rothman, W. H. Chow, F. M. Waldman, and L. E. Moore, "Improved identification of von Hippel-Lindau gene alterations in clear cell renal tumors," *Clin Cancer Res,* vol. 14, pp. 4726-34, Aug 1 2008.

[81] T. E. Hutson, G. Sonpavde, and M. D. Galsky, "Targeting growth factor and antiangiogenic pathways in clear-cell renal cell carcinoma: rationale and ongoing trials," *Clin Genitourin Cancer,* vol. 5 Suppl 1, pp. S31-9, Dec 2006.

[82] G. Sonpavde and T. E. Hutson, "Recent advances in the therapy of renal cancer," *Expert Opin Biol Ther,* vol. 7, pp. 233-42, Feb 2007.

[83] P. E. Clark, "Recent advances in targeted therapy for renal cell carcinoma," *Curr Opin Urol,* vol. 17, pp. 331-6, Sep 2007.

[84] D. G. Duda, T. T. Batchelor, C. G. Willett, and R. K. Jain, "VEGF-targeted cancer therapy strategies: current progress, hurdles and future prospects," *Trends Mol Med,* vol. 13, pp. 223-30, Jun 2007.

[85] A. R. Golshayan, A. J. Brick, and T. K. Choueiri, "Predicting outcome to VEGF-targeted therapy in metastatic clear-cell renal cell carcinoma: data from recent studies," *Future Oncol,* vol. 4, pp. 85-92, Feb 2008.

[86] I. Frank, M. L. Blute, J. C. Cheville, C. M. Lohse, A. L. Weaver, and H. Zincke, "An outcome prediction model for patients with clear cell renal cell carcinoma treated with radical nephrectomy based on tumor stage, size, grade and necrosis: the SSIGN score," *J Urol,* vol. 168, pp. 2395-400, Dec 2002.

[87] A. N. Young, V. A. Master, G. P. Paner, M. D. Wang, and M. B. Amin, "Renal epithelial neoplasms: diagnostic applications of gene expression profiling," *Adv Anat Pathol,* vol. 15, pp. 28-38, Jan 2008.

[88] M. Nogueira and H. L. Kim, "Molecular markers for predicting prognosis of renal cell carcinoma," *Urol Oncol,* vol. 26, pp. 113-24, Mar-Apr 2008.

[89] K. A. Furge, K. A. Lucas, M. Takahashi, J. Sugimura, E. J. Kort, H. O. Kanayama, S. Kagawa, P. Hoekstra, J. Curry, X. J. Yang, and B. T. Teh, "Robust classification of renal cell carcinoma based on gene expression data and predicted cytogenetic profiles," *Cancer Res,* vol. 64, pp. 4117-21, Jun 15 2004.

[90] F. Kosari, A. S. Parker, D. M. Kube, C. M. Lohse, B. C. Leibovich, M. L. Blute, J. C. Cheville, and G. Vasmatzis, "Clear cell renal cell carcinoma: gene expression analyses identify a potential signature for tumor aggressiveness," *Clin Cancer Res,* vol. 11, pp. 5128-39, Jul 15 2005.

[91] H. Zhao, B. Ljungberg, K. Grankvist, T. Rasmuson, R. Tibshirani, and J. D. Brooks, "Gene expression profiling predicts survival in conventional renal cell carcinoma," *PLoS Med,* vol. 3, p. e13, Jan 2006.

[92] C. M. Perou, T. Sorlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lonning, A. L. Borresen-Dale, P. O. Brown, and D. Botstein, "Molecular portraits of human breast tumours," *Nature,* vol. 406, pp. 747-52, Aug 17 2000.

[93] T. Sorlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. Eystein Lonning, and A. L. Borresen-Dale, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proc Natl Acad Sci U S A,* vol. 98, pp. 10869-74, Sep 11 2001.

[94] M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards, "A gene-expression signature as a predictor of survival in breast cancer," *N Engl J Med,* vol. 347, pp. 1999-2009, Dec 19 2002.

[95] S. Paik, S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin, F. L. Baehner, M. G. Walker, D. Watson, T. Park, W. Hiller, E. R. Fisher, D. L. Wickerham, J. Bryant, and N. Wolmark, "A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer," *N Engl J Med,* vol. 351, pp. 2817-26, Dec 30 2004.

[96] G. Alexe, G. S. Dalgin, D. Scanfeld, P. Tamayo, J. P. Mesirov, C. DeLisi, L. Harris, N. Barnard, M. Martel, A. J. Levine, S. Ganesan, and G. Bhanot, "High expression of lymphocyte-associated genes in node-negative HER2+ breast cancers correlates with lower recurrence rates," *Cancer Res,* vol. 67, pp. 10669-76, Nov 15 2007.

[97] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus Clustering: A resampling-based method for class discovery and visualization of gene expression microarray data.," *Machine Learning Journal,* vol. 52, pp. 91-118, 2003.

[98] G. S. Dalgin, G. Alexe, D. Scanfeld, P. Tamayo, J. P. Mesirov, S. Ganesan, C. DeLisi, and G. Bhanot, "Portraits of breast cancer progression," *BMC Bioinformatics,* vol. 8, p. 291, 2007.

[99] J. Gordan, P. Lal, V.R. Dondeti, R. Letrero, K. Parekh, C. Oquendo, R. Greenberg, K. Flaherty, W. Rathmell, B. Keith, M. Simon, and K. Nathanson, "HIF-alpha effects on c-Myc distinguish two subtypes of sporadic VHL-deficient clear cell renal carcinoma," *Cancer Cell* vol. 14, pp. 435-46, 2008.

[100] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc Natl Acad Sci U S A,* vol. 98, pp. 5116-21, Apr 24 2001.

[101] M. Benito, J. Parker, Q. Du, J. Wu, D. Xiang, C.M. Perou, and J. S. Marron, "Adjustment of systematic microarray data biases," *Bioinformatics,* vol. 20, pp. 105-114, 2004.

[102] M. J. de Hoon, S. Imoto, J. Nolan, and S. Miyano, "Open source clustering software," *Bioinformatics,* vol. 20, pp. 1453-4, Jun 12 2004.

[103] A. J. Saldanha, "Java Treeview--extensible visualization of microarray data," *Bioinformatics,* vol. 20, pp. 3246-8, Nov 22 2004.

[104] G. Dennis, Jr., B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki, "DAVID: Database for Annotation, Visualization, and Integrated Discovery," *Genome Biol,* vol. 4, p. P3, 2003.

[105] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed ed. New York: Springer-Verlag, 2002.

[106] M. E. Wall, A. Rechtsteiner, and L. M. Rocha, "Singular value decomposition and principal component analysis.," in *A Practical Approach to Microarray Data Analysis*, D. P. Berrar, W. Dubitzky, M. Granzow, and M. A. Norwell, Eds. Boston, MA: Kluwer Academic Publishers, 2003, pp. 91-109.

[107] B. S. Everitt and G. Dunn, *Applied Multivariate Data Analysis*, 2nd ed ed. London: Hodder Arnold Publication, 2001.

[108] T. Kohonen, *Self-Organizing Maps*, 3rd ed ed. New York: Springer, 2001.

[109] J. R. Vasselli, J. H. Shih, S. R. Iyengar, J. Maranchie, J. Riss, R. Worrell, C. Torres-Cabala, R. Tabios, A. Mariotti, R. Stearman, M. Merino, M. M. Walther, R. Simon, R. D. Klausner, and W. M. Linehan, "Predicting survival in patients with metastatic kidney cancer by gene-expression profiling in the primary tumor," *Proc Natl Acad Sci U S A,* vol. 100, pp. 6958-63, Jun 10 2003.

[110] H. Sultmann, A. von Heydebreck, W. Huber, R. Kuner, A. Buness, M. Vogt, B. Gunawan, M. Vingron, L. Fuzesi, and A. Poustka, "Gene expression in kidney cancer is associated with cytogenetic abnormalities, metastasis formation, and patient survival," *Clin Cancer Res,* vol. 11, pp. 646-55, Jan 15 2005.

[111] M. Takahashi, D. R. Rhodes, K. A. Furge, H. Kanayama, S. Kagawa, B. B. Haab, and B. T. Teh, "Gene expression profiling of clear cell renal cell carcinoma: gene

identification and prognostic classification," *Proc Natl Acad Sci U S A,* vol. 98, pp. 9754-9, Aug 14 2001.

[112] Y. S. Chun, J. Y. Hyun, Y. G. Kwak, I. S. Kim, C. H. Kim, E. Choi, M. S. Kim, and J. W. Park, "Hypoxic activation of the atrial natriuretic peptide gene promoter through direct and indirect actions of hypoxia-inducible factor-1," *Biochem J,* vol. 370, pp. 149-57, Feb 15 2003.

[113] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and Stratton M.R., "A census of human cancer genes," *Nat. Rev. Cancer,* vol. 4, pp. 177-183, 2004.

[114] O. L. Mangasarian, W. N. Street, and W. H. Wolberg, "Breast cancer diagnosis and prognosis via linear programming," *Operations Research,* pp. 570-577, 1995.

[115] M. Sanders, and B.A. Murtagh, "MINOS 5.1 User's Guide," Stanford University 1987.

[116] M. L. Brennan, M. S. Penn, F. Van Lente, V. Nambi, M. H. Shishehbor, R. J. Aviles, M. Goormastic, M. L. Pepoy, E. S. McErlean, E. J. Topol, S. E. Nissen, and S. L. Hazen, "Prognostic value of myeloperoxidase in patients with chest pain," *N Engl J Med,* vol. 349, pp. 1595-604, Oct 23 2003.

[117] P. L. Hammer, T. O. Bonates, and A. Kogan, "Maximum patterns in datasets," *Discrete Appl. Math.,* vol. 156, 2008.

[118] A. Reddy and G. Alexe, "Bagging Logical Analysis of Data," 2007.

[119] D. R. Cox, "Regression Models and Life Tables (with Discussion)," *Journal of Royal Statistical Society,* vol. 34, 1972.

[120] U. B. Kogalur, H. Ishwaran, E. H. Blackstone, M. S. Lauer, "Random survival forests," *Annals of Applied Statistics,* vol. 2, 2008.

[121] R. Agah, S. Ellis, S. Chase, M. Henderson, L. Mlady, G. Murugesan, R. Tubbs, K. Marchant, I. Warshawsky, C. Rouse, K. Hughes, P. Welch, and E. J. Topol, "Creation of a large-scale genetic data bank for cardiovascular association studies," *Am Heart J,* vol. 150, pp. 500-6, Sep 2005.

[122] "The Cancer Genome Atlas (http://cancergenome.nih.gov/)."

[123] "Cancer Genetic Markers of Susceptibility (http://cgems.cancer.gov/)."

[124] "Gene expression omnibus (http://www.ncbi.nlm.nih.gov/geo/)."

Curriculum Vitae

Anupama Rajasekhara Reddy

B.E., Electronics and Communication Vishweshwariah Institute of Technology, India	[1999 - 2003]
Summer Intern Merrill Lynch, New York, NY	[June - September 2004]
R&D Intern Varentas, NY, NY	[September - December 2005]
M.S., Operations Research Rutgers University, New Jersey	[2003 - 2006]
Research Assistant Department of cell biology, Cleveland Clinic, Ohio	[May - August 2007]
Research Assistant Cancer Institute of New Jersey, New Jersey	[June – August 2008]
Ph.D. candidate, Operations Research Rutgers University, New Jersey	[2006 - 2009]

Publications

L.-P. Kronek, A. Reddy: Logical Analysis of Survival Data: prognostic survival models by detecting high degree interactions in right-censored data. *Bioinformatics* 2008 24(16):i248-i253

A. Reddy, H. Wang, H. Yu, T. O. Bonates, V. Gulabani, J. Azok, G. Hoehn, P. L. Hammer, A. E. Baird, K. C. Li: Logical Analysis of Data (LAD) model for the early diagnosis of acute ischemic stroke. *BMC Medical Informatics and Decision Making* 2008, 8:30.

A. Reddy, A. R. Brannon, M. Seiler, J. Irgon, B. Ljungberg, H. Zhao, J. D. Brooks, W. K. Rathmell, S. Ganesan, G. Bhanot. A Predictor for Survival in Intermediate Grade Clear Cell Renal Cell Carcinoma. *The 2009 International Conference on Bioinformatics & Computational Biology* July 2009.

R. Mathew, C. Karp, B. Beaudoin, N. Vuong, G. Chen, H.-Y. Chen, K. Bray, A. Reddy, G Bhanot, C Gelinas, R.S. DiPaola, V. Karantza-Wadsworth and E. White. Autophagy Suppresses Tumorigenesis Through Elimination of p62. *Cell* 2009 137(1).