An Analytics Approach to Designing Clinical Trials for Cancer

by

Stephen L. Relyea

B.S., Duke University (2006)

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Master of Science in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2013

Author .....
Sloan School of Management
May 14, 2013

Certified by .....
Dimitris J. Bertsimas
Boeing Professor of Operations Research
Co-Director, Operations Research Center
Thesis Supervisor

Accepted by .....
Patrick Jaillet
Dugald C. Jackson Professor
Department of Electrical Engineering and Computer Science
Co-Director, Operations Research Center

An Analytics Approach to Designing Clinical Trials for Cancer

by

Stephen L. Relyea

Submitted to the Sloan School of Management
on May 14, 2013, in partial fulfillment of the
requirements for the degree of
Master of Science in Operations Research

Abstract

Since chemotherapy began as a treatment for cancer in the 1940s, cancer drug development has become a multi-billion dollar industry. Combination chemotherapy remains the leading treatment for advanced cancers, and cancer drug research and clinical trials are enormous expenses for pharmaceutical companies and the government. We propose an analytics approach for the analysis and design of clinical trials that can discover drug combinations with significant improvements in survival and toxicity. We first build a comprehensive database of clinical trials. We then use this database to develop statistical models from earlier trials that are capable of predicting the survival and toxicity of new combinations of drugs. Then, using these statistical models, we develop optimization models that select novel treatment regimens that could be tested in clinical trials, based on the totality of data available on existing combinations. We present evidence for advanced gastric and gastroesophageal cancers that the proposed analytics approach a) leads to accurate predictions of survival and toxicity outcomes of clinical trials as long as the drugs used have been seen before in different combinations, b) suggests novel treatment regimens that balance survival and toxicity and take into account the uncertainty in our predictions, and c) outperforms the trials run by the average oncologist to give survival improvements of several months. Ultimately, our analytics approach offers promise for improving life expectancy and quality of life for cancer patients at low cost.

Thesis Supervisor: Dimitris J. Bertsimas
Title: Boeing Professor of Operations Research
Co-Director, Operations Research Center

Acknowledgments

This work was inspired by the personal experience of my advisor Professor Dimitris Bertsimas, following his father's diagnosis with advanced cancer and subsequent passing in 2009. It is a testament to his good will and determination that he was able to transform the great difficulty of managing his father's treatment into a vision for finding better treatments for others. More than all the technical advice, guidance, and support he has provided along the way, it is this lesson of finding work that aspires to make a difference in the world that I will carry with me in the future. This work is dedicated to the memory of his father, John Bertsimas.

This research has been a joint effort with Allison O'Hair and John Silberholz, my fellow collaborators at the Operations Research Center alongside Professor Bertsimas. It has been a privilege working with such a talented team. Without their insights, commitment, and persistence, this project would not have been a success, and I hope our paths cross again in the future. I would also like to acknowledge the contributions of Dr. Natasha Markuzon of Draper Laboratory in exploring automated data extraction techniques and of Jason Acimovic, Cynthia Barnhart, Allison Chang and over twenty MIT undergraduate students in the early stages of this research.

I next would like to thank MIT Lincoln Laboratory for the scholarship which afforded me the opportunity to pursue my degree,[1] and the members of Group 36 for their support of my professional and personal development. In particular, I would like to thank my Lincoln Scholars mentor Sung-Hyun Son for his advice and assistance in pursuing a graduate education.

Finally, I would like to thank my family for all their love and support over the years – my parents, for providing me a sturdy foundation from which to grow, and my wife, Kristin, for her unwavering encouragement and reassurance throughout the challenges and successes of the past two years.

[1]Sponsored by the Department of the Air Force under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

THIS PAGE INTENTIONALLY LEFT BLANK

Contents

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 1

Introduction

Cancer is a leading cause of death worldwide, accounting for 7.6 million deaths in 2008. This number is projected to continue rising, with an estimated 13.1 million deaths in 2030 (World Health Organization 2012). The prognosis for many advanced cancers is grim unless they are caught at an early stage, when the tumor is contained and can still be surgically removed. Often, at the time of diagnosis, the cancer is sufficiently advanced that it has metastasized to other organs and can no longer be surgically removed, leaving chemotherapy and radiation as the only treatment options.

Since chemotherapy began as a treatment for cancer in the 1940s, cancer drug development has become a multi-billion dollar industry. For instance, Avastin alone generated $2.9 billion in revenues for Genentech in 2008. Though most improvements in the effectiveness of chemotherapy treatments have come from new drug development, one of the largest breakthroughs in cancer treatment occurred in 1965, when a team of researchers suggested the idea of combination chemotherapy (Chabner and Roberts 2005). Today, most successful chemotherapy treatments for advanced cancers use multiple drugs simultaneously; specifically, in this work we found that nearly 80% of all chemotherapy clinical trials for advanced gastric and gastroesophageal cancers have tested combined treatments.

Finding effective new combination chemotherapy treatments is challenging — there are a huge number of potential drug combinations, especially when considering different dosages and dosing schedules for each drug. There are some generally accepted guidelines for selecting drug combinations, such as selecting drugs that are known to be active as single agents, and combining drugs with different mechanisms of action (Page and Takimoto 2002, Pratt 1994, Golan et al. 2008).

While these serve as guidelines to clinical trial decision makers, there are still a large number of feasible combinations. In our approach, we adhere to these guidelines while optimizing over drug combinations.

Trials are also expensive, with average costs in many cases exceeding $\$10,000$ per patient enrolled (Emanuel et al. 2003); these costs are often incurred either by pharmaceutical companies or the government. Further, comparing clinical trial results is complicated by the fact that the trials are run with different patient populations; establishing one regimen as superior to another involves running a large randomized study, at a cost of millions of dollars. In this work, we develop low-cost techniques for suggesting new treatment combinations and for comparing results from trials with different patient populations.

Our aspiration in this work is to propose an analytics approach for the analysis and design of clinical trials that provides insights into what is the best currently available drug combination to treat a particular form of cancer and how to design new clinical trials that can discover improved drug combinations. The key contributions of the thesis are:

(1) We developed a database for advanced gastric and gastroesophageal cancers from papers published in the period 1979-2012. Surprisingly and to the best of our knowledge, such a database did not exist prior to this study.

(2) We construct statistical models trained on previous randomized and single-arm clinical trials to predict the outcomes of clinical trials (survival and toxicity) before they are run, when the trials' drugs have been tested before but in different combinations. One of the most important findings of this research is that the survival and toxicity outcomes of clinical trials can to a large extent be predicted in advance ($R^2 = 0.60$ for out-of-sample survival predictions, $AUC = 0.84$ for predicting high toxicity), as long as the drugs used have been seen before in different combinations.

(3) We provide evidence that our analytics based methods a) identify clinical trials that are unlikely to succeed, thus avoiding low-quality experiments, saving money and time and extending patients' lives and b) determine best treatments to date taking into account toxicity and survival tradeoffs as well as patient demographics, thus enabling doctor and patients to make more informed decisions regarding best available treatments.

(4) Perhaps most importantly, we propose an optimization-based methodology of suggesting novel treatment regimens that balances survival and toxicity and takes into account the uncertainty

in our predictions. We demonstrate that the proposed approach can quickly identify promising combinations of drugs, speeding the process of medical discovery. Specifically, we make proposals for promising trials that have been performed years later that have significantly better outcomes than the average trial.

Overall, we feel that this work provides evidence that analytics, that is the combination of data, statistical models and optimization, can offer insights on which are the best treatments today and open new frontiers in the design of promising clinical trials in the future. While the results presented here are for a specific form of cancer (gastric and gastroesophageal), the methodology is widely applicable for other forms of cancer.

Medical practitioners have a rich history of predicting clinical outcomes through the field of medical prognosis. For instance, techniques for prediction of patient survival range from simple approaches like logistic regression to more sophisticated ones such as artificial neural networks and decision trees (Ohno-Machado 2001). Most commonly, these prediction models are trained on individual patient records and used to predict the clinical outcome of an unseen patient, often yielding impressive out-of-sample predictions (Burke 1997, Delen 2005, Jefferson et al. 1997). Areas of particular promise involve incorporating biomarker and genetic information into individualized chemotherapy outcome predictions (Efferth and Volm 2005, Phan et al. 2009). Though individualized predictions represent a useful tool to patients seeking treatment, they do not enable predictions of outcomes for patients treated with previously unseen chemotherapy regimens, limiting their usefulness in planning new clinical trials.

The field of meta-regression involves building models of clinical trial outcomes such as patient survival or toxicities, trained on patient demographics and trial drug information. These regressions are used to complement meta-analyses, explaining statistical heterogeneity between the effect sizes computed from randomized clinical trials (Thompson and Higgins 2002). Though in structure meta-regressions are identical to the prediction models we build, representing trial outcomes as a function of trial properties, to date they have mainly been used as tools to explain differences in existing randomized trials, and evaluations of the predictiveness of the regression models are not performed. Like meta-analyses, meta-regressions are performed on a small subset of the clinical trials for a given disease, often containing just a few drug combinations. Even when a wide range of drug combinations are considered, meta-regressions typically do not contain enough drug-related

variables to be useful in proposing new trials. For instance, Hsu et al. (2012) predicts 1-year overall survival using only three variables to describe the drug combination in the clinical trial; new combination chemotherapy trials could not be proposed using the results of this meta-regression.

To the best of our knowledge, this work presents the first prediction model to date that contains enough detail about the drug combination and dosages used in a clinical trial to enable planning for future combination chemotherapy trials. We attain this by significantly broadening the scope of meta-regression, training our model not only on randomized clinical trials but also on single-arm trials for a given form of cancer and performing out-of-sample validation of the predictions returned. This model enables us to propose new combination chemotherapy clinical trials via optimization. To our knowledge this is a new approach to the design of clinical trials, an important problem facing the pharmaceutical industry, the government, and the healthcare industry.

Throughout the thesis, we focus on gastric and gastroesophageal cancers. Not only are these cancers very important — gastric cancer is the second leading cause of cancer death in the world and esophageal cancer is the sixth (Jemal et al. 2011) — but there is no single chemotherapy regimen widely considered to be the standard or best treatment for these cancers (Wagner 2006, Wong and Cunningham 2009, NCCN 2013).

This thesis is structured as follows. In Section 2, we describe our data collection, including how we built a database of clinical trial results for gastric and gastroesophageal cancers. In Section 3, we describe our statistical models to predict the outcomes of clinical trials. In Section 4, we propose optimization models to improve the design of clinical trials. In Section 5, we present additional applications of our statistical models to identify clinical trials that are unlikely to succeed and determine the best known treatments to date. Lastly, in Section 6, we provide some concluding remarks.

Chapter 2

Data Collection

To train statistical models to predict the results of unseen clinical trials, we first built a database of existing clinical trials for advanced and metastatic gastric and gastroesophageal cancers.

2.1 Inclusion and Exclusion Criteria

In this study, we seek to include a wide range of clinical trials, subject to the following inclusion criteria: (1) Phase I/II, Phase II or Phase III clinical trials for metastatic gastric or gastroesophageal cancer, (2) trials published no later than March 2012, the study cutoff date, (3) trials published in the English language. Notably, these criteria include non-randomized clinical trials; all published meta-analyses we are aware of for gastric cancer (e.g. Hermans (1993), Earle and Maroun (1999), Mari (2000), Wagner (2006)) are limited to randomized controlled trials, in which patients are randomly assigned to one of several arms in the trial. While including non-randomized trials provides us with a significantly larger set of clinical trial outcomes and the ability to generate predictions for a broader range of chemotherapy drug combinations, this comes at the price of needing to control for differences in demographics and other factors between different clinical trials.

Exclusion criteria were: (1) sequential trials, (2) trials that involve the application of radiation therapy,[1] (3) trials that apply curative or adjuvant chemotherapy, and (4) trials to treat gastrointestinal stromal tumors.

[1]Radiotherapy is not recommended for metastatic gastric cancer patients (NCCN 2013), and through PubMed and Cochrane searches for stomach neoplasms and radiotherapy, we only found three clinical trials using radiotherapy for metastatic gastric cancer.

2.2 Data Sources and Queries

To locate candidate papers for our database, we performed searches on PubMed, the Cochrane Central Register of Controlled Trials, and the Cochrane Database of Systematic Reviews. In the Cochrane systems, we searched for either MESH term "Stomach Neoplasms" or MESH term "Esophageal Neoplasms" with the qualifier "Drug Therapy." In PubMed, we searched for "gastr*" or "stomach" in the title and "advanced" or "metastatic" in the title and "phase" or "randomized trial" or "randomised trial" in the title. These searches yielded 350 papers that met the inclusion criteria for this study.

After this search through databases of clinical trial papers, we further expanded our set of papers by searching through the references of papers that met our inclusion criteria. This reference search yielded 56 additional papers that met the inclusion criteria for this study. In total, our systematic literature review yielded 406 papers for gastric cancer that we deemed appropriate for our approach. Since there are often multiple papers published regarding the same clinical trial, we verified that each clinical trial included was unique.

2.3 Manual Data Collection

We manually extracted data from clinical trials, and extracted data values were inputted into a database. Values not reported in the clinical trial report were marked as such in the database. For each drug in a given trial's chemotherapy treatment, the drug name, dosage level for each application, number of applications per cycle, and cycle length were collected. We also extracted many covariates that have been previously investigated for their effects on response rate or overall survival in prior chemotherapy clinical trials for advanced gastric cancer. To limit bias associated with missing information about a clinical trial, we limited ourselves to variables that are widely reported in clinical trials. These variables are summarized in Table 2.1. We chose not to collect many less commonly reported covariates that have also been investigated for their effects on response and survival, including cancer extent, histology, a patient's history of prior adjuvant therapy and surgery, and further details of patients' initial conditions, such as their baseline bilirubin levels or body surface areas (Ajani et al. 2010, Bang et al. 2010, Kang et al. 2009, Koizumi et al. 2008).

We extracted clinical trial outcome measures of interest that capture the efficacy and toxicity

of each treatment. Several measures of treatment efficacy (e.g. tumor response rate, time until tumor progression, survival time) are commonly reported in clinical trials. A review of the primary objectives of the Phase III trials in our database indicated that for the majority of these trials (62%), the primary objective was to demonstrate improvement in terms of the median overall survival (OS) of patients in the treatment group. As a result, this is the metric we have chosen as our measure of efficacy.[2] To capture the toxic effects of treatment, we also extracted the fraction of patients experiencing any toxicity at Grade 3/4 or Grade 4, designating severe, life-threatening, or disabling toxicities (National Cancer Institute 2006).

In Table 2.1, we record the patient demographics we collected as well as trial outcomes. We note that the set of toxicities reported varies across trials, and that the database contains a total of 7,360 toxicity entries, averaging 15 reported toxicities per trial arm. In Section 2.4 below, we present a method for combining this toxicity data into a single score representative of the overall toxicity of each trial.

2.4 An Overall Toxicity Score

As described in Section 2.3, we extracted the proportion of patients in a trial that experience each individual toxicity at Grade 3 or 4. In this section, we present a methodology for combining these individual toxicity proportions into a clinically relevant score that captures the overall toxicity of a treatment.

To gain insight into the rules that clinical decision makers apply in deciding whether a treatment has an acceptable level of toxicity, we referred to guidelines established in Phase I clinical trials. The primary goal of these early studies is to assess drugs for safety and tolerability on small populations and to determine an acceptable dosage level to use in later trials (Golan et al. 2008). These trials enroll patients at increasing dosage levels until the toxicity becomes unacceptable. The Patients and Methods sections of Phase I trials specify a set of so-called dose-limiting toxicities (DLTs). If a patient experiences any one of the toxicities in this set at the specified grade, he or she is said to have experienced a DLT. When the proportion of patients with a DLT exceeds a pre-determined

[2]The full survival distributions of all patients were available for only 340/483 (70.4%) of treatment arms, as opposed to the median which was available for 453/483 (93.8%). Given this limitation as well as the established use of median survival as a primary endpoint in Phase III trials, we have chosen to proceed with the median.

Variable	Average Value	Range	% Reported
	Patient Demographics		
Fraction male	0.72	0.29 – 1.00	97.9
Fraction with prior palliative chemotherapy	0.13	0.00 – 1.00	98.1
Median age (years)	59.6	46 – 80	99.2
Weighted performance status[1]	0.86	0.11 – 2.00	84.1
Fraction with primary tumor in the stomach	0.90	0.00 – 1.00	94.8
Fraction with primary tumor in the GEJ	0.07	0.00 – 1.00	94.2
	Non-drug trial information		
Fraction of study authors from each country (43 different variables)	Country Dependent	0.00 - 1.00	95.6[2]
Fraction of study authors from Asian country	0.43	0.00 – 1.00	95.6
Number of patients	54.4	11 – 521	100.0
Publication year	2003	1979 – 2012	100.0
	Trial outcomes		
Median overall survival (months)	9.2	1.8 – 22.6	93.8
Incidence of each Grade 3/4 or Grade 4 toxicity	–Toxicity Dependent–		

[1] A composite score of the Eastern Cooperative Oncology Group (ECOG) performance status of patients in a clinical trial. See Appendix A.1 for details.

[2] The remaining studies listed affiliated institutions without linking authors to institutions.

Table 2.1: Non-drug variables extracted from gastric and gastroesophageal cancer clinical trials. These variables, together with the drug variables, were inputted into a database.

threshold, the toxicity is considered "too high," and a lower dose is indicated for future trials. From these Phase I trials, we can learn the toxicities and grades that trial designers consider the most clinically relevant, and design a composite toxicity score to represent the fraction of patients with at least one DLT during treatment.

Based on a review of the 20 clinical trials meeting our inclusion criteria which also presented a Phase I study (so-called combined Phase I/II trials), we identified the following set of DLTs to include in calculation of our composite toxicity score (see Appendix A.2 for details):

1. Any Grade 3 or Grade 4 non-blood toxicity, excluding alopecia, nausea, and vomiting.

2. Any Grade 4 blood toxicity.

The threshold for the proportion of patients with a DLT that constitutes an unacceptable level of toxicity ranges from 33% to 67% over the set of Phase I trials considered, indicating the degree of variability among decision makers regarding where the threshold should be set for deciding when a trial is "too toxic."

The fraction of patients with at least one DLT during treatment cannot be calculated directly from the individual toxicity proportions reported in Phase II/III clinical trials. For instance, in a clinical trial in which 20% of patients had Grade 4 neutropenia and 30% of patients had Grade 3/4 diarrhea, the proportion of patients with a DLT might range from 30% to 50%. As a result, there is a need to combine the individual toxicity proportions into an estimate of the fraction of patients with a DLT. We evaluated five different methods for performing this combination, and found the following approach to be the most accurate (see Appendix A.3) :

1. Group the DLT toxicities into the 20 broad anatomical/pathophysiological categories defined by the National Cancer Institute Common Terminology Criteria for Adverse Events v3.0 (National Cancer Institute 2006).

2. For each trial, assign a "group score" for each group equal to the incidence of the most frequently occurring DLT in that group.

3. Compute a final toxicity score for the trial by assuming toxicities from each group occur independently of one another, with probability equal to the group score.

If one or more of the DLT toxicities for a trial arm are mentioned in the text but their values cannot be extracted (e.g. if toxicities are not reported by grade), then the overall toxicity score for that trial is marked as unavailable. This is the case for 106/483 (21.9%) of trial arms in the database.

By taking a maximum over all the reported toxicities in each category, this method has the advantage of being robust to the uneven reporting of infrequently observed toxicities within each category. Moreover, it yields a correlation coefficient of 0.893 between the estimated and actual proportion of patients with any Grade 3/4 toxicity on the 36 trial arms for which this data was available (Appendix A.3). As a result, we have confidence that our overall toxicity score calculated in this manner is a reliable indicator of the fraction of patients experiencing a DLT.

Chapter 3

Statistical Models for Clinical Trials

This section describes the development and testing of statistical models that predict the outcomes of clinical trials. These models are capable of taking a proposed clinical trial involving chemotherapy drugs that have been seen previously in different combinations and generating predictions of patient outcomes. In contrast with meta-analysis and meta-regression of clinical data, whose primary aim is the synthesis and summary of existing trials, our objective is accurate prediction on unseen future trials (out-of-sample prediction).

Key components in the development of our predictive models include (i) the definition of variables that describe a clinical trial, including the patient characteristics, chemotherapy treatment, and trial outcomes described in Table 2.1, (ii) utilization of statistical learning techniques that limit model complexity to improve predictability, and (iii) development of a sequential testing framework to evaluate our models' out-of-sample prediction accuracy. Details of each of these components and results follow in the sections below.

3.1 Data and Variables

The data that we extracted from published clinical trials described in Table 2.1 was used to develop the statistical models. This data can be classified into four categories: patient demographics, non-drug trial information, the chemotherapy treatment, and trial outcomes.

One challenge of developing statistical models using data from different clinical trials comes from the patient characteristic data. The patient populations can vary significantly from one trial

to the next. For instance, some clinical trials enroll healthier patients than others, making it difficult to determine whether differences in outcomes across trials are actually due to different treatments or only differences in the patients. To reduce the impact of such confounding variables, traditional meta-analyses often limit their scope to at most a few dozen reports with similar patient populations. We consider this reduction in scope an unnecessarily large price to pay for the goal of ensuring patient homogeneity, and choose instead to include these confounding variables explicitly in our modeling. In this way, models can be trained on the entirety of historical data, and differences in the underlying patient populations are automatically accounted for in the model. To this end, we include in our model all of the patient demographic and non-drug trial variables listed in Table 2.1, excluding the number of patients in the trial. The reporting frequencies for each of these variables is given in Table 2.1, and missing values are replaced by their variable means before model building.

For each treatment protocol we also define a set of variables to capture the chemotherapy drugs used and their dosage schedules. There exists considerable variation in dosage schedules across chemotherapy trials. For instance, consider two different trials that both use the common drug fluorouracil[1]: in the first, it is administered $3,000\,mg/m^2$ once a week, and in the second, at $200\,mg/m^2$ once a day. To allow for the possibility that these different schedules might lead to different survival and toxicity outcomes, we define variables that describe not only whether or not the drug is used (a binary variable), but we also define variables for both the instantaneous and average dosages for each drug in a given treatment. The instantaneous dose is defined as the dose administered in a single session, and the average dose is defined as the average dose delivered each week.

Lastly, for every clinical trial arm we define outcome variables to be the median overall survival and the combined toxicity score defined in Section 2.4. Trial arms without an outcome variable are removed prior to building or testing the corresponding models.

3.2 Statistical Models

We implement and test several statistical learning techniques to develop models that predict clinical trial outcomes. Information extracted from results of previously published clinical trials serve as

[1]Lutz et al. (2007) and Thuss-Patience et al. (2005)

the training database from which the model parameters are learned. Then, given a vector of inputs corresponding to patient characteristics and chemotherapy treatment variables for a newly proposed trial, the models will produce predictions of the outcomes for the new trial.

The first class of models we consider are those which assume a linear relationship between the input variables and the outcomes. If we let $\mathbf{x}$ represent a vector of inputs for a proposed trial (i.e. patient, trial, and treatment variables) and y represent a particular outcome measure we would like to predict (e.g. median survival), then this class of models assumes a relationship of the form $y = \boldsymbol{\beta}^T\mathbf{x} + \beta_0 + \epsilon$, for some unknown vector of coefficients $\boldsymbol{\beta}$, intercept β_0, and error term ϵ. We assume that the noise terms are independent and identically distributed across trial arms, as tests on the model residuals have indicated only mild heteroskedasticity of variance (see Appendix A.5).

Ordinary least squares is the classical method for learning the values of the model parameters $\boldsymbol{\beta}$ and β_0, by minimizing the sum of the squared prediction errors over the training set. However, it is well known that in settings with a relatively small ratio of data samples to predictor variables, ordinary least squares produces highly variable estimates of $\boldsymbol{\beta}$ and β_0 that are overfit to the training set. To overcome this limitation and to handle issues of collinearity between the predictor variables (see Appendix A.4), we employ two algorithms for estimating β and β_0 by minimization of the following objective:

$$\min_{\boldsymbol{\beta},\beta_0} \sum_{i=1}^{N} (\boldsymbol{\beta}^T(\mathbf{x}_i) + \beta_0 - y_i)^2 + \lambda \|\boldsymbol{\beta}\|_p \tag{3.1}$$

Here, λ is a regularization parameter that limits the complexity of the model and prevents overfitting to the training data, thereby improving prediction accuracy on future unseen trials. In building our models, we choose the value of λ through 10-fold cross-validation on the training set.

The choice of norm p leads to two different algorithms. Setting $p = 2$ yields the more traditional Ridge Regression algorithm (Hoerl and Kennard 1970), popular historically for its computational simplicity. More recently the choice of $p = 1$, known as the Lasso, has gained popularity due to its tendency to induce sparisty in the solution (Tibshirani 1996). We present results for both variants below.

The use of regularized linear models provides significant advantages over more sophisticated models in terms of simplicity, ease of interpretation, and resistance to overfitting. Nevertheless,

there is a risk that they will miss significant nonlinear effects and interactions in the data. Therefore, we also implement and test two additional techniques which are better suited to handle nonlinear relationships: Random Forests and Support Vector Machines.

The Random Forests algorithm makes no underlying assumptions about the relationship between the input and output variables. The building block for the algorithm is the regression tree (Breiman et al. 1984), which is generated by recursively partitioning the input space into regions with similar outputs. When given a new input, a regression tree will produce as a prediction the average training set output from the corresponding region. Random Forests extends this concept by introducing randomization to build an ensemble of hundreds of regression trees, and then averaging the predictions made by the individual trees (Breiman 2001). We use the nominal values recommended by Hastie et al. (2008) for the number of trees to grow (500) and minimum node size (5). The number of variable candidates at each split is chosen through 10-fold cross-validation on the training set from among exponentially spaced candidates centered at $d/3$, where d is the number of input variables.

Regression trees and therefore Random Forests are naturally able to accomodate nonlinearities and variable interactions. In addition, they have been shown to have high predictive accuracy on a number of empirical tests (Caruana and Niculescu-Mizil 2006). Nevertheless, the models and predictions they produce are not as readily interpretable as those generated by regularized linear regression.

The final modeling approach we consider is the use of support vector machines (SVM) for regression. Similar to regularized linear regression, support vector machines use regularization penalities to limit model complexity and avoid overfitting to the training data. In addition, they employ the use of kernels to allow for the implicit modeling of nonlinear and even nonparametric functions. The reader is referred to Vapnik (1998) and Cristianini and Shawe-Taylor (2000) for more details. While support vector regression offers the ability to model more complex relationships than a simple linear model, it introduces additional practical challenges, in particular the selection of an optimal kernel for modeling. Following the approach of Hsu et al. (2003), we adopt the radial basis function kernel and select the regularization parameter C and kernel parameter γ through 10-fold cross validation over an exponentially spaced 2-D grid of candidates ($C = 2^{-5}, 2^{-3}, \ldots, 2^{15}, \gamma = 2^{-15}, 2^{-13}, \ldots, 2^{3}$).

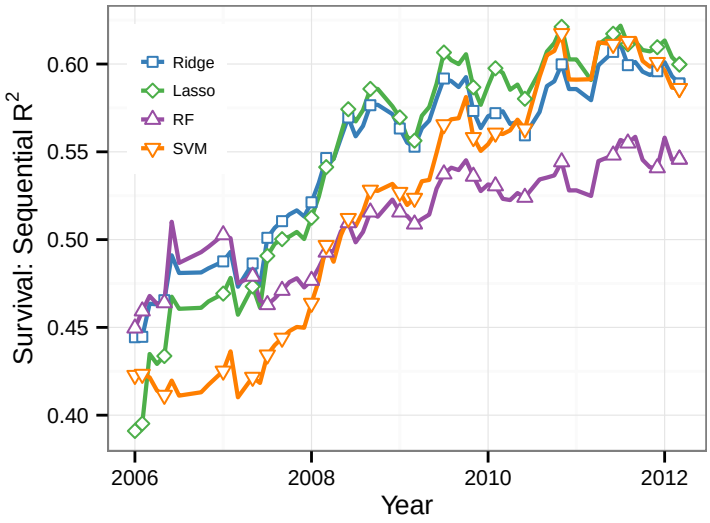
All models were built and evaluated with the statistical language **R** version 2.15.3 (R Core Team 2012) using packages `glmnet` (Friedman et al. 2010), `randomForest` (Liaw and Wiener 2002), and `e1071` (Meyer et al. 2012).

3.3 Results

Following the methodology of Section 2, we collected and extracted data from a set of 406 published journal articles from 1979–2012 describing the treatment methods and patient outcomes for a total of 483 treatment arms of gastroesophageal cancer clinical trials. Within this set, 72 different chemotherapy drugs were used in a wide variety of combinations and dosages, with 19 drugs appearing in five or more trial arms.

To compare our statistical models and evaluate their ability to predict well on unseen trials, we implement a sequential testing methodology. We begin by sorting all of the variables extracted from published clinical trials in order of their publication date. We then only use the data from prior published trials to predict the patient outcomes for each clinical trial arm. Note that we never use data from another arm of the same clinical trial to predict any clinical trial arm. This chronological approach to testing evaluates our model's capability to do exactly what will be required of it in practice: predict a future trial outcome using only the data available from the past. Following this procedure, we develop models to predict the median survival as well as the overall toxicity score. We begin our sequential testing one third of the way through the set of 483 total treatment arms, setting aside the first third (161) to use solely for model building. Of the remaining 322 trials, we predict outcomes only for those using drugs that have been seen at least once in previous trials (albeit possibly in different combinations and dosages).

The survival models are evaluated by calculating the root mean square error (RMSE) between the predicted and actual trial outcomes. They are compared against a naive predictor (labeled "Baseline"), which ignores all trial details and reports the average of previously observed outcomes as its prediction. Model performance is presented in terms of the coefficient of determination (R^2) of our prediction models relative to this baseline. To illustrate changes in model performance over time, we calculate the R^2 for each model over all test points within a 4-year sliding window. These are shown in Figure 3-1, along with the values of the RMSE and R^2 over the most recent 4-year



March 2008 – March 2012

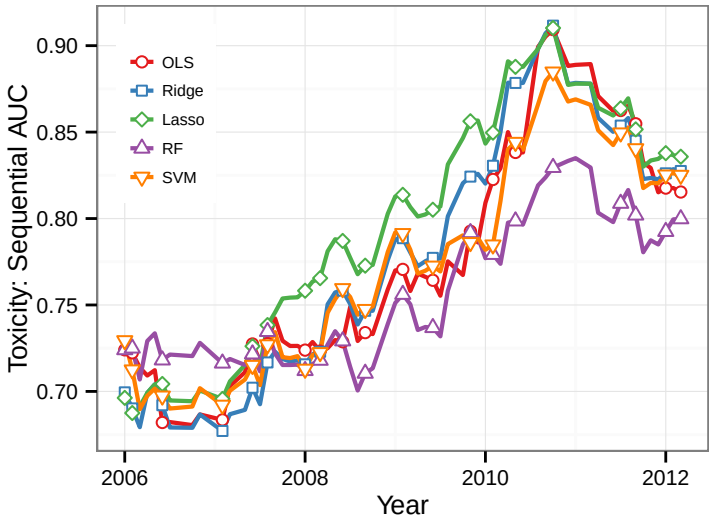
Models	**RMSE (months)**	R^2
Baseline	3.662	0
OLS	4.180	< 0
Ridge	2.348	.589
Lasso	2.317	.600
RF	2.468	.546
SVM	2.356	.586

Figure 3-1: [Left] Sequential out-of-sample prediction accuracy of survival models calculated over 4-year sliding windows ending in the date shown, reported as the coefficient of determination (R^2) of our prediction models. Ordinary least squares is not shown because all values are below 0.4. [Right] Root mean square prediction error (RMSE) and R^2 for the most recent 4-year window of data (March 2008–March 2012), which includes 134 test trial arms.

window of sequential testing.

To evaluate the toxicity models, we recall from the discussion of Section 2.4 that the toxicity of a treatment is considered manageable as long as the proportion of patients experiencing a dose-limiting toxicity is less than a fixed threshold – a typical value used in Phase I studies for this threshold is 0.5. Thus we evaluate our toxicity models on their ability to distinguish between trials with "high toxicity" (score > 0.5) and those with "low toxicity" (score ≤ 0.5). The metric we will adopt for this assessment is the area under the receiver-operating-characteristic curve (AUC). The AUC can be naturally interpreted as the probability that our models will correctly distinguish between a randomly chosen test trial arm with high toxicity against a randomly chosen test trial arm with low toxicity. As was the case for survival, we calculate the AUC for each model over a 4-year sliding window, with the results shown in Figure 3-2.

We see in Figures 3-1 and 3-2 that models for survival and toxicity all show a strong trend of improving predictability over time. This is encouraging, as it indicates our models are becoming more powerful as additional data is added to the training set. We see that the linear Ridge Regression and Lasso models perform the strongest in both cases – with model R^2 values approaching 0.6



March 2008 – March 2012

Models	**AUC**
Baseline	.500
OLS	.815
Ridge	.828
Lasso	.836
RF	.800
SVM	.825

Figure 3-2: [Left] Sequential out-of-sample classification accuracy of toxicity models calculated over 4-year sliding windows ending in the date shown, reported as the area under the curve (AUC) for predicting whether a trial will have high toxicity (score > 0.5). [Right] AUC for the most recent 4-year window of data (March 2008–March 2012), which includes 119 test trial arms. Of these, 21/119 (17.6%) actually had high toxicity.

for survival, and AUC values above 0.825 for predicting high toxicity, we have evidence that the survival and toxicity outcomes for clinical trials can be reasonably well predicted ahead of time, as long as the drugs have been seen before.

Ordinary (unregularized) least squares performs very poorly at predicting survival, which is not surprising given its tendency to overfit given the small ratio of predictor variables to training samples; its stronger performance in predicting toxicity indicates that it still has some ability to rank treatments in terms of their toxicity (which is what the AUC metric measures). Nevertheless, it is outperformed by both of the regularized linear models, and there is no reason to pursue it further. Finally, we note that the performance of the Random Forests algorithm is not competitive with the regularized linear models in terms of predicting either survival or toxicity.

As a result of this performance assessment, we identified the regularized linear models as the best candidates for inclusion in our optimization models. They are both the strongest and simplest of the models we evaluated. Before proceeding, we conducted additional testing to determine whether the explicit inclusion of pairwise interaction terms between variables improved either of the models for survival and toxicity in a significant way. We found this not to be the case (see Appendix A.6),

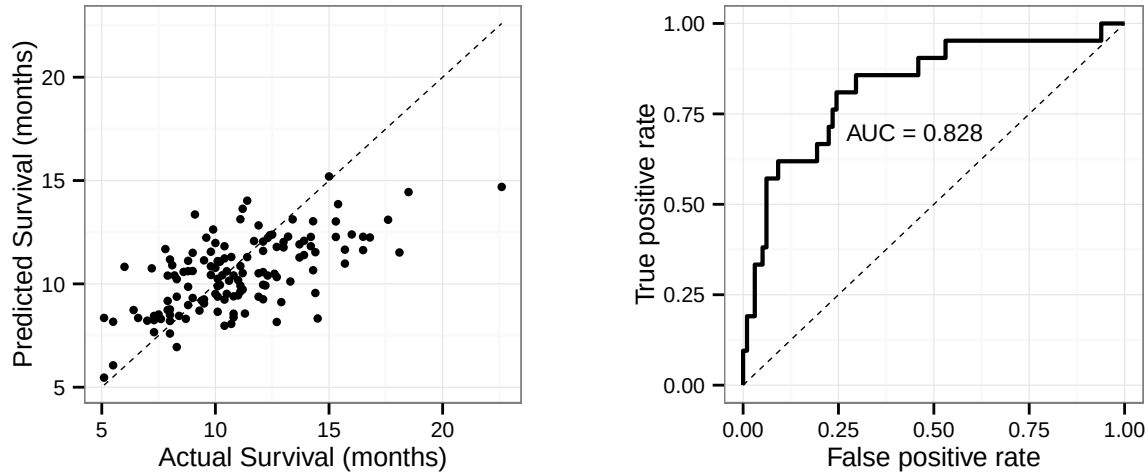


Figure 3-3: Performance of the Ridge Regression models for survival and toxicity over the most recent 4 years of data (March 2008–March 2012) [Left] Predicted vs. actual values for survival model ($n = 134$). [Right] ROC Curve for high toxicity (score > 0.5) predictions ($n = 119$), of which 21 are actually high.

and chose to proceed with the simpler models without interaction terms. Since both the Ridge Regression and Lasso models show very similar performance, and because the Ridge Regression model lends itself directly to the computation of a model uncertainty measure (Section 3.4 below) that we utilize in the design of clinical trials, we selected the Ridge Regression models to carry forward into the optimization. Depictions of the predicted vs. actual values for survival along with the receiver-operating-characteristic (ROC) curve for toxicity are shown for the Ridge Regression models in Figure 3-3.

3.4 Model Uncertainty

We implement data-driven methods for estimating and reporting the level of uncertainty in our models and their predictions. These methods provide an indication of how confident our models are in their predictions, helping to guide the design of new clinical trials with promising outcomes in the presence of uncertainty. We also envision other applications for the confidence estimation procedure developed here, such as helping determine the number of patients required to test a particular therapy before its efficacy can be known to a desired degree of accuracy.

The use of linear models allows us to leverage statistical methods for uncertainty characteriza-

tion that have been developed in this setting. If we denote by $\mathbf{X}$ the matrix of input variables in our training set (rows correspond to trial arms, columns to the demographic and drug variables), and let $\mathbf{Y}$ be a vector of recorded outcome measures for each trial (e.g. median survival), then the uncertainty in the unregularized least squares estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{U}}$ is captured by the covariance matrix $\boldsymbol{\Sigma}_{\mathbf{U}} = \sigma^2(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}$. Here σ^2 is the noise variance, for which an estimate s^2 can be derived from the residuals of the model fit. Note that when we refer to the training set here, we mean the training set up to the moment in time that we are making predictions. We calculate the uncertainty metrics at each prediction in our sequential approach, so the uncertainties are obtained only using prior data.

Variances for each of the model coefficients β_U^i can then be obtained directly from the diagonal elements of $\boldsymbol{\Sigma}_{\mathbf{U}}$, and variances for the predictions at a new trial $\mathbf{x}_0$ are estimated by $(s^2 + \mathbf{x}_0^T \boldsymbol{\Sigma}_{\mathbf{U}} \mathbf{x}_0)$. From these quantities confidence intervals can readily be defined, centered at the model estimates. To adapt these uncertainty estimators to the regularized setting, we use the following equation motivated by Bishop (2006) as an approximate uncertainty matrix for the regularized Ridge estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$:

$$\boldsymbol{\Sigma}_{\mathbf{R}} = \sigma^{\mathbf{2}}(\mathbf{X}^{\mathbf{T}}\mathbf{X} + \lambda\mathbf{I})^{-\mathbf{1}} \tag{3.2}$$

From $\boldsymbol{\Sigma}_{\mathbf{R}}$ we similarly obtain uncertainty estimates for each model coefficient, as well as confidence intervals for predictions on new trials. In addition, we can develop estimates of the average uncertainty in our model associated with each of the individual drugs. Recall that our models include three variables for each drug used in a trial arm: a binary variable indicating whether the drug is included, and two variables encoding the dosage. For a given drug i, consider a vector $\mathbf{x_i}$ whose only nonzero components are the dosage variables corresponding to drug i; for these elements we set the binary variable to 1, and the dosage variables to the average dosage administered for that drug in the training set. Then the effect on the outcome of using drug i at its average dose has an uncertainty given by:

$$\sigma_i^2 = \mathbf{x_i^T} \boldsymbol{\Sigma}_{\mathbf{R}} \mathbf{x_i} \tag{3.3}$$

This quantity provides a representation of our model's uncertainty regarding the effect of each

drug. In Figure 3-4, we provide two examples of how the model's uncertainty evolves over time as new trials are conducted and added to our training database. Shown are the model's estimates and confidence intervals for the impact of two different drugs on patient survival. On the left is the drug Oxaliplatin, first appearing in a clinical trial at the end of 2002 and appearing in 59 trial arms over the next decade. We see that the two-sigma confidence intervals start out with a wide range of nearly ten months, reflecting our large initial uncertainty regarding the drug's efficacy; they then gradually decrease over time as more trials are added to the model and improve its confidence. Moreover, as the model's best estimate of the drug's efficacy evolves over time, it remains well within the confidence intervals of prior years. We contrast this behavior with that of the drug Doxorubicin (right), which appears in 34 trial arms prior to 2004, but in only two trials afterwards. As the model gains no new information about Doxorubicin over this time period, the estimate and uncertainty of its efficacy are largely unchanged.

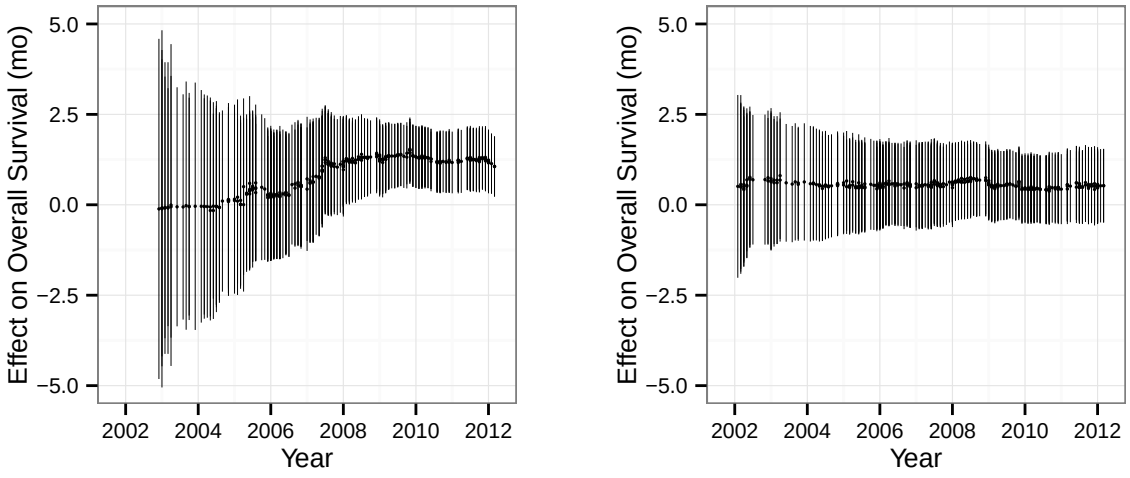


Figure 3-4: Model estimates over time for the effect of Oxaliplatin (left) and Doxorubicin (right) on the median patient survival, when considering the average dosages of each drug. Error bars depict the model's assessment of its uncertainty ($\pm 2\sigma$).

Chapter 4

Design of Clinical Trials

In this section, we present an analytical approach for designing clinical trials using mixed integer and robust optimization. Currently, most clinical trials are designed by researchers who test new therapies in the laboratory and then turn the most promising treatments into clinical trials (Page and Takimoto 2002, Pratt 1994, Golan et al. 2008). These clinical trials are then sponsored or funded by a variety of organizations or individuals such as physicians, medical institutions, foundations, pharmaceutical companies and federal agencies (ClinicalTrials.gov 2007). Using the extracted data and the predictive statistical models we have developed in Sections 2 and 3, we develop an optimization strategy for designing clinical trials. This would help researchers and organizations to select the chemotherapy drugs to use in a new clinical trial by using analytical methods to complement laboratory tests. The proposed strategy uses all the known data collected in past trials to make decisions, which is a key advantage of our approach over current practices. In this section, we will first describe the mixed integer optimization model that we use. We will then present a robust version of this model, which helps to control the stability of the predictions. Lastly, we present some results of the models and compare them to the real selections made by oncologists.

4.1 Model Formulations

Given the current data from clinical trials and the current predictive models that we have constructed, we would like to select the next best trial to perform. The determination of the next "best" trial can be made in different ways. Here, we choose to select the trial that maximizes

survival and limits toxicity, by using the predictive models presented in Section 3. Our reasoning for this is that for the majority of Phase III trials in our database, the stated primary objective was to demonstrate improvement in terms of the median overall survival (OS) of patients in the treatment group.

To use the predictive models in an optimization problem, we first need to define decision variables corresponding to the variables in these models. We first fix the patient characteristic variables described in Table 2.1 to constant values. For overall survival the specific constants do not matter as they affect all suggested combinations equally. For toxicity, the constants we choose affect the value of the rhs t we use in the formulations. We can choose values that are representative for a population of patients or a subpopulation.

We then define decision variables for the chemotherapy variables described in Section 3.1. Suppose there are n possible chemotherapy drugs to select from when designing a clinical trial. We will assume here that we are trying to select a clinical trial to perform using only existing drugs that were used in the predictive models (we start including a drug in the predictive models when it has been seen in at least one previous trial arm). We define three variables for each drug, corresponding to the chemotherapy treatment variables used in the statistical models: a binary indicator variable z_i to indicate whether drug i is or is not part of the trial ($z_i = 1$ if and only if drug i is part of the optimal chemotherapy regimen), a continuous variable u_i to indicate the instantaneous dose of drug i that should be administered in a single session, and a continuous variable v_i to indicate the average dose of drug i that should be delivered each week.

We will then use the regularized linear models from Section 3.2 with these decision variables as inputs. Let the model for overall survival (OS) be denoted by $\boldsymbol{\beta}'(\mathbf{z}, \mathbf{u}, \mathbf{v})$, where $(\mathbf{z}, \mathbf{u}, \mathbf{v})$ denotes the vector of size $3n$, where the first n entries are the binary drug variables $\mathbf{z}$, the second n entries are the instantaneous dose variables $\mathbf{u}$, and the last n entries are the average dose variables $\mathbf{v}$. Similarly, we have a model for overall toxicity, which we will denote by $\boldsymbol{\tau}'(\mathbf{z}, \mathbf{u}, \mathbf{v})$. Note that these models are all linear in the variables.

We can then select the next best trial to perform by using the following mixed integer optimization model:

$$\begin{aligned}
\text{maximize} \quad & \boldsymbol{\beta}'(\mathbf{z}, \mathbf{u}, \mathbf{v}) && && (1)\\
\text{subject to} \quad & \boldsymbol{\tau}'(\mathbf{z}, \mathbf{u}, \mathbf{v}) \leq t, && && (1a)\\
& \sum_{i=1}^{n} z_i = N, && && (1b)\\
& \mathbf{A}\mathbf{z} \leq \mathbf{b}, && && (1c)\\
& c_i z_i \leq u_i \leq C_i z_i, && i = 1, \ldots, n, && (1d)\\
& d_i z_i \leq v_i \leq D_i z_i, && i = 1, \ldots, n, && (1e)\\
& (u_i, v_i) \in \boldsymbol{\Omega}_i, && i = 1, \ldots, n, && (1f)\\
& z_i \in \{0, 1\}, && i = 1, \ldots, n. && (1g)
\end{aligned}$$

The objective of (1) maximizes the predicted overall survival of the selected chemotherapy regimen. Constraint (1a) bounds the predicted toxicity by a constant t. This constant values can be defined based on common values used in Phase I/II trials, or can be varied to suggest trials with a range of predicted toxicity. In Section 4.4, we present results from varying the toxicity limits. Constraint (1b) sets the total number of drugs in the selected trial to N, which can be varied to select trials with different numbers of drugs. We also include constraints (1c) to constrain the drug combinations that can be selected, which incorporate the generally accepted guidelines for selecting combination chemotherapy regimens. These guidelines recommend that drugs with different mechanisms of action be used, which is equivalent in our situation to selecting drugs from different classes (Page and Takimoto 2002, Pratt 1994, Golan et al. 2008). We capture this guideline by limiting the drug combination to contain no more than one drug from the classes of drugs used for gastric cancer. The other guidelines for combination chemotherapy include selecting drugs with different dose limiting toxicities, and drugs with different patterns of resistance. However, we found in our database of clinical trials that classes with the same pattern of resistance or with the same dose-limiting toxicities are often combined. Therefore, we only incorporated guidelines that were violated no more than once in our database.[1]

[1]The following pairs of classes were disallowed from being used together: anthracycline/camptothecin, alkylating

The constraints (1c) can also eliminate or force other combinations of drugs, which may be necessary due to the known toxicities and properties of the drugs. Additionally, these constraints can be used to add preferences of the business or research group running the clinical trial. For example, a pharmaceutical company may want to force a new drug they have developed and only tested a few times to be used in the trial. In this case, the optimal solution will be the best drug combination containing the necessary drug.

Constraints (1d) give a lower bound c_i and a upper bound C_i for each drug's instantaneous dose u_i, given that the drug i has been selected. These bounds have been defined through Phase I clinical trials. Constraints (1e) similarly provide upper and lower bounds for each drug's average dose v_i. Constraints (1f) limit u_i and v_i to belong to a feasible set $\mathbf{\Omega}_i$. This is important since the instantaneous dose and the average dose are often not independent. In the results shown later in this section, we let $\mathbf{\Omega}_i$ be all combinations of instantaneous and average doses that have been used in prior clinical trials. This forces the dosage for a particular drug to be realistic. Lastly, constraints (1g) define $\mathbf{z}$ to be a binary vector of decision variables. For the remainder of the thesis, we will refer to the feasible set of (1), that is the set of all vectors $\mathbf{z}$, $\mathbf{u}$, and $\mathbf{v}$ satisfying constraints (1a)–(1g), as $\hat{\mathbf{W}}$.

While the optimization model (1) finds the single best trial to suggest, we are also interested in building a model to suggest k different trials at once. One reason for this is for recommending trials when multiple trials will be run at once. We would like to suggest different trials that will all provide us with interesting information, before knowing the results of each of the other trials. Additionally, we would also like to see all of the best k trials since there are often several different drug combinations with similar overall survival predictions, and one is not necessarily superior to the others. We can thus alter model (1) to propose k different trials. We do this by including k vectors of binary decision variables, $\{\mathbf{z}^1, \mathbf{z}^2, \ldots, \mathbf{z}^k\}$, k vectors of instantaneous dose decision variables, $\{\mathbf{u}^1, \mathbf{u}^2, \ldots, \mathbf{u}^k\}$, and k vectors of average dose decision variables, $\{\mathbf{v}^1, \mathbf{v}^2, \ldots, \mathbf{v}^k\}$. We then solve the following problem:

agent/taxane, taxane/topoisomerase II inhibitor, antimetabolite/protein kinase, and camptothecin/topoisomerase II inhibitor. If a chemoprotectant drug is used, it must be used with a drug from the antimetabolite class that is not capecitabine.

$$\text{maximize} \quad \sum_{j=1}^{k} \boldsymbol{\beta}'(\mathbf{z}^j, \mathbf{u}^j, \mathbf{v}^j) \tag{2}$$

$$\text{subject to} \quad (\mathbf{z}^j, \mathbf{u}^j, \mathbf{v}^j) \in \hat{\mathbf{W}}, \qquad j = 1, \ldots, k, \tag{2a}$$

$$\mathbf{z}^{j_1} \neq \mathbf{z}^{j_2}, \qquad j_1 = 1, \ldots, k-1, \;\; j_2 = j_1 + 1, \ldots, k. \tag{2b}$$

The objective of (2) aims to maximize the total survival of the k different trials. Constraints (2a) require each selected trial meet the constraints of (1). Constraints (2b) prevent any pair of suggested trials from having identical drugs, and can be implemented using standard techniques; we will not elaborate further here because in practice our models will be solved using the column generation approach described in Section 4.3. In the remainder of the thesis, we will refer to all variables satisfying constraints (2a) and (2b) as the feasible set $\mathbf{W}$. Note that this formulation requires the k trials to all be different, but they could be very similar. The k trials are only required to have one different drug between any two trials. We will see in the next section how the trials are encouraged to be more diverse using robust optimization.

4.2 A Robust Optimization Model

While the models presented in Section 4.1 correctly select the best trials using the predictive models, the optimal solution can be very sensitive to the coefficients of the regularized linear model for survival ($\boldsymbol{\beta}$). To handle this, we use robust optimization to allow $\boldsymbol{\beta}$ to vary in an uncertainty set. As described in Section 3.4, we constructed an uncertainty measure to capture the uncertainty of each drug. In the following optimization formulation, we allow the binary drug coefficients to vary in this uncertainty set, while keeping the instantaneous and average dose coefficients fixed.

Denoting the feasible set of (2) by $\mathbf{W}$, we can rewrite (2) as

$$\max_{(\mathbf{z}^j, \mathbf{u}^j, \mathbf{v}^j) \in \mathbf{W}, \forall j} \quad \sum_{j=1}^{k} [(\boldsymbol{\beta}^z)'\mathbf{z}^j + (\boldsymbol{\beta}^u)'\mathbf{u}^j + (\boldsymbol{\beta}^v)'\mathbf{v}^j] \tag{3}$$

We can then reformulate this as the following robust problem:

$$\max_{(\mathbf{z}^j,\mathbf{u}^j,\mathbf{v}^j)\in\mathbf{W},\forall j} \min_{\boldsymbol{\beta}^z} \quad \sum_{j=1}^{k}[(\boldsymbol{\beta}^z)'\mathbf{z}^j + (\bar{\boldsymbol{\beta}}^u)'\mathbf{u}^j + (\bar{\boldsymbol{\beta}}^v)'\mathbf{v}^j] \tag{4}$$

$$\text{subject to} \quad \frac{|\beta_i^z - \bar{\beta}_i^z|}{\sigma_i} \leq \Gamma, \qquad i = 1,\ldots,n, \tag{4a}$$

$$\sum_{i=1}^{n} \frac{|\beta_i^z - \bar{\beta}_i^z|}{\sigma_i} \leq \Gamma\sqrt{N}, \tag{4b}$$

where $\boldsymbol{\beta}^z$ is now a vector of variables, and $\bar{\boldsymbol{\beta}}^z$, $\bar{\boldsymbol{\beta}}^u$, and $\bar{\boldsymbol{\beta}}^v$ are the coefficient values of the predictive models that have been constructed for the binary variables, instantaneous dose variables, and average dose variables, respectively. For each drug i, σ_i is the uncertainty parameter described in Section 3.4. The parameter Γ controls how conservative we would like to be. Constraints (4a) restrict each coefficient β_i^z to be at most $\Gamma\sigma_i$ larger or smaller than the nominal coefficient $\bar{\beta}_i^z$. Constraint (4b) further restricts the sum of the normalized deviations of β_i^z from $\bar{\beta}_i^z$ to be no more than $\Gamma\sqrt{N}$, where N is the number of drugs that can be selected in a single trial. This constraint prevents the robust model from being too conservative.

For a fixed set of $(\mathbf{z}^j, \mathbf{u}^j, \mathbf{v}^j) \in \mathbf{W}$, the inner optimization problem selects the worst possible vector of coefficients $\boldsymbol{\beta}^z$ that is feasible, given the constraints limiting $\boldsymbol{\beta}^z$ to be close to $\bar{\boldsymbol{\beta}}^z$. The outer optimization problem then tries to find the best set of trials $(\mathbf{z}^j, \mathbf{u}^j, \mathbf{v}^j) \in \mathbf{W}$ given this worst case approach. This problem is thus robust in the sense that we are trying to maximize the worst case scenario in a given uncertainty set. This approach combines Wald's maximin model (Wald 1945) with a parameter to control how conservative the solution is, or the price of robustness (Bertsimas and Sim 2004).

Constraint (4b) also serves to encourage the k trials to be different from each other. If many drugs are selected, many coefficients will contribute to the objective, and constraint (4b) will prevent them all from being pushed to their worst case bound. On the contrary, if only a few drugs are selected, only a few coefficients will contribute to the objective, and constraint (4b) allows all of

them to be closer to their worst case bound. Evidence of this effect will be shown in the results section below.

To solve (4), we first reformulate the problem to eliminate all absolute values, using standard linear optimization techniques (Bertsimas and Tsitsiklis 1997). We then take the dual of the inner problem, resulting in a mixed integer optimization problem that can be solved as before. Note that (3) is a special case of the robust problem (4), where Γ is set to zero.

4.3 A Column Generation Approach

The optimization model (4) solves very slowly when asked for even a modest number of suggestions, due to the symmetry in the formulation of $\mathbf{W}$. Here, we present a fast column generation-based approach to generate approximate solutions to (4).

Define variables δ^t associated with each feasible $(\mathbf{z}^t, \mathbf{u}^t, \mathbf{v}^t) \in \hat{\mathbf{W}}$. Let T be the set of all drug tuples of size N, and let V_R be the set of all feasible treatment indexes of treatments using tuple $R \in T$. Finally let $\mathbf{U}$ be uncertainty set for $\boldsymbol{\beta}^z$, as defined in (4a) and (4b) above. Then (4) can be reformulated as:

$$\max_{\delta^t} \min_{\boldsymbol{\beta}^z \in \mathbf{U}} \quad \sum_t \sum_{i=1}^{n} (\beta_i^z z_i^t \delta^t + \bar{\beta}_i^u u_i^t \delta^t + \bar{\beta}_i^v v_i^t \delta^t) \tag{5}$$

$$\text{subject to} \quad \sum_t \delta^t = k, \tag{5a}$$

$$\sum_{t \in V_R} \delta^t \leq 1, \qquad \forall R \in T, \tag{5b}$$

$$\delta^t \in \{0, 1\}, \qquad \forall t. \tag{5c}$$

Constraint (5a) requires that we select k different treatments, and constraint (5b) prevents us from selecting more than one treatment with the same set of drugs. As an approximation, we consider a relaxed version of (5), where constraint (5c) is replaced with $0 \leq \delta^t \leq 1$. Expanding out the objective by dualizing the inner problem, we have:

$$\max_{\delta^t,\alpha^+,\alpha^-,\rho,\gamma} \quad \sum_{i=1}^{n}(-\bar{\beta}_i^z\alpha_i^+ + \bar{\beta}_i^z\alpha_i^- - \Gamma\rho_i - \Gamma\sqrt{N}\gamma) + \sum_t\sum_{i=1}^{n}(\bar{\beta}_i^u u_i^t\delta^t + \bar{\beta}_i^v v_i^t\delta^t) \tag{6}$$

$$\begin{aligned}
\text{subject to} \quad & \sigma_i\alpha_i^+ + \sigma_i\alpha_i^- - \rho_i - \gamma \le 0, && i = 1,\dots,n, \\
& -\alpha_i^+ + \alpha_i^- = \sum_t z_i^t\delta^t, && i = 1,\dots,n, \quad (\epsilon_i) \\
& \sum_t \delta^t = k, && (f) \\
& \sum_{t\in V_R}\delta^t \le 1, && \forall R \in T, \ (m_R) \\
& 0 \le \delta^t \le 1, && \forall t, \\
& \alpha_i^+, \alpha_i^-, \rho_i \ge 0, && i = 1,\dots n, \\
& \gamma \ge 0.
\end{aligned}$$

The optimal solution to (6) provides an upper bound on the optimal solution to (5). We solve (6) by column generation, adding at each iteration the variable δ^t corresponding to the optimal solution of the following mixed integer program:

$$\max_{\mathbf{z},\mathbf{u},\mathbf{v},\mathbf{s}} \quad \sum_{i=1}^{n}(\epsilon_i z_i + \bar{\beta}_i^u u_i + \bar{\beta}_i^v v_i) - f - \sum_{R\in T} s_R \tag{7}$$

$$\begin{aligned}
\text{subject to} \quad & (\mathbf{z},\mathbf{u},\mathbf{v}) \in \mathbf{W}, \\
& s_R \ge m_R[(\sum_{i\in R} z_i) - (N-1)], && \forall R \in T, \\
& s_R \ge 0, && \forall R \in T.
\end{aligned}$$

To obtain a final set of suggestions, we then solve a restricted version of (5), by considering only the set of δ^t variables that we have collected through column generation. We require the δ^t to be binary, optimally selecting between the current columns. This provides a lower bound on the optimal solution to the problem, since some of the columns in the true optimal solution might not

be in the current master problem. This approach allows us to compute a worst-case gap between approximate solutions to (5) and the optimal solution. Computational experiments show that the approach yields significant computational improvements over direct approaches to solving (4), with a limited cost in terms of suboptimality (Appendix A.7).

4.4 Optimization Results

To evaluate the strength of the optimization and predictive models in designing promising clinical trials, we solve the optimization models sequentially with time, as was done in Section 3.3 with the prediction models. We start making and evaluating our suggestions one third of the way through our database of clinical trials, starting in 2002. For all results, we fix the number of trial recommendations made at any point in time to $k = 20$. Throughout this section, we will present results for triplet drug combinations ($N = 3$). There are several reasons for this. First, it has been shown in the literature that combination chemotherapy is superior to single drug treatment (Wagner 2006). This is supported by our database, in which single drug treatments have a mean overall survival of 6.9 months, compared to a mean overall survival of 10.1 months for combination chemotherapy. Additionally, nearly 80% of all chemotherapy trials for advanced gastric and gastroesophageal cancers have tested combined treatments. Since our goal is to recommend future clinical trials, it is thus logical for us to suggest combination regimens. Additionally, there are many more potential triplet chemotherapy combinations than doublet chemotherapy combinations, so our techniques have more to offer in this space. Furthermore, studies have shown a statistically significant benefit in using triplet regimens compared to doublet regimens (Wagner 2006, Hsu et al. 2012).

We note that evaluating the quality of suggestions made by our optimization model is an inherently difficult task. If a trial that our models suggest at one point in time is actually performed in the future, we can of course use the actual outcome of the trial to evaluate our suggestion. However, given the small number of clinical trials that have been conducted relative to the large number of feasible drug and dosage combinations, the likelihood of a proposed trial matching an actual trial performed in the future is small. To address this challenge, we have developed a set of three metrics to use in evaluating our models over time. Each metric evaluates our models from a

different perspective, and each comes with its own advantages and limitations. But by considering all metrics together and comparing our performance on these metrics against the performance of what we call the "average oncologist," we provide evidence that our methodology indeed has merit. We describe and present results for each of these metrics below.

The first metric we define is the Matching Metric, which compares a trial proposal against all trials that were actually run after the date it was suggested. If the drugs proposed in the trial are the same as the set of drugs in a trial that was actually performed, we consider it a match. Note that we do not require the dosages to be identical to consider the trial a match. If a proposed trial matches one or more future trials, we score the suggestion for survival by taking the average survival over the set of future trials that it matches. For toxicity, we score the suggestion by the fraction of matching trials with low toxicity (DLT score below chosen threshold). If a proposed trial does not match any future trials, it does not get a score. As we slide sequentially through time, we calculate a score for every trial proposal we make (or no score if there are no future matches) and record the result. To obtain an overall score for survival and toxicity over the entire evaluation period (2002 – 2012), we average the scores over all proposals that matched at least one future trial.

We compare our model's survival and toxicity scores for the Matching Metric to the baseline performance of an "average oncologist," defined as follows. At each point in time, we take the set of all drug combinations that were actually run in the future, and which could have been suggested by our optimization model at the time.[2] Then, we score each of these combinations using the same procedure as above (i.e. for survival, average the actual survival of all future trials using that combination, and for toxicity, record the fraction of future trials that use that combination with low toxicity), and add them to the evaluation set. To obtain an overall survival and toxicity score for the "average oncologist," we then average the scores over all trials in the evaluation set. The interpretation of this score is that if our "average oncologist" were to randomly select drug combinations to test out of those which have been actually run in the future, this would be his or her expected score for the Matching Metric. We present results for the Matching Metric in Figure 4-1.

[2]For a trial testing N drugs to be a candidate in the optimization model, all N drugs must have been seen at least once prior to the evaluation date, and the drug combination must not have been seen previously.

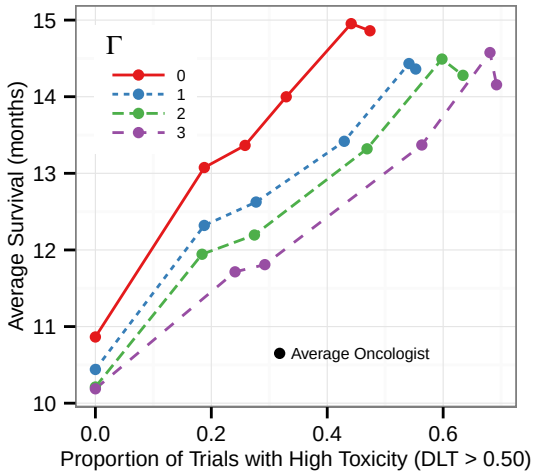


Figure 4-1: Average scores for Matching Metric for optimization suggestions made from 2002–2012. Each line corresponds to a different value of the robustness parameter Γ, and the points correspond to different right-hand-side values t for the toxicity constraint in the set $\{0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$.

There are two parameters that can be adjusted in the optimization model to affect the nature of the combinations we suggest: the threshold t for the right hand side of the toxicity constraint, and the robustness parameter Γ. For values of Γ in {0,1,2,3}, the performance of the average oncologist is dominated by that of the optimization model. In particular, with $(\Gamma = 0, t = 0.5)$, the matching trials suggested by the optimization model have average survival that is 3.3 months greater than the average oncologist, with comparable toxicity. In addition, with $(\Gamma = 0, t = 0.2)$, the matching trials suggested by the optimization model have slightly greater survival than the average oncologist, and no toxicity. These findings are powerful evidence that our methods are indeed selecting high performing clinical trials before they are being run in practice.

Figure 4-1 shows that the best results for the Matching Metric are achieved at $\Gamma = 0$. We note, however, that strong performance is still observed with nonzero values of Γ, and there are several reasons why a more conservative decision maker might decide to select a nonzero Γ. First, the fraction of trials suggested by the optimization that match future trials increases with increasing Γ, as shown in Table 4.1. A higher match rate might provide a conservative decision maker with greater confidence that the performance in the future will be as strong as that indicated by the Matching Metric. Another motivation for the selection of a positive Γ would be to increase the diversity among the set of proposed trials. For example, if a decision maker has a set of 20 trials to

plan over the next two years, he or she may prefer to diversify the trials as much as possible in order to minimize risk. To quantify the amount of diversity in a set of trials, we can define the distance between any pair $(\mathbf{z}_1, \mathbf{z}_2)$ of 3-drug combinations by $d(\mathbf{z}_1, \mathbf{z}_2) = 2 - (\#$ of drugs in common). Then for a set of k suggestions, we define the diversity score as the average distance between all pairs. Average values for the diversity score calculated over the entire evaluation period (2002–2012) are given in Table 4.1, which shows that diversity increases substantially with increasing values of Γ.

$\boldsymbol{\Gamma}$	**Number of Matches / Number of Suggestions**	**Average Diversity Score**
0	493 / 6440 (7.7%)	.611
1	598 / 6440 (9.3%)	.754
2	673 / 6440 (10.5%)	.876
3	712 / 6440 (11.1%)	.981

Table 4.1: Match rate and average trial diversity as a function of the robustness parameter Γ for a fixed toxicity right-hand-side $t = 0.4$.

We have thus far considered average values for the Matching Metric taken over the entire evaluation period. Additional insight can be obtained by evaluating how the quality of proposed trials changes over time. To evaluate performance at a fixed time step, we take the set of $k = 20$ suggestions returned by the optimization model, score them according to the Matching Metric, and average them to get a score for that time. Similarly, we compute an "average oncologist" score by averaging the scores for all combinations that could have been recommended by the optimization at that point. To obtain an upper bound on performance, we calculate the mean of the best m trials that could have been recommended by the optimization at that point (in terms of overall survival), where m is the number of optimization suggestions that actually matched. In Figure 4-2, we present the results for the Matching Metric for survival at $(\Gamma = 0, t = 0.4)$, and note that similar trends are observed at other values of Γ and t.

Prior to 2004, the trials suggested by the optimization are not performing as well as those of the "average oncologist." They begin improving in 2004, and by 2006 the suggestions made by optimization are strongly outperforming those made by the average oncologist. This improvement in performance is not surprising given the improvements we observed in our predictive models over this same time period, as shown in Section 3.3. We note that the average oncologist score does show some improvement in outcomes over this period, but the improvement is not as rapid as that

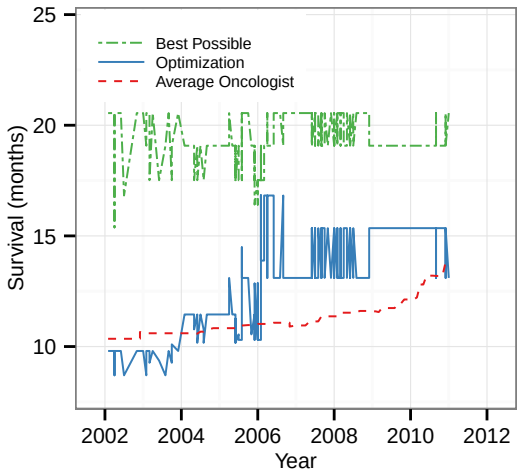


Figure 4-2: Matching Metric evaluated on optimization suggestions made from 2002–2012 ($\Gamma = 0, t = 0.4$). Average actual survival of our optimization suggestions is compared to the best possible score and an average oncologist baseline.

shown by the optimization model. It is also important to point out that the apparent drop-off in performance at the end of the evaluation window is an artifact that can be attributed to the small number of "future trials" in our database that can be matched against after this point in time.

The strength of the Matching Metric is that it directly evaluates the output of our optimization model using real future trial data, when such an evaluation is possible. Unfortunately, it has three limitations: (1) it does not capture the regret of not suggesting trials which are actually run in the future and turn out to be promising, (2) it cannot evaluate the quality of suggestions which have not been run in the future, which as discussed above can be a significant fraction of our suggestions, and (3) it does not take the dosages of our suggested combinations into account. We will address each of these limitations by defining two additional performance metrics, beginning with the Ranking Metric. The motivation behind the Ranking Metric is to assess how well our models can identify the top performing trials out of all those which have actually been run in the future. To calculate the metric, we begin by taking the set of all clinical trials that were actually run after a fixed evaluation date, and which could have been suggested by our optimization model on that date. Then, we use our predictive models built with the data prior to the evaluation date to rank the treatments in order of their expected outcomes. Finally, we calculate our score for the Ranking Metric by taking the top quartile of the treatments on our ranked list, and averaging

their true outcomes. Our performance on the Ranking Metric can again be compared against two baseline scores: the "best possible" score, obtained by ranking the treatments in order of actual outcomes and then computing the average of the top quartile, and the "average oncologist" score, calculated as a mean over all treatments on the list. The interpretation of this score is that if our "average oncologist" were to randomly order the list of all trials that are actually seen in the future, this would be his or her expected score. The Ranking Metric is shown on the left in Figure 4-3. Again the drop-off in performance of the Ranking Metric at the end is attributed to the small number of future trials that can be evaluated at this time.

Neither the Matching nor the Ranking metrics can evaluate the quality of our suggestions that have not been run in the future. This is undoubtedly the most difficult assessment to make, because we cannot conduct actual clinical trials using all these suggestions. As a result, we turn to the only measure of performance that we have available: how well do these suggested trials perform, when evaluated using the final March 2012 regression model trained on the full data set. We call this metric the Final Model Metric. We feel this metric is important as it is the only one capable of evaluating trial suggestions that have not yet been run in the future. To calculate the performance for this metric we take the set of k suggestions made by our optimization model, use the final regression model to estimate their true outcomes, and average the results. There are three baseline scores to compare against for this metric: (1) the "best possible" score, calculated by solving optimization problem (4) using only the drugs and dosages that were known to the model at the evaluation date, but by taking the model coefficients from the March 2012 regression model and using $\Gamma = 0$, (2) the "random trial" score, calculated by taking all the feasible drug combinations that could have been suggested, evaluating them using the final model, and averaging the results, and (3) the "average oncologist" score, calculated by taking all drug combinations that could have been suggested *and were actually run in the future*, evaluating them using the final model, and averaging the results. The Final Model Metric is shown on the right in Figure 4-3. For all metrics, the performance of the optimization model starts out weak, but improves rapidly over the course of the 10-year evaluation period.

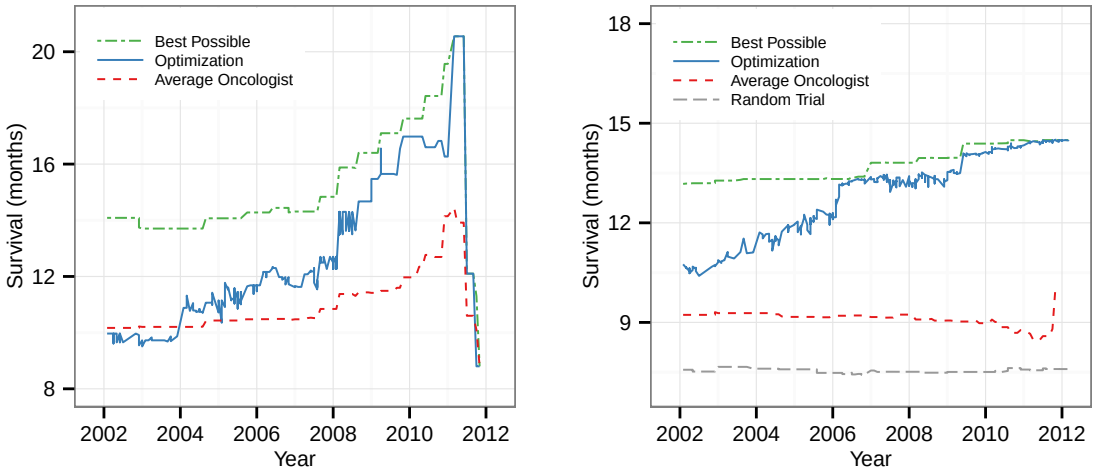


Figure 4-3: Ranking Metric (left) and Final Model Metric (right) evaluations of optimization suggestions made from 2002–2012 ($\Gamma = 0, t = 0.4$).

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 5

Additional Modeling Applications

In this section, we describe two additional applications of our modeling approach that we feel positively contribute to the design of clinical trials. These are just two of many possible ways our models could be used to assist clinical trial decision makers.

5.1 Identifying Clinical Trials that are Unlikely to Succeed

A natural application of the statistical models developed in Section 3 involves determining whether a proposed clinical trial is likely to meet its clinical goals. This is a challenging problem in general, because of the difficulty of predicting clinical trial outcomes, making our models useful for decision makers faced with deciding whether to fund a proposed clinical trial. Avoiding trials that are unlikely to succeed could be beneficial not only to clinical decision makers, who stand to save significant resources, but also to patients.

To determine if our models can be used to identify trials that are unpromising, we performed an out-of-sample experiment in which we predicted the median overall survival of each trial before it was run, based on all previous trial outcomes. Using this prediction along with the computed standard error σ_i^2 described in Section 3.4, we calculated the probability that the current trial's median overall survival exceeds the 5-year rolling average median overall survival. In our experiment, the decision maker uses these results to avoid the trials with the lowest probability of achieving the rolling average overall survival.

In Table 5.1, we see that the proposed model is effective at identifying trials that are unlikely

to outperform recent trials. The 10 trials flagged as least likely to succeed all underperformed the rolling average, and of the 30 least likely to succeed, 22 did not achieve the mean, 6 were above average but not in the fourth quartile for survival, and 2 were in the top quartile of recent trials.

Number Flagged	**Below Average (<50%)**	**Above Average (50%-75%)**	**High (>75%)**
10	10	0	0
20	15	4	1
30	22	6	2
40	30	8	2
50	38	8	4
60	45	11	4

Table 5.1: Out-of-sample accuracy in identifying unpromising trials before they are performed.

Though the decision maker in this experiment is simplistic, ranking trials without regard for their demographics or their toxicity outcomes, our models for predicting clinical trial outcomes can be used in much more sophisticated analyses, as well.

5.2 Determining the Best Chemotherapy Treatments to Date

Identifying the best chemotherapy regimen currently available for advanced gastric cancer is a task that has proven challenging for traditional meta-analysis, but it is one that our methods are well suited to address. Through the use of regression models, which leverage a large database of clinical trial outcomes, we are able to control for differences in demographics and other factors across different clinical trials, enabling direct comparison of results that were not from the same randomized experiment.

To determine the best chemotherapy treatments to date, we first note that selecting a chemotherapy treatment for cancer involves a tradeoff between survival time and toxic effects that affect quality of life. Since individual patients will differ in how they value these competing objectives, the notion of trying to find the a single "best" regimen is not correct. Instead, we seek the set of treatments that make up the "efficient frontier" of chemotherapy treatments for a given cancer: a particular treatment is included in the efficient frontier only if there are no other available treatments with both higher survival and lower toxicity. On the left panel of Figure 5-1, we present the

survival and toxicity of all large trial arms in our database (with the number of patients exceeding the mean of 54.4) for which both outcome variables are available, and highlight those that make up the efficient frontier. A significant concern with this representation is that the demographics of the patient populations differ from one trial to the next, making a direct comparison between them difficult. To control for this effect, we utilize the coefficients from the Ridge Regression models for survival and toxicity trained on the entire data set, which are available at the conclusion of the sequential testing performed in Section 3.3. To give an example, the fraction of patients with prior palliative chemotherapy has a regression coefficient β_i of -0.70 months in the survival model. A trial with 80% prior palliative chemotherapy, instead of the population average of 13% (from Table 2.1), would be expected to have $-0.70 * (0.13 - 0.80) = 0.47$ months lower survival. We correct for this effect by adding 0.47 months to the survival outcome of this trial. After adjusting the survival and toxicity values for all demographic variables in this manner, we present an updated efficient frontier in the right panel of Figure 5-1.

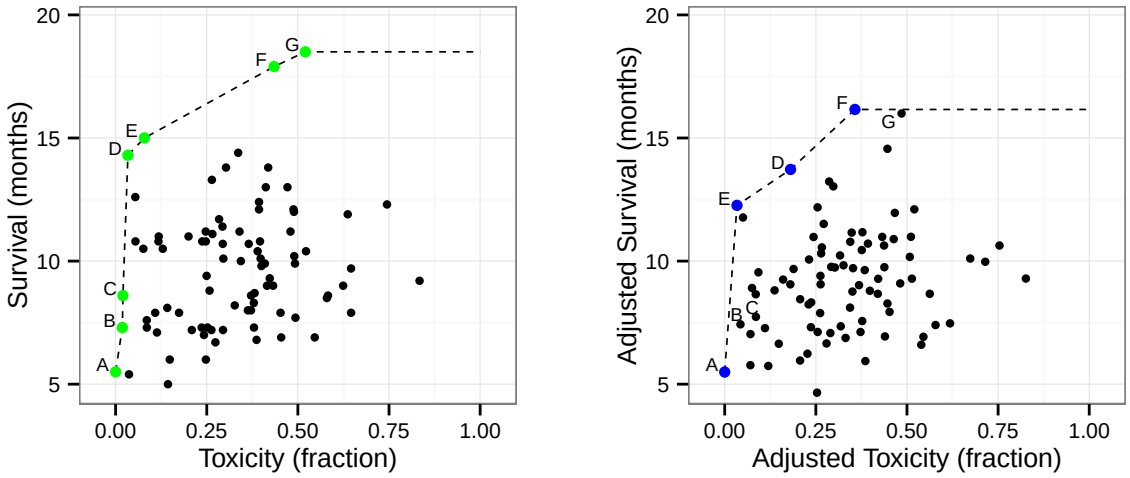


Figure 5-1: Survival and dose-limiting toxicity for clinical trial arms with $\geq$ 55 patients, before (left) and after (right) adjustment for demographic factors. After adjustment, the efficient frontier changes significantly, showing the importance of controlling for demographics when identifying the best available cancer treatments.

In Table 5.2, we report the treatments (as reported in the papers listed) that are in the efficient frontier between overall survival and toxicity with and without adjusting for patient demographics and trial information listed in Table 2.1. We see that the efficient frontier changes considerably when

Trial Arm	Unadjusted			Adjusted		
	Survival (months)	Toxicity (fraction)	EF	Survival (months)	Toxicity (fraction)	EF
(A) Adelstein et al. (2012)	5.5	.000	Yes	5.5	.000	Yes
(B) Kondo et al. (2000), 5'-DFUR	7.3	.018	Yes	7.4	.043	No
(C) Graziano et al. (2003)	8.6	.020	Yes	7.7	.086	No
(D) Wang et al. (2011)	14.3	.034	Yes	13.7	.181	Yes
(E) Iwase et al. (2011)	15.0	.079	Yes	12.3	.034	Yes
(F) Lorenzen et al. (2007)	17.9	.435	Yes	16.2	.357	Yes
(G) Koizumi et al. (2011)	18.5	.520	Yes	16.0	.485	No

Table 5.2: Trials on the efficient frontier (EF) trading off survival and toxicity, before and after adjusting for demographics.

adjustments are made for the trial's demographics — only four of the seven combinations appearing in the adjusted frontier appeared in the unadjusted frontier, and with significantly reduced median overall survival times. This indicates that trials often have better outcomes due to the patient population, and that the outcomes should be adjusted when deciding which treatments are best.

Chapter 6

Concluding Remarks

We believe that our analytics-based approach has the potential to fundamentally change the design process for new chemotherapy clinical trials. This approach can help medical researchers identify the most promising drug combinations for treating different forms of cancer by drawing on previously published clinical trials. This would save researchers' time and effort by identifying proposed clinical trials that are unlikely to succeed and, most importantly, save and improve the quality of patients' lives by improving the quality of available chemotherapy regimens.

The models presented to predict survival and toxicity given demographic information and chemotherapy drugs and dosages represents the first application of data mining techniques to predicting chemotherapy clinical trial outcomes. Our modeling results show that we can predict future clinical trial outcomes using past data, even if the exact combination of drugs being predicted has never been tested in a clinical trial before. The optimization models we proposed will open new frontiers in the design of clinical trials, enabling us to generate new suggestions for clinical trials instead of being limited to providing predictions of the effectiveness of proposed trials.

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix A

Appendices

A.1 Weighted Performance Status

Performance status is a measure of an individual's overall quality of life and well-being. It is reported in our database of clinical trials predominantly using the Eastern Cooperative Oncology Group (ECOG) scale (Oken et al. 1982), and less often using the Karnofsky performance status (KPS) scale (Karnofsky 1949). Table A.1 provides counts of how often the different scales are used in our gastric cancer database.

Scale	**# Trial Arms**
ECOG	414/483 (85.7%)
KPS	68/483 (14.1%)
None	1/483 (0.2%)

Table A.1: Scales used to report performance status in the gastric cancer database.

The ECOG scale runs from 0–5 and is reproduced in Table A.2 for reference. In our database, patients with ECOG score $\geq$3 are rare; nearly 90% of the trial arms have no patients with scores in this range. As a result, we develop a weighted performance status score truncated at 2 as follows: if p_0, p_1 and $p_{\geq 2}$ are the proportions of patients in the trial with ECOG scores of 0, 1, and at least 2, respectively, then our weighted performance status variable is given by $p_1 + 2p_{\geq 2}$.

Of the 414 trial arms that report using ECOG score, only 295 arms report the fraction of patients with each individual score. Of the remaining 119, 15 of these arms report only summary statistics on the ECOG distribution over the patients (min/max/median), and for these arms we

Score	ECOG
0	Fully active, able to carry on all pre-disease performance without restriction.
1	Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature, e.g., light house work, office work.
2	Ambulatory and capable of all selfcare but unable to carry out any work activities. Up and about more than 50% of waking hours.
3	Capable of only limited selfcare, confined to bed or chair more than 50% of waking hours.
4	Completely disabled. Cannot carry on any selfcare. Totally confined to bed or chair.
5	Dead.

Table A.2: Eastern Cooperative Oncology Group (ECOG) performance status scale.

do not compute a score. The remaining 104 arms provide a bucketed ECOG breakdown. For example, 92 arms only report the proportion of patients with either ECOG score 0 or 1, but not the proportions in each category separately. To compute the weighted performance score for these arms, we first obtain a rough estimate the proportion of patients with ECOG score 0 and score 1, based on the proportion of patients in the combined 0-1 bucket. This estimation is done by taking the $n = 292$ trials with full ECOG breakdown and nonzero $p_0 + p_1$, and fitting a linear regression model to estimate $p_0/(p_0 + p_1)$ from the logit of $p_0 + p_1$ (Figure A-1).

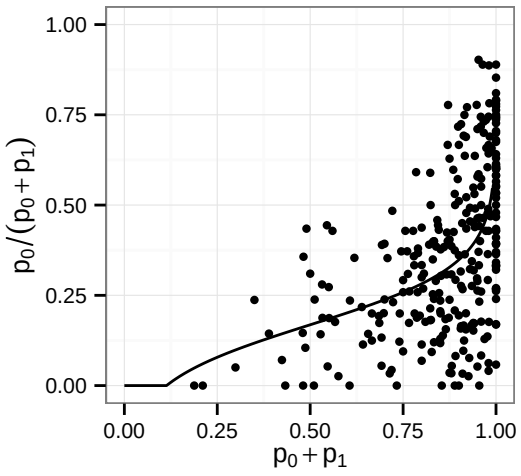


Figure A-1: Data and resulting model used to split the combined ECOG-01 bucket into the proportion of patients with score either 0 or 1 ($n = 292$). Model fit by linear regression to the logit of $p_0 + p_1$.

Using this model to split the combined ECOG-01 bucket, we can compute a weighted perfor-

mance status score for 390/414 of the arms that use the ECOG scale. For the remaining arms (those with only min/max/median reporting or more complex bucketing), the score is marked as unavailable.

Of the 68 trial arms that report using the KPS score, 16/68 provide a full KPS breakdown, 30/68 provide a bucketed breakdown, and 22/68 provide only the min/max/median summary statistics. For the 16/68 arms with full breakdown, we first perform a conversion from the KPS scale to the ECOG scale based on data in (Buccheri et al. 1996), and then calculate the weighted score as before. For trial arms reporting only bucketed KPS or summary statistics, the score is marked as unavailable.

At the end of processing, 406/483 (84.1%) of all trial arms are assigned a weighted performance status score, with the remaining arms marked as missing a score.

A.2 Definition of Dose-Limiting Toxicity

We reviewed the 20 clinical trials meeting our inclusion criteria which also presented a Phase I study (so-called combined Phase I/II trials), to learn:

- Which toxicities are considered dose-limiting toxicities (DLTs), and at which grades?
- What fraction of patients must experience one or more of the DLTs for the trial to be considered "too toxic" (i.e. maximum tolerated dose (MTD) reached)?

18 of 20 trials stated that all Grade 3/4 non-blood toxicities are DLTs, except for some specified toxicities. These excluded toxicities are:

- Alopecia, 18/18 (100%).
- Nausea/vomiting, 12/18 (67%)
- Anorexia, 5/18 (28%)
- Fatigue, 3/18 (17%)
- Diarrhea, 2/18 (11%)
- Abdominal pain, 1/18 (6%)

The remaining two papers defined non-blood toxicities considered DLTs as Grade 3/4 stomatitis and diarrhea. Based on these results, we defined all Grade 3/4 non-blood toxicities except alopecia, nausea, and vomiting to be dose-limiting toxicities.

In this work, we define all Grade 4 blood toxicities as being a DLT. This is in line with the Phase I/II trial results, in which 17/20 trials defined Grade 4 Neutropenia as a DLT, 16/20 defined Grade 4 Thrombocytopenia as a DLT, 7/20 defined Grade 4 Leukopenia as a DLT, and 4/20 defined Grade 4 Anemia as a DLT. Only one trial defined Grade 3 blood toxicities as DLTs, so we chose to exclude this level of blood toxicity from our definition of DLT.

19 of the studies specified the proportion of patients experiencing a DLT that would be defined as the MTD. The studies defined a number of cutoffs:

- 33%: 8/19 (42%)
- 50%: 8/19 (42%)
- 67%: 2/19 (11%)
- 63%: 1/19 (5%)

A.3 Models for Combining Individual Toxicity Levels into an Overall Toxicity Level

Here we compare approaches for combining the proportion of patients experiencing individual dose-limiting toxicities (DLT) into an estimate of the proportion of patients experiencing at least one DLT. We consider five options for combining the toxicities:

- **Max Approach**: Label a trial's toxicity as the proportion of patients with the most frequently occurring DLT. This is a lower bound on the true proportion of patients with a DLT.
- **Independent Approach**: Assume all DLTs in a trial occurred independently of one another, and use this to compute the expected proportion of patients with any DLTs.
- **Sum Approach**: Label a trial's toxicity as the sum of the proportion of patients with each DLT. This is an upper bound on the true proportion of patients with a DLT.

- **Grouped Independent Approach**: Define groups of toxicities, and assign each one a "group score" that is the incidence of the most frequently occurring DLT in that group. Then, compute a toxicity score for the trial by assuming toxicities from each group occur independently, with probability equal to the group score.

- **Grouped Sum Approach**: Using the same groupings, compute a toxicity score for the trial as the sum of the group scores.

For the grouped approaches, we use the 20 broad anatomical/pathophysiological categories defined by the NCI-CTCAE v3 toxicity reporting criteria as the groupings.

To compare these approaches, we evaluate how each of these five approaches do at estimating the proportion of patients with Grade 3/4 toxicities in clinical trials that report this value given the individual Grade 3/4 toxicities. Because there is a strong similarity between the set of Grade 3/4 toxicities and the set of DLTs, we believe this metric is a good approximation of how well the five approaches will approximate the proportion of patients with a DLT. 40/482 (8.3%) of trials report this value, though we can only compute the combined metric for 36/40 (90%) due to missing toxicity data. The quality of each combination approach is obtained by taking the correlation between that approach's results and the combined grade 3/4 toxicities.

Combination Approach	**Correlation**
Grouped Independent	0.893
Independent	0.875
Max	0.867
Grouped Sum	0.843
Sum	0.813

Table A.3: Coefficient of correlation between estimated and actual proportion of patients with a Grade 3/4 toxicity for various toxicity combination approaches ($n = 36$).

As seen in Table A.3, all five combination approaches provide reasonable estimates for the combined toxicity value, though in general grouped metrics outperformed non-grouped metrics. The best approach is the "grouped independent approach," because it allows the best approximation of the combined Grade 3/4 toxicities.

A.4 Multicollinearity

In this section, we analyze the amount of collinearity between our explanatory variables. In standard linear regression, a high degree of correlation among the variables can lead to inflated estimation errors in the model coefficients and poor out-of-sample predictions. It is partially for this reason that we employ regularization techniques (Ridge Regression and Lasso), which can alleviate the impact of collinearity on coefficient errors and improve predictability in this setting (see, for example, the discussion in Hastie et al. (2008), p.61–79). Nevertheless, we acknowledge that the interpretability of the model coefficients can be impaired by severe collinearity, and we include this section as a caution to the reader attempting to interpret any of the model coefficients directly.

We first note that the pairwise correlations between the three dosage variables defined for each drug – binary, average, and instantaneous dose – are all high (in all cases, the correlation coefficient exceeds 0.65). This is primarily due to the fact that these variables have a large number of rows for which all values are zero, corresponding to trial arms in which the drugs are not used. In addition to applying regularization techniques, we have taken two additional steps to limit the impact that this correlation has on our optimization models:

1. Constructing a single uncertainty score for each drug, avoiding any attempt to interpret the individual dosage variable standard errors directly.

2. Restricting the optimization to the set of drug dosages that have been observed previously, so that they have the same general correlation structure as data used to train the model.

While we feel these steps are sufficient to handle the collinearity among the dose variables for the purposes of our optimization models, care must be taken when attempting to interpret the individual model coefficients.

We next consider the pairwise correlation coefficients between the remaining variable pairs. We report any pairs of variables for which the magnitude of the correlation coefficient exceeds 0.5 in Table A.4. We note that correlations between between different drugs are not problematic from the standpoint of our optimization, since we impose constraints on the pairs of drugs that can be selected for our trials. For example, in Table A.4 we see that Leucovorin use is correlated with that of Fluorouracil. Indeed, Leucovorin is in the class of drugs known as chemoprotectants, which

are typically paired with drugs in the antimetaboline class, which includes Fluorouracil. Since this constraint is imposed directly in our optimization, we are able to mitigate the effect of correlation between these variables. In addition, correlations between non-drug variables do not affect the optimization, since we are not optimizing over these variables. We note finally that none of the correlations between a drug and non-drug variable exceeded 0.5.

Category	**Variable 1**	**Variable 2**	**Correlation coefficient**
Variables from different drugs	Leucovorin binary dose	Fluorouracil instantaneous dose	0.583
	PALA binary dose	IFN average dose	0.564
	PALA instantaneous dose	IFN average dose	0.564
	IFN average dose	PALA average dose	0.564
	Fluorouracil binary dose	Leucovorin binary dose	0.563
	Fluorouracil instantaneous dose	Leucovorin instantaneous dose	0.557
Non-drug variables	Trial in Asia	Trial in South Korea	0.569
	Trial in Asia	Trial in Japan	0.512
	Cancer in stomach	Cancer in GEJ	-0.919

Table A.4: Pairs of predictor variables with correlation coefficient exceeding 0.5, excluding pairs corresponding to dosage variables for the same drug.

A.5 Heteroskedasticity

To assess our models for heteroskedasticity of variance, we plot the sequential out-of-sample prediction errors against the parameters of interest to check whether there is a significant trend. In Figure A-2, we depict the absolute value of the sequential prediction errors for the Ridge Regression survival and toxicity models against two parameters: the number of patients in the test trial (on a log-scale to get meaningful separation between the data points), and the predicted value for the test trial.

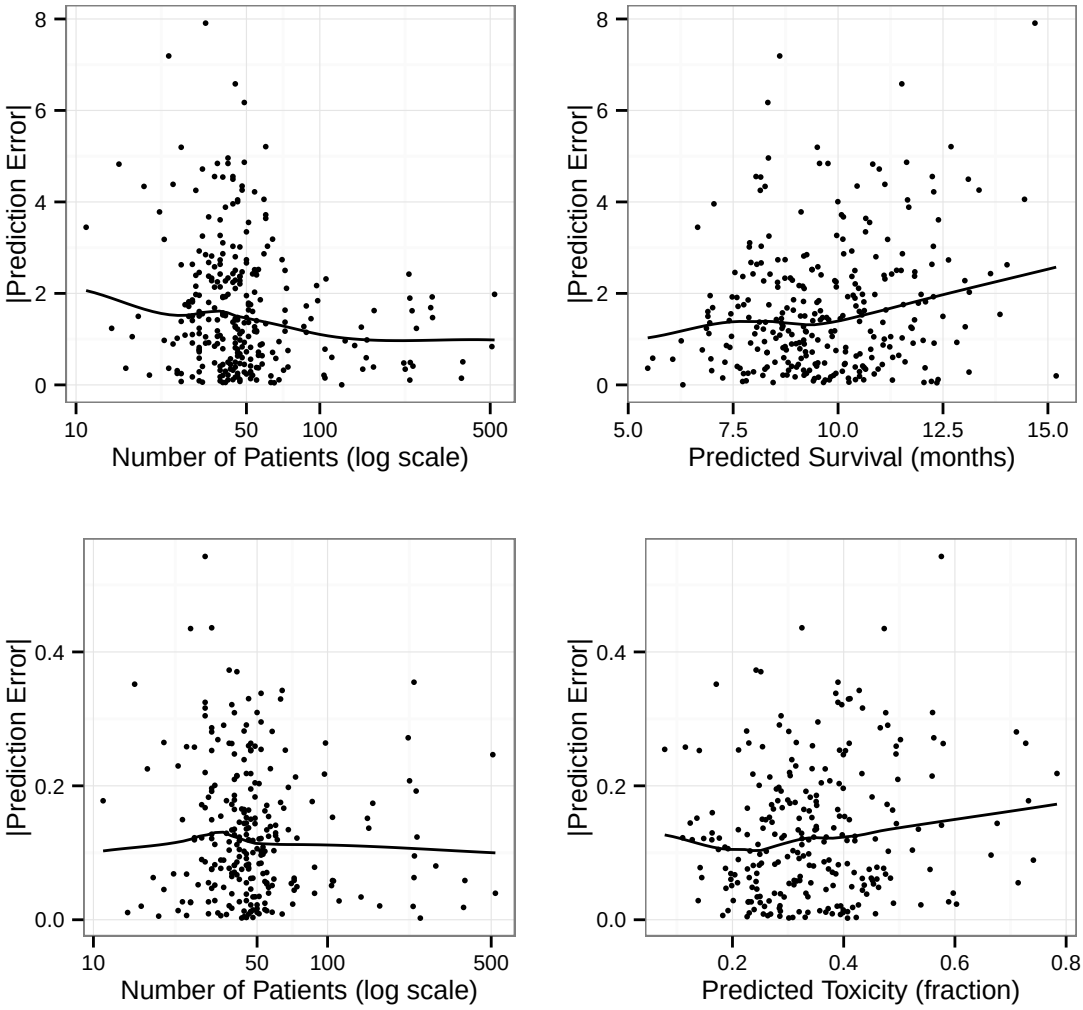


Figure A-2: Heteroskedasticity tests for survival model (top) and toxicity model (bottom). For both models we check the relationship between the sequential out-of-sample error terms and the number of patients in the trial arm (left), and the predicted outcome value (right). On top of all plots we have added a locally linear LOESS smoother (Cleveland et al. 1992) to show the trend.

To quantify magnitude of heteroskedasticity, we calculate how much the errors vary over the range of our data points, by taking the ratio of the maximum and minimum values from the LOESS curves shown in Figure A-2. These values are given in Table A.5.

Model	Number of Patients	Predicted Value
Survival	2.13	2.50
Toxicity	1.31	1.67

Table A.5: Measure of how much the error variance changes as a function of two parameters: the number of patients in a trial arm and the predicted outcome value. Heteroskedasticity factors are calculated as the ratio between the maximum and minimum error values obtained from the LOESS curves shown in Figure A-2.

According to Fox (1991), heteroskedasticity is only worth correcting if the error terms vary by more than a factor of 3 over the data set. We conclude that this is not the case for either our survival or toxicity models, and that heteroskedasticity of variance is mild.

A.6 Interaction Terms

In this section, we test whether the inclusion of interaction terms between our predictor variables improves the out-of-sample accuracy of our linear models for survival and toxicity. We note that Random Forests and SVM with a radial basis function kernel naturally capture interactions between all variables in the model, and there is no need to explicitly define additional interaction variables. As a result, our focus here will be restricted to the regularized linear models: Ridge Regression and the Lasso.

We evaluate two classes of interaction terms separately. First, we consider the interactions between pairs of drugs, which we can account for in our model by adding pairwise products between all of the binary drug variables in our database. Each of the new binary variables represents a different drug pair, and will be set to 1 if and only if that drug pair is used in the treatment arm. There are 72 total drugs in our database, yielding 2,556 possible drug pairs. Of these, only 149 drug pairs appear in one or more treatment arms in our database. To limit the number of uninformative variables added to the model, we impose the additional requirement that a drug pair must appear in at least 3 arms of the database before encoding it with a variable. There are 62 drug pairs meeting this requirement, so we add these 62 binary drug pair variables to our model.

Next, we consider the interactions between drug variables and the non-drug variables (patient demographics, as well as non-drug trial information). We model these interactions by taking prod-

ucts between the binary drug variables and each of the non-drug variables. We again impose the constraint that an interaction variable must take a nonzero value for at 3 arms of the database before it will be included in the model. The result is the addition of 289 variables to the model.

In Table A.6, we compare the sequential model performance with and without each set of interaction variables. We see that the drug/drug interaction terms provide a very slight improvement in the survival predictions, but the improvement to the R^2 is less than 0.01 for both the Lasso and Ridge models. The toxicity predictions are slightly degraded by the inclusion of these variables terms, but again the effect is small. The inclusion of interactions between the drug and non-drug variables decreases the predictive accuracy of all survival and toxicity models.

	No Interactions		**Drug / Drug**		**Drug / Non-drug**	
Models	Survival R^2	Toxicity AUC	Survival R^2	Toxicity AUC	Survival R^2	Toxicity AUC
Ridge	.589	.828	.596	.822	.582	.826
Lasso	.600	.836	.605	.824	.584	.819

Table A.6: Performance of the Ridge and Lasso models for survival and toxicity over the most recent 4 years of data (March 2008–March 2012), with and without the inclusion of interaction terms. Survival performance is presented in terms of the coefficient of determination (R^2) of our prediction models relative to the baseline, and toxicity performance is reported as the area under the curve (AUC) for predicting whether a trial will have high toxicity (score > 0.5).

A.7 Computational Experiments on the Column Generation Approach

To evaluate the efficiency and solution quality of the column generation approach (6), we solved the model at the beginning of 2010 for tuple size $N = 3$ and a range of suggestion counts k and robustness parameters Γ. Results were compared with a mixed integer programming formulation of (4). Experiments were run on an Intel Core i7-860 (2.8 GHz) with 16 GB RAM, and a computational limit of 30 minutes was applied for each model.

Results are given in Table A.7. Internally calculated optimality gaps (labeled "procedure gaps" in the table) for the column generation approach were uniformly small, and runtimes dominated those of the direct MIP formulation of (4) for moderate and large values of k. When possible to verify, the column generation approach returned optimal solutions to the problem.

		Column Generation			Direct MIP		
k	Γ	Runtime	Procedure Gap (%)	Opt. Gap (%)	Runtime	Procedure Gap (%)	Opt. Gap (%)
5	0	0.51	0	0	14.26	0	0
5	1	0.49	0	0	18.75	0	0
5	2	0.49	0.067	0	15.37	0	0
5	3	0.60	0.269	0	17.82	0	0
10	0	1.25	0	0	> 1800	3.305	0
10	1	1.33	0	0	> 1800	4.587	0
10	2	1.39	0	0	> 1800	5.171	0
10	3	1.78	0.091	n/a	> 1800	4.872	n/a
20	0	3.00	0	0	> 1800	14.046	1.542
20	1	2.84	0	0	> 1800	12.688	0.825
20	2	3.19	0	0	> 1800	12.186	0.400
20	3	3.72	0	0	> 1800	12.772	0.832

Table A.7: Computational results for the column generation and direct MIP approaches for tuple size $N = 3$, at a range of suggestion counts k and robustness parameters Γ. Runtimes are in seconds. Procedure gaps are calculated internally for each algorithm between the best solution and internal upper bound. Optimality gaps are between the algorithm solution and true optimal solution, when possible to verify (otherwise labeled n/a).

THIS PAGE INTENTIONALLY LEFT BLANK

Bibliography

Adelstein, David J, Cristina P Rodriguez, Lisa A Rybicki, Denise I Ives, Thomas W Rice. 2012. A phase ii trial of gefitinib for recurrent or metastatic cancer of the esophagus or gastroesophageal junction. *Investigational new drugs* **30**(4) 1684–1689.

Ajani, Jaffer, Wuilbert Rodriguez, Gyorgy Bodoky, Vladimir Moiseyenko, Mikhail Lichinitser, Vera Gorbunova, Ihor Vynnychenko, August Garin, Istvan Lang, Silvia Falcon. 2010. Multicenter phase III comparison of cisplatin/s-1 with cisplatin/infusional fluorouracil in advanced gastric or gastroesophageal adenocarcinoma study: The FLAGS trial. *Journal of Clinical Oncology* **28**(9) 1547–1553.

Bang, Yung-Jue, Eric Van Cutsem, Andrea Feyereislova, Hyun Chung, Lin Shen, Akira Sawaki, Florian Lordick, Atsushi Ohtsu, Yasushi Omuro, Taroh Satoh, Giuseppe Aprile, Evgeny Kulikov, Julie Hill, Michaela Lehle, Josef Rüschoff, Yoon-Koo Kang. 2010. Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of her2-positive advanced gastric or gastro-oesophageal junction cancer (ToGA): a phase 3, open-label, randomised controlled trial. *Lancet* **376** 687–697.

Bertsimas, Dimitris, Melvyn Sim. 2004. The price of robustness. *Operations Research* **52**(1) 35–53.

Bertsimas, Dimitris, John Tsitsiklis. 1997. *Introduction to Linear Optimization*. 1st ed. Athena Scientific.

Bishop, Christopher M. 2006. *Pattern recognition and machine learning*, vol. 1. springer New York.

Breiman, Leo. 2001. Random forests. *Machine Learning* **45** 5–32.

Breiman, Leo, J. H. Friedman, R. A. Olshen, C. J. Stone. 1984. *Classification and Regression Trees*. Statistics/Probability Series, Wadsworth Publishing Company, Belmont, California, U.S.A.

Buccheri, G., D. Ferrigno, M. Tamburini. 1996. Karnofsky and ecog performance status scoring in lung cancer: A prospective, longitudinal study of 536 patients from a single institution. *European Journal of Cancer* **32**(7) 1135 – 1141.

Burke, H.B. 1997. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer* **79**(4) 857–862.

Caruana, Rich, Alexandru Niculescu-Mizil. 2006. An empirical comparison of supervised learning algorithms.

Proceedings of the 23rd international conference on Machine learning. ICML '06, ACM, New York, NY, USA, 161–168.

Chabner, Bruce, Thomas Roberts. 2005. Chemotherapy and the war on cancer. *Nature Reviews Cancer* **5** 65–72.

Cleveland, William S, Eric Grosse, William M Shyu. 1992. Local regression models. *Statistical models in S* 309–376.

ClinicalTrials.gov. 2007. Understanding Clinical Trials. ClinicalTrials.gov.

Cristianini, Nello, John Shawe-Taylor. 2000. *An introduction to support Vector Machines: and other kernel-based learning methods*. Cambridge University Press, New York, NY, USA.

Delen, D. 2005. Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial Intelligence in Medicine* **34**(2) 113–127.

Earle, C.C., J.A. Maroun. 1999. Adjuvant Chemotherapy after Curative Resection for Gastric Cancer in Non-Asian Patients: Revisiting a Meta-analysis of Randomised Trials. *European Journal of Cancer* **35**(7) 1059–1064.

Efferth, Thomas, Manfred Volm. 2005. Pharmacogenetics for individualized cancer chemotherapy. *Pharmacology and Therapeutics* **107** 155–176.

Emanuel, Ezekiel, Lowell Schnipper, Deborah Kamin, Jenifer Levinson, Allen Lichter. 2003. The costs of conducting clinical research. *Journal of Clinical Oncology* **21**(22) 4145–4150.

Fox, John. 1991. *Regression diagnostics: An introduction*, vol. 79. SAGE Publications, Incorporated.

Friedman, Jerome, Trevor Hastie, Robert Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**(1) 1–22. URL `http://www.jstatsoft.org/v33/i01/`.

Golan, David E., Armen H. Tashjian Jr., Ehrin J. Armstrong, April W. Armstring, eds. 2008. *Principles of Pharmacology: The Pathophysiologic Basis of Drug Therapy*. 2nd ed. Lippincott Williams and Wilkins.

Graziano, F, D Santini, E Testa, V Catalano, GD Beretta, S Mosconi, G Tonini, V Lai, R Labianca, S Cascinu. 2003. A phase ii study of weekly cisplatin, 6s-stereoisomer leucovorin and fluorouracil as first-line chemotherapy for elderly patients with advanced gastric cancer. *British journal of cancer* **89**(8) 1428–1432.

Hastie, Trevor, Robert Tibshirani, Jerome Friedman. 2008. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd edition)*. Springer-Verlag.

Hermans, J. 1993. Adjuvant Therapy After Curative Resection for Gastric Cancer: Meta-Analysis of Randomized Trials. *Journal of Clinical Oncology* **11**(8) 1441–1447.

Hoerl, Arthur E., Robert W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**(1) 55–67.

Hsu, Chih-Wei, Chih-Chung Chang, Chih-Jen Lin. 2003. A practical guide to support vector classification.

Hsu, Chiun, Ying-Chun Shen, Chia-Chi Cheng, Ann-Lii Cheng, Fu-Chang Hu, Kun-Huei Yeh. 2012. Geographic difference in safety and efficacy of systemic chemotherapy for advanced gastric or gastroesophageal carcinoma: a meta-analysis and meta-regression. *Gastric Cancer* **15** 265–280.

Iwase, H., M. Shimada, T. Tsuzuki, K. Ina, M. Sugihara, J. Haruta, M. Shinoda, T. Kumada, H. Goto. 2011. A Phase II Multi-Center Study of Triple Therapy with Paclitaxel, S-1 and Cisplatin in Patients with Advanced Gastric Cancer. *Oncology* **80** 76–83.

Jefferson, Miles, Neil Pendleton, Sam Lucas, Michael Horan. 1997. Comparison of a genetic algorithm neural network with logistic regression for predicting outcome after surgery for patients with nonsmall cell lung carcinoma. *Cancer* **79**(7) 1338–1342.

Jemal, Ahmedin, Freddie Bray, Melissa Center, Jacques Ferlay, Elizabeth Ward, David Forman. 2011. Global cancer statistics. *CA: A Cancer Journal for Clinician* **61** 69–90.

Kang, Y.-K., W.-K. Kang, D.-B. Shin, J. Chen, J. Xiong, J. Wang, M. Lichinitser, Z. Guan, R. Khasanov, L. Zheng, M. Philco-Salas, T. Suarez, J. Santamaria, G. Forster, P.-I. McCloud. 2009. Capecitabine/cisplatin versus 5-fluorouracil/cisplatin as first-line therapy in patients with advanced gastric cancer: a randomised phase iii noninferiority trial. *Annals of Oncology* **20** 666–673.

Karnofsky, David A. 1949. The clinical evaluation of chemotherapeutic agents in cancer. *Evaluation of chemotherapeutic agents* .

Koizumi, Wasaburo, Norisuke Nakayama, Satoshi Tanabe, Tohru Sasaki, Katsuhiko Higuchi, Ken Nishimura, Seiichi Takagi, Mizutomo Azuma, Takako Ae, Kenji Ishido, Kento Nakatani, Akira Naruke, Chikatoshi Katada. 2011. A multicenter phase ii study of combined chemotherapy with docetaxel, cisplatin, and s-1 in patients with unresectable or recurrent gastric cancer (kdog 0601). *Cancer Chemotherapy and Pharmacology* **69** 407–13.

Koizumi, Wasaburo, Hiroyuki Narahara, Takuo Hara, Akinori Takagane, Toshikazu Akiya, Masakazu Takagi, Kosei Miyashita, Takashi Nishizaki, Osamu Kobayashi, Wataru Takiyama, Yasushi Toh, Takashi Nagaie, Seiichi Takagi, Yoshitakaa Yamamura, Kimihiko Yanaoka, Hiroyuki Orita, Masahiro Takeuchi. 2008. S-1 plus cisplatin versus s-1 alone for fi rst-line treatment of advanced gastric cancer (SPIRITS trial): a phase III trial. *Lancet* **9** 215–221.

Kondo, K, J Sakamoto, H Nakazato, A Koike, T Kitoh, K Hachisuka, S Saji, J Yura, Y Nimura, N Hama-

jima. 2000. A phase iii randomized study comparing doxifluridine and 5-fluorouracil as supportive chemotherapy in advanced and recurrent gastric cancer. *Oncology reports* **7**(3) 485–490.

Liaw, Andy, Matthew Wiener. 2002. Classification and regression by randomforest. *R News* **2**(3) 18–22. URL `http://CRAN.R-project.org/doc/Rnews/`.

Lorenzen, S., M. Hentrich, C. Haberl, V. Heinemann, T. Schuster, T. Seroneit, N. Roethling, C. Peschel, F. Lordick. 2007. Split-dose docetaxel, cisplatin and leucovorin/fluorouracil as first-line therapy in advanced gastric cancer and adenocarcinoma of the gastroesophageal junction: results of a phase ii trial. *Annals of Oncology* **18** 1673–1679.

Lutz, Manfred P., Hansjochen Wilke, D.J. Theo Wagener, Udo Vanhoefer, Krzysztof Jeziorski, Susanna Hegewisch-Becker, Leopold Balleisen, Eric Joossens, Rob L. Jansen, Muriel Debois, Ullrich Bethe, Michel Praet, Jacques Wils, Eric Van Cutsem. 2007. Weekly infusional high-dose fluorouracil (hd-fu), hd-fu plus folinic acid (hd-fu/fa), or hd-fu/fa plus biweekly cisplatin in advanced gastric cancer: Randomized phase ii trial 40953 of the european organisation for research and treatment of cancer gastrointestinal group and the arbeitsgemeinschaft internistische onkologie. *Journal of Clinical Oncology* **25**(18) 2580–2585.

Mari, E. 2000. Efficacy of adjuvant chemotherapy after curative resection for gastric cancer: A meta-analysis of published randomised trials. *Annals of Oncology* **11** 837–843.

Meyer, David, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, Friedrich Leisch. 2012. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*. URL `http://CRAN.R-project.org/package=e1071`. R package version 1.6-1.

National Cancer Institute. 2006. Common terminology criteria for adverse events v3.0 (ctcae). URL `"http://ctep.cancer.gov/protocolDevelopment/electronic_applications/docs/ctcaev3.pdf"`.

NCCN. 2013. *NCCN Clinical Practice Guidelines in Oncology: Gastric Cancer*. National Comprehensive Cancer Network, 1st ed.

Ohno-Machado, Lucila. 2001. Modeling medical prognosis: Survival analysis techniques. *Journal of Biomedical Informatics* **34** 428–439.

Oken, Martin M, Richard H Creech, Douglass C Tormey, John Horton, Thomas E Davis, Eleanor T McFadden, Paul P Carbone. 1982. Toxicity and response criteria of the eastern cooperative oncology group. *American journal of clinical oncology* **5**(6) 649–656.

Page, Ray, Chris Takimoto. 2002. *Cancer Management: A Multidisciplinary Approach: Medical, Surgical and Radiation Oncology*, chap. Chapter 3: Principles of Chemotherapy. PRR Inc.

Phan, John, Richard Moffitt, Todd Stokes, Jian Liu, Andrew Young, Shuming Nie, May Wang. 2009. Con-

vergence of biomarkers, bioinformatics and nanotechnology for individualized cancer treatment. *Trends in Biotechnology* **27**(6) 350–358.

Pratt, William B. 1994. *The Anticancer Drugs*. Oxford University Press.

R Core Team. 2012. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

Thompson, Simon, Julian Higgins. 2002. How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine* **21** 1559–1573.

Thuss-Patience, Peter C., Albrecht Kretzschmar, Michael Repp, Dorothea Kingreen, Dirk Hennesser, Simone Micheel, Daniel Pink, Christian Scholz, Bernd Dörken, Peter Reichardt. 2005. Docetaxel and continuous-infusion fluorouracil versus epirubicin, cisplatin, and fluorouracil for advanced gastric adenocarcinoma: A randomized phase ii study. *Journal of Clinical Oncology* **23**(3) 494–501.

Tibshirani, Robert J. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**(1) 267–288.

Vapnik, Vladimir N. 1998. *Statistical learning theory*. Wiley.

Wagner, A. 2006. Chemotherapy in Advanced Gastric Cancer: A Systematic Review and Meta-Analysis Based on Aggregate Data. *Journal of Clinical Oncology* **24**(18) 2903–2909.

Wald, Abraham. 1945. Statistical decision functions which minimize the maximum risk. *The Annals of Mathematics* **46** 265–280.

Wang, Fenghua, Zhiqiang Wang, Ningning Zhou, Xin An, Ruihua Xu, Youjian He, Yuhong Li. 2011. Phase ii study of biweekly paclitaxel plus infusional 5-fluorouracil and leucovorin as first-line chemotherapy in patients with advanced gastric cancer. *American Journal of Clinical Oncology* **34** 401–405.

Wong, R., D. Cunningham. 2009. Optimising treatment regimens for the management of advanced gastric cancer. *Annals of Oncology* **20** 605–608.

World Health Organization. 2012. Fact Sheets: Cancer. World Health Organization.