An Analytics Approach to Designing Clinical Trials for Cancer

Dimitris Bertsimas
Sloan School and Operations Research Center, Massachusetts Institute of Technology,
dbertsim@mit.edu

Allison O'Hair, Stephen Relyea, and John Silberholz
Operations Research Center, Massachusetts Institute of Technology,
akohair@mit.edu, srelyea@mit.edu, josilber@mit.edu

Dedicated to the memory of John Bertsimas, 1934-2009

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 for overall 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 in current practice 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.

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, often leaving drug therapy or best supportive care 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. 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.

Our aspiration in this paper 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 paper 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 the paper is that the survival outcomes of clinical trials can to a large extent be predicted ($R^2 = 0.60$ for out-of-sample predictions) in advance, as long as the drugs used have been seen before in different combinations.

(3) We propose an optimization-based methodology of suggesting novel treatment regimens that balance survival and toxicity and take into account the uncertainty in our predictions. We design a number of quality measures to evaluate the suggestions made by our approach, and we demonstrate that the proposed approach outperforms the current clinical practice for selecting new combination chemotherapy treatments.

(4) 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.

Overall, we feel that the paper provides evidence that the combination of data, statistical models and optimization can open new frontiers in the design of clinical trials. While the results presented here are for a specific form of cancer (gastric and gastroesophageal), the methodology is widely applicable to other forms of cancer.

Medical practitioners and researchers in the fields of data mining and machine learning have a rich history of predicting clinical outcomes. 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 et al. 2005, Lee et al. 2003, Hurria et al. 2011, 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). Individualized predictions represent a useful tool to patients choosing between multiple treatment options (Zhao et al. 2012, van't Veer and Bernards 2008), and when trained on clinical trial outcomes for a particular treatment can be used to identify promising patient populations to test that treatment on (Zhao et al. 2011) or to identify if that treatment is promising for a Phase III clinical trial (De Ridder 2005). However, such models do not enable predictions of outcomes for patients treated with previously unseen chemotherapy regimens, limiting their usefulness in planning clinical trials with new drug combinations.

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 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.

Because approaches driven by patient outcomes such as medical prognosis and meta-regression have limited usefulness in planning clinical trials with new drug combinations, other methodologies are more commonly applied in the treatment design process. Studies of drug combinations in animals and *in vitro* are often cited as motivations for combination chemotherapy trials (Chao et al. 2006, Iwase et al. 2011, Lee et al. 2009), and molecular simulation is a well developed methodology for identifying synergism in drug combinations (Chou 2006). Though such approaches are attractive because they can evaluate any proposed treatment regimen, they are limited because they don't incorporate treatment outcomes from actual patients. Further, identifying the best possible combination chemotherapy regimen with either methodology involves enumerating all possible drug combinations, which is resource intensive.

To the best of our knowledge, this paper presents the first prediction model to date that is trained on clinical trial outcomes and contains enough detail about the drug combination and dosages to enable the design of novel combination chemotherapy regimens. We attain this by significantly broadening the scope of meta-regression, training our model not only on randomized clinical trials but also on non-randomized 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 this paper, 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).

2 Data Collection

In this section, we describe the inclusion/exclusion rules we used and the data we collected to build our database. Definitions of some of the common clinical trial terminologies we use are given in Table 1.

Term	**Definition**
Arm	A group or subgroup of patients in a trial that receives a specific treatment.
Controlled trial	A type of trial in which the treatment given is compared to a standard treatment.
Cycle	An interval of time during which the treatment of the trial is consistent.
Exclusion criteria	The factors that prevent a person from participating in a clinical trial.
Inclusion criteria	The factors that allow a person to participate in a clinical study.
Phase	The classification of a clinical trial. Phase I studies identify safe dosages, while Phase II and III studies determine if a treatment is effective.
Randomized trial	A type of trial in which patients are randomly assigned to one of several arms.
Sequential treatment	A treatment regimen in which patients transition from one treatment to another after a pre-specified number of treatment cycles.

Table 1: Definitions of some common clinical trial terminology.

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 advanced or 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. 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) trials testing sequential treatments, (2) trials that involve the application of radiation therapy,[1] (3) trials that apply curative or adjuvant chemotherapy, and (4) trials to treat

[1] Radiotherapy is not recommended for metastatic gastric cancer patients (NCCN 2013), and through PubMed and Cochrane

gastrointestinal stromal tumors.

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.

2.1 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. 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).

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

searches for stomach neoplasms and radiotherapy, we only found three clinical trials using radiotherapy for metastatic gastric cancer.

[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.

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.

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). However, the variables we do collect enable us to control for sources of endogeneity, in which patient and physician decision rules in selecting treatments might limit the generalizability of model results. For example, we collect performance status, a factor used by physicians in selecting treatments (NCCN 2013). Although other factors, such as comorbidities and patient preferences for toxicities, are important in treatment decisions for the general population (NCCN 2013), clinical trials uniformly exclude patients with severe comorbidities and toxicity preferences do not affect actual survival or toxicity outcomes. The only other treatment decision we do not account for in our models is that patients with HER2-positive cancers should be treated with trastuzumab (NCCN 2013), while this treatment is ineffective in other patients. We address this issue by excluding trastuzumab from the clinical trial suggestions we make in Section 4.

In Table 2, 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.

2.2 An Overall Toxicity Score

As described in Section 2.1, we extracted the proportion of patients in a trial who 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. The motivation for an overall toxicity score is that there are 370 different possible averse events from cancer treatments (National Cancer Institute 2006). Instead of building a model for each of these toxicities, some of which are significantly more severe than others, we use an overall toxicity score.

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

Variable	**Average Value**	**Range**	**% Reported**
Patient Demographics			
Fraction male	0.72	0.29 – 1.00	97.9
Fraction of patients 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 of patients with primary tumor in the stomach	0.90	0.00 – 1.00	94.8
Fraction of patients with primary tumor in the gastroesophageal junction	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 an 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 every Grade 3/4 or Grade 4 toxicity	Toxicity Dependent[3]		

[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.

[3] See Appendix A.2 for details on data preprocessing for blood toxicities.

Table 2: Non-drug variables extracted from gastric and gastroesophageal cancer clinical trials. These variables, together with the drug variables, were inputted into a database.

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 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 clinical 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 that also presented a Phase I study (so-called combined Phase I/II trials), we identified the following set of DLTs to include in the calculation of our composite toxicity score:

- *Any Grade 3 or Grade 4 non-blood toxicity, excluding alopecia, nausea, and vomiting.* 18 of 20 trials stated that all Grade 3/4 non-blood toxicities are DLTs, except some specified toxicities. Alopecia was excluded in all 18 trials and nausea/vomiting were excluded in 12 (67%). The next most frequently excluded toxicity was anorexia, which was excluded in 5 trials (28%).
- *Any Grade 4 blood toxicity.* Of the 20 trials reviewed, 17 (85%) defined Grade 4 neutropenia as a DLT, 16 (80%) defined Grade 4 thrombocytopenia as a DLT, 7 (35%) 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.

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 thresholds of 33% and 50% were the most common, each occurring in 8 of 19 (42%) of studies that published a threshold. Details on the computation of the proportion of patients experiencing a DLT are presented in Appendix A.3.

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).

3.1 Data and Variables

We used the data we extracted from published clinical trials described in Table 2 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 account for this, we include in our model all of the patient demographic and non-drug trial variables listed in Table 2, excluding the number of patients in the trial. The reporting frequencies for each of these variables is given in Table 2, 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[3]: 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 of drug d administered in a single session, and the average dose of a drug d 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 2.2. 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 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 regularized linear regression models. 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

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

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. It is well known that in settings with a relatively small ratio of data samples to predictor variables, regularized models help to reduce the variability in the model parameters. We estimate the regression coefficients β and β_0 by minimizing the following objective:

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

where 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. We choose the value of λ from among a set of 50 candidates through 10-fold cross-validation on the training set.[4]

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 sparsity 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 (SVM). For Random Forests (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 to sample at each split is chosen through 10-fold cross-validation on the training set from among exponentially spaced values centered at $d/3$, where d is the total number of input variables. For SVM, 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 ($C = 2^{-5}, 2^{-3}, \ldots, 2^{15}, \gamma = 2^{-15}, 2^{-13}, \ldots, 2^{3}$).

[4] Candidate values of lambda are exponentially spaced between $\lambda_{max}/10^4$ and λ_{max}. For $p = 1$, we take λ_{max} to be the smallest value for which all coefficients $\boldsymbol{\beta}$ are zero. For $p = 2$, we take λ_{max} to be the largest eigenvalue of $\mathbf{X}^{\mathbf{T}}\mathbf{X}$, where $\mathbf{X}$ is the matrix of input variables in the training set.

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).

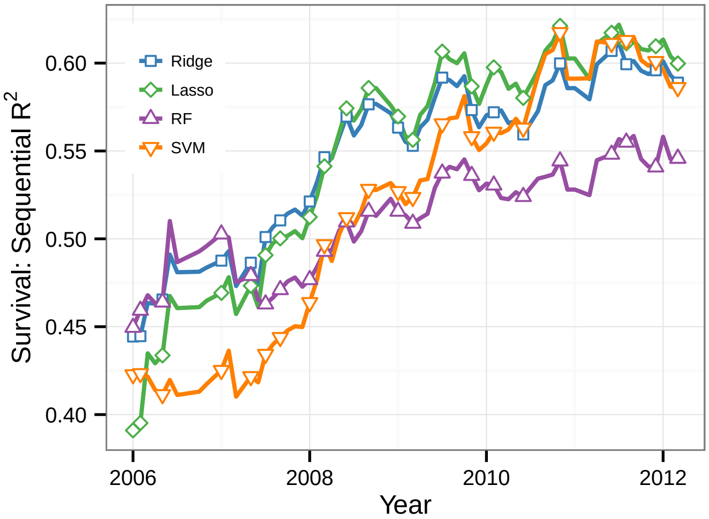
3.3 Results

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 gastric and 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 arms, we first remove those for which the outcome is not available, leaving 315 arms for survival and 278 for toxicity. We then predict outcomes only for those using drugs that have been seen at least once in previous trials (albeit possibly in different combinations and dosages). This provides us with 287 data points to evaluate the survival models, and 253 to evaluate the toxicity models.

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 1, along with the values of the RMSE and R^2 over the most recent 4-year window of sequential testing.

To evaluate the toxicity models, we recall from the discussion of Section 2.2 that the toxicity of a



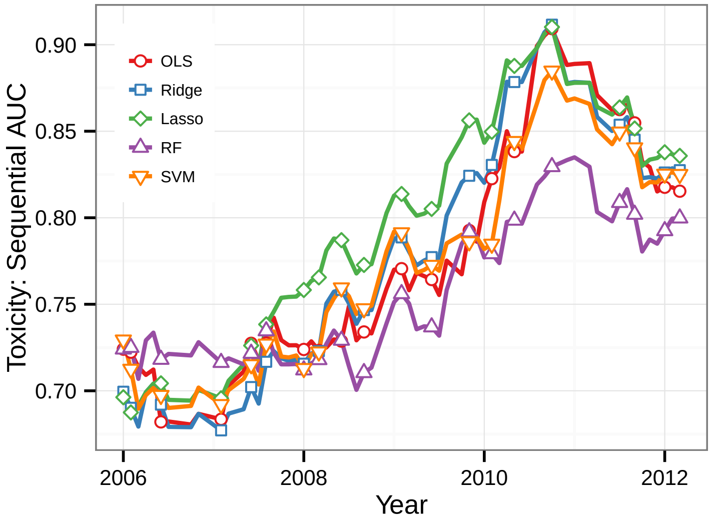
March 2008–March 2012		
Models	RMSE (months)	R^2
Baseline	3.662	0
OLS	4.180	< 0
Ridge	2.348	.589
Lasso	2.317	.600
RF	2.468	.546
SVM	2.356	.586

Figure 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 132 test trial arms.

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 and 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 2.

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



March 2008–March 2012	
Models	AUC
Baseline	.500
OLS	.815
Ridge	.828
Lasso	.836
RF	.800
SVM	.825

Figure 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.

prising 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. 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 that out-of-sample results were not significantly improved by either the addition of drug/drug interaction terms or drug/non-drug interaction terms, and therefore chose to proceed with the simpler models without interaction terms. Since both the Ridge Regression and Lasso models show very similar performance in Figures 1 and 2, and because the Ridge Regression model lends itself directly to the computation of a model uncertainty measure (see Section 4.2) that we use in the design of clinical

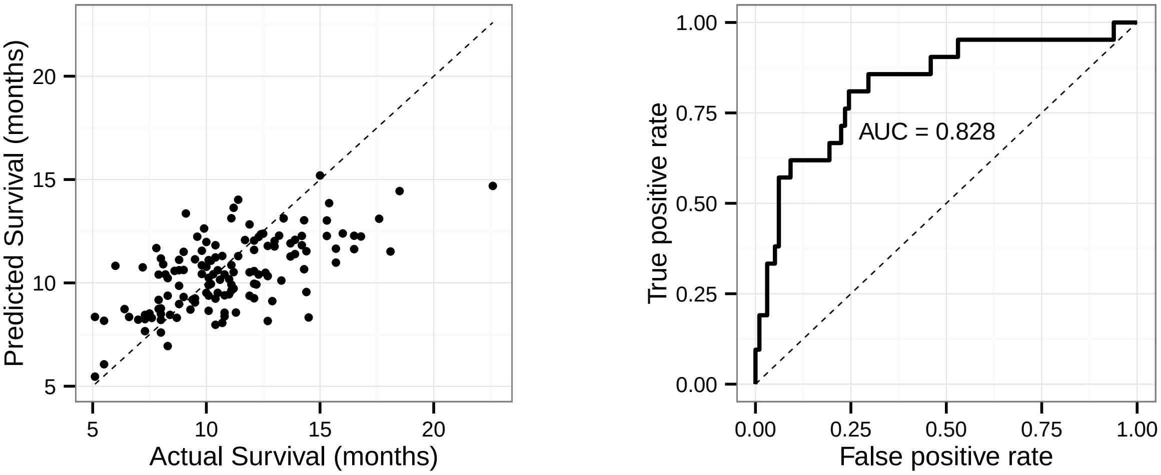


Figure 3: Performance of the Ridge Regression models for survival and toxicity over the most recent 4 years of data (March 2008–March 2012) [Left] Predicted vs. actual values for survival model ($n = 132$). [Right] ROC curve for high toxicity (score > 0.5) predictions, of which 21 are actually high ($n = 119$).

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 model are shown for the Ridge Regression models in Figure 3.

4 Design of Clinical Trials

This section describes an analytical approach for designing clinical trials using mixed integer and robust optimization, using the extracted data and the predictive statistical models we have developed in Sections 2 and 3. We describe the mixed integer optimization model we use and then present a robust version of this model, which provides more conservative suggestions to account for uncertainty in the statistical models. 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 use the predictive models presented in Section 3 to select the trial that maximizes survival, limiting the proportion of patients with a dose-limiting toxicity to no more than some value t. 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 the median overall survival (OS) of patients in the treatment group.

To use the predictive models in an optimization problem, we must provide values for each variable in these models. We fix the patient characteristic variables described in Table 2 to their mean values across all trials. Because our statistical models include no demographic/drug interaction terms, the patient characteristics uniformly affect the predicted overall survival and toxicity of each suggested combination. The suggested treatments for the average patient population with toxicity limit t are therefore identical to the suggested treatments for patient population P with toxicity limit $t + \Delta_P$, for some constant Δ_P. A clinical trial decision maker can optimize for any desired population by appropriately adjusting the toxicity limit. In our database, $|\Delta_P| \leq 0.1$ in 85.7% of treatment arms, so optimization suggestions for the average population are applicable for most clinical trial populations.

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 $(\bar{\boldsymbol{\beta}}^z)'\mathbf{z}+(\bar{\boldsymbol{\beta}}^u)'\mathbf{u}+(\bar{\boldsymbol{\beta}}^v)'\mathbf{v}$, where the superscripts denote the corresponding coefficients for the decision variables (the binary drug variables $\mathbf{z}$, the instantaneous dose variables $\mathbf{u}$, and the average dose variables $\mathbf{v}$). Similarly, we have a model for overall toxicity, which we will denote by $(\bar{\boldsymbol{\tau}}^z)'\mathbf{z} + (\bar{\boldsymbol{\tau}}^u)'\mathbf{u} + (\bar{\boldsymbol{\tau}}^v)'\mathbf{v}$. 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:

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

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 4.3, 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. In our models, we disallow any therapy whose drugs match those in a previous study, and we incorporate the generally accepted guidelines for selecting combination chemotherapy regimens (Page and Takimoto 2002, Pratt 1994, Golan et al. 2008).[5] As discussed in Section 2.1, we also eliminate the drug trastuzumab because it is only indicated for the subpopulation of HER2-positive patients. We leave research into effective treatments for this subpopulation as future work.

Additional requirements could be added to constraints (1c), though we do not do so in this work. Such additional constraints may be necessary due to known toxicities and properties of the drugs, or 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 require a new drug they have developed and only tested

[5] We limit the drug combinations to contain no more than one drug from the classes of drugs used for gastric cancer. Further, we disallow drug combinations that appear no more than once in our database and were discouraged in the guidelines. The following pairs of classes were disallowed from being used together: anthracycline/camptothecin, alkylating 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.

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 $\mathbf{\Omega}_i$, which is defined to be the set of all combinations of instantaneous and average doses that have been used in prior clinical trials. This is important since the instantaneous dose and the average dose are often not independent, so 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 paper, 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:

$$\max_{\mathbf{z}^j, \mathbf{u}^j, \mathbf{v}^j, \forall j} \quad \sum_{j=1}^{k} (\bar{\boldsymbol{\beta}}^z)'\mathbf{z}^j + (\bar{\boldsymbol{\beta}}^u)'\mathbf{u}^j + (\bar{\boldsymbol{\beta}}^v)'\mathbf{v}^j \tag{2}$$

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

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

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 Appendix B. In the remainder of the paper, 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 more diverse trials can be selected via 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 ($\bar{\boldsymbol{\beta}}$). To handle this, we use robust optimization to allow $\bar{\boldsymbol{\beta}}$ to vary in an uncertainty set. For computational simplicity, we allow the binary drug coefficients to vary, while keeping the instantaneous and average dose coefficients fixed.

The use of linear models allows us to parameterize the drug-specific uncertainties. Let $\mathbf{X}$ denote the matrix of input variables in our training set (rows correspond to trial arms, columns to variables), and let $\mathbf{Y}$ be a vector of median overall survivals. We then use the following equation motivated by Bishop (2006) as an approximate uncertainty matrix for the regularized Ridge estimator: $\mathbf{\Sigma_R} = \sigma^{\mathbf{2}}(\mathbf{X'X} + \lambda\mathbf{I})^{\mathbf{-1}}$. 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. 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 of using drug i at its average dose on the median overall survival has an uncertainty $\sigma_i^2 = \mathbf{x_i'\Sigma_R x_i}$.

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

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

We can then reformulate this as the following robust problem:

$$\begin{aligned} \max_{(\mathbf{z}^j,\mathbf{u}^j,\mathbf{v}^j)\in\mathbf{W}} \min_{\boldsymbol{\beta}^z} \quad & \sum_{j=1}^{k}[(\boldsymbol{\beta}^z)'\mathbf{z}^j + (\bar{\boldsymbol{\beta}}^u)'\mathbf{u}^j + (\bar{\boldsymbol{\beta}}^v)'\mathbf{v}^j] && (4)\\ \text{subject to} \quad & \frac{|\beta_i^z - \bar{\beta}_i^z|}{\sigma_i} \le \Gamma, \qquad i = 1,\ldots,n, && (4a)\\ & \sum_{i=1}^{n}\frac{|\beta_i^z - \bar{\beta}_i^z|}{\sigma_i} \le \Gamma\sqrt{N}, && (4b) \end{aligned}$$

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. 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. Our uncertainty set is a simple interval set and is motivated by the Central Limit Theorem (Bandi and Bertsimas 2012).

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 different drugs are selected in the k trials, many coefficients will contribute to the objective and the minimization will try to push all of these coefficients to their worst case bounds. However, constraint (4b) will prevent this from happening since it constrains the total deviation of the coefficients from the nominal values to be less than $\Gamma\sqrt{N}$. On the contrary, if only a few drugs are selected, only a few coefficients will contribute to the objective, and there will be fewer nonzero terms in the sum of constraint (4b). Because the right hand side of the constraint remains the same regardless of how many different drugs are selected, the constraint becomes less restrictive with fewer terms in the sum. This is similar to diversification in financial applications, in that risk is reduced by diversifying a portfolio. Evidence of this will be shown in the results section.

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.

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

4.3 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 $k = 20$, a value that is large enough to ensure several of the suggestions match future trials. 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 "current practice," 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 the "current practice," 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.[6] 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 "current practice," we then average the scores over all trials in the evaluation set. The interpretation of this score is that if a clinical trial decision maker 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.

There are two parameters that can be adjusted in the optimization model to affect the nature of

[6] 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.

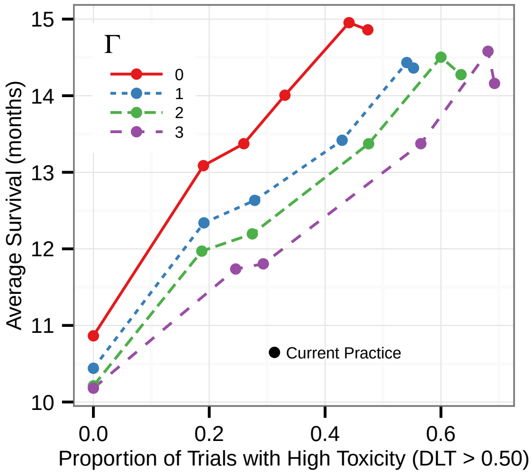


Figure 4: Average scores for Matching Metric for optimization suggestions made from 2002–2012. Each line corresponds to a different value of the robustness parameter Γ, and the points correspond to different right-hand-side values t for the toxicity constraint in the set $\{0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$.

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 current practice performance 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 current practice, with comparable toxicity. In addition, with $(\Gamma = 0, t = 0.2)$, the matching trials suggested by the optimization model have slightly greater survival than those suggested in current practice, and no instances of high toxicity. This evidence suggests our methods are indeed selecting high-performing clinical trials before they are being run in practice.

Figure 4 shows that the best results for the Matching Metric are achieved at $\Gamma = 0$. We note, however, that strong performance is still observed with positive 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 3. 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. We measure the diversity of a set of combinations as the average number of drug in common between each pair of combinations in the set; the higher the number in common, the less diversity. Average number of drugs in common calculated over the entire evaluation period (2002–2012) are given in Table 3, which shows that diversity increases substantially with increasing values of Γ.

Γ	Number of Matches / Number of Suggestions	Average Drugs in Common
0	494 / 6440 (7.7%)	1.389
1	599 / 6440 (9.3%)	1.245
2	673 / 6440 (10.5%)	1.124
3	710 / 6440 (11.0%)	1.020

Table 3: Match rate and average number of drugs in common 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 point in time, 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 a "current practice" 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 future 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 5, 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 "current practice." They begin improving in 2004, and by 2006 the suggestions made by optimization are strongly outperforming those made by current practice. 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 current practice score does show some improvement in outcomes over this period, but the improvement is not as rapid as that shown by the optimization model.

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 that are actually run in the future and turn out to be

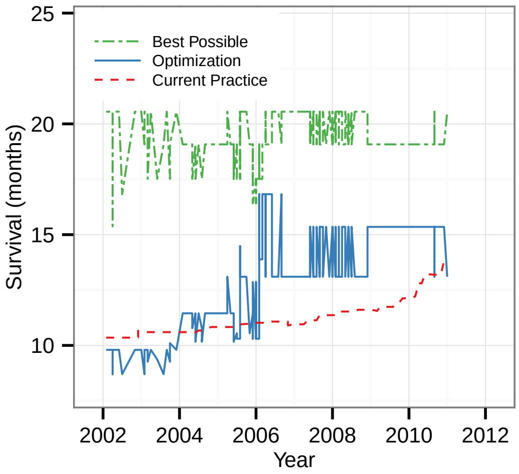


Figure 5: Matching Metric evaluated on optimization suggestions made from 2002–2012 ($\Gamma = 0, t = 0.4$). Average actual survival of our optimization suggestions is compared to the best possible score and a current practice baseline.

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 that 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 "current practice" score, calculated as a mean over all treatments on the list. The interpretation of this score is that if a decision maker randomly ordered 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 6. It is 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 evaluated after this point in 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 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 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 average dosages, and averaging the results, and (3) the "current practice" 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 6. For all metrics, the performance of the optimization model starts out weakly, but improves rapidly over the course of the 10-year evaluation period.

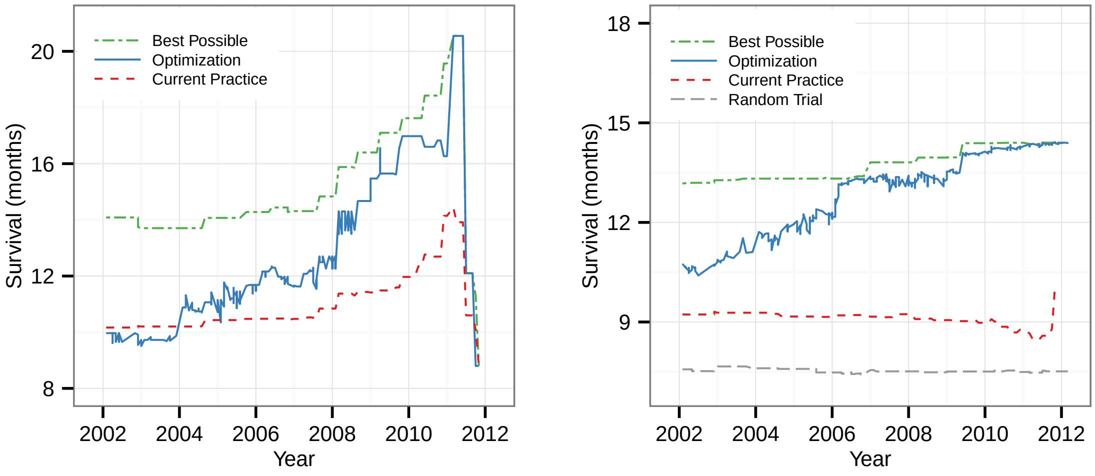


Figure 6: Ranking Metric (left) and Final Model Metric (right) evaluations of optimization suggestions made from 2002–2012 ($\Gamma = 0, t = 0.4$).

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 error $\mathbf{x}'\mathbf{\Sigma_R}\mathbf{x}$, where $\mathbf{x}$ is the trial data and $\mathbf{\Sigma_R}$ is the uncertainty matrix defined in Section 4.2, we calculated the probability that the current trial's median overall survival exceeds the 5-year rolling average median overall survival. In our experiment, we fix a threshold p and flag any trial with probability less than p of exceeding the rolling average overall survival.

Table 4 displays the properties of the flagged trials for a variety of threshold probabilities p. The proposed model is effective at identifying trials that are unlikely to outperform recent trials. At threshold $p = 0.163$, the 10 trials flagged all underperformed the rolling average, while with $p = 0.256$, 30 of 40 flagged trials did not achieve the mean, eight were above average but not in the fourth quartile for survival, and two were in the top quartile of recent trials.

Though the decision maker in this experiment is simplistic, ranking trials without regard for their demographics or their toxicity outcomes, trial outcome predictions and uncertainties provide a useful tool for decision makers in computing the probability that trial objectives will be achieved.

Num. Flagged	p	**Below Average**	**Third Quartile**	**Fourth Quartile**
10	0.163	10	0	0
20	0.223	15	4	1
30	0.238	22	6	2
40	0.256	30	8	2
50	0.278	38	8	4
60	0.305	45	11	4

Table 4: Out-of-sample accuracy in identifying unpromising trials before they are performed.

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 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 7, 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 use 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 -1.91 months in the survival model. A trial with 80% prior palliative chemotherapy, instead of the population average of 13% (from Table 2), would be expected to have $-1.91 * (0.13 - 0.80) = 1.28$ months lower survival. We correct for this effect by adding 1.28 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 7.

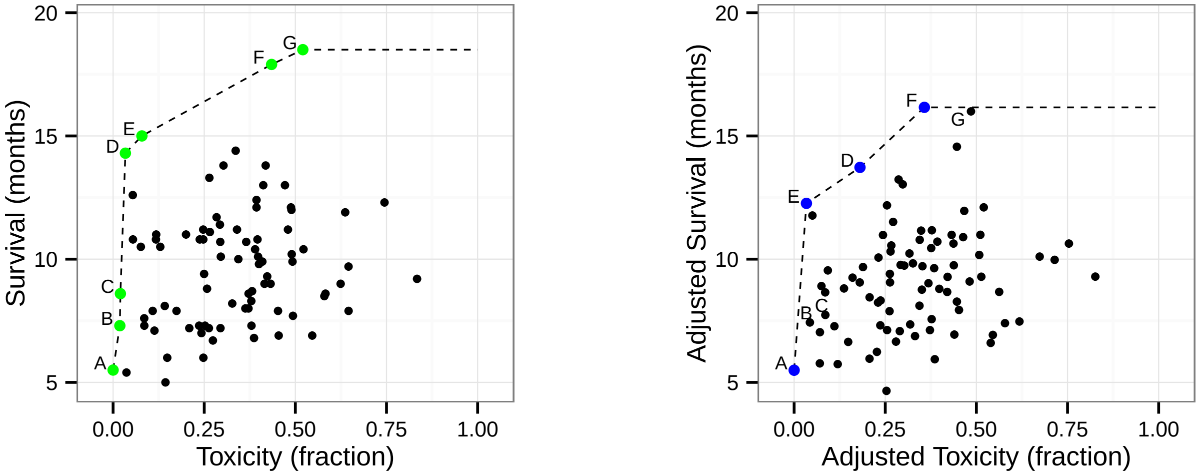


Figure 7: Survival and dose-limiting toxicity for clinical trial arms with ≥ 55 patients, before (left) and after (right) adjustment for demographic factors.

The efficient frontier changes considerably when 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.

6 Discussion and Future Work

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.

There are some limitations to our approach. While we show that we can predict future clinical trial outcomes even if the exact combination of drugs being predicted has never been tested in a clinical trial before, a limitation is that there may be undesired outcomes due to combining drugs for the first time that our models are unable to capture. Another important limitation is that even though we suggest drug and dosage combinations for future trials, a phase I trial would still need to be performed to determine if the combination is safe and we do not consider the costs involved. Additional limitations are that we are not able to use patient-level data, we do not incorporate advanced trial designs (such as sequential trials), we do not generate inclusion/exclusion criteria for patients in the study, and we cannot guarantee that the drug combinations that we suggest are biologically feasible. Finally, our approach learns from and proposes clinical trials, which generally enroll healthier than average patients. Additional work remains to evaluate suggested treatments in the general population.

Despite these limitations, the models presented to predict survival and toxicity given demographic information and chemotherapy drugs and dosages represent the first application of data mining techniques to suggest new clinical trials. Our modeling results show that we can successfully predict future clinical trial outcomes using past data. 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.

We have identified some areas of future work. We see the setting of an online learning problem as an interesting and challenging future direction. Additionally, automated information extraction could be explored as a more scalable way of extracting the data from the clinical trial papers. This work can also be extended to different types of cancers, and could be enhanced by using individual patient data.

Acknowledgments

The authors gratefully acknowledge the contributions of Natasha Markuzon in helping to design semi-automated forms of data extraction and of Jason Acimovic, Cynthia Barnhart, Allison Chang and over twenty MIT undergraduate students in the early stages of this research.

A Data Preprocessing

We performed a series of data preprocessing steps to standardize the data collected from clinical trials.

A.1 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). The ECOG scale runs from 0–5, where 0 is a patient who is fully active and 5 represents death. Among the 483 treatment arms evaluated in this work, 414 (85.7%) reported performance status with the ECOG scale, 68 (14.1%) reported with the KPS scale, and 1 (0.2%) did not report performance status.

In the clinical trials considered, 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 $p_1 + 2p_{\geq 2}$.

In 92 treatment arms, the proportion of patients with ECOG score 0 and 1 was reported as a combined value. To compute the weighted performance score for these arms, we first obtain an estimate of the proportion of patients with ECOG score 0 and score 1, based on the proportion of patients with score 0 or 1. 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 $p_0 + p_1$. For the 16 treatment arms with the proportion of patients with each KPS score reported, 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 treatment arms reporting performance status with other data groupings, the combined score is marked as unavailable. In total, 77 trial arms (15.9%) were assigned a missing score.

A.2 Grade 4 Blood Toxicities

We need to compute the proportion of patients with each Grade 4 blood toxicity to compute the proportion of patients with a DLT, as defined in Section 2.2. Two of the most common blood toxicities, leukopenia and neutropenia, are rarely reported together due to their similarity (neutrophils are the most common type of leukocyte). Because neutropenia is more frequently reported than leukopenia in clinical trial reports, we chose this as the common measure for these two toxicities. We trained a linear model using the proportion of patients experiencing Grade 3/4 leukopenia to predict the proportion of patients experiencing Grade 4 neutropenia, training on the 138 arms that reported both proportions. This model had $R^2 = 0.727$, and we used the model to convert data from 98 treatment arms that reported leukopenia toxicity data but not neutropenia.

Many treatment arms report the proportion of patients with a Grade 3/4 blood toxicity but do not provide the proportion of patients specifically with a Grade 4 blood toxicity. For neutropenia, we built a quadratic model with $R^2 = 0.868$ to predict the proportion of patients with Grade 4 neutropenia given the proportion with Grade 3/4 neutropenia, training on the 211 arms reporting both values. Similarly, we trained linear models to predict the proportion of patients with Grade 4 thrombocytopenia, anemia, and lymphopenia using the proportion of patients with Grade 3/4 levels of these toxicities. These models were trained on 302, 248, and 8 arms, respectively, that reported both values and had R^2 values of 0.669, 0.260, and 0.035, respectively. The models for neutropenia, thrombocytopenia, anemia, and lymphopenia were used to compute the proportion of patients with a Grade 4 toxicity in 118, 93, 99, and 3 treatment arms, respectively.

A.3 Proportion of Patients with a DLT

The fraction of patients with at least one DLT during treatment cannot be calculated directly from the individual toxicity proportions reported. 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%. Here we compare approaches for computing 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 (National Cancer Institute 2006).

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 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 5: Correlation of estimates of total Grade 3/4 toxicity to the true value.

As reported in Table 5, 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. We use this approach to compute the final proportion of patients experiencing a DLT.

If one or more of the DLTs 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 proportion of patients experiencing a DLT for that trial arm is marked as unavailable. This is the case for 106/483 (21.9%) of trial arms in the database.

B Column generation approach

The optimization model (4) in Section 4.2 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 indices of treatments using tuple $R \in T$. Finally let $\mathcal{U}$ be uncertainty set for $\boldsymbol{\beta}^z$, as defined in (4a) and (4b). Then (4) can be reformulated as:

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

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

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

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

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 $0 \leq \delta^t \leq 1$. This provides an upper bound on the optimal solution. Then by taking the dual of the inner problem, we can rewrite (5) as a linear optimization problem. We solve this problem using column generation.

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.

To evaluate the efficiency and solution quality of the column generation approach, 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. A limit of 10 seconds was imposed for solving the restricted version of (5), but this limit is rarely met and was not binding in this analysis. Results were compared with a mixed integer programming formulation of (4). Models were implemented using the Gurobi C++ interface (Gurobi Optimization 2013), and experiments were run on an Intel Core i7-860 (2.8 GHz) with 16 GB RAM. A computational limit of 30 minutes was applied for each model.

Results are given in Table 6. 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 6: 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).

References

Ajani, Jaffer, Wuilbert Rodriguez, Gyorgy Bodoky, et al. 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.

Bandi, Chaithanya, Dimitris Bertsimas. 2012. Tractable stochastic analysis in high dimensions via robust optimization. *Mathematical Programming* **134** 23–70.

Bang, Yung-Jue, Eric Van Cutsem, Andrea Feyereislova, et al. 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.

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.

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

Chao, Y., C. P. Li, T. Y. Chao, et al. 2006. An open, multi-centre, phase II clinical trial to evaluate the efficacy and safety of paclitaxel, UFT, and leucovorin in patients with advanced gastric cancer. *British Journal of Cancer* **95** 159–163.

Chou, Ting-Chao. 2006. Theoretical basis, experimental design, and computerized simulation of synergism and antagonism in drug combination studies. *Pharmacological Reviews* **58**(3) 621–681.

De Ridder, Filip. 2005. Predicting the Outcome of Phase III Trials using Phase II Data: A Case Study of Clinical Trial Simulation in Late Stage Drug Development. *Basic and Clinical Pharmacology and Toxicology* **96** 235 – 241.

Delen, Dursun, Glenn Walker, Amit Kadam. 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, et al. 2003. The costs of conducting clinical research. *Journal of Clinical Oncology* **21**(22) 4145–4150.

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, et al., eds. 2008. *Principles of Pharmacology: The Pathophysiologic Basis of Drug Therapy*. 2nd ed. Lippincott Williams and Wilkins.

Gurobi Optimization, Inc. 2013. Gurobi optimizer reference manual. URL `http://www.gurobi.com`.

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, et al. 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.

Hurria, Arti, Kayo Togawa, Supriya G. Mohile, et al. 2011. Predicting Chemotherapy Toxicity in Older Adults With Cancer: A Prospective Multicenter Study. *Journal of Clinical Oncology* **29**(25) 3457 – 3465.

Iwase, H., M. Shimada, T. Tsuzuki, et al. 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, et al. 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, et al. 2011. Global cancer statistics. *CA: A Cancer Journal for Clinician* **61** 69–90.

Kang, Y.-K., W.-K. Kang, D.-B. Shin, et al. 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, Hiroyuki Narahara, Takuo Hara, et al. 2008. S-1 plus cisplatin versus s-1 alone for first-line treatment of advanced gastric cancer (SPIRITS trial): a phase III trial. *Lancet* **9** 215–221.

Lee, Kyung Hee, Myung Soo Hyun, Hoon-Kyo Kim, et al. 2009. Randomized, multicenter, phase iii trial of heptaplatin 1-hour infusion and 5-fluorouracil combination chemotherapy comparing with cisplatin and 5-fluorouracil combination chemotherapy in patients with advanced gastric cancer. *Cancer Research and Treatment* **41** 12–18.

Lee, Y.-J., O.L. Mangasarian, W.H. Wolberg. 2003. Survival-Time Classification of Breast Cancer Patients. *Computational Optimization and Applications* **25**(1) 151–166.

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

Lutz, Manfred P., Hansjochen Wilke, D.J. Theo Wagener, et al. 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, et al. 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, et al. 2009. Convergence 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, et al. 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.

van't Veer, Laura J., René Bernards. 2008. Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature* **452**(7187) 564–570.

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.

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.

Zhao, Lihui, Lu Tian, Tianxi Cai, Brian Claggett, L.J. Wei. 2011. Effectively selecting a target population for a future comparative study. *Harvard University Biostatistics Working Paper Series* .

Zhao, Yingqi, Donglin Zeng, A. John Rush, Michael R. Kosorok. 2012. Estimating individualized treatment rules using outcome weighted learning. *JASA* **107**(499) 1106 – 1118.