Submitted to *Management Science*
manuscript MS-14-01733.R1

Authors are encouraged to submit new papers to INFORMS journals by means of a style file template, which includes the journal title. However, use of a template does not certify that the paper has been accepted for publication in the named journal. INFORMS journal templates are for the exclusive purpose of submitting to an INFORMS journal and should not be used to distribute the papers in print or online or to submit the papers to another publication.

An Analytics Approach to Designing Combination Chemotherapy Regimens for Cancer

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

Allison O'Hair
Stanford Graduate School of Business, akohair@stanford.edu

Stephen Relyea
Lincoln Laboratory, Massachusetts Institute of Technology, srelyea@ll.mit.edu

John Silberholz
Sloan School and Operations Research Center, Massachusetts Institute of Technology, josilber@mit.edu

Dedicated to the memory of John Bertsimas (1934–2009).

Cancer is a leading cause of death worldwide, and advanced cancer is often treated with combinations of multiple chemotherapy drugs. In this work, we develop models to predict the outcomes of clinical trials testing combination chemotherapy regimens before they are run and to select the combination chemotherapy regimens to be tested in new Phase II and III clinical trials, with the primary objective of improving the quality of regimens tested in Phase III trials compared to current practice. We built a database of 414 clinical trials for gastric and gastroesophageal cancers and use it to build statistical models that attain an out-of-sample R^2 of 0.56 when predicting a trial's median overall survival (OS) and an out-of-sample area under the curve (AUC) of 0.83 when predicting if a trial has unacceptably high toxicity. We propose models that use machine learning and optimization to suggest regimens to be tested in Phase II and III trials. Though it is inherently challenging to evaluate the performance of such models without actually running clinical trials, we use two techniques to obtain estimates for the quality of regimens selected by our models compared with those actually tested in current clinical practice. Both techniques indicate the models might improve the efficacy of the regimens selected for testing in Phase III clinical trials without changing toxicity outcomes. This evaluation of the proposed models suggests they merit further testing in a clinical trial setting.

1. Introduction

Cancer is a leading cause of death worldwide, accounting for 8.2 million deaths in 2012. This number is projected to increase, with an estimated 13.1 million deaths in 2030 (World Health Organization 2012). The prognosis for many solid-tumor cancers is grim unless they are caught at an early stage, when the tumor is contained and can still be surgically removed. At the time of diagnosis, the tumor is often sufficiently advanced that it has metastasized to other organs and can no longer be surgically removed, leaving drug therapy or best supportive care as the best treatment options.

A key goal of oncology research for advanced cancer is to identify novel chemotherapy regimens that yield better clinical outcomes than currently available treatments (Overmoyer 2003, Roth 2003). Phase II clinical trials are used to evaluate the efficacy of novel regimens, with a focus on exploring treatments that have never been previously tested for a disease — in this work we found that 89.4% of Phase II trials for advanced gastric cancer test a new chemotherapy regimen. While some trials evaluate a new drug for a particular cancer, the majority (84.6% for gastric cancer) instead test novel combinations of existing drugs in different dosages and schedules; in this work we focus primarily on this type of chemotherapy regimen. The most effective regimens identified in Phase II trials are then evaluated in Phase III studies, which are large randomized controlled trials comparing one or more experimental regimens against a control group treated with the best available standard chemotherapy regimen (Friedman et al. 2010b). Treatments that perform well against standard treatments in Phase III trials may then be considered new standard regimens for advanced cancer; identifying such regimens and thereby improving the set of treatments available to patients is a key goal of oncology research for advanced cancer (Overmoyer 2003, Roth 2003).

Finding novel, effective chemotherapy treatments for advanced cancer is challenging in part because the most effective chemotherapy regimens often contain more than one drug. Meta-analyses for a number of cancers have demonstrated efficacy gains of combination chemotherapy regimens over single-agent treatments (Delbaldo et al. 2004, Wagner 2006), and in this work we found that 80% of all chemotherapy clinical trials for advanced gastric cancer have tested multi-drug treatments. As a result of the large number of different chemotherapy drugs, there are a huge number of potential drug combinations that could be investigated in a new Phase II trial, especially when considering different dosages and dosing schedules for each drug. However, testing any chemotherapy regimen in a clinical trial is expensive, costing on average more than $10 million for Phase II studies and $20 million for Phase III trials (Sertkaya et al. 2014); these costs are often incurred either by pharmaceutical companies or by the government. Furthermore, even after a Phase II study has been run testing a new regimen, it can be difficult to determine whether this regimen is a good candidate for testing in a larger Phase III study because Phase II trials often enroll patient populations that are not representative of typical advanced cancer patients (Friedman et al. 2010b). For these reasons, it is a challenge for researchers to identify effective new combination chemotherapy regimens.

Our aspiration in this paper is to propose an approach that could serve as a method for selecting the chemotherapy regimens to be tested in Phase II and III clinical trials. Because Phase III clinical trials are used to test the most promising chemotherapy regimens to date and can directly affect standard clinical practice, our central objective in this work is to design tools that can improve the quality of the chemotherapy regimens tested in Phase III trials compared to current practice. The key contributions of the paper are:

Clinical Trial Database We developed a database containing information about the patient demographics, study characteristics, chemotherapy regimens tested, and outcomes of all Phase II and III clinical trials for advanced gastric cancer from papers published in the period 1979–2012 (Section 2). Surprisingly, and to the best of our knowledge, such a database did not exist prior to this study.

Statistical Models Predicting Clinical Trial Outcomes We train statistical models using the results of previous randomized and non-randomized clinical trials (Section 3). We use these models to predict survival and toxicity outcomes of new clinical trials evaluating regimens whose drugs have individually been tested before, but potentially in different combinations or dosages. To our knowledge, this is the first paper to employ statistical models for the prediction of clinical trial outcomes of arbitrary drug combinations and to perform an out-of-sample evaluation of the predictions.

Design of Chemotherapy Regimens We propose and evaluate tools for suggesting novel chemotherapy regimens to be tested in Phase II studies and for selecting previously tested regimens to be further evaluated in Phase III clinical trials (Section 4). Our methodology balances the dual objectives of exploring novel chemotherapy regimens and testing treatments predicted to be highly effective. To our knowledge, this is the first use of statistical models and optimization to design novel chemotherapy regimens based on the results of previous clinical trials.

We summarize the models developed and evaluated in this paper in Table 1. In Section 4.4 we approximate the quality of our suggested chemotherapy regimens using both simulated clinical trial outcomes and the true outcomes of similar clinical trials in our database. In Section 5 we discuss the next step in evaluating our models: using clinical trials to evaluate the quality of the chemotherapy regimens we suggest.

The approach we propose in this work is related to both patient-level clinical prediction rules and meta-regressions, though it differs in several important ways. 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

Model	**Approach**	**Evaluation Techniques**
Prediction of clinical trial efficacy and toxicity outcomes	Statistical models trained on a large database of previous clinical trials (Section 3.2)	Sequential out-of-sample R^2, root mean square error, and area under the curve (Section 3.3), as well as evaluation of whether models could aid planners in avoiding unpromising trials (Section 3.4)
Design of novel chemotherapy regimens for evaluation in Phase II studies	Integer optimization using our statistical models to select novel chemotherapy regimens with high predicted efficacy and acceptable predicted toxicity (Section 4.1)	The *simulation and matching metrics*, which use both simulation and the outcomes of true clinical trials to compare our suggested chemotherapy regimens against those selected in current clinical practice (Section 4.4)
Selection of previously tested chemotherapy regimens for further evaluation in Phase III clinical trials	Using our statistical models to identify previously tested regimens with high predicted efficacy and acceptable predicted toxicity (Section 4.2)	

Table 1 A summary of the models developed and evaluated in this paper.

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. 2009, 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 designing novel chemotherapy regimens.

The technique of meta-regression involves building models using the effect size of randomized trials as the dependent variable and using independent variables such as patient demographics and information about the chemotherapy regimen in a particular trial. These models 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 similar to the prediction models we build, representing trial outcomes as a function of trial properties, they are used to explain differences between existing randomized trials, and study authors generally do not evaluate the out-of-sample predictiveness of the models. Like meta-analyses, meta-regressions are performed on a small subset of the clinical trials for a given disease, often containing just a few drug combinations. Even when a wide range of drug combinations are considered, meta-regressions typically do not contain enough drug-related variables to be useful in proposing new trials. For instance, Hsu et al. (2012) uses 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. Finally, meta-analyses and meta-regressions are typically performed on a subset of published randomized controlled studies, while our approach uses data from both randomized and non-randomized studies.

Because approaches such as patient-level clinical prediction rules, meta-analysis, and meta-regression cannot be readily used to design new chemotherapy regimens to be tested in clinical trials, other methods, collectively termed *preclinical models*, are instead used in the treatment design process. Following commonly accepted principles for designing combination chemotherapy regimens (Page and Takimoto 2002), researchers seek to combine drugs that are effective as single agents and that show synergistic behavior when combined; drugs that cause the same toxic effects and that have the same patterns of resistance are not combined in suggested regimens. Molecular simulation is a well developed methodology for identifying synergism in drug combinations (Chou 2006), and virtual clinical trials, which rely on pharmacodynamic and pharmacokinetic models to analyze different drug combinations, can also be used to suggest new treatments (Kleiman et al. 2009). Animal studies and *in vitro* experimentation can be used to further evaluate novel chemotherapy regimens; results of these preclinical studies are often cited as motivations for Phase II studies of combination chemotherapy regimens (Chao et al. 2006, Iwase et al. 2011, Lee et al. 2009). A key limitation of current preclinical models is that most do not incorporate treatment outcomes from actual patients, while the new models we propose in this work leverage patient outcomes reported in previous clinical trials.

We believe the data-driven approaches we propose in this paper would complement existing preclinical models. For instance, an *in vitro* experiment could be performed to evaluate the anti-tumor activity of a combination chemotherapy regimen suggested by our model from Section 4.1. Such experimentation could be used to evaluate and refine chemotherapy regimens suggested by our models, which would be especially important when our model suggests a regimen that combines drugs that have never been tested together in a prior clinical trial. Because our models cannot accurately predict outcomes of clinical trials testing new drugs, existing preclinical models would need to be used to design therapies incorporating these new drugs. On the other hand, our statistical models could be used to predict the efficacy and toxicity of chemotherapy regimens designed using other preclinical models, identifying the most and least promising suggested regimens.

In this work, we evaluate our proposed approach on gastric cancer. Not only is this cancer important — gastric cancer is the third leading cause of cancer death in the world (Torre et al. 2015) — but there is no single chemotherapy regimen widely considered to be the standard or best treatment for this cancer (Wagner 2006, Wong and Cunningham 2009, NCCN 2013), and

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 an experimental treatment is compared to a standard treatment.
Cycle	The length of time between repeats of a dosing schedule in a chemotherapy treatment.
Exclusion criteria	The factors that make a person ineligible from participating in a clinical trial.
Inclusion criteria	The factors that allow a person to participate in a clinical trial.
Phase I Study	A clinical study focused on identifying safe dosages for an experimental treatment.
Phase I/II Study	A study that combines a Phase I and Phase II investigation.
Phase II Study	A clinical study that explores the efficacy and toxicity of an experimental treatment.
Phase III Trial	A randomized controlled trial that compares an experimental treatment with an established therapy.
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 2 Definitions of some common chemotherapy clinical trial terms.

researchers frequently perform clinical trials testing new chemotherapy regimens for this cancer. We believe, however, that our approach has potential to help in selecting regimens to test in trials for many other diseases, and we discuss this as an area of future work in Section 5.

2. Clinical Trial Database

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 terms we use are given in Table 2.

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 cancer,[1] (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, unlike meta-analyses, which typically only include randomized studies. 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.

[1] Clinical trials for gastric cancer often contain patients with cancer of the gastroesophageal junction or the esophagus due to the similarities between these three types of cancer. We include studies as long as all patients have one of these three forms of cancer.

Exclusion criteria were: (1) trials testing sequential treatments, (2) trials that involve the application of radiotherapy,[2] (3) trials that apply chemotherapy for earlier stages of cancer, when the disease can still be cured, and (4) trials to treat gastrointestinal stromal tumors, a related form of cancer.

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 a combination of the following keywords in the title: "gastr*" or "stomach"; "advanced" or "metastatic"; and "phase" or "randomized trial" or "randomised trial". A single individual reviewed these search results, and these searches yielded 350 clinical trials that met the inclusion criteria for this study.

After this search through medical databases, we further expanded our set of papers by searching through the references of papers that met our inclusion criteria. This reference search yielded 64 additional papers that met the inclusion criteria for this study. In total, our literature review yielded 414 clinical trials testing 495 treatment arms 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

A single individual manually extracted data from clinical trial papers and entered extracted data values 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, median time until tumor progression, median 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 (60%), the primary objective was to demonstrate improvement in the median overall survival (OS) — the length of time from enrollment in the study until death — of patients in the treatment group. As a result, this is the metric we have chosen as our measure of efficacy.[3] To capture the toxic effects of treatment, we also extracted the fraction of patients experiencing any toxicity at Grade 3 or 4, designating severe, life-threatening, or disabling toxicities (National Cancer Institute 2006).

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

[3] The full survival distribution of all patients, which enables the computation of metrics such as 6-month and 1-year survival rates, was available for only 348/495 (70.3%) of treatment arms. Meanwhile, the median OS was available for 463/495 (93.5%) of treatment arms. Given the broader reporting of median OS coupled with the established use of median OS as a primary endpoint in Phase III trials, we have chosen this metric as our central efficacy measure.

For each drug in a given trial's chemotherapy treatment, the drug name, dosage level for each application, number of applications per cycle, and cycle length were collected. We also extracted many covariates that have been previously investigated for their effects on response rate or overall survival in prior chemotherapy clinical trials for advanced gastric cancer. To limit the number of missing values in our database, we limited ourselves to variables that are widely reported in clinical trials. These variables are summarized in Table 3.

We chose not to collect many less commonly reported covariates that have also been investigated for their effects on response and survival in other studies, 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 potential 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 the drug trastuzumab (NCCN 2013), while this treatment is ineffective in other patients. We address this issue by excluding trastuzumab from the combination chemotherapy regimen suggestions we make in Section 4.

In Table 3, 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 9,592 toxicity entries, averaging 19.4 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 refer to guidelines established in Phase I clinical trials. The

	Variable	**Average Value**	**Range**	**% Reported**
Patient Demographics	Fraction male	0.72	0.29 – 1.00	97.8
	Fraction of patients with prior palliative chemotherapy	0.14	0.00 – 1.00	98.2
	Median age (years)	59.60	46 – 80	99.0
	Mean performance status[1]	0.88	0.11 – 3.00	84.2
	Fraction of patients with primary tumor in the stomach	0.89	0.00 – 1.00	94.9
	Fraction of patients with primary tumor in the gastroesophageal junction	0.08	0.00 – 1.00	94.3
Study Characteristics	Fraction of study authors from each country (11 different variables for countries in at least 10 trial arms)[2]	Country Dependent	0.00 – 1.00	95.8
	Fraction of study authors from an Asian country[2]	0.42	0.00 – 1.00	95.8
	Number of patients	53.9	11 – 521	100.0
	Publication year	2003	1979 – 2012	100.0
Outcomes	Median overall survival (months)	9.2	1.8 – 22.6	93.5
	Incidence of every Grade 3/4 or Grade 4 toxicity	Toxicity Dependent[3]		

[1] The mean Eastern Cooperative Oncology Group (ECOG) performance status of patients in a clinical trial, on a scale from 0 (fully active) to 5 (dead). See Appendix A.1 for details.

[2] The studies that did not report this variable instead reported affiliated institutions without linking authors to institutions. The proportion of authors from an Asian country serves as a proxy to identify study populations with patients of Asian descent, who are known to have different treatment outcomes than other populations.

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

Table 3 Patient demographic, study characteristic, and outcome variables extracted from gastric cancer clinical trials. These variables, together with the drug variables, were inputted into a database.

primary goal of these early studies is to assess drugs for safety and tolerability on small populations and to determine an acceptable dosage level to use in later trials (Golan et al. 2008). These trials enroll patients at increasing dosage levels until the toxicity becomes unacceptable. The "Patients and Methods" sections of Phase I trials specify a set of so-called *dose-limiting toxicities* (DLTs). If a patient experiences any one of the toxicities in this set at the specified grade, he or she is said to have experienced a DLT. When the proportion of patients with a DLT exceeds a pre-determined 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.” In this work we use the median value of 0.5 to identify trials with an unacceptably high proportion of patients experiencing a DLT. Details on the computation of the proportion of patients experiencing a DLT are presented in Appendix A.3. The 372 clinical trial arms in the dataset with non-missing median OS and DLT proportion are plotted in Figure 1. This figure shows that a typical clinical trial in our database has a median OS between 5 months and 15 months, and a proportion of patients with a DLT between 0 and 0.75.

3. Statistical Models Predicting Clinical Trial Outcomes

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, whose primary aim is the synthesis 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 3 together with data about the drug therapy tested in each trial arm to develop the statistical models. This data

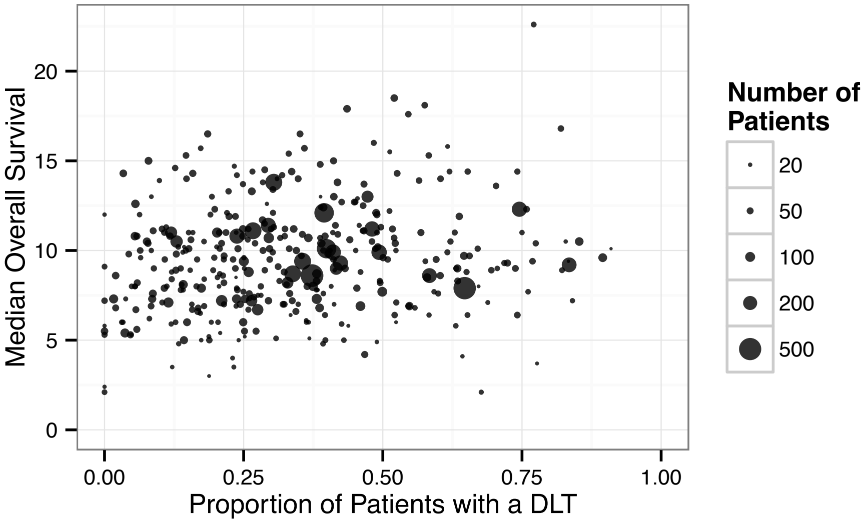


Figure 1 **The results of the 372 clinical trial arms in our database with non-missing median OS and DLT proportion. The size of a point is proportional to the number of patients in that clinical trial arm.**

can be classified into four categories: patient demographics, study characteristics, chemotherapy treatment, and trial outcomes.

One challenge of developing statistical models using data from different clinical trials comes from the patient demographic 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 as independent variables in our model all of the patient demographic and study characteristic variables listed in Table 3. The reporting frequencies for each of these variables is given in Table 3, and missing values are replaced by their variable means or estimates described in Appendix A.1 before model building. In total, we included 20 patient demographic and study characteristic variables in our models.[4]

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 5-fluorouracil[5]: in the first, it is administered at $3,000\,mg/m^2$ 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

[4] Variables include the fraction of patients who are male, the fraction of patients with prior palliative chemotherapy, the median patient age, the mean ECOG performance status of patients, the fraction of patients with a primary tumor in the stomach, the fraction of patients with a primary tumor in the gastroesophageal junction, the fraction of study authors from each country (11 total variables), the fraction of study authors from an Asian country, the number of patients in the study, and the study's publication year.

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

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 each day d is given to patients, and the average dose of a drug d is defined as the average dose of d delivered each week. We do not encode information about loading dosages, which are only given during the first cycle of a chemotherapy regimen, and we use the average instantaneous dosage if a drug is given at different dosages on different days during a cycle. In total, we included 72 drugs in our models, which are listed in the electronic companion of this paper. As a result, we included 216 drug-related independent variables in our models.

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 (and for which we cannot replace the value by estimates described in Appendices A.2 and A.3) 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 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 mean-centered, unit-variance vector of inputs for a proposed trial (i.e. patient, study, and treatment variables) and y represent a particular outcome measure we would like to predict (median OS or DLT proportion), then this class of models assumes a relationship of the form $y = \boldsymbol{\beta}'\mathbf{x} + \beta_0 + \epsilon$, for some unknown vector of coefficients $\boldsymbol{\beta}$, intercept β_0, and error term ϵ. We assume that the noise terms ϵ_i are independent with variance of the form $V(\epsilon_i) = \sigma^2 n_i^{-1}$, where σ is an unknown constant and n_i is the number of patients in trial arm i. We adjust for this expected heteroskedasticity by assigning weight $w_i = n_i/\bar{n}$ to each trial arm i for all linear models, where $\bar{n}$ is the average number of patients in a clinical trial arm. 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 obtain estimates of the regression coefficients $\hat{\boldsymbol{\beta}}$ and $\hat{\beta}_0$ by minimizing the following objective:

$$\min_{\hat{\boldsymbol{\beta}}, \hat{\beta}_0} \sum_{i=1}^{N} w_i (\hat{\boldsymbol{\beta}}' \mathbf{x}_i + \hat{\beta}_0 - y_i)^2 + \lambda \|\hat{\boldsymbol{\beta}}\|_p^p, \tag{1}$$

where N is the number of observations in the training set and 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.[6]

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, as well as results for unregularized linear regression models.

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 (RF) and support vector machines (SVM). For random forests (Breiman 2001), we use the nominal values recommended by Hastie et al. (2009) 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.[7] 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 on the training set.[8]

All models were built and evaluated with the statistical language R version 3.0.1 (R Core Team 2012) using packages `glmnet` (Friedman et al. 2010a), `randomForest` (Liaw and Wiener 2002), and `e1071` (Meyer et al. 2012).

3.3. Statistical Model Results

Following the methodology of Section 2, we collected and extracted data from a set of 414 published journal articles from 1979–2012 describing the treatment methods and patient outcomes for a total of 495 treatment arms of gastric cancer clinical trials.

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 clinical trials in order of their publication date. We then only use the data from prior published trials to predict the

[6] Candidate values of λ are exponentially spaced between $\lambda_{max}/10^4$ and λ_{max}. We take λ_{max} to be the smallest value for which all fitted coefficients $\hat{\boldsymbol{\beta}}$ are (numerically) zero.

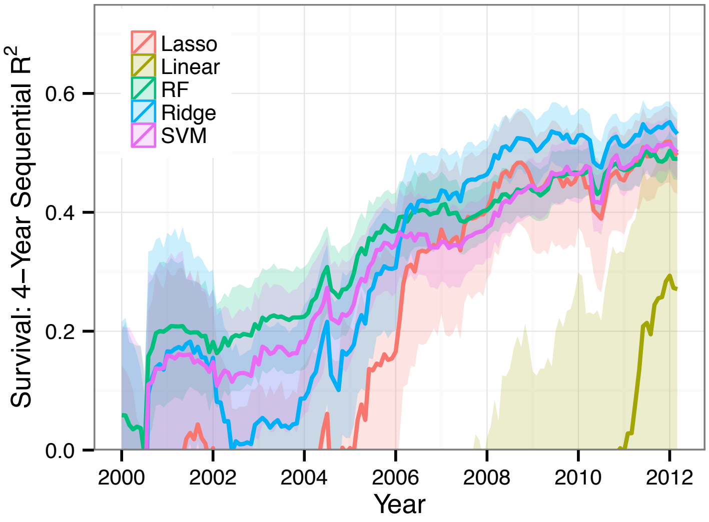
[7] Candidate values are chosen from among exponentially spaced values ($[1.5^{-4}\frac{v}{3}], [1.5^{-3}\frac{v}{3}], \ldots, [1.5^{2}\frac{v}{3}]$), where v is the total number of input variables and $[\cdot]$ denotes rounding to the nearest integer.

[8] Candidate values are chosen from an exponentially spaced 2-D grid of candidates ($C = 2^{-5}, 2^{-3}, \ldots, 2^{15}, \gamma = 2^{-15}, 2^{-13}, \ldots, 2^{3}$).

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 overall survival as well as the overall toxicity score. We begin our sequential testing 20% of the way through the set of 495 total treatment arms, setting aside the first 98 arms to be used solely for model building so that our training set is large enough for the first prediction. Of the remaining 397 arms, we first remove those for which the outcome is not available, leaving 383 arms for survival and 338 for toxicity. We then predict outcomes only for those arms using drugs that have been seen at least once in previous trials (albeit possibly in different combinations and dosages). This provides us with 347 data points to evaluate the survival models and 307 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. This is a standard baseline method used in evaluating sequential prediction models. Model performance is presented in terms of the coefficient of determination (R^2) of our prediction models relative to this baseline. For each 4-year period of time we compute the RMSE and R^2 of each model. To assess statistical fluctuation of these quantities, for each prediction we additionally train 40 models with bootstrap resampled versions of the training set, and for each 4-year period we report the mean, 2.5% quantile, and 97.5% quantile of the RMSE and R^2 obtained when randomly sampling one of the 40 bootstrap model predictions for each of the predictions made during that 4-year period. Figure 2 displays the 4-year sliding-window statistical fluctuation of the out-of-sample R^2 value, along with the values of the RMSE and R^2 over the most recent 4-year window of sequential testing, both for the cross-validation results and the out-of-sample predictions.

To evaluate the toxicity models, recall from the discussion of Section 2.2 that the toxicity of a treatment is considered manageable as long as the proportion of patients experiencing a dose-limiting toxicity (DLT) 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" (DLT proportion > 0.5) and those with "acceptable toxicity" (DLT proportion ≤ 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 unseen trial arm with high toxicity and a randomly chosen unseen trial arm with acceptable toxicity. As was the case for



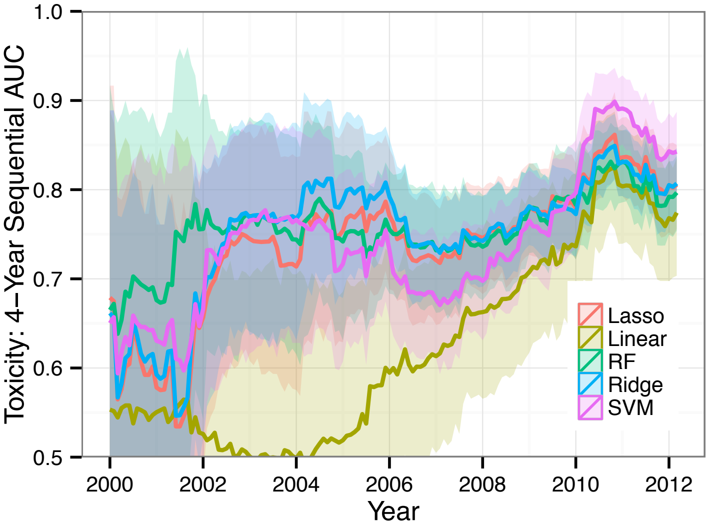
Root Mean Square Error (RMSE)

Model	X-Val	OOS	Bootstrap
Baseline	—	3.75	3.76 [3.73,3.78]
Linear	—	2.72	3.16 [2.67,4.58]
Ridge	2.02	2.50	2.57 [2.47,2.67]
Lasso	2.00	2.44	2.66 [2.50,2.83]
RF	2.05	2.61	2.68 [2.59,2.77]
SVM	2.00	2.60	2.65 [2.54,2.77]

Coefficient of Determination (R^2)

Model	X-Val	OOS	Bootstrap
Baseline	—	0.00	0.00 [-0.01,0.01]
Linear	—	0.48	0.27 [-0.49,0.50]
Ridge	0.49	0.56	0.53 [0.50,0.57]
Lasso	0.50	0.58	0.50 [0.43,0.56]
RF	0.48	0.52	0.49 [0.45,0.53]
SVM	0.50	0.52	0.50 [0.46,0.54]

Figure 2 [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). [Right] Root mean square prediction error (RMSE) and R^2 for the cross-validation set ("X-Val"), for out-of-sample predictions ("OOS"), and for bootstrapped out-of-sample predictions ("Bootstrap") for the most recent 4-year window of data (March 2008–March 2012), which includes 132 out-of-sample predictions.



Area Under the Curve (AUC)

Model	X-Val	OOS	Bootstrap
Baseline	—	0.43	0.49 [0.36,0.61]
Linear	—	0.82	0.77 [0.70,0.84]
Ridge	0.76	0.83	0.81 [0.77,0.85]
Lasso	0.77	0.82	0.81 [0.76,0.85]
RF	0.78	0.81	0.80 [0.75,0.84]
SVM	0.78	0.87	0.84 [0.80,0.89]

Figure 3 [Left] Four-year sliding-window sequential out-of-sample classification accuracy of toxicity models, reported as the area under the curve (AUC) for predicting whether a trial will have high toxicity (DLT proportion $>$ 0.5). [Right] AUC for the cross-validation set ("X-Val"), for out-of-sample predictions ("OOS"), and for bootstrapped out-of-sample predictions ("Bootstrap") for the most recent 4-year window of data (March 2008–March 2012), which includes 119 out-of-sample predictions. Of these, 24 (20.1%) actually had high toxicity.

survival, we calculate the AUC for each model over a 4-year sliding window and assess statistical fluctuation using bootstrapping, with the results shown in Figure 3.

We see in Figures 2 and 3 that models for survival and toxicity all show a trend of improving prediction quality over time, which indicates our models are becoming more powerful as additional data is added to the training set. The decrease in the AUC of the toxicity model toward the end of the testing period might be attributable to the large number of new drugs tested in gastric cancer in recent years — 8% of trial arms from 2007–2009 evaluated a drug that had been tested in fewer than three previous arms, while 21% of arms from 2010–2012 tested such a drug. We see that ridge regression, lasso, SVM, and RF all attain similar performance when predicting both survival and toxicity, with sequential R^2 of more than 0.5 for recent survival predictions and AUC of more than 0.8 for recent toxicity predictions. For both prediction tasks, statistical fluctuations overlap for these four models' performances in the final 48-month window. Especially for earlier predictions, the unregularized linear model is not competitive, likely due to overfitting to the training set.

As a result of this performance assessment, we identified the regularized linear models as the best candidates for inclusion in our optimization models, as they have good prediction quality, are the least computationally intensive, and are the simplest of the models we evaluated. We conducted additional testing to determine whether the explicit inclusion of pairwise interaction terms between variables improved the ridge regression models for survival and toxicity in a significant way. We found that out-of-sample results were not significantly improved by the addition of drug/drug, drug/demographic, or drug/trial information interaction terms, and therefore chose to proceed with the simpler models without interaction terms. The lack of improved out-of-sample performance due to interaction terms may be due to insufficient sample size to identify interaction effects or due to nonlinear interactions that could not be captured by the regularized linear models. We ultimately 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 the toxicity model are shown for the ridge regression models in Figure 4.

While we rely on a naive baseline throughout this section to evaluate our prediction models, it would be challenging to improve this baseline. Clinical trial authors do not publish predictions of trial survival and efficacy outcomes, so we cannot compare our predictions to oncologists' predictions. In Section 3.4 we evaluate whether these models could be used by clinical trial planners to identify unpromising clinical trials before they are run, and in Section 4 we evaluate if our prediction models could help us design effective combination chemotherapy regimens.

3.4. Identifying Unpromising Clinical Trials Before They Are Run

One application of statistical models for predicting a trial's efficacy and toxicity is to identify and eliminate or modify unpromising proposed trials before they are run. Such a tool could assist clinical trial planners in deciding whether to run a proposed trial.

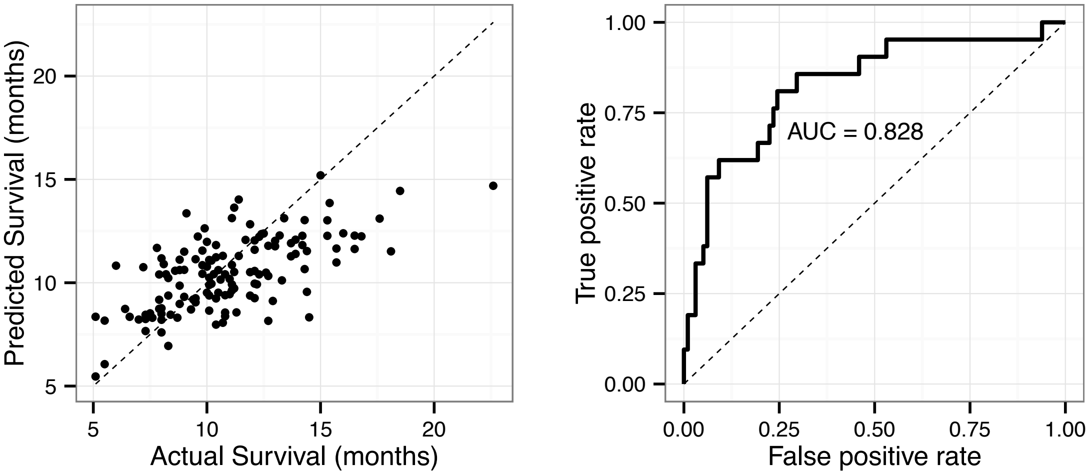


Figure 4 **Performance of the ridge regression models for survival and toxicity over the most recent 4 years of data (March 2008–March 2012) [Left] Predicted vs. actual values for survival model ($n = 132$). [Right] ROC curve for high toxicity (DLT proportion $>$ 0.5) predictions, of which 24 are actually high ($n = 119$).**

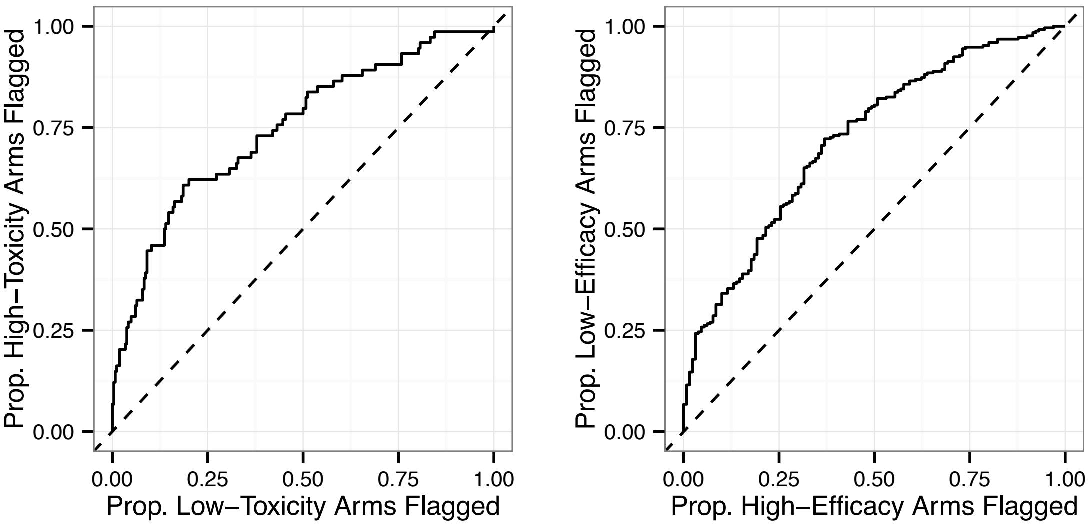


Figure 5 **Performance of the ridge regression models at flagging low-performing trial arms among the 397 arms in the testing set. [Left] Performance flagging trials with unacceptably high toxicity. [Right] Performance flagging trials that do not achieve top-quartile median OS.**

First, clinical trial planners might use the models predicting toxicity to avoid clinical trials predicted to have a high DLT rate or to adjust the dosages of the drugs being tested. The ridge regression model for the proportion of patients with a DLT could be used to rank trials based on their predicted DLT proportion, and trials with predicted values exceeding some cutoff c_{DLT} could be flagged. The left side of Figure 5 plots the proportion of trials with a high DLT proportion (more than 50%) and the proportion of trials with a low DLT proportion that are flagged with various cutoffs c_{DLT} across all 397 studies in the statistic model testing set (studies published since 1997). Overall, the ridge regression model achieves an AUC of 0.75 in predicting if a trial will have a high DLT rate. Further, 10% of all trials with a high DLT rate can be flagged while only flagging 0.4% of trials with a low DLT rate, and 20% of all trials with high DLT rate can be flagged while only flagging 1.9% of trials with a low DLT rate.

Planners might also use the models predicting median OS to identify clinical trials predicted not to attain a high efficacy compared to recent trials in a similar patient population. We stratify trials based on their patient demographics[9] and define a study to have high efficacy if it exceeds the 75th percentile of median OS values reported in trial arms within its strata in the past four years. The ridge regression model for the median OS could be used to rank trials based on the ratio between the predicted median OS and the strata-specific cutoff for high efficacy, and the trials with a ratio below some cutoff c_{OS} could be flagged as being unlikely to achieve high efficacy. The right side of Figure 5 plots the proportion of flagged trials that did and did not achieve the strata-specific cutoff for high efficacy for various cutoffs c_{OS} across the 397 studies in the testing set. Overall, ridge regression model achieves an AUC of 0.72 in predicting if a trial will not achieve high efficacy. Further 10% of all trials that do not achieve high efficacy can be flagged while only flagging 0.8% of trials with high efficacy, and 20% of trials that do not achieve high efficacy can be flagged while only flagging 3.1% of trials with high efficacy.

4. Design of Chemotherapy Regimens

This section describes an approach for designing novel chemotherapy regimens to be tested in Phase II studies using mixed integer optimization, using the extracted data and the statistical models we have developed in Sections 2 and 3. Further, we present a methodology that leverages the statistical models from Section 3 to select the best-performing regimen already tested in a Phase II study for further evaluation in a Phase III trial. Finally, we evaluate the quality of the regimens we suggest for Phase II and Phase III trials against those actually selected by oncology researchers using two evaluation approaches: the *simulation metric* and the *matching metric*.

[9] The first strata is trials for which at least half of patients had received prior palliative chemotherapy. These 55 arms in the test set had an average median OS of 7.6 months. The remaining four strata all consist of trial arms with fewer than half of the patients with prior palliative chemotherapy (or that value not reported) and varying patient health levels as measured by the average ECOG performance status in the trial. We include a strata for arms with good performance status (average value 0 to 0.5; these 57 arms in the test set had an average median OS of 11.8 months), with medium performance status (average value 0.5 to 1.0; these 173 arms in the test set had an average median OS of 10.0 months), with poor performance status (average value of 1.0 or greater; these 77 arms in the test set had an average median OS of 9.0 months), and unreported performance status (these 35 arms in the test set had an average median OS of 9.3 months).

4.1. Phase II Regimen Optimization Model

Given the current data from clinical trials and the current predictive models that we have constructed, we would like to select the next best regimen to test in a Phase II clinical trial. Following the objectives for designing clinical trials laid out in Section 1, we seek to identify trials that have high efficacy, that have acceptable toxicity, and that test novel treatments.

To identify regimens with high efficacy, we include the predicted median OS of patients in the trial in the objective of our optimization model. Our reasoning for this is that for the majority of Phase III trials in our database that clearly stated a primary objective, the objective was to demonstrate improvement in the overall survival (OS) of patients in the treatment group.[10]

To limit our suggestions to regimens with acceptable toxicity, we add a constraint to our optimization model to bound the predicted proportion of patients experiencing a DLT to not exceed some constant t. No Phase III studies in our database listed a toxicity outcome as a primary objective, validating our choice to address toxicity using a constraint instead of as part of the objective. Even in cases where our model predicts that a regimen will be acceptably non-toxic, a Phase I study would still be necessary to ensure patient safety, potentially resulting in changes to the dosage levels suggested by our models.

We seek novel treatments using three approaches. First, we require that all regimens (the combination of drug and dose selections) suggested by our models have never been previously tested in a clinical trial; hence, our model always suggests novel regimens. Secondly, we require that new drugs are tested in a clinical trial as soon as they are available, ensuring we evaluate new drugs as quickly as possible. Finally, our models assign higher weight to regimens containing drugs that have not been extensively tested. Motivated by the standard deviation of the sample mean,[11] we assign weight $u_d = t_d^{-1/2}$ to drug d, where t_d is the number of times drug d has previously been tested in a clinical trial, defining $\mathbf{u}$ to be the vector of all such weights. In the model (1), we increment the objective by Γu_d if drug d is selected for testing in the regimen, where Γ is a parameter that controls the aggressiveness of the exploration.

Our mathematical model includes decision variables for the chemotherapy variables described in Section 3.1. We define three variables for each drug, corresponding to the chemotherapy treatment variables used in the statistical models: a binary indicator variable b_d to indicate whether drug d is or is not part of the trial ($b_d = 1$ if and only if drug d is part of the optimal chemotherapy

[10] Out of the 20 Phase III trials in our database with a clearly stated primary objective, 12 of them listed OS as a primary objective.

[11] Recall that $\text{sd}((\sum_{i=1}^{t} X_i)/t) = \sigma t^{-1/2}$ when X_i are IID random variables with standard deviation σ. If we view each X_i as the observed impact of some drug d on the efficacy or toxicity of a regimen that tests it, then the standard deviation of the sample mean estimate of that drug's effect scales with $t_d^{-1/2}$, where t_d is the number of times the drug has been tested.

regimen), a continuous variable i_d to indicate the instantaneous dose of drug d that should be administered in a single session, and a continuous variable a_d to indicate the average dose of drug d that should be delivered each week. We define $\mathbf{x}$ in our optimization model to be the demographic and trial information variables for which the chemotherapy regimen is being selected; these values are treated as a constant in the optimization process.

We use the ridge regression models from Section 3.2 to parameterize the optimization model. Let the model for overall survival (OS) be denoted by $\hat{\boldsymbol{\beta}}^{\boldsymbol{b}}_{\boldsymbol{OS}}{}'\mathbf{b} + \hat{\boldsymbol{\beta}}^{\boldsymbol{i}}_{\boldsymbol{OS}}{}'\mathbf{i} + \hat{\boldsymbol{\beta}}^{\boldsymbol{a}}_{\boldsymbol{OS}}{}'\mathbf{a} + \hat{\boldsymbol{\beta}}^{\boldsymbol{x}}_{\boldsymbol{OS}}{}'\mathbf{x}$, with drug variables $\mathbf{b}$, instantaneous dose variables $\mathbf{i}$, average dose variables $\mathbf{a}$, demographic and study characteristic constants $\mathbf{x}$, and coefficients corresponding to each set of variables indicated with superscripts. Similarly, we have a model for the proportion of patients with a DLT, which we will denote by $\hat{\boldsymbol{\beta}}^{\boldsymbol{b}}_{\boldsymbol{DLT}}{}'\mathbf{b} + \hat{\boldsymbol{\beta}}^{\boldsymbol{i}}_{\boldsymbol{DLT}}{}'\mathbf{i} + \hat{\boldsymbol{\beta}}^{\boldsymbol{a}}_{\boldsymbol{DLT}}{}'\mathbf{a} + \hat{\boldsymbol{\beta}}^{\boldsymbol{x}}_{\boldsymbol{DLT}}{}'\mathbf{x}$. Note that these models are all linear in the variables.

We can then select the drug therapy to test in the next clinical study using the following mixed integer optimization model:

$$
\begin{aligned}
\max_{\mathbf{b},\mathbf{i},\mathbf{a}} \quad & (\hat{\boldsymbol{\beta}}^{\boldsymbol{b}}_{\boldsymbol{OS}} + \Gamma\mathbf{u})'\mathbf{b} + \hat{\boldsymbol{\beta}}^{\boldsymbol{i}}_{\boldsymbol{OS}}{}'\mathbf{i} + \hat{\boldsymbol{\beta}}^{\boldsymbol{a}}_{\boldsymbol{OS}}{}'\mathbf{a} + \hat{\boldsymbol{\beta}}^{\boldsymbol{x}}_{\boldsymbol{OS}}{}'\mathbf{x} && && (1)\\
\text{subject to} \quad & \hat{\boldsymbol{\beta}}^{\boldsymbol{b}}_{\boldsymbol{DLT}}{}'\mathbf{b} + \hat{\boldsymbol{\beta}}^{\boldsymbol{i}}_{\boldsymbol{DLT}}{}'\mathbf{i} + \hat{\boldsymbol{\beta}}^{\boldsymbol{a}}_{\boldsymbol{DLT}}{}'\mathbf{a} + \hat{\boldsymbol{\beta}}^{\boldsymbol{x}}_{\boldsymbol{DLT}}{}'\mathbf{x} \le t, && && (1a)\\
& \sum_{d=1}^{n} b_d \le N, && && (1b)\\
& \mathbf{A}\mathbf{b} \le \mathbf{c}, && && (1c)\\
& (\mathbf{b},\mathbf{i},\mathbf{a}) \notin P, && && (1d)\\
& (b_d, i_d, a_d) \in \boldsymbol{\Omega}_d, && d = 1,\ldots,n, && (1e)\\
& b_d \in \{0,1\}, && d = 1,\ldots,n. && (1f)
\end{aligned}
$$

The objective of (1) maximizes the predicted overall survival of the selected chemotherapy regimen plus some constant Γ times u_d, the weight capturing how often drug d has previously been tested, for each drug d in the regimen. This “exploration constant” Γ controls how much weight is assigned to exploring drugs that have not been extensively tested in the training set; a large Γ would value exploration of new drugs over identifying a combination with high predicted efficacy, while $\Gamma = 0$ optimizes the efficacy of the regimen with no consideration for exploration. We experiment with a number of Γ values in Section 4.4.

Constraint (1a) bounds the predicted toxicity by a constant t. This constant value 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 toxicities. In Section 4.4, we present results from varying the toxicity limit t. Constraint (1b) limits the total number of drugs in the selected trial to N, which can be varied to select trials with different numbers of drugs. We limit suggested drug combinations to contain no more than $N = 3$ drugs, which encompasses 89.1% of our database. We chose not to select a limit of $N = 4$ or higher both because the average number of drugs tested in combinations in our database is 2.3 and because all preferred regimens in the National Comprehensive Cancer Network (NCCN) guidelines for gastric cancer contain three or fewer drugs (Ajani et al. 2014).

We also include constraints (1c) to constrain the drug combinations that can be selected. In our models, we require a new drug to be included if it has never been evaluated in a previous clinical trial and we incorporate generally accepted guidelines for selecting combination chemotherapy regimens (Page and Takimoto 2002, Pratt 1994, Golan et al. 2008).[12] 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 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) force our selected regimen to differ from the set P of all regimens previously tested in the training set. Constraints (1e) limit the instantaneous and average dose of drug d to belong to a feasible set $\mathbf{\Omega}_d$. This forces i_d and a_d to equal 0 when $b_d = 0$ and to match the instantaneous and average dosages of drug d in some clinical trial in the full database when $b_d = 1$. These constraints force the dosage for a particular drug to be realistic. Lastly, constraints (1f) define $\mathbf{b}$ to be a binary vector of decision variables.

4.2. Phase III Regimen Selection Model

Phase III trials are large randomized controlled trials that evaluate the most promising regimens tested in previous Phase II studies, and treatments that perform well against historical controls in Phase III trials may then be considered new standard therapies for advanced cancer. A relatively small number of Phase III trials are run (7% of trials in our database are Phase III), both because a Phase III trial is only run when a therapy is shown to be particularly effective in a Phase II

[12] We limit the combinations to contain no more than one drug from any drug class. There are 23 classes of drugs used in total in our database, using the classes defined by Golan et al. (2008). The most common classes are: platinum-based, antimetabolites, anthracyclines, taxanes, camptothecins, alkylating agents, and chemoprotectants. We disallow pairs of drug classes from being used together if this pairing appears no more than once in our database and is discouraged in the guidelines for selecting regimens. 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.

study and because Phase III trials enroll more patients than Phase II studies and are therefore more expensive. Both due to the impact of Phase III results on clinical standards and the relatively small numbers of these trials run, it is especially important that we select high-quality regimens to test in the experimental arms of these clinical trials.

One option for selecting a regimen to test in a Phase III trial would be to identify the prior Phase II study with DLT proportion not exceeding parameter t that achieved the highest median OS, selecting the regimen tested in that Phase II study. However, Phase II studies are often performed using patient populations that are not representative of the patients tested in Phase III trials and standard clinical practice (Friedman et al. 2010b). As a result, we instead use the statistical models from Section 3 to select the regimen to be tested in the experimental arms of Phase III clinical trials. Using the demographic and trial variables $\mathbf{x}$ from the new Phase III trial being designed, we select the regimen previously tested in a Phase II study that has the highest predicted median OS in population $\mathbf{x}$, limiting to regimens with predicted DLT proportion not exceeding a toxicity limit parameter t. Each of the prior Phase II studies is already in the training set of the statistical model being used to select Phase III regimens, but using the prediction model enables us to control for the patient population of trials and variability in trial outcomes due to chance. The experimental arms of Phase III clinical trials seek novel chemotherapy regimens, so we limit our selected regimens to those whose set of drugs do not exactly match the set of drugs tested in any arm of a prior Phase III trial.

The effects of using prediction models instead of published clinical trial outcomes are displayed in Figure 6. The figure on the left shows, for each regimen tested in a Phase II study in our database that does not match the drug combination tested in a Phase III clinical trial, the proportion of patients experiencing a DLT and the median overall survival, as they were reported in the Phase II study. The figure on the right shows the predicted performance of each drug regimen, using the average of trial and demographic variables across all Phase III trials in our database. The red points show the five best trials according to the data, and the green points show the five best trials according to the prediction models (we define the best trials here as the ones with the highest overall survival, subject to a DLT proportion of no more than 0.5). These figures show that our method will often suggest different regimens than we would select by just using the data published in the individual trials. These differences occur because our method takes into account patient and trial characteristics, controlling for trials run in particularly healthy populations (where strong efficacy outcomes may be due to demographics) and for trials with fewer enrolled patients (where strong results are more likely to be due to chance). For instance, the trials indicated in red in Figure 6 were smaller on average than the trials in green (37 vs. 51 patients), which may indicate why our models had more confidence in the quality of the trials labeled in green. We will investigate

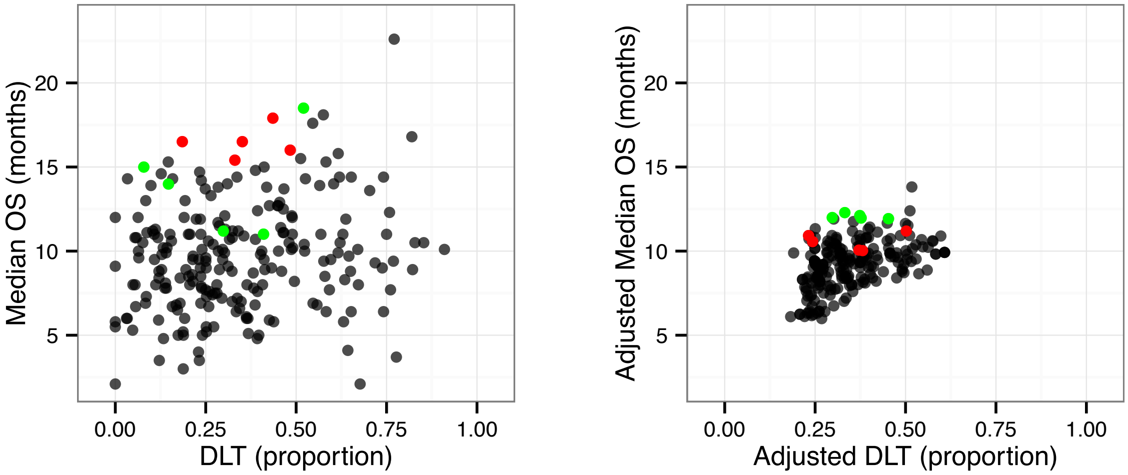


Figure 6 Median OS and DLT proportion for chemotherapy regimens that could potentially be selected for a Phase III trial experimental arm by our models. The left figure plots outcomes of the Phase II study that tested each regimen, and the right figure plots predicted outcomes for the regimen in a typical Phase III clinical trial patient population. The red points show the top five trials according to the data reported in the Phase II trials, and the green points show the top five trials according to the prediction models.

how the regimens suggested by our models compare to those selected in current clinical practice in Sections 4.3–4.4.

4.3. Evaluation Techniques

Evaluating the quality of chemotherapy regimens suggested by our optimization model against those selected for clinical trials by oncologists is an inherently difficult task. The only definitive way to evaluate a chemotherapy regimen in a given population is through a clinical trial, and the only definitive way to evaluate the performance of some regimen A suggested by our models against some regimen B designed without the models is through running a randomized controlled trial in a population, testing A against B.

While ultimately clinical trial evidence is needed to evaluate the effectiveness of the proposed models, it is important to first perform preclinical evaluations of the proposed approach to determine if it shows promise in improving the results of clinical trials compared to current practice. As described in Section 1, preclinical evaluations estimate the quality of an approach before testing it in humans and are used extensively in the design of chemotherapy regimens for advanced cancer. Preclinical evaluation cannot be used to conclusively confirm the effectiveness of a new therapy, but it can be used to rule out unpromising approaches. For instance, a new drug that has a high cell kill rate *in vitro* may or may not be effective at treating cancer in the human body, but a new drug that performs poorly in preclinical testing would likely not be tested in humans. We similarly use preclinical evaluation to determine if our proposed models show promise, which could help in deciding whether clinical evaluation is appropriate.

To perform preclinical evaluation of our proposed models, we use two different techniques to approximate the quality of suggestions from our model compared to those made in real clinical trials: the *simulation metric* and the *matching metric*.

Simulation Metric Through clinical trials, oncologists learn about the efficacy and toxicity of regimens they test and use this information when designing further chemotherapy regimens. To evaluate the ability of our approach to learn through time, we simulate the efficacy and toxicity outcomes of clinical trials that test the chemotherapy regimens proposed by our models. To simulate clinical trial outcomes, we first select plausible coefficient vectors $\boldsymbol{\beta}^*_{\boldsymbol{OS}}$ and $\boldsymbol{\beta}^*_{\boldsymbol{DLT}}$ for our survival and toxicity prediction models, which we use to represent the ground truth impact of the explanatory variables from Section 2 on efficacy and toxicity. We then simulate the median OS and the proportion of patients with a DLT in a proposed clinical trial using $\boldsymbol{\beta}^*_{\boldsymbol{OS}}$ and $\boldsymbol{\beta}^*_{\boldsymbol{DLT}}$, respectively, in both cases adding normally distributed noise with variance based on the number of patients in the clinical trial. To compare the regimens selected by our models with toxicity limit t and exploration parameter Γ against the regimens tested in clinical trials in current practice, we start 20% of the way through the clinical trial database (in 1997), selecting the regimen to be tested in each Phase II study using the optimization model from Section 4.1 and selecting the regimen to be used in each Phase III trial experimental arm using the procedure described in Section 4.2. For each trial, both the regimen selected by our models and the regimen run in current practice are evaluated using $\boldsymbol{\beta}^*_{\boldsymbol{OS}}$ and $\boldsymbol{\beta}^*_{\boldsymbol{DLT}}$. The outcomes for the regimen selected by our models, plus normally distributed noise, are added to the training set and can be used to design subsequent chemotherapy regimens. We evaluate each parameter set (t, Γ) using 40 different sets of coefficients $\boldsymbol{\beta}^*_{\boldsymbol{OS}}$ and $\boldsymbol{\beta}^*_{\boldsymbol{DLT}}$. Each set of coefficients is obtained by drawing a bootstrap sample of the entire database of clinical trials and training ridge regression models for the efficacy and toxicity outcomes as described in Section 3; because these coefficients are obtained using bootstrap resampling of true clinical trial outcomes, they are plausible according to clinical trial data. Details of the simulation metric procedure are provided in Appendix B.

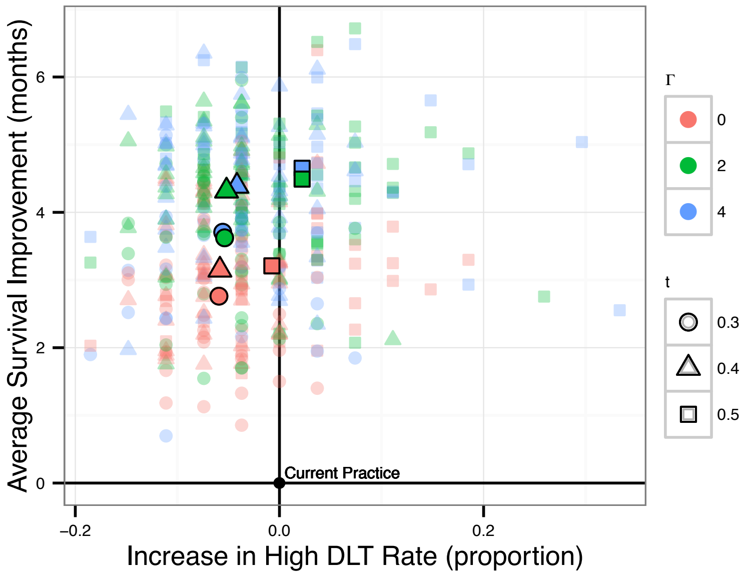
As with many procedures developed to simulate complicated real-world systems, the simulation metric may make a biased evaluation of our proposed model. First, the approach simulates outcomes using the same linear model specification used by the ridge regression model from Section 3, which mean that the statistical models used to make decisions are assumed to be structurally accurate. Model performance might be worse if there were a mismatch between the structure of the ridge regression model and the true structure of the relationship between the covariates and outcomes.

Further, while we include a number of constraints in the optimization models to prevent infeasible regimens from being suggested, there there is no guarantee that all chemotherapy regimens in the feasible set of the optimization model are indeed biologically, legally, or practically feasible. This could lead the optimization model to obtain a strong evaluation for a suggestion that is actually infeasible, which would favorably bias the evaluation of our approach. Due to the potential biases in the simulation metric, results indicating our models improve over current practice might merit further study in a clinical trial setting, while results indicating our models do not improve over current practice would suggest that no further evaluation is warranted.

Matching Metric While the simulation metric enables us to evaluate our ability to learn through time by providing feedback about the chemotherapy regimens selected by our optimization model, a shortcoming of this technique is that it relies on simulated outcomes instead of actual clinical trial outcomes, introducing a number of potential biases. As a result, we also evaluate our model's suggested regimens for Phase III trial experimental arms using the results of similar clinical trials that were run in practice. To compare the regimens selected by our models with toxicity limit t against the regimens tested in clinical trials in current practice, we start 20% of the way through the clinical trial database (in 1997), selecting the regimen to be used in each Phase III trial experimental arm using the procedure described in Section 4.2. For each Phase III experimental arm, the regimen selected by our models is evaluated using the clinical trial in the database testing the most similar chemotherapy regimen, taking into account how well the drug classes, drugs, and dosages match[13] and limiting to chronologically future clinical trials. Meanwhile, the regimens tested in the actual Phase III experimental arms are evaluated using the outcomes of those trials. Details of the matching metric procedure are provided in Appendix B.

A key benefit of the matching metric as defined is that it does not use the statistical models developed in this chapter in any way to evaluate proposed regimens, not even to adjust for the population in the matched clinical trial. As a result, the matching metric may make a biased evaluation of proposed chemotherapy regimens because it evaluates a chemotherapy regimen using the outcomes of a trial run in a different population. If the matched trial used to evaluate a proposed chemotherapy regimen is run in a healthier population than the population for which the regimen was selected, favorable bias would be introduced to the evaluation. Correspondingly if the patients in the matched trial are less healthy, unfavorable bias is introduced. Due to the potential bias in the matching metric's evaluation of our model, any improvements over current practice indicated by the matching metric would require confirmatory testing in a clinical trial setting.

[13] For each drug in our suggested regimen, we assess a penalty of 0 if the same drug is tested at the same dosage in the future regimen, a penalty of 1 if the same drug is tested at a different dosage, a penalty of 10 if a different drug from the same drug class is tested, and a penalty of 100 if no drugs from the same drug class are tested in the future regimen. The penalties are summed for each drug in our suggested regimen, and the future clinical trial with the smallest penalty score is considered the best match.



Γ	t	Δ OS (95% CI)	Δ DLT (95% CI)
0	0.3	2.8 [0.9, 5.0]	-0.06 [-0.15, 0.04]
0	0.4	3.1 [1.8, 4.9]	-0.06 [-0.15, 0.04]
0	0.5	3.2 [1.6, 6.3]	-0.01 [-0.18, 0.18]
2	0.3	3.6 [1.6, 5.9]	-0.05 [-0.15, 0.07]
2	0.4	4.3 [1.8, 5.6]	-0.05 [-0.15, 0.11]
2	0.5	4.5 [2.1, 6.7]	0.02 [-0.18, 0.26]
4	0.3	3.7 [0.7, 6.0]	-0.06 [-0.18, 0.07]
4	0.4	4.4 [2.0, 6.3]	-0.04 [-0.15, 0.07]
4	0.5	4.7 [2.6, 6.5]	0.02 [-0.18, 0.33]

Figure 7 **[Left] Comparison between Phase III experimental arm suggestions (n=27) from our models and from current practice according to the simulation metric. Each point represents a simulation of our approach and highlighted points are averages across simulations. Points to the left of the y-axis improved over current practice in the proportion of trials with unacceptably high toxicity, and points above the x-axis improved over current practice in average median OS. [Right] For each optimization model with parameters Γ and t, the change in average median OS and proportion of trial arms with unacceptably high DLT compared to current practice.**

4.4. Optimization Results

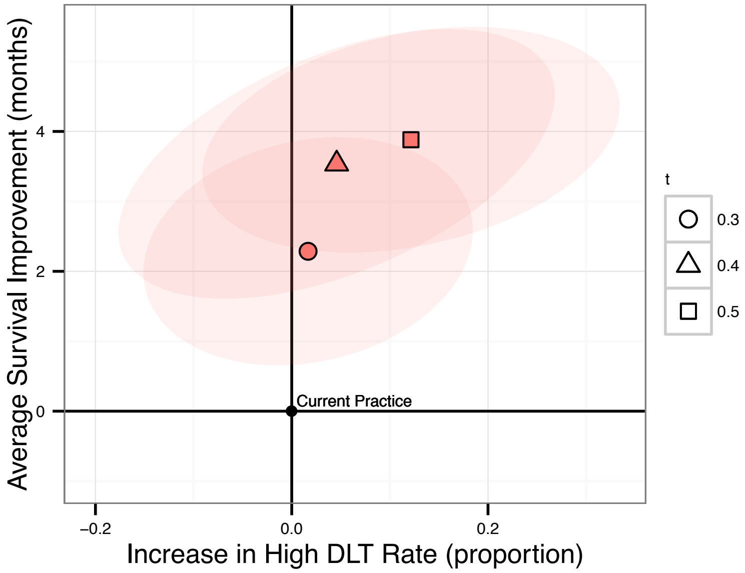
To compare our models' suggested regimens with those of oncologists, we sequentially computed the simulation metric for 40 sets of simulated coefficients $(\boldsymbol{\beta}^*_{\boldsymbol{OS}}, \boldsymbol{\beta}^*_{\boldsymbol{DLT}})$, testing all nine combinations of parameter settings $t \in \{0.3, 0.4, 0.5\}$ and $\Gamma \in \{0, 2, 4\}$. As detailed in Appendix B, for each set of parameter values (t, Γ), we obtain 40 vectors of overall survival values for the 27 Phase III experimental arms for both current practice $(\mathbf{y}^{\mathbf{CP}}_{\mathbf{OS}})$ and for the proposed models $(\mathbf{y}^{\mathbf{Mod}}_{\mathbf{OS}})$. Additionally, we obtain 40 vectors of "high toxicity" indicator values for the Phase III experimental arms both for current practice $(\mathbf{y}^{\mathbf{CP}}_{\mathbf{DLT}})$ and for the proposed models $(\mathbf{y}^{\mathbf{Mod}}_{\mathbf{DLT}})$.

Figure 7 compares the performance of the regimens selected by this paper's models for the 27 Phase III experimental arms against the 27 Phase III experimental arms selected by oncologists, reporting the average differences in outcomes across the 27 arms, $\mathbf{1}'(\mathbf{y}^{\mathbf{Mod}}_{\mathbf{OS}} - \mathbf{y}^{\mathbf{CP}}_{\mathbf{OS}})/27$ and $\mathbf{1}'(\mathbf{y}^{\mathbf{Mod}}_{\mathbf{DLT}} - \mathbf{y}^{\mathbf{CP}}_{\mathbf{DLT}})/27$. The figure plots each individual bootstrap replicate and the average comparative performance across replicates, and the accompanying table additionally reports 95% bootstrap percentile confidence intervals (Davison and Hinkley 1997) of comparative performance, computed using the R boot package (Canty and Ripley 2014). In all 360 simulation runs, the average simulated median OS of the regimens suggested by the proposed models was higher then the average simulated median OS of the regimens tested in current practice, with the average differences ranging from 2.8 months (95% CI 0.9, 5.0) for the model with $(t, \Gamma) = (0.3, 0)$ to 4.7 months (95% CI 2.6, 6.5) for the model with $(t, \Gamma) = (0.5, 4)$. The simulated proportion of trials with a high DLT rate did not significantly differ from current practice for any model. Although on average models with high Γ values have better simulated survival outcomes and models with restrictive toxicity limits have worse simulated survival outcomes and better simulated toxicity outcomes, the bootstrap 95% confidence intervals for all these differences include 0.

The regimens selected by our proposed models for Phase II studies and Phase III trial experimental arms are qualitatively different from the ones tested in current practice, which may explain some of the differences in simulated performance. Due to the nature of the formulation of optimization problem (1), nearly all regimens selected by the proposed models for Phase II studies (100%) and Phase III experimental arms (98%) contain exactly three drugs. While this is similar to the average number of drugs tested in Phase II study arms (2.3) and Phase III trial experimental arms (2.4), these studies test many other combination sizes (only 31% of Phase II study arms and 37% of Phase III trial experimental arms test three-drug combinations). Given that there are benefits to regimens with fewer drugs (e.g. ease of administration and lower cost) and regimens with more drugs (e.g. better combatting drug resistance in the cancer), oncology researchers may prefer to test a range of regimen sizes. For Phase III trial experimental arms, the proposed models selected regimens with newer drugs (newest drug tested in a median of 5 previous trials) than the regimens tested in practice (newest drug tested in a median of 22 previous trials). Given the significant cost of Phase III trials, this could represent risk aversion on the part of clinical trial planners that is not captured in the procedure from Section 4.2. By design 100% of regimens suggested by our models for Phase III experimental arms had never been tested before (even in different dosages), similar to the 89% rate seen in clinical practice. In contrast, the proportion of suggested regimens testing new combinations (ignoring dosages) for Phase II studies was 14% in the proposed models with $\Gamma = 0$, 25% in the proposed models with $\Gamma = 2$, 37% in the proposed models with $\Gamma = 4$, and 41% in current practice. This suggests the proposed models with $\Gamma = 0$ and $\Gamma = 2$ may spend more effort optimizing dosages within a combination and less effort exploring new combinations compared to clinical practice.

To further compare our models' suggested regimens for Phase III experimental arms with those of oncologists, we used the matching metric to evaluate our suggested regimens obtained using each toxicity limit $t \in \{0.3, 0.4, 0.5\}$. As detailed in Appendix B, for each parameter value t we obtain overall survival values for the 27 Phase III experimental arms for both current practice ($\mathbf{y}_{\mathbf{OS}}^{\mathbf{CP}}$) and for the proposed models ($\mathbf{y}_{\mathbf{OS}}^{\mathbf{Mod}}$) as well as "high toxicity" indicator values for the Phase III experimental arms both for current practice ($\mathbf{y}_{\mathbf{DLT}}^{\mathbf{CP}}$) and for the proposed models ($\mathbf{y}_{\mathbf{DLT}}^{\mathbf{Mod}}$).

Using the matching metric, Figure 8 compares the performance of the suggested regimens for Phase III experimental arms from the three models with $t \in \{0.3, 0.4, 0.5\}$ against the regimens



t	Match Drugs	Match Classes	Δ OS	Δ DLT
0.3	56%	82%	2.3	0.02
0.4	59%	85%	3.5	0.05
0.5	59%	89%	3.9	0.12

Figure 8 [Left] Comparison between Phase III experimental arm suggestions ($n = 27$) from our models and from current practice according to the matching metric. [Right] For each optimization model with parameter t, the proportion of model suggestions that match all drugs and all drug classes of the matched future trial arm, as well as the average change in median OS and proportion of trial arms with unacceptably high DLT compared to current practice.

selected by oncologists. The plot uses the same axes as Figure 7 and captures statistical fluctuations with 2-standard deviation ellipses fitted by bootstrap resampling the matching metric results for the 27 Phase III experimental arms. According to the matching metric, the regimens suggested by the proposed models with $t = 0.3, 0.4$, and 0.5 had on average 2.3, 3.5, and 3.9 months higher median OS than current practice and 0.02, 0.05, and 0.12 proportion higher rate of trials with unacceptably high toxicity, respectively. The ellipses for all three models include only positive changes in survival but positive and negative changes in proportion of trials with unacceptably high toxicity.

5. Discussion and Future Work

In this work, we built a database of clinical trials for gastric cancer, built statistical models to predict out-of-sample efficacy and toxicity of clinical trials, and designed models that propose combination chemotherapy regimens to be tested in clinical trials. Out-of-sample evaluation suggests that the statistical models could be used by clinical trial planners to identify 10–20% of the trials with high toxicity or that fail to achieve high efficacy, in both cases with a small number of misclassifications. Further, two preclinical evaluation techniques indicate that the models presented in this work might improve the efficacy of the regimens selected for testing in Phase III clinical trial experimental arms without major changes in toxicity outcomes.

A key limitation of the results presented in this paper is that we do not run clinical trials to evaluate our suggested chemotherapy regimens; instead, we evaluate our suggestions using simulated results or the results of clinical trials testing similar chemotherapy regimens. As detailed in Section 4.3, such preclinical evaluation of proposed approaches is important, as it can be used to eliminate unpromising approaches without needing to run a clinical trial. Given that both the simulation and matching metrics indicated that the proposed approach could improve the efficacy of the regimens tested in Phase III trials, we believe it merits further evaluation in a clinical trial setting. We believe the technique should first be evaluated using a single-arm clinical trial testing a regimen designed by our optimization models from Section 4.1. Key outcomes of this clinical trial would be the acceptability of our tools to clinical trial decision makers and the efficacy and toxicity outcomes of the trial. If this initial evaluation is deemed a success, we could then perform a randomized controlled trial comparing a regimen designed by our approach against a regimen designed using standard clinical trial design methodology.

We believe the models presented in this work have the potential to significantly improve the quality of chemotherapy regimens tested in clinical trials. However, there are a number of promising future directions that have the potential to strengthen the results presented in this work. One opportunity would be to integrate preclinical models, such as *in vitro* experimentation and molecular simulation studies, into our optimization framework. This integration could take the form of adding constraints to eliminate drug combinations found to be biologically infeasible and adding interaction terms between pairs of drugs found to be synergistic or antagonistic in preclinical models. Another avenue of future work is to use more sophisticated techniques such as the Knowledge Gradient algorithm (Frazier et al. 2008) or Q-learning (Dearden et al. 1998) when exploring the space of combination chemotherapy regimens using Phase II studies.

While in this work we focused on designing combination chemotherapy regimens for gastric cancer, we believe data-driven tools leveraging databases of clinical trial results could prove useful in other settings. Combination therapy is used to treat many other cancers as well as other diseases such as hypertension and diabetes, so our models could be applied to design therapies for these diseases. Models trained on a subset of clinical trial results for a specific patient subpopulation, such as HER2-positive patients, could be used to design specialized regimens for these subgroups. Other applications might include improving clinical prognostic models for individual cancer patients, identifying the best available therapies for a particular disease from amongst all those tested in clinical trials, and comparing treatments that have never been compared in a randomized setting.

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. The authors also thank department editor Noah Gans as well as the anonymous associate editor and referees for their helpful feedback that led to significant improvements in the work.

Appendix A: Data Preprocessing

We performed a series of data preprocessing steps to standardize the data collected from clinical trials. Many of these steps involved imputation of some value y from a closely related value x. In all such cases, we consider linear imputation via linear regression equation $y = \beta_0 + \beta_1 x + \epsilon$ and quadratic imputation via linear regression equation $y = \beta_0 + \beta_1 x^2 + \epsilon$, selecting the model that obtains the lowest sum of squared residuals.

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 495 treatment arms evaluated in this work, 423 (85.5%) reported performance status with the ECOG scale, 71 (14.3%) reported with the KPS scale, and 1 (0.2%) did not report performance status. We include mean ECOG score as a variable in our prediction models.

In 94 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 $n = 302$ trials with full ECOG breakdown and nonzero $p_0 + p_1$ and fitting a linear and quadratic imputation models to estimate $p_0/(p_0 + p_1)$ from $p_0 + p_1$; the quadratic model was selected because it had the lowest sum of squared residuals and achieved an R^2 value of 0.264. For the 18 treatment arms with the proportion of patients with each KPS score reported, we perform a conversion from the KPS scale to the ECOG scale based on data in Buccheri et al. (1996). For treatment arms reporting performance status with other data groupings, the combined score is marked as unavailable. In total, 100 trial arms (20.2%) 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. 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, thrombocytopenia, anemia, and lymphopenia, we built linear and quadratic imputation models to predict the Grade 4 toxicity from the Grade 3/4 toxicity, training on arms reporting both values. The models were trained using 217, 311, 254, and 8 treatment arms, respectively. The quadratic imputation model was most effective in predicting grade 4 neutropenia and the linear models were the most effective for the remaining three imputations, with the models obtaining R^2 values of 0.881, 0.669, 0.277, and 0.035, respectively. The models were used to impute the proportion of patients with a Grade 4 toxicity in 119, 93, 101, and 3 treatment arms, respectively.

Two of the most common blood toxicities, leukopenia and neutropenia, are often not both reported 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 quadratic and linear imputation models using the proportion of patients experiencing Grade 3/4 leukopenia to predict the proportion of patients experiencing Grade 4 neutropenia, training on the 142 arms that reported both proportions. The linear model was the most effective with $R^2 = 0.752$, and we used that model to convert data from 99 treatment arms that reported leukopenia toxicity data but not neutropenia. Overall, we used some form of imputation to compute Grade 4 blood toxicities in 230 treatment arms (46.5%).

As a sensitivity analysis to evaluate the effects of imputing these dependent variables, we sequentially evaluated ridge regression models limited to the 265 treatment arms for which no imputation was performed on the dependent variables. The ridge regression model predicting the proportion of patients with a DLT had a test-set sequential AUC of 0.817 (bootstrap 95% CI [0.757,0.866]) on the last four years of the limited dataset and did not significantly differ from the AUC of 0.827 (bootstrap 95% CI [0.770,0.846]) on the last four years of the full dataset.

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, using the 20 broad anatomical/pathophysiological categories defined by the NCI-CTCAE v3 toxicity reporting criteria (National Cancer Institute 2006). Assign each toxicity group 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 as in the Grouped Independent Approach, compute a toxicity score for the trial as the sum of the group scores.

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

(8.1%) trial arms report this value, though we can only compute the combined metric for 36 of them 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	0.855
Sum	0.820

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

As reported in Table 4, 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 104 (21.0%) of trial arms in the database.

Appendix B: Pseudocode of Evaluation Techniques for Proposed Regimens

Figure 9 provides the full pseudocode for the simulation metric evaluation technique, and Figure 10 provides the full pseudocode for the matching metric evaluation technique.

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.

Ajani, Jaffer A., Thomas A. D'Amico, Khaldoun Almhanna, et al. 2014. NCCN guidelines for gastric cancer, version 1.2014. `http://www.nccn.org/professionals/physician_gls/pdf/gastric.pdf`.

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.

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.

```
Input: Clinical trial database $\mathbf{X}$ (rows in chronological order) containing $q$ arms (the first $r := \lceil q/5 \rceil - 1$ will be used as a training set), corresponding trial weight vector $\mathbf{w}$, and corresponding Phase III study indicator $\mathbf{p^{III}}$. Vectors of binary, instantaneous, and average dose variables and demographic/study characteristic variables for trial $k$ are indicated by $\mathbf{b^k}, \mathbf{i^k}, \mathbf{a^k}$, and $\mathbf{x^k}$, respectively. Further input: outcome vectors $\mathbf{y_{OS}}$ and $\mathbf{y_{DLT}}$ and parameters $t$ and $\Gamma$.
$\mathbf{X^{boot}}, \mathbf{y^{boot}_{OS}}, \mathbf{y^{boot}_{DLT}}, \mathbf{w^{boot}} \leftarrow$ bootstrap resampled versions of the $q$ arms in $\mathbf{X}$, $\mathbf{y_{OS}}$, $\mathbf{y_{DLT}}$, and $\mathbf{w}$
$\boldsymbol{\beta^*_{OS}}, \boldsymbol{\beta^*_{DLT}}, \mathbf{r_{OS}}, \mathbf{r_{DLT}} \leftarrow$ coefficients and residuals of ridge regression models with parameters selected via cross-validation, trained with $\mathbf{X^{boot}}$, $\mathbf{y^{boot}_{OS}}$, $\mathbf{y^{boot}_{DLT}}$, and $\mathbf{w^{boot}}$
$\sigma^2_{OS} \leftarrow \mathbf{r'_{OS}} \mathrm{diag}(\mathbf{w^{boot}}) \mathbf{r_{OS}} / (q - f - 1)$, with ridge regression degrees of freedom $f$ (Hastie et al. 2009)
$\sigma^2_{DLT} \leftarrow \mathbf{r'_{DLT}} \mathrm{diag}(\mathbf{w^{boot}}) \mathbf{r_{DLT}} / (q - f - 1)$
$\mathbf{P} \leftarrow \bigcup_{k=1}^{r} \{(\mathbf{b^k}, \mathbf{i^k}, \mathbf{a^k})\}$, $\mathbf{P^{III}} \leftarrow \{\mathbf{b^k} \mid 1 \le k \le r, p_k^{III} = 1\}$
$\mathbf{X^{train}} \leftarrow \mathbf{X_{\{1,\dots,r\}}}$, $\mathbf{y^{train}_{OS}} \leftarrow \mathbf{y_{OS,\{1,\dots,r\}}}$, $\mathbf{y^{train}_{DLT}} \leftarrow \mathbf{y_{DLT,\{1,\dots,r\}}}$, $\mathbf{w^{train}} \leftarrow \mathbf{w_{\{1,\dots,r\}}}$
$\mathbf{y^{CP}_{OS}} \leftarrow [\,], \mathbf{y^{Mod}_{OS}} \leftarrow [\,], \mathbf{y^{CP}_{DLT}} \leftarrow [\,], \mathbf{y^{Mod}_{DLT}} \leftarrow [\,]$
for $k = r + 1$ to $q$ do
    $\hat{\boldsymbol{\beta}}_{\boldsymbol{OS}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{DLT}} \leftarrow$ coefficients of ridge regression models trained using $\mathbf{X^{train}}$, $\mathbf{y^{train}_{OS}}$, $\mathbf{y^{train}_{DLT}}$, and $\mathbf{w^{train}}$
    if Trial $k$ is a control arm of a Phase III trial then
        $\mathbf{b}^* \leftarrow \mathbf{b^k}$, $\mathbf{i}^* \leftarrow \mathbf{i^k}$, $\mathbf{a}^* \leftarrow \mathbf{a^k}$, $\mathbf{P^{III}} \leftarrow \mathbf{P^{III}} \bigcup \{\mathbf{b^k}\}$
    else if Trial $k$ is an experimental arm of a Phase III trial then
        $\tilde{\mathbf{P}} \leftarrow \{(\mathbf{b}, \mathbf{i}, \mathbf{a}) \in \mathbf{P} \setminus \mathbf{P^{III}} \mid {\hat{\boldsymbol{\beta}}^{\boldsymbol{b}}_{\boldsymbol{DLT}}}' \mathbf{b} + {\hat{\boldsymbol{\beta}}^{\boldsymbol{i}}_{\boldsymbol{DLT}}}' \mathbf{i} + {\hat{\boldsymbol{\beta}}^{\boldsymbol{a}}_{\boldsymbol{DLT}}}' \mathbf{a} + {\hat{\boldsymbol{\beta}}^{\boldsymbol{x}}_{\boldsymbol{DLT}}}' \mathbf{x^k} \le t\}$
        $\mathbf{b}^*, \mathbf{i}^*, \mathbf{a}^* \leftarrow \arg\max_{\mathbf{b}, \mathbf{i}, \mathbf{a} \in \tilde{\mathbf{P}}} \ {\hat{\boldsymbol{\beta}}^{\boldsymbol{b}}_{\boldsymbol{OS}}}' \mathbf{b} + {\hat{\boldsymbol{\beta}}^{\boldsymbol{i}}_{\boldsymbol{OS}}}' \mathbf{i} + {\hat{\boldsymbol{\beta}}^{\boldsymbol{a}}_{\boldsymbol{OS}}}' \mathbf{a} + {\hat{\boldsymbol{\beta}}^{\boldsymbol{x}}_{\boldsymbol{OS}}}' \mathbf{x^k}$, with ties broken randomly
        $\mathbf{P^{III}} \leftarrow \mathbf{P^{III}} \bigcup \{\mathbf{b}^*\}$
    else if Trial $k$ is an arm of a Phase II study then
        Solve optimization model (1) using coefficients $\hat{\boldsymbol{\beta}}_{\boldsymbol{OS}}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{DLT}}$, vector $\mathbf{u}$ computed using $\mathbf{X^{train}}$, and patient demographic variables $\mathbf{x^k}$, obtaining optimal binary, instantaneous, and average dosage variable values $\mathbf{b}^*, \mathbf{i}^*$, and $\mathbf{a}^*$, respectively. Use parameter $\Gamma$ in the objective, $t$ in constraint (1a), $N = 3$ in constraint (1b), set $\mathbf{P}$ in constraint (1d), and sets $\boldsymbol{\Omega_d}$ derived from $\mathbf{X}$ for constraints (1e).
    end if
    $\mathbf{P} \leftarrow \mathbf{P} \bigcup \{(\mathbf{b}^*, \mathbf{i}^*, \mathbf{a}^*)\}$
    Append $\mathbf{X^{train}}$ with a row derived from $\mathbf{b}^*, \mathbf{i}^*, \mathbf{a}^*$, and $\mathbf{x^k}$, and append $\mathbf{w^{train}}$ with $w_k$
    Append $\mathbf{y^{train}_{OS}}$ with a sample from $\mathcal{N}({\boldsymbol{\beta}^{*b}_{\boldsymbol{OS}}}' \mathbf{b}^* + {\boldsymbol{\beta}^{*i}_{\boldsymbol{OS}}}' \mathbf{i}^* + {\boldsymbol{\beta}^{*a}_{\boldsymbol{OS}}}' \mathbf{a}^* + {\boldsymbol{\beta}^{*x}_{\boldsymbol{OS}}}' \mathbf{x^k}, \sigma^2_{OS}/w_k)$
    Append $\mathbf{y^{train}_{DLT}}$ with a sample from $\mathcal{N}({\boldsymbol{\beta}^{*b}_{\boldsymbol{DLT}}}' \mathbf{b}^* + {\boldsymbol{\beta}^{*i}_{\boldsymbol{DLT}}}' \mathbf{i}^* + {\boldsymbol{\beta}^{*a}_{\boldsymbol{DLT}}}' \mathbf{a}^* + {\boldsymbol{\beta}^{*x}_{\boldsymbol{DLT}}}' \mathbf{x^k}, \sigma^2_{DLT}/w_k)$
    if Trial $k$ is the experimental arm of a Phase III trial then
        Append $\mathbf{y^{CP}_{OS}}$ with ${\boldsymbol{\beta}^{*b}_{\boldsymbol{OS}}}' \mathbf{b^k} + {\boldsymbol{\beta}^{*i}_{\boldsymbol{OS}}}' \mathbf{i^k} + {\boldsymbol{\beta}^{*a}_{\boldsymbol{OS}}}' \mathbf{a^k} + {\boldsymbol{\beta}^{*x}_{\boldsymbol{OS}}}' \mathbf{x^k}$
        Append $\mathbf{y^{Mod}_{OS}}$ with ${\boldsymbol{\beta}^{*b}_{\boldsymbol{OS}}}' \mathbf{b}^* + {\boldsymbol{\beta}^{*i}_{\boldsymbol{OS}}}' \mathbf{i}^* + {\boldsymbol{\beta}^{*a}_{\boldsymbol{OS}}}' \mathbf{a}^* + {\boldsymbol{\beta}^{*x}_{\boldsymbol{OS}}}' \mathbf{x^k}$
        Append $\mathbf{y^{CP}_{DLT}}$ with $\mathbb{1}_{{\boldsymbol{\beta}^{*b}_{\boldsymbol{DLT}}}' \mathbf{b^k} + {\boldsymbol{\beta}^{*i}_{\boldsymbol{DLT}}}' \mathbf{i^k} + {\boldsymbol{\beta}^{*a}_{\boldsymbol{DLT}}}' \mathbf{a^k} + {\boldsymbol{\beta}^{*x}_{\boldsymbol{DLT}}}' \mathbf{x^k} \ge 0.5}$
        Append $\mathbf{y^{Mod}_{DLT}}$ with $\mathbb{1}_{{\boldsymbol{\beta}^{*b}_{\boldsymbol{DLT}}}' \mathbf{b}^* + {\boldsymbol{\beta}^{*i}_{\boldsymbol{DLT}}}' \mathbf{i}^* + {\boldsymbol{\beta}^{*a}_{\boldsymbol{DLT}}}' \mathbf{a}^* + {\boldsymbol{\beta}^{*x}_{\boldsymbol{DLT}}}' \mathbf{x^k} \ge 0.5}$
    end if
end for
Output: $\mathbf{y^{CP}_{OS}}$, $\mathbf{y^{Mod}_{OS}}$, $\mathbf{y^{CP}_{DLT}}$, and $\mathbf{y^{Mod}_{DLT}}$
```

Figure 9 Pseudocode of the simulation metric procedure.

```
**Input:** Clinical trial database $\mathbf{X}$ (rows in chronological order) containing $q$ arms (the first $r := \lceil q/5 \rceil - 1$ will be used as a training set), corresponding trial weight vector $\mathbf{w}$ from Section 3.2, and corresponding Phase III study indicator $\mathbf{p^{III}}$. Vectors of binary, instantaneous, and average dose variables and demographic/study characteristic variables for trial $k$ are indicated by $\mathbf{b^k}, \mathbf{i^k}, \mathbf{a^k}$, and $\mathbf{x^k}$, respectively. The vector of drug class indicators (Golan et al. 2008) for trial $k$ is indicated by $\mathbf{c^k}$. Further input includes outcome vectors $\mathbf{y_{OS}}$ and $\mathbf{y_{DLT}}$ and parameter $t$.
$\mathbf{P^{III}} \leftarrow \{\mathbf{b^k} \mid 1 \le k \le r, p_k^{III} = 1\}$
$\mathbf{y_{OS}^{CP}} \leftarrow [\,], \mathbf{y_{OS}^{Mod}} \leftarrow [\,], \mathbf{y_{DLT}^{CP}} \leftarrow [\,], \mathbf{y_{DLT}^{Mod}} \leftarrow [\,]$
**for** $k = r + 1$ to $q$ **do**
  **if** Trial $k$ is a control arm of a Phase III trial **then**
    $\mathbf{P^{III}} \leftarrow \mathbf{P^{III}} \bigcup \{\mathbf{b^k}\}$
  **else if** Trial $k$ is an experimental arm of a Phase III trial **then**
    $\mathbf{X^{train}} \leftarrow \mathbf{X_{\{1,\ldots,k-1\}}}$, $\mathbf{y_{OS}^{train}} \leftarrow \mathbf{y_{OS,\{1,\ldots,k-1\}}}$, $\mathbf{y_{DLT}^{train}} \leftarrow \mathbf{y_{DLT,\{1,\ldots,k-1\}}}$, $\mathbf{w^{train}} \leftarrow \mathbf{w_{\{1,\ldots,k-1\}}}$
    $\hat{\boldsymbol{\beta}}_{\boldsymbol{OS}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{DLT}} \leftarrow$ coefficients of ridge regression models with parameters selected via cross-validation (see Section 3), trained using $\mathbf{X^{train}}$, $\mathbf{y_{OS}^{train}}$, $\mathbf{y_{DLT}^{train}}$, and $\mathbf{w^{train}}$
    $\mathbf{P} \leftarrow \bigcup_{j=1}^{k-1} \{(\mathbf{b^j}, \mathbf{i^j}, \mathbf{a^j})\}$
    $\tilde{\mathbf{P}} \leftarrow \{(\mathbf{b}, \mathbf{i}, \mathbf{a}) \in \mathbf{P} \setminus \mathbf{P^{III}} \mid \hat{\boldsymbol{\beta}}_{\boldsymbol{DLT}}^{\boldsymbol{b}\,\prime} \mathbf{b} + \hat{\boldsymbol{\beta}}_{\boldsymbol{DLT}}^{\boldsymbol{i}\,\prime} \mathbf{i} + \hat{\boldsymbol{\beta}}_{\boldsymbol{DLT}}^{\boldsymbol{a}\,\prime} \mathbf{a} + \hat{\boldsymbol{\beta}}_{\boldsymbol{DLT}}^{\boldsymbol{x}\,\prime} \mathbf{x^k} \le t\}$
    $\mathbf{b}^*, \mathbf{i}^*, \mathbf{a}^* \leftarrow \arg\max_{\mathbf{b,i,a} \in \tilde{\mathbf{P}}} \ \hat{\boldsymbol{\beta}}_{\boldsymbol{OS}}^{\boldsymbol{b}\,\prime} \mathbf{b} + \hat{\boldsymbol{\beta}}_{\boldsymbol{OS}}^{\boldsymbol{i}\,\prime} \mathbf{i} + \hat{\boldsymbol{\beta}}_{\boldsymbol{OS}}^{\boldsymbol{a}\,\prime} \mathbf{a} + \hat{\boldsymbol{\beta}}_{\boldsymbol{OS}}^{\boldsymbol{x}\,\prime} \mathbf{x^k}$, with ties broken randomly
    $\mathbf{P^{III}} \leftarrow \mathbf{P^{III}} \bigcup \{\mathbf{b}^*\}$
    $\mathbf{c}^* \leftarrow$ vector of drug class indicators (Golan et al. 2008) from $\mathbf{b}^*$
    $\mathbf{F} \leftarrow \arg\max_{k \le j \le q} 90\mathbf{c}^{*\prime}\mathbf{c^j} + 9\mathbf{b}^{*\prime}\mathbf{b^j} + \mathbf{b}^{*\prime}[\mathbb{1}_{\mathbf{a_1^* = a_1^j} \text{ and } \mathbf{i_1^* = i_1^j}} \ \cdots \ \mathbb{1}_{\mathbf{a_n^* = a_n^j} \text{ and } \mathbf{i_n^* = i_n^j}}]$
    Append $\mathbf{y_{OS}^{CP}}$ with $y_{OS,k}$ and append $\mathbf{y_{DLT}^{CP}}$ with $\mathbb{1}_{y_{DLT,k} \ge 0.5}$
    Append $\mathbf{y_{OS}^{Mod}}$ with $\sum_{j \in \mathbf{F}} y_{OS,j}/|\mathbf{F}|$ and append $\mathbf{y_{DLT}^{Mod}}$ with $\sum_{j \in \mathbf{F}} \mathbb{1}_{y_{DLT,j} \ge 0.5}/|\mathbf{F}|$
  **end if**
**end for**
**Output:** $\mathbf{y_{OS}^{CP}}$, $\mathbf{y_{OS}^{Mod}}$, $\mathbf{y_{DLT}^{CP}}$, and $\mathbf{y_{DLT}^{Mod}}$
```

Figure 10 Pseudocode of the matching metric procedure.

Canty, Angelo, Brian Ripley. 2014. *boot: Bootstrap R (S-Plus) Functions*. Package version 1.3-13.

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.

Davison, A. C., D. V. Hinkley. 1997. *Bootstrap Methods and Their Applications*. Cambridge University Press, Cambridge.

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.

Dearden, Richard, Nir Friedman, Stuart Russell. 1998. Bayesian Q-learning. *AAAI-98 Proceedings* 761–768.

Delbaldo, Catherine, Stefan Michiels, Nathalie Syz, et al. 2004. Benefits of adding a drug to a single-agent or a 2-agent chemotherapy regimen in advanced non-small-cell lung cancer: A meta-analysis. *Journal of the American Medical Association* **292**(4) 470–484.

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.

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

Frazier, Peter I., Warren B. Powell, Savas Dayanik. 2008. A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization* **47**(5) 2410–2439.

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

Friedman, Lawrence M., Curt D. Furberg, David L. DeMets. 2010b. *Fundamentals of Clinical Trials*, vol. 4. Springer.

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.

Hastie, Trevor, Robert Tibshirani, Jerome Friedman. 2009. *The Elements of Statistical Learning (2nd edition)*. Springer-Verlag.

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.

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* .

Kleiman, Marina, Yael Sagi, Naamah Bloch, Zvia Agur. 2009. Use of virtual patient populations for rescuing discontinued drug candidates and for reducing the number of patients in clinical trials. *ATLA* **37**(Supplement 1) 39–45.

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.

Meyer, David, Evgenia Dimitriadou, Kurt Hornik, et al. 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.

Overmoyer, Beth. 2003. Combination chemotherapy for metastatic breast cancer: Reaching for the cure. *Journal of Clinical Oncology* **21**(4) 580–582.

Page, Ray, Chris Takimoto. 2002. *Cancer Management: A Multidisciplinary Approach: Medical, Surgical and Radiation Oncology*, chap. 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.

Roth, A. D. 2003. Chemotherapy in gastric cancer: a never ending saga. *Annals of Oncology* **14** 175–177.

Sertkaya, Aylin, Anna Birkenbach, Ayesha Berlind, et al. 2014. Examination of clinical trial costs and barriers for drug development. Tech. Rep. HHSP23337007T, U.S. Department of Health and Human Services.

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.

Torre, Lindsey A., Freddie Bray, Rebecca L. Siegel, et al. 2015. Global cancer statistics, 2012. *CA: A Cancer Journal for Clinicians* **65**(2) 87–108.

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.

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, et al. 2011. Effectively selecting a target population for a future comparative study. *Harvard University Biostatistics Working Paper Series* .

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

Zhao, Yufan, Michael R. Kosorok, Donglin Zeng. 2009. Reinforcement learning design for cancer clinical trials. *Statist. Med.* **28**(26) 3294–315.